A high-quality Genome and Comparison of Short versus Long Read Transcriptome of the Palaearctic duck Aythya fuligula (Tufted Duck)

Manuscript Number: GIGA-D-21-00072

Full Title: A high-quality Genome and Comparison of Short versus Long Read Transcriptome of the Palaearctic duck Aythya fuligula (Tufted Duck)

Article Type: Data Note

Funding Information:
- European Cooperation in Science and Technology (CA15112)
- Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (BioDATEN)

Abstract:
Background The tufted duck is a non-model organism that suffers high mortality in highly pathogenic avian influenza outbreaks. It belongs to the same bird family (Anatidae) as the mallard, one of the best-studied natural hosts of low-pathogenic avian influenza viruses. Studies in non-model bird species are crucial to disentangle the role of the host response in avian influenza virus infection in the natural reservoir. Such endeavour requires a high-quality genome assembly and transcriptome.

Findings This study presents the first high-quality, chromosome-level reference genome assembly of the tufted duck using the Vertebrate Genomes Project pipeline. We sequenced RNA (cDNA) from brain, ileum, lung, ovary, spleen and testis using Illumina short-read and PacBio long-read sequencing platforms, which was used for annotation. We found 34 autosomes plus Z and W sex chromosomes in the curated genome assembly, with 99.6% of the sequence assigned to chromosomes. Functional annotation revealed 14,099 protein-coding genes that generate 111,934 transcripts, which implies an average of 7.9 isoforms per gene. We also identified 246 small RNA families.

Conclusions
This annotated genome contributes to continuing research into the host response in avian influenza virus infections in a natural reservoir. Our findings from a comparison between short-read and long-read reference transcriptomics contribute to a deeper understanding of these competing options. In this study, both technologies complemented each other. We expect this annotation to be a foundation for further comparative and evolutionary genomic studies, including many waterfowl relatives with differing susceptibilities to the avian influenza virus.

Corresponding Author: Ralf Christian Mueller
Max Planck Institute of Animal Behavior: Max-Planck-Institut fur Verhaltensbiologie Radolfzell, GERMANY

First Author: Ralf Christian Mueller
Patrik Ellström
Kerstin Howe
| Question                                                                 | Response |
|-------------------------------------------------------------------------|----------|
| Are you submitting this manuscript to a special series or article collection? | No       |
| **Experimental design and statistics**                                   | Yes      |
| Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. |          |
| Have you included all the information requested in your manuscript?      | Yes      |
| **Resources**                                                            | Yes      |
| A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the |          |
Methods section. Authors are strongly encouraged to cite [Research Resource Identifiers (RRIDs)](https://rrids.info) for antibodies, model organisms and tools, where possible.

| Have you included the information requested as detailed in our [Minimum Standards Reporting Checklist](#)? |
|---|
| No |

### Availability of data and materials

All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in [publicly available repositories](https://www.ncbi.nlm.nih.gov/SeqR/) (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

| Have you met the above requirement as detailed in our [Minimum Standards Reporting Checklist](#)? |
|---|
| No |

If not, please give reasons for any omissions below.

- as follow-up to "Availability of data and materials"

Raw RNA sequencing data was uploaded to NCBI's sequence read archive but not published yet. Supporting data is available through Figshare (temporary links provided in the manuscript). The source code for scripts used in this study is available through the corresponding author's GitLab project page, as indicated in the manuscript.
This is pdfTeX, Version 3.14159265-2.6-1.40.21 (TeX Live 2020/W32TeX)
(preloaded format=pdflatex 2020.5.12)  3 MAR 2021 08:30
entering extended mode
  restricted \write18 enabled.
  %&-line parsing enabled.
**20210303_rmueller_tufted_duck.tex
  ./20210303_rmueller_tufted_duck.tex
LaTeX2e <2020-02-02> patch level 5
L3 programming layer <2020-05-05>

! LaTeX Error: File `oup-contemporary.cls' not found.

Type X to quit or <RETURN> to proceed,
or enter new name. (Default extension: cls)

Enter file name:
! Emergency stop.
<read *>

1.12 ^^M

*** (cannot \read from terminal in nonstop modes)

Here is how much of TeX's memory you used:
  22 strings out of 480681
  590 string characters out of 5908536
  236875 words of memory out of 5000000
  15943 multiletter control sequences out of 15000+600000
  532338 words of font info for 24 fonts, out of 8000000 for 9000
  1141 hyphenation exceptions out of 8191
  14i,0n,17p,120b,10s stack positions out of
  5000i,500n,10000p,200000b,80000s
! ==> Fatal error occurred, no output PDF file produced!
DATA NOTE

A high-quality Genome and Comparison of Short versus Long Read Transcriptome of the Palaearctic duck Aythya fuligula (Tufted Duck)

Ralf C Mueller¹,²,* , Patrik Ellström³, Kerstin Howe⁴, Marcela Uliano-Silva⁴, Richard I Kuo⁵, Katarzyna Miedzinska⁵, Amanda Warr⁵, Olivier Fedrigo⁶, Bettina Haase⁶, Jacquelyn Mountcastle⁶, William Chow⁴, James Torrance⁴, Jonathan Wood⁴, Josef D Järhult³, Mahmoud M Naguib⁷, Björn Olsen³, Erich D Jarvis⁸, Jacqueline Smith⁵, Lél Eöry⁵ and Robert HS Kraus¹,²

¹Department of Migration, Max Planck Institute of Animal Behavior, Radolfzell, Germany and ²Department of Biology, University of Konstanz, Konstanz, Germany and ³Department of Medical Sciences, Zoonosis Science Center, Uppsala University, Uppsala, Sweden and ⁴Tree of Life, Wellcome Sanger Institute, Cambridge, UK and ⁵The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK and ⁶Vertebrate Genome Laboratory, The Rockefeller University, New York, NY and ⁷Department of Medical Biochemistry and Microbiology, Zoonosis Science Center, Uppsala University, Uppsala, Sweden and ⁸Vertebrate Genome Laboratory and HHMI, The Rockefeller University, New York, NY

*rmueller@ab.mpg.de; ORCID: 0000-0002-7348-9164

Abstract

Background The tufted duck is a non-model organism that suffers high mortality in highly pathogenic avian influenza outbreaks. It belongs to the same bird family (Anatidae) as the mallard, one of the best-studied natural hosts of low-pathogenic avian influenza viruses. Studies in non-model bird species are crucial to disentangle the role of the host response in avian influenza virus infection in the natural reservoir. Such endeavour requires a high-quality genome assembly and transcriptome. Findings This study presents the first high-quality, chromosome-level reference genome assembly of the tufted duck using the Vertebrate Genomes Project pipeline. We sequenced RNA (cDNA) from brain, ileum, lung, ovary, spleen and testis using Illumina short-read and PacBio long-read sequencing platforms, which was used for annotation. We found 34 autosomes plus Z and W sex chromosomes in the curated genome assembly, with 99.6% of the sequence assigned to chromosomes. Functional annotation revealed 14,099 protein-coding genes that generate 111,934 transcripts, which implies an average of 7.9 isoforms per gene. We also identified 246 small RNA families. Conclusions This annotated genome contributes to continuing research into the host response in avian influenza virus infections in a natural reservoir. Our findings from a comparison between short-read and long-read reference transcriptomics contribute to a deeper understanding of these competing options. In this study, both technologies complemented each other. We expect this annotation to be a foundation for further comparative and evolutionary genomic studies, including many waterfowl relatives with differing susceptibilities to the avian influenza virus.

Key words: Tufted duck; Aythya fuligula; Genome annotation; Transcriptome sequencing; Vertebrate Genomes Project; Iso-Seq; PacBio; RNA-Seq; small RNA
Background

The tufted duck (Aythya fuligula) is a non-model organism that has received attention because of its role in the zoonotic ecology of avian influenza A viruses (AIV). As a member of the Anatidae family of ducks, it is closely related to the mallard (Anas platyrhynchos), the primary natural host of AIV [1, 2, 3, 4, 5]. However, in contrast to mallards, which carry this virus largely asymptomatically, tufted ducks are less commonly infected but suffer high mortality in outbreaks of highly pathogenic avian influenza (HPAI) [6, 7, 8, 9]. The tufted duck is a diving duck. Its breeding range is throughout northern Eurasia where it is largely a seasonal short-distance migrant. Although it generally feeds in deeper waters than mallards and other dabbling ducks, it generally shares its habitat and roosting areas with these and many other waterfowl species, even leading to a high rate of interspecific hybridisation [10, 11]. Hence, given a frequent exposure to AIVs circulating in such habitats, the differences in susceptibility to and outcome of AIV infection between these species are likely related to genetic differences affecting, e.g., receptor expression or host response. For example, the resistance of mallards against severe HPAI infection has been partially attributed to the presence of the RIG-I gene and its strong interferon response, in contrast to chickens that lack this gene and develop severe disease upon HPAI infection [12]. Future studies in non-model bird species such as the tufted duck are important to disentangle the role of the host response and other genetic factors in AIV infection and aid in our understanding of the zoonotic ecology of AIV in the natural reservoir. A prerequisite for such studies is a well-assembled and annotated genome and transcriptome [13].

Developments in omic sequencing technologies over the last two decades have revolutionised biology. Instead of studying single genes and their products, whole genomes and transcriptomes can now be readily sequenced and assembled at a lower cost than before. Massive parallelisation and high throughput in next-generation sequencing (NGS) have decreased sequencing costs and ultimately increased sequencing depth [14]. This allows for whole-genome sequencing of any species and opens up new possibilities for in-depth studies related to infection biology and host response to external stressors beyond model species in which a rich genetic toolbox can be deployed [15, 16]. Non-model species are frequently understudied, yet they are exposed to environmental stressors like infectious diseases, which can transmit to livestock and humans [17]. Approximately 70% of human infectious diseases are zoonoses [18, 19]. An in-depth understanding of a pathogen’s zoonotic ecology in the animal reservoir is important to prevent human infections. This, in turn, requires studies of host–pathogen interactions at the genetic level. NGS helps bridge the gap between model and non-model species [20] and with third-generation long-read based sequencing, high-quality reference genomes are now also available for non-model organisms, as in the Vertebrate Genomes Project (VGP) [21]. This is supplemented by affordable long-read RNA sequencing, making de novo assembled transcriptomes unnecessary [22].

In de novo sequencing/assembly, high sequencing depth with short reads (< 250 bp) is used in an attempt to recover whole genomes, which decreases uncertainty during the assembly process. While this results in the possibility of studying previously uncharacterised species without relying on reference genomes, there are still limitations. For instance, short reads make it difficult to assemble and identify very similar genes (e.g., paralogs), regions of low complexity (e.g., tandem repeats), or GC-rich microchromosomes of birds [23, 16]. Illumina sequencing produces short reads, which often fail to assemble these difficult regions leading to artefacts, mis-assemblies or incorrect sequence reconstructions [24, 25, 21]. On the other hand, the high sequencing base call quality accurately recovers nucleotide sequences in critical regions, e.g., splice junction sites. The introduction of PacBio’s Single Molecule, Real-Time (SMRT) sequencing offers the opportunity to sequence very long reads (N50 > 10kb) of unfragmented DNA [26, 27] leading to lower assembly structural errors and ambiguity. As there is no prior amplification needed, there is also no PCR-induced variance of coverage in the sequences. However, error rates can be high for continuous long-read (CLR) technologies, resulting in more single-nucleotide errors. This can be corrected by increasing sequencing depth, using specialised polishing algorithms with long reads or high accuracy short reads [21]. Exploiting each technology’s strength facilitates more in-depth and unbiased genome studies while also enabling us to draw much more robust and detailed conclusions regarding research questions, e.g., identifying genes and gene products that previously escaped detection [24].

Transcriptome annotation of the genome used to be constrained to either low-throughput and costly cDNA cloning or Illumina’s high-throughput short-read RNA sequencing (RNA-Seq). High-quality short-read RNA sequencing combined with a reference genome allows for a reasonable transcriptome reconstruction. However, there are some caveats: Due to alternative splicing, a single gene can have multiple alternative variants (isoforms) and as a consequence can be translated into proteins with different functions [25, 28]. Illumina short reads need to be assembled into transcripts, which can lead to incompleteness and errors in transcript model reconstruction. This biases the correct inference of isoforms and thus misses the transcriptome’s complexity [25, 29]. Furthermore, since short-read sequencing is also limited in GC-rich regions and regions of low sequence complexity [24], not all transcripts are recovered. Instead, in full-length transcript isoform sequencing (e.g., PacBio Iso-Seq) of mRNA, the single direct reads from 5' to 3' usually do not need to be assembled and, thus prevent assembly ambiguities, which facilitates the detection of novel isoforms [30, 31] and accurate reconstruction of transcript structure [24]. This third-generation transcript sequencing technology also allows for a much more detailed functional annotation. It is crucial to recover as many isoforms as possible for functional studies [32, 33, 34]. For example, the immunome comprises a set of genes associated with immune processes and is a heavily complex portion of the genome [35, 22]. Thus, the benefits of 3rd generation sequencing must be exploited to their fullest potential to facilitate studies in bird immunogenomics [36] and improve databases such as the Avian Immunome DB for comparative analyses in non-model bird species [37].

Data Description

This study presents the first high-quality, chromosome-level reference genome assembly of the tufted duck using the VGP pipeline [21], which we annotated with short- and long-read transcripts from six different tissues, and short reads from small RNAs from the same tissues. We further contrasted different sequencing technologies and bioinformatics pipelines, and tissue-specific comparisons of expressed genes. This annotation serves as a foundation for further comparative and evolutionary genomics, and gene expression experiments in
the tufted duck and its many waterfowl relatives with different susceptibilities to AIV [36, 38].

Analyses

Reference genome assembly. The reference genome fully met VGP’s 6.7.P2.Q40.C99 metric standards [21]. This includes contig NG50 > 10^6 bp, scaffold NG50 > 10^7 bp, most of the genome separated into haplotypes (P2), Phred-scaled base accuracy > Q40, and 99.5% of the assembly assigned to chromosomes. The contig NG50 of the tufted duck genome assembly was 17.8 Mb. The curation resulted in 34 removals of misjoins, 34 joins previously missed, and four removals of false duplications. This reduced the primary assembly length by 0.8% and increased the scaffold NG50 by 18.7% to 85.9 Mb whilst decreasing the scaffold number from 123 to 105. Chromosomal-scale units were identified and named by size. Out of the expected 39 chromosome pairs according to the karyotype [39], 34 autosomes plus Z and W could be identified, with 99.6% of the sequence assigned to them (Tab. 1, Fig. 1).

Table 1. Assembly statistics of the tufted duck genome.

| Statistic                                      | Value               |
|-----------------------------------------------|---------------------|
| Genome coverage [x]                          | 64.03               |
| Total sequence length [bp]                   | 1,127,004,725       |
| Ungapped sequence length [bp]                | 1,117,587,328       |
| Number of scaffolds                           | 105                 |
| Scaffolds assigned to chromosomes            | 36 + 1 mitochondrion|
| Unplaced scaffolds                            | 68                  |
| Contig NG50 [bp]                             | 17,816,505          |
| Scaffold NG50 [bp]                           | 85,905,639          |
| Base-call error in 10 kb                     | less than 1 nucleotide |

The raw data has been deposited in the GenomeArk repository [40]. The curated primary assembly has been deposited in NCBI under accession number GCF_009819795.1 [41] and can be browsed in the Genome Data Viewer [42].

In the Illumina short-read RNA-Seq and PacBio long-read RNA-Seq reads. With the Illumina RNA-Seq, an average of 97.65% (SD = 0.51%) of the reads were retained after adapter and quality trimming. For the PacBio Iso-Seq, an average of 80.57% (SD = 3.84%) full-length non-chimeric reads were retained after error correction (Tab. 2).

On average, HISAT2 could map 93.21% (SD = 0.64%) Illumina RNA-Seq short reads and Minimap2 could map 97.39% (SD = 1.34%) PacBio Iso-Seq long reads to the reference genome. The average read length of the long-read data was 1246.0 nt (SD = 262.0), an order of magnitude longer than that of the short reads (average 131.5 nt, SD = 0.7). However, the average number of reads was almost 600-fold higher with short-read data (129,411,992; SD = 11,989,532) compared to long-read data (217,293; SD = 143,219), implying a 60 times higher sequencing depth with the short reads. Not surprisingly, StringTie2 (in the short-read pipeline) assembled more transcript models and inferred more genes and exons than the long-read pipeline. This was true for all tissues except in the lung; more transcript models were inferred in the long-read pipeline. Lung RNA (cDNA) was sequenced on two Zero-mode waveguides (ZMW) and produced the highest numbers of sub-reads and full-length, non-chimeric reads (FLNC) after processing. The highest number of genes in each pipeline was predicted for ovary (Illumina) and brain (PacBio). The highest number of transcript models was predicted for ovary (Illumina) and lung (PacBio). The same pattern applies to predicted exons (Tab. 3).

In the Illumina short-read RNA-Seq annotations, more exons per gene were recovered on average than in the long-read pipeline (Tab. 3). The distribution of exons per gene in both pipelines generally followed the same pattern with a decreasing number of genes as the exons per gene increased. However, there is a remarkable difference in the short-read data, which recovered more two-exon genes than single-exon genes for all tissues (Fig. 2).

In the PacBio long-read Iso-Seq annotations, more transcripts per gene were recovered on average than in the short-
Table 2. Illumina RNA-Seq reads before and after trimming, and PacBio Iso-Seq reads before and after error correction. Zero-mode waveguide (ZMW) yield and the number of full-length, non-chimeric reads (FLNC). Two ZMW were used for lung and ovary.

| Platform | brain | ileum | lung | ovary | spleen | testis |
|----------|-------|-------|------|-------|--------|--------|
| Illumina | 71,648,303 | 72,410,462 | 63,191,981 | 72,442,392 | 68,315,029 | 53,727,262 |
| Number of trimmed (paired) | 70,025,835 | 70,902,259 | 61,059,701 | 70,892,177 | 66,813,691 | 52,652,923 |
| PacBio | 9,249,099 | 12,353,569 | 19,206,097 | 1,279,561 | 8,401,115 | 12,616,953 |
| Number of subreads | 158,698 | 415,314 | 529,108 | 68,112 | 167,077 | 288,914 |
| Number of ccs | 12,353,569 | 19,206,097 | 1,279,561 | 8,401,115 | 12,616,953 | |
| ZMW yield [%] | 7.9 | 11.6 | 7.9 | 11.6 | 11.6 | 11.6 |
| Average number of passes | 58.3 | 41.53 | 26.46 | 3.4 | 16.71 | 28.9 |
| Number of FLNC | 133,684 | 343,634 | 432,817 | 49,887 | 134,124 | 234,423 |

Table 3. Annotation per tissue and pipeline.

| Platform | brain | ileum | lung | ovary | spleen | testis |
|----------|-------|-------|------|-------|--------|--------|
| Illumina | 92.92 | 92.98 | 93.42 | 94.33 | 92.44 | 93.19 |
| PacBio | 96.66 | 98.13 | 97.92 | 99.40 | 95.67 | 96.57 |
| Number of genes | 22,348 | 20,838 | 20,692 | 32,046 | 21,608 | 29,225 |
| PacBio | 15,776 | 10,813 | 12,912 | 8,862 | 6,773 | 11,746 |
| Number of transcripts | 44,808 | 40,968 | 43,877 | 77,997 | 42,000 | 57,758 |
| PacBio | 37,601 | 35,284 | 46,587 | 19,513 | 14,030 | 28,852 |
| Number of exons | 422,741 | 395,719 | 412,225 | 569,566 | 383,679 | 483,444 |
| PacBio | 138,038 | 217,391 | 243,566 | 119,080 | 71,211 | 160,380 |
| Exons per gene (average) | 19.8 | 19.0 | 19.9 | 17.8 | 17.8 | 16.5 |
| PacBio | 8.7 | 20.1 | 18.9 | 13.4 | 10.5 | 13.7 |
| Transcripts per gene (average) | 2.0 | 2.0 | 2.1 | 2.4 | 1.9 | 2.0 |
| PacBio | 2.4 | 3.3 | 3.6 | 2.2 | 2.1 | 2.5 |

Table 4. Functional annotation categorised by different matches. Note: UniRef50 total hits also includes 5’ degraded records, whereas the match classes only include full-length records.

| Platform | genes | transcripts | isoforms |
|----------|-------|-------------|---------|
| Uniref50 total hits | 49,746 | 345,870 | 7.0 |
| Uniref50 full match | 17,911 | 208,274 | 11.6 |
| Uniref50 90% match | 13,024 | 99,737 | 7.7 |
| Uniref50 50% match | 4,937 | 12,197 | 2.5 |
| Uniref50 < 50% match | 3,208 | 6,540 | 2.0 |
| Uniref50 ≥ 90% match | 3,474 | 4,447 | 1.8 |
| Uniref50 ≥ 50% match | 14,731 | 138,674 | 8.0 |
| No hit (but full-length) | 14,099 | 111,934 | 7.9 |

Out of 78,860 full-length transcripts with no UniRef50 hit, 62,147 were flagged as potentially protein-coding (and the remainder as nonsense-mediated decay), and 26,489 as single-exonic while the remainder as composed of two or more exons.

Tissue-specific expression and intersections. The number of supported protein-coding genes was generally higher with Illumina transcriptome reconstruction than with that of PacBio. The short-read pipeline supported 11,165 annotated genes across all tissues. The highest number of exclusively expressed genes was in testis (988) followed by ovary, brain, spleen, ileum and lung (Fig. 4). The long-read pipeline supported 2,475 annotated genes across all tissues. The highest number of exclusively expressed genes was in brain (779) followed by testis, ileum, lung, ovary and spleen (Fig. 5). Overlap of expressed genes from each pipeline can be found in Fig. S5.

Small RNA analyses. For the TruSeq small RNA sequencing data, 95.8% (SD = 3.51%) of the reads were retained (on average) after adapter and quality trimming (Tab. 5). Overall, STAR could map on average 93.91% (SD = 5.28%) of these reads to the reference genome, which divides into 67.99% (SD = 13.75%) uniquely mapped reads and 25.92% (SD = 12.08%) multi-mapped reads (up to 10 loci). Cufflinks predicted the highest number of genes in the spleen (33,133) followed by testis (31,205). The remaining tissues had much lower numbers of genes ranging from 8,442 (ileum) to 13,606 (brain). The same pattern applies to the number of predicted transcripts and exons (Tab. 5).

On average, each transcript was composed of 1.3 exons (SD = 0.2 exons). The distribution of single-exon and multi-exon transcripts shows a clear trend towards single-exon transcripts and a quickly diminishing number of multi-exon transcripts. However, this was less pronounced in spleen but even more so in testis (Fig. 6). The generally decaying length distribution of transcripts shows two clear peaks at 20 bp and 50 bp except for testis with the first peak at 30 bp. Except for spleen and testis data, there are very few transcripts > 200 bp (Fig. S2).
Scanning the genome \textit{(in silico)} for Rfam’s covariance models of RNAs resulted in 1,234 hits. The same scan on the assembled small RNA transcripts \textit{(in vitro)} revealed an average of 346.5 hits (SD = 26.9) across all tissues. After removing lower-scoring overlaps and hits with E-value larger than 5.0E-4 from the cmrscan result, 1,076 distinct features were predicted in the tufted duck genome. In the tissue’s small RNA assemblies, 327.5 (SD = 26.5) features were annotated, on average, with the same filtering (Tab. 6). On average 93.25\% (SD = 1.09\%) of the \textit{in vitro} annotated features were predicted by the \textit{in silico} scan. Furthermore, 22.2 (SD = 2.3) features were annotated on average, which were not predicted \textit{in silico} (Tab. 6).

After further filtering for unique RNA families (Rfam accession numbers/covariance models), 306 distinct RNA families were predicted in the genome, with 246 annotated in all tissues (pooled). The number of predicted and annotated covariance models overlapped for 237 RNA families, while 69 were only identified in the genome scan, and nine were only identified in the pooled tissue annotations.

**Discussion**

In this study, we present the first chromosome-level reference genome assembly of the tufted duck. The genome contiguity is on par with other reference bird genome assemblies that used long reads like mallard [43, 44], chicken [45], and recent VGP-pipeline generated zebra finch [46]. The assembly’s contig NG50 of 17.8 Mbp is comparable to chicken (17.7 Mbp) but considerably higher than in mallard (5.7 Mbp) and zebra finch (4.4 Mbp). The assembly’s scaffold NG50 of 85.9 Mbp is considerably higher than in mallard (76.3 Mbp) and zebra finch (70.9 Mbp), and considerably higher than in chicken (20.8 Mbp). All our mapping results from the Illumina short-read and PacBio long-read RNA pipelines confirm full-adherence to the VGP 6.7.P2.Q40.C99 standard, which also implies transcript mapability > 80\% [21].

The higher numbers of recovered genes, transcripts and exons from the short-read transcript pipeline may be mainly explained by the higher sequencing depth and further reinforced by the different RNA preparation protocol. With Illumina, virtually all trimmed, paired-end reads were kept for...
mapping to the genome while with PacBio, only 5’ cap-selected and full-length, non-chimeric reads were kept. In the short-read pipeline, more two-exon genes than single-exon genes were predicted for all tissues, and it appears as if some transcripts could not be assembled entirely or StringTie2 tried to "avoid" single-exon genes. Transcript model reconstruction in StringTie2 is based on the concept of extending short reads to create so-called super-reads [47], which seems appropriate for whole-genome assemblies. In transcriptomics, however, multiple splice variants are possible, and with the super-read concept in mind, it may thus be possible that StringTie2 discards a single-exon transcript model in favour of an alternative multi-exon splice variant containing the same exon. This, in turn, might have a substantial impact on the functional annotation based on transcript models solely assembled with short reads. Real single-exon transcripts might be missed, or even worse, false-positive multi-exon transcript models might be reconstructed. We merged transcript models of both pipelines with higher priority on transcript end sites for the long-read-inferred transcript models to mitigate this effect. The average number of transcripts in the long-read pipeline often almost matched or even exceeded (lung) the average number of transcripts in the short-read pipeline. Furthermore, the long-read pipeline recovered more transcripts per gene (isoforms) on average. Taken together, this is remarkable considering the 60-fold higher sequencing depth in the short-read pipeline. It also corroborates the strength of PacBio Iso-Seq, which seems to better reflect the complexity of the transcriptome, considering that transcripts did not need to be assembled but were sequenced full-length.

While both the number and quality of published vertebrate genome assemblies are rising, hardly any are complemented by a transcriptome of multiple tissues from the same individual [25, 48]. Automated annotations (e.g., the NCBI Eukaryotic Genome Annotation Pipeline, [49]) provide reasonable in silico predictions of coding potential; however, RNA (cDNA) sequencing adds evidence for expressed genes. Based on the inferred transcripts in this study, a total of 14,099 protein-coding genes could be identified in the UniRef50 database after conservative filtering (at least 90% match), which is comparable to NCBI’s
Figure 4. In the short-read data set, the highest total number of supported genes was found in testes (left panel, bottom), followed by ovary, brain, spleen, ileum and lung. All six tissues intersected in 11,165 genes (main panel, left). The highest number of exclusively supported genes was also found in testes (988), and followed the same order as the total number of genes (main panel, yellow).

Table 5. Small RNA read processing and assembly stats.

| Tissue       | brain                  | ileum                  | lung                   | ovary                  | spleen                  | testis                  |
|--------------|------------------------|------------------------|------------------------|------------------------|-------------------------|-------------------------|
| Number of raw reads (PE) | 78,078,195             | 58,021,264             | 70,381,224             | 79,425,103             | 65,767,638              | 67,436,837              |
| Number of trimmed reads (PE) | 73,753,404             | 57,021,189             | 69,326,112             | 77,333,115             | 58,716,681              | 65,395,835              |
| Mapped uniquely [%] | 51.82                  | 88.04                  | 72.04                  | 71.01                  | 72.43                   | 52.59                   |
| Mapped multi [%] | 44.90                  | 9.48                   | 24.42                  | 26.59                  | 18.18                   | 31.97                   |
| Number of genes | 13,606                 | 8,441                  | 11,899                 | 9,903                  | 33,133                  | 31,205                  |
| Number of transcripts | 13,685                 | 8,520                  | 11,995                 | 9,954                  | 33,761                  | 31,342                  |
| Number of exons | 17,276                 | 12,650                 | 15,397                 | 11,588                 | 54,504                  | 35,026                  |

Table 6. Results of cmscan on assembled small RNA transcripts after filtering. Intersection refers to small RNAs predicted by the in silico genome scan. Additional refers to annotated small RNAs which were not detected by cmscan in the reference genome.

| Tissue       | brain                  | ileum                  | lung                   | ovary                  | spleen                  | testis                  |
|--------------|------------------------|------------------------|------------------------|------------------------|-------------------------|-------------------------|
| in vitro     | 328                    | 294                    | 317                    | 312                    | 345                     | 369                     |
| intersection | 310                    | 274                    | 295                    | 293                    | 315                     | 345                     |
| additional   | 18                     | 20                     | 22                     | 19                     | 30                      | 24                      |

in silico prediction of 15,578 protein–coding genes [50]. The number of identified protein–coding genes in tufted duck is also in line with the prediction in other bird species like mallard (16,816; [51]), chicken (17,477; [52]) or zebra finch (16,197; [53]). CPC2 predicted 84.2% of the potentially protein–coding genes found in UniRef50, which would serve as a conservative estimate of the protein–coding potential by just looking at the reference genome. However, beyond the 17,911 genes annotated by UniRef, the annotation contains an additional 27,787 genes with protein–coding potential according to the TAMA ORF/NMD prediction pipeline, with these being potential candidates for further analyses. Gene and transcript identification in non–model species relies on annotations of preferably closely related model organisms. However, protein–coding genes are mainly described by a single transcript and predominantly built on short–read and comparative data [25].

Besides a high–quality transcriptome for the tufted duck, this study also provides a tissue–specific expression atlas. In the short–read pipeline, there is a large drop in numbers of transcripts supported in all tissues to transcripts exclusively expressed in a single tissue or a few tissues. This distribution is much more balanced in the long–read pipeline and may indicate a better recovery of isoforms or a higher number of false–positive assembled transcripts in the short–read pipeline.

The number of 306 Rfam small RNA families predicted (with 246 annotated) for the tufted duck in this study is comparable to 352 families predicted in chicken (Gallus gallus, version 5) in Rfam 14.4 [54]. The peaks of transcript length at 20 bp and 50 bp are in the area of microRNA (miRNA), which are usually 18 – 23 nt [55, 56, 57], and pre–miRNA which are in the range of 55 – 70 nt. The substantial variance across tissues in our data (many more genes, transcripts and exons in spleen and testis than in the other tissues) might be explained by Illumina sequencing bias introduced at the adapter–ligation step of cDNA library constructions [58, 59]. Furthermore, according to Illumina’s TruSeq Small RNA library preparation protocol, small RNA populations can vary significantly between different tissue types and species, and types and coverage vary depending on which bands are selected during gel excision. This is additional support of a strategy to sequence multiple tissues to
Figure 5. In the long-read data set, the highest total number of supported genes was found in brain (left panel, top), followed by lung, ileum, testes, ovary and spleen. All six tissues intersected in 2,475 genes (main panel, left). The highest number of exclusively supported genes was found in brain (779), followed by testes, ileum, lung, ovary and spleen (main panel, blue).

Figure 6. Distribution of single-exon and multi-exon small RNA transcripts for each tissue.

obtain a fuller picture of small RNA expression in an organism. Gene duplications might explain the significant difference between unique mappings and multi mappings of small RNA to the genome across tissues. Gene duplications might explain the significant difference between unique mappings and multi mappings of small RNA to the genome across tissues. The in silico scan of small RNA in the genome could predict almost all small RNAs in the assembled transcripts. More importantly, however, 22.2 additional small RNAs were discovered on average in the in vitro scan which would otherwise have been unnoticed. Small (non-coding) RNA play an essential role in gene regulation, translation and chromosome structure [60, 61], and are often associated with diseases [62, 63]. This vital fraction of the genome is rarely validated in vitro in the genome and transcriptome annotation literature. However, small RNA studies have been continuously rising over the last 20 years from 1,966 publications in 1999 to 8,034 in 2019 (searching for "small RNA" on Web of Science [64]). Relying on in silico prediction of small RNA alone can lead to misinterpretation of pathways and gene regulation, and sequencing small RNA of non-model organisms is, therefore, an advancement for genome annotation [65]. We strongly encourage in vitro sequencing of small RNA in de novo genome and transcriptome studies, or otherwise, the scientific community will miss such detail that will be important to decipher relevant differences in the biology between species.

Potential implications

This study presents the first high-quality reference genome assembly of the non-model tufted duck species. It is complemented by coding and small non-coding RNA transcriptome annotation from six different tissues. The genome assembly contributes to the VGP’s ongoing mission to generate near-error-free and complete genome assemblies of all extant vertebrate species. By utilising, comparing and combining the strengths of low error rates and high sequencing depth in Illumina RNA sequencing, and the full-length transcript sequencing in PacBio’s Iso-Seq, this annotation culminates in a merged transcriptome with functional annotation and an expression atlas. Evidence from small RNA of the same tissues sequenced using the Illumina platform revealed small RNAs that would have otherwise remained undetected. Our findings from a comparison between short-read and long-read reference transcriptomics contribute to a deeper understanding of these competing options. In this study, both technologies complemented each other. While short-read data was sufficient to annotate protein-coding genes, long-read data recovered more transcripts per gene and potentially further protein-coding genes that could not be annotated. With the ongoing improvement of base call quality in long-read sequencing, short-read transcriptome sequencing might become expendable. Together, the genome and transcriptome annotation of the tufted duck is
an excellent resource for public omics databases and a foundation for downstream studies, e.g., regarding disease response. The dataset’s high quality for a non-model species allows for a much finer resolution of genetic differences and commonalities in closely related species, which is crucial while studying the reservoirs of zoonotic pathogens.

Methods

Sampling and dissection of tissues. Captive-bred, wild tufted ducks were kept at the animal breeding facility (Swedish National Veterinary Institute, Uppsala, Sweden). The ducks were obtained from Snavelhof breeding farm, Veeningen, the Netherlands, in May 2017. Tissue samples were obtained from five females and five males (12 months old) after euthanasia with an injection of 1 mL of pentobarbital (100 mg) in the wing vein. The following tissues were collected from the birds: brain, ileum, spleen, lung and gonads (ovary or testis). Tissues were immediately snap-frozen in liquid nitrogen and stored at −80°C until shipment on dry ice to the Roslin Institute, Edinburgh, UK. All animal experiments were carried out in strict accordance with a protocol legally approved by the regional board of the Uppsala animal ethics committee, Sweden (permission number 5.8.18−07998/2017). The animal experiments were conducted in biosafety level two (BSL-2) animal facilities at the Swedish National Veterinary Institute.

Genomic DNA: library preparation, sequencing and assembly. To obtain both sex chromosomes, DNA was extracted from lung tissue of a female tufted duck. Library preparation and sequencing was conducted as in [21], using four types of sequencing data and the Vertebrate Genomes Project (VGP) assembly pipeline 1.6 (all details given in Rhiie et al. [21]). The sequence data consisted of PacBio CLR reads, 10X Genomics chromosome linked-reads, Bionano Genomics optical maps created by direct labelling and staining (DLS), and chromatin conformation capture coupled with high throughput sequencing (Hi-C) (Arima Genomics, San Diego, CA, USA). In brief, PacBio reads were input to the diploid−aware long-read assembler FALCON and its haplotype-resolving tool FALCON-Unzip [66]. The resulting primary contigs were input to the Purge-Dups pipeline [67] to identify and remove remaining haplotigs in the primary set. In the next step, primary−purged contigs were subject to two rounds of scaffolding using the toX long molecule linked−reads. Further, pre-assembled DLS Bionano cnaps were applied for further scaffolding and ordering using the Solve pipeline (Bionano Genomics, San Diego, CA, USA). The resulting scaffolds were then further scaffolded into chromosome-scale scaffolds using Hi−C data and the Salsa2 pipeline [68]. Finally, the primary assembly plus the Falcon−phased haplotigs were concatenated for three rounds of base call polishing: first with PacBio reads and Arrow software [69], and subsequently two rounds of polishing with 10X linked−reads and FreeBayes software [70]. The genome was decontaminated and manually curated as described in Howe et al. [71].

Tissue preparation, DNA and RNA extraction. For disruption and homogenisation of tissues, snap−frozen samples were ground to a fine powder under liquid nitrogen using a mortar and pestle. Samples were transferred to 1.5 mL frozen tubes and kept on dry ice until further processing. Total RNA was obtained following a standard TRIzol protocol with DNase treatment and column purification. Small RNA was prepared according to the miRNeasy kit protocol 217004 (Qiagen, Venlo, Netherlands). Integrity and quality of the RNA were confirmed by electrophoresis on an Agilent 2200 Tapestation using appropriate screen tapes. The concentration was determined using the Nanodrop ND−1000 (Thermo Fisher Scientific, Waltham, MA, USA) (Tab. S1, Tab S2). For DNA extraction and sequencing, powdered lung tissue was sent to the Vertebrate Genomes Lab (VGL) at Rockefeller University, New York, USA.

Illumina cDNA library preparation and sequencing. RNA was sent to Edinburgh Genomics, Edinburgh, UK, for library preparation and sequencing on an Illumina Hiseq 4000 platform with 2 x 150 bp paired−end reads using the TrueSeq library preparation protocol (stranded). Median insert size was 137 bp − 148 bp (SD: 67 bp − 81 bp), yielding at least 290 M + 290 M reads per sample. Small RNA was also sent to Edinburgh Genomics for TrueSeq small RNA library preparation and sequencing using a NovaSeq 6000 platform with 2 x 50 bp paired−end reads. Median insert size was 141 bp − 144 bp yielding at least 225 M + 225 M reads per sample.

PacBio cDNA library preparation and sequencing. Two micrograms of total RNA from each sample in four parallel reactions were converted to cDNA using the Teloprimse full−length cDNA amplification kit (013, vt) according to the manufacturer’s instructions (Lexogen, Vienna, Austria). After end−point PCRs, all samples were tested for quality and quantity. The product size distribution was visualised using an Agilent 2200 Tapestation using D5000 screen tape. The library concentration was measured on a Qubit 3 (Thermo Fisher Scientific, Waltham, MA, USA) with high−sensitivity DNA reagents (Tab. S3). Technical replicates were pooled and selected for PacBio Iso−Seq assays. The samples were sequenced at Edinburgh Genomics using Sequel (version 2.1) chemistry.

Gene/transcript annotation. Illumina raw RNA−Seq reads were quality checked and filtered using FastQC (0.11.8) [72] and Trimmomatic (0.38) [73], respectively. Corrected reads were mapped to the genome using HISAT2 (2.2.0) [74, 75] and transcript models assembled using StringTie2 (2.1.1) [47]. The resulting annotation file was converted with tama_format_gtf_to_bed12_stringtie.pl. In the remainder of this manuscript, all tools described as starting with "tama_" are part of the software suite Transcriptome Annotation by Modular Algorithms [76], except for tama_merge_report_parser.pl [77]. PacBio raw Iso−Seq reads were pre−processed using the IsoSeq3 pipeline to obtain full−length, non−chimeric reads (FLNC; first three steps in [78]; ccs 3.0, lima 1.8.0, refine 3.1.0). Afterwards, fasta sequences were extracted from bam files using Bambus 2.5.1 [79] and poly−A tails were trimmed using tama_flnc_polya_cleanup.py (20191022). These FLNC were mapped to the reference genome with the splice site−aware mapper Minimap2 2.17−r974−dirty [80]. Redundant transcript models were collapsed with the capped flag (−x capped) using tama Collapse.py when coverage was at least 95% (−c 95). Additionally, 5’t threshold and 3’ threshold (tolerance in bp for grouping reads) were set to 100 (−a 100 −z 100).

Functional annotation. Transcriptomes from all six tissues inferred by the short−read and long−read pipelines were merged based on similarity using tama_merge.py (options −a 100 −z 100 −d merge_dup) with different priorities for splice junctions and transcript end sites. Short−read inferred transcript models were given higher priority on splice junctions, whereas long−read inferred transcript models were given higher priority on transcript end sites. Nucleotide sequences based on coordinates in the merged transcriptome were extracted from the reference genome using Bedtools (2.29.0) [81]. The protein−coding potential was predicted with CPC2 (0.1) [82] based on the transcripts’ sequence features. Open reading frames (ORFs) were predicted and translated into amino acid sequences us-
small RNA analyses. Illumina raw reads were quality checked with FastQC (0.11.8) [72] and adapters removed with Cutadapt (2.10) [87]. Corrected reads were mapped to the reference genome using STAR (2.7.3a) [88] and assembled with Cufflinks (2.2.1) [89, 90, 91, 92]. Nucleotide sequences were extracted from the reference genome at the assembled transcripts’ coordinates using Bedtools getfasta –split (2.29.2), and transcript lengths extracted from the output of Samtools faidx [93]. All plots were created using the package ggplot2 [94] in RStudio [86, 95].

The tool cmscan (1.1.3) from the software suite Infernal [96] was used to predict structural RNAs in the reference genome (in silico) and to annotate assembled small RNA transcripts (in vitro) based on Rfam [97, 98] covariance models (CM) downloaded from ftp://ftp.ebi.ac.uk/pub/databases/Rfam/14.2. The output of cmscan (tblout) was converted to gff3 annotation files using tblout2gff3.pl [77]. Shared intervals between in silico and in vitro RNA annotations were identified with the intersect option of Bedtools (2.29.2) [81].

Availability of source code and requirements
Perl and R scripts used in this study are available on GitLab at https://gitlab.com/rcmueller/ta2ed_duck_annotation under MIT license.

Availability of supporting data and materials
The data sets supporting the results of this article are available in NCBI and Figshare. The curated assembly of the tufted duck genome has been deposited in NCBI under accession number GCF_000819795.1 [41]. The Illumina RNA-Seq and small RNA-Seq, and PacBio Iso-Seq raw reads have been deposited in NCBI under accession numbers from-to (uploaded but not set to public yet). Supporting data has been deposited on Figshare: Illumina RNA-Seq transcript models [99], PacBio Iso-Seq transcript models [100], functional annotation [101], expression atlases [102], illumina small RNA-Seq annotation [103], additional scripts [77]. (The links in these references are temporary and will get a DOI once set to public)

Declarations

List of abbreviations
AIV – Avian influenza A virus | BSL – Biosafety level | CLR – Continuous long reads | DLS – Direct label and stain | FLNC – Full-length, non-chimeric reads | HPAI – Highly pathogenic avian influenza | NGS – Next-generation sequencing | ORF – Open reading frame | SMRT – Single-molecule, real-time | TAMA – Transcriptome annotation by modular algorithms | VGL – Vertebrate Genomes Lab | VGP – Vertebrate Genomes Project | ZMW – Zero-mode waveguide

Ethical Approval (optional)
All animal experiments were carried out in strict accordance with a protocol legally approved by the regional board of the animal ethics committee, Sweden (permission number 5.8.18-07998/2017). The animal experiments were conducted in biosafety level two (BSL–2) animal facilities at the Swedish National Veterinary Institute (Uppsala, Sweden).

Consent for publication
Not applicable

Competing Interests
The authors declare that they have no competing interests.

Funding
The authors gratefully acknowledge funding of a Short-Term Scientific Mission (STSM) through COST Action CA15112 – Functional Annotation of Animal Genomes – European network (FAANG-Europe), and funding from the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) during the Science Data Centre project BioDATEN.

Author’s Contributions
RCM: Conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, visualisation, writing—original draft, writing—review & editing | PE: Resources, writing—original draft | KH: Formal analysis, methodology, resources, validation, visualisation, writing—original draft | MUS: Formal analysis, methodology, validation, writing—original draft | RHSK: Conceptualisation, methodology, software | KM: Methodology, writing—original draft | AW: Resources | OF: Formal analysis | BH: Formal analysis | JM: Formal analysis | JC: Formal analysis | JW: Formal analysis | JD: Resources | MMN: Resources | BO: Funding acquisition | EDJ: Project administration, resources | JS: Project administration, resources, supervision | LE: Formal analysis, investigation, visualisation | RHSK: Conceptualisation, Funding acquisition, project administration, resources, supervision
All authors read and approved the final manuscript.

Acknowledgements
The authors acknowledge support by the local HPC resources through the core facility SCCKN at the University of Konstanz. The authors also acknowledge continued support by the Max Planck Institute of Animal Behavior and its director Martin Wikelski.

List of abbreviations
AIV – Avian influenza A virus | BSL – Biosafety level | CLR – Continuous long reads | DLS – Direct label and stain | FLNC – Full-length, non-chimeric reads | HPAI – Highly pathogenic avian influenza | NGS – Next-generation sequencing | ORF – Open reading frame | SMRT – Single-molecule, real-time | TAMA – Transcriptome annotation by modular algorithms | VGL – Vertebrate Genomes Lab | VGP – Vertebrate Genomes Project | ZMW – Zero-mode waveguide
References

1. Atkinson PW, Clark JA, Delany S, Diagana CH, du Feu C, Fiedler W, et al. Urgent Preliminary Assessment of Ornithological Data Relevant to the Spread of Avian Influenza in Europe. Wageningen: Wetlands International; 2006.

2. Huang ZXY, Xu C, van Langevelde F, Ma Y, Lancaster T, Mundkur T, et al. Contrasting Effects of Host Species and Phylogenetic Diversity on the Occurrence of HPAI H5N1 in European Wild Birds. Journal of Animal Ecology 2019;88(7):1044–1053. https://beijounes.library.wiley.com/doi/abs/10.1111/1365-2666.12997.

3. Kraus RHS. The Role of Mallard (Anas Platyrhynchos) in the Spread of Avian Influenza: Genomics, Population Genetics, and Flyways. PhD thesis; 2011.

4. Olsen B, Munster VJ, Wallensten A, Waldenström J, Osterhaus ADME, Fouchier RAM. Global Patterns of Influenza A Virus in Wild Birds. Science 2006 Apr;312(5772):384–388. http://science.sciencemag.org/content/312/5772/384.

5. Webster RG, Yakhno M, Hinshaw VS, Bean WJ, Cox NJ. Pathogenic Avian Influenza A Virus H5N1 of Asian Lineage in Wild Birds in Germany, 2006 and 2007. Transboundary and Emerging Diseases 2009;56(3):57–72. https://www-library.wiley.com/doi/10.1111/j.1865-1682.2008.01061.x.

6. Abdo W, Haridy M, Katou Y, Goto M, Mizoguchi T, Sakoda Y, et al. Pathological and Immunohistochemical Findings of Natural Highly Pathogenic Avian Influenza Infection in Tufted Ducks during 2010–2011 Outbreaks in Japan. Journal of Veterinary Diagnostic Investigation 2014;70(9):1285–1290.

7. Bröjer C, van Amerongen G, van de Bildt M, van Run P, Osterhaus A, Gavier-Widén D, et al. Pathogenicity and Tissue Tropism of Currently Circulating Highly Pathogenic Avian Influenza A Virus (H5N1; Clade 2.3.2) in Tufted Ducks (Aythya Fuligula). Veterinary Microbiology 2015 Nov;180(3):268–278. http://www.sciencedirect.com/science/article/pii/S0378113515300298.

8. Bröjer C, Ägren EO, Uhlhorn H, Bermot K, Mörner T, Désirée et al. Pathology of Natural Highly Pathogenic Avian Influenza H5N1 Infection in Wild Tufted Ducks (Aythya Fuligula). Journal of Veterinary Diagnostic Investigation 2009 Sep;21(3):579–587. https://doi.org/10.1177/10406370902100501.

9. Fiedler W, Bauer HG, “Massive” Outbreak of High Pathogenic Avian Influenza among Wild Ducks at Lake Constance in Autumn 2016. Isle of Great Cumbrae, Scotland; 2018. https://ducksg.org/events/peds5/.

10. Kraus RH, Kerstens HH, van Hooft P, Megens HJ, Elmendorf T, Mundkur T, et al. Contrasting Effects of Host Species and Phylogenetic Diversity on the Occurrence of HPAI H5N1 in European Wild Birds. Journal of Animal Ecology 2019;88(7):1044–1053. https://beijounes.library.wiley.com/doi/abs/10.1111/1365-2666.12997.

11. Kraus RHS. The Role of Mallard (Anas Platyrhynchos) in the Spread of Avian Influenza: Genomics, Population Genetics, and Flyways. PhD thesis; 2011.

12. Evseev D, Magor KE. Innate Immune Responses to Avian Influenza Virus. Dev Genomics. Heredity 2011 Jul;107(1):1–15. https://www-ncbi-nlm-nih-gov/pmc/articles/PMC31088493/.

13. Globig A, Staabach C, Beer M, Köppen U, Fiedler W, Nieburg M, et al. Epidemiological and Ornithological Aspects of Outbreaks of Highly Pathogenic Avian Influenza Virus H5N1 of Asian Lineage in Wild Birds in Germany, 2006 and 2007. Transboundary and Emerging Diseases 2009;56(3):57–72. https://www-library.wiley.com/doi/10.1111/j.1865-1682.2008.01061.x.

14. van Doorn HR. Emerging Infectious Diseases. Medicine 2017 Dec;45(12):798–801. http://www.sciencedirect.com/science/article/pii/S1578319117302426.

15. Elblom R, Galindo J. Applications of next Generation Sequencing in Molecular Ecology of Non-Model Organisms. Heredity 2011 Jul;107(1):1–15. https://www-ncbi-nlm-nih-gov/pmc/articles/PMC39730152/.

16. Rahi A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Korren S, et al. Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species. bioRxiv 2020 May;p. 2020.05.22.110833. https://www-biorxiv-org/content/10.1101/2020.05.22.110833v1.

17. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized Long Read RNA Sequencing in Chicken Normal Long Read RNA Sequencing. Genome Biology 2013;14:405. http://dx.doi.org/10.1186/gb-2013-14-6-405.

18. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized Long Read RNA Sequencing in Chicken Normal Long Read RNA Sequencing. Genome Biology 2013;14:405. http://dx.doi.org/10.1186/gb-2013-14-6-405.

19. Kraus RH. Avian Transcriptomics: From the Lab into the Wild Cham: Springer International Publishing; 2019.

20. Kraus RH. Avian Transcriptomics: Opportunities and Challenges. Journal of Ornithology 2018 Jul;159(3):599–629. https://doi.org/10.1007/s10336-018-1532-5.

21. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized Long Read RNA Sequencing in Chicken Normal Long Read RNA Sequencing. Genome Biology 2013;14:405. http://dx.doi.org/10.1186/gb-2013-14-6-405.

22. Roberts RJ, Carneiro MO, Schatz MC. The Advantages of SMRT Sequencing. Genome Biology 2013;14:405. http://dx.doi.org/10.1186/gb-2013-14-6-405.

23. Warren WC, Hillier LW, Tomlinson C, Minx P, Krimtzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3: Genes, Genomes, Genetics 2017 Jan;7(1):109–117. https://www-g3journal-org/content/7/1/109.
29. Steijger T, Abril JF, Engström PG, Kokocinski F, The RGASP Consortium, Hubbard TJ, et al. Assessment of Transcript Reconstruction Methods for RNA-Seq. Nature Methods 2013 Dec;10(12):1177–1184. http://www.nature.com/nmeth/journal/v10/n12/abs/nmeth.2714.html.

30. Au RF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, et al. Characterization of the Human ESC Transcriptome by Hybrid Sequencing. Proceedings of the National Academy of Sciences of the United States of America 2013 Dec;110(50):E4821–E4830. http://resolver.caltech.edu/CaltechAUTHORS:20141013-101222866.

31. Sharon D, Tiligner H, Grubert F, Snyder M. A Single-Molecule Long-Read Survey of the Human Transcriptome. Nature Biotechnology 2013 Nov;31(10):1009–1014. http://www.nature.com/nbt/journal/v31/n11/abs/nbt.2705.html.

32. Li Y, Dal C, Hu C, Liu Z, Kang C. Global Identification of Alternating Splicing via Comparative Analysis of SMRT- and Illumina-Based RNA-Seq in Strawberry. The Plant Journal 2017 Apr;90(1):164–176. http://onlinelibrary.wiley.com/doi/10.1111/tpj.13462/abstract.

33. Reixachs-Solé M, Ruiz-Orera J, Albà MM, Eyras E. Ribosome Profiling at Isoform Level Reveals an Evolutionary Conserved Impact of Differential Splicing on the Proteome. bioRxiv 2019 Oct;p. 582031. https://www.biorxiv.org/content/10.1101/582031v3.

34. Zhang H, Jain C, Aluru S. A Comprehensive Evaluation of Long Read Error Correction Methods. bioRxiv 2019 May;p. 519330. https://www.biorxiv.org/content/10.1101/519330v2.

35. Chapman JR, Hellgren O, Helin AS, Kraus RHS, Cromie RL, Wäldenström J. The Evolution of Innate Immune Genes: Purifying and Balancing Selection on α-Defensins in Waterfowl. Molecular Biology and Evolution 2016 Jan;33(12):3075–3087. http://mbe.oxfordjournals.org/content/33/12/3075.

36. Karawita A, Cheng Y, Tong M, Mueller R, Bielefeldt-Ohmann H, Chew K, et al. Comparative Genomics and Transcriptomics Help Unravel Why Australian Black Swans Are Uniquely Susceptible to Highly Pathogenic Avian Influenza (HPAI)? 2020. .

37. Mueller RC, Mallig N, Smith J, Eöery L, Kuo RI, Kraus RHS. Purification and Transcriptome Profiling at Isoform Level Reveals an Evolutional 2017 Apr;90(1):164–176. https://www.ncbi.nlm.nih.gov/pubmed/28489064.

38. Naguib MM, Eriksson P, Tilgner H, Grubert F, Snyder M. A Single-Molecule Long-Read Survey of the Human Transcriptome. Nature Methods 2013 Dec;10(12):1177–1184. http://www.nature.com/nmeth/journal/v10/n12/abs/nmeth.2714.html.

39. Hammar B. The Karyotypes of Nine Birds. Hereditas 1966;55(2-3):367–385. http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1601-0722.1966.tb00266.x.

40. GenomeArk, Aythya fuligula; 2019. https://www.ncbi.nlm.nih.gov/assembly/GCF_008822105.2.

41. GVGP, bTaeGut2.Pat.W.v2 - Genome - Assembly - NCBI; 2020. https://www.ncbi.nlm.nih.gov/assembly/GCF_008822105.2.

42. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome Assembly from Long-Read RNA-Seq Alignments with StringTie2. bioRxiv 2019 Jul;p. 694554. https://www.biorxiv.org/content/10.1101/694554v1.

43. Yin Z, Zhang F, Smith J, Kuo R, Hou ZC. Full-Length Transcriptome Sequencing from Multiple Tissues of Duck, Anas Platyrynchos. Scientific Data 2019 Nov;6(1):1–9. https://www.nature.com/articles/s41597-019-0293-1.

44. NCBI, The NCBI Eukaryotic Genome Annotation Pipeline; 2018. https://www.ncbi.nlm.nih.gov/assembly/annotation_euk/process/.

45. NCBI, Aythya fuligula Annotation Report; 2020. https://www.ncbi.nlm.nih.gov/assembly/annotation_euk/Aythya_fuligula/100/TranscriptAlignmentStats.

46. NCBI, Anas Platyrynchos Annotation Report; 2020. https://www.ncbi.nlm.nih.gov/assembly/annotation_euk/Anas_platyrynchos/104/.

47. NCBI, Gallus gallus Annotation Report; 2018. https://www.ncbi.nlm.nih.gov/assembly/annotation_euk/Gallus_gallus/104/.

48. NCBI, Taeniopygia guttata Annotation Report; 2020. https://www.ncbi.nlm.nih.gov/assembly/annotation_euk/Taeniopygia_guttata/105/.

49. Kalvari I, Nawrocki EP, Onitíeros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: Expanded Coverage of Metagenomic, Viral and microRNA Families. Nucleic Acids Research 2021 Jan;p. 192–200. https://doi.org/10.1093/nar/gkaa1047.

50. Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell 2004 Jan;116(2):281–297. http://www.sciencedirect.com/science/article/pii/S009286740000456X.

51. Lee RC, Feinbaum RL, Ambros V. The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14. 1993 Dec;75(5):834–854. doi:10.1016/0027-5120(93)90083-4.

52. LGT II, LGTII, LGT-2, LGT-2 - Genome - Assembly - NCBI; 2020. https://www.ncbi.nlm.nih.gov/assembly/GCF_008822105.2.

53. NCBI, Taeniopygia guttata Annotation Report; 2020. https://www.ncbi.nlm.nih.gov/assembly/annotation_euk/Taeniopygia_guttata/105/.

54. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Sander C, Marks D, Snyder M, et al. Small Non-Coding RNA Transcription of Four High-Altitude Vertebrates and Their Low-Altitude Relatives. Scientific Reports 2019 Oct;9(1):192. https://www.nature.com/articles/s41598-019-0204-9.

55. Baroin–Tourancheau A, Jaszczyszn Y, Benigni X, Amar L. Evaluating and Correcting Inherent Bias of microRNA Expression in Illumina Sequencing Analysis. Frontiers in Molecular Biosciences 2019;6. https://www.frontiersin.org/articles/10.3389/fmolb.2019.00017/full.

56. Fuchs RT, Sun Z, Zhuang F, Robb GB. Bias in Ligation-Based Small RNA Sequencing Library Construction Is Determined by Adaptor and RNA Structure. PLOS ONE 2015 May;10(5):e0126049. https://www.ncbi.nlm.nih.gov/assembly/GCF_008822105.2.

57. NCBI, Anas platyrhynchos Annotation Report; 2020. https://www.ncbi.nlm.nih.gov/assembly/annotation_euk/Anas_platyrhynchos/104/.

58. NCBI, Gallus gallus Annotation Report; 2018. https://www.ncbi.nlm.nih.gov/assembly/annotation_euk/Gallus_gallus/104/.

59. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Sander C, Marks D, Snyder M, et al. Small Non-Coding RNA Transcription of Four High-Altitude Vertebrates and Their Low-Altitude Relatives. Scientific Reports 2019 Oct;9(1):192. https://www.nature.com/articles/s41598-019-0204-9.

60. Baroin–Tourancheau A, Jaszczyszn Y, Benigni X, Amar L. Evaluating and Correcting Inherent Bias of microRNA Expression in Illumina Sequencing Analysis. Frontiers in Molecular Biosciences 2019;6. https://www.frontiersin.org/articles/10.3389/fmolb.2019.00017/full.

61. Fuchs RT, Sun Z, Zhuang F, Robb GB. Bias in Ligation-Based Small RNA Sequencing Library Construction Is Determined by Adaptor and RNA Structure. PLOS ONE 2015 May;10(5):e0126049. https://www.ncbi.nlm.nih.gov/assembly/GCF_008822105.2.
ronment for R. Boston, MA: RStudio, Inc.; 2016. http://www.rstudio.com/.
96. Nawrocki EP, Eddy SR. Infernal 1.1: 100-Fold Faster RNA Homology Searches. Bioinformatics 2013 Nov;29(22):2933–2935. https://academic.oup.com/bioinformatics/article/29/22/2933/318439.
97. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, et al. Non-Coding RNA Analysis Using the Rfam Database. Current protocols in bioinformatics 2018 Jun;62(1):e51. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6754622/.
98. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: Shifting to a Genome-Centric Resource for Non-Coding RNA Families. Nucleic Acids Research 2018 Jan;46(D1):D335–D342. https://academic.oup.com/nar/article/46/D1/D335/4588106.
99. Mueller RC, Illumina RNA-Seq, Transcript Models; 2021. https://figshare.com/s/c77434c60056b4ab7389.
100. Mueller RC, PacBio Iso-Seq, Transcript Models; 2021. https://figshare.com/s/a3f0b13080b9beaf2c21.
101. Mueller RC, Functional Annotation, Illumina and PacBio Transcript Models; 2021. https://figshare.com/s/01044d0b063b31fd6c.
102. Mueller RC, Expression Atlas; 2021. https://figshare.com/s/3e737f88d36afa738a2ca.
103. Mueller RC, Illumina Small RNA-Seq; 2021. https://figshare.com/s/2c5eab450186103ad32.
Exons per transcript

- **brain**
- **ileum**
- **lung**
- **ovary**
- **spleen**
- **testis**
