# TEACHING THE GENOME GENERATION

## PROTOCOL 6: SEQUENCE ANALYSIS

The Jackson Laboratory
*Leading the search for tomorrow's cures*

## BEFORE YOU BEGIN

**Download the sequence data files:**
*Download and uncompress your .zip file received from The Jackson Laboratory.*

**Download the necessary software:**
*For Macs - Four Peaks: free download nucleobytes.com/4peaks/index.html*
*For PC – Chromas: free download technelysium.com.au/wp/chromas*
*For Netbooks - Create a free account at Benchling benchling.com*

# PRE-REQUISITES & GOALS

## STUDENT PRE-REQUISITES

Prior to implementing this lab, students should understand:

- All previous pre-requisites
- The connection between the genotypes and the DNA sequence
- The benefits of knowing a DNA sequence and its applications
- Recommended: Sanger sequencing method and sequence trace interpretation. Watch Sanger Sequencing of DNA: *www.youtube.com/watch?v=nudG0r9zL2M*

Additionally, students should have completed the Bioinformatics Exercises and understand:

- The DNA sequence for a gene contains both introns and exons and both can harbor sequence variants.
- The effect of different types of mutations.
- How DNA sequences code for proteins and how DNA mutations can affect amino acid sequence.
- What information can be provided by using analysis tools on the NCBI website and BLAST tool.

## STUDENT LEARNING GOALS

1. Interpret sequence quality and genotypes for ACTN3, TAS2R38 and CYP2C19 Exon 5 among sequenced individuals.
2. Locate the specific SNP variant within the DNA sequences.
3. Correlate restriction enzyme results for CYP2C19 data with sequence data, demonstrating that two techniques can be used to genotype.

## NOTES

Watch our video tutorials to take you through the sequencing analysis protocol step-by-step

*For Macs* - Four Peaks
*https://www.youtube.com/playlist?list=PLWNp6Z5dXDZ6pDkLBG7Su1-hldLsC0Yx_*

*For PC* - Chromas
*https://www.youtube.com/playlist?list=PLWNp6Z5dXDZ4TWiDzqfg3SzhCk4X2v9Yx*

*For Netbooks* - Benchling
*https://www.youtube.com/playlist?list=PLWNp6Z5dXDZ6wvvyQdoiD1XaGwu5DFjkv*

# CURRICULUM INTEGRATION

Use the planning notes space provided to reflect on how this protocol will be integrated into your classroom. You'll find every course is different, and you may need to make changes in your preparation or set-up depending on which course you are teaching.

Course name:

**1. What prior knowledge do the students need?**

**2. How much time will this lesson take?**

**3. What materials do I need to prepare in advance?**

**4. Will the students work independently, in pairs, or in small groups?**

**5. What might be challenge points for students during this lesson?**

# MATERIALS

## REQUIRED LAB MATERIALS

Computers with downloaded software

## PROVIDED BY JAX

*For these materials please contact ttgg@jax.org*

DNA sequence files

# PROTOCOL STRUCTURE for Macs and PCs *or Netbooks*

| | |
|---|---|
| **STEPS 1-6** *or 1-14* | 15 minutes |
| | Break point if needed |
| **STEPS 7-9** *or 15-17* | 35 minutes |
| | Break point if needed |
| **STEP 10-16** *or 18-23* | 40 minutes |

# PROCEDURE

## FOR PCs AND MACs

### ☐ STEP 1

Open and log into appropriate software.

### ☐ STEP 2

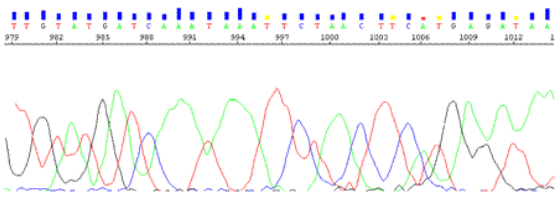
Open the (F)orward sequence (the file should have a .ab1 extension) provided by JAX with the software.

NOTE: The file names are long. Look for your sample name (example HSF) after the JAX unique identifier and before the sequence date. C09_7119_530413_1_HSF_075_2016-03-07.ab1

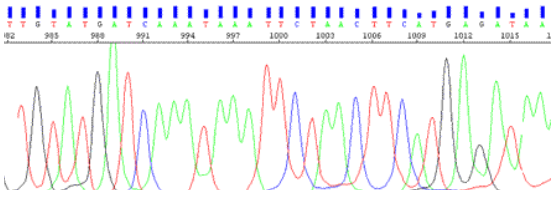
### ☐ STEP 3

Review the sequence and determine where the high quality begins. This high quality area is the beginning of the sequence you will analyze.

NOTE: The first 20-30 bp and the last 10 bp of almost every sequencing reaction are typically of low quality so they should be trimmed off from the data file.



Example of low quality sequence trace data: peaks are broad, jagged and overlapping



Example of high quality sequence trace data: peaks are steep, smooth and non-overlapping

## PLANNING NOTES

## ☐ STEP 4

Isolate the high quality sequence to analyze. To do this either:

1. Highlight and delete the low quality sequence data in first 20-30 bp of the sequence read OR
2. Highlight and copy the high quality section of the sequence after the first 20-30 bases.

## ☐ STEP 5

Copy and paste final edited sequence into a Word or Text document using the following format:

>filename or identifier
Paste here the raw text of edited sequence, either F or R (reversed and complemented)

FOR EXAMPLE:

>ACTN3_DNA1 F
GGACTTAATTTGCGCAATTGNGGCCATATGTTTAAA

## ☐ STEP 6

Save the file as TEXT only. This will enable you to have the sequence for future activities.

BREAK POINT IF NEEDED.

## ☐ STEP 7

Compare your high quality sequence to the NCBI database via BLAST. This can be done by following the:

a. Links for nucleotide sequences in the software that you are using OR
b. Bioinformatics exercises previously completed.

Write down the top hit listed on your BLAST query. FOR EXAMPLE: Homo sapien actinin, alpha 3) (gene/pseudogene) (ACTN3)

## PLANNING NOTES

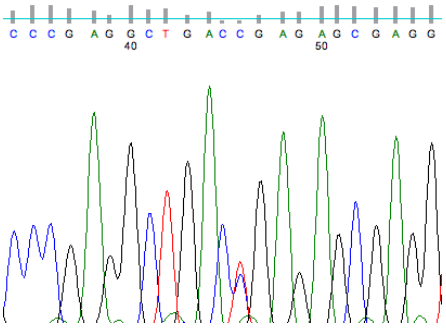
## ☐ STEP 8

Return to Four Peaks or Chromas and find the SNP through visual inspection of the sequence trace (see below).

> NOTE: Even if a particular base is "called" at the SNP location, evaluating the raw trace file may reveal two different bases at that location.

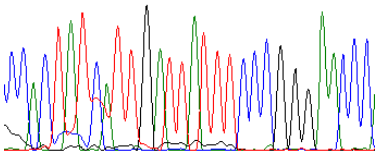
## FOR ACTN3

Look for the GGCTGAC sequence (by using Ctrl or Command F) around the 40th base in the original sequence. The SNP is the next downstream base and will either be a C (wild type) or a T (variant). The T creates a premature stop codon (TGA). This example is heterozygous (C/T) as it shows both bases.




## FOR CYP2C19

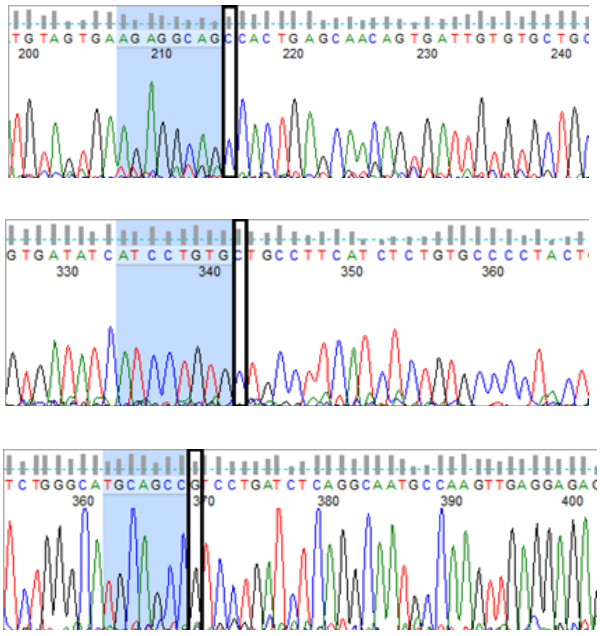
Look for the ATTTCCC sequence (Ctrl or Command F) in the original sequence. The SNP is the next downstream base and will either be a G (wild type) or an A (variant). The A variant creates an aberrant splice site that causes a frameshift. This example is homozygous wild type (G/G).

## PLANNING NOTES

## FOR TAS2R38

To find the first locus SNP, look for the AGAGGCAG signature, second locus: ATCCTGTG signature and third locus: TGCAGCC signature (Ctrl or Command F) in the original sequence. The SNP is the next downstream base and will either be a G (wild type) or a C (variant) for the first locus, either a T (wild type) or a C (variant) for the second locus, and either an A (wild type) or a G (variant) for the third locus.  Example below is homozygous CCG.




## ☐ STEP 9

Determine the genotype of your sample.

Gene name:

Sample name:

Genotype:

BREAK POINT IF NEEDED.

## PLANNING NOTES

## ☐ STEP 10

Open the (R)everse file of the same DNA sample provided by JAX with the software.

## ☐ STEP 11

The software that you are using will have a way to **reverse** or **flip sequence**. Since DNA strands are anti-parallel the (R)everse read must be reoriented to compare it to the (F)orward read.

NOTE: This is process is done for two reasons:

1. As a double check of the (F)orward read
2. In the event that the (F)orward read is of such poor quality that it cannot be analyzed.

## ☐ STEP 12

Repeat Steps 5-9 with (R)everse sequence file.

## Gene name:
## Sequence name:
## Genotype:

## ☐ STEP 13

Create a consensus sequence with data from the (F) orward and (R)everse reads by navigating to *www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html*

NOTE: A consensus sequence will align the overlapping regions of the two reads to double check sequence calls. Any PCR product greater than 650 bp in length will have significant overlap, if not complete overlap.

## ☐ STEP 14

In the first window, enter your (F)orward sequence from your text file by using **Ctrl** or **Command C** to copy the sequence and **Ctrl** or **Command V** to paste the sequence in the window.

## ☐ STEP 15

Do the same in the second window with the (R)everse sequence.

## PLANNING NOTES

## ☐ STEP 16

Default settings are fine for this purpose, so click **Submit**.

NOTE: The (F)orward read will be on the top, the (R)everse read on the bottom for each segment of alignment. A line ( | )between two bases indicates consensus between two sequences, and a dot ( . ) indicates a difference.

### How much overlap is there between the (F)orward and (R)everse read? Hint: Look at the Identity ratio

### Do the (F)orward and (R)everse reads indicate the same genotype? If not, provide a possible explanation.

## FOR NETBOOKS

## ☐ STEP 1

Create a Benchling account at *benchling.com* by clicking **Join Benchling**.

## ☐ STEP 2

On your lab notebook page, click the **Create** drop down menu in the top right hand corner.

## ☐ STEP 3

From the drop down menu select **Import sequences**.

## ☐ STEP 4

Click **Choose folder** to select where your sequence will be uploaded. The **My Project** folder is fine.

## ☐ STEP 5

Either drag and drop your "F" or "Forward" sequence into the box, or click **Choose a file** and

## PLANNING NOTES

navigate to your forward sequence.

NOTE:

1. To find your sequence in the data files consult (See STEP 2 in For PCs and Macs)

2. The sequence will upload automatically. Click **Close** when upload is complete.

## ☐ STEP 6

Double click the imported file to open it. You will see the sequence in FASTA form.

NOTE: The first time you log on you will be taken through a short tutorial.

## ☐ STEP 7

In order to view the trace file and trim the sequence for quality, click the alignment button  on the right hand tool bar and select the first file under **Saved alignments**.

## ☐ STEP 8

You should now be able to see the trace of the base calls (red (T), green (A), blue (C) and black (G) lines) and the quality of those calls (indicated by gray bars).

NOTES:

a. You can scan through your sequence by clicking on areas of the dark gray bars at the bottom of the screen.

b. High quality is generally any base with a score >40.

c. Heterozygous regions will have low quality scores, but that is expected.

## ☐ STEP 9

Trim your sequence for quality by sliding the black vertical bar or by right clicking on the base where

## PLANNING NOTES

high quality begins and select **Trim** and **to start**.

NOTES:

a. You should select the beginning of a large chunk of high quality scores.

b. The trimmed sequence will become shaded gray.

## ☐ STEP 10

Trim your sequence for quality by right clicking on the base where high quality ends. Select **Trim** and **to end**.

NOTE: Your high quality sequence should be unshaded.

## ☐ STEP 11

Select your high quality sequence by left clicking and dragging through the entire region.

Note: Use the dark gray bars as the bottom of the screen to change the viewing window.

## ☐ STEP 12

Right click on your highlighted sequence and select **Copy**.

## ☐ STEP 13

Copy (as **Sequence**) and paste into a Word or Text document using the following format:

>filename or identifier
Paste here the raw text of edited sequence,
either F or R (reversed and complemented)

FOR EXAMPLE:

>ACTN3_DNA1 F
GGACTTAATTTGCGCAATTGNGGCCATATGTTTAAA

## ☐ STEP 14

Save the file as TEXT only. This will enable you to have the sequence for future activities.

## PLANNING NOTES

BREAK POINT IF NEEDED.

## ☐ STEP 15

Compare your high quality sequence to the NCBI database via BLAST. This can be done by following the:

a. Links for nucleotide sequences in the software that you are using OR

b. Bioinformatics exercises previously completed.

Write down the top hit listed on your BLAST query. FOR EXAMPLE: Homo sapien actinin, alpha 3) (gene/pseudogene) (ACTN3)

## ☐ STEP 16

Return to Benchling and find the SNP polymorphism by sequence comparison.

NOTES:

a. To locate the reference sequence of interest select all on your trace file by hitting Ctrl+A and then Ctrl+F to bring up the find toolbar. Type sequence appropriate for the gene of interest (See STEP 8 For PCs and Macs) into window and program will automatically navigate to the region.

b. You can zoom in on the trace to assess the base call by using the vertical scroll bar on the left hand side of the trace window and shortening the viewing window on the gray bars at the bottom of the screen.

## ☐ STEP 17

Determine the genotype of your sample.
See STEP 8 For PCs and Macs.

Gene name:
Sequence name:
Genotype:

## PLANNING NOTES

BREAK POINT IF NEEDED.

## ☐ STEP 18

Repeat Steps 3-14 with the "R" or "Reverse" sequence file.

## ☐ STEP 19

Right click on your highlighted high quality sequence and copy (as **reverse complement**) and paste into your text file. This takes the "Reverse" read and re-orients it to match the "Forward" read.

NOTE: This is done for two reasons:

a. As a double check of the "Forward" read

b. In the event that the "Forward" read is of such poor quality that it cannot be analyzed, which is often the case with ACTN3.

## ☐ STEP 20

Align the full length forward and reverse reads to create a consensus sequence by opening one of the sequences and clicking the alignment button on the right hand tool bar and **Create new alignment**.

## ☐ STEP 21

**Search for a sequence** using the name of the opposing sequence (i.e.: if you opened forward, you should select reverse).

## ☐ STEP 22

Click on the first sequence name to turn off the **Template** button. This will create a new consensus alignment between the two sequences.

NOTE: Under Algorithm, use the MAFFT for this alignment as it will automatically reverse your R sequence.

## ☐ STEP 23

Click **Create alignment**.

NOTES:

a. You should see three horizontal bars at the bottom of your screen. The top is the consensus

## PLANNING NOTES

bar indicating the agreement between the two sequences. The middle is the first sequence file and the bottom is the second sequence file. Gray indicates agreement between the two files, whereas red indicates that only one file has that specific sequence. There should be a region of gray overlap between the files somewhere in the middle with red on the ends.

b. Any PCR product less than ~650 bp will have significant overlap, if not complete overlap.

## How much overlap is there between the (F)orward and (R)everse read?

## Do the (F)orward and (R)everse reads indicate the same genotype? If not, provide a possible explanation.

## Sources of Potential Error:

The most common errors in PROTOCOL 6 include:

- Copying the wrong sequence
- Using the low quality section of the read (first 20-30 bases)
- Low quality of entire sequence, which would be due to:
  - Pipetting errors in previous protocols
  - Target gene not amplified
  - Contamination of DNA samples or reagents

## PLANNING NOTES

## NEED HELP?

Email the experts – *ttgg@jax.org*