Received: 10 January 2024 | Revised: 19 June 2024 | Accepted: 24 June 2024

DOI: 10.1002/ece3.11694

Ecology and Evolution Open Access WILEY

GENETICS NOTES

De-novo assembly of four rail (Aves: Rallidae) genomes: A resource for comparative genomics

Julien Gaspar[1,2] | Steve A. Trewick[1] | Gillian C. Gibb[1]

[1]School of Food Technology and Natural Sciences, Wildlife and Ecology Group, Massey University, Palmerston North, New Zealand

[2]Royal Belgian Institute of Natural Sciences, Brussels, Belgium

Correspondence
Julien Gaspar and Gillian C. Gibb, School of Food Technology and Natural Sciences, Wildlife and Ecology Group, Massey University, Private Bag 11-222, Palmerston North, New Zealand.
Email: julien.gaspar93@gmail.com and g.c.gibb@massey.ac.nz

Funding information
Marsden Fund, Grant/Award Number: MAU1601

Abstract

Rails are a phenotypically diverse family of birds that includes 130 species and displays a wide distribution around the world. Here we present annotated genome assemblies for four rails from Aotearoa New Zealand: two native volant species, pūkeko *Porphyrio melanotus* and mioweka *Gallirallus philippensis*, and two endemic flightless species takahē *Porphyrio hochstetteri* and weka *Gallirallus australis*. Using the sequence read data, heterozygosity was found to be lowest in the endemic flightless species and this probably reflects their relatively small populations. The quality checks and comparison with other rallid genomes showed that the new assemblies were of good quality. This study significantly increases the number of available rallid genomes and will enable future genomic studies on the evolution of this family.

KEYWORDS
Gallirallus australis, *Gallirallus philippensis*, genome assemblies, heterozygosity, *Porphyrio hochstetteri*, *Porphyrio melanotus*, rails, Rallidae

TAXONOMY CLASSIFICATION
Genomics

1 | INTRODUCTION

Rails (Aves: Rallidae) are a phenotypically diverse family of primarily terrestrial birds with relatively short wings and strong, variably elongated bills (Livezey, 2003; Ripley et al., 1977; Taylor, 1998). Despite the terrestrial lifestyle of the majority of the species (Taylor, 1998), this bird family displays remarkable dispersal capacity resulting in broad distribution and the colonisation of numerous oceanic islands (Garcia-R et al., 2017; Olson, 1973; Ripley et al., 1977). At the same time, more than 30 flightless rail species are known (Kirchman, 2012; Steadman, 1995) and a large proportion of them are endemic to single oceanic islands, demonstrating that their ancestors had been volant (Trewick, 1997a, 1997b). The high proportion of flightless species as well as the fact that flightlessness evolved many times among extant rails provides a suitable system with which to study genomic changes associated with maintenance and loss of flight in birds.

Rallidae has its origin during the Eocene around 40 million years ago (Garcia-R et al., 2014a, 2014b) and has diversified into over 130 extant species (Garcia-R et al., 2014a, 2014b; Kirchman, 2012; Steadman, 1995). Rails are part of the order Gruiformes that includes two suborders; the Gruoidea containing, among others, the cranes (family Gruidae), and the Ralloidea that is dominated by the rails (family Rallidae) (Boast et al., 2019; Fain et al., 2007). The rails are further divided into around 40 genera in 9 tribes (Kirchman et al., 2021).

Despite their phylogenetic diversity (Figure 1), flightless rails typically exhibit smaller sterna and wings than volant taxa along with wider pelves and more robust femora (Gaspar et al., 2020; Livezey, 2003). Moreover, it has been shown that these differences are independent of phylogeny and instead demonstrate convergent evolution associated with a walking ecology (Gaspar et al., 2020). Despite some research using short markers at the population level (Garcia-R et al., 2017;

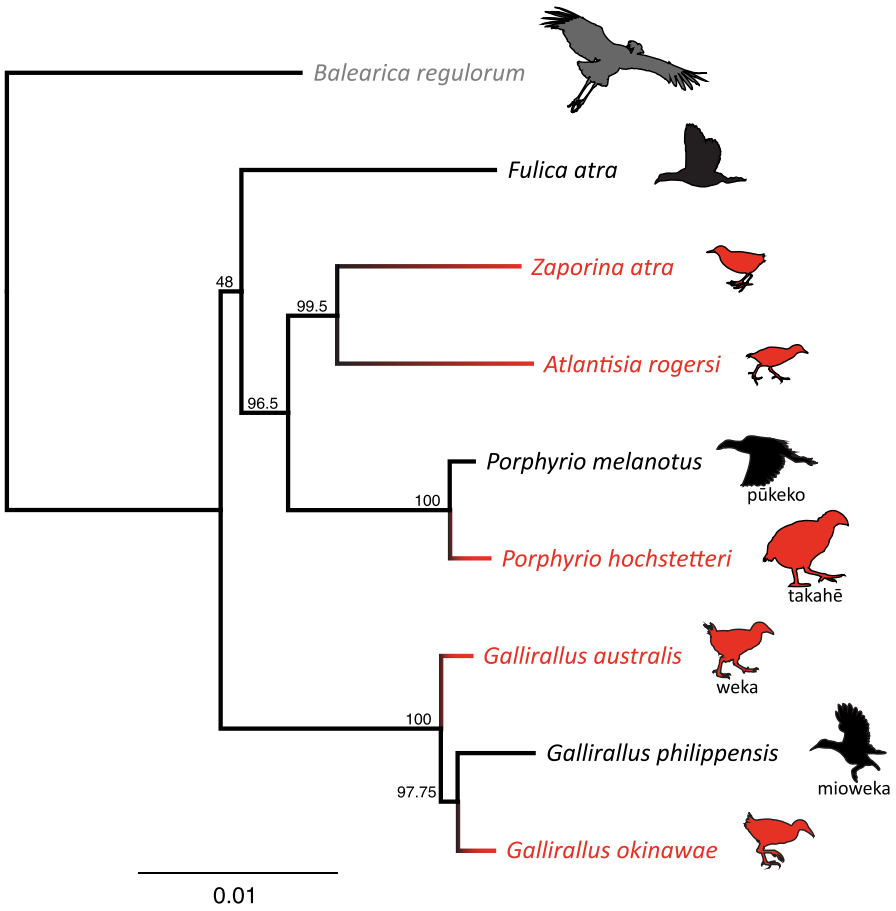

FIGURE 1 Maximum likelihood (RAxML V.8) phylogeny of three volant (black) and five flightless (red) rail lineages (Aves: Rallidae) based on 10 concatenated nuclear genes analysed with *Balearica regulorum* crane (Aves: Gruidae) (grey) as outgroup; bootstrap supports are indicated for each node.

Garcia-R & Trewick, 2014; Trewick et al., 2017), the molecular basis underlying the convergent evolution of flightless rails remains unknown. To investigate that question, more genomic data are needed. Here we present new, annotated rail genome assemblies of four rail species from Aotearoa New Zealand; two volant, purple swamphen (called pūkeko in Aotearoa New Zealand) *Porphyrio melanotus melanotus* (Temminck, 1820), and buff-banded rail (also called mioweka and moho pererū) *Gallirallus philippensis assimilis* (Linnaeus, 1766), and two flightless species, takahē *Porphyrio hochstetteri* (Meyer, 1883), and weka *Gallirallus australis australis* (Sparrman, 1786) (Clements et al., 2023). These four genome assemblies were generated to provide two volant-flightless pairs of closely related living species that will enable future genomic comparisons to highlight the differences and similarities in evolutionary trends between rails with and without the ability to fly (Figure 1).

2 | METHODS

2.1 | DNA extraction and sequencing

DNA was extracted from muscle tissue samples of four rails sampled in Aotearoa New Zealand: *Porphyrio melanotus*, *Gallirallus philippensis*, *Porphyrio hochstetteri*, and *Gallirallus australis*. Permission to obtain roadkill specimens is given by Department of Conservation Authority WA-17590-DOA. Full sample details can be found in Table 1. Extraction used the Geneaid© Genomic DNA Mini Kit following the kit instructions and eluted in 100 μL. DNA quality was then verified by gel electrophoresis and quantified using Qubit 2.0. Library preparation using the TruSeq Nano DNA kit and quality check were performed by the Massey University Genome Service (New Zealand) with sequencing by Novagene (Hong Kong). Libraries were sequenced on the Illumina HiSeq™ X platform generating non-overlapping 150 bp paired-end reads with an insert size of 550 bp. Fastp V0.19.4 (Chen et al., 2018) was used with default settings for paired-end data to trim the adapters as well as filter and assess the read quality.

2.2 | Genome assembly

De novo assembly was performed for each of the genomes using Meraculous (Chapman et al., 2011). Average insert size, standard deviation, and average read lengths were estimated using sequence reads mapped to a nuclear gene of a close species. Following the Meraculous manual instructions, a range of k-mer sizes were analysed using KmerGenie V1.7051 (Chikhi & Medvedev, 2014). The k-mer frequency histograms were reviewed and we selected k that had a main haploid peak with at least 30× coverage and a distinct trough to its left that was at most 1/10 of the peak height. These were 61, 87, 61, and 57 for respectively *Porphyrio melanotus*, *Porphyrio hochstetteri*, *Gallirallus philippensis*, and *Gallirallus australis* (Figure 2). High heterozygosity for *G. philippensis* meant that ideal peak height/trough specs could not be met but the assembly was still successful. See supplementary data for full details of settings used in all Meraculous runs.

TABLE 1 Sampling information for four New Zealand rails including the location, date of collection and Massey University museum ID (when applicable).

Species	Sampling	Sex	ID
Porphyrio melanotus melanotus	Roadkill, Turitea Valley near Palmerston North, North Island, New Zealand, within the rohe (area) of Rangitāne o Manawatū. October 2018	Male	MUNZ12900
Porphyrio hochstetteri	Provided by the Department of Conservation via Massey University Veterinary Pathology. A translocated individual on Maud Island, Marlborough Sounds, New Zealand	Male	NA
Gallirallus philippensis assimilis	Roadkill, Whananāki estuary, Northland, North Island, New Zealand, within the rohe of Ngatiwai. Retrieved March 2011	Male	MUNZ12901
Gallirallus australis australis	Roadkill Granity, West Coast, South Island, New Zealand, within the rohe of Ngāi Tahu. Retrieved July 2012	Male	MUNZ12767

Note: Taxonomy follows Clements et al. (2023).

Meraculous (Chapman et al., 2011) was implemented using a docker container we created, which is publicly available at both Github and docker (https://github.com/GenomicsForAotearoaNewZealand/genomics-tools, https://hub.docker.com/r/gfanz/meraculous). The assembly was run through the Catalyst Cloud server (https://catalystcloud.nz) using a cloud instance with 32 vCPU and 256 GB RAM.

2.3 | Additional genomes

In order to assess the quality of our genome assemblies, we compared them to a selection of additional rail genomes, Okinawa rail *Gallirallus okinawae* (*also known as Hypotaenidia okinawae*), GenBank assembly accession: GCA_027925045.1, Henderson crake *Zapornia atra* (formerly *Porzana atra*) GCA_013400835.1, Eurasian coot *Fulica atra* GCA_013372525.1, Inaccessible Island rail *Atlantisia rogersi* GCA_013401215.1, and takahē *Porphyrio hochstetteri* GCA_020800305.1. The chromosome-level takahē genome assembly was released while this work was in preparation and is included in the comparative analysis. The genome of a grey crowned crane *Balearica regulorum* (order Gruiformes, family Gruidae; Bennett, 1834) GCA_011004875.1 was used as a reference for the gene annotations.

2.4 | Quality assessment

Meraculous outputs were used to compare the sequence length of the shortest scaffold at 50% of the total genome length (N50) and the smallest number of scaffolds whose total length makes up half of the genome size (L50) values as well as the assembly length and the number of contigs and scaffolds. Busco v4 (Seppey et al., 2019) was implemented using a Docker (Merkel, 2014) container (default parameters, mode: genome) on the genomes using the aves_odb10 dataset to assess the assembly completeness.

2.5 | Genome annotation

Geneious R.11 (https://www.geneious.com) was used to extract the coding sequences (CDS) from *B. regulorum* genome (GCA_000709895) and these were filtered to retain only the longest CDS per gene where multiple annotations existed. Gmap (version 2019-09-12) (Wu & Watanabe, 2005) was used to annotate the newly assembled genomes. Each assembly was first indexed using the gmap_buil function, and then *B. regulorum* CDS were mapped to it with the setting -f 2 to obtain a GFF3 formatted annotation.

2.6 | Extracting coding regions

During the assembly process, exons from the same gene are sometimes assembled into different scaffolds. To obtain a sequence list containing the entire coding region for each gene, the exons were extracted using Geneious R.11 and remapped to the *B. regulorum* CDS with BWA (0.7.17-r1188) using BWA-mem with the default settings (Li, 2013).

To assess the size and quality of the extracted CDS for each genome they were compared to the *B. regulorum* reference. The quality (complete or partial) of coding regions retrieved was assessed using the samtools V.1.9 (Li et al., 2009) faidx tool (to obtain the length of each sequence) and a custom R script to compare the CDS sequences with the reference (see supplementary data).

2.7 | Heterozygosity

Read depth, coverage, and heterozygosity of the newly assembled genomes were estimated using a random selection of 20 genes (*ADA*, *DHX40*, *ENPEP*, *EXOG*, *FAM196B*, *FUBP3*, *GOLGA7B*, *GRHL3*, *KCNK5*, *LEMD3*, *LOC104630315*, *LOC104633950*, *LOC104643156*, *MLNR*, *MMS19*, *PIANP*, *THOC3*, *ZCCHC2*, *ZNF410*, and *ZRANB1*) with a total length of 266,456 bp and the paired reads for each

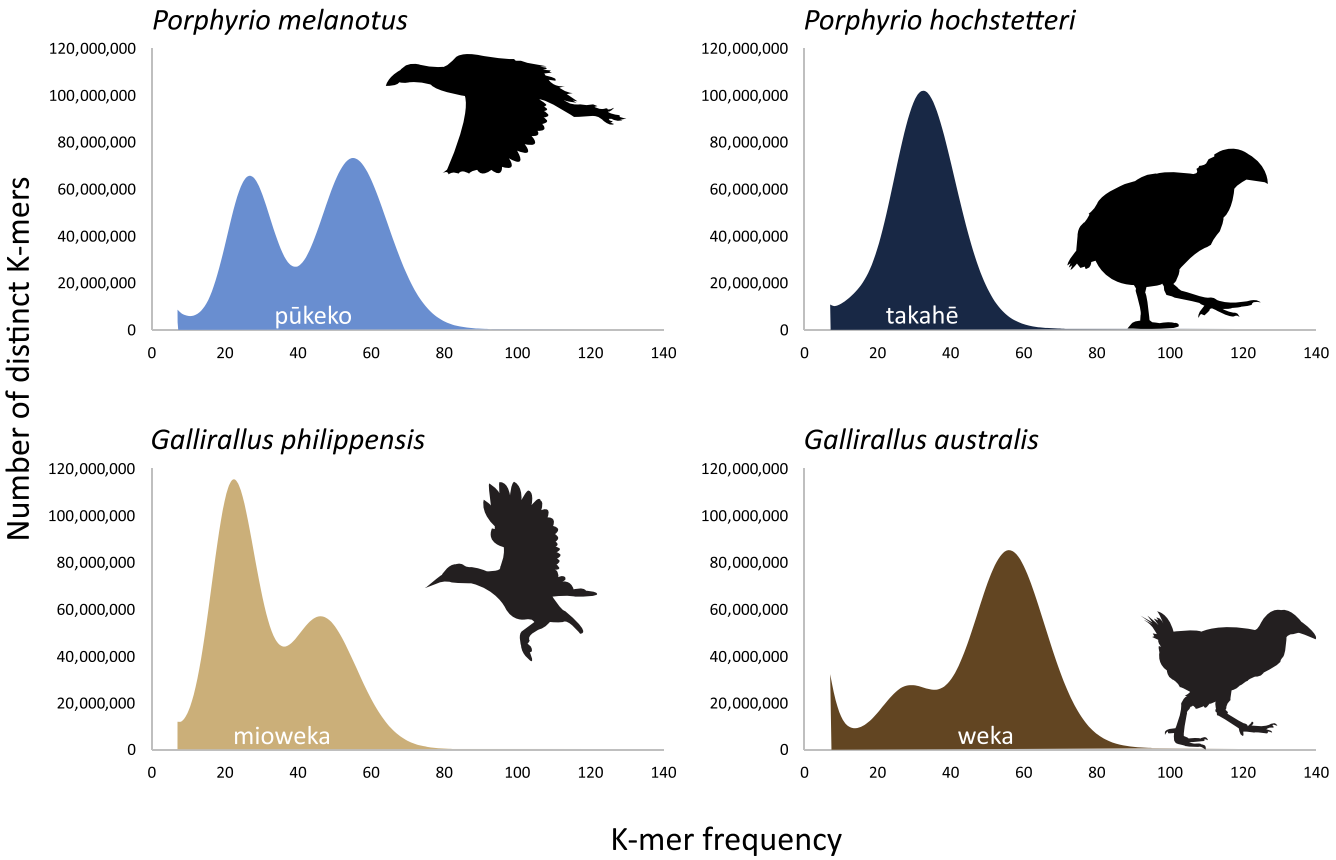

FIGURE 2 K-mer frequency in four rails from Aotearoa New Zealand. K-mer (nucleotide sequence of a certain length) were 57, 61, 61, 87 for *Gallirallus australis*, *Gallirallus philippensis*, *Porphyrio melanotus* and *Porphyrio hochstetteri* respectively. In each distribution, two main peaks correspond to the genomic K-mers for the heterozygous (left) and homozygous (right) parts of the genome. The single main peak of *P. hochstetteri* indicates high homozygosity. Low depth peaks corresponding to erroneous K-mer populations have been masked for clarity. Icons indicate flightless and volant species.

genome mapped to them in Geneious R.11 with low sensitivity/fast mapping settings. The Geneious 'Find variations/SNPs' tool in the 'Annotate & Predict' section was used with the following settings: a minimum coverage of 50 and a minimum variant frequency of 0.3 to locate the heterozygous sites. Heterozygosity was then estimated by dividing the number of heterozygous sites by the total length of the concatenated gene sequences. This method, despite not using the whole genome to assess the heterozygosity level of each species, generates reliable estimates that can be compared between lineages.

2.8 | Phylogeny

A basic phylogenetic inference was performed to show relative relationships between the four new genomes and other selected rails with *Balearica regulorm* as outgroup. Ten genes selected from a set of universal nuclear markers suitable for avian phylogenetic reconstruction (Liu et al., 2018) were used to construct the phylogenetic tree. The genes were *ADNP*, *BEGAIN*, *INO80D*, *KBTBD8*, *NCOA6*, *RHOBTB1*, *S1PR3*, *SPECC1L*, *ZNF618*, and *ZNF654*. These 10 CDS alignments were concatenated into a 21,390 bp alignment using Phyluce v1.7.1 (Faircloth, 2016) with the default settings and the best-fit partitioning scheme was determined using PartitionFinder2 (Lanfear et al., 2017) via the CIPRES Science Gateway (Miller et al., 2010). A list of genes and partitions can be found in the supplementary data. Maximum Likelihood (ML) analyses were implemented in RaxML v8.2.10 (Stamatakis, 2014) via the CIPRES Science Gateway with bootstrapping automatically stopped employing the majority rule criterion. The consensus tree was then visualised in Geneious (Figure 1).

3 | RESULTS

3.1 | DNA extraction and sequencing

The raw data comprised between 780 million (*G. philippensis*) and 936 million (*G. australis*) paired reads per species. Most of these were retained after the filtering and cleaning step (Table 2). Fastp generates a Phred quality score (Q score) for each of the species that represents the ratio of bases with a probability of containing no more than 1/100 (Q20) or in 1/1000 (Q30) errors (Ewing et al., 1998; Ewing & Green, 1998; Richterich, 1998). These scores range between 97.37%

TABLE 2 Fastp outputs after the sequencing of four rail species indicating the number of reads before and after filtering as well as the quality assessment.

	Before fastp filtering	After fastp filtering			
Species	Total reads	Total reads	% reads retained	Q20 bases (%)	Q30 bases (%)
Porphyrio melanotus	894.570034 M	881.624382 M	98.55	97.73	94.46
Porphyrio hochstetteri	845.999032 M	817.395666 M	96.62	97.37	93.84
Gallirallus philippensis	781.084610 M	760.059606 M	97.31	97.39	93.74
Gallirallus australis	936.861886 M	917.214824 M	97.90	98.05	95.23

and 98.5% for Q20 and between 93.74% and 95.23% for Q30 implying high sequencing quality for all four species.

K-mer frequency plots can be used to estimate the level of heterozygosity for each individual and by proxy each species. Indeed, k-mers from the heterozygous regions (left peak on Figure 2) will have half the sequencing coverage (i.e., K-mer frequency) compared to the homozygous regions (right peak). The higher the left peak the higher the heterozygosity. The two volant species *G. philippensis* and *P. melanotus* exhibited high heterozygosity with the left peak being higher than the right for *G. philippensis*. A very low left peak was found for the *G. australis* data and only one peak was observed for *P. hochstetteri*. This implies a much lower level of heterozygosity for both of the endemic, flightless species that have limited populations.

In addition to the K-mer frequencies, the relative heterozygosities among the newly assembled genomes were compared by mapping the paired reads to a set of 20 genes for each species. The ratio of heterozygous sites divided by the total sequence length was calculated (Figure 3). The two volant species showed a higher heterozygosity level than the two flightless species. Based on the paired reads mapping, the mean depth of coverage was calculated for each species (Table 3) with the overall average being 96.4x.

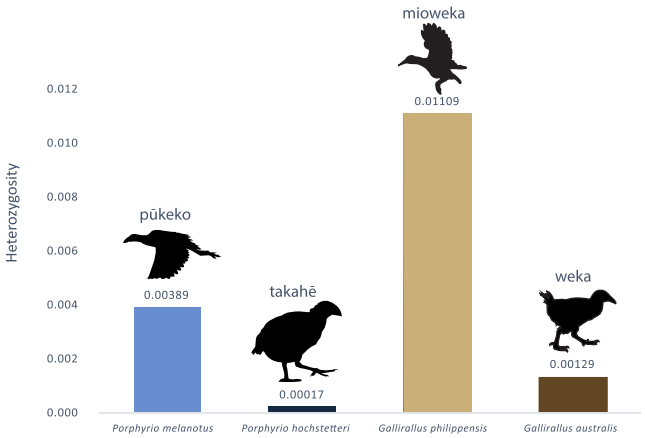

FIGURE 3 Average heterozygosity at 20 randomly selected genes from four newly assembled and annotated rail genomes (average total length 266,456 bp). Heterozygosity is the proportion of total nucleotide sites per individual site having two bases. Icons indicate flightless and volant species.

3.2 | Genome assembly

Meraculous de novo assemblies yielded scaffold N50 between 126 kb (*G. australis*) and 30 kb (*P. hochstetteri*) and scaffold L50 between 2365 (*G. australis*) and 6047 (*G. philippensis*) (Table 2). The total genome assembly size of the four newly assembled rails differed little with a range from 1.07 Gb (*G. philippensis*) to 1.16 Gb (*G. australis*). This was similar to the previously assembled rails (between 1.11 and 1.27 Gb, see Table 2) and slightly shorter than the crane *B. regulorum* (1.22 Gb).

BUSCO scores were similar for *P. melanotus*, *P. hochstetteri*, and *G. australis* with close to 80% of single copy genes which were found complete. In contrast, *G. philippensis* comprised 69% of "Complete single copy" and had a higher proportion (17%) of missing genes (Table 3). A chromosome-level assembly of *P. hochstetteri* became available while this work was in progress. Unsurprisingly, the comparison between both assemblies shows that the long-read sequencing technology yields a higher BUSCO score than the short-read one. Nevertheless, the new assemblies have comparable BUSCO scores to other short-read rail genome assemblies.

3.3 | Extracting coding regions

For the four new rail assemblies, the coding regions of each gene were extracted based on the annotations and compared with the respective *B. regulorum* CDS. Over 9000 gene CDSs were retrieved near-complete (above 95% of the reference CDS nucleotide sequence length) for the two *Porphyrio* species and *Gallirallus australis* (Figure 4). *G. philippensis* exhibited a slightly lower proportion (8259 CDS over 95%) which was consistent with the BUSCO results. The CDSs present in the reference genome but not in the rail data ("Not found" in Figure 4) represent less than 7.5% of the CDS for all species.

3.4 | Phylogeny

A maximum likelihood phylogenetic inference was made based on 10 genes selected from a set of universal avian markers (Liu et al., 2018) (Figure 1). The generated tree is consistent with the previously published rallid phylogenies (Garcia-R et al., 2014a, 2014b; Kirchman

TABLE 3 De novo genome assembly metrics among 8 rail species and one crane (*Balearica regulorum*).

Species	Genome assembly size (Gb)	Largest scaffold	Number of scaffolds	Scaffold N50	Scaffold L50	Number of contigs	Contig N50 (kb)	Contig L50	Depth of coverage	BUSCO score (%)	Sequencing technology
Porphyrio melanotus[a]	1.11	1,068,914	34,563	82,204	3707	159,218	16.5	16,605	102	C:82,3[S:82.2,D:0.1], F:8.8,M:8.8	Illumina HiSeq
Porphyrio hochstetteri[a]	1.12	1,015,032	30,278	30,278	3641	76,213	40.8	7446	95	C:79.8[S:79.7,D:0.1], F:7.5,M:12.6	Illumina HiSeq
Gallirallus philippensis[a]	1.07	722,688	55,205	46,737	6047	209,712	9.6	28,118	103	C:69.1[S:69,D:0.1], F:14,M:16.8	Illumina HiSeq
Gallirallus australis[a]	1.16	1,692,012	36,524	126,032	2365	96,978	41.2	7224	86	C:80.8[S:80.6,D:0.2], F:6.7,M:12.3	Illumina HiSeq
Atlantisia rogersi	1.17	1,015,111	159,311	36,139	8295	160,845	36,1	8295	41	C:65[S:64.9,D:0.1], F:15,M:19.9	Illumina HiSeq
Zapornia atra	1.12	1,795,565	58,849	134,191	2049	113,655	44,130	6609	45	C:80.3[S:80.2,D:0.1], F:7.8,M:11.8	Illumina HiSeq
Fulica atra	1.17	27,139,163	17,827	6,390,841	46	31,348	2461	1314	53	C:92.3[S:91.6,D:0.7], F:1.7,M:5.3	Illumina NovaSeq
Porphyrio hochstetteri	1.27	224,114,340	173	71.6 MB	5	500	13.5 MB	31	367	C:95.5[S:95.0,D:0.5], F:0.7,M:3.8	PacBio Sequel II HiFi; Bionano Genomics DLS; Illumina HiSeq; Arima Genomics Hi-C v2
Galirallus okinawae	1.18	218,223,205	258	101.8 MB	4	440	20,700	16	100	C:94.6[S:94.2,D:0.4], F:1.2,M:3.8	Illumina Novaseq6000; ONT PromethION
Balearica regulorum	1.22	219,267,915	104	82,577,926	5	248	23.3	14	60	C:96.2[S:95.6,D:0.6], F:0.6,M:3.2	PacBio Sequel I CLR; Ilumina NovaSeq; Arima Genomics Hi-C; Bionano Genomics DLS

Note: BUSCO score out of $n = 8338$ BUSCO genes: Complete (C), Complete and single-copy (S), Complete and duplicated (D), Fragmented (F), Missing (M).

[a]New assemblies.

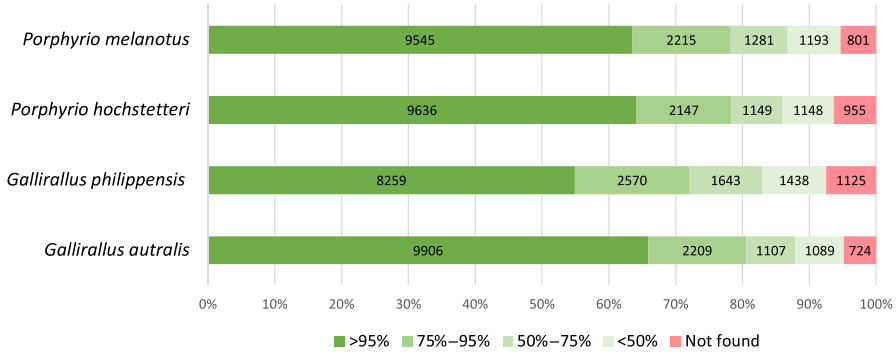

FIGURE 4 Completeness of CDSs retrieved from eight rail genomes compared to the reference crane *Balearica regulorum* genome that has a total of 15,035 annotated CDSs. Colours indicate the proportion of genes retrieved from a sample at various scales of completeness.

et al., 2021) and shows the relative relationships of the species under consideration.

4 | DISCUSSION

We present a straightforward method, based on limited sequencing resources, to generate genomic data for comparative analyses. Although this method does not result in chromosome level assemblies, it does not require a reference genome, is easily reproducible, and retrieves a significant proportion of the coding regions.

Considerable variation was observed between species heterozygosity (Figures 2 and 3). Indeed, the two volant species were more heterozygous than the flightless ones (Figure 3) with big differences being observed between the most heterozygous species, *G. philippensis* (frequency of heterozygous site of 0.01) and the least heterozygous species *P. hochstetteri* (0.0002). Those observations were consistent with the K-mer frequencies (Figure 2). The low heterozygosity in flightless species probably reflects their much reduced populations owing to habitat loss and invasive predators (Baker et al., 1995; Burga et al., 2017; White et al., 2018). The takahē *P. hoschetteri* is a critically endangered flightless species with a population of only 500 in 2023 (www.doc.govt.nz), all derived from a remnant discovered in the 1950s that may have numbered as low as two individuals (Wallace, 2002). The resulting inbreeding depression likely explains its extremely low level of heterozygosity (Grueber et al., 2010). *Gallirallus philippensis* on the other hand is a relatively abundant species with a geographic range that includes the islands of Aotearoa New Zealand and the western Pacific (Garcia-R et al., 2017; Trewick, 1997b) which is likely to maintain high heterozygosity at the species level.

The four newly assembled genomes have similar or better characteristics than the other rail genomes assembled from Illumina HiSeq data (Table 3) with N50 and L50 scaffolds within the same range as these other rails. The BUSCO results (Table 3) and CDS extractions (Figure 4) showed similar trends and add to our confidence that the genome assemblies are of good quality with limited assembly errors. Despite being naturally more fragmented than those assembled using long-read sequencing technology (Table 3), a significant proportion of full-length coding regions were identified and extracted showing good utility for future comparative analyses (Figure 4).

Among the four newly assembled rallid genomes, *G. philippensis* had the lowest proportion of complete genes according to both the BUSCO (Table 3) and extracted CDS comparison (Figure 4). This can be attributed to the high heterozygosity level which generally makes the assembly process more challenging due to the increased complexity of the de Bruijn graph structure (Kajitani et al., 2014). Nonetheless, the *G. philippensis* genome is a good quality assembly that can be used to investigate evolutionary processes along with the three other assembled genomes. For all four genomes, a large majority of the genes were retrieved. In all species, over 70% of the genes were identified with greater than 75% completeness.

To conclude, we provide here four new avian assemblies which represent valuable genomic resources to investigate evolutionary processes within the rail family. The quality checks that were performed showed that the generated assemblies are reliable. Comparing the newly assembled genomes showed lower levels of heterozygosity in flightless species which likely reflects their relatively small populations. This study significantly increases the number of available rallid genomes, targeting flying-flightless pairs; this creates new opportunities to investigate the evolution of avian flightlessness.

AUTHOR CONTRIBUTIONS

Julien Gaspar: Conceptualization (equal); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); software (equal); visualization (lead); writing – original draft (lead); writing – review and editing (equal). **Steve A. Trewick:** Conceptualization (supporting); data curation (supporting); formal analysis (supporting); funding acquisition (supporting); methodology (supporting); supervision (supporting); writing – original draft (supporting); writing – review and editing (equal). **Gillian C. Gibb:** Conceptualization (equal); data curation (supporting); formal analysis (supporting); funding acquisition (lead); investigation (supporting); methodology (equal); project administration (lead); software (equal); supervision (lead); writing – original draft (supporting); writing – review and editing (equal).

ACKNOWLEDGEMENTS

This study was supported by the New Zealand Marsden Fund Council from Government funding, managed by the Royal Society Te Apārangi, grant MAU1601 to GCG. Thanks to Roger Moraga for initial bioinformatic discussions and assistance using the software Meraculous. The genome assemblies were generated with the help of Genomics for Aotearoa New Zealand (GFANZ, genomics.nz), thanks

to Rob Elshire. The authors would like to thank Richard Witehira who provided the helpful local knowledge about the mioweka name. Thanks also to Jonathan Proctor (Rangitāne o Manawatū) for ongoing consultation around the role of Rangitāne o Manawatū as kaitiaki (guardians) of the museum collection samples held by Massey University Palmerston North. Open access publishing facilitated by Massey University, as part of the Wiley - Massey University agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA782688/ and https://figshare.com/s/3a89eea20c4607abbefe.

DATA AVAILABILITY STATEMENT

The genomes and annotations are available on NCBI, BioProject PRJNA782688. The configuration files, command lines used, CDS lists, and R scripts are all publicly available under CC BY 4.0 licence as supplementary data in the Figshare data repository (https://figshare.com/s/3a89eea20c4607abbefe).

ORCID

Julien Gaspar https://orcid.org/0000-0003-3985-4436
Gillian C. Gibb https://orcid.org/0000-0002-4283-9790

REFERENCES

Baker, A. J., Daugherty, C. H., Colbourne, R., & McLennan, J. L. (1995). Flightless brown kiwis of New Zealand possess extremely subdivided population structure and cryptic species like small mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *92*, 8254–8258.

Boast, A. P., Chapman, B., Herrera, M. B., Worthy, T. H., Scofield, R. P., Tennyson, A. J. D., Houde, P., Bunce, M., Cooper, A., & Mitchell, K. J. (2019). Mitochondrial genomes from New Zealand's extinct adzebills (Aves: Aptornithidae: Aptornis) support a sister-taxon relationship with the Afro-Madagascan Sarothruridae. *Diversity*, *11*, 24.

Burga, A., Wang, W., Ben-David, E., Wolf, P. C., Ramey, A. M., Verdugo, C., Lyons, K., Parker, P. G., & Kruglyak, L. (2017). A genetic signature of the evolution of loss of flight in the Galapagos cormorant. *Science*, *356*, eaal3345.

Chapman, J. A., Ho, I., Sunkara, S., Luo, S., Schroth, G. P., & Rokhsar, D. S. (2011). Meraculous: De novo genome assembly with short paired-end reads. *PLoS One*, *6*, e23501.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*, i884–i890.

Chikhi, R., & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, *30*, 31–37.

Clements, J. F., Rasmussen, P. C., Schulenberg, T. S., Iliff, M. J., Fredericks, T. A., Gerbracht, J. A., Lepage, D., Spencer, A., Billerman, S. M., Sullivan, B. L., & Wood, C. L. (2023). *Clements checklist of birds of the world: v2023*. https://www.birds.cornell.edu/clementschecklist/download/.

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Research*, *8*, 186–194.

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces UsingPhred. I. Accuracy assessment. *Genome Research*, *8*, 175–185.

Fain, M. G., Krajewski, C., & Houde, P. (2007). Phylogeny of "core Gruiformes" (Aves: Grues) and resolution of the Limpkin–Sungrebe problem. *Molecular Phylogenetics and Evolution*, *43*, 515–529.

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*, 786–788.

Garcia-R, J. C., Gibb, G. C., & Trewick, S. A. (2014a). Eocene diversification of crown group rails (Aves: Gruiformes: Rallidae). *PLoS One*, *9*, e109635.

Garcia-R, J. C., Gibb, G. C., & Trewick, S. A. (2014b). Deep global evolutionary radiation in birds: Diversification and trait evolution in the cosmopolitan bird family Rallidae. *Molecular Phylogenetics and Evolution*, *81*, 96–108.

Garcia-R, J. C., Joseph, L., Adcock, G., Reid, J., & Trewick, S. A. (2017). Interisland gene flow among populations of the buff-banded rail (Aves: Rallidae) and its implications for insular endemism in Oceania. *Journal of Avian Biology*, *48*, 679–690.

Garcia-R, J. C., & Trewick, S. A. (2014). Dispersal and speciation in purple swamphens (Rallidae: Porphyrio). *The Auk*, *132*, 140–155.

Gaspar, J., Gibb, G. C., & Trewick, S. A. (2020). Convergent morphological responses to loss of flight in rails (Aves: Rallidae). *Ecology and Evolution*, *10*, 6186–6207.

Grueber, C. E., Laws, R. J., Nakagawa, S., & Jamieson, I. G. (2010). Inbreeding depression accumulation across life-history stages of the endangered takahe. *Conservation Biology*, *24*, 1617–1625.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., & Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, *24*, 1384–1395.

Kirchman, J. J. (2012). Speciation of flightless rails on islands: A DNA-based phylogeny of the typical rails of the Pacific. *The Auk*, *129*, 56–69.

Kirchman, J. J., Rotzel McInerney, N., Giarla, T. C., Olson, S. L., Slikas, E., & Fleischer, R. C. (2021). Phylogeny based on ultra-conserved elements clarifies the evolution of rails and allies (Ralloidea) and is the basis for a revised classification. *Ornithology*, *138*, ukab042.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2017). PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, *34*, 772–773.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*:1303.3997.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Liu, Y., Liu, S., Yeh, C.-F., Zhang, N., Chen, G., Que, P., Dong, L., & Li, S. (2018). The first set of universal nuclear protein-coding loci markers for avian phylogenetic and population genetic studies. *Scientific Reports*, *8*, 15723.

Livezey, B. C. (2003). *Evolution of flightlessness in rails*. American Ornithologists' Union.

Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. *Linux Journal*, *2*, 2.

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop (GCE)* (pp. 1–8).

Olson, S. L. (1973). Evolution of the rails of the South Atlantic islands (Aves: Rallidae). *Smithsonian Contributions to Zoology*, *152*, 1–53.

Richterich, P. (1998). Estimation of errors in "raw" DNA sequences: A validation study. *Genome Research*, *8*, 251–259.

Ripley, S. D., Lansdowne, J. F., & Olson, S. L. (1977). *Rails of the world: A monograph of the family Rallidae*. M. F. Feheley Publishers.

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. In M. Kollmar (Ed.), *Gene prediction: Methods and protocols, methods in molecular biology* (pp. 227–245). Springer.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313.

Steadman, D. W. (1995). Prehistoric extinctions of Pacific Island birds: Biodiversity meets zooarchaeology. *Science*, *267*, 1123–1131.

Taylor, B. (1998). *Rails: A guide to the rails, crakes, gallinules and coots of the world*. Bloomsbury Publishing.

Trewick, S. A. (1997a). Flightlessness and phylogeny amongst endemic rails (Aves: Rallidae) of the New Zealand region. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences*, *352*, 429–446.

Trewick, S. A. (1997b). Sympatric flightless rails *Gallirallus dieffenbachii* and *G. modestus* on the Chatham Islands, New Zealand; morphometrics and alternative evolutionary scenarios. *Journal of the Royal Society of New Zealand*, *27*, 451–464.

Trewick, S. A., Pilkington, S., Shepherd, L. D., Gibb, G. C., & Morgan-Richards, M. (2017). Closing the gap: Avian lineage splits at a young, narrow seaway imply a protracted history of mixed population response. *Molecular Ecology*, *26*, 5752–5772.

Wallace, G. E. (2002). The takahe: Fifty years of conservation management and research. *The Auk*, *119*, 291–293.

White, D. J., Ramón-Laca, A., Amey, J., & Robertson, H. A. (2018). Novel genetic variation in an isolated population of the nationally critical Haast tokoeka (Apteryx australis 'Haast') reveals extreme short-range structure within this cryptic and flightless bird. *Conservation Genetics*, *19*, 1401–1410.

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*, 1859–1875.

How to cite this article: Gaspar, J., Trewick, S. A., & Gibb, G. C. (2024). De-novo assembly of four rail (Aves: Rallidae) genomes: A resource for comparative genomics. *Ecology and Evolution*, *14*, e11694. https://doi.org/10.1002/ece3.11694