Wu *et al. BMC Genomics* (2017) 18:884
DOI 10.1186/s12864-017-4245-x

BMC Genomics

RESEARCH ARTICLE Open Access

Assembling large genomes: analysis of the stick insect (*Clitarchus hookeri*) genome reveals a high repeat content and sex-biased genes associated with reproduction

Chen Wu[1,2,3*], Victoria G. Twort[1,2,4], Ross N. Crowhurst[3], Richard D. Newcomb[1,3] and Thomas R. Buckley[1,2]

Abstract

Background: Stick insects (Phasmatodea) have a high incidence of parthenogenesis and other alternative reproductive strategies, yet the genetic basis of reproduction is poorly understood. Phasmatodea includes nearly 3000 species, yet only the genome of *Timema cristinae* has been published to date. *Clitarchus hookeri* is a geographical parthenogenetic stick insect distributed across New Zealand. Sexual reproduction dominates in northern habitats but is replaced by parthenogenesis in the south. Here, we present a de novo genome assembly of a female *C. hookeri* and use it to detect candidate genes associated with gamete production and development in females and males. We also explore the factors underlying large genome size in stick insects.

Results: The *C. hookeri* genome assembly was 4.2 Gb, similar to the flow cytometry estimate, making it the second largest insect genome sequenced and assembled to date. Like the large genome of *Locusta migratoria*, the genome of *C. hookeri* is also highly repetitive and the predicted gene models are much longer than those from most other sequenced insect genomes, largely due to longer introns. Miniature inverted repeat transposable elements (MITEs), absent in the much smaller *T. cristinae* genome, is the most abundant repeat type in the *C. hookeri* genome assembly. Mapping RNA-Seq reads from female and male gonadal transcriptomes onto the genome assembly resulted in the identification of 39,940 gene loci, 15.8% and 37.6% of which showed female-biased and male-biased expression, respectively. The genes that were over-expressed in females were mostly associated with molecular transportation, developmental process, oocyte growth and reproductive process; whereas, the male-biased genes were enriched in rhythmic process, molecular transducer activity and synapse. Several genes involved in the juvenile hormone synthesis pathway were also identified.

Conclusions: The evolution of large insect genomes such as *L. migratoria* and *C. hookeri* genomes is most likely due to the accumulation of repetitive regions and intron elongation. MITEs contributed significantly to the growth of *C. hookeri* genome size yet are surprisingly absent from the *T. cristinae* genome. Sex-biased genes identified from gonadal tissues, including genes involved in juvenile hormone synthesis, provide interesting candidates for the further study of flexible reproduction in stick insects.

Keywords: Phasmatodea, Genome assembly, RNA-Seq, Reproduction, Parthenogenesis, *Clitarchus hookeri*

* Correspondence: chen.wu@plantandfood.co.nz
[1]School of Biological Sciences, The University of Auckland, Auckland, New Zealand
[2]Landcare Research, Auckland, New Zealand
Full list of author information is available at the end of the article

Background

The insect order Phasmatodea, commonly known as stick insects or walking sticks, contains approximately 3000 species distributed worldwide [1]. At least 10% of stick insect species can reproduce parthenogenetically and for this reason have attracted much attention [1]. *Clitarchus hookeri* is one of the most common New Zealand stick insect species and is distributed across a wide range of habitats on both the North and South Islands with a lower population density at higher latitudes and altitudes [2]. Its reproductive biology is interesting as it displays extreme sexual dimorphism and geographical parthenogenesis [2, 3]. On the upper half of the North Island sexual reproduction dominates, which produces offspring with relatively equal numbers of both sexes, whereas, obligate parthenogenesis is widespread on the lower North Island and South Island, forming all-female populations [2, 3]. In addition, *C. hookeri* is also thought to have hybridized with the obligate parthenogenetic genus *Acanthoxyla* [4–7]. All these features make *C. hookeri* an ideal species for the study of geographical parthenogenesis, hybridisation, and mating behaviour [2–5, 7–13]. However, at the molecular level, little is known about these processes, impeding further understanding of their reproductive biology. Currently, except for some genes encoding male accessory gland proteins [12] from *C. hookeri*, the genes involved in other female and male reproductive traits, such as oogenesis, spermatogenesis, egg and sperm maturation are largely unknown. It is critical that these genes are characterised to advance our understanding of the evolution of parthenogenesis and other alternative reproductive processes within the Phasmatodea.

RNA-Seq based transcriptome profiling has been successfully employed to identify genes expressed in the germline tissues from both model and non-model insect species, such as *Aedes aegypti* (mosquito) [14], *Nasonia vitripennis* (wasp) [15] and *Periplaneta americana* (cockroach) [16]. By sampling gonadal tissues from males and females, a digital gene expression profile can be obtained from the RNA-Seq data for the identification of the genes predominantly expressed in one sex (sex-biased genes). These genes frequently encode proteins essential to the sex-linked characteristics that are critical to the study of reproduction and its evolution [17]. Coupling genome and transcriptome sequencing allows the use of a reference-guided approach to studying patterns of gene expression. This is thought to be more accurate than the de novo method, because it can reduce transcript redundancy, reveal strand orientation to tease apart overlapping transcripts and enhance the representation of lowly expressed genes [18].

Another interesting feature of the stick insects is their large genome size. Currently, only the genome of *Timema cristinae*, sister lineage to all other stick insects (Euphasmatodea), has been sequenced and assembled [19]. However, compared with the 1.5 Gb genome of *T. cristinae*, even larger genomes occur in Euphasmatodea. Multiple species from the European stick insect genus *Bacillus* are ~2 Gb [20], while the South American species *Anisomorpha buprestoides*, Lord Howe Island species *Dryococelus australis* and the Australian species *Extatosoma tiaratum* have even larger genomes, up to ~3 [21], ~4 [22] and ~8 Gb (reported from the Animal genome size database: http://www.genomesize.com/), respectively. Without genome sequences from euphasmids, we are unable to determine the main causes for the formation of these large genomes. High-throughput sequencing has been used to sequence the whole genomes of some model and non-model species, especially those from the more derived Holometabola with small or moderate genome sizes [23–27]. Recently, the *Locusta migratoria* (Orthoptera) genome sequence comprising 6.5 Gb has become available. The analysis of this genome revealed a large numbers of repetitive elements (~58.9% of the genome assembly) and gene copy expansion within some gene families (e.g. detoxification) [28]. This resource has enhanced our understanding of the causes and consequences of large insect genomes, but whether similar patterns also occur in other insect lineages with large genomes is unknown.

Here we have sequenced, assembled and analysed the genome of *C. hookeri*, and through comparison with other insects, provide insights into the evolution of stick insect genomes. We particularly focused on repetitive elements and predicted gene models, both potential factors underlying large euphasmid genome sizes. We also performed an analysis of the RNA-Seq data produced from sequencing the female reproductive tract and the male testis to identify genes that are essential to gamete production and maturation in *C. hookeri*. The candidate genes predominantly expressed in the gonad of one sex and the genes involved in the juvenile hormone pathway provide candidates for the further study of stick insect reproductive flexibility.

Methods

Sample collection and preparation

All samples used for sequencing in this study were collected from Totara Park, Auckland, New Zealand (37°0.111 S, 174° 55.039 E). A female *C. hookeri* (CLI739) was collected in 2013 and leg tissues were used for the estimation of genome size. Three female insects (CLI525, 600 and 654) were collected in 2012 for genome sequencing and leg tissues were used. Three males (CLI755, CLI757 and CLI760) and three females (CLI765, CLI767 and CLI768) were collected in 2014 for gonadal transcriptome sequencing. Females were nymphs when collected and separately reared from males until they laid the first egg (reached maturation). Live insects were snap frozen in liquid nitrogen on capture and stored at −80 °C after collection. Female reproductive tract (approximately 18 to 20 ovarioles, early developing eggs and oviducts), and male testicle pairs were dissected in ethanol (100%) for RNA extraction.

Genome size estimation

The genome size of *C. hookeri* was estimated using flow cytometry following the Otto two-step method with the substitution of propidium iodine for DAPI [29]. A female *L. migratoria* was used as an internal standard. Approximately 15 mm^2 of leg tissue with a standard was co-chopped in a few drops of ice cold Otto buffer 1 with a stainless steel razor blade and then incubated for approximately 2 min. The sample was then filtered through a 20 μm Celltrics filter (PARTEC GmbH) before adding 2.5 ml of Otto buffer 2 with 1 mg/ml propidium iodine. The sample was run on a PARTEC CyFlow Space with a 488 nm laser as the excitation source. The 2C content of the *L. migratoria* standard was determined to be 18.31 pg, using *Pisum sativum* Citrad (9.09 pg 2C content; 4445 Mbp) [30] internal standard, and the gain adjusted as required. The total amount of DNA in the sample was determined as the ratio of the average channel number of the sample 2 N to the average channel number of the standard 2 N times the 1C amount of DNA in the standard.

DNA extraction, library preparation and sequencing

Genomic DNA was extracted from *C. hookeri* leg tissue using the DNeasy Plant Mini kit (Qiagen) with the following modifications: frozen leg tissue was chopped and incubated for 1 h with digestion buffer before DNA elution columns were incubated with digestion buffer for 1 h, and after the addition of elution buffer columns were incubated at room temperature for 30 min prior to centrifugation. DNA was quantified by a Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific) and quality checked by running a sample on 0.5% *w*/*v* agarose gel stained with 1X GelRed (Huntingtree Ltd.). The resulting nuclear genomic DNA was sent to New Zealand Genomics Limited (NZGL: http://www.nzgenomics.co.nz/) Otago, Dunedin, for library construction and sequencing. Seven paired-end (PE) sequencing libraries with average insert sizes of 200 bp (2 libraries), 350 bp, 500 bp and 720 bp (3 libraries) (Table 1) were prepared using the TruSeq® DNA LT Sample Prep Kit v2 (catalogue ID: FC-121-2001) and sequenced with 11 lanes on an Illumina HiSeq2000™. The three mate-pair (MP) libraries with insert sizes of 5 k bp and 8 k bp (2 libraries) (Table 1) were constructed using the Nextera® Mate Pair Sample Prep Kit (catalogue ID: FC-132-1001) and sequenced with eight lanes on the same platform. The number of libraries per insert size and the number of lanes sequenced per library are described in Table 1. All reads were 101 bases in length. For RNA-Seq, total RNA extraction and library preparation were performed as described in [12]. The extractions were barcoded and then pooled together for sequencing on the HiSeq2000 platform for one lane to generate 100 bp PE reads at NZGL.

Table 1 Illumina sequencing output for *Clitarchus hookeri* whole genome assembly

Insert size (bp)	Sample × library × lane	Sequencing output (Gb)	Estimated Genome coverage (×)
200	CLI525 × 1 × 5, CLI600 × 1 × 1	261.2	59.4
350	CLI525 × 1 × 2	88.5	20.1
500	CLI600 × 1 × 1	39.7	9.0
720	CLI525 × 1 × 1, CLI600 × 2 × 2	72.3	16.4
5000	CLI654 × 1 × 4	126.4	28.7
8000	CLI525 × 1 × 2, CLI654 × 1 × 2	145.2	33.0
	Total	733.3	166.6

Genome coverage estimates were obtained using the flow cytometry genome size

De novo genome assembly and quality assessment

The raw PE reads were preprocessed to remove duplicate sequence pairs, possible contaminants, and low-quality bases as follows: 1) reads containing ambiguities (Ns) and duplicates were filtered using PRINSEQ (v0.20.3) [31] and FastUniq (v1.1) [32], respectively; 2) reads with adapters and low quality ends (Phred < 30) were trimmed using Cutadapt (v1.3) [33]; 3) a read pair with overlapping ends more than 10 bp was merged into a single read using "abyss-mergepairs" from ABySS (v1.5.1) [34]; 4) reads shorter than 50 bp and orphan reads (single pair) were discarded; 5) remaining reads (83.3%) were error corrected using "ErrorCorrectReads.pl" from ALLPATH-LG (v46436) [35]. The raw MP reads were preprocessed following 1), 2), 4), and then trimmed to retain the first 36 bases of the 5 prime end in order to minimise inclusion of Nextera® adapters resulting from library preparation.

De bruijn graph and initial contigs were constructed and assembled using "pregraph" and "contig" commands from SOAPdenovo2 (vR223) [36] with Kmer 75 on the PE data derived from a single insect (CLI525; Table 1). The resulting contigs were used to construct scaffolds using "map" and "scaff" commands from SOAPdenovo2 based on their relationships implied by mapping all PE reads to the contigs from the shortest to the largest insect size libraries. Sequence gaps were then filled by GapCloser (v1.12-r6) [36] with PE reads from CLI525, and the resulting sequences were re-scaffolded again by non-CLI525 PE reads using SSPACE-basic (v2.0) [37] with default options (−z 0, −k 5, −a 0.7 and −n 15) and CLI525 PE reads with additional option "−X 1" for extending scaffolds. SSPACE-basic was also used to post-scaffold the resulting scaffolds using MP reads from the shortest to the largest insert size libraries with default options, followed by a step of gap filling using CLI525 reads, as described above, to produce the final genome assembly.

Approximately 1.0 billion genomic PE reads and 54.7 million RNA-Seq reads [12] were mapped to the genome assembly using Bowtie2 (v2.2.0) [38] with paired-end mode (−q −1 R1.fastq −2 R2.fastq) to produce a proportion of the coverage of the assembly from the short reads. The assembly was then quality evaluated with Core Eukaryotic Genes Mapping Approach (CEGMA; v2.4) [39] with large genome size mode (−mam) and Benchmarking Universal Single-Copy Orthologs (BUSCO; v2.0.1) [39] with searching database "arthropoda_odb9" to detect the presence of a core protein set of 248 highly conserved eukaryotic genes and 1066 highly conserved arthropoda genes. Scaffolds were also searched against GenBank *nt* database (Release 212) to estimate the percentage of archea, bacteria and virus sequences.

Repeat identification

RepeatModeler (v1.0.8) [40] with default options was used to predict and classify repetitive elements. It employs two de novo repeat finding programs, RECON and RepeatScout, to identify repeat element boundaries, followed by an assignment to the repeat classes based on the sequence feature.The resulting repeat models were searched against GenBank non-redundant (*nr*) protein database (evalue $<10^{-5}$) using Blastx (v2.2.28) [41] to exclude potential protein-coding genes. An additional repeat classification was conducted using PASTEClassifier (v1.0) [42] to assign miniature inverted repeat transposable elements (MITEs). The abundances of all predicted repeats were estimated in the genome assembly with RepeatMasker (v4.0.5) [43] (−no_is, −gff and −lib RepeatModels.gff).

Gene model annotation

Structural gene annotation was performed using MAKER2 (v2.31.3) [44] on scaffolds longer than 2000 bp. Before annotation, meta parameters of *C. hookeri* protein-coding genes, including those determining intron and exon length distributions, splice site patterns, and translation start codon patterns were generated using AUGUSTUS (v3.0.2) [45] (optimize_augustus.pl). Spliced alignments of protein sequences as inputs for AUGUSTUS were generated by aligning *C. hookeri* protein sequences including CEGMA predicted proteins and those that were identified from the head and prothorax transcriptome [9] to the assembly using Scipio (v1.4.1) [46] with default options. The workflow of MAKER2 involves: 1) producing *ab initio* gene predictions using trained *C. hookeri* meta parameters, 2) aligning de novo transcripts collected from [9], [12] and Arthropoda conserved protein sequences (OrthoDB: v7) [47] to the assembly followed by the identification of intron-exon boundaries and splice forms as evidence, 3) producing evidence-informed gene predictions, computing quality scores and selecting the gene models best supported by the evidence. The resulted *C. hookeri* gene models were searched for the presence of core proteins (BUSCO) and homology matches against *nr* (evalue $<10^{-5}$) for gene annotation. These gene models were also identified with *T. cristinae* orthologues from a reciprocal blast method using a custom python script. The *T. cristinae* gene models (v0.2) were downloaded from: http://nosil-lab.group.shef.ac.uk/?page_id=25.

Transcript construction and annotation

Raw reads were trimmed 5′ end (8 bases), adapter sequences, low quality 3′ ends and filtered reads containing ambiguous bases (Ns) using PRINSEQ (v0.20.3) [31] and CUTADAPT (v1.3) [33]. The program STAR (v2.5) [48] with options "−−outFilterType BySJout −−outFilterIntronMotifs RemoveNoncanonical −−outSAMstrandField intronMotif −−outSAMtype BAM SortedByCoordinate −−outReadsUnmapped Fastx" was used to align reads to the genome assembly (scaffolds longer than 10 kbp). We did not include the putative gene models generated from the in silico gene annotation above as the reference gene set for annotating the gonad-expressed genes; instead, the regions mapped by all the RNA-Seq reads were counted as the gene loci. Cufflinks (v2.2.1) [49] with default options was used to generate these gene loci on the genome assembly according to the alignments, followed by producing merged gene loci using an embedded command "cuffmerge". The gene loci present with intron-exon boundaries were stranded on the genomic scaffolds. The transcript set was constructed by extracting sequences from annotated gene loci using "gffread". For multiple isoforms detected from a single gene locus, the first (longest) isoform present was chosen as the representative transcript that was subjected to a search against the SwissProt (release-2015_12) [50], UniProt (release-2015_12) [51] and Flybase (*Drosophila melanogaster*: dmel_r6.08) [52] protein databases using BLASTx (v2.2.28, E-value cut-off: 10^{-5}, keeping the top hit). These isoforms were also searched against GenBank *nt* (release14) using Blastn (evalue $< 10^{-10}$) to screen for contamination from bacteria, fungi and virus.

Read quantification and differential expression analysis

The RNA-Seq reads were aligned to the annotated gene loci followed by quantification. The number of read pairs aligned to the stranded loci was calculated using htseq-count from HTSeq (v0.6.1) [53] with default options and the unstranded gene loci with an additional option "−s no". Differential expression comparison was performed in R (v3.1.1) [54] using the DESeq2 Bioconductor package [55]. This program takes read counts to estimate sample size factors, followed by estimating dispersions with expected mean values from the maximum likelihood estimate of log2 fold changes, and then fits a negative binomial distribution [55]. The principle component analysis from DESeq2 and an R package pheatmap were used to visualise global similarities and differences. Transcripts with an adjusted *p* value less than 0.05 and a minimum fold change (FC) of 2 were reported as differently expressed.

Gene ontology and pathways analysis

The predicted gene loci showing significantly sex-biased expression were used to detect enriched gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The matched *D. melanogaster* FlyBase gene IDs were imported into the Database for Annotation, Visualization, and Integrated Discovery (DAVID, v6.8) [56, 57] for functional annotation and enrichment tests. The default FlyBase database was used as the background. The level one GOs (GOTERM_1) and the GOs directly mapped (GOTERM_DIRECT) from the source database (without parental terms) were generated with *p* values below 0.05, and the significantly enriched ones were defined according to the *p* values adjusted by the Benjamini and Hochberg (BH) procedure (<0.05). They were visualised using the R package GOplot [58]. The enriched KEGG pathways (KEGG_PATHWAY) were generated with a *p* value less than 0.1.

Results

Genome sequence and analysis

The genome size of *C. hookeri* was estimated at approximately 4.4 Gb using flow cytometry. Illumina reads derived from ten libraries with various insert sizes shown in Table 1 were subject to de novo assembly. Sequencing of multiple libraries derived from DNA of a single female (CLI525) yielded 400.4 Gb of PE reads for contig construction with an estimated coverage of 90.9 × the estimated genome size and the initial contigs constructed from these data had a N50 of 3715 bp. The final assembly had a total size of 4.2Gb with a N50 of 255.7 kb (Table 2). Nearly all genomic reads (>99%) were mapped back to the assembly while the map back rate of RNA-Seq reads was ~86%. BUSCO and CEGMA analysis suggested the whole genome sequence was reasonably complete with 91.6% complete and 3.6% partial BUSCO proteins present and 76.6% complete and 22.3% partial CEGMA genes present. The "scaffold13492" comprising 19,561 bp was identified as the *C. hookeri* mitochondrial genome. The total number of estimated archea, bacteria and virus scaffolds is only 1387 (Additional file 1).

Table 2 *Clitarchus hookeri* genome assembly statistics

	Size (bp)	Number of scaffolds
N90	188	859,481
N80	5446	48,684
N70	30,351	11,639
N60	148,642	5894
N50	255,691	3749
Longest	4,944,527	–
Total (>100 bp)	4,244,875,252	4,114,148
Total (>2 kb)	3,503,002,174	78,458

Repetitive elements

The genome assembly of *C. hookeri* is highly repetitive, with approximately half the genome (51.6%) predicted as repeats. Among these, a total of 3210 repeat models were determined, 1404 (43.7%) of which were classified as different groups of interspersed transposon elements (TEs). To compare with *T. cristinae*, we used the same repeat identification method to identify repeats from the published genome assembly. In *T. cristinae*, 1288 repeat models were detected, including 433 (33.6%) assigned to the known repeat groups. The proportions of different repeat groups from the genomes of the two stick insects and the other two Polyneoptera species (*Zootermopsis nevadensis* and *L. migratoria*) [28, 59] are shown in Table 3.

In the *C. hookeri* genome, miniature inverted repeat transposable element (MITE) was the most abundant repeat type, which was identified with 87 putative models and 1,214,018 copies, covering 5.79% of the assembly. The short sequence "rnd-1_family-8" was detected as the most frequent MITE repeat in the genome. Similarly, MITEs were also reported to be highly abundant in the stick insect *Bacillus rossius*, *B. grandii* and *B. atticus* partial genomes [20]. However, they were absent from the *T. cristinae* genome. In comparison, Maverick was the most abundant repeat type in the *T. cristinae* genome, predicted with 12 repeats comprising 145,675 copies, which cover 3.13% of the genome, and the most frequent repeat copy was "rnd-2_family-6". Maverick was ranked as the third most abundant in the *C. hookeri* genome. The second most abundant DNA transposon in *C. hookeri* was TcMar-Tc1, the most abundant repeat type in the *L. migratoria* genome.

Class I TEs containing long terminal repeat retrotransposon (LTR), short interspersed element (SINE) and LINE were much less abundant than Class II TEs in the two stick insect genomes. Class I TEs constitute 5.74% and 8.08% of the *C. hookeri* and the *T. cristinae* genomes, respectively. LTR gypsy was the most abundant repeat

Table 3 Comparisons of repeats among four Polyneoptera genomes

	Zootermopsis nevadensis		*Timema cristinae*		*Clitarchus hookeri*		*Locusta migratoria*	
Genome size (Mb)	490		1030		4240		6500	
Repeat types	Length (Mb)	P%	Lengh (Mb)	P%	Lengh (Mb)	P%	Length (Mb)	P%
DNA	12.78	2.60	75.13	7.30	612.35	14.43	1480.54	22.69
LINE	22.13	4.50	34.34	3.34	55.84	1.32	1332.72	20.42
LTR	0.72	0.15	5.89	0.57	103.59	2.44	508.68	7.80
SINE	9.78	2.00	42.88	4.17	84.08	1.98	141.18	2.16
Simple repeat	1.39	0.28	9.00	0.87	96.82	2.28	13.03	0.20
Other	0.00	0.00	1.00	0.10	0.74	0.02	0.03	0.00
Unknown	81.32	16.50	243.05	23.61	1237.41	29.15	406.10	6.22
Total	128.41	26.00	411.29	39.96	2190.84	51.60	3840.81	58.86

type in the *C. hookeri* genome, comprising 194,787 copies, covering 1.44% of the genome. The short sequence "rnd-1_family-196" was detected with the highest copy number. In comparison, the LINE repeat RTE-BovB was the most abundant repeat type in the *T. cristinae* genome and the most frequent model was "rnd-3_family-716". This repeat type was predicted to occupy 2.33% of the *T. cristinae* genome, comprising 75,215 copies. It is also one of the most frequent repeat types in the *L. migratoria* genome. There were only a few repeat types that were uniquely present in one of the stick insect genomes when compared with the other. The repeats RTE-RTE (LINE), ERV1 (LTR) and TRIM (LTR) were detected only from the *C. hookeri* genome, whereas Ngaro (LTR) was only present in the *T. cristinae* genome. The SINE repeats appear to be slightly more abundant in the *C. hookeri* than the *T. cristinae* genome.

Protein-coding genes

The current predicted 66,470 *C. hookeri* gene models include 10,266 models revised from transcript and protein sequences after in silico prediction. This gene set includes 779 (73.1%) complete predicted BUSCO proteins with 747 (95.9%) single-copy genes, and 154 (14.4%) partial proteins. The proportion of gene models that have Blast matches with *nr* database proteins was 36.2% (24,085), the majority of which hit the eusocial termite *Z. nevadensis* (42.3%), followed by the red flour beetle *Tribolium castaneum* (5.7%) (Additional file 2). There were 8478 predicted gene orthologues between the two sets of stick insect gene models using a reciprocal blast approach. We compared the average values of a variety of gene model features across *C. hookeri*, *T. cristinae*, *Holacanthella duospinosa* (New Zealand giant collembolan) [60] and a wide range of arthropods with these values available [28, 59] (Fig. 1). This analysis demonstrates that insects from Polyneoptera frequently have genome sizes in the gigabase range and that increases in genome size are positively correlated with increasing transcript and intron sizes.

Gonadal transcriptome assembly and annotation

Approximately 95.8% of the 600 million raw RNA-Seq reads passed the cleaning criteria and were mapped to the genome assembly. The mapping ratios ranged from 89.8% to 94.9% across individuals and the ratios of uniquely mapped reads were all above 70% (Table 4). We then annotated only those gene loci with mapped RNA-seq reads, which can maximise the number of genes predicted to be expressed in the tissues compared with reusing the in silico produced gene models. In total, the mapped reads generated 39,940 putative gene loci, including 23,778 (59.5%) genes containing more than one exon, with strand orientation determined according to the intron-exon boundaries. Their transcribed sequences show an N50 length of 3828 bp, and a minimum and maximum length of 64 and 37,398 bp, respectively. A total of 36,072 (90.3%) transcripts have lengths longer than 500 bp. Only 45 sequences had the best homologous matches from bacteria and virus sequences within the Genbank *nt* database. This suggests an extremely low level (0.1%) of xenobiotic RNA contamination in this predicted gene set compared with our previous dataset generated from the same laboratory procedure [12].

The number of transcripts having matches from the UniProt, SwissProt and *D. melanogaster* protein databases were 20,841 (52.24%), 14,270 (35.77%), and 12,443 (31.19%), respectively (Additional file 3). Among the UniProt blast hits, 37% were from the eusocial termite *Z. nevadensis*, followed by 7.5% and 4.5% from the pea aphid *Acyrthosiphon pisum* and the araneomorph spider *Stegodyphus mimosarum*, respectively. Notably, the proportion of single-exon transcripts having homologous matches was much lower than the multi-exon transcripts (Fig. 2), likely due to the fact that many of them

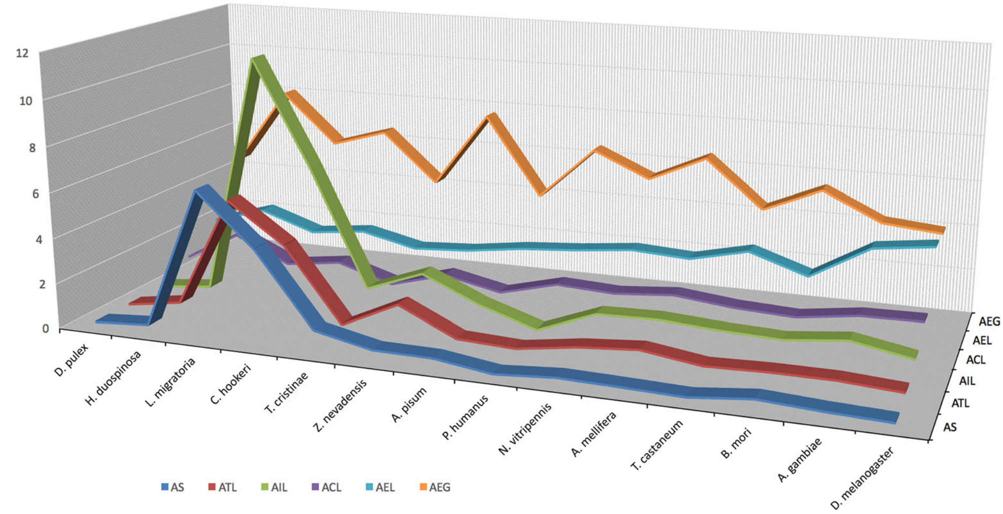


Fig. 1 Correlation between genome size and gene model features across insects (*Holacanthella duospinosa*, *Locusta migratoria*, *Clitarchus hookeri*, *Timema cristinae*, *Zootermopsis nevadensis*, *Acyrthosiphon pisum*, *Pediculus humanus*, *Nasonia vitripennis*, *Tribolium costaneum*, *Bombyx mori*, *Anopheles gambiae* and *Drosophila melanogaster*) and *Daphnia pulex*. AS: genome assembly size; ATL: average of transcript length; AIL: average of intron length; ACL: average of coding region length; AEL: average of exon length; AEG: average number per gene. Scale numbering was adjusted by dividing ATL, AIL, ACL and AEL by 10^4, 10^4, 10^3 and 10^2, respectively

were derived from non-coding regions, thus lacking similarity with sequences from protein databases.

Gonadal sex-biased genes

The female reproductive tract had 3262 gene loci that were uniquely expressed in this tissue, whereas approximately four times of this number (12,516) were detected in testis (ECS > 5). Comparative gene expression analysis between female and male gonadal samples revealed a large set of transcripts that were significantly differentially expressed (FC > 2 and BH adjusted $p < 0.05$). There were 6308 genes significantly overexpressed in the female, whereas the genes displaying male-biased expression had more than two-fold abundance (15,889) (Fig. 3 and Additional file 3). Notably, compared with female-biased transcripts, a much larger proportion of transcripts over-expressed in testis were lacking matches, especially when compared with the *D. melanogaster* protein database (Fig. 3).

The top 20 significantly over-expressed genes having *D. melanogaster* protein matches from each of the sexes are shown in Table 5. The genes with orthologues that were also highly expressed in *D. melanogaster* ovary were *CG1077*, *yellow-g*, *yellow-g2* and *Acph-1*, and testis were *CG12020*, *CG5458*, *CG32392*, *CG14838*, *CG13442*, *CG31068* and *CG17377*, respectively (data retrieved from the FlyAtlas Anatomy Microarray and modENCODE Anatomy RNA-Seq data on http://flybase.org).

Enriched gene ontology terms and pathways

The matched *D. melanogaster* protein hits from the *C. hookeri* sex-biased genes were used to detect enriched gene ontology (GO) terms. In summary, a total of 2405 *D. melanogaster* FlyBase IDs were subjected to multiple tests of GO enrichment using DAVID. Within level one GOs, most of the significantly enriched terms were female-biased. Some of these terms closely represent female features of reproduction, such as developmental

Table 4 Genomic DNA mapping statistics

Sample ID	Read pairs	Uniquely mapped	No. of splices	Multiply mapped	Total mapped (%)
CLI765	47,699,036	37,219,558 (78.0%)	12,199,348	7,414,350 (15.5%)	93.6
CLI767	50,546,345	35,708,416 (70.6%)	11,153,462	12,243,359 (24.2%)	94.9
CLI768	45,978,461	34,815,789 (75.7%)	10,547,208	8,859,828 (19.3%)	95.0
CLI755	44,052,337	34,808,276 (79.0%)	10,253,623	6,884,797 (15.6%)	94.6
CLI757	45,000,013	32,997,927 (73.3%)	10,140,927	8,889,222 (19.8%)	93.1
CLI760	52,261,126	40,301,287 (77.1%)	13,724,663	6,631,109 (12.7%)	89.9

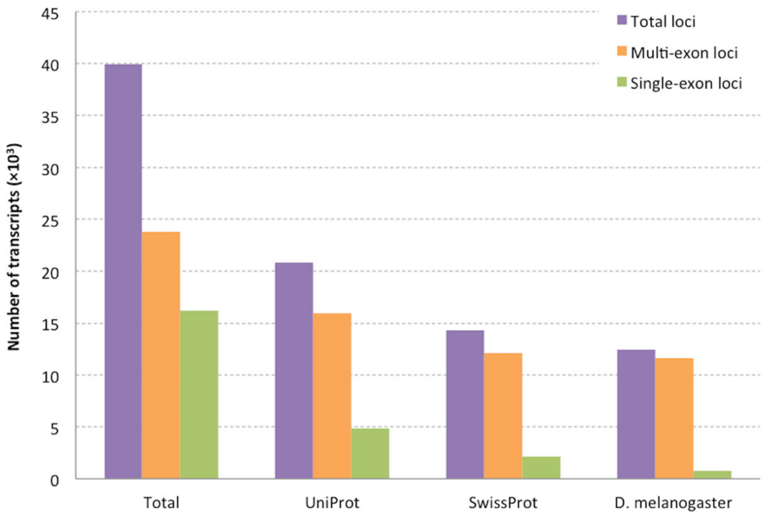


Fig. 2 Distribution of homologous sequence matches from the Uniprot, SwissProt and *Drosophila melanogaster* protein databases, respectively

process, signaling, reproductive process and immune system process from the biological process category (BP). In comparison, rhythmic process (BP), molecular transducer activity (MF) and synapse (CC) were significantly enriched among male-biased genes (Fig. 4a).

After removing all the parental terms, the more specific GOs were revealed (Fig. 4b). The significantly enriched terms for the male-biased genes include ATP binding, motor activity, microtubule binding and ATPase activity (MF), corresponding to sperm maturation and movement; whereas, the terms for the female-biased genes include calcium ion, protein and actin binding and ATPase activity (MF). In addition, females also overexpressed genes involved in glucose transmembrane transporter activity, oxidoreductase activity and carboxylic ester hydrolase activity (MF) (Fig. 4b). Furthermore, we also found GOs associated with oocyte maturation and development, such as imaginal disc-derived wing morphogenesis, open tracheal system development (BP), myofibril assembly, neuron projection morphogenesis and dorsal closure (CC) that were enriched within female-biased genes.

The enriched KEGG pathways include many related to carbohydrate metabolism, including pentose and glucuronate interconversions and starch and sucrose metabolism enriched within female-biased genes, and glycolysis/gluconeogenesis, glycan degradation, glycosaminoglycan degradation and galactose metabolism pathways enriched within male-biased genes (Fig. 5). Other enriched pathways within female-biased genes were ECM-receptor

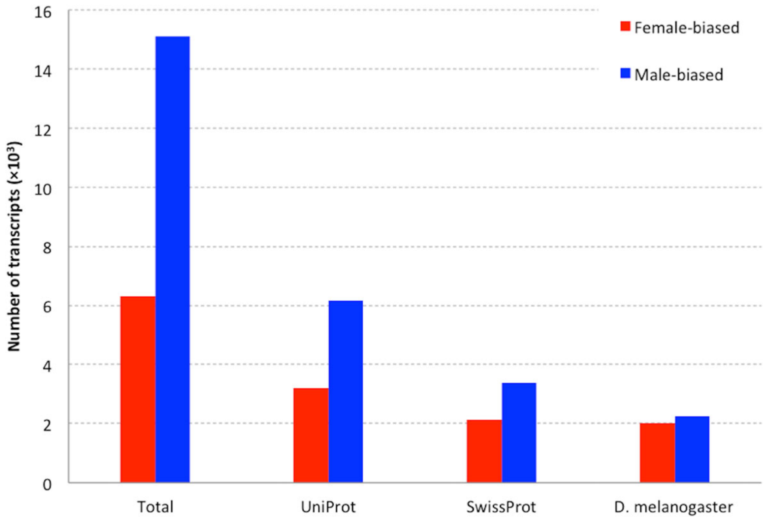
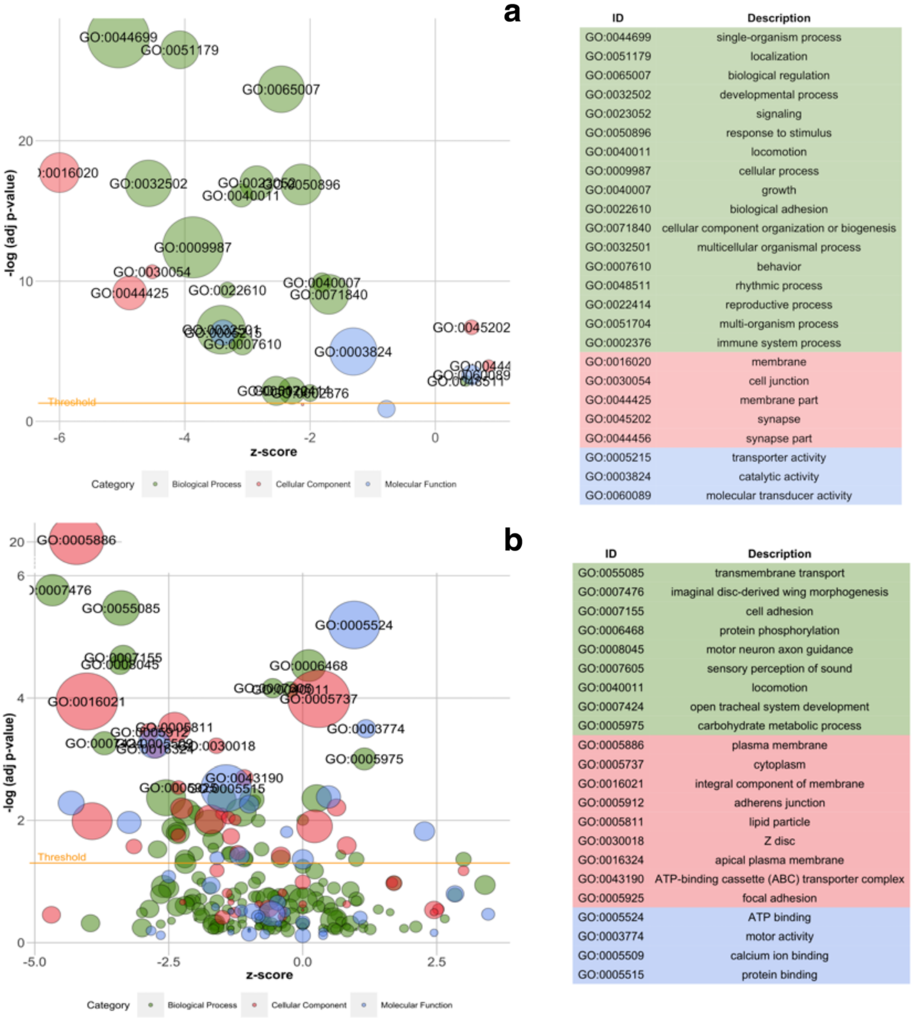


Fig. 3 Distribution of homologous sequence matches from the Uniprot, SwissProt and *Drosophila melanogaster* protein databases for female-biased and male-biased genes, respectively

Table 5 Summary of top 20 differentially expressed genes from female reproductive tract and testis, respectively

	Gene locus	Hit ID	Hit Name	E-value	Log2 FC	Adjusted p
Female-biased	XLOC_018823	FBgn0037405	CG1077[a]	2.38E-06	−13.06	1.58E-12
	XLOC_015696	FBgn0041709	yellow-g[a]	5.04E-19	−12.65	3.91E-11
	XLOC_032619	FBgn0266435	CG45065[a]	1.72E-151	−12.36	1.64E-10
	XLOC_030196	FBgn0035328	yellow-g2[a]	4.20E-65	−12.06	2.09E-36
	XLOC_007922	FBgn0035089	Phk-3	1.59E-08	−12.01	1.45E-91
	XLOC_009044	FBgn0034145	CG5065	1.75E-13	−11.93	2.21E-45
	XLOC_007550	FBgn0039084	CG10175	8.89E-79	−11.9	1.73E-50
	XLOC_021021	FBgn0038799	MFS9	1.96E-09	−11.89	1.59E-09
	XLOC_021022	FBgn0038799	MFS9	3.49E-45	−11.85	1.88E-21
	XLOC_025391	FBgn0013680	mt:ND2	6.59E-09	−11.81	1.74E-283
	XLOC_000499	FBgn0032433	Oatp33Ea	3.76E-32	−11.54	2.71E-09
	XLOC_003735	FBgn0034145	CG5065	3.59E-14	−11.52	6.18E-32
	XLOC_026733	FBgn0000032	Acph-1[a]	4.73E-69	−11.46	2.48E-41
	XLOC_002774	FBgn0000261	Cat	1.66E-34	−11.28	2.37E-08
	XLOC_005024	FBgn0039896	yellow-h	7.54E-111	−11.26	1.75E-18
	XLOC_024658	FBgn0030452	MFS10	5.42E-19	−11.23	2.92E-08
	XLOC_016572	FBgn0038799	MFS9	3.93E-29	−11.12	4.62E-08
	XLOC_032950	FBgn0000032	Acph-1[a]	3.86E-21	−10.99	6.21E-18
	XLOC_036184	FBgn0259247	laccase2	2.23E-09	−10.94	3.44E-09
	XLOC_018465	FBgn0026314	Ugt35b	6.37E-96	−10.91	4.59E-18
Male-biased	XLOC_037165	FBgn0014869	Pglym78	7.44E-70	13.61	1.81E-13
	XLOC_034828	FBgn0035273	CG12020[a]	9.64E-35	13.59	2.02E-13
	XLOC_033633	FBgn0020412	JIL-1	1.21E-19	13.47	4.23E-13
	XLOC_003842	FBgn0032478	CG5458[a]	3.75E-45	13.31	1.07E-12
	XLOC_039688	FBgn0038385	Fbxl7	6.21E-07	13.29	1.15E-12
	XLOC_035079	FBgn0019982	Gs1l	2.23E-53	13.29	1.19E-12
	XLOC_019177	FBgn0020412	JIL-1	2.99E-16	13.25	1.45E-12
	XLOC_039094	FBgn0052392	CG32392[a]	2.03E-08	13.12	3.14E-12
	XLOC_010029	FBgn0005612	Sox14	3.45E-08	13.05	4.59E-12
	XLOC_016670	FBgn0033635	CG7777	2.72E-26	12.98	2.59E-27
	XLOC_035177	FBgn0004380	Klp64D	1.96E-09	12.94	1.30E-35
	XLOC_024216	FBgn0264574	Glut1	1.89E-43	12.88	8.30E-25
	XLOC_004823	FBgn0035799	CG14838[a]	1.28E-72	12.84	1.01E-25
	XLOC_010324	FBgn0034546	CG13442[a]	2.57E-10	12.84	1.45E-11
	XLOC_019341	FBgn0036211	CG5946	9.38E-16	12.77	2.02E-11
	XLOC_002845	FBgn0020412	JIL-1	6.64E-13	12.75	2.79E-24
	XLOC_035152	FBgn0051068	CG31068[a]	1.24E-33	12.73	2.62E-26
	XLOC_038804	FBgn0031988	CG8668	1.07E-63	12.65	7.43E-25
	XLOC_020898	FBgn0031859	CG17377[a]	2.27E-10	12.58	1.47E-24
	XLOC_038151	FBgn0039396	CCAP-R	6.78E-28	12.57	5.83E-11

“[a]” indicates the transcript is also highly expressed in the *Drosophila melanogaster* ovary and testicle respectively



ID	Description
GO:0044699	single-organism process
GO:0051179	localization
GO:0065007	biological regulation
GO:0032502	developmental process
GO:0023052	signaling
GO:0050896	response to stimulus
GO:0040011	locomotion
GO:0009987	cellular process
GO:0040007	growth
GO:0022610	biological adhesion
GO:0071840	cellular component organization or biogenesis
GO:0032501	multicellular organismal process
GO:0007610	behavior
GO:0048511	rhythmic process
GO:0022414	reproductive process
GO:0051704	multi-organism process
GO:0002376	immune system process
GO:0016020	membrane
GO:0030054	cell junction
GO:0044425	membrane part
GO:0045202	synapse
GO:0044456	synapse part
GO:0005215	transporter activity
GO:0003824	catalytic activity
GO:0060089	molecular transducer activity

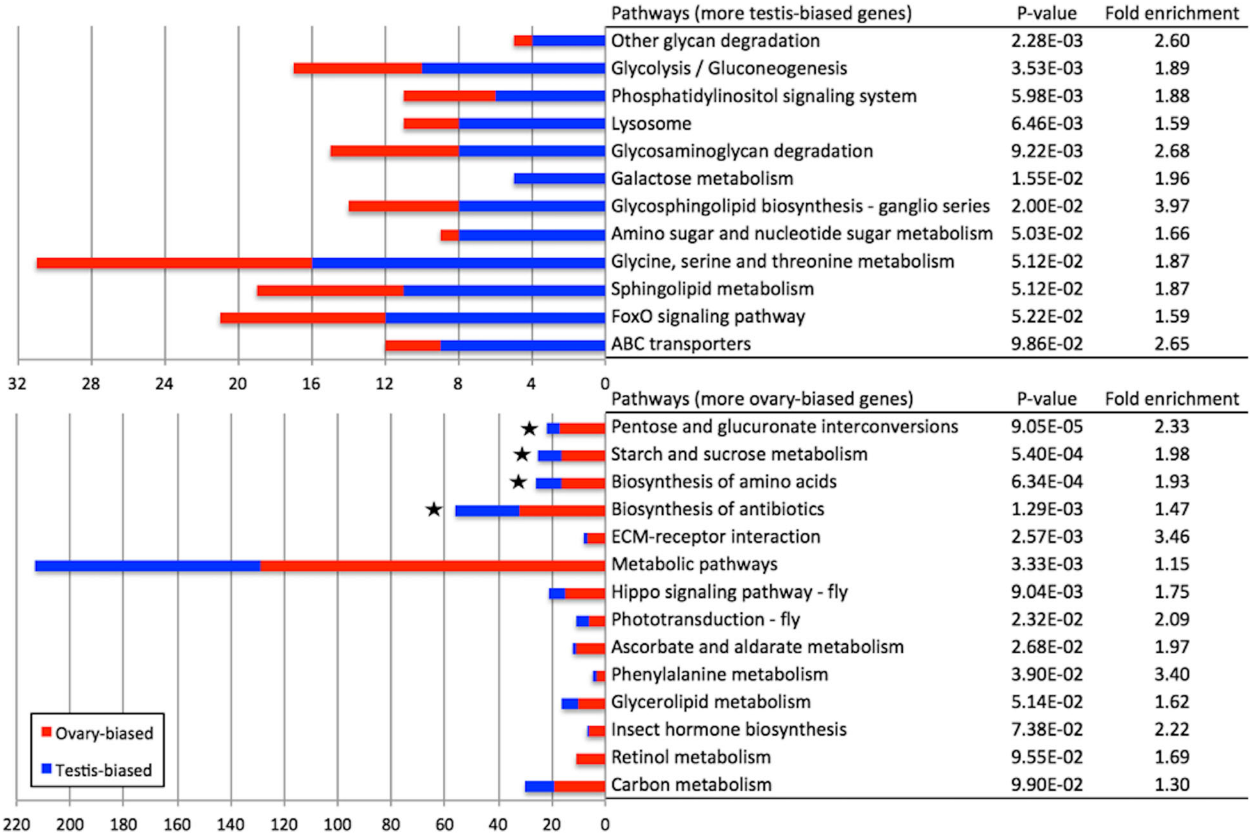
ID	Description
GO:0055085	transmembrane transport
GO:0007476	imaginal disc-derived wing morphogenesis
GO:0007155	cell adhesion
GO:0006468	protein phosphorylation
GO:0008045	motor neuron axon guidance
GO:0007605	sensory perception of sound
GO:0040011	locomotion
GO:0007424	open tracheal system development
GO:0005975	carbohydrate metabolic process
GO:0005886	plasma membrane
GO:0005737	cytoplasm
GO:0016021	integral component of membrane
GO:0005912	adherens junction
GO:0005811	lipid particle
GO:0030018	Z disc
GO:0016324	apical plasma membrane
GO:0043190	ATP-binding cassette (ABC) transporter complex
GO:0005925	focal adhesion
GO:0005524	ATP binding
GO:0003774	motor activity
GO:0005509	calcium ion binding
GO:0005515	protein binding

Fig. 4 Bubble plots showing enriched GO terms generated from sex-biased genes. "Z-score" was calculated for each term using the formula: male-biased minus female-biased gene number divided by the square root of sex-biased gene number. The significant GOs are indicated above the yellow line. The bubbles calculated with minus z-scores represent GOs containing more female-biased genes, while GOs with more male-biased genes have z-scores larger than zero. The bubble size represents the number of genes. The GO descriptions on the right are listed from the highest to lowest significance of enrichment. (**a**) Level one terms; (**b**) after removing all parental terms

interaction, hippo signaling pathway, phototransduction, insect hormone biosynthesis and retinol metabolism, and within male-biased genes were lysosome, phosphatidylinositol signaling system, glycosphingolipid biosynthesis, foxO signaling pathway and ABC transporters (Fig. 5). Within the juvenile hormone (JH) biosynthesis pathway (belonging to the insect hormone biosynthesis pathway), JH epoxide hydrolase 2 (*Jheh2*) and *CG9360* were over-expressed in females.

Discussion

The *C. hookeri* genome is the second largest insect genome published to date after *L. migratoria* [28]. It is also the second stick insect genome to be reported and the only one sequenced from the Euphasmatodea. Thus, the availability of this genome sequence provides valuable information to investigate the evolution of large insect genomes and the diversification of the Phasmatodea. In particular, parthenogenesis and other alternative reproductive strategies have played an important role in phasmid evolution. However, the molecular mechanisms underlying parthenogenesis have been poorly studied. Characterisation of genes essential to stick insect reproduction, especially from species with different reproductive modes, is the first step towards understanding reproductive flexibility. In this study, we constructed a *C. hookeri* genome assembly, which was then used as a reference genome to detect genes that are present in the female reproductive tract and testicle transcriptomes.



Pathways (more testis-biased genes)	P-value	Fold enrichment
Other glycan degradation	2.28E-03	2.60
Glycolysis / Gluconeogenesis	3.53E-03	1.89
Phosphatidylinositol signaling system	5.98E-03	1.88
Lysosome	6.46E-03	1.59
Glycosaminoglycan degradation	9.22E-03	2.68
Galactose metabolism	1.55E-02	1.96
Glycosphingolipid biosynthesis - ganglio series	2.00E-02	3.97
Amino sugar and nucleotide sugar metabolism	5.03E-02	1.66
Glycine, serine and threonine metabolism	5.12E-02	1.87
Sphingolipid metabolism	5.12E-02	1.87
FoxO signaling pathway	5.22E-02	1.59
ABC transporters	9.86E-02	2.65

Pathways (more ovary-biased genes)	P-value	Fold enrichment
Pentose and glucuronate interconversions	9.05E-05	2.33
Starch and sucrose metabolism	5.40E-04	1.98
Biosynthesis of amino acids	6.34E-04	1.93
Biosynthesis of antibiotics	1.29E-03	1.47
ECM-receptor interaction	2.57E-03	3.46
Metabolic pathways	3.33E-03	1.15
Hippo signaling pathway - fly	9.04E-03	1.75
Phototransduction - fly	2.32E-02	2.09
Ascorbate and aldarate metabolism	2.68E-02	1.97
Phenylalanine metabolism	3.90E-02	3.40
Glycerolipid metabolism	5.14E-02	1.62
Insect hormone biosynthesis	7.38E-02	2.22
Retinol metabolism	9.55E-02	1.69
Carbon metabolism	9.90E-02	1.30

Fig. 5 KEGG pathways enriched in gonadal sex-biased genes. Enriched pathways with modified Fisher exact test *p* value below 0.1 are shown here. Stars indicate pathways with BH adjusted *p* value below 0.05

Evolution of repeats and gene length expansion

The *C. hookeri* genome assembly comprising 4.2 Gb, was very similar to the flow cytometry estimate, and approximately four times larger than *T. cristinae* [19]. Similar to *L. migratoria*, the analysis of the *C. hookeri* genome also revealed a large volume of repetitive elements. DNA transposon (Class II) was the most dominant repeat group in the stick insect and the *L. migratoria* genomes; whereas, the genome of *Z. nevadensis*, only around one third of *T. cristinae* genome size, was mostly enriched with Non-LTR retroid long interspersed element (LINE). The comparison also shows that the increasing genome size from *Z. nevadensis* to *L. migratoria* is consistent with the growing absolute amount and proportion of DNA transposons. Relative to *T. cristinae*, the *C. hookeri* genome is expanded with a variety of repeat types. Notably, MITE, the most abundant group of DNA transposon in the *C. hookeri* genome assembly, was also reported to be present at [61] high frequency in the partial *Bacillus* genome sequences [20], but absent in the *T. cristinae* genome. Either the invasion and proliferation of MITEs occurred along the evolution of euphasmids or the loss of MITEs occurred along the *Timema* lineage. A survey of MITEs across a wide range of stick insect lineages may shed light on the evolution of this repeat group in stick insects. A relatively large proportion of un-classified repeats were also revealed in the two stick insect genomes, whether they indeed represent novel repeat families requires further analysis.

The annotated gene models from *L. migratoria* and *C. hookeri* were roughly 6 and 5 times longer than other insects with moderate to small genome sizes. This suggests there was an intron length expansion during the formation of large insect genomes. There are several consequences to an increase in intron length. First, long introns are associated with an increased metabolic cost as introns are transcribed together with exons [61]. Second, they are thought to be associated with the large genome chromosomal compactness [61]. Third, they may be negatively correlated with recombination [62]. The *C. hookeri* genome provides a novel resource to test these three hypotheses in insects. In addition to longer introns, the *C. hookeri* genes also exhibited a higher number of introns and exons per gene and a slightly longer coding sequence length compared with *T. cristinae*. However, whether these patterns hold across the Euphasmatodea requires the availability of more stick insect genome sequences.

Sex-biased genes

The comparative transcriptome analysis between female and male gonads revealed a large set of genes with sex-biased expression. The male-biased genes were more than twice as abundant as those displaying female-biased expression, and a much greater proportion of male-biased genes were lacking blast matches from the known proteins. Also, *Clitarchus hookeri* females sometimes reproduce by parthenogenesis, which might lead to less selective pressure on males because the male-biased genes are used less frequently than the female-biased genes [63, 64]. It has also been reported that the sexually-derived *Timema* parthenogenetic lineages have experienced sexual trait decline, such as shrunken spermatotheca and the loss of male attractiveness [65]. This has also likely occurred in the *C. hookeri* parthenogens. It is possible that this sexual trait decay correlates with female-biased gene expression change. All these hypotheses require further investigation of the sequence and expression divergence across euphasmids and the gene expression between sexual and parthenogenetic *C. hookeri*.

The *C. hookeri* genes that are over-expressed in female reproductive tract are enriched in development, signalling, growth, behaviour and reproductive process. The top five female-biased DE genes had *D. melanogaster* matches with a chorion-containing eggshell formation protein (*CG1077*) predicted to have anti-microbial activity [66], *yellow-g* and *yellow-g2* proteins essential to eggshell integrity [67], *CG45065* protein responding to mating [68], and *Phk-3* protein associated with metamorphosis [69]. Stick insects show interesting traits relating to egg morphology [70, 71]. Eggs have in a variety of shapes, often mimicking plant seeds, many of which contain a knob-like capitulum that resembles an elaiosome to attract ants for burial [72]. Recently, a species (*Korinninae* sp.) was found to produce an ootheca containing numerous eggs in a highly ordered arrangement, which is distinct from other stick insects that produce eggs singly by dropping them to the ground or inserting them into crevices or soil [73]. The genes identified in our study may be essential to eggshell formation and maturation and are candidates to further investigate egg variation and adaptations in stick insects.

In comparison, the male-biased genes were significantly enriched for rhythmic process, molecular transducer activity and synapse. Rhythmic processes play important roles in temporally coordinating the release of sequential sperm, the acidification of the vas deferens, and contractile activity [74–77]. The enrichment of the molecular transducer activity was also found in crab and a sex-changing fish testis [78, 79]. The top five male-biased genes with blast matches include a glycolysis protein *Pglym78* that is present in the semen and seminal vesicle tissue of a honey bee [80], *CG12020* enriched in the sperm proteome with a role in protein folding [81, 82], *JIL-1* essential for chromosomal organisation [83], *CG5458* involved in sperm axoneme assembly [81, 82] and *Fbxl7* regulating mitosis through Aurora A [84].

Notably, a large number of sex-biased transcripts showed no matches with any of the databases. These transcripts may include: 1) highly diverged genes; 2) unknown genes; and 3) non-coding elements. Stick insects contain panoistic ovaries where oogonia eventually differentiate into oocytes [85]. However, insect ovary-biased genes have mostly been identified from the meroistic ovaries frequently present in the derived Holometabola, where oogonia differentiate into an oocyte and several nurse cells [86, 87]. Thus, some of the novel female-biased genes found in this study may be playing unique functions in the panoistic ovary.

Interestingly, we found some *C. hookeri* genes with matches to *D. melanogaster* proteins involved in the JH synthesis pathway. These genes include *Jheh1* and *Jheh2* involved in the JH catabolic process [88] and *CG9360* having an oxidoreductase activity [89], *Jheh2* and *CG9360* over-expressed in female *C. hookeri*. It has been suggested that in the cyclical parthenogenetic aphids, higher levels of JH induce parthenogenetic reproduction [90–93]. However, whether the level of JH has an impact in the differentiation of the southern parthenogenetic *C. hookeri* is unknown. This could be examined by assessing levels of JH or indirectly by measuring expression levels of the genes involved in JH hormone biosynthesis. In addition, we also found some enriched pathways were related in carbohydrate metabolism (e.g. starch and sucrose metabolism and glycan degradation), which are likely to contain genes playing important roles in nutrition and energy support for the production of gametes and the maintenance of reproductive organs.

Conclusions

The analysis of the *C. hookeri* genome assembly revealed a large, repetitive genome, likely resulting from the accumulation of DNA transposons along with an increase of intron length. MITE, the most abundant repeat type in the *C. hookeri* genome assembly, contributed significantly to the growth in genome size. Using a reference-guided approach coupled with differential expression analysis, a large number of sex-biased genes were identified by comparing gonadal transcriptomes between females and males. Female reproductive tract over-expressed genes were involved in development, signalling, growth, behaviour and reproductive process, whereas, testicle over-expressed genes were involved in rhythmic process, transducer activity and synapse. We also identified several genes involved in JH synthesis that were over-expressed in the female. These genes are an important resource for furthering understanding of the evolution of reproductive strategies within Phasmatodea.

Additional files

Additional file 1: Scaffolds match archea, bacteria and virus sequences. (XLSX 140 kb)

Additional file 2: BLAST results from gene models. (XLSX 5388 kb)

Additional file 3: BLAST results from transcripts, raw counts from gene expression and differential expression results. (XLSX 13185 kb)

Abbreviations

Bp: Base pair; BP: Biological process; CC: Cellular component; Gb: Gigabases; GO: Gene ontology; JH: Juvenile hormone; Jheh2: JH epoxide hydrolase; LTR: Long terminal repeat retrotransposon; MF: Molecular function; MITEs: Miniature inverted repeat transposable elements; MP: Mate pair; PE: Paired end; SINE: Short interspersed element; TEs: Interspersed transposon elements

Acknowledgments

We thank New Zealand eScience Infrastructure (NeSI) for high-performance computing facilities; Elena Hilario, Gary Houliston, Dagmar Goeke, Robyn Howitt, Julia Allwood, Ana Ramon-Laca, and Duckchul Park for laboratory support. We thank John Parsch and two anonymous reviewers for comments that improved the manuscript.

Funding

This research was funded by the Alan Wilson Centre (RM 13790/24947), the University of Auckland, and Strategic Science Investment Fund to Landcare Research from the Ministry of Business, Innovation and Employment's Science and Innovation Group.

Availability of data and materials

Data from this study are deposited in the National Centre for Biotechnology Information (NCBI) under the project ID PRJNA395945 including whole genome sequencing (WGS) and RNA sequencing (RNA-Seq) reads and the draft whole genome assembly (contigs >200 bp).

Authors' contributions

TRB and RDN conceived and designed the project and advised on analyses. CW performed the genome and transcriptome assembly, analysis. RNC provided valuable suggestions and participated in the bioinformatics analysis. VGT performed flow cytometry experiment. VGT and CW performed DNA and RNA extractions. CW prepared the manuscript draft. TRB, RDN and RNC edited and commented extensively on the draft manuscript. All authors reviewed and approved the final manuscript.

Ethics approval and consent to participate

Specimens of *C. hookeri* were collected from Totara Park under permits issued by Auckland Council.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]School of Biological Sciences, The University of Auckland, Auckland, New Zealand. [2]Landcare Research, Auckland, New Zealand. [3]New Zealand Institute for Plant & Food Research Ltd, Auckland, New Zealand. [4]Department of Biology, Lund University, Lund, Sweden.

Received: 21 August 2017 Accepted: 31 October 2017
Published online: 16 November 2017

References

1. Scali V: Metasexual stick insects: model pathways to losing sex and bringing it back. In: Schön I, Martens K, Dijk PV, editors. *Lost sex*. Springer; 2009:317–345.
2. Buckley TR, Marske K, Attanayake D. Phylogeography and ecological niche modelling of the New Zealand stick insect *Clitarchus hookeri* (white) support survival in multiple coastal refugia. J Biogeogr. 2010;37(4):682–95.
3. Morgan-Richards M, Trewick SA, Stringer IA. Geographic parthenogenesis and the common tea-tree stick insect of New Zealand. Mol Ecol. 2010;19(6):1227–38.
4. Buckley TR, Attanayake D, Park D, Ravindran S, Jewell TR, Normark BB. Investigating hybridization in the parthenogenetic New Zealand stick insect *Acanthoxyla* (Phasmatodea) using single-copy nuclear loci. Mol Phylogenet Evol. 2008;48(1):335–49.
5. Morgan-Richards M, Trewick SA. Hybrid origin of a parthenogenetic genus? Mol Ecol. 2005;14(7):2133–42.
6. Trewick SA, Morgan-Richards M, Collins LJ. Are you my mother? Phylogenetic analysis reveals orphan hybrid stick insect genus is part of a monophyletic New Zealand clade. Mol Phylogenet Evol. 2008; 48(3):799–808.
7. Morgan-Richards M, Hills SF, Biggs PJ, Trewick SA. Sticky genomes: using NGS evidence to test hybrid speciation hypotheses. PLoS One. 2016;11(5):e0154911.
8. Myers SS, Trewick SA, Morgan-Richards M. Multiple lines of evidence suggest mosaic polyploidy in the hybrid parthenogenetic stick insect lineage *Acanthoxyla*. Insect Conserv Diver. 2013;6(4):537–48.
9. Dennis AB, Dunning LT, Sinclair BJ, Buckley TR. Parallel molecular routes to cold adaptation in eight genera of New Zealand stick insects. Sci Rep. 2015;5:13965.
10. Myers SS, Buckley TR, Holwell GI. Mate detection and seasonal variation in stick insect mating behaviour (Phamatodea: *Clitarchus hookeri*). Behaviour. 2015;152(10):1325–48.
11. Myers SS, Buckley TR, Holwell GI. Male genital claspers influence female mate acceptance in the stick insect *Clitarchus hookeri*. Behav Ecol Sociobiol. 2016;70(9):1547–56.
12. Wu C, Crowhurst RN, Dennis AB, Twort VG, Liu S, Newcomb RD, Ross HA, Buckley TR. 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.
13. Myers SS, Holwell GI, Buckley TR. Genetic and morphometric data demonstrate alternative consequences of secondary contact in *Clitarchus* stick insects. J Biogeogr. 2017;44(9):2069–81.
14. Akbari OS, Antoshechkin I, Amrhein H, Williams B, Diloreto R, Sandler J, Hay BA. The developmental transcriptome of the mosquito *Aedes aegypti*, an invasive species and major arbovirus vector. G3: Genes Genomes Genetics. 2013;3(9):1493–509.
15. Ferree PM, Fang C, Mastrodimos M, Hay BA, Amrhein H, Akbari OS. Identification of genes uniquely expressed in the germ-line tissues of the Jewel wasp *Nasonia vitripennis*. G3: Genes Genomes Genetics. 2015; 5(12):2647–53.
16. Chen W, Liu Y-X, Jiang G-F. *De novo* assembly and characterization of the testis Transcriptome and development of EST-SSR markers in the cockroach \textitPeriplaneta americana. Sci Rep. 2015;5:11144.
17. Parsch J, Ellegren H. The evolutionary causes and consequences of sex-biased gene expression. Nat Rev Genet. 2013;14(2):83–7.
18. Lu B, Zeng Z, Shi T. Comparative study of *de novo* assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. Sci China Life Sci. 2013;56(2):143–55.
19. Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, Buerkle CA, Feder JL, Bast J, Schwander T, et al. Stick insect genomes reveal natural selection's role in parallel speciation. Science. 2014; 344(6185):738–42.
20. Ricci M, Luchetti A, Bonandin L, Mantovani B. Random DNA libraries from three species of the stick insect genus bacillus (Insecta: Phasmida): repetitive DNA characterization and first observation of polyneopteran MITEs. Genome. 2013;56(12):729–35.
21. Hanrahan SJ, Johnston JS. New genome size estimates of 134 species of arthropods. Chromosom Res. 2011;19(6):809–23.

22. Mikheyev AS, Zwick A, Magrath MJL, Grau ML, Qiu L, Su YN, Yeates D. Museum genomics confirms that the Lord Howe Island stick insect survived extinction. Curr Biol. 2017; In press
23. Consortium HGS, et al. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature. 2006;443(7114):931.
24. Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Group NGW, et al. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science. 2010;327(5963):343–8.
25. Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, et al. Draft genome of the globally widespread and invasive argentine ant (*Linepithema humile*). Proc Natl Acad Sci U S A. 2011; 108(14):5673–8.
26. Zhan S, Merlin C, Boore JL, Reppert SM. The monarch butterfly genome yields insights into long-distance migration. Cell. 2011;147(5):1171–85.
27. You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, et al. A heterozygous moth genome provides insights into herbivory and detoxification. Nat Genet. 2013;45(2):220–5.
28. Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, Li B, Cui F, Wei J, Ma C, et al. The locust genome provides insight into swarm formation and long-distance flight. Nat Commun. 2014;5:2957.
29. Otto F. DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA. Methods Cell Biol. 1990;33:105–10.
30. Lysak MA, Dolezel J. Estimation of nuclear DNA content in *Sesleria* (Poaceae). Caryologia. 1998;51(2):123–32.
31. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4.
32. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. FastUniq: a fast *de novo* duplicates removal tool for paired short reads. PLoS One. 2012; 7(12):e52249.
33. Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–12.
34. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, et al. *De novo* transcriptome assembly with ABySS. Bioinformatics. 2009;25(21):2872–7.
35. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011;108(4):1513–8.
36. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1(1):18.
37. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578–9.
38. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9.
39. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061–7.
40. Smit A, Hubley R: RepeatModeler Open-1.0. <http://wwwrepeatmaskerorg/> 2008-2015.
41. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
42. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H. PASTEC: an automatic transposable element classification tool. PLoS One. 2014;9(5):e91929.
43. Smit A, Hubley R, Green P: RepeatMasker Open-4.0. <http://wwwrepeatmaskerorg/> 2013-2015.
44. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
45. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19(Suppl 2):ii215–25.
46. Keller O, Odronitz F, Stanke M, Kollmar M, Waack S. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics. 2008;9:278.
47. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic Acids Res. 2011;39(Database issue):D283–8.
48. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
49. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat Protoc. 2012;7(3):562–78.
50. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform. 2002;3(3):275–84.
51. UniProt C. The universal protein resource (UniProt). Nucleic Acids Res. 2008; 36(Database issue):D190–5.
52. Drysdale RA, Crosby MA, FlyBase C. FlyBase: genes and gene models. Nucleic Acids Res. 2005;33(Database issue):D390–5.
53. Anders S, Pyl PT, Huber W. HTSeq–a python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
54. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5(3):299–314.
55. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
56. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
57. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
58. Walter W, Sánchez-Cabo F, Ricote M. GOplot: an R package for visually combining expression data with functional analysis. Bioinformatics. 2015; 31(17):2912–4.
59. Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, Chen Z, Childers CP, Glastad KM, Gokhale K, et al. Molecular traces of alternative social organization in a termite genome. Nat Commun. 2014;5:3636.
60. Wu C, Jordan MD, Newcomb RD, Gemmell NJ, Bank S, Meusemann K, Dearden PK, Duncan EJ, Grosser S, Rutherford K, et al. Analysis of the genome of the New Zealand giant Collembola (*Holacanthella duospinosa*) sheds light on hexapod evolution. BMC Genomics. 2017; In press
61. Vinogradov AE. Intron-genome size relationship on a large evolutionary scale. J Mol Evol. 1999;49(3):376–84.
62. Comeron JM, Kreitman M. The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. Genetics. 2000;156(3):1175–90.
63. Huylmans AK, Ezquerra AL, Parsch J, Cordellier M. *De novo* transcriptome assembly and sex-biased gene expression in the cyclical parthenogenetic *Daphnia galeata*. Genome Biol Evol. 2016;8(10):3120–39.
64. Purandare SR, Bickel RD, Jaquiery J, Rispe C, Brisson JA. Accelerated evolution of morph-biased genes in pea aphids. Mol Biol Evol. 2014; 31(8):2073–83.
65. Schwander T, Crespi BJ, Gries R, Gries G. Neutral and selection-driven decay of sexual traits in asexual stick insects. Proc Biol Sci. 2013; 280(1764):20130823.
66. Tootle TL, Williams D, Hubb A, Frederick R, Spradling A. *Drosophila* eggshell production: identification of new genes and coordination by Pxt. PLoS One. 2011;6(5):e19943.
67. Claycomb JM, Benasutti M, Bosco G, Fenger DD, Orr-Weaver TL. Gene amplification as a developmental strategy: isolation of two developmental amplicons in *Drosophila*. Dev Cell. 2004;6(1):145–55.
68. McGraw LA, Gibson G, Clark AG, Wolfner MF. Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. Curr Biol. 2004;14(16):1509–14.
69. Sabatier L, Jouanguy E, Dostert C, Zachary D, Dimarcq JL, Bulet P, Imler JL. Pherokine-2 and -3. Eur J Biochem. 2003;270(16):3398–407.
70. Buckley TR, Bradler S. *Tepakiphasma ngatikuri*, a new genus and species of stick insect (Phasmatodea) from the far north of New Zealand. New Zealand Entomologist. 2010;33(1):118–26.
71. Sellick JC. Descriptive terminology of the phasmid egg capsule, with an extended key to the phasmid genera based on egg structure. Syst Entomol. 1997;22(2):97–122.
72. Hughes L, Westoby M. Capitula on stick insect eggs and elaiosomes on seeds: convergent adaptations for burial by ants. Funct Ecol. 1992;6(6): 642–8.
73. Goldberg J, Bresseel J, Constant J, Kneubühler B, Leubner F, Michalik P, Bradler S. Extreme convergence in egg-laying strategy across insect orders. Sci Rep. 2015;5:7825.

74. Beaver L, Gvakharia B, Vollintine T, Hege D, Stanewsky R, Giebultowicz J. Loss of circadian clock function decreases reproductive fitness in males of *Drosophila melanogaster*. Proc Natl Acad Sci. 2002;99(4):2134–9.
75. Bebas P, Cymborowski B, Giebultowicz J. Circadian rhythm of sperm release in males of the cotton leafworm, *Spodoptera littoralis*: in vivo and in vitro studies. J Insect Physiol. 2001;47(8):859–66.
76. Giebultowicz J, Riemann J, Raina A, Ridgway R. Circadian system controlling release of sperm in the insect testes. Science. 1989;245(4922):1098–100.
77. Kotwica-Rolinska J, Gvakharia BO, Kedzierska U, Giebultowicz JM, Bebas P: Effects of period RNAi on V-ATPase expression and rhythmic pH changes in the vas deferens of *Spodoptera littoralis* (Lepidoptera: Noctuidae). Insect Biochem Mol Biol 2013, 43(6):522-532.
78. He L, Wang Q, Jin X, Wang Y, Chen L, Liu L, Wang Y. Transcriptome profiling of testis during sexual maturation stages in \textitEriocheir sinensis using Illumina sequencing. PLoS One. 2012;7(3):e33735.
79. Liu H, Lamm MS, Rutherford K, Black MA, Godwin JR, Gemmell NJ. Large-scale transcriptome sequencing reveals novel expression patterns for key sex-related genes in a sex-changing fish. Biol Sex Differ. 2015;6(1):1.
80. Collins AM, Caperna TJ, Williams V, Garrett WM, Evans JD. Proteomic analyses of male contributions to honey bee sperm storage and mating. Insect Mol Biol. 2006;15(5):541–9.
81. Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL. Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. Nat Genet. 2006;38(12):1440–5.
82. Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, Wilkin E, Polpitiya A, Petritis K, Karr TL: The *Drosophila melanogaster* sperm proteome-II (DmSP-II). J Proteome 2010, 73(11):2171-2185.
83. Ivaldi MS, Karam CS, Corces VG. Phosphorylation of histone H3 at Ser10 facilitates RNA polymerase II release from promoter-proximal pausing in *Drosophila*. Genes Dev. 2007;21(21):2818–31.
84. Coon TA, Glasser JR, Mallampalli RK, Chen BB. Novel E3 ligase component FBXL7 ubiquitinates and degrades aurora a, causing mitotic arrest. Cell Cycle. 2012;11(4):721–9.
85. Taddei C, Chicca MV, Maurizii MG, Scali V. The germarium of panoistic ovarioles of *Bacillus rossius* (Insecta Phasmatodea): larval differentiation. Invertebr Reprod Dev. 1992;21(1):47–56.
86. Büning J: The insect ovary: ultrastructure, previtellogenic growth and evolution: Springer Science & Business Media; 1994.
87. Irles P, Bellés X, Piulachs MD. Identifying genes related to choriogenesis in insect panoistic ovaries by suppression subtractive hybridization. BMC Genomics. 2009;10(1):1.
88. Campbell PM, Healy MJ, Oakeshott JG. Characterisation of juvenile hormone esterase in *Drosophila melanogaster*. Insect Biochem Mol Biol. 1992;22(7):665–77.
89. Mayoral JG, Nouzova M, Navare A, Noriega FG. NADP+−dependent farnesol dehydrogenase, a corpora allata enzyme involved in juvenile hormone synthesis. Proc Natl Acad Sci. 2009;106(50):21091–6.
90. Mittler T, Nassar S, Staal G. Wing development and parthenogenesis induced in progenies of kinoprene-treated gynoparae of *Aphis fabae* and *Myzus persicae*. J Insect Physiol. 1976;22(12):1717–25.
91. Hardie J, Baker FC, Jamieson GC, Lees AD, Schooley DA. The identification of an aphid juvenile hormone, and its titre in relation to photoperiod. Physiol Entomol. 1985;10(3):297–302.
92. Corbitt TS, Hardie J. Juvenile hormone effects on polymorphism in the pea aphid, *Acyrthosiphon pisum*. Entomol Exp Appl. 1985;38(2):131–5.
93. Liu L-J, Zheng H-Y, Jiang F, Guo W, Zhou S-T. Comparative transcriptional analysis of asexual and sexual morphs reveals possible mechanisms in reproductive polyphenism of the cotton aphid. PLoS One. 2014;9(6):e99506.