# SC18 CLOUD HPC HACK

Team chaos

Mentors: Tim & David

# Red flour beetle (Tribolium castaneum)

- Pest that infests stored food grains
- A higher-quality genome assembly could help scientists find better ways to reduce its economic impact
- Project: assemble high-quality genome (~205 Megabases) from ~30 billion bases of sequence reads using the Canu assembler
  - Requires minimum of ~10s of GBs of memory and possibly thousands of CPU hours
  - Data provider expects it to run

Source: wikipedia

# Cloud Computing Resources

- Compute Nodes
  - 10 c5.18xlarge (Intel skylake processor)
    - Parts of the application (Canu) run on a single core, and we want the latest fastest core.
    - AWS cost scales linearly with the virtual cpu and memory, so we are using the biggest c5 instance type.
    - CANU needs big memory for the part of the application, and we don’t exactly know how much of a memory it needs, so we are using the biggest.
- Data Storage
  - 3 TB OrangeFS on General Purpose SSD (gp2) for working data storage
    - 4 File Servers (c4.2xlarge)
    - Canu is I/O intensive
  - Amazon Elastic File System (EFS)
    - Home & software directories accessible from all nodes

# Challenges

- Technical hiccups provisioning CloudyCluster
  - Brandon & crew worked tirelessly to resolve
- Default AWS EC2 instance limit disallowed use of c5.18xlarge instances
  - Contacted AWS support; they upped the limit from 0 to 10

# Progress

- Completed E. coli bacteria test data set assembly in ~21 minutes
- Red flour beetle assembly partially complete
    - Canu has 3 stages; currently in 1st stage

# Future Work

- Tune Canu parameters for better performance on c5.xlarge instances
  - Defaults don’t use all memory and processors on each node
  - Useful for researchers who want to use Canu on AWS
- Complete this red flour beetle genome assembly