Supplementary information

Decay of sexual trait genes in an asexual parasitoid wasp

Ken Kraaijeveld

Yahya Anvar

Jeroen Frank

Arnoud Schmitz

Jens Bast

Jeanne Wilbrandt

Malte Petersen

Tanja Ziesmann

Oliver Niehuis

Peter de Knijff

Johan T. den Dunnen

Jacintha Ellers

Functional annotation of the MAKER2 gene models

The scan of the 50,004 protein sequences yielded a total of 133,527 matches distributed as follows: 791 hits for TIGRFAM, 2 for ProDom, 15,822 for SMART, 0 for HAMAP, 5,961 for ProSitePatterns, 18,142 for SuperFamily, 20,181 for PANTHER, 18,700 for Gene3D, 521 for PIRSF, 20,878 for Pfam-A, 13,570 for ProSiteProfiles and 8,238 for Coils.

Transposable elements

De novo repeat detection was done using RepeatModeler v1.0.7(Smit & Hubley) on the *de novo* assembly. Output sequences larger than 500 bp were clustered based on a 95% identity threshold using UCLUST with the centroid option to join fragments and reduce redundancy (Edgar 2010). To identify transposable elements in the clustered repeat library, REPCLASS (Feschotte et al. 2009) and homology repeat searches were run using RepeatMasker(Smit et al. 2013), tBLASTx and BLASTn against RepBase (Jurka et al. 2011) and non-redundant NCBI entries (keywords: retrotransposon, transposase, reverse transcriptase, transposon, transposable element)(E-value > 10^{-30}).

Nucleotide sequences were discarded if all annotation methods labelled library entries as 'unknown'. Ambiguous repetitive elements were double-checked with the online version of CENSOR (Kohany et al. 2006). To remove sequences with high similarity to non-TE entries, library entries were searched against the NCBI database using BLAST. Headers of remaining TE sequences were reformatted to match RepeatMasker naming standards. The final 'strict' TE library contained 162 entries, representing 25 super-families. These sequences had a combined size of 305 kb, a mean length of 1886 bp, with a maximum size of 11311 bp.

Supplementary figures

Figure S1. Base-coverage frequency distribution of Illumina Hiseq reads mapped to the *Leptopilina clavipes* draft genome assembly.

Figure S2. Flow cytometric genome size estimation using heads of *Drosophila melanogaster* (left peak, genome size = 175 Mb) and *Leptopilina clavipes* (right peak, estimated genome size = 321 Mb).

Figure S3. Variants predicted to result in less stable proteins in the sexual *Leptopilina clavipes* lineage than in the asexual lineage are overrepresented in reproductive tissues. Change in protein stability was predicted using MUpro and the orthologs of the genes in which they were found were searched against the genome of *Drosophila melanogaster*. The tissue in which each of these orthologs show highest expression was identified in Flyatlas(Chintapalli et al. 2007).

Supplementary tables

Table S1. Summary statistics of the *Leptopilina clavipes* draft genome assembly.

Metric	**Value**
Initial number of scaffolds	41,161
Number of scaffolds >200bp (*Wolbachia* removed)	36,601
Total bp in scaffolds	255,388,375
Longest scaffold	419,800
N50	13,759

Table S2. Sequencing statistics after quality trimming (Illumina) and read correction (PacBio). Genome coverage was estimated after mapping the reads to the reference assembly.

Sequencing platform	Illumina HiSeq (paired-end)	Pacific Biosciences (corrected)
Total reads	283 M	754 K
Total bases	27.5 Gb	742 Mb
Mean read length	97	984
Median read length	100	862
Size range	36 - 100	500 - 6,109
GC content	30.1 %	30.7 %
Estimated coverage	137 x	3.7 x

Supplementary references

Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat. Genet. 39:715–720. doi: 10.1038/ng2049.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 26:2460–1. doi: 10.1093/bioinformatics/btq461.

Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D. 2009. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol. Evol. 1:205–20. doi: 10.1093/gbe/evp023.

Jurka J, Bao W, Kojima KK. 2011. Families of transposable elements, population structure and the origin of species. Biol. Direct. 6:44. doi: 10.1186/1745-6150-6-44.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 7:474. doi: 10.1186/1471-2105-7-474.

Smit AFA, Hubley R. RepeatModeler Open-1.0. <http://www.repeatmasker.org>.

Smit AFA, Hubley R, Green P. 2013. RepeatMasker Open-4.0. www.repeatmasker.org.

Fig. S1

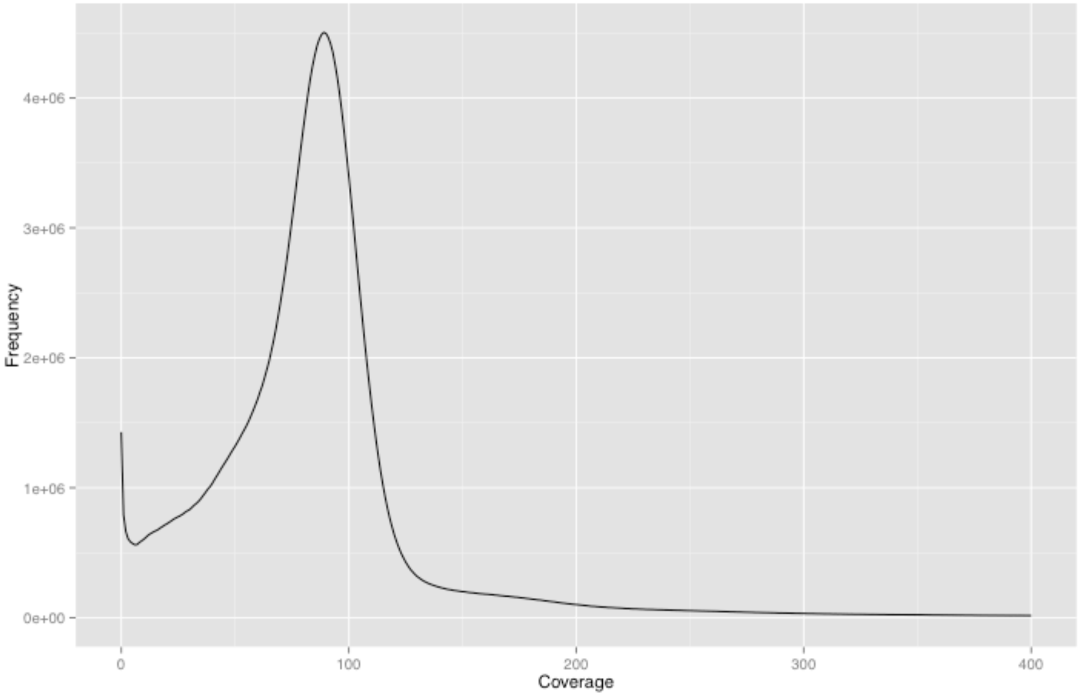

Frequency
4e+06
3e+06
2e+06
1e+06
0e+00
0
100
200
300
400
Coverage

Fig. S2

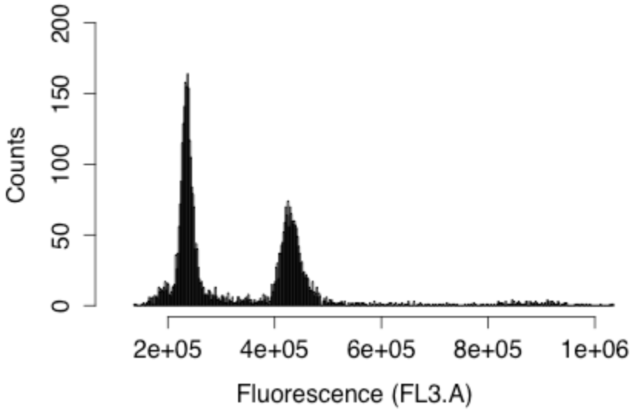
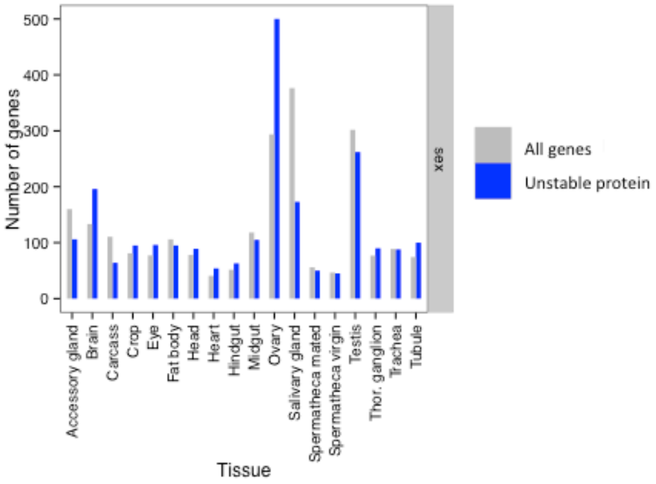

Counts
0
50
100
150
200
2e+05
4e+05
6e+05
8e+05
1e+06
Fluorescence (FL3.A)

Fig. S3


500
400
300
200
100
0
Number of genes
Accessory gland
Brain
Carcass
Crop
Eye
Fat body
Head
Heart
Hindgut
Midgut
Ovary
Salivary gland
Spermatheca mated
Spermatheca virgin
Testis
Thor. ganglion
Trachea
Tubule
Tissue
sex
All genes
Unstable protein