GigaScience

Genome sequencing of the winged midge, Parochlus steinenii, from the Antarctic Peninsula

--Manuscript Draft--

Manuscript Number:	GIGA-D-16-00062R1
Full Title:	Genome sequencing of the winged midge, Parochlus steinenii, from the Antarctic Peninsula
Article Type:	Data Note
Funding Information:	Korea Polar Research Institute (PE16070) — Not applicable Korea Polar Research Institute (PE16080) — Not applicable
Abstract:	Background: In the Antarctic, only two species of Chironomidae occur naturally—the wingless midge, Belgica antarctica, and the winged midge, Parochlus steinenii. B. antarctica is an extremophile with unusual adaptations. The larvae of B. antarctica are desiccation- and freeze-tolerant and the adults are wingless. Recently, the compact genome of B. antarctica was reported and it is the first Antarctic eukaryote to be sequenced. Although P. steinenii occurs naturally in the Antarctic with B. antarctica, the larvae of P. steinenii are cold-tolerant but not freeze-tolerant and the adults are winged. Differences in adaptations in the Antarctic midges are interesting in terms of evolutionary processes within an extreme environment. Herein, we provide the genome of another Antarctic midge to help elucidate the evolution of these species. Results: The draft genome of P. steinenii had a total size of 138 Mbp, comprising 9513 contigs with an N50 contig size of 34,110 bp, and a GC content of 32.2 %. Overall, 13,468 genes were predicted using the MAKER annotation pipeline, and gene ontology classified 10,801 (80.2 %) predicted genes to a function. Compared with the assembled genome architecture of B. antarctica, that of P. steinenii was approximately 50 Mbp longer with 6.2-fold more repeat sequences, whereas gene regions were as similarly compact as in B. antarctica. Conclusions: We present an annotated draft genome of the Antarctic midge, P. steinenii. The genomes of P. steinenii and B. antarctica will aid in the elucidation of evolution in harsh environments and provide new resources for functional genomic analyses of the order Diptera.
Corresponding Author:	Seung Chul Shin, Ph.D KOREA, REPUBLIC OF
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Sanghee Kim
First Author Secondary Information:	
Order of Authors:	Sanghee Kim
	Mijin Oh
	Woongsic Jung
	Joonho Park
	Han-Gu Choi
	Seung Chul Shin, Ph.D
Order of Authors Secondary Information:	

Response to Reviewers:

Dear Editor:

We are pleased to have an opportunity to revise our manuscript, entitled, “Genome sequencing of the winged midge, Parochlus steinenii, from the Antarctic Peninsula". In revising the paper, we carefully considered your comments and suggestions, as well as those offered by the reviewers.
As instructed, we explained how we revised this manuscript based on the comments and recommendations. We greatly appreciate the time and effort put forth to provide us insightful guidance.
The revision includes a number of changes:
-We changed the background information in the abstract, as requested by reviewers.
-We added a more detailed description of methods and parameters used.
-We clarified portions of the methodology.
-We added the results of the BUSCO and CEGMA analyses.

In rebuttal letter, we offer detailed responses to your comments, as well as those of the reviewers.
We hope that our revisions improved the quality of this manuscript and thank you again for your consideration of our manuscript.

Sincerely,
Seung Chul Shin

Reviewer reports:

Reviewer #1: This paper describes the genome assembly and annotation of the winged midge, Parochlus steinenii. This species is of particular interest, as it co-occurs in Antarctica with another midge species, Belgica antarctica, but is cold-tolerant; comparative analyses of these two genomes may yield insights into the origins of freeze-tolerance in Belgica antarctica. The data generated and the analyses performed are useful, and should be valuable to the insect comparative genomics community. However, there are a number of uncertainties with the manuscript. Specifically, many details of the analyses are left out, which will make it difficult for others to 1) understand and 2) replicate the analyses performed. It is possible that these details are contained within the 'Supporting data' in the GigaDB database, but I do not have access to these records, and there are no mentions of the Supporting data within the manuscript, except under 'Availability of supporting data'.

Specific Comments:

Abstract - Background
l. 28: "with a compact genome as a result of adaptation to an extreme environment": As far as I know, there are no studies yet that have determined that B. antarctica's small genome size is a result of the extreme environment this insect lives in (the 2014 genome paper doesn't go that far). Please soften the language, or provide a citation that demonstrates a causal relationshiop.
> We agree with the reviewer’s comments and have rewritten the sentence as follows: “B. antarctica is an extremophile with unusual adaptations. The larvae of B. antarctica are desiccation- and freeze-tolerant and the adults are wingless.”

l. 31: change "are cold, but not freeze, tolerant" to "are cold- but not freeze-tolerant,"
> We have changed “are cold, but not freeze, tolerant” to “are cold-tolerant but not freeze-tolerant” in the revised manuscript.

Abstract - Conclusions
l. 44: Please change "cold, but not freeze, tolerant" to "cold- but not freeze- tolerant"
> We have rewritten the conclusions as follows:
“We present an annotated draft genome of the Antarctic midge, P. steinenii. The genomes of P. steinenii and B. antarctica will aid in the elucidation of evolution under harsh environments and provide new resources for functional genomic analyses of the order Diptera.”

Data description - Sequencing
l. 49. How many individuals? Did you determine the sex, or was it a mixed collection?

If there were too many to count, did you weigh them? What life history stage?
> We have added the description in the revised manuscript as follows:
“Twenty adults were used for genome sequencing, regardless of gender.”

l. 61: What life history stage? How many insects?
> We have changed “whole body of P. steinenii using the Qiagen kit” to “whole body of 10 adults in three different groups using the RNeasy mini kit (Qiagen, Valencia, CA, USA)”

l. 61: Which Qiagen kit was used?
> We have changed “Qiagen kit” to “RNeasy mini kit” in the manuscript.

l. 67: Which Fastx program was used?
> We have changed “using the FASTX-Toolkit” to “using the fastq_qulity_trimmer in the FASTX-Toolkit” in the revised manuscript.

l. 69: I don't understand what you mean by "data from paired-end trimmed reads with 14 gigabase pairs (Gbp) were obtained". I couldn't find a table legend that explained this, either.
Table 1:
- I couldn't find a table legend.
> We have changed “data from paired-end trimmed reads with 14 gigabase pairs (Gbp) were obtained (Table1)” to “yields after quality trimming for the fragment library totaled 14.8 gigabase pairs (Gbp).”

We have added a table legend as follows:
“Tree-type libraries were constructed in this study. A PE400 library was constructed as a fragment library for ALLPATHS-LG. Mate-pair (MP3K and MP5K) libraries were also constructed for ALLPATHS-LG assembly. Three PE300 libraries (PE300A, PE300B, and PE300C) were constructed from RNA for gene annotation.”

Additionally, we have added the library name from the table into the manuscript as follows:
“One was a fragment library, which was a paired-end type with an insert size of 400 bp (PE400), whereas others were jumping libraries, which were mate-pair types with insert sizes of 3 kbp (MP3K) and 5 kbp (MP5K).”
“Three paired-end libraries with an insert size of 300 bp (PE300) were constructed using the TruSeq Stranded mRNA Library Prep Kit (Illumina, San Diego, CA, USA)”

- the column 'Read lengths' doesn't make sense to me - is this the combined length of all reads?
> We have changed “Read lengths” to “Total read lengths”

Data description - Genome assembly
l. 78: "the fragment library should be designed to overlap": Do you mean that the reads from paired-end library overlapped, and were thus combined to generate one longer read?
> Yes. We have rewritten the sentence as follows:
“For better assembly in ALLPATHS-LG, a larger k-mer size was used with one longer read generated from the paired-end library [4]. As a result, the paired-end reads from the fragment library were designed to overlap, and the insert size of the paired-end library was slightly less than twice the read size [4]. In this assembly, 93.8% of the paired-end reads from the fragment library overlapped and merged into one longer read.”

l. 79: I don't understand what this means: "In this assembly, 93.8% of the fragment library was full".
> We have rewritten the sentence as follows:
“93.8% of paired-end reads from the fragment library overlapped and merged into one longer read.”

l. 81: "The resulting assembly had a total size of 137 Mb" - In table 2, you list 130.6 Mb for the contigs and 138 Mb for the scaffolds. Where do you get the number 137 Mb from?

> We have corrected 137 Mb with 138 Mb.

l. 83: How did you calculate the coverage?
> We estimated contig coverage by total read lengths from the fragment library, but we used the coverage in the assembly report from ALLPATHS-LG in the revised manuscript.
Thus, we have changed “revealed contig coverage of approximately 108.5 ×” to “revealed contig coverage of approximately 89 × total read lengths from the fragment library.”

Data description - Gene annotation
l. 93: "For proper gene annotation, RNA and protein evidence alignment were used".
- Which RNAs were used? At what step in the Maker program?
> For RNA evidence, we extracted total RNA from the whole body of adults, sequenced, and assembled the resulting reads into contigs. The resulting contigs were used for the MAKER2 annotation pipeline to find the best gene model using RNA evidence with the alignment results of proteins. We have clarified the RNA evidence in the revised manuscript as shown in the response to the next comment.

- You list the proteins that were used to train Maker in the next paragraph (l. 99-104). Please list them here, instead.
> We have rewritten the paragraph as follows:
“To find the best possible gene model for the given region, RNA and protein evidence alignment were considered in MAKER2 [17]. Transcriptome assembly results were used for RNA evidence, the paired-end reads resulting from mRNA of the whole body of adults were trimmed using the fastq_qulity_trimmer in the FASTX-Toolkit (Ver. 0.0.11) (http://hannonlab.cshl.edu/fastx_toolkit) with the parameters -t 30, -l 80, and -Q 33, and they were assembled with CLC Genomics Workbench (Ver. 8.0.0) with default parameters. In all, 68,392 contigs with an N50 contig size of 435 bp and an average contig size of 407 bp, were generated and used for RNA evidence. Protein sequences from six species, given in NCBI reference sequences, were used for protein evidence—Drosophila melanogaster (fruit fly, GCF_000001215.4), Ceratitis capitata (Mediterranean fruit fly, NC_000857.1), Bactrocera dorsalis (oriental fruit fly, NC_008748.1), Anopheles gambiae (African malaria mosquito, NZ_AAAB00000000.1), Aedes aegypti (yellow fever mosquito, AAGE00000000.2), and Culex quinquefasciatus (southern house mosquito, AAWU01000000). Alignment of transcriptome assembly with BLASTn and alignment of homologous protein information from tBLASTx were considered as evidence for annotation.”

l. 94: What ESTs were used?
> We used the transcriptome assembly as RNA evidence instead of EST. Thus, we have changed “expressed sequencing tags” to “transcriptome assembly” in the revised manuscript.

l. 96: What transcriptome assembly? Does this line describe the assembly of the RNA data that were generated? Were reads trimmed prior to the assembly? This needs more detail.
> To clarify the transcriptome assembly in the manuscript, we have added the details for transcriptome assembly as follows
“To find the best possible gene model for the given region, RNA and protein evidence alignment were considered in MAKER2 [17]. Transcriptome assembly results were used for RNA evidence, the paired-end reads resulting from mRNA of the whole body of adults were trimmed using the fastq_qulity_trimmer in the FASTX-Toolkit (Ver. 0.0.11) (http://hannonlab.cshl.edu/fastx_toolkit) with the parameters -t 30, -l 80, and -Q 33, and they were assembled with CLC Genomics Workbench (Ver. 8.0.0) with default parameters. In all, 68,392 contigs with an N50 contig size of 435 bp and an average contig size of 407 bp, were generated and used for RNA evidence.”

l. 111: "This was annotated with the BLASTp results and InterproScan [9]." I'm confused by this sentence - what is "This"? which BLASTp results - is this output from the Blast2Go program, or another analysis? Also, are the InterproScan results part of the Blast2Go analysis?
> We have changed “gene ontology (GO) classified 10,801 (80.2%) of the predicted genes to a function. This was annotated with the BLASTp results and InterproScan [9].”

with “gene ontology (GO) classified 10,801 (80.2%) of the predicted genes to a function using the BLASTp and InterproScan results [9].”

Data description - Gene annotation for B. antarctica
l. 125 - are the six other species used for protein evidence the same that are listed on l. 99-104?
> We have changed “We matched proteins from P. steinenii to those from six species for protein evidence.” to “We used the same protein sequence from the six species used for gene annotation in P. steinenii and predicted proteins of P. steinenii for protein evidence.”

l. 143: How did the methods for repeat analysis differ between this paper and the B. antarctica genome paper (cited in [5])?
> In the case of repeat analysis for B. antarctica, RepeatMasker and the T-lex2 de novo pipeline were used for repeat analysis. We only used RepeatMasker for repeat annotation. Thus, it might be improper to compare the results of repeat analysis directly.

Table 4 - This could probably be moved to a supplement, or condensed.
> As suggested by reviewer, we have made Table 4 a Supplementary Table.

Data description - Ortholog analysis
Table 5 - This should be moved to a supplement, or condensed. Also, is there supporting data that lists what genes are in which group? It is interesting that D. melanogaster has so many unique proteins, compared to the other 5 species. Can you speculate why?
> We have made Table 5 a Supplementary Table and added the number of unique orthologous groups of six species to Figure 1a. D. melanogaster belongs to the suborder Brachycera and the other five species belong to a different suborder, Nematocera, in Diptera. This might be the reason why D. melanogaster showed so many unique orthologous groups.

Data description - Gene structure of Orthologous groups
l. 174 - Are you using the genome size for B. antarctica that was calculated by flow cytometry for this comparison? Perhaps you should use the range calculated from genome sequencing, instead, since this is how you are estimating the P. steinenii genome size.
> We used the assembled genome size for B. antarctica and for P. steinenii. Thus, we have changed “Despite 39 Mbp difference in genome size ……” to “Despite approximately 50 Mbp difference in the assembled genome size between B. antarctica and P. steinenii, ……’.

Data description - GO enrichment test
l. 180: "statistically represented" - do you mean over-represented?
> We have changed from “represented” to “overrepresented.”

l. 180: "437 orthologous groups" - please change to "437 orthologous groups that are unique to P. steinenii"
> We have changed from "437 orthologous groups" to "437 groups that were unique to P. steinenii."

l. 181: AgriGO has several analysis tools. Which one did you use, with what parameters?
> We have rewritten the sentence to clarify the methods as follows:
“AgriGO is a web-based tool for GO analysis, we selected “Fisher’s exact test” for the statistical test method and selected “Hochberg FDR” as the multiple test adjustment method. GO terms were tested with a significance level of p < 0.05.”

l. 182: "significant levels of p = 0.05" - do you mean "significance levels of p < 0.05"?
> We have corrected “=” with “<.”

l. 185: Can you 1) explain what an unfolded protein response is, and how this may be biologically interesting for P. steinenii, and 2) elaborate on why you think an enrichment for genes associated with 'unfolded protein response under stress

conditions' in the orthomcl groups that are unique to P. steinenii implies that they evolved independently? Finally, why did you only single out these categories, when many more were enriched for genes in orthomcl groups unique to P. steinenii?
> We have added the description to the unfolded protein response. Because 14 GO terms among 26 GO term were associated with the UPR in the GO hierarchy, we singled out these categories as representative. It is hard to explain how they evolved independently. We have rewritten the section as follows:
“It is noteworthy that 14 GO terms among 26 GO terms in biological processes were associated with the unfolded protein response (UPR). The UPR is a stress response that occurs in the lumen of the endoplasmic reticulum (ER) [22]. When unfolded or misfolded proteins were accumulated in the ER lumen under stress conditions, the UPR is activated to improve protein folding by increasing the production of chaperones [22].”

Data description - Likelihood analysis of gene gain and loss
l. 198 - The URL doesn't work.
> We have deleted the URL and kept only a reference for café3.0.

l. –99 - How did you generate the tree? From what datasets?
> To clarify the methods used to generate the tree, we have added a section to the manuscript as follows:
“we estimated differences in the size of 15,633 orthologs using the program CAFE3.0 [25]. A Newick description of a rooted and bifurcating phylogenetic tree was needed for this analysis. Therefore, we performed phylogenetic analyses among six insects with the protein-coding gene in the orthologous groups. We selected 4,814 orthologous gene sets from the orthologous groups from OrthoMCL using the reciprocal best BLASTP hit criteria. Protein-coding gene sequences were aligned using PRANK (Ver. 130820) under a codon model with the “-DNA and –codon” option [26], poor alignment sites were eliminated using Gblock (Ver. 0.91) under a codon model with the “-t = c” option [27], and the remaining alignment regions were concatenated to be used in the phylogenetic analyses. The phylogenetic tree was constructed using the neighbor-joining method [28] in the MEGA version 6 program [29]. With the resulting phylogenetic tree, we prepared the ultrametric tree of the species, including branch lengths in units of time through TimeTree [30], for the analysis (Figure 1C).”

l. 218 - Are you going to deposit the genome assembly in a public repository?
> If the genome assembly is suitable to be deposited in the GigaDB, we do not plan to deposit it in other public repositories.

General
- I couldn't find any Table legends. These are essential, as the tables by themselves are not descriptive enough.
> We have added legends to all tables.

- It is not clear what supporting data are in the GigaDB database - is there a way to make this obvious to the reader in the manuscript?
> We have added a better description for supporting data as follows:
“Supporting data (sequence files for CDS, protein, transcript, and the draft genome, and the general feature format for genes and repeats) are available in the GigaDB database, and the raw data were deposited in the PRJNA284858 (SRX1976250–5).”

Reviewer #2: A few comments:
- Table 2: Even though scaffolding greatly improved the assembly there is still a great number of scaffolds (>4,000) and a relatively low scaffold N50, compared to the genome size of this midge (~138 Mbp). In addition, I would say that this unexpected given the amount of sequencing data generated for this insect, which resulted in >100x average contig coverage. I would suggest that the authors comment on it and mention some probable causes for this (e.g. increased repeat content, increased heterozygosity, no mate-pair libraries with an insert of >5 Kbp?).
> Approximately 89 × total reads lengths from the fragment library was used and two mate-pair libraries were used in this assembly. A total of 57.2% of the 3-kb jumping library and 33.1% of the 5-kb jumping library were used, and 9.6 kb of the N50 contig size was increased to 157kb of the N50 scaffold size. More jumping libraries and long jumping libraries might improve this assembly and another fragment library might

improve it by increasing the randomness of reads in the library.

We have added the sentences as follows:
“A total of 57.2% of the 3-kb jumping library and 33.1% of the 5-kb jumping library were used and 9.6 kb of the N50 contig size was increased to 157 kb of the N50 scaffold size. If more jumping libraries or long jumping libraries (the insert size was larger than 20 kbp) were used, the scaffolding might improve the assembly.”

- Lines 73-83: While the authors mention all the tools and parameters used for genome assembly, there is no mention about the tool they used for scaffolding. I think it would be nice to add this important information, especially since scaffolding contributes to a significant improvement of the assembly.
> Assembly was performed using ALLPATHS-LG. This assembler linked the contigs into scaffolds with two mate-pair libraries. A total of 57.2% of the 3kb jumping library and 33.1% of the 5kb jumping library were used. Thus, we did not perform additional scaffolding.

- Line 92: Why did you only use SNAP for gene prediction? Augustus is known to perform better and can be run from the MAKER pipeline.
> Augustus showed better performance than SNAP in ab initio predictions, but in the MAKER pipeline, SNAP and Augustus showed similar results in evidence-based annotation (Holt et al., 2011; MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects).

- The authors haven't performed an evaluation of their genome assembly or their predicted gene set. Such evaluations are usually done by tools such as BUSCO [Simao et al. 2015], that search for conserved genes in the assembly/gene set.
> To assess the annotated gene set and genome assembly, we ran BUSCO and CEGMA analyses and have added the results to the manuscript as follows:
“The assembled genome size was similar to the predicted genome size (143.8 Mb). We also validated this assembly using CEGMA [7] and BUSCO [8]. CEGMA evaluation showed that gene completeness of this assembly was 85.08% and BUSCO analysis using arthropod databases showed 67% completeness (Tables 3 and 4). If partially matched genes were considered, 92.34% and 89.6% of the genes were identified in CEGMA and BUSCO, respectively (Tables 3 and 4).”
“To assess the annotated gene set, we ran a BUSCO analysis with the “-m OGS” options for gene set completeness and identified 70.7% genes considered to be complete with the expanded gene set, and 16.5% of the gene set was classified as missed [8].”

- Table 5 is unnecessarily long and complicated. First of all, I think that not all the different combinations are necessary to show. I would only include the largest groups and also the most biologically important (certainly no more than 10 groups). I would also suggest that the authors find descriptive names of each group, such as "P. steinenii-specific", or Antarctic midge-specific, or mosquito-specific. I find group names such as "ABC", "ABCDF", "BCEF" to not be human-readable. Last, instead of showing numbers of orthologous groups it would be more meaningful to show number of genes (and maybe show how many of them are transcribed).
> As suggested by this reviewer, we have made Table 5 a Supplementary Table and added the number of unique groups in the six species to Figure 1a.

Some more, minor comments:

- Line 33: The sentence "In this study..." is isolated does not say much. I would suggest to either delete it, or develop it to something more informative.
The ms focused on "reproducibility of analyses",
while more on the biology of this midge would make the story more exciting.
> We have deleted the sentence “In this study….” on line 33.
We have rewritten this sentence as follows:
“Differences in adaptations in the Antarctic midges are interesting in terms of evolutionary processes under an extreme environment.”

A few comments:

- Table 1: the first two lines refer to libraries "PE400trim" and "PE400". It's not clear to me if these two libraries are different or the same (with PE400tim simply being the trimmed PE400 library). The authors should clarify this.

> We have removed the row “PE400trim” and added a legend to Table 1 as follows: “Tree-type libraries were constructed in this study. A PE400 library was constructed as a fragment library for ALLPATH-LG. Mate-pair (MP) libraries was also constructed for ALLPATH-LG assembly. Three PE300 libraries (PE300A, PE300B, and PE300C) were constructed from RNA for gene annotation.”

- Table 1: Is the column named "Read lengths" showing total read lengths? If so, please rename it.

> We have changed “read lengths” to “total read lengths.”

- Line 180: The authors refer to 437 orthologous groups but it is not clear what these are. Are they the P. steinenii-specific groups? If so, it should be clearly mentioned in the sentence to avoid confusion.

> We have changed "437 orthologous groups" to "437 groups specific in P. steinenii genes were not identified in any other species."

- Table 6: I'm not sure what the columns "number of target genes in term" and "number of genes in terms" mean. Is the former representing the number of genes with a GO term in the species-specific orthologous groups, while the latter represents the same number in the whole gene set? You should make the descriptions more clear.

> We have changed “number of target genes in term” to “the number of genes with GO terms in the P. steinenii-specific groups” and changed “number of gene in terms” to “the number of genes with GO terms in P. steinenii’s entire gene set.”

- Some typos:
  - line 90: "...using MAKER2..." --> "using the MAKER2..."

> We have changed "...using MAKER2..." to "using the MAKER..."

- lines 119-124: Please rephrase the sentence "For RNA evidence...". The first part of the sentence (up to "pipe lines") does not make sense.

> We have deleted the sentence “For RNA evidence…… ……pipe lines” and added the sentence “The resulting contigs were aligned to the genome sequence of B. antarctica with BLASTn in MAKER2 for RNA evidence” as follows:
> “The reads in various experimental conditions with B. antarctica (SRR566981, SRR567289, SRR567164~7, and SRR567169~71) were downloaded from SRA databases in NCBI and we assembled the reads into 38,017 contigs with an N50 contig size of 1,799 bp and an average contig size of 913 bp through CLC Genomics Workbench (Ver. 8.0.0). The resulting contigs were aligned to the genome sequence of B. antarctica with BLASTn in MAKER2 for RNA evidence.”

- line 168: "...of orthologous gene..." --> "of orthologous genes"

> We have corrected “gene” with “genes” in line 204.

- line 184: Is there something missing in "...and 18 GO terms were identified..."?

> We have added “in molecular functions” in line 223 as follows:
> “…18 GO terms in molecular functions were…”

- Table 6: in the title of the table delete "were".

> We have deleted “were”

- line 195: The sentence "The size..." is not informative at all. I suggest you merge it with the next one.

> We have rewritten the sentence as follows:
> “To estimate the average gene expansion/contraction rate and to identify gene families that have undergone significant size changes through evolution”

- lines 199-200: "We performed the program" --> "We ran the program".

> We have changed “performed” to “ran” in line 257

- line 205: delete "there"

> We have deleted the word “there.”

Reviewer #3: The article describes the genome assembly of the winged midge Parochlus steineii. There are comparisons to other insect genome assemblies. The genome assembly will provide a useful resource for comparative genome studies. My comments and concerns are listed below.

There are statements in the manuscript that are factually incorrect and must be remedied. The statement that B. antarctica "adults lose their wings" (line 29) is an inaccurate description of the adult wing status. The adults never have wings. Also, the genome assembly of B. antarctica proposes that the small genome is likely due to adaptation to cold environment, however, given that there is no comparative data or study of adaptation and genome size, there is not conclusive evidence that the small genome is itself adaptive.
> We have rewritten the sentence as follows:
“B. antarctica is an extremophile with unusual adaptations. The larvae of B. antarctica are desiccation- and freeze-tolerant and the adults are wingless.”

The concluding sentences of the manuscript state that the mechanisms of freeze tolerance are unknown in B. antarctica. The mechanisms are known and have been explored extensively (see 1979 paper http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3032.1979.tb00171.x/abstract and http://www.ncbi.nlm.nih.gov/pubmed/16424090).
> We have rewritten the sentence as follows:
“We present an annotated draft genome of the Antarctic midge, P. steinenii. The genomes of P. steinenii and B. antarctica will aid in the elucidation of evolution under harsh environments and provide new resources for functional genomic analyses of the order Diptera.”

How many individuals were used in the genome assembly?
> Twenty adults were used for genome sequencing, regardless of gender. Thus, we have added this sentence to the revised manuscript as follows:
“Twenty adults were used for genome sequencing, regardless of gender. Genomic DNA was extracted using a DNeasy Tissue Kit (Qiagen, Valencia, CA, USA).”

What is the k-mer analysis estimate of genome size and heterozygosity?
> We have estimated the genome size and heterozygosity using the kmer analysis and added it to the revised manuscript as follows:
“Before assembly, we estimated the genome size and heterozygosity using a kmer analysis with sequencing reads. The software Jellyfish (Ver 1.1.10) [5] and GenomeScope (http://qb.cshl.edu/genomescope/) [6] were used. The 17-mers were counted in the reads from the PE400 library and the resulting histogram of 17-mers occurrence was used as a query for GenomeScope [6]. The estimated genome size was 143.8 Mb and the estimated heterozygosity was 0.613%.”

Table 2: Later in the manuscript there is a re-analysis of the B. antarctica genome for direct comparison - that data should be presented in Table 2 instead of quoting from the previous paper.
> We have added the results from the re-analysis of the B. antarctica genome to Table 2.

An additional estimate of genome size - flow cytometry, k-mer analysis, etc - would allow the actual comparison of absolute genome size between the species. The sum of scaffolds in the B. antarctica paper was ~10Mb smaller than the flow cytometry estimated size.
> We have added the estimate genome size as an additional estimate to the manuscript and we have compared the gene structure based on the assembled genome size.

The methods for the ortology analysis are missing (lines 125-127).
> This ortholog analysis was the same analysis as in the next paragraph. Thus, we have removed for clarity.

The authors should run BUSCO to present results on the putative completeness of the genome assembly.

> To assess the annotated gene set and genome assembly, we ran BUSCO and CEGMA analysis and have added the results to the revised manuscript as follows:
“The assembled genome size was similar to the predicted genome size (143.8 Mb). We also validated this assembly using CEGMA [7] and BUSCO [8]. CEGMA evaluation showed that gene completeness of this assembly was 85.08% and BUSCO analysis using arthropod databases showed 67% completeness (Tables 3 and 4). If partially matched genes were considered, 92.34% and 89.6% of the genes were identified in CEGMA and BUSCO, respectively (Tables 3 and 4).”
“To assess the annotated gene set, we ran a BUSCO analysis with the “-m OGS” options for gene set completeness and identified 70.7% genes considered to be complete with the expanded gene set, and 16.5% of the gene set was classified as missed [8].”

Was the repeat content re-estimated for the B. antarctica assembly?
> The statistics of repeats of B. antarctica are quoted from a previously published paper (Kelley et al, 2014). Thus, we have replaced the data from a previously reported paper with the results of the repeat analysis for the B. antarctica from this study.

GO enrichment test (section starting on line 179) is missing crucial information that the 437 orthologous groups are those genes that are unique to P. steinenii. Please note that they are not orthologs if they are unique to P. steinenii. In this analysis, how many genes had no term? What is the correction for genes belonging to multiple GO levels in the GO enrichment analysis?
> We have added more description to the methods for the GO enrichment test and have made corrections as follows:
“We identified which GO terms of the 437 groups that were unique to P. steinenii were statistically overrepresented versus GO terms of all genes of P. steinenii using AgriGO [21]. A total of 1,352 genes comprised 437 groups and therein were 717 genes with GO terms. AgriGO is a web-based tool for GO analysis, we selected “Fisher’s exact test” for the statistical test method and selected “Hochberg FDR” as the multiple test adjustment method. GO terms were tested with a significance level of p < 0.05.”

What dataset was used for Timetree?
> TimeTree looks for two queried species in the searchable tree of life scaled by time and produces the time estimate available. We used the estimated divergence time from TimeTree.

What is the estimated divergence time of the two species?
> The estimated divergence time of the two species is 199 millions years and the estimated divergence time was added in the legend to Figure 1c.
“…Numbers on each branch denote the number of gained, lost, and stable genes. AE denotes the average expansion. The number below each corresponding node denotes the estimated divergence time based on TimeTree.”

Minor comments:
Table 1: why are both the pre-and post-trimmed datasets shown for the PE400 data but not the other datasets? Why are three lines shown for the RNAseq PE300 data? Are the three lines from the same library? If so, the amounts can be summed and presented on one line, if not, the differences need to be explained in the table and in the text.
> We have removed the row “PE400trim” and have added a legend to Table 1 as follows:
“Tree-type libraries were constructed in this study. A PE400 library was constructed as a fragment library for ALLPATHS-LG. Mate-pair libraries (MP3K and MP5K) were also constructed for ALLPATHS-LG assembly. Three PE300 libraries (PE300A, PE300B, and PE300C) were constructed from RNA for gene annotation.”
We have added more description to the methods in the revised manuscript as follows:
“, total RNA was extracted from the whole body of 10 adults in three different groups using the RNeasy mini kit (Qiagen, Valencia, CA, USA), according to the manufacturer’s instructions.”

Table 2 and the text are not consistent with regards to the reported size of the genome.
> We have corrected “137 Mb” with “138 Mb” in the manuscript.

What does "93.8% of the fragment library was full" mean? It is not clear in the text.
> We have rewritten the sentence as follows:
“93.8% of the paired-end reads from the fragment library overlapped and merged into one longer read.”

Several of the tables should be included as supplementary tables and not in the main text (Table 4 and Table 5, for example).
> As the reviewer suggested, we have moved the Tables 4 and Table 5 and made them as Supplementary Tables.

Table 7 is unclear, what are the 1*, 2* etc?

> We have added this to the legend of the Table, which is now Table S3.
“* The numbers are identifiers for internal branches of the phylogeny (Figure 1C)”

The use of "expressed sequence tags" to describe RNAseq is incorrect.
> We have changed “expressed sequencing tags” to “RNA evidence” on line 62 and changed “expressed sequencing tags” to “transcriptome assembly” on line 165.

The sentence on lines 68-69 does not make sense.
> We have rewritten the sentence as follows:
“Finally, yields after quality trimming for the fragment library totaled 14.8 gigabase pairs (Gbp).”

Blast2Go is Blast2GO. Similarly, maker should be capitalized.
> We have corrected “Blast2Go” with “Blast2GO”.

Reviewer #4: The authors describe their sequencing of this midge genome and a pretty typical set of metrics of evaluating it, but not much more. I understand this is acceptable for a Data Note, so instead have focused this review on making the work more readable and interpretable.

L28. In the Background of the Abstract, the authors says that "B. antartica has unusual characteristics with a compact genome as a result of adaptation to an extreme environment". I don't think there is any evidence that the compactness of that genome has anything to do with the extreme Antarctic environment, it could just be a coincidence. Perhaps other members of that genus or that lineage of midges has similarly tiny genomes, and even if they don't, one would require study of many independent origins of cold-hardiness to say small genomes result from adaptation to extreme environments.
> We have rewritten the sentence as follows:
“B. antarctica is an extremophile with unusual adaptations. The larvae of B. antarctica are desiccation- and freeze-tolerant and the adults are wingless.”

L31. Here and elsewhere the authors say that their subject, P. steinenii, could be a good species for comparative analysis with B. antactica, however that would depend on how close a relative it is. From their phylogeny in Figure 1C is appears that they are very distantly related to each other so presumably these are two independent examples of adaptation to a cold environment. In this case it would be hard to come to much of a conclusion as their routes to cold-resistance might be completely different. This affects the final sentence of the Conclusions too.
> We have rewritten the sentence as follows:
“We present an annotated draft genome of the Antarctic midge, P. steinenii. The genomes of P. steinenii and B. antarctica will aid in the elucidation of evolution under harsh environments and provide new resources for functional genomic analyses of the order Diptera.”

L49. The authors use the singular sense to describe the "Specimen of Parachlus steinenii was collected", implying that the entire genome sequence was obtained from a single specimen, however they then describe at least three libraries constructed for the project and it is hard to imagine doing that from a single midge. I presume they mean to say "Specimens …. were collected". Even so it would be good to specify how

many individuals were used for each of the three libraries, especially the fragment or paired-end library, because that determines how many different haplotypes might be represented in the assembly. Presumably the jumping or mate-pair libraries were from multiple specimens.
> We used the genomic DNA extracted from twenty adults for genomic libraries and have added the sentence in the manuscript:
“Twenty adults were used for genome sequencing, regardless of gender.”

L53. As written this does not make sense as there were apparently two jumping or mate-pair libraries with inserts originally 3 and 5kb long, so it should be plural.
> We have changed “while the other was a jumping library” to “whereas others were jumping libraries.”

L54. Again, the authors say "Paired-end libraries were sequenced…", however they describe only a single paired-end library as being constructed.
> We have corrected “libraries” with “library” on line 54.

L60. While technically "expressed sequencing tags", this term is generally not used for modern RNAseq libraries sequenced on ILLUMINA machines, instead these are generally entire transcriptomes. The term ESTs went out with Sanger sequencing.
> We have changed “expressed sequencing tags” to “RNA evidence” on lines 62 and have changed “expressed sequencing tags” to “transcriptome assembly” on line 165.

In Table 1, the authors list three PE300 libraries for RNAseq, however in the text they only mention a "whole body" extraction, so were all three libraries from the same whole body extraction? If so, why three libraries?
> We have added a legend to Table 1 as follows:
“Tree-type libraries were constructed in this study. A PE400 library was constructed as a fragment library for ALLPATHS-LG. Mate-pair libraries (MP3K and MP5K) were also constructed for ALLPATHS-LG assembly. Three PE300 libraries (PE300A, PE300B, and PE300C) were constructed from RNA for gene annotation.”
We have changed the manuscript as follows:
“total RNA was extracted from the whole body of 10 adults in three different groups using the RNeasy mini kit (Qiagen, Valencia, CA, USA), according to the manufacturer’s instructions.”

The English is the description of the Genome Assembly, lines 74-83 is again poor with singular and plural mixed up repeatedly. And what does it mean that "In this assembly, 93.8% of the fragment library was full."?
> We have corrected the singular and plural mix up and we changed “93.8% of the fragment library was full” to “93.8% of the paired-end reads from the fragment library overlapped and merged into one longer read.”
Thus, we have rewritten the section as follows:
“Assembly was performed using ALLPATHS-LG for both the fragment library (400 bp) and the jumping libraries (3 kbp and 5 kbp) [4]. This was performed on a 96-processor workstation with an Intel Xeon X7460 2.66 GHz processor, 1 terabyte RAM, and default parameters. For better assembly in ALLPATHS-LG, a larger k-mer size was used with one longer read generated from the paired-end library [4]. As a result, the paired-end reads from the fragment library were designed to overlap, and the insert size of the paired-end library was slightly less than twice the read size [4]. In this assembly, 93.8% of the paired-end reads from the fragment library overlapped and merged into one longer read. The resulting assembly had a total size of 138 Mb, comprising 9,513 contigs, with an N50 contig size of 34,110 bp, and an N50 scaffold size of 168 kb (Table 2). The GC content was 32.2% and the assembly revealed contig coverage of approximately 89 × total read length from the fragment library.”

At this point I would suggest a slight reorganization of the manuscript, placing the Repeat Analysis and Non-coding RNA section before the Gene Annotation sections, which makes more sense as the repeats were then masked for the gene annotation.
> As the reviewer suggested, we have placed the repeat analysis and non-coding RNA section before the gene annotation section.

L151-154 are redundant.
> We have removed the sentence “coding sequences (CDS) … … In this study,”

	There is something unsettling about the gene family expansion analysis reported in L194-208 and Table 7. Perhaps it is just that the families identified, such as ID PS0074 for "serine protease" are just one particular family of proteases, but it certainly seems very unlikely, for example, that P steinenii would have no serine proteases. A little more elaboration of these results would be useful. > We have added more description to the revised manuscript as follows: “To estimate the average gene expansion/contraction rate and to identify gene families that have undergone significant size changes through evolution [23, 24], we estimated differences in the size of 15,633 orthologs using the program CAFE3.0 [25]. A Newick description of a rooted and bifurcating phylogenetic tree was needed for this analysis. Therefore, we performed phylogenetic analyses among six insects with the protein-coding gene in the orthologous groups. We selected 4,814 orthologous gene sets from the orthologous groups from OrthoMCL using the reciprocal best BLASTP hit criteria. Protein-coding gene sequences were aligned using PRANK (Ver. 130820) under a codon model with the “-DNA and –codon” option [26], poor alignment sites were eliminated using Gblock (Ver. 0.91) under a codon model with the “-t = c” option [27], and the remaining alignment regions were concatenated to be used in the phylogenetic analyses. The phylogenetic tree was constructed using the neighbor-joining method [28] in the MEGA version 6 program [29]. With the resulting phylogenetic tree, we prepared the ultrametric tree of the species, including branch lengths in units of time through TimeTree [30], for the analysis (Figure 1C). We ran the program using $p < 0.05$, and estimated birth (λ) and death (μ) rates were calculated using the program LambdaMu with the “–s” option. We calculated the number of gene gains and losses on each branch of the tree with the “–t” option. Average expansion size of the two Antarctic midges were lower than that of other insects (Figure 1C), and average expansion size of D. melanogaster exhibited the highest score among the six insects. Using $p < 0.0001$ for the family-wide significance value, we expected approximately one significant result by chance and calculated the exact p-values for transitions over every branch. We called individual branches significant at $p < 0.005$ [31]. We identified three and two gene families that were significantly expanded in P. steinenii and B. antarctica, respectively (Table S3).” In addition, we have changed “serine protease” to “serine protease gd-like.”
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals	Yes

and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Data Note

Genome sequencing of the winged midge, *Parochlus steinenii*, from the Antarctic Peninsula

Sanghee Kim[1], Mijin Oh[2], Woongsic Jung[1], Joonho Park[3], Han-Gu Choi[1] and Seung Chul Shin[4]*****

[1]Division of Life Sciences, Korea Polar Research Institute (KOPRI), Incheon 21990, South Korea

[2]LabGenomics Clinical Research Institute, LabGenomics, Seongnam, Korea

[3]Department of Fine Chemistry, Seoul National University of Science and Technology, Seoul 01811, South Korea

[4]Unit of Polar Genomics, Korea Polar Research Institute (KOPRI), Incheon 21990, South Korea

*Correspondence: biotech21@gmail.com

Abstract{1st level heading}

Background: In the Antarctic, only two species of Chironomidae occur naturally—the wingless midge, *Belgica antarctica,* and the winged midge, *Parochlus steinenii*. *B. antarctica* is an extremophile with unusual adaptations. The larvae of *B. antarctica* are desiccation- and freeze-tolerant and the adults are wingless. Recently, the compact genome of *B. antarctica* was reported and it is the first Antarctic eukaryote to be sequenced. Although *P. steinenii* occurs naturally in the Antarctic with *B. antarctica*, the larvae of *P. steinenii* are cold-tolerant but not freeze-tolerant and the adults are winged. Differences in adaptations in the Antarctic midges are interesting in terms of evolutionary processes within an extreme environment. Herein, we provide the genome of another Antarctic midge to help elucidate the evolution of these species.

Results: The draft genome of *P. steinenii* had a total size of 138 Mbp, comprising 9513 contigs with an N50 contig size of 34,110 bp, and a GC content of 32.2 %. Overall, 13,468 genes were predicted using the MAKER annotation pipeline, and gene ontology classified 10,801 (80.2 %) predicted genes to a function. Compared with the assembled genome architecture of *B. antarctica*, that of *P. steinenii* was approximately 50 Mbp longer with 6.2-fold more repeat sequences, whereas gene regions were as similarly compact as in *B. antarctica*.

Conclusions: We present an annotated draft genome of the Antarctic midge, *P. steinenii*. The genomes of *P. steinenii* and *B. antarctica* will aid in the elucidation of evolution in harsh environments and provide new resources for functional genomic analyses of the order Diptera.

Keywords{1st level heading}

Parochlus steinenii, Cold-tolerant, Complete mitochondrial genome, Antarctic winged midge

Data description{1st level heading}

Sequencing{2nd level heading}

Parochlus steinenii specimens [1-3] were collected from King George Island, West Antarctica (62° 14′ S, 58° 47′ W) during 2014 and 2015. Twenty adults were used for genome sequencing, regardless of gender. Genomic DNA was extracted using a DNeasy Tissue Kit (Qiagen, Valencia, CA, USA). For genome sequencing and assembly using ALLPATHS-LG [4], two types of libraries were prepared. One was a fragment library, which was a paired-end type with an insert size of 400 bp (PE400), whereas others were jumping libraries, which were mate-pair types with insert sizes of 3 kbp (MP3K) and 5 kbp (MP5K). The paired-end library was sequenced with the MiSeq platform (Illumina, San Diego, CA, USA) using a read-length configuration of 2 × 300 bp, and the mate-pair libraries were sequenced with the HiSeq platform (Illumina, San Diego, CA, USA) using a read-length configuration of 2 × 150 bp (see Table 1). Library preparation and sequencing were performed according to the manufacturer’s instructions.

For gene annotation with RNA evidence, total RNA was extracted from the whole bodies of ten adults in three different groups using the RNeasy Mini Kit (Qiagen, Valencia, CA, USA), according to the manufacturer’s instructions. Three paired-end libraries with an insert size of 300 bp (PE300) were constructed using the TruSeq Stranded mRNA Library Prep Kit (Illumina, San Diego, CA, USA) and sequenced with the HiSeq platform (Illumina, San Diego, CA, USA) using a read-length configuration of 2 × 150 bp (Table 1).

Before assembly using ALLPATHS-LG, the paired-end reads resulting from the fragment library were trimmed using the fastq_quality_trimmer in the FASTX-Toolkit (Ver. 0.0.11) [5] with the parameters “-t 30”, “-l 200” and “-Q 33”. Paired sequences from the trimmed Illumina reads were then selected. Finally, after quality trimming, yields for the fragment library totaled 14.8 giga base pairs (Gbp).

Tree-type libraries were constructed in this study, as shown in Table 1. A PE400 library was constructed as a fragment library for ALLPATHS-LG. Mate-pair libraries (MP3K and MP5K) were also constructed for ALLPATHS-LG assembly. Three PE300 libraries (PE300A, PE300B, and PE300C) were constructed from RNA for gene annotation.

Genome assembly{2nd level heading}

Before assembly, we estimated the genome size and heterozygosity using a k-mer analysis with sequencing reads. Jellyfish (Ver. 1.1.10) [6] and GenomeScope [7, 8] software were used. The 17-mers were counted in the reads from the PE400 library and the resulting histogram of 17-mer occurrence was used as a query for GenomeScope [8]. The estimated genome size was 143.8 mega base pairs (Mbp) and the estimated heterozygosity was 0.613 %. Assembly was performed using ALLPATHS-LG for both the fragment library (400 bp) and the jumping libraries (3 kbp and 5 kbp) [4]. This was performed on a 96-processor workstation with Intel Xeon X7460 2.66 GHz processors, 1 TB of RAM, and default parameters. For better assembly in ALLPATHS-LG, a larger k-mer size was used with one longer read generated from the paired-end library [4]. As a result, the paired-end reads from the fragment library were designed to overlap, and the insert size of the paired-end library was slightly less than twice the read size [4]. In this assembly, 93.8 % of the paired-end reads from the fragment library overlapped and merged into one longer read. The resulting assembly had a total size of 138 Mbp, comprising 9513 contigs with an N50 contig size of 34,110 bp and an N50 scaffold size of 168 kbp (Table 2). The GC content was 32.2 % and the assembly revealed contig coverage of approximately 89 × total read length from the fragment library. A total of 57.2 % of the 3-kbp jumping library and 33.1 % of the 5-kbp jumping library were used to improve scaffolding. If more jumping libraries or long jumping libraries (with insert size larger than 20 kbp) were used, the scaffolding might improve the assembly. The assembled genome size was similar to the predicted genome size (143.8 Mbp). We also validated this assembly using CEGMA [9] and BUSCO [10]. CEGMA evaluation showed that the gene completeness of this assembly was 85.08 %, and BUSCO analysis using arthropod databases showed 67.2 % completeness (Tables 3 and 4). If partially matched genes were considered, 92.34 % and 87.5 % of the genes were identified in CEGMA and BUSCO, respectively (Tables 3 and 4).

The statistics for gene annotation for *B. antarctica* were from a reanalysis for comparison of the percentage of the genome created, based on the assembled genome size. From a previous report [11], the assembled genome size of *B. antarctica* was 89.6 Mbp.

CEGMA analysis (Table 3) was performed to validate the genome assembly of *P. steinenii*. The genome sequence of *B. antarctica* (JPYR00000000.1) from the National Center for Biotechnology Information (NCBI) was also analyzed for comparison.

BUSCO analysis was performed to validate genome assembly and gene annotation. For *B. antarctica*, the genome sequence (JPYR00000000.1) from NCBI and the gene set annotated in this study were used. Table 4 shows the numbers and percentages of BUSCO groups.

Repeat analysis and non-coding RNA{2nd level heading}

Interspersed repeats for *P. steinenii* were predicted using RepeatMasker (Ver. 3.3.0) [12] with a *de novo* repeat library. The *de novo* repeat library for *P. steinenii* was constructed using RepeatModeler (Ver. 1.0.3) [13], including the RECON (Ver. 1.07) [13] and RepeatScout (Ver. 1.0.5) [14] software, with default parameters. Tandem repeats, including simple repeats, satellites and low-complexity repeats, were predicted using TRF [15]. Putative tRNA genes were identified using tRNAscan-SE (Ver. 1.3.1) [16] with option “-H”. The repeat content for *B. antarctica* was re-estimated for comparison using RepeatMasker (Ver. 3.3.0) [12] with the Repbase library (Ver. 20140131) [17, 18]. The total coverage of repeat sequences in *P. steinenii* was approximately six times greater than that of repeat sequences in *B. antarctica* (Table 2), and the percentage of the genome was approximately three times higher than that of *B. antarctica*, based on the assembled genome size. Most statistics for repeats were higher in the *P. steinenii* library (Table 5). A total of 186 tRNAs were predicted through tRNAscan-SE [16] (Additional file 1: Table S1).

Gene annotation{2nd level heading}

Gene annotation was accomplished using the MAKER annotation pipeline [19, 20]. RepeatMasker (Ver. 3.3.0) [12] was used to identify repetitive elements against a *de novo* repeat library, and the SNAP gene finder [21] was selected to perform *ab initio* gene prediction from the masked genome sequence. To find the best possible gene model for the given region, evidence of RNA and protein alignments were considered in MAKER2 [20]. Transcriptome assembly results were used for RNA evidence; the paired-end reads resulting from mRNA of the whole body of adults were trimmed using the fastq_quality_trimmer in the FASTX-Toolkit (Ver. 0.0.11) [5] with the parameters “-t 30”, “-l 80” and “-Q 33”, and they were assembled with CLC Genomics Workbench (Ver. 8.0.0) using default parameters. In all, 68,392 contigs, with an N50 contig size of 435 bp and an average contig size of 407 bp, were generated and used for RNA evidence. Protein sequences from six species, given in NCBI reference sequences, were used for protein evidence—*Drosophila melanogaster* (fruit fly, GCF_000001215.4), *Ceratitis capitata* (Mediterranean fruit fly, NC_000857.1), *Bactrocera dorsalis* (oriental fruit fly, NC_008748.1), *Anopheles gambiae* (African malaria mosquito, NZ_AAAB00000000.1), *Aedes aegypti* (yellow fever mosquito, AAGE00000000.2) and *Culex quinquefasciatus* (southern house mosquito, AAWU01000000). Alignment of transcriptome assembly with BLASTn and alignment of homologous protein information from tBLASTx were considered as evidence for annotation. To assess the annotated gene set, we ran a BUSCO analysis in the "OGS" mode for gene set completeness and identified 70.7 % genes to be considered complete with the expanded gene set; 16.5 % of the gene set was classified as missing [10].

Blast2GO (Ver. 2.6.0) assigned preliminary functions for 13,468 genes, and gene ontology (GO) classified 10,801 (80.2 %) of the predicted genes to a function using the BLASTp and InterproScan results [22]. GO annotation described the classified proteins as those required for biological processes (7434; 55.2 %), molecular functions (9576; 71.1 %) and cellular components (4871; 36.2 %). Enzyme Commission (EC) numbers were obtained for 987 proteins.

Gene annotation for *B. antarctica*{2nd level heading}

To investigate the difference in gene content between *P. steinenii* and *B. antarctica,* we also annotated the genome of *B. antarctica* with the same methods used for *P. steinenii*. The reads in various experimental conditions for *B. antarctica* (SRR566981, SRR567289, SRR567164–SRR567167 and SRR567169–SRR567171) were downloaded from the NCBI Sequence Read Archive and we assembled them into 38,017 contigs, with an N50 contig size of 1799 bp and an average contig size of 913 bp, through CLC Genomics Workbench (Ver. 8.0.0). For RNA evidence, the resulting contigs were aligned to the genome sequence of *B. antarctica* with BLASTn in MAKER2. For protein evidence, we used the same protein sequence from the six species used for gene annotation in *P. steinenii* and predicted proteins of *B. antarctica*. From

MAKER2, 11,005 genes were predicted in the *B. antarctica* genome. The annotated gene set in this analysis was assessed using BUSCO [10]. Gene set completeness was 86.6 % including the expanded gene set, and 8.3 % of the gene set was missing (Table 4).

Ortholog analysis{2nd level heading}

Orthologous groups were identified using OrthoMCL (Ver. 2.0.5) [23]. We used the standard parameters and options of OrthoMCL for all steps. In this analysis, coding sequences (CDS) from four genome assemblies (BDGP6 for *D. melanogaster*, AgamP4 for *A. gambiae*, AaegL3 for *A. aegypti* and CpipJ2 for *C. quinquefasciatus*) were collected from Ensemble Metazoa [24] and the CDS from MAKER2 were used for *B. antarctica* and *P. steinenii*. Total proteins were categorized into 15,633 groups—4814 orthologous groups were identified as common to all six insects, 437 groups specific to *P. steinenii* genes were not identified in any other species, and 349 groups were identified only in the two Antarctic midges (Fig. 1A and Additional file 1: Table S2).

Gene structure of orthologous groups{2nd level heading}

B. antarctica showed a reduction in intron length with very low repeat sequences [11]. Therefore, we compared intron lengths of orthologous genes among the six insects to identify whether the intron lengths of the genes in *P. steinenii* were also reduced. We used the information from gene structures of the four genome assemblies (BDGP6 for *D. melanogaster*, AgamP4 for *A. gambiae*, AaegL3 for *A. aegypti* and CpipJ2 for *C. quinquefasciatus*) and the information from MAKER2 annotation of *B. antarctica* and *P.*

steinenii. Among the six insects, the average intron length of *B. antarctica* (302 bp) was the smallest, although that of *P. steinenii* (319 bp) was similar (Fig. 1B). Despite a difference in the assembled genome size between *B. antarctica* and *P. steinenii* of approximately 50 Mbp, the average length of gene regions and CDS were also similar in the two. However, the average intron number in orthologous genes was higher in *P. steinenii*, which was the highest of all six insects (Fig. 1B).

GO enrichment test{2nd level heading}

We used AgriGO [25] to identify which GO terms of the 437 groups that were unique to *P. steinenii* were statistically overrepresented relative to the GO terms of all genes of *P. steinenii*. A total of 1352 genes comprised these 437 groups, and therein were 717 genes with GO terms. AgriGO is a web-based tool for GO analysis: we selected “Fisher’s exact test” for the statistical test method and “Hochberg FDR” as the multiple test adjustment method. GO terms were tested with a significance level of $p < 0.05$. Complete hierarchies of GO terms for each gene were examined. GO enrichment analysis identified 49 GO terms as statistically overrepresented: 26 GO terms in biological processes, five in cellular components and 18 in molecular functions (Table 6). It is noteworthy that 14 of the 26 significant GO terms in biological processes were associated with the unfolded protein response (UPR). The UPR is a stress response that occurs in the lumen of the endoplasmic reticulum (ER) [26]. When unfolded or misfolded proteins are accumulated in the ER lumen under stress conditions, the UPR is activated to improve protein folding by increasing the production of chaperones [26]. Representative GO terms in biological processes related to the UPR were mRNA splicing via endonucleolytic cleavage and ligation (GO:0070054), response to unfolded protein

(GO:0006986), and endoplasmic reticulum unfolded protein response (GO:0030968).

Likelihood analysis of gene gain and loss{2nd level heading}

To estimate the average gene expansion/contraction rate and to identify gene families that have undergone significant size changes through evolution [27, 28], we estimated differences in the size of 15,633 orthologs using the program CAFE3.0 [29]. A Newick description of a rooted and bifurcating phylogenetic tree was needed for this analysis. Therefore, we performed phylogenetic analyses among six insects with the protein-coding gene in the orthologous groups. We selected 4814 orthologous gene sets from the orthologous groups from OrthoMCL using the criterion of reciprocal best BLASTP hit. Protein-coding gene sequences were aligned using PRANK (Ver. 130820) under a codon model with the “-dna -codon” option [30], poor alignment sites were eliminated using Gblock (Ver. 0.91) under a codon model with the “-t = c” option [31], and the remaining alignment regions were concatenated for use in the phylogenetic analyses. The phylogenetic tree was constructed using the neighbor-joining method [32] in the MEGA (Ver. 6) program [33]. From the resulting phylogenetic tree, we prepared the ultrametric tree of the species, including branch lengths in units of time through TimeTree [34], for the analysis (Fig. 1C). We ran the program using $p < 0.05$, and estimated rates of birth (λ) and death (μ) were calculated using the program LambdaMu with the “-s” option. We calculated the number of gene gains and losses on each branch of the tree with the “-t” option. The average expansion (AE) sizes of the two Antarctic midges were lower than those of the other four insects (Figure 1C), and *D. melanogaster* exhibited the highest AE score among the six. Using $p < 0.0001$ for the family-wide significance value, we expected approximately one significant result by chance and calculated the exact p-values for transitions over every branch. We called individual branches significant at $p < 0.005$ [35]. We identified three gene families that were significantly expanded in *P. steinenii* and two in *B. antarctica*, (Additional file 1: Table S3).

Availability of supporting data{1st level heading}

Supporting data (sequence files for CDS, proteins, transcripts and the draft genome, and the general feature format for genes and repeats) are available in the *GigaScience* GigaDB database [36] and the raw data were deposited in the NCBI BioProject repository PRJNA284858 (SRX1976250–SRX1976255).

Abbreviations{1st level heading}

CDS, coding sequence; Gbp, giga base pairs; GO, gene ontology; Mbp, mega base pairs.

Competing interests{1st level heading}

The authors declare that they have no competing interests.

Authors’ contributions{1st level heading}

SK, HGC and SCS designed the study. SK, WJ and HGC collected the samples and performed the experiments. SCS and JP analyzed the data. All authors participated in the writing of the manuscript.

Acknowledgements{1st level heading}

This study was supported by grant PE16020, grant PE16070, a Polar Genomics 101 Project grant, and a Modeling responses of terrestrial organisms to environmental changes on King

George Island grant funded by the Korea Polar Research Institute (KOPRI).

References{1st level heading}

1. Convey P, Block W: Antarctic Diptera: ecology, physiology and distribution. *European Journal of Entomology* 1996, 93:1-14.
2. EDWARDS M, USHER MB: The winged Antarctic midge Parochlus steinenii (Gerke)(Diptera: Chironomidae) in the South Shetland Islands. *Biological Journal of the Linnean Society* 1985, 26(1):83-93.
3. Shimada K, Ohyama Y, Pan C: Cold-hardiness of the Antarctic winged midge Parochlus steinenii during the active season at King George Island. *Polar Biology* 1991, 11(5):311-314.
4. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S: High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* 2011, 108(4):1513-1518.
5. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit. Accessed 22 November 2016.
6. Marçais G, Kingsford C: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011, 27(6):764-770.
7. GenomeScope. http://qb.cshl.edu/genomescope. Accessed 22 November 2016.
8. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz M: GenomeScope: Fast reference-free genome profiling from short reads. *bioRxiv* 2016:075978.
9. Parra G, Bradnam K, Korf I: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 2007, 23(9):1061-1067.
10. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015:btv351.
11. Kelley JL, Peyton JT, Fiston-Lavier A-S, Teets NM, Yee M-C, Johnston JS, Bustamante CD, Lee RE, Denlinger DL: Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nature communications* 2014, 5.
12. Tarailo- Graovac M, Chen N: Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* 2009:4.10. 11-14.10. 14.
13. Bao Z, Eddy SR: Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research* 2002, 12(8):1269-1276.
14. Price AL, Jones NC, Pevzner PA: De novo identification of repeat families in large genomes. *Bioinformatics* 2005, 21(suppl 1):i351-i358.
15. Benson G: Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* 1999, 27(2):573.
16. Lowe TM, Eddy SR: tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* 1997, 25(5):955-964.
17. Repbase library. http://www.girinst.org/repbase/index.html. Accessed 22 November 2016.

18. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* 2005, 110(1-4):462-467.

19. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M: MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* 2008, 18(1):188-196.

20. Holt C, Yandell M: MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *Bmc Bioinformatics* 2011, 12(1):1.

21. Korf I: Gene finding in novel genomes. *Bmc Bioinformatics* 2004, 5(1):1.

22. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, 21(18):3674-3676.

23. Li L, Stoeckert CJ, Roos DS: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 2003, 13(9):2178-2189.

24. Ensemble Metazoa. http://metazoa.ensembl.org/index.html. Accessed 22 November 2016.

25. Du Z, Zhou X, Ling Y, Zhang Z, Su Z: agriGO: a GO analysis toolkit for the agricultural community. *Nucleic acids research* 2010:gkq310.

26. Chen Y, Brandizzi F: IRE1: ER stress sensor and cell fate executor. *Trends in cell biology* 2013, 23(11):547-555.

27. Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. *Science* 2000, 290(5494):1151-1155.

28. Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N: Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome research* 2005, 15(8):1153-1160.

29. De Bie T, Cristianini N, Demuth JP, Hahn MW: CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 2006, 22(10):1269-1271.

30. Loytynoja A, Goldman N: An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* 2005, 102(30):10557-10562.

31. Talavera G, Castresana J: Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology* 2007, 56(4):564-577.

32. Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* 1987, 4(4):406-425.

33. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution* 2013, 30(12):2725-2729.

34. Hedges SB, Dudley J, Kumar S: TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 2006, 22(23):2971-2972.

35. Hahn MW, Han MV, Han S-G: Gene family evolution across 12 Drosophila genomes. *PLoS Genet* 2007, 3(11):e197.

36. Kim, S; Oh, M; Jung, W; Park, J; Choi, H; Shin, S, C (2016): Supporting data for "Genome sequencing of the winged midge, Parochlus steinenii, from the Antarctic Peninsula" GigaScience Database. http://dx.doi.org/10.5524/100256.

{Tables}

Table 1 Sequencing libraries and respective yield used for genome assembly of *Parochlus steinenii*

Library	Mode	Insert size	Library type	Reads	Total read lengths (Gbp)	Source
PE400	2×300	400	paired-end	51,892,430	15.6	Genomic DNA
MP3K	2×150	3000	mate-pair	170,887,140	25.6	Genomic DNA
MP5K	2×150	5000	mate-pair	157,622,418	23.6	Genomic DNA
PE300A	2×150	300	paired-end	27,663,170	3.5	RNA
PE300B	2×150	300	paired-end	27,782,288	3.5	RNA
PE300C	2×150	300	paired-end	30,806,804	3.9	RNA

Table 2 Global statistics of the *Parochlus steinenii* genome assembly

Assembly results	Number	N50 (kbp)[a]	Size (Mbp)
Contig	9513	34.1	130.6
Scaffold	4151	168.1	138.0

Annotation	Number	Total length (kbp)	Percentage of the assembled genome
Genes	13,468	36,239.1	26.3
Coding regions (Coding regions in *B. antarctica*)	13,468 (11,005)	17,967.6 (17,518.0)	13.0 (19.6)

Introns (Introns in *B. antarctica*)	69,960 (43,577)	24,191.6 (15,494.9)	17.5 (17.2)
Repeats (Repeats in *B. antarctica*)	37,507 (10,084)	2252.6 (361.4)	1.6 (0.40)

[a]Minimum sequence length in which half of the assembled bases were found

Table 3 CEGMA analysis of two Antarctic midges

	CEG set	Complete proteins	Percentage complete	Total observed	Average copy number	Percentage of orthologs
P. steinenii	Complete	211	85.08	247	1.17	14.22
	Partial	229	92.34	283	1.24	19.65
B. antarctica	Complete	241	97.18	283	1.17	12.03
	Partial	247	99.6	311	1.26	18.18

CEG core eukaryotic gene

Table 4 BUSCO analysis of two Antarctic midges

	Genome assembly		Gene set	
	P. steinenii	***B. antarctica***	***P. steinenii***	***B. antarctica***
Complete BUSCOs (%)	1798 (67.2)	2310 (86.4)	1890 (70.7)	2316 (86.6)
Complete and single-copy BUSCOs (%)	1648 (61.6)	2170 (81.1)	1620 (60.6)	2074 (77.5)
Complete and duplicated BUSCOs (%)	150 (5.6)	140 (5.2)	270 (10.1)	242 (9.0)
Fragmented BUSCOs (%)	543 (20.3)	270 (10.1)	343 (12.8)	137 (5.1)
Missing BUSCOs (%)	334 (12.5)	95 (0.04)	442 (16.5)	222 (8.3)
Total BUSCO groups searched	2675 (100)			

Table 5 Repeat content in Antarctic midges

	P. steinenii		*B. antarctica*	
	Total coverage (bp)	**Number of sequences**	**Total coverage (bp)**	**Number of sequences**
Low complexity	404,490	8661	290,095	8812
Simple repeats	1,105,449	26,336	40,475	1066
Transposon elements				
Class I/LTR	289,059	1075	945	13
Class I/Non-LTR	169,298	675	18,003	271
Class II/DNA elements	216,807	649	5247	83
Small RNA	67,503	111	6425	13
Totals	2,252,606	37,507	361,370	10,258

LTR long terminal repeat

Table 6 GO terms statistically overrepresented only in *Parochlus steinenii*-specific groups

GO ID	GO tree	Term	No. of genes[a]	No. of genes[b]	*p*-value	FDR
GO:0006508	P	proteolysis	106	632	8.60E-13	2.60E-10
GO:0006397	P	mRNA processing	32	120	6.80E-10	1.00E-07
GO:0070054	P	mRNA splicing, via endonucleolytic cleavage and ligation	8	8	1.40E-09	1.40E-07
GO:0016071	P	mRNA metabolic process	32	130	5.80E-09	4.50E-07
GO:0000394	P	RNA splicing, via endonucleolytic cleavage and	8	11	1.90E-07	1.10E-05

		ligation				
GO:0006986	P	response to unfolded protein	6	7	1.50E-06	7.60E-05
GO:0019538	P	protein metabolic process	173	1506	2.20E-06	9.40E-05
GO:0051789	P	response to protein stimulus	6	8	5.50E-06	0.00021
GO:0006950	P	response to stress	50	330	8.00E-06	0.00027
GO:0006468	P	protein amino acid phosphorylation	42	272	2.40E-05	0.00074
GO:0080135	P	regulation of cellular response to stress	9	24	4.80E-05	0.0013
GO:0006396	P	RNA processing	34	210	5.00E-05	0.0013
GO:0051347	P	positive regulation of transferase activity	8	22	0.00016	0.0031
GO:0033674	P	positive regulation of kinase activity	8	22	0.00016	0.0031
GO:0045860	P	positive regulation of protein kinase activity	8	22	0.00016	0.0031
GO:0034620	P	cellular response to unfolded protein	4	5	0.00017	0.0031
GO:0030968	P	endoplasmic reticulum unfolded protein response	4	5	0.00017	0.0031
GO:0042246	P	tissue regeneration	6	13	0.00024	0.0041
GO:0031099	P	regeneration	6	14	0.00039	0.0063
GO:0071445	P	cellular response to protein stimulus	4	6	0.00049	0.0071
GO:0071216	P	cellular response to biotic stimulus	4	6	0.00049	0.0071
GO:0034976	P	response to endoplasmic reticulum stress	4	7	0.0011	0.015
GO:0061053	P	somite development	3	4	0.0018	0.024
GO:0006984	P	ER-nuclear signaling pathway	4	8	0.002	0.026
GO:0006379	P	mRNA cleavage	4	9	0.0034	0.041
GO:0016310	P	phosphorylation	49	421	0.0041	0.049
GO:0031463	C	Cul3-RING ubiquitin ligase complex	5	5	2.90E-06	0.00019
GO:0031461	C	cullin-RING ubiquitin ligase complex	5	12	0.0014	0.047
GO:0005789	C	endoplasmic reticulum membrane	11	55	0.0032	0.063
GO:0042175	C	nuclear envelope–endoplasmic reticulum network	11	57	0.0042	0.063
GO:0044432	C	endoplasmic reticulum part	11	58	0.0049	0.063
GO:0004252	F	serine-type endopeptidase activity	76	292	3.70E-20	5.50E-18

GO:0004540	F	ribonuclease activity	30	54	1.90E-19	1.40E-17
GO:0008236	F	serine-type peptidase activity	76	318	6.90E-18	2.50E-16
GO:0017171	F	serine hydrolase activity	76	318	6.90E-18	2.50E-16
GO:0004175	F	endopeptidase activity	84	416	5.70E-15	1.70E-13
GO:0070011	F	peptidase activity, acting on L-amino acid peptides	103	570	1.60E-14	4.00E-13
GO:0008233	F	peptidase activity	103	595	2.40E-13	5.10E-12
GO:0004518	F	nuclease activity	30	102	1.70E-10	3.10E-09
GO:0031072	F	heat shock protein binding	10	17	1.00E-07	1.60E-06
GO:0004672	F	protein kinase activity	47	300	5.90E-06	8.70E-05
GO:0008234	F	cysteine-type peptidase activity	15	59	3.70E-05	0.00049
GO:0016787	F	hydrolase activity	171	1580	5.00E-05	0.00061
GO:0016773	F	phosphotransferase activity, alcohol group as acceptor	49	363	0.00018	0.002
GO:0042802	F	identical protein binding	10	38	0.00052	0.0055
GO:0031625	F	ubiquitin protein ligase binding	5	12	0.0014	0.014
GO:0005515	F	protein binding	229	2357	0.0015	0.014
GO:0016301	F	kinase activity	48	405	0.0032	0.027
GO:0003676	F	nucleic acid binding	144	1469	0.0055	0.045

A total of 49 GO terms were statistically overrepresented: 26 in biological processes (P), five in cellular components (C) and 18 in molecular functions (F) were identified as significant by GO enrichment analysis. Fisher's exact test was performed and the resulting *p*-values were adjusted using the Hochberg FDR for multiple comparisons. GO terms associated with the unfolded protein response are shown in *bold font*

ER endoplasmic reticulum, *FDR* false discovery rate, *GO* gene ontology

[a]The number of genes with GO terms in the *P. steinenii*-specific groups

[b]The number of genes with GO terms in *P. steinenii*'s entire gene set

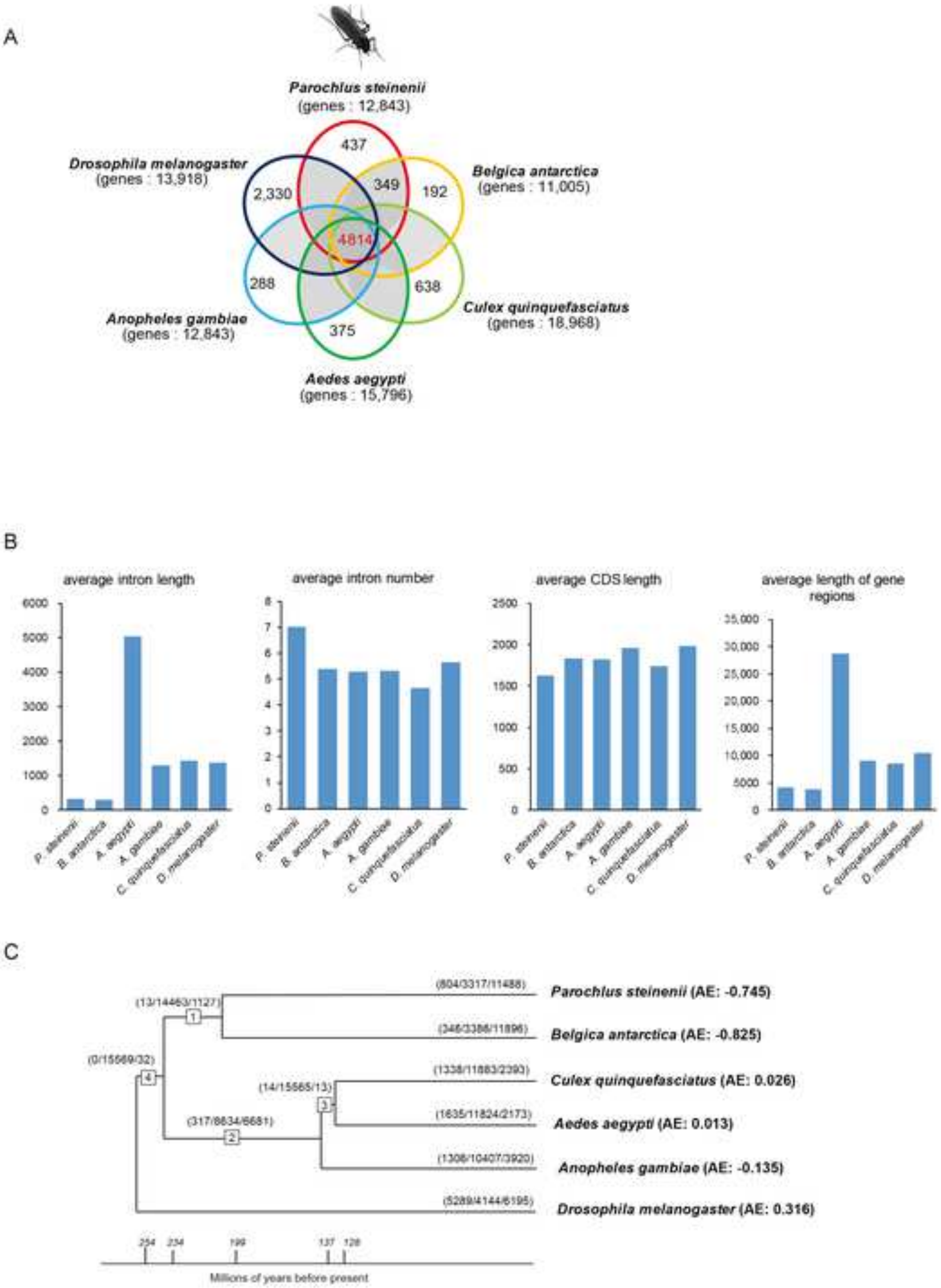
{Figure legends}

Fig. 1 Genome-wide analysis of protein-coding genes in *Parochlus steinenii*. a Venn diagram displaying the overlap in orthologous genes of six insect species and the number of unique groups in each species. **b** The statistics of gene structures of the six insects. **c** Lineage-specific gene gains and losses among the six insects. The numbers in the *boxes* are identifiers for internal branches of the phylogeny. Numbers on each *branch* denote the number of gained, lost and stable genes, respectively. AE denotes the average expansion. The numbers on the *bottom line* denote the estimated divergence time of the corresponding tree nodes above, based on TimeTree

{Additional files}

Additional file 1. DOCX file. **Supplementary tables.**

Figure 1


A
Parochlus steinenii
(genes : 12,843)
Drosophila melanogaster
(genes : 13,918)
Belgica antarctica
(genes : 11,005)
Anopheles gambiae
(genes : 12,843)
Culex quinquefasciatus
(genes : 18,968)
Aedes aegypti
(genes : 15,796)
437
2,330
349
192
4814
288
638
375
B
average intron length
6000
5000
4000
3000
2000
1000
0
average intron number
8
7
6
5
4
3
2
1
0
average CDS length
2500
2000
1500
1000
500
0
average length of gene regions
35,000
30,000
25,000
20,000
15,000
10,000
5000
0
P. steinenii
B. antarctica
A. aegypti
A. gambiae
C. quinquefasciatus
D. melanogaster
C
(804/3317/11488)
Parochlus steinenii (AE: -0.745)
(13/14463/1127)
1
(346/3386/11896)
Belgica antarctica (AE: -0.825)
(0/15569/32)
4
(1338/11883/2393)
Culex quinquefasciatus (AE: 0.026)
(14/15565/13)
3
(1635/11824/2173)
Aedes aegypti (AE: 0.013)
(317/8634/6681)
2
(1306/10407/3920)
Anopheles gambiae (AE: -0.135)
(5289/4144/6195)
Drosophila melanogaster (AE: 0.316)
254
234
199
137
128
Millions of years before present