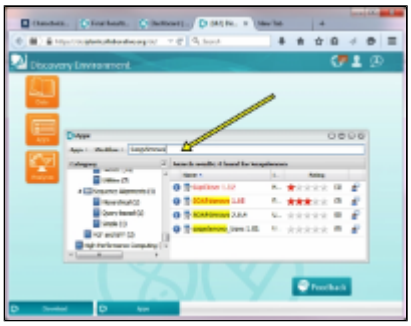
Assemble a Genome Using SOAPdenovo (Workflow Tutorial)



The App Store is currently being restructured, and apps are being moved to an HPC environment. During this transition, users may occasionally be unable to locate or use apps that are listed in our tutorials. In many cases, these apps can be located by searching them using the search bar at the top of the Apps window in the DE. To increase the chance for search success, try not searching the entire app name and version number but only the portion that refers to the app's function or origin (e.g. 'SOAPdenovo' instead of 'SOAPdenovo-Trans 1.01'). In critical cases, please report your concern to the iPlant Ask forum or to support@iplantcollaborative.org. Thank you for your patience.

Tutorial under review

Please work through the tutorial and add your comments on the bottom of this page. Or send comments per email to bmlau@email.arizona.edu. Thank you.

Introduction and overview

Author: Dr. Roger Barthelson, iPlant Collaborative/University of Arizona

Goal

The goal of this tutorial is to gain familiarity with a commonly used procedure for *de novo* whole genome assembly of Illumina reads using the iPlant Discovery Environment (DE).

This procedure will include assembly of paired and unpaired Illumina reads with SOAPdenovo2, followed by an analysis of the assembly quality.

Method

1. Workflow: De novo Assembly I: Genome assembly with SOAPdenovo2.
2. The procedure begins with reads previously trimmed with Scythe to remove extraneous sequence, and Sickle to remove low quality reads and low quality portions of the remaining reads. After assembly with the SOAPdenovo2 App in the DE, the resulting assembly will be analyzed for basic quality statistics, and compared to a reference genome for more in depth analysis of the assembly, specifically for assembly fidelity.

Rationale and Background

***De Novo* Sequencing** A process in which a novel genome is sequenced for the first time and requires specialized assembly of sequencing reads.

For this tutorial the assembler SOAPdenovo2 will be used to assemble the genome. A recommended approach will be followed in testing different kmer settings.

SOAPdenovo SOAPdenovo is a novel short-read assembly method that can build draft assemblies *de novo* for human-sized genomes. The program is designed to assemble Illumina GA short reads. It creates new opportunities for building reference sequences and carrying out accurate analyses of unexplored genomes in a cost effective way. SOAPdenovo aims for large plant and animal genomes, although it also works well on bacteria and fungi genomes.

Prerequisites

1. An iPlant account. (Register for an iPlant account at user.iplantcollaborative.org.)
2. The DE Quick Start tutorial provides an introduction to basic DE functionality and navigation.

Test/sample data

This tutorial uses *Rhodobacter sphaeroides* Illumina sequencing data that are stored in the Data Store in the iPlant Discovery Environment. They were downloaded from the GAGE project, which, in turn, retrieved them from the NCBI Sequence Read Archive (SRA) to test genome assembly methods. The data was trimmed and cleaned up with the applications Scythe and Sickle, as described in a separate tutorial. During the process of trimming, many culled reads left unpaired mates behind, which were moved into a separate single-reads file. The data to be used in this tutorial is available in the Data Store at Community Data > iplant_training > genome_assembly_soapdenovo > A_Assemble_Reads.

More information on the GAGE project is available here.

Results

Application	Output	Output Type	Content
SOAPdenovo2	*.scafSeq	fasta format text file	scaffold fasta file, which represents the final output for the assembly sequences (includes large gaps filled with N's)
SOAPdenovo2	*.contig	fasta format text file	contig fasta file, the final assembly of continuous sequence (only small gaps)
SOAPdenovo2	*.scafStatistics	text	text formatted list of statistics about the scaffold sequences including the N50 value
Assess assembly	*.report	text	text formatted list of statistics about the scaffold sequences (or any input fasta file), including N50, but also comparisons to a reference genome (e.g. representation value)

Workflow

A. Assemble reads (app: *SOAPdenovo 2.0.4*)
B. Assess assembly (app: *Assess assembly vs whole genome*)

Approximate analysis durations for the iPlant sample data are provided in each step. Other datasets, depending on size, could take less or more time. Using the sample data, users can skip through the workflow (a la 'cooking show'), returning later to examine the results of their own analysis.