SYMPHONY

Highly Contiguous Genomes Improve the Understanding of Avian Olfactory Receptor Repertoires

Robert J. Driver and Christopher N. Balakrishnan

Department of Biology, East Carolina University, E 5th Street, Greenville, NC 27858, USA

From the symposium “Manakin genomics: comparative studies of evolution and behavior in a unique clade of birds” presented at the annual meeting of the Society for Integrative and Comparative Biology virtual annual meeting, January 3–February 28, 2021.

1E-mail: driverr16@students.ecu.edu

Synopsis  Third-generation (long-read–based) sequencing technologies are reshaping our understanding of genome structure and function. One of the most persistent challenges in genome biology has been confidently reconstructing radiations of complex gene families. Olfactory receptors (ORs) represent just such a gene family with upward of thousands of receptors in some mammalian taxa. Whereas in birds olfaction was historically an overlooked sensory modality, new studies have revealed an important role for smell. Chromosome-level assemblies for birds allow a new opportunity to characterize patterns of OR diversity among major bird lineages. Previous studies of short-read–based (second-generation) genome assemblies have associated OR gene family size with avian ecology, but such conclusions could be premature especially when new assembly methods reshape our understanding of avian OR evolution. Here we provide a fundamental characterization of OR repertoires in five recent genome assemblies, including the most recent assembly of golden-collared manakin (Manacus vitellinus). We find that short read-based assemblies systematically undercount the avian-specific gamma-c OR subfamily, a subfamily that comprises over 65% of avian OR diversity. Therefore, in contrast to previous studies, we find a high diversity of gamma-c ORs across the avian tree of life. Building on these findings, ongoing sequencing efforts and improved genome assemblies will clarify the relationship between OR diversity and avian ecology.

Introduction

Our understanding of avian sensory biology has progressed substantially in recent years. Studies have discovered fantastic ways in which birds experience the world, including the visual detection of nonspectral colors, the detection of sugar via a repurposed umami taste receptor in nectivorous species, and amphibious hearing in cormorants (Baldwin et al. 2014; Larsen et al. 2020; Stoddard et al. 2020). However, while studies investigating most senses, particularly bird vision, have received considerable attention, research into olfaction has lagged behind. Misconceptions about bird olfaction date back nearly 200 years, when John James Audubon falsely concluded that turkey vultures (Cathartes aura) could not smell carrion (Audubon 1826). Darwin also performed behavioral experiments on Andean condors (Vultur gryphus) to conclude that they could not smell meat (Darwin 1891). An examination of olfactory bulb size across a diversity of bird species concluded that birds could not have anything more than a rudimentary sense of smell (Hill 1905). In response to these conclusions, bird olfaction remained relatively unexplored until behavioral studies showed odor recognition in pigeons (Michelsen 1959). Following this study, there have been a wealth of morphological and behavioral studies testing for olfaction in both captive and wild birds (Bang and Cobb 1968; Hagelin 2007; Gwinner and Berger 2008; Nevitt et al. 2008; Krause et al. 2012; Van Huynh and Rice 2019).

To follow this appreciation for the behavioral and ecological roles of olfaction in birds, researchers have characterized bird olfactory receptors at a genomic level. Olfactory receptors (ORs) are seven transmembrane domain rhodopsin-like G protein-coupled recep-
tors that detect odors when expressed in the olfactory sensory neurons of the nasal epithelium (Buck and Axel 1991; Mombaerts 2004). In the protruding cilia of the olfactory sensory neurons, ORs recognize specific volatile compounds in their transmembrane domain binding pocket, which creates a signaling cascade that depolarizes the cell membrane and sends an action potential to the olfactory bulb glomeruli and later the brain (summarized in Breer 2003). Each OR may recognize one or more odorants, and each odorant may be detected by one or more ORs, and so in this way, species may perceive a wide array of odors (Saito et al. 2009).

ORs constitute one of the largest gene families in vertebrates. For example, the elephant genome contains about 2000 intact ORs (Niimura et al. 2014). In birds, there are three major subfamilies of ORs, the alpha, gamma, and gamma-c (Niimura and Nei 2005). The alpha and gamma subfamilies are shared between all avian lineages (Niimura and Nei 2005; Steiger et al. 2009). The third subfamily, gamma-c, is unique to birds (Niimura and Nei 2005; Silva et al. 2020). The gamma-c subfamily is common in some bird genomes comprising over 65% of the OR repertoire (Steiger et al. 2009; Khan et al. 2015). Gamma-c ORs are similar in sequence and sequences cluster by species rather than by orthologs among species (Steiger et al. 2009). Gamma-c ORs likely evolve with a high level of birth and death rates and gene conversion to maintain the species-specific clustering (Niimura and Nei 2005; Steiger et al. 2009).

The first genomic investigations to determine bird OR repertoire counts provided further evidence for the potential of birds to recognize a wide variety of odors. A total of 214 intact ORs were reported in chicken (Gallus gallus) and 134 reported in zebra finch (Taeniopygia guttata) by Steiger et al. (2009). The finding of hundreds of ORs in the chicken and zebra finch genomes was replicated using multiple OR identifying pipelines (Wang et al. 2013; Khan et al. 2015; Vandewege et al. 2016). All of these studies identified similar OR counts as well as similar proportions of each OR subfamily, with the gamma-c family dominating the OR repertoire. The majority of ORs, however, were located on unmapped contigs, including over 90% of gamma-c ORs in chicken and zebra finch.

Second-generation (Illumina, short-read–based) genomes greatly broadened genome sampling across the tree of life, including birds (Jarvis et al. 2014). However, in these assemblies, intact OR numbers were significantly lower than had been observed in the Sanger-based chicken and zebra finch assemblies (Steiger et al. 2009; Khan et al. 2015). Particularly absent from analyses was the distinct avian radiation of the gamma-c OR subfamily, with 45 of 46 species with short-read assemblies yielding fewer than 25 gamma-c ORs (Khan et al. 2015). Despite sequencing technology being a common thread in the 45 assemblies with lower OR counts, technical explanations were ruled out in favor of evolutionary explanations for the observed patterns of diversity (Khan et al. 2015).

Chromosome-level reference genomes, using long-read sequencing technology, should provide more reliable information about OR repertoire diversity in birds. The Vertebrate Genomes Project recently expanded chromosome-scale assembly methods from model organisms across the vertebrate tree of life (Rhie et al. 2021). Combining these and other new assemblies, we are now able to characterize OR diversity in five bird genomes in which long-read approaches have been deployed (Feng et al. 2020; Liu et al. 2021; Rhie et al. 2021). Included in our species analyses is the new assembly of golden-collared manakin (Manacus vitellinus) that was sequenced as part of a collaborative effort within the National Science Foundation supported Research Coordination Network for biologists studying manakins (Pipridae). We directly compare Sanger, Illumina, hybrid, and Pac-Bio based assemblies to examine the ways in which our understanding of bird OR family repertoire, and our comprehension of avian olfactory capabilities, is shaped by these higher-quality assemblies.

**Methods**

**Assembly selection**

We sought to compare OR discovery rates and assembly quality by using select bird species with multiple publicly available genome assemblies on GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Assemblies for each species varied in the sequencing technology employed and assembly software used (Table 1). In order to examine how variation in genome sequencing methods impacts OR discovery and description, we included one genome of each species with long-read–sequencing technology (for example, PacBio [Pacific Biosciences], Menlo Park, California) as well as one genome without long-read sequencing. We obtained two assemblies from five bird species: emu (Dromaius novaehollandiae), chicken (G. gallus), Anna’s hummingbird (Calypte anna), golden-collared manakin (M. vitellinus), and zebra finch (T. guttata) (Warren et al. 2010; Zhang et al. 2014; Feng et al. 2020; Liu et al. 2021; Rhie et al. 2021, Table 1). In addition to the availability of multiple genomes, these five species are representative across the three major groupings of extant birds, including the Paleognathae and two groups within the Neognathae, Galloanseres,
and Neoaves, represent diverse ecology, and include two important avian models, chicken and zebra finch (Fig. 1A).

**OR identification**

To detect putatively functional ORs in the selected genomes, we created a BLAST query with a set of 2110 OR protein sequences from 6 mammals (Ornithorhynchus anatinus, Didelphis virginiana, Bos taurus, Canis lupus, Rattus norvegicus, Macaca mulatta), 2 birds (G. gallus, T. guttata), and 1 crocodilian (Gavialis gangeticus). We obtained this query OR set by combining previously published OR subgenomes (Niimura and Nei 2007; Niimura 2009; Vandewege et al. 2016). Using this query file, we performed TBLASTN searches against all 11 bird genomes with a threshold of E < 1e–20. To remove pseudogenized and truncated ORs, we filtered for hits > 250 amino acids long. For any single location on the genome, we filtered out hits within 100 bp of each other, and selected the lowest E-value associated with that location.

After obtaining unique BLAST hits, we extracted the associated nucleotide sequence from the genome as well as 300-bp regions flanking the hit both upstream and downstream. We used a modified Perl script from Beichman et al. (2019) to detect open reading frames (ORFs) within each extracted region (Montague et al. 2014; Beichman et al. 2019). We then aligned these ORFs to each other as well as to the human Olfactory Receptor Family 2 Subfamily J Member 3 (OR2J3) sequence using the E-INS-I default parameters in MAFFT (Katoh and Standley 2013). Using the previously characterized transmembrane domains of OR2J3 as a guide, we removed any sequences that had five or more amino acid insertions or deletions within a transmembrane domain in the alignment (McRae et al. 2012; Beichman et al. 2019). This included ORFs with stop codons appearing prior to the end of the seventh transmembrane domain.

Using this alignment, we recorded the position of the first amino acid in the first transmembrane domain. To estimate the location of the ORF start codon, we used modified Perl scripts from Beichman et al. (2019) to find the most appropriate methionine upstream of this recorded transmembrane start position (Montague et al. 2014; Beichman et al. 2019). ORF sequences were then truncated at the S' ends to begin with this methionine. This set of ORFs was then aligned using the E-INS-I parameters in MAFFT to a set of T. guttata reference ORs as well as 11 non-OR rhodopsin-like G-protein coupled receptors (non-OR GPCRs) that functioned as an outgroup (Katoh and Standley 2013; Niimura 2013; Vandewege et al. 2016; Beichman et al. 2019). We then used clustalW to generate a neighbor-joining tree from this alignment with 1000 bootstraps, gaps removed, and Kimura’s distance correction (Kimura 1980; Goujon et al. 2010). We then removed any ORFs that were phylogenetically more closely related to the non-OR GPCRs.

**Classification of final OR set**

We classified all remaining ORFs as functional ORs. Using this final set, we ran a maximum-likelihood tree using IQ-TREE with automatic model selection and 1000 Shimodaira–Hasegawa-like approximate likelihood ra-
Fig. 1 Maximum-likelihood phylogenetic reconstruction of OR repertoire from long-read assemblies. Phylogenetic trees were assembled using IQ-TREE and only nodes with >50% support following a likelihood ratio test are shown. (A) Topological phylogeny of species used in this study is shown with red branches indicating Neoaves species. OR counts from long-read genomes for each species are given. The five species and assemblies shown are (B) M. vitellinus Mvit3, (C) C. anna Cann2, (D) T. guttata Tgut2, (E) G. gallus Ggut6, and (F) D. novaehollandiae Dnov2. The three OR subfamilies were assigned based on putative orthology to previously described bird ORs (Niimura and Nei 2005; Vandewege et al. 2016). Images not to scale. Bird images in (E) by Jim Bendon and in (F) by Daniel Naish. Images (E) and (F) were rasterized by Michael Keesey and shared under CC by 3.0. Image (A) is modified artwork by Kristen Orr.

to the 11 non-OR GPCRs (Letunic and Bork 2019). We classified bird ORs into subfamilies alpha, gamma, and gamma-c based on the subfamily of the query sequence used to identify the OR and the location of the OR in
Table 2 OR counts obtained from each assembly search

| Abbreviation | Total ORs | Alpha ORs | Gamma ORs | Gamma-c ORs | Proportion gamma-c | contigN50 (Mb) | OR total in literature |
|--------------|-----------|-----------|-----------|-------------|-------------------|---------------|-----------------------|
| Mvit1        | 9         | 1         | 8         | 0           | 0.000             | 0.04          | 9^a                   |
| Mvit3        | 117       | 2         | 18        | 97          | 0.829             | 0.29          |                       |
| Ggal4        | 272       | 10        | 33        | 229         | 0.842             | 0.30          | 266^b                 |
| Ggal6        | 355       | 11        | 41        | 303         | 0.854             | 17.65         |                       |
| Tgut1        | 184       | 2         | 3         | 179         | 0.973             | 0.038         |                       |
| Tgut2        | 69        | 3         | 6         | 60          | 0.870             | 12.00         | 190^b                 |
| Dnov1        | 57        | 17        | 33        | 7           | 0.123             | 0.86          |                       |
| Dnov2        | 296       | 26        | 75        | 195         | 0.659             | 14.09         |                       |
| Cann1        | 27        | 2         | 23        | 2           | 0.071             | 0.03          | 21^a                  |
| Cann2        | 109       | 0         | 23        | 86          | 0.761             | 14.52         |                       |

Notes: OR counts are further divided into the three avian OR subfamilies. OR counts reported in literature reflect any previously published OR scans from the same assembly we report on here.

^aKhan et al. (2015).
^bVandewege et al. (2016).

Results
OR totals

We identified a total of 1496 ORs across all 10 bird assemblies from 5 species (Table 2). Of these ORs, the gamma-c subfamily constituted 77% (1158) of the total, while 18% (263) of the identified ORs were gamma, and 5% (74) were alpha subfamily ORs. For assemblies with long-read sequencing, we found 946 ORs, with 42 alpha (4%), 162 gamma (17%), and 741 gamma-c (78%). Within a single assembly, the chicken Ggal6 (see assembly abbreviations in Table 1) yielded the largest number of ORs, 355 in total, 303 (85%) of which were gamma-c ORs (Fig. 1E). Gamma-c represented 97% (179/184) of the ORs in zebra finch Tgut1, the highest percentage of gamma-c out of the total OR repertoire for any assembly.

Emu, Dnov2, yielded both the highest gamma counts (75) and the highest alpha counts (26), in addition to 195 gamma-c ORs (Fig. 1F). In the OR maximum-likelihood phylogenies, we elected to present each species separately for clarity of visualization (Fig. 1B–E). We noted that in these analyses the gamma OR subfamily for D. novaehollandiae was not recovered as monophyletic (Fig. 1F). In other multispecies analyses we have done, this is not the case, and this is also not the case in our Dnov2 neighbor-joining tree (analyses not shown). This unusual pattern here seems to be driven by the long branch at the base of the Dnov2 alpha OR clade.

OR counts relative to previous studies

Overall, reanalysis of previously analyzed genomes was consistent with previous findings (Khan et al. 2015; Vandewege et al. 2016, Table 2). For Ggal4, we recovered 272 ORs, 6 more than previously recovered (Vandewege et al. 2016). Two previous searches of the Tgut1 assembly yielded 182 and 190 ORs, similar to our search of 184 ORs (Wang et al. 2013; Vandewege et al. 2016, respectively). We also found similar subfamily diversity of gamma and gamma-c ORs in zebra finch and chicken (Niimura 2009; Steiger et al. 2009; Wang et al. 2013; Khan et al. 2015; Vandewege et al. 2016). Similar patterns emerged for the other short-read assemblies in our analysis (Table 2), giving us confidence in our methods of recovering ORs.

Assembly effects on OR subgenome

We found that assembly had substantial effects on the ability to reconstruct OR subgenomes. In four of the five surveyed species, inclusion of long-read sequencing (Pacific Biosciences) increased OR counts (Table 2). The most pronounced effect on OR repertoire was in the gamma-c family, which also constitutes the majority of known avian ORs. Between D. novaehollandiae assemblies, contigN50 improved from 0.86 Mb in Dnov1 with Illumina sequencing to 14.09 Mb in Dnov2. We detected an additional 239 ORs in Dnov2, of which 188 (78.6%) were from the gamma-c family. In the case of M. vitellinus, no gamma-c representatives were recovered from Mvit1 in our analysis or a previous analysis (Khan et al. 2015), yet our search of Mvit3 yielded 97 gamma-c ORs (Fig. 1B).

Improved assemblies also resulted in the identification of additional ORs in the alpha and gamma sub-
families as well, but these effects were less pronounced (Table 2). The gamma subfamily of *D. novaehollandiae* and in *M. vitellinus* more than doubled in count in Dnov2 compared to Dnov1 (42 new ORs) and in Mvit3 compared to Mvit1 (10 new ORs), respectively. In other species, the relative increase in gamma was smaller (Table 2). Alpha OR counts were similar between within-species assemblies (Table 2). Unexpectedly, we identified no alpha ORs in the *C. anna* Cann2 assembly, though two were previously reported based on Cann1.

The Sanger-sequencing–based *T. guttata* genome Tgut1 unexpectedly yielded a higher number of ORs than Tgut2, which was assembled using multiple technologies. This species was the only case in which a newer assembly based on long-read technology yielded fewer ORs than the assembly without long-read sequencing. Despite an overall higher OR count in Tgut1, we found more alpha ORs (three vs. two) and more gamma ORs (six vs. three) in Tgut2 than in Tgut1. However, the overall count of ORs in Tgut2 was substantially lower as there were 119 fewer gamma-c ORs in Tgut2 than in Tgut1. Therefore, the lower OR count in Tgut2 is entirely due to differences in gamma-c OR recovery.

**Physical location of ORs in avian genomes**

Although new, chromosome-scale assemblies have assigned the vast majority of genomic sequence data to chromosomes (Rhie et al. 2021), gamma-c OR regions remain primarily assigned on long-read assembled scaffolds (Fig. 2). The exception to this rule is the Ggal6 assembly in which 302 out of 303 identified ORs have been assigned to chromosomes. Most of these (274 ORs) map to a single microchromosome 33 (Lee et al. 2020, Fig. 3). The remaining ORs are divided between two other microchromosomes, chromosome 16 (8 ORs) and chromosome 31 (16 ORs), with a single receptor on an unmapped scaffold. For *C. anna* Cann2, only 20% of gamma-c receptors were on scaffolds assigned to chromosomes. The main clusters of ORs were a group of 10 assigned to the W chromosome and another 18 assigned to a single scaffold (RRCD01000065.1). For *D. novaehollandiae* Dnov2, only 8% of gamma-c ORs are assigned to chromosomes, with a large cluster of 108 on scaffold JABVCD010000554.1 (Fig. 2). The remaining gamma-c loci were distributed among 29 chromosomes and scaffolds. Finally, only 3% (2/60) of *T. guttata* Tgut2 gamma-c ORs were assigned to chromosomes with a cluster of 24 loci on scaffold VOHI02000029.1.

**Discussion**

**Long-read sequencing is critical for characterizing avian OR repertoire**

Our results show that Illumina short-read–based approaches were not successful in accurately characterizing the gamma-c OR subfamily. The three Illumina-based assemblies we assessed undercounted gamma-c diversity, revealing fewer than 10 gamma-c ORs in each assembly. In all five of the assemblies sequenced with long reads, we found that gamma-c ORs constituted at least 66% of the OR subgenome, and, on average, there were 148 gamma-c ORs per species. The hybrid assembly approach using MaSuRCA (Zimin et al. 2013) also substantially increased gamma-c recovery in Mvit3.

In the most phylogenetically comprehensive analysis of avian ORs to date, Khan et al. (2015) analyzed OR repertoire from 48 bird species. Forty-six of the assemblies surveyed were sequenced and assembled using short-read–based methods. The two other species included were the Sanger-based chicken and zebra finch, and those had the highest OR diversity, which was attributed to ecological adaptations of these
Genome quality and gene family analyses

The image shows a magnification of the largest OR cluster in the chicken genome on chromosome 33. Numerous ORs are present in the displayed region and are flanked by repetitive elements, potentially contributing to the difficulty of gamma-c OR subfamily discovery. Image from the UCSC Genome Browser (Lee et al. 2020).

Two particular species (Khan et al. 2015). Other reports in the literature also interpret a lack of gamma-c ORs as biologically meaningful without considering the shortcomings of short-read–based assemblies (Zhan et al. 2013). Importantly, these issues extend beyond gamma-c ORs to other complex gene families such as the major histocompatibility complex (MHC). Long-read–based studies are also improving our understanding of the avian MHC (He et al. 2021). Select MHC genes are linked to ORs in mammals, providing further evidence that this family repertoire may also be obscured in Illumina-sequenced assemblies (Santos et al. 2010).

Prior to our analysis, chicken (Ggal3, Ggal4) and zebra finch (Tgut1) were the only Sanger-based assemblies analyzed (Niimura 2009; Steiger et al. 2009; Wang et al. 2013; Khan et al. 2015; Vandewege et al. 2016). Analyses of these assemblies, which involve longer read lengths and Bacterial Artificial Chromosome (BAC)-based scaffolding, provided the only previous evidence of substantial gamma-c OR diversity. Whereas the incorporation of long-read sequencing methods greatly increased the count of gamma-c ORs relative to Illumina-based assemblies, they also reduced the count of gamma-c in T. guttata compared to the previous Sanger-based assembly. We propose two potential reasons for this disparity. It is possible that the original Tgut1 assembly resolved alternative alleles as separate loci, artificially inflating the total gamma-c count with duplicate loci. Tgut2 and many third-generation assemblies are haplotype phased, mitigating this problem. Additionally, however, filtering steps at the end of the Tgut2 assembly curation process used for quality control may aggressively remove repetitive regions that harbor tandem gamma-c loci.

Phylogenetic implications of updated OR counts

Our finding of high OR diversity for D. novaehollandiae has potentially important implications for broad-scale patterns of OR evolution in birds. High OR diversity in this paleognath genome contrasts with lower diversity found in previous analyses (Le Duc et al. 2015) and now suggests that the avian ancestor had high gamma-c OR diversity. Based on our assessment, all three OR subfamilies have lower diversity in the three neoavian lineages tested relative to D. novaehollandiae and G. gallus. Due to our limited taxonomic sampling, it remains unclear whether the differences reflect broader patterns of phylogenetic change or lineage-specific adaptations. For example, D. novaehollandiae and G. gallus are both omnivorous and therefore may retain ORs as a variety of different odorants are relevant when foraging. The three Neoaves species we selected are not generalists, instead C. anna is mostly nectivorous, M. vitellinus is mostly frugivorous, and T. guttata is a granivore. Our understanding of phylogenetic patterns of OR diversity will continue to change with improved genome assembly and taxonomic sampling. For example, by sequencing additional manakin species, we will be able to determine whether OR repertoire is elaborated within a frugivorous family or if counts vary substantially within individual lineages.

Toward a better understanding of the avian OR subgenome

Many key features of OR subgenomes and olfaction generally are well-characterized in mammals, but not in birds (Olender et al. 2008; Niimura et al. 2014). There are no previous reports of using assemblies with long-read sequencing to search for an OR repertoire in birds,
and until now, no previous reports of an expansive gamma-c OR repertoire outside of G. gallus and T. guttata assemblies. To date, there has been only one report of transcriptome sequencing an avian olfactory epithelium, a critical undertaking considering that ORs are expressed in nonolfactory tissues and some identified ORs may be nonfunctional despite having open reading frames (Pluznick et al. 2009; Sin et al. 2019). Although Sin et al. (2019) do show the expression of at least three gamma-c ORs in the Oceanodroma leucorhoa olfactory epithelium, it still leaves the expression of a potentially large number of gamma-c genes uncharacterized. A current priority would be to sequence the olfactory epithelium transcriptome from species with high-quality genome assemblies to understand the extent to which a high number of gamma-c ORs recovered from the genome are functionally expressed in the olfactory epithelium.

Information about which ORs are expressed in the olfactory epithelium will also help with functional testing to identify the binding properties of avian ORs in a process known as “deorphanizing.” To date, no avian ORs have been deorphanized, and so it remains unclear what ligands are bound by any avian ORs. The diverse gamma-c ORs, unique to birds, are of particular interest. Avian olfaction is in many fundamental ways a frontier in the field of sensory biology.

There is also a great deal left to be learned about the molecular evolutionary mechanisms that have given rise to the diversity of gamma-c genes in birds. First, enhanced spatial information on the physical location of OR genes will be informative for understanding the genetic processes at play. With most loci still scattered among unmapped scaffolds, it is somewhat unclear how clustered these loci are. That said, the G. gallus Ggal6 OR repertoire is highly spatially localized (Figs 2 and 3), a pattern that is likely to be found in other species as well. Given spatial clustering, and extremely short branch-lengths, gamma-c species-specific clades could be a result of gene conversion among loci, but further study is needed. In the passerine bird MHC, endogenous retroviral elements may have played a role in generating gene family diversity in MHC Class II (Balakrishnan et al. 2010), and the same could be the case in the gamma-c radiation across birds. We note, however, that high repeat content is a general pattern across avian microchromosomes and is not restricted to OR and MHC regions (Fig. 3, Burt 2002).

Long-read-assembly methods better characterize complex gene families

Our increased OR counts in long-read assemblies contribute to the growing literature quantifying the advantages of third-generation (long-read) sequencing technology over second-generation (short-read) sequencing technology. Third-generation sequencing has greatly improved the detection of tandem repeats generated by long terminal repeat retrotransposons, microsatellites, homonucleotide stretches, and repetitive regions (Mason et al. 2016; Kapusta and Suh 2017; Korlach et al. 2017). Large gene families found in clusters like the ORs described here may show the greatest improvement following the incorporation of long-read data. Long-read sequencing has already led to better characterization of MHC loci in birds, and also greatly increased the resolution and count of vomeronasal receptors in mammals (Larsen et al. 2014; He et al. 2021).

Even with long-read sequencing, the chromosomal location of all ORs remains largely unresolved. With the exception of the chicken, ORs in the long-read assemblies that we analyzed largely mapped to unassigned scaffolds, including the largest OR cluster in each assembly (Fig. 2). The inability to assign these scaffolds to a chromosome is likely related to the expansion of duplicate regions that contain OR loci and the high repeat content found in microchromosomes (Fig. 3, Burt 2002). Indeed the assignment of an OR-containing region to the hummingbird female-specific W chromosome is likely spurious and driven by the repetitive sequences in these regions. Bird chromosomes are highly syntenic, suggesting that the large OR cluster on the chicken microchromosome 33 likely matches homologous chromosomes (Nanda et al. 2011). Manual curation of these regions may be required to resolve the remaining uncertainties. Other solutions to this complexity include physically mapping OR loci to chromosomes, and/or using approaches less dependent on assembly. Sin and colleagues (2019) incorporated assessment of depth of coverage in their OR quantification pipeline, an approach that should be informative in the face of varying assembly quality.

Acknowledgments

We would like to thank the editors, two anonymous reviewers, and Dr Maude Baldwin for helpful discussion of the analyses presented here. In addition, we thank Drs Michael Brewer and John Stiller, who are serving on R.J.D.’s dissertation committee. The D. novaehollandiae, C. anna, and T. guttata genomes were generated as part of the Vertebrate Genomes Project (https://vgp.github.io/genomereark/), and prepublication public access to these, now published, data made these analyses possible. Likewise, we thank the Manakin Genomics Research Coordination Network (https://www.manakinsrcn.org/) for generating public genomic resources for the manakin species analyzed here. This ma-
terial is based upon work conducted while C.N.B. was serving at the National Science Foundation.

**Funding**

This work was supported by National Science Foundation awards [grant numbers RCN-1457541, IOS-1456612].

**Supplementary data**

Supplementary data available at ICB online.

**Data availability**

No new data were generated or analyzed in support of this research.

**References**

Audubon JJ. 1826. Account of the habits of the turkey buzzard, *Vultur aura*, particularly with the view of exploding the opinion generally entertained of its extraordinary power of smelling. Edinburgh New Phil J 2:172–84.

Balakrishnan CN, Ekkblom R, Völker M, Westerdahl H, Godinez R, Kotkiewicz H, Burt DW, Graves T, Griffin DK, Warren WC et al. 2010. Gene duplication and fragmentation in the zebra finch major histocompatibility complex. BMC Biol 8: 29.

Baldwin MW, Toda Y, Nakagita T, O’Connell MJ, Klasing KC, Misaka T, Edwards SV, Liberles SD. 2014. Evolution of sweet taste perception in hummingbirds by transformation of the ancestral umami receptor. Science 345:929–33.

Bang BG, Cobb S. 1968. The size of the olfactory bulb in 108 species of birds. Auk 85:55–61.

Beichman AC, Koeplli K-P, Li G, Murphy W, Dobrynin P, Kliver S, Tinker MT, Murray MJ, Johnson J, Lindblad-Toh K et al. 2019. Aquatic adaptation and depleted diversity: a deep dive into the genomes of the sea otter and giant otter. Mol Biol Evol 36:2631–55.

Breer H. 2003. Olfactory receptors: molecular basis for recognition and discrimination of odors. Anal Bioanal Chem 377:427–33.

Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65:175–87.

Burt DW. 2002. Origin and evolution of avian microchromosomes. Cytogenet Genom Res 96:97–112.

Darwin C. 1891. Journal of researches into the natural history and ecology of the countries visited during the voyage of H.M.S. Beagle round the world, under the command of Capt. Fitz Roy, R.N. London: Ward, Lock & Co.

Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC et al. 2020. Dense sampling of bird diversity increases power of comparative genomics. Nature 587:252–7.

Goujon M, McWilliam H, Li W, Valentin F, Squirizzato S, Paern J, Lopez R. 2010. A new bioinformatics analysis tools framework at EMBL–EBI. Nucleic Acids Res 38:W695–9.

Gwinner H, Berger S. 2008. Starling males select green nest material by olfaction using experience-independent and experience-dependent cues. Anim Behav 75: 971–6.

Hagelin JC. 2007. The citrus-like scent of crested auklets: reviewing the evidence for an avian olfactory ornament. J Ornithol 148:195–201.

He K, Minias P, Dunn PO. 2021. Long-read genome assemblies reveal extraordinary variation in the number and structure of MHC loci in birds. Genome Biol Evol 13: evaa270

Hill A. 1905. Can birds smell? Nature 71:318–9.

Huynh AV, Rice AM. 2019. Conspecific olfactory preferences and interspecific divergence in odor cues in a chickadee hybrid zone. Ecol Evol 9:9671–83.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–31.

Kapusta A, Suh A. 2017. Evolution of bird genomes—a transposon’s-eye view. Ann N Y Acad Sci 1389: 164–85.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–80.

Khan I, Yang Z, Maldonado E, Li C, Zhang G, Gilbert MTP, Jarvis ED, O’Brien SJ, Johnson WE, Antunes A. 2015. Olfactory receptor subgenomes linked with broad ecological adaptations in Sauropsida. Mol Biol Evol 32:2832–43.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–20.

Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, Jarvis ED. 2017. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. GigaScience 6:1–16

Krause ET, Krüger O, Kohlmeier P, Caspers BA. 2012. Olfactory kin recognition in a songbird. Biol Lett 8:327–9.

Larsen ON, Wahlberg M, Christensen-Dalsgaard J. 2020. Amphibious hearing in a diving bird, the great cormorant (*Phalacrocorax carbo sinensis*). J Exp Biol 223: jeb217265

Larsen PA, Heilman AM, Yoder AD. 2014. The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. BMC Genomics 15:720.

Le Duc D, Renaud G, Krishnan A, Almén MS, Huynen L, Prohaska SJ, Ongyerth M, Bitarello BD, Schüth HB, Hofreiter M et al. 2015. Kiwi genome provides insights into evolution of a nocturnal lifestyle. Genome Biol 16:147.

Lee CM, Barber GP, Casper J, Clawson H, Diekhans M, Gonzalez JN, Hinrichs AS, Lee BT, Nassar LR, Powell CG et al. 2020. UCSC Genome Browser enters 20th year. Nucleic Acids Res 48:D756–61.

Letunic I, Bork P. 2019. Interactive tree of life (iTOl) v4: recent updates and new developments. Nucleic Acids Res 47:W256–9.

Liu J, Wang Z, Li J, Xu L, Liu J, Feng S, Guo C, Chen S, Ren Z, Rao J et al. 2021. A new emu genome illuminates the evolution of genome configuration and nuclear architecture of avian chromosomes. Genome Res 31:497–511.
Mason AS, Fulton JE, Hocking PM, Burt DW. 2016. A new look at the LTR retrotransposon content of the chicken genome. BMC Genomics 17:688.

McRae JF, Mainland JD, Jaeger SR, Adipietro KA, Matsunami H, Newcomb RD. 2012. Genetic variation in the odorant receptor OR2J3 is associated with the ability to detect the “grassy” smelling odor, cis-3-hexen-1-ol. Chem Senses 37:585–93.

Michelsen WJ. 1959. Procedure for studying olfactory discrimination in pigeons. Science 130:630–1.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–4.

Mombaerts P. 2004. Odorant receptor gene choice in olfactory sensory neurons: the one receptor–one neuron hypothesis revisited. Curr Opin Neurobiol 14:31–6.

Montague MJ, Li G, Gandolfi B, Khan R, Aken BL, Searle SMJ, Minx P, Hillier LW, Koboldt DC, Davis BW et al. 2014. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. Proc Natl Acad Sci 111:17230–5.

Nanda I, Benisch P, Fetting D, Haaf T, Schmid M. 2011. Synteny conservation of chicken macrochromosomes 1–10 in different avian lineages revealed by cross-species chromosome painting. Cytogenet Genome Res 132:165–81.

Nevitt GA, Losekoot M, Weimerskirch H. 2008. Evidence for olfactory search in wandering albatross, Diomedea exulans. Proc Natl Acad Sci 105:4576–81.

Niimura Y, Matsui A, Touhara K. 2014. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res 24:1485–96.

Niimura Y, Nei M. 2005. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. Proc Natl Acad Sci 102:6039–44.

Niimura Y, Nei M. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One 2:e708.

Niimura Y. 2009. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. Genome Biol Evol 1:34–44.

Niimura Y. 2013. Identification of olfactory receptor genes from mammalian genome sequences. In: Crasto CJ, editor. Olfactory receptors. Methods in molecular biology (methods and protocols). Totowa (NJ): Humana Press. p. 39–49.

Olender T, Lancet D, Nebert DW. 2008. Update on the olfactory receptor (OR) gene superfamily. Hum Genomics 3:87.

Olender T, Lancet D, Nebert DW. 2008. Update on the olfactory receptor (OR) gene superfamily. Hum Genomics 3:87.

Olender T, Lancet D, Nebert DW. 2008. Update on the olfactory receptor (OR) gene superfamily. Hum Genomics 3:87.

Pluznick JL, Zou D-J, Zhang X, Yan Q, Rodriguez-Gil DJ, Eisner C, Wells E, Greer CA, Wang T, Firestein S et al. 2009. Functional expression of the olfactory signaling system in the kidney. Proc Natl Acad Sci 106:2059–64.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Gedman GL et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592:737–46.

Saito H, Chi Q, Zhuang H, Matsunami H, Mainland JD. 2009. Odor coding by a mammalian receptor repertoire. Sci Signal 2:ra9.

Santos PSC, Kellermann T, Uchanska-Ziegler B, Ziegler A. 2010. Genomic architecture of MHC-linked odorant receptor gene repertoires among 16 vertebrate species. Immunogenetics 62:569–84.

Silva MC, Chibucos M, Munro JB, Daughtery S, Coelho MM, Silva JC. 2020. Signature of adaptive evolution in olfactory receptor genes in Cory’s Shearwater supports molecular basis for smell in procellariiform seabirds. Sci Rep 10:543.

Steiger SS, Kuryshiev VY, Stensmyr MC, Kempenaers B, Mueller JC. 2009. A comparison of reptilian and avian olfactory receptor gene repertoires: species-specific expansion of group γ genes in birds. BMC Genomics 10:446.

Stoddard MC, Eyster HN, Hogan BG, Morris DH, Soucy ER, Inouye DW. 2020. Wild hummingbirds discriminate nonspectral colors. Proc Natl Acad Sci 117:15112–22.

Vandeweghe MW, Mangum SF, Gabaldon T, Castoe TA, Ray DA, Hoffmann FG. 2016. Contrasting patterns of evolutionary diversification in the olfactory repertoires of reptile and bird genomes. Genome Biol Evol 8:470–80.

Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, Li C, White S, Xiong Z, Fang D et al. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat Genet 45:701–6.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S et al. 2010. The genome of a songbird. Nature 464:757–62.

Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, P Ni, Hu L, Liu Y, Hou H et al. 2013. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. Nature Genetics 45:563–566.

Zhang G, Li B, Li C, Gilbert MTP, Jarvis ED, Wang J, The Avian Genome Consortium. 2014. Comparative genomic data of the Avian Phylogenomics Project. GigaScience 3.26

Zimin AV, Marçais G, Puig D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. Bioinformatics 29:2669–77.