Deep, Staged Transcriptomic Resources for the Novel Coleopteran Models *Atrachya menetriesi* and *Callosobruchus maculatus*

Matthew A. Benton¹⁺, Nathan J. Kenny²⁺, Kai H. Conrads¹, Siegfried Roth¹* and Jeremy A. Lynch³*

¹ Institute for Developmental Biology, University of Cologne, Zülpicherstrasse 47b, Cologne 50674, Germany

² Simon F.S. Li Marine Science Laboratory of School of Life Sciences and Center for Soybean Research of the State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong

³ Department of Biological Sciences, University of Illinois at Chicago, Chicago, Illinois, United States of America

⁺Authors contributed equally

* Co-corresponding authors

Matthew A. Benton: matthewabenton@gmail.com

Nathan J. Kenny: nathanjameskenny@gmail.com

Kai Conrads: kai.conrads@gmx.de

Siegfried Roth: siegfried.roth@uni-koeln.de

Jeremy A. Lynch: jlynch42@uic.edu
Abstract:

Background: Despite recent efforts to sample broadly across metazoan and insect diversity, current sequence resources in the Coleoptera do not adequately describe the diversity of the clade. Here we present deep transcriptomic data generated with Illumina platform sequencing of a number of key stages in the development of two coleopteran species, the false melon beetle *Atrachya menetriesi* (Faldermann 1835) and the cowpea weevil *Callosobruchus maculatus* (Fabricius 1775). These species are both agricultural pests and are of great interest to developmental biology, as well as providing additional coleopteran datapoints for comparative genomic analysis.

Results: By sampling at a range of timepoints covering ovary development and several crucial phases of embryonic development we assembled five and three timed transcriptomes for *A. menetriesi* and *C. maculatus*, respectively. We utilized this sequence data to build a transcriptomic resource combining all reads into a single, combined assembly for each species, and we analysed each of these resources in detail. The combined *A. menetriesi* assembly consists of 228,096 contigs with an N50 of 1,598 bp, while the combined *C. maculatus* assembly consists of 128,837 contigs with an N50 of 2,263 bp. For each of these assemblies, we identified 78,879 (34.6%) and 41,744 (32.4%) of contigs by BLASTx against the nr database using Blast2GO, and we found that 97% and 98.3% of the BUSCO set of metazoan orthologs were present in these assemblies. We also carried out manual analysis of 38 key developmental genes, and found that nearly all expected genes were present in our transcriptomes. Each of these analyses showed that these transcriptomes are of very high quality. Lastly, we performed read mapping using the individual timed RNA samples, allowing us
to identify up- and down-regulated contigs at key stages in the development of these beetles for further analysis.

Conclusions: The data presented here represent a significant increase in the publicly available transcriptomic resources in the Coleoptera. They have also allowed us to note significant changes in expression at important embryonic stages, and will provide a firm basis for a variety of experimentation, both in developmental biology and in wider exploration of the genomic basis of growth and differentiation.

Key Words: Beetle, Coleopteran, Transcriptome, Ovary, Embryonic, Staged, Atrachya menetriesi, Callosobruchus maculatus, Chrysomelidae

Background:

The order Coleoptera is the most speciose clade of animals currently known. Despite the best efforts of generations of biologists, its species are only sparsely sampled and are yet to be comprehensively described, with approximately 90% of coleopteran diversity as yet uncategorized (e.g. [1, 2]). A similar discrepancy exists at the molecular level; while several genomic resources are available in this clade, their number and phylogenetic distribution is only just beginning to accurately sample that of the Coleoptera as a whole. A wide range of transcriptomic data is available in whole organisms (for example [3–5], among others), specific body parts (such as [6, 7]), and in several cases for staged embryos following RNA interference-mediated gene knock-down [8, 9]. The i5K project [10] will also greatly advance our knowledge of the genomic complement of Coleopterans, with 69 species of this Order listed on that database as nominated for genomic sequencing as of the 14/09/15 (url: http://arthropodgenomes.org/wiki/i5K_nominations) and at least one genome publically available
However, the majority of this information is largely still outside of the public domain, the Coleoptera are still relatively undersampled compared to the Diptera and Hymenoptera, and, in particular, timed embryonic resources (e.g. [12]) are rare in the literature.

The true phylogeny of the Coleoptera is still under investigation but, in general, four suborders, 17 superfamilies, and 168 families are recognised [13]. The structure of the coleopteran clade can be seen summarised in Fig 1A. Coleopterans have long been used for research into embryology, and in the pre-molecular era, the Chrysomelidae (summarised phylogeny shown in Fig 1B) was one of the best studied superfamilies. For example, the first functional embryonic experiments in any insect were carried out in the Colorado potato beetle, *Leptinotarsa decemlineata* (sub-family Chrysomelinae, see Fig 1B), leading to the discovery of the function of pole cells and the existence of germ plasm in insects [14, 15]. Further, the larval cuticle preparation technique, so vital for arthropod developmental biology, was first perfected in the bean beetle, *Callosobruchus maculatus* [16]. Chrysomelid beetles are also interesting from an ecological and economic viewpoint, as members of the group are usually pest species, perhaps most famously the aforementioned Colorado potato beetle, which is a major pest of potato crops in America, Asia and Europe.

The Chrysomelidae are represented in public databases by a number of ongoing transcriptomic and genomic projects. In particular, both *L. decemlineata* (Bioproject PRJNA171749) and *C. maculatus* [17] are the subject of ongoing genomic sequencing. However, while a number of transcriptomes are planned or published in this clade (e.g. [5, 6, 18–22]) none yet sample across embryonic time points in a fashion allowing insight into the genetic mechanisms behind key developmental stages and deep sampling of developmentally important genes. Here we present transcriptomic sequences for two species of chrysomelid beetle, the false melon beetle, *Atrachya menetriesi*, and the aforementioned *C. maculatus* (each of which are pictured along with their relative phylogenetic positions in Fig 1).
Compared to *C. maculatus*, the false melon beetle *A. menetriesi* (Faldermann), is a relatively unknown and understudied species. It is native to Japan, where it is an agricultural pest, and feeds on a variety of plants such as clover and lettuce. Although there has only been a small amount of research carried out on this beetle, the work that has been done has highlighted several interesting developmental traits, including the possibility of generating twin embryos after egg bisection [23], or up to four, seemingly complete, embryos following treatment with low temperatures [23, 24]. Another interesting trait exhibited by this beetle is that almost all eggs enter diapause at a certain stage, and this is only broken in the wild by winter conditions. However, a small proportion of eggs skip this diapause and continue to adulthood. The ratio of diapause to non-diapause eggs varies in different parts of Japan [25], and is a heritable trait [26]. Further, *A. menetriesi* embryos undergo a very short germ band mode of development [27], contrasting strongly with beetles such as *C. maculatus* (see below). As the last common ancestor of these two species is estimated to have existed only 80 million years ago [28], how their developmental mechanisms have diverged so greatly is a potentially fascinating area for future study. Research on these topics would greatly benefit from modern molecular studies.

*C. maculatus* (Fabricius) is native to West Africa [29] but is now found worldwide, and is a common pest of stored legumes. It is also known as the southern cowpea weevil, however, it is not a true weevil. As noted above, this beetle was the focus of active developmental research in the pre-molecular era, with special focus on segmentation [30, 31]. Segmentation in *C. maculatus* has also been studied more recently via immunohistochemistry for the even-skipped protein [32]. This recent work confirmed previous reports that the embryos undergo the long germ mode of development, similar to dipterans like *Drosophila melanogaster* and hymenopterans like *Apis mellifera* [33]. It is commonly believed that the long germ mode of development evolved independently in dipterans and hymenopterans, and given the phylogenetic distribution of short and intermediate germ development within the Coleoptera [34,
it seems likely that the long germ mode of development also evolved independently in the clade to which *C. maculatus* belongs. A comparison of the molecular basis of development in *C. maculatus* and other long germ insects, plus with more closely related species that feature short germ development, such as well studied flour beetle, *Tribolium castaneum* (super-family Tenebrionoidea, see Fig. 1 A) and *A. menetriesi*, would yield crucial information on how developmental pathways have evolved to generate the long germ mode of development. *C. maculatus* has also been studied in other fields and is a useful system for undergraduate lab teaching [17].

In order to facilitate research on *A. menetriesi* and *C. maculatus*, as well as wider investigations in the Coleoptera and beyond, we present here deep, multi-stage transcriptomic resources from a range of key time points in the development of these two beetle species. Using RSEM-based methods we have compared transcript abundance across these life stages, which will allow the investigation of genes that play key roles in developmental changes in these species, particularly at the maternal/zygotic transition. We have carried out extensive searches for key genes in developmental patterning and cell signalling pathways, and from our analyses we conclude that the transcriptomes for both species are of very high coverage, with almost all expected genes being present with full open reading frames. We have already made extensive use of these resources for our own studies on the embryonic development of these two beetles, and are confident that they will be of broad utility to a range of fields in genomics and developmental biology.

**Results and Discussion**

**RNA Extraction and Sample Selection**

Ovaries and embryos were collected as described in Materials and Methods, and as seen in Fig 2. The chosen time-windows cover a variety of important stages in the development of these
species, and can be expected to contain the majority of embryonically active transcripts in their expressed complement. Briefly, RNA was collected from dissected ovaries from each species and from four embryonic time windows for *A. menetriesi* and two embryonic time windows for *C. maculatus*. Samples were sequenced on the Illumina HiSeq platform (one lane per species), adaptor trimming was performed, along with preliminary assessment of read quality, and data was made available for download from an external server.

**Read Quality and Assembly Metrics:**

Table 1 summarises initial read information, and these reads are available from the NCBI SRA, Bioproject Accession numbers: PRJNA293391 and PRJNA293393. To confirm quality of read data, FastQC [36] was run on all read data. This showed Phred quality scores were high, with median scores always exceeding 28 through to the 101st base, and generally in the mid-30s. A slight bias was found in initial nucleotide sequence. To ensure that Illumina adaptors were removed in their entirety, Trimmomatic v.0.32 [37] was used to check for residual adaptor sequence, but none were observed. We therefore posit that this bias is due to known biases in Illumina hexamer binding [38] rather than sequencing-based artifacts.

Using Trinity [39] as an assembler, Trimmomatic-corrected read assemblies were then compared with assemblies based on uncorrected reads. Initial assays of Trimmomatic-treated read assemblies empirically found them to be less well assembled than those using un-trimmed reads, with a shorter N50 and less overall sequence recovery. Trimmed assemblies were therefore discarded in favour of untrimmed assemblies, which were used for all further analyses.

Assays of initial assemblies noted a small amount of fungal contamination in the data for both species. DeconSeq [40] was therefore run on the Trinity output for both species, with high stringency as noted in the methods. A total of 336 and 206 contigs with some homology to fungal sequence was removed from the *A. menetriesi* and *C. maculatus* total assemblies as a result, before read mapping was performed. In the *A. menetriesi* assembly, we observed
particularly high contamination with the protists, notably *Dictyostelium discoideum* and *Naegleria gruberi*, and some of this almost certainly remains in the transcriptome. Metrics for final assemblies after the removal of contamination can be seen in Table 2, alongside the results of assemblies for individual time points.

The final, combined timepoint, contamination-removed assemblies, comprising 228,096 and 128,837 contigs for *A. menetriesi* and *C. maculatus* respectively, contain a large number of well-assembled transcripts, with the number of contigs greater than 1 kb in length (52,217 in *A. menetriesi*, 33,250 in *C. maculatus*) and a high N50 (1,598 bp *A. menetriesi*, 2,263 bp *C. maculatus*) indicating a very well assembled dataset. This size is sufficient to span most protein coding domains, allowing easy inference of homology, and will be the full length of many transcripts. It is important to note that many of the contigs in our assembly will represent splice variants of single genes, and some genes will have multiple splice variants, which will affect these statistics. However, excellent recovery of splicing variation itself will be useful for a range of later analysis. GC content of the final assembled transcriptomes closely mirrors that of reads (32.82% in *A. menetriesi*, 38.7% in *C. maculatus*; reads 36-38%/40-43% respectively depending on library).

To gain an understanding of whether full-length transcripts were present in our data, we ran TransDecoder v 2.01 [41] to identify open reading frames (ORFs) and filtered for the results that were at least 100 amino acids long. For the *A. menetriesi* assembly, this analysis yielded 71,961 raw and 51,912 filtered contigs, while for the *C. maculatus* assembly, the analysis yielded 65,433 raw and 36,535 filtered contigs. The mean average length of the predicted polypeptides, after the filtering step, was 351 (*A. menetriesi*) and 448 (*C. maculatus*) amino acids. This is long enough for us to be confident that our assembly adequately spans coding regions, as the average eukaryotic protein length is 361 amino acids [42]. Together, these analyses suggest that we have recovered the vast majority of coding sequence in our combined assemblies, with sufficient length to adequately span ORFs, a conclusion further supported by
gene annotation data as discussed further below. The numbers of ORFs presented here are considerably more than the 16,404 gene models observed in the *T. castaneum* genome (Tribolium Genome Sequencing Consortium, 2008), and this is likely due to both spurious ORFs in our dataset and to multiple splice variants. All assemblies are available from Figshare online (A. menetriesi DOI: 10.6084/m9.figshare.2056464, C. maculatus DOI: 10.6084/m9.figshare.2056467).

**Timed RNA Expression - Differential Expression Analysis**

Key developmental stages for both *A. menetriesi* and *C. maculatus* can be easily observed following fixation and nuclei staining (data not shown). Briefly, egg lay to uniform blastoderm stage takes approximately 24 hours in *A. menetriesi*, and 7 hours in *C. maculatus*. Germband formation, gastrulation and elongation occur from 24-100 hours in *A. menetriesi*, and from 7-24 hours in *C. maculatus*. The latter period was subdivided in *A. menetriesi* according to characteristic stages of short germ development pertinent to our research interests. These stages, along with assayed sample periods, are shown diagramatically in Fig 2. As well as being used for combined assemblies as described earlier, RNA extracted from mature ovaries and from embryos collected during the aforementioned time periods was also individually assembled using Trinity.

The timed transcriptome assemblies for each species often possess better assembly quality when compared to the combined assemblies by metrics such as N50 and mean contig length (Fig 2, Table 2). As a result of being made up of a subset of the total reads, the individual assemblies do not possess the breadth of the combined assemblies, with fewer contigs, especially at long contig length (e.g. 1kb <). As such, the staged transcriptomes were used solely for comparison of expression levels across time, while the combined assemblies were used for gene family analyses. We emphasise that no technical replicate was performed for these comparisons, and any conclusions drawn from them should hold this consideration in
With this limitation in mind, we carried out differential expression analyses and observed broad trends in expression, as can be seen in Fig 3. First, we generated matrices from general comparison between time points in order to find the most similar samples (Fig 3 A,B). Generally, these results are congruent with steady changes in gene expression across the course of development, with most time points being most similar to those immediately preceding and following them. However, a split can be seen in *A. menetriesi* (Fig 3 A,B) between the ovary and 0-24 hour samples and all others, with the three later samples resembling each other more closely than the 0-24 hour dataset resembles the 24-48 hour sample. This could be due to the maternal:zygotic transition, which will occur at some point in this time frame. This cannot be seen for *C. maculatus*, and could be due to more admixture of RNA within the last 7-24 hour sample. Focused analysis on when the maternal:zygotic transition occurs in each species is required to resolve this question. Next, we clustered the results of our differential expression calculations (Fig 3 C,D). The results shown are those with RSEM considering each isoform separately (rather than taking into account clustering into genes performed by Trinity). Numerous up- and down-regulated contigs can be seen at each time point, with some time points more obviously possessing or lacking a subset of genes found in the combined transcriptomes. While replicates have not been performed and we have not analysed up and down regulated transcripts in detail, these data are available to download from Additional Files 6 and 7 attached to this document online. These data will act as a good initial guide for those interested in tracking differential expression of specific genes across development and will be useful for hypothesis building.

**Basic Gene Annotation**

To gain an understanding of the depth of coverage of our datasets we used the BUSCO library of well-annotated genes [43], which are known to be highly conserved in single copy across the
Metazoa, as a basis for comparison with our combined, assembled transcriptomes. Of the
BUSCO set of 843 metazoan orthologs, the A. menetriesi assembly possesses 801 (95%)
complete (of which 225, 26%, are duplicated), 16 fragmented (1.8%) and 26 missing genes
(3.0%), for a total recovery of 97% of the BUSCO dataset. The C. maculatus assembly contains
815 (96%) complete (with 202, 23%, duplicated), 13 (1.5%) fragmented and 15 (1.7%) missing
genes (98.3% recovery). This extremely high level of recovery gives us confidence that at least
all housekeeping genes expected to be present in these species are found in our datasets,
which strongly suggests that these transcriptomes contain the vast majority of the gene cassette
of these species.

The number of duplicate genes in our BUSCO analyses likely reflects the construction of
our transcriptomes from mixed RNA samples, with the allelic variation that this implies.
Discerning true duplicates from allelic and splice variant data is largely contingent on the
availability of well-assembled genomes. However, the recovery of these putative duplicates in
our assemblies underlines that our RNA sequencing and assembly was of good depth and
quality (respectively). With this information in-hand, a range of investigations will be made
possible, particularly into the regulation and expression of developmentally important genes.

A further understanding of the content of our assemblies was gained from Blast2GO
analysis [44, 45]. Genes were annotated using b2gpipe, on the basis of BLASTx (E value cutoff,
$10^{-3}$) comparisons made against the nr database as downloaded on the 26 January 2015. Of
228,096 (A. menetriesi) and 128,837 (C. maculatus) contigs in each assembly, 78,879 (34.6%)
and 41,744 (32.4%) possessed a hit in the nr databases above the threshold. After further
annotation with ANNEX and Interproscan, a total of 36,315 (15.9%) and 13,096 (10.2%) contigs
were assigned to one or more GO categories. These numbers, while only a fraction of the total
number of contigs within our transcriptomes, more closely reflect the expected eukaryotic
protein complement in number. Blast2GO annotations are given in Additional Files 2 and 3.
Fig 4A shows the distribution of species best hit by BLASTx comparison of contigs from A. mentriesi and C. maculatus with the nr database. For both species, T. castaneum is the best represented species - a reflection of both the phylogenetic position of this species and its well annotated genome. The fact that contigs in our transcriptomes match T. castaneum and other coleopteran and insect species more closely than those of other species suggests that gene orthology will be easy to assign in many cases, and that a higher than normal rate of molecular evolution is not observed in our species.

The distribution of GO terms within our datasets are shown in Fig 4B, alongside those of the well-annotated D. melanogaster and Mus musculus proteomes. In general, our datasets resemble one another more than they mirror that of the two sequenced genomes noted. Our transcriptomic resources empirically seem under-represented relative to D. melanogaster and M. musculus in developmentally interesting categories such as ‘Protein Binding’ (Molecular Function), ‘Multicellular Organism Development’ and ‘Cell Differentiation’ (Biological Process). Given our results for more targeted investigations, found below, we feel this is likely a result of poor annotation of these by Blast2GO, rather than true absence from the transcriptome.

At a gross level, in the intracellular ‘Cellular Component’ GO categories, our data appears more similar to that seen in the two ‘model’ species than in Molecular Function or Biological Process categories, although this has not been tested statistically. This would suggest these well-conserved structural components were more readily assigned GO categories than the developmentally interesting categories listed above. While not all GO categories are as well-annotated by this process as may be desired, the broad classification of our data into a wide range of GO categories of all levels of GO distribution demonstrates that Blast2GO annotations of our data are a useful starting point for more focussed investigations and identification of specific genes and pathways of interest.

Gene Family Recovery
To extend the semi-automated analyses presented above, we performed more targeted analysis of individual, developmentally important gene families. Both the Hox family of transcription factors and the TGF-$\beta$ cassette were exceptionally well recovered in our dataset. The Hox genes, and in particular the ANTP HOXL class, which pattern the anterior-posterior body axis, are recovered almost in their entirety in both species when compared to well-catalogued databases (e.g. [46]). This can be seen in Fig 5, which shows the phylogenetic distribution of recovered ANTP HOXL class sequences from our transcriptomes alongside previously annotated members of this class.

In *A. menetriesi*, sequences for all the ANTP HOXL class genes are recovered in our transcriptome, as can be seen in Fig 5, with sequences and alignments in Additional File 1. We note, however, that the *A. menetriesi* Hox 2 / maxillopaedia (mxp)/ proboscipedia (pb) sequence bears some BLAST similarity with Hox 4 / Dfd sequences, while the putative Hox 4 / Deformed (Dfd) recovered does not include the homeodomain sequence - whether it has lost this crucial domain, or if this portion of the sequence is simply not recovered in our assembly is at present unknown. The putative *A. menetriesi* Hox 4 / Dfd sequence is given in Additional File 1, although it is not shown in the phylogeny in Fig 5 due to its truncated length.

While 2 (*A. menetriesi*) and 4 (*C. maculatus*) Hox 3 / zerknüllt (zen) variants are seen in our species, these are identical at the coding level, and therefore seem to be splice or allelic variants, rather than the two paralogous genes seen in *T. castaneum* [47]. Similarly, no evidence of the caudal (cad) duplication seen in *T. castaneum* [48] can be found in our transcriptomic assemblies, suggesting this is perhaps specific to the flour beetle lineage. Both *zen* and *caudal* are important embryonic patterning genes, and comparison of these genes in our two species and *Tribolium* would be an excellent situation in which to study how sub- and neo-functionalisation occurs. Likely allelic or splice variants are also observed for other HOXL genes in both *A. menetriesi* and *C. maculatus*. It should be noted that these could represent very recent duplications, or the effect of gene conversion, although this can only be tested fully
with the advent of a complete genomic resource. No Pdx/Xlox gene was seen, adding further circumstantial evidence for the broad scale loss of this gene across the Arthropoda. Our C. maculatus transcriptome also contains the full complement of ANTP HOXL genes, although, similar to the case of Hox 4 / Dfd in A. menetriesi, the 4 recovered abdominal-A (abd-A) homologs lack the whole homeodomain sequence, with several residues missing. These truncations excluded them from the phylogeny shown in Fig 5, but the sequences for these putative homologs are given in Additional File 1.

Even more so than in A. menetriesi, a remarkable diversity of potential splice/allelic variants are noted in C. maculatus, particularly for the Hox 6/8 superfamily and Hox 9-13 / Abdominal-B (Abd-B) gene family. Of the Hox 6/8 gene superfamily, normally represented by four genes in the beetle (prothoraxless (ptl), fushi tarazu (ftz), Ultrathorax (Utx) and abd-A), 12 ptl, 1 ftz, 4 Utx and 4 abd-A representatives were found in our analysis. Furthermore, up to 26 different potential allelic or splice variants of abdB are recorded. As our transcriptome is made of mixed embryonic samples, it is perhaps not surprising that a diversity of splice/allelic variants are observed, but the excellent recovery of this data confirms the deep coverage provided by our sequencing and assembly.

The TGF-β cassettes of the insects have been very well described previously (e.g. [49, 50]). Our datasets recover almost the full expected complement of the Coleoptera. A slight exception to this is Activin (Act), a partial sequence of which is recovered for both species: a portion of the propeptide which does not span the mature signal peptide sequence. Whether this is a consequence of loss of the mature domain in these species or low levels of expression at the sampled timepoints remains to be established. A BMPx ortholog of clear homology to genes of that family can be found in C. maculatus, but has been excluded from the tree seen in Fig 6 as it is incomplete in length. Its sequence can be found in Additional File 1, and we have no doubt as to its identity due to high levels of sequence conservation between it and the T. castaneum and A. menetriesi orthologs of this gene.
The glass bottom boat (gbb) duplication observed in *T. castaneum* cannot be found in our data, but we can recover a range of splice or allelic variants for other genes, especially in *A. menetriesi*. These do not differ in the protein coding regions, which leads us to suspect that these are not from gene duplications (unless the duplication(s) occurred very recently). The phylogeny shown in Fig 6 confirms the homology of all genes, and splice/allelic variant numbers observed are given there in brackets, with all sequences available in Additional File 1.

We also note the discovery of an additional TGF-β ligand in *C. maculatus*. This gene has been previously automatically annotated as derriere in *T. castaneum*, (XP_008191586.1) and if it is truly of this family, which is also known as GDF1/3/Univin/Vg1, this would be a surprise, as its presence outside the Deuterostomia is controversial [51]. If proof could be found for this being a bona fide GDF1/3/Univin/Vg1, the presence of this gene in more than one coleopteran could suggest that this might in fact be ancestrally present in all bilaterian species, but further investigation is warranted before strong conclusions can be drawn in this regard. We could gain no phylogenetic support for placing either of these beetle sequences in the GDF1/3 clade, and it may well be that these sequences instead represent a coleopteran novelty.

Our recovery of not only the full expected complements of these vital developmental genes, but also a remarkable diversity of alternative variants, demonstrates the depth of our assemblies as a resource. Whether used as the basis for simple cloning or more sophisticated analysis of patterns of gene variation and diversification, these transcriptomes will be of wide utility to the field of coleopteran and insect developmental biology.

Pathway Recovery

As well as examining specific gene families, we investigated a number of broader pathways commonly studied in insects [52]. This allows us to both note how well-recovered such pathways are in our species as a measure of transcriptome utility, as well as note interesting differences between these pathways in our species when compared to others. We did this using
both automated (KEGG KAAS mapping) and manual (BLAST based) methods. Some representative results of KEGG KAAS mapping can be seen in Fig 7, and all KEGG annotations can be downloaded from Additional Files 4 and 5.

KEGG KAAS mapping uses BLAST results to annotate known pathways, and gives a rapid overview of the recovery of these. Here we have shown the well-known Wnt, Notch and Hedgehog pathways to indicate the depth of our transcriptomes, and show how they may be useful for future research. However, these maps often use terminology based on vertebrate nomenclature, and contain genes known to be absent from particular clades. We have therefore indicated in Fig 7 (using unshaded boxes as shown in the Key) genes that may be absent ancestrally in the Coleoptera, based on their absence from the T. castaneum pathway. Of genes expected to be present in the Coleoptera we find almost total recovery in our transcriptomes. In the three pathways examined, only three genes noted to be present in T. castaneum were noted as absent from both of our transcriptome datasets, all in the Wnt cascades (Fig 7A). The expected Hedgehog cassette was recovered in toto (Fig 7B) and in the Notch signalling cascade (Fig 7C), only APH-1 was noted to be absent, and only from the C. maculatus transcriptome. We must note that these may not be true absences - KEGG KAAS mapping is based on automatic BLAST assignation, and if these sequences are divergent in our transcriptomes they may have been missed by this analysis.

We also examined pathways manually, using reciprocal BLAST hits and closer manual investigation to confirm identity of individual genes, the results of which can be seen in Table 3. The anterior-posterior patterning genes cad (mentioned earlier) and hunchback (hb) are present in both species. Of the germline establishment and localization genes examined, nanos was surprisingly absent in both species, while bruno (bru; also known as arrest), exuperantia (exu), tudor (tud; 2 copies in A. menetriesi), oskar (osk), vasa (vas) and valois (vls) were present. Interestingly, pumilio (pum) is present in C. maculatus in single copy (although it is divided across two contigs), while A. menetriesi possesses a total of seven copies. The different A.
menetriesi pum copies varied both at the nucleotide level and in their amino acid sequences, strongly suggesting that they are in fact paralogs. In depth analysis of these genes is required to uncover why they have undergone several rounds of duplication. Orthologs of the Drosophila gene swallow (swa) could not be found in either of our transcriptome resources, nor is it present in several other insects (data not shown) and we suggest it may therefore be a schizophoran novelty.

Canonical gap genes Krüppel (Kr), knirps (kni), giant (gt), huckebein (hkb), tailless (tll; 2 paralogs in C. maculatus), buttonhead (btd), empty spiracles (ems) and both orthodenticle orthologs (Otd and Otd2) were recovered in both species examined here. The C. maculatus specific duplication of tll would be another excellent opportunity to study gene duplication and evolution. The pair rule genes even skipped (eve), hairy (h), fushi tarazu (ftz), odd paired (opa), odd skipped (odd), paired (prd), runt (run) and sloppy paired 1 (slp1) were present in single copy. The segment polarity gene cassette was also present in both species, with the notable absence of gooseberry-neuro (gsb-n) from our datasets. The genes armadillo (arm), cubitis interruptus (ci), Engrailed (En), fused (fu), gooseberry (gsb), hedgehog (hh), pangolin (pan), patched (ptc) and wingless (wg) were all present, and their sequences can be found in Additional File 1.

All of these pathways are commonly studied in insects, and the annotations provided here, along with preliminary timed expression data, will provide a basis for a wide range of targeted investigations into the embryonic development of these two species, and how these pathways have changed over the course of evolution. Furthermore, the excellent recovery of these pathways by both automated (KEGG-KAAS) and manual annotation gives us high confidence in the completeness of our transcriptomic resources. This confirms the results of our BUSCO analysis, and our datasets are therefore likely to contain the vast majority of transcribed genes in these two species, with only lowly expressed and temporally restricted genes absent from these transcriptome resources.
Conclusions:

Our production of deep transcriptomic sequence data for *A. menetriesi* and *C. maculatus* will assist in the inference of character gain and loss across the Coleoptera, aid in future phylogenetic efforts, and allow a range of investigations into the embryonic development of these species at the molecular level. The status of these organisms as common agricultural pests also suggests that such resources may allow targeted control mechanisms to be developed for these species. This data will be another key building block in our understanding of the transcriptomic basis to embryological development, and provide a window into the basic biology of the most successful clade of animals.

Methods:

Animal Husbandry

*A. menetriesi* eggs were kindly provided by Dr Yoshikazu Ando, and were reared at 25°C on wet sand or soil and fed fresh lettuce. *C. maculatus* beetles were kindly provided by Dr Joel Savard, and were reared at 30°C on dry black-eyed peas.

RNA Extraction and Sequencing

RNA was extracted from ovary and timed embryonic samples using a TRIzol RNA extraction kit according to the manufacturer’s protocols. RNA quantity and quality was checked using a Thermo Scientific Nanodrop 2000C Spectrophotometer and 1 µg was sent for sequencing by the Cologne Centre for Genomics. Adaptor trimming and initial quality control was performed by the commercial provider according to their internal standards. This cleaned data was then made available to us for download from an external server. Paired end read quality after sequencing was assessed using the FastQC program [36].
Transcriptome Assembly and Comparative Expression Analyses

Assemblies used in our final analysis were made using Trinity version 2013_08_14 [39], with the default settings. Full assemblies were made using reads from all time points, and individual assemblies were then constructed using each sampled time point individually. All assemblies are available from Figshare online (A. menetriesi DOI: 10.6084/m9.figshare.2056464, C. maculatus DOI: 10.6084/m9.figshare.2056467). DeconSeq standalone version 0.4.3 [40] was run on full assemblies with settings -i 95 -c 95, using the bact, fungi, hsref, and prot databases. Comparative expression analyses were performed using RSEM [53] as packaged in the Trinity module, and results shown here are the ‘as-isof orm’ data, although ‘as-gene’ data is provided in Additional Files. BUSCO v1.1b1 [43] was used to assess gene complement completeness.

Functional Annotation

Our combined assemblies were automatically assigned homologs and annotated according to gene ontology (GO) terms using Blast2GO [44, 45]. Initially, BLASTx was run using BLAST 2.2.29+ against the NCBI nr database as downloaded to a local server on the 17/01/2015, with settings -evalue 0.001 -max_target_seqs 5 -outfmt 5. GO term distribution within the D. melanogaster and H. sapiens genomes were downloaded from B2GO-FAR [54] and calculated using the Combined Graph function of Blast2GO. KEGG KAAS mapping was automatically performed using the KEGG KAAS tool (http://www.genome.jp/tools/kaas/), single-directional best hit with default BLAST settings, and with the eukaryote dataset as a basis for annotation.

Gene Identification

Gene sequences were manually identified and their homology confirmed by independently using tBLASTn [55] searches using gene sequences of known homology downloaded from the NCBI nr database as queries against standalone databases created on a local server using BLAST
2.2.29+ or the CLC Main workbench. Genes putatively identified using this method were reciprocally BLASTed against the online NCBI nr database using BLASTx to confirm their identity. Where identity was uncertain, phylogenetic analysis was used to confirm identity.

**Phylogenetic Tree Construction**

Sequences were aligned using MAFFT 7 [56] unless otherwise stated under the L-INS-i strategy. Alignments were then saved and exported to MEGA 6, where regions of poor alignment were manually excluded and maximum likelihood phylogenetic trees were constructed using the LG model, 1000 bootstrap replicates as indicated, 4 gamma categories and invariant sites, and all other default prior settings [57].

**List of Abbreviations Used:**

- BLAST: Basic Local Alignment Search Tool
- BUSCO: Benchmarking Universal Single-Copy Orthologs
- GDF: Growth Differentiation Factor
- KEGG-KAAS: Kyoto Encyclopaedia of Genes and Genomes - Automatic Annotation Server
- MAFFT: Multiple Alignment using Fast Fourier Transform
- NCBI: National Centre for Biotechnology Information
- ORF: Open Reading Frame
- RNA: Ribonucleic Acid
- RSEM: RNA-Seq by Expectation-Maximization
- SRA: Short Read Archive
- TGF: Transforming Growth Factor

**Competing Interests:**

The authors declare that they have no competing interests.

**Authors' contributions:**
MAB, KC and JAL maintained animals, extracted RNA and arranged sequencing. MAB and NJK assembled transcriptomes, analysed raw data and wrote the manuscript. MAB, NJK, KC, JAL and SR designed experiments and contributed to gene family analyses presented in this manuscript. All authors read and approved the final manuscript.

Acknowledgements:
The authors would like to thank the members of their laboratories for their support and discussions. In addition, we thank Dr Y Ando for providing *A. menetriesi* eggs and advice on establishing cultures, Dr J Savard for providing *C. maculatus*, and Dr F Marletaz for help in running BUSCO analyses. The efforts of editors and reviewers in considering this manuscript are gratefully acknowledged. MAB was supported by a Humboldt Fellowship for Postdoctoral Researchers. KHC, SR and JAL were supported by the DFG Collaborative Research Center grant 680.

Availability of data and materials
The datasets supporting the conclusions of this article are available in the NCBI SRA repository [Bioproject Accession numbers: PRJNA293391 and PRJNA293393] and in the Figshare repository [DOIs: 10.6084/m9.figshare.2056464 and 10.6084/m9.figshare.2056467]. NOTE: DATA EMBARGOED UNTIL PUBLICATION

Bibliography:
3. 1KITE: *1000 Insect Transcriptome Evolution* [http://www.1kite.org/]


16. Meer JM Van Der: Optical clean and permanent whole mount preparation for phase-contrast microscopy of cuticular structures of insect larvae. *Dros Inf Serv* 1977, **52**.


22. Chi YH, Salzman R a, Balfe S, Ahn J-E, Sun W, Moon J, Yun D-J, Lee SY, Higgins TJ V,


29. Tran BMD, Credland PF: **Consequences of inbreeding for the cowpea seed beetle, Callosobruchus maculatus (F)(Coleoptera: Bruchidae).** *Biol J Linn Soc* 1995, **56**:483–503.


51. Kenny NJ, Namigai EKO, Dearden PK, Hui JHL, Grande C, Shimeld SM: The


**Figure legends**

*Figure 1*: A) Cladogram of Coleoptera simplified from that determined by Hunt et al. (2007) [13]. Red asterisk indicates the superfamily in which the Chrysomelidae are located. B) Cladogram of Chrysomelidae simplified from that determined by Gomez-Zurita et al. (2007) [28]. Black and grey asterisks indicate the sub-families in which *A. menetriesi* and *C. maculatus* are located (respectively). C) *A. menetriesi* and D) *C. maculatus* adults.

*Figure 2*: Summary of RNA sources, life stages sampled and sequencing results, with *A. menetriesi* data presented at left and *C. maculatus* data at right.
Figure 3: Overview of results of differential expression analysis performed by RSEM within the Trinity framework. A) and B) show the sample correlation matrix for *A. menetriesi* and *C. maculatus* respectively. C) and D) show relative expression of each differentially expressed contig, considered as isoforms, across time.

Figure 4: Blast2GO results A) Distribution of BLASTx best hits by species, showing metazoans only, for *A. menetriesi* in orange, *C. maculatus* in blue. B) Distribution of GO terms expressed as a percentage of annotated contigs which were assigned a term within each of the three (Molecular Function, Cellular Component, Biological Process) categories of GO ID.

Figure 5: Phylogenetic inter-relationships of ANTP HOXL class genes, as reconstructed by MEGA 6 using the LG+Freqs model with 4 gamma categories and invariant sites, based on a 59 amino acid alignment spanning the homeodomain. Numbers at base of nodes represent bootstrap percentages of 1000 replicates. Scale bar at base of phylogeny gives substitutions per site at given unit distance. Red underline indicates *A. menetriesi* and *C. maculatus* sequences, coloured boxes are used to delineate known gene families (and in the case of Hox 6/7/8, a superfamily).

Figure 6: Maximum Likelihood Phylogeny of TGF-β ligands, as determined using MEGA under the LG+Freqs model with 4 gamma categories and invariant sites, on the basis of a 72 amino acid alignment of mature peptide sequences. The given scale depicts the number of substitutions per site per unit length. Bootstrap percentages (of 1000 replicates) are given at base of nodes. *A. menetriesi* and *C. maculatus* sequences are underlined in red. Coloured boxes represent known gene families with representatives in our transcriptomic resources, while all gene families, including those not found in our datasets, are indicated at right.
Figure 7: KEGG style pathway maps showing recovery in our transcriptome resources of A) the Wnt signalling pathway in canonical and non-canonical contexts, B) the Hedgehog signalling pathway and C) the Notch signalling pathway. Coloration of genes indicates presence, absence or ancestral absence from the Coleoptera as detailed in the key, which also gives other information as noted.

Tables

Table 1: Basic read data metrics

<table>
<thead>
<tr>
<th>Library: (hpf= hours post fertilization)</th>
<th>A. menetriesi</th>
<th>C. maculatus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovary</td>
<td>44,695,199</td>
<td>67,026,913</td>
</tr>
<tr>
<td>0-24 hpf</td>
<td>56,139,626</td>
<td>61,060,357</td>
</tr>
<tr>
<td>24-48 hpf</td>
<td>39,911,618</td>
<td>55,273,927</td>
</tr>
<tr>
<td>48-72 hpf</td>
<td>46,217,614</td>
<td></td>
</tr>
<tr>
<td>72-100 hpf</td>
<td>45,213,862</td>
<td></td>
</tr>
<tr>
<td>Ovary</td>
<td>101</td>
<td>101</td>
</tr>
<tr>
<td>Read Length (bp)</td>
<td>101</td>
<td>101</td>
</tr>
<tr>
<td>Amount of Data (Gigabytes)</td>
<td>7.3</td>
<td>10.8</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>38</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>40</td>
</tr>
</tbody>
</table>
### Table 2: Assembly metrics for individual time point and combined transcriptomic resources

<table>
<thead>
<tr>
<th></th>
<th>A. menetriesi</th>
<th></th>
<th>C. maculatus</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ovary</td>
<td>0-24 hpf</td>
<td>24-48 hpf</td>
<td>48-72 hpf</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>100,084</td>
<td>106,548</td>
<td>120,497</td>
<td>140,868</td>
</tr>
<tr>
<td>Max contig length(bp)</td>
<td>20,759</td>
<td>17,765</td>
<td>18,398</td>
<td>17,591</td>
</tr>
<tr>
<td>Mean contig length(bp)</td>
<td>983.95</td>
<td>1027.36</td>
<td>924.27</td>
<td>879.03</td>
</tr>
<tr>
<td>Median contig length(bp)</td>
<td>471</td>
<td>490</td>
<td>453</td>
<td>426</td>
</tr>
<tr>
<td>N50 contig length(bp)</td>
<td>1,894</td>
<td>1,993</td>
<td>1,721</td>
<td>1,643</td>
</tr>
<tr>
<td># contigs in N50</td>
<td>14,196</td>
<td>15,200</td>
<td>17,511</td>
<td>19,814</td>
</tr>
<tr>
<td># contigs &gt; 1kb</td>
<td>28,255</td>
<td>31,944</td>
<td>32,307</td>
<td>34,520</td>
</tr>
<tr>
<td># bases in contigs &gt; 1kb</td>
<td>68,737,074</td>
<td>78,526,348</td>
<td>75,208,107</td>
<td>80,796,004</td>
</tr>
<tr>
<td>Pathway/Gene</td>
<td>A. menetriesi</td>
<td>C. maculatus</td>
<td>Pathway/Gene</td>
<td>A. menetriesi</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------</td>
<td>--------------</td>
<td>-------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Maternal Effect:</td>
<td></td>
<td></td>
<td>Pair rule:</td>
<td></td>
</tr>
<tr>
<td>caudal</td>
<td>present</td>
<td>present</td>
<td>even skipped</td>
<td>present</td>
</tr>
<tr>
<td>hunchback</td>
<td>present</td>
<td>present</td>
<td>fushi tarazu</td>
<td>present (see Hox)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>hairy</td>
<td>present</td>
</tr>
<tr>
<td>Germline:</td>
<td></td>
<td></td>
<td>odd paired</td>
<td>present</td>
</tr>
<tr>
<td>Bruno/arrest</td>
<td>present</td>
<td>present</td>
<td>odd skipped</td>
<td>present</td>
</tr>
<tr>
<td>exuperantia</td>
<td>present</td>
<td>present</td>
<td>paired</td>
<td>present</td>
</tr>
<tr>
<td>nanos</td>
<td>absent</td>
<td>absent</td>
<td>runt</td>
<td>present</td>
</tr>
<tr>
<td>oskar</td>
<td>present</td>
<td>present</td>
<td>sloppy paired 1</td>
<td>present</td>
</tr>
<tr>
<td>pumilio</td>
<td>present - 7 copies</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>swallow</td>
<td>absent</td>
<td>absent</td>
<td>Segment Polarity:</td>
<td></td>
</tr>
<tr>
<td>tudor</td>
<td>present - 2 copies</td>
<td>present</td>
<td>armadillo</td>
<td>present</td>
</tr>
<tr>
<td>valois</td>
<td>present</td>
<td>present</td>
<td>cubitis interruptus</td>
<td>present</td>
</tr>
<tr>
<td>vasa</td>
<td>present</td>
<td>present</td>
<td>Engrailed</td>
<td>present</td>
</tr>
<tr>
<td>Gene</td>
<td>A. menetriesi</td>
<td>C. maculatus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fused</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gooseberry</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buttonhead</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gooseberry-neuro</td>
<td>absent</td>
<td>absent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empty spiracles</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hedgehog</td>
<td>present</td>
<td>present (but on 3 contigs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Giant</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hucklebein</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patched</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knirps</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wingless</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krüppel</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orthodenticle 1</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orthodenticle 2</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tailless</td>
<td>present</td>
<td>present - 2 copies</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Additional Files:**

Additional File 1: Sequences of all genes referred to in text, and alignments used in phylogenetic analyses (.xls)

Additional File 2: Annotations of transcriptome (Blast2GO .annot file) A. menetriesi

Additional File 3: Annotations of transcriptome (Blast2GO .annot file) C. maculatus

Additional File 4: KEGG data, A. menetriesi

Additional File 5: KEGG data, C. maculatus

Additional File 6: Comparative expression data, A. menetriesi

Additional File 7: Comparative expression data, C. maculatus
Atrachya menetriesi

Sample 1: Mixed Ovarian  
44,695,199 paired end reads  
100,084 contigs  

Sample 2: 0-24 hpf  
56,139,626 paired end reads  
106,548 contigs  

Sample 3: 24-48 hpf  
39,911,618 paired end reads  
120,497 contigs  

Sample 4: 48-72 hpf  
46,217,614 paired end reads  
140,868 contigs  

Sample 5: 72-100 hpf  
45,213,862 paired end reads  
126,400 contigs  

Samples 1-5 Combined:  
232,177,919 paired end reads  
228,096 contigs

Callosobruchus maculatus

Sample 1: Mixed Ovarian  
67,026,913 paired end reads  
98,082 contigs  

Sample 2: 0-7 hpf  
61,060,357 paired end reads  
57,879 contigs  

Sample 3: 7-24 hpf  
55,273,927 paired end reads  
80,665 contigs  

Sample 3: 7-24 hpf  
55,273,927 paired end reads  
80,665 contigs

Samples 1-3 Combined:  
183,361,197 paired end reads  
128,837 contigs
**A)**

Blast2GO BLASTX Top Hit by Species

- Tribolium castaneum
- Dendroctonus ponderosae
- Acyrthosiphon pisum
- Stegodyphus mimosarum
- Bombyx mori
- Camponotus floridanus
- Hydra vulgaris
- Microplitis demolitor
- Harpegnathos saltator
- Danaus plexippus
- Zootermopsis nevadensis
- Nasonia vitripennis
- Aedes aegypti
- A. menetriesi
- C. maculatus
- Drosophila melanogaster
- Mus musculus
- Other

**B)**

Percentage of GO Category (BP, CC, or MF) with Given Annotation

- Catalytic Activity
- Signal Transducer Activity
- Receptor Activity
- G Protein Binding Activity
- Transporter Activity
- Enzyme Activity
- Oxidoreductase Activity
- Lyase Activity
- Channel Activity
- Extracellular Region
- Extracellular Space
- Ion Transmembrane Transporter Activity
- Cell
- Nucleus
- Cytoplasm
- Membrane
- Macromolecular Complex
- Transmembrane Transporter Activity
- Protein Activity
- Gene Regulatory Activity
- Oxidation-Reduction Process

**Molecular Function** | **Cellular Component** | **Biological Process**
A) Wnt Signalling:

- Canonical Pathway:
  - Frizzled
  - LRPs
  - GSK3β
  - beta-catenin
  - AKT
  - MAPK

- Hedgehog Signalling:
  - Dvl
  - Prickle
  - Gli

- Wnt5a/Plensin Pathway:
  - Frizzled
  - PLC
  - Ca2+

B) Hedgehog Signalling:

- Hedgehog
  - Gli
  - Ptc
  - Smo
  - Rag23
  - Ci

- Wnt
  - Fzd
  - Dvl
  - Axin
  - APC

C) Notch Signalling:

- Notch
  - CSL
  - HES1/5
  - Hairless
  - SMRT
  - MAML
  - CIR
  - Groucho

- Delta
  - PSEN
  - NCSTN

- other Pathway
  - MAPK Signalling

- Notch Signalling:
  - Intracellular Domain
  - Via MAPK/RAS

- Others
  - Nuclear Targets

Key:
- Activated
- Constitutive
- Inhibited
- Up-regulation
- Down-regulation
- State Change
- Disassociation
- Drug