GENOME ASSEMBLY WEEK 1

DATA WRANGLING

PRE-COURSE SURVEY

- Please complete this now:
- https://docs.google.com/forms/d/1nA8JDRm_e5PpRMvxWA3Qhb1YJV585cYm9jha4hCW3CE/viewform

OVERALL SCHEDULE

- 1. 4/28: Data wrangling
- 2. 5/5: Resource allocation and configuration files
- 3. 5/12: Monitoring assembly progress and troubleshooting errors
- 4. 5/19: Generating assembly metrics
- 5. 5/26: Comparing multiple assemblies
- 6. 6/2: Visualization and next steps

WORKSHOP EXPECTATIONS

- At the end of the 6 weeks, I hope you:
  - Feel comfortable with the basic steps of the assembly process and able to perform them on your own.
  - Are empowered to make decisions about what kind of assembly software you might need for a given project.
  - Begin to appreciate that this is more of an art than a science!

OUTCOMES

- You will become versed in three* major assemblers using a mixed bag of public data.
- We will perform all the steps from raw data to assembly comparison and visualization.

*If everyone becomes super proficient, we can advance to bigger genomes and additional assemblers.

TEST DATASETS

- NIST (National Institute of Standards and Technology)
  - *Salmonella enterica* RM 8375
    - MiSeq
    - PacBio

WHICH ASSEMBLER DO I NEED?

- Depends on:
  - Data (sequencing platform, libraries)
  - Genome size
- Compute resources at your disposal

ASSEMBLERS

- **SPAdes** (small genomes, any data)
- **DISCOVAR** (Illumina 2X250bp reads only)
- **MaSuRCA** (hybrid: pretty much any data, any genome size)

ASSEMBLIES WE WILL RUN

- Salmonella:
  - DISCOVAR (Illumina only)
  - SPAdes (Illumina only)
  - SPAdes (Illumina and PacBio)
  - MaSuRCA (Illumina and PacBio)

LET'S MOVE TO THE TUTORIAL

- https://github.com/SmithsonianWorkshops/GenomeAssembly/blob/master/week1-data_wrangling.md