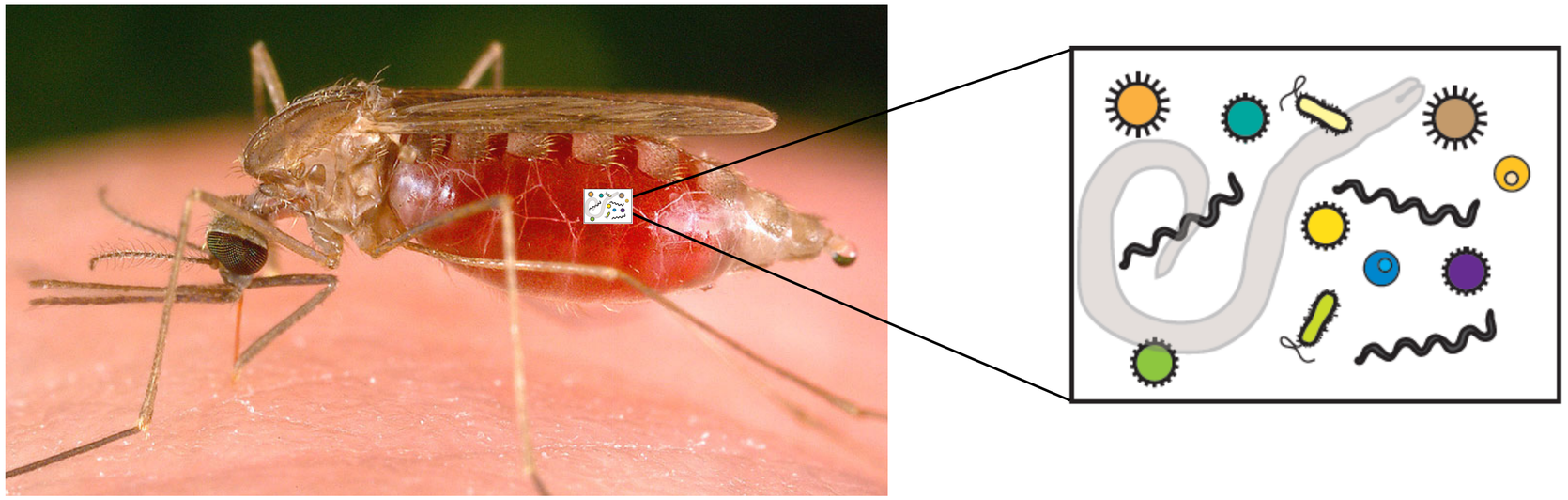
Final Project!

Mark Stenglein, MIP 280A4

National Geographic Channel

Your goal is to identify and characterize virus sequence(s) in datasets from pools of flies or mosquitoes



Your goal is to identify and characterize virus sequence(s) in datasets from pools of flies or mosquitoes

- This exercise combines skills you've learned throughout this semester
- These are real datasets that include not yet published virus sequences
- Shotgun Illumina libraries were made from total RNA from pools of 12 male or female flies or mosquitoes
- Libraries were sequenced on an Illumina NextSeq sequencer to produce single-end 150 base reads
- Library preparation and sequencing done by MIP undergraduates Tillie Dunham and Kai Chase
- You will submit a written report and present to the rest of us what you found.

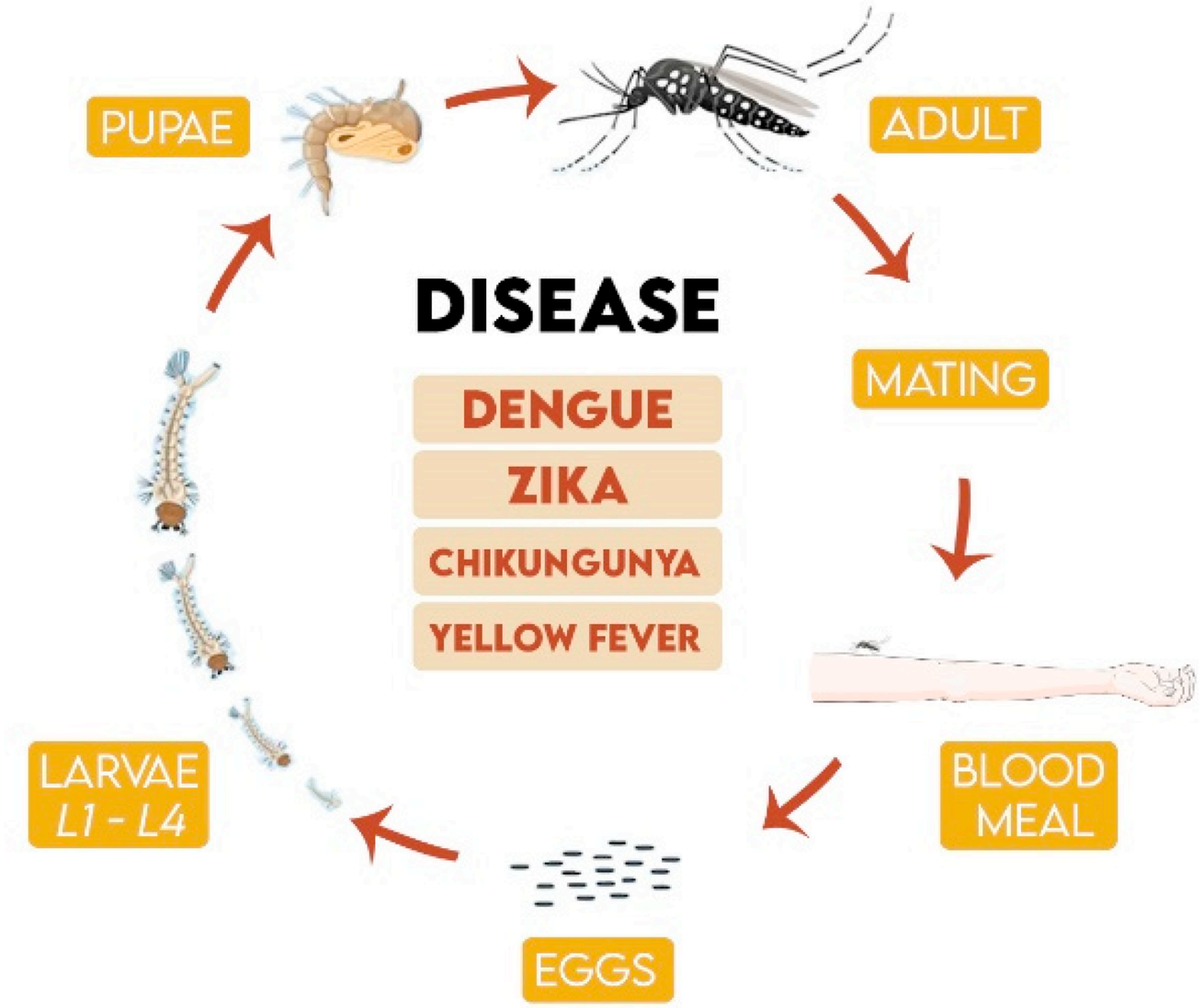
Dataset #1 - Female *Aedes aegypti* originally from Recife, Brazil, now in a colony at the Center for Vector-Borne Infectious Diseases (CVID)

Image: Surfer43, CC-BY-SA3.0 LInk

Female *Aedes aegypti*

Image: James Gathany, Frank Hadley Collins, Public Health Image Laboratory, Public Domain LInk

Aedes aegypti are an important vector of arboviruses, including dengue virus, Zika virus, and yellow fever virus.

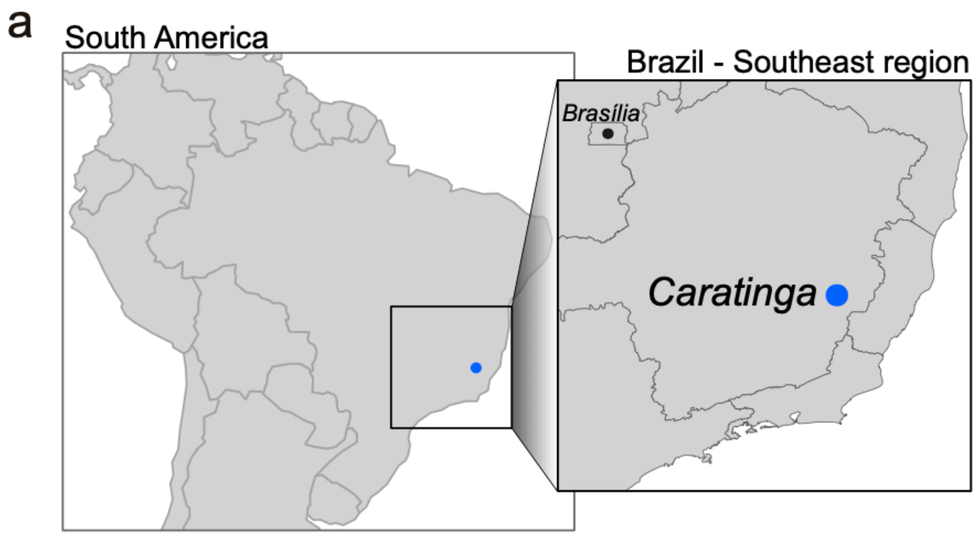


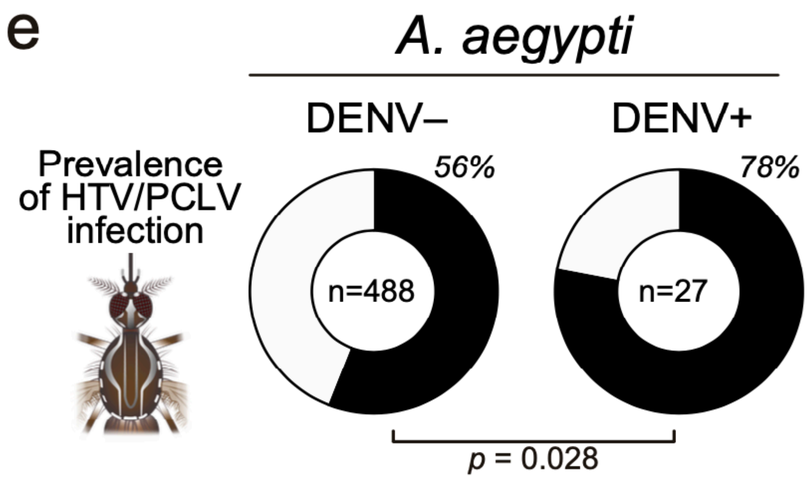
Silvério et al (2020) Molecules

Viruses that persistently infect mosquitoes can change the ability of the mosquito to vector pathogenic viruses

Insect-specific viruses regulate vector competence in *Aedes aegypti* mosquitoes via expression of histone H4

Roenick P. Olmo[1,2]*, Yaovi M. H. Todjro[1]*, Eric R. G. R. Aguiar[1,3]*, João Paulo P. de Almeida[1], Juliana N. Armache[1], Isaque J. S. de Faria[1], Flávia V. Ferreira[1], Ana





bioRxiv preprint doi: https://doi.org/10.1101/2021.06.05.447047; this version posted June 5, 2021.

Dataset #2 - Male *Drosophila melanogaster* from Fort Collins. Pool of wild-caught flies.

Tillie Dunham collecting flies

Male *Drosophila melanogaster*

Image: Darren Obbard LInk

Drosophila melanogaster is an important model organism. Studying host-pathogen interactions in Drosophila has produced fundamental advances including the identification of the role of Toll (-like) receptors in immunity

Cell, Vol. 86, 973–983, September 20, 1996,

The Dorsoventral Regulatory Gene Cassette *spätzle/Toll/cactus* Controls the Potent Antifungal Response in Drosophila Adults

Bruno Lemaitre, Emmanuelle Nicolas, Lydia Michaut, Jean-Marc Reichhart, and Jules A. Hoffmann
Institut de Biologie Moléculaire et Cellulaire
UPR 9022 du Centre National de la Recherche Scientifique
15 rue René Descartes
67084 Strasbourg Cedex
France

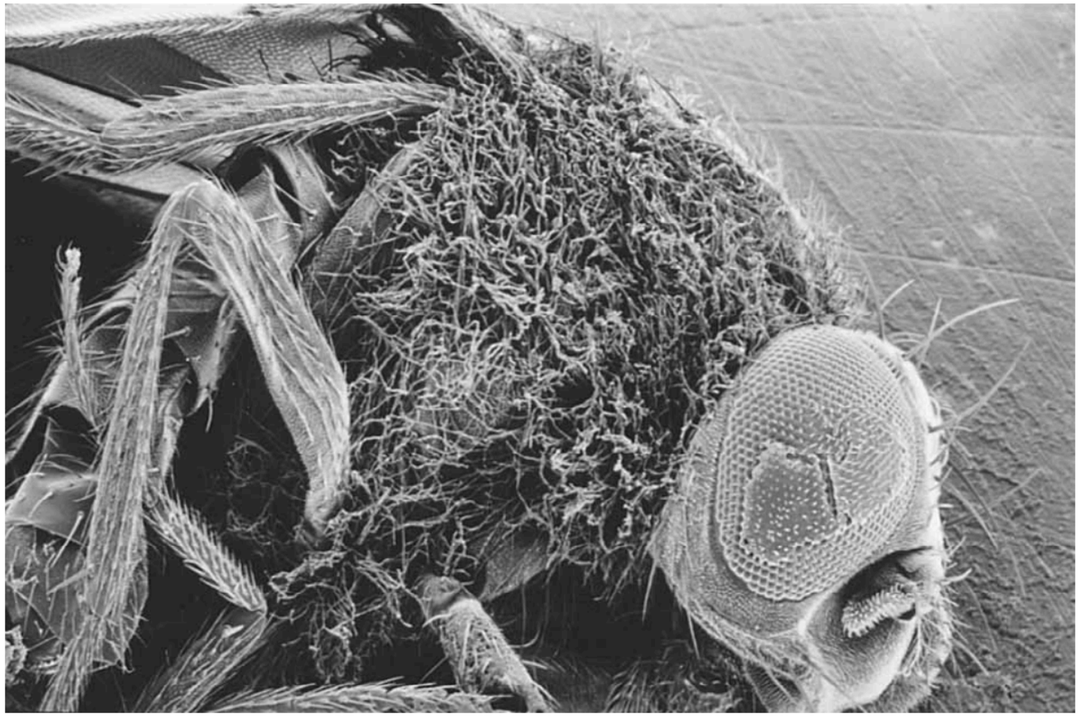


Figure 5. Germinating Hyphes of A. fumigatus on a Dead Drosophila

Scanning electron micrograph of a Drosophila adult that succumbed to infection by A. fumigatus and is covered with germinating hyphae (200× magnification).

Dataset #3 - Female *Aedes aegypti* originally from Guerrero, Mexico, now in a colony at the Center for Vector-Borne Infectious Diseases (CVID)

Image: Surfer43, CC-BY-SA3.0 LInk

Female *Aedes aegypti*

Image: James Gathany, Frank Hadley Collins, Public Health Image Laboratory, Public Domain LInk

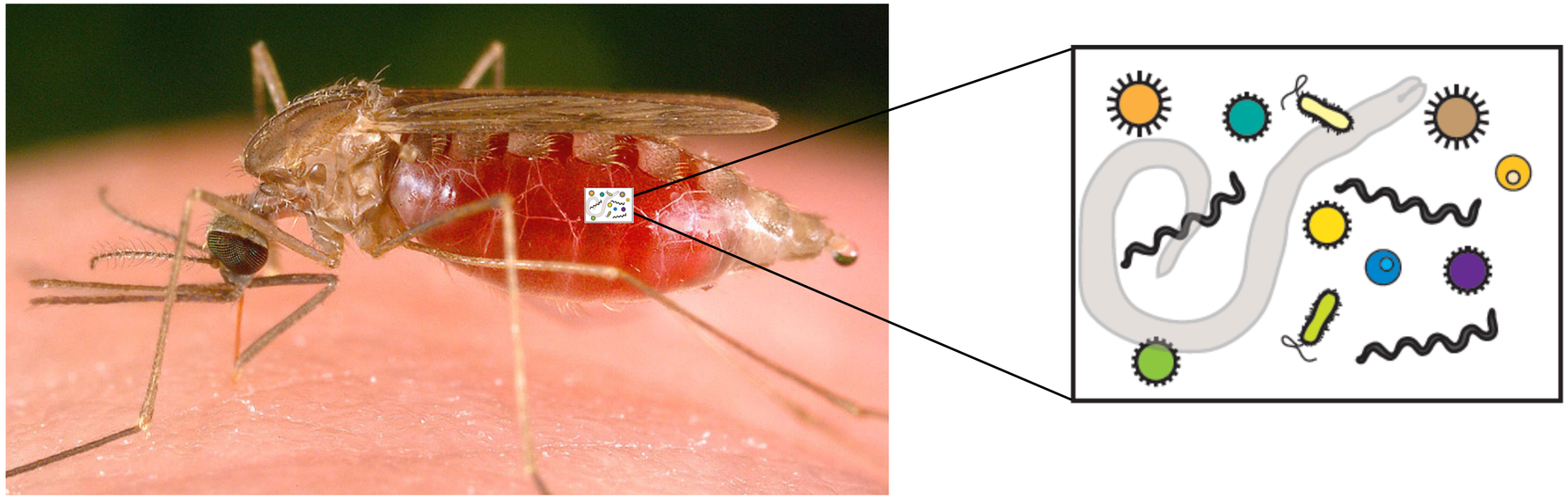
Dataset #4 - *Drosophila virilis* from Fort Collins.
Pool of wild-caught flies of unknown sex.

Tillie Dunham collecting flies

Male *Drosophila virilis*

Image: Darren Obbard LInk

What steps should you take to identify viral sequences in these datasets?



Datasets have between 1.4M and 5M reads

Your goal is to identify and characterize virus sequence(s) in datasets from pools of flies or mosquitoes

- Identify and characterize virus sequences in your assigned dataset using a computational workflow like we did in class.
  - You can restrict your analysis to the 12 longest contigs from assembly of host-filtered reads
  - We will discuss in additional detail what it means to characterize the virus sequences

- Document what you've done in a computational lab notebook, in the format of a GitHub-hosted markdown document.
  - Well-organized
  - Record software used and versions
  - Record code used to accomplish analysis (commands run)
  - Record (copy and paste) key output of commands
  - Record your interpretation of each step in the workflow

- Present your findings to all of us next Thursday in class.

Datasets are located on thoth01 in: /home/data_for_classes/2022_MIP_280A4/final_project_datasets

```
(base) mdstengl@thoth01:/home/data_for_classes/2022_MIP_280A4/final_project_datasets$ ls -lrth
total 4.9G
-rw-r--r-- 1 mdstengl mdstengl 508M Oct 20 13:49 Aedes_Guerrero_R1.fastq
-rw-r--r-- 1 mdstengl mdstengl 1.5G Oct 20 13:49 Aedes_Recife_R1.fastq
-rw-r--r-- 1 mdstengl mdstengl 587M Oct 20 13:49 FoCo_melanogaster_R1.fastq
-rw-r--r-- 1 mdstengl mdstengl 576M Oct 20 13:49 FoCo_virilis_R1.fastq
-rw-r--r-- 1 mdstengl mdstengl 463M Nov 29 09:14 Planococcus_Illumina_R1.fastq
-rw-r--r-- 1 mdstengl mdstengl 463M Nov 29 09:14 Planococcus_Illumina_R2.fastq
-rw-rw-r-- 1 mdstengl mdstengl  18M Nov 29 09:14 Planococcus_Nanopore.fastq
-rw-r--r-- 1 mdstengl mdstengl 443M Nov 29 09:14 Paenibacillus_Illumina_R1.fastq
-rw-r--r-- 1 mdstengl mdstengl 443M Nov 29 09:14 Paenibacillus_Illumina_R2.fastq
-rw-rw-r-- 1 mdstengl mdstengl  21M Nov 29 09:14 Paenibacillus_Nanopore.fastq
```