**Host Ecological Impact on SNP and Indel Formation in *Buchnera aphicidola* Strains**

Copeland HD, Lynch BG, Pollio AR, Spencer NJ

***1. Background and Significance***

*Buchnera aphidicola* (hereafter *Buchnera*), the primary endosymbiont for many species of aphid, nutritionally provisions amino acids to its host (*1–3*). The endosymbiont has specialized independently through millions of years of evolution, resulting in an extremely small genome (*4*). Each species of *Buchnera* has unique traits that are thought to be host-dependent. The divergent nature of aphid lineages has resulted in variation in *Buchnera* genome size and capabilities(*1*). In some lineages the gene reduction has resulted in the acquisition of a secondary endosymbiont. Though the difference in the smallest and largest *Buchnera* genomes has been well documented very little has been discussed in the differences between closely related species. This presents a critical gap in knowledge as the host-specific fine-tuning of genomes could have major implications for our understanding of host-microbe interactions and their effects on microbial evolution. Understanding the host’s effect on endosymbiotic evolution in the short term would also give insights into how organisms manipulate symbionts over an extended evolutionary scale.

Endosymbionts from many organisms show similar gene size, but limited similarity in genetic capability. For example, *Carsonella* and its host psyllid vary considerably based on host diet (*5, 6*). This is most apparent in genomic tools available to the psyllid’s secondary symbiont. Much like the psyllid, the aphid ecology varies depending on plant host, and environment. Due to the similarities between aphid and psyllid ecology we propose exploring the genomic dissimilarities between two known closely related *Buchnera* species. We intend to map raw metagenome sequence reads containing a known *Buchnera* strain from the aphid *Stegophylla sp*. to the genome of *Buchnera* from the pea aphid *Acyrthosiphon pisum* (*Buchnera* strain APS). By identifying SNPs and Indels we can make a more general statement on the effects the host has on endosymbiont genome conservation.

***2. Methods***

*2.1 Data Retrieval*

A FASTA file of the closed reference genome of *Buchnera* strain APS will be downloaded from NCBI Genome database (NC\_002528.1), and paired-end Illumina sequencing reads for the *Buchnera* endosymbiont of the aphid *Stegophylla sp*.will be acquired from the Sequence Read Archive (SRA) in a FASTQ file format (SRR7796613). While this dataset is large and metagenomic in nature, the SRA data set was used to build a complete genome with 256X coverage (ASM508078), justifying the presence of sufficient *Buchnera* sequence for ample coverage during mapping.

*2.2 Quality Control*

To ensure proper mapping of the reads to the reference genome, it is important to ensure that the data acquired are appropriate. This means being in the correct format (a FASTQ file with read pairs separated into individual reads) and being consistent with our expectations based on the associated information provided on the SRA (length and number of reads, sequencing depth based on total amount of sequence). It will also be important to ensure that all reads used have high Phred quality scores (which represent the sequencer’s confidence in the base call) uniformly across the reads and that extraneous sequences like Illumina adapters are absent.

To accomplish this, we will run the forward and reverse end reads through FastQC, a program that analyzes the sequences for quality of reads, trends in quality across the lengths of reads, GC content distribution, read length distribution and overrepresented sequences like adapters in addition to providing general summary statistics like average read length and total base pairs of sequence data. GC content distribution, specifically, will provide important insights into the nature of our sample because our *Buchnera* target has an AT-rich genome.

If the basic summary statistics do not match those described on the SRA, it may be necessary to return to the data acquisition step and ensure that the appropriate data sets are being used. Otherwise, our main concern will be with quality scores. Our chosen data set contains 3.9Gb of metagenomic sequence. Despite the presence of insect sequence and other non-*Buchnera* data, the large number of sequences are relatively small size of the *Buchnera* genome means that ample coverage can be obtained based on the previously cited genome assembly. Even if only 1% of the reads mapped to *Buchnera*, they would still offer 50X coverage of the genome as a whole.

Based on the trends we observe in our FastQC analysis, we will use Trimmomatic to trim reads back from the ends until average Phred scores across reads are above our threshold of 28. We will verify this by analyzing the trimmed reads through a second round of FastQC quality control.

*2.3 Mapping*

Quality-controlled reads will be mapped to the reference genome using the Burrows-Wheeler Aligner (BWA) (*7*). BWA and other mapping software take advantage of the Burrows-Wheeler transform, which can be used to losslessly compress strings of characters, allowing for rapid and computationally inexpensive analyses of larger data sets. This process is critical to our first step of mapping, in which we will use BWA-MEM (the fastest and most accurate algorithm in the BWA software package for Illumina reads >70bp) to construct an index for each of the reference genomes. The sequencing reads can then be mapped to the reference genome, outputting a sequence alignment map (SAM) file.

Based on the relationship of our genomes, we expect a high proportion of reads to map, although the metagenomic nature of our reads will inevitably lead to some proportion of reads corresponding to other organisms not mapping. If the BWA output indicates that unexpectedly few reads mapped, it may be necessary to revisit quality control or to adjust the mapping parameters to allow greater sequence differences during mapping. It’s also important to note that the AT bias and other intrinsic features of the *Buchnera* genome will theoretically prevent the mapping of other sequences to the reference genome, which would lead to chimeric mapping and produce false positives during variant calling. The previous experience of our group using BWA and troubleshooting issues associated with metagenomic datasets will be of utility in troubleshooting any issues that arise as a result of the nature of the data. BWA Mapping and variant identification analysis will make use of WVU’s Spruce High Performance Compute Cluster.

*2.4 Variant Identification*

To compare these symbiont genomes and determine patterns of co-conservation or covariation, we must identify genotypic changes in the primary and secondary symbionts from each system in the form of SNPs and indels. First, the SAM files will be converted to a compressed binary format (BAM) using Samtools (*8*). Next, another distribution from the Samtools suite, BCFtools, will be used to call the variants themselves by generating genotype frequencies at each position and then calling true variants based on likelihood. Our coverage cutoffs for variant identification will be informed by the mapping rate observed in the BWA analysis.

The output will be a binary variant call format (BCF) file, which can be converted to an uncompressed variant call format (VCF) file. The VCF file can then be visualized with the Integrative Genomics Viewer (IGV), which will visually put the identified variants between symbiont strains in the contexts of the complete reference genomes. This lends itself to downstream analysis of pathways and biological functions associated with the adaptation of these endosymbiont lineages to diverging host systems.

***3. Expected Outcomes***

We expect to have a defined SNP map with locations within the genome. Further, we expect to have information regarding the specific sequences that have been deleted or inserted as compared to our reference genome and their distribution. We anticipate very few SNPs or indels in amino acid synthesis due to the importance of this function in all aphid symbioses. The visualization of the mapping and variant calls will provide a solid basis for any and all downstream analysis, although full functional annotation of all variants will likely not be feasible within our timescale.

***4. Investigator Roles***

Noah Spencer and Brianna Lynch will be responsible for data acquisition and quality control, including FastQC analysis and trimming. Brianna was previously involved in quality control, and Noah has important background in these organisms that will inform expectations during FastQC analysis. Together, they will be able to handle the technical aspects of quality control and the critical decisions involved in assessing the data.

Haylee Copeland will be responsible for writing scripts for BWA mapping to expand on her current skill set, with Adam providing assistance based on his previous experience and Noah providing troubleshooting support.

Adam Pollio will be responsible for variant identification and visualization, drawing from his previous experience in visualization and analysis of mapping data and his background in comparative genomics. Any further downstream analysis, report composition, and large-scale troubleshooting will be completed as a collective.

***References***

1. L. Klasson, S. G. E. Andersson, Evolution of minimal-gene-sets in host-dependent bacteria. *Trends Microbiol.* **12**, 37–43 (2004).

2. N. A. Moran, A. Mira, The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. *Genome Biol.* **2**, RESEARCH0054 (2001).

3. A. E. Douglas, Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria Buchnera. *Annu. Rev. Entomol.* **43**, 17–37 (1998).

4. H. Charles, D. Mouchiroud, J. Lobry, I. Goncalves, Y. Rahbe, Gene size reduction in the bacterial aphid endosymbiont, Buchnera. *Mol. Biol. Evol.* **16** (1999), pp. 1820–1822.

5. D. B. Sloan, N. A. Moran, Genome Reduction and Co-evolution between the Primary and Secondary Bacterial Symbionts of Psyllids Research article. **29**, 3781–3792 (2012).

6. M. L. Thao, M. A. Clark, L. Baumann, E. B. Brennan, N. A. Moran, P. Baumann, Secondary Endosymbionts of Psyllids Have Been Acquired Multiple Times. **41**, 300–304 (2000).

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

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