Molecular Phylogenetics and Evolution 158 (2021) 107044

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Unraveling the history of the genus *Gallus* through whole genome sequencing

Mahendra Mariadassou [a,*], Marie Suez [a], Sanbadam Sathyakumar [b], Alain Vignal [c], Mariangela Arca [a], Pierre Nicolas [a], Thomas Faraut [c], Diane Esquerré [c,d,2], Masahide Nishibori [e], Agathe Vieaud [f], Chih-Feng Chen [g], Hung Manh Pham [h,1], Yannick Roman [i], Frédéric Hospital [f], Tatiana Zerjal [f], Xavier Rognon [f], Michèle Tixier-Boichard [f]

[a] *Université Paris Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France*
[b] *Wildlife Insitute of India, Chandrabani, Dehradun, India*
[c] *GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet Tolosan, France*
[d] *Get-PlaGe, INRAE, 31326 Castanet Tolosan, France*
[e] *Lab. of Animal Genetics, Department of Animal Life Science, Graduate School of Integrated Sciences for Life, Hiroshima University, Higashi-Hiroshima 739-8528, Japan*
[f] *Université Paris Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France*
[g] *Department of Animal Science, iEGG and Animal Biotechnology Center, National Chung-Hsing University, Taichung 40227, Taiwan*
[h] *Faculty of Animal Science, Vietnam National University of Agriculture, Trau Quy Town, Gia Lam District, Ha Noi City, Viet Nam*
[i] *Le Parc de Clères, 76690 Clères, France*

ARTICLE INFO

Keywords:
Chicken genomics
Hybridization
Introgression
Domestication

ABSTRACT

The genus *Gallus* is distributed across a large part of Southeast Asia and has received special interest because the domestic chicken, *Gallus gallus domesticus*, has spread all over the world and is a major protein source for humans. There are four species: the red junglefowl (*G. gallus*), the green junglefowl (*G. varius*), the Lafayette's junglefowl (*G. lafayettii*) and the grey junglefowl (*G. sonneratii*). The aim of this study is to reconstruct the history of these species by a whole genome sequencing approach and resolve inconsistencies between well supported topologies inferred using different data and methods.

Using deep sequencing, we identified over 35 million SNPs and reconstructed the phylogeny of the *Gallus* genus using both distance (BioNJ) and maximum likelihood (ML) methods. We observed discrepancies according to reconstruction methods and genomic components. The two most supported topologies were previously reported and were discriminated by using phylogenetic and gene flow analyses, based on ABBA statistics. Terminology fix requested by the deputy editor led to support a scenario with *G. gallus* as the earliest branching lineage of the *Gallus* genus, instead of *G. varius*. We discuss the probable causes for the discrepancy. A likely one is that *G. sonneratii* samples from parks or private collections are all recent hybrids, with roughly 10% of their autosomal genome originating from *G. gallus*. The removal of those regions is needed to provide reliable data, which was not done in previous studies. We took care of this and additionally included two wild *G. sonneratii* samples from India, showing no trace of introgression. This reinforces the importance of carefully selecting and validating samples and genomic components in phylogenomics.

1. Introduction

The domestic chicken, *Gallus gallus domesticus*, is the world's most ubiquitous bird and domestic animal; outnumbering humans by a factor of 3 to 1 (Lawler, 2014). Chickens being a cheap and easily available animal protein source, are important to human (Miao et al., 2013) and also play a significant role in religion, entertainment (e.g. cockfighting), ornamental breeding and in biomedical research (e.g. embryogenesis; Crawford, 1990). Domestication has led to an impressive diversification of chicken breeds with more than one thousand local chicken breeds

* Corresponding author.
E-mail address: mahendra.mariadassou@inrae.fr (M. Mariadassou).
[1] Present address: Private Company, Trau Quy Town, Gia Lam District, Ha Noi City, Viet Nam.
[2] Present address: DYNAFOR, Université de Toulouse, INRA, Castanet-Tolosan, France.

https://doi.org/10.1016/j.ympev.2020.107044
Received 19 June 2020; Received in revised form 23 October 2020; Accepted 14 December 2020
Available online 17 December 2020
1055-7903/© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

across the world (Malomane et al., 2019). Chickens were domesticated in Asia, where four species belonging to the genus *Gallus* are identified: the red junglefowl (*Gallus gallus*), the grey junglefowl (*Gallus sonneratii*), the Lafayette's junglefowl (*Gallus lafayettii*) and the green junglefowl (*Gallus varius*). *Gallus gallus* has the widest distribution area from southeast India and southern China to Indonesia, whereas the three other species occupy smaller areas, with *G. varius* in the Java Island and surrounding islands, *G. lafayettii* in Sri Lanka and *G. sonneratii* in a region spanning from central to south of India. *Gallus gallus* habitat has an overlap with *G. sonneratii* habitat in central east India, and with *G. varius* habitat on Java island. Main morphological differences between species involve the color and structure of plumage, comb and wattles (particularly in males). The phenotype of *G. gallus* is the closest to the one of the domestic chicken.

The red jungle fowl (*G. gallus*) is generally considered as the main ancestor of the domestic chicken based on mitochondrial analysis (Fumihito et al., 1994, 1996). However, possibility that different species of the *Gallus* genus may have contributed to genetic makeup of the domestic chicken is supported by different molecular studies (Eriksson et al., 2008; Lawal et al., 2020; Nishibori et al., 2005). In particular, the yellow skin mutation described in many chicken breeds is thought to have originated from the grey jungle fowl (*G. sonneratii*) (Eriksson et al., 2008). To date several DNA sources and strategies were used to resolve the relationship among species of the *Gallus* genus: mitochondrial DNA (mtDNA) (Fumihito et al., 1996; Kan et al., 2010; Meiklejohn et al., 2014; Nishibori et al., 2005; Shen and Dai, 2014; Shen et al., 2010; Stein et al., 2015), UCE (Ultra Conserved Elements) loci (Hosner et al., 2016), mtDNA and nuclear DNA (Kimball and Braun, 2014; Wang et al., 2013), whole genome data (Lawal et al., 2020; Tiley et al., 2020). These studies often produced discordant topologies that can be explained considering the use of different data sources and/or methods often based on different hypotheses.

The present study was undertaken to resolve these inconsistencies and to improve the knowledge of the history of the genus *Gallus*. A whole genome approach was used to reveal the phylogenetic histories of the different genomic components (autosomes, W chromosome and mtDNA) by comparing classical distance-matrix methods with more comprehensive methods such as the maximum likelihood, multispecies-coalescent and the ABBA-BABA statistics. Using whole genome sequencing of 26 *Gallus* individuals (16 wild *G. gallus* subspecies, 9 other *Gallus* species and 1 African village chicken) and one individual from the close relative *Bambusicola thoracicus* as an outgroup, we identified more than 35 million high quality biallelic SNPs that were used to reconstruct the genetic relationship of these birds. This has allowed us to propose a reference data set to map the ancestry of the domestic chickens.

2. Material and methods

2.1. Bird collection and sample choice for analysis

We collected 54 samples representing the 4 species (*G. gallus*, *G. varius*, *G. lafayettii* and *G. sonneratii*) of the genus *Gallus*. For *G. gallus*, the sampling involved the subspecies: *G. g. gallus*, *G. g. bankiva*, *G. g. spadiceus* and *G. g. murghi* but left out the subspecies *G. g. jabouillei*, restricted to Northern Vietnam and for which no specimen could be secured. They were obtained either from Zoological Parks located in France, Japan, Taiwan and Vietnam, or sampled in the forest of Northern Thailand in the frame of the AvianDiv collection (Hillel et al., 2003, Table S1). In the context of a collaboration agreement between INRAE and the Wildlife Institute of India (WII), WII independently collected 7 samples from wild Gallus in India, including 2 *G. sonneratii* (1 male, 1 female), 3 *G. gallus murghi* (males) and 2 *G. gallus spadiceus* (males). Finally, we added an African village chicken (Cameroun) to the dataset to represent a population which had not been submitted to organized management or selection for industrial purposes.

To validate our first batch of 54 samples, we performed a preliminary analysis by genotyping all our samples with the 60 K SNP Illumina iSelect chicken array (Groenen et al., 2011). The results confirmed the species assignment of the tested samples, since there were clearly 4 groups, one for each species (Fig. S1). A subset of 18 individuals was then selected for resequencing according to the following criteria: their position in the Neighbor-Joining tree for the 54 wild animals, the quality and quantity of DNA available for sequencing, and the sex, with a priority given to females in order to collect data on the W chromosome and to facilitate the phasing of haplotypes on Z chromosome. There were 16 females out of 18 samples retained.

The final set of samples included 16 *G. gallus* (with 4 *G. gallus gallus* (GGg), 7 *G. gallus spadiceus* (GGs), 2 *G. gallus bankiva* (GGb) and 3 *G. gallus murghi* (GGm)), 5 *G. sonneratii* (GS), 2 *G. lafayettii* (GL) and 2 *G. varius* (GV) (Table 1 and S1).

Finally, a *Bambusicola thoracicus* female sample provided by the Parc Zoologique de Clères, located in France, was sequenced to serve as an outgroup.

2.2. Library construction and DNA sequencing

High molecular weight genomic DNA was extracted from whole blood samples.

The samples from the AvianDiv collection and from the French zoological park were obtained as follows: hemolysis of 80 µl blood was done at 4 °C followed by incubation with 200 µg/mL proteinase K, precipitation with 4.5 mL dimethylformamide/acetone (5:95 vxol/vol), resuspension into TE buffer, and a second precipitation with 100% ethanol. Genomic DNA was finally resuspended in 2 mL of TE buffer. DNA samples from Japan were extracted with Phenol/Choloroform as described by Nishibori et al. (2005). DNA samples from Vietnam were extracted by the Bioneer kit (AccuPrep® Genomic DNA Extraction Kit) and stored at −20 °C before shipment. DNA samples from Taiwan were extracted with a commercial DNA extraction kit (DNeasy® Blood and Tissue kit) and diluted to 50 ng/µl. At WII, approximately, 500 µl of blood sample was collected and stored in DNA zol BD (Invitrogen™, Carlsbad, CA, USA) and genomic DNA was extracted following Mackey et al. (1998).

The DNA concentration was determined by spectrophotometer for each sample and the ratio of OD260/OD280 had to be above 1.8 for

Table 1
Bird sampling information.

ID	Species (subspecies)	Country	Place	sex
GGb_04992	*G. gallus bankiva*	France	zoological park	F
GGb_04322	*G. gallus bankiva*	France	zoological park	F
GGg_01072	*G. gallus gallus*	Thailand	forest, Chiang-Mai	F
GGg_02152	*G. gallus gallus*	Thailand	forest, Chiang-Mai	F
GGg_02172	*G. gallus gallus*	Thailand	forest, Chiang-Mai	F
GGg_01042	*G. gallus gallus*	Thailand	forest, Chiang-Mai	F
GGm_46	*G. gallus murghi*	India	Jammu & Kashmir	M
GGm_67	*G. gallus murghi*	India	Dehradun	M
GGm_120	*G. gallus murghi*	India	Uttar Pradesh	M
GGs_01112	*G. gallus spadiceus*	Thailand	forest, Chiang-Mai	F
GGs_01132	*G. gallus spadiceus*	Thailand	forest, Chiang-Mai	F
GGs_02211	*G. gallus spadiceus*	Thailand	forest, Chiang-Mai	M
GGs_03032	*G. gallus spadiceus*	Vietnam	zoological park	M
GGs_03082	*G. gallus spadiceus*	Vietnam	zoological park	F
GGs_114	*G. gallus spadiceus*	India	Nagaland	M
GGs_113	*G. gallus spadiceus*	India	Mizoram	M
GL_04252	*G. lafayetii*	France	zoological park	F
GL_04372	*G. lafayetii*	France	zoological park	F
GS_06012	*G. sonneratii*	Japan	zoological park	F
GS_04252	*G. sonneratii*	France	zoological park	F
GS_04572	*G. sonneratii*	France	zoological park	F
GS_113	*G. sonneratii*	India	Andhra Pradesh	M
GS_349	*G. sonneratii*	India	Andhra Pradesh	F
GV_05682	*G. varius*	Taiwan	zoological park	F
GV_06022	*G. varius*	Japan	zoological park	F
GGd_Cameroun	*G. gallus domesticus*	Cameroun	village	F

further processing. Prior to sequencing, the quality of the DNA samples was also checked with a gel picture and with Qubit fluorometer.

2.3. DNA sequencing

The 7 samples collected by WII were sequenced using paired-end sequencing (2x150 bp) on an Illumina NextSeq, with a mean coverage of 30X. The paired-end sequencing library was prepared using NEBNext Ultra DNA Library Preparation Kit. 200 ng g-DNA were fragmented by Covaris sonication. Covaris shearing generates dsDNA fragments with 3′ or 5′ overhangs. The fragments were then subjected to end-repair. This process converts the overhangs resulting from fragmentation into blunt ends using End Repair Mix. The 3′ to 5′ exonuclease activity of this mix removes the 3′ overhangs and the 5′ to 3′ polymerase activity fills in the 5′ overhangs. A single 'A' nucleotide is added to the 3′ ends of the blunt fragments to prevent them from ligating to one another during the adapter ligation reaction. A corresponding single 'T' nucleotide on the 3′ end of the adapter provides a complementary overhang for ligating the adapter to the fragment. This strategy ensures a low rate of chimera (concatenated template) formation. Indexing adapters were ligated to the ends of the DNA fragments, preparing them for hybridization onto a flow cell. The ligated products were purified using Ampure XP beads. The product was PCR amplified as described in the kit protocol. The amplified library was analyzed in Bioanalyzer 2100 (Agilent Technologies) using High Sensitivity (HS) DNA chip as per manufacturer's instructions.

A similar protocol was used for the 19 individuals collected by INRA except that the paired-end sequencing (2x100b) was performed on an 'Illumina HiSeq 2000/2500 with a target mean coverage of 25 $\sim$ 30x ($\sim$400 million reads per sample).

Finally the Chinese bamboo-partridge genome (*Bambusicola thoracicus*) was sequenced using three libraries (paired-end, 3 kb and 6 kb mate-pairs), each targeting a coverage of 50$\times$ on the same sequencing platforms.

Sequences have been deposited to NCBI-SRA (BioProject PRJNA552030).

2.4. SNP calling

Joint SNP calling for *Gallus* samples was performed following *GATK Best Practices* (DePristo et al., 2011) from the Broad Institute with default parameters. Briefly, reads were aligned to the *G. gallus* reference genome (galGal5 assembly) using bwa-mem (v 0.7.12) and filtered to remove PCR-duplicates and low quality reads (MAQ $<$ 20). Reads were then locally realigned around indels and base quality scores recalibrated using a set of 600,000 high quality SNPs from the 600 K SNP chip as known variants (Kranis et al., 2013). Variants were then detected using a pair-HMM model (HaplotypeCaller). The previous set of 600 K known variants was then used to train a learning model, with the Variant Quality Score Recalibration (VQSR) tool, and classify detected variants into true and false positives. False positives were filtered with option "ts_filter_level = 99.0", corresponding to a recall rate of 99% for known SNPs, and only biallelic SNPs were kept.

For the *B. thoracicus* sample, we used a more complex, three steps strategy. Reads were first mapped against the chicken genome to partition them according to the chicken chromosome they had highest sequence similarity with. More than 95% of the reads aligned with identity 95% or higher against a chromosome. *Bambusicola* reads were then assembled *de novo* using the MaSuRCA genome assembler (Zimin et al., 2013) with default parameters within each partition. The resulting assembly totalizes 920 Mb. of sequences assembled into 1035 scaffolds ($>$10 kb) with a N50 scaffold length of 2.5 Mb. Finally, one-to-one alignments between the *B. thoracicus* and *G. gallus* genomes were obtained using LAST software (Frith and Kawaguchi, 2015; Kiełbasa et al., 2011) and all nucleotides within 5 bp of an alignment gap were discarded. The resulting alignments covered 92.5% of the *G. gallus* reference (chromosomes only) with 95.6% identity. Finally, substitutions were identified based on these alignments, resulting in a total of 41,150,984 SNPs between *G. gallus* and *B. thoracicus*. We used the assembly-based approach instead of the direct approach because the two species are more divergent: mapping quality is sufficient for read partitioning but below typical scores observed for *Gallus* samples. The assembly step also allowed us to take advantage of the mate pair libraries to limit the impact of structural variations, expected to be higher between the *B. thoracicus* and *G. gallus* genomes than between two *Gallus* genomes, on SNP calling.

2.5. Additional mitochondrial sequences

Sixty-five additional mitochondrial sequences (from *G. gallus*, *G. sonneratii* and *B. thoracicus*) were extracted from Nishibori et al. (2005), Miao et al. (2013) and Meiklejohn et al. (2014). These studies performed sequence assembly to reconstruct full length mitochondrial sequences whereas we used a genotyping by sequencing approach and kept only SNPs with very high quality score. Mitochondrial sequence of *B. thoracicus* was produced by Shen et al. (2009). Accession numbers are shown in Table S2.

Mitochondrial sequences for our 26 *Gallus* samples were obtained by projecting the mitochondrial SNPs found in the previous step to the mitochondrial sequence of the *G. gallus* reference genome (galGal5 assembly).

To make sequences comparable between studies and avoid bias in genetic distances, we aligned all sequences against the *G. gallus* reference using MUSCLE aligner (Edgar, 2004) and used only positions corresponding to a variant detected using the mapping approach (n = 810 nucleotides) to compute the haplotype network and the phylogenetic tree.

2.6. Introgressed regions

Recent genetic flow from *G. gallus* to any of the other wild species was assessed using 20 kb sliding windows as shown in Fig. 1. For example, for *G. sonneratii* birds, we computed for every bird in every window (i) the local divergence to the *G. gallus* population (d_G), (ii) the local divergence to the *G. sonneratii* population (d_S), (iii) the ratio $r = d_S/(d_S + d_G)$, (iv) the density of heterozygous SNPs (normalized by its average in the *G. sonneratii* birds) (Het) and (v) quality metrics (mapping coverage, number of typed snps).

Windows with low quality metrics (coverage $<$12x, less than 100 genotyped SNPs) were discarded, as they could not reliably be classified as introgressed or not. Windows were then classified as Introgressed whenever the 'r' ratio was either high ($r > 0.5$) or moderately high with high local heterozygosity ($r > 0.38$, Het > 1.5) or Normal otherwise. The values used as thresholds were based on the empirical distribution of (r, Het) in the 5 *G. Sonneratii* samples (Fig. S6). Three types of windows were expected according to the number of *G. Gallus* haplotypes found in each window: pure *G. sonneratii* (two *G. sonneratii* haplotypes), mixed (one haplotype of each species) and pure *G. gallus* (two *G. gallus* haplotypes).

Contiguous windows of the same type were then merged into regions, and adjacent or nearly adjacent regions (less than 1 Mb apart) were manually inspected to assess whether or not they should be merged. In cases when these adjacent regions were separated by a stretch of low-coverage sequence, they were merged.

Finally introgressed regions were further classified based on their normalized heterozygosity as *mixed*, or *heterozygous* (Het > 1) if only one *G. gallus* haplotype was introgressed and *pure gallus*, or *homozygous* (Het <1) if the two *G. gallus* haplotypes were introgressed. The threshold value was once again chosen empirically based on the distribution of 'Het' in introgressed regions of introgressed samples (Fig. S7).

The same analyses were also performed on *Gallus varius* and *Gallus lafayettii* samples to find introgressions from *Gallus gallus*.

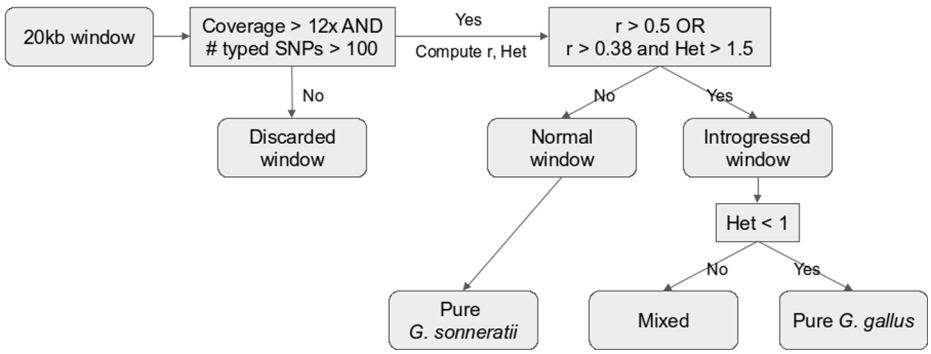

Fig. 1. Workflow for detecting introgressed regions. The various thresholds used are justified by the emprical distribution of windows in the (r, Het) space which shows a clear trimodal distribution.

2.7. Phylogenetic analyses

We considered three distinct genomic components for phylogenetic analyses mitochondrial DNA, W chromosome and autosomes. Due to the haploid (female) or diploid (male) status of the Z chromosome in our dataset, it was discarded in the phylogenetic analyses. Sequence alignments were created by projecting SNPs on the *G. gallus* reference genome (galGal5 assembly) and include both variant and invariant sites. For the autosomal compartment, we removed *G. sonneratii* sequences detected as introgressed from *G. gallus* (see Introgressed Regions) and replaced them with N prior to inference. Since the W chromosome and the mitochondrial are transmitted as a single block, we conserved them in introgressed samples.

2.8. Distance-based method

Genetic distances were computed using Nei's D_{XY} distance measure (Nei, 1987) with the R software. The D_{XY} distance captures the average number of differences between sequences from two birds (or a bird and a population) at a random locus. A drawback of Nei's distance for cross-species studies is its inability to distinguish between alleles shared through identity by descent and through homoplasy. We nevertheless used it here since (i) it accommodates diploid samples, (ii) the genetic distances considered here are quite small (<1% between any two *Gallus* samples) so that homoplasy is unlikely to be a problem and (iii) to make our results comparable to Lawal et al. (2020). Distances were computed in a pairwise fashion, meaning that sites were ignored only if they were missing in one of the two samples being compared.

Phylogenetic trees were reconstructed using BioNJ (Gascuel, 1997) on the pairwise genetic distance matrices. Bootstrap values were computed over 100 replicates using either standard bootstrap, for the mitochondrial and the W tree, or block bootstrap, with 20 kb blocks, for the autosomal tree.

2.9. Maximum likelihood method

We used the annotation of galGal5 to select the regions corresponding to 20,946 gene sequences, genomic region from the start of the gene to its end (including the 3′ and 5′ UTR) as defined by the genome annotation provided by NCBI (annotation 103) and hereafter called GS, in all chromosomes. We reconstructed a multiple sequence alignment for each GS by extracting the consensus sequence of each bird from the VCF file. We then discarded all GS with (i) less than 100 SNPs, (ii) less than 30 of those SNPs genotyped as N for *Bambusicola* (for example due to low coverage) or (iii) a genetic divergence between galGal5 and *B. thoracicus* lower than 3.5% (chosen based on the genome-wide average divergence of 4.4%). The later two criteria were used to remove GS that had no equivalent in Bambusicola or were atypical in terms of divergence between *Gallus-Bambusicola*, and thus rooting with an outgroup might fail. We ended up with 10,574 GS. For all genomic compartments (mtDNA, W chromosome and autosomes), we considered a supermatrix approach by concatenating all GS to create a superalignment. For the autosomes, we performed an additional supertree approach. Our assumption for the use of GS is that all exons and introns in a gene share the same evolutionary history. We are aware that intra-locus recombination may break that assumption for some genes but it allows us to keep as many sites as possible for a single gene and increase the accuracy of gene tree estimation.

For the autosomal tree, RAxML was run once on the unpartitioned matrix (n = 420,305,533 sites) with model GTR + CAT and 100 bootstrap replicates. We also ran RAxML on the partitioned matrix (one partition per GS) using the same model. In parallel, we used a supertree approach to infer the species tree from single gene trees using a multi-species coalescent model. A gene tree was inferred on each of the 10,574 autosomal GS using RAxML (v. 8.2.11) (Stamatakis, 2014) with GTR + Γmodel and 100 bootstrap replicates (-m GTRGAMMA -f aT -N 100 -p 1234 -x 1234). The inferred gene trees were then fed to Astral III (v. 5.6.3) (Zhang et al., 2018). Astral III was executed once on gene trees, as inferred by RAxML, and once after collapsing the low support branches (bootstrap <= 70%) of those trees.

The tree of W chromosome was reconstructed using the same supermatrix strategy: RAxML with model GTR + Γon the unpartitioned matrix of the GS (n = 2,066,132 sites). We used GTR + Γinstead of GTR + CAT because of the low number of alignment patterns (only 732 different patterns) making it hard to accurately fit a GTR + CAT model.

The tree of the mtDNA genome was reconstructed using RAxML with model GTR + Γ. The alignment was obtained by projecting our 26 samples and the additional 65 mitochondrial sequences to the reference genome on the 810 SNPs detected with the mapping approach to obtain full length sequences, as explained in Section 2.5. GTR + Γ was used instead of GTR + CAT because of the low number of sites in that alignment (n = 16,813).

2.10. Mitochondrial haplotype network

A haplotype network for the mitochondrial genome was reconstructed for all *G. gallus* and the *G. sonneratii* from zoological parks, using the HaploNet function from the Pegas R package (Paradis, 2010).

2.11. ABBA statistics and gene flow

Gene flow between the different gallus species was evaluated using the f_d (Martin et al., 2014) statistic on all SNPs with no missing value and MAF > 5% (n = 23,276,367). f_d was used instead of the original Patterson's *D*-statistic (Green et al., 2010) as it is more robust to variation in nucleotide diversity across the genome. f_d evaluates unidirectional gene flow from A to C by evaluating the number of SNPs with pattern *ABBA* and *BABA* in species related by the tree (((Pop1, Pop2), Pop3), O) where

O is an outgroup, here *B. thoracicus*. Due to uncertainties in the phylogenetic tree of the Gallus species, f_d was evaluated for the two scenarios obtained from the two phylogenetic methods tested. For each gene flow, the variance of f_d was computed using block-jackknife with blocks of size 5 Mbp.

2.12. Genetic differentiation of G. gallus subspecies

F_{ST} for all pairs of subpecies were computed on the autosomal genome using PopGenome (Pfeifer et al., 2014).

SNP of the *G. gallus* samples were extracted and filtered according to the following criteria: (i) at most 5% of missing data, (ii) MAF > 5%. The SNPs were then clumped to keep only SNPs with weak Linkage Disequilibrium: we forced all pairs of SNPs within 500 bp of each other in the subset to have $R^2 \leq 0.2$. The 12,496,171 remaining SNPs were then used to examine the genetic structure of the subspecies using PCA with bigsnpr (Privé et al., 2018).

2.13. Demographic history

Effective population sizes were estimated backward in time based on individual whole genome sequences under the Pairwise Sequential Markovian Coalescent (psmc) model (Li and Durbin, 2011). Briefly, the analyzed individual whole genome sequences were obtained from the bam files following the requirements of psmc, i.e creating mpileup files that were processed (individually) with bcftools using the –c option before being passed on to the vcf2fq utility of samtools, discarding sites covered by less than 10 or more than 60 reads (-d 10 -D 60 options). The resulting fastq files were then converted to psmcfa files, a reduced version of (autosomal) genomes, using the fq2psmcfa utility of psmc. Consecutive sites were grouped into consecutive bins of 30 nucleotides and marked as "K" (at least one heterozygote), "T" (no heterozygote site) or "N" (less than 27 called sites). Effective population sizes were estimated from psmcfa files using psmc with default parameters except for the pattern of atomic time intervals (-p option) that was set to "4 + 50 * 1 + 4 + 6". This pattern was chosen to have low resolution for both the recent and distant past, as demographic history is notoriously difficult to infer in those zones. Finally, psmc scales all quantities to $2N_0$. To transform them back to effective sizes and real time (in years), we considered a generation time of $T = 1$ year and a mutation rate per year of $\mu = 1.91 \times 10^{-9}$ (Nam et al., 2010).

In addition to the psmc analysis, we also estimated population sizes using smc++ (Terhorst et al., 2017) based on unphased whole-genome sequence data. Unlike psmc, smc++ can analyze genotyped sequences from multiple individuals at the same time. Briefly, the population level data of autosomal chromosomes were obtained from the vcf file using the vcf2smc utility of smc++. To avoid spurious runs of homozygosity caused by large uncalled regions, we masked all regions where the median coverage was lower than 7x using the -m option. We then ran smc++ with default parameters and a per generation mutation rate of $\mu = 1.91 \times 10^{-9}$ (Nam et al., 2010), like in the psmc analysis. Finally, we used smc++ in split mode to estimate divergence times between *G. sonneratii*, *G. lafayettii* and *G. gallus*. *G. varius* was left out of split-analyses as there are only two samples including the one with lowest coverage and highest fraction of missing data.

3. Results

3.1. Sequencing data

The average sequencing depth was 34.5X (20.5–41.6). We obtained 34X coverage on all autosomes, 68X for the mitochondria, 20X for the Z chromosome and 7X for the W chromosome (Table S1). We identified more than 35 million high quality biallelic SNPs. The total number of SNPs is shown per species and per genomic component in Table 2 and the number of segregating sites for each species can be found in Table S3. The fraction of missing (non-genotyped due to insufficient depth) SNPs varied from 0.029% to 3.995%. The only sample exceeding 0.5% of missing data was a *G. varius*, which also showed the lowest depth (20.5X, Table S1).

Table 2
Number of samples and SNPs in each Gallus species for autosomes, chromosomes W and Z and mitochondrial genome.

Species	Size	Auto (Snps)	W (Snps)	Z (Snps)	MT (Snps)
G. varius	2	9,954,704	3,252	695,496	213
G. lafayettii	2	8,890,248	3,886	635,316	375
G. sonneratii	5	11,743,067	4,305	718,759	402
G. gallus	16	17,463,031	1,904	879,872	127
All	25	33,552,454	8,625	2,028,872	810

The SNP density varied from ~ 3.64/100 bps for autosomes to 4.83/100 bps for the mitochondria. The 20 kb windows contained on average 728 ± 203 SNPs for the autosomes, 505 ± 191 in the Z chromosome Z and 46.9 ± 38.6 in the W. Most windows had at least 93 SNPs.

The high number of high quality SNPs and their high density along the genome (~3.64/100 bp in the autosomes and >0.46/100 bp in 99% of the 20 kb windows) justifies both the mapping strategies and the computation of local statistics on sliding windows to detect local introgressions.

The assembly of *B. thoracicus* reads and alignment of resulting scaffolds against the *G. gallus* reference ended up with a 92.5% coverage of the reference with 95.6% identity. Among the 41,150,984 SNPs found between galGal5 and *B. thoracicus*, 1,803,275 were also variable in at least one *Gallus* sample.

The high coverage and identity of *B. thoracicus* scaffolds aligned against galGal5 also justify the pseudo-mapping strategy to find variants and their coordinates in the reference galGal5, although some regions were poorly covered (only 7 positions for the mitochondrial genome) by that strategy.

The high nucleotide divergence of *B. thoracicus* with *G. gallus* on the covered regions (~4.4%) is in line with its position relative to the *Gallus* genus in the tree of galliforms (Hosner et al., 2016).

3.2. Phylogenetic analysis

3.2.1. Distance-based method

The neighbor joining tree obtained from the autosomal data (Fig. 2a) revealed a first separation of *G. varius* from the other species and a second separation that divided the *G. sonneratii* and *G. lafayettii* from the *G. gallus*. All nodes were well supported with bootstrap values (BP) of 100% and each *Gallus* species formed a monophyletic clade. The three *G. sonneratii* samples issued from zoological parks clustered with the wild ones. The phylogenetic trees of the W chromosome (Fig. 2b) and of the whole mitochondrial genome (Fig. 2c) both differed from the autosomal tree with the *G. varius* branching with the *G. gallus* cluster (100% BP). The trees revealed also that in our dataset the three *G. sonneratii* samples collected in zoological parks were clustered with the *G. gallus*, while the wild *G. sonneratii* birds branch on a separate clade.

3.2.2. Maximum-Likelihood method

The ML phylogenetic analyses of the autosomal genome (i.e. partitioned and unpartitioned analyses of the supermatrix and supertree of the gene trees) all resulted in the same species-level topology and we only show results from the unpartitioned RAxML analysis (Fig. 2d). The phylogeny showed a separation of the G. gallus from the other three species among which the *G. lafayettii* and *G. sonneratii* appeared as sister species and *G. varius* as a more distant relative. Each *Gallus* species formed a monophyletic clade and inner branches had extremely high non-parametric boostrap values (>99).

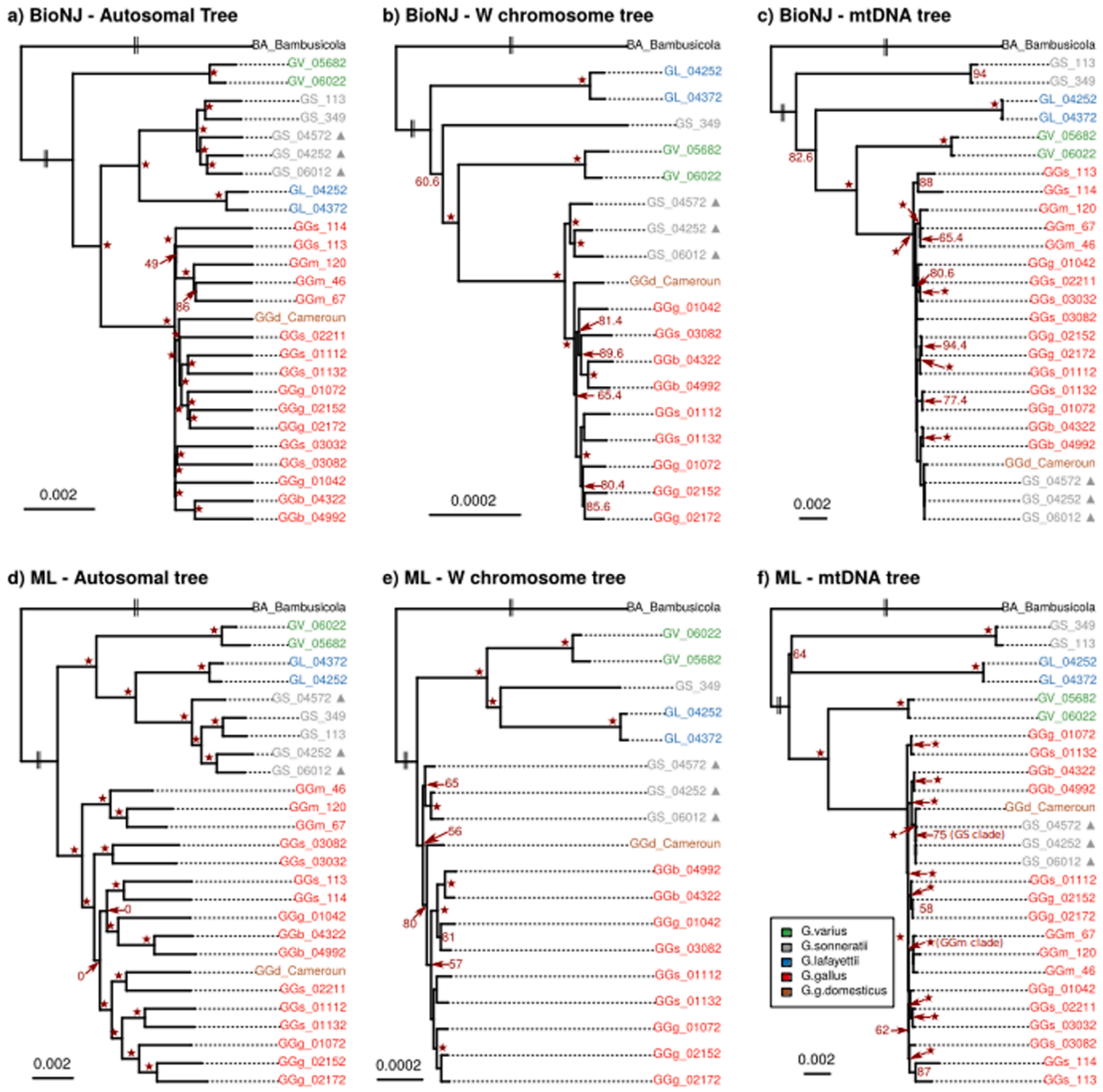

Fig. 2. Phylogenetic tree of the autosomal genome (left), the W chromosome (middle) and the mitochondrial genome (right) reconstructed using distance-based (top, BioNJ) and maximum likelihood (bottom, RAxML with GTR + CAT for the autosomal tree and GTR + GAMMA for the others) methods. All trees are rooted using Bambusicola as an outgroup. G. sonneratii individuals from zoological parks are highlighted with a triangle (▲). Bootstrap values higher than 95% are replaced by a star (★) and values lower than 50% are omitted. Branches marked with ‖ were shortened for legibility purposes. Regions identified as introgressed (see text for details) were removed prior to tree construction.

The W chromosome phylogenetic tree (Fig. 2e) also identified the separation of the *G. gallus* from the other three species with the exception that the three *G. sonneratii* samples from zoological parks clustered together with *G. gallus* (91% BP).

The mitochondrial tree (Fig. 2f) showed a different topology from the autosomal tree with the first node separating the wild *G. sonneratii* and *G. lafayettii* from the *G. varius* and *G. gallus*. The three *G. sonneratii* samples from zoological parks formed instead a clade with *G. gallus*, in accordance to what observed for the W chromosome tree.

3.2.3. Haplotype network

To further explore this discrepancy among the *G. sonneratii* samples, we reconstructed the mitochondrial haplotype. The network (Fig. 3) revealed that the most frequent haplotype cluster among domestic chickens also contained all the *G. sonneratii* sequences.

3.3. Gene flows between gallus species

We estimated the gene flow among species for the topologies obtained with the two phylogenetic methods: the BioNJ one (*B. thoracicus*, (*G. varius*, (*G. gallus*, (*G. sonneratii*, *G. lafayettii*)))) and the ML one (*B. thoracicus*, (*G. gallus*, (*G. varius*, (*G. sonneratii*, *G. lafayettii*)))). Results are summarized in Table S5 and Fig. 4. Both topologies suggested direct gene flow from *G. gallus* and *G. varius* to *G. lafayetti* (1.8% and 2.8% respectively). In addition, the ML topology (*G. gallus* basal) suggested a gene flow of 2.7% from *G. gallus* to the common ancestor of *G. lafayetti* and *G. sonneratii* whereas the BioNJ (*G. varius* basal) topology suggested a gene flow of 27.6% from *G. varius* to the same branch.

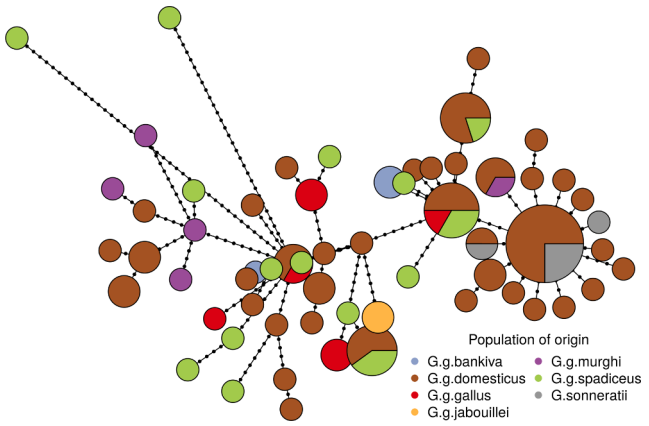

Fig. 3. Haplotype network of the mtDNA of G. gallus birds (including additional mtDNA from the literature) and G. sonneratii birds from zoological parks. Each circle corresponds to a haplotype. Haplotype surface is proportional to the number of sequences in that haplotype, and the color pie corresponds to population of origin of those sequences. Edge lengths are proportional to the number of mutation between haplotypes.

3.4. Analyses of autosomal introgression from G. gallus to the other species

We analyzed 20 kb sliding windows to look for local introgression. We identified three types of introgressed regions (example in Fig. 5). The first (top panel) were the mixed regions (one *G. gallus* haplotype and one *G. sonneratii* haplotype) characterized by a very high density of heterozygous SNPs and similar local distances to both *G. gallus* and *G. sonneratii*. The second (middle panel) were the pure *G. gallus* regions (two *G. gallus* haplotypes) characterized by a very low density of heterozygous SNPs and genetic distance to G. gallus lower than or similar to *G. sonneratii*. The third (bottom panel) were the composite regions, characterized by a succession of mixed and pure regions.

Dozens of Mb-long introgressed regions were identified in the *G. sonneratii* samples from zoological parks (Table 3, Table S4 and Fig. S4). Detected introgressed regions contribute between 6.5 and 13.4% of the total genome of these samples.

Two regions were introgressed in all three G. sonneratii individuals from parks: Chr4:46,540,001–46,880,000 (Region 1) and Chr19:2,200,001–2,700,000 (Region 2). Region 1 overlaps 20 genes with unknown functions whereas Region 2 overlaps no gene. Region 1 is pure *G. gallus*, with no heterozygote SNPs and almost identical sequences in individuals GS_06012 and GS_04252. The corresponding haplotypes have an average nucleotide diversity of 5.81×10^{-7}. Assuming a mutation rate of $\mu = 1.91 \times 10^{-9}$ per site per year (Nam et al., 2010), the haplotypes started diverging 150 years ago.

Nine short regions were flagged as introgressed in the two wild *G. sonneratii samples*. They are all small in size (all but one <160 kb) and contribute less than 0.05% of the total genome of GS_113 and GS_349. Two of the nine regions found in GS_113 (Chr4:46,540,001–46,640,000 and Chr4:46,780,001–46,880,000) fall inside Region 1 mentioned above. This can be explained considering that the consensus *G. sonneratii* sequence used to compute the summary is heavily biased in this region by the triple introgression to GS_06012, GS_04572 and GS_04252. Considering their size distribution, and the number of windows analyzed, the other regions can be considered as false positive and do not constitute evidence of recent gene flow from *G. gallus* to GS_113 and GS_349.

Likewise, no evidence of recent genetic flow from *G. gallus* to *G. varius* or *G. lafayettii* birds conserved in zoological parks was found.

3.5. Nucleotide divergence within the genus gallus

Upon removal of introgressed regions in *G. sonneratii* individuals from zoological parks, the different species exhibited mean heterozygosity (Ho) and average nucleotide diversity (π) ranging from 0.08% to 0.35% and from 0.1% to 0.4%, respectively (Table 4). *G. lafayettii* and *G. varius* exhibited the lowest diversity and the unselected domestic Cameroun chicken had similar genetic diversity to its wild *G. gallus* counterparts.

The autosomal nucleotide divergence between galGal5 and the three *Gallus* species (*lafayettii*, *sonneratii* and *varius*) ranged from 0.82% for *G. sonneratii* to 0.95% for *G. varius*. It was markedly higher in all cases than the intra-species genetic diversities (as measured by π).

Finally, the nucleotide divergence was 9.5 times smaller for W chromosome than for autosomes and 2.2 smaller for autosomes than for the mitochondrial genome.

3.6. Demographic history

The demographic histories reconstructed using psmc analysis (Fig. S5) showed similar trends for all individuals of the same species, with one exception for the *G. gallus* where the subspecies *G. g. murghi* did not show the same pattern as the other subspecies. The population sizes estimated by PSMC for the different species (Fig. S5) converged toward each other around 2–5 MYA.

At the species level, smc++ identified pronounced differences among species (Fig. 6) with *G. gallus* displaying a 3 to 5 folds higher effective population size than the other species. The result also suggested a much smaller population size for *G. varius* compared to *G. gallus*. Population sizes fluctuation did not appear to be related with recent glaciation periods, except maybe for *G. lafayettii* during the last two ice ages (Fig. 6). Steep increase (viewed backward in time) in population size for *G. gallus* were observed starting at 30 KYA reaching a maximum around 400 KYA with 900 000 individuals before decreasing. *G. sonneratii* and *G. varius* experienced a similar dynamic, with overall moderate changes in effective population sizes throughout time. The *G. lafayettii* is the only species with fluctuating population sizes with two

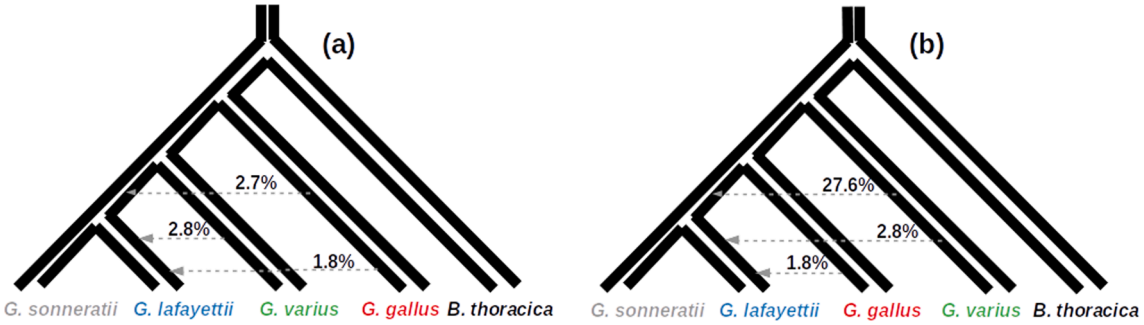

Fig. 4. Unidirectional introgression levels estimated using the statistic on the two competing topologies of the Gallus genus, obtained following (a) maximum likelihood and (b) distance-based methods.

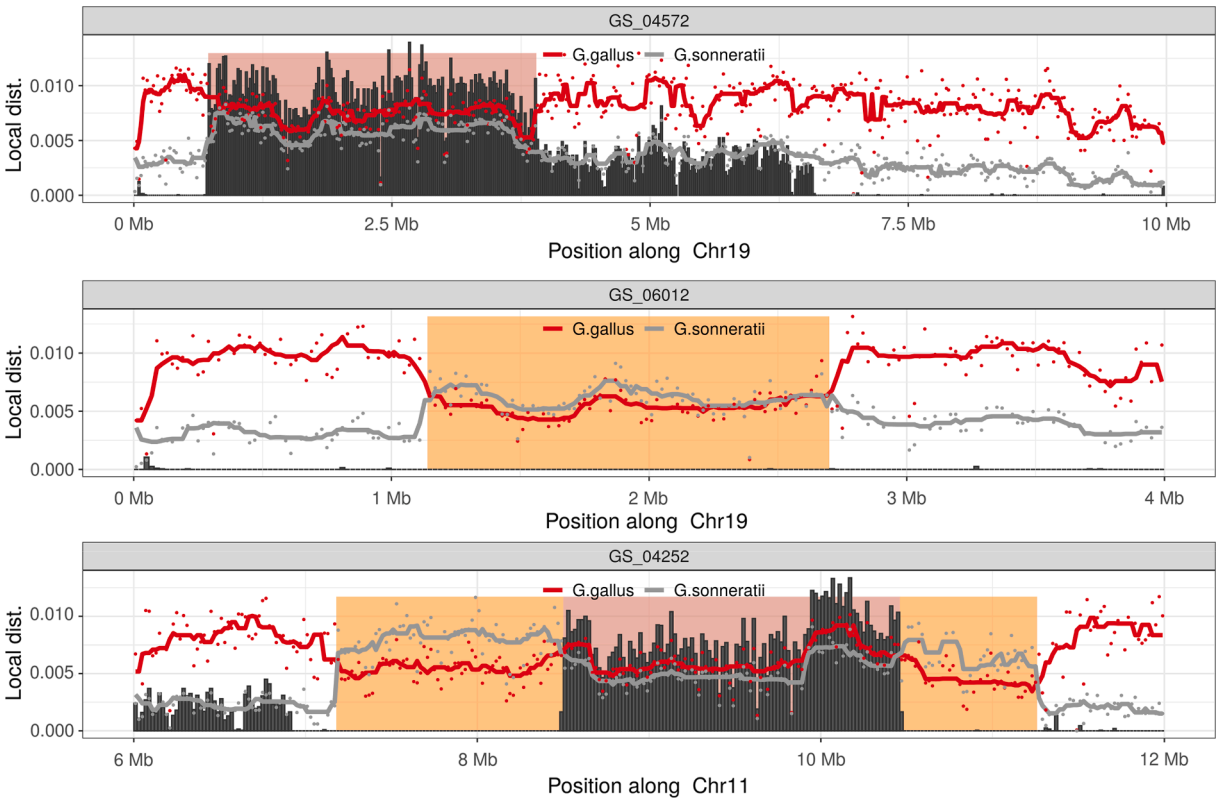

Fig. 5. Examples of a mixed (top), pure G. gallus (middle) and composite (bottom) introgressed regions. Red and grey points correspond to distances to either G. gallus (red) or G. sonneratii (grey) populations, and thick lines to smoothed version of those (running medians). Black bars are scaled density of heterozygote SNPs. Finally, shaded regions correspond to mixed (pale red) or pure G. gallus (pale orange) introgressed regions. All quantities are computed on non-overlapping 20 kb windows. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Summary of introgressed regions in G. sonneratii samples. ID: animal identifier, Fraction: percentage of introgressed genome, Count: number of introgressed regions, Total: total length of introgressed regions, Mean: average introgressed region size.

ID	Fraction (%)	Count	Total	Mean
GS_04252	13.4	37	139,596,603	3,772,881
GS_04572	9.17	29	96,916,775	3,341,958
GS_06012	6.52	24	68,216,775	2,842,366
GS_113	<0.1	5	560,000	112,000
GS_349	<0.1	4	840,000	210,000

episodes of population size increase at 36 KYA and 190 KYA.

The split analysis (Fig. S6) suggested divergence time as recent as 300 KYA between *G. sonneratii* and *G. lafayettii*, and 900 KYA between *G. sonneratii* and *G. gallus*.

3.7. Genetic structure in gallus gallus subspecies.

The PCA showed that the first axis (17.6%) separated the *G. g murghi* from the others while the second axis (15%) separated *G. g. bankiva* from the others (Fig. 7). No clear separation was observed between *G. g. gallus* and *G. g. spadiceus*. The values observed between all pairs of subspecies (Fig. 7) ranged from 0.04 to 0.30. The subspecies F_{ST} pairwise values confirmed the PCA results with *G. g. bankiva* and *G. g murghi* being more differentiated (F_{ST} with all others subspecies higher than 0.17) than G. g. gallus and G. g. spadiceus (F_{ST} = 0.037).

4. Discussion

Out of the 15 possible topologies for the four *Gallus* species, no less than seven have been reported with good support over the past 20 years as reviewed by Tiley et al. (2020). Here, we report that the most likely topology is a *G. gallus*-basal topology, first reported by Kimball and Braun (2014). Our strong argument relies on the high gene flow (above 27%) from *G. varius* to the common ancestor of *G. lafayetti* and *G. sonneratii* that was estimated with the ABBA statistics when applied to the *G. varius*-basal topology. This high value is extremely unlikely and leads to invalidate the *G. varius*-basal topology. In addition to this argument, we are confident in our results because we have corrected for

Table 4
Summary statistics of diversity in each species. From left to right: average nucleotide diversity (π), average heterozygosity (Ho), average nucleotide divergence to the G. gallus reference in autosomes (Div Auto), in the Z sexual chromosome Z (Div Z), in the hemizygous sexual chromosome W (Div W) and in the mitochondrial genome (Div MT). Male samples were excluded when computing Div W.

Species	π	Ho	Div Auto.	Div W	Div Z	Div MT
G. g. domesticus		3.06e−03	3.99e−03	1.70e−04	2.49e−03	9.54e−04
G. gallus	4.12e−03	3.53e−03	4.33e−03	1.74e−04	2.67e−03	8.64e−04
G. lafayettii	1.10e−03	1.05e−03	8.56e−03	1.01e−03	7.54e−03	2.24e−02
G. sonneratii	2.60e−03	1.70e−03	8.35e−03	9.98e−04	7.33e−03	2.31e−02
G. varius	1.02e−03	8.04e−04	9.50e−03	8.09e−04	7.68e−03	1.24e−02

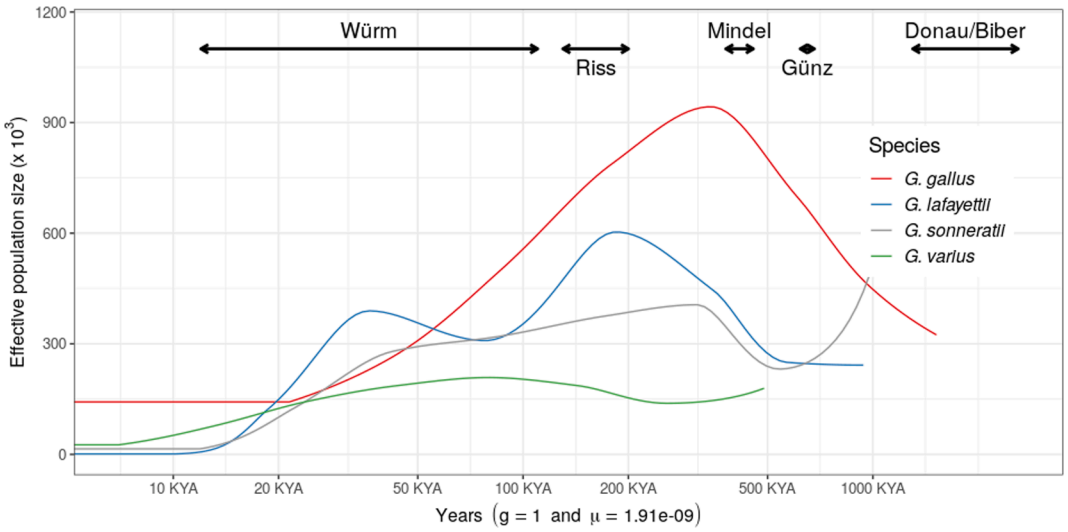

Fig. 6. Demographic histories of the 4 Gallus species reconstructed using smc++ (Terhorst et al. 2017). Glaciation periods are indicated by black arrows.

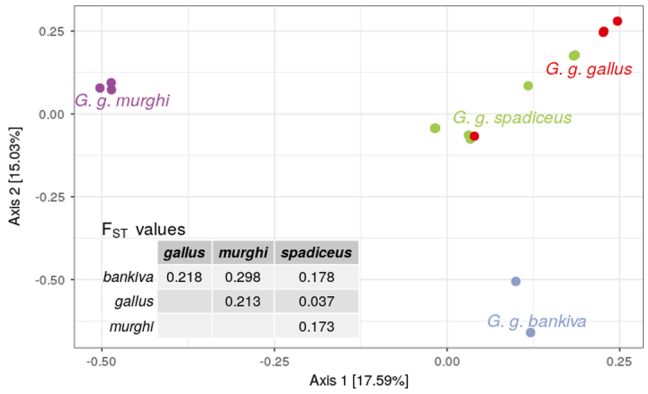

Fig. 7. F_{ST} values and PCA plot of the G. gallus subspecies.

an important source of error by removing the genomic regions found to be introgressed from *G. gallus* to the *G. sonneratii* birds sampled in the parks, representing 10% of the genome on average. At the same time, we could clarify the status of four putative subspecies of *G. gallus*: *G. g. murghi* and *G. g. bankiva* form well delineated subspecies with low genetic diversities, whereas *G. g. gallus* and *G. g. spadiceus* do not differentiate clearly, which may be partly due to their large within-population diversity. This is the first time that these subspecies are compared with all types of genetic information, i.e. from autosomes, mitochondria and sexual chromosomes. Finally, the demographic history showed that effective population sizes reflected quite well the distribution area of the four species, at the exception of *G. lafayettii*, and supported the hypothesis that *G. varius* speciated by isolation whereas *G. gallus* retained most of the genetic variability of the genus.

Considering the fact that the most recent studies favored a G. varius-basal topology (Hosner et al., 2016; Lawal et al., 2020; Tiley et al., 2020), we will now discuss the possible factors explaining this discrepancy between these studies and our conclusion.

4.1. Inferring the whole genome gallus topology

4.1.1. Autosomal compartments

The three main factors to be considered when comparing topologies obtained by different studies are (1) sampling strategy, (2) methodology and (3) type of molecular data.

Most phylogenetic studies of Galliforms using genomic data in addition to mitochondrial DNA data included only one individual to represent a species (Hosner et al., 2016; Kimball and Braun, 2014; Tiley et al., 2020). Furthermore, samples often came from private collections, rather than from the wild or from zoological parks. Only Lawal et al. (2020) and more recently Wang et al. (2020) used several individuals to represent each of the 4 wild species. In both studies, *G. sonneratii* samples came either from private collection or from unspecified locations. Shallow taxon sampling will result in longer branches in the intree and amplify potential misplacement of the outgroup, especially since the crown branch of *G. sonneratii* and *G. lafayettii* is very short compared to branch leading to the outgroup (Holland et al., 2003), To test this effect, we randomly selected one sample from each species and estimated the phylogeny on the 5 samples subset using both BioNJ and unpartitioned RAxML strategies. Repeating the process 100 times consistently returned the same topology: *G. varius* basal for the BioNJ tree and *G. gallus* basal for the RAxML tree. BioNJ trees all had perfect bootstrap support however, the RAxML bootstrap values (Table S6) for the (*G. sonneratii, G. lafayettii*) was only 50 (a random coin flip) for 12 subsets. In 10 of those, the *G. sonneratii* sample came from the zoological parks and the *G. gallus* sample was either a *murghi* or a *spadiceus* from Thailand. Likewise, the bootstrap values for the (*G. sonneratii, G. lafayettii, G. varius*) clade went down to 50 for 27 of the 100 subsets (all distinct from the previous 12). Note that for those subsets, the *G. varius* and *G. gallus* basal topologies have equal bootstrap support. In 20 out of the 27 subsets, the *G. gallus* sample was again a wild *murghi* or *spadiceus* sample but no similar peculiarity was identified for other species. Overall, in those 39 subsets, the best scoring RAxML tree coincides with the full topology but is very sensitive to small differences in the input super matrix. This seems to be especially true when using samples from zoological parks as *G. sonneratii* representative.

Obviously, the method used to establish the phylogenetic tree is of key importance, since we obtained very different trees according to the method used. Using both ML and distance methods on several genomic compartments (autosomes, W chromosome, mtDNA), we recovered topologies with well supported inner nodes that corresponded to three conflicting phylogenetic situations: *G. varius* and *G. gallus* as sisters species according to BioNJ tree with W and mtDNA (Fig. 2b and 2c) and to ML tree with mtDNA (Fig. 2f), a *G. varius*-basal topology according to BioNJ tree with autosomal data (Fig. 2a) and a *G. gallus*-basal topology according to ML tree with autosomal data (Fig. 2d). Our autosomal BioNJ topology was also found by Lawal et al. (2020) using distance methods, and by Hosner et al. (2016) and Tiley et al. (2020) using ML methods, whereas our ML topology was also found by Kimball and Braun (2014) using ML methods. In all instances, those topologies had close to

perfect support for their respective inference methods. In fact, high bootstrap values become irrelevant as the sheer size of the dataset means that systematic errors dominate stochastic and sampling errors (Young and Gillung, 2020). There are some modeling differences between our approach and Lawal et al. (2020). First, Lawal et al. (2020) reconstruct a neighbor joining tree on whole-genome data after doing SNP thinning (at most 1 SNP/kb). Note that we find the same topology when using the same method without SNP thinning (Fig. 2a). Second their ML tree is reconstructed using the GTR model on exon SNPs whereas we use GTR + CAT on full gene sequences. The CAT variant of GTR models rate heterogeneity across sites (RAS) and downweights rapidly evolving sites. With a parameter α estimated equal to 0.02, the sites evolve at markedly different rates. Using models without RAS often results in the same trees as distance method and may explain why Lawal et al. (2020) reconstructed the same topology with ML and NJ methods. Differences with Tiley et al. (2020) are more suprising as we used the same inference strategy (RAxML with no partitions and GTR + CAT models) but they could be due to differences in sample types and number.

The type of molecular data also plays an important role. Most previous studies used only mitochondrial or a combination of mitochondrial and nuclear markers. We relied instead on whole genome sequencing, like Lawal et al. (2020) and Tiley et al. (2020). We however used a different processing pipeline and set of base pairs. Tiley et al. (2020) used shallow sequencing, genome assembly, ab initio gene prediction and blastN to extract GS and used only exon/intron per GS. This resulted in smaller alignments (32 Mbp vs 420 Mbp) and less gene trees (3,406 vs 10,574) than our study. To test the effect of shorter alignments, we randomly selected 1000 genes (from the full set of 10,574) and ran an unpartitioned RAxML analyses on the resulting supermatrix. Repeating the process 100 times with different sets of regions consistently returned the same *G. varius* basal topology, with 100% bootstrap for the nodes corresponding to speciation events. The 100 corresponding supermatrices were in the range 34–44 Mbp in line with the one used by Tiley et al. (2020). This suggests that, past a certain point, the data processing pipeline has more impact on phylogenomic inference than the alignment length. Lawal et al. (2020) used the same "deep-sequencing and direct mapping" approach as us but performed SNP thinning (down to a density of 1 SNP / kb) prior to phylogenetic inference, whereas we used all SNPs. Note also that the gene tree spectrum reconstructed by Lawal et al. (2020) [Fig. 3] identified the two topologies discussed here (*G. varius*-basal and *G. gallus*-basal) as the most likely along the genome with almost equal frequencies (~close to 20% for each). Their ML tree favored the *G. varius*-basal tree in the end but they only used 3 *G. sonneratii* samples from a private collection. This is problematic as these samples suffered from recent *G. gallus* to *G. sonneratii* introgression, as shown by their position in the mtDNA topology. We found in this study that recent introgression accounts for around 10% of the genome, which can pull *G. gallus* closer to *G. sonneratii*. The addition of wild samples (like GS_113 and GS_349) could lead to a larger fraction of the genome favoring the *G. gallus*-basal and tip the balance towards our autosomal-ML tree. To test this hypothesis, we removed the non-introgressed *G. sonneratii* from the analyses and computed the ML tree again using an unpartitioned RAxML analyses.

In view of these results, including multiple samples per species and cleaning recent footprints of introgression appear to have a strong impact on phylogenomic inference. All methods used in Tiley et al. and Lawal et al. favored the *G. varius*-basal tree but the *G. gallus*-basal one was always a close second when looking at signal heterogeneity along the chromosome or when using different subsets. Since bootstrap values are uninformative in that context, we looked for further clues supporting one or the other topology. We estimated gene flows from *G. varius* to *G. sonneratii* and *G. lafayetti* under both topologies using the ABBA-BABA statistics. We found estimates in the 4.5–5.5% range for the ML topology and 29.4–30.4% for the BioNJ topology. The latter are not realistic estimates and constitutes strong evidence against the BioNJ topology and in favor of the ML one. This is further corroborated by results from the ASTRAL multispecies-coalescent analysis, based on results from individual gene trees, which led to the same topology with all inner branches having posterior probabilities equal to or close to 1.

Since we used whole genome sequencing, we could separately study the phylogenetic trees according to the genetic origin of the polymorphism and not surprisingly, each of them tells a different story.

4.1.2. Other compartments

Differences between nuclear trees and mitochondrial trees have been observed in Lawal et al. (2020) and can arise from many mechanisms, including introgression and inadequate taxon sampling (illustrated here by the location of the *G. sonneratii* samples from parks among the *G. gallus* birds), incomplete lineage sorting (ILS) or on the different properties of the mtDNA compared to autosomes. The mtDNA sequence represents one unique haplotype that does not recombine and that is maternally transmitted. It has higher mutation rates, as well as a smaller effective population size, than typical autosomal loci and hence has a shorter coalescence time, thus telling evolutionary histories that can differ from those obtained from autosomal data. Furthermore, the mitochondrial genome has a complex secondary structure that can affect mutation rates along the molecule (Meiklejohn et al., 2014) and would require dedicated models of sequence evolution, not available in RaxML. It is remarkable that ML and BioNJ trees for mtDNA consistently positioned G. varius as a sister species to G. gallus with perfect bootstrap support. Such a discrepancy was described previously and could be explained, beyond the error in phylogenetic reconstruction due to poor model fit as aforementionned, by two alternative hypotheses: introgressive hybridization or retention of ancestral mtDNA through ILS (Avise, 1993; Maddison, 1997; Funk and Omland, 2003).

In the first scenario, we can hypothesize the complete replacement of mtDNA of one species by that of another species. Natural introgressive hybridization has been described within birds (Grant and Grant, 1992; Ottenburghs et al., 2015, 2017). Following this hypothesis, the discordance between mitochondrial and autosomal genomes could be explained by mtDNA transfer from *G. gallus* to *G. varius*. This would be possible since Sawai et al. (2010) described interfertile hybrids between *G. gallus domesticus* and *G. varius* in Java Island. Yet, if such introgression took place, it must have occurred quite a long time ago, since we found no evidence for such an event at the whole genome level, as we did for *G. sonneratii*.

In the last scenario, we can hypothesize the existence of several mtDNA lineages present in the ancestral species of the *Gallus* group that do no segregate according to the species tree. Note that Tiley et al. (2020 Fig. 5) reported that ~15% of gene trees reconstructed from the autosomal genome had a topology where *G. gallus* and *G. varius* were sister species, in line with prediction from a multispecies coalescent. Similarly, Lawal et al. (2020) found a high proportion of trees (~15%) supporting a (*G. gallus*, *G. varius*) clade using a sliding windows approach along the chromosome. The discrepancy between the mtDNA and the full autosome therefore merely reflects the diversity of phylogenetic signals encoded in the autosome and is compatible with ILS. Cases of peripheral speciation have previously been shown to favor such a situation (Funk and Omland, 2003).

Yet, the specific situation of W chromosome phylogeny in the *Gallus* genus also deserves a specific analysis as the ML-W tree coincides with the ML-autosomal tree (albeit for the 3 introgressed sonneratii) but the BioNJ-W tree coincides with the BioNJ-mitochondrial tree, although the differences correspond to some branches with medium support, We expected the W-tree to have the same topology as the mtDNA tree as both are inherited maternally and do not recombine, except for the pseudo-autosomal region (PAR) of the W chromosome. This is indeed the case for the BioNJ trees but not for the ML ones, although the differences are not well supported (~60% bootstrap values). The difference between mtDNA and W trees could stem from the presence of genes in the PAR, bringing the W tree closer to the autosomal tree. However, our

estimates suggest that the PAR is only 5 Mb long (Tixier-Boichard et al., 2016) and the PAR should affect both the BioNJ and the ML tree. The difference could also arise from different evolutionary pressures on the W and the mtDNA. This was suggested by Smeds et al. (2015) in the flycatcher, where W genes are under stabilizing selection pressure, but it was not enough to obtain different topologies for the W-tree and the mtDNA-tree in their studies. Finally, the discrepancy between the ML-W tree and the ML-mtDNA tree on our data could also be due to a relatively low amount of data, as compared to the autosomal tree.

4.2. Evolutionary dynamics of gallus

The smc++ and psmc results suggested a much smaller population size for *G. varius* compared to *G. gallus*, in agreement with a scenario of speciation by isolation. The lack of correlation between population sizes fluctuations and glaciation periods, usually observed for domestic species, can be related to the tropical nature of *Gallus* species habitats, making them impervious to glaciations. Overall, effective population sizes are ordered according to the distribution area of species, with the notable exception of *G. lafayettii* for which both smc++ and PSMC estimated large population sizes, up to twice higher than for *G. sonneratii*.

At least two explanations can be proposed to explain it. First, we used the same value, extracted from the literature and computed from *G. gallus* sequences (Nam et al., 2010), and generation time (1 year) for all *Gallus* species, whereas species-specific values might be more appropriate and give different scalings. Second, and more likely, it is possible that *G. lafayettii* suffers from population fragmentation when compared to *G. sonneratii*. Indeed, Mazet et al. (2016) showed that in the presence of structured population and decreased gene flow between subpopulations, psmc/smc++ will reconstruct increased N_e and cannot distinguish between genuine changes in the effective population size and population fragmentation. The rapid increase in effective population size of *G. lafayettii* during the last ice ages, when the sea level decreased and birds could cross the Palk Strait is also reminiscent of changes in habitat size.

Finally, the speciation times estimated by smc++ are much lower than those reported in the literature. All estimates of speciation times based on fossil (Kan et al., 2010), molecular (Jetz et al., 2012) or geological evidence (Sawai et al., 2010) suggest speciation times in the 2–5 MYA range. Smc++ is based on an idealized two-population split model after which effective migration between populations becomes negligible. It is thus expected to underestimate divergence times when post-divergence gene flow has taken place. The discrepancy between fossil estimates and smc++ estimates suggests significant post-speciation gene flow between the *Gallus* species, in line with the results of the ABBA-BABA analyses.

It must be noticed that the trajectory of *G. gallus murghi* exhibited early divergence from the other *G. gallus*, supporting its classification as a sub-species.

4.3. Gallus sonneratii from zoological parks

Introgression from domestic chickens to *G. sonneratii* individuals from zoological parks was documented in Nishibori et al. (2005) and recovered in our mtDNA tree (Fig. 2c, 2f). Our results suggest that *G. sonneratii* individuals from zoological parks are in fact hybrids between *G. sonneratii* and *G. gallus*, in contrast to individuals sampled in the Indian rainforest and to the one individual kept in the zoological park of New Delhi. This is corroborated by the evidence of recent genetic flow from *G. gallus* to those samples. The length of those regions (up to a few Mb) and their composite nature (succesion of windows with two *G. gallus* haplotypes and windows with only one *G. gallus* haplotype) is also suggestive of cross-over events breaking the *G. gallus* genetic material initially transferred.

The haplotype network (Fig. 3) suggests that the original mating event leading to hybrids took place between a male *G. sonneratii* and a domestic female *G. gallus* followed by backcross of the female progeny to *G. sonneratii* males. This is compatible with previous findings that (i) to mate *G. gallus* males with *G. sonneratii* females seems harder than the reverse (Danforth, 1958; Ghighi, 1922; Morejohn, 1968), (ii) backcrosses are vigorous (Crawford, 1990) and (iii) backcrosses with G. sonneratii are morphologically similar to pure G. sonneratii due to re-appearance of the phenotypic traits (Danforth, 1958; Ghighi, 1922). Furthermore, experimental crosses showed a better hatchability for backcross chicks obtained from an F1 dam mated to a *G. sonneratii* male, than for F2 chicks (Morejohn, 1968), as observed more generally within Galliform species (Arrieta et al., 2013).

Considering the high similarity of the mitochondrial haplotypes found in *G. sonneratii* park birds (3 are identical and the other 2 differ by a single nucleotide) and their high frequency among *G. g. domesticus*, a single cross-species mating event could be responsible for the introgression observed. Indeed, Region 1 (Chr4:46,540,001–46,880,000) was introgressed in all three individuals, and pure *G. gallus* in GS_06012 and GS_04252. Since these two individuals were sampled in different zoological parks (France and Japan), it is likely that there was only one *G. sonneratii* to *G. gallus* mating and that hybrid individuals were then exchanged between parks. Most *G. sonneratii* samples from previous studies, with the notable exception of the sample from the New Delhi park used in Nishibori et al. (2005) and then later in Meiklejohn et al. (2014), also possess a *G. gallus* mtDNA haplotype. Since those samples, when documented, come from private collections, breeding facilities or zoological parks, the problem appears no be limited to our two parks but rather widespread. We estimate that the haplotypes of Region 1 in GS_06012 and GS_04252 started diverging 150 years ago and that the original mating therefore took place before that, and possibly before *G. sonneratii* started being conserved in parks and samples exchanged between conservation sites.

Our results highlight the utmost importance of taxon sampling and careful validation of the taxa for phylogenetic analyses.

5. Conclusions and perspectives

Our study highlight the importance of verifying the genetic status of samples which originate from zoological parks or private collections to avoid using animals that have undergone some level of hybridization, as for *G. sonneratii*. Considering that the zoological parks (at least the ones reported in genetic studies) do not own *G. sonneratii* birds that are not pure representative of the species, its preservation in the wild in India is of utmost importance. The other two wild species (*G. varius* and *G. lafayettii*) appear not to suffer from the same problem.

Although conflicting, all topologies discussed in this work had strong support, which is a predictable by-product of the huge number of markers used in each inference procedure. Differences arise not so much from weak phylogenetic signal and uncertainty in the sequence data (variance) rather than from systematic error induced by invalid modeling assumptions, inadequate evolution models and insufficient taxon sampling. This highlights once again the need to use enough loci and enough taxa of high quality when performing phylogenomics inference (Shen et al., 2017). We can add that the ABBA statistics was very useful to discriminate between two possible topologies and we are fairly confident that the *G. gallus*-basal topology is the most likely according to our whole genome data.

CRediT authorship contribution statement

Mahendra Mariadassou: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing - original draft. **Marie Suez:** Conceptualization, Formal analysis, Writing - original draft. **Sathya Sathyakumar:** Resources. **Alain Vignal:** Investigation, Writing - review & editing. **Mariangela Arca:** Formal analysis. **Pierre Nicolas:** Formal analysis, Investigation, Writing - review & editing. **Thomas**

Faraut: Investigation, Writing - review & editing. **Diane Esquerré:** Data curation. **Masahide Nishibori:** Resources. **Agathe Vieaud:** Resources, Data curation. **Chih-Feng Chen:** Resources. **Hung Manh Pham:** Resources. **Yannick Roman:** Resources. **Frédéric Hospital:** Investigation, Writing - review & editing. **Tatiana Zerjal:** Conceptualization, Investigation, Validation, Writing - review & editing. **Xavier Rognon:** Conceptualization, Investigation, Validation, Writing - review & editing. **Michèle Tixier-Boichard:** Conceptualization, Investigation, Validation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

This work was supported by the French research agency (ANR) [grant name Domestichick, grant number ANR-12-BSV6-0018]. We are grateful to the INRAE MIGALE bioinformatics platform (http://migale.inrae.fr) for providing computational resources and Jean-Claude Fotsa for providing the sample of a village chicken from Cameroun. Marie Suez was also partially supported by the CRB-Anim infrastructure project, ANR-11-INBS-0003, funded by the French National Research Agency in the frame of the 'Investing for the Future' program.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2020.107044.

References

Arrieta, R.S., Lijtmaer, D.A., Tubaro, P.L., 2013. Evolution of postzygotic reproductive isolation in galliform birds: Analysis of first and second hybrid generations and backcrosses. Biol. J. Linn. Soc. 110, 528–542. https://doi.org/10.1111/bij.12153.

Avise, J.C., 1993. Molecular markers, natural history and evolution. Chapman; Hall, New York. https://doi.org/10.1007/978-1-4615-2381-9.

Crawford, R.D., 1990. Poultry Breeding and Genetics. Elsevier, Amsterdam; New York.

Danforth, C., 1958. Gallus sonnerati and the domestic fowl. J. Hered. 49, 167–170.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Angel, G. del, Rivas, M.A., Hanna, M., et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genet. 43, 491–498. https://doi.org/10.1038/ng.806.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340.

Eriksson, J., Larson, G., Gunnarsson, U., Bed'hom, B., Tixier-Boichard, M., Strömstedt, L., Wright, D., Jungerius, A., Vereijken, A., Randi, E., Jensen, P., Andersson, L., 2008. Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. PLOS Genetics 4, 1–8. https://doi.org/10.1371/journal.pgen.1000010.

Frith, M.C., Kawaguchi, R., 2015. Split-alignment of genomes finds orthologies more accurately. Genome Biol. 16, 106. https://doi.org/10.1186/s13059-015-0670-9.

Fumihito, A., Miyake, T., Sumi, S., Takada, M., Ohno, S., 1994. One subspecies of the red junglefowl (Gallus gallus gallus) suffices as the matriarchic ancestor of all domestic breeds. PNAS 91, 12505–12509.

Fumihito, A., Miyake, T., Takada, M., Shingu, R., Endo, T., Gojobori, T., Kondo, N., Ohno, S., 1996. Monophyletic origin and unique dispersal patterns of domestic fowls. PNAS 93, 6792–6795.

Funk, D.J., Omland, K.E., 2003. Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial dna. Annu. Rev. Ecol. Evol. Syst. 34, 397–423. https://doi.org/10.1146/annurev.ecolsys.34.011802.132421.

Gascuel, O., 1997. BIONJ: An improved version of the nj algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14, 685–695.

Ghighi, A., 1922. L'hybridisme dans la génèse des races domestiques d'oiseaux. Genetica 4, 364–374.

Grant, P.R., Grant, B.R., 1992. Hybridization of bird species. Science 256, 193–197. https://doi.org/10.1126/science.256.5054.193.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., Pääbo, S., 2010. A draft sequence of the neandertal genome. Science 328, 710–722. https://doi.org/10.1126/science.1188021.

Groenen, M.A., Megens, H.-J., Zare, Y., Warren, W.C., Hillier, L.W., Crooijmans, R.P., Vereijken, A., Okimoto, R., Muir, W.M., Cheng, H.H., 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genom. 12 https://doi.org/10.1186/1471-2164-12-274.

Hillel, J., Groenen, M.A.M., Tixier-Boichard, M., Korol, A.B., David, L., Kirzhner, V.M., Burke, T., Barre-Dirie, A., Crooijmans, R.P., Elo, K., Feldman, M.W., Freidlin, P.J., Maki-Tanila, A., Oortwijn, M., Thomson, P., Vignal, A., Wimmers, K., Weigend, S., 2003. Biodiversity of 52 chicken populations assessed by microsatellite typing of dna pools. Genet. Selection Evol. 35, 533.

Holland, B.R., Penny, D., Hendy, M.D., 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - a simulation study. Syst. Biol. 52, 229–238. https://doi.org/10.1080/10635150390192771.

Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2016. Avoiding missing data biases in phylogenomic inference: An empirical study in the landfowl (aves: Galliformes). Mol. Biol. Evol. 33, 1110–1125. https://doi.org/10.1093/molbev/msv347.

Jetz, W., Thomas, G., Joy, J., Hartmann, K., Mooers, A., 2012. The global diversity of birds in space and time. Nature 491, 444–448.

Kan, X.-Z., Li, X.-F., Lei, Z.-P., Chen, L., Gao, Z.-Y., H. Yang, Yang, J.-K., Guo, Z.-C..., Yu, L., Zhang, L.-Q., Qian, C.-J., 2010. Estimation of divergence times for major lineages of galliform birds: Evidence from complete mitochondrial genome sequences. African Journal of Biotechnology 9, 3073–3078.

Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., Frith, M.C., 2011. Adaptive seeds tame genomic sequence comparison. Genome Res. 21, 487–493. https://doi.org/10.1101/gr.113985.110.

Kimball, R.T., Braun, E.L., 2014. Does more sequence data improve estimates of galliform phylogeny? Analyses of a rapid radiation using a complete data matrix. PeerJ 2, e361. https://doi.org/10.7717/peerj.361.

Kranis, A., Gheyas, A.A., Boschiero, C., Turner, F., Yu, L., Smith, S., Talbot, R., Pirani, A., Brew, F., Kaiser, P., Hocking, P.M., Fife, M., Salmon, N., Fulton, J., Strom, T.M., Haberer, G., Weigend, S., Preisinger, R., Gholami, M., Qanbari, S., Simianer, H., Watson, K.A., Woolliams, J.A., Burt, D.W., 2013. Development of a high density 600K snp genotyping array for chicken. BMC Genomics 14, 59. https://doi.org/10.1186/1471-2164-14-59.

Lawal, R.A., Martin, S.H., Vanmechelen, K., Vereijken, A., Silva, P., Al-Atiyat, R.M., Aljumaah, R.S., Mwacharo, J.M., Wu, D.-D., Zhang, Y.-P., Hocking, P.M., Smith, J., Wragg, D., Hanotte, O., 2020. The wild species genome ancestry of domestic chickens. BMC Biol. 18 https://doi.org/10.1186/s12915-020-0738-1.

Lawler, A., 2014. Why did the chicken cross the world?: The epic saga of the bird that powers civilization. Atria Books.

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496. https://doi.org/10.1038/nature10231.

Mackey, K., Steinkamp, A., Chomczynski, P., 1998. DNA extraction from small blood volumes and the processing of cellulose blood cards for use in polymerase chain reaction. Mol. Biotechnol. 9, 1–5. https://doi.org/10.1007/BF02752692.

Maddison, W.P., 1997. Gene Trees in Species Trees. Syst. Biol. 46, 523–536. https://doi.org/10.1093/sysbio/46.3.523.

Malomane, D.K., Simianer, H., Weigend, A., Reimer, C., Schmitt, A.O., Weigend, S., 2019. The SYNBREED chicken diversity panel: a global resource to assess chicken diversity at high genomic resolution. BMC Genom. 20, 345. https://doi.org/10.1186/s12864-019-5727-9.

Martin, S.H., Davey, J.W., Jiggins, C.D., 2014. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol. 32, 244–257. https://doi.org/10.1093/molbev/msu269.

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., Chikhi, L., 2016. On the importance of being structured: Instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? Heredity 116, 362–371, 10.1038/hdy.2015.104.

Meiklejohn, K.A., Danielson, M.J., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2014. Incongruence among different mitochondrial regions: A case study using complete mitogenomes. Mol. Phylogenet. Evol. 78, 314–323. https://doi.org/10.1016/j.ympev.2014.06.003.

Miao, Y.-W., Peng, M.-S., Wu, G.-S., Ouyang, Y.-N., Yang, Z.-Y., Yu, N., Liang, J.-P., Pianchou, G., Beja-Pereira, A., Mitra, B., al., 2013. Chicken domestication: An updated perspective based on mitochondrial genomes. Heredity 110, 277–282. https://doi.org/10.1038/hdy.2012.83.

Morejohn, G.V., 1968. Breakdown of isolation mechanims in two species of captive junglefowl (gallus gallus and gallus sonneratii). Evolution 22, 576–582. https://doi.org/10.1111/j.1558-5646.1968.tb03993.x.

Nam, K., Mugal, C., Nabholz, B., Schielzeth, H., Wolf, J.B., Backström, N., Künstner, A., Balakrishnan, C.N., Heger, A., Ponting, C.P., Clayton, D.F., Ellegren, H., 2010. Molecular evolution of genes in avian genomes. Genome Biol. 11, R68. https://doi.org/10.1186/gb-2010-11-6-r68.

Nei, M., 1987. Molecular evolutionary genetics. Columbia University Press.

Nishibori, M., Shimogiri, T., Hayashi, T., Yasue, H., 2005. Molecular evidence for hybridization of species in the genus Gallus except for Gallus varius. Anim. Genet. 36, 367–375. https://doi.org/10.1111/j.1365-2052.2005.01318.x.

Ottenburghs, J., Kraus, R.H.S., van Hooft, P., van Wieren, S.E., Ydenberg, R.C., Prins, H.H.T., 2017. Avian introgression in the genomic era. Avian Res. 8 https://doi.org/10.1186/s40657-017-0088-z.

Ottenburghs, J., Ydenberg, R.C., Van Hooft, P., Van Wieren, S.E., Prins, H.H., 2015. The avian hybrids project: Gathering the scientific literature on avian hybridization. Ibis 157, 892–894.

Paradis, E., 2010. Pegas: An R package for population genetics with an integrated–modular approach. Bioinformatics 26, 419–420.

Pfeifer, B., Wittelsbuerger, U., Ramos-Onsins, S.E., Lercher, M.J., 2014. PopGenome: An efficient swiss army knife for population genomic analyses in r. Mol. Biol. Evol. 31, 1929–1936. https://doi.org/10.1093/molbev/msu136.

Privé, F., Aschard, H., Ziyatdinov, A., Blum, M.G.B., 2018. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics 34, 2781–2787. https://doi.org/10.1093/bioinformatics/bty185.

Sawai, H., Kim, H.L., Kuno, K., Suzuki, S., Gotoh, H., Takada, M., Takahata, N., Satta, Y., Akishinonomiya, F., 2010. The origin and genetic variation of domestic chickens with special reference to junglefowls gallus g. Gallus and g. Varius. PLOS ONE 5, 1–11. https://doi.org/10.1371/journal.pone.0010639.

Shen, K.A.C., Dai, Yong-Yi AND, 2014. The updated phylogenies of the phasianidae based on combined data of nuclear and mitochondrial dna. PLoS ONE 9, 1–5. https://doi.org/10.1371/journal.pone.0095786.

Shen, X.-X., Hittinger, C.T., Rokas, A., 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1 https://doi.org/10.1038/s41559-017-0126.

Shen, Y.-Y., Liang, L., Sun, Y.-B., Yue, B.-S., Yang, X.-J., Murphy, R.W., Zhang, Y.-P., 2010. A mitogenomic perspective on the ancient, rapid radiation in the galliformes with an emphasis on the phasianidae. BMC Evol. Biol. 10, 132. https://doi.org/10.1186/1471-2148-10-132.

Shen, Y.-Y., Shi, P., Sun, Y.-B., Zhang, Y.-P., 2009. Relaxation of selective constraints on avian mitochondrial DNA following the degeneration of flight ability. Genome Res. 19, 1760–1765. https://doi.org/10.1101/gr.093138.109.

Smeds, L., Warmuth, V., Bolivar, P., Uebbing, S., Burri, R., Suh, A., Nater, A., Bureš, S., Garamszegi, L.Z., Hogner, S., et al., 2015. Evolutionary analysis of the female-specific avian w chromosome. Nat. Commun. 6, 7330.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Stein, R.W., Brown, J.W., Mooers, A.Ø., 2015. A molecular genetic time scale demonstrates cretaceous origins and multiple diversification rate shifts within the order galliformes (aves). Mol. Phylogenet. Evol. 92, 155–164, 10.1016/j.ympev.2015.06.005.

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

Tiley, G.P., Pandey, A., Kimball, R.T., Braun, E.L., Burleigh, J.G., 2020. Whole genome phylogeny of gallus: Introgression and data-type effects. Avian Res. 11 https://doi.org/10.1186/s40657-020-00194-w.

Tixier-Boichard, M., Mariadassou, M., Nicolas, P., Marca, M., Esquerre, D., Vignal, A., 2016. The domestichick project: From 57K to whole genome sequence data., in: 26th Plant and Animal Genome Conference.

Wang, N., Kimball, R.T., Braun, E.L., Liang, B., Zhang, Z., 2013. Assessing phylogenetic relationships among galliformes: A multigene phylogeny with expanded taxon sampling in phasianidae. PLoS ONE 8, 1–12. https://doi.org/10.1371/journal.pone.0064312.

Wang, M., Thakur, M., Peng, M., Jiang, .Y, Frantz, L., Li, M., Zhang, J., Wang, S., Peters, J., Otecko, N., Suwannapoom, C., Guo, X., Zheng, Z., Esmailizadeh, A., Hirimuthugoda, N., Ashari, H., Suladari, S., Zein, M., Kusza, S., Sohrabi, S., Kharrati-Koopaee, H., Shen, Q., Zeng, L., Yang, M., Wu, Y., Yang, Y., Lu, X., Jia, X., Nie, Q., Lamont, S., Lasagna, S., Ceccobelli, S., Gunwardana, H., Senasige, T., Feng, S., Si, J., Zhang, H., Jin, J., Li, M., Liu, Y., Chen, H., Ma, C., Dai, S., Bhuiyan, A., Khan, M., Pradeepa Silva, G., Le, T., Mwai, O., Ibrahim, M., Supple, M., Shapiro, B., Hanotte, O., Zhang, G., Larson, G., Han, J., Wu, D, Zhang, Y., 2020. 863 genomes reveal the origin and domestication of chicken. Cell Research 30, 693–701. https://doi.org/10.1038/s41422-020-0349-y.

Young, A.D., Gillung, J.P., 2020. Phylogenomics - principles, opportunities and pitfalls of big-data phylogenetics. Syst Entomol 45, 225–247. https://doi.org/10.1111/syen.12406.

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-iii: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinf. 19, 153. https://doi.org/10.1186/s12859-018-2129-y.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A., 2013. The masurca genome assembler. Bioinformatics 29, 2669–2677. https://doi.org/10.1093/bioinformatics/btt476.