Supplemental File 1. Additional Material and Methods

Phagmid cDNA library construction, EST sequencing and analysis

The total RNA isolated from the salivary glands was reverse transcribed using PowerScript reverse transcriptase and CDS III primer provided in the SMART cDNA library construction kit (Clontech). Second strand synthesis was performed using the PCR-based protocol using SMART III and CDS III primers provided in the kit following manufacturer's instructions. Double strand synthesis was followed by proteinase K digestion; thereafter the double strand cDNA was ligated into a Lambda TriplEx2 vector (Clontech). Finally the resulting ligation reaction was packaged using Gigapack Gold III (Stratagene) following the manufacturer's specifications.

EGassembler

The assembly pipeline consists of several steps: sequence cleaning (short and/or poor quality sequences are removed), repeat masking, vector masking, organelle masking and sequence assembly. Most of the options used in the pipeline were the default options. The *Drosophila* RepBase repeats library was used for repeat masking, while the NCBI's core vector library was used for vector masking. Sequences that were too short after masking vectors and organelles were removed. The 'overlap percent identity' cutoff for sequence assembly was 80%. The resulting assembly consisted of 834 contigs and 2189 singletons, (unassembled reads).

LC/MS-MS and MALDI-TOF/MS

75 µg of salivary gland homogenates electrophoresed on a SDS PAGE gel and visualized with the PlusOne Silver Staining Kit (GE Healthcare, UK). Gels were digitalized using a GS-800

calibrated densiotometer coupled with the Discovery Series™ QuantOne software (v 4.4; BioRad, Sweden). The gel lane was cut into 25 bands were transferred to 1.5 mL microfuge tubes with 10 µl water to prevent dehydration.

Sample preparation used the Ettan™ Digester (Amersham Biosciences, GE Healthcare, UK) following the manufacturer's instructions for MS compatible silver stained gels[1]. Proteins were reduced, acetylated and trypisn digested as outlined in Carolan et al. [2]. Samples were dried under vacuum in an Eppendorf Concentrator 5301 (Sigma-Aldrich, Germany) and pellets were resuspended in 12 µl 0.1% formic acid and subjected to LC MS/MS on a Finnigan LTQ mass spectrometer (Thermo Fisher Scientific, UK) connected to a Surveyor chromatography system incorporating an auto-sampler. All data were acquired with the mass spectrometer operating in automatic data-dependent switching mode. A zoom scan was performed on the five most intense ions to determine charge state prior to MS/MS analysis. Protein identification from the MS/MS data was performed using the TurboSEQUEST [3] algorithm in BioWorks v. 3.2 (Thermo Fisher Scientific, UK) to correlate the data against NCBI non-redundant database and the official consensus protein databank (ACYPIprotein, searched September 2010) originating from the *A. pisum* genome sequencing project (http://www.aphidbase.com/aphidbase/). The following search parameters were used: precursor-ion mass tolerance of 1.5 Da, fragment ion tolerance of 1.0 Da with methionine oxidation and cysteine carboxyamidomethylation specified as differential modifications and a maximum of 2 missed cleavage sites allowed. In order to account for the high false positive rate that accompanies single peptide matched proteins (see [4-6]) stringent post probabilistic filters and search cutoffs were applied to SEQUEST search and result processing.

Three filters were applied: XCorr vs. charge state (1, 2, and 3= 1.90, 3.00, and 3.50 respectively), >60% ion match and peptide probability ($p< 0.001$).

400 µg of salivary gland homogenate was diluted in rehydration solution (8 M urea,0.5% w/v CHAPS, 0.2% w/v DTT, and 0.2% w/v Pharmalyte, pH 3–10) added to a 24 cm non linear pH gradient IPG strip of pH 3-10 (GE, Healthcare, UK) and rehydrated overnight in an Immobiline™ DryStrip Reswelling Tray (GE Healthcare, UK). Strips were focused using an Ettan IPGphor II IEF unit (GE Healthcare, UK) following the method described by Rabiloud et al., [7]. IEF was performed using an IPGphor™ apparatus (GE, Healthcare, UK) for 75,000Vh at 20°C and maximum current setting of 50 A per strip. Once IEF was completed, strips were equilibrated by reduction and alkylation following the method of Gorg *et al*., [8]. Equilibrated strips were loaded onto 12% polyacrylamide gels and sealed with 1% agarose for second dimension separation using the Ettan DALT12 system (GE, Healthcare, UK) overnight at 1W/gel at 15°C. Gels were silver stained and spots were removed manually using a 1.5 mm spot picker and transferred to a 96 well plate. Spot plugs were destained and subjected to purification and in-gel trypsin digestion as previously described [1]. A 0.5 µl aliquot of each digest was applied directly to the MALDI target plate and subjected to MALDI-TOF/MS using an Applied Biosystems 4800 Proteomics Analyzer (Applied Biosystems, Foster City, CA, USA). All spectra were acquired with a solid-state laser (355 nm) at a laser repetition rate of 200 Hz. After measuring all samples in the MS mode, a maximum of 15 precursors per spot were selected for subsequent fragmentation by CID. The resulting spectra were processed and analyzed using the Global Protein Server software interface (v3.6; Applied Biosystems), using internal MASCOT (Matrix Sciences) software for matching MS and MS/MS data against databases of in silico digested proteins (same as those used in the LC/MS-MS searches). Search criteria included:

Maximum missed cleavages, 1; Three variable modifications, (NEM, PNGase F (conversion of Asn to Asp) and oxidation of methionine); Precursor tolerance, 30 ppm; MS/MS tolerance, 0.2 Da. Peptides were considered correct calls when the confidence interval was greater than 95%.

Mass spectral data obtained in batch mode were searched against the databases using a locally-running copy of the Mascot program (Matrix Science Ltd., version 2.1). Batch-acquired MS and MS/MS spectral data were submitted to a combined peptide mass fingerprint and MS/MS ion search through the Applied Biosystems GPS Explorer software interface (version 3.6) to Mascot against the same databases described for SEQUEST searches. Search criteria included: Maximum missed cleavages, 1; Variable modifications, Oxidation (M), Carbamidomethyl; Peptide tolerance, 100 ppm; MS/MS tolerance, 0.1 Da. Only proteins with a MASCOT score >50 are reported here and redundant hits were determined and removed from the 2DE/MALDI-Tof/MS dataset

BLAST searches, functional annotation and secretion signal prediction

This software suite was used for four procedures: 1) to assign preliminary annotation to assembled and singleton ESTs and the proteins identified by mass spectrometry by conducting BLASTX and BLASTP searches against the NCBI nr database (as of May, 2010) reporting matches with a HSP cutoff value $\geq$ 33 and an $E \leq 10^{-3}$ value, 2) to match EST sequences to pea aphid proteins available in the same databases (by reporting the top BLAST match), 3) to assign functional annotation to transcript-supported, MS-identified proteins and the final predicted salivary gland secretome by assigning GO terms using the InterProScan, ANNEX and GOSlim options in the Blast2GO suite and 4) to identify EST open reading frames, transcript supported

and MS-identified proteins with a potential peptide secretion signal. EST contigs and singletons were also BLASTn searched against gene databanks arising from the pea aphid genome sequencing project (BLAST facility of AphidBase; http://genoweb1.irisa.fr/AphidBase/Blast/Blast.php) as the non -redundant databases used by Blast2Go did not contain the most up-to-date official gene/protein lists. BLAST hits with an $E \leq 10^{-3}$ value and >90% sequence similarity were reported. The corresponding protein sequence for each of the returned top mRNA BLAST hit was obtained from the official protein consensus set (ACYPIprotein.fasta available from AphidBase). These proteins are hereafter described as transcript-supported proteins and were analyzed further. MS-identified or transcript-supported proteins were deemed non-annotatable if their current annotation or best BLAST hit comprised the text hypothetical (or conserved hypothetical) or if the best BLAST hit to a protein from another organism had no associated annotation, assigned GO term or identifiable InterPro domains or motifs. All proteins deemed non-annotatable (based on criteria described above) were augmented manually after searching the gene report page for each "hypothetical" aphid protein identifier and the GenBank submission for the best BLAST match for proteins originating from non aphid organisms.

Phylogenetic analysis of candidate secreted salivary proteins

Phylogenetic trees were reconstructed based on parsimony analysis of these coding sequence alignments in MEGA4 [9]. Relative rates of evolution between salivary gland genes and their nearest paralogs/orthologs (relative to an outgroup sequence), were estimated using relative rate tests in MEGA4. Aphid homologs were always preferentially selected for this analysis when available; alternatively the nearest homolog from other insects, towards the base of the tree, was

selected. Rates of synomymous (dS) and non-synonymous (dN) substitution were analysed using Codeml (PAML4.2, [10]), using a branch-site model to test for positive selection on the foreground branch containing all the known salivary paralogs [11]. When the likelihood ratio test (LRT) indicated significant positive selection, a Bayes empirical Bayes (BEB) test was used to identify the specific sites under positive selection [12]. Phylogenetic analysis was not conducted on genes without homologs in other insects, even if multiple paralogs were present in the pea aphid genome, because (1) relative rate tests require an outgroup, and (2) dN/dS is not valid as an intraspecific measure of selection [13].

REFERENCES

1. Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M., Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels. *Annals of Chemistry* **1996,** 68, (5), 850-858.
2. Carolan, J., C. ; Fitzroy, C., I. J. ; Ashton, P., D. ; Douglas, A., E. ; Wilkinson, T., L. , The secreted salivary proteome of the pea aphid *Acyrthosiphon pisum* characterised by mass spectrometry. *Proteomics* **2009,** 9, (9), 2457-2467.
3. Eng, J. K.; McCormack, A. L.; Yates, J. R. I., An approach to correlate MS/MS data to amino acid sequences in a protein database. . *J. Am. Soc. Mass Spectrom.* **1994,** 5, 976-989.
4. Anderson, K.; Monroe, M.; Daly, D., Estimating probabilities of peptide database identifications to LC-FTICR-MS observations. *Prot. Sci.* **2006,** 4, (1), 1.
5. Nesvizhskii, A. I.; Aebersold, R., Analysis, statistical validation and dissemination of large-scale proteomics data sets generated by tandem MS. *Drug Discov. Today* **2004,** 9, 173-181.
6. Nesvizhskii, A. I.; Vitek, O.; Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Meth.* **2007,** 4, (10), 787-797.
7. Rabilloud, T.; Valette, C.; Lawrence, J. J., Sample application by in-gel rehydration improves the resolution of 2-dimensional electrophoresis with immobilized pH- gradients in the first-dimension. *Electrophoresis* **1994,** 15, (12), 1552-1558.
8. Gorg, A.; Postel, W.; Weser, J.; Gunther, S.; Strahler, J. R.; Hanash, S. M.; Somerlot, L., Elimination of point streaking on silver stained two-dimensional gels by addition of iodoacetamide to the equilibration buffer. *Electrophoresis* **1987,** 8, (2), 122-124.
9. Tamura, K.; Dudley, J.; Nei, M.; Kumar, S., MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol. Biol. and Evol.* **2007,** 24, (8), 1596-1599.
10. Yang, Z., PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **2007,** 24, (8), 1586-1591.
11. Zhang, J.; Nielsen, R.; Yang, Z., Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **2005,** 22, (12), 2472-2479.

12. Yang, Z.; Wong, W. S. W.; Nielsen, R., Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **2005,** 22, (4), 1107-1118.
13. Kryazhimskiy, S.; Plotkin, J. B., The Population Genetics of dN/dS. *PLoS Genet.* **2008,** 4, (12), e1000304.