iPlant Collaborative™

Cyberinfrastructure for the Life Sciences

Fast and free Data Storage, Computation, and Bioinformatics Platforms for Research

DNA Subway: A Simple, Powerful Bioinformatics Workflow for RNA-Seq Analysis and Distributed Genome Annotation

J. Williams*,†, S. McKay‡, M. Khalfan*,†, C. Ghiban*,†, U. Hilgert†,§, Sue Lauter*,†, Eun-Sook Jeong*,†, and D. Micklos*,†
*Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, †iPlant Collaborative, T.W. Keating Bioresearch Building, ‡Ontario Institute for Cancer Research, MaRS Centre, Toronto, ON, Canada, §BIO5 Institute, T.W. Keating Bioresearch Building, U. Arizona, Tucson, AZ

DNA Subway - Classroom Friendly Bioinformatics

DNA Subway bundles research-grade bioinformatics tools, high-performance computing, and databases into workflows with an easy-to-use interface. "Riding" DNA Subway lines, students can predict and annotate genes in up to 150kb of DNA (Red Line), identify homologs in sequenced genomes (Yellow Line), identify species using DNA barcodes and phylogenetic trees (Blue Line), and examine RNA-Seq datasets for differential transcript abundance (Green Line). With support for plant and animal genomes, DNA Subway engages students in their own learning, bringing to life key concepts in molecular biology and genetics. DNA barcoding and RNA extraction wet-lab experiments support a variety of inquiry-based learning experiences using student-generated data. Products of student research can be exported, published, and utilized in follow-up experiments.

Yellow Line

- Probe whole (unmasked) plant genomes
- Search for transposons/elements with TARGeT
- View sequence alignments for match results
- View trees of match results

Blue Line

- View and process sequences from trace files
- Identify species from DNA barcodes
- Generate alignments and phylogenetic trees
- Export DNA sequences to GenBank

RNA-Seq and High Performance Computing

The Green Line supports genome-scale science in the classroom. Users can store data up to their iPlant allocation (~100GB) and run analyses using iPlant APIs to access XSEDE (eXtreme Science and Engineering Discovery Environment) supercomputing resources.

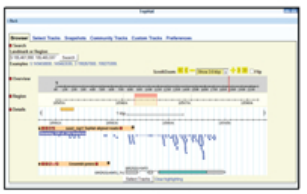
Project Creation

Projects begin by selecting from supported Plant/Animal genomes. Users work with their own transcriptome data or use sample data provided. Version 1 of the Green Line supports differential expression experiments using single reads. Support for paired end reads and transcriptome assembly is in progress

Manage Data/Quality

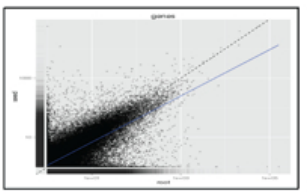
Select FASTQ reads from your *iPlant* Data Store. Quality assessment and quality-based filtering of sample data use programs from the FASTX toolkit. The Green Line allows users to run analyses using a "Basic" mode set with reasonable defaults, or adjust parameters using an "Advanced" mode.

Assemble and Align



TopHat aligns reads to a reference genome in a splice-aware fashion. Results are examined directly in GBrowse. Cufflinks uses aligned reads to infer transcripts, and an independent transcriptome assembly allows novel genes and isoforms to be discovered.

Examine Abundance



At the CuffDiff stop, users select combinations of read datasets to test for differential expression; a variety of graphs from the CummeRbund package and a sortable list of transcripts with annotation and quantitation are available and exportable.

Building Capacity for Maize Annotation

"Students are overwhelmed by their first introduction to genome sequences viewed on a genome browser. Students who used DNA Subway needed little or no guidance when they moved on to use MaizeGDB and had an easier time transitioning to genomes depicted in different genome browsers"[1]

Developing Red Line and Web Apollo

DNA Subway's Green Line and Red Line updated with WebApollo will offer a robust way for distributed classroom projects annotate genomes. The number of sequenced genomes will outpace the available curators for the foreseeable future. Pushing the development of crowd-sourced models can engage students and contribute to science.

Connecting Students to Data

Students have limited patience for computer work and want a wet bench "hook."[2] Advances in technology mean that RNA-Seq datasets are now within classroom reach. A 2014 Working Group at Cold Spring Harbor Laboratory's DNA Learning Center will develop datasets from Maize and other organisms as mainstream components of undergraduate education.

Student Generated RNA-Seq Data

↓

Red Line and Existing Annotations

↓

Community Validated Annotations

The *iPlant Collaborative* is funded by a grant from the National Science Foundation Plant Cyberinfrastructure Program (#DBI-0735191 and #DBI-1265383)

References:
1. Brent Buckner, Truman State University, MO. Personal Communication
2. iPlant Genomics in Education Faculty Workshop, Washington University of St. Louis - 2006