G3, 2021, 11(2), jkab009
DOI: 10.1093/g3journal/jkab009
Advance Access Publication Date: 23 January 2021
Genome Report

The assembled and annotated genome of the pigeon louse *Columbicola columbae*, a model ectoparasite

James G. Baldwin-Brown [1,*] Scott M. Villa [1,2] Anna I. Vickrey,[1] Kevin P. Johnson,[3] Sarah E. Bush,[1] Dale H. Clayton,[1] and Michael D. Shapiro [1,*]

[1]School of Biological Sciences, University of Utah, Salt Lake City, UT 84112, USA
[2]Department of Biology, O. Wayne Rollins Research Center, Emory University, Atlanta, GA 30322, USA
[3]Illinois Natural History Survey, Prairie Research Institute, University of Illinois, Champaign, IL 61820, USA

*Corresponding authors: School of Biological Sciences, University of Utah, 257 South 1400 East; Salt Lake City, UT 84112, USA. jgbaldwinbrown@gmail.com (J.G.B.-B.); mike.shapiro@utah.edu (M.D.S.)

Abstract

The pigeon louse *Columbicola columbae* is a longstanding and important model for studies of ectoparasitism and host-parasite coevolution. However, a deeper understanding of its evolution and capacity for rapid adaptation is limited by a lack of genomic resources. Here, we present a high-quality draft assembly of the *C. columbae* genome, produced using a combination of Oxford Nanopore, Illumina, and Hi-C technologies. The final assembly is 208 Mb in length, with 12 chromosome-size scaffolds representing 98.1% of the assembly. For gene model prediction, we used a novel clustering method (*wavy_choose*) for Oxford Nanopore RNA-seq reads to feed into the *MAKER* annotation pipeline. High recovery of conserved single-copy orthologs (BUSCOs) suggests that our assembly and annotation are both highly complete and highly accurate. Consistent with the results of the only other assembled louse genome, *Pediculus humanus*, we find that *C. columbae* has a relatively low density of repetitive elements, the majority of which are DNA transposons. Also similar to *P. humanus*, we find a reduced number of genes encoding opsins, G protein-coupled receptors, odorant receptors, insulin signaling pathway components, and detoxification proteins in the *C. columbae* genome, relative to other insects. We propose that such losses might characterize the genomes of obligate, permanent ectoparasites with predictable habitats, limited foraging complexity, and simple dietary regimes. The sequencing and analysis for this genome were relatively low cost, and took advantage of a new clustering technique for Oxford Nanopore RNAseq reads that will be useful to future genome projects.

Keywords: genome assembly; genome annotation; insect genomics; ectoparasitism; phthiraptera; ischnocera

Introduction

Parasites represent a large proportion of eukaryotic biodiversity, and it is estimated that 40% of insect diversity is parasitic (de Meeûs and Renaud 2002). Parasitic lice (Insecta: Phthiraptera) comprise a group of about 5000 species that parasitize all orders of birds and most orders of mammals (Mullen and Durden 2009; Clayton *et al.* 2015). Two thirds of louse species are associated with only a single host species (Durden and Musser 1994; Smith 2004). The genus *Columbicola* comprises 91 known species, all found on pigeons or doves (Bush *et al.* 2009; Gustafsson *et al.* 2015; Adly *et al.* 2019); most of these louse species are found on a single host species (Johnson *et al.* 2007, 2009).

Like all feather lice (suborder Ischnocera), *Columbicola* are "permanent" parasites that complete their entire life cycle on the body of the host (Marshall 1981). Feather lice feed primarily on feathers, which they metabolize with the assistance of endosymbiotic bacteria (Fukatsu *et al.* 2007; Smith *et al.* 2013). The feather damage caused by lice has a chronic effect that leads to reduced host survival (Clayton *et al.* 1999) and mating success (Clayton 1990). Birds are able to defend themselves against feather lice by preening them with the beak. However, *Columbicola* lice escape from preening by hiding in grooves between feather barbs, and the sizes of these grooves scale with host body size. In microevolutionary time, the result is stabilizing selection on body size of lice (Clayton *et al.* 1999; Bush and Clayton 2006). In macroevolutionary time, the result is that host defense (preening) and body size interact to reinforce the host specificity and size matching of *Columbicola* species to their hosts (Clayton *et al.* 2003). Similarly, selection for visual crypsis drives the evolution of color similarities between *Columbicola* species their hosts (Bush *et al.* 2010, 2019).

Within the feather lice, the biology of *C. columbae* (Figure 1) is better known than that of any other louse species, including details about its morphology, physiology, ecology, and behavior (Martin 1934; Stenram 1956; Rakshpal 1959; Nelson and Murray 1971; Eichler *et al.* 1972; Rudolph 1983; Clayton 1990, 1991; Clayton and Tompkins 1995; Clayton *et al.* 1999, 2003, 2008; Bush and Clayton 2006; Bush *et al.* 2006; Harbison and Clayton 2011). A unique feature of the *C. columbae* study system is that its host, the rock pigeon *Columba livia*, has been under artificial selection

Received: October 14, 2020. **Accepted:** December 13, 2020
© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022

Figure 1 Slender pigeon louse (*Columbicola columbae*)—about 2 mm in length—clinging to a rock pigeon feather. The thumblike processes on the antennae of the male louse shown here are used to grasp a female when mating. Dubbed "the bird louse par excellence" (Eichler *et al.* 1972), *C. columbae* has long been a model for studies of host-parasite coevolution. Photo by Scott Villa and Juan Altuna.

by pigeon breeders for millennia, resulting in dramatic phenotypic variation (Darwin 1868; Shapiro and Domyan 2013), similar to that seen across the 300+ other species of pigeons and doves (Gibbs *et al.* 2001). This variation makes it possible to transfer *C. columbae* among diverse size and color phenotypes within the single native host species. Recently, we showed that switching lice to pigeons of different sizes and colors elicits rapid population-level changes in louse size and color (Bush *et al.* 2019; Villa *et al.* 2019). Despite the wealth of phenotypic data about real-time adaptation of *C. columbae* to changes in host environment, the underlying molecular mechanisms remain unknown.

A deeper understanding of louse evolution and genetics is limited largely by a paucity of genomic resources. The louse with the best available genomic resources is the blood-feeding human body louse *Pediculus humanus*, the draft genome of which was assembled using low-coverage shotgun sequencing (Kirkness *et al.* 2010). *Pediculus humanus* had the smallest insect genome known at that time (108 Mb), with a repertoire of 10,773 annotated genes. Presently, what we know about the genomic signatures of parasitism in Phthiraptera is largely limited to this one species. Prior work on *C. columbae* has generated valuable DNA sequence datasets for phylogenetics (Sweet and Johnson 2018; Sweet *et al.* 2018) and studies of mitochodrial evolution (Sweet *et al.* 2021), but whole-genome data are still lacking.

Here, we report a high-quality draft genome assembly and annotation for *C. columbae* that incorporates short-read Illumina (Bennett 2004) sequences, long-read Oxford Nanopore (Jain *et al.* 2016) sequences, and scaffolding using Hi-C data (van Berkum *et al.* 2010). These new resources will enable genomic approaches to understanding the molecular basis of rapid adaptation in *C. columbae*. More generally, the *C. columbae* genome provides comparative genomic data to understand the molecular basis of traits associated with parasitism that are shared among lice.

Materials and methods

Animal tissue samples

All lice used in this study were drawn from natural populations infesting wild-caught feral rock pigeons (*C. livia*) from Salt Lake City, UT. We maintained 15–20 infested pigeons in cages to provide a reliable source of *C. columbae* for sequencing.

We reduced the nucleotide heterozygosity of our colony by creating a partially inbred population of lice. Initially, a single pair of lice (1 male and 1 female) was arbitrarily drawn from the pigeon colony and allowed to reproduce on a new, individually caged louse-free feral pigeon. After a period of 21 days, all immature lice were removed from the pigeon using CO_2. At this point, these F1 lice were all full siblings. All offspring were then individually placed in glass vials with pigeon feathers for food, and allowed to mature. Rearing lice individually in vials ensured that F1s could not mate. Once mature, a single pair of unmated F1 adults (1 male and 1 female) were arbitrarily chosen and placed on a new, louse-free feral pigeon to mate and reproduce. Thus, all offspring on this new pigeon were the product of full-sibling mating and represented the first generation of inbreeding. These methods were repeated for eight generations.

After eight rounds of full-sibling inbreeding, the partially inbred lice were transferred to a new louse-free pigeon and left to mature and produce offspring. We left the lice on this pigeon for 4 months, which allowed the population to grow enough to provide sufficient numbers for sequencing. The lice used for Illumina genomic DNA sequencing were derived from this partially inbred population. Reduced heterozygosity should have resulted in higher quality polishing of the Oxford Nanopore-derived contigs with our Illumina data (see below). We pooled 100 adult lice for Illumina genomic DNA sequencing, 2000 adult lice for Oxford Nanopore genomic DNA sequencing, and 100 adult lice for Illumina RNA sequencing. All lice were drawn from the same partially inbred laboratory population, except for the lice used in Oxford Nanopore DNA sequencing, which were drawn from the main laboratory population from which the inbred population was derived. Ultimately, the inbred louse population was too small to provide sufficient material for Oxford Nanopore DNA sequencing. We generated Oxford Nanopore RNA sequencing reads from four different life stages of lice, all drawn from the inbred population (100 lice each from nymphal instars 1, 2, and 3 adults).

Isolation of genetic material

DNA was isolated by grinding with a disposable homogenizer pestle (VWR, Radnor PA, USA) on ice for 30 min followed by DNA extraction with the Qiagen DNeasy extraction kit (Qiagen, Germantown, MD, USA). DNA for long read sequencing was extracted using the Qiagen DNA Blood and Tissue Midi kit. RNA was isolated using the Qiagen Oligotex mRNA mini kit.

Illumina genomic DNA and RNA sequencing

Illumina DNA sequencing was performed using an Illumina HiSeq 2500 sequencer at the University of Utah High Throughput Genomics Core. We generated four libraries with mean insert sizes of 180, 500, 3500, and 8200 bp. Genomic DNA was sequenced with paired-end 125-bp reads. cDNA sequencing was also performed on the Illumina HiSeq 2500 sequencer, producing paired end reads with a read length of 125 bp.

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022

Oxford Nanopore genomic DNA and RNA sequencing

We generated long read genomic data using Oxford Nanopore MinION sequencers and a custom library preparation designed to increase read length. This protocol followed the standard procedure for producing 1d[2] reads with kit LSK308 (Oxford Nanopore community, https://community.nanoporetech.com/protocols/), with the following modifications. (1) During all alcohol washes of magnetic SPRI beads, an additional wash was performed using Tris-EDTA to remove small DNA fragments. This step was performed quickly and without disturbing the beads to avoid dissolving all available DNA into solution. (2) All elutions from magnetic SPRI beads were performed after an incubation in elution buffer at 37° for 30 min. These practices improve the length of Oxford Nanopore sequencing reads (Urban *et al.* 2015).

We generated long mRNA reads using Oxford Nanopore MinION sequencers and a standard cDNA PCR-based sequencing method (PCS109, Oxford Nanopore community, https://community.nanoporetech.com/protocols/).

Genome size estimation

We used the following formula (Liu *et al.* 2020) to estimate genome size from 21-mers counted from the Illumina sequencing data using *jellyfish* (Marçais and Kingsford 2011):

$$G = \frac{n_{k-mer}}{c_{k-mer}} = \frac{n_{base}}{c_{base}} = \frac{n_{base}}{c_{k-mer} \cdot \frac{L}{L-K+1}} \quad (1)$$

where G is the genome size, n is the total number of sequenced bases, c is the expected sequence coverage depth, L is the average sequencing read length, and K is the k-mer length.

Genome assembly

We used *Trimmomatic* version 0.36 (Bolger *et al.* 2014) to trim Illumina input reads using the following settings: ILLUMINACLIP: adapters.fa: 2:30:10 LEADING: 20 TRAILING: 20 MINLEN: 30 CROP: 85. We then used *fastq-join* from *ea-utils* version 1.1.2-537 (Aronesty 2011) to join all short reads into pair joined reads, and used these throughout the assembly process. We used *Canu* version 1.6 (Koren *et al.* 2017) with the parameter genome Size = 220m to assemble Oxford Nanopore genomic DNA reads, then polished the assembled contigs using *pilon* v1.22 (Walker *et al.* 2014) and the Illumina genomic DNA reads. The *pilon* software was run with the following switches: –changes –vcf –vcfqe –tracks –fix all.

The polished draft assembly was scaffolded by Phase Genomics using their proprietary scaffolding software (Burton *et al.* 2013; Bickhart *et al.* 2017; Peichel *et al.* 2017). We supplied Phase Genomics with approximately 1600 lice preserved at −80° for high molecular weight DNA extraction, Hi-C library preparation, and sequencing (Belton *et al.* 2012).

Transcript selection and assembly

Standardized pipelines do not yet exist for selecting transcripts from raw Oxford Nanopore RNAseq reads. Therefore, we produced a custom pipeline that identifies putatively full-length transcripts to serve as evidence for genome annotation. In short, we aligned all RNAseq reads using *Minimap* (Li 2018), then clustered these alignments into sets that represent a gene using *Carnac-LR* (Marchet *et al.* 2019). We wrote a program, *wavy_choose*, that extracts the aligned reads from the original data, then identifies reads that likely represented full-length transcripts using *scipy*'s function scipy.signal.find_peaks_cwt() (Du *et al.* 2006; Figure 2). A more detailed version of the pipeline follows, below.

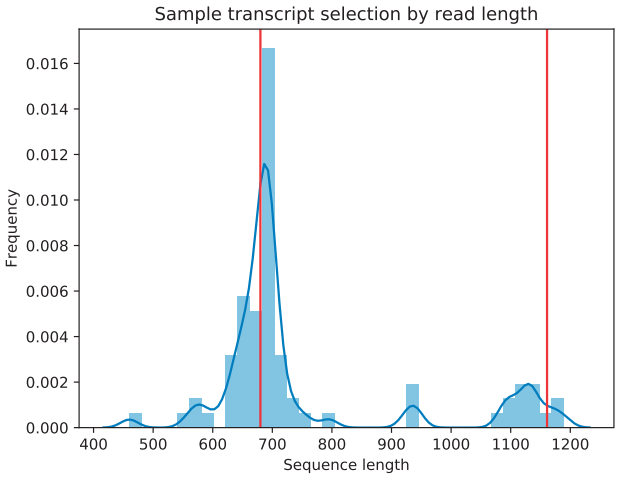

Figure 2 *wavy_choose* identifies likely full-length transcripts from clustered Oxford Nanopore reads. Depicted here is a histogram of read lengths (blue) for one *carnac-LR*-clustered set of reads. *wavy_choose* is able to identify two length peaks (red lines) in this transcript set, and discards all reads of other lengths. This process simplifies the transcriptome evidence dataset for MAKER, which uses the identified reads for gene annotation.

Minimap aligns long, low-quality reads against one another, and can do so in an all-by-all comparison. *Carnac-LR* then clusters these reads into groups according to their alignments. Each *Carnac-LR*-clustered group of mRNA sequencing reads should represent all of the reads associated with a single gene, but if a gene has multiple alternative transcripts, *Carnac-LR* will not distinguish between them. The custom tool *wavy_choose* takes all of the clustered reads identified by *Carnac-LR* and identifies clusters within clusters that are most similar in both length and sequence. Because Oxford Nanopore reads are generally long enough to span an entire mRNA transcript, *wavy_choose* identifies the reads most likely to be complete transcripts by identifying the most common read lengths. It then removes all nonfull-length reads from the analysis. This tool is especially well suited to transcript discovery, as multiple alternative transcripts may be identified from a single cluster of reads with overlapping sequence, and *wavy_choose* makes no assumptions as to the number of transcripts to identify.

The function find_peaks_cwt() uses continuous wavelet transformation, a technique from signal processing (Grossmann and Morlet 1984) to identify peaks in a 2-dimensional dataset. It does this by first convolving (transforming) the dataset to amplify the portion of the dataset that matches a wavelet with specified parameters (here, the default "Mexican hat" wavelet) and dampens the portions of the dataset that do not match the wavelet. The program then identifies local relative maxima that appear at the specified peak widths (here, 50–200 bp) and have sufficiently high signal-to-noise ratio (here 1.0). This widely accepted technique is straightforward to apply in this context, but it is limited to detecting transcripts that have unique lengths. Two alternative transcripts of matching lengths would appear as a single peak in the length histogram. In these cases, reads from both alternative transcripts were retained in the final dataset. We kept at least one read per cluster of reads.

Untrimmed Illumina cDNA reads were assembled using *Trinity* with the –jaccard_clip setting (Grabherr *et al.* 2011).

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022

Genome annotation

We used a combination of *wavy_choose*-selected Oxford Nanopore-derived transcripts, Illumina RNAseq-derived *Trinity* assemblies, and orthology information from Swissprot as evidence for gene models in *MAKER* (Cantarel *et al.* 2007), a widely used genome annotation tool. We used *AUGUSTUS* 3.3.1 to perform the gene finding portion of the *MAKER* pipeline. *BUSCO* (Simão *et al.* 2015) trains *AUGUSTUS* as part of the *BUSCO* pipeline, so we ran *BUSCO* on the genome assembly and used its *AUGUSTUS* training model during gene finding. We used both WU BLAST (Chao *et al.* 1992) and InterProScan (Jones *et al.* 2014) to match genes to their orthologs in the Uniprot-Swissprot database, and to provide the GO terms associated with genes in the final annotation set.

Feature density analysis

We used *bedops* (Neph *et al.* 2012) to generate a.bed file of sliding windows across all chromosomes, then used *bedmap* (Neph *et al.* 2012) to count genes and repetitive sequences in these windows. Sliding windows were 1 Mb in width with a step length of 100 kb. For genes, we counted the total number of features identified by *MAKER* as "gene"s in its output.gff file. For repeats, we counted all *MAKER*-identified *repeatmasker* "match"es.

Detection of bacterial contaminants

After assembly and annotation, we manually checked the louse genome for contamination with bacterial genomic sequences by identifying regions with unusually high gene density, *repeatmasker*-identified (Chen 2004) artifacts, and contiguous runs of bacterial genes. We also used *kraken* (Wood and Salzberg 2014) with the DustMasked MiniKraken DB database (https://ccb.jhu.edu/software/kraken/dl/minikraken_20171101_8GB_dustmasked.tgz) to identify known bacterial kmer contaminants.

We identified two sections of the genome that likely contained bacterial contamination, and removed them from the final assembly. The first section, at the beginning of chromosome 4, had a higher density of genes than any other region of the genome (280 genes per 10 kb, vs. 64 genes per 10 kb in the bacteria-free genome). It also had a paucity of repetitive elements (262 repeats per 10 kb, as opposed to 800 elsewhere). *MAKER*'s annotation (see below) indicated that the majority of the region's genes were bacterial in origin, and *BLASTn* searches (Zhang *et al.* 2000) against the NCBI *nr* database (https://blast.ncbi.nlm.nih.gov/) confirmed this, as did *kraken*. The region also contained the annotation's only instance of an explicit bacterial artifact identified by *repeatmasker*. The second region, on chromosome 8, was flagged as containing bacterial content by *kraken*. Both the chromosome 4 and 8 regions contained genes annotated by *MAKER* as similar to genes from the *Sodalis* clade, which contains the endosymbiont of the tsetse fly and a known bacterial endosymbiont of *C. columbae* (Fukatsu *et al.* 2007; Smith *et al.* 2013). Two hundred nineteen of the 554 genes in the chromosome 4 section are annotated as being *Sodalis*-related, as are 3 of the 4 genes in the chromosome 8 section. Thus, the totality of evidence led us to conclude that these regions on chromosomes 4 and 8 of our preliminary *C. columbae* genome assembly were bacterial contaminants from a known *Sodalis*-clade endosymbiont.

Lice were starved for 24 h and the transparent gut was checked visually for content before DNA and RNA extraction, reducing the likelihood of contamination due to eukaryotic tissue in the gut. Nevertheless, these measures do not completely rule out sequence contamination from the pigeon host, humans, or other eukaryotes. We searched for contamination from eukaryotes by performing *BLASTn* searches (Zhang *et al.* 2000) against the human reference genome (Schneider *et al.* 2017), the *C. livia* reference genome (Holt *et al.* 2018), and the NCBI *nr* nucleotide database. We did not find any regions in the *C. columbae* genome greater than 3 kb in length and with identity greater than 90% to any of the target sequence databases. Therefore, we concluded that there is not substantial eukaryotic contamination in the final assembly.

Data availability

Raw sequence data for this project are publicly available through NCBI SRA (SAMN16076762-SAMN16076765). All analysis scripts are available through GitHub at https://github.com/jgbaldwinbrown/jgbutils. The genome assembly and annotation are available at NCBI GenBank (PRJNA662097).

Results and discussion

Genome size estimation

We generated 2.92×10^{10} bases of genomic sequence using the Illumina short-read platform (mean read length after trimming = 107.2 bp). We estimated the genome size via *k*-mer counting (Liu *et al.* 2020) using *jellyfish* (Marçais and Kingsford 2011) (Figure 3). Using a k-mer size of 21, we estimate the genome size of *C. columbae* to be 230 Mb, within the range expected for insects.

Genome assembly summary

We generated a high-quality draft genome assembly using a combination of Illumina and Oxford Nanopore sequencing data, and Hi-C scaffolding (Table 1). Our initial, unscaffolded assembly with *Canu* consists of 1193 contigs with a total length of 206 Mb, and an N50 contig length of 511 kb. We scaffolded the assembly using Hi-C data, producing chromosome-size scaffolds from the initial contigs. The final assembly comprises 12 chromosome-sized scaffolds and 380 small scaffolds, totaling 208 Mb of sequence. The N50 scaffold length for the final assembly is 17.7 Mb. Karyotyping evidence (Ries 1932) indicates that the *C. columbae* genome consists of 12 holocentric chromosomes. Based on this physical evidence, and the striking difference in size between the 12 largest scaffolds and all other scaffolds in the assembly

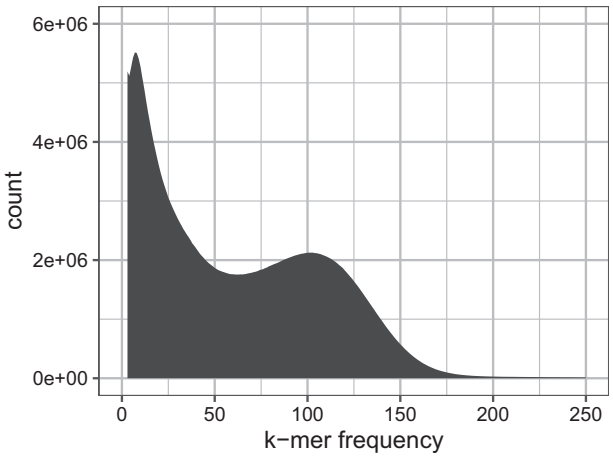

Figure 3 *Jellyfish*-derived (Marçais and Kingsford 2011) 21-mer histogram based on Illumina reads from the *C. columbae* genome.

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022

Table 1 Assembly and annotation statistics

Genome size	**208 Mb**
Illumina sequencing coverage	102
Oxford nanopore sequencing coverage	35
Pre-scaffolding contigs	—
Total number of contigs	1,193
Contig N50	511 kb
Contig N90	93 kb
Contig L50	103
Contig L90	466
Scaffolds	—
Chromosome-size scaffolds (≥ 12 Mb)	12
Total number of scaffolds	386
Scaffold N50	17.6 Mb
Scaffold N90	13.7 Mb
Scaffold L50	6
Scaffold L90	11
Annotation	—
Annotated genes	13362
Annotated transcripts	19140
Annotated genes (AED ≥ 0.5)	1972
Repeat content	9.70%
BUSCO score	96.4%

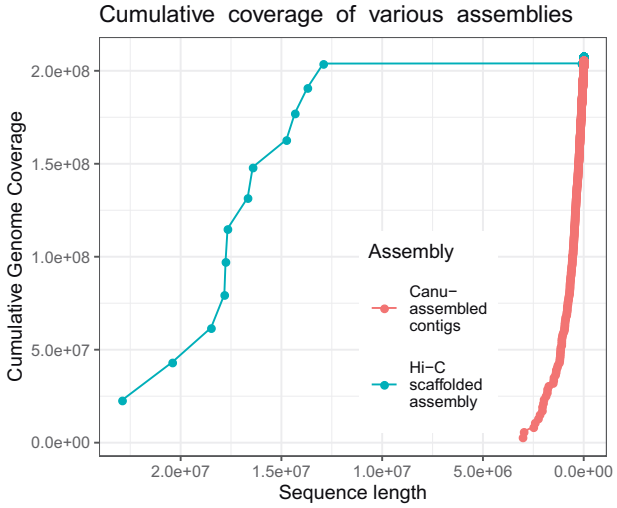

Figure 4 Cumulative coverage of initial and final (scaffolded) *C. columbae* genome assemblies, illustrating the improvement in contiguity by scaffolding with Hi-C data. All scaffolds in the assembly are plotted largest to smallest, from left to right. The x-axis indicates cumulative length of an assembly, and the y-axis corresponds to the cumulative portion of the genome covered by initial contigs (red dots) and final scaffolds (blue dots).

(Figure 4), we predict that each of the 12 largest scaffolds in the assembly represents one of the 12 karyotyped chromosomes.

Annotation

We annotated the genome using the *MAKER* pipeline, with transcriptome evidence from *Trinity*-assembled Illumina RNAseq reads and *wavy_choose*-selected Oxford Nanopore RNAseq reads (Figure 2). We identified 19,139 transcripts from 13,362 genes. 13,246 of these genes are functionally annotated by BLAST using the Swissprot database, 8354 are functionally annotated by similarity to InterPro or Pfam, and 13,248 are functionally annotated by either Swissprot, InterPro, or Pfam. *MAKER* produces a combined quality statistic called Annotation Edit Distance (AED); Eilbeck *et al.* 2009; Holt and Yandell 2011). Perhaps owing to our use of long-read transcriptome sequencing, 10.3% of our annotated transcripts have ideal AED scores of 0 (Figure 5), and only 5.6% of annotated transcripts have low-quality AED scores above 0.5. The abundance of low AED scores and relative dearth of low scores are indicators of a high-quality annotation.

Table 2 BUSCO results for genome completeness for the reference genome assembly, the annotated transcriptome, and the predicted proteome

Count	**Genome**	**Transcriptome**	**Proteome**
Complete, single-copy BUSCOs	1,593	1,440	1,438
Complete, duplicated BUSCOs	6	50	12
Fragmented BUSCOs	25	54	55
Missing BUSCOs	34	114	153
Complete BUSCOs (%)	96.44	89.86	87.45
Complete, single-copy BUSCOs (%)	96.07	86.85	86.73
Complete, duplicated BUSCOs (%)	0.36	3.01	0.72
Fragmented BUSCOs (%)	1.50	3.25	3.31
Missing BUSCOs (%)	2.05	6.87	9.22
Total BUSCO groups searched	1,658	1,658	1,658

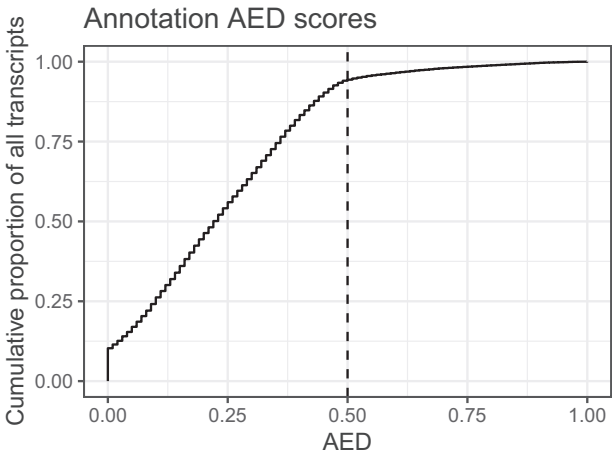

Figure 5 Cumulative annotation edit distance (AED) for all genes in the MAKER-derived annotation. 10.3% of genes have an AED of 0, while only 5.6% of genes have an AED above 0.5 (vertical dashed line).

Genome completeness

We used BUSCO (Simão *et al.* 2015) to measure completeness of the genome by counting the number of highly conserved, single copy genes that should be present in insects (Table 2). The reference genome, transcriptome, and translated transcriptome contain complete copies of 96%, 90%, and 87% of insect BUSCOs, respectively. These high values indicate that the *C. columbae* genome assembly is sufficiently complete for downstream comparative genomic analyses.

Repetitive elements

Repeatmasker identified 20.2 Mb (9.70%) of the genome as repetitive content. Of this 20.2 Mb, 65.8% is DNA transposons, 14.8% is LINEs, 8.6% is simple repeats, and 5.7% is LTR transposons (Wicker et al. 2007). The remainder (5.1%) is an assortment of transposable elements, low-complexity regions, and satellites (Table 3). Repetitive content in the *C. columbae* genome is, therefore, considerably higher than in *P. humanus*. In the latter species, only 1% of the genome is annotated as class I (retrotransposons, including LTR, LINE, and SINE) or class II (DNA transposons) transposable elements, and 6.9% is tandem repeats (Kirkness

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022

Table 3 Repetitive elements in the *C. columbae* genome

Identity	Number of bases	Percent of all bases	Percent of repetitive elements
DNA	13,283,184	6.39	65.9
LINE	2,980,785	1.43	14.8
Low_complexity	506,296	0.244	2.51
LTR	1,156,688	0.556	5.74
Other	684	0.000329	0.00339
RC	126,151	0.0607	0.626
Retroposon	749	0.000360	0.00371
RNA	16,720	0.00804	0.0829
rRNA	33,934	0.0163	0.168
Satellite	48,279	0.0232	0.239
Simple_repeat	1,745,924	0.840	8.660
SINE	112,911	0.0543	0.560
snRNA	24,864	0.0120	0.123
tRNA	54,873	0.0264	0.272
Unknown	66,922	0.0322	0.332

et al. 2010). One caveat to this conclusion is the lower contiguity of the earlier *P. humanus* assembly. Because genomes often fail to assemble at repetitive sites, the *P. humanus* assembly may have captured a smaller proportion of repetitive sequences than the more contiguous *C. columbae* assembly.

Kirkness *et al.* (2010) predicted that the monophagous, permanently parasitic lifestyle of lice should lead to reduced genomes due to the reduced need to seek food and avoid enemies compared to free-living species. While Kirkness et al. identified a reduction in gene families related to sensing, their conclusion that overall genome size is affected by lifestyle is not supported by the genome size of *C. columbae*, which has a genome size and number of genes that are more typical for a free-living insect. Indeed, both *C. columbae* and *P. humanus* appear to have a full complement of genes, and while *P. humanus* has a small genome and a reduction in transposable elements, *C. columbae* has neither of these. The pattern of reduced genome size and reduction in TE content without loss of genes is characteristic of high-population-size species (Lefébure *et al.* 2017). However, a robust estimate of the population size of *P. humanus*, combined with evidence ruling out alternative hypotheses, would be necessary to demonstrate that population size drove the reduced genome size in *P. humanus*. Other authors (Oliver *et al.* 2007) have hypothesized that large populations may not actually be under selection to have smaller genomes.

Genomic evidence for the lack of centromeres

Centromeres are characterized by a depletion of genic content and an increase in repetitive content (Jain *et al.* 2018). Based on these criteria (Figure 6), we find no evidence for centromeres in any of the *C. columbae* chromosomes. Presence of genes is moderately anti-correlated with presence of simple repetitive sequences ($r = -0.28$, 1 Mb sliding windows). Still, the overall repeat density is not correlated with gene density, and both measures are relatively consistent across the genome. Many chromosomes (cf., Figure 6, chromosomes 6 and 7) have a twin-peaked pattern of simple repeats, in which chromosome ends and centers have high genic content and low repeat content, but the genomic segments between the ends and the center have high repeat content and low-genic content. It is possible that these twin peaks of simple repeat content are the centromeres in a polycentromeric chromosome, and that the chromosomes were actually misclassified as holocentric based on karyotyping evidence.

Comparisons to the closest sequenced relative

The closest relative of *C. columbae* with an assembled genome is the human body louse *P. humanus*. *C. columbae*, and *P. humanus* are thought to have diverged 65 million years ago (Johnson *et al.* 2018). *Pediculus humanus* has five metacentric chromosomes and one telocentric chromosome (Kirkness *et al.* 2010), in contrast to the 12 putatively holocentric chromosomes described here. *Pediculus humanus* has a genome assembly size of 108 Mb, approximately half that of the 208-Mb *C. columbae* genome assembly. The *C. columbae* genome has a typical genome-wide GC content of 36%, while *P. humanus* has an extremely AT-rich genome with 28% GC content, making *C. columbae* the more typical insect genome of the two.

Synteny analysis

We used the default settings of *SynIma* (Farrer 2017) to identify synteny between *C. columbae* and *P. humanus* (Figure 7). We were unable to test for chromosome-scale syntenic blocks between *P. humanus* and *C. columbae* due to the low contiguity of the *P. humanus* genome. However, we found very few locations in which synteny is broken between a *P. humanus* scaffold and a *C. columbae* scaffold, showing that short-range synteny is almost entirely conserved between these species.

Functional annotation reveals depletion of environmental sensing and metabolic genes

Pediculus humanus has a small complement of opsins (3, as opposed to 275 in *D. melanogaster*) and G protein-coupled receptors (GPCR, 104, as opposed to 408 in *D. melanogaster*) (Kirkness *et al.* 2010; Thurmond *et al.* 2019). Similarly, we find that only 2 annotated genes in *C. columbae* are associated with the opsin gene ontology term (GO:00007602) and only 107 genes are associated with the GPCR GO category (GO:00004930). This reduced repertoire of sensory system genes supports the hypothesis that the relatively static environments encountered by lice and other ectoparasites relaxes selection on the ability to sense and respond to stimuli in more variable environments (Kirkness *et al.* 2010). *Columbicola columbae* is incapable of surviving off of its obligate host, so there might be little selection to retain complex visual, olfactory, or other complex sensory acuity. We find support for the hypothesis that specific gene families, such as those relating to sensory capabilities and metabolism, are reduced in obligate parasites (Jackson 2015).

Pediculus humanus is massively depleted in terms of odorant receptors, gustatory receptors, and chemosensory proteins, and

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022

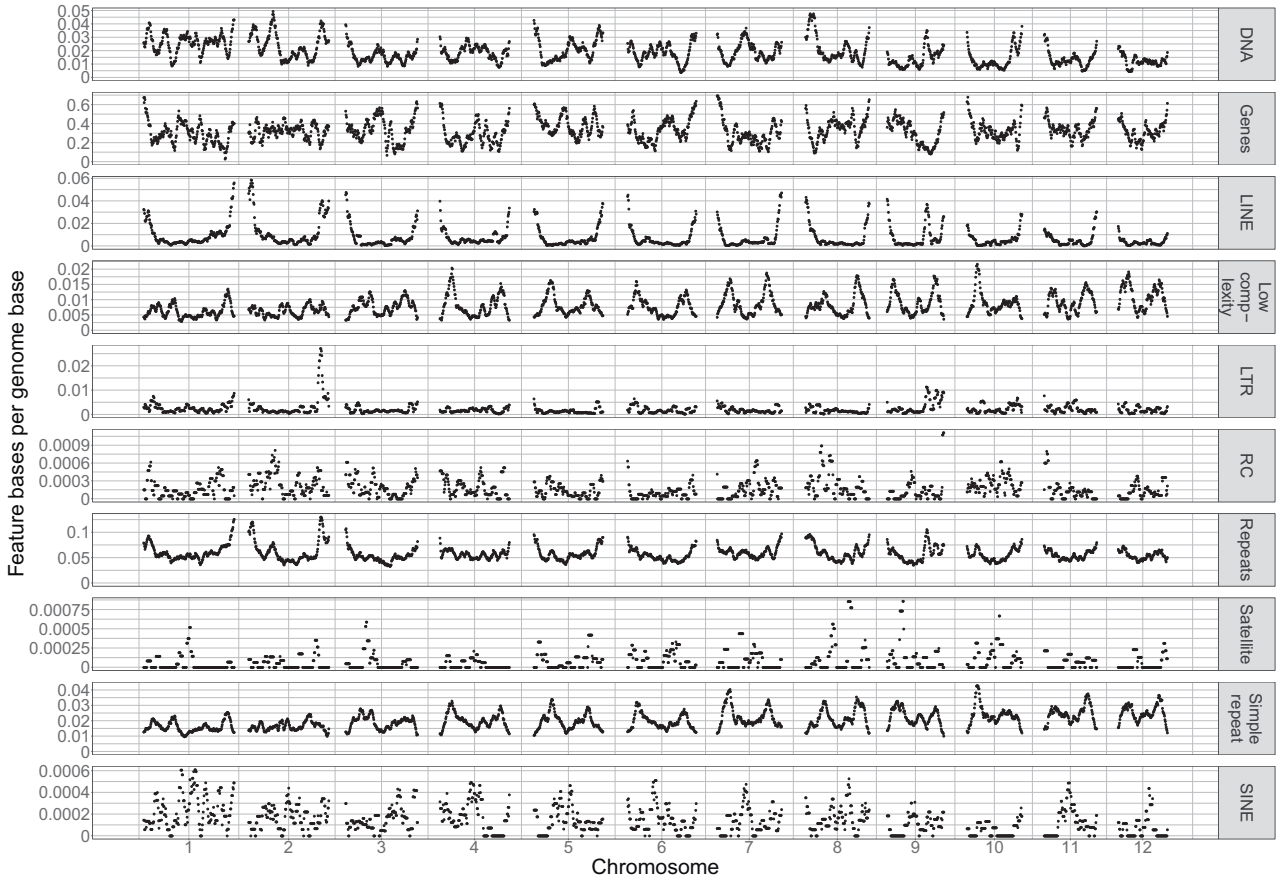

Figure 6 Chromosome-wide feature density in the *C. columbae* genome. Gene and repeat density in 1 Mb-wide sliding windows across the *C. columbae* genome show that there are no clear centromeres, and gene and simple repeat density are negatively correlated.

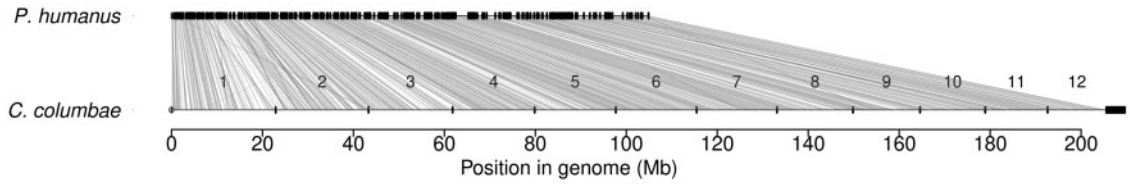

Figure 7 Short range synteny is largely conserved between *C. columbae* (bottom) and *P. humanus* (top) genomic scaffolds. Lines connecting scaffolds from each genome assembly represent the positions of orthologous genes. *P. humanus* contigs were aligned to the *C. columbae* genome in order and orientation using *SynIma*. Chromosome-size scaffolds in *C. columbae* are labeled 1–12.

C. columbae shows the same pattern. For example, *C. columbae* has only 13 genes with olfactory receptor activity (GO: 0004984) and *P. humanus* has only 10, compared with 152 in *D. melanogaster* (Thurmond *et al.* 2019). *Columbicola columbae* has 2 genes associated with taste receptor activity (GO: 0008527) and *P. humanus* has 6, yet *D. melanogaster* has 150. We speculate that this dramatic depletion of taste receptor genes is due to the homogeneous diet of ectoparasitic lice. The diet of *C. columbae*, for instance, consists entirely of pigeon feathers and flakes of dead skin (Ash 2008; Nelson and Murray 1971; Singh *et al.* 2010).

Another highly depleted gene functional category in *P. humanus* is the insulin signaling/TOR pathway. Kirkness *et al.* (2010) show that the canonical pathway appears nonfunctional in *P. humanus*, with only one gene having *P. humanus* EST-derived evidence for its expression. BLAST evidence indicates that other TOR pathway genes are reduced to a single copy in *P. humanus*, including genes such as *dilps* and *eIF-4E* (class I), which respectively have 6 and 7 copies in *D. melanogaster* (Kirkness *et al.* 2010). We find the same qualitative result in *C. columbae*, with no annotated genes associated with the insulin receptor signaling pathway (GO:0008286). Finally, the complement of detoxification genes is depleted in both *P. humanus* and *C. columbae*, with *C. columbae* having no annotated genes associated with detoxification (GO:0098754).

The striking reduction in sensory and metabolic gene categories in *C. columbae* and *P. humanus* could be due to independent gene loss in each lineage, inheritance of a depleted repertoire from a common ancestor, or a combination of the two. Loss of the same suite of genes in each species would be consistent with inheritance of a reduced sensory repertoire from a common ancestor, while loss of different genes in each species would indicate independent reductions. Reciprocal best BLAST hits of *C. columbae* and *P. humanus* genes to a shared outgroup, *Drosophila melanogaster*, indicate that the identities of the lost and retained

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022

Table 4 Reciprocal best-hit BLAST of the proteomes of *C. columbae* and *P. humanus* against *D. melanogaster* reveals the identity of the retained genes in depleted gene families is largely the same in both species

Gene family	*D. melanogaster* genes	*C. columbae* hits	*P. humanus* hits	Shared hits
Opsin	275	40	40	28
GPCR	408	66	69	50
Olfactory receptor activity	152	6	8	6
Taste	150	4	3	3
Odorant binding	248	6	7	4
Insulin	349	59	61	46
Tor	225	51	57	47
Chemosensory behavior	441	54	57	44
Detoxification	132	16	18	16

The first column is the tested family of genes. The second column is the number of genes assigned the corresponding GO term in the *D. melanogaster* proteome. The third and fourth columns, respectively, are the numbers of reciprocal best BLAST hits with *D. melanogaster* genes by genes from either *C. columbae* or *P. humanus*. The fifth column is the number of reciprocal best BLAST hits that had the same *D. melanogaster*-derived identity when BLASTing against *C. columbae* or *P. humanus*.

genes are mostly the same between the two louse species (Table 4), thereby supporting the hypothesis of ancestral loss. We note the possibility that these "missing" genes are not actually absent from the genomes of *C. columbae* and *P. humanus*, but are simply not annotated in their respective genomes. However, the BUSCO completeness score of 96.4% for the *C. columbae* genome renders large-scale incompleteness and misannotation less likely.

Acknowledgments

We thank Juan C. Altuna for assistance in creating the louse inbreeding protocol and for maintenance of the captive populations of pigeons and lice. We gratefully acknowledge support and resources from the Center for High Performance Computing at the University of Utah. We also thank Mark Yandell for providing computational resources, and Carson Holt for technical advice and assistance with MAKER.

Funding

This work was funded by the National Science Foundation Dimensions of Biodiversity grants NSF DEB-1342604 and DEB-1342600.

Conflicts of interest: None declared.

Literature cited

Adly E, Nasser M, Soliman D, Gustafsson DR, Shehata M. 2019. New records of chewing lice (Phthiraptera: Amblycera, Ischnocera) from Egyptian pigeons and doves (Columbiformes), with description of one new species. Acta Trop. 190:22–27.

Aronesty E. 2011. Comparison of Sequencing Utility Programs. The Open Bioinformatics Journal. 7.

Ash JS. 2008. A study of the mallophaga of birds with particular reference to their ecology. Ibis. 102:93–110.

Belton J-M, McCord RP, Gibcus J, Naumova N, Zhan Y, *et al.* 2012. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 58:268–276.

Bennett S. 2004. Solexa Ltd. Pharmacogenomics. 5:433–438.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, *et al.* 2017. Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. Nat Genet. 49:643–650.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, *et al.* 2013. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol. 31:1119–1125.

Bush S, Kim D, Reed M, Clayton D. 2010. Evolution of cryptic coloration in ectoparasites. Am Nat. 176:529–535.

Bush SE, Clayton DH. 2006. The role of body size in host specificity: reciprocal transfer experiments with feather lice. Evolution. 60: 2158–2167.

Bush SE, Price RD, Clayton DH. 2009. Descriptions of eight new species of feather lice in the genus *Columbicola* (Phthiraptera: Philopteridae), with a comprehensive world checklist. J Parasitol. 95:286–294.

Bush SE, Sohn E, Clayton DH. 2006. Ecomorphology of parasite attachment: experiments with feather lice. J Parasitol. 92:25–31.

Bush SE, Villa SM, Altuna JC, Johnson KP, Shapiro MD, *et al.* 2019. Host defense triggers rapid adaptive radiation in experimentally evolving parasites. Evol Lett. 3:120–128.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, *et al.* 2007. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188–196.

Chao K-M, Pearson WR, Miller W. 1992. Aligning two sequences within a specified diagonal band. Bioinformatics. 8:481–487.

Chen N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinform. 5:4.10.1–4.10.14.

Clayton DH. 1990. Mate choice in experimentally parasitized rock doves: lousy males lose. Am Zool. 30:251–262.

Clayton DH. 1991. Coevolution of avian grooming and ectoparasite avoidance. In: JE Loye and M. Zuk, editors. Bird–Parasite Interactions: Ecology, Evolution and Behaviour, Oxford Univ. Press, London. Vol. 14. p. 258–289.

Clayton DH, Adams RJ, Bush SE. 2008. Parasitic Diseases of Wild Birds. In: Carter T. Atkinson, Nancy J. Thomas, and D. Bruce Hunter, editors. Phthiraptera, the chewing lice. Hoboken, NJ: John Wiley & Sons. p. 515–526.

Clayton DH, Bush SE, Goates BM, Johnson KP. 2003. Host defense reinforces host–parasite cospeciation. Proc Natl Acad Sci USA. 100:15694–15699.

Clayton DH, Bush SE, Johnson KP. 2015. Coevolution of Life on Hosts: Integrating Ecology and History. Chicago, IL: University of Chicago Press.

Clayton DH, Lee PLM, Tompkins DM, Brodie ED, III. 1999. Reciprocal natural selection on host–parasite phenotypes. Am Nat. 154: 261–270.

Clayton DH, Tompkins DM. 1995. Comparative effects of mites and lice on the reproductive success of rock doves (*Columba livia*). Parasitology. 110:195–206.

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022

Darwin C. 1868. The Variation of Animals and Plants Under Domestication, Vol. 1. New York, NY: Orange Judd & Co.
de Meeûs T, Renaud F. 2002. Parasites within the new phylogeny of eukaryotes. Trends Parasitol. 18:247–251.
Du P, Kibbe WA, Lin SM. 2006. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics. 22:2059–2065.
Durden LA, Musser GG. 1994. The sucking lice (Insecta, Anoplura) of the world: a taxonomic checklist with records of mammalian hosts and geographical distributions. Bull Am Museum Nat History. 218:6–77.
Eichler W, Zlotorzycka J, Ludwig L, Wolfgang H, Stenram H. 1972. The pigeon louse *Columbicola columbae*. Angewandte Parasitol. 13: 1–18.
Eilbeck K, Moore B, Holt C, Yandell M. 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics. 10:67.
Farrer RA. 2017. Synima: a Synteny imaging tool for annotated genome assemblies. BMC Bioinformatics. 18:507.
Fukatsu T, Koga R, Smith WA, Tanaka K, Nikoh N, *et al.* 2007. Bacterial endosymbiont of the slender pigeon louse, *Columbicola columbae*, allied to endosymbionts of grain weevils and tsetse flies. Appl Environ Microbiol. 73:6660–6668.
Gibbs D, Barnes E, Cox J. 2001. Pigeons and Doves: A Guide to the Pigeons and Doves of the World, Vol. 13. London: A&C Black.
Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, *et al.* 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 29:644–652.
Grossmann A, Morlet J. 1984. Decomposition of hardy functions into square Integrable wavelets of constant shape. SIAM J Math Anal. 15:723–736.
Gustafsson DR, Tsurumi M, Bush SE. 2015. The chewing lice (Insecta: Phthiraptera: Ischnocera: Amblycera) of Japanese pigeons and doves (Columbiformes), with descriptions of three new species. J Parasitol. 101:304–313.
Harbison CW, Clayton DH. 2011. Community interactions govern host-switching with implications for host–parasite coevolutionary history. Proc Natl Acad Sci USA. 108:9525–9529.
Holt C, Campbell M, Keays DA, Edelman N, Kapusta A, *et al.* 2018. Improved genome assembly and annotation for the rock pigeon (*Columba livia*). G3 (Bethesda). 8:1391–1398.
Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 12:491.
Jackson AP. 2015. The evolution of parasite genomes and the origins of parasitism. Parasitology. 142:S1–S5.
Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17:239.
Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, *et al.* 2018. Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol. 36:321–323.
Johnson KP, Malenke JR, Clayton DH. 2009. Competition promotes the evolution of host generalists in obligate parasites. Proc Biol Sci. 276:3921–3926.
Johnson KP, Nguyen N-P, Sweet AD, Boyd BM, Warnow T, *et al.* 2018. Simultaneous radiation of bird and mammal lice following the K-Pg boundary. Biol Lett. 14:20180141.
Johnson KP, Reed DL, Hammond Parker SL, Kim D, Clayton DH. 2007. Phylogenetic analysis of nuclear and mitochondrial genes supports species groups for *Columbicola* (Insecta: Phthiraptera). Mol Phylogenet Evol. 45:506–518.
Jones P, Binns D, Chang H-Y, Fraser M, Li W, *et al.* 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236–1240.
Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, *et al.* 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci USA. 107:12168–12173.
Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al.* 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27:722–736.
Lefébure T, Morvan C, Malard F, François C, Konecny-Dupré L, *et al.* 2017. Less effective selection leads to larger genomes. Genome Res. 27:1016–1028.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.
Liu B, Shi Y, Yuan J, Hu X, Zhang H, *et al.* 2020. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. arXiv:1308.2012 [q-Bio].
Marchet C, Lecompte L, Silva CD, Cruaud C, Aury J-M, *et al.* 2019. De novo clustering of long reads by gene from transcriptomics data. Nucleic Acids Res. 47:e2.
Marshall AG. 1981. The Ecology of Ectoparasitic Insects. Cambridge, MA: Academic Press.
Martin M. 1934. Life history and habits of the pigeon louse (*Columbicola columbae* [Linnaeus]). Can Entomol. 66:6–16.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27: 764–770.
Mullen GR, Durden LA. 2009. Medical and Veterinary Entomology. Cambridge, MA: Academic Press.
Nelson BC, Murray MD. 1971. The distribution of mallophaga on the domestic pigeon (*Columba livia*). Int J Parasitol. 1:21–29.
Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, *et al.* 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics. 28:1919–1920.
Oliver MJ, Petrov D, Ackerly D, Falkowski P, Schofield OM. 2007. The mode and tempo of genome size evolution in eukaryotes. Genome Res. 17:594–601.
Peichel CL, Sullivan ST, Liachko I, White MA. 2017. Improvement of the threespine stickleback genome using a Hi-C-based proximity-guided assembly. J Hered. 108:693–700.
Rakshpal R. 1959. On the behaviour of pigeon louse, *Columbicola columbae* Linn. (Mallophaga). Parasitology. 49:232–241.
Ries E. 1932. Die prozesse der eibildung und des eiwachstums bei pediculiden und mallophagen. ZZellforsch. 16:314–388.
Rudolph D. 1983. The water-vapour uptake system of the phthiraptera. J Ins Physiol. 29:15–25.
Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, *et al.* 2017. Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27:849–864.
Shapiro MD, Domyan ET. 2013. Domestic pigeons. Curr Biol. 23: R302–R303.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31: 3210–3212.
Singh SK, Arya S, Singh SK, Khan V. 2010. Feeding and reproductive behaviour of pigeon slender louse, *Columbicola columbae* (Phthiraptera, Insecta, Ischnocera). J Appl Nat Sci. 2:126–133.
Smith VS. 2004. The chewing lice: world checklist and biological overview. Syst Biol. 53:666–668.

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022

Smith WA, Oakeson KF, Johnson KP, Reed DL, Carter T, *et al.* 2013. Phylogenetic analysis of symbionts in feather-feeding lice of the genus Columbicola: evidence for repeated symbiont replacements. BMC Evol Biol. 13:109.

Stenram H. 1956. The ecology of *Columbicola columbae* L. (Mallophaga). Opusculata Entomol. 21:170–190.

Sweet AD, Boyd BM, Allen JM, Villa SM, Valim MP, *et al.* 2018. Integrating phylogenomic and population genomic patterns in avian lice provides a more complete picture of parasite evolution. Evolution. 72:95–112.

Sweet AD, Johnson KP. 2018. The role of parasite dispersal in shaping a host–parasite system at multiple evolutionary scales. Mol Ecol. 27:5104–5119.

Sweet AD, Johnson KP, Cao Y, de Moya RS, Skinner RK, *et al.* 2021. Structure, gene order, and nucleotide composition of mitochondrial genomes in parasitic lice from Amblycera. Gene. 768: 145312.

Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, *et al.* 2019. FlyBase 2.0: the next generation. Nucleic Acids Res. 47: D759–D765.

Urban JM, Bliss J, Lawrence CE, Gerbi SA. 2015. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. bioRxiv. 019281.

van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, *et al.* 2010. Hi-C: a method to study the three-dimensional architecture of genomes. J Vis Exp. e1869. [10.3791/1869]

Villa SM, Altuna JC, Ruff JS, Beach AB, Mulvey LI, *et al.* 2019. Rapid experimental evolution of reproductive isolation from a single natural population. Proc Natl Acad Sci USA. 116:13440–13445.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, *et al.* 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, *et al.* 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.

Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15: R46.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol. 7:203–214.

Communicating editor: J. Udall

Downloaded from https://academic.oup.com/g3journal/article/11/2/jkab009/6114464 by University of Mount Olive - Moye Library user on 28 July 2022