**Indel Detection in Divergent *Buchnera aphidicola* Lineages and the Challenges of Evaluating Depth**

Copeland HD, Lynch BG, Pollio AR, Spencer NJ

***1. Response to reviewers***

The reviewers were unclear with the specific goals of our project, and how our specific goals are addressed by our methods. They specifically pointed out that the foundation of our project did not have a clear evolutionary relevance based on the information provided within our proposal. Further, our methods did not clearly address the gap presented in the proposal. We have attempted to restate our gap and significance with additional information from the field in order to help relieve the dicidence that was presented by our reviewers.

*1.1 Introductions*

*Buchnera* is a common nutritional endosymbiont that has been conserved over thousands of years within many aphid lineages (Brinza et al. 2009; Douglas 1998; Charles et al. 1999; Klasson and Andersson 2004). The resulting phylogeny is congruent with that of the host. All *Buchnera* species have undergone gene reduction, as is common among nutritional symbionts, resulting in closer aphid species having symbionts that appear to be more similar (Chong et al. 2019). Further, specific genes are conserved across all *Buchnera* species. Suggesting that there are necessary genes for this organism, either due to its highly specialized role or due to the nature of bacterial life (Klasson and Andersson 2004). Some of these genes are known to be responsible for amino acid synthesis within the microecology of aphids (Charles et al. 1999; Douglas 1998). In specific aphid species, some amino acid synthesis is further supported by more recent symbiont acquisition (Chong and Moran 2018; von Dohlen et al. 2017). The combination of host diversification and local microecology puts novel pressure on the genome. In these unique environments it is yet unknown how well these ubiquitous genes are conserved is yet unknown, presenting a critical gap.

In order to address this gap we compared two distant species of *Buchnera.* Our goal was to identify variants between the two lineages within genes that are conserved in all *Buchnera* species.

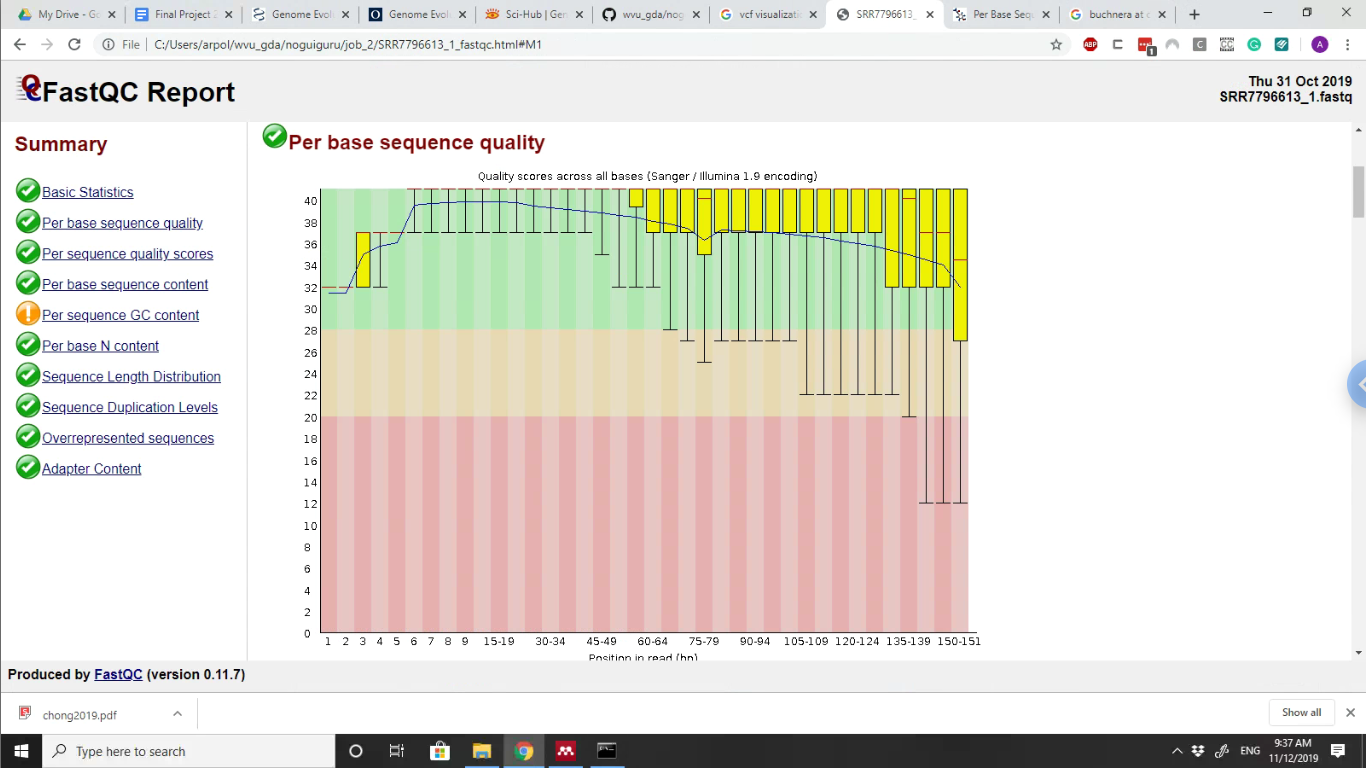
***2. Results***

*2.1 Data Acquisition, Quality Control, and Processing*

Raw sequence reads from the chosen metagenomic dataset representing the aphid *Stegophylla sp.* and its associated microbes (SRR7796613) were downloaded from the SRA using the prefetch utility from the SRA Toolkit. The SRA Toolkit program fastq-dump was used to convert the resulting interleaved SRA file into two FastQ files.

Analysis with FastQC verified that set of reads contained 9398774. The FastQC report also showed that per base sequence quality declined at read ends, with mean Phred quality scores ranging from 30-35 and interquartile ranges from 26-40. While we initially planned to assess the composition of this metagenomic dataset in part by examining trends in GC content distribution, the distribution ultimately resembled a normal distribution around a %GC of 27, which, given our reference *Buchnera*’s GC content of ~26% (Shigenobu et al. 2000) and the aphid host’s gene content of ~29% (Consortium 2010), made it difficult to visually distinguish the relative representation of the insect and microbial partners.

Based on these findings, we sought a baseline level of mapping to the reference genome by running a preliminary mapping of the raw reads to the reference genome from *Buchnera aphidicola* strain APS (NC\_002528.1) with BWA-MEM. Issues with the initial run revealed that the use of the -I argument during the fastq-dump file format conversion had appended characters to the read IDs in the resulting FastQ files based on whether they represented forward or reverse reads. This made the read pairs intolerable to BWA, so a simple sed command was used to remove all of the appended identifiers based on a regular expression that identified the tags at the ends of the read identifiers.

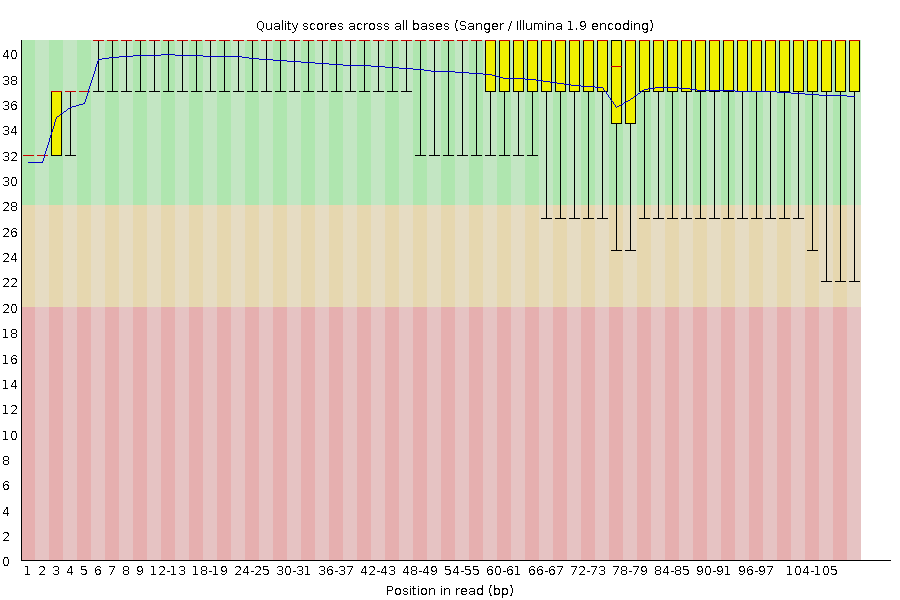
**

**Figure 1. Preliminary FastQC Per-Base Sequence Quality.** The distribution of Phred quality scores at each position for all forward reads is shown. Per-base quality scores for reverse reads are not shown, but follow a similar trend.

*2.2 Preliminary Mapping and Trimming*

Our initial mapping of the untrimmed forward and reverse reads to our reference genome (NC\_002528.1) used the BWA-MEM algorithm with 12 threads, a maximum MEM occurance of 50000, and a mismatch penalty of three. For this assembly, the other mapping values and penalties and the final mapping score threshold were left at the default settings. The “-P” parameter was also used, which prevents BWA-MEM from trying to find hits that fit a proper pair. The SamTools view and sort functions were used to convert the SAM file output of this mapping to an ordered BAM file so that the mapping rate could be assessed using SamTools flagstat, which reads the flags in a BAM file to output mapping rates, among other useful statistics.

Initially, no mapping rate outputs were produced. This was fixed by setting the output of SamTools sort to “buchnera.map.sorted” as opposed to “buchnera.map.sorted.bam,” as SamTools already assumes a .bam output and appends those characters to the specified output. This ultimately resulted in an extremely low mapping rate of 0.28% or 28376 reads. We suspected that low quality at read ends could be responsible in part for our low mapping rate, since previous work had achieved over 200X coverage of a *Buchnera* genome with the data set used (ASM508078). To test whether this was the case, we used Trimmomatic to crop the reads (taking care to use the processed read set with appended tags removed) to a conservative length of 110. This threshold is somewhat arbitrary relative to the observed per-base quality statistics, but FastQC analysis of the trimmed reads indicated that low quality read ends were removed by this trimming (Figure 2).



**Figure 2. Per-base sequence quality after trial trimming.** The The distribution of Phred quality scores at each position for all forward reads is shown, as in Figure 1. Notably, all IQRs are above 30. This is also true of the reverse read set (not shown).

To determine the effects of this trimming, trimmed reads were mapped in the same manner as the untrimmed reads, and the mapping rate was assessed. This still showed only 37146 sequences mapping, and this increase came at the cost of nearly 27% of overall sequencing depth. To guide the next set of adjustments, we first determined whether a few high-coverage regions in these mapping might be usable for calling some variants by running variant calling with BCFTools despite the low mapping rate.

*2.3 Preliminary Variant Calling*

To assess the potential efficacy of the preliminary maps from untrimmed and trimmed sequence data, we conducted variant calling using BCFTools. First, the input sorted BAM files were indexed with samtools index. Genotype likelihoods were then generated for each map using bcftools mpileup. Orphaned reads not matched with a paired read were counted. Maximum depth was set to 10000, although no coverage even approaching such a value was anticipated.

During the actual variant calling with bcftools call, the variant call format (VCF) file output was set to display only the variants (-v) and to display all alternate alleles for each variant, including those not found in any of the genotypes (-A). Importantly, this round of variant calling used the multiallelic calling model (-m) as opposed to bcftools’s original consensus calling method (-c). A Perl script was used to produce a plaintext table of these variants and the associated features in the target genome using a general feature format (GFF) file for the reference genome.

The variant calling using the untrimmed reads produced a total of three variant calls of which all were indels rather than SNPs. Variant calling using the trimmed data had similar results, calling those three variants as well as one additional indel. Two of these indel calls showed high values for DP (depth) and IMF (the maximum fraction of reads supporting an indel) with values of 99 and 92 for DP and 1 and 0.978261 for IMF. These calls were present in both of the preliminary variant calling tests. The other two calls had considerable lower values for both of these metrics with 8 and 5 for DP and 0.5 and 0.6 for IMF.

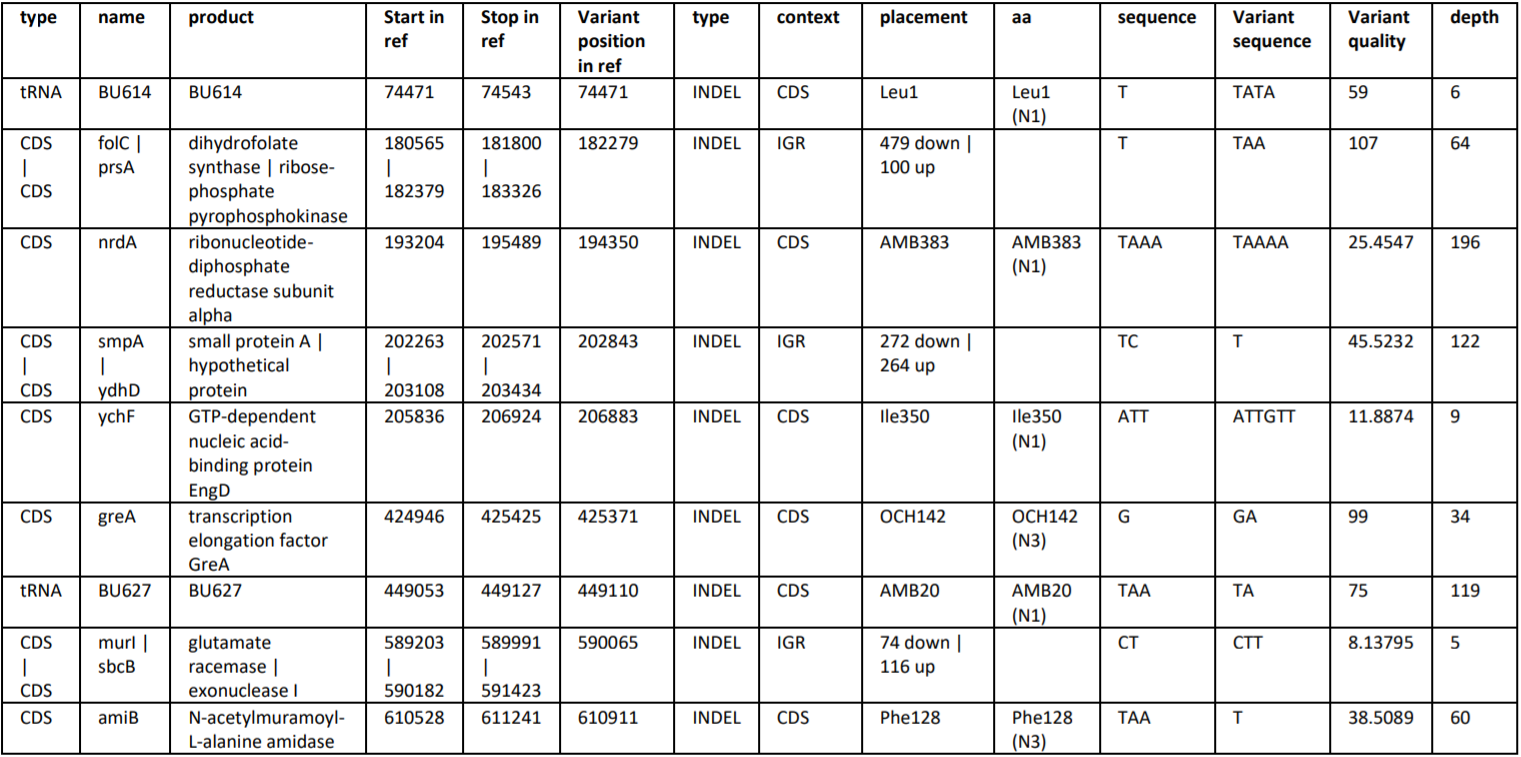
Based on the evolutionary time and differences in genome size between these organisms, high conservation of *Buchnera* genes was insufficient to describe these results. To ensure that the variant calling process was not responsible for our lack of depth and few identified variants, we ran the variant calling portion of the pipeline again for the trimmed read set using the original consensus calling method along with very lax parameters (p value threshold = 0.1). This had no effect. In an attempt to rescue some of the expected depth and identify additional variants, we trimmed the data instead to 125bp to maintain high quality (seen across that range of bases in Figure 1) without sacrificing as much coverage and then mapping with more lax parameters to increase depth for variant calling.

*2.4 Final Trimming and Variant Identification*

As expected, the FastQC report showed no decrease in quality in the read set cropped to 125 bases relative to that cropped to 110 bases. These reads were mapped as before, but this time increasing the sensitivity of the mapping by changing the parameter -T (the minimum mapping “score” for the output alignment) to 25 and 20 as opposed to the default 30. The mapping with -T 25 resulted in 71903 sequences mapping, and the -T 20 mapping resulted in a huge increase with 762787 sequences mapping. These were then used to perform variant calling as before, and the vcf output was used to generate the final table of identified variants as described.

Despite the massive discrepancy in mapped sequence between these two BWA-MEM runs, the latter only yielded one additional variant call relative to the former (nine in total versus eight). While some showed low depth as before, six of the nine variants showed depths greater than 30. All called variants were still indels (Table 1). Variants with higher depth occurring in coding sequences were of particular interest. Annotations of the associated genes were largely to do with basic cellular functions like transcription and translation, although too few variants were ultimately called to identify any major trends.

**Table 1. Summary of final variants called.** In the final round of variant calling, nine variants were called, all of which were indels. Only six of these showed depth greater than 30. The four of those that occur in coding sequences are found in *nrdA, greA,* tRNA BU627, and *amiB*.



***3. Discussion***

Aphids are an ecologically diverse species of insect. Their endosymbiont has been evolving within the diversification of aphid species. Much about the differences in the genome structure of these divergent lineages has been explored (Chong et al 2018). Our study aimed to identify the genes shared by aphid species, by comparing two distantly related *Buchnera.* Our work showed little differences between the species. While the metagenomic nature of our data did lead to a considerable cost to coverage, some of our later mappings indicated that usable overall depth may have been achieved. The identified indels in these basic genes, including a tRNA, may indicate ongoing evolutionary trends in these highly-reduced organisms.

The presence of only indels was of note but may simply reflect another effect of low overall coverage that made SNP-calling impossible. Even when we were able to increase our overall mapping rate considerably, few variants could be identified. As a result, we omitted our proposed map of variants in the genome. This low variant number may be due in part to an uneven distribution of coverage across the genome, resulting in only a small number of high-coverage regions being useful for variant calling. Unique features of the *Buchnera* genomes like its AT-richness may have played a role in coverage bias. The lineages of *Buchnera* used were also separated by considerable evolutionary time and had evolved different genome sizes, while may have led to difficulties in mapping due to larger, structural changes that this approach would not be sensitive to. Ultimately, analysis of per-sequence mapping rates across the reference genome could help further develop our understanding of this problem and inform future directions and changes.

Despite not outputting a large or high-quality set of genomic variants, this investigation highlighted some important lessons and challenges associated with these types of analysis. Broadly, they can be understood as the problem of understanding arguments. Bioinformatics, despite the computational nature of the techniques, involved human input at nearly every step. Decisions like including the -I argument when running fastq-dump have implications in subsequent steps, and the outputs of later steps like mapping or variant calling cannot automatically be considered representative of what is in the data. A bioinformatics pipeline is, in reality, a non-linear structure that requires constant tweaking and investigation to determine the effects of changes to the code on the outputs by something resembling a scientific approach as well as a reasoned perspective on the implications of those changes and whether they will result in meaningful outputs.

Still, lessons on the science of choosing data sets are apparent here as well. While a data set may have been used for an acceptable genome assembly, those reads cannot be guaranteed to be useful for other applications like mapping variants against a target genome from a different isolate. This has to do both with the issue of uneven coverage as described before as well as the issue of the relative levels of coverage required for these two tasks in general. This can be a cause of great frustration in the attempt to achieve some output, but the most important lessons are learned in the process when deciding what that output is expected to look like and what constitutes a meaningful result.

***4. Pipeline Code***

*4.1 Data File Format Conversion with SRA Toolkit*

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

prefetch -v SRR7796613

fastq-dump --outdir /opt/fastq/ --split-files /home/hdc0001/ncbi/public/sra/SRR7796613.sra

fastq-dump -A SRR7796613 -D SRR7796613.sra

fastq-dump -split-files -I SRA7796613

*4.2 Quality Control Checking with FastQC*

Script submitted to spruce using qsub fastqc.sh:

#PBS -q debug

#PBS -lwalltime=00:05:00

#PBS -nodes=1:ppn=12

#PBS -N fastqc\_buchenera1

#PBS -m arp -M hdc0001@mix.wvu.edu

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

fastqc SRR7796613\_1.fastq SRR7796613\_2.fastq

conda deactivate

*4.3 Modification of Read IDs with sed*

Script submitted to spruce using qsub readids.sh

#PBS -q debug

#PBS -lwalltime=00:05:00

#PBS -lnodes=1:ppn=12

#PBS -N readids

#PBS - m arp -M njs0012@mix.wvu.edu

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

sed -E "s/^((@|\+)SRR[^.]+\.[^.]+)\.(1|2)/\1/" SRR7796613\_1.fastq > SRR7796613\_1.fixed.fastq

sed -E "s/^((@|\+)SRR[^.]+\.[^.]+)\.(1|2)/\1/" SRR7796613\_2.fastq > SRR7796613\_2.fixed.fastq

conda deactivate

*4.3 Trimming Preliminary Mapping with BWA-MEM*

Script submitted to spruce using qsub maptest.sh

#PBS -q comm\_mmem\_week

#PBS -lwalltime=10:00:00

#PBS -lnodes=1:ppn=12

#PBS -N mappingtest

#PBS - m arp -M njs0012@mix.wvu.edu

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

bwa index buchnera\_ref.fasta

bwa mem -t 12 -c 50000 -P -B 3 buchnera\_ref.fasta SRR7796613\_1.fixed.fastq SRR7796613\_2.fixed.fastq > buchnera.map.sam

samtools view -o buchnera.map.bam -bS buchnera.map.sam

samtools sort buchnera.map.bam buchnera.map.sorted

samtools flagstat buchnera.map.sorted.bam

exit 1

Script submitted to spruce using qsub maptest.110.sh

trimmomatic PE -threads 6 SRR7796613\_1.fixed.fastq

SRR7796613\_2.fixed.fastq -baseout buchnera.110.fastq CROP:110 MINLEN:110

fastqc buchnera.110.fastq

bwa mem -t 12 -c 50000 -P -B 3 buchnera\_ref.fasta

buchnera.110\_1P.fastq buchnera.110\_2P.fastq > buchnera.110.map.sam

samtools view -o buchnera.110.map.bam -bS buchnera.110.map.sam

samtools sort buchnera.110.map.bam buchnera.110.map.sorted

samtools flagstat buchnera.110.map.sorted.bam

exit

*4.4 Preliminary Variant Calling with BCFTools*

Script submitted to spruce using qsub varcalluntrimmed.sh

#PBS -q debug

#PBS -lwalltime=00:05:00

#PBS -lnodes=1:ppn=12

#PBS -N varcalluntrimmed

#PBS - m arp -M njs0012@mix.wvu.edu

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

samtools index buchnera.map.sorted.bam buchnera.map.sorted.bai

bcftools mpileup --threads 12 -6 -A -d 10000 -f buchnera\_ref.fasta

buchnera.map.sorted.bam -O b -o buchnera.bcf

bcftools call -o buchnera.vcf -O v --threads 12 -v -A --ploidy 1 -m buchnera.bcf

./vcf2table.pl buchnera\_ref.gff3 -g buchnera\_ref.fasta < buchnera.vcf > buchnera.variant\_table.txt

exit 1

Script submitted to spruce using qsub varcall.110.sh

#PBS -q debug

#PBS -lwalltime=00:05:00

#PBS -lnodes=1:ppn=12

#PBS -N varcall.110

#PBS - m arp -M njs0012@mix.wvu.edu

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

samtools index buchnera.110.map.sorted.bam buchnera.110.map.sorted.bai

bcftools mpileup --threads 12 -6 -A -d 10000 -f buchnera\_ref.fasta

buchnera.110.map.sorted.bam -O b -o buchnera.110.bcf

bcftools call -o buchnera.110.vcf -O v --threads 12 -A -v -m --ploidy 1 buchnera.110.bcf

./vcf2table.pl buchnera\_ref.gff3 -g buchnera\_ref.fasta < buchnera.110.vcf > buchnera.110.variant\_table.txt

exit 1

Script submitted to spruce using qsub varcall.110c.sh

#PBS -q debug

#PBS -lwalltime=00:05:00

#PBS -lnodes=1:ppn=12

#PBS -N varcall.110c

#PBS - m arp -M njs0012@mix.wvu.edu

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

samtools index buchnera.110.map.sorted.bam buchnera.110.map.sorted.bai

bcftools mpileup --threads 12 -6 -A -d 10000 -f buchnera\_ref.fasta buchnera.110.map.sorted.bam -O b -o buchnera.110.bcf

bcftools call -o buchnera.110.c.vcf -O v --threads 12 -A -v -c -g 3 -p 0.1 --ploidy 1 buchnera.110.bcf

./vcf2table.pl buchnera\_ref.gff3 -g buchnera\_ref.fasta < buchnera.110.c.vcf > buchnera.110.c.variant\_table.txt

exit 1

*4.4 Final Trimming and Variant Identification*

Script submitted to spruce using qsub map125.25.20.sh

#PBS -q comm\_mmem\_day

#PBS -lwalltime=24:00:00

#PBS -lnodes=1:ppn=12

#PBS -N map125.25.20

#PBS - m arp -M njs0012@mix.wvu.edu

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

fastqc buchnera.125\_1P.fastq buchnera.125\_2P.fastq

bwa mem -t 12 -c 50000 -P -B 3 -T 25 buchnera\_ref.fasta

buchnera.125\_1P.fastq buchnera.125\_2P.fastq > buchnera.125.25.map.sam

samtools view -o buchnera.125.25.map.bam -bS buchnera.125.25.map.sam

samtools sort buchnera.125.25.map.bam buchnera.125.25.map.sorted

samtools flagstat buchnera.125.25.map.sorted.bam

samtools index buchnera.125.25.map.sorted.bam buchnera.125.25.map.sorted.bai

bcftools mpileup --threads 12 -6 -A -d 10000 -f buchnera\_ref.fasta

buchnera.125.25.map.sorted.bam -O b -o buchnera.125.25.bcf

bcftools call -o buchnera.125.25.vcf -O v --threads 12 -A -v -m --ploidy 1 buchnera.125.25.bcf

./vcf2table.pl buchnera\_ref.gff3 -g buchnera\_ref.fasta < buchnera.125.25.vcf > buchnera.125.25.variant\_table.txt

bwa mem -t 12 -c 50000 -P -B 3 -T 20 buchnera\_ref.fasta buchnera.125\_1P.fastq buchnera.125\_2P.fastq > buchnera.125.20.map.sam

samtools view -o buchnera.125.20.map.bam -bS buchnera.125.20.map.sam

samtools sort buchnera.125.20.map.bam buchnera.125.20.map.sorted

samtools flagstat buchnera.125.20.map.sorted.bam

samtools index buchnera.125.20.map.sorted.bam buchnera.125.20.map.sorted.bai

bcftools mpileup --threads 12 -6 -A -d 10000 -f buchnera\_ref.fasta

buchnera.125.20.map.sorted.bam -O b -o buchnera.125.20.bcf

bcftools call -o buchnera.125.20.vcf -O v --threads 12 -A -v -m --ploidy 1 buchnera.125.20.bcf

./vcf2table.pl buchnera\_ref.gff3 -g buchnera\_ref.fasta < buchnera.125.20.vcf > buchnera.125.20.variant\_table.txt

exit 1

***References***

Brinza, Lilia, José Viñuelas, Ludovic Cottret, Federica Calevro, and Yvan Rahbé. 2009. “Systemic Analysis of the Symbiotic Function of Buchnera Aphidicola , the Primary Endosymbiont of the Pea Aphid Acyrthosiphon Pisum.” *C. R. Biologies* 332 (11): 1034–49. https://doi.org/10.1016/j.crvi.2009.09.007.

Charles, H, D Mouchiroud, J Lobry, I Goncalves, and Y Rahbe. 1999. “Gene Size Reduction in the Bacterial Aphid Endosymbiont, Buchnera.” *Molecular Biology and Evolution*. United States. https://doi.org/10.1093/oxfordjournals.molbev.a026096.

Chong, Rebecca A., and Nancy A. Moran. 2018. “Evolutionary Loss and Replacement of Buchnera, the Obligate Endosymbiont of Aphids.” *The ISME Journal*. https://doi.org/10.1038/s41396-017-0024-6.

Consortium, Internation aphid. 2010. “Genome Sequence of the Pea Aphid Acyrthosiphon Pisum” 8 (2). https://doi.org/10.1371/journal.pbio.1000313.

Dohlen, Carol D. von, Usha Spaulding, Kistie B. Patch, Kathryn M. Weglarz, Robert G. Foottit, Nathan P. Havill, and Gaelen R. Burke. 2017. “Dynamic Acquisition and Loss of Dual-Obligate Symbionts in the Plant-Sap-Feeding Adelgidae (Hemiptera: Sternorrhyncha: Aphidoidea).” *Frontiers in Microbiology* 8 (JUN): 1–15. https://doi.org/10.3389/fmicb.2017.01037.

Douglas, A E. 1998. “Nutritional Interactions in Insect-Microbial Symbioses: Aphids and Their Symbiotic Bacteria Buchnera.” *Annual Review of Entomology* 43: 17–37. https://doi.org/10.1146/annurev.ento.43.1.17.

Klasson, Lisa, and Siv G E Andersson. 2004. “Evolution of Minimal-Gene-Sets in Host-Dependent Bacteria.” *Trends in Microbiology* 12 (1): 37–43.

Shigenobu, Shuji, Hidemi Watanabe, and Masahira Hattori. 2000. “Genome Sequence of the Endocellular Bacterial Symbiont of Aphids Buchnera Sp . APS,” 81–86.