2017 de novo Assembly Tutorial

Robert I. Colautti

February 6, 2017

## *de novo* assembly and annotation of NGS data

In this tutorial you will learn how to work with work with next-generation sequencing data. Beginning with short-read sequence data from Illumina, you will run a series of quality assurance and quality control steps, assemble short-reads into contigs, assemble a genome, and then annotate the genome. To keep things simple and computation short, we will assemble a small chloroplast genome (~150kb) but the same principal is used to assemble much larger genomes.

## Exploring short-read data

As noted in lecture, the NCBI's short-read archive (SRA) database is the go-to repository for short-read sequences. These contain the raw output of next-generation sequencers. Data come primarily from Roche 454, Ion Torrent, Illumina, and Oxford Nanopore platforms. You can explore the database at your leisure: <https://trace.ncbi.nlm.nih.gov/Traces/sra/>

Regardless of whether your data comes from the SRA or something you've sequenced yourself, the first step is to assess the quality of your data. Remember the mantra: garbage in == garbage out

sh test.sh  
echo test