Cold Spring Harbor Laboratory
DNA Learning Center

DNALC Live

Barcoding Bioinformatics
Part II

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

Cold Spring Harbor Laboratory

DNA LEARNING CENTER

Summer & Public Programs ˅ School Programs ˅ Student Research Educator Training Resources ˅ DNALC & Partners ˅

DNA LEARNING CENTER

Preparing students and families to thrive in the gene age

LEARN MORE

dnalc.cshl.edu/dnalc-live

DNALC Website and Social Media

youtube.com/DNALearningCenter

facebook.com/cshldnalc

@dnalc

@dna_learning_center

Barcoding Bioinformatics
Part II

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 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)

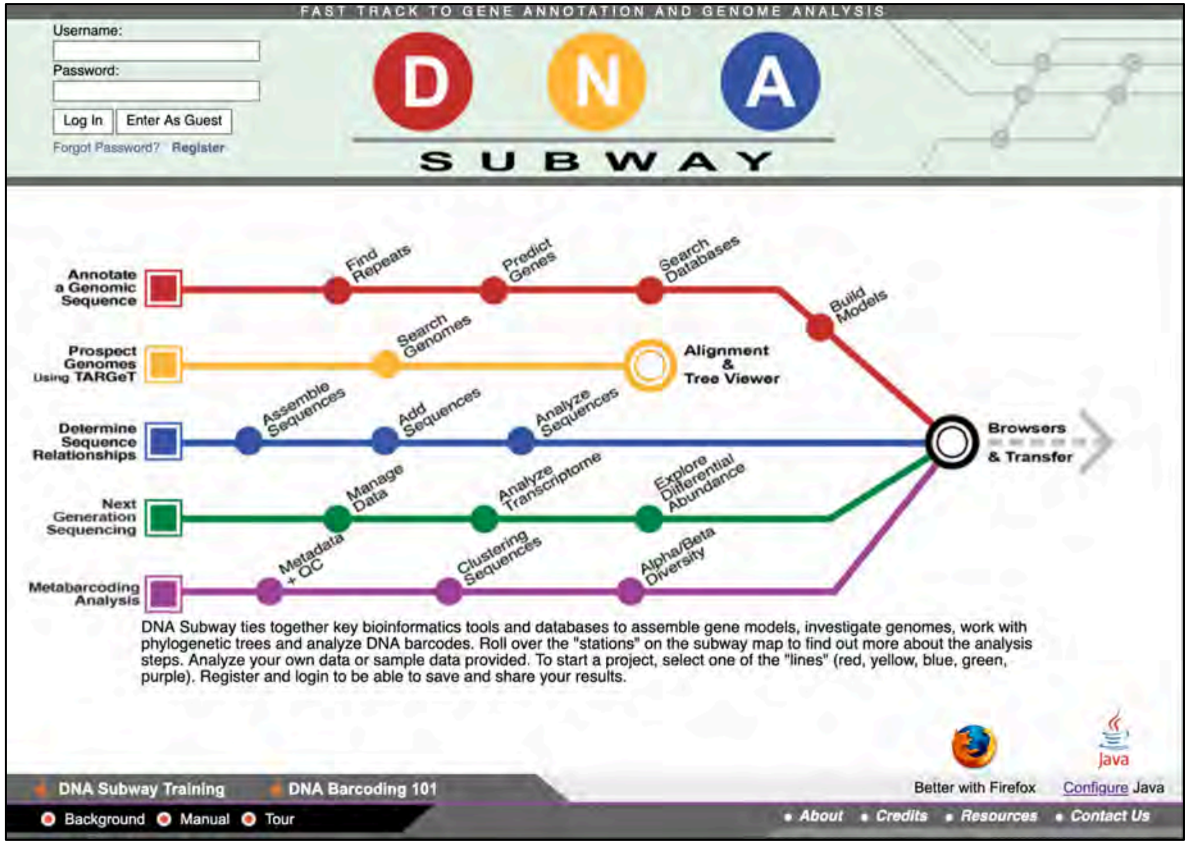
Course Learning Goals

- 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)



DNA Subway ties together key bioinformatics tools and databases to assemble gene models, investigate genomes, work with phylogenetic trees and analyze DNA barcodes. Roll over the "stations" on the subway map to find out more about the analysis steps. Analyze your own data or sample data provided. To start a project, select one of the "lines" (red, yellow, blue, green, purple). Register and login to be able to save and share your results.

Barcoding Bioinformatics Part II

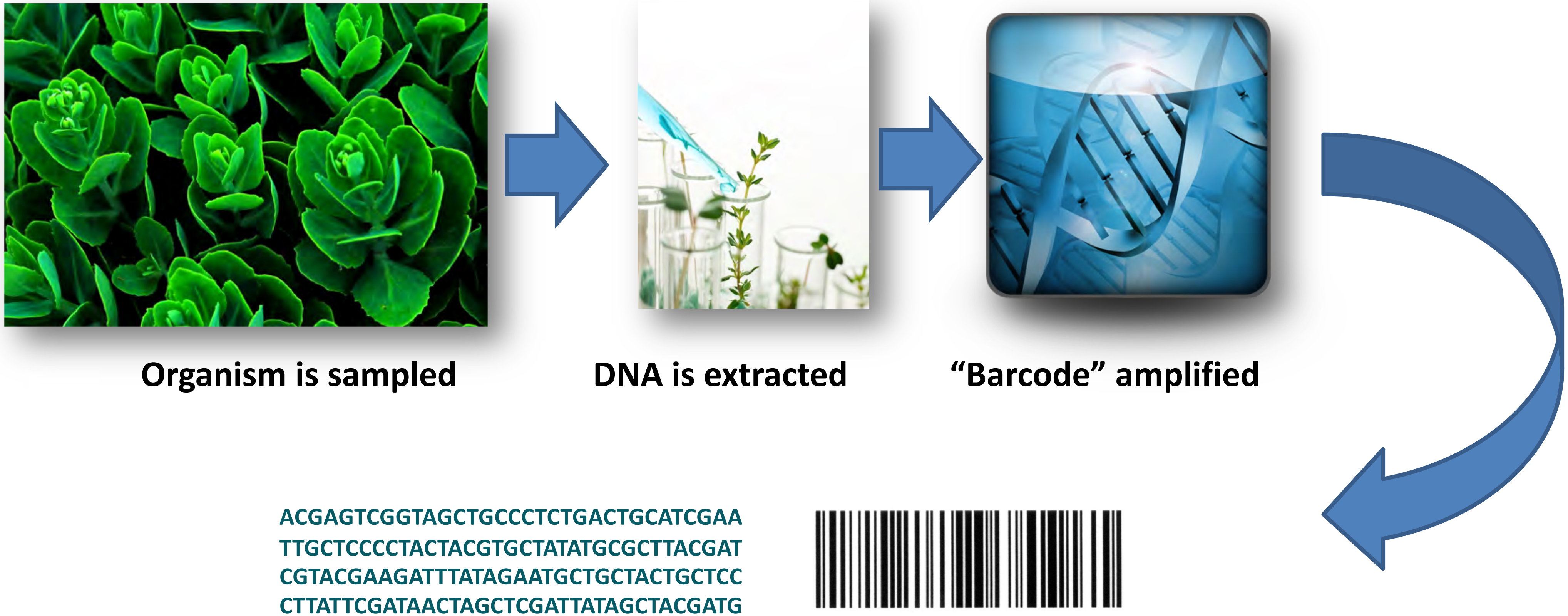
(Sequence cleaning and BLAST)

Steps for today’s session

- Recap on our experimental dataset
- Review of sequence quality
- Sequence cleaning and pairing
- Introduction to BLAST

Recap of the dataset

Steps to DNA Barcoding



Organism is sampled

DNA is extracted

“Barcode” amplified

ACGAGTCGGTAGCTGCCCTCTGACTGCATCGAA
TTGCTCCCCTACTACGTGCTATATGCGCTTACGAT
CGTACGAAGATTTATAGAATGCTGCTACTGCTCC
CTTATTCGATAACTAGCTCGATTATAGCTACGATG

Sequenced DNA is compared with DNA in a barcode database

Example barcoding experiment

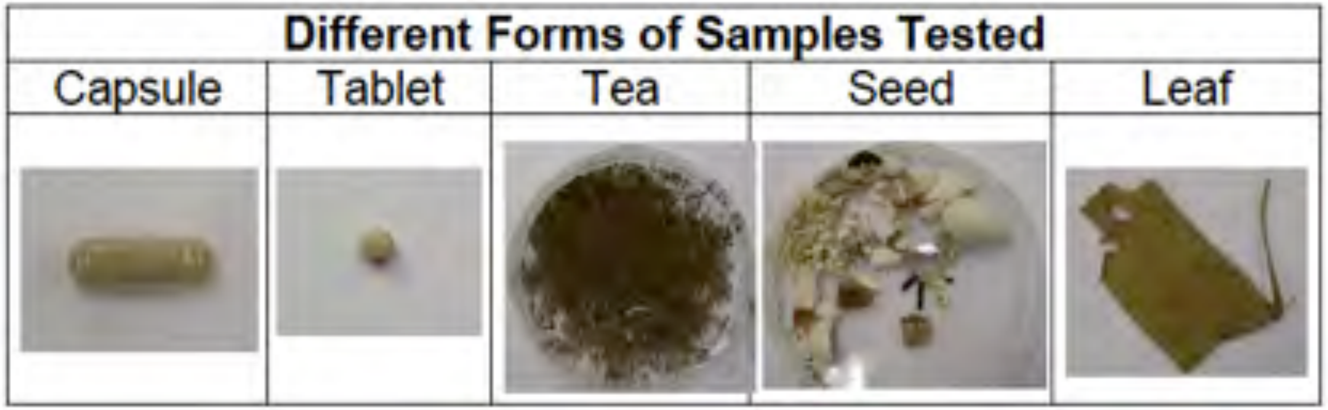
Traditional Chinese Medicine
Using Barcode to Identify the Content of Ginkgo Products

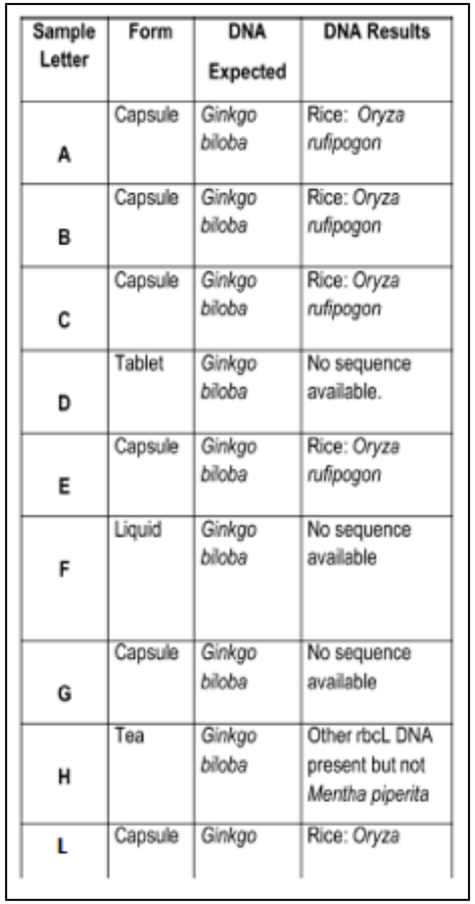
Mary Acheampong,
Bobby Glover, and Marisa VanBrakle

Mentor: Allison Granberry
Hostos-Lincoln Academy of Science,
The Bronx

2012 UBP Grand Prize Winners

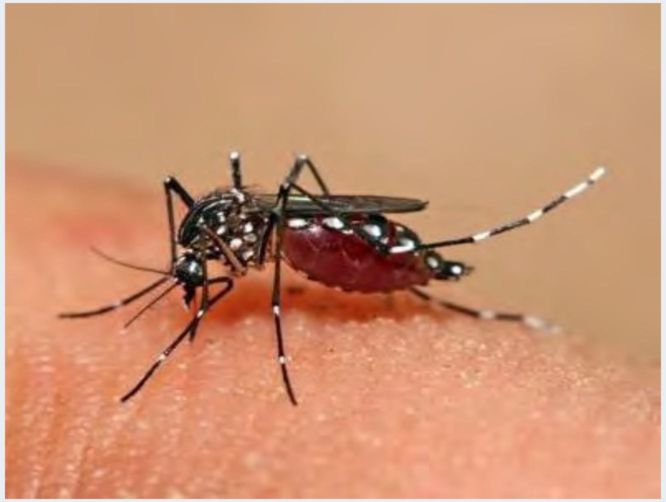
Example barcoding experiment





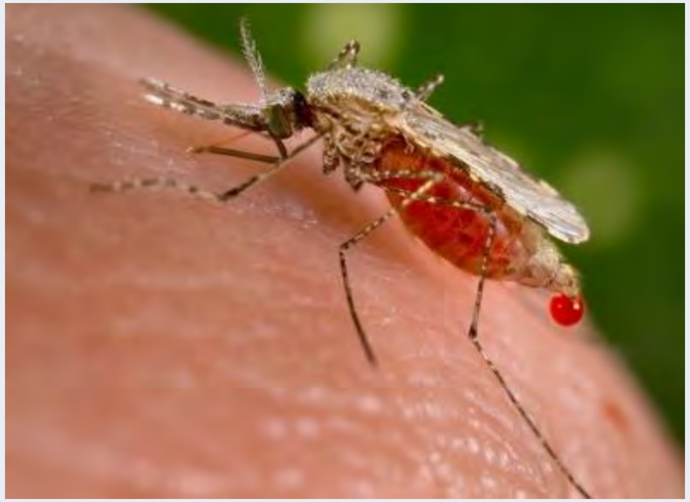
Sample Letter	Form	DNA Expected	DNA Results
A	Capsule	*Ginkgo biloba*	Rice: *Oryza rufipogon*
B	Capsule	*Ginkgo biloba*	Rice: *Oryza rufipogon*
C	Capsule	*Ginkgo biloba*	Rice: *Oryza rufipogon*
D	Tablet	*Ginkgo biloba*	No sequence available.
E	Capsule	*Ginkgo biloba*	Rice: *Oryza rufipogon*
F	Liquid	*Ginkgo biloba*	No sequence available
G	Capsule	*Ginkgo biloba*	No sequence available
H	Tea	*Ginkgo biloba*	Other rbcL DNA present but not *Mentha piperita*
L	Capsule	*Ginkgo*	Rice: *Oryza*

Aedes adult



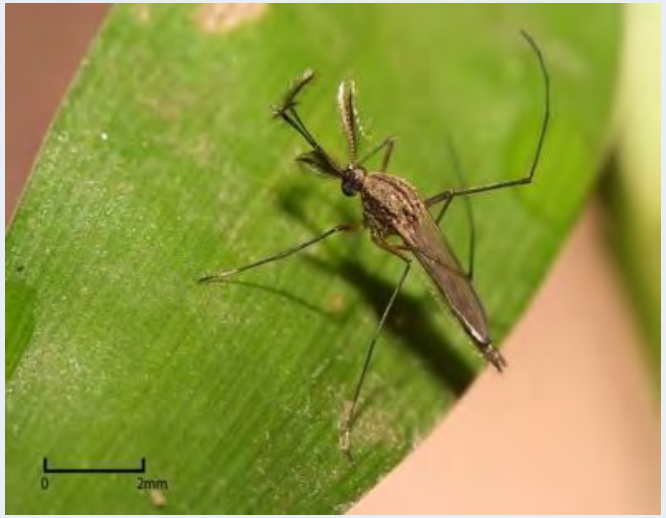
By Muhammad Mahdi Karim - Own work, GFDL 1.2, https://commons.wikimedia.org/w/index.php?curid=11185617

Anopheles adult



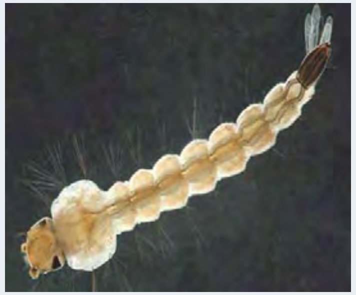
By Jim Gathany - (PHIL), ID #5814. https://commons.wikimedia.org/w/index.php?curid=799284

Culex adult



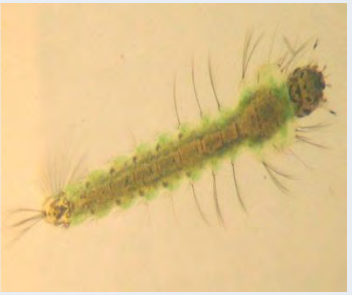
By Muhammad Mahdi Karim - Own work, GFDL 1.2, https://commons.wikimedia.org/w/index.php?curid=7673048

Aedes larva

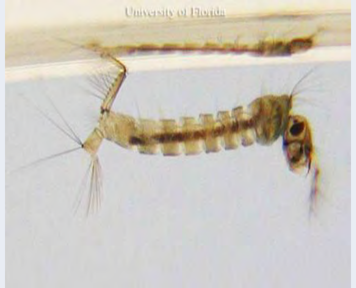


Photograph by Michele M. Cutwa, University of Florida.

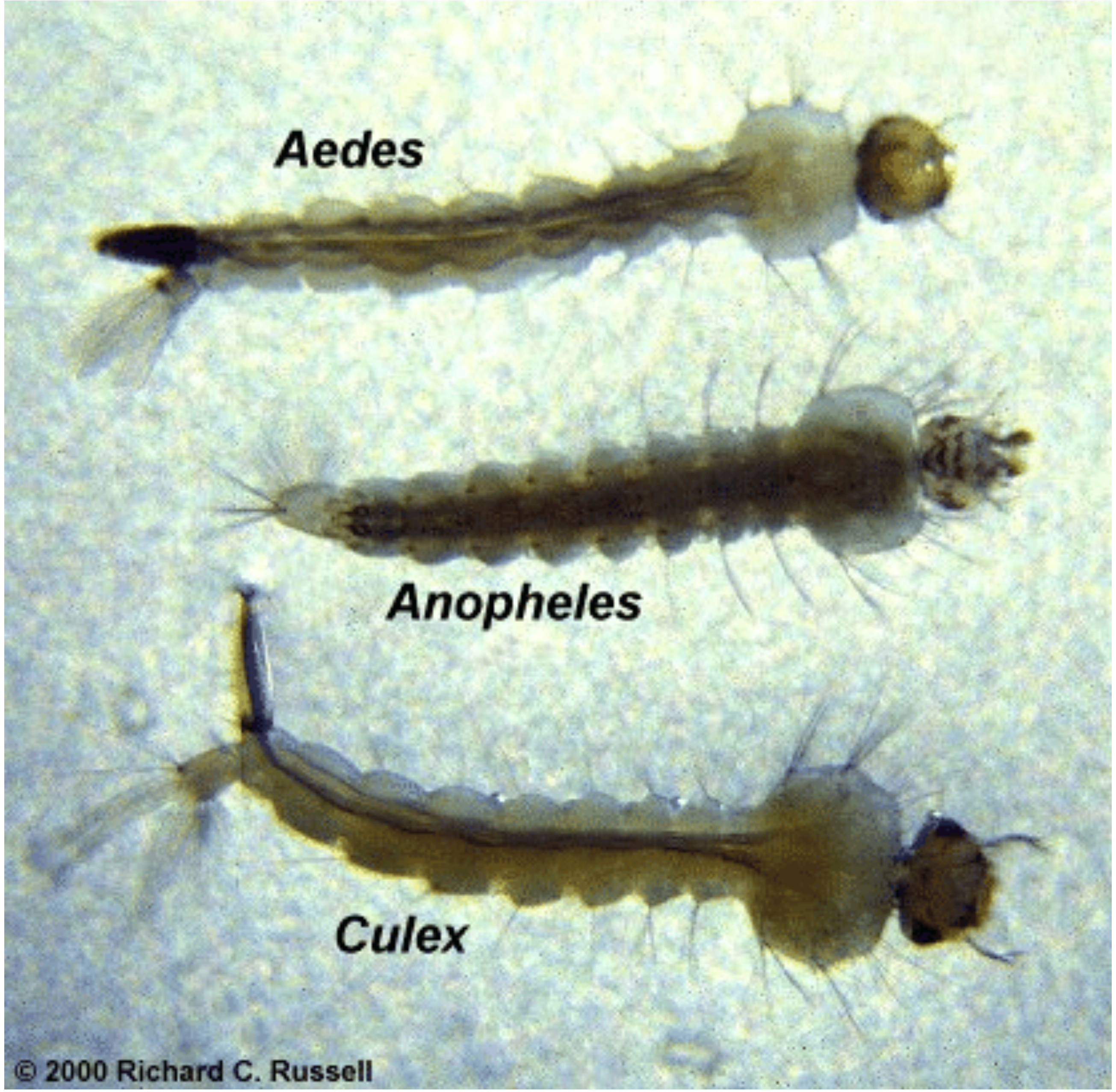
Anopheles larva



Culex larva



Photograph by Michelle Cutwa-Francis, University of Florida.



© 2000 Richard C. Russell

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

Experimental components/design

Materials

- We have DNA from unknown mosquito samples
- We can obtain DNA from known samples

Experimental components/design

Materials

- We have DNA from unknown mosquito samples
- We can obtain DNA from known samples

Hypothesis

- We can use computational methods (BLAST/phylogenetic analysis) to infer the species

Experimental components/design

Materials

- We have DNA from unknown mosquito samples
- We can obtain DNA from known samples

Hypothesis

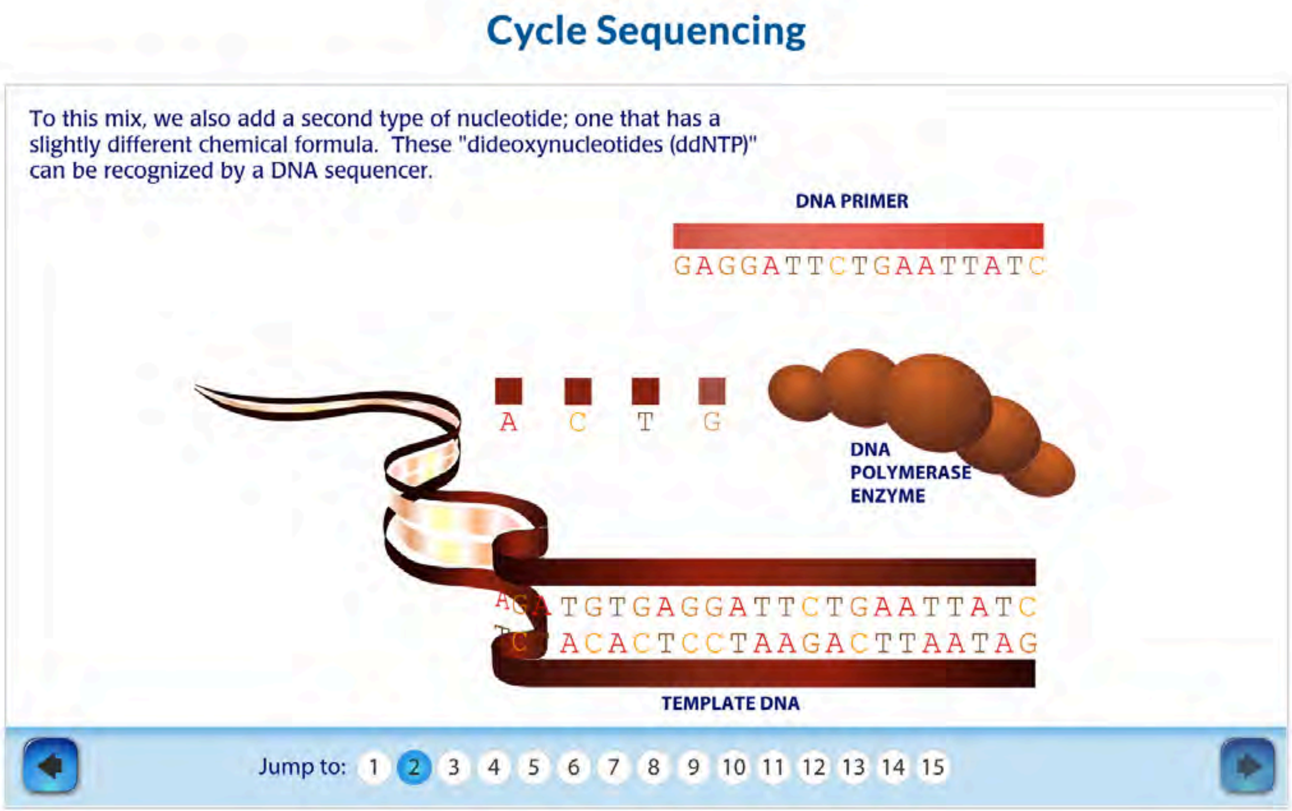
- We can use computational methods (BLAST/phylogenetic analysis) to infer the species

Controls

- We have sensitivity controls (sequence quality, BLAST parameters)
- We have outgroup sequences (non-mosquito, negative controls) and known samples (positive controls)

Review of sequencing and quality

DNA Sequencing

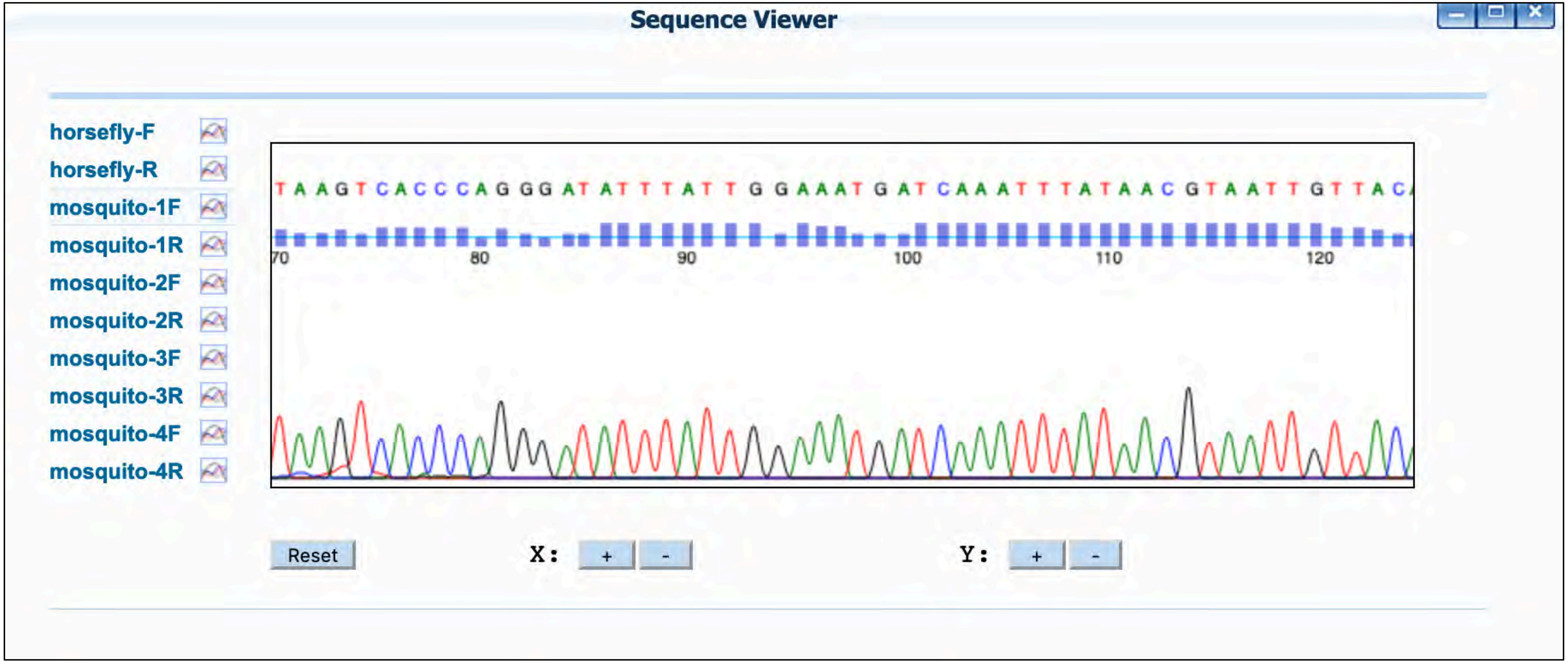


Cycle Sequencing

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.

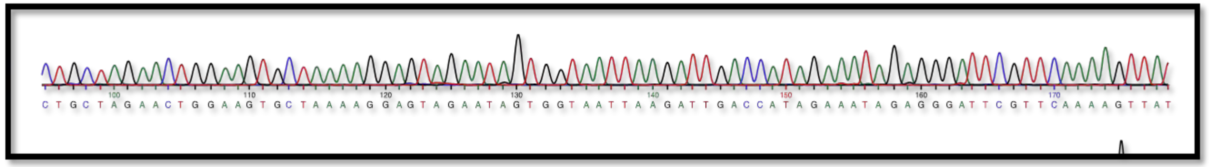
https://dnalc.cshl.edu/resources/animations/cycseq.html

Chromatogram/Electropherogram

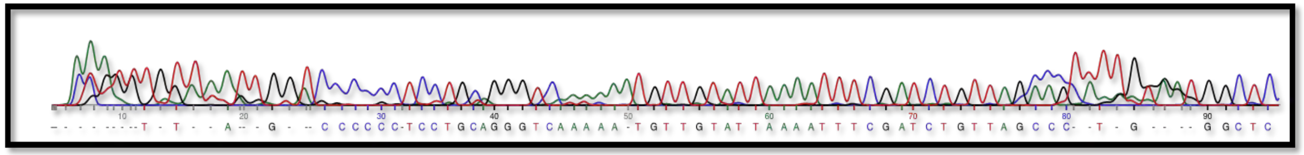


Some sequence examples…

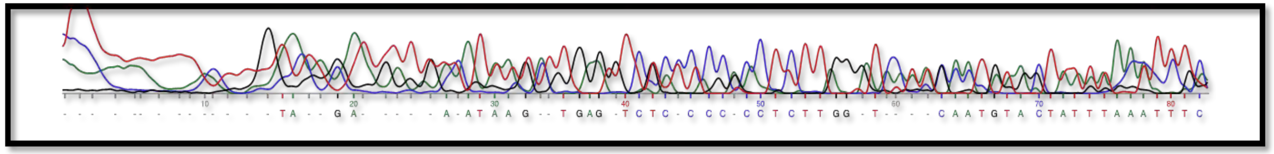
High Quality Sequence



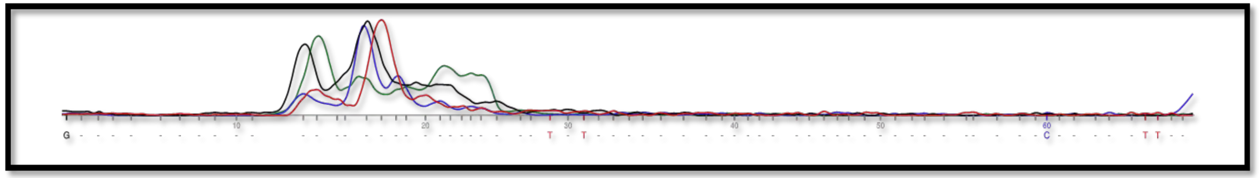
Acceptable Quality Sequence



Low Quality Sequence (multiple base calls per position)



Low Quality Sequence (no base calls)

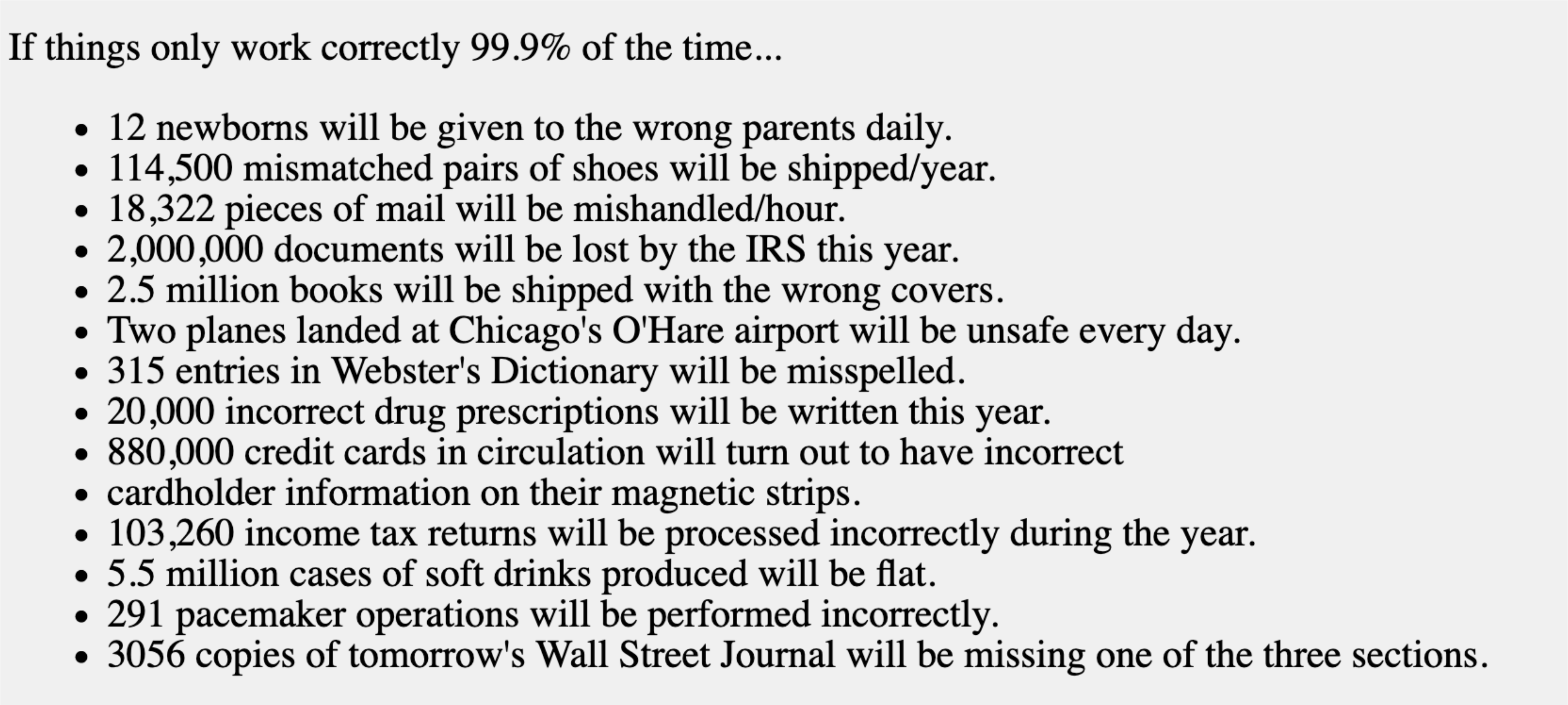


Cold Spring Harbor Laboratory
DNA LEARNING CENTER

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%

If 99% was good enough



If things only work correctly 99.9% of the time...

- 12 newborns will be given to the wrong parents daily.
- 114,500 mismatched pairs of shoes will be shipped/year.
- 18,322 pieces of mail will be mishandled/hour.
- 2,000,000 documents will be lost by the IRS this year.
- 2.5 million books will be shipped with the wrong covers.
- Two planes landed at Chicago's O'Hare airport will be unsafe every day.
- 315 entries in Webster's Dictionary will be misspelled.
- 20,000 incorrect drug prescriptions will be written this year.
- 880,000 credit cards in circulation will turn out to have incorrect
- cardholder information on their magnetic strips.
- 103,260 income tax returns will be processed incorrectly during the year.
- 5.5 million cases of soft drinks produced will be flat.
- 291 pacemaker operations will be performed incorrectly.
- 3056 copies of tomorrow's Wall Street Journal will be missing one of the three sections.

Photo credit
http://www.personal.psu.edu/sxt104/class/99percent.html

A note on controls

At what temperature does ice (H_2O) + Chemical “X” melt?

A note on controls

Positive control: What does the effect look like if present?

A note on controls

Positive control: What does the effect look like if present?

Negative control: What does the effect look like if absent?

A note on controls

Positive control: What does the effect look like if present?

Negative control: What does the effect look like if absent?

Sensitivity control: Across what range of values can I measure the effect?

A note on controls

Positive control

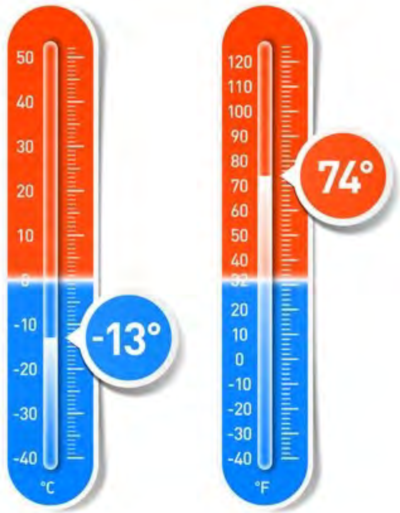
Negative control

Sensitivity control

Photo credits

https://commons.wikimedia.org/wiki/File:Water_in_a_beaker.JPG
http://www.chem.uiuc.edu/webfunchem/temperature/Temp10.htm
https://www.dreamstime.com/photos-images/alcohol-thermometer.html

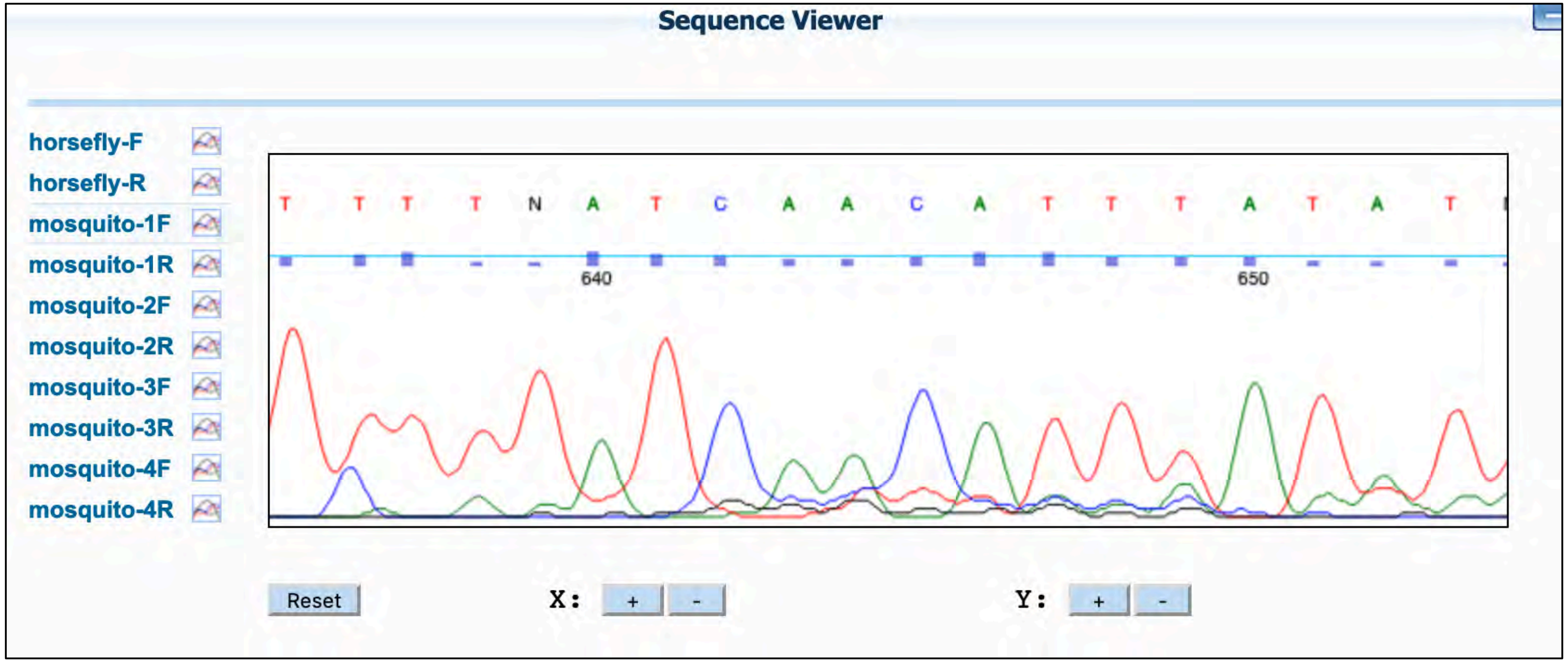
A note on controls



Positive control **Negative control** **Sensitivity control**

Photo credits
https://commons.wikimedia.org/wiki/File:Water_in_a_beaker.JPG
http://www.chem.uiuc.edu/webfunchem/temperature/Temp10.htm
https://en.clipdealer.com/vector/media/A:17494508?

Phred are our measure of quality (signal/noise)



Lower score = more noise than signal

Bi-directional sequencing

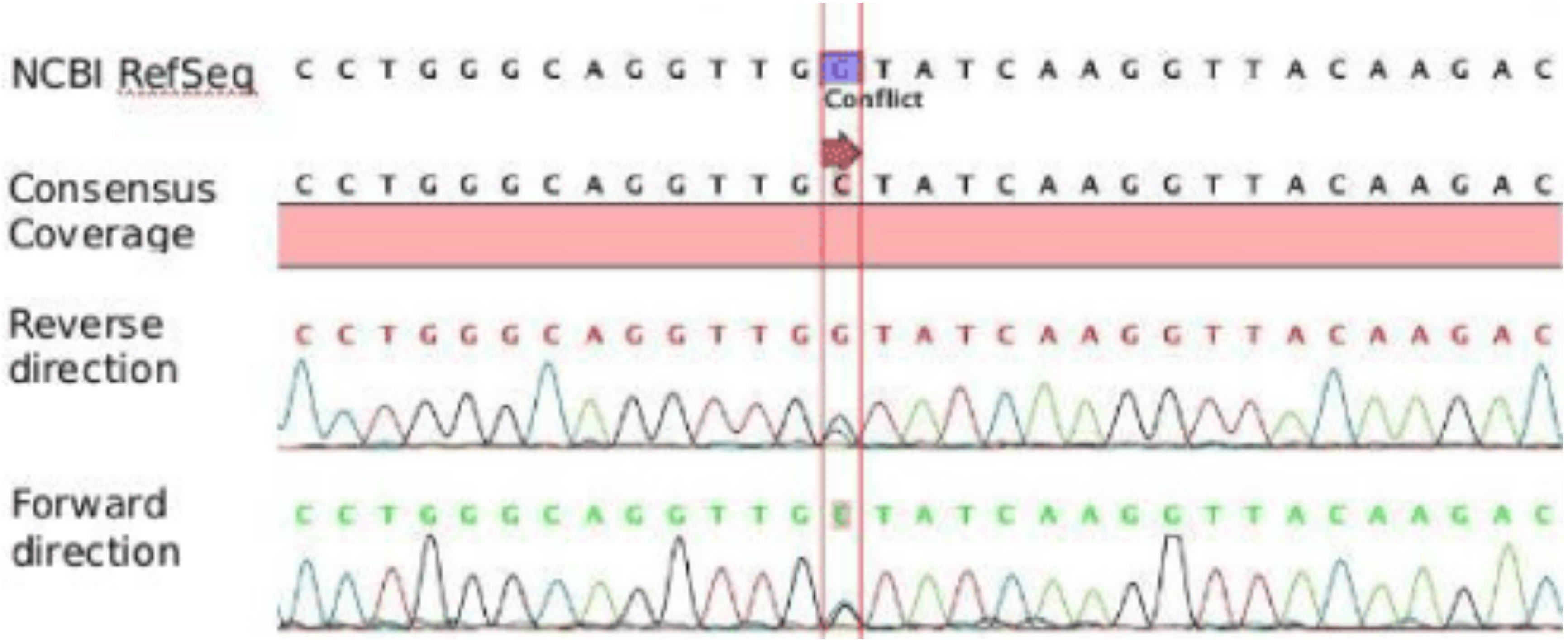


Photo credit
https://www.omicsonline.org/articles-images/CMBO-2-108-g003.html

Reverse complementation

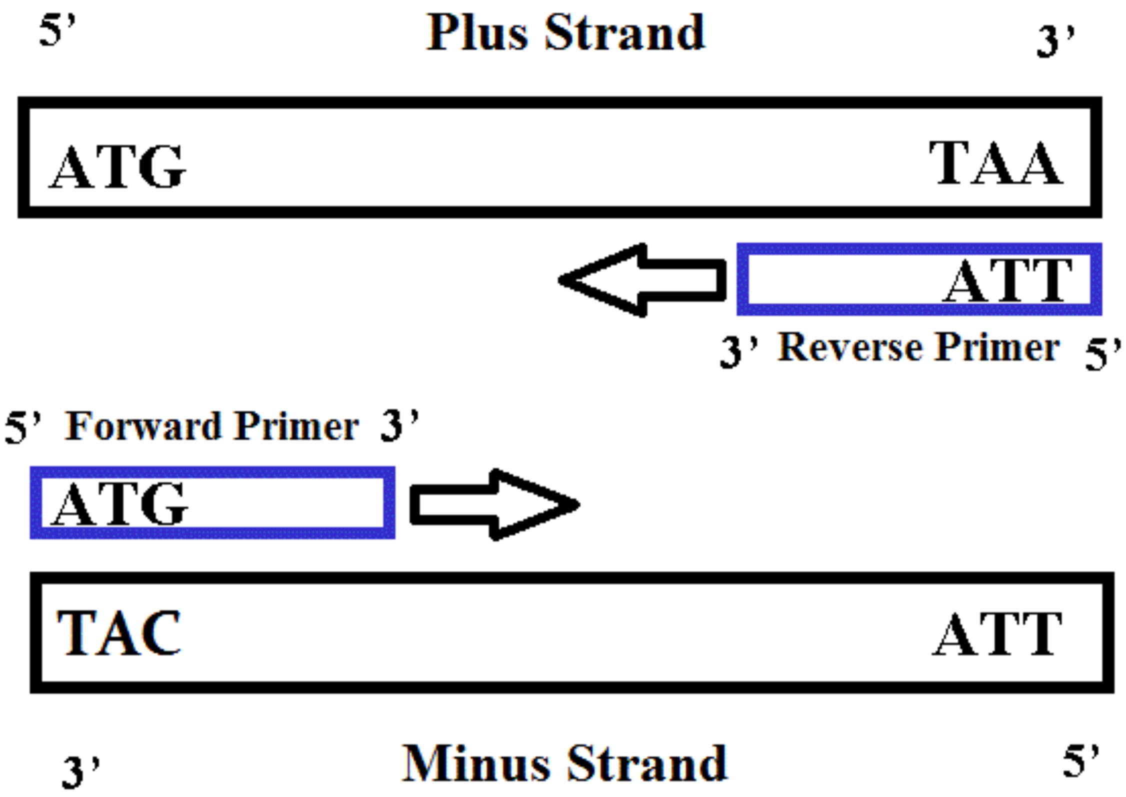
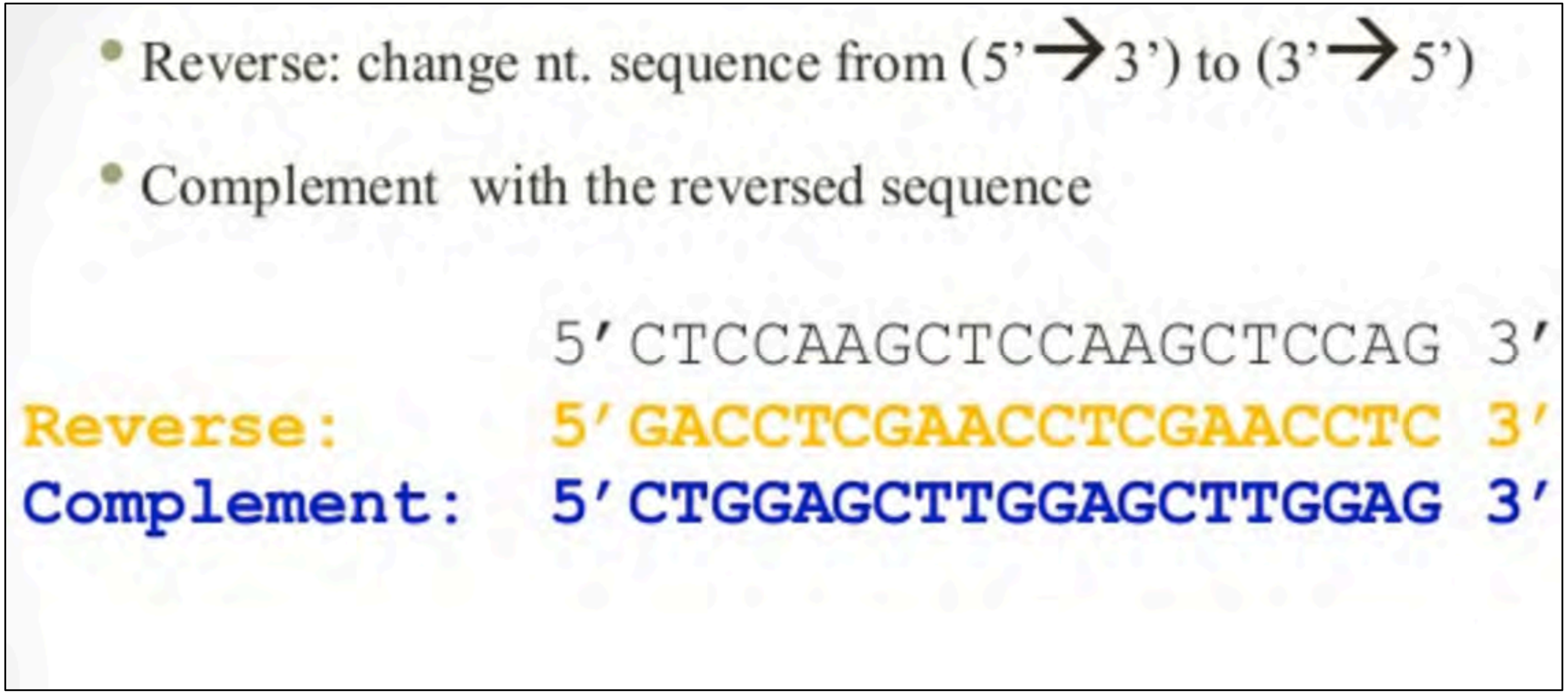


Photo credit
https://biology.stackexchange.com/questions/56304/manual-primer-design-for-a-gene-on-the-reverse-strand

Reverse complementation

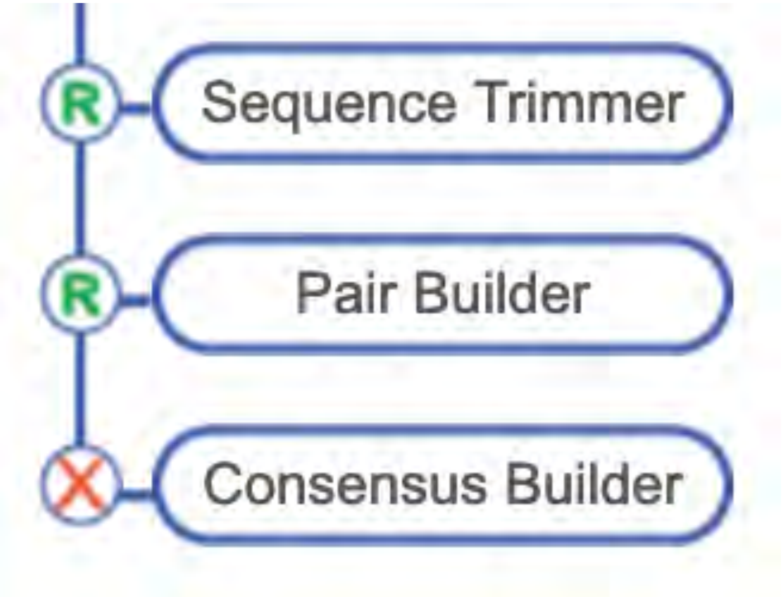


- Reverse: change nt. sequence from (5’→3’) to (3’→5’)
- Complement with the reversed sequence

```
            5′CTCCAAGCTCCAAGCTCCAG 3′
Reverse:    5′GACCTCGAACCTCGAACCTC 3′
Complement: 5′CTGGAGCTTGGAGCTTGGAG 3′
```

Photo credit
https://image.slidesharecdn.com/pcrprimerdesignenglishversion-160317161103/95/pcr-primer-design-english-version-10-638.jpg?cb=1458231192

Clean up and consensus



Introduction to BLAST

Basic Local Alignment Search Tool

- An algorithm for searching a database of sequences

Basic Local Alignment Search Tool

- An algorithm for searching a database of sequences

- “Google for DNA” (although works with any biological sequence, and started before Google ~1985)

Basic Local Alignment Search Tool

- An algorithm for searching a database of sequences
- “Google for DNA” (although works with any biological sequence, and started before Google ~1990 vs 1998)
- NCBI is the most popular interface, but this is software that can be run anywhere (including Subway)

Warning: Analogy

(useful for discussion but not the whole picture)

BLAST algorithm analogy

Query sequence

ACTGACATCGGGGTGCTACG

Database

BLAST algorithm analogy

Query sequence

ACTGACATCGGGGTGCTACG

Database

BLAST algorithm analogy

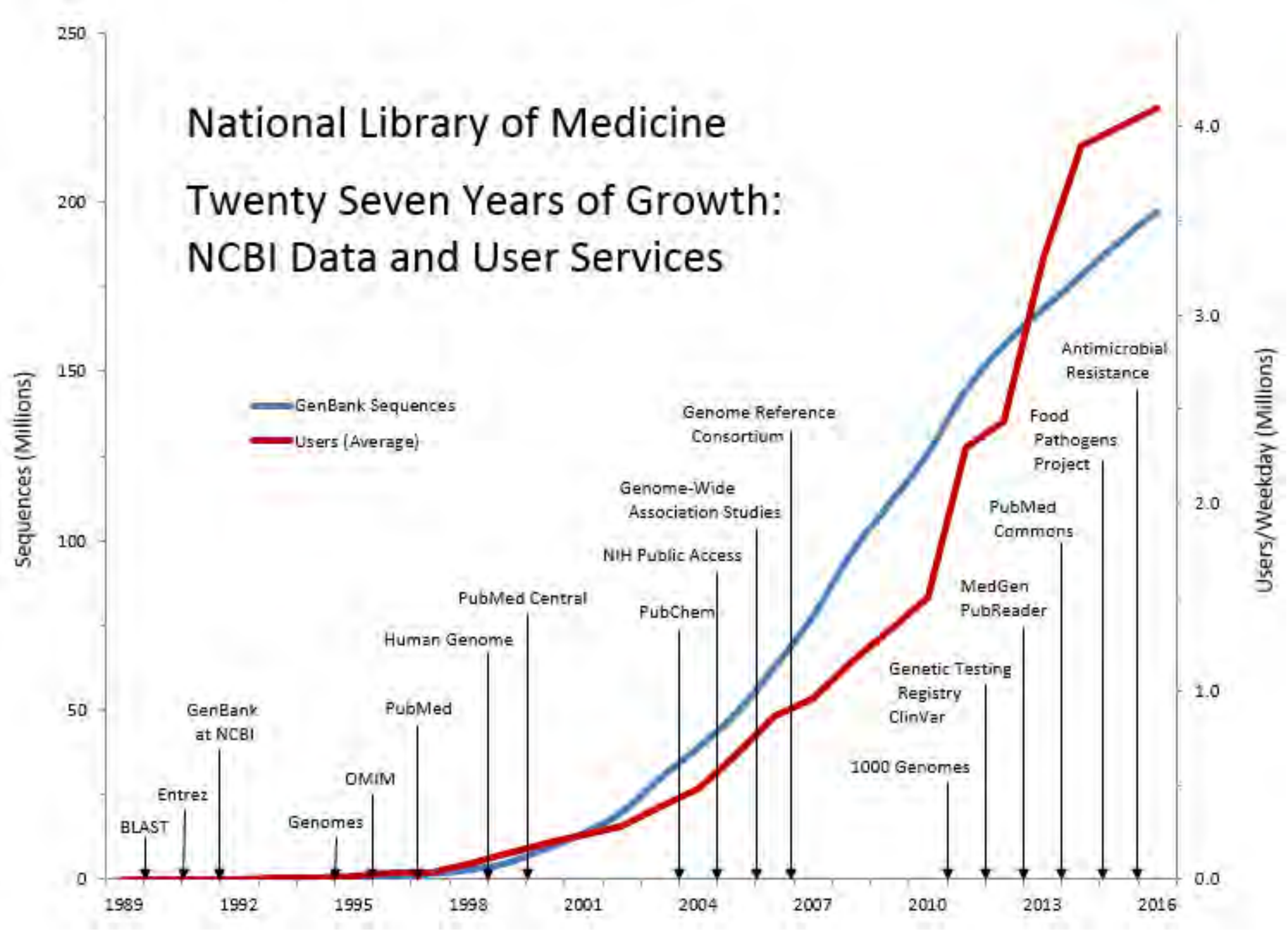


Photo credit
https://www.nlm.nih.gov/about/2018CJ.html

BLAST algorithm analogy – searching by "word"

Break the *Query sequence*
Into "words" (k-mers)

ACT GAC ATC GGG GTG CTA CG

Database

BLAST algorithm analogy – searching by “word”

Break the *Query sequence*
Into “words” (k-mers)

ACT GAC ATC GGG GTG CTA CG

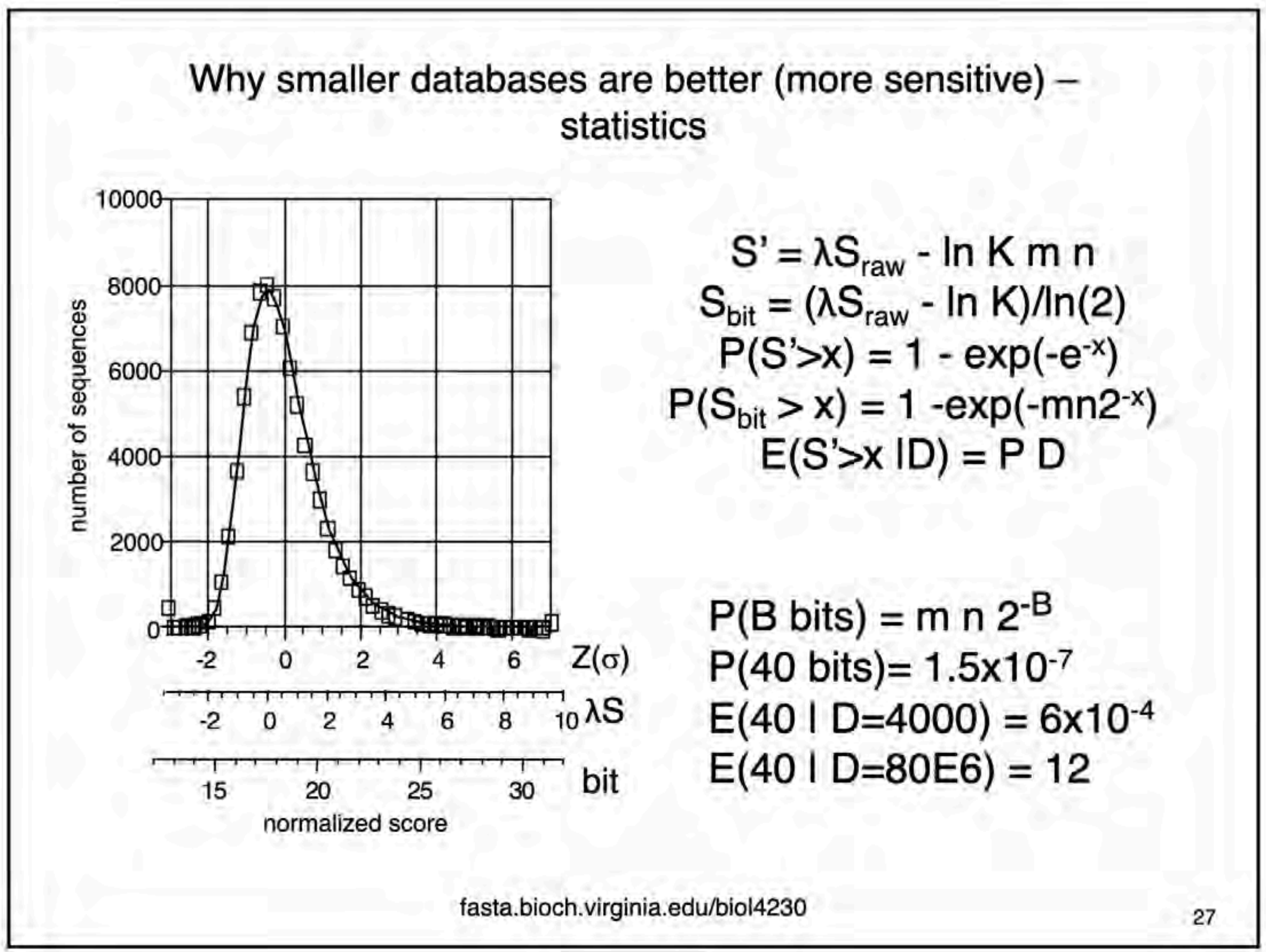
Database

Let’s BLAST a sequence

```
>mosquito-1F
CTTTAAGTATATTAATTCGTGCTGAATTAAGTCACCCAGGGATATTTAT
TGGAAATGATCAAATTTATAACGTAATTGTTACAGCTCATGCATTTATT
ATAATTTTTTTTATAGTAATACCAATTATAATTGGAGGATTTGGAAATT
GATTAGTTCCTTTAATATTAGGAGCTCCTGATATAGCATTTCCTCGAAT
AAATAATATAAGTTTTTGAATATTACCTCCTTCTTTAACTCTACTACTTT
CTAGTTCAATAGTAGAAAATGGAGCAGGGACAGGATGAACAGTTTA
TCCTCCTCTTTCATCAGGAACAGCACATGCTGGAGCTTCTGTTGATTT
AGCAATTTTCTCTCTTCATTTAGCAGGGATTTCATCTATTTTAGGAGC
AGTAAATTTTATTACTACTGTTATTAATATACGATCATCTGGAATTACTT
TAGATCGATTACCTTTATTTGTTTGATCTGTAGTAATTACTGCTATTTTA
TTACTTTTATCTCTTCCTGTATTAGCTGGAGCTATTACTATATTATTAACT
GATCGAAATTTAAATACTTCCTTCTTTGACCCAATTGGAGGAGGAGA
```

https://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST and controls



Why smaller databases are better (more sensitive) – statistics

$$S' = \lambda S_{raw} - \ln K\, m\, n$$
$$S_{bit} = (\lambda S_{raw} - \ln K)/\ln(2)$$
$$P(S'>x) = 1 - \exp(-e^{-x})$$
$$P(S_{bit} > x) = 1 - \exp(-mn2^{-x})$$
$$E(S'>x \mid D) = P\, D$$

$$P(B \text{ bits}) = m\, n\, 2^{-B}$$
$$P(40 \text{ bits}) = 1.5\text{x}10^{-7}$$
$$E(40 \mid D=4000) = 6\text{x}10^{-4}$$
$$E(40 \mid D=80E6) = 12$$

fasta.bioch.virginia.edu/biol4230

27

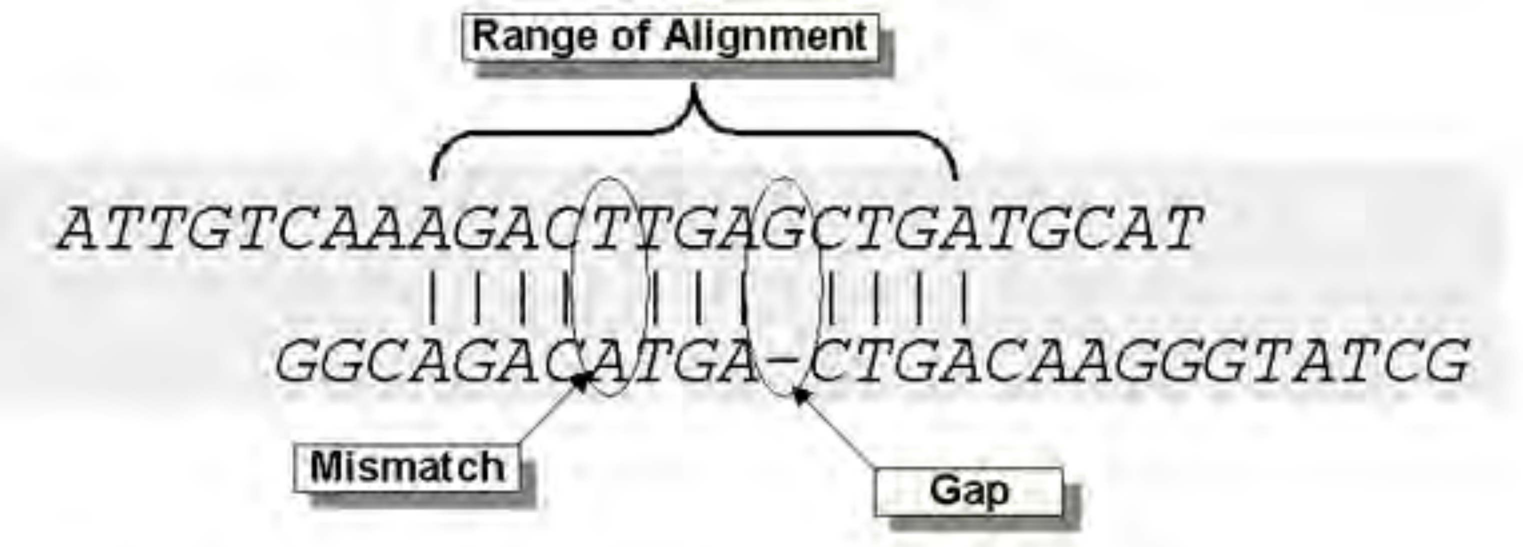
Photo credit
https://fasta.bioch.virginia.edu/biol4230/lects/biol4230_4_blast2.pdf

BLAST algorithm analogy – searching by "word"

The *Query sequence*
Is aligned to a *Subject* (a sequence in the database)

```
Q: ACTGAC–ATCGGGGTGCTACG
    ||| |||  ||||| |   || ||||| ||
S:  ACTGACCATCGGAGTGCTACG
```

BLAST algorithm analogy – alignment

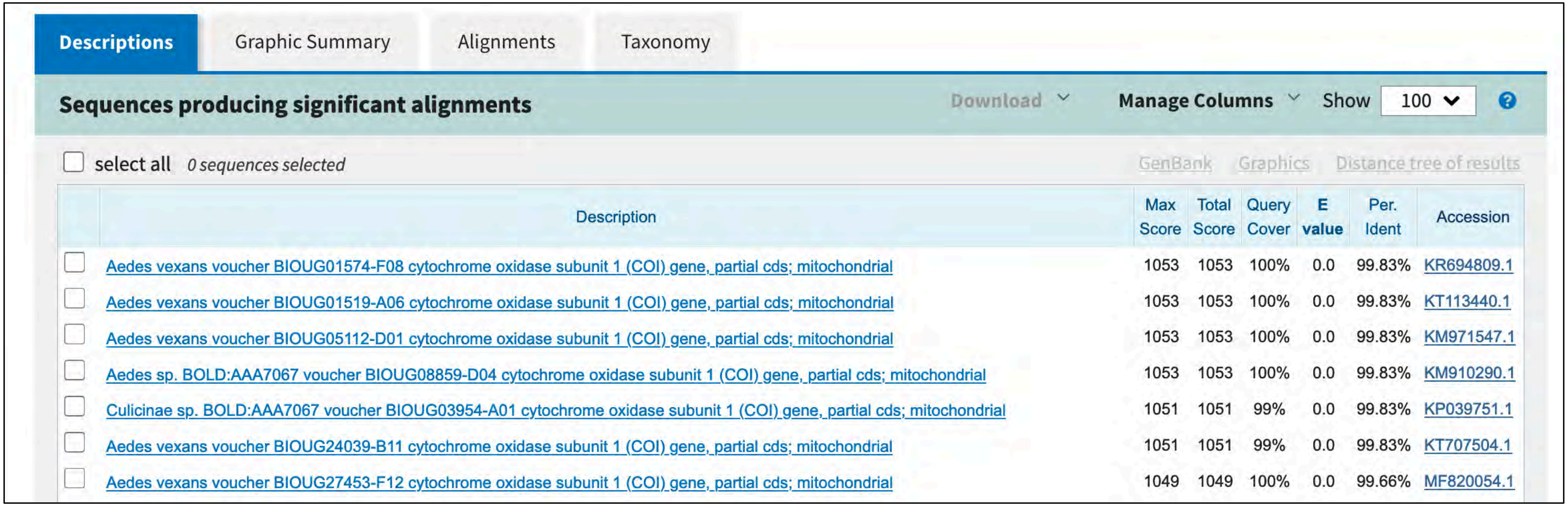


$$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$$

$$\text{Score} = \text{Max}(S)$$

Photo credit
https://www.ncbi.nlm.nih.gov/books/NBK62051/

Let’s do a BLAST



Sequences producing significant alignments

□ select all *0 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
□	Aedes vexans voucher BIOUG01574-F08 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1053	1053	100%	0.0	99.83%	KR694809.1
□	Aedes vexans voucher BIOUG01519-A06 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1053	1053	100%	0.0	99.83%	KT113440.1
□	Aedes vexans voucher BIOUG05112-D01 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1053	1053	100%	0.0	99.83%	KM971547.1
□	Aedes sp. BOLD:AAA7067 voucher BIOUG08859-D04 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1053	1053	100%	0.0	99.83%	KM910290.1
□	Culicinae sp. BOLD:AAA7067 voucher BIOUG03954-A01 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1051	1051	99%	0.0	99.83%	KP039751.1
□	Aedes vexans voucher BIOUG24039-B11 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1051	1051	99%	0.0	99.83%	KT707504.1
□	Aedes vexans voucher BIOUG27453-F12 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	1049	1049	100%	0.0	99.66%	MF820054.1

Some BLAST definitions

- **Max Score**: Highest alignment score (according to a formula)

Some BLAST definitions

- **Max Score**: Highest alignment score (according to a formula)
- **Query Cover**: % of the query length included in aligned segment

Some BLAST definitions

- **Max Score**: Highest alignment score (according to a formula)
- **Query Cover**: % of the query length included in aligned segment
- **E value**: The number of alignments expected by chance with the calculated score or better

Some BLAST definitions

- **Max Score**: Highest alignment score (according to a formula)

- **Query Cover**: % of the query length included in aligned segment

- **E value**: The number of alignments expected by chance with the calculated score or better

- **Per. Identity**: Highest % identity for a set of aligned segments to the same subject sequence.

Does BLAST tell me what species I have identified?

No*

(Some) Limitations to BLAST

- **Homology**: BLAST is trying to indicate which homologous (related by ancestry) sequences are found in the database

(Some) Limitations to BLAST

- **Homology**: BLAST is trying to indicate which homologous (related by ancestry) sequences are found in the database

- **Data base coverage**: BLAST returns its best result; that is not guaranteed to be the true result

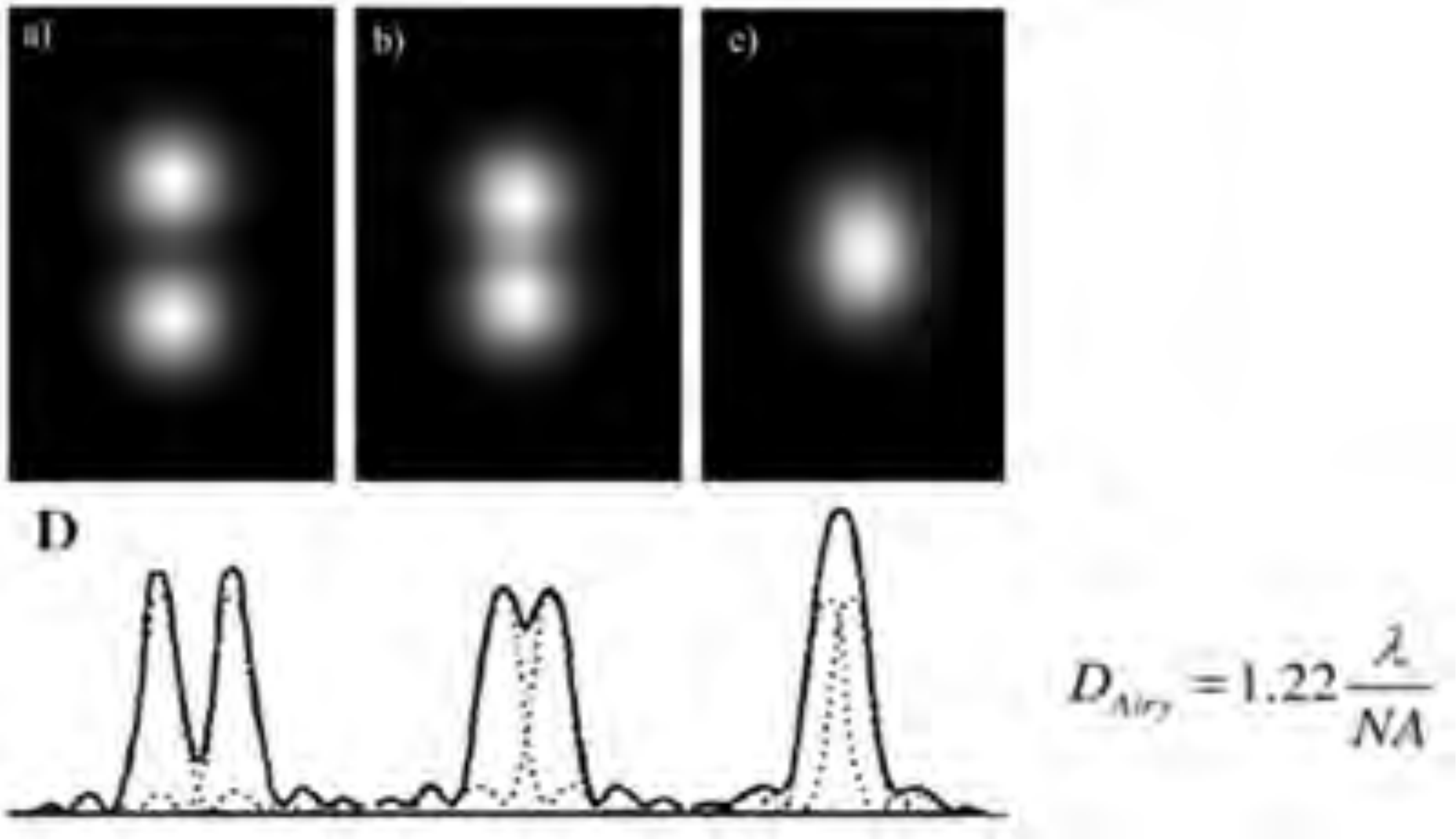
(Some) Limitations to BLAST

- **Homology**: BLAST is trying to indicate which homologous (related by ancestry) sequences are found in the database

- **Data base coverage**: BLAST returns its best result; that is not guaranteed to be the true result

- **Locus resolution**: Barcodes are often good for **genus**-level resolution

A note on resolution (and controls)



$$D_{Airy} = 1.22\frac{\lambda}{NA}$$

Photo credit
http://physwiki.apps01.yorku.ca/index.php?title=Main_Page/BPHS_4090/microscopy_I

Next time:

Multiple sequence alignments and phylogenetics

DNALC Website and Social Media

dnalc.cshl.edu

Cold Spring Harbor Laboratory

DNA Learning Center

Summer & Public Programs · School Programs · Student Research · Educator Training · Resources · DNALC & Partners

DNA LEARNING CENTER

Preparing students and families to thrive in the gene age

LEARN MORE

dnalc.cshl.edu/dnalc-live