Received: 8 December 2021 | Revised: 11 April 2022 | Accepted: 17 April 2022

DOI: 10.1111/mec.16484

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Satellite DNA evolution in Corvoidea inferred from short and long reads

Valentina Peona[1] | Verena E. Kutschera[2] | Mozes P. K. Blom[3,4] | Martin Irestedt[3] | Alexander Suh[1,5]

[1]Department of Organismal Biology – Systematic Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[2]Department of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University, Solna, Sweden

[3]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden

[4]Museum für Naturkunde, Leibniz Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany

[5]School of Biological Sciences—Organisms and the Environment, University of East Anglia, Norwich, UK

Correspondence
Valentina Peona and Alexander Suh, Department of Organismal Biology – Systematic Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden.
Emails: valentina.peona@gmail.com (V. P.); alexander.suh@ebc.uu.se (A. S.)

Funding information
Vetenskapsrådet, Grant/Award Number: 2016-05139, 2019-03900, 2020-04436 and 621-2014-5113; Svenska Forskningsrådet Formas, Grant/Award Number: 2017-01597; Knut and Alice Wallenberg Foundation

Handling Editor: Polina Novikova

Abstract

Satellite DNA (satDNA) is a fast-evolving portion of eukaryotic genomes. The homogeneous and repetitive nature of such satDNA causes problems during the assembly of genomes, and therefore it is still difficult to study it in detail in nonmodel organisms as well as across broad evolutionary timescales. Here, we combined the use of short- and long-read data to explore the diversity and evolution of satDNA between individuals of the same species and between genera of birds spanning ~40 millions of years of bird evolution using birds-of-paradise (Paradisaeidae) and crow (*Corvus*) species. These avian species highlighted the presence of a GC-rich Corvoidea satellitome composed of 61 satellite families and provided a set of candidate satDNA monomers for being centromeric on the basis of length, abundance, homogeneity and transcription. Surprisingly, we found that the satDNA of crow species rapidly diverged between closely related species while the satDNA appeared more similar between birds-of-paradise species belonging to different genera.

KEYWORDS
base composition, birds, birds-of-paradise, comparative genomics, crow, genomic dark matter, satellitome

1 | INTRODUCTION

Satellite DNA (satDNA) comprises homogeneous tandemly repeated genomic sequences that can extend for megabases (Plohl et al., 2012) while repetitive units (monomers) range from hundreds to thousands of base pairs. satDNA monomers are commonly arranged in a head-to-tail fashion but can also form more complex structures like higher-order repeats where the tandemly repeated

unit consists of multiple different monomers. Arrays of satDNA monomers are mainly distributed in heterochromatinised areas of the chromosomes like (peri)centromeres and telomeres, and less often in euchromatin (Brajković et al., 2012; Kuhn et al., 2012; Larracuente, 2014; de Lima et al., 2017; Palacios-Gimenez et al., 2017; Ruiz-Ruano et al., 2016). Centromeric satDNA is involved in the recognition of the centromere itself and the establishment of the correct heterochromatin profile that guides the kinetochore assembly and attachment of the spindle during cell division (Leclerc & Kitagawa, 2021). The transcription and the potential to form secondary structures seem to be key for the establishment, recognition, and maintenance of centromeres (Grenfell et al., 2017; Hartley & O'Neill, 2019; Leclerc & Kitagawa, 2021). While some satDNA arrays outside the centromeres have been shown to participate in meiotic drive or chromosome segregation (Ferree, 2014; Joshi & Meller, 2017; Larracuente, 2014), most have not been linked to any function and probably evolve under relaxed selective constraints, leading to fast sequence divergence between species (Lower et al., 2018; Plohl et al., 2012). satDNA sequences can diverge so fast that no similarity can be found at all between the satellitomes (collection of all the satDNA sequences in a genome, species, or group) of different taxonomic groups (Ruiz-Ruano et al., 2018). Though on shorter evolutionary timescales (e.g., between species of the same genus), it may be possible to retrieve shared satDNA sequences (i.e., satDNA families) (Smalec et al., 2019). Even satDNA arrays without any function can exert several effects at genomic and evolutionary levels (Pezer et al., 2012). The presence/absence of arrays and their population variation in length can lead to different genome-wide epistatic effects by modulating the heterochromatin landscape (Jiang et al., 2010). Upon hybridisation, a different satellite composition between the two parental species' chromosomes can lead to heterochromatic instability (Ferree & Barbash, 2009), epistatic effects, problems in the correct establishment of centromeres (Dion-Côté & Barbash, 2017), the establishment of aberrant chromocenters with repercussions on chromosome segregation and cell death at meiosis (Jagannathan & Yamashita, 2021), and in general to genetic incompatibilities (Dion-Côté & Barbash, 2017).

satDNA arrays are thought to originate, expand, and diverge through mechanisms of unequal crossing-over (Smith, 1976), rolling-circle replication (Charlesworth et al., 1994), and polymerase slippage (Charlesworth et al., 1994; Raz et al., 2019). The origin of new satDNA families is often linked to transposable elements (TEs) (Dias et al., 2014; Hartley & O'Neill, 2019; Meštrović et al., 2015; Plohl et al., 2012). Since some TEs already contain tandem repeats within their sequences, active TEs can scatter these satDNA seeds across the genome where they can subsequently expand (Cheng & Murata, 2003; Dias et al., 2014). It has also been observed that when active TEs insert directly adjacent to one another, they can rapidly form dimers that could further expand into arrays, turning the TE itself into a satDNA as in the case of the DNA transposon *hobo* in *Drosophila melanogaster* (McGurk & Barbash, 2018).

The characterisation of satDNA solely from genomic data is still very challenging because sequencing technologies are not able to produce genomic reads long enough to span entire satDNA arrays (Miga, 2020; Peona et al., 2018). Typical next-generation sequencing (NGS) short reads are much shorter than the lengths of arrays and often of individual monomers. Therefore, satDNA arrays collapse during the assembly process, leaving arrays significantly fragmented into only a few or even single satDNA monomers (Peona, Blom, et al., 2021; Peona et al., 2018). Although short reads are not particularly useful to assemble satellite DNA, they can be useful to characterise a great part of the diversity and quantity of satellite DNA. Because of the incomplete assembly of satDNA arrays, cytogenetic verification is still necessary to understand the chromosomal locations (e.g., centromeric, telomeric) of the satDNA families of interest (Deakin et al., 2019). In the past few years, thanks to telomere-to-telomere complete assemblies of human chromosomes, the first complete investigation of satDNA (and other repetitive elements) became possible in a vertebrate (Jain et al., 2018; Logsdon et al., 2021; Miga, 2020; Sergey et al., 2022). This is not feasible yet for the vast majority of species because most genome assemblies, even those of model organisms, are not even close to being telomere-to-telomere (Peona et al., 2018). Since some satDNAs can have significant genomic and fitness effects on the organisms, it is important to fully describe their sequences, abundance, structure, and evolution across the Tree of Life. To investigate such elusive genomic regions in the absence of telomere-to-telomere assemblies, the combination of different types of genomic data and bioinformatic approaches is essential.

While some cytogenetic (Brown & Jones, 1972; Cacciò et al.,1994; Deryusheva et al., 2007; Liangouzov et al., 2002; Saifitdinova et al., 2001; Tsuda et al., 2007; Yamada et al., 2002, 2004) or bioinformatic (Liu et al., 2021; Vontzou, 2021; Weissensteiner et al., 2017; Westerberg, 2020) studies of satDNA familes have been done, our understanding of satDNA evolution in birds remains relatively limited given that they are the most species-rich group of land vertebrates. In this study, we focus on the sequence and structure characterisation of satDNA evolution in multiple species of Paradisaeidae (birds-of-paradise, BOPs) and *Corvus* spp. (crows and relatives) that both belong to the songbird superfamily Corvoidea (Jønsson et al., 2016; Oliveros et al., 2019). Paradisaeidae is a family of about 40 species that started to diversify from a crow-like ancestor ~20–25 million years ago (Ma) in the Australo-Papuan region which is divided into a small clade of monomorphic crow-like species and a core clade that includes the colourful species (Irestedt et al., 2009) (core BOPs; Figure 2). Indeed, BOPs are a prime example for species that evolved under strong and prolonged sexual selection (Irestedt et al., 2009). Despite the strong sexual selection, careful evaluation of museum specimens suggests that hybridization may be widespread among birds-of-paradise even between species of different genera (Frith & Frith, 1996a, 1996b; Fuller, 1979; Mayer & Peckover, 1991; Scott, 2013). The genus *Corvus* encompasses 40–44 species that started to diversify about 18 Ma (Jønsson et al., 2016) and their species are spread all over the world. Although the exact extent of hybridisation among BOPs is still unknown, hybridisation in the *Corvus* genus seems to occur only between a few species (Ottenburghs et al., 2015) but these hybrids became an important model for speciation in

evolutionary biology. The *Corvus* genus presents one of the most famous examples of hybrid zones following post-glacial recolonization of Europe between the black carrion crow (*Corvus corone corone*) and the black-and-grey hooded crow (*Corvus corone cornix*) (Meise, 1928).

Given the contrasting evolutionary paths that *Corvus* spp. and BOPs experienced, and the vast availability of genomic data for many of their species, we chose these groups to carry out the, to our knowledge, first broadscale comparative analysis of the avian satellitome. This sampling encompasses species across multiple evolutionary timescales with comparisons between closely related species where we expect the satellitome to be rather similar, and more distantly related and morphologically differentiated species where we expect the satellitome to be rather different. We collected genomic libraries and genome assemblies for 16 species of BOPs (16 short-read or linked-read libraries and two long-read libraries) and eight species of *Corvus* (8 short-read libraries). Moreover, the sampling included multiple individuals of *Lycocorax pyrrhopterus*, *Corvus corone cornix*, *Corvus corone corone*, and *Corvus corone cornix* × *corone* hybrids. This large multispecies data set allowed us to investigate and compare the evolution (presence/absence, abundance, and array structures) of satDNA families ranging from species that rapidly differentiated into a kaleidoscope of dimorphisms and behaviours (i.e., BOPs) to monomorphic species that maintained a largely similar morphology throughout their evolution (i.e., *Corvus* spp.). We did so by combining a well-established method to detect satellite DNA from raw short reads (RepeatExplorer2; Novák et al., 2020) and a new long-read assembly-based approach. Thanks to this combination of approaches and additional simulations, we found that the satellitome of these birds looks more GC-rich than the genomic average (contrary to what usually happens in other organisms; Talbert & Henikoff, 2020), and that the monomers are mostly arranged in a head-to-tail conformation and lack apparent secondary structures. Interestingly, despite a long history of sexual selection in BOPs, these birds share more similar satellitomes than the more recently diverged *Corvus* spp. and we hypothesise this similarity has been maintained in BOPs because of either a satDNA reservoir on the female-specific W chromosome or the possible gene flow between species (Blom et al., 2021). To our knowledge, this is the first study that reveals avian satellite evolution across deep and recent evolutionary timescales while integrating both short-read and long-read sequencing technologies.

2 | MATERIALS AND METHODS

2.1 | Samples

To search for satDNA in BOPs and *Corvus* spp. genomes, we collected genomic data for 16 species of birds-of-paradise (BOPs) and eight *Corvus* spp. as (linked) short-read libraries, long-read libraries, linked-read genome assemblies and long-read genome assemblies, together with RNA-seq libraries for one species of BOPs and two *Corvus* spp.

First, we gathered 10X Genomics Chromium linked reads (barcoded short reads for haplotype phasing; Weisenfeld et al., 2017) and corresponding Supernova2 pseudohaploid linked-read genome assemblies of 13 BOPs from a previous study (Peona et al., 2022). *Cicinnurus magnificus*, *Cicinnurus regius*, *Paradisaea rubra*, *Astrapia rothschildi*, *Epimachus meyeri*, *Ptiloris intercedens*, *Ptiloris magnificus*, *Parotia helenae*, *Parotia lawesi*, *Manucodia keraudrenii*, and *Manucodia chalybata* were females; *Drepanornis albertisi* and *Paradigalla brevicauda* were males. We also collected the 10X Genomics Chromium linked reads and genome assemblies of four females and one male of *Lycocorax pyrrhopterus* (Peona, Blom, et al., 2021; Peona, Palacios-Gimenez, et al., 2021) together with an RNA-seq library of pectoral muscle for one of the females (Peona, Blom, et al., 2021), and Illumina short-read resequencing libraries for *Astrapia stephaniae* and *Paradisaea raggiana* (Blom et al., 2021; Xu et al., 2019). Finally, we used the multiplatform chromosome-level genome assembly of *Lycocorax pyrrhopterus* (Peona, Blom, et al., 2021), the PacBio long-read genome assembly of *Ptiloris intercedens* (Peona et al., 2022), and the corresponding raw long reads (subreads) from each data set.

Regarding the *Corvus* spp., we collected publicly available Illumina short-read resequencing libraries from Kutschera et al. (2020) and Weissensteiner et al. (2017) for *Corvus brachyrhynchus*, *Corvus corone cornix* (3 individuals of unspecified sex), *Corvus corone corone* (3 individuals of unspecified sex), *Corvus corone cornix* × *corone* (3 hybrid females and 3 hybrid males), *Corvus dauuricus*, *Corvus monedula loides*, *Corvus splendens*, and *Corvus woodfordi*. Finally, one RNA-seq library for *Corvus corone cornix* (forebrain) and one for *Corvus corone corone* (brain hypothalamus and pituitary) were retrieved from Poelstra et al. (2014). All the accession numbers for the libraries and assemblies of *Corvus* spp. and BOPs are given in Table S1.

2.2 | De novo satellite DNA characterisation

To de novo characterise satellite DNA sequences from the (linked) short-read libraries of BOPs and *Corvus* spp., we ran RepeatExplorer2 (Novák et al., 2020) on the Galaxy server (https://repeatexplorer-elixir.cerit-sc.cz). Reads from each genomic library were filtered for quality (Q > 30), trimmed for low-quality bases, and adapters were removed following the recommended procedure from the RepeatExplorer2 protocol (Novák et al., 2020). Then, the processed libraries were randomly sampled to generate sublibraries of one million paired-end reads corresponding to ~0.2X genomic coverage using seqkt (https://github.com/lh3/seqtk). The subsampled libraries were uploaded onto the Galaxy server online, and the RepeatExplorer2 analysis was then run separately for each species selecting the REXdb Metazoa version 3.0 database (collection of TE-related proteins) and providing a custom avian repeat library (Peona, Palacios-Gimenez, et al., 2021). This avian repeat library included curated consensus sequences of TEs from birds-of-paradise and crows.

The candidate satDNA consensus sequences generated by RepeatExplorer2 were downloaded and manually curated when a genome assembly was available for the species of interest following the procedure described in Peona, Blom, et al. (2021). If a genome assembly was not available for the species (i.e., *Astrapia stephaniae*,

Paradisaea raggiana, *Corvus splendens*, *Corvus dauuricus*, *Corvus corone corone*), the monomers were not further curated after generation by RepeatExplorer2. The curation procedure consisted in aligning the consensus sequences back to the genome from the same species with BLAST (Camacho et al., 2009) collecting the best 20 hits, and manually inspecting the final alignments produced with MAFFT (Katoh et al., 2018) to better determine the sequence and length of the satellite monomers. Sequences that were not found tandemly repeated in the genome assemblies were discarded from the library. Similarities with TEs were detected by running RepeatMasker on the satDNA monomers with the avian repeat library from Peona, Palacios-Gimenez, et al. (2021).

The curated consensus sequences were divided into different families by aligning all sequences versus all with BLAST. Sequences that shared 80% of their length with others with at least 85% of identity were considered belonging to the same family. All the satellite families were named as "sat" followed by a number (identifier of the family) and a suffix of the six-letter abbreviation of the species where the consensus was found (e.g., sat1_astRot). When two or more consensus sequences from the same family were found in the same species, a letter was added as further suffix after the family name (e.g., sat1_a_lycPyr and sat1_b_lycPyr), thus defining subfamilies.

The nomenclature of crowSat and bopSat monomers followed a different naming system. Three different crowSat monomers were previously described in Weissensteiner et al. (2017, 2020) and belong to three separate families. Two of the three different bopSat monomers (bopSat1 and bopSat2) were first found in Peona, Blom, et al. (2021) and they are part of the same family as crowSat1. bopSat3 found in this study belongs to the same family as bopSat1 and bopSat2.

Long satellite DNA arrays are one of the major causes for assembly (contig) fragmentation (Peona, Blom, et al., 2021; Peona et al., 2018), therefore contig extremities even of long-read assemblies are expected to be enriched for satDNA sequences and this feature can be useful to characterise otherwise overlooked satDNAs. With this in mind, we designed PipeSat, a Snakemake (Köster & Rahmann, 2012) pipeline that collects the 1-kb sequences directly adjacent to contig and scaffold extremities or to gaps longer than 10 "N" nucleotides. PipeSat then runs a de novo repeat annotation on this set of gap-adjacent sequences using RepeatModeler2. PipeSat is available on GitHub (https://github.com/ValentinaBoP/CorvidesSat/PipeSat). Note that while we expect PipeSat to work on any genome assembly of interest, we recommend running it on long-read assemblies where gaps can be expected to be enriched for satDNAs and other large tandem repetitive regions (Peona, Blom, et al., 2021; Peona et al., 2018).

The consensus sequences resulting from PipeSat can contain any type of sequences that are enriched at contig extremities, therefore we removed all sequences with homology to genes or TEs after aligning them to the NCBI nonredundant nucleotide database with BLAST (Camacho et al., 2009; Wheeler et al., 2003) and masking them with the Repbase library on the CENSOR webserver (Bao et al., 2015). Then, we aligned the rest of the sequences back to the genome to produce an alignment of the 20 best hits for each raw consensus sequence as described above for the consensus sequences produced by RepeatExplorer2. Each alignment was carefully inspected for evidence of tandem repetitions homologous to the raw consensus, and a curated consensus sequence was generated after determining the precise sequence and length of the satDNA monomer.

Self-dotplots of the satDNA monomers were generated with Flexidot (Seibt et al., 2018) and the plots were annotated with the results of a further alignment between the monomers and the avian transposable element library. Finally, the phylogeny of the monomers was produced by the alignment-free clustering tool Alfpy (Zielezinski et al., 2017).

2.3 | Secondary/tertiary structures

All satDNA consensus sequences were analysed for the presence of palindromes (sequences that potentially can form secondary structures), potential folding structures and for the presence of G-quadruplex motifs (Sahakyan et al., 2017).

EMBOSS palindrome (Rice et al., 2000) was run to find palindromes using the same parameters as in Kasinathan & Henikoff (2018) (minimum palindrome length 5; maximum palindrome length 100; maximum gap length between repeated regions 20 and allowing for overlapping results). The curated consensus sequences were also analysed with RNAfold 2.4.17 (Gruber et al., 2008; Lorenz et al., 2011) to find secondary structures with the same options used in Kasinathan & Henikoff (2018) (–noGU –noconv –noPS –paramFile=dna_mathews2004.par –p –g). Finally, Quadron (Sahakyan et al., 2017) was run to find G-quadruplex motifs with default settings and filtering for G-quadruplex motifs with a score higher than 19 as suggested by the developers of the tool (Sahakyan et al., 2017) and as implemented before in birds (Peona, Blom, et al., 2021).

2.4 | Structure of satellite DNA arrays in long reads

To investigate the structure of satDNA arrays, we directly searched for arrays in the PacBio long subreads available for *Lycocorax pyrrhopterus* and *Ptiloris intercedens*. To do so, we masked the long reads with RepeatMasker (Smit et al., 2015) using the custom satDNA library produced here and filtered the RepeatMasker output for reads where >80% of their length were homologous to the satDNA library. Afterwards, we counted the number of monomers in each read and classified the arrays on the basis of the monomer content. Then, we calculated the frequency of occurrence of each type of array in the two species within the satDNA-containing reads (3,388 reads for *L. pyrrhopterus* and 12,192 reads for *P. intercedens*).

2.5 | Abundance, divergence, and transcription of satellite DNA

To investigate the satDNA content in all sampled species, we ran RepeatMasker with the custom satDNA library (consensus sequences in form of satDNA dimers) on (linked) short-read libraries, linked-read genome assemblies, and long-read genome assemblies.

First, we masked the sampled (linked) short-read libraries of all the 24 avian species used for the analysis with RepeatExplorer2, calculated the abundance of each satDNA family as the proportion of base pairs masked in the reads over the total size of the sampled libraries. Then we also ran RepeatMasker on the 10X Genomics linked-read genome assemblies, PacBio long-read libraries, and PacBio long-read genome assemblies of *Lycocorax pyrrhopterus* and *Ptiloris intercedens*.

The RepeatMasker output was then processed with the calcDivergenceFromAlign.pl script from the RepeatMasker suite (https://github.com/rmhubley/RepeatMasker/blob/master/util/calcDivergenceFromAlign.pl) to recalculate the 2-parameter Kimura distance (divergence from consensus) by correcting for the presence of CpG sites. The processed RepeatMasker output was then visualised as divergence landscapes by calculating the abundance (percentage of reads) of each satDNA family at the different levels of divergence (bin size of 1%). Only satDNA families that masked more than 50 kb of reads in at least one species were kept for the abundance analysis. Pairwise comparisons between the sets of abundances estimated in *Lycocorax pyrrhopterus* and *Ptiloris intercedens* with different sequencing technologies were conducted by calculating linear regression models to understand if the abundances estimated with one technology can predict the abundances found by another.

Finally, we investigated the presence of satDNA families in the transcriptome by mapping one RNA-seq library each of *Lycocorax pyrrhopterus* pectoral muscle, *Corvus corone cornix* forebrain, and *Corvus corone corone* brain hypothalamus and pituitary (Table S1) to the satDNA consensus sequences with BWA (Li & Durbin, 2009). The mapping was performed on dimers rather than on monomers to allow reads to map at the interface of two monomers. Then, we filtered the resulting BAM file for alignments with a quality score higher than 30 using samtools (Li et al., 2009) and quantified the transcribed satDNA by calculating the reads per kilobase million (RPKM) value for each satellite DNA consensus sequence.

2.6 | Simulations

To test the robustness of the satellite DNA quantification using short reads, we simulated genomes with controlled diversity and percentages of satellite DNA. We also simulated short-read sequencing on those genomes. The simulated sequencing data were used to characterise and compare observed and expected satDNA quantifications.

First, we simulated three genomes with 1%, 3%, and 5%, respectively, of satellite content using the *Lycocorax pyrrhopterus* satDNA library and its genome assembly as baseline. All the satDNA sequences annotated in the genome assembly by RepeatMasker were hard-masked as N nucleotides and subsequently all the N nucleotides, including assembly gaps, were removed. This assembly devoid of satDNA and gaps, together with the *Lycocorax pyrrhopterus* satDNA library of monomers, were used as input for the script genomeSimulation.R (https://github.com/ValentinaBoP/CorvidesSat) that introduced satDNA arrays of random lengths at random locations until reaching the established percentage of the genome occupied by satDNA (1%, 3%, 5%).

Then, we simulated short-read sequencing data at 30X of each of the three simulated genomes with wgsim (https://github.com/lh3/wgsim). The three simulated short-read libraries were subsampled at 0.3X, 0.5X, 1X and 2X coverage to be used by RepeatExplorer2 for de novo characterisation of satDNA monomers (as done for the real genomic data). In total, we collected 12 libraries of satDNA monomers, one for each percentage of satDNA present in the genome and for each level of library coverage. This way, we were able to test whether the coverage of the short-read libraries influenced satDNA identification and abundance estimations. The coverages of 0.3X and 0.5X were selected because these values lie within the coverage range suggested by the authors of RepeatExplorer2 (Novák et al., 2020), while the rest was chosen to further test the analysis steps.

Finally, we annotated satDNA of each subsampled library with their respective satDNA monomers (as found by RepeatExplorer2) and with the original *Lycocorax pyrrhopterus* library (used for simulating the genomes). In addition to the subsampled libraries, we also annotated satDNA in the simulated genomes using each of the 12 satDNA libraries separately as well as the original satDNA library.

3 | RESULTS

3.1 | Sequence characteristics of the avian satellitome

We ran RepeatExplorer2 (Novák et al., 2020) on raw (linked) short reads and our new PipeSat pipeline on long-read assemblies to get an as complete as possible de novo characterisation of the satellitome of the 24 sampled Corvoidea species. We obtained, in total, 165 consensus sequences of candidate satDNAs from RepeatExplorer2 and 20 from PipeSat. All consensus sequences were manually inspected to ensure that satDNA monomer sequences were correctly identified. The manual inspection was an important step for both RepeatExplorer2 and PipeSat outputs. RepeatExplorer2 bases its classification on a graph-based clustering of sequencing reads, which resulted here in the occasional misclassification of LTR retrotransposons as satDNA. PipeSat relies on RepeatModeler2 (Flynn et al., 2020) repeat prediction of only sequences at the contig extremities of an assembly, therefore its output often contained sequences of multicopy genes and interspersed repeats. After the curation process, 150 out of 165 RepeatExplorer2 consensus sequences and four out of 20 PipeSat consensus sequences were kept as satellite

monomers in the Corvoidea satellite library (Table S2) while the rest was discarded as probably being multicopy genes or interspersed repeats. We merged this collection of satDNA consensus sequences with three satDNA consensus sequences already characterised in *Lycocorax pyrrhopterus* (Peona, Blom, et al., 2021) and three in *Corvus corone cornix* (Weissensteiner et al., 2017), for a total of 160 sequences. Of the 154 new consensus sequences detected here, 69 were first found in *Corvus* spp. and 85 in BOPs. The satDNA consensus sequences that shared more than 85% similarity over 80% of their lengths were clustered, obtaining 61 distinct satDNA families (Table S2) that were named numerically and with a suffix indicating the species they were first identified in Methods.

After the detection, curation, and clustering steps, we investigated sequence characteristics of the monomers of satDNA consensus sequences such as their length distribution, base composition, potential presence of palindromes, and secondary structures. The length of the obtained monomers ranged from 20 bp to 4 kb, and most showed a length ranging between 130 and 200 bp (Figure 1a,b). The previously characterised bopSat1, bopSat2 and bopSat3 monomers shared a high degree of similarity with the crowSat1 monomer. The crowSat1 consensus sequence was previously reported to be >14 kb long (Weissensteiner et al., 2017) but actually represented a heterologous higher-order repeat in which the smallest monomer was ~1.4 kb (Figures S1–S3). The three monomers bopSat1, bopSat2, and bopSat3 were part of the same family but their nomenclature differs from the rest of the monomers curated here because the family was previously described elsewhere (Peona, Blom, et al., 2021). The homologous regions between the three different bopSat monomers and crowSat1 corresponded to a rearranged version of the crowSat1 monomer (Figure S2). The three different bopSat monomers present in *Lycocorax pyrrhopterus* were 1.2–4.7 kb long but only bopSat1 was found throughout the BOP phylogeny (with different levels of abundance; Figures 2 and 3).

Regarding base composition, all the satDNA consensus sequences had a GC content equal or higher than the mean genomic GC content (40%–42%) of the sampled genomes of *Corvus* spp. and BOPs (dotted lines in Figure 1c, Table S3). 119 out of 160 consensus sequences had a GC content between 50% and 60%, while the remaining ranged from a minimum of 43% to a maximum of 70%. We then checked for the presence of palindromes and secondary structures within the monomers with EMBOSS palindrome (Rice et al., 2000), RNAfold (Gruber et al., 2008; Lorenz et al., 2011), and Quadron (Sahakyan et al., 2017). These analyses did not reveal the presence of either palindromic sequences or secondary structures due to the nonsignificant scores for stable secondary structures and G-quadruplexes (Figure S4). However, it must be noted that solely bioinformatic approaches may fall short in properly detecting such structures.

Afterwards, we explored the structure of the satDNA arrays in species with available long reads (*Lycocorax pyrrhopterus* and *Ptiloris intercedens*). We did so by running RepeatMasker on the PacBio subreads with the satDNA library and filtering for those reads masked as satDNA for at least 80% of their length. From this approach, we found that the most common arrays in *Lycocorax pyrrhopterus* were made up of bopSat1, a composite unit of sat1+sat13, bopSat2, and a composite unit of sat6+sat15 (Table S4) which formed arrays up to 33, 36, 27, and 18 kb length in the reads, respectively. In *Ptiloris intercedens*, the most common arrays were made of a composite unit of sat1+sat13, sat6, and bopSat1 (Table S4) which formed arrays up to 55, 43, and 41 kb length, respectively. The lengths of these arrays are bound to the maximum length of the long reads, therefore they may not reflect the true maximum lengths of the arrays. Note that the read length N50 for the PacBio libraries was 20 kb for *Lycocorax pyrrhopterus* (Peona, Blom, et al., 2021) and 10 kb for *Ptiloris intercedens* (Peona et al., 2022). All the arrays found had a head-to-tail arrangements of the monomers (Table S5). In addition to this, we also annotated the PacBio reads of *Corvus corone cornix* (Weissensteiner et al., 2017) to check for the presence of satDNA arrays showing an association between crowSat1 and other monomers that could indicate a transition between the candidate pericentromeric crowSat1 and the centromeric repeats. We found a common association of crowSat1 with several monomers: sat17, sat20, sat30, sat34, sat35, sat37, sat50, sat53 and, sat55 (here listed in alphabetical order; Table S6).

Finally, we looked for homology between the satDNA sequences and previously described TEs of birds (Peona, Blom, et al., 2021; Peona, Palacios-Gimenez, et al., 2021; Prost et al., 2019; Weissensteiner et al., 2020). We found that many satDNA sequences contained pieces of retrotransposons, namely Chicken Repeat 1 long interspersed elements (CR1 LINEs) and endogenous retroviruses (ERVs; Table S7, Figure S3).

3.2 | Abundance, divergence, and transcription of the satellitome

To detect the presence and abundance of each satDNA family in *Corvus* spp. and BOPs, we ran RepeatMasker (Smit et al., 2015) with our satDNA library on subsampled short-read libraries for all species, either in the form of 10X Genomics Chromium linked reads (i.e., short reads linked by unique barcodes; Methods) or short reads from Illumina resequencing. The abundance of every satellite DNA family detected in raw reads (Figure 2, Table S8) ranged from 0% to 3%. There were 17 satDNA families shared between *Corvus* spp. and BOPs (sat1, sat9, sat17, sat30, sat32, sat35, sat36, sat37, sat47, sat50, sat54, sat55, sat57, bopSat1, crowSat1, crowSat2, crowSat3). In addition, 14 families were present only in BOPs or some species thereof (sat2, sat6, sat8, sat13, sat14, sat15, sat16, sat24, sat26, sat28, sat56, sat58, bopSat2, bopSat3) and 10 families were specific to the *Corvus* genus (sat18, sat19, sat20, sat36, sat38, sat39, sat44, sat45, sat49, sat60). Only those satDNA families that occupied at least 50 kb of reads in at least one species were reported (41 families; Figure 2) while the remaining (20 families) were discarded from the abundance analysis. All these families are widely present on all chromosome models of *Lycocorax pyrrhopterus* with a few exceptions that can be found only on a couple of chromosomes (Table S9).

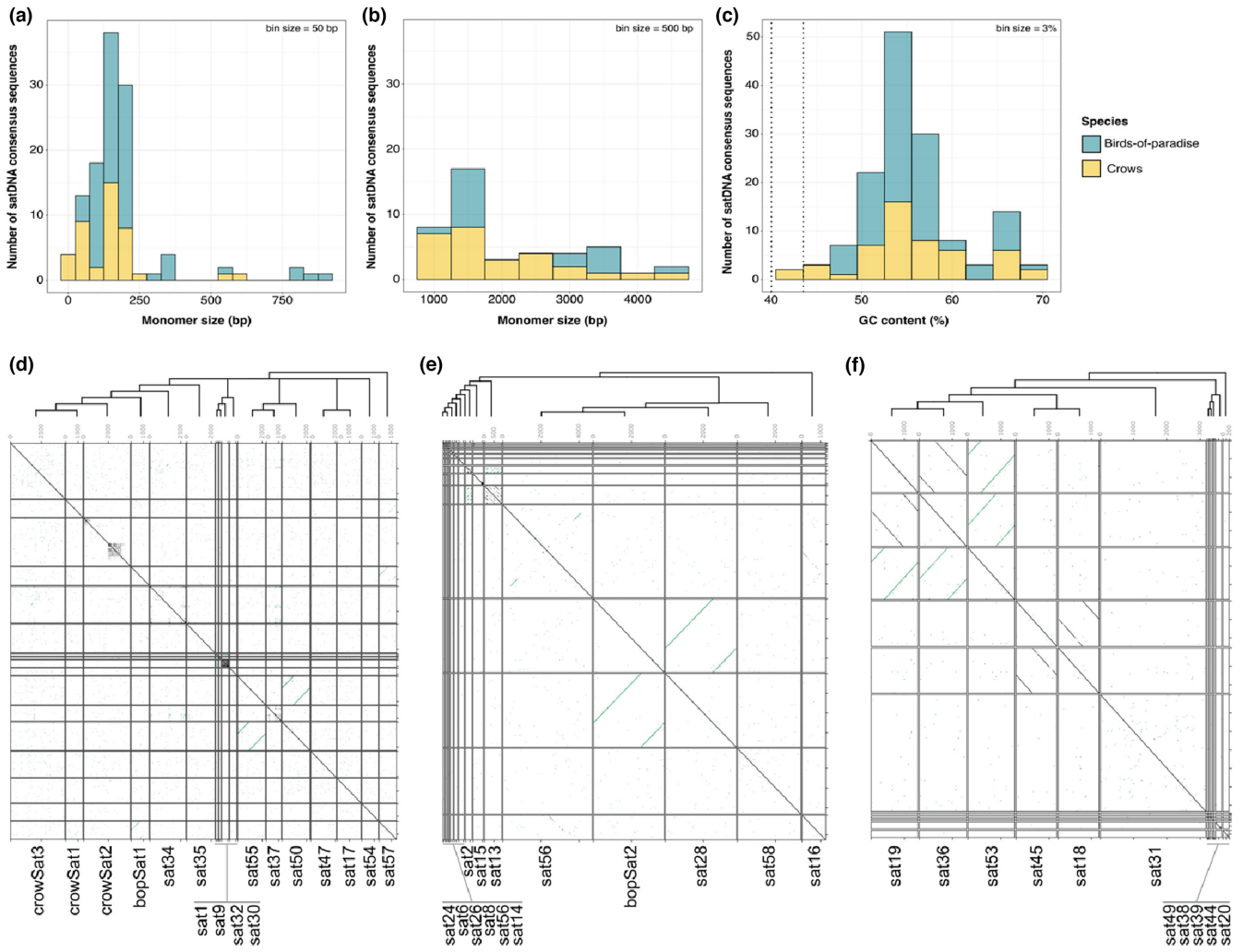

FIGURE 1 Sequence and length characteristics of monomers of the satDNA consensus sequences detected in crow and birds-of-paradise species. (a) Size distribution of monomers shorter than 1000 bp (bin size of 50 bp). (b) Size distribution of monomers longer than 1000 bp (bin size of 500 bp). (c) GC content distribution of the monomers (bin size of 3%). The black dotted lines represent the range of the mean genomic GC content in crow and birds-of-paradise species. (d–f) Self-dotplots and clustering of the consensus sequences of the satDNA families annotated for at least 50 kb in at least in one species divided by shared (d), birds-of-paradise-specific (e), and Corvus-specific (f) families

The total satDNA content in the (linked) short reads of each species ranged between a minimum of 1.12% in *Paradisaea raggiana* and a maximum of 9.7% in *Corvus splendens*, with an average content of 3.8% (Figure 3 and Table S10). Using the RepeatMasker output, we visualised the landscape of divergence of the total satDNA content as bins of Kimura two-parameter distance between the repeat copies and their consensus sequences (Figure 3). *Corvus* spp. (Figure S5) showed a landscape dominated by crowSat1, sat20, and sat37 especially at low levels of divergence (0%–5%). BOPs (Figure S5) instead showed a landscape dominated by sat1 and sat6, while crowSat families appeared at low abundance and high levels of divergence. The landscapes of *Manucodia chalybatus* and *Lycocorax pyrrhopterus* showed a high proportion of bopSat1 repeats at low divergence levels. In general, the landscapes of *Corvus* spp. were left-skewed while BOPs (except for the left-skewed *Lycocorax pyrrhopterus*) had similar relative abundances of satDNA families with a range of divergence levels.

Finally, we used RNA-seq libraries of *Lycocorax pyrrhopterus*, *Corvus corone cornix*, and *Corvus corone corone* to investigate whether any satellite DNA monomers were transcribed and in what quantity. Since centromeric satellite DNA is actively transcribed in cells (Lyn et al., 2012), RNA-seq data may provide an additional hint to identify candidate satellite DNA monomers that constitute the centromeres. After mapping the RNA-seq data to the satellite DNA library and filtering for mapping quality, we found evidence of transcription for sat1 in *Lycocorax pyrrhopterus* pectoral muscle (RPKM value of 123.31), sat1, sat9, sat20, and sat54 in *Corvus corone corone* brain hypothalamus and pituitary (RPKM values of 5.7, 50.7, 2.9, and 11.1), and sat1, sat9, sat20, and crowSat1 in *Corvus corone cornix* forebrain (RPKM values of 4.31, 30.7, 10.88, and 0.7; Table S11).

3.3 | Satellite DNA abundance variance within species, hybrids, and sequencing technologies

After investigating the presence, absence, and abundance of satDNA families between species, we focused on detecting abundance differences thereof between individuals of the same species, that is, in (linked) short-read sequencing libraries of five *Lycocorax pyrrhopterus*, three *Corvus corone cornix*, three *Corvus corone corone*, and five *Corvus corone cornix* × *corone* individuals (Figure 4, Table S12).

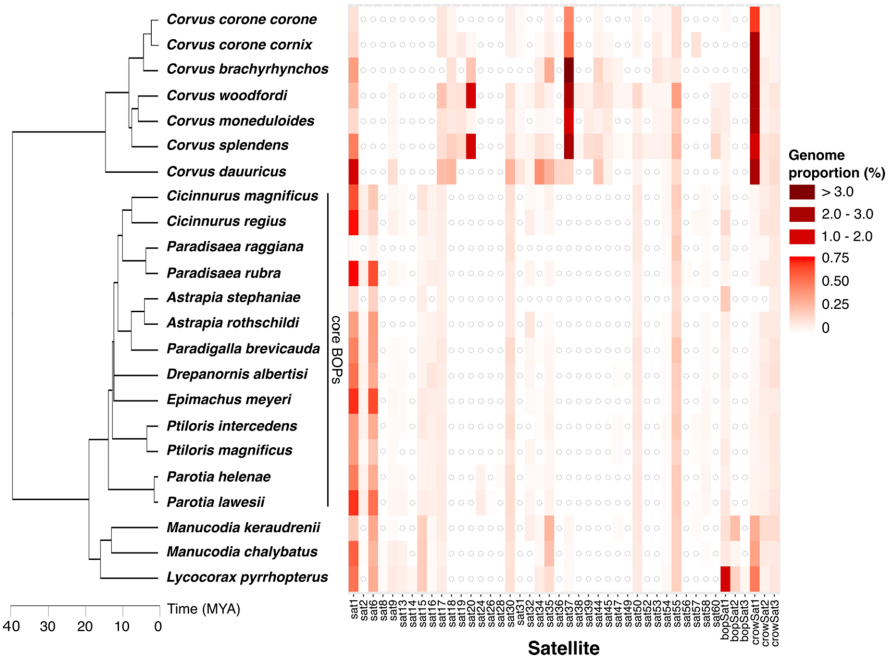

FIGURE 2 Presence/absence and abundance of the different satDNA families across crow and birds-of-paradise (BOPs). Abundances were calculated as the proportion of raw (linked) short reads masked as satellite DNA over the total length of the sampled reads. The satDNA families that are completely absent are marked with a grey circle on a white background. Only the 41 out of 61 satDNA families that occupied at least 50 kb of reads in at least one species are shown. BOP species belonging to the core BOP clade are marked by a vertical black line. Dated phylogeny obtained from Timetree.org (Kumar et al., 2017)

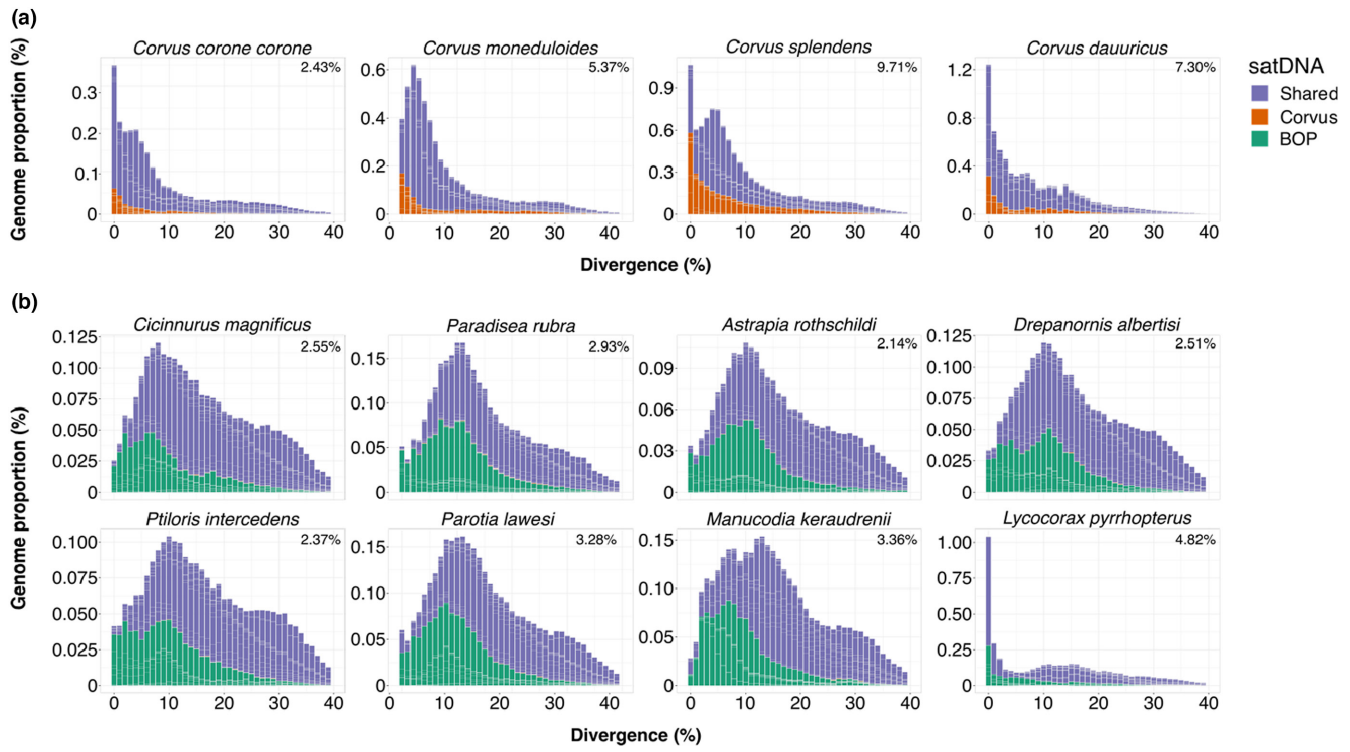

FIGURE 3 Divergence landscapes of the satDNA families detected in the raw short reads of a selection of crow species (a) and birds-of-paradise (b). Only 41 out of 61 satDNA families that occupied at least 50 kb of reads in at least one species are shown. The divergence between the masked satellite sequences and their consensus is shown on the x-axis as a Kimura two-parameter distance and the genome proportion of the satellite sequences is shown as percentage of the sampled reads on the y-axis. The percentage reported over each plot indicates the total satDNA content for that genome. The landscapes of individual satDNA families in all sampled species can be found in Figure S5

satDNA abundance differences between *Lycocorax pyrrhopterus* individuals were mostly limited to sat1, bopSat1, and crowSat1. The families sat14 and sat54 were present at low frequencies in only one or two individuals and completely absent in the others (Figure 4a) suggesting satDNA presence/absence between individuals. In *Corvus* spp. (Figure 4b), the *Corvus corone cornix* (corCon) individuals appeared to be homogeneous for the presence and abundance of satDNA families, except for six families present in only a subset of individuals and at very low abundances. *Corvus corone corone* (corCor) individuals showed a similar pattern where the majority of satDNA families were present in all individuals in similar quantities, but six low-abundance families showed a patchy distribution among the individuals. Finally, the satellitome of hybrid individuals appeared similar to their parental species with only minor differences in quantity (Figure 4b).

Since we had 10X Genomics linked-read libraries, 10X Genomics linked-read assemblies, PacBio long subreads, and PacBio long-read assemblies for both *Lycocorax pyrrhopterus* and *Ptiloris intercedens*, we were also able to detect differences in satDNA presence and abundance in the same individual depending on the sequencing technologies adopted (Figure 4c, Table S13). In both species, the number and types of satDNA families largely agreed between the subsampled linked short-read libraries ("lycPyr F1 reads" and "ptiInt reads") and the PacBio assemblies ("lycPyr F1 PB" and "ptiInt PB"), even though the relative abundances varied. On the other hand, the 10× Genomics assemblies ("lycPyr F1 10X" and "ptiInt 10X") showed a drastically lower relative abundance for most of the satDNA families. Some families were masked at very low abundance in the PacBio assemblies while they were absent in the linked short reads, namely sat2, sat10, sat18, sat23, sat31, sat32, sat 37, sat47, sat54 in "lycPyr F1 PB" versus "lycPyr F1 10X reads", and sat12, sat28, bopSat2, bopSat3 in "ptiInt PB" versus "ptiInt 10X reads". In other cases, families were more represented in the PacBio assembly (sat1, sat13, sat30, sat37) or in the linked short reads (sat1, sat58, crowSat1) of *Lycocorax pyrrhopterus*. Instead, in *Ptiloris intercedens*, five families (sat6, sat13, sat30, sat37, sat50) were more abundant in the PacBio assembly and sat1 was more present in the linked short reads. In addition, sat1 and bopSat1 were more present in the PacBio subreads than in the respective PacBio assemblies, suggesting that these satDNA families probably collapsed during the assembly process. In general, the sets of abundances found by the different technologies do not show significant correlations ($R^2 \sim 0$ and $p > 0.5$; Table S14) between one another.

Given that the comparisons of the proportions of the satDNA families with different sequencing technologies showed that there were discrepancies (Figure 4c), we carried out simulations with the objective of testing how reliable our quantifications with short reads are (Figure 5, Table S15). Briefly, we simulated three *Lycocorax pyrrhopterus* genomes with 1%, 3%, and 5% of satDNA from which we simulated short-read sequencing data (compare to ~4% identified in the empirical *L. pyrrhopterus* genome). The short reads were subsampled to 0.3X, 0.5X, 1X, and 2X depth of coverage and satDNA monomers were identified using RepeatExplorer2. We annotated the simulated genomes and short reads with the satDNA libraries derived from each of the subsampled simulated data. We found that the satDNA abundance in the subsampled short-read libraries was underestimated in the annotation when using the new satDNA libraries (Figure 5a) but the use of the complete and original satDNA library yielded the correct abundances (Figure 5b). A similar pattern was observed when the genome assemblies were annotated (Figure 5c).

After estimating the proportions of the satDNA families in 16 species of BOPs and seven *Corvus* spp., we carried out a principal component analysis (PCA) to summarise the genomic differentiation of satDNA abundances in these species (Figure 6). We found that BOP species and *Corvus* spp. were well differentiated on the first principal component (PC1). All the BOP species clustered together on the PC1, though *Lycocorax pyrrhopterus* was separated from the rest of the BOPs on the PC2 (Figure 6a). While *Corvus* spp. were separated from the BOP cluster, they were scattered more along both principal components and no geographical clustering was observed for *Corvus* spp. and BOPs (Heads, 2001; Jønsson et al., 2016), except for *Corvus corone corone* and *Corvus corone cornix* that share a hybrid zone (Figure 6a). To detect possibly significant differences in the general satellitome between parental and hybrid species, the same analysis was carried out for *Corvus corone cornix*, *Corvus corone corone*, and hybrids thereof (Figure 6b). In this sampling, corCor_I1 was from Spain, corCor_I2 and corCor_I3 from Germany; corCon_I1 and corCon_I2 from Poland, and corCon_I3 from Sweden. All the hybrids were sampled from Ireland (Table S1). The parental individuals used here came from allopatric populations far from the hybrid zones, therefore on the basis of SNP diversity, the two species were expected to cluster separately (Vijay et al., 2016). Furthermore, the Spanish individual was expected to differ from the German individuals because of the higher degree of population differentiation (F_{ST}) previously found between the two populations (Vijay et al., 2016). In line with these expectations, in the PCA *Corvus corone cornix* individuals clustered closely together in the first two principal components while *Corvus corone corone* individuals were spread along both components. The two German *Corvus corone corone* individuals were close on the PC1 and farther from the Spanish individual. The hybrid individuals were clustered into two smaller groups at the extremities of the spread of the parental species along PC1, while remaining one cluster on PC2. The two hybrid clusters did not separate based on sex.

4 | DISCUSSION

The repetitive genomic portion represented by satDNA still remains vastly unexplored in non-model organisms. This undercharacterisation is mostly due to the fact that satellite DNA is a main component of the so-called "genomic dark matter", therefore largely missing from genome assemblies (Peona, Blom, et al., 2021; Peona et al., 2018; Thomma et al., 2016) including, as we quantified previously, avian genomes (Peona, Blom, et al., 2021). The study of

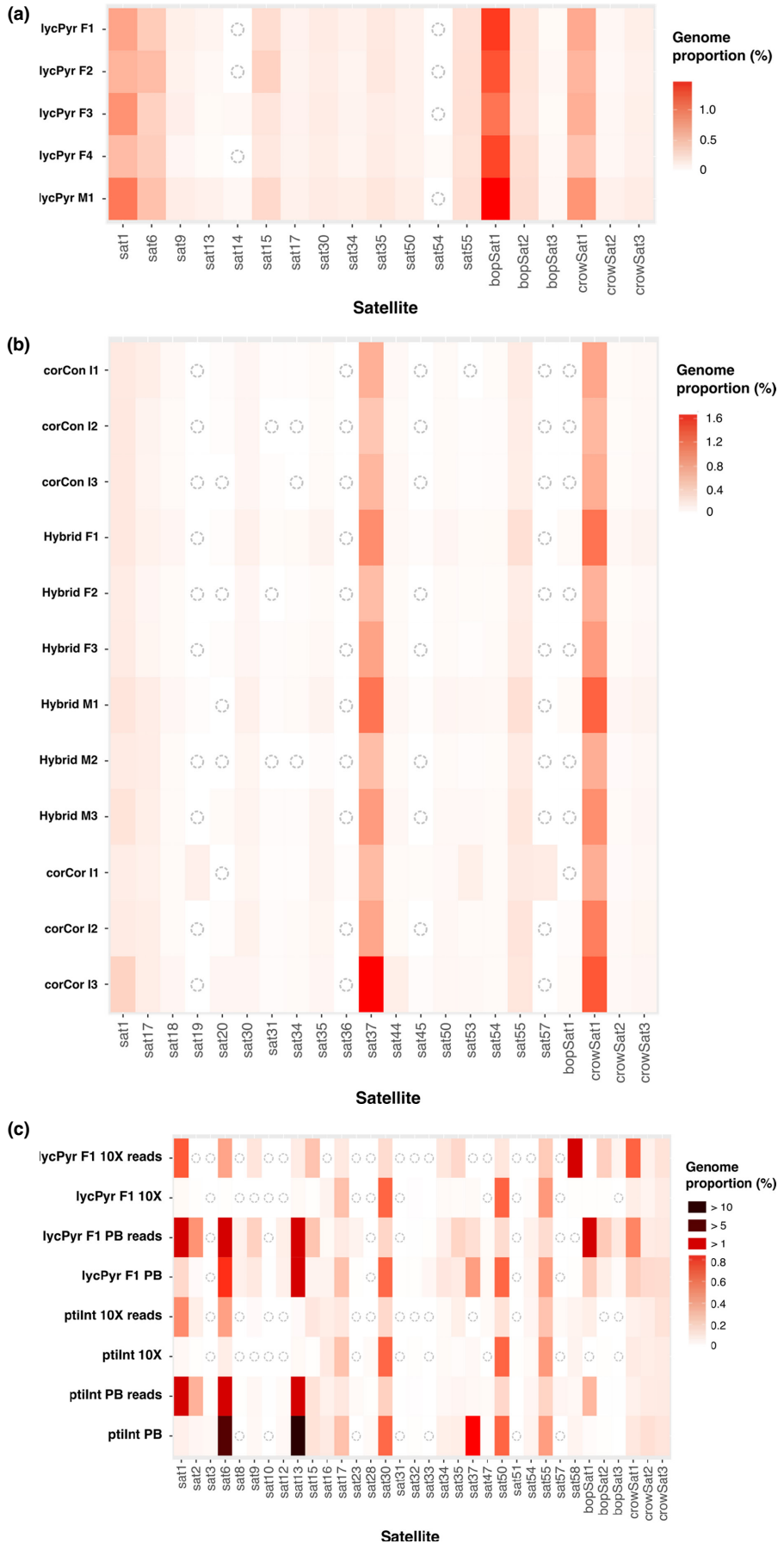
(a)
lycPyr F1
lycPyr F2
lycPyr F3
lycPyr F4
lycPyr M1
sat1 sat6 sat9 sat13 sat14 sat15 sat17 sat30 sat34 sat35 sat50 sat54 sat55 bopSat1 bopSat2 bopSat3 crowSat1 crowSat2 crowSat3
Satellite
Genome proportion (%)
1.0
0.5
0
(b)
corCon I1
corCon I2
corCon I3
Hybrid F1
Hybrid F2
Hybrid F3
Hybrid M1
Hybrid M2
Hybrid M3
corCor I1
corCor I2
corCor I3
sat1 sat17 sat18 sat19 sat20 sat30 sat31 sat34 sat35 sat36 sat37 sat44 sat45 sat50 sat53 sat54 sat55 sat57 bopSat1 crowSat1 crowSat2 crowSat3
Satellite
Genome proportion (%)
1.6
1.2
0.8
0.4
0
(c)
lycPyr F1 10X reads
lycPyr F1 10X
lycPyr F1 PB reads
lycPyr F1 PB
ptiInt 10X reads
ptiInt 10X
ptiInt PB reads
ptiInt PB
sat1 sat2 sat3 sat6 sat8 sat9 sat10 sat12 sat13 sat15 sat16 sat17 sat23 sat28 sat30 sat31 sat32 sat33 sat34 sat35 sat37 sat47 sat50 sat51 sat54 sat55 sat57 sat58 bopSat1 bopSat2 bopSat3 crowSat1 crowSat2 crowSat3
Satellite
Genome proportion (%)
> 10
> 5
> 1
0.8
0.6
0.4
0.2
0

FIGURE 4 SatDNA abundance differences between individuals and data types. Abundance was calculated as the proportion of the genome (either raw reads or genome assembly) masked as satDNA. satDNA families that are completely absent in a given data set are marked with a grey circle on a white background. (a) Abundance of satDNA families in four females and one male of *Lycocorax pyrrhopterus* ("lycPyr"). (b) Abundance of satDNA families in individuals of *Corvus corone corone* ("corCor"), *Corvus corone cornix* ("corCon") and *Corvus corone cornix* × *corone* (hybrids between the former two subspecies; "Hybrid"). The sex of the individuals is indicated as F (female), M (male), or I (unspecified sex). (c) Abundance of satDNA families in different types of sequencing data produced for the same individual of *Lycocorax pyrrhopterus* ("lycPyr") and *Ptiloris intercedens* ("ptiInt"), respectively. 10X reads: subsampled raw 10X Genomics linked short reads; 10X: genome assemblies based on 10X Genomics linked reads; PB reads: PacBio long subreads; PB: genome assemblies based on PacBio long reads

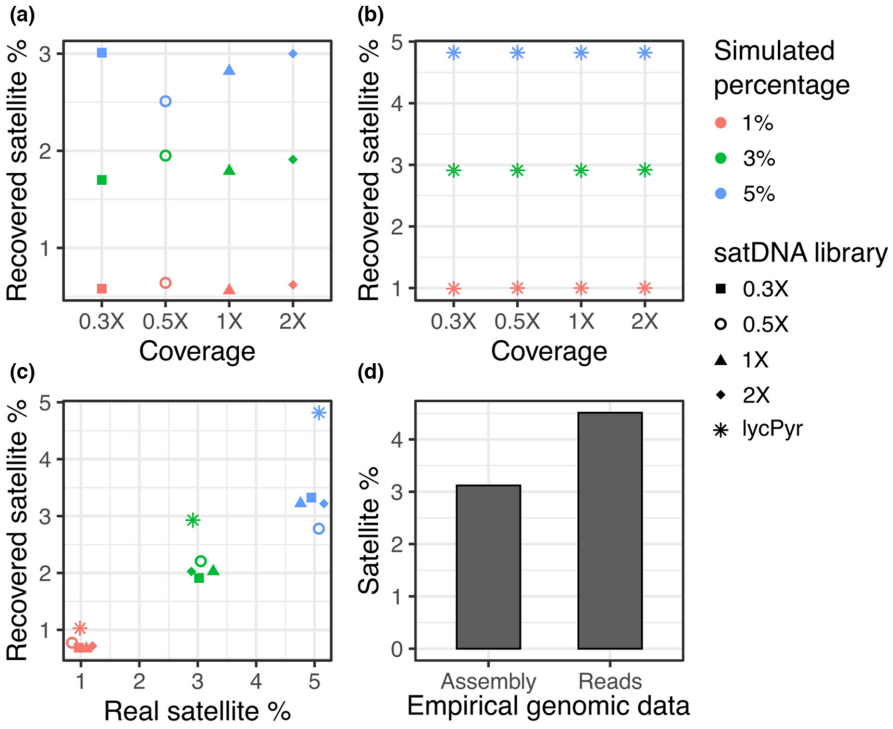

FIGURE 5 Satellite DNA abundances retrieved from simulated and empirical genomic data of *Lycocorax pyrrhopterus* (lycPyr). Three genomes were simulated with 1%, 3% and 5% satellite DNA content. Short-read sequencing libraries from the simulated genomes were produced at 0.3X, 0.5X, 1X and 2X coverage from which satDNA libraries were generated that were in turn used to annotate the simulated genomes and sequencing libraries. (a) Percentages of satellite DNA masked in the simulated sequencing libraries using the respectively derived satellite DNA libraries. (b) Percentages of satellite DNA masked in the simulated sequencing libraries using the lycPyr satellite DNA library used to simulate the genomes. (c) Percentages of satellite DNA masked in the simulated genomes using the lycPyr satellite DNA library. (d) Percentages of satellite DNA masked in the empirical genome assembly and sequencing libraries of lycPyr

satDNA in birds so far has been limited to individual species or individual satDNA families (Brown & Jones, 1972; Deryusheva et al., 2007; Liangouzov et al., 2002; Vontzou, 2021; Weissensteiner et al., 2017; Westerberg, 2020; Yamada et al., 2002, 2004), thus an evolutionary perspective on the avian satellitome is currently lacking. In this study, we widen the characterisation of satellite DNA sequences and their evolution to over 40 million years of bird evolution by analysing genomic data from 24 Corvoidae species from the *Corvus* genus and the family Paradisaeidae, by comparing within-genus and within-species data using complementary sequencing data, and by combining existing and new satDNA detection tools.

We applied two orthogonal methods to de novo characterise the satellitome of the 24 species by running RepeatExplorer2 (Novák et al., 2020) on (linked) short-read libraries and PipeSat on long-read genome assemblies. RepeatExplorer2 is a tool to that uses graph-based clustering of next-generation sequencing reads for assembling and characterisation of repetitive DNA. PipeSat is a newly developed pipeline that runs directly on the genome assemblies and takes advantage of current assembly limitations to find satDNA sequences. satDNA is one of the main causes for assembly fragmentation and therefore it is likely that satDNA sequences are found enriched at contig extremities in long-read genome assemblies (Peona, Blom, et al., 2021; Peona et al., 2018). PipeSat collects the flanking regions to contig extremities ("N" gaps or scaffold start/end positions) and runs RepeatModeler2 (Flynn et al., 2020) on such an enriched sequence library. The output of the pipeline consists of consensus sequences for repetitive regions that, due to their gap adjacency, probably cause assembly fragmentation including satDNA. We manually curated all the consensus sequences predicted by

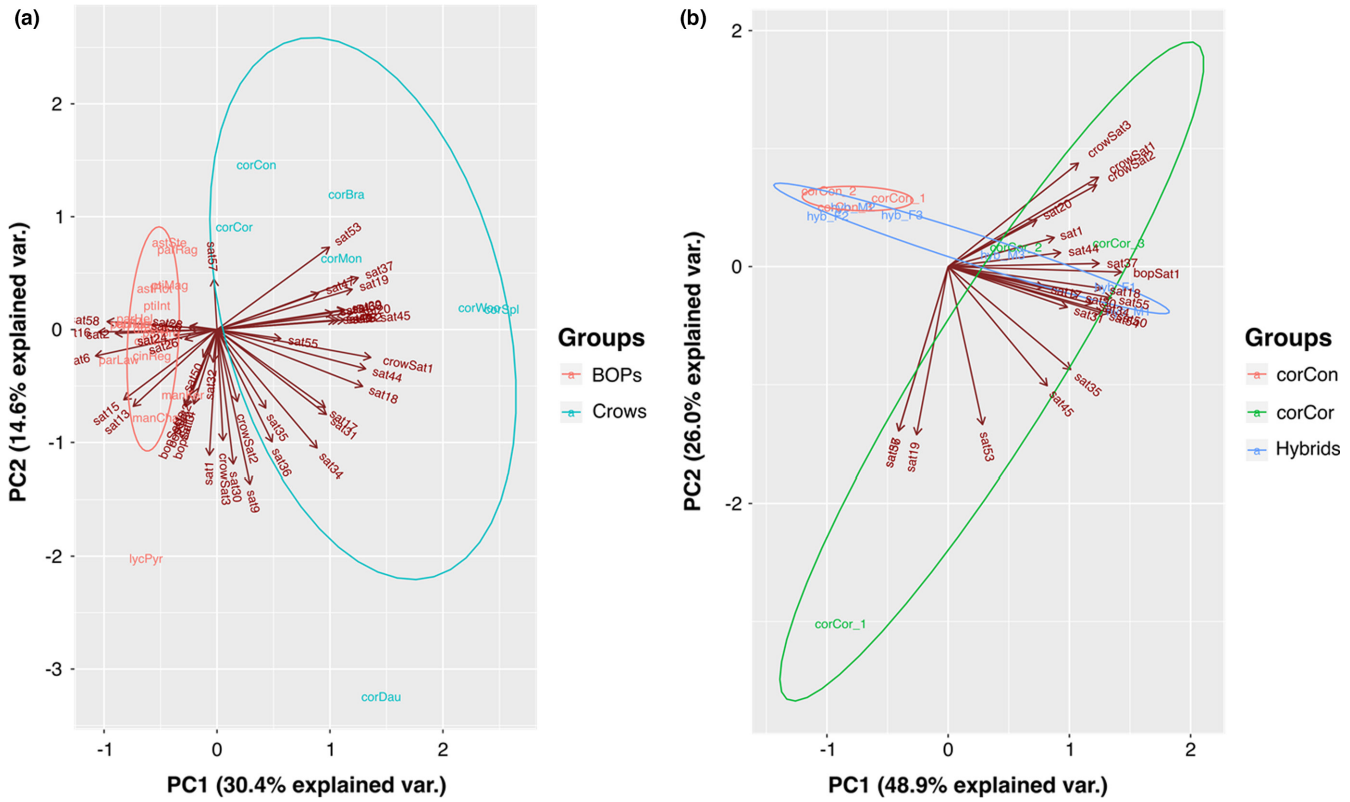

FIGURE 6 Principal component analysis (PCA) based on abundance estimates of satDNA families. (a) Comparison of birds-of-paradise (BOPs) and crow species and (b) comparison of individuals of *Corvus corone corone* ("corCor"), *Corvus corone cornix* ("corCon"), and hybrids thereof ("Hybrids"). Arrows indicate the direction of each variable (satDNA family abundance)

RepeatExplorer2 and PipeSat to make sure that they were part of the satellitome and not other types of repetitive elements.

The two methods combined yielded distinct consensus sequences for a total of 154 new satDNA monomers that were merged with satDNAs previously detected in *Lycocorax pyrrhopterus* (Peona, Blom, et al., 2021) and *Corvus corone cornix* (Weissensteiner et al., 2017), respectively, for a total of 160 consensus sequences which clustered into 61 distinct satDNA families. The length of satDNA monomers ranged from 20 bp to 4 kb (Figure 1a and b) and most had sizes between 130 and 200 bp (Figure 1a). The longest monomers (>1000 bp) were mostly identified in *Corvus* spp. (Figure 1b) while BOPs show monomers <500 bp with the exception of the three different bopSat monomers. Many of the monomers showed homologies to TEs suggesting that TEs could have acted as seeds for the formation of new satDNA. The crowSat1 monomer was first discovered in *Corvus corone cornix* (Weissensteiner et al., 2017) and found to be arranged in tandem for megabases (estimated from optical mapping data) at genomic locations corresponding to valleys of recombination rate, suggesting these arrays are associated with (peri)centromere positions. However, in absence of ChIP-seq data for centromeric proteins (e.g., CENP-A), it remains unclear whether crowSat1 arrays form the actual centromere or are rather part of the pericentromere. We hypothesise that the best candidate for being a centromeric satDNA in *Corvus* spp. is sat20 because it is the most abundant satellite at low divergences after crowSat1 and it is 190 bp long, therefore close to a nucleosome length (~147 bp) as found in centromeric satDNA from other organisms (Luger et al., 1997; Melters et al., 2013; Talbert & Henikoff, 2020). The association between crowSat1 and sat20 is one of the most commonly associations found in the *Corvus corone cornix* long reads, suggesting a possible transition from pericentromeric to centromeric regions. However, cytogenetic experiments need to confirm this hypothesis. In BOPs, sat1 (179 bp) and sat6 (76 bp) are the most abundant satDNA families and present at low levels of divergence. In long-read data, long (>20 kb) satDNA arrays (arranged in a head-to-tail fashion) formed by sat1 and sat6 are the most frequent in both *L. pyrrhopterus* and *P. intercedens*. In addition, sat1 is the only satDNA sequence found to be transcribed in the RNA-seq data for *Lycocorax pyrrhopterus*. For these reasons, we consider the two satDNA families sat1 and sat6 as the best candidates for being centromeric in BOPs.

All the satDNA consensus sequences identified here showed a GC content higher (>50%) than the mean genomic GC content (40%–42%), which is unusual for satDNA that generally tends to be AT-rich, especially if centromeric (e.g., in humans, mice, rice, *Neurospora* fungi) (Garrido-Ramos, 2015; Talbert & Henikoff, 2020). For example, the centromeric satDNA candidates sat1 and sat6 show a GC content of 55% and 56%, respectively. Similarly to *Corvus* spp. and BOPs, the centromeric repeats found in chicken (Shang et al., 2010) are not AT-rich but very balanced in their base composition with a GC content around 50%, yet higher than the mean genomic GC content of 42%.

From the chicken data and our new satellite DNA library it seems that avian satDNAs have different sequence features (Vontzou, 2021) with respect to organisms like plants, fungi and nonavian vertebrates (Talbert & Henikoff, 2020) (but see Naish et al., 2021) for very recent evidence of GC-rich centromeric satellites in *Arabidopsis thaliana*). A gradual increase in AT of individual satDNA monomers over time would be expected as a consequence of the heavy state of heterochromatinization, especially of methylation (deamination of cytosine into thymine), of the arrays (Ruiz-Ruano et al., 2016, 2019). The tendency of avian satDNA to maintain a high GC content may be linked to the pronounced GC-biased gene conversion phenomenon observed in bird genomes (Bolívar et al., 2019; Mugal et al., 2013, 2015) that could contrast the gradual loss in GC content given by time and methylation. In the Japanese quail (Brown & Jones, 1972), satellite DNA has been found to be highly enriched in microchromosomes (defined as chromosomes <20 Mb) that are in turn the most affected by the GC-biased gene conversion phenomenon mentioned above because of their increased recombination rate with respect to bigger chromosomes (Burt, 2002; Griffin & Burt, 2014). This accumulation pattern has yet to be studied at a broad phylogenetic scale, so it remains unknown if it holds for any avian species other than quail, but it could be considered as a potential explanatory factor for the high GC content of satDNAs observed in birds here.

In this study, we uncovered satellitome diversity of avian species at different evolutionary timescales and taxonomic levels. The findings showed that birds belonging to closely related families showed vastly different percentages of total satDNA content and satellitome landscapes (Figure 3 and Figure S5). *Corvus* spp. showed a higher satDNA content range (2.43%–9.71%) than BOP species (2.41%–4.82%), but the latter presented a higher similarity of landscapes between species (Figure 3). The satellitome landscapes of *Corvus* spp., in general, were dominated by fewer satDNA families with longer monomers which were largely absent in BOPs (e.g., crowSat1, sat20, sat37). In BOPs, our results suggest that bopSat1 replaced crowSat1 monomers and remained at a relatively high abundance in *Lycocorax pyrrhopterus* and *Manucodia* spp., but only very few monomers were present in the core BOP species (Figures 2 and 3). Indeed, while bopSat1 arrays were commonly found in *Lycocorax pyrrhopterus* long-read data, crowSat1 monomers in *Lycocorax pyrrhopterus* were found as short fragments scattered within other satDNA arrays (Tables S4 and S5) and thus might be remnants of ancient arrays. The *Lycocorax pyrrhopterus* satDNA divergence landscape appears more similar to the landscape of *Corvus* spp. than to the ones of the other BOPs (Figure 3) because of its left skewedness and low number of satellite families present. The *Manucodia* and *Lycocorax* genera contain species that much resemble crows in morphology and lack the mating behaviours and colourful plumage typical of core BOPs (Marshall, 1951). It is possible that these monomorphic and crow-like species, especially *Lycocorax*, have retained both morphological and genomic features similar to the last common ancestor of the Paradisaeidae family that is considered to have looked like a crow (Marshall, 1951). However, considering the results of the PCA (Figure 6a), the total satellitome of *Lycocorax pyrrhopterus* separates from the other BOPs on the second principal component but without clustering with *Corvus* spp., suggesting that the two Corvoidea groups are well differentiated. This difference between *Lycocorax pyrrhopterus* and *Corvus* spp. is probably due to the large difference in presence/absence of satDNA families, that is, of the 24 satDNA families present in *L. pyrrhopterus*, nine are not shared with any of the *Corvus* spp. (Figure 2).

One of the classic models of satellite DNA evolution is described by the "library hypothesis" (Fry & Salser, 1977; Palacios-Gimenez et al., 2020; Ruiz-Ruano et al., 2016; Salser et al., 1976), which states that related species share a common pool of satDNA sequences that expands or contracts (even to the point of disappearing) independently during the species' evolution. According to the library hypothesis, one can expect to see drastic changes in abundance and even the disappearance of satDNA families with increasing time. Instead, in the core BOP species the satDNA landscape plots (Figure 3 and Figure S5) showed similar diversity and divergence of satDNA families. We speculate that this similarity of satDNA landscapes in core BOP species might have been maintained through events of hybridisation that "replenished" the satDNA pool of the species. However, this hypothesis does not completely explain the pattern since *Manucodia* spp. are not known to have hybridised with the core BOP species (Marshall, 1951). Another speculative explanation is that the probable accumulation of satDNA families on the female-specific W chromosome in *Manucodia*, *Epimachus*, *Parotia*, and *Cicinnurus* genera (Peona et al., 2022) acted as a pool of satDNA arrays that was mobilised across the remaining genome by the many intact TEs present on the W chromosome (Peona et al., 2022; Peona, Palacios-Gimenez, et al., 2021). BOP and *Corvus* groups, according to the library hypothesis, are expected to show very different satellitome compositions given that these groups shared an ancestor >40 MYA, and indeed these species share only a portion of the satDNA families which also differ in their abundances. Surprisingly, within each of the two groups the satellitomes resulted in being more similar among the more deeply diverged species of BOPs and less similar among the more recently diverged *Corvus* spp.

Thanks to the large species data set, we were able to investigate the diversity and composition of satellitomes over long evolutionary timescales. The presence of genomic data from multiple individuals of *Lycocorax pyrrhopterus* and *Corvus* spp. additionally allowed us to get an overview of the satDNA diversity within species (Figures 4a,b and 6b). In general, satellitomes of individuals from the same species were very similar. The very few differences, for example, in *Lycocorax pyrrhopterus* (Figure 4a) where sat14 and sat54 were present only in a few samples, suggested that these satDNA families either originated very recently or were on the verge of extinction. In the case of sat14, all its monomers had a mean divergence from the consensus of ~2%, suggesting that this satDNA family originated very recently in *Lycocorax* (Figure 2) and remained at low frequencies in the population. Instead, for sat54 monomers we observed a mean divergence from the consensus of 17%, suggesting that this satDNA family is probably present as old monomers in some *Lycocorax* individuals and at low frequencies in the other BOP species (Figure 2).

Next, we investigated the satDNA content dynamics in *Corvus corone corone*, *Corvus corone cornix*, and hybrids thereof (Figures 4b and 6b). The *Corvus corone* species complex is characterised by an extensive gene flow between the two subspecies across Europe and only very few fixed SNP differences which are almost all clustered on chromosome 18 (Poelstra et al., 2014; Vijay et al., 2016; Weissensteiner et al., 2017). The abundances of the satDNA families were similar between both parental and hybrid individuals (Figure 4b). When comparing the total satDNA content of this sampling, we observed no drastic differences in abundance, probably because of the continuous gene flow with the parental subspecies (Poelstra et al., 2014), suggesting that the satDNA regions are undergoing the same pattern of gene flow as the vast majority of the *Corvus corone* genome. This hypothesis is also supported by the PCA on these species (Figure 6b) in which the hybrid individuals are scattered at the extremities of the variability of the parental subspecies, in line with asymmetric backcrossing with either of the parental subspecies (Poelstra et al., 2014; Vijay et al., 2016).

Finally, we showed that different sequencing technologies provide different satellite abundances (Figure 4c), therefore it is important to understand which characteristics of the sequencing are influencing such patterns. Our short-read sequencing simulations highlighted that, under ideal conditions, short reads and coverage do not influence the abundance quantifications if the satDNA monomer library is complete (Figure 5b,c). However, to get a complete library, methods based on both short and long reads may be necessary as well as the use of multiple species. It has been shown for TEs that the repeat library curation of two closely related species improves the annotation of both species (Boman et al., 2019), and probably the same holds for satDNA libraries. In conclusion, short reads can be more useful than assemblies (Figure 5d) to estimate satDNA abundances but, since sequencing is not necessarily evenly spread across the genome, multiple technologies can help refining abundances as well as presence/absence patterns of satDNA families. We therefore emphasise that it is key to have independent sources of genomic data to be able to uncover the diversity, quantity, and structure of satDNA monomers and arrays in a species.

5 | CONCLUSIONS

The rapid technological advances in genome sequencing have begun to provide the fascinating possibility of exploring even the most complex genomic regions. Satellite DNA is certainly part of those complex regions, indeed it is a main component of the avian genomic dark matter even of long-read assemblies as we previously found for *Lycocorax pyrrhopterus* (Peona, Blom, et al., 2021) and confirmed here for *Ptiloris intercedens*. By being part of the genomic dark matter, satDNA is systematically underrepresented in genome assemblies, therefore complicating its investigation with solely bioinformatic approaches. From the results of this study, the diversity of satDNAs can be retrieved with a combination of long and short reads; however, long-read libraries and genome assemblies are essential to study the structure of any satDNA arrays. Inferences on presence/absence and expansion/contraction of satDNA sequences can be wrongly estimated by the sequencing technologies when sequencing biases exist. In absence of sequencing biases, the completeness of the satDNA library is key for downstream analysis (Figure 5). The consistent use of the same sequencing technology throughout the sampling and analysis already helps to reduce these different technological biases.

By integrating different genomic data types, we were able to describe some satDNA features that make the avian satellitome differ from other vertebrates. For example, our species sampling suggests that the avian satellitome is GC-rich, while it is usually AT-rich in other animals (Henikoff et al., 2015; Kipling et al., 1991) and plants (Ambrožová et al., 2011; Ruiz-Ruano et al., 2019; Yan et al., 2008), and none of the satDNA consensus sequences show secondary structures as expected for vertebrate centromeric repeats (Kasinathan & Henikoff, 2018). Considering sequence characteristics such as the monomer length and array abundance, we pinned down key candidate centromeric satDNA families in *Corvus* spp. and BOPs for future cytogenetic and/or epigenetic validation.

Finally, our evolutionary analyses of the avian satellitome in *Corvus* spp. and BOPs highlighted two different modes of evolution. The satDNA landscapes in *Corvus* are more diverged from one another than those among birds-of-paradise even though *Corvus* spp. are more closely related. On the other hand, BOPs show satDNA families that remained at similar frequencies throughout the phylogeny. A global characterisation of satDNA across all major clades of birds is therefore necessary to understand which of these models is the rule versus the exception and will help identify the consequences that satDNA dynamics have for avian (genome) evolution.

AUTHOR CONTRIBUTIONS

Valentina Peona and Alexander Suh designed the study. Valentina Peona analysed the data and wrote the first manuscript draft. Valentina Peona and Alexander Suh wrote and revised the subsequent drafts. Verena E. Kutschera helped developing the Snakemake pipeline PipeSat. Mozes P. K. Blom and Martin Irestedt established the birds-of-paradise taxon sampling and provided unpublished genomic data.

ACKNOWLEDGEMENTS

We thank Marco Ricci, Francisco J. Ruiz-Ruano, Julie Blommaert, and Octavio M. Palacios-Gimenez for comments on earlier versions of this manuscript. We also thank Francisco J. Ruiz Ruano and Octavio M. Palacios-Gimenez for their help with running RepeatExplorer2. This study was supported by the Swedish Research Council Vetenskapsrådet (2016-05139 and 2020-04436 to Alexander Suh; 621-2014-5113 and 2019-03900 to Martin Irestedt), and the Swedish Research Council Formas (2017-01597 to Alexander Suh). Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through Uppsala Multidisciplinary Centre for Advanced Computational Science (UPPMAX). We also

thank the Bioinformatic Advisory Programme from Science for Life Laboratory for the support in designing the initial analysis. Verena E. Kutschera is financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All the code has been made available on GitHub (https://github.com/ValentinaBoP/CorvidesSat).

ORCID

Valentina Peona https://orcid.org/0000-0001-5119-1837
Verena E. Kutschera https://orcid.org/0000-0002-8930-534X
Mozes P. K. Blom https://orcid.org/0000-0002-6304-9827
Martin Irestedt https://orcid.org/0000-0003-1680-6861
Alexander Suh https://orcid.org/0000-0002-8979-9992

REFERENCES

Ambrožová, K., Mandáková, T., Bureš, P., Neumann, P., Leitch, I. J., Koblížková, A., Macas, J., & Lysak, M. A. (2011). Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of Fritillaria lilies. *Annals of Botany*, *107*(2), 255–268. https://doi.org/10.1093/aob/mcq235

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11. https://doi.org/10.1186/s13100-015-0041-9

Blom, M. P. K., Peona, V., Xu, L., Prost, S., Christidis, L., Benz, B. W., & Irestedt, M. (2021). A genomic perspective on intergeneric hybridization and sexual selection as drivers of extreme phenotypic change in Birds-of-Paradise. In Preparation.

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., & Mugal, C. F. (2019). GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology*, *20*(1), 5. https://doi.org/10.1186/s13059-018-1613-z

Boman, J., Frankl-Vilches, C., dos Santos, M. D. S., de Oliveira, E. H. C., Gahr, M., & Suh, A. (2019). The genome of blue-capped cordon-bleu uncovers hidden diversity of LTR retrotransposons in zebra finch. *Genes*, *10*(4), 301. https://doi.org/10.3390/genes10040301

Brajković, J., Feliciello, I., Bruvo-MadWarić, B., & Ugarković, D. W. (2012). Satellite DNA-like elements associated with genes within euchromatin of the beetle tribolium castaneum. *G3: Genes, Genomes, Genetics*, *2*(8), 931–941. https://doi.org/10.1534/g3.112.003467

Brown, J. E., & Jones, K. W. (1972). Localisation of satellite DNA in the microchromosomes of the Japanese quail by in situ hybridization. *Chromosoma*, *38*(3), 313–318. https://doi.org/10.1007/BF00290928

Burt, D. W. (2002). Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research*, *96*(1–4), 97–112. https://doi.org/10.1159/000063018

Cacciò, S., Perani, P., Saccone, S., Kadi, F., & Bernardi, G. (1994). Single-copy sequence homology among the GC-richest isochores of the genomes from warm-blooded vertebrates. *Journal of Molecular Evolution*, *39*(4), 331–339. https://doi.org/10.1007/BF00160265

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Chan, F. L., Marshall, O. J., Saffery, R., Won Kim, B. O., Earle, E., Choo, K. H. A., & Wong, L. H. (2012). Active transcription and essential role of RNA polymerase II at the centromere during mitosis. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(6), 1979–1984. https://doi.org/10.1073/pnas.1108705109

Charlesworth, B., Sniegowski, P., & Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, *371*(6494), 215–220. https://doi.org/10.1038/371215a0

Cheng, Z. J., & Murata, M. (2003). A centromeric tandem repeat family originating from a part of Ty3/gypsy-retroelement in wheat and its relatives. *Genetics*, *164*(2), 665–672. https://doi.org/10.1093/genetics/164.2.665

de Lima, L. G., Svartman, M., & Kuhn, G. C. S. (2017). Dissecting the satellite DNA landscape in three cactophilic Drosophila sequenced genomes. *G3: Genes, Genomes, Genetics*, *7*(8), 2831–2843. https://doi.org/10.1534/g3.117.042093

Deakin, J. E., Potter, S., O'Neill, R., Ruiz-Herrera, A., Cioffi, M. B., Eldridge, M. D. B., Fukui, K., Marshall Graves, J. A., Griffin, D., Grutzner, F., Kratochvíl, L., Miura, I., Rovatsos, M., Srikulnath, K., Wapstra, E., & Ezaz, T. (2019). Chromosomics: Bridging the gap between genomes and chromosomes. *Genes*, *10*(8), 627. https://doi.org/10.3390/genes10080627

Deryusheva, S., Krasikova, A., Kulikova, T., & Gaginskaya, E. (2007). Tandem 41-bp repeats in chicken and Japanese quail genomes: FISH mapping and transcription analysis on lampbrush chromosomes. *Chromosoma*, *116*(6), 519–530. https://doi.org/10.1007/s00412-007-0117-5

Dias, G. B., Svartman, M., Delprat, A., Ruiz, A., & Kuhn, G. C. S. (2014). Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of Drosophila virilis. *Genome Biology and Evolution*, *6*(6), 1302–1313. https://doi.org/10.1093/gbe/evu108

Dion-Côté, A. M., & Barbash, D. A. (2017). Beyond speciation genes: an overview of genome stability in evolution and speciation. *Current Opinion in Genetics and Development*, *47*, 17–23. https://doi.org/10.1016/j.gde.2017.07.014

Ferree, P. M. (2014). Mitotic misbehavior of a drosophila melanogaster satellite in ring chromosomes: Insights into intragenomic conflict among heterochromatic sequences. *Fly*, *8*(2), 101–107. https://doi.org/10.4161/fly.29488

Ferree, P. M., & Barbash, D. A. (2009). Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. *PLoS Biology*, *7*(10), e1000234. https://doi.org/10.1371/journal.pbio.1000234

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Frith, C. B., & Frith, D. W. (1996a). Description of the unique Parotia lawesii Paradisaea rudolphi hybrid bird of paradise (Aves: Passeriformes: Paradisaeidae). *Records of the Australian Museum*, *48*, 111–116.

Frith, C. B., & Frith, D. W. (1996b). The unique type specimen of the bird of paradise Lophorina superba pseudoparotia Stresemann 1934 (Paradisaeidae): A hybrid of Lophorina superba x Parotia carolae. *Journal Für Ornithologie*, *137*(4), 515–521. https://doi.org/10.1007/BF01661105

Fry, K., & Salser, W. (1977). Nucleotide sequences of HS-α satellite DNA from kangaroo rat dipodomys ordii and characterization of similar sequences in other rodents. *Cell*, *12*(4), 1069–1084. https://doi.org/10.1016/0092-8674(77)90170-2

Fuller, E. (1979). Hybridization amongst the Paradisaeidae. *Bulletin of the British Ornithologists' Club*, *99*, 145–152.

Garrido-Ramos, M. A. (2015). Satellite DNA in plants: More than just rubbish. *Cytogenetic and Genome Research*, *146*(2), 153–170. https://doi.org/10.1159/000437008

Grenfell, A. W., Strzelecka, M., & Heald, R. (2017). Transcription brings the complex(ity) to the centromere. *Cell Cycle*, *16*(3), 235–236. https://doi.org/10.1080/15384101.2016.1242962

Griffin, D., & Burt, D. W. (2014). All chromosomes great and small: 10 years on. *Chromosome Research*, *22*(1), 1–6. https://doi.org/10.1007/s10577-014-9413-0

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., & Hofacker, I. L. (2008). The Vienna RNA Websuite. *Nucleic Acids Research*, *36*(Web Server), W70–W74. https://doi.org/10.1093/nar/gkn188

Hartley, G., & O'Neill, R. (2019). Centromere repeats: Hidden gems of the genome. *Genes*, *10*(3), 223. https://doi.org/10.3390/genes10030223

Heads, M. (2001). Birds of paradise, biogeography and ecology in New Guinea: A review. *Journal of Biogeography*, *28*(7), 893–925. https://doi.org/10.1046/j.1365-2699.2001.00600.x

Henikoff, J. G., Thakur, J., Kasinathan, S., & Henikoff, S. (2015). A unique chromatin complex occupies young a-satellite arrays of human centromeres. *Science Advances*, *1*(1), e1400234. https://doi.org/10.1126/sciadv.1400234

Irestedt, M., Jønsson, K. A., Fjeldså, J., Christidis, L., & Ericson, P. G. (2009). An unexpectedly long history of sexual selection in birds-of-paradise. *BMC Evolutionary Biology*, *9*(1), 235. https://doi.org/10.1186/1471-2148-9-235

Jagannathan, M., & Yamashita, Y. M. (2021). Defective satellite DNA clustering into chromocenters underlies hybrid incompatibility in Drosophila. *Molecular Biology and Evolution*, *38*(11), 4977–4986. https://doi.org/10.1093/molbev/msab221

Jain, M., Olsen, H. E., Turner, D. J., Stoddart, D., Bulazel, K. V., Paten, B., Haussler, D., Willard, H. F., Akeson, M., & Miga, K. H. (2018). Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology*, *36*(4), 321–323. https://doi.org/10.1038/nbt.4109

Jiang, P. P., Hartl, D. L., & Lemos, B. (2010). Y not a dead end: Epistatic interactions between Y-linked regulatory polymorphisms and genetic background affect global gene expression in Drosophila melanogaster. *Genetics*, *186*(1), 109–118. https://doi.org/10.1534/genetics.110.118109

Jønsson, K. A., Fabre, P. H., Kennedy, J. D., Holt, B. G., Borregaard, M. K., Rahbek, C., & Fjeldså, J. (2016). A supermatrix phylogeny of corvoid passerine birds (Aves: Corvides). *Molecular Phylogenetics and Evolution*, *94*, 87–94. https://doi.org/10.1016/j.ympev.2015.08.020

Joshi, S. S., & Meller, V. H. (2017). Satellite repeats identify X chromatin for dosage compensation in Drosophila melanogaster males. *Current Biology*, *27*(10), 1393–1402.e2. https://doi.org/10.1016/j.cub.2017.03.078

Kasinathan, S., & Henikoff, S. (2018). Non-B-form DNA is enriched at centromeres. *Molecular Biology and Evolution*, *35*(4), 949–962. https://doi.org/10.1093/molbev/msy010

Katoh, K., Rozewicki, J., & Yamada, K. D. (2018). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

Kipling, D., Ackford, H. E., Taylor, B. A., & Cooke, H. J. (1991). Mouse minor satellite DNA genetically maps to the centromere and is physically linked to the proximal telomere. *Genomics*, *11*(2), 235–241. https://doi.org/10.1016/0888-7543(91)90128-2

Köster, J., & Rahmann, S. (2012). Snakemake-A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kuhn, G. C. S., Küttler, H., Moreira-Filho, O., & Heslop-Harrison, J. S. (2012). The 1.688 repetitive DNA of drosophila: Concerted evolution at different genomic scales and association with genes. *Molecular Biology and Evolution*, *29*(1), 7–11. https://doi.org/10.1093/molbev/msr173

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, Timetrees, and divergence times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Kutschera, V. E., Poelstra, J. W., Botero-Castro, F., Dussex, N., Gemmell, N. J., Hunt, G. R., Ritchie, M. G., Rutz, C., Wiberg, R. A. W., & Wolf, J. B. W. (2020). Purifying selection in corvids is less efficient on islands. *Molecular Biology and Evolution*, *37*(2), 469–474. https://doi.org/10.1093/molbev/msz233

Larracuente, A. M. (2014). The organization and evolution of the Responder satellite in species of the Drosophila melanogaster group: Dynamic evolution of a target of meiotic drive. *BMC Evolutionary Biology*, *14*(1), 233. https://doi.org/10.1186/s12862-014-0233-9

Leclerc, S., & Kitagawa, K. (2021). The role of human centromeric RNA in chromosome stability. *Frontiers in Molecular Biosciences*, *8*, 170. https://doi.org/10.3389/fmolb.2021.642732

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liangouzov, I. A., Derjusheva, S. E., Saifitdinova, A. F., Malykh, A. G., & Gaginskaya, E. R. (2002). Monomers of a satellite DNA sequence of chaffinch (Fringilla coelebs L., Aves: Passeriformes) contain short clusters of the TTAGGG repeat. *Russian Journal of Genetics*, *38*(12), 1359–1364. https://doi.org/10.1023/A:1021679520236

Liu, J., Wang, Z., Li, J., Xu, L., Liu, J., Feng, S., Guo, C., Chen, S., Ren, Z., Rao, J., Wei, K., Chen, Y., Jarvis, E. D., Zhang, G., & Zhou, Q. I. (2021). A new emu genome illuminates the evolution of genome configuration and nuclear architecture of avian chromosomes. *Genome Research*, *31*(3), 497–511. https://doi.org/10.1101/GR.271569.120

Logsdon, G. A., Vollger, M. R., Hsieh, P. H., Mao, Y., Liskovykh, M. A., Koren, S., & Eichler, E. E. (2021). The structure, function and evolution of a complete human chromosome 8. *Nature*, *593*(7857), 101–107. https://doi.org/10.1038/s41586-021-03420-7

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26. https://doi.org/10.1186/1748-7188-6-26

Lower, S. S., McGurk, M. P., Clark, A. G., & Barbash, D. A. (2018). Satellite DNA evolution: Old ideas, new approaches. *Current Opinion in Genetics and Development*, *49*, 70–78. https://doi.org/10.1016/j.gde.2018.03.003

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, *389*(6648), 251–260. https://doi.org/10.1038/38444

Marshall, A. J. (1951). Birds of paradise and bower birds. *Nature*, *168*(4265), 135. https://doi.org/10.1038/168135a0

Mayer, F. W. S., & Peckover, W. S. (1991). Eggs of hybrid Shaw Mayer's bird of paradise: The ribbontail *Astrapia mayeri x? Emu*, *91*(3), 189. https://doi.org/10.1071/MU9910189

McGurk, M. P., & Barbash, D. A. (2018). Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. *Genome Research*, *28*(5), 714–725. https://doi.org/10.1101/gr.231472.117

Meise, W. (1928). Die Verbreitung der AaskrÃ¤he (Formenkreis ~Corvus corone~ L.). *Journal of Ornithology*, *76*, 1–203.

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J., DeRisi, J. L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., & Chan, S. W. L. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, *14*(1), 1–20. https://doi.org/10.1186/gb-2013-14-1-r10

Meštrović, N., Mravinac, B., Pavlek, M., Vojvoda-Zeljko, T., Šatović, E., & Plohl, M. (2015). Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome Research*, *23*(3), 583–596. https://doi.org/10.1007/s10577-015-9483-7

Miga, K. H. (2020). Centromere studies in the era of 'telomere-to-telomere' genomics. *Experimental Cell Research*, *394*(2), 112127. https://doi.org/10.1016/j.yexcr.2020.112127

Mugal, C. F., Arndt, P. F., & Ellegren, H. (2013). Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. *Molecular Biology and Evolution*, *30*(7), 1700–1712. https://doi.org/10.1093/molbev/mst067

Mugal, C. F., Weber, C. C., & Ellegren, H. (2015). GC-biased gene conversion links the recombination landscape and demography to genomic base composition: GC-biased gene conversion drives genomic base composition across a wide range of species. *BioEssays*, *37*(12), 1317–1326. https://doi.org/10.1002/bies.201500058

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Lambing, C., Henderson, I. R. (2021). The genetic and epigenetic landscape of the Arabidopsis centromeres. BioRxiv, *Science*, *374*(6569), eabi7489. https://doi.org/10.1101/2021.05.30.446350

Novák, P., Neumann, P., & Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols*, *15*(11), 3745–3776. https://doi.org/10.1038/s41596-020-0400-y

Oliveros, C. H., Field, D. J., Ksepka, D. T., Barker, F. K., Aleixo, A., Andersen, M. J., Alström, P., Benz, B. W., Braun, E. L., Braun, M. J., Bravo, G. A., Brumfield, R. T., Chesser, R. T., Claramunt, S., Cracraft, J., Cuervo, A. M., Derryberry, E. P., Glenn, T. C., Harvey, M. G., … Faircloth, B. C. (2019). Earth history and the passerine superradiation. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(16), 7916–7925. https://doi.org/10.1073/pnas.1813206116

Ottenburghs, J., Ydenberg, R. C., Van Hooft, P., Van Wieren, S. E., & Prins, H. H. T. (2015). The Avian Hybrids Project: Gathering the scientific literature on avian hybridization. *Ibis*, *157*(4), 892–894. https://doi.org/10.1111/ibi.12285

Palacios-Gimenez, O. M., Dias, G. B., De Lima, L. G., Kuhn, G. C. E. S., Ramos, É., Martins, C., & Cabral-De-Mello, D. C. (2017). High-throughput analysis of the satellitome revealed enormous diversity of satellite DNAs in the neo-Y chromosome of the cricket Eneoptera surinamensis. *Scientific Reports*, *7*(1), 6422. https://doi.org/10.1038/s41598-017-06822-8

Palacios-Gimenez, O. M., Milani, D., Song, H., Marti, D. A., López-León, M. D., Ruiz-Ruano, F. J., Camacho, J. P. M., & Cabral-de-Mello, D. C. (2020). Eight million years of satellite DNA evolution in grasshoppers of the genus Schistocerca illuminate the ins and outs of the library hypothesis. *Genome Biology and Evolution*, *12*(3), 88–102. https://doi.org/10.1093/gbe/evaa018

Peona, V., Blom, M. P. K., Frankl-Vilches, C., Milá, B., Ashari, H., Thébaud, C., Suh, A. (2022). The hidden structural variability in avian genomes. *BioRxiv*, 2021.12.31.473444. https://doi.org/10.1101/2021.12.31.473444

Peona, V., Blom, M. P. K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T., Jønsson, K. A., Zhou, Q. I., Irestedt, M., & Suh, A. (2021). Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources*, *21*(1), 263–286. https://doi.org/10.1111/1755-0998.13252

Peona, V., Palacios-Gimenez, O. M., Blommaert, J., Liu, J., Haryoko, T., Jønsson, K. A., Irestedt, M., Zhou, Q. I., Jern, P., & Suh, A. (2021). The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1833), 20200186. https://doi.org/10.1098/rstb.2020.0186

Peona, V., Weissensteiner, M. H., & Suh, A. (2018). How complete are "complete" genome assemblies?—An avian perspective. *Molecular Ecology Resources*, *18*(6), 1188–1195. https://doi.org/10.1111/1755-0998.12933

Pezer, Ž., Brajković, J., Feliciello, I., & Ugarković, D. (2012). Satellite DNA-mediated effects on genome regulation. *Genome Dynamics*, *7*, 153–169. https://doi.org/10.1159/000337116

Plohl, M., Meštrović, N., & Mravinac, B. (2012). Satellite DNA evolution. *Genome Dynamics*, *7*, 126–152. https://doi.org/10.1159/000337122

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., & Wolf, J. B. W. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, *344*(6190), 1410–1414. https://doi.org/10.1126/science.1253226

Prost, S., Armstrong, E. E., Nylander, J., Thomas, G. W. C., Suh, A., Petersen, B., Dalen, L., Benz, B. W., Blom, M. P. K., Palkopoulou, E., Ericson, P. G. P., & Irestedt, M. (2019). Comparative analyses identify genomic features potentially involved in the evolution of birds-of-paradise. *GigaScience*, *8*(5), giz003. https://doi.org/10.1093/gigascience/giz003

Raz, O., Biezuner, T., Spiro, A., Amir, S., Milo, L., Titelman, A., & Shapiro, E. (2019). Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. *Nucleic Acids Research*, *47*(5), 2436–2445. https://doi.org/10.1093/nar/gky1318

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The european molecular biology open software suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Ruiz-Ruano, F. J., Castillo-Martínez, J., Cabrero, J., Gómez, R., Camacho, J. P. M., & López-León, M. D. (2018). High-throughput analysis of satellite DNA in the grasshopper Pyrgomorpha conica reveals abundance of homologous and heterologous higher-order repeats. *Chromosoma*, *127*(3), 323–340. https://doi.org/10.1007/s00412-018-0666-9

Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., & Camacho, J. P. M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, *6*(1), 28333. https://doi.org/10.1038/srep28333

Ruiz-Ruano, F. J., Navarro-Domínguez, B., Camacho, J. P. M., & Garrido-Ramos, M. A. (2019). Characterization of the satellitome in lower vascular plants: The case of the endangered fern Vandenboschia speciosa. *Annals of Botany*, *123*(4), 587–599. https://doi.org/10.1093/aob/mcy192

Sahakyan, A. B., Chambers, V. S., Marsico, G., Santner, T., Di Antonio, M., & Balasubramanian, S. (2017). Machine learning model for sequence-driven DNA G-quadruplex formation. *Scientific Reports*, *7*(1), 14535. https://doi.org/10.1038/s41598-017-14017-4

Saifitdinova, A. F., Derjusheva, S. E., Malykh, A. G., Zhurov, V. G., Andreeva, T. F., & Gaginskaya, E. R. (2001). Centromeric tandem repeat from the chaffinch genome: Isolation and molecular characterization. *Genome*, *44*(1), 96–103. https://doi.org/10.1139/g00-098

Salser, W., Bowen, S., Browne, D., el-Adli, F., Fedoroff, N., Fry, K., Heindell, H., Paddock, G., Poon, R., Wallace, B., & Whitcome, P. (1976). Investigation of the organization of mammalian chromosomes at the DNA sequence level. *Federation Proceedings*, *35*(1), 23–35. http://europepmc.org/abstract/MED/1107072

Scott, P. (2013). The birds of paradise. *The Birds of Paradise*, *261*(6), 116–123. https://doi.org/10.7208/chicago/9780226088099.001.0001

Seibt, K. M., Schmidt, T., & Heitkam, T. (2018). FlexiDot: Highly customizable, ambiguity-aware dotplots for visual sequence analyses. *Bioinformatics*, *34*(20), 3575–3577. https://doi.org/10.1093/bioinformatics/bty395

Sergey, N., Sergey, K., Arang, R., Mikko, R., Andrey, V. B., Alla, M., … Adam, M. P. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Shang, W.-H., Hori, T., Toyoda, A., Kato, J., Popendorf, K., Sakakibara, Y., Fujiyama, A., & Fukagawa, T. (2010). Chickens possess centromeres

with both extended tandem repeats and short non-tandem-repetitive sequences. *Genome Research*, *20*(9), 1219–1228. https://doi.org/10.1101/gr.106245.110

Smalec, B. M., Heider, T. N., Flynn, B. L., & O'Neill, R. J. (2019). A centromere satellite concomitant with extensive karyotypic diversity across the Peromyscus genus defies predictions of molecular drive. *Chromosome Research*, *27*(3), 237–252. https://doi.org/10.1007/s10577-019-09605-1

Smit, A. F. A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. http://www.repeatmasker.org

Smith, G. P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science*, *191*(4227), 528–535. https://doi.org/10.1126/science.1251186

Talbert, P. B., & Henikoff, S. (2020). What makes a centromere? *Experimental Cell Research*, *389*(2), 111895. https://doi.org/10.1016/j.yexcr.2020.111895

Thomma, B. P. H. J., Seidl, M. F., Shi-Kunne, X., Cook, D. E., Bolton, M. D., van Kan, J. A. L., & Faino, L. (2016). Mind the gap; seven reasons to close fragmented genome assemblies. *Fungal Genetics and Biology*, *90*, 24–30. https://doi.org/10.1016/j.fgb.2015.08.010

Tsuda, Y., Nishida-Umehara, C., Ishijima, J., Yamada, K., & Matsuda, Y. (2007). Comparison of the Z and W sex chromosomal architectures in elegant crested tinamou (Eudromia elegans) and ostrich (Struthio camelus) and the process of sex chromosome differentiation in palaeognathous birds. *Chromosoma*, *116*(2), 159–173. https://doi.org/10.1007/s00412-006-0088-y

Vijay, N., Bossu, C. M., Poelstra, J. W., Weissensteiner, M. H., Suh, A., Kryukov, A. P., & Wolf, J. B. W. (2016). Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nature Communications*, *7*(1), 13195. https://doi.org/10.1038/ncomms13195

Vontzou, N. (2021). *Comparative genomics of satellite DNA and putative centromere positions in birds*. Stockholm University.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757–767. https://doi.org/10.1101/gr.214874.116

Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., Warmuth, V. M., & Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, *11*(1), 3403. https://doi.org/10.1038/s41467-020-17195-4

Weissensteiner, M. H., Pang, A. W. C., Bunikis, I., Höijer, I., Vinnere-Petterson, O., Suh, A., & Wolf, J. B. W. (2017). Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Research*, *27*(5), 697–708. https://doi.org/10.1101/gr.215095.116

Westerberg, I. (2020). *Deciphering the formation of evolutionary new centromeres in a microchromosome of birds*. Uppsala University.

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., & Wagner, L. (2003). Database resources of the national center for biotechnology. *Nucleic Acids Research*, *31*(1), 28–33. https://doi.org/10.1093/nar/gkg033

Xu, L., Auer, G., Peona, V., Suh, A., Deng, Y., Feng, S., Zhang, G., Blom, M. P. K., Christidis, L., Prost, S., Irestedt, M., & Zhou, Q. I. (2019). Dynamic evolutionary history and gene content of sex chromosomes across diverse songbirds. *Nature Ecology and Evolution*, *3*(5), 834–844. https://doi.org/10.1038/s41559-019-0850-1

Yamada, K., Nishida-Umehara, C., & Matsuda, Y. (2002). Characterization and chromosomal distribution of novel satellite DNA sequences of the lesser rhea (Pterocnemia pennata) and the greater rhea (Rhea americana). *Chromosome Research*, *10*(6), 513–523. https://doi.org/10.1023/A:1020996431588

Yamada, K., Nishida-Umehara, C., & Matsuda, Y. (2004). A new family of satellite DNA sequences as a major component of centrometric heterochromatin in owls (Strigiformes). *Chromosoma*, *112*(6), 277–287. https://doi.org/10.1007/s00412-003-0267-z

Yan, H., Talbert, P. B., Lee, H. R., Jett, J., Henikoff, S., Chen, F., & Jiang, J. (2008). Intergenic locations of rice centromeric chromatin. *PLoS Biology*, *6*(11), 2563–2575. https://doi.org/10.1371/journal.pbio.0060286

Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, *18*(1), 1–17. https://doi.org/10.1186/s13059-017-1319-7

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Peona, V., Kutschera, V. E., Blom, M. P. K., Irestedt, M., & Suh, A. (2022). Satellite DNA evolution in Corvoidea inferred from short and long reads. *Molecular Ecology*, *00*, 1–18. https://doi.org/10.1111/mec.16484