UNIL | Université de Lausanne

Genetics of sex determination in a parasitoid wasp

Master Project
Master in Molecular Life Sciences, Bioinformatics

Cyril Matthey-Doret
Supervised by: Casper Van der Kooi
Directed by: Tanja Schwander

Department of Ecology and Evolution
University of Lausanne - Switzerland

November 9, 2017

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Introduction

New (from scratch)

Sex determination is the upstream signal that will determine whether an organism develops into a male or female. A wide variety of sex determination have been identified across animal species, they can be either environmental, genetic or a combination of both. An active area of research in evolutionary biology focuses on the reasons why so many different strategies have evolved and what would favour the emergent of a particular mechanism. Because of the abundant diversity of sex determination mechanisms they harbour, insects are a good system to investigate these fundamental questions.

Here we are interested in complementary sex determination (CSD) a particular mechanism of sex determination which was proposed to be the ancestral system in Hymenoptera, one of the largest orders of insects comprising over 230,000 described species. All Hymenoptera are haplodiploid; females lay eggs which develop into haploid males if left unfertilized, or into diploid females if fertilized.

Although haploidy is a well described phenomenon, the genetics underlying the detection of haplodiploidy in the embryo are still very little studied. A particular mechanism named complementary sex determination (CSD) has been described and is known to occur in more than 60 hymenopteran species. In CSD species, only one or a few loci underlie haplodiploidy and assume the role of sex determiners. If two different alleles are present in the egg for at least one CSD locus, the organism develops into a female. Haploid eggs therefore develop into males as they are hemizygous for all loci. In wild populations, on the other hand, there is usually a large allelic diversity at CSD, which usually result in heterozygous CSD loci triggering female development. However, when populations are artificially inbred in laboratory strains, the low allelic diversity will eventually result in diploid individuals with all CSD loci homozygous. Such individuals will develop into abnormal diploid males, which, depending on the species, are not viable, sterile or have reduced fitness. The presence of such diploid males is typically how the presence of CSD is inferred in a species, and the frequency of these males has been used in breeding experiments to determine the number of CSD loci in a species from mathematical models.

Lysiphlebus fabarum is a haplodiploid parasitoid wasp with reproductive polymorphism (both sexual and asexual lineages) where CSD is known to be the mechanism underlying sex determination. As it has both reproductive polymorphism and a simple sex determination system, it has great potential as a model species for the study of questions relating to sex determination and the evolutionary maintenance of sex. In this study, we investigate the genetic basis of CSD in *L. fabarum*. The genetics of CSD is only known in the honeybee (*Apis mellifera*) where the *csd* gene stems from a duplication of the ancestral *fem*, a *transformer* homolog. The *csd* gene underwent strong positive

selection which resulted in neo functionalization and it now acts as an RNA splicing protein acting on its paralog, *fem*. Uncovering the mechanisms underlying CSD in a distant species will provide valuable insight into the evolution of sex determination in insects. Here, we show that *L. fabarum* has ml-CSD and that the loci are close to the centromeres. Our results also suggest a different mechanism from the *Apis mellifera* CSD which might have evolved independently.

Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Figures, tables and legends

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Materials and methods

Crossing experiments

Performed prior to master project.
A haploid male coming from an asexual family of *Lysiphlebus fabarum* was crossed with an inbred sexual line. The offspring of asexual females at the 4th generation were used.

RAD-seq protocol

Performed prior to master project.
N samples of 4th generation *Lysiphlebus fabarum* coming from X different asexual females kept in ethanol at -XX°C for XX months. They were sexed visually and prepared in 6 separate libraries, all following the same protocol from XXX. ddRAD-seq was performed on all samples using EcorI and

MseI restriction enzymes and a fragments of 200-450 bp were size selected on agarose gel. Single-end sequencing was performed using Illumina Miseq (or Hiseq 2500???). Samples were multiplexed in each library following the TruSeq multiplexing design and libraries were pooled pairwise on the same Illumina lane using different adapters (iA06 or iA12).

STACKS pipeline

We use the STACKS software (version 1.30) **CITATION** to process RAD-seq data as it is versatile, lends itself well to the analysis of non-model species and is well documented. Following quality control using fastqc (version 0.11), the raw reads were trimmed and demultiplexed using the "process radtags" module from the STACKS suite and two mismatches were allowed to detect adapters. The 96bp demultiplexed reads were mapped to the latest assembly of the *L. fabarum* genome using BWA aln (version 0.7.2) with 4 mismatches allowed. Only uniquely mapped alignments were extracted using samtools (version 1.3). Stacks were generated from SAM files of unique hits using the Pstacks module, requiring a minimum stack depth (-m) of 3. The catalogue of loci was built with Cstacks allowing for a distance (-n) of 3 mismatches between samples at each locus. Individuals with less than 10% total radtags compared to the average across all samples were excluded from the analysis. Populations was run pooling all samples together, requiring each locus to be present (-r) in at least 80% of samples. The different STACKS parameters were selected following guidelines in **Paris et al. 2017**.

Ploidy separation

Genome wide heteroyzgosity per individual was computed on all variant sites using the output VCF file from a first populations run with more stringent parameters. Only high confidence loci were included by requiring a minimum sequencing depth (-m) of 20 reads in populations, to minimize chances of assigning the wrong ploidy to any individual. A conservative threshold of 77% heterozygosity among variant sites was determined empirically **(Figure SX)** and individuals above that threshold were considered haploid and excluded from the main populations run used the other analyses.

Categorizing families

The proportion of males among diploid offspring was used to group families by number of heterozygous CSD loci in the mother. The total number of diploid males among non-sequenced individuals was inferred from sequenced individuals by extrapolating the rate of haploidy from sequenced individuals to the total number of males as follows:

$$N_{DM} = N_M * \frac{n_{dm}}{n_{dm}+n_{hm}}$$

Where upper-case letters represent total individuals count in each family and lower-case letters represent sequenced individuals count in the family. N_M is the number of males, N_{HM} is the number of haploid males and N_{DM} the number of diploid males.

The proportion of males among diploid offspring was then computed using the inferred number of diploid males. The families were classified using 1-dimensional k-means clustering on the proportion of males among diploid offspring and 2 scenarios were considered to decide the number of categories of families; either 2 CSD loci, resulting in 3 categories (k=3), or 3 CSD loci resulting

in 7 categories (k=7).

Finding centromeres

All fixed and variant sites were extracted using the –genomic parameter of the populations STACKS module. For families where the mother was available, we excluded all sites that were either homozygous or missing in the mother from the family. For other families, we excluded sites that were missing of homozygous in all offspring in the family. The proportion of heterozygous sites was computed along the chromosomes in two ways: 1) computing mean heterozygosity in a sliding window containing 30 sites with a step size of 1 site and 2) Using a local regression model with a span of 0.6 to obtain a smooth estimate curve. In each chromosome, the approximate centromere location was considered to be the minimum of the local regression curve.

Association mapping

Case-control association mapping was used to score CSD-candidate SNPs based on their segregation pattern. The number of heterozygous males, heterozygous females, homozygous males and homozygous females was computed for every SNP and a one-sided Fisher exact-test was performed on the 2X2 contingency table. The alternative hypothesis was that the proportion of homozygous males at the SNP is higher than the proportion of homozygous females. P-values were corrected for multiple-testing using Benjamini-Hochberg correction.

Acknowledgements

Jens Bast: Linkage map Casper Van der Kooi: Supervision Tanja Schwander: ideas and direction French team: genome assembly + annotations

Supplementary Material

Code hosted at: [INSERT Github URL] Data hosted at: [INSERT DB URL]

plan

intro: sex-det (envi-gen), role in development, evolutionary biology, arthropods, haplodiploidy, CSD, honeybee, CFA results: ploidy, centromeres, GWAS, homology discussion: hits significance, tra, annotations methods: quick lab stuff, RAD-processing, populations, filters, ploidy, centromeres, homology