2023

A standardized pipeline for isolation and assembly of genomes from symbiotic bacteria in whole louse genomic sequence data.

Mohammad Mikail I. Bala
Virginia Commonwealth University

A standardized pipeline for isolation and assembly of genomes from symbiotic bacteria in whole louse genomic sequence data.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

By

Mohammad Mikail Bala

M.S. Bioinformatics – Virginia Commonwealth University

Mentor:

Bret M. Boyd, Ph.D.

Assistant Professor, Center for Biological Data Science

Virginia Commonwealth University

Richmond, Virginia

March 2023

Acknowledgements

I'd like to thank:

- Dr. Allison Johnson, my graduate advisor, who encouraged me to pursue my master's thesis, and for her advice and encouragement.
- Dr. Bret Boyd, my mentor, for his patience and encouragement. Throughout the years, he had been generous with his editorial advice and enviable erudition. A big thanks to him for the opportunity to develop my programming skills, for providing wisdom along with a sharp critical eye, as well as for rallying round to help me through the hectic final months.
- Dr. Michael Rosenberg, for providing indispensable feedback on my work and teaching the fundamentals of programming, which were essential to completing my work.
- My mom, for her unwavering love. I would have given up long ago without her steady support.
- My dad, for his endless patience, as he went through a lot and did a lot
- Faizal Surti, my uncle, for his support throughout, wisdom, and advice
- Latika Meena and Musaddiq Lodi, for being true friends who kept me going despite my frequent sidetracks.

Table of Contents

List of Abbreviations

Abbreviation	Meaning
AT	Adenine and Thymine nucleotide bases
Blastn	Basic Local Alignment Search Tool for Nucleotides
C	sequencing depth coverage
CDS	Protein coding DNA sequences
CWE	Columbicola wolffhuegeli endosymbiont
DNA	Deoxyribonucleic Acid
fusA	Elongation factor G gene
gDNA	Genome Deoxyribonucleic Acid
GL	Genome length
HiSeq	Illumina Hight-throughput Sequencing platform
L	Length of whole genome sequencing reads
N50	sequence length of the shortest contig at 50% of the total assembly length
NCBI	National Center for Biotechnology Information
NovaSeq	efficient and cost-effective whole-genome sequencing platform by Illumina
ORFs	Open Reading Frames
PGR	Percent genome recovery
RC	Correct recovered contigs or true positive contigs
SdBG	succinct de Bruijn graph
WGS	Whole genome sequencing

List of Figures

Abstract

Many insects are known to harbour intracellular and heritable bacteria (endosymbionts), which provide their hosts with adaptive traits. Whole insect gDNA shotgun sequencing projects often sequence the genome of endosymbiont, in addition to the insect's genome. There are approximately 600 whole genome shotgun libraries from insects available on the public repository (NCBI), which can be mined to obtain endosymbiont genomes. The assembly and annotation of endosymbiont genomes can contribute towards the exploration of their role as obligate symbiotic partners. However, *de novo* assembly of an endosymbiont genome, continues to be challenging, when the host and/or enteric bacterial gDNA is present in the library as well. So far, whole genome sequence data has been mined by investigators, who manually interrogate the data at multiple steps, a process that is time consuming and difficult to replicate. Here I developed and evaluated a novel strategy that reduces intervention required by the researcher. The strategy consists of two steps: 1) filtering of *de novo* assembled endosymbiont contigs using Blastn search against a custom reference database and 2) reconstruction of the genome through *de novo* assembly of reads associated with filtered contigs. Illumina HiSeq libraries were simulated *in silico* and the pipeline was deployed using the simulated data to test the efficacy of the method. The mean endosymbiont genome recovery from simulated data was 91.27% with a range of 100%-76%. When the method was tested with "real" whole louse shotgun sequencing libraries obtained from a public repository, the results were mixed. The strategy was accurate in contig selection when the louse contained an endosymbiont which had a small genome enriched for AT bases, with a mean percent genome recovery of 98.38% and the range of 100% - 95.88%. However, in other cases involving symbionts with larger genomes, the resulting genomes appeared to be incomplete and require further evaluation.

Introduction

Approximately 10% to 20% insect species are host to heritable microbial partners known as host-beneficial endosymbionts (Buchner, 1965). These bacteria provide a wide range of beneficial functions for their hosts, including defence against predators (Oliver et al., 2003, 2009; Nakabachi et al., 2013; Kaltenpoth et al., 2005), insecticide resistance (Kikuchi et al., 2012), and stress resistance (Heyworth et al., 2016). However, most endosymbionts provide their hosts with nutrients absent in their host's diet (Puchta et al., 1955; Eberle et al., 1982, 1983; Douglas, 1989; Perotti et al., 2007; Fukatsu et al., 2007; McCutcheon et al., 2019; Alickovic et al., 2021). Numerous insect species are known to harbour endosymbiotic bacteria in specialized cells called bacteriocytes (Buchner, 1965). While it is technically possible to extract endosymbionts from bacteriocytes, their *in vitro* culture has proven to be challenging (Nadal-Jimenez et al., 2019; Masson et al., 2020). The inability of endosymbionts to synthesize many essential metabolites and dependence on the host cell can hinder their growth in culture (Shigenobu et al., 2000; Moya et al., 2008; McCutcheon et al., 2012; Moran et al., 2014; Chevignon et al. 2018), limiting our ability to examine the diversity of endosymbionts and the services they provide to the host. Fortunately, insect whole genome sequencing (WGS) projects often capture the genome of their respective endosymbionts, providing an opportunity to study the diversity of insect-endosymbiont relationships and infer function and phylogenetic relationships.

Isolation and assembly of the endosymbiont genomes from insect WGS projects can be a time consuming process that is potentially difficult to replicate. Numerous short reads obtained from whole louse shotgun sequencing projects are assembled *(in silico)* to produce contiguous sequences (contigs) of the endosymbiont genome (*de novo* assembly), host genome, and contaminate genomes of varying lengths and quality. The presence of these unwanted contigs necessitate manual parsing to identify and isolate the endosymbiont contigs. While insects and bacteria develop obligate symbiotic relationships that are sustained over evolutionary time scales (Douglas, 1998; McCutcheon et al., 2012;Moran et al., 2014), insects have repeatedly acquired endosymbionts from non-endosymbiotic progenitors, and the genome of endosymbionts may vary dramatically in both size and composition (Smith et al., 2013; Sudakaran et al., 2017). Therefore, we may not know the phylogenetic position or genomic makeup of an endosymbiont prior to the start of an assembly, further complicating the process of genome isolation.

In this study I focused on a novel approach for streamlining and automating the assembly of endosymbiont genome from whole insect gDNA libraries and used parasitic lice (Insecta; Phthiraptera). Lice were selected as they provide a simple test to the novel approach and develop tools. The benefits of using lice includes, (1) lice comparatively have smaller genomes than most other insects, leading to manageable size datasets (e.g., *Pediculus humanus* (human louse) has a genome that is 110.78 Mb in length (Kirkness et al., 2010), whereas *Magicicada septendecim* (seventeen-year cicada) is ~1.12 Gb (Du et al., 2019)); (2) most lice contain a single endosymbiont species (e.g., the human louse is a host to *Ca*. Riesia Pediculicola (Kirkness et al., 2010)); (3) lice are wingless ectoparasites that complete their entire life cycle on the host, meaning environmental variables (e.g., temperature) are stable.

Lice are divided into over 4500 species of chewing lice (Ischnocera and Amblycera) and approximately 500 species of blood sucking lice (Anoplura) (Puchta et al. 1955; Durden et al., 1994; Price et al., 2003). Feather lice (Ischnocera), a subgroup of chewing lice, comprise over 3000 species (Price et al. 2003; Johnson et al., 2012). Blood sucking lice and their endosymbionts are well characterized (Puchta et al., 1955; Eberle et al., 1982, 1983; Perotti et al., 2007; Kirkness et al., 2010; Boyd et al., 2014, 2016, 2017), however, much less is known about the symbiotic bacterial partners of the closely related feather (wing) lice (Fukatsu et al., 2007; Alickovic et al., 2021). Previous studies have demonstrated the presence and inheritance of endosymbiotic bacteria in pigeon lice, *Columbicola* sp. (Insecta: Phthiraptera) and the results reveal that pigeon louse species carry one of three different bacterial endosymbionts (Ries et al., 1931; Fukatsu et al., 2007; Smith et al., 2013). The biological roles and molecular functions of these bacteria as obligate intracellular symbionts are relatively less known and there is a lack of an automated computational method that could be used to obtain complete or even near to complete endosymbiont genomes in isolation.

As mentioned above, the process of endosymbiont genome isolation and assembly from a WGS library is commonly carried out with manual and iterative processes. (Perotti et al., 2007; Kirkness et al., 2010; Boyd et al. 2014, 2016, 2017; Alickovic et al. 2021, Park et al., 2021; Říhová et al., 2022). During the development of my thesis research, a study was published reporting complete genome assemblies of endosymbionts, *Rickettsia and 'Candidatus Megaira',* using a standardized pipeline, including isolates from lice (Davison et al., 2022). In order to assess the viability of the aforementioned pipeline and its potential to overlap with my objectives, I evaluated their approach using whole louse genomic data. However, within my test, the approach yielded incomplete endosymbiont assembly, given my dataset. The approach was not evaluated in additional datasets.

Here I present a novel method that can be used to produce an endosymbiont genome in isolation using the gDNA sequence libraries generated from whole insects. The following steps were implemented to replace the manual processes involved in the genome assembly process, including parsing and isolation of target endosymbiont genome from the pool of contigs and filtering to extract selected reads of interest and assembly in isolation. The method was named ***Co***ntig filtering method based on ***B***lastn out format 6 metadata and ***R***econstruction through a *de novo* ***A***ssembly (CoBRA) and provides a critical test of methods that could be translated into an automated process.

Methods

Strategy to produce an endosymbiont genome in isolation:

The method developed in this study has 7 steps (Figure 1): 1) whole louse genomic DNA reads were either obtained from NCBI, or simulated; 2) genomic DNA reads were filtered through a quality control step to exclude any questionable base calls; 3) reads were assembled *de novo*; 4) the resulting set of *de novo* assembled contigs were searched against a custom nucleotide database using blastn to identify candidate endosymbiont contigs; 5) the results of the search were parsed to identify candidate endosymbiont contigs; 6) the candidate endosymbionts contigs and associated reads were isolated; 7) and the reads were assembled in isolation, using existing *de novo* assembly packages. Each step is discussed in detail below.

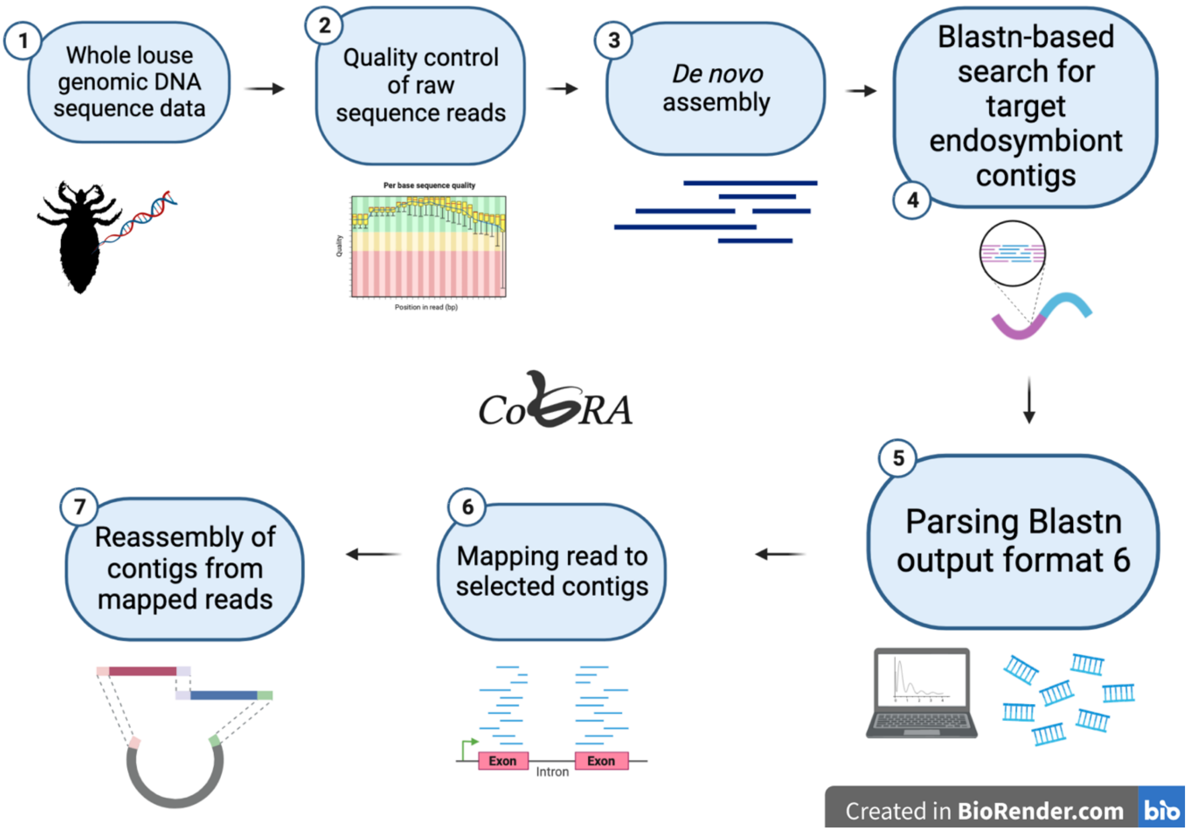
1
Whole louse genomic DNA sequence data
2
Quality control of raw sequence reads
Per base sequence quality
Quality
Position in read (bp)
3
De novo assembly
Blastn-based search for target endosymbiont contigs
4
CoBRA
5
Parsing Blastn output format 6
6
Mapping read to selected contigs
Exon
Intron
Exon
7
Reassembly of contigs from mapped reads
Created in BioRender.com bio

Figure 1. Outline of endosymbiont genome assembly procedure using the CobRA method.

Step 1, simulated data: Insect whole genome sequencing projects not only sequence the genome of the insect, but also associated microbes, which may include endosymbionts (Smith et. al 2013; Boyd et. al 2014, 2016). When high-throughput technologies, such as Illumina HiSeq or NovaSeq are used, the resulting sequence libraries often sequence the endosymbiont genome along with the whole insect genome. Additionally, sequencing depth may be sufficient to fully assemble a draft genome of the endosymbiont. In order to develop a tool that could be used to isolate an endosymbiont genome from a WGS library (whole insect reads), it was necessary to simulate HiSeq data that would emulate whole insect genome sequencing results, while allowing for the assemblies to be evaluated. Therefore, simulated libraries were used to test and evaluate the method's performance, and the results were used as feedback to improvise the method further.

Simulated libraries included a host insect (represented by one of two louse species, for which genome sequences were available, PRJNA16223, PRJNA662097; Kirkness et al., 2010; Baldwin-Brown et al. 2021), endosymbiont (multiple species were selected to capture the diversity of known endosymbionts previously identified in lice; CP001085.1, CP006569.1, CP038613.1, CP010907.1; Kirkness et al., 2010; Johnson et al. 2007; Smith et al., 2013; Boyd et al., 2016, 2017; Alickovic et al., 2021), and additional sources of DNA, such as bacteria present in the insect's gut or on its surface. *Burkholderia* species have been detected in several whole louse sequencing projects and therefore (Per. Com. B. M. Boyd), a species of *Burkholderia* (GCA_002223275.1) was used (table 1) to emulate a surface or enteric source of DNA. Libraries were simulated using InSilicoSeq v. 1.5.4 (flags --model hiseq --cpus 64; Gourlé et al., 2019), yielding multiple simulated libraries that reflected known louse-endosymbiont pairs (Kirkness et al., 2010; Johnson et al. 2007; Smith et al., 2013; Boyd et al., 2016, 2017; Alickovic et al., 2021), described below. InSilicoSeq requires the exact amount of reads to be simulated from a specific genome. The number of reads were quantified using the following formulae: (C * GL) / L; where C is Sequencing Depth Coverage, GL is Genome length, and L is the length of reads. The numeric values pertaining to coverage, genome length and length of reads, i.e., all the required values to calculate the number of reads, were based on genome size and expected sequencing depth. The genomes constituting each specific library were simulated in isolation and then the reads were combined to create libraries. InSilicoSeq also incorporates substitution, insertion, and deletion errors and its default error model was automatically applied to all the simulated reads.

In each simulated library, the species included were chosen specifically to imitate known host-endosymbiont pairs, in both blood sucking and feather feeding lice (Allen et al., 2007; Novakova et al., 2009; Kirkness et al., 2010; Smith et al., 2013; Boyd et al., 2016, 2017; Kashkouli et al., 2021, 2021; Říhová et al., 2022). Many blood sucking lice are suspected of having smaller genomes when compared to other insect species (e.g. *P. humanus*, genome size is 110.78 Mb; Kirkness et al., 2010) and are host to endosymbionts with small AT-rich genome (0.7 Mb; Kirkness et al., 2010; Boyd et al., 2014, 2017; Říhová et al., 2022). The first library was designed to reflect a known host-endosymbiont pair of a blood sucking lice and an endosymbiont with a highly reduced genome, *P. humanus* and *Ca.* Riesia pediculicola (Kirkness et al., 2010). Conversely, some feather lice have comparatively larger genomes (e.g. *C. columbae*, genome size is 207.89 Mb; Baldwin-Brown et al. 2021) and are host to a range of bacterial clades including *Sodalis* sp. (Fukatsu et al., 2007; Smith et al., 2013). At the moment, it was not possible to simulate the real endosymbiont pair *C. columbae* and *Sodalis* sp*.,* as the *Sodalis* endosymbiont genome is not available on a public repository; therefore, in order to represent a similar host-endosymbiont pair through a simulated library, the "ccol_sp_bc" library (representing *Sodalis* sp. in feather louse) was simulated using *Sodalis praecaptivus,* a closely related species with a fully sequenced and published genome (Chari et al*.,* 2015; Clayton et al., 2012; Chrudimský et al., 2012; Snyder et al., 2011). Four other libraries were simulated to account for other possible real host-endosymbiont pairs: "Ph_sp_bc" (representing *Sodalis sp.* in blood sucking louse; as demonstrated by Boyd et al. 2016), "Ph_a_bc" (representing *Arsenophonus sp.* in blood sucking louse), "Ccol_pantoae_bc" (representing *Pantoea sp.* in feather louse) and lastly, "Ccol_a_bc" (representing *Arsenophonus sp.* in feather louse) (table 1).

Table 1. Simulated insect whole genome HiSeq libraries and associated data.

Library name	**Host name & genome length**	**Endosymbiont name & genome length**	**Contaminate name & genome length**	**Total simulated reads**
Ph_r_bc	*P. humanus corporis* (110.78 Mb)	*Ca.* Riesia pediculicola USDA (0.58 Mb)	*B. cepacia* (8.2 Mb)	7,231,103
Ph_sp_bc	*P. humanus corporis* (110.78 Mb)	*S. praecaptivus HS1* (5.16 Mb)	*B. cepacia* (8.2 Mb)	10,467,659
Ph_a_bc	*P. humanus corporis* (110.78 Mb)	*A. nasoniae* (4.9 Mb)	*B. cepacia* (8.2 Mb)	11,699,016
Ccol_p_bc	*C. columbae* (207.89 Mb)	*Ca.* Pantoea carbekii (1.2 Mb)	*B. cepacia* (8.2 Mb)	68,501,894
Ccol_a_bc	*C. columbae* (207.89 Mb)	*A. nasoniae* (4.9 Mb)	*B. cepacia* (8.2 Mb)	68,160,696
Ccol_sp_bc	*C. columbae* (207.89Mb)	*S. praecaptivus HS1* (5.16 Mb)	*B. cepacia* (8.2 Mb)	66,929,339

"Ph" = *Pediculus humanus corporis*. "Ccol" = *Columbicola columbae*. "r" = *Candidatus* Riesia pediculicola USDA. "sp" = *Sodalis praecaptivus*. "a" = *Arsenophonus nasoniae*. "p" = *Candidatus* Pantoea carbekii. "bc" = *Burkholderia cepacia*.

In addition to simulated libraries, I sought to evaluate the method using data generated from the whole louse gDNA sequencing projects (Boyd et al. 2017). In total, eleven "real" gDNA libraries were selected, including three blood sucking and eight feather (wing) louse libraries. The three blood sucking lice species, *Pediculus schaeffi* species (chimpanzee louse), *P. humanus* (Human head louse), and *Pedicinus badii*, for all of which, already have their respective endosymbiont's genome available on public repositories; namely, *Ca.* Riesia pediculischaeffi, *Ca.* Riesia pediculicola and an undescribed bacterial gammaproteobacterial species with a small AT-rich genome (24-32% GC), respectively. The feather lice (*Columbicola* species) are suspected of containing endosymbionts closely related to *Enterobacter* species (Smith et al., 2013) (Table 2). The Data was obtained directly from the short read archive (https://www.ncbi.nlm.nih.gov/sra) (table 2).

Table 2. Illumina HiSeq 2500 libraries obtained from NCBI for evaluation

Sample library name	Host Genus Sp.	SRA sample identifier	Number of raw reads
Pdhum	*Pediculus humanus*	SRX2405460	83,689,331
Pdscf	*Pediculus schaeffi*	SRX390495	64,020,102
Pnbad	*Pedicinus badii*	SRX2609261	24,549,545
Coalt	*Columbicola altamimiae*	SRS1284168	56,593,624
Cowil	*Columbicola eowilsoni*	SRS1284129	88,862,786
Cokoo	*Columbicola koopae*	SRS1284131	109,164,924
Comac1	*Columbicola macrourae 1*	SRS1284175	60,041,496
Comas	*Columbicola masoni*	SRS1284153	58,281,528
Cothe	*Columbicola theresae*	SRS1284174	69,941,006
Cotri	*Columbicola triangularis*	SRS1284126	121,706,960
Cowom	*Columbicola wombeyi*	SRS1284146	52,700,590

Step 2, quality control: Both simulated and real libraries were checked for the presence of adapters and low quality regions using FastQC v0.11.9 (Andrews, 2010). FastQC reported an average quality score of 30 (q30) or greater and did not detect any adapters within the simulated libraries. The reported base quality was above q30 on average for all the *Columbicula* libraries with minimal TruSeq adapter content. Quality control was performed via Trimmomatic v0.32. (Bolger et al., 2014) using the flags: LEADING: 20 TRAILING:20 (trims low quality bases on either ends if the base quality is below 20), SLIDINGWINDOW:20:20 (implements a the sliding window, trimming once the average quality within the window of 20 bases falls below the threshold of q20), ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 (removes the specified adapters), and lastly MINLEN:40, to remove any reads shorter than 40 bases in length.

Step 3, whole genome assembly: Due to the high probability of encountering a novel endosymbiont species in the metagenomic reads, coupled with the unavailability of a reference genome in most cases, the only feasible option remains to be a *de novo* assembly of the genome. In order to determine the sequence assembler for this pipeline, three *de novo* assemblers including ABySS (Simpson et al., 2009), Megahit (Li et al., 2015), and metaSPAdes (Nurk et al., 2017) were employed to assemble our Illumina reads. The k-mer length for ABySS was set to 85 (as the same k-mer length using this assembler successfully produced a complete genome of *C. wolffhuegeli* endosymbiont; Alickovic et al., 2021). In Megahit, a single node assembler for large and complex metagenomic reads, the default parameters were used. Megahit uses a succinct de Bruijn graph (SdBG) in order to accomplish the lowest memory assembly possible. MetaSPAdes v. 3.14.0, again a De Bruijn graph based assembler designed and tailored specifically for metagenomic reads, offers an option (-k) to specify multiple k-mer sizes for a single assembly. A range of different k-mer sizes (-k 29, 57, 71, 91, 111, 127) were used to produce *de novo* assemblies.

From each assembler, the quality of draft assembly was evaluated based on the best combination of N50 and the least fragmented contiguous sequences that belong to the endosymbiont genome. MetaSPAdes was selected, as it proved to be more efficient and also allowed me to execute *de novo* assemblies with a range of different k-mer sizes. In addition to MetaSPAdes's efficiency, it achieved assemblies overall with N50 greater than assemblies from the other two assemblers with comparatively longer endosymbiont contigs.

Step 4, Identification of candidate endosymbiont contigs: As the libraries were simulated with known depths and compositions, I was able to evaluate the quality of the resulting assemblies by quantifying the percent of the true genome captured (percent genome recovery or PGR). Here, PGR refers to the percentage of true bacterial genome that was recovered from a gDNA read library; quantified through: (RC/GL * 100), where RC refers to the Total Recovered Correct genome length (*i.e.* combined length of contigs identified correctly as part of the endosymbiont genome) and GL stands for the Target Genome length of the reference. To this end, success or failure could be evaluated by PGR of the endosymbiont genome following its isolation. The "missing regions" can be identified as due to, either assembly error or resulting from an issue with the filtering of the contigs. For each simulated library, NCBI BLASTn v. 2.10.0 was used with flag -outfmt 6, to identify contigs that might represent the endosymbiont genome (Altschul et al., 1990). A custom blastn database was used, which consisted of specific bacterial genomes (*Ca.* Riesia pediculicola USDA, *S. praecaptivus* HS1, *Arsenophonus nasoniae, Ca.* Pantoea carbekii, *C. wolffhuegeli* endosymbiont (CWE)). It can be argued that a tailored blast database can be crucial in identifying desired contigs. Being the closest relatives to the known bacterial endosymbionts found in the chewing (feather) lice and blood sucking lice, these specific bacterial genomes were chosen as they were our best options to identify contigs representing endosymbionts (Park et al., 2021; Alickovic et al., 2021, Boyd et al., 2012, 2014, 2016; Kashkouli et al., 2021; Nováková et al., 2009).

Consequently, in order to introduce a second test case scenario, the aforementioned custom blastn database was replaced with a new database solely consisting *S. praecaptivus HS1* genome (a closely related free living bacteria). After the database was replaced in the pipeline, two of the simulated libraries, "Ph_r_bc" and "Ccol_p_bc" were used (Per. Com. B. Boyd) as an input. The results were used as feedback to accommodate the pipeline with necessary changes before deploying it on real data.

Step 5, parsing blastn output to filter endosymbiont contigs: To evaluate the most efficient combination of BLAST output format 6 metrics for the identification of endosymbiont contigs, multiple combinations of metrics and alignment statistics were evaluated. The test combinations were comprised of metrics including length (length of contigs), alignment length (sequence overlap), E-value (expect value), pident (percentage of identical base matches) and lastly contig length greater than N50 (length > N50). The combination of the following metrics set to proper thresholds were found to be adequate: E-value, length (length of contigs), and pident. The thresholds for each of these parameters were set to specific values based on the combination that provided the highest PGR.

Numerous small contigs may result from a *de novo* genome assembly. As roughly one gene/1000 bases was expected, contigs smaller than 1000 bases were rejected, as they are not likely informative to subsequent studies that may employ the technique. Next, as our sequence of interest is a closely related sequence to the reference provided through the custom blastn database, e-values in the range of 10^{-1} to 10^{-50} were tested to gain the highest PGR. Smaller e-values within the range of 10^{-15} to 10^{-50} tend to reject contigs (true positives) which may appear less significant mostly due to the differences in the sequence arising from hits to comparatively more distantly related bacterial sequence in the custom blastn database described above. Hence, the e-value threshold was set to 10^{-3}, i.e., a balanced e-value from the remaining range of 10^{-1} to 10^{-15} was selected in order to accept true positives, simultaneously keeping the amount of false positives comparatively low. Initially, while testing the simulated libraries, the percent identity threshold was set to 97.5% (pident > 97.5) with a rational of rejecting every contig that is not high in sequence similarity. In order to identify and isolate endosymbiont contigs, the combination of aforementioned threshold values for node length, e-value and percent identity, was used to filter the *de novo* assembled contigs through a custom python script (GitHub link). The percent of true endosymbiont genome recovered from their respective de novo assemblies, are denoted by percent genome recovery (table 3-6). The mean genome recovery with this threshold was 98.73% (excluding *Arsenophonus* libraries, more explanation on this later). However, with the change in blastn database, i.e., a database solely consisting of the genome of closely related free living species (*S. praecaptivus*), the genome recovery immediately dropped by a considerable amount (figure 2).

In order to obtain the optimal percent identity threshold (pident), values from 50 to 100 were tested through a custom python test script (GitHub link, Figure 2). When the percent identity threshold was raised above 90% or 97%, the genome recovery for *Ca*. Riesia (red line, figure 2) and *Pantoea* (green line, figure 2) assemblies dropped significantly. More importantly, if the percent identity threshold was raised above 71, the percent genome recovery for all genomes tested dropped simultaneously (red highlighted region, Figure 2). Hence, the percent identity threshold for my tool was switched from 97.5% to 71%. More discussion on the change in percent identity threshold and its effects are considered in the results and discussion section.

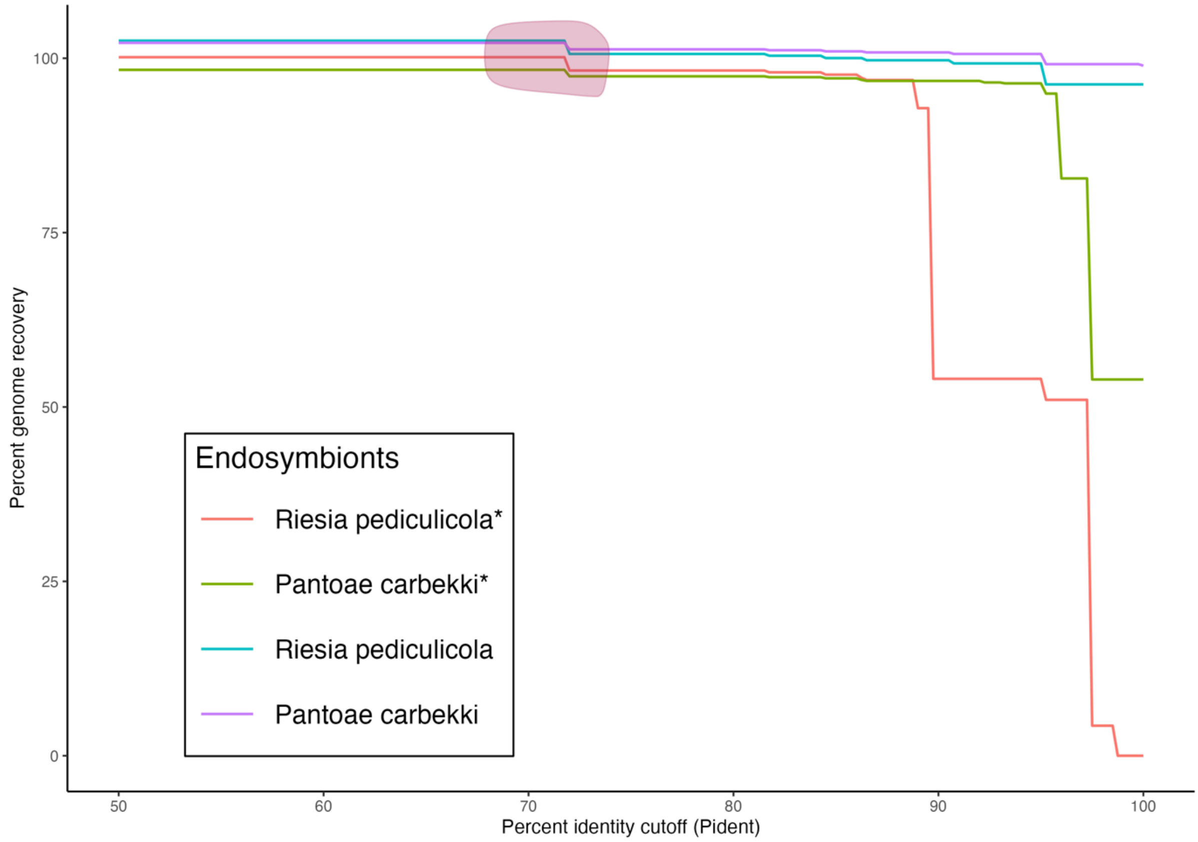

Figure 2. Percent genome recovery and percent identity cutoff. *An asterisk inside the figure legend indicates the change of database.*

Step 6, isolation of potential endosymbiont reads: In order to isolate the potential endosymbiont reads, the resulting filtered contigs from the previous step were used to map all the input reads using the end to end mode of Bowtie2 v.2.2.6 (Langmead et al., 2012), and the unaligned reads were suppressed using the "–no-unal" flag. The paired end reads that aligned to the filtered contigs were collected from the dataset stored separately.

Step 7, reassembly of endosymbiont genomes: In order to reassemble the endosymbiont contigs, the mapped reads were then used as an input into the new assembly using metaSPAdes, thereby contributing to the elongation and/or merging of contigs. Lastly, in order to investigate the missing endosymbiont contigs and to get a visual mapping of the contigs representing specific parts of the endosymbiont genome, all the genome assemblies were plotted using Circos v 0.69-8 (Krzywinski et al., 2009) (Figures 3-4 & 6-9).

Results and discussion

In this study, I evaluated a novel method to select *de novo* assembled contigs that represent the endosymbiont genome when starting with whole louse genomic DNA library (Figure 1). The mean endosymbiont genome recovery from simulated data was 91.27%, with a range of 100%-76%. From the total recovered genome, the average true positive rate for the true genome recovered was 99.14%, with an average false positive rate of 0.85% (table 3). Furthermore, regarding the two test case scenarios, despite the change in database (i.e. a database with the genome of a closely related free living bacteria, *S. praecaptivus*), the mean endosymbiont genome recovery was 99.17% (table 4). When the method was tested with "real" whole louse shotgun sequencing data, the results were mixed (table 6-8). The strategy was accurate in contig selection for libraries obtained from blood sucking lice gDNA, with a mean percent genome recovery of 98.38%, from the range of 100% - 95.88%. In comparison, libraries built on gDNA from feather lice produced questionable assemblies that, unfortunately, could not be evaluated as easily as the other test cases.

Parsing simulated *de novo* genome assemblies

As the query (*de novo* assembled contigs) and reference genomes (blastn database) were identical, I expected that the percentage of genome captured would be identical in composition and similar in length with the input genome. In order to check the accuracy of my expectations, percent identity threshold for my tool was initially set to a higher value (97.5%) with a rationale that a searching criterion should be restricted enough to reject all contigs that are not analogous. Next, in order to identify target contigs, all the *de novo* assemblies were filtered using the aforementioned filter thresholds. As expected, the average percent genome recovery was 91.27% with an average true positive rate of 99.14% and a negligible, 0.85% false positive rate (table 3-4, figure 3-4).

Table 3. *De novo* assembly results for simulated libraries

Library name	Target genome length	Recovered correct	Recovered incorrect	Total recovered	Percent true positive	Percent false positive	True genome recovery
ph_r_bc	0.582 Mb	0.56 Mb	0 Mb	0.56 Mb	100%	0%	96%
ph_sp_bc	5.16 Mb	5.16 Mb	0.13 MB	5.29 Mb	97.54%	2.46%	100%
ph_a_bc	4.98 Mb	3.83 Mb	0 Mb	3.83 Mb	100%	0%	76%

ccol_a_bc	4.98 Mb	3.84 Mb	0 Mb	3.84 Mb	100%	0%	76.72%
ccol_p_bc	1.19 Mb	1.18 Mb	0 Mb	1.18 Mb	100%	0%	98.90%
ccol_sp_bc	5.15 Mb	5.15 Mb	0.14 Mb	5.29 Mb	97.35%	2.64%	100%

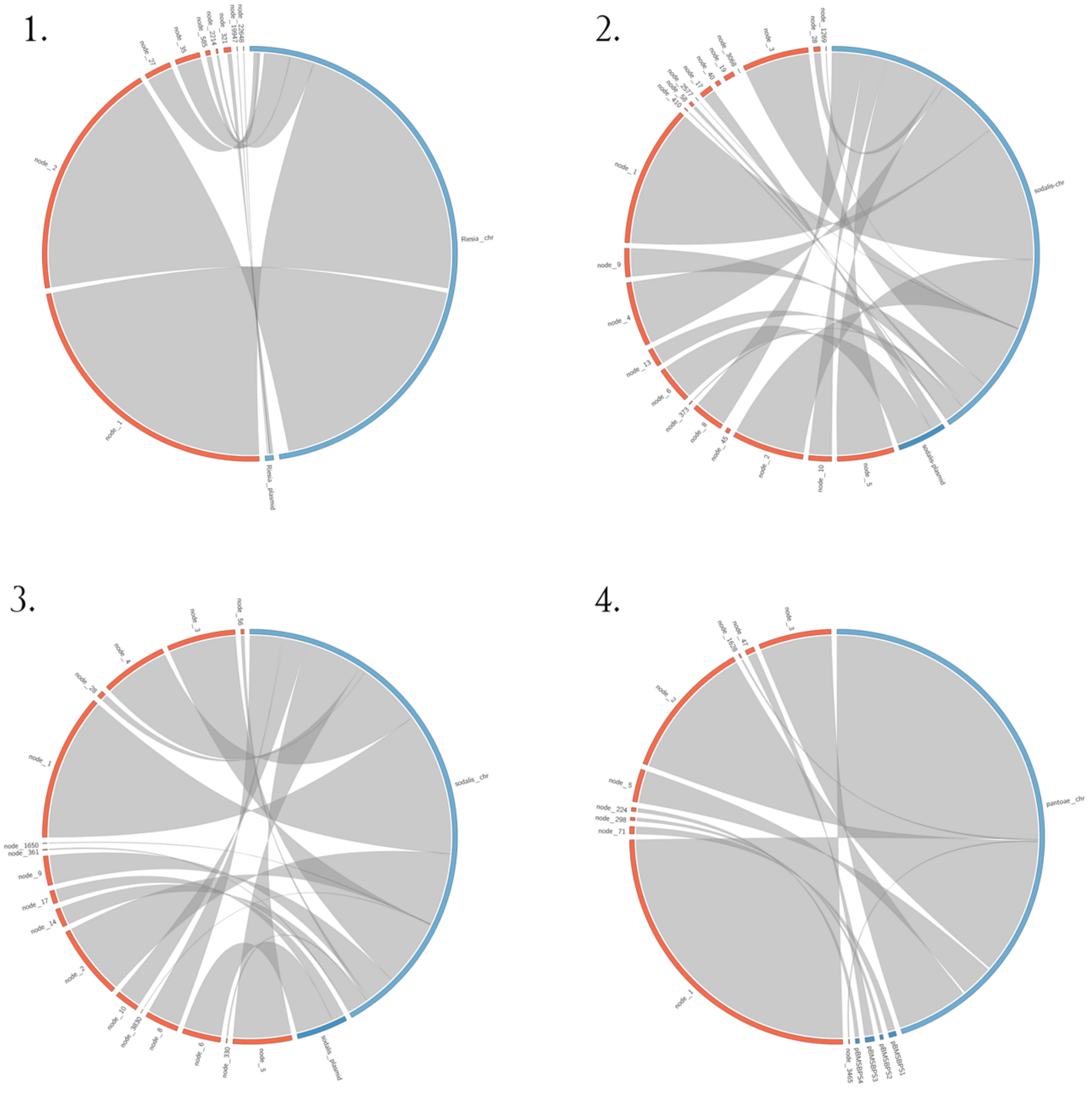

Figure 3. ***Comparison of the filtered contigs identified and the expected genome assembly, given a percent identity of 97.5 used to exclude false positives.*** *In each panel, the right hemisphere (blue arcs) of the plot represents an endosymbiont genome assembly used to simulate a library and the left hemisphere*

(red arcs) represents the draft assembly produced by my method. Shared regions of the reference genome and the draft assembly are identified by connecting ribbons (gray). Comparison of four assemblies with target endosymbiont genomes including, 1) Ca. Riesia and "ph_r_bc" draft assembly, produced using simulated reads of P. humanus, Ca. Riesia, and Burkholderia genomes; 2) S. praecaptivus and "ph_sp_bc" draft assembly, produced using simulated reads of P. humanus, S. praecaptivus, and Burkholderia genomes; 3) S. praecaptivus and "ccol_sp_bc" draft assembly, produced using simulated reads of C. columbae, S. praecaptivus, and Burkholderia genomes; 4) Pantoea carbekki genome and "ccol_pantoea_bc" draft assembly, produced using simulated reads of C. columbae, Pantoea carbekki, and Burkholderia genomes.

The majority of the nodes/contigs from the draft assemblies connected to specific portions (98.73% on average for all draft assemblies) of their respective reference genomes (Figure 3). However, both *A. nasoniae* assemblies (*A. nasoniae* in blood sucking and feather lice respectively) had a lower genome recovery rate of 76%. Here I found that the nodes/contigs from the *de novo* assembly only connect partially (~76%) to the *A. nasoniae* chromosome, leaving all the plasmids vacant. The disconnected regions of the circos plots could be an indicator that both *A. nasoniae* assemblies lack the contigs that should have represented the remaining genome; i.e., some parts of chromosome, and all the plasmids.

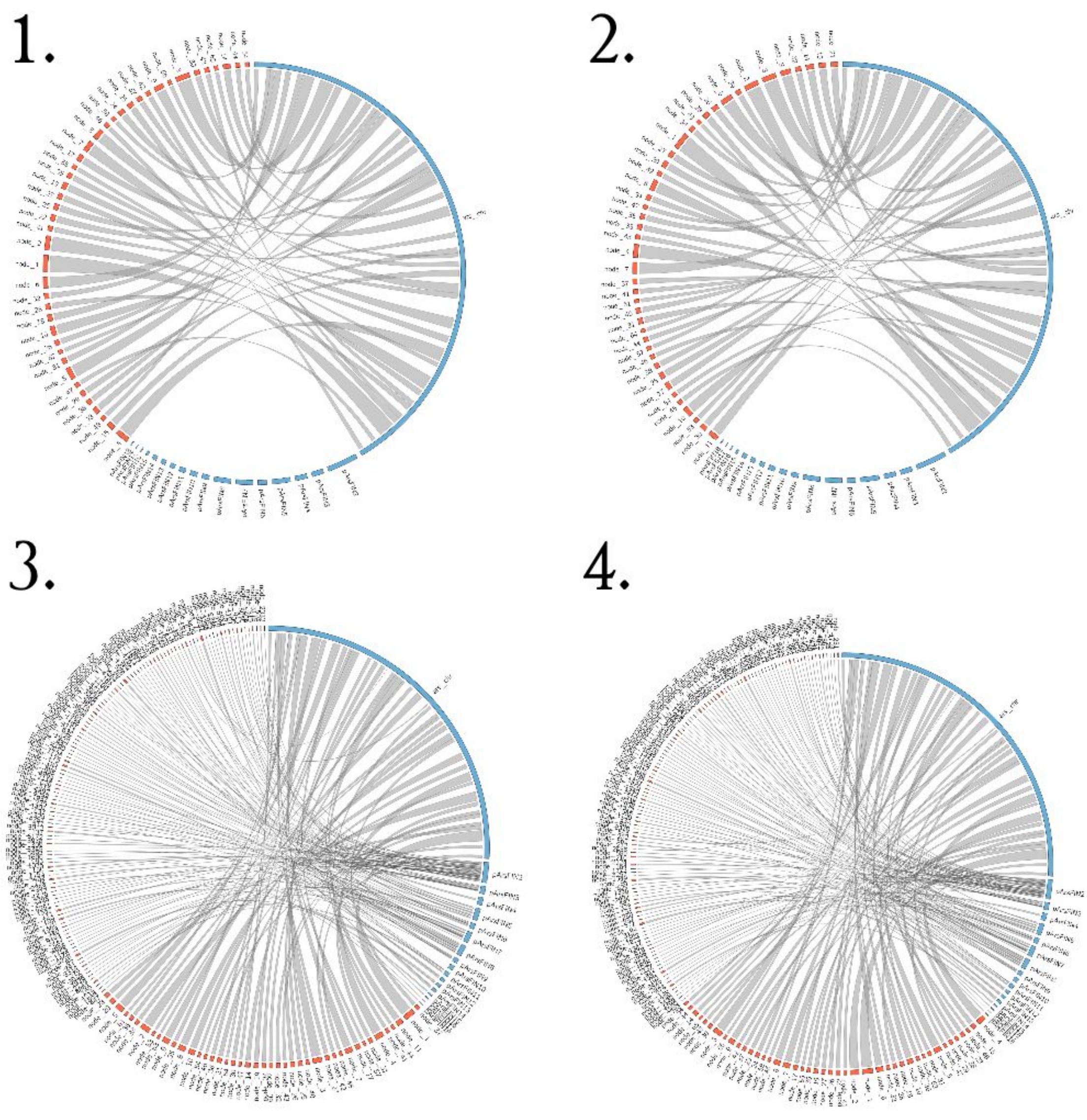

Figure 4. ***Comparison of the filtered contigs identified by my method and the expected genome assembly, given a percent identity of 97.5 used to exclude false positives.*** *In each panel, the right hemisphere (blue arcs) of the plot represents an endosymbiont genome assembly used to simulate a library and the left hemisphere (red arcs) represents the draft assembly produced by my method. Shared regions of the reference genome and the draft assembly are identified by connecting ribbons (gray). Comparison of four assemblies with target endosymbiont genomes including: 1) A. nasonia and "ph_a_bc" draft assembly, produced using simulated reads of P. humanus, A. nasonia, and Burkholderia*

genomes; 2) A. nasonia and "ccol_a_bc" draft assembly, produced using simulated reads of C. columbae, A. nasonia, and Burkholderia genomes; 3) A. nasonia and revised "ccol_a_bc" draft assembly, produced using simulated reads of C. columbae, A. nasonia, and Burkholderia genomes; 4) A. nasonia and revised "ph_a_bc" draft assembly, produced using simulated reads of P. humanus, A. nasonia, and Burkholderia genomes.

To further investigate the missing endosymbiont contigs, both *A. nasonia* genome assemblies (*A. nasoniae* in blood sucking and feather lice respectively) were queried against a custom database solely consisting of *A. nasoniae* plasmids. The contigs representing the endosymbiont plasmids were successfully isolated. Next, these isolated contigs were then included in the revised genome assemblies. As expected, all the contigs that should have represented the *A. nasoniae* plasmids were captured successfully (3rd & 4th panel, Figure 4). However, this type of manual recovery with a custom plasmid database was counter to the goal, i.e. to minimize manual steps that are difficult to replicate. Therefore, the contigs captured after querying the custom *A. nasonia* "plasmids only" database, were excluded from the revised *A. nasonia* assemblies. The contigs were not included in the draft assemblies that were used to report percent genome recoveries of both *A. nasonia* assemblies.

After the sequencing of total genomic DNA, plasmids that have a different copy number or repetitive elements of the bacterial genome may result in genomic DNA reads with biased or irregular sequencing depths. Due to the same issue with the published *A. nasonia* genome on SRA (18 plasmids), the successful recovery of a complete *A. nasonia* genome from a gDNA sample remains to be a challenge. The aforementioned reasons could be a plausible explanation for both the *A. nasonia* assemblies ("Ph_a_bc" and "Ccol_a_bc") having a comparatively lower true genome recovery rate (i.e. 77.4%) than the remaining simulated libraries (mean genome recovery of 98.73%).

Despite the successful recovery of the endosymbiont genome from the majority of *de novo* assemblies (average PGR without *A. nasonia* assembly was 98.73%), it is important to note that the query and subject were identical. However, the scenario described above does not reflect a real world situation, where the query is at best, closely related to the subject database. In order to intensify the simulated scenario, the aforementioned custom blastn database was replaced with a new database solely consisting of *S. praecaptivus* HS1 genome. After the database was replaced in the pipeline, two of the simulated libraries, “Ph_r_bc” and “Ccol_p_bc” were used as an input. The two test cases resulted with a percent genome recovery of 4.30% and the latter with 53.8%. These inconsistent results were expected, because of the highly restrictive nature of the threshold (Percent identity greater than or equal to 97.5%). Consequently, it was necessary to optimize the filter thresholds in order to improve the genome recovery rate and accommodate the pipeline for real data.

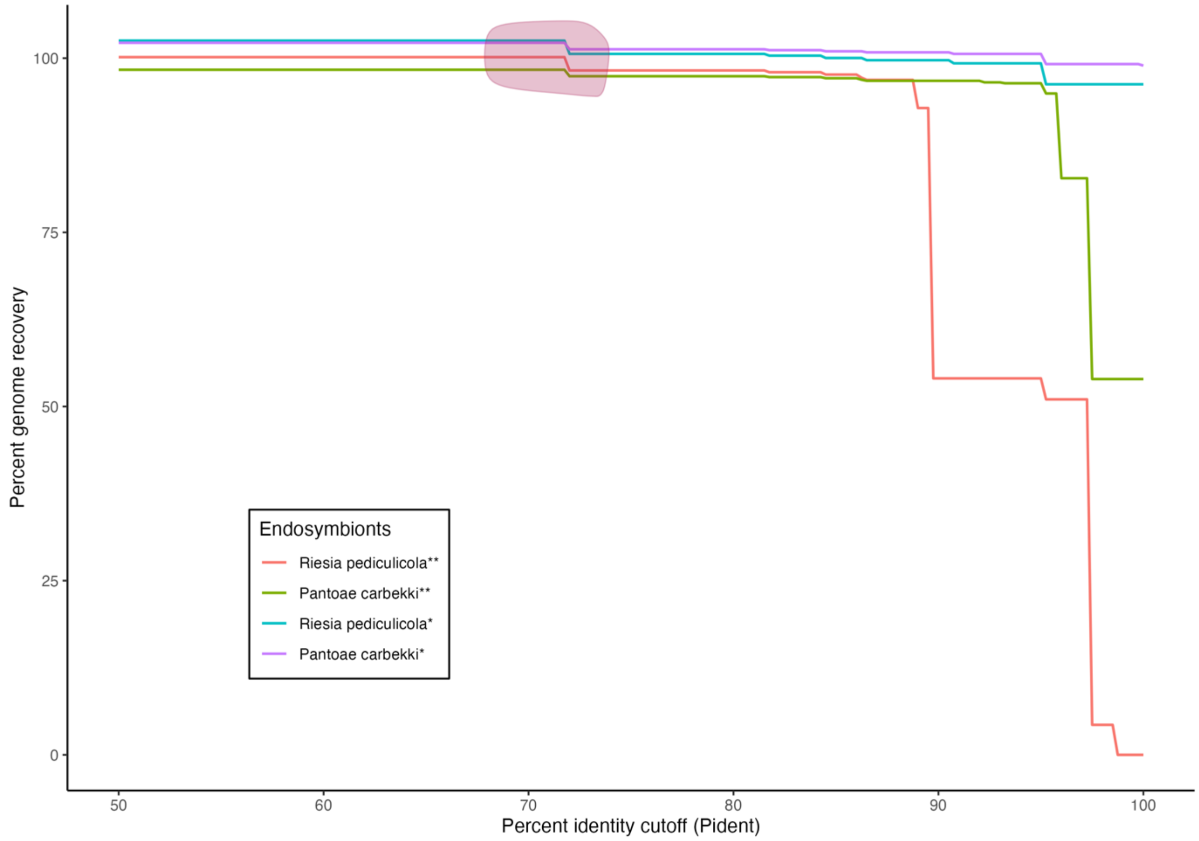

Figure 5. ***Percent genome recovery and percent identity cut-off.***

In order to obtain the optimal percent identity threshold (pident), values from 50 to 100 were tested (Figure 5). When the percent identity threshold was raised above 90% or 97%, the genome recovery for *Ca*. Riesia (red line, Figure 5) and *Pantoea* (green line, Figure 5) assemblies declined rapidly. More importantly, if the percent identity threshold was raised above 71, the percent genome recovery for all genomes tested dropped simultaneously (red highlighted region, Figure 5). Hence, the percent identity threshold was lowered from 97.5% to 71%. The genome recovery for both test cases surged to 100% and 98.33% respectively with minimal false positive rates (Table 4).

Table 4. Simulated libraries and *de novo* assembly results using a database that does not contain the same genome.

Library name	Target genome length	Recovered correct	Recovered incorrect	Total recovered	Percent true positive	Percent false positive	True genome recovery
ph_r_bc	0.582 Mb	0.582 Mb	0.00086 Mb	0.582 Mb	99.85 %	0.001%	100%
ccol_p_bc	1.19 Mb	1.17 Mb	0 Mb	1.17 Mb	100%	0%	98.33%

The decrease in percent identity threshold should intuitively increase the amount of false positives for a case where the database is made of genomes identical to the query. In order to observe and visualize the outcome, the draft assemblies were parsed with an updated python script with the newly adjusted thresholds (percent identity >= 71). The results demonstrated an increase in false positive contigs for each assembly (Table 5, Figure 6).

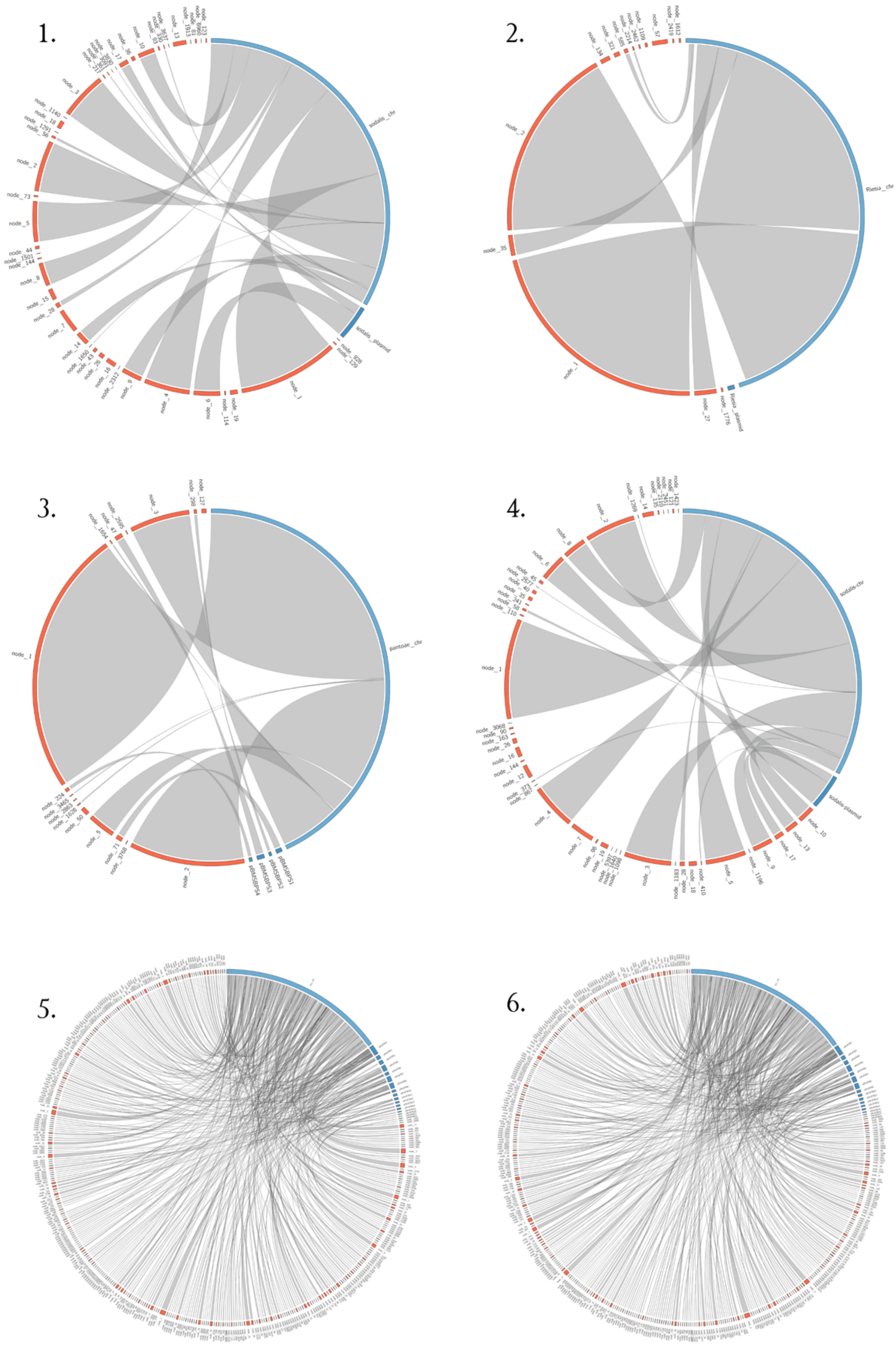

Figure 6. *Comparison of the filtered contigs identified by my method and the expected genome assembly, given a percent identity of 71 used*

to exclude false positives. *In each panel, the right hemisphere (blue arcs) of the plot represents an endosymbiont genome assembly used to simulate a library and the left hemisphere (red arcs) represents the draft assembly produced by my method. Shared regions of the reference genome and the draft assembly are identified by connecting ribbons (gray). Comparison of six different draft assemblies with target endosymbiont genomes including: 1) S. praecaptivus and "ccol_sp_bc" draft assembly, produced using simulated reads of C. columbae, S. praecaptivus, and Burkholderia genomes; 2) Ca. Riesia and "ph_r_bc" draft assembly, produced using simulated reads of P. humanus, Ca. Riesia, and Burkholderia genomes; 3) Pantoea carbekki and "ccol_p_bc" draft assembly, produced using simulated reads of C. columbae, Pantoea carbekki, and Burkholderia genomes; 4) S. praecaptivus and "ph_sp_bc" draft assembly, produced using simulated reads of P. humanus, S. praecaptivus, and Burkholderia genomes. 5) A. nasonia and "ccol_a_bc" draft assembly, produced using simulated reads of C. columbae, A. nasonia, and Burkholderia genomes; 6) A. nasonia and "ph_a_bc" draft assembly, produced using simulated reads of P. humanus, A. nasonia, and Burkholderia genomes.*

Table 5. Simulated libraries and *de novo* assembly metadata after the percent identity threshold was lowered (Percent identity threshold set to >= 71.0)

Library name	**Target genome length**	**Recovered correct**	**Recovered incorrect**	**Total recovered**	**Percent true positive**	**Percent false positive**	**True genome recovery**
ph_r_bc	0.582 Mb	5.7 Mb	0.02 Mb	5.96 Mb	95.61%	4.44%	98.03%
ph_sp_bc	5.16 Mb	4.9 Mb	1.4 Mb	6.37 Mb	77.97%	22.02%	96.30%

ph_a_bc	4.98 Mb	3.81 Mb	0.05 Mb	3.87 Mb	98.61%	1.38%	76.54%
ccol_a_bc	4.98 Mb	3.82 Mb	0.05 Mb	3.87 Mb	98.65%	1.34%	76.60%
ccol_p_bc	1.19 Mb	1.91 Mb	0.03 Mb	1.22 Mb	97.37%	2.62%	99.52%
ccol_sp_bc	5.15 Mb	4.9 Mb	1.4 Mb	6.37 Mb	77.97%	22.02%	96.30%

Implementation with libraries from previous studies

In the case of libraries produced from blood sucking lice, as the published reference genomes for endosymbionts were available, the draft assemblies produced using the pipeline were compared to their respective reference genomes from NCBI. The average percent genome recovery was 98.38% with an average true positive rate of 99.63 %. In the case of *Pediculus humanus* library, the genome recovery of its known endosymbiont species *Ca*. Riesia pediculicola, was 95.88% with 100% true positives and 0% false positives (Table 6, Figure 7).

Table 6. Endosymbiont genome assembly statistics generated from whole genome blood sucking louse libraries.

Library name	*Ca.* Riesia pediculicola	Endosymbiont of *Pedicinus badii*	*Ca.* Riesia pediculischaeffi
Target genome length	0.582124 Mb	0.558122 Mb	0.571864 Mb
Recovered correct	0.558192 Mb	0.558122 Mb	0.567703 Mb
Recovered incorrect	0 Mb	0.006269 Mb	0 Mb
Total recovered	0.558192 Mb	0.564391	0.567703 Mb
Percent true positive	100%	98.89%	100%

Percent false positive	0%	1.11%	0%
target genome recovery	95.88%	100%	99.27%

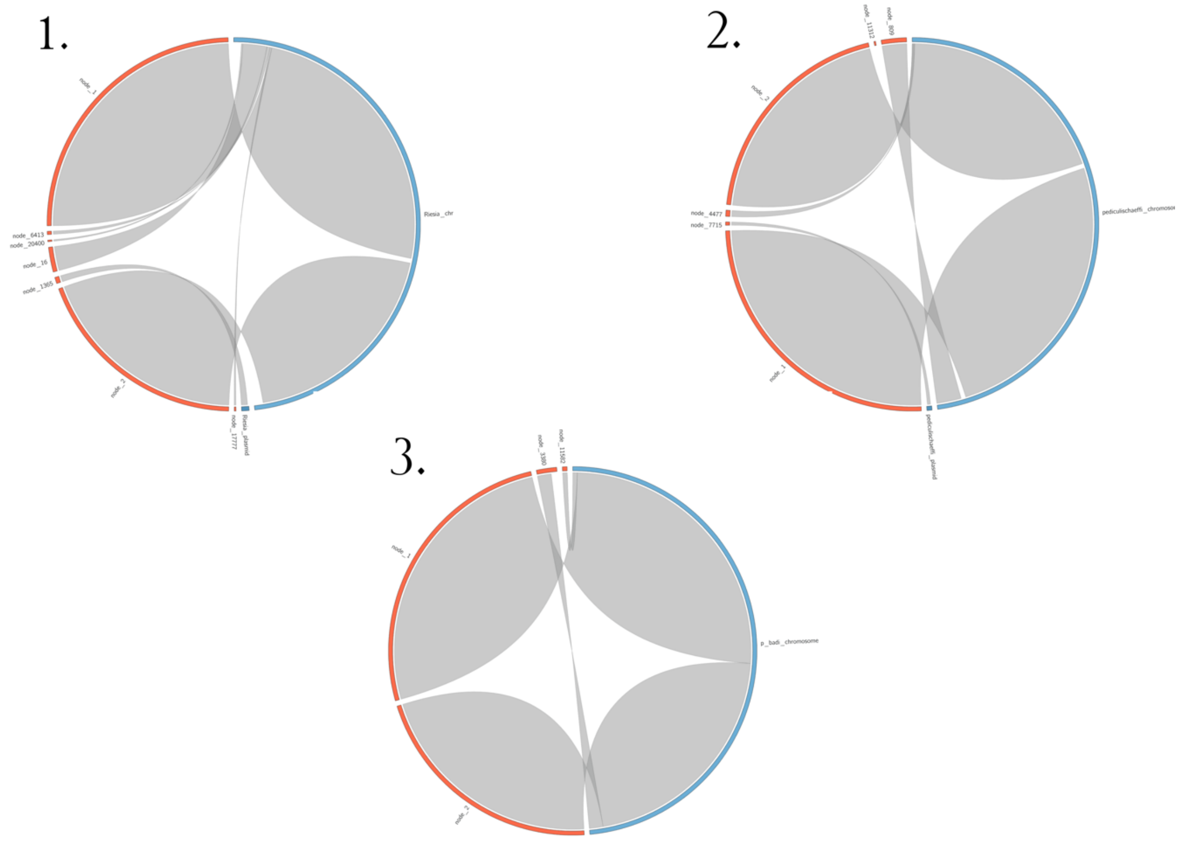

Figure 7. *Comparison of the filtered contigs identified by my method and the expected genome assembly, given a percent identity of 71 used to exclude false positives.* *The de novo assembly was produced using a whole blood sucking louse gDNA library downloaded from NCBI. In each panel, the right hemisphere (blue arcs) of the represents their respective published endosymbiont genome and the left hemisphere (red arcs) represents the draft assembly produced by my method. Shared regions of the reference genome and the draft assembly are identified by connecting ribbons (gray). Comparison of three different draft assemblies with target endosymbiont genomes including: 1) Ca. Riesia pediculicola*

and Ca. Riesia pediculicola draft assembly; 2) Ca. Riesia pediculishaeffi and Ca. Riesia pediculishaeffi draft assembly; 3) Endosymbiont of Pedicinus badii and its respective draft assembly.

For all three circos plots (Figure 7) that represent the draft assemblies recovered from whole blood sucking louse reads, the majority of nodes connected and covered the whole reference genome (98.73% on average). The method was accurate in selection of contigs that represent the endosymbiont genome; hence producing draft assemblies, similar in length and composition to the reference genomes.

Finally, draft assemblies were produced from the whole wing louse gDNA (*Columbicola* species) sequencing projects and the assemblies were evaluated. These feather louse endosymbionts were suspected to be closely related to *Enterobacter*, likely congeneric. In a study by Smith et al. in 2013, a part of *fusA* gene from the endosymbiont of *Columbicola macourae* (Comac1) was obtained and used for the phylogenetic evaluations in their study. The same *fusA* sequence was mapped to the *Enterobacter* draft assemblies produced by the method from this study. The results demonstrated 100% match, confirming successful isolation of Comac1 contigs from the *de novo* assembly (Figure 8). The homologous region was identified in the assemblies from additional endosymbionts from *Columbicola* species and was compared as well (Figure 8).

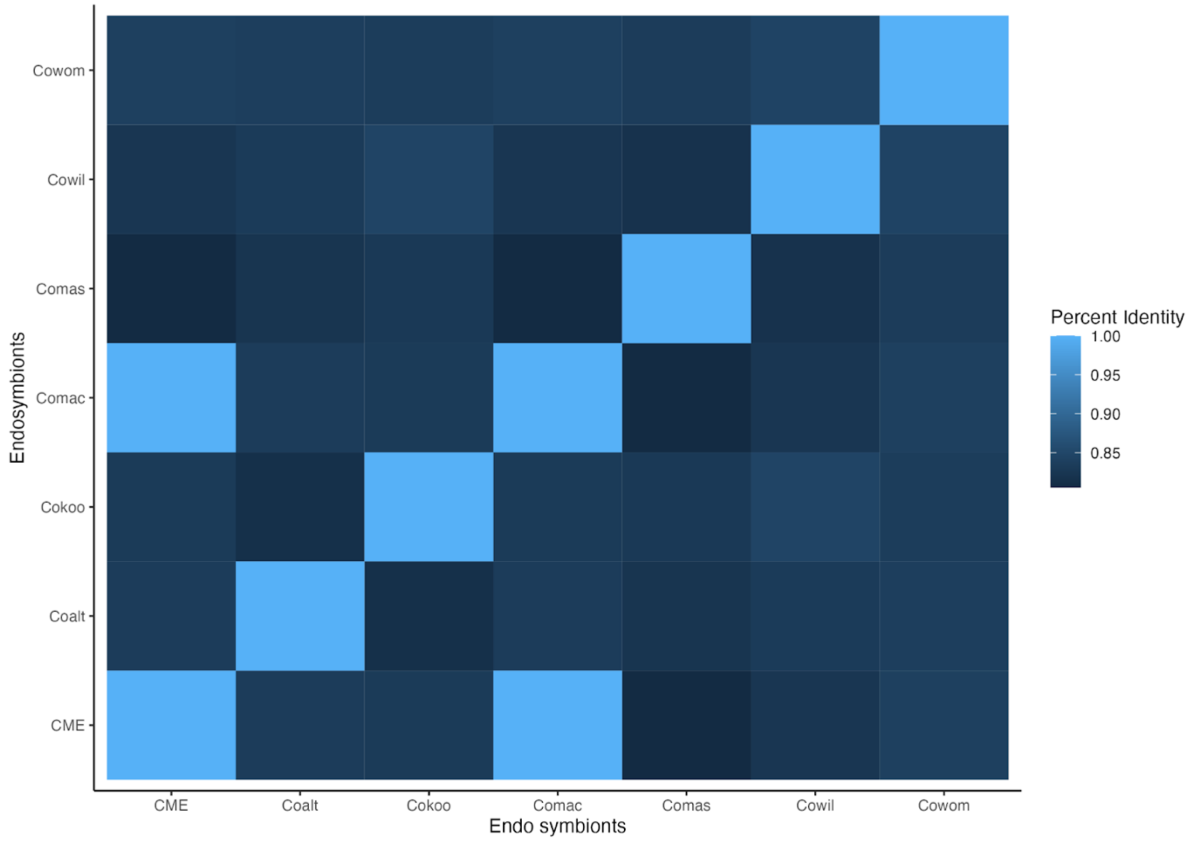

Figure 8. ***A heatmap comparing the fusA gene previously isolated and sequenced from the endosymbiont of C. macourae (Smith et al., 2013) and homologous genes identified in genome assemblies of endosymbionts from C. macourae along with the other species in the genus Columbicola (based on uncorrected p-distance).***

Enterobacter cloacae has a genome length of 5.04 Megabases with 55% GC content. Likewise, the genome of *Ca*. Riesia pediculicola, the endosymbiont of the human head louse, is 0.58 Mb. Therefore, the approximate "hard" lower base limit should be around 0.6 Mb. However, in the case of *Enterobacter* endosymbionts, considering they are suspected of being "nascent" endosymbionts and are still undergoing the process of genome reduction (McCutcheon et al., 2012), therefore I expected their genome length to be in the range of ~1 Mb to 2 Mb. As expected, all the draft assemblies produced through the method had their genome length in the range of 1.1-2 Mb with approximately 34-44% GC base composition. Next, genome statistics of all the draft endosymbiont genome assemblies were compared to a range of closely related free living along with some endosymbiotic bacteria (Figure 9).

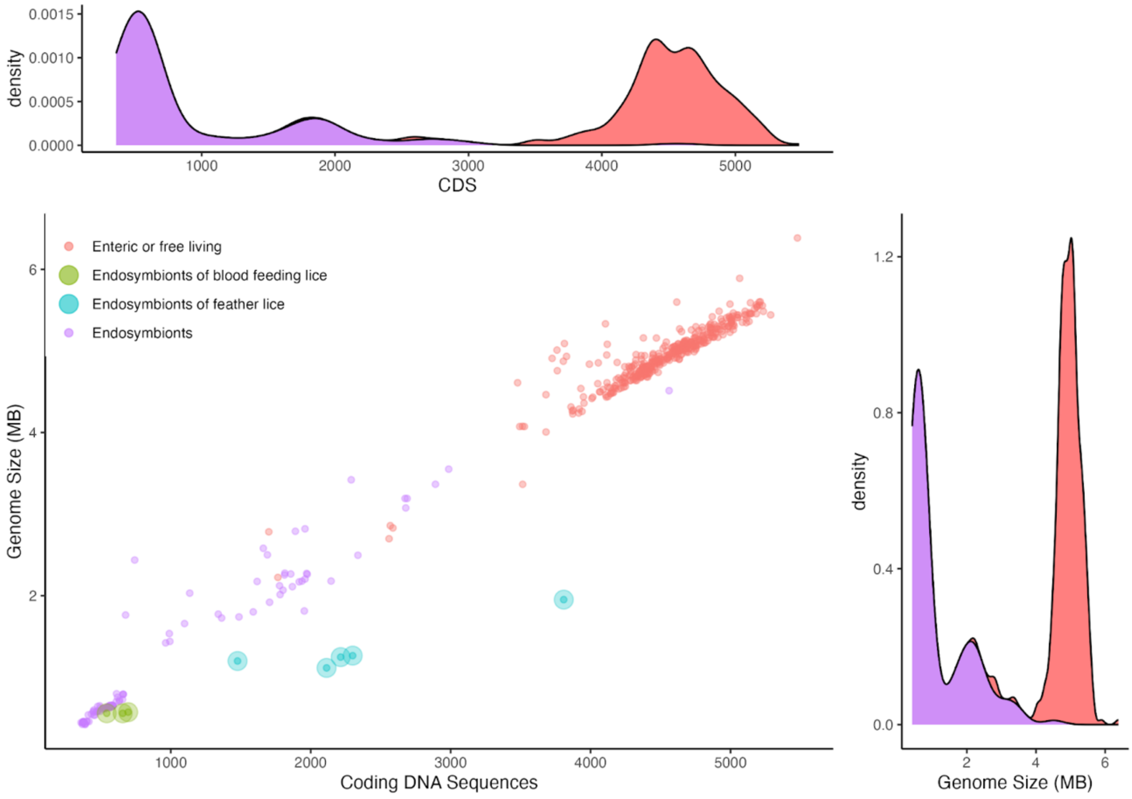

Figure 9. *Comparison of genome length and number of Protein coding DNA sequences (CDS).* *The central panel (enlarged) represents the comparison of Coding DNA sequences (X axis) and bacterial genome size (Y axis), whose colours correspond to the type of bacteria (i.e. symbionts or enteric/free living). The points highlighted with enlarged radius represent endosymbionts of specific blood sucking (green) and/or feather (blue) louse. The top panel demonstrates the density of coding DNA sequences from symbionts (purple) and enteric/free living bacteria (red); likewise, the panel on right demonstrates the density of their genome sizes.*

The draft genome assemblies of endosymbionts from blood sucking lice, namely: *Ca.* Riesia pediculicola, *Ca.* Riesia pediculischaeffi, and an undescribed species, which were produced by the method, yielded genomes of similar size and base composition to the reference (Figure 7). However, gene count was higher than the reference, suggesting the methods could still be further evaluated (green points, Figure 9; Table 6). However, the gene counts (CDS) for all *Enterobacter* assemblies were comparatively high, with the exception of Comac1 assembly. To get an idea of its extremity, the gene counts can be compared to a free living bacteria *E.coli* K12, and a well-studied endosymbiont, *Ca.* Riesia pediculicola. *E-coli* K12 has a genome length of 5.10 Megabases and produces 4723 proteins, therefore, the gene/1000 bases ratio here is 0.9:1000. Similarly, *Ca.* Riesia pediculicola has a genome length of 0.582124 Megabases producing 468 proteins, hence, the gene/1000 bases ratio here is 0.8:1000. Therefore, the aforementioned facts set the expectation to be approximately around 1 gene/1000 bases. The coding DNA sequences (CDS) for most endosymbionts (purple points, Figure 9) including endosymbiont of blood sucking louse (green points, Figure 9) as well as enteric/free living species (red points, Figure 9) which were included in the analysis, are directly proportional to their genome size. Conversely, the results for *Enterobacter* assemblies (blue points, Figure 9) did not fit the trend (Table 7).

Table 7. *Enterobacter de novo* assemblies and their genome characteristics.

Libraries (sample names)	Genome length (Megabases)	Number of Coding DNA sequences (CDS)
Cowom	1.114332 Mb	2113

Comas	1.951348 Mb	3808
Cowil	1.246816 Mb	2214
Cokoo	1.264745 Mb	2300
Comac1	1.198779 Mb	1477

To gain more insight regarding the high-coding density, all the assemblies were examined for completeness using CheckM (Parks et al., 2015). CheckM reports the completeness of the genome based on a core set of genes. It is phylogenetically informed, therefore, the program gains accuracy by comparing closely related species. The assembly of the *C. wolffhuegeli* endosymbiont genome (Alickovic et al., 2021) was determined to be near complete (94.58%) using CheckM and no contamination was detected, providing a baseline expectation when implementing the program with endosymbionts containing small genomes. On the contrary, all the *Enterobacter de novo* assemblies were determined to be in the range of 25-35% complete using CheckM. This suggests the assemblies are incomplete, despite the high coding density.

Here I explore two possibilities that attempt to explain the observed coding density in *Enterobacter de novo* assemblies, (1) gene count is true, or (2) methodological (assembly) error result in erroneous gene prediction. The first scenario, gene count is accurate, could be due to the presence of small pseudogenes that are reported as genes. The second scenario, inaccurate, but well supported base calls could be incorporated into the *de novo* assemblies, leading to breaks in real genes. Both scenarios could result in additional short ORFs that impede the process of gene annotation, hence explaining the high gene count and the suggested incomplete genome report by CheckM.

To investigate and improve on the resulting assemblies, additional steps were taken to measures the impact on genome assembly. First the gDNA sequence reads were reisolated by mapping their respective *de novo* assemblies and the *de novo* assemblies were repeated from the isolated sequence reads hoping for longer contiguous sequences in the resulting assembly. In order to isolate only the aligned reads this time, the assemblies were mapped to the original trimmed gDNA sequence reads using Bowtie2. Subsequently, the isolated gDNA sequence reads were used afresh to produce *de novo* assemblies. Furthermore, multiple k-mer sizes ranging from 31 to 121, were used in the production of *de novo* assemblies and the assembly jobs were executed individually this time using separate bash scripts. Consequently, the last approach was to verify proper implementation of quality control and confirm if any good quality bases were trimmed erroneously, or if any low quality bases remained. To avoid either of the former prospects, quality control was revisited by using Trimmomatic's sliding window option with the parameter set to 4:20, i.e., implementing a sliding window of 4 bases and trimming the leftmost base of that window if the average quality drops below 20. Despite all the aversions, the results remained the same overall and negligible improvements were observed.

Table 8. Real Illumina HiSeq 2500 Columbicola Columbae whole louse libraries & their assembly statistics.

Sample library name	**SRA sample identifier**	**Number of raw reads**	***De novo* assembly length (in bases)**	**Number of Contigs**	**N50**	**GC content**
Coalt	SRS1284168	56,593,624	1.536,389 Mb	90	290,12	35%
Cowil	SRS1284129	88,862,786	1.246,816 Mb	30	77,692	43.6%
Cokoo	SRS1284131	109,164,924	1.264,745 Mb	23	130,495	41%
Comac1	SRS1284175	60,041,496	1.148,974 Mb	21	139,497	34.4%
Comas	SRS1284153	58,281,528	1.951,348 Mb	29	101,748	35.3%
Cothe	SRS1284174	69,941,006	0.636,347 Mb	41	30,539	32.9%
Cotri	SRS1284126	121,706,960	0.097,080 Mb	33	3,343	33.3%
Cowom	SRS1284146	52,700,590	1.114,332 Mb	54	30,962	40.3%

The close match between reference and assembly in case of the blood louse libraries may be partly due to the lack of pseudogenes, lack of repeats, or the genome size of *Ca*. Riesia pediculicola*,* *Ca*. Riesia pediculischaeffi, and the *Pedicinus badii* endosymbiont (Kirkness et al., 2010; Boyd et al., 2017). In the case of nascent symbionts, which may contain transposon derived sequences and tandem repeats, obtaining complete genome assemblies may be difficult from short-read data. Further examination of the method would benefit from comparisons with previously assembled genomes, that are larger in size and contain complex structures, which would require examination of insect-endosymbiont pairs outside of lice.

Conclusions

In this study, a method was developed and evaluated to assemble endosymbiont genomes; and was demonstrated how this approach was effective and less time and labour intensive than previous studies. At this stage, no official pipeline exists, instead the method is composed of bash and python scripts. However, since the workflow is straightforward, the pipeline could be easily automated either through python, e.g. Snakemake or PyBuilder. The results of this study demonstrate that the genomic sequences of a comparatively distant relative which can be either enteric/free living or symbiotic bacteria, may be used as a reference to identify and isolate endosymbiont genome sequences. Conversely, the potential limitations of this method were demonstrated through *Enterobacter* libraries. Therefore, the results presented in this study may contribute substantially towards the development of more efficient methods.

References:

1. Alickovic, Leila, Kevin P. Johnson, and Bret M. Boyd. "The reduced genome of a heritable symbiont from an ectoparasitic feather feeding louse." *BMC Ecology and Evolution* 21.1 (2021): 1-11.
2. Allen, Julie M., et al. "Evolutionary relationships of "Candidatus Riesia spp.," endosymbiotic Enterobacteriaceae living within hematophagous primate lice." *Applied and Environmental Microbiology* 73.5 (2007): 1659-1664.
3. Altschul, Stephen F., et al. "Basic local alignment search tool." *Journal of molecular biology* 215.3 (1990): 403-410.
4. Andrews, Simon. "FastQC: a quality control tool for high throughput sequence data." (2010).
5. Baldwin-Brown, James G., et al. "The assembled and annotated genome of the pigeon louse Columbicola columbae, a model ectoparasite." *G3* 11.2 (2021): jkab009.
6. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30.15 (2014): 2114-2120.
7. Boyd, Bret M., et al. "Genome sequence of Candidatus Riesia pediculischaeffi, endosymbiont of chimpanzee lice, and genomic comparison of recently acquired endosymbionts from human and chimpanzee lice." *G3: Genes, Genomes, Genetics*4.11 (2014): 2189-2195.
8. Boyd, Bret M., et al. "Primates, lice and bacteria: speciation and genome evolution in the symbionts of hominid lice." *Molecular Biology and Evolution* 34.7 (2017): 1743-1757.
9. Boyd, Bret M., et al. "Two bacterial genera, Sodalis and Rickettsia, associated with the seal louse Proechinophthirus fluctus (Phthiraptera: Anoplura)." *Applied and Environmental Microbiology* 82.11 (2016): 3185-3197.

10. Boyd, Bret M. et al. (2017), Data from: Phylogenomics using target-restricted assembly resolves intra-generic relationships of parasitic lice (Phthiraptera: Columbicola), Dryad, Dataset, https://doi.org/10.5061/dryad.4812p
11. Brinza, Lilia, et al. "Systemic analysis of the symbiotic function of Buchnera aphidicola, the primary endosymbiont of the pea aphid Acyrthosiphon pisum." *Comptes rendus biologies* 332.11 (2009): 1034-1049.
12. Buchner, Paul. "Endosymbiosis of animals with plant microorganisms." (1965).
13. Bush, Sarah E., and Dale H. Clayton. "The role of body size in host specificity: reciprocal transfer experiments with feather lice." *Evolution* 60.10 (2006): 2158-2167.
14. Bush, Sarah E., and Jael R. Malenke. "Host defence mediates interspecific competition in ectoparasites." *Journal of Animal Ecology* (2008): 558-564.
15. Camacho, Christiam, et al. "BLAST+: architecture and applications." *BMC bioinformatics* 10.1 (2009): 1-9.
16. Chari, Abhishek, et al. "Phenotypic characterization of Sodalis praecaptivus sp. nov., a close non-insect-associated member of the Sodalis-allied lineage of insect endosymbionts." *International journal of systematic and evolutionary microbiology* 65.Pt 5 (2015): 1400.
17. Chevignon, Germain, et al. "Culture-facilitated comparative genomics of the facultative symbiont Hamiltonella defensa." *Genome Biology and Evolution* 10.3 (2018): 786-802.
18. Chrudimský, Tomáš, et al. "Candidatus Sodalis melophagi sp. nov.: phylogenetically independent comparative model to the tsetse fly symbiont Sodalis glossinidius." *PLoS One* 7.7 (2012): e40354.

19. Clayton, Adam L., et al. "A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect–bacterial symbioses." *PLoS genetics* 8.11 (2012): e1002990.
20. Davison, Helen R., et al. "Genomic diversity across the Rickettsia and 'Candidatus Megaira'genera and proposal of genus status for the Torix group." *Nature communications* 13.1 (2022): 1-14.
21. Davison, Helen R., et al. "Genomic diversity across the Rickettsia and 'Candidatus Megaira'genera and proposal of genus status for the Torix group." *Nature Communications* 13.1 (2022): 2630.
22. Douglas, A. E., and W. A. Prosser. "Synthesis of the essential amino acid tryptophan in the pea aphid (Acyrthosiphon pisum) symbiosis." *Journal of insect physiology* 38.8 (1992): 565-568.
23. Douglas, A. E. "Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria Buchnera." *Annual review of entomology* 43.1 (1998): 17-37.
24. Douglas, AE2696562. "Mycetocyte symbiosis in insects." *Biological Reviews* 64.4 (1989): 409-434.
25. Du, Zhenyong, et al. "Mitochondrial genomics reveals shared phylogeographic patterns and demographic history among three periodical cicada species groups." *Molecular biology and evolution* 36.6 (2019): 1187-1200.
26. Durden, Lance A., and Guy G. Musser. "The sucking lice (Insecta, Anoplura) of the world: a taxonomic checklist with records of mammalian hosts and geographical distributions. Bulletin of the AMNH; no. 218." (1994).
27. Eberle, M. W., and D. L. McLean. "Initiation and orientation of the symbiote migration in the human body louse Pediculus humanus L." *Journal of insect physiology* 28.5 (1982): 417-422.

28. Eberle, M. W., and D. L. McLean. "Observation of symbiote migration in human body lice with scanning and transmission electron microscopy." *Canadian Journal of Microbiology* 29.7 (1983): 755-762.
29. Fukatsu T, Koga R, Smith WA, Tanaka K, Nikoh N, Sasaki-Fukatsu K, Yoshizawa K, Dale C, Clayton DH. Bacterial endosymbiont of the slender pigeon louse, Columbicola columbae, allied to endosymbionts of grain weevils and tsetse flies. Appl Environ Microbiol. 2007 Oct;73(20):6660-8. doi: 10.1128/AEM.01131-07. Epub 2007 Aug 31. PMID: 17766458; PMCID: PMC2075037.
30. Gourlé, Hadrien, et al. "Simulating Illumina metagenomic data with InSilicoSeq." *Bioinformatics* 35.3 (2019): 521-522.
31. Heyworth, Eleanor R., and Julia Ferrari. "Heat stress affects facultative symbiont-mediated protection from a parasitoid wasp." *PLoS One* 11.11 (2016): e0167180.
32. Johnson, Kevin P., et al. "Simultaneous radiation of bird and mammal lice following the K-Pg boundary." *Biology letters* 14.5 (2018): 20180141.
33. Johnson, Kevin P., Scott M. Shreve, and Vincent S. Smith. "Repeated adaptive divergence of microhabitat specialization in avian feather lice." *BMC biology* 10.1 (2012): 1-11.
34. Johnson, Kevin P. "Genomic Approaches to Uncovering the Coevolutionary History of Parasitic Lice." *Life* 12.9 (2022): 1442.
35. JOHNSON, KP, and DH CLAYTON. "The biology, ecology, and evolution of chewing lice (pp. 449–476, 4 figs)." *PRICE RD, HELLENTHAL RA, PALMA RL, JOHNSON KP & DH CLAYTON (2003), The chewing lice: world checklist and biological overview. Illinois Natural History Survey Special Publication* 24 (2003).

36. Kaltenpoth, Martin, et al. "Symbiotic bacteria protect wasp larvae from fungal infestation." *Current Biology* 15.5 (2005): 475-479.

37. Kashkouli, M., Y. Fathipour, and M. Mehrabadi. "The Crucial Role of the Endosymbiont Pantoea sp. in Morphology and Mating of the Pistachio Green Stink Bug, Brachynema germari (Hemiptera: Pentatomidae)." *Journal of Agricultural Science and Technology* 23.1 (2021): 137-148.

38. Kashkouli, Marzieh, et al. "Characterization of a novel Pantoea symbiont allows inference of a pattern of convergent genome reduction in bacteria associated with Pentatomidae." *Environmental Microbiology* 23.1 (2021): 36-50.

39. Kikuchi, Yoshitomo, et al. "Symbiont-mediated insecticide resistance." *Proceedings of the National Academy of Sciences* 109.22 (2012): 8618-8622.

40. Kirkness, Ewen F., et al. "Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle." *Proceedings of the National Academy of Sciences* 107.27 (2010): 12168-12173.

41. Krzywinski, Martin, et al. "Circos: an information aesthetic for comparative genomics." *Genome research* 19.9 (2009): 1639-1645.

42. Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." *Nature methods* 9.4 (2012): 357-359.

43. Li, Dinghua, et al. "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph." *Bioinformatics* 31.10 (2015): 1674-1676.

44. Martin ML. Life history and habits of the pigeon louse (*Columbicola columbae* [Linnaeus])/ Thesis, William Marsh Rice Institute. 1933.

45. Masson, Florent, and Bruno Lemaitre. "Growing ungrowable bacteria: overview and perspectives on insect symbiont culturability." *Microbiology and Molecular Biology Reviews* 84.4 (2020): e00089-20.
46. McCutcheon, John P., and Nancy A. Moran. "Extreme genome reduction in symbiotic bacteria." *Nature Reviews Microbiology* 10.1 (2012): 13-26.
47. McCutcheon, John P., Bret M. Boyd, and Colin Dale. "The life of an insect endosymbiont from the cradle to the grave." *Current Biology* 29.11 (2019): R485-R495.
48. Moran, Nancy A., and Gordon M. Bennett. "The tiniest tiny genomes." *Annual review of microbiology* 68 (2014): 195-215.
49. Moran, Nancy A., John P. McCutcheon, and Atsushi Nakabachi. "Genomics and evolution of heritable bacterial symbionts." *Annual review of genetics* 42 (2008): 165-190.
50. Morrow, Jennifer L., Aidan AG Hall, and Markus Riegler. "Symbionts in waiting: the dynamics of incipient endosymbiont complementation and replacement in minimal bacterial communities of psyllids." *Microbiome* 5.1 (2017): 1-23.
51. Moya, Andrés, et al. "Learning how to live together: genomic insights into prokaryote–animal symbioses." *Nature Reviews Genetics* 9.3 (2008): 218-229.
52. Nadal-Jimenez, Pol, et al. "Genetic manipulation allows in vivo tracking of the life cycle of the son-killer symbiont, Arsenophonus nasoniae, and reveals patterns of host invasion, tropism and pathology." *Environmental microbiology* 21.8 (2019): 3172-3182.
53. Nakabachi, Atsushi, et al. "Defensive bacteriome symbiont with a drastically reduced genome." *Current biology* 23.15 (2013): 1478-1484.

54. Nelson, B. C., and M. D. Murray. "The distribution of Mallophaga on the domestic pigeon (Columba livia)." *International Journal for Parasitology* 1.1 (1971): 21-29.

55. Nováková, E., Hypša, V. & Moran, N.A. *Arsenophonus*, an emerging clade of intracellular symbionts with a broad host distribution. *BMC Microbiol* 9, 143 (2009). https://doi.org/10.1186/1471-2180-9-143

56. Nurk, Sergey, et al. "metaSPAdes: a new versatile metagenomic assembler." *Genome research* 27.5 (2017): 824-834.

57. Oliver, Kerry M., et al. "Bacteriophages encode factors required for protection in a symbiotic mutualism." *Science* 325.5943 (2009): 992-994.

58. Oliver, Kerry M., et al. "Facultative bacterial symbionts in aphids confer resistance to parasitic wasps." *Proceedings of the National Academy of Sciences* 100.4 (2003): 1803-1807.

59. Park, Jongsun, Si Hyeock Lee, and Ju Hyeon Kim. "Complete genome sequence of the endosymbiotic bacterium "Candidatus Riesia pediculicola"." *Microbiology Resource Announcements* 10.18 (2021): e01181-20.

60. Parks, Donovan H., et al. "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." *Genome research* 25.7 (2015): 1043-1055.

61. Perotti, M. Alejandra, et al. "Host-symbiont interactions of the primary endosymbiont of human head and body lice." *The FASEB Journal* 21.4 (2007): 1058-1066.

62. Price RD, Hellenthal RA, Palma RL, Johnson KP, Clayton DH. The chewing lice: world checklist and biological overview. Illinois Nat Hist Sur. 2003;Special Publication 24. X-501pp.

63. Puchta, Otto. "Experimentelle Untersuchungen über die Bedeutung der Symbiose der Kleiderlaus Pediculus vestimenti Burm." *Zeitschrift für Parasitenkunde* 17.1 (1955): 1-40.

64. Ries, Erich. "Die symbiose der läuse und federlinge." *Zeitschrift für Morphologie und Ökologie der Tiere* (1931): 233-367.

65. Říhová, Jana, et al. "Lightella neohaematopini: A new lineage of highly reduced endosymbionts coevolving with chipmunk lice of the genus Neohaematopinus." *Frontiers in microbiology* 13 (2022): 900312.

66. Shigenobu, Shuji, et al. "Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS." *Nature* 407.6800 (2000): 81-86.

67. Simpson, Jared T., et al. "ABySS: a parallel assembler for short read sequence data." *Genome research* 19.6 (2009): 1117-1123.

68. Smith, W.A., Oakeson, K.F., Johnson, K.P. *et al.* Phylogenetic analysis of symbionts in feather-feeding lice of the genus *Columbicola*: evidence for repeated symbiont replacements. *BMC Evol Biol* 13, 109 (2013). https://doi.org/10.1186/1471-2148-13-109

69. Snyder, Anna K., et al. "The phylogeny of Sodalis-like symbionts as reconstructed using surface-encoding loci." *FEMS microbiology letters* 317.2 (2011): 143-151.

70. Sudakaran, Sailendharan, Christian Kost, and Martin Kaltenpoth. "Symbiont acquisition and replacement as a source of ecological innovation." *Trends in Microbiology* 25.5 (2017): 375-390.

71. Villa, Scott M., et al. "Body size and fecundity are correlated in feather lice (Phthiraptera: Ischnocera): Implications for Harrison's Rule." *Ecol. Entomol* 43 (2018): 394-396.