



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

Barcode Bioinformatics

Part I

Jason Williams

Cold Spring Harbor Laboratory, DNA Learning Center

williams@cshl.edu



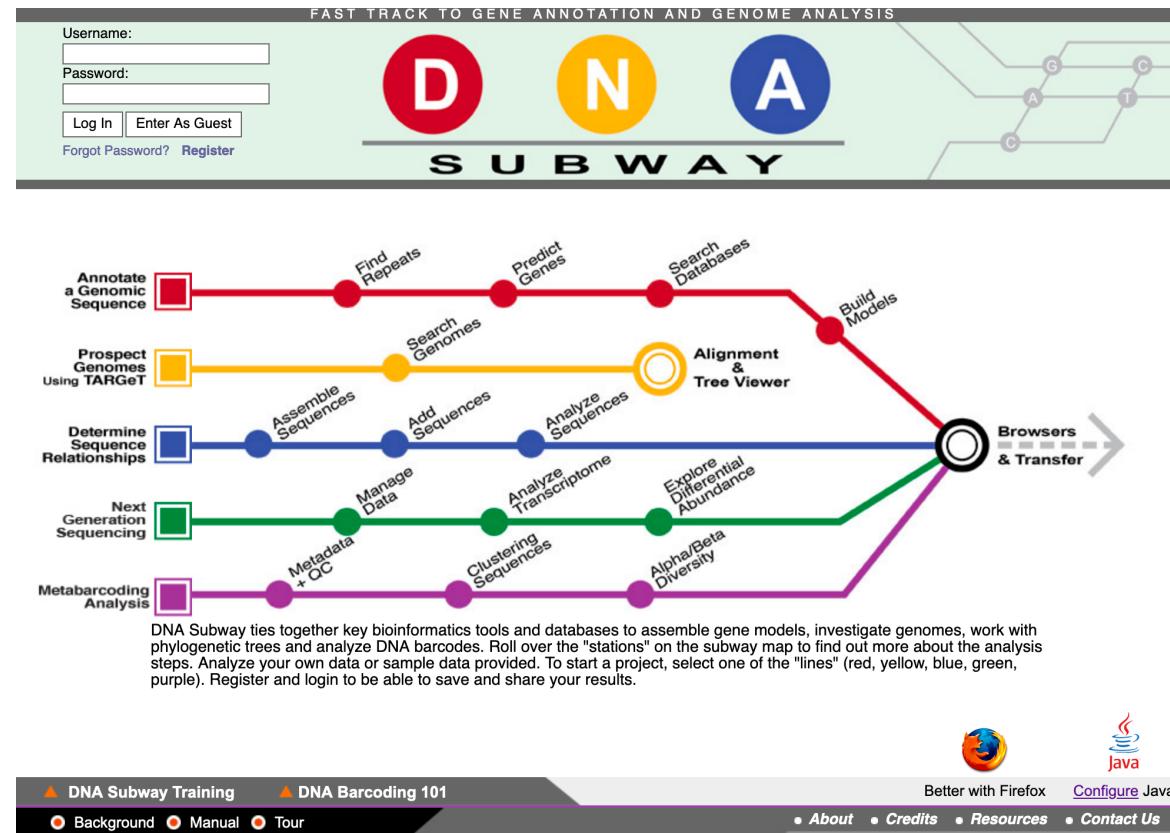
[@JasonWilliamsNY](https://twitter.com/JasonWilliamsNY)



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

Lab Setup

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



Barcode Bioinformatics

Part I

(Background and Sequence Quality)



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



National
Center for
Biotechnology
Information

Quick Tour

- NCBI Homepage <https://www.ncbi.nlm.nih.gov/>
- Human Genome: <https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml>
- Corona Virus Genome: <https://mra.asm.org/content/9/11/e00169-20>



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

Introduction to Barcoding

How can you identify an organism using just its DNA?

What is DNA Barcoding?



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

Steps to DNA Barcoding



Organism is sampled



DNA is extracted



“Barcode” amplified

ACGAGTCGGTAGCTGCCCTCTGACTGCATCGAA
TTGCTCCCTACTACGTGCTATATGCGCTTACGAT
CGTACGAAGATTATAGAACATGCTGCTACTGCTCC
CTTATTGATAACTAGCTGATTATAGCTACGATG



Sequenced DNA is compared with DNA in a barcode database



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

The Start of Barcoding



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

How many species can you name?

How many Animals did you name?

How many mammals?

How many plants?

How many insects?

“Dog”

Canis lupus familiaris



“Cat”

Felis catus

“Shark”

Ginglymostoma cirratum



“Oak Tree”

Quercus alba



“Beetle”

Popillia japonica





Problem 1: No one know how many species there are.

How many species are there?

Vertebrates	Species
Mammals	5,490
Birds	9,998
Reptiles	9,084
Amphibians	6,433
Fishes	31,300
Total	62,305

Invertebrates	Species
Insects	1,000,000
Mollusks	85,00
Crustaceans	47,000
Corals	2,175
Arachnids	102,248
Total (+others)	1,305,250

Plants	Species
Angiosperms	281,821
Gymnosperms	1,021
Ferns and Allies	12,000
Mosses	16,236
Green and Red Algae	10,134
Total	321,212

- 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.



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

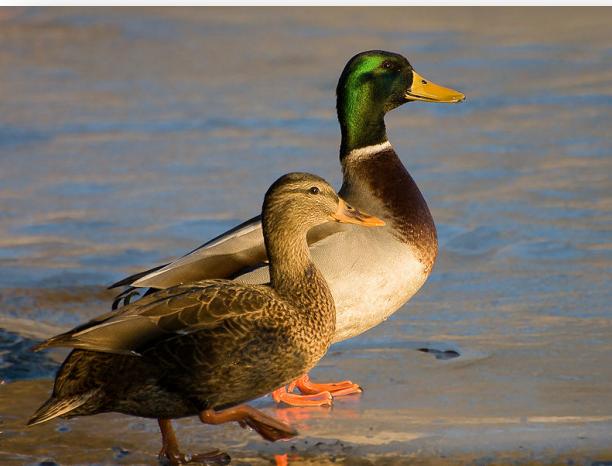
What is a species?



Canis lupus



Canis lupus (familiaris)



Anas platyrhynchos

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**



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

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 ca. 1 cm



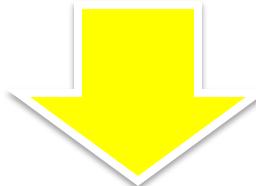
The body form ranges from hemispherical (e.g., *Cleidostethus*) to elongate oval (e.g., *Clypastraea*) to latridiid-like (e.g., *Foadia*). Corylophids are typically dull brown, but some species have contrasting yellowish-brown patches on the pronotum or elytra. The integument is often densely punctured 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



Sequence 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**

```
ATTTAAATTATGTGTCAGATATTAATACCCATCCCCTGGAAATCCTGGTCAAATACTCAATGCTGGACTCAAGATGTTCTT  
TGCATTTATTGCGATTCTTCCACGAATATCATAATTGAAT AGTTTCATTACTCCGAAAAAACCTATTTACGTGATTCAATTCAAAGAAA  
ATAAAAGATTTTCGAT TCCTATATAATTCTTATGTATTGAATGTGAATTGTATTAGTTTTTCTATAAGCAATCCTTTACGATCAA  
GGTCCTCTGGAGTCTTCTTGAGCGAACACATTCTATGGAAAATGGGCATTTTAGTAGTGTGTTGTAATTATTCAGAACCCAATG  
GTTCTCAAAGATCCTTCTGCATTATGTCGATATC AAGGAAAAGCAATTCTGGTCAAAGGGAACTCGTCTTGATGAGGAAATGGAGA  
TCTTACCTTGTCCATTGGCAATTATTTCAATTGGTCTCATCCGCATAGGATTATCAAACCAATTCAAATTATCCTTCTGTTTC  
TGGGTATTTCAAATGTACTAATAAAATTCCGTGGTAAGGAGTCAAATGTAGAAAATTCTTGTAAATAGATACTCTACTAAGAAATT  
TGATACCAGAGTTCAGTTATTGCTTATTG ATCATTGTCTAAAGCGAAATTGTACCGTATCCGGGCATCCTATTAGTAAGTCATATGGA  
CAAATTTCAGATTGGATATTATCGATTGGGGATATGTAGAA
```

**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?

Aedes adult



By Muhammad Mahdi Karim - Own work, GFDL 1.2, <https://commons.wikimedia.org/w/index.php?curid=11185617>

Anopheles adult



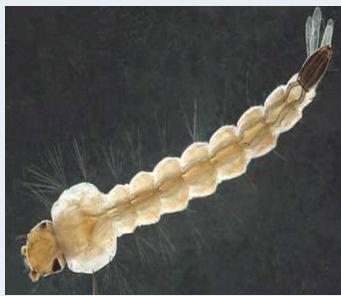
By Jim Gathany - (PHIL), ID #5814. <https://commons.wikimedia.org/w/index.php?curid=799284>

Culex adult



By Muhammad Mahdi Karim - Own work, GFDL 1.2, <https://commons.wikimedia.org/w/index.php?curid=7673048>

Aedes larva



Photograph by Michele M. Cutwa, University of Florida.

Anopheles larva

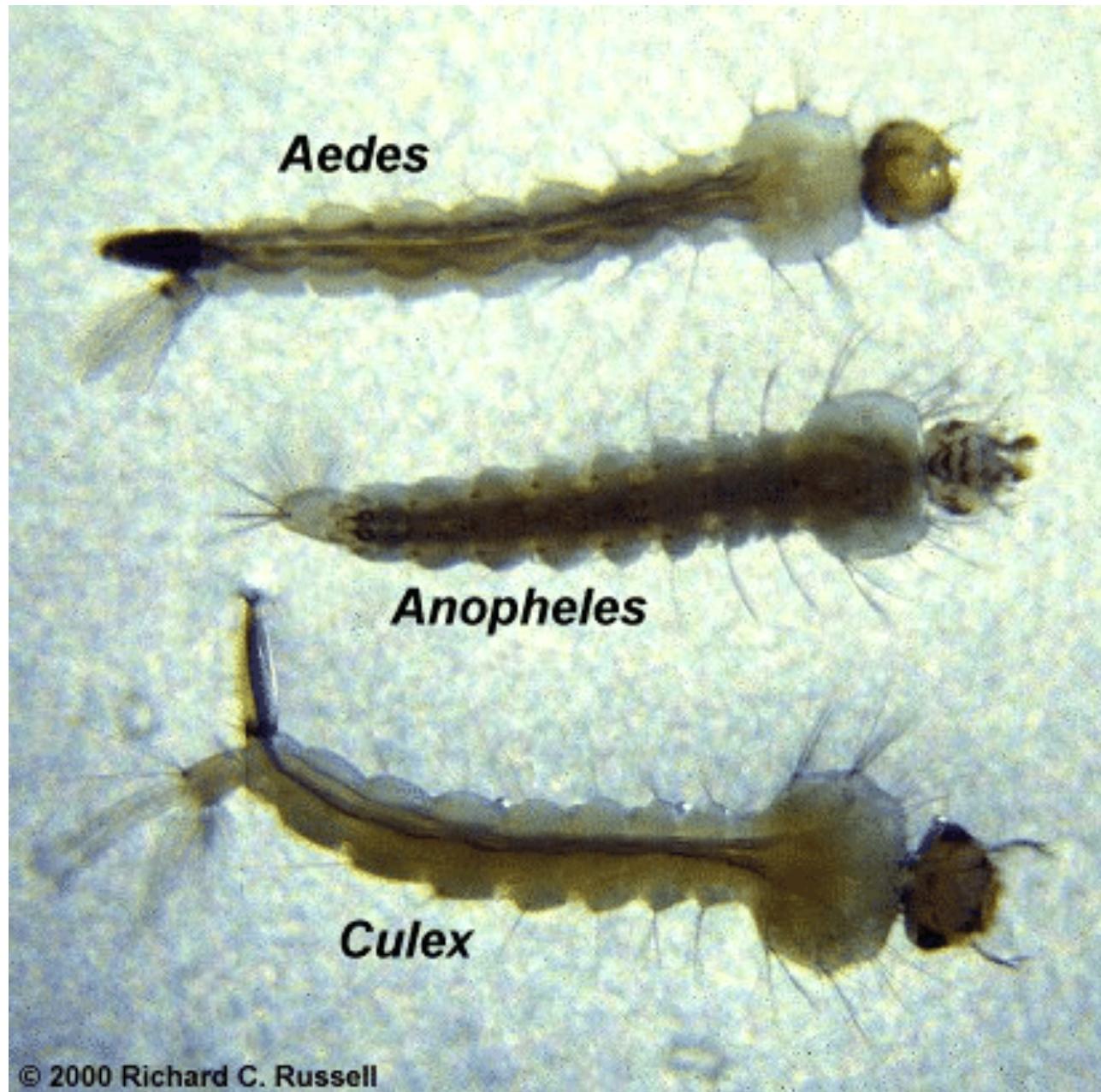


Culex larva



Photograph by Michelle Cutwa-Francis, University of Florida.





© 2000 Richard C. Russell



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

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



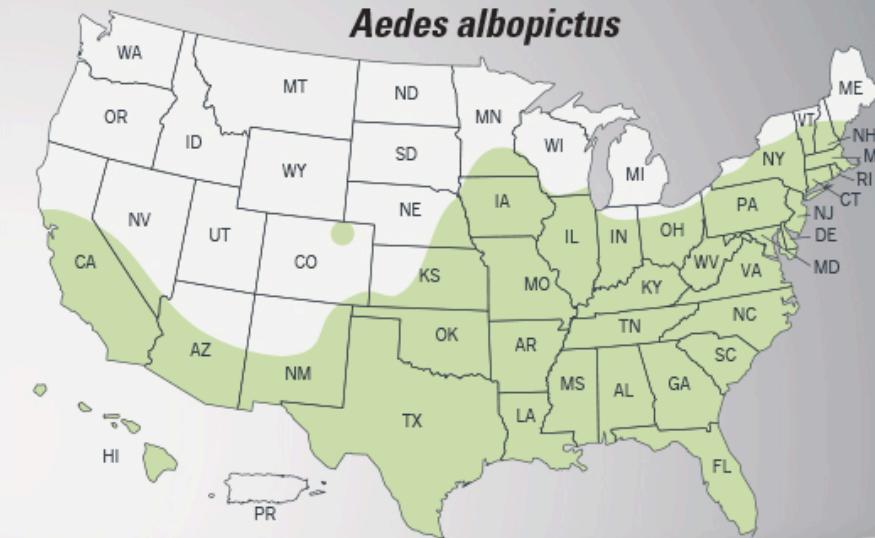
Why does this matter?

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

Aedes aegypti



Aedes albopictus



***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 meant to represent risk for spread of disease.

SOURCE: Zika: Vector Surveillance and Control. www.cdc.gov/zika/vector/index.html

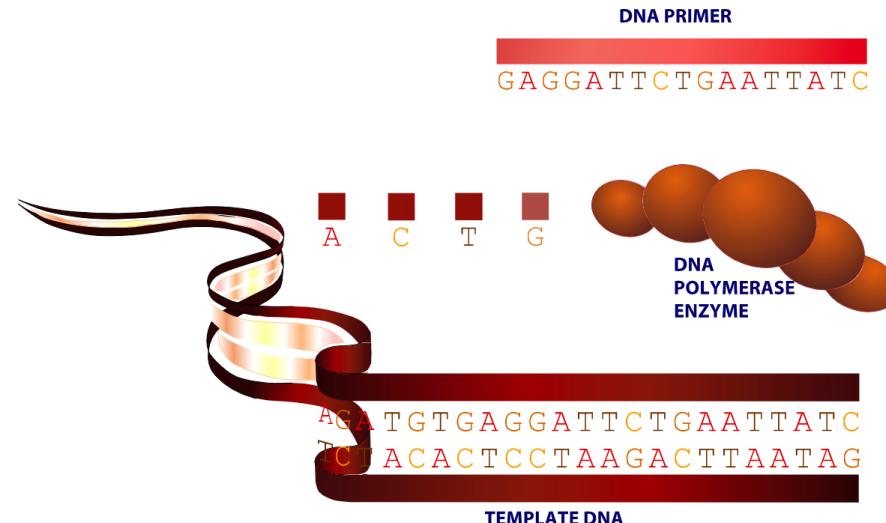


Lab: DNA Sequencing Background

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.



Jump to: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

<https://dnalc.cshl.edu/resources/animations/cycseq.html>



Cold Spring Harbor Laboratory
DNA LEARNING CENTER

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

[GenBank](#) [Graphics](#)

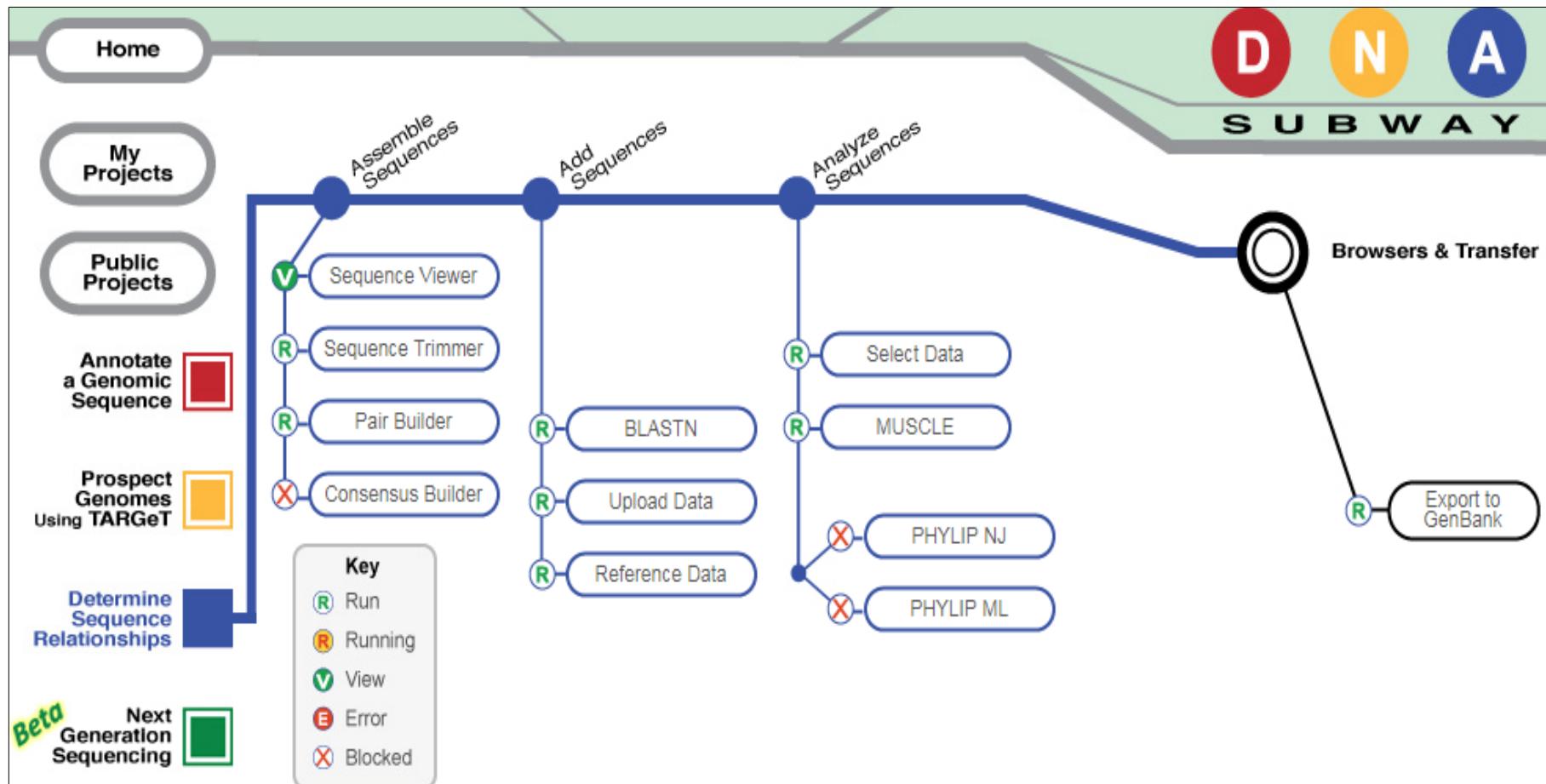
```
>MK592083.1 Anopheles gambiae isolate 10016 5.8S ribosomal RNA gene, partial
sequence; internal transcribed spacer 2, complete sequence; and large subunit
ribosomal RNA gene, partial sequence
ACACATGAACATTGATAAGTTAACGCATATGGCGCATCGGACGTTAACCGATGCACACATT
TTGAGTGCCTACTAATTACCAAAGTCTCATTAGTTAACTACAGTGGCCGTCCGCGAAGGTGCCCGGGTC
ATCCGACGCAGTGGCGGTGCTGTGCATAATGACGTGCTGGTCCCCGTCTGGGTCTCGGGCGTTG
AAAGTGGACACTCTCGAGCGTATGGATGCGTTCTGTGTTGGTGTGTTGATGCGTAGGGCTTGTGG
TGTGTGTCAGCCGCATGGTTCGAACTAATGCTACGTCGTTCCGATGCCACCGGCAGTCTACTCTCCA
GGCTAAAGTCGGCTCGTCTAGGGATTGGAAAGCTAACGTGCTGTAACTCATGTGGGCCATACACGGCG
TTGCGCTACCACGCTAACGTTAGCCCTACATACACAAGCATCAACCCACGGCACGGCGTAGCTGTAATAC
TTACGTCTCGGTTATACCACGTAGGCCTAACGTGATGTGACTACCCCC
```



Lab: Creating a DNA Subway Project

(follow along in the packet)

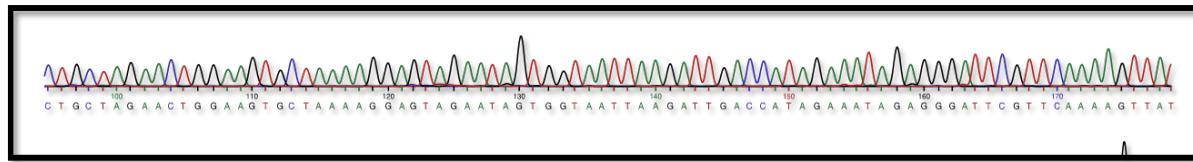
Working on DNA Subway Blue Line



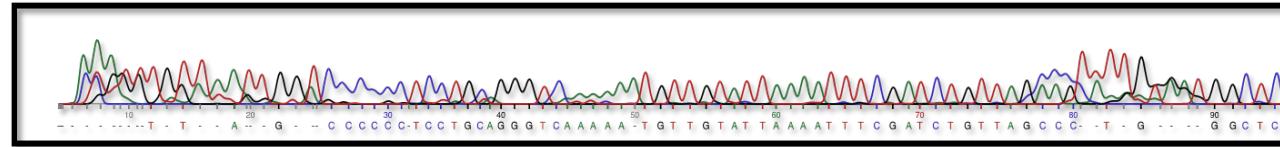
Key Concept: Data Quality

Some sequence examples...

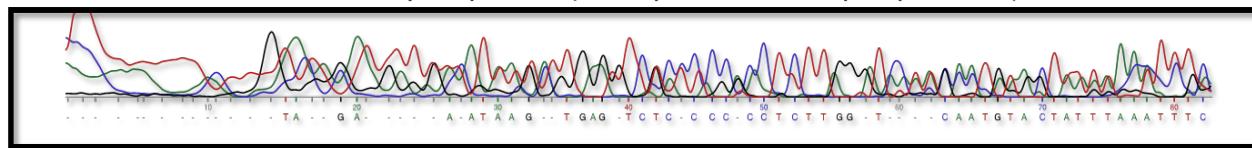
High Quality Sequence



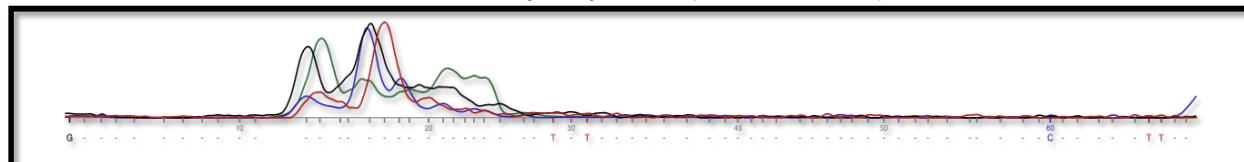
Acceptable Quality Sequence



Low Quality Sequence (multiple base calls per position)



Low Quality Sequence (no base calls)



Phred scores...

Phred Score	Error (bases miscalled)	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Next time:
Comparing sequences with BLAST