DNA Subway – An Educational Bioinformatics Platform for Gene and Genome Analysis: DNA Barcoding, and RNA-Seq

***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

ABSTRACT: *DNA Subway* is an educational bioinformatics platform developed by the *iPlant Collaborative* (NSF #DBI–0735191). *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. *DNA Subway* is freely accessible online at dnasubway.iplantcollaborative.org.

Keywords: DNA barcoding, RNA-Seq, Undergraduate education

Introduction

High-throughput sequencing (HTS) and the related progress of computational biology have revolutionized nearly every aspect of life science investigation. However, the transition of this technology into undergraduate classrooms faces many obstacles. The sense that much of the undergraduate biology curriculum is in need of an update is summarized in the National Research Council's *BIO2010* report: "In contrast to biological research, undergraduate biology education has changed relatively little during the past two decades. The ways in which most future research biologists are educated are geared to the biology of the past, rather than to the biology of the present or future" (NRC (2003), p.1).

Updating curricula in light of new technologies can be challenging given the speed at which technologies like HTS advance. Additionally, textbooks and professional development resources needed to equip educators with the knowledge, tools, and confidence to address new topics necessarily take additional time to develop. For HTS applications in particular, teaching bioinformatics using genome-scale datasets depends on resources (e.g., software, high-performance computing, and data storage) that are often a limiting factor, both in availability and expertise.

Fortunately, advances in technologies and good timing have produced promising solutions to these challenges. The cost of HTS has become reasonable – more than 1000-fold reduction since 2004 (NHGRI (2013) – and the amount of data freely available for students presents real opportunities for them to contribute to a biology paradigm that operates along a continuum of research and education. This paper outlines how *DNA Subway* and other *iPlant* related resources enable educators to take advantage of these opportunities while bringing HTS to their students.

Driven by educational design principles. *DNA Subway* was conceived to address a need not just for powerful tools, but a "classroom friendly" user interface – the lack of which is an acknowledged barrier to bioinformatics instruction (Cummings and Temple (2010)). In 2006, the *iPlant Collaborative* held a meeting on "Genomics in Education" at Washington University in St. Louis, at which 44 faculty identified guiding requirements to shape the development of *iPlant's* educational platforms: 1) *Mix lecture and lab* – students have limited patience for computer work and want a wet bench "hook"; 2) *Enable student-scientist partnerships* – someone has to care about the data generated by students; 3) *Co-investigation* – projects should potentially lead to publications; and 4) *Scale* – platforms should support distributed projects that multiple classrooms can join. From these principles, 25 collaborators at 11 institutions aided the development of *DNA Subway*. To accomplish our objectives, we integrated existing open-source tools (and additional novel software to fill gaps such as viewing trace files, alignments, or file export utilities) into an approachable graphical user interface. The purpose was to make it possible for educators to prepare streamlined lessons utilizing the same tools and data that are accessible to researchers. Pipelines based on standard bioinformatics processes were progressively released for genome annotation (Red Line; launched 2010), identification of gene homologues and non-coding DNA using *TARGeT* (Yellow Line; 2010) (Han, Burnette, and Wessler (2009)), DNA barcoding and phylogenetics (Blue Line; 2011), and RNA-Seq analysis of transcript abundance (Green Line; beta launched 2013). *DNA Subway* makes more than 30 tools available through a web interface that mitigates pedagogical barriers to adoption, such as moving data (and students) from one platform to another, or using tools that

primarily exist as command line/Linux applications. Wherever possible, tools run using reasonable defaults, increasing the likelihood students can derive scientifically valid results while abstracting away complexities not central to an introductory treatment of bioinformatics.

Addressing limited access to computing resources. Computing infrastructure is a major barrier to addressing bioinformatics education, particularly at institutions for underserved faculty at community colleges and primarily undergraduate institutions (Cummings and Temple (2010)). *DNA Subway* was built using *iPlant*-developed resources; any student or educator can obtain a free *iPlant* account, which includes storage allocations that allow work to be saved automatically. *DNA Subway* is accessed through a web browser and analyses take place on non-local computing resources that support multiple concurrent users/classrooms. For the RNA-Seq analysis pipeline (Green Line) users can store very large datasets (up to ~100GB) in the *iPlant* Data Store. Compute for RNA-Seq uses *iPlant* application programming interfaces (APIs) to run jobs on XSEDE (eXtreme Science and Engineering Discovery Environment) – this means student access to supercomputing can be enabled by a broader field of biology educators lacking highly specialized programming expertise.

Educational applications relevant to animal and livestock researchers. This paper features possible applications of *DNA Subway* of interest to the animal breeding community – identification of foods and food ingredients by DNA barcoding (Blue Line) and analysis of RNA-Seq data (Green Line). Both applications involve wet lab exercises and student-generated data. Step-by-step documentation is available on the *DNA Subway* website, and a general overview is presented in Methods.

Methods

Method 1: DNA Barcoding – Blue Line

1.1 Barcoding for species identification – Blue Line. DNA barcoding involves DNA extraction then PCR amplification of unique (barcode) regions of sequence with universal primers (Hebert and Gregory (2005)). Amplification of mitochondrial Cytochrome c oxidase subunit 1 (*COI*) generates sequence suitable for high confidence identifications of unknown animal samples. This technique has already been identified as a helpful tool in food safety (see Yancy, Zemlak, Mason, et. al. (2008)). Wet lab steps are fairly simple – high school students have published results of barcoding foods such as sushi and tea (Stoeckle et al. (2011), and *DNA Subway* has already been used to support hundreds of distributed student research projects (see www.urbanbarcodeproject.org).

1.2 Sample collection, DNA extraction, and sequencing. DNA is extracted from a sample such as plant leaves, meat, or processed food. When collecting samples, users wishing to submit resultant sequence data to GenBank via *DNA Subway* directly or via BOLD (www.boldsystems.org) should become familiar with required metadata (see www.dnabarcoding101.org). Any extraction protocol producing template suited for PCR amplification is acceptable; our suggest protocols and primer sets are on the *DNA Barcoding 101* website. The online protocol includes instructions to send samples for discounted PCR cleanup and sequencing at GENEWIZ Inc. (www.genewiz.com). GENEWIZ data is typically available on *DNA Subway* in 24–48 hours; users may also start projects with their own trace (AB1) files or FASTA files.

1.3 Create *DNA Subway* project and upload data. Log into *DNA Subway* (www.dnasubway.org) and click the blue *Determine Sequence Relationships* square. Select the project type (*COI* for animals or *rbcL* for plants). If sequencing was performed at GENEWIZ users can access their data by selecting "Import trace files from DNALC." Otherwise users upload AB1 trace files or upload/paste FASTA sequences. Projects may also be created using sample sequences provided.

1.4 Assemble sequences – *Sequence Viewer* and *Sequence Trimmer*. Using *Sequence Viewer* (Figure 1) users can view trace files and nucleotide sequences; poor quality sequences are flagged for inspection. Individual quality scores are displayed for each nucleotide. Sequences can also be downloaded. *Sequence Trimmer* automatically removes low quality sequence at either end of the reads.

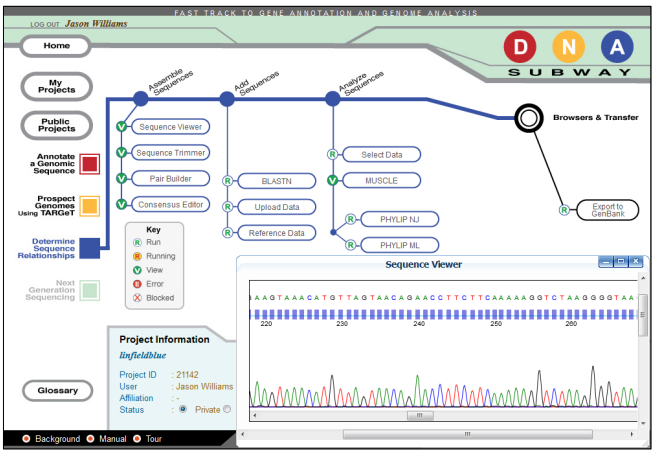

Figure 1. Viewing sequence trace files on the Blue Line.

1.5 Assemble sequences – *Pair Builder* and *Consensus Builder*. *Pair Builder* allows specified or automatic annotation of bidirectional reads belonging to the same PCR amplicon, as well as sequencing orientation (forward/reverse). *Consensus Builder* generates consensus sequences, highlights discrepancies, and allows additional editing.

1.6 Add sequences – *BLASTN*, *Upload Data*, and *Reference Data*. *BLASTN* returns formatted BLAST results from NCBI. Results for particular species are linked to automatically-retrieved images and other online content. Additional sequences can be added to a project from the *BLASTN*, *Upload Data*, or *Reference Data* "subway" stops; data from these stops aid in creating phylogenetic trees.

1.7 Analyze sequences – *MUSCLE* multiple alignment and *PHYLIP* tree generation. *Select Data* allows users to indicate sequence data to create an alignment or tree; data can also be exported from this stop. *MUSCLE* stop is used to generate a multiple alignment, and supports automated trimming. *PHYLIPNJ* and *PHYLIPML* use the trimmed alignment to generate neighbor-joining and maximum-likelihood trees respectively.

1.8 Export sequences to GenBank and share projects. At *Export to GenBank* novel, high quality sequences can be exported directly to GenBank; valid sequences and metadata are published to a "BioProject" at NCBI (Accession: PRJNA206193). Projects can also be shared with other users by selecting the *Public* button.

Method 2: RNA-Seq Differential Expression – Green Line

2.1 RNA-Seq using Green Line. Genome-wide association studies (GWAS) and annotation are just some of the RNA-Seq applications of interest to animal breeders (Pérez-Enciso and Ferretti (2010)). The Green Line workflow is based on the Tuxedo Protocol (Trapnel, Roberts, Goff, et al. (2012)) and can identify significant expression differences between samples and also support Red Line annotation projects. Log into *DNA Subway* (www.dnasubway.org) and click the green *Next generation sequencing* square. Users upload FASTQ files into the *iPlant* Data Store or select a species and use a sample dataset provided at the *Manage Data* stop.

2.2. Data management and quality control. Quality assessment and quality-based filtering of sample data use programs from the FASTX toolkit (Hannon Lab (2010)). Throughout the Green Line, users can select a "Basic" mode with reasonable defaults, or adjust parameters using an "Advanced" mode.

2.3 Analyze transcriptome. Reads are aligned to a reference genome using *TopHat*, and assembled using *CuffLinks*; results are visualized locally in *GBrowse* and available in the *iPlant* Data Store. 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.

2.4 Export to Red Line. RNA-Seq data is a valid class of evidence for genome annotation. A GFF file is generated and can be exported to a Red Line project.

Discussion and Evaluation

Educators can use *DNA Subway* to complete a DNA barcoding workflow that would normally require several independent software and web applications. Ease of use is especially important for more complex analyses, such as RNA-Seq. The Green Line user interface for example is not dramatically more complicated than the Blue Line; however it incorporates a sophisticated system of scripts and wrappers to migrate data through software originally designed for command line, using *iPlant* APIs to push jobs and results between high-performance computing systems and the *iPlant* Data Store. To date, we are unaware of any educational platform for biology that furnishes such easy access to high-performance computing. We deliberately committed to a simple interface design for *DNA Subway* while acknowledging that teaching bioinformatics by focusing on tools risks downplaying the computational (Pevzner (2004)) and biological concepts behind how and why we use these tools. However, user feedback and evaluation supports our approach that starting students out with easy-to-use tools, such as our integrated genome browsers, significantly increases student confidence in approaching the same tools elsewhere: "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" (Brent Buckner, Truman State University, pers. comm.).

While *DNA Subway* is a valid tool for research, we anticipate that its widest application will be to lower the barriers to entry that would keep some educators and students from attempting to tackle these sophisticated (but highly relevant) bioinformatics workflows. Since 2010 we have trained more than 1500 educators in the US and abroad, and we continue to produce learning materials that assist educators to utilize *DNA Subway* in their teaching. Long-term evaluations of attendees of our two-day "Genomics in Education" workshops demonstrate significant adoption – a 2013 survey revealed 43% educators trained in *DNA Subway* use had adopted it into their teaching, with over 24,000 student exposures; in addition, 67% had shared the resource with other educators.

Literature Cited

Cummings, M.P., and Temple, G.G. (2010). *Brief. Bioinform*. 11(6):537–543.

Han, Y., Burnette, J.M., and Wessler, S.R. (2009). *Nucl. Acids Res.* 37(11):e78.

Hannon Lab. (2010). *Website:* http://hannonlab.cshl.edu/fastx_toolkit/index.html. Accessed Feb. 2014.

Hebert, P.D.N., and Gregory, R.T. (2005). *System. Bio.* 54(5):852–859.

National Research Council. (2003). *BIO2010*. Nat. Acad. Press, Washington, D.C.

National Human Genome Research Institute. (2014). *Website*: www.genome.gov/sequencingcosts/. Accessed on Feb. 2014.

Pevzner, P. (2004). *Bioinform*. 20(14):2159–2161.

Pérez-Enciso, M., and Ferretti, L. (2010). *Anim. Genet.* 41(6):561–569.

Stoeckle, M.Y., Gamble, C.C., Kirpekar, R., et al. (2011). *Sci. Reports.* 1(42):1–7.

Trapnel, C., Roberts, S., Goff, L., et al. (2012). *Nat. Protoc.* 7:562–578.

Yancy, H.F., Zemlak, T.S., Mason, J.A., et al. (2008). *J. Food Protect.* 1(8):210–217.