

RESEARCH ARTICLE

Positive selection and comparative molecular evolution of reproductive proteins from New Zealand tree weta (Orthoptera, *Hemideina*)

Victoria G. Twort^{1,2#a*}, Alice B. Dennis^{2#b}, Duckchul Park², Kathryn F. Lomas³, Richard D. Newcomb^{1,4}, Thomas R. Buckley^{1,2}

1 School of Biological Sciences, The University of Auckland, Auckland, New Zealand, **2** Landcare Research, Auckland, New Zealand, **3** CSIRO, Melbourne, Australia, **4** The New Zealand Institute for Plant and Food Research Ltd, Auckland, New Zealand

#a Current address: Department of Biology, Lund University, Lund, Sweden

#b Current address: Evolutionary Biology & Systematic Zoology, Institute for Biochemistry & Biology, University of Potsdam, Potsdam, Germany

* vtwort@gmail.com



OPEN ACCESS

Citation: Twort VG, Dennis AB, Park D, Lomas KF, Newcomb RD, Buckley TR (2017) Positive selection and comparative molecular evolution of reproductive proteins from New Zealand tree weta (Orthoptera, *Hemideina*). PLoS ONE 12(11): e0188147. <https://doi.org/10.1371/journal.pone.0188147>

Editor: J. Joe Hull, USDA Agricultural Research Service, UNITED STATES

Received: May 22, 2017

Accepted: November 1, 2017

Published: November 13, 2017

Copyright: © 2017 Twort et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Raw sequencing data are available from GenBank Short Read Archive (BioProject: PRJNA353021). Assembled transcriptomes are available from TSA (accession numbers: GFBX00000000 and GFBW00000000). Sanger sequences are available from Genbank (accession numbers: KY999988 - KY999999, MF000001 - MF000301).

Funding: Funding was provided by the Allan Wilson Centre and core funding for Crown

Abstract

Animal reproductive proteins, especially those in the seminal fluid, have been shown to have higher levels of divergence than non-reproductive proteins and are often evolving adaptively. Seminal fluid proteins have been implicated in the formation of reproductive barriers between diverging lineages, and hence represent interesting candidates underlying speciation. RNA-seq was used to generate the first male reproductive transcriptome for the New Zealand tree weta species *Hemideina thoracica* and *H. crassidens*. We identified 865 putative reproductive associated proteins across both species, encompassing a diverse range of functional classes. Candidate gene sequencing of nine genes across three *Hemideina*, and two *Deinacrida* species suggests that *H. thoracica* has the highest levels of intra-specific genetic diversity. Non-monophyly was observed in the majority of sequenced genes indicating that either gene flow may be occurring between the species, or that reciprocal monophyly at these loci has yet to be attained. Evidence for positive selection was found for one lectin-related reproductive protein, with an overall omega of 7.65 and one site in particular being under strong positive selection. This candidate gene represents the first step in the identification of proteins underlying the evolutionary basis of weta reproduction and speciation.

Introduction

Reproductive associated proteins have been shown to have increased evolutionary rates, and diverge rapidly between related taxa [1–6]. In particular, proteins present in the seminal fluid (SFPs) have been identified as often evolving under positive selection [4, 7–9]. In insects, the synthesis and secretion of SFPs occurs within male reproductive tract secretory tissues, such as accessory glands and testis [10, 11]. SFPs encompass a diverse range of functional classes and

Research Institutes from the Ministry of Business, Innovation and Employment's Science and Innovation Group. The New Zealand Institute for Plant and Food Research Ltd provided support in the form of a salary to RDN. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The New Zealand Institute for Plant and Food Research Ltd provided support in form of a salary to RDN. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

are involved in the modulation or induction of post mating responses in females [12–18]. In addition, SFPs have been identified as playing a key role in reproductive isolation between diverging lineages [2, 19–22].

Many *Drosophila* SFPs show increased evolutionary rates when compared to non-seminal proteins [1, 23, 24] and show evidence of positive selection [7–9, 25, 26]. However, not all SFPs exhibit rapid evolution; some show signs of evolutionary conservation [27–30], while others exhibit both conservation and rapid evolution in different regions of a single protein [31–33]. *Drosophila* SFPs have been shown to play a key role in reproductive isolation through species-specific gamete use, whereby SFPs need to have a specific structure and binding affinity for successful reproduction to occur [34]. Studies of orthopteran taxa have revealed similar patterns. In particular, studies on crickets have shown that SFPs have a higher level of divergence when compared with non-seminal proteins, with a significant proportion being under positive selection [7, 8, 25, 35]. Despite these studies in crickets, very little is known about SFP evolution in other orthopteran taxa.

Insects from the orthopteran family Anostomatidae are collectively known in New Zealand by their Māori name, weta, and represent an important component of the native forest ecosystem [36]. All Tree (*Hemideina*) and Giant (*Deinacrida*) weta species are endemic to New Zealand, and include both relatively widespread and threatened species. The larger *Deinacrida* species have limited distributions and abundance, with 10 of the 11 species being under conservation management [37–40]. There are seven species of *Hemideina* distributed throughout New Zealand, most of which are abundant [37–41]. *Hemideina crassidens* (Blanchard) has the largest distribution of all weta species, with populations distributed in the south of the North Island, as well as in the north and west coast of the South Island in New Zealand [42]. *Hemideina crassidens* has two chromosomal races (15 and 19) that are morphologically identical and successfully produce offspring in laboratory crosses [43, 44]. *Hemideina thoracica* (White) is found in the upper three quarters of the North Island [39]. All populations are morphologically similar despite there being eight chromosomal races with diploid chromosome numbers ranging from 11 to 24 [45]. Individual populations have been shown to exhibit only a single karyotype, with interbreeding occurring in the narrow regions of contact [45–47]. The presence of multiple chromosomal races within these species indicates that chromosomal differences are insufficient to lead to reproductive isolation [45, 48]. *Hemideina trewicki* has the smallest distribution of the three species, occurring only in southern and central Hawke's Bay. In the northern parts of the range, *H. trewicki* is sympatric with *H. thoracica* [42, 49] and has one known chromosomal race [48]. Although much progress has been made in revealing patterns of speciation and hybridisation in tree weta, the molecular basis of mate recognition, fertilisation and other reproductive processes are not known. The present study describes the male reproductive transcriptomes for *H. crassidens* and *H. thoracica*. We identify putative male reproductive associated proteins, and investigate patterns of divergence of nine genes between three *Hemideina* species. We test the hypothesis that male reproductive associated proteins have elevated rates of positive selection compared to general metabolic, or housekeeping, genes.

Materials and methods

Sample collection

Hemideina thoracica, *H. crassidens* and *Deinacrida mahoenui* (Gibbs) specimens were collected across their known distributions between 2010 and 2012 (S1 Table), by day and night searching. All samples were collected under a permit issued by the Department of Conservation

(CA-31615-OTH). Insects were transported live to Landcare Research, Auckland, and then snap frozen and stored at -80°C .

Transcriptome sequencing

Accessory gland and testis tissue from one adult male *H. thoracica* and *H. crassidens* (S1 Table) were dissected under a dissecting microscope in 100% ethanol. Total RNA was extracted from each tissue using TRIzol RNA extraction reagent (Life Technologies) according to the manufacturer's protocol. A further RNA clean-up was performed using the RNeasy Mini Kit (Qiagen). RNA quality and quantity was determined using a Nanodrop (ThermoScientific) and an Agilent 2100 bioanalyzer (Agilent Technologies). High quality total RNA was used to synthesise cDNA using the SMARTerTM PCR cDNA synthesis kit (Clontech) with a modified oligo-dT primer (Cap-TRSA-CV) [50]. Double stranded cDNA was purified using AMPure beads (Agencourt). Library quality was assessed using an Agilent 2100 bioanalyzer (Agilent Technologies) and quantified with the Quanti-iTM Picogreen[®] assay (Life Technologies). Cleaned cDNA was fed into the Rapid Library Preparation Protocol (Roche, GS Junior Titanium Series, June 2010) at the fragment end repair step, with each tissue sample being MID barcoded. The resulting libraries were pooled by tissue and sequenced in two runs on a 454 GS Junior (Roche) at Landcare Research (Auckland).

Pre-processing, assembly and annotating RNA-seq data

Raw sequences were split by MID barcode using Geneious V5.4.6 [51]. Reads with ambiguous bases and low quality sequences were removed using SnoWhite 1.1.14 [52]. The primer and adaptor sequences were removed using CUTADAPT V1.1 [53]. Poly A/T tails longer than 15 bp from either end of the reads, and reads shorter than 50 bp were removed using PRINSEQ LITE V0.16 [54]. Cleaned reads were *de novo* assembled with Newbler GS *de novo* Assembler (V. 2.5.3), with default parameters, a minimum overlap of 25 bp and a minimum overlap identity of 95%. Redundancy in the alignment was removed using cd-hit-est V. 4.5.6 [55]. Poor *de novo* assembly of the *H. thoracica* dataset was observed, due to a lower sequencing quality of the testis library run. Therefore, in order to obtain a more representative dataset additional assembly steps were undertaken. The cleaned, trimmed reads from both *H. thoracica* libraries were reference assembled against the *H. crassidens* transcriptome using the Roche GS reference mapper (version 2.5.3, default parameters, except for a 25 bp overlap). The purpose of this reference assembly was to obtain contigs that were unassembled in the *de novo* assembly due to inadequate coverage. The reference and *de novo* assemblies were combined and subjected to a second round of redundancy removal with cd-hit-est. Preliminary tests on the combined assembly showed similar levels of blast homology, GO annotation and assembly statistics; therefore, this assembly replaced the *de novo* assembly for *H. thoracica* in downstream analyses. Singletons (unassembled reads) were excluded from downstream analysis. The assemblies were annotated using Blastx V2.6.0 [56] (e-value $< 1e^{-5}$) against the GenBank non-redundant (nr) protein database (downloaded July 2017). Transcripts were searched for conserved protein domains with InterProScan [57] and GO terms were assigned using Blast2GO v2.8 [58]. Full-length transcripts were identified using Full-Lengther [59].

Identification of reproductive associated, orthologous, and candidate genes

Orthologous genes were identified using a bidirectional best hit method, which has been shown to outperform more complex algorithms for orthology predication [60]. A pair-wise

reciprocal blastn approach was carried out in Geneious (e-value threshold $1e^{-3}$) with orthologues being called if the best blast hit was identical in both directions.

Putative reproductive associated genes were identified in a two-step process. First, transcripts were identified based on mapping counts of *D. fallai* muscle RNA-seq reads downloaded from SRA (SRA accession: SRR5965744) using RSEM [61]. Second, transcripts unique to the reproductive transcriptomes in *H. thoracica* and *H. crassidens* were identified by mapping the *D. fallai* short reads to each assembly using Bowtie2 [62] and identifying the transcripts with no counts. Within these candidates, signal peptides, cellular location and the presence of trans-membrane domains were identified with SignalP (v4.1) [63], ProtComp v9.0 (<http://linux1.softberry.com>) and TMHMM v2.0 [64], respectively. Transcripts were retained as putative reproductive proteins if they had one of the following: (i) signal peptide, (ii) cellular localisation as extracellular and/or plasma membrane, (iii) transmembrane helix.

Candidate genes for downstream evolutionary analysis were chosen from among the contigs identified in the reproductive and orthologous gene screen. Candidates from the reproductive gene search were chosen based on their annotation, level of similarity (cut-off of 60%) and sequence length (minimum 400 bp). Orthologous candidates were based on a minimum transcript overlap of 200 bp between *H. thoracica* and *H. crassidens* transcripts, and the level of similarity at the amino acid and nucleotide level. Lastly, general metabolic control genes were chosen based on a minimum contig length of 300 bp and their involvement in general cellular processes, thereby ensuring tissue wide expression.

Sequencing of candidate genes

For Sanger sequencing samples, total RNA extractions from testis tissue followed the methods described above for the RNA-seq samples. Contaminating DNA was removed from RNA extractions prior to cDNA synthesis using TURBO DNase (Invitrogen). The first strand cDNA synthesis used the SuperScript III First Strand Kit (Invitrogen) following the manufacturer's protocol. cDNA libraries were subsequently amplified using 5 μ L first strand cDNA, 0.8 μ M random hexamer primer, 2X PCR buffer (Roche), 2.5 mM MgCl₂ (Roche), 0.2 mM dNTP (Roche), 1 U FastStart Taq DNA polymerase (Roche) in a total volume of 53 μ L. Amplifications were performed on a GeneAmp PCR system 9700 thermal cycler (Applied Biosystems) using the following parameters: 5 min at 95°C, 3 min at 50°C, 40 sec 72°C; 40 cycles of 40 sec at 94°C, 40 sec at 65°C and 40 sec at 72°C; and 10 min at 72°C.

Primers for nine candidate genes (*COI*, *Protease*, *Sflag*, *EFdelta*, *Unk2*, *Tkinase*, *Acp3*, *Acp4*, *Acp5*) were designed using Primer 3 [65] implemented within Geneious. Primer pairs were designed to amplify products of 200–1500 bp in length, have TMs of 60°C ($\pm 3^\circ$ C) and to have a GC content of 40–60% (S2 Table). Target genes were PCR amplified with reactions consisting of approximately 5 ng DNA, 1X PCR buffer (Roche), 2 mM MgCl₂ (Roche), 0.2 mM dNTP (Roche), 0.1–0.2 μ M forward and reverse primers (Sigma-Aldrich), 1 U FastStart Taq DNA polymerase (Roche), in a total volume of 25 μ L. Amplification were performed on a GeneAmp PCR system 9700 thermal cycler (Applied Biosystems) using the following parameters: 5 mins at 95°C; 40 cycles of 15 sec at 95°C, 30 sec at primer specific annealing temperature and 1 min 30 sec at 72°C; and 5 min at 72°C.

PCR products were sequenced using BigDye Terminator Cycle Sequencing Ready Reaction Mix v3.1 (Applied Biosystems). Cycle sequencing products were cleaned using the BigDye Xterminator Purification Kit (Applied Biosystems) and sequenced in both directions on the ABI Prism 3100 Genetic Analyzer (Applied Biosystems). Sequences were subsequently cleaned, trimmed and aligned using Geneious. In addition, the *D. mahoenui* and *D. fallai* transcriptomes (unpublished) were searched for *COI* and all candidate gene orthologues, respectively using a

bidirectional tblastx approach, and included in downstream analysis (sequences given in [S1 File](#)).

Haplotype reconstruction for sequences that exhibited heterozygosity was performed using PHASE V. 2.1 [66, 67] prior to calculating descriptive statistics in DnaSP V. 5 [68]. Tajima's D [69] and the McDonald-Kreitman test [70] were calculated in DNAsp. Haplotype networks were constructed using TCS V. 1.2.1 (Clement, Posada, Crandall 200), with gaps being considered as the 5th state and a 95% connection limit. The substitution model for the *COI* phylogeny was selected using the corrected Akaike information criterion [71] generated by jModel Test v.0.0.1 [72, 73]. A maximum likelihood phylogeny was constructed in Garli v2.0 [74] using 100 search and 1000 bootstrap replicates.

Inferring positive selection

Genes were identified for selection tests based on the number of non-synonymous changes. The three candidates with the most non-synonymous changes (*Acp3*, *Protease*, *Unk2*) were chosen for downstream analysis. For these three, a neighbour-joining phylogeny was generated for selection tests in Geneious. To screen for positive selection ω was estimated by maximum likelihood, using codon-based substitution models implemented in the CODEML package of PAML V. 4.5 [75]. The models implemented (M0, M1a, M2a, M3, M7, M8, M8a) are extensively described elsewhere [76–78]. Complex models (M2a, M3, M8) allow more than one category of ω , thereby allowing individual codons to be identified as under positive selection when the average ω across the whole gene indicates purifying selection. Likelihood ratio tests (LRTs) between nested models allows inference of positive selection acting on a sequence [75]. Codons under positive selection were identified using the Bayes empirical Bayes (BEB) method under the M8 model.

Results and discussion

Transcriptome assembly and characterisation

454 sequencing of *H. thoracica* and *H. crassidens* cDNA libraries resulted in a total of 254,628 reads, of which 73,012 and 59,465 reads were from *H. thoracica*, and 37,384 and 84,767 from *H. crassidens* testis and accessory gland tissue libraries, respectively. Raw sequences have been submitted to the GenBank Short Read Archive (BioProject: PRJNA353021). After trimming to remove bases with low quality scores, adapters and MID barcodes, 97% of the data remained. *De novo* assemblies were generated for each species as described above. The *H. crassidens* assembly generated 1,759 unigenes with an N50 of 608 bp and a maximum transcript size of 2,861 bp. In comparison, the *H. thoracica* assembly generated 2,691 unigenes with an N50 of 576 bp and a maximum transcript size of 1,860 bp. Both assemblies have been submitted to TSA under GFBX00000000 and GFBW00000000. A total of 890 and 1,537 transcripts were identified as being full length, respectively. Approximately 45% of the unigenes present in each assembly were functionally annotated using a tblastx search against the NCBI non-redundant database, with the species distribution of top matches overlapping among the two assemblies ([S1 Fig](#)). Functional annotation (GO) was similar across the two species, with the highest number of annotated transcripts related to cellular processes (GO:0009987) and metabolic process (GO:000812) ([Fig 1](#)). All unigenes were screened against the InterPro database, from which 2,690 and 1,753 *H. thoracica* and *H. crassidens* transcripts, respectively, were identified as containing conserved protein domains. The top 20 most frequent entries are shown in [Table 1](#). These top entries show a diverse range of predicted functions, including proteins associated with general house-keeping roles and other that have been linked to reproductive functions. Domains identified include ubiquitin (IPR000626, IPR029071) and translation protein

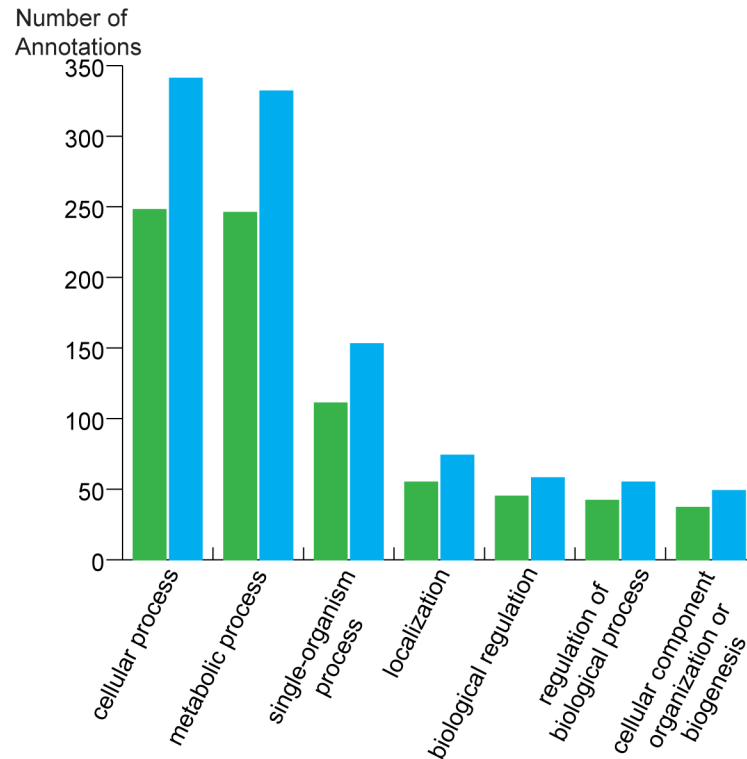


Fig 1. Distribution of biological function annotation of two *Hemideina* transcriptomes. Green bars: *H. crassidens* and blue bars: *H. thoracica*.

<https://doi.org/10.1371/journal.pone.0188147.g001>

(IPR008991) domains. Many of the genes represented in these groups are probably highly conserved genes involved in the processes of transcription and protein degradation. Other domains identified, such as proteases (IPR001254, IPR018114) and protease inhibitors (IPR00215, IPR023796), have been associated with a number of reproductive functions, such as the modifying postmating changes in females [79], and are frequently identified in the study of insect SFPs [9, 80, 81].

Identification of orthologous transcripts and reproductive proteins

The bidirectional best hit method identified 113 pairs of sequences that were putatively orthologous between the two species (S3 Table). For simplicity, these genes are hereafter referred to as orthologous. Some orthologues will have been missed using this approach, as in *D. melanogaster* the evolutionary rate of some SFPs has been shown to be so rapid that they lack any detectable similarity with their homologues from other *Drosophila* species [1, 24, 82, 83].

Putative reproductive associated transcripts were identified by mapping *D. fallai* muscle RNA-seq reads to each transcriptome. Transcripts unique to the reproductive transcriptomes (those lacking mapped reads) were further analysed to identify putative reproductive associated proteins. Of the transcripts unique to the reproductive transcriptome, 258 and 337 meet the criteria of having a signal peptide, transmembrane helix or localisation at the plasma membrane or extracellular for *H. thoracica* and *H. crassidens*, respectively (S4 Table). Roughly 19% of these had Blast hits, indicating that those lacking homology might be novel proteins or proteins highly diverged in weta. Among the genes with Blast hits, the most common molecular function GO terms were serine-type endopeptidase activity (GO:0004252), serine-type

Table 1. The 20 most encountered InterPro accessions present in two *Hemideina* transcriptomes.

| <i>H. thoracica</i> | | | <i>H. crassidens</i> | | |
|---------------------|--|-------------------|----------------------|---|-------------------|
| InterPro Entry | InterPro Description | Number of Contigs | InterPro Entry | InterPro Description | Number of Contigs |
| IPR012336 | Thioredoxin-like fold | 23 | IPR016187 | C-type lectin fold | 17 |
| IPR011991 | Winged helix-turn-helix DNA-binding | 18 | IPR016186 | C-type lectin-like/link domain | 16 |
| IPR029071 | Ubiquitin-related domain | 17 | IPR001304 | C-type lectin-like | 14 |
| IPR027417 | P-loop containing nucleoside triphosphate | 17 | IPR012336 | Thioredoxin-like fold | 13 |
| IPR000504 | RNA recognition motif domain | 15 | IPR001254 | Serine proteases, trypsin domain | 13 |
| IPR000626 | Ubiquitin domain | 15 | IPR009003 | Peptidase S1, PA clan | 13 |
| IPR016187 | C-type lectin fold | 12 | IPR011991 | Winged helix-turn-helix DNA-binding | 10 |
| IPR010987 | Glutathione S-transferase, C-terminal-like | 12 | IPR011992 | EF-hand-like domain | 9 |
| IPR014756 | Immunoglobulin E-set | 11 | IPR023796 | Serpin domain | 8 |
| IPR015943 | WD40/YVTN repeat-like-containing domain | 11 | IPR027417 | P-loop containing nucleoside triphosphate | 8 |
| IPR016186 | C-type lectin-like/link domain | 11 | IPR029277 | Single domain Von Willebrand factor type C domain | 8 |
| IPR011992 | EF-hand domain pair | 10 | IPR013783 | Immunoglobulin-like fold | 7 |
| IPR004045 | Glutathione S-transferase, N-terminal | 10 | IPR008037 | Pacifastin domain | 7 |
| IPR000477 | Reverse transcriptase domain | 10 | IPR000477 | Reverse transcriptase domain | 7 |
| IPR032675 | Leucine-rich repeat domain, L domain-like | 9 | IPR002048 | EF-hand domain | 7 |
| IPR008991 | Translation protein SH3-like domain | 9 | IPR029071 | Ubiquitin-related domain | 7 |
| IPR005203 | Hemocyanin, C-terminal | 9 | IPR016040 | NAD(P)-binding domain | 7 |
| IPR001304 | C-type lectin-like | 9 | IPR000626 | Ubiquitin domain | 6 |
| IPR016040 | NAD(P)-binding domain | 8 | IPR007110 | Immunoglobulin-like domain | 6 |
| IPR013766 | Thioredoxin domain | 7 | IPR013766 | Thioredoxin domain | 6 |

<https://doi.org/10.1371/journal.pone.0188147.t001>

endopeptidase inhibitor activity (GO:0004867) and ATP binding (GO:0005524) (S2 Fig). Overall, the GO categories identified in the reproductive gene search are similar to categories commonly seen when studying insect SFPs [9, 16, 81, 84]. Various peptidase and peptidase regulators are among the reproductive proteins identified, and are believed to be essential for the regulation of reproduction through proteolytic cascades [28]. These types of proteins constitute a large proportion of the *D. melanogaster* [80], *Anopheles gambiae* [85], *Aedes aegypti* [86], *Lutzomyia longipalpis* [87] and *Clitarchus hookeri* [84] identified SFP and accessory gland proteins. The reproductive proteins identified in this screen are similar when compared with other insects, however a large proportion lack Blast hits. These unknown transcripts indicate the presence of novel or highly divergent proteins, and provide a large resource for the study of sexual reproduction and speciation in the New Zealand weta [80, 85–90].

Candidate gene identification and sequencing

To study the patterns of molecular evolution of weta reproductive proteins, alignments of the candidates generated from the transcriptome sequencing were used to design PCR primers from nine genes (Table 2). Six putative reproductive proteins (*Acp3*, *Acp4*, *Acp5*, *Sflag*, *Tkinase*, and *Protease*) were identified as interesting candidates for downstream evolutionary analysis based on their blast homologies. The contig *Unk2*, despite lacking significant blast homology, was included for further analysis based on the interesting amino-acid pairwise identity

Table 2. Candidate genes for downstream evolutionary analysis.

| | Gene | Annotation | <i>H. thoracica</i> | <i>H. crassidens</i> | AA % identity | Nuc % identity |
|----------------------------|-----------------|--------------------------------|---------------------|----------------------|---------------|----------------|
| General metabolic controls | <i>COI</i> | Cytochrome oxidase subunit I | contig01355 | contig00784 | 80.2 | 81.3 |
| | <i>Efdelta</i> | Elongation factor 1 delta | refmapcontig00958 | contig00992 | 99.2 | 97.5 |
| Reproductive proteins | <i>Protease</i> | Serine protease snake-like | contig02289 | contig01653 | 90 | 94.1 |
| | <i>Tkinase</i> | Testis specific ser/thr kinase | contig01847 | — | — | — |
| | <i>Sflag</i> | Sperm flagellar protein | contig01892 | — | — | — |
| | | | | | | |
| | <i>Acp5</i> | Accessory gland protein | refmapcontig01086 | contig00834 | 93.5 | 98.1 |
| | <i>Acp4</i> | Accessory gland protein | refmapcontig00651 | contig01386 | 90.4 | 97.3 |
| | <i>Acp3</i> | Accessory gland protein | refmapcontig00312 | contig01836 | 88.5 | 90.8 |
| | <i>Unk2</i> | —N/A— | refmapcontig00670 | contig1351 | 93.8 | 93.7 |

AA, amino acid

Nuc, nucleotide

<https://doi.org/10.1371/journal.pone.0188147.t002>

observed during the orthologue gene screen. In addition, one nuclear (*Efdelta*) and one mitochondrial (*COI*) gene were included as general metabolic controls due to their tissue-wide expression. All nine genes were successfully amplified and sequenced from cDNA from, 19 *H. thoracica*, 11 *H. crassidens*, 5 *H. trewicki* and 1 *D. mahoenui* (outgroup) individuals. In addition, transcripts were identified within our unpublished RNA-seq data for *D. fallai* for all nine genes. All sequences have been submitted to NCBI GenBank (accession numbers: KY999988—KY999999, MF000001—MF000301).

Polymorphism, divergence and molecular evolution

Sequence data was obtained for *COI* from the majority of *Hemideina* individuals, resulting in a 672 bp alignment. The maximum likelihood phylogeny (Fig 2) supports each *Hemideina* species as monophyletic, with *H. trewicki* being sister to *H. crassidens*. The pairing of *H. crassidens* with *H. trewicki* is consistent with previous genetic and allozyme studies [41, 91].

Haplotype networks were constructed for every gene, except *COI*, rather than phylogenetic trees. Very little sequence divergence was present within these genes, thereby reducing the statistical power of phylogenetic reconstruction [92, 93]. Two genes (*Unk2*, *Protease*, Fig 3A and 3D) showed monophyletic groupings of alleles, whereas the remaining six genes showed the presence of shared alleles between at least two of the species sequenced (Figs 3 and 4). Generally speaking, both the reproductive and general metabolic control genes showed similar patterns. Previous work has shown that at both a genetic [39, 94] and karyotypic [48] level *H. crassidens* and *H. trewicki* are more genetically similar to each other than either is to *H. thoracica*, and hence are more likely to produce fertile hybrids. Of the 8 genes sequenced, only *Protease* and *Unk2* show a complete lack of allele sharing among the species. The genes *Tkinase*, *Sflag*, *Acp4*, and *Acp5*, show sharing of alleles between *H. thoracica* and *H. crassidens*. The geographically restricted *H. trewicki* shares alleles with *H. crassidens* (*Efdelta*, *Acp3*, *Acp5*) and *H. thoracica* (*Acp5*). The two *Deinacrida* species are well differentiated from the three sampled *Hemideina* species at all loci, in agreement with previous studies [39, 91]. Overall these results demonstrate genetic differentiation among the three tree weta species, in agreement with McKean *et al.* [48].

A summary of intraspecific sequence variation is shown in Table 3. Intraspecific variation within *H. thoracica* was greater than that observed in *H. crassidens* and *H. trewicki* for the majority of genes examined. This is consistent with allozyme and mitochondrial DNA studies

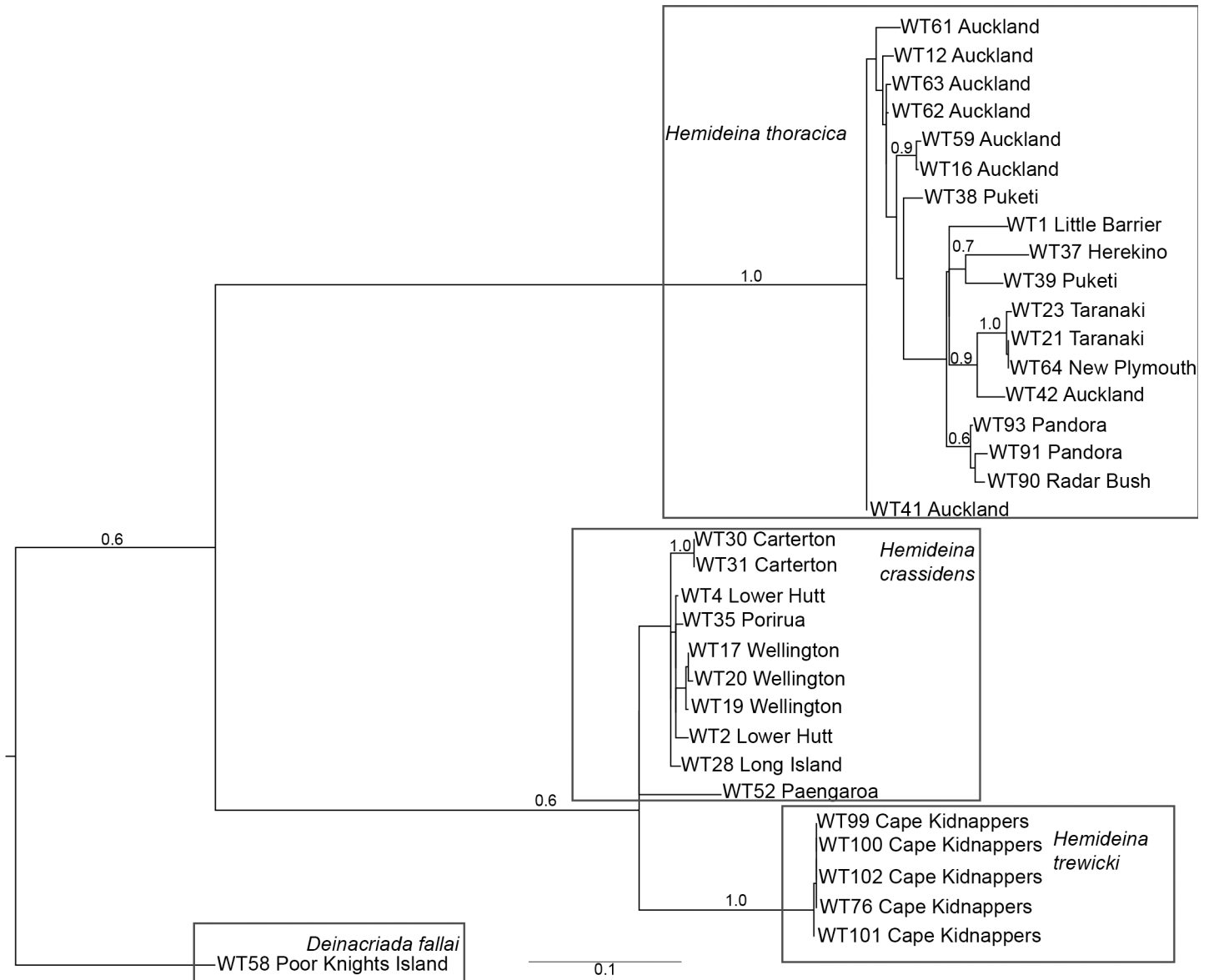


Fig 2. Maximum likelihood phylogeny constructed using mitochondrial cytochrome oxidase subunit I (COI) DNA sequences from individuals representing three *Hemideina* and one *Deinacrida* species. Bootstrap support values greater than 0.5 are indicated. Scale bar represents the number of substitutions per site.

<https://doi.org/10.1371/journal.pone.0188147.g002>

that show *H. thoracica* has the highest levels of intraspecific diversity of all *Hemideina* species [39, 41, 91]. This signature is consistent with inferences that the range of *H. thoracica* has recently expanded southwards [46, 95], while in comparison, little to no diversity was observed within the *H. trewicki* samples. This is not an unexpected result as all samples originated from a single population.

The reproductive-associated candidate genes tended to display higher levels of within species diversity than the general metabolic controls. However, in some reproductive genes, especially *Acp5* and *Tkinase*, the observed level of genetic diversity was at the same or similar levels as the general metabolic controls. The relatively lower levels of diversity in these two

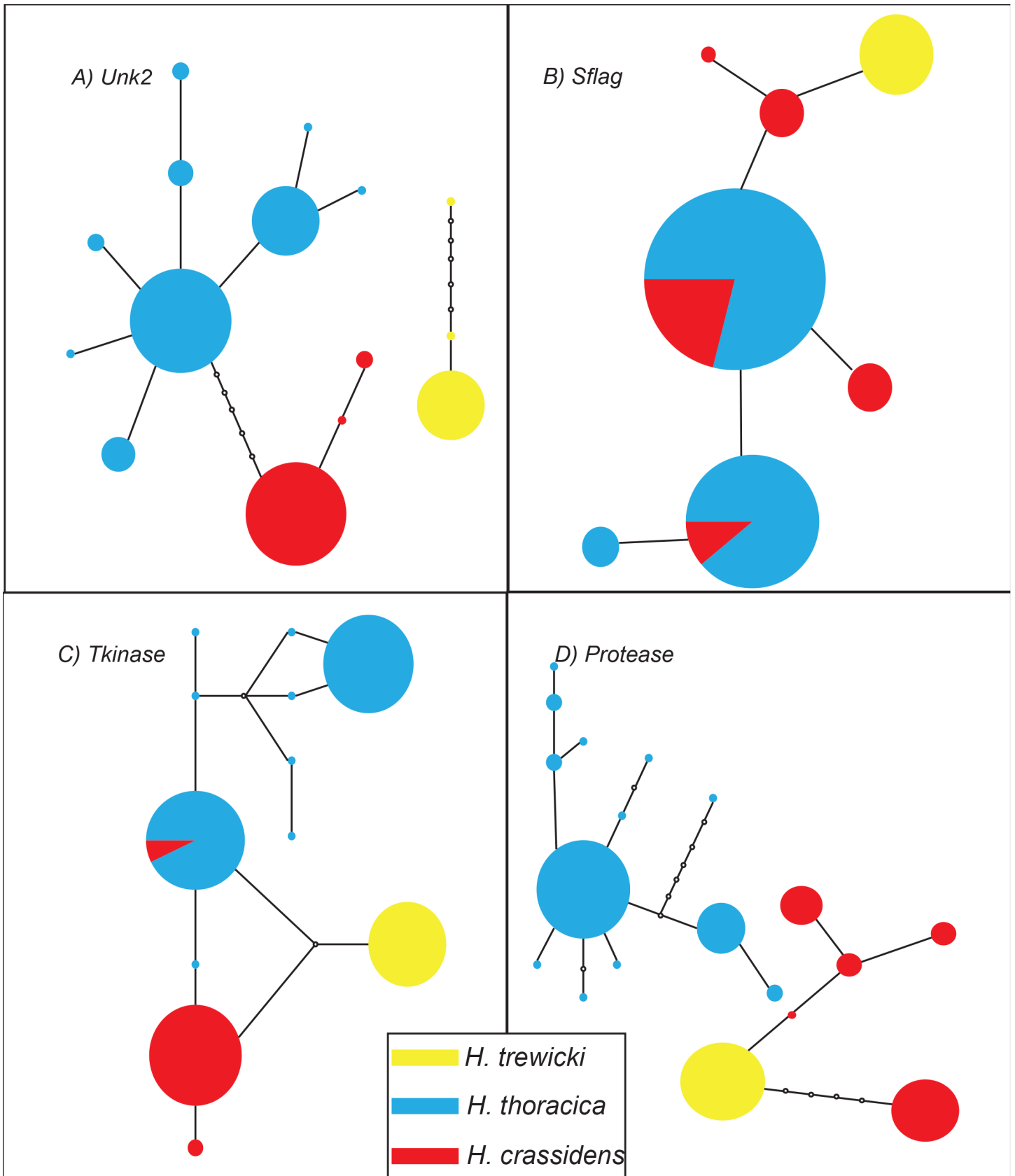


Fig 3. Haplotype network of A) *Unk2*, B) *Sflag*, C) *Tkinase*, and D) *Protease* gene regions. Circles represent different haplotypes, with the circles area being proportional to the frequency of each haplotype. Lines between haplotypes represent mutational steps between sequences. The empty circles represent inferred unsampled haplotypes. Colours correspond to species: Red, *H. crassidens*; Blue, *H. thoracica*; Yellow, *H. trewicki*.

<https://doi.org/10.1371/journal.pone.0188147.g003>

reproductive genes suggest they are functionally constrained. However, this requires further investigation, as only two metabolic controls were included in this study, and only partial transcripts were sequenced. Possible explanations for the lower diversity, include the sequenced region may be in a functionally constrained region of the protein, with relaxed selection occurring upstream or downstream of the sequenced region, or these genes may be located in regions of low recombination.

Within the *Acp3* alignment an allelic variant containing a 24 bp indel was identified. All *H. thoracica* and two *H. crassidens* individuals have the insertion, while the remainder of *H. crassidens* and all *H. trewicki* samples lack the insertion. The 24 bp indel, appears to be a true allelic variant rather than the effect of preferential amplification of a paralogous gene as some individuals had only one copy of either the deletion or complete variant. If the deletion variant was in fact paralogous amplification then both copies would have been expected in all individuals of *H. crassidens*. InterProScan analysis revealed the presence of C-lectin type domains within the coding sequence. Lectins and lectin-related proteins have been shown to be involved in carbohydrate binding and the mediation of sperm-egg interactions [88–90], suggesting that this is an interesting reproductive candidate gene family.

At the species level, Tajima's D was significant for the serine protease gene (*Protease*) within *H. crassidens*, ($D = 2.17$, Table 4) which may indicate balancing selection or demographic influences. In addition, for *Acp4*, *Acp3*, *Acp5*, *Unk2*, and *Tkinase* Tajima's D statistics were negative for at least one species, thereby indicating an excess of rare or recent mutations that may be due to purifying selection or a recent demographic expansion, the latter of which has been observed for *H. crassidens* and *H. thoracica* [95]. Under the McDonald-Kreitman test no departures from neutrality were detected for any of the genes (S5 Table).

To study the patterns of molecular evolution of weta reproductive proteins, the ratio of nonsynonymous to synonymous substitution rates of protein-coding sequences (ω) was calculated for two reproductive genes (*Acp3*, *Protease*) and the unknown gene, *Unk2*. The candidate *Unk2* was included as part of the selection tests, as initial screening identified an ORF, which showed higher numbers of nonsynonymous than synonymous substitutions (Table 3). In the case of *Acp3* the mean ω ratio was >1 , indicating an excess of nonsynonymous changes across the protein coding region as a whole (Table 5). In contrast, *Unk2* and *Protease* had an $\omega < 1$ (Table 5). Omega ratios averaged over an entire protein coding region are typically < 1 , due to positive selection commonly acting on specific domains or residues [96], with evidence suggesting genes with mean ω ratios above 0.5 are experiencing episodes of adaptive evolution [1, 7, 97, 98]. Individual amino acid residues likely to have been influenced by selection were identified using site-based ω calculations, with likelihood ratio tests and chi-squared distributions used to assess the goodness of fit for a given model [77, 99, 100]. For comparison, we used six codon-substitution models to assess the mode of selection acting on each amino acid residue in our candidate genes. We found that for *Acp3* and *Protease* there is significant among-site variation in ω , with the M3 model permitting three ω values providing a significantly better fit to the data than the M0 model ($p\text{-value} < 0.05$; $df = 4$; M0:M3; Table 5, S6 Table). The M1a:M2a comparison was insignificant for all three genes. A more conservative approach for testing for positive selection is the M8:M8a comparison, under this model *Acp3* was identified as being under positive selection with a mean ω of 7.9 (Tables 5 and S6).

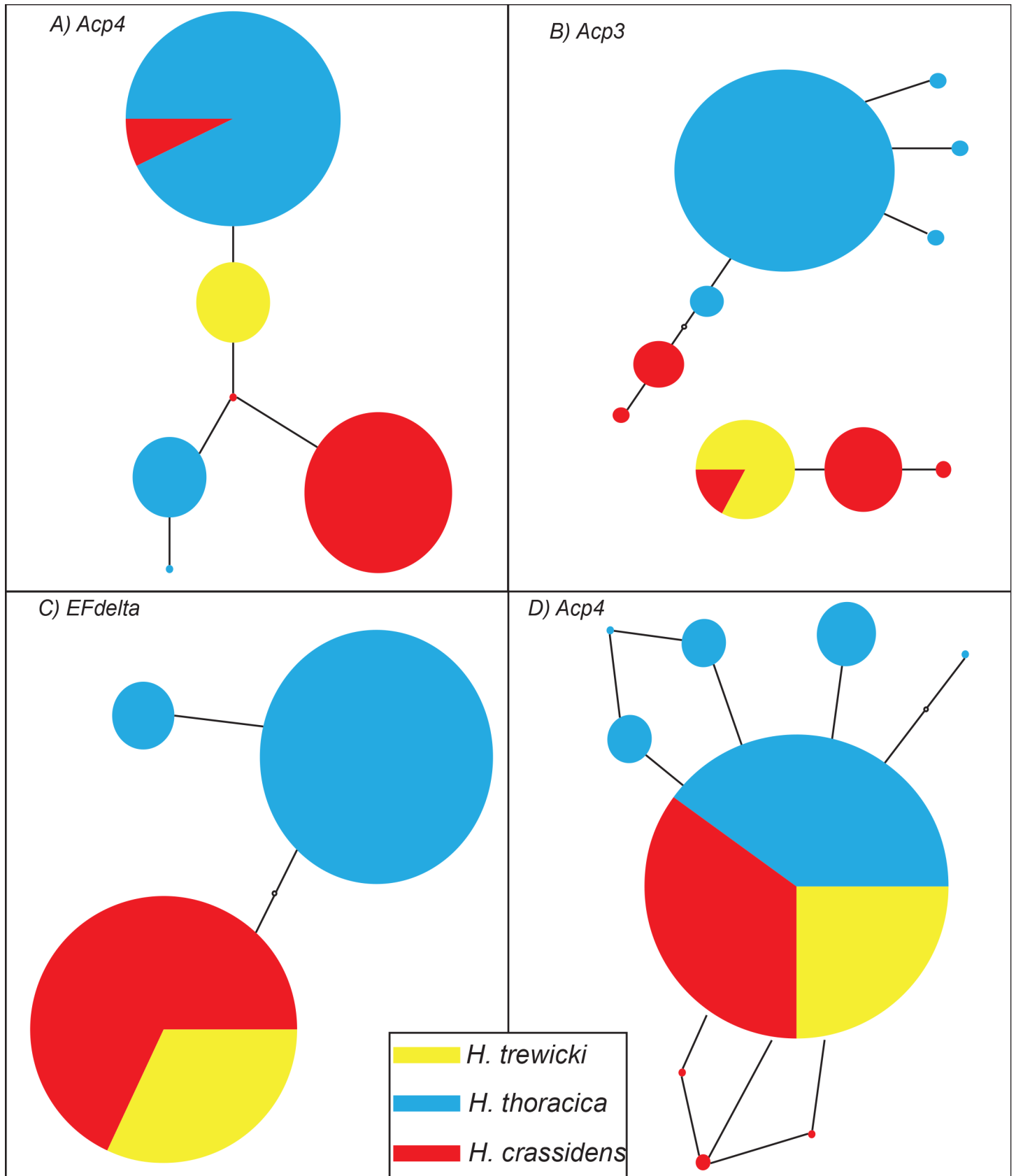


Fig 4. Haplotype network of A) *Acp4*, B) *Acp3*, C) *EFdelta*, and D) *Acp5* gene regions. Circles represent different haplotypes, with the circles area being proportional to the frequency of each haplotype. Lines between haplotypes represent mutational steps between sequences. The empty circles represent inferred unsampled haplotypes. Colours correspond to species: Red, *H. crassidens*; Blue, *H. thoracica*; Yellow, *H. trewicki*.

<https://doi.org/10.1371/journal.pone.0188147.g004>

A Bayes Empirical Bayes computation [101] implemented in PAML was used to assess the significance of the ω ratio at each codon position. Under the M8 model five sites were assigned to the positively selected class ($\omega > 1$), only one of which had a probability $> 95\%$ ($\omega = 6.366$, Table 5, Fig 5). Observed changes at three of the five sites were non-conservative (S7 Table). It is acknowledged that likelihood-based methods can produce high levels of false positives [102, 103], however the alternative parsimony-based models tend to be very conservative and have low power detecting true positives, particularly in small datasets such as this one [104, 105]. The five sites presented here represent testable hypotheses for functionally important regions under selection within *Acp3* that could be examined with a larger dataset and other analysis methods. However, it should be noted that the function of *Acp3* is unknown, as is the exact position of these sites within the overall protein structure.

Table 3. Summary statistics of intra-specific sequence variation within three *Hemideina* species.

| | Gene | Species | N | n | s | H | Hd | Eta | S | π | k | π_s | π_a | π_a/π_s | S | NS |
|----------------------------|----------------------|----------------------|----|-----|-----|-------|-------|-----|-------|-------|-------|---------|---------|---------------|---|----|
| Reproductive proteins | <i>Protease</i> | <i>H. thoracica</i> | 18 | 36 | 423 | 13 | 0.798 | 19 | 19 | 0.005 | 2.175 | 0.008 | 0.004 | 0.5 | 8 | 11 |
| | | <i>H. crassidens</i> | 10 | 20 | 423 | 5 | 0.768 | 8 | 8 | 0.008 | 3.711 | 0.118 | 0.007 | 0.0593 | 3 | 5 |
| | | <i>H. trewicki</i> | 5 | 10 | 423 | 1 | 0.000 | 0 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | — | 0 | 0 |
| | <i>Tkinase</i> | <i>H. thoracica</i> | 19 | 38 | 438 | 9 | 0.680 | 8 | 7 | 0.005 | 2.241 | 0.015 | 0.002 | 0.133 | 5 | 3 |
| | | <i>H. crassidens</i> | 11 | 22 | 438 | 3 | 0.255 | 3 | 3 | 0.001 | 0.355 | 0.002 | 0.000 | 0.000 | 2 | 1 |
| | | <i>H. trewicki</i> | 5 | 10 | 438 | 1 | 0.000 | 0 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | — | 0 | 0 |
| | <i>Sflag</i> | <i>H. thoracica</i> | 19 | 38 | 333 | 3 | 0.619 | 2 | 2 | 0.002 | 0.747 | 0.010 | 0.000 | 0.000 | 2 | 0 |
| | | <i>H. crassidens</i> | 11 | 22 | 333 | 5 | 0.797 | 5 | 5 | 0.005 | 1.662 | 0.014 | 0.002 | 0.142 | 3 | 2 |
| | | <i>H. trewicki</i> | 5 | 10 | 333 | 1 | 0.000 | 0 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | — | 0 | 0 |
| | <i>Acp3</i> | <i>H. thoracica</i> | 18 | 36 | 258 | 5 | 0.470 | 4 | 4 | 0.002 | 0.527 | 0.000 | 0.003 | — | 0 | 4 |
| | | <i>H. crassidens</i> | 11 | 22 | 258 | 5 | 0.727 | 12 | 11 | 0.020 | 4.762 | 0.031 | 0.017 | 0.548 | 4 | 8 |
| | | <i>H. trewicki</i> | 5 | 10 | 234 | 1 | 0.000 | 0 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | — | 0 | 0 |
| | <i>Acp4</i> | <i>H. thoracica</i> | 19 | 38 | 237 | 3 | 0.437 | 3 | 3 | 0.005 | 1.24 | 0.017 | 0.002 | 0.118 | 2 | 1 |
| | | <i>H. crassidens</i> | 11 | 22 | 237 | 3 | 0.177 | 3 | 3 | 0.001 | 0.355 | 0.005 | 0.000 | 0.000 | 2 | 1 |
| | | <i>H. trewicki</i> | 5 | 10 | 237 | 1 | 0.000 | 0 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | — | 0 | 0 |
| <i>Acp5</i> | <i>H. thoracica</i> | 18 | 36 | 459 | 6 | 0.763 | 5 | 5 | 0.002 | 1.111 | 0.011 | 0.000 | 0.000 | 5 | 0 | |
| | <i>H. crassidens</i> | 10 | 20 | 459 | 4 | 0.363 | 3 | 1 | 0.001 | 0.363 | 0.004 | 0.000 | 0.000 | 3 | 0 | |
| | <i>H. trewicki</i> | 5 | 10 | 459 | 1 | 0.000 | 0 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | — | 0 | 0 | |
| <i>Unk2</i> | <i>H. thoracica</i> | 17 | 34 | 372 | 9 | 0.813 | 8 | 7 | 0.004 | 1.303 | 0.005 | 0.003 | 0.600 | 3 | 5 | |
| | <i>H. crassidens</i> | 11 | 22 | 372 | 3 | 0.279 | 2 | 2 | 0.001 | 0.458 | 0.000 | 0.002 | — | 0 | 2 | |
| | <i>H. trewicki</i> | 5 | 10 | 372 | 3 | 0.378 | 7 | 7 | 0.004 | 1.556 | 0.008 | 0.003 | 0.375 | 3 | 4 | |
| General metabolic controls | <i>EFdelta</i> | <i>H. thoracica</i> | 19 | 38 | 363 | 2 | 0.341 | 1 | 1 | 0.001 | 0.341 | 0.000 | 0.001 | — | 0 | 1 |
| | | <i>H. crassidens</i> | 11 | 22 | 363 | 5 | 0.644 | 8 | 8 | 0.004 | 1.608 | 0.000 | 0.000 | — | 0 | 0 |
| | | <i>H. trewicki</i> | 5 | 10 | 363 | 1 | 0.000 | 0 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | — | 0 | 0 |
| | <i>COI</i> | <i>H. thoracica</i> | 18 | 18 | 672 | 17 | 0.993 | 134 | 119 | 0.050 | 34.22 | — | — | — | — | — |
| | | <i>H. crassidens</i> | 10 | 10 | 672 | 9 | 0.978 | 60 | 59 | 0.024 | 16.33 | — | — | — | — | — |
| | | <i>H. trewicki</i> | 5 | 5 | 672 | 2 | 0.400 | 1 | 1 | 0.001 | 0.400 | — | — | — | — | — |

N, number of individuals; n, number of alleles; s, number of sites; H, number of haplotypes; Hd, haplotype diversity; Eta, number of mutations; S, number of segregating sites; π , nucleotide diversity; k, average number of nucleotide differences; π_s , nucleotide diversity at synonymous sites; π_a , nucleotide diversity at nonsynonymous sites; S, total number of synonymous substitutions; NS, total number of nonsynonymous substitutions

<https://doi.org/10.1371/journal.pone.0188147.t003>

Table 4. Tajima's D test results for three *Hemideina* gene datasets.

| Gene | Species | D-statistic |
|----------|----------------------|-------------|
| Protease | <i>H. thoracica</i> | 2.175* |
| | <i>H. crassidens</i> | -1.759 |
| | <i>H. trewicki</i> | — |
| Tkinase | <i>H. thoracica</i> | 0.511 |
| | <i>H. crassidens</i> | -1.471 |
| | <i>H. trewicki</i> | — |
| EFdelta | <i>H. thoracica</i> | 0.629 |
| | <i>H. crassidens</i> | — |
| | <i>H. trewicki</i> | — |
| Acp3 | <i>H. thoracica</i> | -1.111 |
| | <i>H. crassidens</i> | 1.564 |
| | <i>H. trewicki</i> | — |
| Acp4 | <i>H. thoracica</i> | 0.637 |
| | <i>H. crassidens</i> | -1.471 |
| | <i>H. trewicki</i> | — |
| Acp5 | <i>H. thoracica</i> | -0.240 |
| | <i>H. crassidens</i> | -1.529 |
| | <i>H. trewicki</i> | — |
| Unk2 | <i>H. thoracica</i> | -0.913 |
| | <i>H. crassidens</i> | -0.440 |
| | <i>H. trewicki</i> | -1.573 |

*, p-value < 0.05.

<https://doi.org/10.1371/journal.pone.0188147.t004>

Conclusions

Here we present the first male reproductive transcriptomes for *H. crassidens* and *H. thoracica*, resulting in putative gene sets of 1,754 and 2,691 non-redundant gene sets, respectively. We identified 865 putative reproductive associated proteins, and 113 orthologs, from which nine candidates were used for downstream evolutionary analyses. Our results suggest that positive selection may be acting on some *Hemideina* SFPs; in contrast, we were unable to detect positive selection on the general metabolic control genes. The lectin-related *Acp3* gene shows evidence for selection acting along the gene as a whole and on particular amino acids. This

Table 5. Likelihood ratio tests of positive selection using PAML site-specific models.

| Gene (#sequences) | dN/dS | 2ΔI M0:M3 | 2ΔI M1a:M2a | 2ΔI M8:M8a | Positive Selection (%) |
|-------------------|-------|--------------|----------------|---------------|------------------------|
| <i>Acp3</i> (72) | 1.11 | 10.62* | 7.65 | 7.65* | 6.8 (1.4) |
| <i>Prot</i> (70) | 0.51 | 10.57* | 0.00 | 1.20 | 11–47 (0–47) |
| <i>Unk2</i> (68) | 0.72 | 1.22 | 0.17 | 0.17 | 4.4–16 (0) |

dN/dS is the average omega (ω) across all sites and branches calculated under M0. 2ΔI is given for each model comparison (M0:M3; M1a:M2a; M8:M8a), which is twice the difference between the log likelihood of the two nested site-specific models implemented in PAML. Models are judged to have a significantly better fit (* = P-value<0.05) based on the chi2 distribution with degrees of freedom proportional to the difference in the number of parameters between models; M0:M3 = 4; M1a:M2a = 4; M8:M8a = 50:50 mixture of point mass 0 and 1).

Parameters indicating positive selection are in bold. Percent positive selection indicates the proportion of sites across the gene predicted to have experienced positive selection, while the percentage in bracketed represents the proportion of sites identified with a >95% probability.

<https://doi.org/10.1371/journal.pone.0188147.t005>

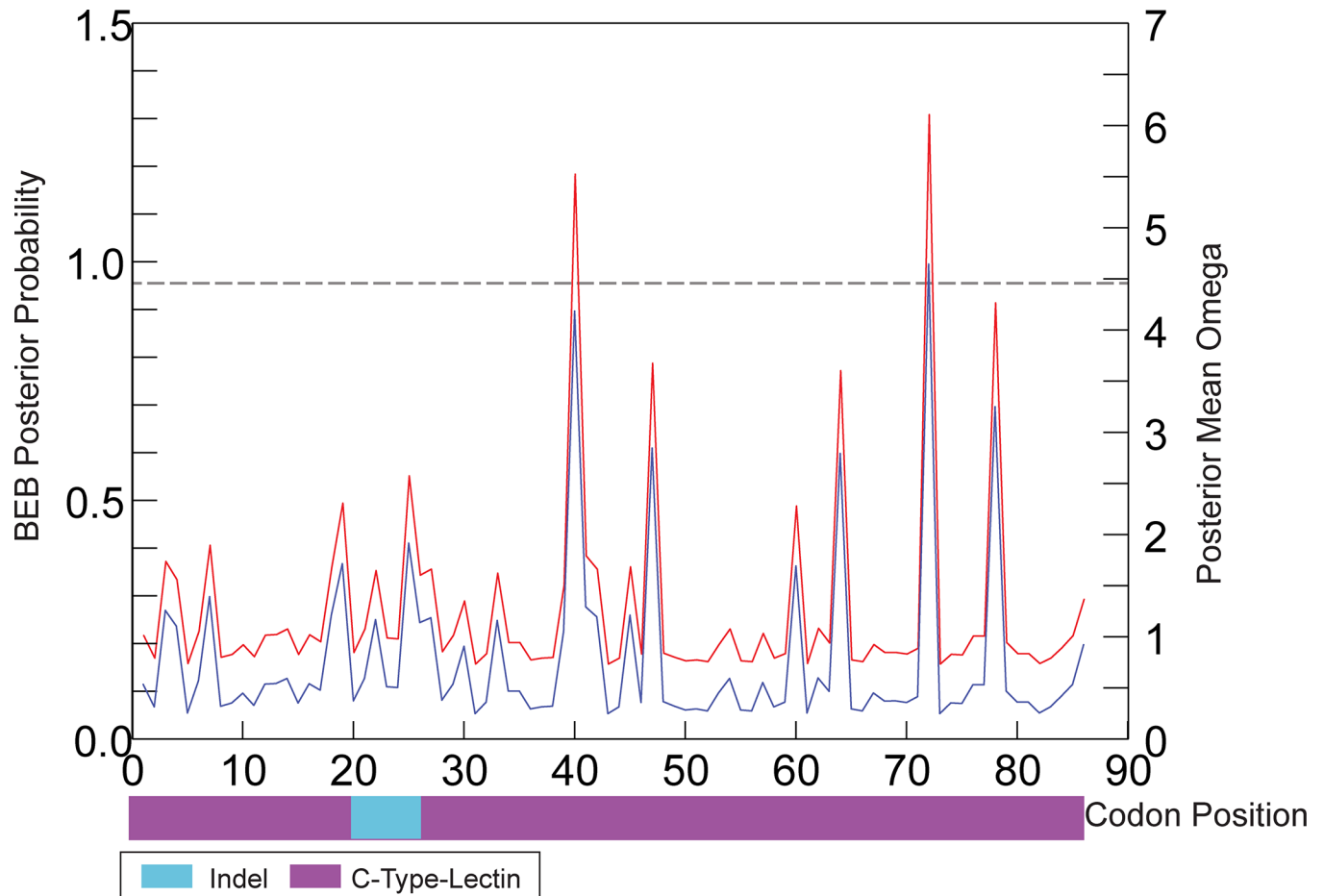


Fig 5. Positive selection within *Acp3*. Red line represents the mean posterior omega, and the blue line represents the probability of each codon being under positive selection. The values were calculated using a Bayes Empirical Bayes analysis under the M8 site-specific model in Paml. Codon position based on full-length alignment. The annotations identified using InterProScan and the position of the 24 bp indel are shown in relation to codon position.

<https://doi.org/10.1371/journal.pone.0188147.g005>

presents a testable hypothesis into what selection may be occurring on weta reproductive proteins. A better understanding can be achieved by incorporating functional and population genomics with candidate gene approaches to reveal the relationship between the evolution of these genes and mate recognition and speciation. In addition, the transcriptome data generated represents a first step in the identification of reproductive associated proteins in weta. These transcriptomic sequences will provide a valuable resource for further research into the evolution of reproductive proteins and speciation of New Zealand weta.

Supporting information

S1 Table. Sample collection details. **Samples used for 454 sequencing.
(XLSX)

S2 Table. Primer sequences for candidate genes. Annotations from tblastx against non-redundant database.
(XLSX)

S3 Table. Orthologous contigs pairs identified in the bidirectional tblastx search.
(XLSX)

S4 Table. Contigs identified in the reproductive protein search.
(XLSX)

S5 Table. McDonald Kreitman test results for each gene. ^a Yates correction is applied to G tests.
(XLSX)

S6 Table. Likelihood values and parameter estimates for site-specific models for each gene. Parameters in brackets are not free, and therefore are not counted when considering difference in the number of parameters for nested model comparisons.
(XLSX)

S7 Table. Sites identified as putatively under selection and their posterior probabilities under the M3 and M8 models. ^a Amino acid site in the original alignment, including gaps; BEB, Bayes Empirical Bayes posterior probability; NEB, Naïve Empirical Bayes; **Probability > 99%.
(XLSX)

S1 Fig. Top-Hit species distribution for tblastx results for A) *Hemideina thoracica* and B) *Hemideina crassidens*.
(PDF)

S2 Fig. Gene ontology categories identified in the reproductive gene screen for A) *Hemideina thoracica* and B) *Hemideina crassidens*.
(PDF)

S1 File. *Deinacrida fallai* and *D. mahoenui* orthologues identified from an unpublished *de novo* assembled illumina transcriptome data.
(FASTA)

Acknowledgments

We would like to thank George Gibbs, Robert Hoare, George Twort, Owen Twort, Ian Twort, Genesta Twort, Dave Seldon, Peter Richie, Shelley Meyers, Chrissie Painting, Luke Dunning, Robin Howarth, Kathy Howarth, Paulina Giraldo Perez, Corrine Watts and Katy Hill for assistance with field work and providing samples, the New Zealand eScience Infrastructure (NeSI) high-performance computing facilities and the staff at the Centre for eResearch at the University of Auckland, for bioinformatic support and the Department of Conservation for access to field sites. We thank Joe Hull and three anonymous reviewers for comments that improved the manuscript.

Author Contributions

Conceptualization: Victoria G. Twort, Richard D. Newcomb, Thomas R. Buckley.

Formal analysis: Victoria G. Twort, Alice B. Dennis.

Investigation: Victoria G. Twort, Alice B. Dennis, Duckchul Park, Kathryn F. Lomas.

Supervision: Richard D. Newcomb, Thomas R. Buckley.

Writing – original draft: Victoria G. Twort.

Writing – review & editing: Victoria G. Twort, Alice B. Dennis, Duckchul Park, Kathryn F. Lomas, Richard D. Newcomb, Thomas R. Buckley.

References

1. Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proceedings of the National Academy of Sciences*. 2001; 98(13):7375–9. <https://doi.org/10.1073/pnas.131568198> PMID: 11404480
2. Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. *Nature Reviews Genetics*. 2002; 3(2):137–44. <https://doi.org/10.1038/nrg733> PMID: 11836507
3. Clark NL, Aagaard JE, Swanson WJ. Evolution of reproductive proteins from animals and plants. *Reproduction*. 2006; 131(1):11–22. <https://doi.org/10.1530/rep.1.00357> PMID: 16388004
4. Walters JR, Harrison RG. Combined EST and proteomic analysis identifies rapidly evolving seminal fluid proteins in *Heliconius* butterflies. *Molecular Biology and Evolution*. 2010; 27(9):2000–13. <https://doi.org/10.1093/molbev/msq092> PMID: 20375075
5. Walters JR, Harrison RG. Decoupling of rapid and adaptive evolution among seminal fluid proteins in *Heliconius* butterflies with divergent mating systems. *Evolution*. 2011; 65(10):2855–71. <https://doi.org/10.1111/j.1558-5646.2011.01351.x> PMID: 21967427
6. Wilburn DB, Swanson WJ. From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. *Journal of Proteomics*. 2016; 135:12–25. <https://doi.org/10.1016/j.jprot.2015.06.007> PMID: 26074353
7. Andrés JA, Maroja LS, Bogdanowicz SM, Swanson WJ, Harrison RG. Molecular evolution of seminal proteins in field crickets. *Molecular Biology and Evolution*. 2006; 23(8):1574–84. <https://doi.org/10.1093/molbev/msl020> PMID: 16731569
8. Braswell WE, Andrés JA, Maroja LS, Harrison RG, Howard DJ, Swanson WJ. Identification and comparative analysis of accessory gland proteins in Orthoptera. *Genome*. 2006; 49(9):1069–80. <https://doi.org/10.1139/g06-061> PMID: 17110987
9. Ahmed-Braimah YH, Unckless RL, Clark AG. Evolutionary Dynamics of Male Reproductive Genes in the *Drosophila virilis* Subgroup. *G3: Genes|Genomes|Genetics*. 2017. <https://doi.org/10.1534/g3.117.1136> PMID: 28739599
10. Wolfner MF. Tokens of love: functions and regulation of *Drosophila* male accessory gland products. *Insect Biochemistry and Molecular Biology*. 1997; 27(3):179–92. PMID: 9090115
11. Gillott C. Male accessory gland secretions: modulators of female reproductive physiology and behavior. *Annual Review of Entomology*. 2003; 48(1):163–84.
12. Neubaum DM, Wolfner MF. Mated *Drosophila melanogaster* Females Require a Seminal Fluid Protein, Acp36DE, to Store Sperm Efficiently. *Genetics*. 1999; 153(2):845–57. PMID: 10511562
13. Lange AB, Loughton BG. An oviposition-stimulating factor in the male accessory reproductive gland of the locust, *Locusta migratoria*. *General and Comparative Endocrinology*. 1985; 57(2):208–15. [https://doi.org/10.1016/0016-6480\(85\)90265-5](https://doi.org/10.1016/0016-6480(85)90265-5) PMID: 3979802
14. Chen PS, Stumm-Zollinger E, Aigaki T, Balmer J, Bienz M, Böhlen P. A male accessory gland peptide that regulates reproductive behavior of female *D. melanogaster*. *Cell*. 1988; 54(3):291–8. [https://doi.org/10.1016/0092-8674\(88\)90192-4](https://doi.org/10.1016/0092-8674(88)90192-4) PMID: 3135120
15. Marshall JL, Huestis DL, Hiromasa Y, Wheeler S, Oppert C, Marshall SA, et al. Identification, RNAi Knockdown, and Functional Analysis of an Ejaculate Protein that Mediates a Postmating, Prezygotic Phenotype in a Cricket. *PLoS ONE*. 2009; 4(10):e7537. <https://doi.org/10.1371/journal.pone.0007537> PMID: 19851502
16. Bayram H, Sayadi A, Goenaga J, Immonen E, Arnqvist G. Novel seminal fluid proteins in the seed beetle *Callosobruchus maculatus* identified by a proteomic and transcriptomic approach. *Insect Molecular Biology*. 2017; 26(1):58–73. <https://doi.org/10.1111/imb.12271> PMID: 27779332
17. Al-Wathiqui N, Dopman EB, Lewis SM. Postmating transcriptional changes in the female reproductive tract of the European corn borer moth. *Insect Molecular Biology*. 2016; 25(5):629–45. <https://doi.org/10.1111/imb.12249> PMID: 27329655
18. Himuro C, Ikegawa Y, Honma A. Males Use Accessory Gland Substances to Inhibit Remating by Females in West Indian Sweetpotato Weevil (Coleoptera: Curculionidae). *Annals of the Entomological Society of America*. 2017; 110(4):374–80. <https://doi.org/10.1093/aesa/sax036>
19. Andrés JA, Arnqvist G. Genetic divergence of the seminal signal—receptor system in houseflies: the footprints of sexually antagonistic coevolution? *Proceedings of the Royal Society of London Series B: Biological Sciences*. 2001; 268(1465):399–405. <https://doi.org/10.1098/rspb.2000.1392> PMID: 11270437

20. Rice WR. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature*. 1996; 381(6579):232–4. <https://doi.org/10.1038/381232a0> PMID: 8622764
21. Turner LM, Hoekstra HE. Causes and consequences of the evolution of reproductive proteins. *International Journal of Developmental Biology*. 2008; 52(5):769.
22. Marshall JL, Huestis DL, Garcia C, Hiromasa Y, Wheeler S, Noh S, et al. Comparative Proteomics Uncovers the Signature of Natural Selection Acting on the Ejaculate Proteomes of Two Cricket Species Isolated by Postmating, Prezygotic Phenotypes. *Molecular Biology and Evolution*. 2011; 28(1):423–35. <https://doi.org/10.1093/molbev/msq230> PMID: 20805188
23. Kern AD, Jones CD, Begun DJ. Molecular Population Genetics of Male Accessory Gland Proteins in the *Drosophila simulans* Complex. *Genetics*. 2004; 167(2):725–35. <https://doi.org/10.1534/genetics.103.020883> PMID: 15238524
24. Mueller JL, Ram KR, McGraw LA, Bloch Qazi MC, Siggia ED, Clark AG, et al. Cross-Species Comparison of *Drosophila* Male Accessory Gland Protein Genes. *Genetics*. 2005; 171(1):131–43. <https://doi.org/10.1534/genetics.105.043844> PMID: 15944345
25. Andrés JA, Maroja LS, Harrison RG. Searching for candidate speciation genes using a proteomic approach: seminal proteins in field crickets. *Proceedings of the Royal Society B: Biological Sciences*. 2008; 275(1646):1975–83. <https://doi.org/10.1098/rspb.2008.0423> PMID: 18495616
26. Almeida FC, DeSalle R. Genetic differentiation and adaptive evolution at reproductive loci in incipient *Drosophila* species. *Journal of Evolutionary Biology*. 2017; 30(3):524–37. <https://doi.org/10.1111/jeb.13021> PMID: 27883252
27. Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ravi Ram K, Sirot LK, et al. Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in *Drosophila*. *Genetics*. 2007; 177(3):1321–35. <https://doi.org/10.1534/genetics.107.078865> PMID: 18039869
28. LaFlamme BA, Ravi Ram K, Wolfner MF. The *Drosophila melanogaster* Seminal Fluid Protease “Seminase” Regulates Proteolytic and Post-Mating Reproductive Processes. *PLoS Genet*. 2012; 8(1):e1002435. <https://doi.org/10.1371/journal.pgen.1002435> PMID: 22253601
29. Dean MD, Clark NL, Findlay GD, Karn RC, Yi X, Swanson WJ, et al. Proteomics and Comparative Genomic Investigations Reveal Heterogeneity in Evolutionary Rate of Male Reproductive Proteins in Mice (*Mus domesticus*). *Molecular Biology and Evolution*. 2009; 26(8):1733–43. <https://doi.org/10.1093/molbev/msp094> PMID: 19420050
30. Findlay GD, Yi X, MacCoss MJ, Swanson WJ. Proteomics Reveals Novel *Drosophila* Seminal Fluid Proteins Transferred at Mating. *PLoS Biol*. 2008; 6(7):e178. <https://doi.org/10.1371/journal.pbio.0060178> PMID: 18666829
31. Tsauro SC, Ting CT, Wu CI. Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of *Drosophila*: II. Divergence versus polymorphism. *Molecular Biology and Evolution*. 1998; 15(8):1040–6. PMID: 9718731
32. Tsauro SC, Wu CI. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of *Drosophila*. *Molecular Biology and Evolution*. 1997; 14(5):544–9. PMID: 9159932
33. Wong A, Albright SN, Wolfner MF. Evidence for structural constraint on ovulin, a rapidly evolving *Drosophila melanogaster* seminal protein. *Proceedings of the National Academy of Sciences*. 2006; 103(49):18644–9. <https://doi.org/10.1073/pnas.0601849103> PMID: 17130459
34. Fuyama Y. Species-specificity of paragonial substances as an isolating mechanism in *Drosophila*. *Cellular and Molecular Life Sciences*. 1983; 39(2):190–2. <https://doi.org/10.1007/bf01958894>
35. Andrés JA, Larson EL, Bogdanowicz SM, Harrison RG. Patterns of Transcriptome Divergence in the Male Accessory Gland of Two Closely Related Species of Field Crickets. *Genetics*. 2012. <https://doi.org/10.1534/genetics.112.142299> PMID: 23172857
36. Griffin MJ, Treweek SA, Wehi PM, Morgan-Richards M. Exploring the concept of niche convergence in a land without rodents: The case of weta as small mammals. *New Zealand Journal of Ecology*. 2011; 35(3).
37. Gibbs G. *New Zealand Weta*: Kyodo Printing Co Ltd, Singapore; 1998.
38. McGuinness CA. The conservation requirements of New Zealand’s nationally threatened invertebrates. In: Department of Conservation, editor. Department of Conservation, Wellington, New Zealand. 2001.
39. Treweek SA, Morgan-Richards M. After the deluge: Mitochondrial DNA indicates Miocene radiation and Pliocene adaptation of tree and giant weta (Orthoptera: Anostostomatidae). *Journal of Biogeography*. 2005; 32(2):295–309. <https://doi.org/10.1111/j.1365-2699.2004.01179.x>
40. White DJ, Watts C, Allwood J, Prada D, Stringer I, Thornburrow D, et al. Population history and genetic bottlenecks in translocated Cook Strait giant weta, *Deinacrida rugosa*: recommendations for future

- conservation management. *Conservation Genetics*. 2016;1–12. <https://doi.org/10.1007/s10592-016-0916-4>
41. Treweek SA, Morgan-Richards M. Phylogenetics of New Zealand's tree, giant and tusked weta (Orthoptera: Anostostomatidae): evidence from mitochondrial DNA. *Journal of Orthoptera Research*. 2004; 13(2):185–96. [https://doi.org/10.1665/1082-6467\(2004\)013\[0185:ponztg\]2.0.co;2](https://doi.org/10.1665/1082-6467(2004)013[0185:ponztg]2.0.co;2)
 42. Treweek SA, Morgan-Richards M. On the distribution of tree weta in the North Island, New Zealand. *Journal—Royal Society of New Zealand*. 1995; 25(4):485–93.
 43. Morgan-Richards M. Robertsonian translocations and B chromosomes in the Wellington tree weta, *Hemideina crassidens* (Orthoptera: Anostostomatidae). *Hereditas*. 2000; 132(1):49–54. PMID: [10857259](https://pubmed.ncbi.nlm.nih.gov/10857259/)
 44. Morgan-Richards M, King T, Treweek S. The evolutionary history of tree weta; A genetic approach. In: Field LH, editor. *The biology of wetas, king crickets and their allies*: Wallingford, Oxon., UK; New York, N.Y., USA: CABI Pub., c2001; 2001.
 45. Morgan-Richards M. Intraspecific karyotype variation is not concordant with allozyme variation in the Auckland tree weta of New Zealand, *Hemideina thoracica* (Orthoptera: Stenopelmatidae). *Biological Journal of the Linnean Society*. 1997; 60(4):423–42.
 46. Morgan-Richards M, Treweek SA, Wallis GP. Chromosome races with Pliocene origins: Evidence from mtDNA. *Heredity*. 2001; 86(3):303–12.
 47. Morgan-Richards M, Wallis GP. A comparison of five hybrid zones of the weta *Hemideina thoracica* (Orthoptera: Anostostomatidae): Degree of cytogenetic differentiation fails to predict zone width. *Evolution*. 2003; 57(4):849–61. PMID: [12778554](https://pubmed.ncbi.nlm.nih.gov/12778554/)
 48. McKean NE, Treweek SA, Morgan-Richards M. Comparative cytogenetics of North Island tree wētā in sympatry. *New Zealand Journal of Zoology*. 2015:1–12. <https://doi.org/10.1080/03014223.2015.1032984>
 49. Morgan-Richards M. A new species of tree weta from the North Island of New Zealand (Hemideina Stenopelmatidae: Orthoptera). *New Zealand Entomologist*. 1995; 18(1):15–23. <https://doi.org/10.1080/00779962.1995.9721996>
 50. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, et al. Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFix. *BMC Genomics*. 2009;10.
 51. Drummond A, Ashton B, Buxton S, Cheung M, Heled J, Kearse M, et al. Geneious. 5.4.6 ed: Available from <http://www.geneious.com>; 2011.
 52. Dlugosch KM, Rieseberg LH. SnoWhite: A pipeline for aggressive cleaning of next-generation sequence reads. In prep.
 53. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet-Journal*. 2011; 17(1):10–2. <http://dx.doi.org/10.14806/ej.17.1.200>.
 54. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011. <https://doi.org/10.1093/bioinformatics/btr026> PMID: [21278185](https://pubmed.ncbi.nlm.nih.gov/21278185/)
 55. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006; 22(13):1658–9. <https://doi.org/10.1093/bioinformatics/btl158> PMID: [16731699](https://pubmed.ncbi.nlm.nih.gov/16731699/)
 56. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 1997; 25(17):3389–402. <https://doi.org/10.1093/nar/25.17.3389> PMID: [9254694](https://pubmed.ncbi.nlm.nih.gov/9254694/)
 57. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. *Nucleic Acids Research*. 2005; 33(suppl_2):W116–W20. <https://doi.org/10.1093/nar/gki442> PMID: [15980438](https://pubmed.ncbi.nlm.nih.gov/15980438/)
 58. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005; 21(18):3674–6. <https://doi.org/10.1093/bioinformatics/bti610> PMID: [16081474](https://pubmed.ncbi.nlm.nih.gov/16081474/)
 59. Lara A, Pérez-Trabado G, Villalobos D, Díaz-Moreno S, Cantón F, Claros M. A Web Tool to Discover Full-Length Sequences—Full-Lengther Innovations in Hybrid Intelligent Systems. In: Corchado E, Corchado J, Abraham A, editors. *Advances in Soft Computing*. 44: Springer Berlin / Heidelberg; 2007. p. 361–8.
 60. Altenhoff AM, Dessimoz C. Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLoS Comput Biol*. 2009; 5(1):e1000262. <https://doi.org/10.1371/journal.pcbi.1000262> PMID: [19148271](https://pubmed.ncbi.nlm.nih.gov/19148271/)
 61. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011; 12(1):323. <https://doi.org/10.1186/1471-2105-12-323> PMID: [21816040](https://pubmed.ncbi.nlm.nih.gov/21816040/)

62. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Meth.* 2012; 9(4):357–9. <https://doi.org/10.1038/nmeth.1923> PMID: 22388286
63. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Meth.* 2011; 8(10):785–6. <https://doi.org/10.1038/nmeth.1701> PMID: 21959131
64. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 2001; 305(3):567–80. Epub 2001/01/12. <https://doi.org/10.1006/jmbi.2000.4315> PMID: 11152613.
65. Rozen S, Skaletsky H. Primer3 on the WWW for General Users and for Biologist Programmers. 1321999. p. 365–86.
66. Stephens M, Donnelly P. A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. *American journal of human genetics.* 2003; 73(5):1162–9. <https://doi.org/10.1086/379378> PMID: 14574645
67. Stephens M, Smith NJ, Donnelly P. A New Statistical Method for Haplotype Reconstruction from Population Data. *American journal of human genetics.* 2001; 68(4):978–89. <https://doi.org/10.1086/319501> PMID: 11254454
68. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* 2009; 25(11):1451–2. <https://doi.org/10.1093/bioinformatics/btp187> PMID: 19346325
69. Tajima F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics.* 1989; 123(3):585–95. PMID: 2513255
70. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature.* 1991; 351(6328):652–4. <https://doi.org/10.1038/351652a0> PMID: 1904993
71. Hurvich CM, Tsai C-L. Regression and time series model selection in small samples. *Biometrika.* 1989; 76(2):297–307.
72. Posada D. jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution.* 2008; 25(7):1253–6. <https://doi.org/10.1093/molbev/msn083> PMID: 18397919
73. Guindon S, Gascuel O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic biology.* 2003; 52(5):696–704. PMID: 14530136
74. Zwickl D, Hillis D. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion: University of Texas at Austin; 2006.
75. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution.* 2007; 24(8):1586–91. <https://doi.org/10.1093/molbev/msm088> PMID: 17483113
76. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution.* 2000; 17(1):32–43. PMID: 10666704
77. Yang Z, Nielsen R, Goldman N, Pedersen A-MK. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics.* 2000; 155(1):431–49. PMID: 10790415
78. Swanson WJ, Nielsen R, Yang Q. Pervasive Adaptive Evolution in Mammalian Fertilization Proteins. *Molecular Biology and Evolution.* 2003; 20(1):18–20. PMID: 12519901
79. Ravi Ram K, Ji S, Wolfner MF. Fates and targets of male accessory gland proteins in mated female *Drosophila melanogaster*. *Insect Biochemistry and Molecular Biology.* 2005; 35(9):1059–71. <https://doi.org/10.1016/j.ibmb.2005.05.001> PMID: 15979005
80. Ravi Ram K, Wolfner MF. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. *Integrative and Comparative Biology.* 2007; 47(3):427–45. <https://doi.org/10.1093/icb/icm046> PMID: 21672851
81. Sirot LK, Hardstone MC, Helinski MEH, Ribeiro JMC, Kimura M, Deewatthanawong P, et al. Towards a Semen Proteome of the Dengue Vector Mosquito: Protein Identification and Potential Functions. *PLOS Neglected Tropical Diseases.* 2011; 5(3):e989. <https://doi.org/10.1371/journal.pntd.0000989> PMID: 21423647
82. Schully S, Hellberg M. Positive Selection on Nucleotide Substitutions and Indels in Accessory Gland Proteins of the *Drosophila pseudoobscura* Subgroup. *Journal of Molecular Evolution.* 2006; 62(6):793–802. <https://doi.org/10.1007/s00239-005-0239-4> PMID: 16752217
83. Begun DJ, Lindfors HA, Thompson ME, Holloway AK. Recently evolved genes identified from *Drosophila yakuba* and *Drosophila erecta* accessory gland expressed sequence tags. *GENETICS.* 2006.
84. Wu C, Crowhurst RN, Dennis AB, Twort VG, Liu S, Newcomb RD, et al. De Novo Transcriptome Analysis of the Common New Zealand Stick Insect *Clitarchus hookeri* (Phasmatodea) Reveals Genes Involved in Olfaction, Digestion and Sexual Reproduction. *PLOS ONE.* 2016; 11(6):e0157783. <https://doi.org/10.1371/journal.pone.0157783> PMID: 27336743

85. Rogers DW, Baldini F, Battaglia F, Panico M, Dell A, Morris HR, et al. Transglutaminase-Mediated Semen Coagulation Controls Sperm Storage in the Malaria Mosquito. *PLoS Biol.* 2009; 7(12): e1000272. <https://doi.org/10.1371/journal.pbio.1000272> PMID: 20027206
86. Sirot LK, Poulson RL, Caitlin McKenna M, Girnary H, Wolfner MF, Harrington LC. Identity and transfer of male reproductive gland proteins of the dengue vector mosquito, *Aedes aegypti*: Potential tools for control of female feeding and reproduction. *Insect Biochemistry and Molecular Biology.* 2008; 38(2):176–89. <https://doi.org/10.1016/j.ibmb.2007.10.007> PMID: 18207079
87. Azevedo RVDM, Dias DBS, Bretãs JAC, Mazzoni CJ, Souza NA, Albano RM, et al. The Transcriptome of *Lutzomyia longipalpis* (Diptera: Psychodidae) Male Reproductive Organs. *PLoS ONE.* 2012; 7(4): e34495. <https://doi.org/10.1371/journal.pone.0034495> PMID: 22496818
88. Romero A, Romao MJ, Varela PF, Kolln I, Dias JM, Carvalho AL, et al. The crystal structures of two spermadhesins reveal the CUB domain fold. *Nat Struct Biol.* 1997; 4(10):783–8. <https://doi.org/10.1038/nsb1097-783> PMID: 9334740
89. Mueller JL, Ripoll DR, Aquadro CF, Wolfner MF. Comparative structural modeling and inference of conserved protein classes in *Drosophila* seminal fluid. *Proceedings of the National Academy of Sciences of the United States of America.* 2004; 101(37):13542–7. <https://doi.org/10.1073/pnas.0405579101> PMID: 15345744
90. Takemori N, Yamamoto M-T. Proteome mapping of the *Drosophila melanogaster* male reproductive system. *PROTEOMICS.* 2009; 9(9):2484–93. <https://doi.org/10.1002/pmic.200800795> PMID: 19343724
91. Morgan-Richards M, Gibbs G. A phylogenetic analysis of New Zealand giant and tree weta (Orthoptera: Anostomatidae: Deinacrida and Hemideina) using morphological and genetic characters. *Invertebrate Taxonomy.* 2001; 15(1):1–12.
92. Schierup MH, Hein J. Consequences of Recombination on Traditional Phylogenetic Analysis. *Genetics.* 2000; 156(2):879–91. PMID: 11014833
93. Posada D, Crandall KA. Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology & Evolution.* 2001; 16(1):37–45. [https://doi.org/10.1016/s0169-5347\(00\)02026-7](https://doi.org/10.1016/s0169-5347(00)02026-7)
94. Bulgarella M, Treweek SA, Godfrey AJR, Sinclair BJ, Morgan-Richards M. Elevational variation in adult body size and growth rate but not in metabolic rate in the tree weta *Hemideina crassidens*. *Journal of Insect Physiology.* 2015; 75:30–8. <https://doi.org/10.1016/j.jinsphys.2015.02.012> PMID: 25753546
95. Minards NA. Physiological Ecology of Two Tree Weta Species [Master's dissertation]. Palmerston North—New Zealand: Massey University; 2011.
96. Nielsen R, Yang Z. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics.* 1998; 148(3):929–36. PMID: 9539414
97. Clark NL, Swanson WJ. Pervasive Adaptive Evolution in Primate Seminal Proteins. *PLoS Genet.* 2005; 1(3):e35. <https://doi.org/10.1371/journal.pgen.0010035> PMID: 16170411
98. Swanson WJ, Wong A, Wolfner MF, Aquadro CF. Evolutionary Expressed Sequence Tag Analysis of *Drosophila* Female Reproductive Tracts Identifies Genes Subjected to Positive Selection. *Genetics.* 2004; 168(3):1457–65. <https://doi.org/10.1534/genetics.104.030478> PMID: 15579698
99. Yang Z, Nielsen R. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Molecular Biology and Evolution.* 2002; 19(6):908–17. PMID: 12032247
100. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences: CABIOS.* 1997; 13(5):555–6. <https://doi.org/10.1093/bioinformatics/13.5.555>. PMID: 9367129
101. Yang Z, Wong WSW, Nielsen R. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Molecular Biology and Evolution.* 2005; 22(4):1107–18. <https://doi.org/10.1093/molbev/msi097> PMID: 15689528
102. Suzuki Y, Nei M. Simulation Study of the Reliability and Robustness of the Statistical Methods for Detecting Positive Selection at Single Amino Acid Sites. *Molecular Biology and Evolution.* 2002; 19(11):1865–9. PMID: 12411595
103. Suzuki Y, Nei M. Reliabilities of Parsimony-based and Likelihood-based Methods for Detecting Positive Selection at Single Amino Acid Sites. *Molecular Biology and Evolution.* 2001; 18(12):2179–85. PMID: 11719567
104. Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution.* 1999; 16(10):1315–28. PMID: 10563013
105. Wong W, Yang Z, Goldman N, Nielsen R. Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. *Genetics.* 2004; 168(2):1041–51. <https://doi.org/10.1534/genetics.104.031153> PMID: 15514074