**Hoppin’ Out of the Metagenome: *De Novo* Genome Assembly of the Obligate Leafhopper Endosymbiont *Candidatus* Sulcia muelleri**

Copeland HC, Lynch BG, Pollio AR, Spencer NJ

***1. Background***

*Candidatus* Sulcia muelleri (hereafter *Sulcia*)is an ancient bacterial symbiont of the sap-feeding insects of the suborder *Auchenorrhyncha* (*1*, *2*). This bacterium primarily colonizes specialized host cells, providing the host with amino acids to supplement nutritional deficiencies associated with a strict diet of plant xylem or phloem (*2*). *Sulcia* genomes, highly reduced from countless generations of isolation and bottlenecking within the host lineage, range in sizes from 191 to 227 kb and vary in genomic content, often biased toward the biosynthesis of seven or eight amino acids (*1*). Their insect hosts often supplement the degraded *Sulcia* symbiont with acquisition of lineage-specific secondary partners (*3*, *4*) .

Leafhoppers are common agricultural pests representing the major vector for many phytopathogens (*5*–*7*). For example, leafhopper genera associated phytopathogen transmission like *Dalbulus maidis* are responsible for significant impacts on corn yields. We aim to focus on the *Sulcia* symbiont associated with the sharpshooter leafhopper *Kolla paulula (K. paulula),* the recently discovered and efficient vector *Xylella fastidiosa*, the causative agent of Pierce’s disease in grapevines (*8*). The insect has become a major problem in all vineyards and distributed from medium to low altitude of its native range. Pierce’s disease has also become endemic in the United States, appearing in California and the southeast (*8*).

We propose the re-assembly of this vector’s endogenous *Sulcia* strain in order to compare total gene length to identify specific changes that might allow this leafhopper to become a more efficient vector. This will also provide insights into the effects of the host system on the evolution of genome-reduced obligate endosymbionts in a host group that is associated with diverse approaches to symbiosis.

***2. Methods***

*2.1 Data Retrieval and Sorting*

This genome assembly will be performed using publicly available sequencing reads from the NCBI Sequence Read Archive (SRA) under accession number SRX5063605. Sequencing was performed with Illumina MiSeq, resulting in 15.3 megabases of 151bp reads with quality scores averaging >30 (*9*). Reads were obtained from a single *K. paulula* individual, resulting in a mixture of metagenomic data from the leafhopper system.

This data set is composed of a mixture of both symbiont and host reads. Thus, in order to create a genomic assembly for either the host, or the symbiont we must bin, or separate, reads by origin. Many programs accomplish this goal by comparing to a reference genome. These programs utilize the ordered nature of any string of characters by mapping the suffix or ending characters of a long string. The resulting map, called a suffix trie, can more easily be compared based on node and leaf structure. This method is ideal because it allows smaller character strings to be found or not found within a longer string of characters.

Recent methods utilize new transformation techniques to simplify and compress a potentially very large suffix-trie. Allowing modern tools for read mapping alignments to be extremely fast, and not extremely computationally expensive. The most efficient way to do this is to take advantage of the Burrows-Wheeler Transformation (*10*, *11*). This transformation turns any long string of characters into a simpler more repetitive set. The resulting string is more easily compressed, thus is significantly less computationally taxing than other older methods. BWA, bowtie, and SOAPv2 utilize this transformation for mapping reads onto a given reference genome.

Here we will be using BWA due to its ease of use, and the familiarity we have with this tool. It has been shown that the difference in efficiency (I.E. speed) is an advantage not a disadvantage due to the higher accuracy of resulting output (*10*). This, as well as all subsequent computational analyses, will be run on WVU’s Spruce Knob High Performance Computing cluster to optimize runtime. The resulting SAM files will be converted to Fastq files using SAMtools (*10*, *12*, *13*). Fastq format is required for downstream steps.

*2.2 Quality Control and Sequence Trimming*

Once the *Sulcia* reads have been separated from the metagenome sample, they will be trimmed and filtered using the FastX software. Trimming is considered a necessary step to build confidence in the final assembly due to the removing of low quality reads or low quality ends (*14*). FastX can also remove adapter and linker sequences left over from the sequencing process. The resulting DNA reads will then undergo additional quality control analysis using FastQC to ensure that they are ready for assembly. While the chosen reads are 150 bases and have an average Phred quality score higher than 30, making quality drops at the end of each read a relatively minor concern, FastQC will provide additional confidence in data quality by identifying any specific regions, reads, or flow cells of questionable quality. FastQC will also identify any biases in the frequency or positioning of specific k-mers in the sequencing data to identify problems like high adapter content. If a substantial number of reads are of unusable quality and must be discarded, the small size of the *Sulcia* genome should provide sufficient flexibility for achieving acceptable coverage.

*2.3 Genome Assembly and Analysis*

Following trimming and quality control analysis of the raw reads, the *Sulcia* reads will be assembled *de novo* using the August 2019 release of ABySS (version 2.2.1). ABySS uses de Bruijn graphs, a common method for short-read sequence assembly that involves finding a path between each k-mer that only touches each edge once (*15*). Because of its parallel assembly capabilities, ABySS is a suitable choice for assembling the *Sulcia* genome and optimizing the necessary parameters within the relatively short time frame of our assembly project. For the initial assembly, an odd value of k (to avoid issues with palindromic sequences) at roughly half the read length will be used. Because of the miniscule nature of our genome of interest and the rapid parallel assembly capabilities of ABySS, we will run the actual assembly multiple times to optimize the value of k before analyzing the final assembly.

Once the final assembly is complete, it will be analyzed for quality and completeness. Comparison to assembled *Sulcia* reference genomes offers a straightforward approach. However, since we will not have access to additional laboratory resources for generating sequence to fill gaps, our result will be a set of contiguous sequences rather than a single closed genome. Our primary measures of assembly analysis will be the N50 value (the minimum length met by at least half of the contigs output by the assembler) and the length of the largest contig relative to the target genome size.

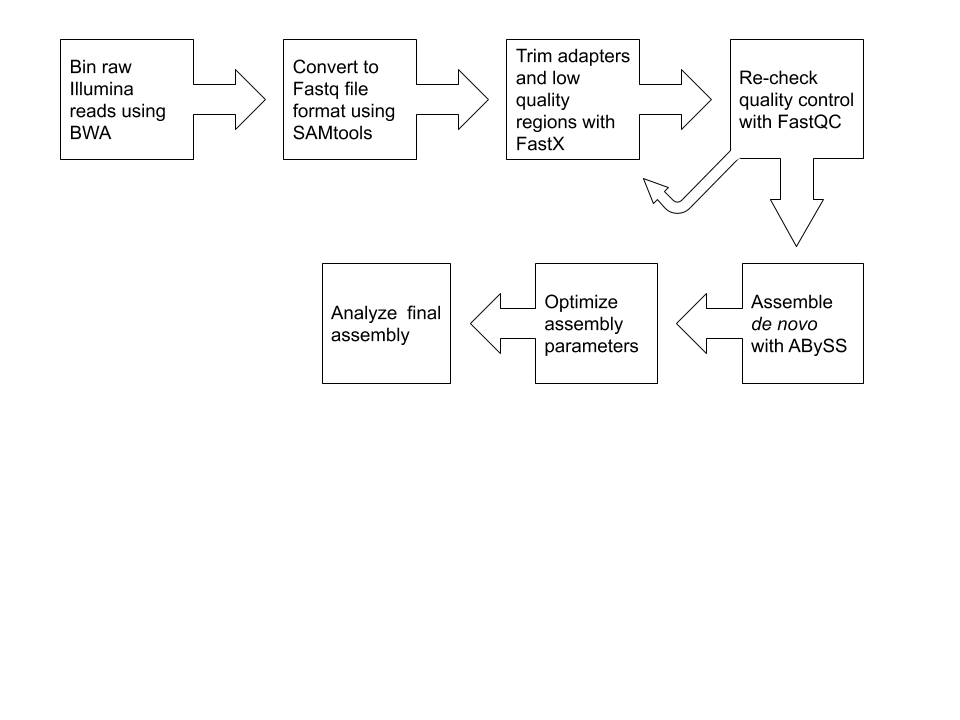
***3. Expected Outcomes***

As previously stated, our assembly alone will not result in a closed *Sulcia* genome. However, we expect our contiguous sequences to approach a total length of roughly 250kb as previously observed in *Sulcia* (*9*). Further alignment and annotation of the assembled genome would be expected to reveal genes associated with amino acid synthesis for host nutrition as well as relatively high conservation relative to other symbiotic lineages as is characteristic of ancient symbioses. However, genomic insights into host immunes processes and the evolutionary consequences of different microbiome dynamics will be the most useful in addressing our questions about *Xyella fastidiosa* transmission and *Sulcia* evolution.

***4. Investigator Roles***

Haylee Copeland will be responsible for writing and running all necessary scripts to bin the reads by organism with BWA and separate out *Sulcia*-associated reads. Adam Pollio will be responsible for writing and running scripts to convert BWA outputs to a Fastq file format with SAMtools and to trim reads based with FastX. Brianna Lynch will be responsible for quality control with FastQC, writing and running scripts as well as analyzing results and coordinating with Adam for additional or differential trimming to ensure adequate depth and quality of data for the final assembly. Noah Spencer will be responsible for designing and performing the final assembly with ABySS and optimizing parameters like k values to produce the best assembly based on the previously identified metrics for quality and completeness.

Collectively, the entire research team will be responsible for composing the final report on the success of the assembly pipeline. Troubleshooting any issues with the chosen bioinformatics software or the computing resources used will also be done collectively and internally. Resolution of any conflicts with regards to methodology will be deferred to the conventions of the literature and the individual strengths and experience of the researchers involved.

***5. Figures***

**Figure 1. *Sulcia* genome assembly pipeline.** A graphical representation of the proposed assembly workflow.

***References***

1. G. M. Bennett, N. A. Moran, Small , Smaller , Smallest : The Origins and Evolution of Ancient. **5**, 1675–1688 (2013).

2. J. P. McCutcheon, B. R. McDonald, N. A. Moran, *Proc. Natl. Acad. Sci.*, in press (available at http://www.pnas.org/content/106/36/15394.abstract).

3. H. Chang, S. Cho, M. C. Canale, S. T. Mugford, J. R. S. Lopes, S. A. Hogenhout, Complete Genome Sequence of “ Candidatus Sulcia muelleri ” ML , an Obligate Nutritional Symbiont of Maize Leafhopper ( Dalbulus maidis ). **3**, 1–2 (2015).

4. M. E. Brentassi, E. Franco, P. Balatti, R. Medina, F. Bernabei, A. M. Marino, D. R. Lenicov, Bacteriomes of the corn leafhopper , Dalbulus maidis ( DeLong Deltocephalinae ) harbor Sulcia symbiont : molecular characterization , ultrastructure , and transovarial transmission. *Protoplasma* (2016), doi:10.1007/s00709-016-1033-4.

5. S. A. Miller, T. Meulia, S. A. Hogenhout, Infection and replication sites of Spiroplasma kunkelii ( Class : Mollicutes ) in midgut and Malpighian tubules of the leafhopper Dalbulus maidis. **82**, 167–175 (2011).

6. E. A. G. E. Gasparich, D. G. Hall, S. A. Hogenhout, Spiroplasma-like organisms closely associated with the gut in five leafhopper species ( Hemiptera : Cicadellidae ), 35–44 (2011).

7. A. O. Jackson, R. G. Dietzgen, M. M. Goodin, J. N. Bragg, M. Deng, B IOLOGY OF P LANT R HABDOVIRUSES ∗ (2005), doi:10.1146/annurev.phyto.43.011205.141136.

8. S. Tuan, F. Hu, H. Chang, P. Chang, Y. Chen, Arthropods in Relation to Plant Disease Xylella fastidiosa Transmission and Life History of Two Cicadellinae Sharpshooters , Kolla paulula and Bothrogonia ferruginea ( Hemiptera : Cicadellidae ), in Taiwan, 1–7 (2016).

9. D. Genome, C. Sulcia, crossm, 1–2 (2019).

10. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows – Wheeler transform. **25**, 1754–1760 (2009).

11. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows – Wheeler transform. **26**, 589–595 (2010).

12. H. Li, A statistical framework for SNP calling , mutation discovery , association mapping and population genetical parameter estimation from sequencing data. **27**, 2987–2993 (2011).

13. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, G. P. Data, T. Sam, The Sequence Alignment / Map format and SAMtools. **25**, 2078–2079 (2009).

14. C. Del Fabbro, S. Scalabrin, M. Morgante, F. M. Giorgi, An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. **8**, 1–13 (2013).

15. S. D. Jackman, B. P. Vandervalk, S. Yeo, L. Coombe, ABySS 2 . 0 : Resource-Efficient Assembly of Large Genomes using a Bloom Filter. **50**, 1–24 (2016).