DNALC Live

Barcoding Bioinformatics
Part I

Jason Williams
Cold Spring Harbor Laboratory, DNA Learning Center
williams@cshl.edu
@JasonWilliamsNY
DNALC Live

This is an experiment, give us feedback on what you would like to see!
DNALC Live

- Provide genetics, molecular biology, and bioinformatics learning resources
- Laboratory and computer demos, short online courses for middle school, high school, and the general public
- Interviews with scientists, help for teachers
- At-home activities, social media contests, and more
DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live
DNALC Website and Social Media

youtube.com/DNALearningCenter
facebook.com/cshldnalc
@dnalc
@dna_learning_center
Licensed Centers
Programs Modeled on the DNALC
Teacher Training Sites (US States and Countries)
Cold Spring Harbor Laboratory
Barcoding Bioinformatics
Part I
Who is this course for?

- **Audience(s):** US AP Biology (high school grades 10-12) AND Intro undergraduate biology
- **Format:** 3 sessions (1 per week); ~ 45-60 minutes each
- **Exercises:** Follow along with our online bioinformatics tool DNA Subway
- **Learning resources:** Slides and packet available (teachers can also request the teacher edition)
Course Learning Goals

• Learn how DNA can be used to identify unknown organisms

• Understand how we obtain DNA Sequence and access its quality

• Use BLAST* to compare an unknown DNA Sequence to known sequences

• Compare DNA Sequences using phylogenetics

*AP Bio (Lab 3 – Comparing DNA Sequences)
Lab Setup

- We will be using DNA Subway – You can get a free account at cyverse.org (optional)
Barcoding Bioinformatics

Part I

(Background and Sequence Quality)
Steps for today’s session

• Introduction to Bioinformatics

• Get background on DNA Barcoding

• Learn about our example experiment

• Start an experiment

• Examine DNA sequence quality
Introduction to Bioinformatics
What is Bioinformatics?

In biology, **Bioinformatics** is an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing **biological data**. A major activity in bioinformatics is to develop software tools to generate useful biological knowledge.

Bioinformatics is about data
Bioinformatics is about data
DNA Structure
Bioinformatics is about data

Often, we are speaking about data in biology, we are talking about DNA Sequence
Bioinformatics is about data

Often, when we are speaking about data in biology, we are talking about DNA Sequence (there are lots of other data, we just won’t be talking about that today)
DNA Sequence

Quick Tour

- Corona Virus Genome: https://mra.asm.org/content/9/11/e00169-20
Introduction to Barcoding
How can you identify an organism using just its DNA?
What is DNA Barcoding?
Steps to DNA Barcoding

Organism is sampled                DNA is extracted            “Barcode” amplified

ACGAGTGGTACGCTGCCCTCTGACTGCATCGAA
TTGCTCCCTACTACGTGCTATATGCGCTTACGAT
CGTACGAAAGATTATAGAATGCTGCTACTGCTCC
CTTATTCGATAACTAGCTCGATTATAGCTACGATG

Sequenced DNA is compared with DNA in a barcode database
The Start of Barcoding
How many species can you name?

How many Animals did you name?
  How many mammals?
  How many plants?
  How many insects?

“Cat”  *Felis catus*

“How many”  *Canis lupus familiaris*

“Oak Tree”  *Quercus alba*

“Shark”  *Ginglymostoma cirratum*

“Beetle”  *Popillia japonica*
Problem 1: No one know how many species there are.
How many species are there?

<table>
<thead>
<tr>
<th>Vertebrates</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammals</td>
<td>5,490</td>
</tr>
<tr>
<td>Birds</td>
<td>9,998</td>
</tr>
<tr>
<td>Reptiles</td>
<td>9,084</td>
</tr>
<tr>
<td>Amphibians</td>
<td>6,433</td>
</tr>
<tr>
<td>Fishes</td>
<td>31,300</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>62,305</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Invertebrates</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insects</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Mollusks</td>
<td>85,000</td>
</tr>
<tr>
<td>Crustaceans</td>
<td>47,000</td>
</tr>
<tr>
<td>Corals</td>
<td>2,175</td>
</tr>
<tr>
<td>Arachnids</td>
<td>102,248</td>
</tr>
<tr>
<td><strong>Total (+others)</strong></td>
<td>1,305,250</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Plants</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angiosperms</td>
<td>281,821</td>
</tr>
<tr>
<td>Gymnosperms</td>
<td>1,021</td>
</tr>
<tr>
<td>Ferns and Allies</td>
<td>12,000</td>
</tr>
<tr>
<td>Mosses</td>
<td>16,236</td>
</tr>
<tr>
<td>Green and Red Algae</td>
<td>10,134</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>321,212</td>
</tr>
</tbody>
</table>

- There are currently between 1.5 and 2 million described species
- It is estimated that this number may represent as little as half of the true number of species
- Perhaps more than 1/3 of all species are threatened
  
  *(IUCN Red list version 2010.1)*
Problem 2: Even though there are millions of species, there is also a lack of agreement on what a “species” means.
Defining what species are is a complex task

Dependent on many factors

• Interbreeding capabilities
• Morphological variation
• Ecological context
• Genetic similarities
Problem 3: Current taxonomic methods may be inadequate (or at least too slow) to capture vanishing biodiversity
Traditional taxonomies

Leaves alternate proximally, opposite and ultimately decussate distally, 6–16 × 4–13 cm; petiole ca. as long as blade, winged, base clasping, basal lobes stipulate, growing as extensions of wings, less than 1 mm wide; blade 5–7-veined, ovate, glabrous, base typically sagittate, margins entire, apex acute to acuminate. Staminate inflorescences axillary, 1–2 per axil, paniculate, fasciculate; panicles bearing flowers singly, bracteolate, in a zigzag pattern along rachis, internodes less than 2 mm; rachis to 25 cm, secondary axes 1–3(–6), fasciculate, less than 3 cm, each subtended by deltate-ovate bracteole shorter than 1 mm. Pistillate inflorescences solitary, 4–8(–20)-flowered, 6–35 cm, internodes 1 cm

The body form ranges from hemispherical (e.g., Cleidostethus) to elongate oval (e.g., Clypastraea) to latridiid-like (e.g., Foadia). Corylophids are typically dull brown, but some species have contrasting yellowish-brown patches on the pronotum or elytra. The integument is often densely punctured and may be glabrous or bear short, fine recumbent setae. Most corylophid adults can be diagnosed using the following morphological features: Maxilla with single apical lobe; Mesotrochanter short and strongly oblique; Head usually covered by pronotum; Frontoclypeal suture absent; Antennae elongate with 3-segmented club; Procoxal cavities closed externally; Tarsal formula 4-4-4; Pygidium exposed
Sequencing is less complex

Leaves alternate proximally, opposite and ultimately decussate distally, 6–16 × 4–13 cm; petiole ca. as long as blade, winged, base clasping, basal lobes stipulate, growing as extensions of wings, less than 1 mm wide; blade 5–7-veined, ovate, glabrous, base typically sagittate, margins entire, apex acute to acuminate. Staminate inflorescences axillary, 1–2 per axil, paniculate, fasciculate; panicles bearing flowers singly bracteolate, in a zigzag pattern along rachis, internodes less than 2 mm; rachis to 25 cm, secondary axes 1–3(–6), fasciculate, less than 3 cm, each subtended by deltate-ovate bracteole shorter than 1 mm. Pistillate inflorescences solitary, 4–8(–20)-flowered, 6–35 cm, internodes ca. 1 cm.

>Dioscorea alata (matK) gene, partial

>ATTTAAATTATGTGTCAGATATATTAATACCCCATCCCATCCATCTGGAAATCCTGGTTCAAATACTTCAATGCTGGACTCAAGATGTTTCCTCTT
TGCTCATTTATTGCGATTCTTTCTCCACGAATATCATAATTCGAAT AGTTTCATTACTCCGAAAAACCTATTTACGTGATTTCAATTTCAAAAGAAA
ATAAAAGATTTTTTCGAT TCCTATATAATTCTTATGTATTTGAATGTGAATTTGTATTAGTTTTTTTTCATAAGCAATCCTCTTATTT ACGATCAA
GGTCCTCTGGAGTCTTTCTTGAGCGAACACATTTCTATGGAAAAATGGGGCATTTTTTAGTAGTGTGTTGTAATTATTTTCAGAAGACCCAATG
GTTCTTCAAAGATCCTTTTCTGCATTATGTTCGATATCATAGAAAAGCAATTCTGGTGTCAAAGGGAACTCGTCTTTTGATGAGGAAATGGAGA
TCTTACCTTGTCCATTTTTGGCAATATTATTTTCAATTTTGGTCTCATCCGCATAGGATTCATATAAACCAATTATCAAATTATTCCTTCTGTTTTC
TGGGTTATCTTTCAAATGTACTAATAAATTTTTCCGTGGTAAGGAGTCAAATGTTAGAAAATTCATTTGTAATAGATACTCTTACTAAGAAATT
TGATACCAGAGTTTCTAGTTATGATCGTGAGTTTCTTTTTCTTCGGTAAAGGAGCTAAATATAATTTTTGATAGTAAGATAGATAGATCTTAAATCGG
TACATCGATATCGATATCGATATCG

Complex and somewhat objective

Simple (A,T,G, or C) and objective
Lab: Mosquito Identification
Can we tell the difference between larvae that look (nearly) identical?
<table>
<thead>
<tr>
<th>Aedes adult</th>
<th>Anopheles adult</th>
<th>Culex adult</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Aedes larva</th>
<th>Anopheles larva</th>
<th>Culex larva</th>
</tr>
</thead>
</table>

Photograph by Michele M. Cutwa, University of Florida.
Why does this matter?

**Aedes:**
- Chikungunya
- Dengue fever
- Lymphatic filariasis
- Rift Valley fever
- Yellow fever
- Zika

**Anopheles:**
- Malaria
- Lymphatic filariasis

**Culex:**
- Japanese encephalitis
- Lymphatic filariasis
- West Nile fever
Why does this matter?

Estimated range of *Aedes aegypti* and *Aedes albopictus* in the United States, 2016*

*Aedes aegypti* mosquitoes are more likely to spread viruses like Zika, dengue, chikungunya than other types of mosquitoes such as *Aedes albopictus* mosquitoes.

- These maps show CDC’s best estimate of the potential range of *Aedes aegypti* and *Aedes albopictus* in the United States.
- These maps include areas where mosquitoes are or have been previously found.
- Shaded areas on the maps do not necessarily mean that there are infected mosquitoes in that area.

*Maps have been updated from a variety of sources. These maps represent CDC’s best estimate of the potential range of *Aedes aegypti* and *Aedes albopictus* in the United States. Maps are not means to represent risk for spread of disease.*

Lab: DNA Sequencing Background
DNA Sequencing

Cycle Sequencing

To this mix, we also add a second type of nucleotide: one that has a slightly different chemical formula. These "dideoxynucleotides (ddNTP)" can be recognized by a DNA sequencer.

https://dnalc.cshl.edu/resources/animations/cycseq.html
Some Anopheles DNA...

Anopheles gambiae isolate 10016 5.8S ribosomal RNA gene, partial sequence; internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence

GenBank: MK592083.1

ACACATGAACATTGATAAAGTGAACGCATATGGCGCATCGGACGTTTTAATTCGAGCCACGATGCAACAATTC
TTGAGTGCCCTACTAAATTACCAAAAGTCTCATTTATTAGTTAACTACAGTGCGCTCCGCAAGGTGCGCCGCT
ATCCGACGACTGGCGCTCCTGCTGCTCTAAATGACGCTGTTGGCTCCCGTGCTGGTCCTCGGCGCTTG
AAAGTGCGACCTCTCGAGCGATGATTGGATCGGTTTCGTTTTGTTGGGTGTTGTGTTTGATCGGTAGGGGTTGG
TGTCGCTCAAGCGCACTGGTTCCGAACATGCTAAGCTCGTCTCCCGATGCCACCGCGAGTCTACTCCTCCA
GGTCAGGTGCTCGCTCTAGGGATTCCGAAAGCTAAGTCGCTGTAACTCATGTGGGCCCATACACCGCG
TTGCGCTACCCAGCGTAAGTTAGCCTACATACCAAGCCATCAACCCACCGGACCGGCGTAGCTGTAATAC
TTACGTCTCGGTTATACACCGTAGGCTCAAGTGATGTGCACTACCC
Lab: Creating a DNA Subway Project
(follow along in the packet)
Working on DNA Subway Blue Line
Key Concept: Data Quality
Some sequence examples...

High Quality Sequence

Acceptable Quality Sequence

Low Quality Sequence (multiple base calls per position)

Low Quality Sequence (no base calls)
Phred scores...

<table>
<thead>
<tr>
<th>Phred Score</th>
<th>Error (bases miscalled)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1,000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>
Next time:
Comparing sequences with BLAST
DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live