bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Neo-sex chromosomes, genetic diversity and demographic history in the Critically Endangered Raso lark

Elisa Dierickx[1,2], Simon Sin[3,5], Pieter van Veelen[6,7], M. de L. Brooke[1], Yang Liu[4], Scott Edwards[3] and Simon Martin[1,8]

[1] Department of Zoology, University of Cambridge

[2] Fauna & Flora International

[3] Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University

[4] Department of Ecology, Sun Yat-sen University

[5] School of Biological Sciences, University of Hong Kong

[6] Wetsus, European Centre of Excellence for Sustainable Water Technology

[7] Groningen Institute for Evolutionary Life Sciences, University of Groningen

[8] Institute of Evolutionary Biology, University of Edinburgh

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

Small effective population sizes could expose island species to inbreeding and loss of genetic variation. Here we investigate factors shaping genetic diversity in the Raso lark, which has been restricted to a single islet for ~500 years, with a population size of a few hundred. We assembled a reference genome for the related Eurasian skylark and then assessed diversity and demographic history using RAD-seq data (75 samples from Raso larks and two related mainland species). We first identify broad tracts of suppressed recombination in females, indicating enlarged neo-sex chromosomes. It is plausible that these regions might inadvertently and temporarily preserve pre-existing allelic variation in females that would otherwise be lost through genetic drift. We then show that genetic diversity across autosomes in the Raso lark is lower than in its mainland relatives, but inconsistent with long-term persistence at its current population size. Finally, we find that genetic signatures of the recent population contraction are overshadowed by an ancient expansion and persistence of a very large population until the human settlement of Cape Verde. Our findings show how genome-wide approaches to study endangered species can help avoid confounding effects of genome architecture on diversity estimates, and how present day diversity can be shaped by ancient demographic events.

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Island species have suffered 89% of all recorded avian extinctions, despite only representing 20% of all bird species [1–3]. This reflects multiple factors, including vulnerability to alien invasive species [3] and intrinsic geographical characteristics of islands such as isolation and small distributional area, which can increase vulnerability to climate change [4] and limit dispersal into alternative habitat. Species with a small effective population size (N_e) are also subject to three types of genetic risk: inbreeding depression through the exposure of deleterious recessive alleles and loss of heterozygote advantage [5,6]; accumulation of deleterious alleles due to increased drift ("mutational meltdown") [7,8]; and loss of potentially adaptive genetic variation limiting future adaptive potential [9,10]. Previous studies have found that some, but not all, island species show reduced genetic diversity and increased inbreeding compared to their mainland counterparts [11,12]. However, this pattern may be driven more by the bottleneck associated with colonisation rather than long-term reduction in N_e [12]. For most species, poor census data makes it difficult to assess whether island existence *per se* is likely to expose species to the genetic risks above.

The Raso lark *Alauda razae* is endemic to the uninhabited 7km^2-islet of Raso in Cape Verde [13]. Irregular counts since 1965 and yearly counts since 2002 indicate a small population fluctuating from 20 breeding pairs to 1558 individuals [14–16] (Table S1). Sub-fossils indicate a larger past distribution encompassing neighbouring islands: Santa Luzia (35km^2), São Vicente (227km^2) and Santo Antão (779km^2) ([13] Figure 1). Raso larks disappeared abruptly from the neighbouring islands following the arrival of humans along with cats, dogs and rodents in 1462 [13]. Today, Raso is the last of the larger islets of Cape Verde that remains free of these mammals.

A previous study raised the intriguing possibility that a change to genome architecture might prevent the loss of pre-existing allelic variation through genetic drift in Raso larks [17]. The genus *Alauda* has enlarged neo-sex chromosomes which appear to derive from ancestral autosomes [18,19]. Cessation of recombination should ultimately lead to sequence divergence between the neo-Z and neo-W chromosomes, and degeneration of the latter. However, in the short-term, before significant divergence has accumulated, suppressed recombination can effectively cause retention of distinct pre-existing alleles at homologous loci on the neo-W and neo-Z. This would artificially elevate estimates of heterozygosity in females if the analysed loci are treated as diploid autosomes rather than two haploid gametologues. Indeed, Brooke et al. [17] observed microsatellite loci that would ordinarily be autosomal showing sex-linked segregation and excessive heterozygosity in female Raso larks. The authors hypothesised that the enlarged neo-

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sex chromosomes might therefore allow the retention of genetic variability in the face of population contraction. If the neo-sex chromosomes are indeed very large (as suggested by cytogenetic analysis; [19]) and if recombination is suppressed across much of their length, they could theoretically impact not only estimates of genetic diversity in Raso larks, but potentially also their fitness. These ideas have hitherto not been explored with genome-wide data.

We therefore set out to investigate the contrasting effects of population contraction, which should reduce genetic diversity, and that of neo-sex chromsomes, which might preserve allelic variation, in the Raso lark. We produced a high-quality draft genome assembly for the related Eurasian skylark *Alauda arvensis* and used restriction site associated DNA sequencing (RAD-seq) of 75 individuals from four lark species: Raso lark, Eurasian skylark, Oriental skylark *A. gulgula* and the crested lark *Galerida cristata*. Our findings reveal the risk of confounding effects of neo-sex chromosomes on commonly-used measures of diversity, as well as unexpected demographic events, and therefore highlight the value of genome-wide approaches to study diversity and genetic risks in endangered species.

METHODS

All custom scripts and commands for all computational analyses are provided at https://github.com/simonhmartin/Raso_lark_diversity.

Sample collection

Twenty-six Raso lark blood samples were collected on Raso between 2002 and 2014. From colleagues, we also obtained blood and tissue samples for related species: 29 Eurasian skylarks, that we group into three populations, 15 Oriental skylarks from two locations, and 5 crested larks from Saudi Arabia (Figure 1, Table S2). None of the sampled birds was likely to be a migrant based on the sampling dates (Table S2) and/or the migration pattern of the species. For samples with unknown sex, sex was determined using PCR [20] and/or by examination of heterozygosity on the Z-linked scaffolds (see Results).

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

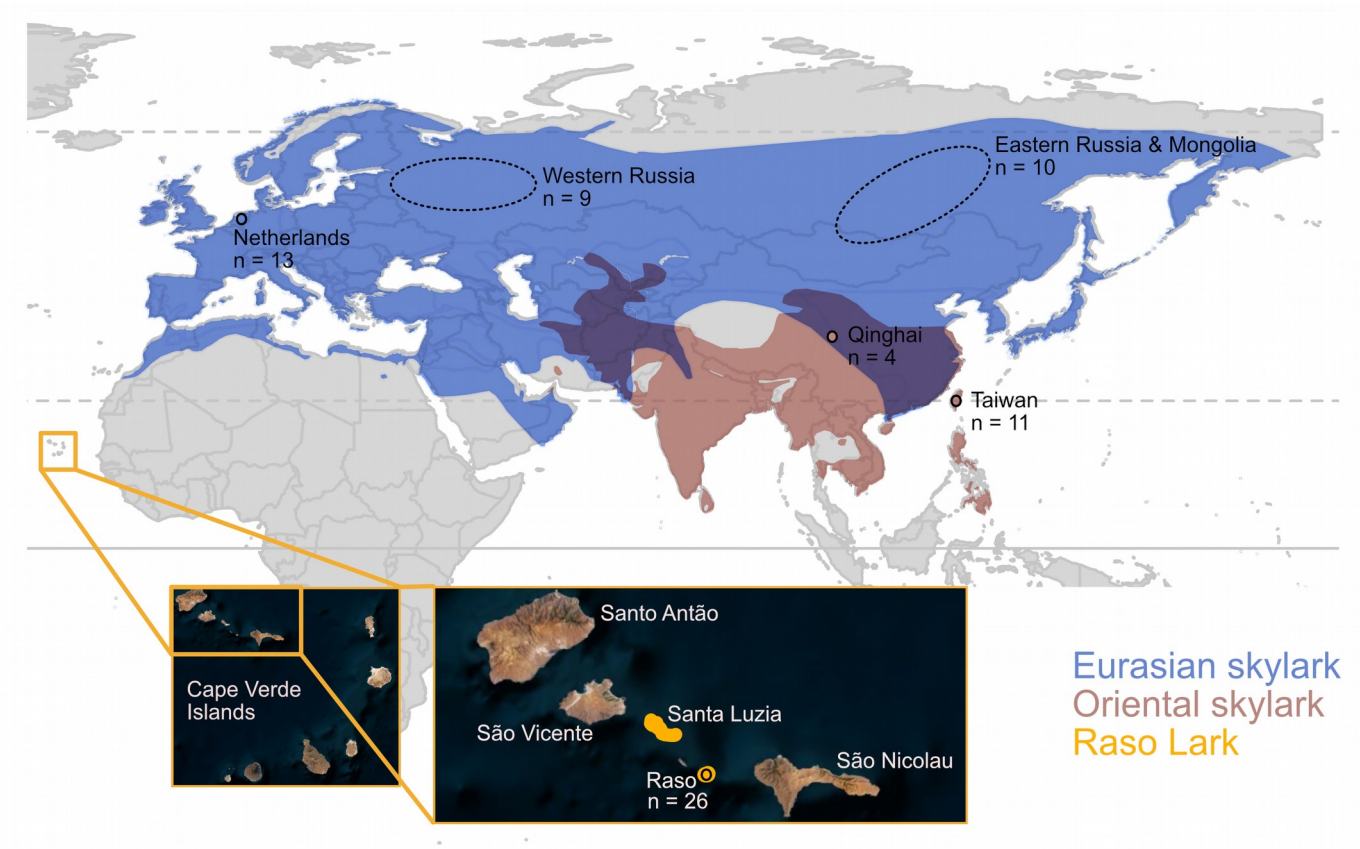

Figure 1. Species ranges, sampling locations and samples sizes. The insert shows the location of Raso in Cape Verde. Sampling areas for the each of the six *Alauda* populations are indicated. See Table S2 for sample details.

Whole genome sequencing and assembly

A draft reference genome was obtained through whole-genome sequencing of a male Eurasian skylark (individual 0) collected in Mongolia (Table S2). Whole genomic DNA was isolated using DNeasy Blood & Tissue Kit (Qiagen, Venlo, The Netherlands) following the manufacturer's protocol. Two libraries were prepared: a 220bp insert size fragment library using a PrepX ILM 32i DNA Library Kit for an Apollo 324 robot, following the manufacturer's protocol (TaKaRa, Kusatsu, Japan), and a 3 kb mate-pair library using an Illumina Nextera Mate Pair Sample Preparation kit and following the manufacturer's protocol (Illumina, San Diego, CA, USA). Both libraries were sequenced on an Illumina HiSeq 2500, producing 125 bp paired-end reads. Adaptor trimming and quality assessment were performed using Trimmomatic 0.32 [21] and FastQC [22], respectively and Allpaths-LG [23] was used for assembly with option '--HAPLOIDIFY=TRUE'. Summary statistics were calculated with Allpaths-LG.

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RAD library preparation

DNA for RAD-seq was extracted as above. Single digest RAD-seq libraries were prepared using the *PstI*. Each individual was assigned an 8-base pair (bp) inline barcode, and equimolar concentrations of 16 uniquely barcoded individuals were pooled and double-indexed by 16 cycles of high-fidelity PCR using Phusion High-Fidelity PCR Master Mix (Thermo Fisher Scientific, Waltham, MA, USA) with Illumina barcodes. The PCR products were pooled in equimolar quantities and sequenced on an Illumina HiSeq 1500, producing 100bp single-end reads.

Sequence processing and alignment

We used *process_radtags* in Stacks 1.35 [24] without quality filters to sort sequence reads by barcode. We then used Trimmomatic to trim restriction sites (6bp) and remove all trimmed reads shorter than 95bp. *Process_radtags* was then used again to filter for quality. Reads were aligned to the Eurasian skylark genome using Bowtie 2 [25], and reads with multiple significant hits were removed.

Pseudo-chromosomal assembly

We inferred the approximate chromosomal location and orientation of each Eurasian skylark scaffold larger than 1 megabase (Mb) based on homology with the zebra finch *Taeniopygia guttata* genome [26] (i.e. a pseudo-chromosomal assembly). We used the *nucmer* tool in MUMmer v3.23 [27] to identify regions of strong homology.Alignments shorter than 5kb were discarded using the *delta-filter* tool. We used *mummerplot* to visualise all alignments and determine the optimal scaffold order and orientation. Additional manual changes were made based on visual inspection of the scaffold arrangement.

Identification of sex-linked regions based on heterozygous genotypes

To identify putative sex-linked regions, we examined how the proportion of heterozygous sites varies across the genome in each individual. We called genotypes using GATK version 3.4 [28,29] HaplotypeCaller (in GVCF mode) and GenotypeGVCFs, with default parameters. Thirteen individuals with poor coverage (< 3 million sites with ≥5x coverage, Table S2) were excluded. The proportion of heterozygous genotype calls for each individual was computed for 250 kb windows across each scaffold considering only sites with >5x coverage using the Python script popgenWindows.py (github.com/simonhmartin/genomics_general). Windows with fewer than 100 sites (with ≥5x coverage) genotyped in the population (both variant and invariant) were excluded.

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Relatedness

We estimated relatedness among individuals using two methods suited to low coverage genomic data. The first was NgsRelate [30], which considers genotype likelihoods. These were computed using GATK as described above. NgsRelate was run considering sites covered by at least one read in each individual in the population. The second approach was KGD [31], which is designed for GBS data such as RAD-seq data, and also accounts for low sequencing depth. The input file was generated using ANGSD [32]. Only sites covered by at least 100 reads across the 26 Raso larks or 50 reads across the 10 Eurasian skylarks from the Netherlands were included. Following [31], we explored filtering options to find SNP subsets that gave realistic values of self-relatedness (~1). The chosen filter was to use only SNPs with a Hardy-Weinberg disequilibrium value between 0 and 0.1.

Allele frequency spectra and genetic diversity

We used a two-step pipeline in ANGSD [32] to infer allele frequency spectra from the RAD-seq reads mapped to the reference genome, accounting for uncertainty in low-coverage sequencing data. Genotype likelihoods were inferred using ANGSD with likelihood method 2, and only sites with a base alignment quality (baq) ≥1 and SNP quality ≥20 were considered. A mapping quality adjustment was applied for reads with multiple mismatches ('-C 50'), following the author's recommendation. The *realSFS* tool was then used to infer the allele frequency spectrum with a maximum of 100 iterations, with 20 bootstraps. Nucleotide diversity (π) was computed from the frequency spectrum as the sum of the weighted products of the major and minor allele counts for each allele count category, including the zero category (invariant sites).

Demographic inference

We applied two different approaches to investigate historical demographic changes in the Raso lark based on the frequency spectrum (averaged across 20 bootstrap replicates). First we used δaδi [33] to compare four different models of increasing complexity. The first model imposes a constant population size. Since δaδi only optimises the shape of the frequency spectrum, this model has no free parameters. The second model adds a single change in population size at some point in the past (two free parameters: time and relative size of the new population). The third and fourth models each added an additional change (along with two free parameters). Model optimisation was performed using grid sizes of 50, 60 and 70, and repeated 10 to 50 times to confirm optimisation.

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In the second approach we used Stairway Plot [34] to infer the optimal population size history given the SFS. We used the "two-epoch" model, with the recommended 67% of sites for training and 200 bootstraps. We tested four different numbers of random breakpoints: 12, 25, 37 and 50.

To convert inferred population sizes and times to numbers of individuals and years, respectively, we used the collared flycatcher *Ficedula albicollis* mutation rate estimate of 4.6×10^{-9} per site per generation [35]. We estimated the generation time of the Raso lark, defined as the mean age of the parents of the current cohort at age at first breeding + (1/mean annual mortality) [36]. This gave a generation time of 6.5 years.

RESULTS

Enlarged neo-sex chromosomes in Alauda larks

We generated a high-quality draft genome assembly for a male Eurasian skylark based on 154,342,128 paired-end read and 263,949,984 mate paired reads. The resulting 5714 scaffolds total 1.06 Gb with a scaffold N50 length of 1.44 Mb (71.5 Kb for contigs) and BUSCO [37] completeness score for the Aves gene set of 93%. We were able to arrange 311 scaffolds of ≥1 Mb in length, totalling 648 Mb (63% of the genome), into pseudo-chromosomes based on homology with the zebrafinch genome (Figure S1).

RAD-seq reads for 75 individuals were mapped to the reference genome, yielding a high density of RAD loci sequenced to low depth (min = 1.5x, max = 6.9x, mean = 2.8x) (Table S2). The average proportion of reads mapped was 88% for Eurasian and Oriental skylarks, 89% for Raso larks and 82% for the outgroup crested larks. Using an arbitrary threshold of at least 100 reads across the dataset to designate a shared RAD locus yields 62 million genotyped sites across the dataset, of which 4 million are SNPs.

To investigate the composition of the neo-sex chromsomes, we examined the density of heterozygous genotypes across the genome in each individual. As females carry a single copy of the Z, they should have no heterozygous sites on this chromosome (with the exception of repetitive elements at which mis-mapping can occur, and possibly pseudo-autosomal regions where the W and Z retain homology). The same should be true for neo-sex chromosomes, unless recombination suppression is recent or incomplete, such that the gametologues retain a high level of homolgy. This would allow RAD-seq reads from the neo-W to map to scaffolds representing the neo-Z (there should be no neo-W scaffolds in our male reference genome). This could potentially lead to an elevation (rather than a reduction) in the density of heterozygous genotype calls in females. Indeed, we observe strong elevations in the density of heterozygous

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genotypes across large portions of chromosomes 3, 4a and 5 in all female Raso larks and not in males (Figure 2A). This is consistent with the formation of neo-sex chromosomes and subsequent recombination suppression involving these three autosomes. The gametologues have evidently not diverged to the point that sequence homology has been eliminated. Indeed they retain around 99% similarity (1.15 differences per 100 bp on average). It is important to note that the chromosome map used here - based on the zebra finch genome - does not reflect the true karyotype for larks. While previous work indicates that only a fragment of chromosome 4a has become sex-linked [18], we cannot currently say whether chromosomes 3 and 5 have undergone similar fragmentation. One scaffold appeared to bridge between regions of suppressed and normal recombination (Figure S1), but the two parts mapped to different chromosomes (3 and 2), so we concluded that this is most likely a misassembly. We also lack any direct evidence that chromosomes 3, 4a and 5 (or fragments thereof) have fused to the Z chromosome. One scaffold appears to bridge chromosomes 3 and 5, but this too may be a misassembly, as it maps to the centre of the two regions of suppressed recombination (Figure S1).

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

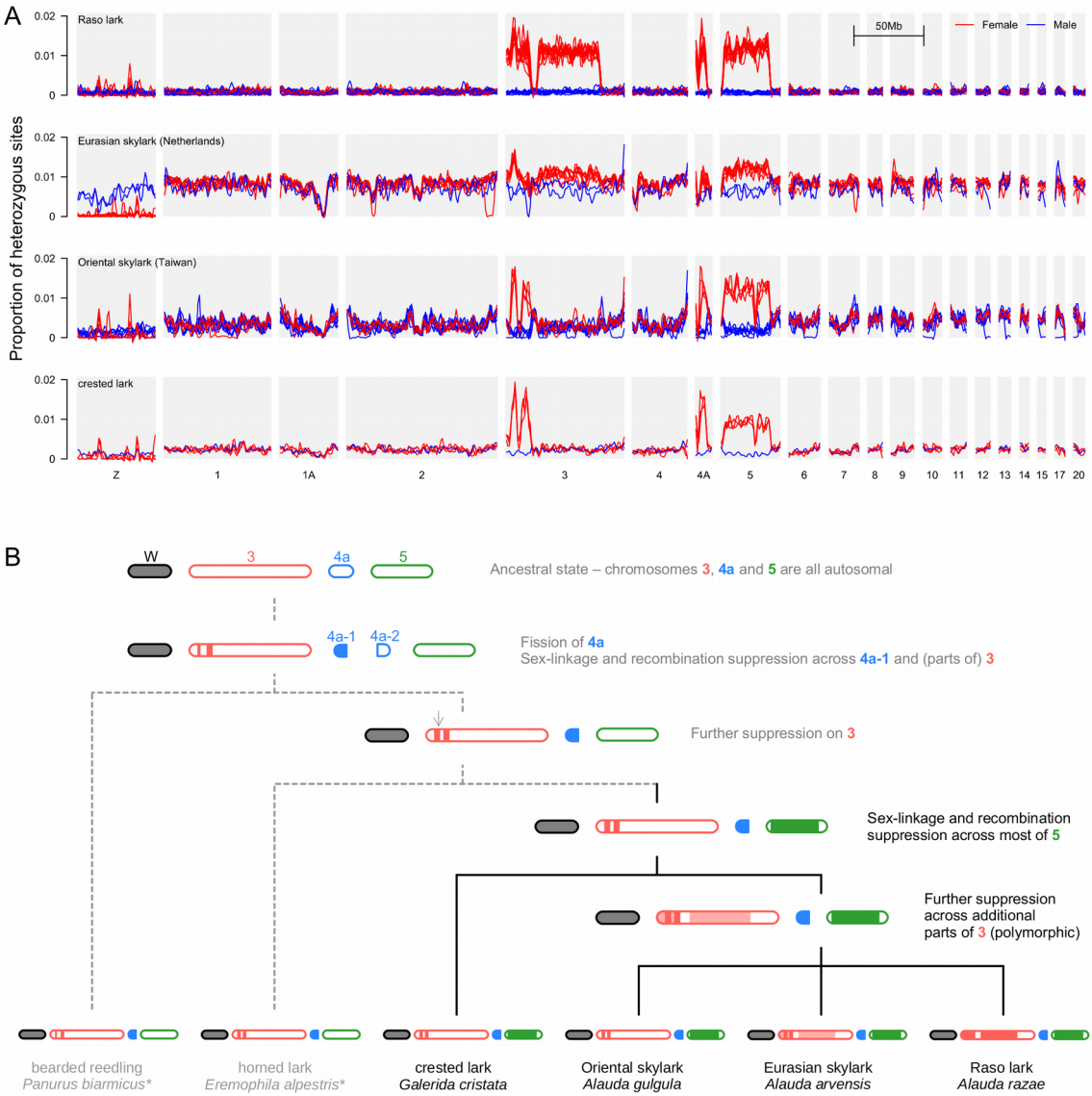

Figure 2. Evidence for largeIndividual heterozygosity reveals neo-sex chromosomes. (**A**) The proportion of heterozygous sites per 100250 kb window in each individual, plotted across each pseudo-chromosome (assembled based on homology with the zebra finch karyotype), with locally-weighted smoothing (loess, span = 10 Mb). Females and males are indicated by red and blue lines, respectively. One population from each species is shown (see Figure S2 for all populations). A lack of heterozygous genotypes in females on the ancestral Z chromosome indicates that it is haploid due to divergence and degeneration of the ancestral W (the narrow peaks most likely reflect collapsed repeats in our assembly). A high density of heterozygous genotypes in females on autosomes is indicative of more-recent recombination suppression between neo-W and neo-Z, without significant degradation of the neo-W (**B**). Model for the progression of recombination suppression between lark neo-sex chromosomes. Shading indicates inferred neo-W regions with suppressed recombination in females based on our analysis (panel A and Figure S2) as well as reference [38] (the latter are shown in grey and indicated with an asterisk). Note that chromosomes are not shown as fused as we do not have evidence for or against physical linkage in this study. The fragmentation of chromosome 4a is based on previous linkage mapping [18]. Light shading indicates polymorphic recombination suppression (i.e. in some populations but not others; Figure S2)

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Although we lack information on how chromosome structure has evolved, comparison with the other species (Figure 2A, S2), gives insights into the progression of recombination suppression through time. An increased density of heterozygous genotypes in females exists across the same three chromosomes in Eurasian skylarks from the Netherlands, though it is not as easily distinguishable from the higher genome-wide background heterozygosity in this population (Figure 2A). Within the regions of suppressed recombination, the density of heterozygous sites is similar to that in Raso lark females (1.09%). Female Oriental skylark and crested lark also show elevated densities of heterozygous sites, but over smaller portions of chromosomes 3 and 5 (Figure 2A). Using our results in combination with those from a recent study of two additional outgroup species, the bearded reedling *Panurus biarmicus* and the horned lark *Eremophila alpestris* [38], we are able to partially reconstruct the stepwise progression of recombination suppression (Figure 2B). We infer that the initial suppression occurred in two narrow 'strata' on chromosomes 3 and 4a-1 [38]. This increased marginally on chr. 3 in the common ancestor of the larks, and was followed by the formation of a large new stratum across most of chr. 5 in the common ancestor of the crested lark and the *Alauda* larks. Further recombination suppression in several parts of chr. 3 then probably occurred in the ancestor of *Alauda*. However, this was likely polymorphic, because it is lacking in the Oriental skylark, and also in the eastern populations of the Eurasian skylark, but it is present in the western populations of the Eurasian skylark as well as in the Raso lark (Figure S2). These species differences cannot easily be explained by artefacts such as allele dropout, as the trends we describe are at the scale of multiple megabases, and are therefore supported by tens to hundreds of separate RAD loci.

Genetic diversity

To study genetic diversity and demography, we used only the genomic scaffolds that show no evidence of suppressed recombination (Figure S3). We first screened for relatedness and found that most of the 26 Raso lark samples show little or no detectable relatedness, but three pairs of individuals have levels of relatedness of ~0.5, indicating either sibling or parent-offspring relationships, while several other pairs have non-zero relatedness of up to ~0.2 (Figure S4). We chose not to exclude these from the downstream diversity analyses as this level of relatedness likely reflects the true composition of the population rather than a sampling artefact. Indeed, one of the pairs of close relatives consists of individuals sampled three years apart, in 2011 and 2014. Sampling dates of the other pairs could not be verified due to a labelling error of some

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DNA samples. The equivalent analyses performed for the 13 Eurasian skylarks from the Netherlands found no consistent evidence of high relatedness (Figure S4).

Average nucleotide diversity (π) across the 26 Raso larks based on inferred allele frequency spectra is 0.001 (Table S3). This is just less than 10% of that in Eurasian skylark from the Netherlands (0.0114), and 25% of that in the Oriental skylark from Taiwan (0.0041). Assuming an equilibrium population with $\Theta = 4N_e\mu$ and a per-generation mutation rate equivalent to that of the collared flycatcher, this translates to an effective population size (N_e) of ~50,000 in the Raso lark, compared to ~500,000 in the Eurasian skylark. Therefore, genetic diversity in the Raso lark, despite being one tenth of that in the Eurasian skylark, appears to reflect a far larger historical effective population size than its current census size of ~1000 individuals.

We also considered how different the estimated genetic diversity of Raso larks would be if we did not account for the enlarged neo-sex chromosomes (as may easily occur using a RAD-seq approach without a reference assembly). Although the regions of suppressed recombination represent 12% of the genome, estimated genetic diversity is nearly doubled (0.0019) in our dataset of 15 females and 11 males when these regions are not excluded (Table 1). Although the neo-Z and neo-W chromosomes are gametologues, making it arguably incorrect to consider them as contributing to heterozygosity in the conventional sense, their retention of fairly similar sequences, potentially reflecting ancestral standing variation that pre-dates recombination suppression, means that they might still contribute to functional allelic variation and therefore fitness in female Raso larks, as hypothesised previously [17].

Genomic signatures of ancient and recent bottlenecks

Although autosomal genetic diversity in Raso larks is lower than that in related species, it remains far greater than would predicted under long-term persistence at its small census size of typically under 1000. We therefore hypothesised that population contraction from a larger ancestral size occurred recently enough that some of the preexisting genetic diversity has been retained. To test this hypothesis, and investigate to what extent a recent population collapse has impacted Raso lark genetic variation, we examined the allele frequency spectrum, which carries information about historical population size changes [39]. Surprisingly, the frequency spectrum (computed using only scaffolds showing no evidence of suppressed recombination; see Figure S3) is skewed towards an excess of rare variants (Figure S5), which is also captured in the negative Tajima's D value of -0.69. An excess of rare variants is typical of population expansion rather than contraction and is therefore not consistent with our

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

understanding of the recent history of the Raso lark. This skew was consistent across 20 bootstrap replicates and 26 'drop-one-out' replicates, in which a single individual was excluded in each case (Figure S5), which demonstrates that it is not a sampling artefact.

We therefore used two related approaches to explore the historical demography of this species based on the frequency spectrum. First, we used δaδi [33] to compare the fit of simple models allowing zero, one, two or three changes in population size in the past. The simplest model imposes a constant population size, and shows a very poor fit to the data, as expected (Figure S6). A model that allows a single change in population size in the past shows a far better fit and a greatly improved composite likelihood (Figure S6). The inferred model involves an ancient population expansion rather than a recent contraction (Figure S6). This one-change model is able to recreate the excess of singleton variants in Raso larks, but still shows a fairly poor fit for to the frequency of other rare variants. The model allowing two changes in population size again shows a major increase in likelihood and better fit to the frequency spectrum (Figure S6). However, again the inferred demographic model does not include a recent contraction. Instead, it consists of an ancient bottleneck followed by an expansion ~110,000 years ago. These findings show that the distribution of genetic variation in the Raso lark can be explained very accurately by a few ancient demographic changes, without inclusion of a recent population contraction.

Given the recent census estimates of ~1000 individuals for the Raso lark, we attempted to fit a final model that allows for a recent population contraction. As δaδi failed to optimise a model with a third change in population size (i.e. two additional free parameters), we instead used a fixed population contraction, and performed a manual search to approximate this parameter estimate by re-running the optimization across a range of final N_e values from 0.1% to 1% of the ancestral size. The resulting best fit model once again shows an improved fit to the data over models without a recent contraction, although it results in minimal detectable difference to the expected frequency spectrum (Figure S6). The model again involves an ancestral bottleneck that ended over 100,000 years ago, expanding to ~3 million individuals thereafter, with a recent contraction down to about 3000 individuals that occurred just 550 years (around 85 generations) ago (Figure 3A). Although parameter estimates are not to be to be interpreted as exact, we note that this date matches nearly perfectly with the arrival of human settlers in Cape Verde in 1462. In summary, these models show that the distribution of genetic variation in the Raso lark is consistent with a recent and dramatic

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

population contraction, but that the overall distribution of genetic diversity in this species is more significantly shaped by expansion after an ancient bottleneck.

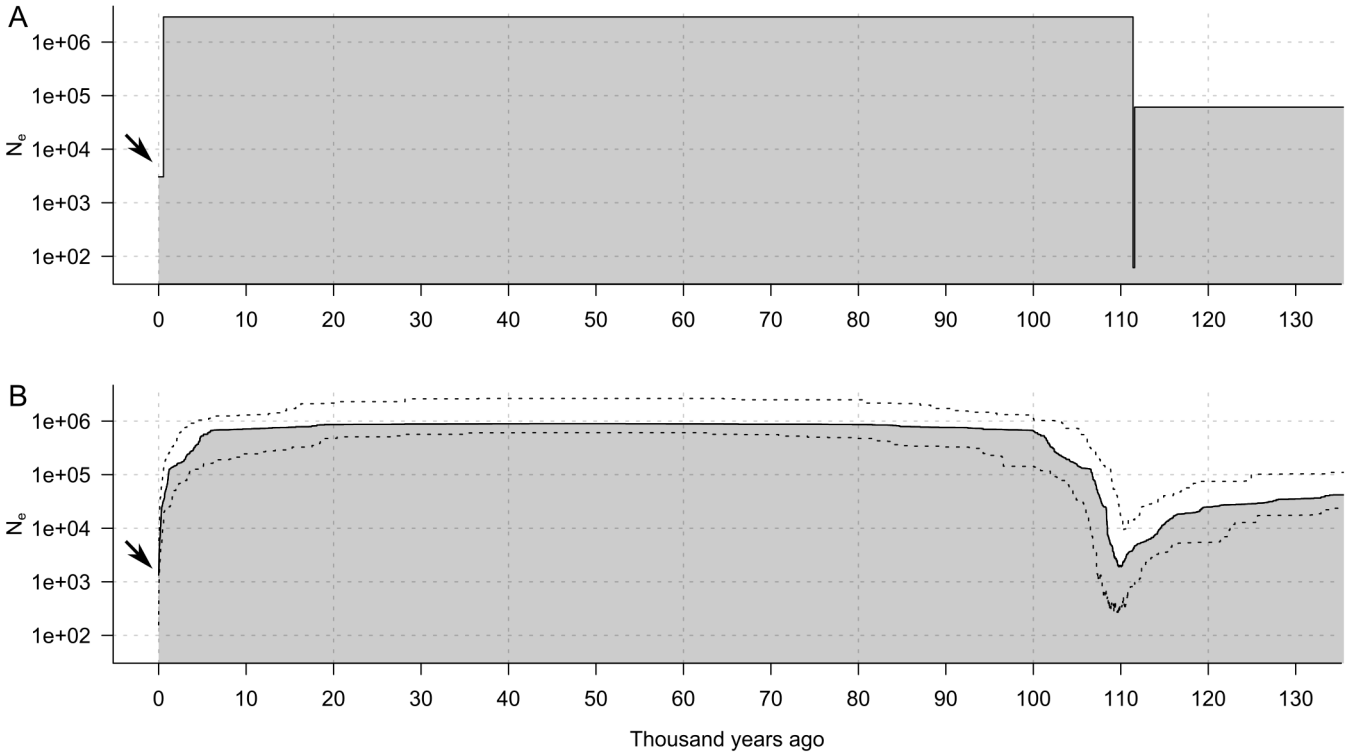

Figure 3. Demographic model fitting supports ancient and recent bottlenecks. The best fit models for inferred N_e (log scale) over time (with the present on the left), based on inference using δaδi [33] (A) and Stairwayplot [34] (B). Dashed lines indicate the 95% confidence interval. Arrows indicate the recent population contraction inferred using both methods.

Our second approach used Stairway Plot [34] to estimate the optimal population size history given the frequency spectrum. The inferred history is remarkably similar in its general structure to the 3-change model inferred using δaδi, with a strong ancestral bottleneck around 110,000 years ago followed by expansion to a large population of nearly a million individuals, with a sharp recent contraction (Figure 3B). The contraction is most pronounced in the most recent past, but it is inferred to have initiated further in the past, ~5,000 years ago. Given the δaδi results showing that any signal of the most recent population contraction in the frequency spectrum is fairly weak, it is likely that the exact timing of this event would be difficult to infer. Nevertheless, both approaches indicate that the Raso lark population was probably very large up until fairly recently, thus agreeing with our second hypothesis that the population still retains much genetic variation that pre-dates its recent contraction.

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DISCUSSION

Genetic markers have long been used to investigate the genetic risks facing species thought to have small N_e, such as island endemics [10–12]. Genomic approaches now allow us to address these questions at far greater resolution, revealing how different parts of the genome have been affected, and inferring past demography. We find that the island-endemic Raso lark has reduced genetic diversity compared to its widespread *Alauda* relatives, but that this difference is much smaller than the difference in census population sizes. Diversity in the Eurasian skylark is high for a passerine, being similar to that in the zebra finch (0.01 [40]) and double that in the housefinch (0.005 [41]). Diversity in the Raso lark is approximately one tenth of that in the skylark , but is similar to that in the Hihi, another island-restricted species (0.00095 [42]). There are an estimated one million breeding pairs of Eurasian skylarks in the United Kingdom alone [43], and usually fewer than 1000 Raso larks on Raso. Our findings indicate that the Raso larks retain some ancestral diversity due to the recency of their population contraction from a much larger size.. In addition, our study demonstrates the importance of understanding genome structure for studies of genetic diversity. The enlarged neo-sex chromosomes of Raso larks might have led to erroneous conclusions had they not been accounted for, but they are also intriguing as a potential source of functional allelic variation.

Previous work suggested that neo-W and neo-Z chromosomes arose at the base of the *Sylvioidea* about 40 million years ago through fusion of part of chromosome 4a to both the W and Z [18]. Our results indicate that two other large chromosomes, 3 and 5, have also become sex-linked in the genera *Alauda* and *Galerida*. Another recent study [38] supports these observations and further indicates that at least chromosome 3 is also sex-linked in two additional outgroup species (horned lark, bearded reedling). Without a chromosomal assembly we cannot currently determine whether homologs of chromosomes 3 and 5 have fused to the W and Z (as is inferred to be the case for chromosome 4a-1; [18]), or whether they segregate as multiple sex chromosomes, as is seen in some other taxa [44,45], but the visually obvious enlargement of both the W and Z chromosomes in larks [19] is consistent with fusions.

Despite the deep age of the neo-sex chromosomes, our results indicate fairly strong homology between the neo-W and neo-Z in the regions of recombination suppression, with around 99% similarity. This value may be somewhat underestimated, as the most divergent parts would be excluded due to poor read mapping [38]. Nevertheless, even a divergence of 2% would translate to ~2.2 million generations to coalescence - much less than the age of the neo-sex chromosomes. Importantly, this coalescence time

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reflects not the age of the fusions but rather the age and extent of recombination suppression. Recombination and gene conversion may continue to occur at low levels even between divergent parts of the gametologues. Moreover, the differences in extent of recombination suppression among species indicates that it has accumulated in a stepwise manner. Indeed, we even find evidence for variation in the extent of recombination among populations of Eurasian skylarks, implying that recombination suppression is recent and not necessarily complete.

The evolutionary causes and consequences of the enlarged neo-sex chromsoomes requires further study. It is hypothesised that sexually antagonistic selection might favour recombination suppression between sex chromosomes [38,46]. Whether this is true in the *Alauda* larks is currently unclear. Nonetheless suppressed recombination might inadvertently result in the maintenance of pre-existing allelic variation between the gametologues that could theoretically contribute to functional diversity and therefore fitness in females [17]. This is analogous to the retention of heterozygosity in functionally asexual hybrid plants [47], except that here it applies only to part of the genome, and only to females. Future work should address whether homologous genes on both gametologues are still expressed and contribute to female fitness.

When we consider only the autosomes, we we find that levels of diversity in Raso larks are higher than would be predicted with long-term persistence at their current population size. Such a mismatch between differences in genetic diversity and census population size is a well-documented phenomenon [48]. Research in this field is typically aimed at explaining lower-than-expected diversity in larger populations, either through a mismatch between census and effective population sizes [49], or through increased action of selection affecting linked sites [50]. Higher-than-expected diversity in smaller populations requires a different explanation, such as gene flow from other populations, increased mutation rate, or recent contraction from a larger ancestral size [12]. In the case of the Raso lark, gene flow is unlikely, given that the other two *Alauda* species are largely restricted to Eurasia. We can also confidently rule out an increase in mutation rate because the level of divergence between the neo-Z and neo-W is similar in all three *Alauda* species. A recent reduction from a larger size is therefore the most likely explanation. Assuming this coincided with the settlement of Cape Verde in 1462, the fairly long average generation time of Raso larks of 6.5 years means that this would correspond to just 85 generations ago. Under a simple model of loss of 1/2N times the ancestral variation through drift each generation, with a current N_e of 1000, this translates to a retention of over 95% of the pre-existing genetic diversity (Figure S7). Even in the more extreme case of N_e = 100, >60% of the diversity is retained.

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

While it may be unsurprising that the recent contraction has failed to eliminate genetic diversity, we were surprised by the skew towards rare variants in the allele frequency spectrum, indicating that the genetic make-up of this population is largely shaped by a major population expansion that occurred deeper in the past (estimated at ~120 thousand years ago). One possibility is that this could coincide with their colonisation of Cape Verde, but it is likely that divergence from the Eurasian skylark occurred much earlier. This question is the focus of an ongoing study. Summing the areas of the surrounding islands, Santa Luzia, São Vicente and Santo Antão [13], we can predict a minimum ancestral range of 1,048km^2, which is ~150 times the area of Raso (although this rough estimate does not consider habitat suitability). Extrapolating from the current population of ≈1000 predicts an ancestral population size of 150,000. Our modelling estimates a much larger N_e of nearly a million prior to the recent collapse. It is possible that Raso larks were more abundant due to higher population density in the past, or that their colonisation of Cape Verde from a larger mainland population occurred more recently.

Despite the considerable genetic diversity in Raso larks relative to their population size, continued existence at this size will inevitably increase their genetic risks. We found three pairs of closely related Raso larks out of 26 sampled. Although we cannot rule out a chance effect, finding related individuals is to be expected given that the population dropped to just 57 individuals in 2004, around two generations ago. With sufficient sample sizes, individual-level relatedness may provide more sensitive detection of recent population collapse than population-level diversity, which can take multiple generations to change appreciably. The remaining genetic diversity in Raso Larks is likely to be of critical importance for their future adaptive potential [10]. This is especially relevant given the vulnerability of Cape Verde to environmental changes such as climate change [51] and the ongoing reintroduction of Raso larks to the nearby island of Santa Luzia, initiated in April 2018.

ACKNOWLEDGEMENTS

Nick Horrocks and Per Alström shared crested lark and Eurasian skylark samples. Allison Shultz provided feedback over the course of this project. Marco van der Velde and Sarah Barker advised on genetic sexing and on RAD library preparation. Chris Jiggins provided use of his laboratory and computational resources. Paul Sunnucks and Roberta Bergero provided helpful comments. Thanks to the field assistants on Raso: Mark Bolton, Ewan Campbell, Simon Davies, Mike Finnie, Tom Flower, Lee Gregory, Sabine Hille, Mark Mainwaring, Jason Moss and Justin Welbergen.

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

1. Johnson TH, Stattersfield AJ. 1990 A global review of island endemic birds. *Ibis (Lond. 1859).* **132**, 167–180. (doi:10.1111/j.1474-919X.1990.tb01036.x)

2. Manne LL, Brooks TM, Pimm SL. 1999 Relative risk of extinction of passerine birds on continents and islands. *Nature* **399**, 258–261. (doi:10.1038/20436)

3. Butchart SHM, Stattersfield AJ, Brooks TM. 2006 Going or gone: defining 'Possibly Extinct' species to give a truer picture of recent extinctions. *Bull. Br. Ornithol. Club* **126**, 7–24.

4. Ferreira MT, Cardoso P, Borges PAV, Gabriel R, de Azevedo EB, Reis F, Araújo MB, Elias RB. 2016 Effects of climate change on the distribution of indigenous species in oceanic islands (Azores). *Clim. Change* **138**, 603–615. (doi:10.1007/s10584-016-1754-6)

5. O'Grady JJ, Brook BW, Reed DH, Ballou JD, Tonkyn DW, Frankham R. 2006 Realistic levels of inbreeding depression strongly affect extinction risk in wild populations. *Biol. Conserv.* **133**, 42–51. (doi:10.1016/j.biocon.2006.05.016)

6. Charlesworth D, Willis JH. 2009 The genetics of inbreeding depression. *Nat. Rev. Genet.* **10**, 783–796. (doi:10.1038/nrg2664)

7. Lynch M, Conery J, Burger R. 2006 Mutational Meltdowns in Sexual Populations. *Evolution* **49**, 1067–1080. (doi:10.2307/2410432)

8. Reed DH. 2005 Relationship between population size and fitness. *Conserv. Biol.* **19**, 563–568. (doi:10.1111/j.1523-1739.2005.00444.x)

9. Frankham R. 2008 Inbreeding and Extinction: Island Populations. *Conserv. Biol.* **12**, 665–675. (doi:10.1111/j.1523-1739.1998.96456.x)

10. de Villemereuil P, Rutschmann A, Lee KD, Ewen JG, Brekke P, Santure AW. 2019 Little Adaptive Potential in a Threatened Passerine Bird. *Curr. Biol.* **29**, 889-894.e3. (doi:10.1016/j.cub.2019.01.072)

11. Frankham R. 1997 Do island populations have less genetic variation than mainland populations? *Heredity* **78**, 311–327. (doi:10.1038/hdy.1997.46)

12. James JE, Lanfear R, Eyre-walker A. 2016 Molecular Evolutionary Consequences of Island Colonization. *Genome Biol. Evol.* **8**, 1876–1888. (doi:10.1093/gbe/evw120)

13. Mateo JA, Jurado LFL, Geniez P. 2009 Historical distribution of Razo Lark Alauda razae in the cape Verde archipelago. *Alauda* **77**, 309–312.

14. Ratcliffe N, Monteiro LR, Hazevoet CJ. 2010 Status of Raso Lark Alauda razae with notes on threats and foraging behaviour. *Bird Conserv. Int.* **9**, 43–46. (doi:10.1017/s0959270900003336)

15. Donald PF, Brooke M de L. 2006 An unlikely survivor: The peculiar natural history of the Raso Lark. *Br. Birds* **99**, 420–430.

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

16. Brooke MDL. 2018 Extreme rainfall-related clutch size variation in the Critically Endangered Raso Lark Alauda razae of the arid Cape Verde islands. *Bird Conserv. Int.* (doi:10.1017/S0959270918000163)

17. Brooke MDL, Welbergen JA, Mainwaring MC, Van Der Velde M, Harts AMF, Komdeur J, Amos W. 2010 Widespread translocation from autosomes to sex chromosomes preserves genetic variability in an endangered lark. *J. Mol. Evol.* **70**, 242–246. (doi:10.1007/s00239-010-9333-3)

18. Pala I, Naurin S, Stervander M, Hasselquist D, Bensch S, Hansson B. 2012 Evidence of a neo-sex chromosome in birds. *Heredity* **108**, 264–272. (doi:10.1038/hdy.2011.70)

19. Bultova NS. 1973 Unusually Large Sex Chromosomes in Some Larks (Aves: Alaudidae). *Mamm. Chromosom. Newsl.* **14**, 150–151.

20. Griffiths R, Double MC, Orr K, Dawson RJG. 1998 A DNA test to sex most birds. *Mol. Ecol.* **7**, 1071–1075. (doi:10.1046/j.1365-294x.1998.00389.x)

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

22. Andrews S, Babraham Bioinformatics. 2010 FastQC: A quality control tool for high throughput sequence data. *Manual*. (doi:citeulike-article-id:11583827)

23. Gnerre S *et al.* 2011 High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci.* **108**, 1513–1518. (doi:10.1073/pnas.1017351108)

24. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013 Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140. (doi:10.1111/mec.12354)

25. Langmead B, Salzberg SL. 2012 Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359. (doi:10.1038/nmeth.1923)

26. Warren WC *et al.* 2010 The genome of a songbird. *Nature* **464**, 757–762. (doi:10.1038/nature08819)

27. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004 Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12. (doi:10.1186/gb-2004-5-2-r12)

28. DePristo M a *et al.* 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–8. (doi:10.1038/ng.806)

29. Van der Auwera GA *et al.* 2013 From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-11.10.33. (doi:10.1002/0471250953.bi1110s43)

30. Korneliussen TS, Moltke I. 2015 NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* **31**, btv509. (doi:10.1093/bioinformatics/btv509)

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

31. Dodds KG, McEwan JC, Brauning R, Anderson RM, Stijn TC, Kristjánsson T, Clarke SM. 2015 Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics* **16**, 1–15. (doi:10.1186/s12864-015-2252-3)

32. Korneliussen TS, Albrechtsen A, Nielsen R. 2014 ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* (doi:10.1186/s12859-014-0356-4)

33. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695. (doi:10.1371/journal.pgen.1000695)

34. Liu X, Fu YX. 2015 Exploring population size changes using SNP frequency spectra. *Nat. Genet.* **47**, 555–559. (doi:10.1038/ng.3254)

35. Smeds L, Qvarnström A, Ellegren H. 2016 Direct estimate of the rate of germline mutation in a bird. *Genome Res.* **26**, 1211–1218. (doi:10.1101/gr.204669.116)

36. Dierickx EG, Robinson R, Brooke MDL. In press. Survival of a Long-Lived Single Island Endemic, the Raso lark Alauda razae, in Relation to Fluctuating Population and Rainfall. *Rev.*

37. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212. (doi:10.1093/bioinformatics/btv351)

38. Sigeman H, Ponnikas S, Chauhan P, Dierickx E, Brooke M de L, Hansson B. 2019 Genomics of expanded avian sex chromosomes shows that certain chromosomes are predisposed towards sex-linkage in vertebrates. *bioRxiv* , 585059. (doi:10.1101/585059)

39. Polanski A, Kimmel M. 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165**, 427–436.

40. Balakrishnan CN, Edwards S V. 2009 Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (Taeniopygia guttata). *Genetics* **181**, 645–660. (doi:10.1534/genetics.108.094250)

41. Shultz AJ, Baker AJ, Hill GE, Nolan PM, Edwards S V. 2016 SNPs across time and space: population genomic signatures of founder events and epizootics in the House Finch (Haemorhous mexicanus). *Ecol. Evol.* **6**, 7475–7489. (doi:10.1002/ece3.2444)

42. de Villemereuil P, Rutschmann A, Lee KD, Ewen JG, Brekke P, Santure AW. 2019 Little Adaptive Potential in a Threatened Passerine Bird. *Curr. Biol.* **29**, 889–894. (doi:10.1016/j.cub.2019.01.072)

43. Donald PF. 2004 *The Skylark*. London: T & AD Poyser.

44. Sahara K, Yoshido A, Traut W. 2012 Sex chromosome evolution in moths and butterflies. *Chromosom. Res.* **20**, 83–94. (doi:10.1007/s10577-011-9262-z)

bioRxiv preprint doi: https://doi.org/10.1101/617563; this version posted December 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

45. Rens W *et al.* 2007 The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. *Genome Biol.* **8**, 1–21. (doi:10.1186/gb-2007-8-11-r243)

46. Wright AE, Dean R, Zimmer F, Mank JE. 2016 How to make a sex chromosome. *Nat. Commun.* **7**, 12087. (doi:10.1038/ncomms12087)

47. Hollister JD, Greiner S, Johnson MTJ, Wright SI. 2019 Hybridization and a loss of sex shape genome-wide diversity and the origin of species in the evening primroses (*Oenothera* , Onagraceae). *New Phytol.* **224**, 1372–1380. (doi:10.1111/nph.16053)

48. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M. 2012 Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10**, e1001388. (doi:10.1371/journal.pbio.1001388)

49. Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC. 1987 Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annu. Rev. Ecol. Syst.* **18**, 489–522. (doi:10.1146/annurev.es.18.110187.002421)

50. Corbett-Detig RB, Hartl DL, Sackton TB. 2015 Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLOS Biol.* **13**, e1002112. (doi:10.1371/journal.pbio.1002112)

51. 'Ministry of Environment Housing and Territory Planning of Cape Verde'. 2011 Climate Change Vulnerability Assessment of Cape Verde: Summary for Policy Makers.