
Plan Overview

A Data Management Plan created using DMPonline

Title: Evolution and Genomics of Sexual Dimorphism in Stalk-Eyed Flies

Creator: Benjamin Alston

Affiliation: The University of Sheffield

Template: Postgraduate Research DMP (The University of Sheffield)

Project abstract:

Sexual selection acts upon variation in an organism's mating and fertilisation success, rather than its survival. This evolutionary process is responsible for a range of ecologically important traits, and is known to contribute to the evolution of pronounced sexual dimorphism and differentiated sex chromosomes. How sexual selection contributes to the evolution and turnover of sexual dimorphism is debated, as is the way in which sex specific traits are encoded in a genome mostly shared between sexes. The aim of this project is to understand the relationship between sexual selection and heteromorphic sex chromosome evolution, and to investigate the genetic architecture underlying sexual dimorphism. To achieve this, we will use stalk-eyed flies (family Diopsidae) as a model system. This study system is uniquely situated to answer these evolutionary questions due to its Neo-XY chromosome system and the presence of pronounced sexual dimorphism underpinned by strong sexual selection. My project will involve the construction of reference genomes for several stalk-eyed fly species, followed by characterization and comparative genomics of the sex chromosomes. I will also investigate the structure and evolution of an X-Linked meiotic drive locus, a class of selfish genetic element that has contributed to both sex chromosome evolution and sexual selection. Following this I will use comparative transcriptomics, taking advantage of single-cell sequencing technology to identify candidate genes with potentially important roles in sexual ornament development to understand the genomic architecture underlying sexual dimorphism.

ID: 154742

Start date: 01-10-2023

End date: 30-09-2027

Last modified: 08-07-2024

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Evolution and Genomics of Sexual Dimorphism in Stalk-Eyed Flies

Defining your data

- What digital data (and physical data if applicable) will you collect or create during the project?
- How will the data be collected or created, and over what time period?
- What formats will your digital data be in? (E.g. .doc, .txt, .jpeg)
- Approximately how much digital data (in GB, MB, etc) will be generated during the project?
- Are you using pre-existing datasets? Give details if possible, including conditions of use.

Data generated for this project:

Raw sequence data for three stalk-eyed fly species - *Teleopsis dalmanni*, *Teleopsis whitei*, and *Diasemopsis meigenii* will be generated over the course of my thesis.

Data Type	Species	Description	File size	File Format	Status
Pacbio HiFi long read (DNA)	<i>T. dalmanni</i> <i>T. whitei</i> <i>D. meigenii</i>	male, 1 female for each species	~180GB	.fastq.gz	sequenced
Illumina Short read (DNA)	<i>T. dalmanni</i> <i>T. whitei</i> <i>D. meigenii</i>	5 males, 5 females for each species	~260GB	.fastq.gz	sequenced
Iso-seq (RNA)	<i>T. dalmanni</i> <i>T. whitei</i> <i>D. meigenii</i>	1 male, 1 female for each species	~4.2GB	.fasta .bam	sequenced
Omni-C (DNA)	<i>T. whitei</i> <i>D. meigenii</i>	1 male, 1 female for each species	~100GB (approx)	.fastq.gz	exp. Aug 2024
Omni-C (DNA)	<i>T. whitei</i>	1 Drive Male	~50GB (approx)	.fastq.gz	exp. early 2025
RNA-seq	<i>T. dalmanni</i> <i>T. whitei</i> <i>D. meigenii</i>	15 males, 10 females per species. Two tissues per individual (wing discs & antenna imaginal discs)	~200GB approx	.fastq.gz	exp. 2025

Fly samples from all species were obtained from breeding stocks maintained by the Pomiankowski Research Group (University College London), and all extraction and sequencing has been carried out by NEOF.

Existing Data sets used in this project:

An existing dataset generated by Pomiankowski Research Group (University College London) will be used during chapter 2 of my thesis.

Data Type	Species	Description	File size	File Format	Status
Illumina Short read (DNA)	<i>T. dalmanni</i>	50 drive males, 50 non-drive males	290GB	.tar.gz	sequenced

Looking after data during your research

- Where will you store digital data during the project to ensure it is secure and backed up regularly? (E.g. [University research storage](#), or University Google drive)
- How will you name and organise your data files? (An example filename can help to illustrate this)
- If you collect or create physical data, where will you store these securely?
- How will you make data easier to understand and use? (E.g. include file structure and methodology in a README file)
- Will you use extra security precautions for any of your digital or physical data? (E.g. for sensitive and/or personal data)

Raw data is stored on the wright lab shared drive on the University of Sheffield HPC, with an accompanying spreadsheet with details regarding each dataset. Metadata spreadsheets are kept within the directory of each dataset.

Genomic Resources (e.g., reference genomes and annotations), when generated, will also be stored on the wright lab shared drive.

Any data stored on my local machine is backed up to google drive.

Storing data after your research

- Which parts of your data will be stored on a long-term basis after the end of the project?
- Where will the data be stored after the project? (E.g. University of Sheffield repository [ORDA](#), or a subject-specific repository)
- How long will the data be stored for? (E.g. standard TUoS retention period of minimum 10 years after the project)
- Who will place the data in a repository or other long-term storage? (E.g. you, or your supervisor)
- If you plan to use long-term data storage other than a repository, who will be responsible for the data?

Following my project, I will upload raw sequence data to Short Read Archive and genomic resources (including reference genomes) to genbank.

Sharing data after your research

- How will you make data available outside of the research group after the project? (E.g. openly available through a repository, or on request through your department)
- Will you make all of your data available, or are there reasons you can't do this? (E.g. personal data, commercial or legal restrictions, very large datasets)
- If there are reasons you can't share all of your data, how might you make as much of it available as possible? (E.g. anonymisation, participant consent, sharing analysed data only)
- How will you make your data as widely accessible as possible? (E.g. include a data availability statement in publications, ensure published data has a DOI)
- What licence will you apply to your data to say how it can be reused and shared? (E.g. one of the [Creative Commons](#) licences)

All data generated during my project will be made available on public repositories (see above).

Putting your plan into practice

- Who is responsible for making sure your data management plan is followed? (E.g. you with the support of your supervisor)
- How often will your data management plan be reviewed and updated? (E.g. yearly and if the project changes)
- Are there any actions you need to take in order to put your data management plan into practice? (E.g. requesting [University research storage](#) via your supervisor.)

I am responsible for ensuring my data management plan is followed, with the support of my supervisor (Dr Alison Wright).

The data management plan will be updated when new data is received or my project changes.

A review of my plan with my supervisor towards the end of my project will help ensure that data is stored and made publicly available in the most suitable repositories.