Population genomics of the critically endangered spoon-billed sandpiper

Mateusz Konczal[1], Luis Zapata[2,3,4], Francisco Camara[3,4], Anna Vlasova[3,4], Carla Bello[3,4], Romain Derelle[5], Maria N. Tutukina[6], Maria Plyuscheva[3,4], Claudia Fontsere[4], Pavel S. Tomkovich[7], Nikolay N. Yakushev[8], Ivan A. Shepelev[8], Vladimir Yu. Arkhipov[9,10], Christoph Zöckler[11,12], Roland Digby[13], Egor Y. Loktionov[14], Elena G. Lappo[15], Stephan Ossowski[3,4,16], Tomas Marques[4], Roderic Guigo[3,4], Evgeny E. Syroechkovskiy[17], Fyodor A. Kondrashov[18]

[1]*Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland.*

[2]*Department of Genetics, Evolution and Environment, University College London, London, UK.*

[3]*Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG) 88 Dr. Aiguader, 08003 Barcelona, Spain.*

[4]*Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.*

[5]*Unite´ d'Ecologie, Syste´matique et Evolution, Centre National de la Recherche Scientifique (CNRS), Universite´ Paris-Sud/Paris-Saclay, AgroParisTech, Orsay, France.*

[6]*Department of Functional Genomics and Cellular Stress, Institute of Cell Biophysics of Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation.*

[7]*Zoological Museum, M.V. Lomonosov Moscow State University Zoological Museum, 6 Bolshaya Nikitskaya ulStr. 6, 123009, Moscow, 123009, Russian Federation.*

[8]*ROSIP, 45-105 Chernyshevskogo, st., Saratov 410017, Russian Federation.*

[9]*Institute of Theoretical & Experimental Biophysics, Pushchino, Moscow Oblast, 142290 Russian Federation.*

[10]*State Nature Reserve Rdeysky, Kholm, Novgorod Oblast, 175270 Russian Federation.*

[11]*Spoon-billed Sandpiper Task Force, ArcCona Consulting, 30 Eachard Rd. Cambridge CB3 0HY, United Kingdom.*

[12]*Goebenstrasse 1, 28209 Bremen, Germany.*

[13]*Wildfowl & Wetlands Trust, Bowditch, Slimbridge, Gloucestershire. GL11 5NP United Kingdom.*

[14]*State Lab for Photon Energetics, Bauman Moscow State Technical University, 5/1 Second Baumanskaya St., 105005, Moscow, Russian Federation.*

[15]*Institute of Geography, Russian Academy of Sciences, 29 Staromonetny per., 119017, Moscow, Russian Federation.*

[16]*Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany.*

[17]*All-Russian Institute for Nature Conservation of Ministry of Natural Resources and Ecology of Russian Federation.*

[18]*IST Austria (Institute of Science and Technology Austria), Am Campus 1, 3400, Klosterneuburg, Austria.*

The nature of genetic changes contributing to species extinction and impeding population recovery remains poorly understood. We study the population genetic history of the critically endangered spoon-billed sandpiper and its sister species, the red-necked stint, which is of least concern. We found that while the red-necked stint population was relatively constant across 500,000 years, the spoon-billed sandpiper population peaked 15,000-25,000 years ago during the last glacial maximum, when suitable breeding habitat was likely abundant, and has been declining since. The increase of the population prior to the ongoing decline led to accumulation of recessive deleterious polymorphisms, imposing an additional burden on the spoon-billed sandpiper population. Thus, complex demographic changes leading to gain of deleterious genetic diversity pose an additional risk to species survival and recovery by increasing the cost of inbreeding. Specifically, species that had greater habitat availability during the last glacial maximum may be especially prone to this effect.

Wildlife is experiencing a global decline in population size[1,2] possibly on a way to a massive extinction event[3]. Birds are not an exception, especially migratory birds[4] including populations of the East Asian Australia Flyway (EAAF). Birds along the EAAF undertake exceptionally long migration routes from the Arctic in Eastern Asia and Alaska to Australia and New Zealand, with up to 90% of species with known population trends declining in number[5,6]. A case in point is the critically endangered spoon-billed sandpiper (*Calidris pygmaea*), which breeds in the coastal sub-Arctic regions of Chukotka and Kamchatka, Russia and migrates along the EAAF to South East Asia **(Fig. 1a**). The spoon-billed sandpiper is one of the world's rarest species with an estimated 100-200 breeding pairs in the wild[7] under an eminent threat of extinction due to continuing population decline[8,9]. Crucial factors contributing to the rapid decline of the spoon-billed sandpiper population include habitat loss along the migratory route and human predation[10], which are common threat factors for many waders along the EAAF[5,11]. To investigate the genetic component of population decline we performed a largescale population genomic comparison of this flagship migratory species to its well-faring sister species, the red-necked stint (*C. ruficollis*).

Genetic characterization of fixed changes in the genome of a critically endangered species has the potential to reveal factors that defined its recent evolution and phenotype. Additionally, the study of genetic variation within the population can provide insight on the population history[12]

and potential genetic risks, such as loss of heterozygosity and the accumulation of deleterious mutations[13-23]. The spoon-billed sandpiper is a good model species to address both questions, because its sister species, the red-necked stint, is abundant in the wild[24], providing a crucial reference for a comparative analysis of its polymorphic and fixed substitutions. Previously published genomic analyses contributed to our understanding of population genetics of endangered species. However, the studies published to date focused either on extinct species[25], or those in which the natural variation was artificially affected by hybridization, captive breeding or reintroduction programs[26,27]. The analysis of the spoon-billed sandpiper population on a genomic scale provides an opportunity to study the population genomics of a rapidly declining population not influenced by ongoing conservation efforts, serving as a model for migratory birds of EAAF and other endangered species.

Genome sequence and annotation

We selected samples collected in several locations of Chukotka from female individuals, including 10 spoon-billed sandpipers, nine red-necked stints and a single individual of the red knot (*C. canutus*), long-toed stint (*C. subminuta*) and the little stint (*C. minuta*) (**Supplementary Table 1**). We sequenced DNA extracted from one of the spoon-billed sandpipers in pair-end and mate-pair libraries creating a *de novo* 35x genome assembly with scaffold N50 = 2.8 Mb with an annotation effort and mapping of RNA-seq data (see **Supplementary Methods**) resulting in 21,145 protein-coding and 5,425 lncRNA gene predictions (**Supplementary Table 2**). 93% of the genes in the core eukaryotic genes set were found in the assembled genome (**Supplementary Table 3**).

To allow for a comparative pairwise sequentially Markovian coalescent (PSMC) analysis[28] we sequenced one red-necked stint and the red knot to 30x coverage with the remaining 9 individuals of the spoon-billed sandpiper and 8 red-necked stints were sequenced to 15x and 10x coverage, respectively. The long-toed stint and little stint were sequenced to 5x coverage. The sequences from these individuals were mapped to the reference spoon-billed sandpiper genome calling the single nucleotide polymorphisms (SNPs). Due to the small population size of the spoon-billed sandpiper, we calculated the kinship coefficient between individuals of the same species identifying two closely related pairs. No close relatives were found in the red-necked stint individuals. The final dataset included the comparison of SNPs from 7 spoon-billed sandpiper and 9 red-necked stint genomes, with ~5.2 and ~9.2 million SNPs, respectively (**Table 1**).

Population structure and demographic history

Using the sequence data, we studied the evolution and variation of the spoon-billed sandpiper genome. There were no genetic differences between spoon-billed sandpiper and red-necked stint individuals from Meinopylgino and the Belyaka Spit, two locations 650 km apart at the core areas of the modern spoon-billed sandpiper breeding range (**Fig. 1a**), suggesting a lack of population structure ($\bar{F}_{ST}$=0.002, **Fig. 1b,c, Extended Data Fig. 1a**). We then reconstructed the phylogeny of the sequenced *Calidris* species, confirming that the red-necked stint is the sister species to the spoon-billed sandpiper (**Fig. 1d**). Furthermore, no evidence of genetic admixture

between the spoon-billed sandpiper and the red-necked stint was found (**Extended Data Fig. 2**), consistent with only two records of possible hybrids over several decades of observation[29]. The lack of population structure and absence of genetic admixture with the closest related species simplifies further analysis allowing for direct comparison of their species-specific variability.

To compare population size dynamics of these species we performed PMSC[28] and SMC++ [ref. 30] analyses using the three high-coverage genomes of spoon-billed sandpiper, red-necked stint and the red knot. These methods rely on reconstructed coalescence times of the haplotype blocks in the genome to predict changes in population size. The size of the red-necked stint population was inferred to have been relatively constant throughout the last 500,000 years (**Fig. 1e, Extended Data Fig. 3**). By contrast, the smaller spoon-billed sandpiper population peaked ~15-25 thousand years ago (tya), roughly corresponding to the last glacial maximum[31], transitioning into a continuous decline (**Fig. 1f, Extended Data Fig. 3**). The red knot also experienced a sharp decline in population size[32], although the recent establishment of red knot subpopulations[33] can also lead to a rapid change of haplotype structure in the population that may be interpreted by PMSC as a population decline (**Fig. 1g**). The latter scenario, however, is not applicable to the spoon-billed sandpiper population because it lacks population structure (**Fig. 1c**).

Breeding habitat availability during the last glacial maximum

The consistent and gradual decline of the spoon-billed sandpiper population since the last glacial age suggests a concomitant decline in appropriate habitat. We considered possible distribution of the spoon-billed sandpiper breeding range at the time of the reconstructed population peak 19-25 tya[31]. At the time of maximum glaciation Eurasia and North America were connected by the land bridge Beringia, due to the regression of sea levels, up to 100-130 meters lower than at present[34]. The regression exposed vast areas of the flat continental shelf in what is now the northern Bering Sea, which in combination with a colder climate, provides two major factors contributing to a more widespread spoon-billed sandpiper breeding habitat during the last glacial maximum. First, the exposed continental shelf created large flat areas of rugged coastal line[35], including lagoons with crowberry spits, salt marshes and tundra vegetation surrounded by shallow sea[36] likely resulting in large paleo-Anadyr and paleo-Yukon–Kuskokwim deltas. Second, the colder climate at the time of the last glacial maximum allowed for the presence of acceptable breeding habitat over a wider area extending far beyond the modern breeding range (**Fig. 1h**). The large territory of coastal tundra and deltas of paleo-rivers correspond to breeding habitats of the spoon-billed sandpiper[24], which suggests that the peak of the spoon-billed sandpiper population may have occurred at that time due to wider availability of breeding habitat. The current critical state of the spoon-billed sandpiper population may thus be caused by it being especially sensitive to anthropogenic pressures on the flyway[37] because of a pre-existing vulnerability associated with deglaciation-driven breeding habitat loss.

Genetic variability in the spoon-billed sandpiper population

Nucleotide diversity, π, in the spoon-billed sandpiper population was 0.0015, less than twofold smaller than in the red-necked stint population at $\pi = 0.0022$ (**Fig. 2a**). Thus, the observed decline of the spoon-billed sandpiper population did not have a strong impact on the overall level

of nucleotide diversity, which depends primarily on common polymorphisms. Indeed, population decline is expected to have eliminated rare variants from the population[13,14,17]. We calculated Tajima's D, a statistic that measures the relative contribution of polymorphisms of different frequencies to overall population variation[38], for several functional classes of sites in the two species. A negative value of Tajima's D can be caused by recent population growth or by negative selection acting against mutations, while a positive Tajima's D indicates the action of balancing selection or population growth. The size of the red-necked stint population has been approximately constant over the last half a million years (**Fig. 1e**), thus, Tajima's D in that species would be mostly influenced by selection and not changes of population size. In the red-necked stint Tajima's D was negative for polymorphisms in coding regions and close to zero for polymorphisms in intergenic regions (**Fig. 2b**). These data indicate selective constraint in coding regions, while intergenic regions were mostly neutral, congruent with observations from other species[39]. Tajima's D in the spoon-billed sandpiper population was higher across all functional classes of polymorphisms (**Fig. 2b**), with intergenic polymorphisms showing a substantially positive Tajima's D. The positive Tajima's D and the skew towards more common polymorphisms for all classes of sites in the spoon-billed sandpiper compared to the red-necked stint (**Extended Data Fig. 4**) are consistent with the inferred population histories of the two species.

Conservatively assuming that intergenic polymorphisms are mostly neutral[39] (**Fig. 2b**), we compared their prevalence to the prevalence of polymorphisms at other functional sites between the two species. A population bottleneck is expected to remove rare polymorphisms without a strong influence on common ones[13,14,17]. Therefore, the spoon-billed sandpiper population was expected to have a lower ratio of rare deleterious to common neutral polymorphism. Surprisingly, we found a proportional increase of upstream, lncRNA and nonsynonymous polymorphisms in the spoon-billed sandpiper population (**Table 2**), which are expected to be under negative selection and to have low frequency. Synonymous and intron sites had fewer polymorphisms than intergenic sites in the spoon-billed sandpiper genome, while the proportion of nonsense and downstream polymorphisms was not significantly different (**Table 2**). The relative differences in the number of polymorphisms was greater for polymorphisms with lower frequency (**Extended Data Fig. 5**).

These data seem counter intuitive because nonsynonymous polymorphisms are more likely to be deleterious than other types and, therefore, we expected the spoon-billed sandpiper to have proportionally fewer such polymorphisms due to the ongoing population decline[13,14,17]. The population genetics of genetic bottlenecks[13-15,17,19] and subsequent recovery[14], including the impact of genetic dominance[40,41] is well understood. By contrast, the impact of population growth prior to a bottleneck on segregating polymorphisms has not been considered (but see [ref. 42,43]). To understand the causes behind the observed excess of deleterious polymorphisms we modelled the change of polymorphisms density with varying selection coefficients under the demographic changes observed in the spoon-billed sandpiper population. We found that spoon-billed sandpiper-like demographic changes had no effect on the relative frequency of semidominant deleterious polymorphisms (**Extended Data Fig. 6a**). By contrast, recessive deleterious polymorphisms showed a marked increase in frequency relative to neutral variants,

increasing in prevalence at the onset of population growth and remaining at higher than pre-demographic change levels even after the population decline (**Fig. 3a**).

In a small population, random genetic drift can either eliminate rare deleterious recessive alleles or, occasionally, increase their frequency[44]. In the latter case these alleles become visible to selection, which reduces their frequency. Thus, in a small population genetic drift reduces the frequency of deleterious recessive alleles[40]. In an increasing population genetic drift slows down[45,46], which allows for a relatively faster accumulation of recessive deleterious than of neutral alleles (**Fig. 3a,b**). The accumulated deleterious recessive alleles begin to be lost once the population starts to decline, remaining overabundant for some time (**Fig. 3a**). Consistent with this explanation, our modeling shows that a population decline without a preceding increase does not lead to an increase of deleterious recessive alleles (**Extended Data Fig. 6b**) and the effect is substantially weaker in a larger population with proportional demographic changes (**Extended Data Fig. 6c**). Nearly neutral alleles, on the border of being visible by selection with $s \sim 1/2N$ [ref. 46] decrease in frequency during population growth (**Fig. 3a**) because after the increase of population size (N -> N') they become visible to selection (s > 1/2N'). The decline of synonymous polymorphisms relative to intergenic polymorphisms in the spoon-billed sandpiper population indicate that the selection coefficient associated with synonymous polymorphisms is ~0.0001 (**Fig. 3a**), consistent with previous estimates[47].

Adaptive footprints in the spoon-billed sandpiper genome

The McDonald-Kreitman test has not revealed a strong signature of positive selection in either the spoon-billed sandpiper or the red-necked stint population (see **Methods**). Thus, we searched for specific regions in the genomes of the two species which show substantial genetic differentiation, possibly indicating the action of species-specific positive selection. We calculated Z-scored $\bar{F}_{ST}$ values in 25kbp windows, identifying regions that had many differences between the spoon-billed sandpiper genome relative to the orthologous red-necked stint regions, and *vice versa*. We identified an excess of genomic regions with a high Z-scored $\bar{F}_{ST}$ values relative to a normal distribution (**Extended Data Fig. 1b**) with 761 windows (251 regions) showing a Z-scored $\bar{F}_{ST}$ value >2.5 (corresponding to a p-value < 0.006), of which 129 had the excess of substitutions in the spoon-billed sandpiper lineage (620 in the red-necked stint). The 128 protein-coding genes found in these 114 spoon-billed sandpiper-accelerated regions were significantly enriched for various biological processes (**Supplementary Table 4**), mainly those related to positive regulation of cell growth (**Extended Data Fig. 1c**). Interestingly, two genes, secreted frizzled-related protein and E3 ubiquitin ligase SMURF1/2, are involved in negative regulation of BMP signalling pathway, the primary pathway involved in beak development[48] and is primarily responsible for the difference in the bill shape between chicken and duck[49]. These changes may reflect the molecular basis of the distinctive bill shape of the spoon-billed sandpiper that evolved since its divergence from the red-necked stint. By contrast, the regions with increased accumulation of substitutions in the red-necked stint did not include genes related to beak development or to positive regulation of cell growth (**Supplementary Table 5; Extended Data Fig. 1d**).

Conclusion

Our data indicate that the spoon-billed sandpiper population is harbouring ~2500 extra nonsynonymous recessive deleterious variants (**Table 2**) due an increase in the population size increase prior to the ongoing decline. Given that only polymorphisms with a selection coefficient of > 0.001 were predicted to be influenced by this demographic history (**Fig. 3a**), we can calculate the lower bound of the genome contamination by extra recessive polymorphisms as 2500*0.001=2.5 expressed lethal equivalents[50], which is a substantial portion of the ~10 expressed lethal equivalents that is expected in a vertebrate genome[51]. Recessive alleles do not contribute to fitness in heterozygous state and their immediate effect may be small. However, inbreeding depression is mostly dependent on the impact of recessive deleterious alleles, thus, the tiny surviving spoon-billed sandpiper population is under an elevated risk of extinction if inbreeding becomes more frequent. We estimated that 0.008%-0.396% of genome showed low level of heterozygosity in spoon-billed sandpiper individuals (**Supplementary Table 6**) and its the longest genome scaffold lacked long runs of homozygosity (**Supplementary Table 6**). Both observations suggest a currently panmictic spoon-billed sandpiper population. Nevertheless, proportionally to intergenic polymorphisms, an average spoon-billed sandpiper genome has 14%, or ~500 homozygous nonsynonymous polymorphisms more than an average red-necked stint genome (**Table 3**). Therefore, if the recessive deleterious polymorphisms that accumulated during the increase of the population act multiplicatively they decrease fitness by ~50% (500 * 0.001) and may impede population recovery.

Loss of heterozygosity is thought to be among the main risk factors of extinction. We identify an opposite risk to the spoon-billed sandpiper, with its population harbouring elevated levels of deleterious variation. Initially small populations that experienced a population increase before a population crash appear to be at particular risk due to overabundance of segregating deleterious alleles. Furthermore, a population recovery, if followed by another population decline, may exacerbate the problem of excess deleterious recessive alleles, perhaps leading to a different type of genetic extinction vortex[52] based on a continuous cycle of population increases and decreases. Other species adapted to Polar conditions that, like the spoon-billed sandpiper, may have had greater availability of habitat during the last glacial maximum may be particularly vulnerable to inbreeding depression due to recessive deleterious polymorphisms.

Methods

Samples We selected samples collected between 2002 and 2013 at three locations in Chukotka (**Supplementary Table 1**). One sample of spoon-billed sandpiper was selected to provide the reference genome, other samples to provide information about genomic variation in the spoon-billed sandpiper and red necked stint populations. Single individuals from three other species (red knot, little stint and long-toed stint) were used for phylogenic and outgroup-based analyses. DNA from all these samples were extracted with Gentra Purgene Tissue Kit following standard protocol. We extracted also a RNA from single embryo sample, to facilitate gene prediction and genome annotation. All individuals used for population analysis were sampled prior to 2011, the year a head-starting program for the spoon-billed sandpiper was initiated. RNA was extracted from samples collected after 2013 to aid in gene prediction.

Library preparation and sequencing To produce a reference genome, 0.5, 3 and 4.5 kb average insert sizes libraries were prepared from a single individual. In addition, a 450 bp fragment PCR-free library was prepared from DNA of the same bird. The 500bp fragment, and the two mate-pair libraries were sequenced on an Illumina HiSeq 2000 at the Center for Genomic Regulation (Barcelona, Spain) Genomics Unit. The 450 bp library was sequenced on an Illumina MiSeq to obtain 250bp overlapping reads. All reads were manually inspected using FASTQC and used for genome assembling. An overview of the samples used in this study can be found in **Supplementary Table 1**.

De novo assembly We ran DiscovarDeNovo assembler[54] (DDN V. r52488, default parameters) using the 450bp-fragment. The scaffolds obtained covered 1.27GB of the spoon-billed sandpiper genome (1.2GB expected genome size) and had an N50 of ~100kb. To improve the assembly, we excluded contigs shorter than 500bp. To extend and scaffold the contigs produced by DDN we ran SSPACE-SR[55] (v3.0) in two steps. First, we extended the contigs using the single 500bp library. Later, we scaffolded the extended assembly using both mate-pair libraries with SSPACE-SR (default parameters, -x = 0, -z = 0, -k = 5, -g = 0, -a = 0.7, -n = 15, -p = 0). The statistics for each assembly were calculated using the assemblathon pipeline[55] described in **Supplementary Table 2**.

Genome annotation The reference genome was annotated using a pipeline derived from EVidence Modeler[57], software that combines gene predictions, spliced- protein and transcript alignments. Prior to the gene-finding step the assembly's complex-repeats were hard-masked using RepeatMasker[58] (v4.0.5) with *G. gallus*-derived library of repeat elements. After that several ab-initio and evidence-based gene predictors were applied to the reference genome. Additionally, we aligned to the reference genome PASA-derived[59] transcriptomic sequences, highly curated vertebrate sequences and chicken protein models. All these sources of evidence were combined using specific weights. Functional annotation of a final set of genes was performed with a pipeline utilizing Interproscan[60], KEGG[61] and Blast2GO[62] software. Detailed information about genome annotation can be found in **Supplementary Materials**.

lncRNA annotation We annotated the long non-coding RNA (lncRNA) transcripts with a combination of tools, including the Coding Potential Calculator[63] and the Coding Potential Assessment Tool[64]. The required models for de novo prediction using CPAT were trained with 7,839 annotated long intergenic non-coding RNAs (lincRNAs) from the chicken, and an equal randomly selected number of coding sequences. These sequences were used as a proxy to detect the coding and non-coding potential of identified transcripts. The estimation of the optimum cutoff value (=0.59), to determine whether a transcript was either a protein coding or non-coding gene was assessed by using a ROC curve using data from 8161 CDS and 7839 lincRNAs (total=16000) from the chicken and a modification of the R script given by CPAT. The output of CPAT was fed to CPC and the remaining sequences were blasted[65] against the SBS protein-coding sequences predicted previously with EVidenceModeler[57]. Sequences with an overlap below 10% with a protein-coding gene were retained as well as those without any significant hit with a protein in the RefSeq database. Only transcripts >=200 nucleotides and with a ratio of transposable elements less than 0.4 were considered as lncRNAs.

Mapping and SNP calling Samples other than used for genome assembly were sequenced with Illumina HiSeq2000 platform (2 x 125bp reads). The quality of reads was assessed with FASTQC, and low quality reads were trimmed with Trimmomatic[66] (version 0.32). Reads were then mapped to the reference genome with bowtie2 [ref. 67] (version 2.2.3). PCR duplicates were marked, indels were realigned and sam/bam files were manipulated with SAMtools[68] and GATK[69], SNPs were called with SAMtools mpileup (with additional options: -C50 -R -t DP,ADF,ADR) and filtered with bcftools filter and vcftools. Based on the empirical distributions we decided to remove sites with QUAL ≤ 30, DP ≤ 40, DP ≥ 250, MQSB ≤ 0.001, and within 5 bp of indels. Indels were normalized and left aligned. We inferred ancestral state for spoon-billed sandpiper and red-necked stint with mrbayes[70] (version 3.2.6). SNPs were polarized, only if the probability of ancestral state was higher than 0.5. The missing and low quality genotypes were inferred separately for each species using BEAGLE[71] (version 4.1).

Relatedness and phylogeny reconstruction To infer relationship between individuals within each species we used the KING software[72] assuming no population structure. From each group of individuals related up to first degree we randomly selected one individual for future analyses avoiding samples that were sequenced with other technology (MiSeq) or to different depth (samples sequenced for genome assembly) (**Supplementary Table 1**). The principal component analyses on final samples was performed with PLINK[73] and phylogeny was inferred with SNPhylo[74] software using following options -m 0.001 -l 0.1 -p 20 -M 0.2 -b -B 100 -a 30000.

Population statistics We scanned the contigs longer than 50kb in non-overlapping 25kb windows to estimate nucleotide diversity and Tajima's D statistics within spoon-billed sandpiper and red-necked stint populations and to estimate genetic divergence ($\bar{F}_{ST}$) between species. VCFtools[75] (0.1.15) was used for these calculations and GO enrichment analyses for selected windows were performed with topGO package[76] and visualized with REVIGO[77]. To calculate Tajima's D statistic per specific region we used PopGenome package[78] within R. To count the number of polymorphisms and species-specific substitutions we excluded tri-allelic sites and sites polymorphic in both species. Using the reconstructed ancestral states, we counted number of polymorphisms and substitutions.

Mutation rate We used protein coding sequences from spoon-billed sandpiper and killdeer, to estimate mutation rate. The reciprocal blast was performed to identify orthologous sequences between species. We used 12,398 orthologous pairs to estimate rate of synonymous substitutions. Each pair of sequences was aligned based on the translated protein sequence. The alignment was then used to estimate rate of synonymous substitutions and number of synonymous sites with PAML[79] (yn00). The divergence time between spoon-billed sandpiper and killdeer was assumed to be 76.0 million years, according to the www.timetree.org[80]. The calculated mutation rate was 2.40×10^{-9} as a rate of mutation per nucleotide per year and 1.20×10^{-8} as a rate of mutation per nucleotide per generation (1 generation = 5 years), similar to the direct estimate obtained for the collared flycatcher[81].

Population history To perform PSMC[28] analyses the mpileup file was generated with SAMtools, and sites were filtered to have minimum mapping quality 25, consensus quality higher than 20 and depth in a range between 10 and 2 x d, where d is average depth[82]. PSMC was

then run with the following options -N50 -t5 -r5 -b -p "4+30*2+4+6+10" and output was generated with the mutation rate calculated as described above. The same procedure was performed for three species: spoon-billed sandpiper, red-necked stint and red knot using spoon-billed sandpiper genome as a reference. To validate PSMC results, we run analyses on three spoon-billed sandpiper samples sequenced to high coverage, and we performed the same analyses by mapping reads to the ruff genome[83]. To assess how population history inferences are affected by used software (PSMC) we performed similar analyses with SMC++ software[30]. The SMC++ was run for a single individual of spoon-billed sandpiper, red-necked stint and red knot (genotype depth > 6), for the interval ranges between 2000 and 200000 generations, which corresponds to 10 tya and 1mya.

Forward simulations We used evolutionary simulation package SLiM 2 [ref. 84] to run individual-based forward simulations. A representative sequence (200kb fragment) was divided into neutral and non-neutral parts, with sizes corresponding to the lengths of introns and exons respectively. Before simulations, population was evolving for at least 200,000 generations to reach mutation-selection-drift equilibrium. Simulations were then performed 5,000 times per every scenario. To improve simulations efficiency population sizes (N) were reduced by the factor of ten, while time (t), mutation rate (μ), recombination rate (r) and selection coefficient (s) were rescaled in a such way that the product of N/t, Nμ, Nr and Ns remain the same.

McDonald-Kreitman test

The preferential maintenance of deleterious recessive nonsynonymous alleles at higher frequency in the spoon-billed sandpiper population complicates a possible search for positive selection in its genome[85]. We used intergenic polymorphisms instead of synonymous ones in the McDonald-Kreitman test[86], such that the fraction of amino acid substitutions driven by positive selection, α, was calculated as $\alpha = 1 - Dinter*Pn/Dn*Pinter$, where Dinter and Dn is the number of fixed intergenic and nonsynonymous substitutions, respectively, and Pinter and Pn is the number of segregating intergenic and nonsynonymous sites, respectively. We found $\alpha = 1 - 1,217,193*29,992/10,687*2,809,956 = -0.22$ and $\alpha = 1 - 2,141,855*38,523/16,399*4,492,302 = -0.12$ for the spoon-billed sandpiper and the red-necked stint, respectively, suggesting a prevalence of low-frequency deleterious non-synonymous polymorphisms in both populations. Considering only common alleles[87], those found in >4 genomes, α, the measure of positive selection remained negative for the spoon-billed sandpiper but became slightly positive for the red-necked stint ($\alpha = 1 - 1,217,193*12,822/10,687*1,405,028 = -0.04$ and $\alpha = 1 - 2,141,855*12,881/16,399*1,755,833 = 0.04$, respectively).

Acknowledgements

We are indebted to Roman Belogorodtsev and Svetlana Belogorodtseva for invaluable on-site support of our field work. We thank Alex Kondrashov and Nick Barton for feedback on the impact of population dynamics on deleterious recessive alleles, Jochen Hecht and the CRG Genomics Unit for the next-generation sequencing, Yun Song for assistance with SMC++

application and Krystyna Nadachowska-Brzyska for help with PSMC analyses. We thank Tyler Alioto of the CNAG (http://www.cnag.cat/institution/) in Barcelona for insightful technical assistance and development of some of the scripts used in the analysis. This work was partially supported by the Instituto Nacional de Bioinformática (INB) from ISCIII in Spain. Mateusz Konczal was supported by the Foundation for Polish Science and the Polish National Science Centre (2016/20/S/NZ8/00208). Luis Zapata was supported by the International PhD scholarship program of La Caixa at CRG. Pavel Tomkovich was supported by the Russian Science Foundation grant (RSF No. 14-50-00029). The work was supported by Royal Society for Protection of Birds (UK), Wildfowl and Wetlands Trust (UK), Manfred Hermsen Foundation (Germany), NABU (Germany), Keidanren Foundation (Japan), the administration of Chukotka Autonomous District of Russian Federation, Bird Life International, Japan Ramsar Network, Heritage Expeditions (New Zealand), HHMI International Early Career Scientist Program (55007424), the MINECO (BFU2012-31329 and BFU2015-68723-P), Spanish Ministry of Economy and Competitiveness Centro de Excelencia Severo Ochoa 2013-2017 grant (SEV-2012-0208), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat's AGAUR program (2014 SGR 0974), the CERCA Programme of the Generalitat de Catalunya, and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013, ERC grant agreement 335980_EinME).

References

1. Pimm, S. L. et al. The biodiversity of species and their rates of extinction, distribution, and protection. *Science* **344**, 1246752 (2014).

2. Tilman, D., Clark, M., Williams, D. R., Kimmel, K., Polasky, S., & Packer, C. Future threats to biodiversity and pathways to their prevention. *Nature* **546**, 73–81 (2017).

3. Barnosky, A. D. et al. Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51–57 (2011).

4. Bairlein, F. Migratory birds under threat. *Science* **354**, 547–548 (2016).

5. Hua, N., K. Tan, Y. Chen & Ma, Z. Key research issues concerning the conservation of migratory shorebirds in the Yellow Sea region. *Bird Cons. Int*. **25**, 38–52 (2015).

6. Studds, C. E., et al. Rapid population decline in migratory shorebirds relying on Yellow Sea tidal mudflats as stopover sites. *Nature Commun.* **8**, 14895 (2017).

7. Clark, N. A., Anderson, G. Q. A., Li, J., Syroechkovskiy, E. E., Tomkovich, P. S., Zöckler, C., Lee, R. & Green, R. E. First formal estimate of the world population of the Critically Endangered spoon-billed sandpiper Calidris pygmaea. *Oryx* 1–10 (2016).

8. Syroechkovski, E. E., Tomkovich, P. S., Kashiwagi, M., Taldenkov, I. A., Buzin, V. A., Lappo, E. G. & Zöckler, C. Population decline in the spoon-billed sandpiper (Eurynorhynchus pygmeus) in northern Chukotka based on monitoring on breeding grounds. *Biol. Bull*. **37**, 941–951 (2010).

9. Zöckler, C., Syroechkovskiy, E. E. & Atkinson, P. W. Rapid and continued population decline in the Spoon-billed Sandpiper Eurynorhynchus pygmeus indicates imminent extinction unless conservation action is taken. *Bird Cons. Int.* **20**, 95–111 (2010).

10. Zöckler, C., T. Htin Hla, N. Clark, E. Syroechkovskiy, N. Yakushev, S. Daengphayon & Robinson R. Hunting in Myanmar is probably the main cause of the decline of the Spoon-billed Sandpiper Calidris pygmeus. *Wader Study Group Bull.* **117**, 1–8 (2010).

11. Piersma, T., Chan, Y., Mu, T., Hassell, C. J., Melville, D. S., Peng, H., Ma, Z., Zhang, Z. & Wilcove, D. S. Loss of habitat leads to loss of birds: Reflections on the jiangsu, China, coastal development plans. *Wader Study* **124**, 93–98 (2017).

12. Schraiber, J. G., & Akey, J. M. Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* **16**, 727–740 (2015).

13. Lande R. Genetics and demography in biological conservation. *Science* **241**, 1455–1460 (1988).

14. Lynch, M., Conery, J., & Burger, R. Mutation accumulation and the extinction of small populations. *Am. Nat.* **146**, 489–518 (1995).

15. Frankham, R. Conservation genetics. *Annu. Rev. Genet*. **29**, 305–327 (1995).

16. Saccheri, I., Kuussaari, M., Kankare, M., Vikman, P., Fortelius, W. & Hanski, I. Inbreeding and extinction in a butterfly metapopulation. *Nature* **392**, 491–494 (1998).

17. Kirkpatrick, M. & Jarne, P. The Effects of a Bottleneck on Inbreeding Depression and the Genetic Load. *Am. Nat*. **155**, 154–167 (2000).

18. Acevedo-Whitehouse, K., Gulland, F., Greig, D. & Amos, W. 2003. Inbreeding: disease susceptibility in California sea lions. *Nature* **422**, 35 (2003).

19. Frankham, R. Genetics and conservation biology. *C. R. Biol*. **326,** S22–S29 (2003).

20. Spielman, D., Brook, B. W. & Frankham, R. Most species are not driven to extinction before genetic factors impact them. *Proc. Natl. Acad. Sci. USA.* **101**, 15261–15264 (2004).

21. Polishchuk, L. V., Popadin, K. Y., Baranova, M. A. & Kondrashov, A. S. A genetic component of extinction risk in mammals. *OIKOS* **124**, 983–993 (2015).

22. Robinson, J. A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., Marsden, C. D., Lohmueller, K. E., & Wayne, R. K. Genomic flatlining in the endangered island fox. *Current Biol*. **26**, 1183–1189 (2016).

23. Hedrick, P. W. & Garcia-Dorado, A. Understanding inbreeding depression, purging, and genetic rescue. *Trends Ecol. Evol.* **31**, 940–952 (2016).

24. Lappo, E. G., Tomkovich, P. S. & Syroechkovskiy, E. E. *Atlas of breeding waders in the Russian Arctic* (Moscow, Russia, 2012).

25. Rogers, R. L. & Slatkin, M. Excess of genomic defects in a woolly mammoth on Wrangel island. *PLoS Genet*. **13**, e1006601 (2017).

26. Li, S. et al. Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biol.* **15**, 557 (2014).

27. Abascal, F. et al. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biol*. **17**, 251 (2016).

28. Li, H. & Durbin, R. (2011) Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).

29. Red'kin, Y. A., Tomkovich, P. S. & Zdorikov, A. I. Unusual specimen of the Spoon-billed Sandpiper Eurynorhynchus pygmeus. *Wader Study Group Bull.* **119**, 56–59 (2012).

30. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).

31. Clark, P. U. et al. The Last Glacial Maximum. *Science* **325**, 710–714 (2009).

32. Piersma, T. Using the power of comparison to explain habitat use and migration strategies of shorebirds worldwide. *J. Ornithology* **148**, S45–S59 (2007)

33. Buehler, D. M., Baker, A. J., Piersma, T. Reconstructing palaeoflyways of the late Pleistocene and early Holocene Red Knot Calidris canutus. *Ardea* **94**, 485–498 (2006).

34. Yokoyama, Y., Lambeck, K., De Deckker, P., Johnston, P., Fifield, L.K., Timing of the Last Glacial Maximum from observed sea-level minima. *Nature* **406**, 713–716 (2000).

35. Clark J., Mitrovica, J. X. Alder J. Coastal paleogeography of the California-Oregon-Washington and Bering Sea continental shelves during the latest Pleistocene and Holocene: implications for archaeological records. *J. Arch. Science* **52**, 12–23 (2014).

36. Erlandson, J. M., Braje, T. J., Gill, K. M. & Graham, M. H. Ecology of the kelp highway: did marine resources facilitate human dispersal from northeast Asia to the Americas? *J. Isl. Coast. Arch.* **10**, 1–20, (2015).

37. Larson, C. Hostile shores. *Science* **350**, 150–152 (2015).

38. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

39. Andolfatto, P. Adaptive revolution of non-coding DNA in Drosophila. *Nature* **437**, 1149–1152 (2005).

40. Balick, D. J., Do, R., Cassa, C. A., Reich, D. & Sunyaev, S. R. Dominance of deleterious alleles controls the response to a population bottleneck. *PLoS Genet*. **11**, e1005436 (2015).

41. Amorim, C. E. G., Gao, Z., Baker, Z., Diesel, J. F., Simons, Y. B., Haque, I. S., Pickrell, J. & Przeworski, M. The population genetics of human disease: The case of recessive, lethal mutations. *PLoS Genet.* **13**, e1006915 (2017).

42. Gazave, E., Chang, D., Clark, A.G. & Keinan, A. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics* **195**, 969–978 (2013).

43. Peischl, S. & Excoffier, L. Expansion load: recessive mutations and the role of standing genetic variation. *Mol. Ecol.* **24**, 2084–2094 (2015).

44. Kimura, M., Maruyama, T. & Crow, J. F. The mutation load in small population. *Genetics* **48**, 1303–1312 (1963).

45. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).

46. Crow, J. F. & Kimura, M. *An introduction to population genetics theory*. (Harper & Row, New York, 1970).

47. Kondrashov, F. A., Ogurtsov, A. Y, & Kondrashov, A. S. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *J. Theor. Biol.* **240**, 616–626 (2006).

48. Abzhanov, A., Protas, M., Grant, B. R., Grant, P. R. & Tabin, C. J. Bmp4 and morphological variation of beaks in Darwin's finches. *Science* **305**, 1462–1465 (2004).

49. Wu, P., Jiang, T. X., Suksaweang, S., Widelitz, R. B. & Chuong, C. M. Molecular shaping of the beak. *Science* **305**, 1465–1466 (2004).

50. Morton, N. E., Crow, J. F. & Muller, H. J. An estimate of the mutational damage in man from data on consanguineous marriages. *Proc. Natl. Acad. Sci. USA* **42**, 855–863 (1956).

51. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating Mutation Load in Human Genomes. *Nat. Rev. Genet*. **16**, 333–343 (2015).

52. Gilpin, M. E. & Soulé, M. E. Minimum viable populations: processes of species extinction. in *Conservation Biology: The Science of Scarcity and Diversity* (ed. Soulé, M. E.) 19–34 (Sinauer, Sunderland, Mass, 1986).

53. Frenzel, B., Pecsi, M. & A.A.Velichko. *Atlas of paleoclimates and paleoenvironments of the northern hemisphere*. (Gustav Fischer, New York, 1992).

54. Love, R. R., Weisenfeld, N. I., Jaffe, D. B., Besansky, N. J. & Neafsey, D. E. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics* **17**, 187 (2016).

55. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).

550 56. Earl, D., Bradnam, K., John, J. S., Darling, A., Lin, D., Fass, J., Yu, H. O. K., Buffalo, V.,
551 Zerbino, D. R. & Diekhans, M. Assemblathon 1: A competitive assessment of de novo short read
552 assembly methods. Genome *Res.* **21**, 2224–2241 (2011).

553 57. Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C.
554 R. & Wortman, J. R. Automated eukaryotic gene structure annotation using EVidenceModeler
555 and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).

556 58. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in
557 genomic sequences. *Cur. Protoc. Bioinformatics.* **25**, 4–10 (2009).

558 59. Haas, B. J., et al. Improving the Arabidopsis genome annotation using maximal transcript
559 alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).

560 60. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*
561 **30**, 1236–1240 (2014).

562 61. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acid Res.*
563 **28**, 27–30 (2000).

564 62. Conesa, A, et al. Blast2GO: a universal tool for annotation, visualization and analysis in
565 functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).

566 63. Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L. & Gao, G. CPC: assess the
567 protein-coding potential of transcripts using sequence features and support vector machine.
568 *Nucleic Acids Res.* **35**, W345–W349 (2007).

569 64. Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P. & Li, W. CPAT: Coding-Potential
570 Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74
571 (2013).

572 65. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
573 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

574 66. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
575 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

576 67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods*
577 **9**, 357–359 (2012).

578 68. Li, H., et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–
579 2079 (2009).

580 69. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing
581 next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

582 70. Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget,
583 B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. MrBayes 3.2: efficient Bayesian phylogenetic
584 inference and model choice across a large model space. *Syst. Biol.* 61, 539–542 (2012).

71. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

72. Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M. & Chen, W. M. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26** 2867–2873 (2010).

73. Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M. & Lee, J. J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7 (2015).

74. Lee, T. H., Guo, H., Wang, X., Kim, C., & Paterson, A. H. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**, 162 (2014).

75. Danecek, P., et al. 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

76. Alexa, A., & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. R package version 2, (2010).

77. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE* **6**, e21800 (2011).

78. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).

79. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

80. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

81. Smeds, L., Qvarnström, A. & Ellegren, H. Direct estimate of the rate of germline mutation in a bird. *Genome Res.* **26**, 1211–1218 (2016).

82. Nadachowska-Brzyska, K., Burri, R., Smeds, L. & Ellegren, H. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Mol. Ecol.* **25**, 1058–1072 (2016).

83. Lamichhaney, S. et al. Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet.* **48**, 84–88 (2016).

84. Haller, B. C. & Messer, P. W. SLiM 2: Flexible, interactive forward genetic simulations. *Mol. Biol. Evol.* **34**, 230–240 (2017).

85. Eyre-Walker, A. Changing effective population size and the McDonald-Kreitman test. *Genetics* **162**, 2017–2024 (2002).

86. McDonald, J. H. & Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652–654 (1991).

87. Charlesworth, J. & Eyre-Walker, A. The McDonald-Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol.* **25**, 1007–1015 (2008).

88. Parra, G., Blanco, E. & Guigó, R. GeneID in Drosophila. *Genome Res.* **10**, 511-515 (2000).

89. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics* **4**, 4.3. (2007).

90. Parra, G. et al. Comparative gene prediction in human and mouse. *Genome Res.* **13**, 108-117 (2003).

91. Consortium ICGS. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).

92. Eyras. E., et al. Gene finding in the chicken genome. *BMC Bioinformatics* **6**, 131 (2005).

93. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

94. Gibbs, R. A., et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).

95. Dobin, A., et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

96. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–34 (2014).

97. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).

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

99. Borodovsky, M., et al. Eukaryotic gene prediction using GeneMark.hmm. *Curr. Protoc. Bioinformatics* Chapter **4**, 4.6 (2003).

100. Lomsadze, A., et al. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).

101. Iwata, H. & Gotoh, O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res.* **40**, e161 (2012).

102. Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 1–11 (2005).

652 103. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining,
653 estimating and interpreting F_{ST}. *Nat. Rev. Genet.* **10**, 639–650 (2009).

Tables

Table 1. Polymorphisms and fixed substitutions in the spoon-billed sandpiper and the red-necked stint genomes.

	Spoon-billed sandpiper	**Red-necked stint**
Total fixed substitutions	2,296,680	3,978,785
intergenic	1,217,193	2,141,855
nonsynonymous	10,687	16,399
synonymous	22,900	26,597
Total SNPs	5,218,932	9,194,008
SNPs without singletons	3,681,367	6,288,764
Indel polymorphisms	691,673	1,471,359
Indel polymorphisms without singletons	480,457	1,225,860

Table 2. Average number of variants per diploid genomes divided into functional categories of SNPs.

	Average number in a diploid spoon-billed sandpiper genome (proportion relative to intergenic polymorphisms in the same genome)	**Average number in a diploid red-necked stint genome (proportion relative to intergenic polymorphisms in the same genome)**	**Ratio of the proportions of polymorphism category to intergenic polymorphisms**
Intergenic	1,638,659 (1)	2,359,340 (1)	1 (N/A)
Intronic	1,236,189 (0.7544)	1,863,653 (0.7899)	0.955 ($p < 0.0001$)
Synonymous	28,947 (0.0177)	45,326 (0.01921)	0.920 ($p < 0.0001$)
Upstream	83,772 (0.0511)	113,732 (0.0482)	1.061 ($p < 0.0001$)
Downstream	87,056 (0.0531)	125,504 (0.0532)	0.999 ($p = 0.77$)
lncRNA	220,976 (0.1349)	312,433 (0.1324)	1.019 ($p < 0.0001$)
Nonsynonymous	16,657 (0.0102)	20,319 (0.0086)	1.180 ($p < 0.0001$)
Nonsense	201 (1.23×10^{-4})	249 (1.06×10^{-4})	1.160 ($p = 0.12$)

659 **Table 3** Average number of derived homozygotes per diploid genome divided into functional categories of SNPs.

	Average number of derived homozygotes in a spoon-billed sandpiper genome (proportion relative to intergenic polymorphisms in the same genome)	Average number of derived homozygotes in a red-necked stint genome (proportion relative to intergenic polymorphisms in the same genome)	Ratio of the proportions of polymorphism category to intergenic polymorphisms (p-value of Chi-square test)
Intergenic	432,317 (1)	636,708 (1)	1 (N/A)
Intronic	325,652 (0.7533)	504,096 (0.7917)	0.951 ($p < 0.0001$)
Synonymous	7,476 (0.0173)	11,599 (0.0182)	0.949 ($p < 0.001$)
Upstream	22,652 (0.0524)	32,014 (0.0551)	1.042 ($p < 0.0001$)
Downstream	23,133 (0.0535)	35,064 (0.0551)	0.972 ($p < 0.001$)
lncRNA	61,121 (0.1414)	88,994 (0.1398)	1.012 ($p < 0.05$)
Nonsynonymous	4,130 (0.0096)	5,324 (0.0084)	1.143 ($p < 0.0001$)
Nonsense	39 (9.0×10^{-5})	60 (9.4×10^{-5})	0.957 ($p = 0.92$)

660

661

Figures

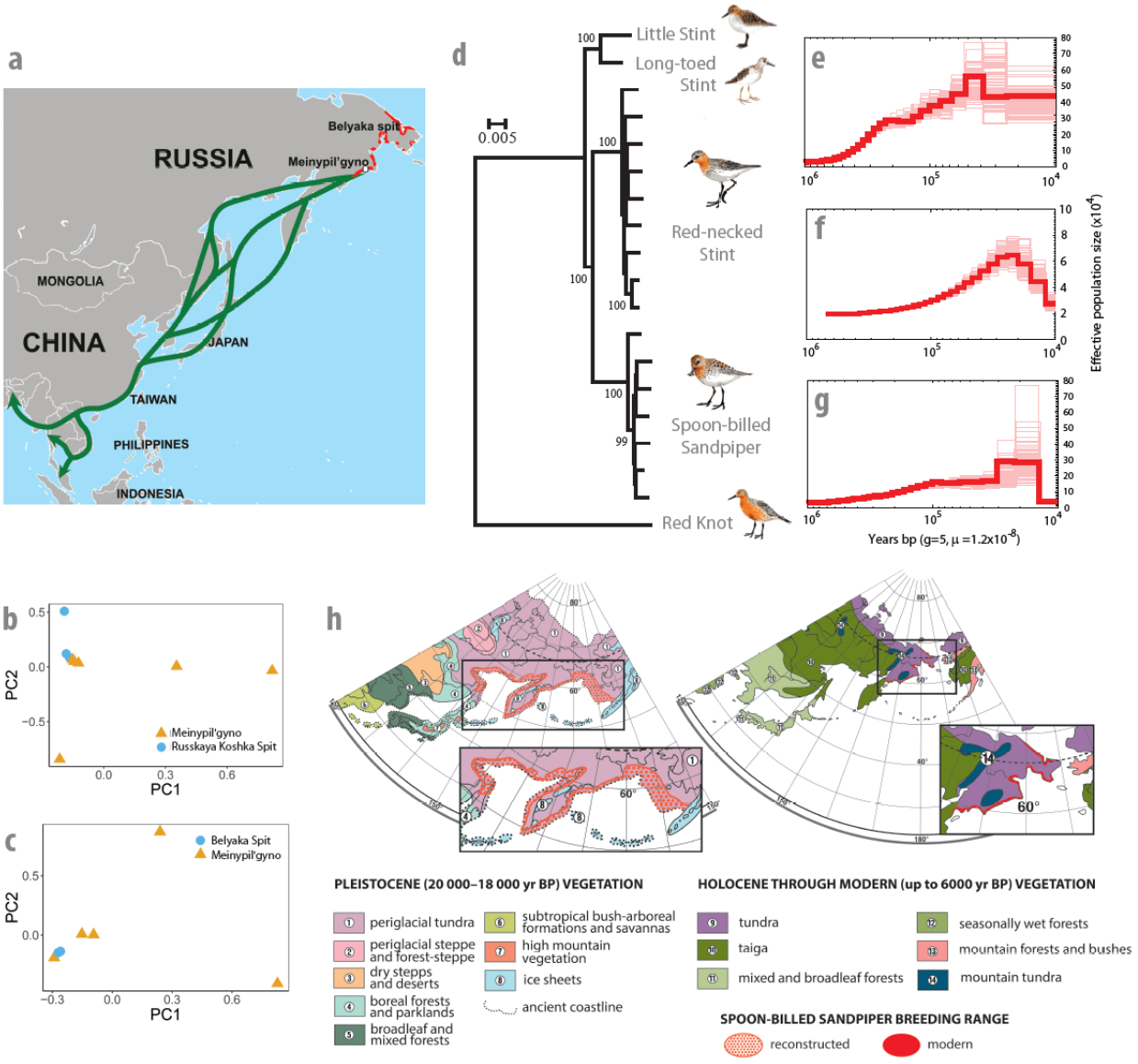


Figure 1. Breeding range, history and genetic structure of the spoon-billed sandpiper and red-necked stint populations. a, The flyway and the breeding range of the spoon-billed sandpiper[24], in green arrows and red, respectively. Spoon-billed sandpiper samples were collected in Meinypil'gyno and the Belyaka Spit. **b**,**c** Principal Component Analyses of the red-necked stint (**b**) and spoon-billed sandpiper (**c**) populations based on variance-standardized relationship matrix. **d**, The phylogenetic relations between spoon-billed sandpiper, red-necked stint, little stint, long-toed stint and red knot reconstructed based on SNP data. **e**,**f**,**g** Demographic history PSMC reconstruction for red-necked stint (**e**), spoon-billed sandpiper (**f**) and red knot (**g**). **h**, Modern and reconstructed Pleistocene breeding habitat[53] of the spoon-billed sandpiper.

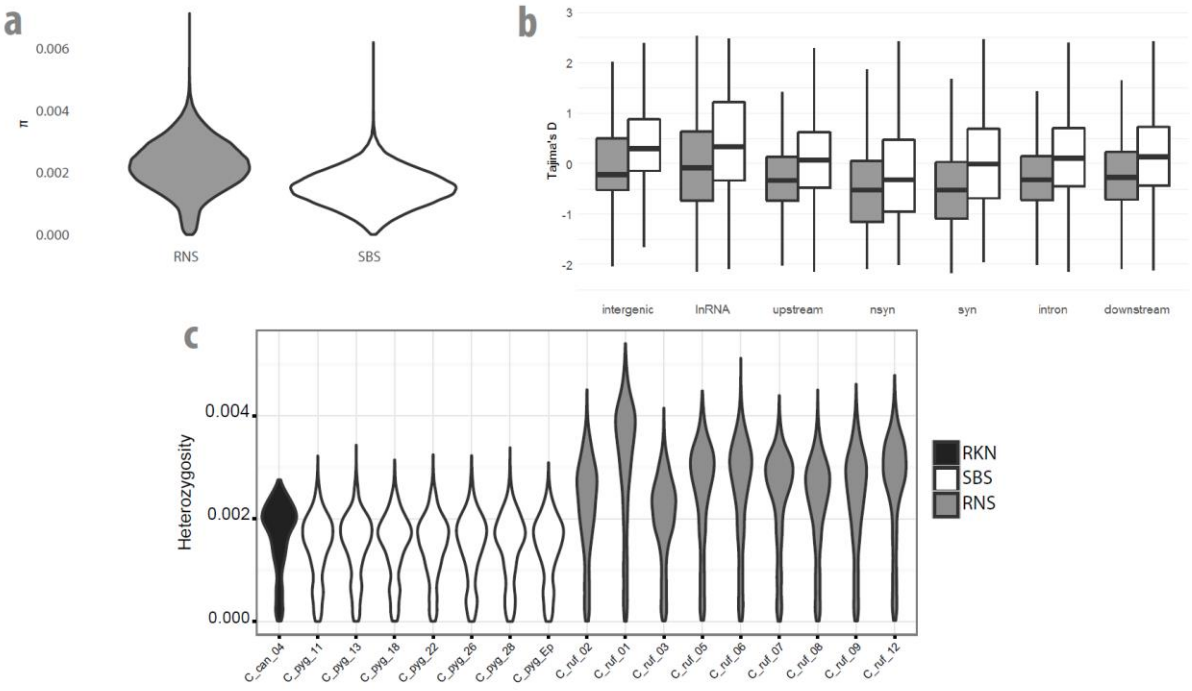


Figure 2. Patterns of genetic variation of red-necked stint and spoon-billed sandpiper populations. a, Distributions of nucleotide diversity values calculated in 50kb windows; **b**, Tajima's D values calculated for functional classes of sites; **c**, Distributions of heterozygosity values calculated in 1 Mbp windows for red knot (black), spoon-billed sandpiper (white)and red-necked stint (grey).

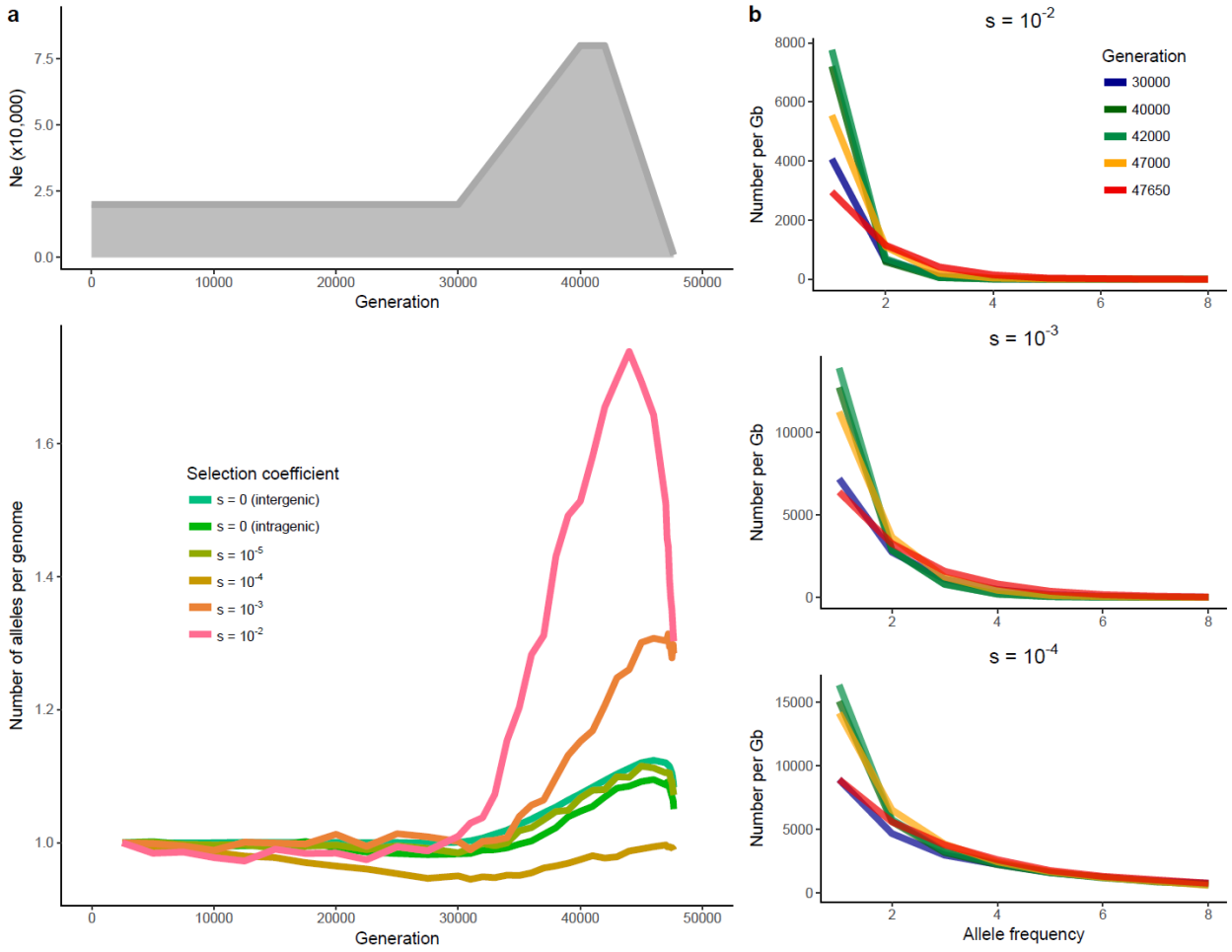
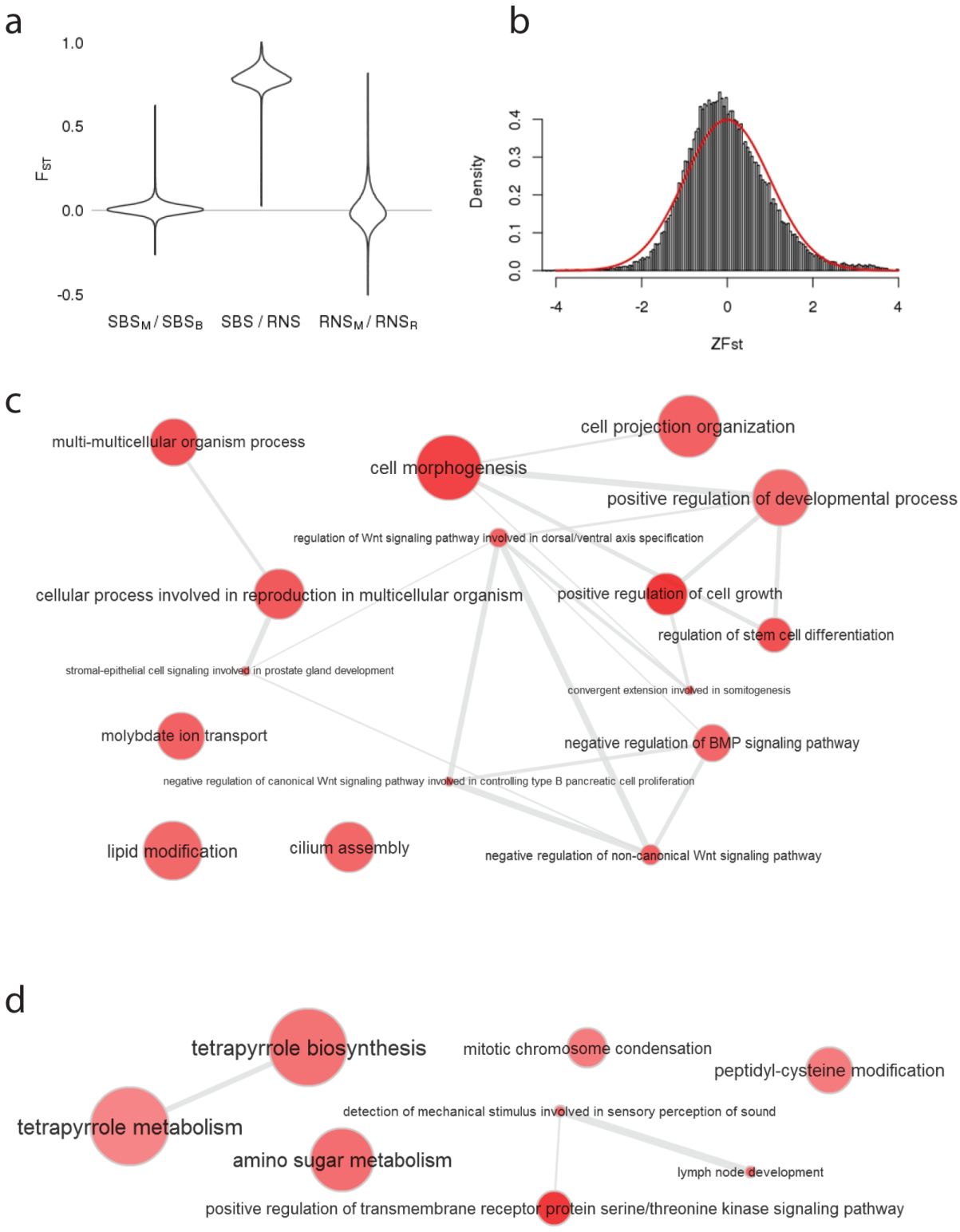


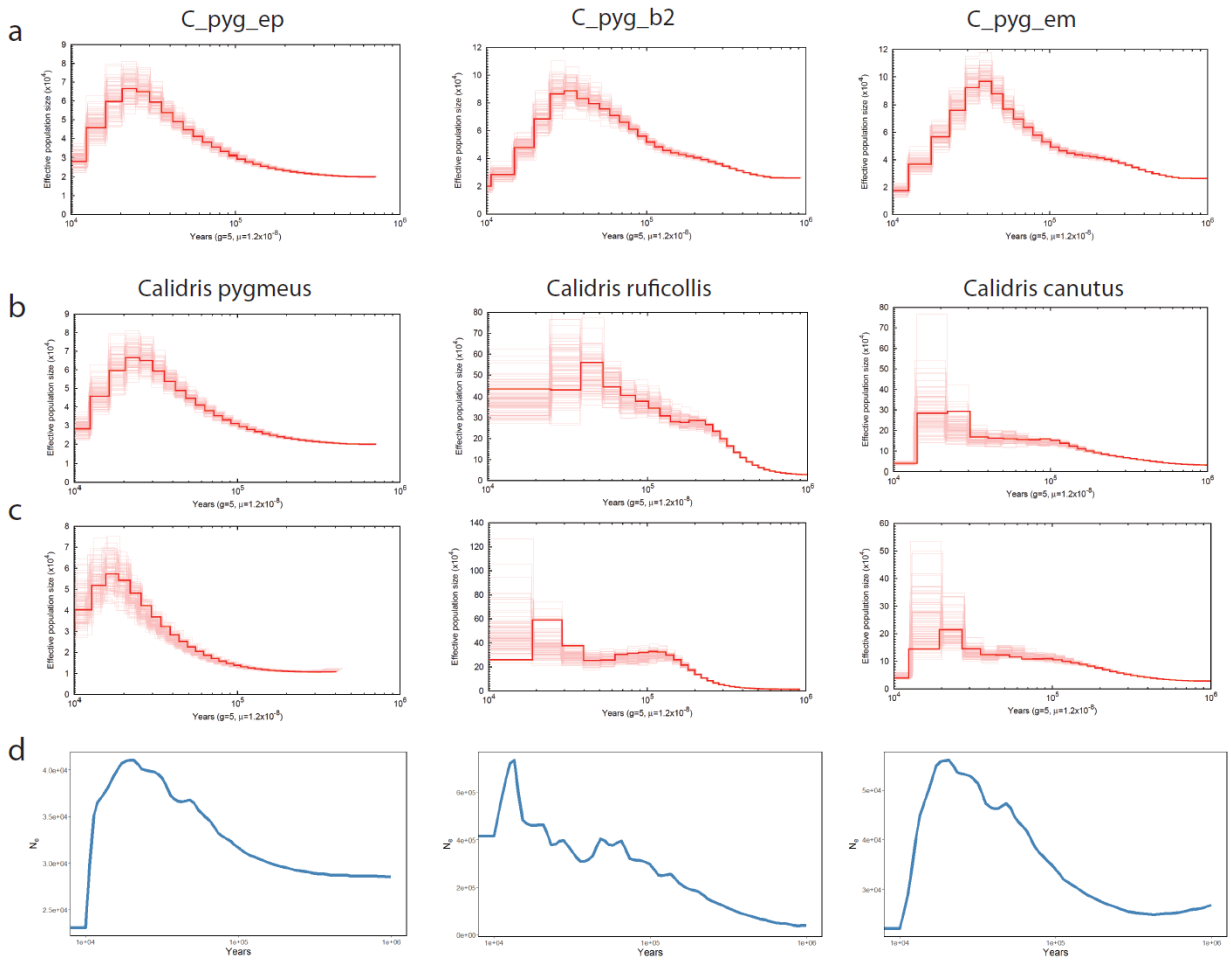
Figure 3. Simulations of demographic changes and relative prevalence of polymorphisms in the population. a, The grey graph shows the simulated population demographic changes following a 1,000,000 generation burn in. The number of recessive polymorphisms (h = 0) per genome relative to the number at generation 0 with a spoon-billed sandpiper-like demographic history are shown as a function of the strength of selection (colour-coded). **b**, The allele frequency distribution of recessive polymorphisms in different selection categories across different time points (colour-coded) of the simulation of the spoon-billed sandpiper-like demographic history.

Extended Data Figures

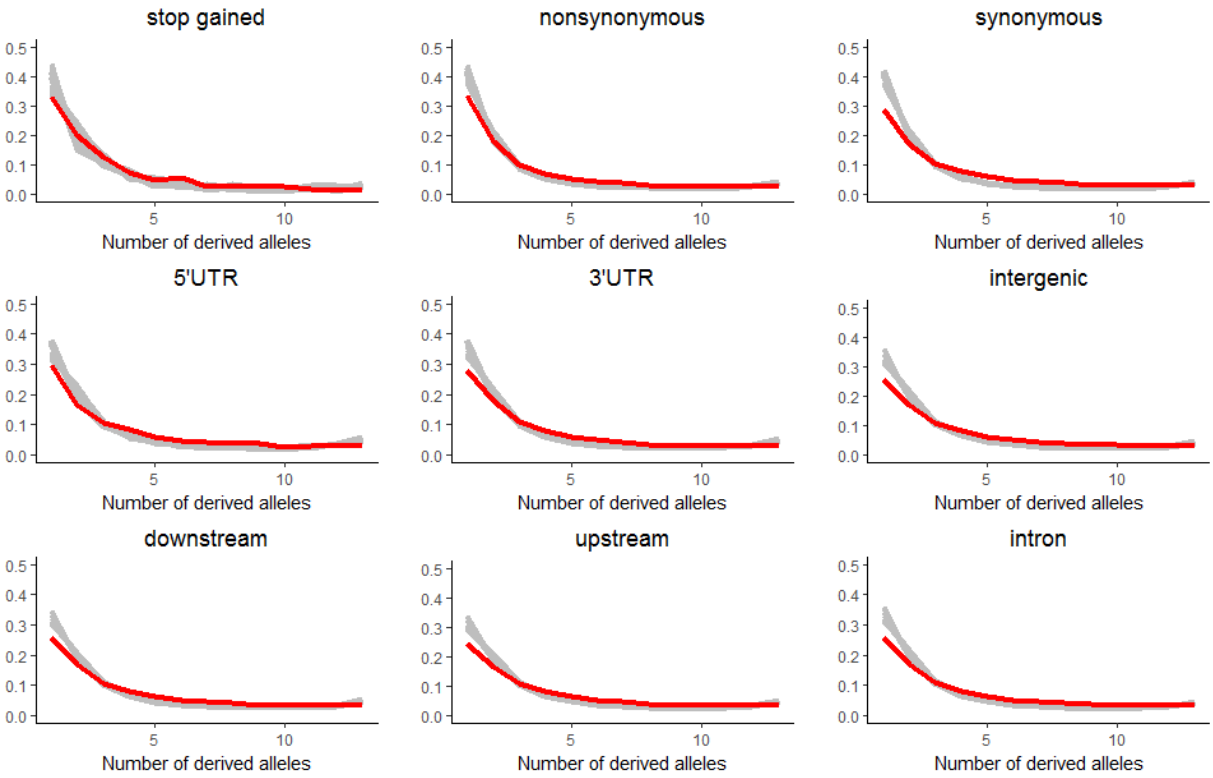


Extended Data Figure 1. Genetic differentiation between spoon-billed sandpiper and red-necked stint. a, Distributions of F_{ST} values, calculated in 25kbp windows, between spoon-billed sandpipers from Meinypil'gyno and the Belyaka Spit, between red-necked stints from the same two locations and between the two species. **b,** Distribution of Z-scored F_{ST} values[103] between the spoon-billed sandpiper and the red-necked stint compared with the normal distribution. **c,** Representation of GO terms overrepresented in regions of high differentiation ($ZF_{ST} > 2.5$) and faster rate of substitutions in the spoon-billed sandpiper compared to the red-necked stint. Bubble colour indicates the p-value; bubble size indicates the frequency of the GO term in the underlying GOA database. Highly similar GO terms are linked by edges in the graph, where the line width indicates the degree of similarity. **d:** Graphic representation of GO terms overrepresented in regions of high differentiation ($ZF_{ST} > 2.5$) and faster rate of substitutions in the red-necked stint compared to the spoon-billed sandpiper.

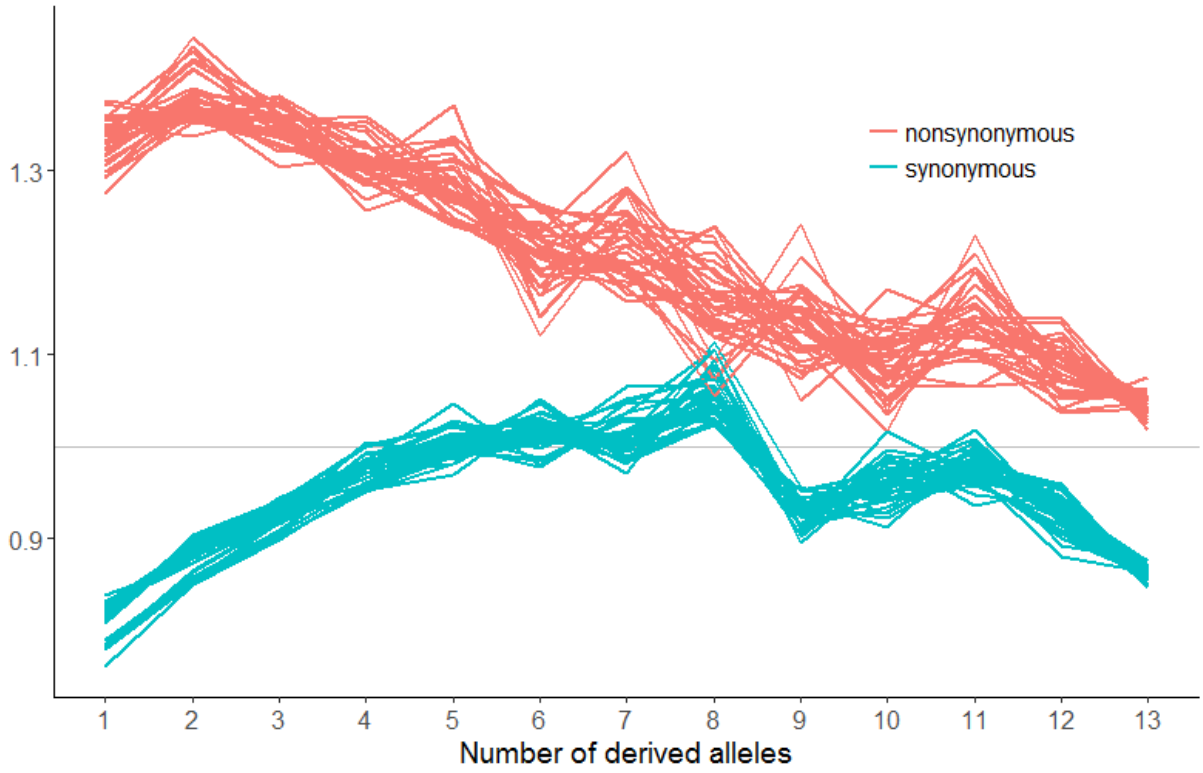
Extended Data Figure 2. Admixture proportion for 7 spoon-billed sandpiper (green) and 9 red-necked stint individuals (red), and cross-validation plot for number of groups from 1 to 5 (right).



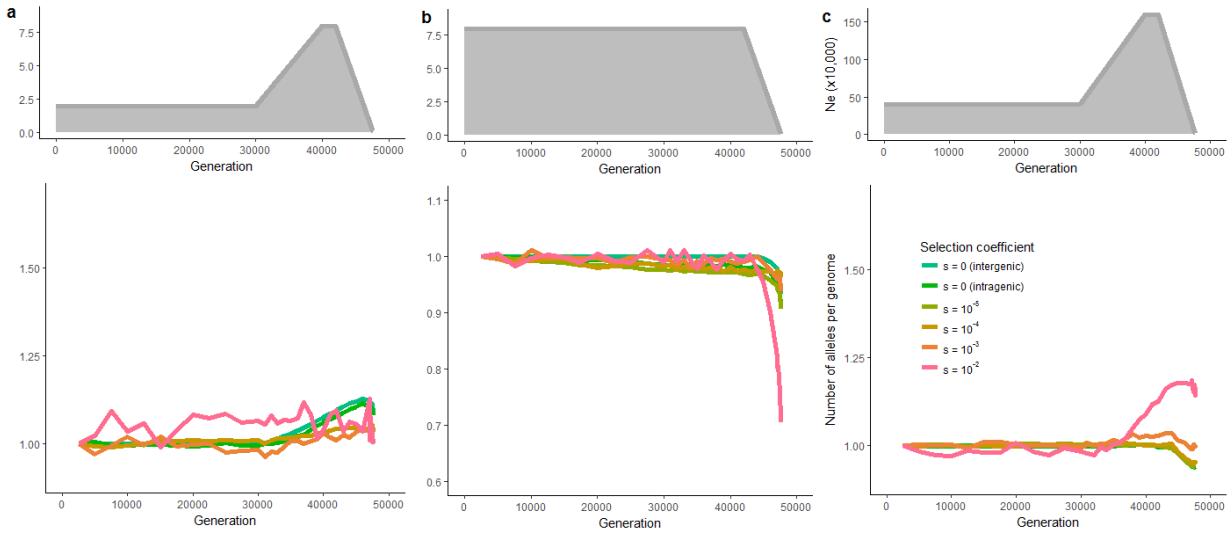
Extended Data Figure 3. Demographic history reconstructions. a, History reconstructions for spoon-billed sandpiper population based on three independent samples based on PSMC. **b,c,** History reconstructions for spoon-billed sandpiper (*C. pygmaea*), red-necked stint (*C. ruficollis*) and red knot (*C. canutus*) inferred based on PSMC model and sequencing reads mapped to the spoon-billed sandpiper reference genome (**b**) and to the ruff (*Philomachus pugnax*) reference genome (**c**). **d,** History reconstruction for spoon-billed sandpiper, red-necked stint and red knot inferred based on SMC++ model.



Extended Data Figure 4. Allele frequency spectra. The allele spectra for 7 spoon-billed sandpipers (red line) and all 7 out of 9 combinations of red-necked stint individuals (grey) populations are shown.



Extended Data Figure 5. Ratio of the number of alleles as a function of the allele frequency. The ratio of the number of nonsynonymous and intergenic alleles in the spoon-billed sandpiper divided by the same ratio from the red-necked stint (red). The ratio of the number of synonymous and intergenic alleles in the spoon-billed sandpiper divided by the same ratio from the red-necked stint (blue).



Extended Data Figure 6. Simulations of demographic changes and prevalence of polymorphisms in the population. The grey graphs show the simulated population demographic changes following a 200,000 generation burn in. For all panels the number of polymorphisms is reported relative to the number observed at the first generation. **a,** Relative change in the number of semidominant polymorphisms (h = 0.2) per genome with a spoon-billed sandpiper-like demographic history. **b,** Relative change in the number of recessive polymorphisms (h = 0) per genome in a population undergoing a simple bottleneck, as a function of the strength of selection (colour-coded). **c,** Relative changes in the number of recessive polymorphisms (h = 0) per genome with a relatively large population, as a function of the strength of selection (colour-coded).

SUPLEMENTARY MATERIALS

I. TABLES

746 **Supplementary Table 1. Samples origin (M – Meinopylgino, BS - Belyaka Spit, RKS – Russkaya Koshka Spit) and description.**

Sample name	Museum specimen	Species	Age	Collection year	Location	Sequenc. mode	Number of reads (mln)	% mapped reads	Mean coverage	Used for	Relative of
C_can_04	N238	*C. canutus*	fled.	2010	-	2 x 125	280.8	85.50%	25.08	SNP calling	
C_min_18	R118443	*C. minuta*	ad.	11/06/2002	M	2 x 125	45.7	88.90%	4.22	SNP calling	
C_pyg_09	E	*C. pygmaea*	?	2013	M	2 x 125	181	96.60%	18.05		C_pyg_28, C_pyg_29, C_pyg_b1, C_pyg_b2
C_pyg_11	F50 173	*C. pygmaea*	ad.	09/06/2002	BS	2 x 125	152	95%	14.88	SNP calling	
C_pyg_12	F50 0130	*C. pygmaea*	ad.	18/06/2002	BS	2 x 125	135.4	93.50%	13.05		C_pyg_13
C_pyg_13	F50 0153	*C. pygmaea*	ad.	26/06/2002	BS	2 x 125	146.6	93.30%	14.09	SNP calling	
C_pyg_18	F50 1414	*C. pygmaea*	ad.	04/07/2003	M	2 x 125	156.5	92.60%	14.93	SNP calling	
C_pyg_22	234 #10	*C. pygmaea*	?	12/07/2010	M	2 x 125	175.6	94.20%	17.16	SNP calling	
C_pyg_26	F50 1493	*C. pygmaea*	ad.	24/06/2003	M	2 x 125	168.3	94.60%	16.46	SNP calling	
C_pyg_27	F50 0142	*C. pygmaea*	ad.	21/06/2002	BS	2 x 125	127.8	95.10%	12.51		C_pyg_11
C_pyg_28	-	*C. pygmaea*	juv.	30/07/2013	M	2 x 125	161.6	95.90%	16	SNP calling	
C_pyg_29	-	*C pygmaea*	fled.	25/07/2013	M	2 x 125	170.7	96.80%	17.08		C_pyg_28, C_pyg_9, C_pyg_b1, C_pyg_b2
C_pyg_b1		*C. pygmaea*	emb.		M	2 x 250	211.8	95.30%	38.43	Geme assembly	C_pyg_9 C_pyg_28, C_pyg_29
C_pyg_b2		*C. pygmaea*	emb.		M	2 x 250	204.3	95.40%	36.48	Geme assembly	C_pyg_9 C_pyg_28, C_pyg_29
C_pyg_Ep		*C. pygmaea*	?		M	2 x 150	356.4	80.70%	35	SNP calling	
C_ruf_01	NNY061	*C. ruficollis*	ad.	14/07/2006	RKS	2 x 125	107.5	92.40%	10.37	SNP calling	
C_ruf_02	347	*C. ruficollis*	?	2012	-	2 x 125	263.8	93.70%	25.76	SNP calling	
C_ruf_03	R126355	*C. ruficollis*	fled.	08/07/2009	M	2 x 125	98.3	94.30%	9.63	SNP calling	
C_ruf_05	NNY062	*C. ruficollis*	adult	14/07/2006	RKS	2 x 125	120.2	93.60%	11.76	SNP calling	
C_ruf_06	No285	*C. ruficollis*	ad.	24/06/2011	M	2 x 125	108.7	94.80%	10.74	SNP calling	
C_ruf_07	No464	*C. ruficollis*	juv.	04/08/2014	M	2 x 125	130	94.70%	12.85	SNP calling	
C_ruf_08	No231	*C. ruficollis*	ad.	20/06/2010	M	2 x 125	98.1	93.90%	9.6	SNP calling	
C_ruf_09	NNY056	*C. ruficollis*	ad.	07/07/2006	M	2 x 125	120	93.00%	11.66	SNP calling	
C_ruf_12	zmmu777	*C. ruficollis*	fled.	09/07/2012	M	2 x 125	107.7	95.50%	10.72	SNP calling	

C_sub_11	No361	*C. subminuta*	ad.	04/06/2013	-	2 x 125	43.2	89.50%	4.03	SNP calling	

747

Supplementary Table 2. Spoon-billed sandpiper genome statistics.

Genome length (Gbases)	1.178
Number of scaffolds	29,994
N50 scaffold (Mbases)	2.781
L50 scaffold	108
No of scaffolds > 100kb	775
N% in scaffolds	0.69
Number of contigs	44,007
N50 contigs (Kbases)	192.8
L50 contigs	1,667
No of contigs > 100 kb	3,560
GC content	42.55%
% of the genome assembly hard-masked	9.4%
Number of scaffolds with annotated protein-coding genes	4,193
Number of protein coding genes	21,145
Avg. length of gene (bp)	23,321
Number of CDS exons	247,861
Number of introns	221,646

Supplementary Table 3. BUSCO and CEGMA analysis of the spoon-billed sandpiper genome and the chicken genome.

	Spoon-billed sandpiper BUSCO	Chicken BUSCO	Spoon-billed sandpiper CEGMA	Chicken CEGMA
Complete single-copy genes	320	325	221/248	213/248
Completely duplicated genes	4	7	N.A.	N.A.
Fragmented genes	17	12	9/248	9/248
Missing genes	92	92	18	26

Supplementary Table 4. GO Biological Processes for which SBS-accelerated regions were enriched.

GO.ID	Term	p-value
GO:0030307	positive regulation of cell growth	0.0014
GO:0022604	regulation of cell morphogenesis	0.0019
GO:0000902	cell morphogenesis	0.002
GO:0045773	positive regulation of axon extension	0.0022
GO:0032989	cellular component morphogenesis	0.0029
GO:0044706	multi-multicellular organism process	0.0035
GO:2000736	regulation of stem cell differentiation	0.0036
GO:0022412	cellular process involved in reproduction in multicellular organism	0.0047
GO:0045927	positive regulation of growth	0.0053
GO:0060560	developmental growth involved in morphogenesis	0.0058
GO:0014034	neural crest cell fate commitment	0.0064
GO:0015689	molybdate ion transport	0.0064
GO:0044345	stromal-epithelial cell signaling involved in prostate gland development	0.0064
GO:0044458	motile cilium assembly	0.0064
GO:0090246	convergent extension involved in somitogenesis	0.0064
GO:2000040	regulation of planar cell polarity pathway involved in axis elongation	0.0064
GO:2000041	negative regulation of planar cell polarity pathway involved in axis elongation	0.0064
GO:2000051	negative regulation of non-canonical Wnt signaling pathway	0.0064
GO:2000053	regulation of Wnt signaling pathway involved in dorsal/ventral axis specification	0.0064
GO:2000054	negative regulation of Wnt signaling pathway involved in dorsal/ventral axis specification	0.0064
GO:2000080	negative regulation of canonical Wnt signaling pathway involved in controlling type B pancreatic cell proliferation	0.0064
GO:0007286	spermatid development	0.0065
GO:0030516	regulation of axon extension	0.0065
GO:0030030	cell projection organization	0.0066
GO:0050772	positive regulation of axonogenesis	0.0072
GO:0060271	cilium morphogenesis	0.0077
GO:0030514	negative regulation of BMP signaling pathway	0.0079
GO:0048515	spermatid differentiation	0.0079
GO:0048468	cell development	0.0081
GO:0030258	lipid modification	0.0084
GO:0032504	multicellular organism reproduction	0.0085
GO:0051094	positive regulation of developmental process	0.0092
GO:0016358	dendrite development	0.0098
GO:0071559	response to transforming growth factor beta	0.0098
GO:0071560	cellular response to transforming growth factor beta stimulus	0.0098

Supplementary Table 5. GO Biological Processes for which RNS-accelerated regions were enriched.

GO.ID	Term	p-value
GO:0090100	positive regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	0.00044
GO:0090092	regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	0.00152
GO:0048535	lymph node development	0.00299
GO:0006040	amino sugar metabolic process	0.00429
GO:0050910	detection of mechanical stimulus involved in sensory perception of sound	0.00491
GO:0033014	tetrapyrrole biosynthetic process	0.0052
GO:0007076	mitotic chromosome condensation	0.00725
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	0.00725
GO:0018198	peptidyl-cysteine modification	0.00725
GO:0033013	tetrapyrrole metabolic process	0.0099

Supplementary Table 6. Regions of low level of heterozygosity found for spoon-billed sandpiper and red-necked stint individuals.

Individual	Scaffold	Region
C_pyg_11	scaffold16	2030000-2130000
C_pyg_13	scaffold106	1550000-1610000
C_pyg_13	scaffold26	2660000-3840000
C_pyg_13	scaffold31	40000-90000
C_pyg_18	scaffold54	4000000-4200000
C_pyg_22	scaffold16	7210000-7720000
C_pyg_22	scaffold168	1910000-2000000
C_pyg_22	scaffold44	990000-1340000
C_pyg_22	scaffold44	2000000-2230000
C_pyg_26	scaffold106	1540000-1610000
C_pyg_26	scaffold3	2690000-2920000
C_pyg_26	scaffold54	4110000-4170000

C_pyg_26	scaffold58	1620000-1710000
C_pyg_26	scaffold8	790000-840000
C_pyg_28	scaffold118	10000-300000
C_pyg_28	scaffold14	1370000-1610000
C_pyg_28	scaffold15	10650000-10870000
C_pyg_28	scaffold16	6300000-6460000
C_pyg_28	scaffold183	2520000-2700000
C_pyg_28	scaffold2	2300000-2390000
C_pyg_28	scaffold21	2500000-2550000
C_pyg_28	scaffold29	1410000-1600000
C_pyg_28	scaffold29	2620000-2690000
C_pyg_28	scaffold4	5900000-6100000
C_pyg_28	scaffold44	2530000-2580000
C_pyg_28	scaffold45	1610000-1680000
C_pyg_28	scaffold45	4080000-4360000
C_pyg_28	scaffold46	2650000-2820000
C_pyg_28	scaffold48	550000-720000
C_pyg_28	scaffold5	10730000-10850000
C_pyg_28	scaffold50	480000-770000
C_pyg_28	scaffold52	4110000-4210000
C_pyg_28	scaffold54	3900000-3970000
C_pyg_28	scaffold59	1380000-1610000
C_pyg_28	scaffold6	5770000-5910000
C_pyg_28	scaffold63	3210000-3590000
C_pyg_28	scaffold63	3830000-4180000
C_pyg_28	scaffold7	1210000-1450000
C_pyg_28	scaffold8	8420000-8580000
C_pyg_28	scaffold9	1060000-1120000

C_pyg_28	scaffold9	5250000-5340000
C_pyg_Ep	scaffold2	5960000-6290000
C_pyg_Ep	scaffold13	1140000-1210000
C_pyg_Ep	scaffold114	1200000-1520000
C_pyg_Ep	scaffold22	5360000-5530000
C_pyg_Ep	scaffold22	6280000-6430000
C_pyg_Ep	scaffold33	1720000-1950000
C_pyg_Ep	scaffold41	0-1300000
C_pyg_Ep	scaffold59	1470000-1830000
C_pyg_Ep	scaffold70	4420000-4500000
C_ruf_02	scaffold106	1540000-1610000
C_ruf_01	scaffold106	1540000-1610000
C_ruf_09	scaffold106	1540000-1610000

763

Supplementary Table 6. Accuracy of gene prediction**.**

Program/**param**	**SN**	**SP**	**SNe**	**SPe**	**SNg**	**SPg**
GeneID **vertb**	0.88	0.89	0.69	0.78	0.10	0.13
GeneID+intron **vertb**	**0.96**	0.95	0.87	**0.87**	0.24	0.29
SGP2 **ggallus/hs**	0.91	0.96	0.77	0.81	0.12	0.16
SGP2+intron **ggal/hs**	0.95	**0.97**	**0.88**	**0.87**	0.23	0.28
augustus+hints **ggallus**	0.93	0.95	0.83	0.86	**0.29**	**0.34**
augustus **ggallus**	0.90	0.95	0.77	0.84	0.16	0.23
GeneMark **hs**	0.91	0.59	0.64	0.45	0.07	0.02

Accuracy of gene prediction on an *C. pygmeus* artificial evaluation "scaffold" consisting of 323 concatenated *C. pygmeus* test sequences (with 800 nucleotides of sequence between each of the gene models) using the *ab initio* programs geneid (with a pre-existing vertebrate/mammalian-specific "vertb" parameter file), augustus (with a pre-existing chicken-specific "ggallus" parameter file) and GeneMark version 2 with their pre-existing *Mammalian/H. sapiens* parameter files (*i.e.* "hs"). The accuracy of SGP2 (homology evidence-based prediction tool that used the genome of *H. sapiens* as reference and previously optimized for prediction in the chicken genome) and that of augustus (using RNASeq and transcript evidence i.e. "augustus+hints") were also tested for accuracy on the same set of sequences. GeneID (GeneID+introns) and SGP2 (SGP2+introns) using introns as external intron evidence were also evaluated. (SN & SP: sensitivity & specificity at nucleotide level; SNe & SPe: sensitivity & specificity at exon level; SNg & SPg: sensitivity & specificity at gene level). The highest values of each of these accuracy parameters among all programs and matrices employed were **in bold.**

Supplementary Table 7. Weights used by EVM to create a consensus CDS model for *C. pygmeus.*

Type	Source	Weight
ABINITIO_PREDICTION	Augustus	1.5
ABINITIO_PREDICTION	AugustusHints	2
ABINITIO_PREDICTION	geneid	1
ABINITIO_PREDICTION	sgp2	1.5
ABINITIO_PREDICTION	geneid+introns	2.25
ABINITIO_PREDICTION	sgp2+introns	2.25
ABINITIO_PREDICTION	GeneMarkHMM	0.25
OTHER_PREDICTION	transdecoder	3
PROTEIN	SPALN2 uniRef90	6
PROTEIN	SPALN2 Ggallus	7
PROTEIN	SPALN2 uniprot-swissprot	7
PROTEIN	exonerate uniprot-swissprot	7
TRANSCRIPT	PASA	10

SPLAN2 against UniRef90 proteins; SPALN2 uniprot-swissprot: SPALN2 against *vertebrate* uniprot/swissprot curated proteins; Exonerate uniprot-swissprot: exonerate against against *vertebrate* uniprot/swissprot curated proteins; SPALN2 Ggallus: SPALN2 against GNOMON-generated NCBI chicken annotated genes.

Supplementary Table 8. Break down of the types of evidence used to build the 22,798 gene-model EVM consensus set (prior to filtering for weakly supported gene models).

Type of source of evidence	Number of consensus gene models supported by the type of source of evidence (% of total number of EVM reference gene models supported)
PASA transcript alignments	12,440 (54.57%)
Protein alignments	16,685 (73.19%)
Protein OR PASA alignments	18,330 (80.40%)
Protein AND PASA alignments	11,574 (50.77%)
Protein or PASA and at least one source of *ab initio* predictions	16,763 (73.53%)
Exclusively *ab initio* evidence (GeneID, geneidi,sgp2,sgp2i,augustus,augustushints or genemark)	4,467 (19.60%)
Only one source of *ab initio* predictions (No protein or transcript evidence)	3,184 (13.97%)
At least two sources of *ab initio* evidence (No protein or transcript evidence)	1,638 (7.18%)
All sources of *ab initio* evidence (No protein or transcript evidence)	553 (2.43%)
just GeneID,GeneID/spg2,sgp2i *(No protein or transcript evidence)	**1,130 (4.96%)**
just augustus or augustushints** (No protein or transcript evidence)	**514 (2.25%)**
just GeneMark*** (No protein or transcript evidence)	**1,423 (6.24%)**
singleEXON genes	4,997 (21.92%)

(*) We removed 355 *ab initio* GeneID-supported genes in scaffolds smaller than 2 Kbases when there was no other evidence for the EVM gene model. (**) We removed 4 *ab initio* augustus-supported genes in scaffolds smaller than 2 Kbases when there was no other evidence for the EVM gene model. (***) We removed 1,423 *ab initio* genemark-supported genes when there was no other evidence for the EVM gene model.

Supplementary Table 9. Statistics for the EVM-generated the protein-coding reference annotation for *C. pygmeus* (v1c).

Annotation versions	***Sandpiper v1c (EVM-generated)***
Genome length (Gbases)	**1.178**
% of the genome assembly hard-masked with RepeatMasker (complex repeats)	**9.4%**
number of scaffolds	**29,994**
number of scaffolds with annotations	**4,193**
Number of protein-coding genes	**21,145**
Gene density (genes/Kbase)	**0.018**
Number of protein-coding transcripts	**26,284**
Transcripts/gene (range) (% genes with more than 1 transcript)	**1 – 37 (12.55%)**
Avg. length of genes	**23,321.5 (SD 45,174.9)**
scaffolds with annotations that are smaller than the average gene length (%)	**3,062 (73.03%)**
Number of transcripts with UTRs	**13,400 (50.98%)**
Number of proteins	**25,015**
Number of complete proteins (%)	**20,618 (82.42%)**
Number/(%) proteins with similarity to sequences in the NCBI NR database (E=10-2; min. identity=25%)	**21,930 (88%)**
Avg. length of proteins	**532.69 aa. SD 592.86**
Avg. length of full-length proteins	**586.14 aa. SD 610.03**
Number of partial proteins (not starting with "M")	**3,109 (12.43%)**
Avg. length of partial proteins (not starting with "M")	**269.89 SD 416.07**
Number of partial proteins (no terminal STOP codon)	**3,119 (12.47%)**

Avg. length of partial proteins (no terminal STOP codon)	**223.65 SD 328.53**
Number of partial proteins (not starting with an M -and- no terminal STOP codon)	**1,831 (7.32%)**
Avg. length of partial proteins (not starting with an M -and- no terminal STOP codon)	**161.89 SD 202.34**
Number of partial proteins (not starting with an M -or- no terminal STOP codon)	**4,397 (17.58%)**
Avg. length of partial proteins (not starting with an M -or- no terminal STOP codon)	**282.06 SD 422.41**
Number of protein-coding CDS exons	**247,861**
Number of introns	**221,646**
Number of UTRs (spliced)	**28,793**
Number of single-exon genes	**4,003**
Number of multi-exonic transcripts (genes)	**22,281 (21,145)**
Exons/transcript (range) (excludes single-exon genes)	**10.95 SD 10.21 (2 – 286)**
Introns/transcript (range)	**9.95 SD 10.21 (1 – 285)**
“spliced” UTRs/transcript (range)	**2.15 SD 0.86 (1 - 7)**
Avg. length of introns (range)	**2,688.87 SD 7,266.45 (21 – 470,524)**
Avg. length of mono-exonic genes	**589.50 SD 546.04**
Avg. length of exons (excludes mono-exonic genes)	**163.29 SD 235.38**
Avg. length of first exons	**214.30 SD 357.25**
Avg. length of internal exons	**151.17 SD 196.27**
Avg. length of terminal exons	**220.75 SD 356.35**
Avg. length of CDS (range)	**1,604.7 SD 1,753.41 (126 – 98,418)**
Avg. length of UTRs (range)	**525.18 SD 886.83 (1 - 14,017)**

G+C content exonic (mono-exonic genes)	**51.22% SD 15.52%**
G+C content exonic (excludes mono-exonic genes)	**51.41% SD 8.49%**
G+C content exonic (first exons)	**53.14% SD 13.07%**
G+C content exonic (internal exons)	**49.20% SD 9.79%**
G+C content exonic (terminal exons)	**53.07% SD 13.07%**
G+C content intronic	**41.20% SD 12.00%**
G+C content genomic	**42.55%**
G+C content UTRs	**49.11% SD 18.48%**

794

II. SUPPLEMENTARY METHODS

1 Protein-coding gene annotation of the C. pygmeus genome

1.1 Obtaining GeneID and SGP2 gene predictions using pre-existing vertebrate/mammalian-specific parameter files

GeneID[88,89] is *an ab initio* gene prediction program used to find potential protein-coding genes in anonymous genomic sequences. In the context of GeneID training basically consists of computing position weight matrices (PWMs) or Markov models of order 1 for splice sites and start codons, and deriving a model of coding DNA (generally a Markov model of order 5). Furthermore, once a preliminary species-specific matrix is obtained it is further optimized by adjusting two internal matrix parameters: the cutoff of the scores of the predicted exons (eWF) and the ratio of signal to coding statistics information to be used (oWF). SGP2 [ref. 90] is a syntenic gene prediction tool that combines *ab initio* gene prediction (GeneID) with TBLASTX[65] searches between two or more genome sequences to provide both sensitive and specific gene predictions, usually showing an improvement to GeneID's performance, especially by reducing the number of false-positive predictions. SGP2 requires a reference genome to which the target genome (in this case *C. pygmeus*) is TBLASTX-compared. We decided to use the genome of *H. sapiens* (assembly hg38) as our "reference genome" based on previous studies that employed SGP2 to predict sequences in chicken using human homology evidence[91,92], and on the premise that *C. pygmeus* likely to be at an appropriate evolutionary distance from *H. sapiens* for optimal SGP2 performance. An "appropriate" evolutionary distance in the context of SGP2 means that the coding portions of the target/reference genomes being aligned would have a higher conservation than their intergenic and intronic regions. This feature would in turn contribute to a higher accuracy of SGP2 when compared to GeneID, especially at the level of specificity. SGP2 (using *H. sapiens* has the informant genome) has been previously used to successfully predict sequences in the chicken (*G. gallus*) genome[91,92]. To predict sequence in the sandpiper genome we used a modified version of *G.gallus*-optimized vertebrate-specific SGP2 matrix developed in the course of these two studies.

Geneid using its *mammalian/vertebrate*-specific parameter file has also been used in the past to accurately produce gene predictions in several different mammalian genomes[93,94]. The accuracy of the GeneID generic vertebrate parameter file and of the existing chicken-optimized SGP2 matrix for predicting sequences in *C. pygmeus* was tested on an artificial "scaffold" of 8.125 Mbases consisting of the 323 evaluation-set (*i.e.* not used in training) concatenated gene models with 800 nucleotides of intervening sequence between each of the genes (refer to **Supplementary Table 6**). This artificial test scaffold was built using one of the modules of the geneid training pipeline (*i.e.* geneidTrainer). The protein-coding gene sequences embedded into the artificial "evaluation" scaffold were previously selected from a set of sandpiper PASA pipeline[59] v2.0.2 – refer to supplementary section 2 for additional information – generated protein-coding transcripts longer than 600 nucleotides (200 amino acids) overlapping with

UniRef90 database (http://www.uniprot.org/) proteins over 100% of these sequences' length. The accuracy of GeneID (v1.4.4) and SGP2 (v1.0) (using *H. sapiens* as the reference genome) in predicting sandpiper sequences on the artificial evaluation scaffold is shown in **Supplementary Table 6**. The homology-based SGP2 is 3 and 8 % more sensitive, plus 7 and 3% more specific than *ab initio* GeneID at nucleotide and exon levels respectively.

GeneID and SGP2 can also use external information, such as the coordinates of known introns, to improve the accuracy of their gene predictions. In order to take advantage of this feature we 1) aligned all available-sandpiper derived RNASeq data (Epp6_5234) to the genome assembly (and evaluation scaffold) using the STAR RNASeq aligner[95]. Subsequently we 2) processed the alignment splice-junctions (SJ) file produced by STAR converting it into a list of potential introns in a format so that it can be used by GeneID and SGP2 as "external" evidence. 3) We then obtained *ab initio* GeneID (v1.4.4) and homology-derived SGP2 (v1.0) predictions on the genome assembly of *C. pygmeae* and used the BEDtools software suite[96] to filter out of the "intron" evidence file those introns not overlapping with the predictions generated by either program *not* using the external evidence (*i.e.* those likely to correspond to non-coding genes or simply generated from low-quality STAR alignments). Finally, we 4) re-ran both GeneID (GeneID and introns) and SGP2 (SGP2 and introns) on the evaluation "scaffold" created by the *geneidTrainer* pipeline as well as on the full *C. pygmeus* assembly, using the pre-processed "introns" file as external evidence. Importantly, we considered a predicted gene model as belonging to the evidence source "GeneID and intron" or "SGP2 and intron" *only* if introns were used as external evidence to build the gene models. This resulted in a smaller number of GeneID/SGP2 (and intron) gene predictions than GeneID/SGP2 models not using STAR aligner-derived intronic data. We then calculated the accuracy of GeneID and SGP2 (and introns) on the evaluation scaffold. Results show there's a significant improvement in the performance of either program when using introns as evidence. SGP2 (and introns) has 37% higher sensitivity and 17% higher specificity than SGP2 at nucleotide and exon levels, respectively. The advantage of using introns as external evidence is even more apparent with the program GeneID. GeneID (and introns) is 8 and 18% more sensitive plus 6 and 9% more specific than *ab initio* GeneID at the levels of nucleotides and exons respectively.

1.2 Obtaining augustus gene predictions using a pre-existing chicken-specific parameter file

We also produced *C. pygmeus* protein-coding gene annotations using the gene prediction tool augustus[97]. Augustus is a program that predicts genes in eukaryotic genomic sequences and is "re-trainable". The program is based on a Hidden Markov Model and integrates a number of known methods and submodels[97]. In order to predict gene sequences in *C. pygmeus* using augustus (v3.2) we used the program's pre-existing chicken parameter file. We estimated the program's accuracy in predicting *C. pygmeus* sequences by evaluating it on the same 8.125 Mbase artificial "scaffold" consisting of the 323 concatenated gene models with 800 nucleotides of intervening sequence used to evaluate the *in-house* GeneID and SGP2 programs (**Supplementary Table 6**).

We also took advantage of augustus' potential to use external evidence to improve its performance. We did this by obtaining a set of predictions that used the mammal/*H. sapiens* augustus parameter file in combination with PASA-derived transcript evidence obtained for this species (refer to supplementary section 2 for additional information). Briefly, for *C. pygmeus* this transcript evidence set consisted of 50,001 PASA transcripts built from an initial set of 1) 50,737 NCBI-derived mRNA sequences from two birds related to *C. pygmeus* (*C. pugnax* and *Charadrius vociferus*) and 2) 64,777 gene models generated running the program StringTie[98] on the RNASeq data previously aligned to the sandpiper assembly using the STAR RNASeq aligner. Refer to supplementary section 3 of this document for further details on how the PASA transcripts were generated.

Our strategy for taking advantage of the large set of transcripts followed the methodology described in (http://augustus.gobics.de/binaries/readme.rnaseq.html) and in Stanke *et al.* [ref. 97] and allowed us to obtain a higher-accuracy evidence-based set of Augustus(+hints) predictions on the *C. pygmeus* assembly (refer to **Supplementary Table 6**). In order for the *C. pygmeus* augustus matrix to take advantage of the external data we first had to optimize some internal parameters of the new augustus parameter file; the *exonpart* bonus for hints corresponding to PASA-evidence ("E") was given a bonus of $1xE^3$. Furthermore, complete *exons* predicted by augustus that perfectly matched the exons in the external hints were given a bonus of $1xE^4$. The *intron* bonus for (PASA) hints of source E was set to $1xE^4$, meaning that a predicted intron would get this bonus when being exactly as in the PASA "hint".

1.3 Obtaining GeneMarkHMM predictions using a pre-existing H. sapiens-specific matrix

Our final source of *ab initio* gene predictions to be used by the EVidence Modeler combiner[56] (EVM-1.1.1) was obtained using the program GeneMarkHMM[99]. The GeneMarkHMM algorithm was developed for finding protein-coding genes in eukaryotic genomes without using training sets consisting of gene models. GeneMarkHMM determines species-specific gene finding parameters using a self-training algorithm based on the species of interest genomic sequence. We attempted generating an *C. pygmeus*-specific matrix for this program using sandpiper genomic sequence by following the self-training instructions found both at the program developer's web site (http://opal.biology.gatech.edu/) and as indicated by Lomsadze *et al.* [ref. 100]. However self-training generated matrices with very low accuracy and therefore we decided to use a pre-built GeneMarkHMM *H. sapiens*-derived matrix developed using a prior version of the program (v2.2a). This GeneMarkHMM matrix was also evaluated on the same artificial scaffold used to evaluate other gene prediction parameter files used in this study (**Supplementary Table 6**).

These programs were then used as input to the combiner EVM, which was the program utilized to obtain the reference annotation for this genome.

2 EVM-based genome annotation of the C. pygmeus assembly by combining different sources of evidence using weights.

A combination of the Program to Assemble Spliced Alignments[59] (PASApipeline-2.0.2) and EVidence Modeler[56] (EVM-1.1.1) was used to obtain consensus coding sequence (CDS) models using three main sources of evidence: aligned transcripts, aligned proteins, and gene predictions (refer to supplementary section 1).

2.1 Gene predictions

GeneID/SGP2 (with or without introns), augustus (with or without "hints"), and GeneMark, using the parameter files described above were subsequently used to predict genes on the latest repeat-masked assembly of the *C. pygmeus* genome (*Sandpiper1.0; 2016/02/29)*. This assembly is made up of 29,994 scaffolds. Prior to the gene-finding step the assembly's complex-repeats were hard-masked using the program RepeatMasker[57] (v4.0.5) with its 20150807 *G. gallus*-derived library of repeat elements.

Given its vertebrate/mammalian-specific parameter file GeneID (v1.4.4) predicted 56,241 protein-coding genes without external evidence and 15,415 sequences when using intronic data (considering *only* those genes using intron data; refer to section 1.1). SGP2 (v1.0) generated 22,070 predictions on the assembly and 13,544 gene models when using introns as external evidence (SGP2+introns; considering *only* those gene models incorporating intron data; refer to section 1.1). The program augustus (v3.2) predicted 21,439 genes on the assembly while its evidence-based variation [augustus(+hints)] produced 24,335 gene models. GeneMark (v2.2a) with its human matrix generated 132,621 gene models on the genome assembly.

2.2 PASA transcript alignments

The C. pygmeus RNA sequences processed by the PASA pipeline were obtained as described in section 1.3. As previously mentioned this process resulted in 50,001 PASA transcript assemblies. The transcriptome was subsequently added to the PASA database (which uses GMAP/BLAT as the alignment engines). PASA was set to be quite stringent. Input sequences with less than 95% identity to the genomic sequence over 90% of their length were discarded.

The PASA pipeline contains a module, which we used to generate a potential training/evaluation used to train/test the different *ab initio* programs used in this study (refer to supplementary section 1.1). This module incorporates TransDecoder, a program that identifies candidate coding regions within transcript sequences based on a number of criteria: 1) a minimum length open reading frame (ORF); 2) a coding Markov log-likelihood score similar to what is computed by the program GeneID is > zero; 3) this coding score is greatest when the ORF is scored in the 1st reading frame as compared to scores of all other reading frames; 4) if a candidate ORF is found contained within another candidate ORF, the longer one is reported. TransDecoder found 36,359 sequences meeting the above criteria in *C. pygmeus*. This PASA TransDecoder alignment sub-set

945 comprising genes very likely to be protein-coding was used as a separate evidence source for the
946 gene combiner EVM.

947 *2.3 Protein alignments*

948 Furthermore, UniRef90 [17/02/2016], 17,892 *Uniprot-Swissprot* highly curated protein
949 vertebrate-derived sequences
950 (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/
951 uniprot_sprot_vertebrates.dat.gz; 17/03/2016) and 46,346 *G. gallus* NCBI GNOMON-generated
952 protein models (ftp://ftp.ncbi.nih.gov/genomes/Gallus_gallus/protein) were split-mapped to the
953 *C. pygmeus* genome by using the program SPALN2 [ref. 101] (v2.1.4) using its Tetrapoda-
954 specific parameters.

955 In addition to SPALN2 we used the spliced-protein alignment tool exonerate[102] (v2.2.0) to also
956 map the *Uniprot-Swissprot* protein sequences to the scaffolds of *C. pygmeus* (**Supplementary**
957 **Table 7**).

958 *2.4 Combining the different EVM sources*

959 The resulting alignments were then filtered as suggested in the EVM documentation
960 (http://evidencemodeler.sourceforge.net/).

961 Gene predictions were obtained as described in supplementary sections 1.1 - 1.3 and also
962 modified as recommended (http://evidencemodeler.sourceforge.net/)) and added to the EVM
963 pipeline. Subsequently the transcript alignments, protein alignments and the *ab initio* gene
964 models were combined into consensus CDS models by EVM using different weights. These
965 weights (shown in **Supplementary Table 7**) were selected following the EVM documentation
966 (*i.e.* http://evidencemodeler.sourceforge.net/) and, with regard to the *ab initio* predictions, also
967 based on the accuracy of the different programs in predicting sequences in the evaluation
968 *artificial* "scaffold" for this species (refer to supplementary section 1 and **Supplementary Table**
969 **6**).

970 We also determined which sources of evidence (*ab initio*, protein or transcript) EVM used to
971 build each of the 22,798 consensus gene models (**Supplementary Table 8**). In total 18,883
972 (82.83%) original consensus EVM reference annotation genes were built from
973 transcript/protein/transdecoder or alternatively, using all available gene prediction sources of
974 evidence (**Supplementary Table 8**). We proceeded to remove several of EVM consensus gene
975 models perceived to be of lower quality from the initial reference set. We filtered-out 1) the
976 1,423 reference gene models supported exclusively by GeneMarkHMM predictions (taking into
977 consideration the low accuracy of this prediction tool – refer to table 1) as well as the 2)
978 consensus gene models derived from predictions solely supported by geneid and augustus if
979 found in scaffolds smaller than 2 Kbases (355 and four respectively). This resulted in a final set
980 of 21,015 consensus CDS models which were then updated with UTRs and alternative exons
981 through five rounds of PASA's routine to update annotations. The resulting transcripts were

grouped into genes and then a pre-selected species-specific identifier was assigned to the genes, transcripts and protein products derived from them.

2.5 EVM consensus annotation statistics

Finally, and as a quality control, the protein products obtained from the reference annotation of this species was aligned against either the exhaustive NCBI non-redundant (NR-201512) database using the "protein vs. protein" BLASTP "flavor" of the sequence comparison tool BLAST ($E=10^{-2}$ with a minimum identity of 25%) to determine what percentage of the annotated genes matched a sequence of this large biological-sequence public databases. Results showed that 21,930 (~88%) of our consensus EVM reference of this species matched an NR protein given the criteria above (**Supplementary Table 9**). Furthermore, table 4 contains a wide-range of statistics obtained both from the *C. pygmeus* assembly and the consensus EVM protein-coding consensus protein-coding annotation. The PASA-updated EVM reference annotation set comprises 21,145 genes, 26,284 transcripts and 25,015 proteins.