Deep Sequencing of the Transcriptomes of Soybean Aphid and Associated Endosymbionts

Sijun Liu  
*Iowa State University*, sliu@iastate.edu

Nanasheb P. Chougule  
*Iowa State University*, chougule@iastate.edu

Diveena Vijayendran  
*Iowa State University*, diveena@iastate.edu

Bryony C. Bonning  
*Iowa State University*, bbonning@iastate.edu

Follow this and additional works at: [http://lib.dr.iastate.edu/ent_pubs](http://lib.dr.iastate.edu/ent_pubs)

Part of the [Entomology Commons](http://lib.dr.iastate.edu/ent_pubs)

The complete bibliographic information for this item can be found at [http://lib.dr.iastate.edu/ent_pubs/4](http://lib.dr.iastate.edu/ent_pubs/4). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).
Deep Sequencing of the Transcriptomes of Soybean Aphid and Associated Endosymbionts

Abstract

Background

The soybean aphid has significantly impacted soybean production in the U.S. Transcriptomic analyses were conducted for further insight into leads for potential novel management strategies.

Methodology/Principal Findings

Transcriptomic data were generated from whole aphids and from 2,000 aphid guts using an Illumina GAII sequencer. The sequence data were assembled de novo using the Velvet assembler. In addition to providing a general overview, we demonstrate (i) the use of the Multiple-k/Multiple-C method for de novo assembly of short read sequences, followed by BLAST annotation of contigs for increased transcript identification: From 400,000 contigs analyzed, 16,257 non-redundant BLAST hits were identified; (ii) analysis of species distributions of top non-redundant hits: 80% of BLAST hits (minimum e-value of 1.0-E3) were to the pea aphid or other aphid species, representing about half of the pea aphid genes; (iii) comparison of relative depth of sequence coverage to relative transcript abundance for genes with high (membrane alanyl aminopeptidase N) or low transcript abundance; (iv) analysis of the Buchnera transcriptome: Transcripts from 57.6% of the genes from Buchnera aphidicola were identified; (v) identification of Arsenophonus and Wolbachia as potential secondary endosymbionts; (vi) alignment of full length sequences from RNA-seq data for the putative salivary gland protein C002, the silencing of which has potential for aphid management, and the putative Bacillus thuringiensis Cry toxin receptors, aminopeptidase N and alkaline phosphatase.

Conclusions/Significance

This study provides the most comprehensive data set to date for soybean aphid gene expression: This work also illustrates the utility of short-read transcriptome sequencing and the Multiple-k/Multiple-C method followed by BLAST annotation for rapid identification of target genes for organisms for which reference genome sequences are not available, and extends the utility to include the transcriptomes of endosymbionts.

Disciplines

Entomology

Comments

This article is from PLoS ONE 7, no. 9 (2012): e45161, doi:10.1371/journal.pone.0045161.

Rights

This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

This article is available at Iowa State University Digital Repository: http://lib.dr.iastate.edu/ent_pubs/4
Deep Sequencing of the Transcriptomes of Soybean Aphid and Associated Endosymbionts

Sijun Liu, Nanasaheb P. Chougule, Diveena Vijayendran, Bryony C. Bonning*

Department of Entomology, Iowa State University, Ames, Iowa, United States of America

Abstract

Background: The soybean aphid has significantly impacted soybean production in the U.S. Transcriptomic analyses were conducted for further insight into leads for potential novel management strategies.

Methodology/Principal Findings: Transcriptomic data were generated from whole aphids and from 2,000 aphid guts using an Illumina GAII sequencer. The sequence data were assembled de novo using the Velvet assembler. In addition to providing a general overview, we demonstrate (i) the use of the Multiple-k/Multiple-C method for de novo assembly of short read sequences, followed by BLAST annotation of contigs for increased transcript identification: From 400,000 contigs analyzed, 16,257 non-redundant BLAST hits were identified; (ii) analysis of species distributions of top non-redundant hits: 80% of BLAST hits (minimum e-value of 1.0-E3) were to the pea aphid or other aphid species, representing about half of the pea aphid genes; (iii) comparison of relative depth of sequence coverage to relative transcript abundance for genes with high (membrane alanyl aminopeptidase N) or low transcript abundance; (iv) analysis of the Buchnera transcriptome: Transcripts from 57.6% of the genes from Buchnera aphidicola were identified; (v) identification of Arsenophonus and Wolbachia as potential secondary endosymbionts; (vi) alignment of full length sequences from RNA-seq data for the putative salivary gland protein C002, the silencing of which has potential for aphid management, and the putative Bacillus thuringiensis Cry toxin receptors, aminopeptidase N and alkaline phosphatase.

Conclusions/Significance: This study provides the most comprehensive data set to date for soybean aphid gene expression: This work also illustrates the utility of short-read transcriptome sequencing and the Multiple-k/Multiple-C method followed by BLAST annotation for rapid identification of target genes for organisms for which reference genome sequences are not available, and extends the utility to include the transcriptomes of endosymbionts.
essential amino acids that are not synthesized by the host aphid [17]. The development of genomic resources for other aphid species has facilitated a more complete understanding of the interaction between *Buchnera* and the host aphid [18,19]. In addition, aphids harbor secondary or facultative endosymbionts such as *Hamiltonella*, *Rickettsia*, *Arsenophonus*, *Regiella*, *Serratia* and *Wölbachia*. These symbionts function in aphid defense against pathogens and parasitoid wasps, and may be involved in resistance to host plant defense resulting in formation of aphid “biotypes” [20,21,22]. Secondary endosymbionts may be lost, or gained via both vertical and horizontal transmission [23].

Given the economic importance of the soybean aphid, genomic sequence resources for this agricultural pest are essential for (i) increased understanding of the biology and physiology of this species, (ii) identification of potential targets in the gut for novel aphidical technologies (as the gut is readily accessible to ingested control agents, it provides a primary focus for novel pest control strategies), and (iii) monitoring of *A. glycines* biotypes in North America. Silencing of *C002* [24,25], and the potential use of Bt-derived toxins against aphids [26] are of particular interest. We employed next-generation sequencing technology (Illumina Genome Analyzer II) to increase the molecular resources available for the soybean aphid. In addition to demonstrating the use of the *Multiple-k/Multiple-C* method for *de novo* assembly of short read sequences following by BLAST annotation of contigs, we addressed (i) analysis of species distributions of top hits, (ii) gene ontology analysis and comparison of whole aphid (WA) and gut transcriptomes, (iii) comparison of the soybean aphid transcriptome with pea aphid gene sets, (iv) comparison of relative depth of sequence coverage to relative transcript abundance for genes with high or low transcript abundance, (v) analysis of the *Buchnera* transcriptome, (vi) identification of *Wölbachia* and *Arsenophonus* as potential secondary endosymbionts of the soybean aphid, (vii) alignment of full length sequences from RNA-seq data. Our dataset has more than doubled the number of unique sequences reported for the soybean aphid [27], and provides valuable datasets for further analyses of the soybean aphid gut and endosymbiont transcriptomes.

**Results and Discussion**

**De novo assembly of Illumina short read sequences**

Analysis of RNA-seq short read sequences presents a challenge for organisms for which genomic sequence data are not available. For *de novo* assembly, the Velvet program was used to generate contiguous sequences (contigs) [28]. In order to acquire maximum information from the RNA-seq data, we used the *Multiple-k* (hash length *k*-mer) method [29] combined with the multiple C (coverage cutoff) to generate multiple sets of contigs. The contig sets were depleted using the CD-HIT program [30] to reduce redundancy, and the resulting contigs for each sample (WA or gut) were combined. The two sets of pooled samples were again depleted with CD-HIT, and the numbers of contigs in each set (gut and WA) reduced to about 16% of the original number of contigs.

The final number of contigs for the soybean aphid gut transcriptome was 141,532 (*≥ 100 nt*; Table 1) with the longest contig being 11,376 nt in length, and the average length being 424 nt. Twenty-five % (35,000) of the contigs were equal or greater than 500 nt in length. The final number of contigs for the whole soybean aphid (WA) transcriptome was 253,603 with an average contig length of 312 nt. Around 15.5% (39,600) of the contigs were equal to or longer than 500 nt, with the longest being 6,350 nt. These final contig sets covered about 80% of the reads from the gut sample and 64% of the reads from the WA sequences. The contig set for the gut transcriptome with the highest N50 was created by using *k* = 31 and *C* = 6. BLASTx analysis of this set of contigs resulted in identification of 3,931 non-redundant top hits. In comparison, by combining multiple contig sets, 10,640 non-redundant hits were identified (Table 1). Thus, the use of multiple contig sets with varying parameters, allowed for identification of 65% more soybean aphid transcripts than use of the single “optimal” set. Two sets of contigs (soybean aphid gut, whole aphid) have been deposited to AphidBase [http://www. aphidbase.org/aphidbase/].

**BLAST annotation of soybean aphid contigs**

The final contig sets for the gut and WA transcriptomes were annotated with BLASTx against the NCBI nr database. Contigs without hits from BLASTx analysis were then annotated with BLASTn for detection of additional gene sequences (Table 1). The majority of the contigs (90.7% for the gut, and 90.8% for the WA) had hits with either BLASTX or BLASTn. Of these, hits were identified for 70.8% of the gut and 73.2% of the WA contigs by BLASTX. Analysis of contigs without BLASTX hits showed that 19.8% of the gut and 17.6% of the WA contigs hit nucleotide sequences on analysis with BLASTn (Table 1). The majority of the contigs that did not align with either protein or nucleotide sequences on BLAST analysis were short contigs: 75% of the contigs had no hits were less than 200 nt in length.

After removing redundant hits, we identified 10,640 and 14,861 non-redundant proteins from the gut and WA transcripts, respectively. Among the non-redundant hits, 9,244 (56.9%) were identified from both the gut and WA transcriptomes, while 1,396 (8.6%) were unique to the gut transcriptome, and 5,617 (34.6%) were unique to the WA transcriptome (Table 1). In total 16,257 unique protein hits were identified by BLASTX. Notably, as a result of both the sequencing and assembly methods employed, the number of non-redundant genes identified using the short read transcriptome sequencing approach was more than double the number reported using Roche-454 and Illumina GA II 31 bp – paired end reads [27].

**Table 1. Summary of BLAST analysis and annotation of soybean aphid sequences.**

<table>
<thead>
<tr>
<th>Contig sets</th>
<th>Gut</th>
<th>Whole Aphid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total contigs (<em>≥ 100 nt</em>)</td>
<td>141,532</td>
<td>253,603</td>
</tr>
<tr>
<td>Total BLASTx hits</td>
<td>100,230 (70.8%)*</td>
<td>185,650 (73.2%)</td>
</tr>
<tr>
<td>Total BLASTn hits</td>
<td>28,071 (19.8%)</td>
<td>44,788 (17.6%)</td>
</tr>
<tr>
<td>No hits</td>
<td>13,231 (9.4%)</td>
<td>23,185 (9.1%)</td>
</tr>
<tr>
<td>Non-redundant top hits (BLASTx)</td>
<td>10,640</td>
<td>14,861</td>
</tr>
<tr>
<td>Non-redundant BLASTX top hits to each set</td>
<td>1,396</td>
<td>5,617</td>
</tr>
<tr>
<td>Non-redundant top hits (BLASTn)</td>
<td>6,396</td>
<td>9,862</td>
</tr>
<tr>
<td>Non-redundant EC (Enzyme code)</td>
<td>527</td>
<td>641</td>
</tr>
<tr>
<td>No-redundant EC numbers for each set</td>
<td>68</td>
<td>182</td>
</tr>
<tr>
<td>% of contigs with at least 1 GO term</td>
<td>18.23</td>
<td>21.04</td>
</tr>
<tr>
<td>Inter Pro (Protein signatures)</td>
<td>1,775</td>
<td>1,478</td>
</tr>
<tr>
<td>Non-redundant Inter Pro to each set</td>
<td>808</td>
<td>440</td>
</tr>
</tbody>
</table>

*% of total number of contigs* doi:10.1371/journal.pone.0045161.t001
Examination of the species distributions of the non-redundant top hits from both BLASTx and BLASTn showed that 83.0/91.1% (BLASTx/BLASTn) of the hits from the gut transcriptome and 75.7/91.4% of the hits from the WA transcriptome aligned to genes of the pea aphid and other aphid species (Table 2). A total of 4.1/1.8% of the WA top hits were genes of the endosymbiotic bacterium Buchnera.

Comparison of soybean aphid and pea aphid genes

To conduct a functional analysis of the soybean aphid genes, we tested various databases for gene annotation, including the NCBI database, FlyBase (FlyBase http://flybase.org/) [31] and Swiss-Prot. Mapping of the soybean aphid transcriptome contigs against the protein sequences in the Swiss-Prot protein database by BLAST2GO resulted in identification of the most GO terms. Overall, only 18.2% of the gut contigs and 21.0% of the WA contigs were assigned at least one GO term (Table 1). Analysis of GO distributions showed similar GO distribution patterns between the gut and WA sequences (Figure 1). GO-enzyme code mapping assigned 709 non-redundant EC codes. Of those, 68 (9.6%) of the enzymes were unique to the gut and 182 (25.7%), were only identified in the WA samples (Table 1).

BLAST analysis of the soybean aphid transcriptome resulted in identification of more than 16,000 potential transcripts from the soybean aphid, which included transcripts from both the aphid and associated endosymbionts. Although some 35,000 genes are predicted from the pea aphid genome [6], it is unknown how many of the predicted genes are transcribed. Identification of sequences in the soybean aphid transcriptome homologous to predicted pea aphid genes supports transcription of these hypothetical genes. The pea aphid genome is remarkable in having a high level of gene duplication and expansion of some gene families. Such gene duplication and gene expansion events could impact the quality of de novo transcript assembly. The impact of this on the transcript assembly reported herein will become apparent once the soybean aphid genome sequence is available.

We used pea aphid genes as reference genes to search the soybean aphid transcriptome for genes homologous to those predicted or identified from the pea aphid genome. Seventeen groups of annotated pea aphid genes were selected for analysis (Table 3), which have a total of 1,430 genes with assigned IDs. Examination of the genes revealed that 1,145 (80.1%) of the 1,430 pea aphid genes had putative homologs in the soybean aphid sequences. Genes functioning in amino acid transport and sugar

<table>
<thead>
<tr>
<th>Species</th>
<th>Gut No. of non-redundant top hits</th>
<th>Gut %</th>
<th>Whole aphid No. of non-redundant top hits</th>
<th>Whole aphid %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemiptera</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acyrthosiphon pisum</td>
<td>8,730</td>
<td>82.05</td>
<td>11,141</td>
<td>74.97</td>
</tr>
<tr>
<td>Other aphid species</td>
<td>104</td>
<td>0.98</td>
<td>110</td>
<td>0.74</td>
</tr>
<tr>
<td>Coleoptera</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tribolium castaneum</td>
<td>120</td>
<td>1.13</td>
<td>223</td>
<td>1.50</td>
</tr>
<tr>
<td>Diptera</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mosquitos</td>
<td>131</td>
<td>1.23</td>
<td>233</td>
<td>1.57</td>
</tr>
<tr>
<td>Drosophila spp.</td>
<td>151</td>
<td>1.42</td>
<td>262</td>
<td>1.76</td>
</tr>
<tr>
<td>Anoplura</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pediculus humanus</td>
<td>95</td>
<td>0.89</td>
<td>163</td>
<td>1.10</td>
</tr>
<tr>
<td>Hymenoptera</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harpegnathos saltator</td>
<td>56</td>
<td>0.53</td>
<td>85</td>
<td>0.57</td>
</tr>
<tr>
<td>Camponotus floridanus</td>
<td>49</td>
<td>0.46</td>
<td>84</td>
<td>0.57</td>
</tr>
<tr>
<td>Nasania vitripennis</td>
<td>64</td>
<td>0.60</td>
<td>84</td>
<td>0.57</td>
</tr>
<tr>
<td>Solenopsis invicta</td>
<td>41</td>
<td>0.39</td>
<td>67</td>
<td>0.45</td>
</tr>
<tr>
<td>Acromyrmex echinatior</td>
<td>31</td>
<td>0.29</td>
<td>50</td>
<td>0.34</td>
</tr>
<tr>
<td>Apis mellifera</td>
<td>43</td>
<td>0.40</td>
<td>64</td>
<td>0.43</td>
</tr>
<tr>
<td>Bombus impatiens</td>
<td>30</td>
<td>0.28</td>
<td>49</td>
<td>0.33</td>
</tr>
<tr>
<td>Bombus terrestris</td>
<td>35</td>
<td>0.33</td>
<td>36</td>
<td>0.24</td>
</tr>
<tr>
<td>Lepidoptera</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danaus plexippus</td>
<td>34</td>
<td>0.32</td>
<td>65</td>
<td>0.44</td>
</tr>
<tr>
<td>Endosymbiotic bacteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buchnera aphidicola</td>
<td>20</td>
<td>0.19</td>
<td>602</td>
<td>4.05</td>
</tr>
<tr>
<td>Wolbachia</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0.02</td>
</tr>
<tr>
<td>Others</td>
<td>918</td>
<td>8.63</td>
<td>1,530</td>
<td>10.30</td>
</tr>
<tr>
<td>Total non-redundant hits</td>
<td>106,40</td>
<td>100</td>
<td>14,861</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2. Species distribution of non-redundant top BLASTx hits for soybean aphid transcripts.
transport had the highest sequence identity between the two aphid species, with 95.7% of the amino acid transporter genes and 94.7% of the sugar transporter genes identified in the soybean aphid transcriptomes. In contrast, only 52.7% of the cathepsin genes (an important protease superfamily) of the pea aphid were identified in the soybean aphid transcriptomes. This result may indicate either that the putative cathepsin genes are not all expressed, possibly because of the high level of gene duplication in aphids and loss of function in some cases, or may reflect the tight regulation of expression of tissue specific cathepsin genes [32].

Specific analysis to identify transcripts of digestive enzymes in the soybean aphid gut transcriptome resulted in identification of transcripts for alpha-amylase (8 BLASTx hits), aminopeptidase (17), carboxypeptidase-like (13), cysteine protease (2), and oligopeptidase (1); Transcripts potentially involved in detoxification included those for cytochrome P450-like (22 BLASTx hits),
Examination of relative transcript abundance

RNA-seq can be used for measuring relative transcript levels [33]. Expression levels are determined by comparing the relative depth of sequence coverage to assembled contigs, followed by qRT-PCR to confirm the relative abundance of selected transcripts. Because no genomic and only limited gene sequence information is available for the soybean aphid, it was not appropriate to determine the relative gene expression level by RPKM value (i.e. reads per kilobase of exon model per million mapped reads). To assess the relative abundance of transcripts in the gut and WA samples, we mapped the 75 nt Illumina reads to the assembled contigs from the gut and WA using the MAQ program. The 10 contigs from the gut and WA samples with the highest depth of reads (and implied highest transcript abundance) are listed in Table 4. There is no overlap between the 10 most abundant transcripts from the soybean aphid WA and gut transcriptomes (Table 4). The RNA-Seq - predicted most abundant transcripts in the gut were for genes involved in amino acid and sugar metabolism. Of the five most highly expressed transcripts from the gut, three encode membrane alanyl aminopeptidase N (APN). This result is consistent with examination of APN expression in the pea aphid gut, which showed that APN is the most abundant protein comprising an estimated 16% of the total gut protein [34]. In that study, only one APN protein was isolated, while our gut transcriptome analysis showed that at least three APN-like genes were highly transcribed.

The depth of reads per putative gene for the 10 mostly highly expressed genes in the gut sample varied 4.5 fold (14,523 to 63,316 reads assembled). In contrast, the numbers of reads per gene for the most highly expressed transcripts in the WA sample, varied only 1.4 fold (4,316 to 6,741). Considering that the whole aphid RNA samples included all tissues and aphids in different developmental stages, it is not surprising to see reduced depth of coverage compared to the tissue specific transcriptome.

To confirm that the number of short reads assembled for a particular cDNA [mRNA-Seq] provided an indication of relative transcript abundance, we conducted qRT-PCR on total aphid RNA for four genes with high or low transcript abundance: two aminopeptidases, which were among the most abundant transcripts in the gut transcriptome, and two randomly selected genes of unknown function, with low transcript abundance (Table 5).

<table>
<thead>
<tr>
<th>Gene groups</th>
<th>No. genes in Pea aphid</th>
<th>Gut (% of PA)</th>
<th>Whole aphid (% of PA)</th>
<th>Total (% of PA)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid biosynthesis</td>
<td>82</td>
<td>69 (84.15)</td>
<td>70 (85.37)</td>
<td>71 (85.57)</td>
<td>[18]</td>
</tr>
<tr>
<td>Amino acid degradation</td>
<td>119</td>
<td>99 (83.19)</td>
<td>99 (83.19)</td>
<td>101 (84.87)</td>
<td>[18]</td>
</tr>
<tr>
<td>Amino acid transporters</td>
<td>47</td>
<td>35 (74.47)</td>
<td>44 (93.62)</td>
<td>45 (95.74)</td>
<td>[18]</td>
</tr>
<tr>
<td>Cathepsins</td>
<td>74</td>
<td>35 (47.3)</td>
<td>39 (52.7)</td>
<td>39 (52.7)</td>
<td>NCBI database</td>
</tr>
<tr>
<td>Chitinase-like proteins</td>
<td>9</td>
<td>7 (77.78)</td>
<td>8 (88.89)</td>
<td>8 (88.89)</td>
<td>[61]</td>
</tr>
<tr>
<td>Chromatin remodeling proteins</td>
<td>145</td>
<td>71 (48.97)</td>
<td>94 (64.83)</td>
<td>94 (64.83)</td>
<td>[62]</td>
</tr>
<tr>
<td>Clock genes</td>
<td>14</td>
<td>11 (78.57)</td>
<td>11 (78.57)</td>
<td>12 (85.71)</td>
<td>[63]</td>
</tr>
<tr>
<td>Developmental</td>
<td>315</td>
<td>187 (59.37)</td>
<td>268 (85.08)</td>
<td>270 (85.71)</td>
<td>[64]</td>
</tr>
<tr>
<td>Homeobox</td>
<td>55</td>
<td>15 (27.27)</td>
<td>43 (78.18)</td>
<td>44 (80)</td>
<td>[64]</td>
</tr>
<tr>
<td>Immune and stress</td>
<td>98</td>
<td>54 (55.1)</td>
<td>68 (69.39)</td>
<td>68 (69.39)</td>
<td>[65]</td>
</tr>
<tr>
<td>Ion channels</td>
<td>85</td>
<td>30 (35.29)</td>
<td>66 (77.65)</td>
<td>66 (77.65)</td>
<td>[66]</td>
</tr>
<tr>
<td>Meiosis and cell cycle</td>
<td>80</td>
<td>43 (53.75)</td>
<td>64 (80)</td>
<td>65 (81.25)</td>
<td>[67]</td>
</tr>
<tr>
<td>Nuclear receptors</td>
<td>14</td>
<td>8 (57.14)</td>
<td>11 (78.57)</td>
<td>11 (78.57)</td>
<td>[68]</td>
</tr>
<tr>
<td>Purine metabolism and urea cycle</td>
<td>52</td>
<td>34 (65.38)</td>
<td>44 (84.62)</td>
<td>44 (84.62)</td>
<td>[19]</td>
</tr>
<tr>
<td>Sugar transporters</td>
<td>75</td>
<td>51 (68)</td>
<td>69 (92)</td>
<td>71 (94.67)</td>
<td>[69]</td>
</tr>
<tr>
<td>Transcytosis</td>
<td>146</td>
<td>94 (64.38)</td>
<td>115 (78.77)</td>
<td>117 (80.14)</td>
<td>[70]</td>
</tr>
<tr>
<td>Wing development</td>
<td>20</td>
<td>6 (30)</td>
<td>19 (95)</td>
<td>19 (95)</td>
<td>[71]</td>
</tr>
<tr>
<td>Total</td>
<td>1430</td>
<td>849 (59.37)</td>
<td>1132 (79.16)</td>
<td>1145 (80.07)</td>
<td>[75]</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0045161.t003

Table 3. Putative soybean aphid homologs to pea aphid gene sets.
when comparing treatments or tissues, correlates strongly with qRT-PCR results for a given gene ($r = 0.966$, $n = 714$ genes; Illumina RNA Analysis data sheet).

**Buchnera aphidicola transcriptome**

The genomes of symbiotic bacteria in the genus *Buchnera* are highly reduced. The *Buchnera* genome size is 14% that of the *E. coli* genome [35,36] and is predicted to encode only 583 genes (*Buchnera* sp. APS) [37], which is only 3-fold the core sequence of a minimal bacterial gene set [38]. Because of the importance of these endosymbionts to aphid survival, we also examined the transcript profiles of the soybean aphid endosymbionts.

One of the WA RNA samples underwent a single polyA RNA purification step, rather than the two recommended by the Illumina RNA sample preparation protocol. As a result of this change in the protocol, approximately 30% of the RNA reads generated lacked a 3' polyA tail.

A total of 1,068 contigs (0.72% of the WA contigs with BLASTx hits) had BLASTx hits to *Buchnera* sequences. An additional 1,058 contigs (1.78% of the contigs with BLASTn hits) has BLASTn hits to *Buchnera* sequences. Only 91 (20 from BLASTx and 71 from BLASTn) contigs from the gut transcriptome were derived from *Buchnera* sequences and most of these were molecular chaperone sequences (e.g. GroEL), or rRNA genes. Analysis of the BLAST and annotation data for the 1,068 contigs identified by BLASTx resulted in identification of 602 non-redundant hits out of the 1,068 top hits obtained from BLASTx (Table 6, Table S2). A total of 334 distinct protein types were found from the non-redundant hits.

**Table 4.** Ten most abundant transcripts in the gut and whole aphid (WA) transcriptomes based on depth of reads assembled into contigs.

<table>
<thead>
<tr>
<th>Gut</th>
<th>Putative Genes</th>
<th>Reads assembled</th>
<th>Top hit ID</th>
<th>Species</th>
<th>E-value</th>
<th>Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Membrane alanyl aminopeptidase N</td>
<td>65316</td>
<td>XP_001944286.1</td>
<td>A. pisum</td>
<td>7.00E-52</td>
<td>75</td>
</tr>
<tr>
<td>2</td>
<td>Sugar transporter 1</td>
<td>42057</td>
<td>ACTI0281.1</td>
<td>Sitobion avenae</td>
<td>2.00E-134</td>
<td>92</td>
</tr>
<tr>
<td>3</td>
<td>Membrane alanyl aminopeptidase N</td>
<td>31032</td>
<td>NP_001119606.1</td>
<td>A. pisum</td>
<td>1.00E-104</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>Membrane alanyl aminopeptidase N</td>
<td>29218</td>
<td>XP_001948350.1</td>
<td>A. pisum</td>
<td>1.00E-174</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>putative cathepsin B-5</td>
<td>22217</td>
<td>AAU84936.1</td>
<td>Toxoptera citricida</td>
<td>1.00E-87</td>
<td>86</td>
</tr>
<tr>
<td>6</td>
<td>Putative ADP/ATP translocase</td>
<td>21521</td>
<td>XP_001948359.1</td>
<td>A. pisum</td>
<td>2.00E-164</td>
<td>98</td>
</tr>
<tr>
<td>7</td>
<td>glutamine synthetase 2</td>
<td>20208</td>
<td>NP_001153848.1</td>
<td>A. pisum</td>
<td>3.00E-53</td>
<td>97</td>
</tr>
<tr>
<td>8</td>
<td>Alkaline phosphatase homologues</td>
<td>19840</td>
<td>XP_001943535.1</td>
<td>A. pisum</td>
<td>2.00E-54</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>Ac1147-like protein</td>
<td>16889</td>
<td>ABG74714.1</td>
<td>Diaphorina citri</td>
<td>4.00E-141</td>
<td>96</td>
</tr>
<tr>
<td>10</td>
<td>cathepsin B-16A</td>
<td>14523</td>
<td>NP_001119617.1</td>
<td>A. pisum</td>
<td>4.00E-122</td>
<td>86</td>
</tr>
</tbody>
</table>

**WA**

1 similar to cytochrome P450              | 6741             | XP_001944205.1   | A. pisum | 4.00E-162 | 92           |
2 similar to AGAP010734-PA, partial       | 6057             | XP_001949485.1   | A. pisum | 0          | 97           |
3 similar to Collagen                     | 5934             | XP_001944753.1   | A. pisum | 0          | 76           |
4 DnaJ-like protein                       | 5554             | NP_001119620.1   | A. pisum | 8.00E-148 | 95           |
5 similar to cement precursor protein 3B  | 5029             | XP_001945547.1   | A. pisum | 1.00E-113 | 67           |
6 ATP synthase-beta                       | 5028             | NP_001119645.2   | A. pisum | 4.00E-95  | 99           |
7 similar to paramyosin, long form         | 5023             | XP_001948420.1   | A. pisum | 0          | 95           |
8 similar to Mitochondrial phosphate carrier protein | 5009 | XP_001945337.1 | A. pisum | 2.00E-166 | 95           |
9 similar to tyrosine hydroxylase          | 4616             | XP_001944964.1   | A. pisum | 0          | 83           |
10 similar to proteophosphoglycan ppg1    | 4316             | XP_001948991.1   | A. pisum | 4.00E-115 | 74           |

**Table 5.** qRT-PCR analysis of relative transcript abundance compared to mRNA-Seq data*.

<table>
<thead>
<tr>
<th>Transcripts</th>
<th>Relative abundance (qRT-PCR)</th>
<th>Reads assembled</th>
<th>Length of contigs</th>
<th>Average sequence coverage (depth*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT1</td>
<td>1.00</td>
<td>2,775</td>
<td>934</td>
<td>222.80</td>
</tr>
<tr>
<td>FT2</td>
<td>8.04</td>
<td>1,725</td>
<td>520</td>
<td>248.79</td>
</tr>
<tr>
<td>APN3</td>
<td>25.46</td>
<td>31,032</td>
<td>2,725</td>
<td>854.09</td>
</tr>
<tr>
<td>APN4</td>
<td>24.84</td>
<td>29,218</td>
<td>1,471</td>
<td>1,489.70</td>
</tr>
</tbody>
</table>

*Selected cDNAs with abundant (APN3, APN4) and few (FT1, FT2) transcripts in the soybean aphid gut transcriptome were amplified by qRT-PCR with GAPDH as internal control. The relative abundance of each transcript as determined by qRT-PCR is shown alongside the numbers of short read sequences assembled for each cDNA.

*depth (coverage per nt) = number of reads x 75 nt/contig length.

---

Soybean Aphid Transcriptomes
hits (Table S3), indicating that transcripts for more than half of the *Buchnera* genes were present in the WA aphid transcriptome. Among the non-redundant top BLASTx hits, 41.2% showed homology to sequences of *Buchnera* associated with the spring grain aphid (also known as the greenbug), *Schizaphis graminum* (*Buchnera aphidicola* str. Sg), 22.2% to sequences of *Buchnera* associated with the pea aphid (*Buchnera aphidicola* str. A, str. Tuc7, str. LSR1 APS, JF96, and JF99), and the rest to *Buchnera* sequences from other aphid species. This result indicates that the *Buchnera* strain in the soybean aphid has diverged and is more closely related to that in the spring grain aphid, consistent with the phylogenetic relatedness of the host species: *Aphis glycines* and *Schizaphis graminum* belong to the tribe Aphidini while the pea aphid belongs to the tribe Macrosiphini.

Gene annotation revealed that 43.4% of the *Buchnera* genes identified contain motifs that function in metabolic processes and 36% have a role in cellular processes (Figure 2). In molecular functions, 42% have catalytic activity and 42% are predicted to function in binding (Figure 2). The most highly expressed bacterial genes are the essential genes encoding ribosomal, cell division and chaperone/protease proteins [39], many of which were identified in the soybean aphid *Buchnera* transcripts. For instance, we identified the transcripts of 27 50S ribosomal protein L and 20 30S ribosomal protein S (Table S2), which were 69.2% of the annotated 50S ribosomal protein L and 74.1% of annotated 30S ribosomal protein S from the *Buchnera* associated with the pea aphid (str. A and APS). We also identified eight transcripts related to cell division functions (MtmG, DaeL, FtsA, H, J, W and Z) and chaperone/heat shock proteins (e.g., dnaJ, dnak, groE, groES, HtpX, hspG, hscA, hslU).

**Wolbachia** is a potential secondary endosymbiont in the soybean aphid

In addition to the primary endosymbiont *Buchnera*, aphids often harbor facultative or secondary endosymbionts in their hemolymph, bacteriocytes and/or reproductive tissues [40]. Several different secondary symbionts have been identified in aphids [41, 42], with the most common species being *Serratia symbiotica*, *Hamiltonellia defense*, and *Regiella insecticola* [43]. A recent study on the symbiotic bacteria of soybean aphids isolated from Illinois, USA, failed to find the secondary endosymbionts that are commonly found in aphids: PCR evidence was presented for the presence of *Arsenophonus*, a symbiont of whiteflies (Hemiptera: Aleyrodidae) [20]. Transcript sequence for soybean aphids isolated from Ohio, USA provided evidence for the presence of *H. defense*, which is closely related to *Arsenophonus* [27]. In searching for the secondary symbionts of soybean aphids isolated in Iowa, no significant hits were obtained by BLASTx or BLASTn to *Serratia*, *Hamiltonellia*, or *Regiella*. However, contigs of *Arsenophonus* 16S RNA were identified. PCR detection by using secondary symbiont universal 16–23S primers [44] confirmed the presence of *Arsenophonus* in our soybean aphid colony (data not shown). In addition, we identified two contigs with BLASTx and 63 contigs with BLASTn, ranging from 100–771 nt in size, with similarity to *Wolbachia* sequences. *Wolbachia* is an obligatory intracellular 2- proteobacterium detected in parasitic nematodes (filarial worms), mites and many insects including aphids [45]. *Wolbachia* sequences have been detected in multiple aphid species including *Teospera citricida*, *Aphis craccivora*, *Cinara cedri*, *Sitobion miscanthi* and *Sitobion miscanthi* [21, 22, 41, 46, 47]. Table 7 lists the 15 contigs with the highest similarity to *Wolbachia* sequences. The corresponding contig sequences (WS1–WS15) are listed in Sequence data S1, S1, WS1 and WS2 identified by BLASTx have homology to WwAna1270 and Scaffold protein (NifU) of *Wolbachia*, respectively. Most of the contigs identified by BLASTn are similar to either 16S or 23S ribosomal RNA with high levels of similarity (92–100%). In total, 1,070 nt of the 16S RNA (71% of the 1,505 nt 16S rRNA of the *Wolbachia* wRi strain) and 1,686 nt of the 23S rRNA (76% of the 2,746 nt 23S of the *Wolbachia* wRi strain) were assembled into the contigs. Interestingly and consistent with previous reports [21], the 16S rRNA-like sequences of the soybean aphid-derived contigs appear to be quite diverse: The top hits of the 16S rRNA contigs were from various *Wolbachia* strains, including strains detected in filarial nematodes (*Brugia sp.* and *Dirofilaria immitis*), a mite (*Bryobia*), the Asian citrus psyllid, *Diaphorina citri* and the aphid *Cinara cedri*. 16S rDNA is commonly used for identification and classification of *Wolbachia* strains [41, 47]. The diversity of the *Wolbachia* 16S rRNA in the soybean aphid transcriptome may reflect co-infection of the soybean aphid with multiple *Wolbachia* strains, as observed in *Drosophila* [48] and the wheat aphid, *Sitobion miscanthi* [47].

In contrast to the diversity of 16S rRNA sequences, the top hit for the *Wolbachia* 23S rRNA was from *Wolbachia* sp. wRi, an endosymbiont of *Drosophila simulans*. The second hit of 23S rRNA was from strain W16m isolated from *D. melanogaster*. The sequences of the soybean aphid *Wolbachia* 23S rRNA contigs and the 23S rRNA of W16m differed only slightly from those of strain wRi, indicating that the strain of *Wolbachia* in soybean aphid may belong to *Wolbachia* group A [47]. To verify the presence of *Wolbachia* in the Iowa isolate of the soybean aphid, primers were designed based on the contig sequences to amplify 23S rDNA (Table S4). A single DNA band of the expected size (2,102 bp) was observed (Figure 3). The PCR fragment was isolated from the gel and sequenced. The sequences (two non-overlapping sequences of 915 and 1,093 bp) were subjected to BLASTx analysis with the NCBI nr database. The top five hits were all *Wolbachia* 23S rDNA sequences with the top hit being to the wRi strain, with 96% and 97% identity to the 915 and 1,093 bp fragments respectively (Sequence data S2). We also designed primers to amplify *Wolbachia* 16S rDNA, *Wsp* (two different reverse primers; Table S4) and 16S rDNA genes. Similar to previous efforts to amplify *Wolbachia* sequences from aphids [21], no product was generated by PCR using *Wsp* and 16S rDNA primers. Primers that were designed for amplification of 16S rDNA based on the contigs that hit the 16S rDNA of *Wolbachia*, resulted in amplification of *Buchnera* 16S rDNA.

It is important to note that there is a precedent for lateral transfer of *Wolbachia* sequences into host genomes, with *Wolbachia* genome fragments encoding multiple genes present in a host beetle [49], transfer of genome segments into the nematode *Onchocerca* [50, 51], transfers into the genomes of four insect and four

---

**Table 6. Summary of annotation of *Buchnera* sequences from whole soybean aphid transcriptome.**

| Number of non-redundant *Buchnera* hits with BLASTx | 602 |
| Buchnera of pea aphid, *Acyrthosiphon pismum* | 134 (22.2%) |
| Buchnera of *Schizaphis graminum* | 248 (41.2%) |
| Buchnera of *Acyrthosiphon kondoi* | 107 (17.8%) |
| Buchnera of other aphids | 113 (18.8%) |
| Distinct protein hits | 329 |
| Number of contigs that hit *Buchnera* with BLASTn | 1,058 |
| Distinct Inter Pro (Protein signatures) | 39 |
| EC (enzymes) | 121 |
| GO-terms | 35 |

doi:10.1371/journal.pone.0045161.t006
nematode species, including one case of transfer of almost the entire \textit{Wolbachia} genome [52], and transfer of \textit{Wolbachia} genes into the genome of the tse-tse fly [53]. Hence, confirmation of the presence of \textit{Wolbachia} in the soybean aphid and in other aphid species by using techniques other than transcript and PCR-based methods is required.

Based on the secondary endosymbionts described for soybean aphids isolated from Illinois (\textit{Arsenophonus}) [20], Ohio (\textit{H. defensa}) [27] and Iowa (\textit{Arsenophonus, Wolbachia}), the secondary endosymbionts of the soybean aphid vary with geographical location.

**Full-length soybean aphid gene sequences**

To investigate the feasibility of using RNA-seq for discovery of full-length genes, we looked for the transcript sequences for homologs of three types of genes that are relevant to potential novel soybean aphid management strategies: \textit{C002}, a salivary gland (SG) gene which is essential for aphid feeding on the host plant [24], and two proteins that are putative secondary receptors for \textit{Bacillus thuringiensis} Cry toxins: membrane alanyl aminopeptidase N (APN) and alkaline phosphatase (ALP)[54]. apn transcripts

---

**Figure 2. Distribution of \textit{Buchnera} sequences by gene ontology.** (GO: level 2; filtered by sequence number cutoff = 5) for biological processes, cellular components, and molecular functions.
doi:10.1371/journal.pone.0045161.g002
ALP of the pea aphid were between 524–1039 aa and 513–565 aa, annotated from the pea aphid genome. The sizes of the APN and including isoforms and transcript variants, were predicted and for the pea aphid [25].

More than 10 APN- and six full-length ALP-like genes, were abundant in the soybean aphid transcriptome, while the putative C002 and alp transcripts were moderately expressed. C002 is a 219 amino acid (aa) peptide, which was originally discovered from the pea aphid SG EST library. C002 was primarily expressed in the SG of the pea aphid, but transcripts of C002 were also detected in the gut at a level of 1% that in the SG [24]. This protein was predicted from the pea aphid genome as a hypothetical protein (XP_001948358.2, LOC100167863). By [24]. This protein was predicted from the pea aphid genome as a hypothetical protein (XP_001948358.2, LOC100167863). By

### Table 7. Wolbachia sequences identified in the soybean aphid transcriptome.

<table>
<thead>
<tr>
<th>Contigs</th>
<th>Seq_length</th>
<th>Seq_description</th>
<th>Hit_ACC</th>
<th>E-value</th>
<th>Alig_length</th>
<th>Positive Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS1</td>
<td>771</td>
<td>protein WwAnd1270, Wolbachia endosymbiont of Drosophila melanogaster</td>
<td>ZP_00373202</td>
<td>5.00E-17</td>
<td>42 (AA)</td>
<td>39 (AA) 93</td>
</tr>
<tr>
<td>WS2</td>
<td>154</td>
<td>Scaffold protein, NlID domain protein, Wolbachia sp. wH</td>
<td>ZP_00373458</td>
<td>3.00E-25</td>
<td>51 (AA)</td>
<td>49 (AA) 96</td>
</tr>
<tr>
<td>WS3</td>
<td>225</td>
<td>strain trs of Brugia complete genome</td>
<td>AE017321</td>
<td>1.31E-86</td>
<td>226</td>
<td>209 92</td>
</tr>
<tr>
<td>WS4</td>
<td>129</td>
<td>Dirofilaria immitis strain dax 16s rRNA partial seq</td>
<td>AF487892</td>
<td>1.12E-58</td>
<td>129</td>
<td>129 100</td>
</tr>
<tr>
<td>WS5</td>
<td>178</td>
<td>Wolbachia endosymbiont of Diaphorina citri isolate sz 16s rRNA partial seq</td>
<td>GU565892</td>
<td>2.18E-82</td>
<td>178</td>
<td>176 98</td>
</tr>
<tr>
<td>WS6</td>
<td>343</td>
<td>Culex quinquefasciatus pel strain Wpip complete genome</td>
<td>AM999887</td>
<td>1.42E-164</td>
<td>343</td>
<td>336 97</td>
</tr>
<tr>
<td>WS7</td>
<td>270</td>
<td>Cinara cedri 16s rRNA partial seq</td>
<td>AE62043</td>
<td>3.33E-133</td>
<td>270</td>
<td>269 99</td>
</tr>
<tr>
<td>WS8</td>
<td>193</td>
<td>Bryobia v vidu-2008 strain ita11 16s rRNA partial seq</td>
<td>EU499316</td>
<td>5.31E-78</td>
<td>193</td>
<td>182 94</td>
</tr>
<tr>
<td>WS9</td>
<td>139</td>
<td>Wolbachia (from New Caledonia) 23s rDNA</td>
<td>X65683</td>
<td>3.54E-59</td>
<td>138</td>
<td>135 97</td>
</tr>
<tr>
<td>WS10</td>
<td>152</td>
<td>Pentatrichinidus leporinus partial 16s rRNA gene</td>
<td>FN428797</td>
<td>1.93E-69</td>
<td>152</td>
<td>151 99</td>
</tr>
<tr>
<td>WS11</td>
<td>297</td>
<td>Wolbachia sp. wH, 235 rRNA</td>
<td>CP001391</td>
<td>2.33E-156</td>
<td>297</td>
<td>295 99</td>
</tr>
<tr>
<td>WS12</td>
<td>271</td>
<td>Wolbachia sp. wH, 235 rRNA</td>
<td>CP001391</td>
<td>4.00E-99</td>
<td>224</td>
<td>216 96</td>
</tr>
<tr>
<td>WS13</td>
<td>673</td>
<td>Wolbachia sp. wH, 235 rRNA</td>
<td>CP001391</td>
<td>0.00E+00</td>
<td>673</td>
<td>656 97</td>
</tr>
<tr>
<td>WS14</td>
<td>771</td>
<td>Wolbachia sp. wH, 235 rRNA</td>
<td>CP001391</td>
<td>0.00E+00</td>
<td>771</td>
<td>748 97</td>
</tr>
<tr>
<td>WS15</td>
<td>107</td>
<td>Oncocerca Wolbachia seq fragment ow3</td>
<td>CU062463</td>
<td>7.56E-47</td>
<td>107</td>
<td>107 100</td>
</tr>
</tbody>
</table>

Sequences are provided in Supporting Information, Sequence Data S1. doi:10.1371/journal.pone.0045161.t007

were abundant in the soybean aphid transcriptome, while the putative C002 and alp transcripts were moderately expressed. C002 is a 219 amino acid (aa) peptide, which was originally discovered from the pea aphid SG EST library. C002 was primarily expressed in the SG of the pea aphid, but transcripts of C002 were also detected in the gut at a level of 1% that in the SG [24]. This protein was predicted from the pea aphid genome as a hypothetical protein (XP_001948358.2, LOC100167863). By conducting local BLAST analysis with the C002 sequence, we identified a full-length copy of the putative C002 homolog (see Figure S1; [GenBank: JN135246]) from a single contig assembled from the WA reads with about 43-fold coverage, and a partial C002 sequence was assembled from the gut Illumina reads, reflecting the lower expression of C002 in the gut. The putative soybean aphid C002 is 214 aa, 5 aa shorter than that of the pea aphid C002. Alignment of the soybean C002 homolog with the pea aphid C002 showed less than 50% sequence identity at the protein level (Figure S1b). C002 is secreted into the host plant and plays an important role in feeding, and hence may be involved in host plant selection [24]. The lower identity between the soybean aphid and pea aphid C002 may reflect the differences in the host plant preferences of the two species and selection for divergent protein sequences to deal with some aspect of survival on the host plant. Functional analysis is required to confirm that silencing of this gene in the soybean aphid has similar effects to those reported for the pea aphid [25]. More than 10 APN- and six full-length ALP-like genes, including isoforms and transcript variants, were predicted and annotated from the pea aphid genome. The sizes of the APN and ALP of the pea aphid were between 524–1039 aa and 513–565 aa, respectively. To identify APN-like and ALP-like genes from the soybean aphid, we analyzed BLAST data and identified >600 hits with contigs from the gut and 247 hits with contigs from WA against pea aphid APN genes. However, only 71 of the contigs were >1000 nt with the longest contig being >3200 nt. From these contigs, we found only two with the predicted full-length APN sequences. For ALP genes, 200 soybean aphid contigs were similar to the ALP genes of the pea aphid. The longest ALP contig was 1,830 nt, and two putative full-length ALP genes were identified. Notably, none of these predicted full-length genes were assembled by using the same k and C combinations. On further analysis of the contigs, one additional APN and one additional ALP full-length genes were identified by aligning the contigs and re-assembling the overlapping fragments. In addition, fragments of APN and ALP genes were also identified. To verify the presence of the full-length genes in the soybean aphid, RT-PCR was carried out to amplify the potential full-length APN transcripts (see Table S4 for primer sequences). cDNA was generated with polyT oligo, and primers specific to the four APN genes were used for PCR. cDNA of the four APNs of the correct sizes were successfully detected. Sequencing of the PCR-amplified APN cDNA showed that only 10 nucleotides (0.03%) differed from the APN4 sequences generated by the Illumina reads. The sequences for soybean aphid APN and ALP were submitted to GenBank [GenBank: ALP1 JN135238; ALP2 JN135239; ALP3 JN135240; ALP4 JN135241 (partial sequence); APN1 JN135242; APN2 JN135243 (partial sequence); APN3 JN135244; APN4 JN135245]. APN and ALP are important receptors for Cry toxins derived from the bacterium Bacillus thuringiensis (Bt) [54]. As Cry toxins are not particularly effective against aphids [26], we sought to address...
whether divergence of the putative receptor proteins could contribute to the low toxicity. Phylogenetic analysis of APN sequences between aphids (the soybean aphid and the pea aphid) and lepidopteran species [55] showed that aphid APNs are distinct from other classes of insect APN and form their own clade (Figure 4). The aphid ALPs were compared with those derived from mosquito, lepidopteran species, Drosophila and Tribolium castaneum. The ALPs of aphids divide into three groups (Fig 5). Divergence of the putative Bt receptor proteins in aphids may contribute to the relatively low toxicity of Bt-derived toxins against aphids [26].

Conclusions

In this study, we analyzed ~400,000 contigs generated by de novo assembly from RNA-seq reads of the soybean aphid gut and WA transcriptomes. The use of multiple sets of contigs with varying k and C parameters, and BLAST analysis significantly increased the number of transcripts identified, and the acquisition of full length gene sequences. This can be explained by the fact that contigs with fewer reads in the data set contain valuable transcript information that would otherwise be excluded when a higher coverage cutoff threshold is used.

Annotation of the contigs by BLAST allowed for identification of almost half of the pea aphid gene homologs from the soybean aphid transcriptome, and more than 50% of the Buchnera transcripts. This approach also allowed for identification of full-length aphid genes and the discovery of a potential new secondary endosymbiont, Wolbachia from the soybean aphid. Our results significantly increase the genomic resources available for the soybean aphid, and demonstrate use of the Multiple-k/Multiple-C methodology on a short read sequence data set for enhanced data mining. These results highlight the potential of RNA-seq for genomics and functional genomics studies on organisms for which genomic sequence data are not available, and extend the potential utility to endosymbiont transcriptomes. This work will provide the foundation for future analyses of soybean aphid biotype formation, the role of facultative endosymbions in aphid adaptation, and for development of novel technologies for soybean aphid management.

Materials and Methods

Insect rearing

A colony of soybean aphids, Aphis glycines Matsumura, was established from aphids collected in soybean fields in Iowa. The colony was maintained on soybean Glycine max (Variety 92M91, Pioneer Hi-Bred International, Inc, Johnston, IA) at 24±1°C with a 12 h light/12 h dark cycle and only produced viviparous parthenogenetic females.

RNA isolation and transcriptome sequencing

Three RNA samples were prepared, one from aphid guts, and two from whole aphids. For isolation of RNA from soybean aphid guts, the entire digestive tract was removed under a dissection microscope (Nikon SMZ 1500) from fourth and fifth instar nymphs, with approximately one-tenth of the sample derived from adults. Approximately 2,000 guts were pooled and stored in TRIzol reagent (Invitrogen). RNA was isolated and purified according to the TRIzol protocol. Total RNA was isolated from whole aphids (WA) (300 mg, all instars, winged and wingless nymphs, and adults).

Two steps of poly-A RNA purification were conducted for two samples (WA and gut) using oligo (dT) magnetic beads and further processed according to Illumina protocols. For the second WA sample, a single polyA purification step was carried out, resulting in increased representation of Buchnera sequences within the transcriptome. RNA integrity was confirmed using a 2100 Bioanalyzer (Agilent Technologies). The purified RNA was used to prepare samples for sequencing by using the Illumina truSeq RNA sample preparation kit. Sequencing on an Illumina GAIII sequencing platform (Illumina Corporation) at the Iowa State University DNA Facility resulted in approximately 8 million single-end reads for each lane, mostly 75 nt in length for each sample. In total, approximately 24 million reads were obtained. Adapter sequences and low quality sequences were removed prior to further analysis.

Bioinformatics

Aphid transcriptome sequences were mapped to the draft 207 genome (Acyr_1.0) of the pea aphid, A. pisum [http://www. aphidbase.com/aphidbase] [56] using the Eland (Illumina Inc.) and MAQ programs [http://maq.sourceforge.net/] with a maximum of 2 mismatches for Eland and 3 mismatches for MAQ. The Illumina reads were assembled using the Velvet assembler (1.0)[28], run on an Apple Mac Pro computer with 8-core Two 2.93GHz Quad-Core Intel Xeon/16GB RAM. Assembly was performed by using various combinations of k and C parameters and according to the program manual. Use of the multiple-k method significantly improves assembly efficiency [29]. By combining the multiple-k and multiple-C methods for assembly, followed by depleting redundant contigs, the numbers of assembled contigs was greatly increased. The selected contigs with a length cutoff of 100 nt were used for annotation by searching against the GenBank non-redundant database (including the A. pisum genome Acyr_2.0) using BLASTx algorithms (Number
of BLAST hits = 1 (return only top hit); minimum e-value = 1.0E-3, BLAST model: QBLAST-NCBI; HSP length cut off = 33; lower capacity filer = yes). Contigs without BLASTx hits were then annotated by using BLASTn algorithms using similar parameter settings to those used for BLASTx analyses. For optimal assignment of annotation quality and BLAST result analysis, only the top hits from BLAST were used for further data analyses. Gene Ontology (GO) annotation was conducted by using the Swiss-Prot database (http://www.uniprot.org/) and the protein signatures were annotated by using InterProScan [57]. All annotation programs were performed using the BLAST2GO platform [58]; http://www.blast2go.org/start_blast2go For annotation of combined contig sets, the contigs were purged for removal of redundant sequences using CD-HIT [30]; http://weizhong-lab.ucsd.edu/cd-hit/

The data sets are available at the NCBI Short Read Archive (SRA) with accession number: SRA038331.

Full length gene assembly and data analysis

For assembly of putative full length soybean aphid genes, contigs (> 300 nt) were aligned using BioEdit 7.0.9; http://www.mbio.ncsu.edu/BioEdit/bioedit.html The assembled cDNA fragments were translated and aligned to the genes of the pea aphid. The putative full length genes were then used for phylogenetic analysis. The multiple sequence alignments and phylogenetic trees (maximum-likelihood trees) were generated using MEGA 5.0 with a bootstrap value of 500 [59].

Assessment of relative transcript abundance

The depth of reads assembled into a contig was used to assign relative transcription levels within the transcriptome. Reads were mapped to the reference contigs using MAQ. The depth of mapping was recorded and the 50 contigs with the highest number of reads were analyzed.

qRT-PCR was used to validate the relative expression levels as determined by RNASeq, of APN3, APN4, and two contigs with

Figure 4. Phylogenetic relatedness of soybean aphid aminopeptidase N (APN) derived from the gut transcriptome with lepidopteran APN. The phylogenetic tree drawn to scale was generated by using the maximum-likelihood method using MEGA 5.0 with a bootstrap value of 500. Soybean aphid (SA), and pea aphid (PA) sequences are boxed. GenBank accession numbers: Bombyx mori: BmAPN1, AAC33301, BmAAPN2, BAAN32140, BmAAPN3, AAL83943, BmAAPN4, BAA33715; Epiphyas postvittana, EpaPN, AAFF99701; Helicoverpa armigera, HaAPN1, AAW72993, HaAPN2, AAN04900, HaAPN3, AAM44056, HaAPN4, AAK85339; Helicoverpa punctigera: HpaPN1, AAF37558, HpaPN2, AAF37560; Heliitis virescens: HvAPN1, AAFO8254, HvAPN2, AAK58066; Lymnaea dispar: LdAPN1, AAD31183, LdAPN2, AAD31184, LdAPN3, AAL26894; LdAPN4, AAL26895; Plutella xylostella: PxAPN1, AAB70755, PxAPN2, CA66467, PxAPN3, AAF01259, PxAPN4, CAA10950; Manduca sexta: MsAPN1, CAA61452, MsAPN2, CAA66466, MsAPN3, AAM13691, MsAPN4, AAM18718; Spodoptera exigua: SeAPN1, AAP44964, SeAPN2, AAP44965, SeAPN3, AAP44966, SeAPN4, AAP44967; Spodoptera litura: SlAPN, AAK69605; Trichoplusia ni, TnAPN1, TnAPN2, AAX39863, TnAPN2, AAX39864, TnAPN3, AAX39865, TnAPN4, AAX39866; Tribolium castaneum: TcAPN1, EEZ99298; TcAPN2, XP_001812439; TcAPN3, XP_00182497; TcAPN4, XP_00182498; TcAPN5, XP_00182499; the pea aphid, A. pisum: PAAPN1, NP_001119606, PAAPN2, XP_001946370, PAAPN3, XP_001946370, PAAPN4, XP_001946370, PAAPN5, XP_001946370; SAAPN1, JN135242, SAAPN2, JN135243, SAAPN3, JN135244, SAAPN4, JN135245.

doi:10.1371/journal.pone.0045161.g004
low transcript abundance (Few Transcripts, FT1, FT2). Total RNA from soybean aphid guts (0.5 mg) was isolated by using Trizol reagent (Invitrogen) according to the manufacturer’s directions. Precipitated RNA was resuspended in DEPC-treated, autoclaved water and stored at 280°C until further use. qRT-PCR was performed in two steps: In the first step, a 20 μl RT reaction was set up using 5 μg of soybean aphid gut total RNA, oligo dT12-18 primers and Superscript reverse transcriptase to synthesize the first strand cDNA according to the recommended protocol (Invitrogen). qRT-PCR primers for all four genes (apn3, apn4, FT1, and FT2: See Table S4 for primer sequences) were tested by PCR to confirm amplification of a single product of the correct size (200 bp). Twenty μl qRT-PCR reactions to amplify all four genes and GAPDH (internal control [60]) were set up in a 96-well plate using IQ Syber Green supermix (Bio-Rad). Two sets of negative controls, the no template control and the total RNA template (to control for contamination with genomic DNA) were set up for each primer pair. For amplification of sequences from all five genes, PCR reactions were performed using the following thermal cycle conditions: 95°C for 3 min, followed by 95°C for 15 s, 52°C for 30 s, and 72°C for 30 s for 40 cycles. PCR reactions were performed with two biological and three technical replicates, and analyzed on a Bio-Rad iCycler™ iQ Optical system using Software Version 3.0a. Values for relative transcript abundance for each of the four genes were calculated and normalized with reference to transcript abundance for the internal control. The relative expression levels of the four genes were compared by one-way ANOVA.

Confirmation of Wolbachia 23S rDNA sequence

Total DNA was extracted from 50 soybean aphids using DNaZol™ (Invitrogen) according to the manufacturer’s protocol, and dissolved in nuclease free water. The primers (Table S4) were designed based on the assembled contig from the soybean aphid transcriptome that had homology to the nearest 5′ and 3′ ends of the Wolbachia 23S rDNA. PCR was performed using Choice Taq™ DNA Polymerase and with 1 cycle of 94°C for 2 min, 35 cycles of 94°C for 30sec, 53 or 55°C (see Table S4) for 30sec, 72°C for 2 min and 1 cycle of 72°C for 5 min. The amplified PCR product (2,102 bp) was run on a 1% agarose gel. The PCR product was

Figure 5. Phylogenetic analysis of insect alkaline phosphatases (ALP). Phylogenetic tree drawn to scale for the soybean aphid (SA), pea aphid (PA) ALP, and mosquito (Aedes aegypti, Aa; Anopheles gambiae, Ag; Culex quinquefasciatus, Cq) lepidopteran (Bombyx mori, Bm; Heliothis virescens, Hv), Drosophila melanogaster (Dm) and Tribolium castaneum (Tc) ALPs. Soybean aphid (SA), and pea aphid (PA) sequences are boxed. GenBank accession numbers: AaALP1, XP_001663478, AaALP2, XP_001649092,AaALP3 XP_001668006, AaALP4, XP_001663538, AaALPXP_001663535; AgALP1, XP_313890, AgALP2, XP_001688180,AgALP3, XP_316433, AgALP4, XP_308522 AgALP5, XP_321411, AgALP6,AgALP7, XP_309345BmALP1, NP_001037536, BmALP2, NP_001036856, CqALP1, XP_001842934, CqALP2, XP_001842932; DmAALP1, NP_001034040, DmAALP12, NP_524601; HvALP1, ACP39712, HvALP2, ACP39713, HvALP3, ACP39714, HvALP4, ACP39715, HvALP5, ABR88230; TcALP1, XP_975050, TcALP2, XP_973094, TcALP3, EFA08950, TcALP4, EFA08951, TcALP5, EFA08952, TcALP6, XP_968925, TcALP7, EEF90048, TcALP8, EEE90049, TcALP10, EFA01926, TcALP11, XP_971418, TcALP12, XP_971358; TcALP13, XP_971482; The pea aphid, A. pisum: PAALP1, XP_001944129, PAALP2, XP_001943536, PAALP3, XP_001943259, PAALP4, XP_001943482, PAALP5, XP_001943555; SAALP1 JN135238; SAALP2, JN135239; SAALP3, JN135240.
doi:10.1371/journal.pone.0045161.g005
removed from the gel and purified using the Qiagen gel extraction kit (Qiagen). The purified PCR product was eluted in nuclease free water and submitted to the Iowa State University DNA Facility for sequencing using both forward and reverse primers.

**Supporting Information**

**Table S1** BLASTn hits for contigs identified by BLASTx to have similarity to barley sequences. The top two BLASTn hits are indicated. (XLSX)

**Table S2** Non-redundant hits of *Buchnera* genes by BLASTx. The list contains the contig ID, length, and BLAST hit descriptions for 602 sequences. (XLSX)

**Table S3** *Buchnera* proteins identified from the soybean aphid transcriptome. The list contains the contig ID, contig length, and 334 corresponding protein and genes names, and descriptions. (XLSX)

**Table S4** Primer sequences. Sequences are provided for primers used for PCR amplification of *apm-3, apm-4, FT1, FT2* and *Wolbachia* (*fts*, *wsp, 16S, 23S*) gene fragments. (PDF)

**Figure S1** Soybean aphid putative homolog of salivary protein C002. A. Sequence of the putative pea aphid C002 homolog from the soybean aphid; B. Clustal W alignment of the C002 amino acid sequences from the pea aphid and the soybean aphid. (PDF)

**Sequence data S1** Sequences from soybean aphid transcriptome contigs derived from the secondary endosymbiont *Wolbachia*. Sequences were derived from the whole aphid transcriptome (WA). Fifteen sequences are provided (WS1-WS15). (PDF)

**Sequence data S2** Additional evidence for the presence of *Wolbachia* in the soybean aphid. The *Wolbachia* 23S rDNA sequences derived from the soybean aphid were PCR-amplified and sequenced. The alignment of the soybean aphid (SA) PCR-amplified sequence with the sequence of *Wolbachia* sp. wRi (wRi) is provided. (PDF)

**Acknowledgments**

The authors thank Hui-Hsien Chou, Iowa State University for bioinformatics advice; John VanDyk, Iowa State University for IT support; Andy Michel, Ohio State University and Nick Miller, University of Nebraska, for helpful discussions; Adam Liu for writing scripts for data analysis; Amy Toth, Iowa State University for critical reading of the manuscript.

**Author Contributions**

Conceived and designed the experiments: SL NPC BB. Performed the experiments: SL NPC DV. Analyzed the data: SL NPC DV. Wrote the paper: SL NPC BB.

---

**References**


