TEACHING THE GENOME GENERATION

Identify a Variant using BLAST

The Jackson Laboratory

Leading the search for tomorrow's cures

NIH National Institutes of Health
Turning Discovery Into Health

SEPA SCIENCE EDUCATION PARTNERSHIP AWARD

SUPPORTED BY THE NATIONAL INSTITUTES OF HEALTH

Identify a Variant using BLAST

Introduction

Background

The researchers sequenced your daughters' DNA and identified differences between their genomes and the human reference genome. Based on the variants in your daughters' DNA variants, they eventually identified ten people with similar symptoms and similar variants from four other families. The researchers sequenced those patients' genomes too.

In this activity, you will follow the researchers' process to learn how they identified a genetic variant that could be contributing to your daughters' disorder. You will also investigate the types of sequences that researchers can compare and discuss why these researchers might choose to compare certain types of sequences over others.

Central Dogma Review

You may have heard of the central dogma of biology: DNA is transcribed into RNA, which is translated into protein. We can compare any of these sequences—DNA, RNA, or protein—across individuals. In this exercise, you will compare individual DNA, RNA, and protein sequences to a reference sequence. Before you begin, let's review core ideas and important vocabulary from the central dogma.

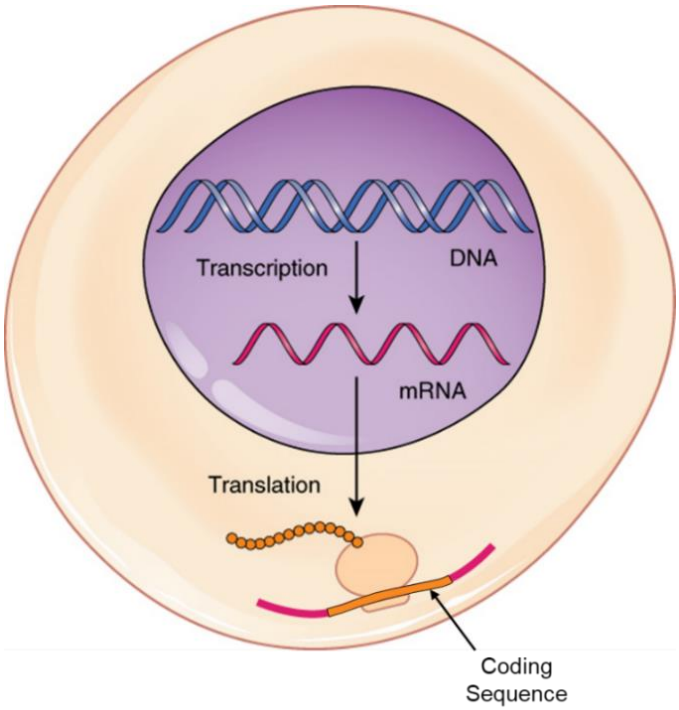

Figure 1. **Central Dogma**[1].

A **gene** is a stretch of DNA, parts of which code for a particular protein. The information contained in the DNA sequence needs to get transported from the nucleus to the cytoplasm in order to be used by the cell to make protein. To accomplish this, the DNA gets transcribed into an RNA molecule, which is processed to form **messenger RNA** (mRNA).

When the cell translates the mRNA into protein, a ribosome will scan the mRNA until it finds a series of three nucleotides called a **start codon**. The start codon tells the ribosome where in the mRNA the protein coding sequence starts. The ribosome will then add amino acids in a chain until it reaches the series of three nucleotides called the **stop codon**. This section from the start codon to the stop codon, which gets translated into protein, is called the **coding sequence**. The coding sequence does not include the sections of the mRNA before the start codon or after the stop codon.

[1] Adapted from OpenStax under the Creative Commons Attribution 4.0 International license.

Knowledge Check

1. Order the following sequences from longest to shortest, based on the number of nucleotides.
    a. mRNA, coding sequence, gene

2. What are the two main processes of the central dogma?

Glossary

Gene – a DNA sequence that codes for a protein.

mRNA – a molecule that includes the information from a gene that can be used to create protein.

Coding sequence – the portion of an mRNA sequence that codes for a protein. The coding sequence starts with a start codon and ends with a stop codon.

Protein sequence – the sequence of amino acids that make up a fully translated protein.

Codon Chart

First Base	Second Base: T		Second Base: C		Second Base: A		Second Base: G		Third Base
T	TTT	Phe (F)	TCT	Ser (S)	TAT	Tyr (Y)	TGT	Cys (C)	T
	TTC		TCC		TAC		TGC		C
	TTA	Leu (L)	TCA		TAA	**Stop**	TGA	**Stop**	A
	TTG		TCG		TAG	**Stop**	TGG	Trp (W)	G
C	CTT	Leu (L)	CCT	Pro (P)	CAT	His (H)	CGT	Arg (R)	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Gln (Q)	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Ile (I)	ACT	Thr (T)	AAT	Asn (N)	AGT	Ser (S)	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lys (K)	AGA	Arg (R)	A
	ATG	**Start** Met (M)	ACG		AAG		AGG		G
G	GTT	Val (V)	GCT	Ala (A)	GAT	Asp (D)	GGT	Gly (G)	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glu (E)	GGA		A
	GTG		GCG		GAG		GGG		G

Figure 2. Codon Chart.

When ribosomes translate mRNA into protein, they read the mRNA in series of three base pairs called a **codon**. A **codon chart** helps us to determine which amino acid each mRNA codon will get translated into.

To read a codon chart, find the first base pair in the codon from the left side of the chart to select a row. Then find the second base pair from the top of the chart to select a column. Finally, select the third base pair from the right side of the chart to find your specific codon. The corresponding amino acid will be directly to the right of the codon. Amino acids can be abbreviated in two ways: with a three-letter code or with a one-letter code.

Try it! The codon ACG codes for the amino acid threonine (Thr, T). Can you find this codon on the chart?

Did you know? RNA molecules use a nucleotide called uracil (U) instead of thymine (T). However, most sequence databases use T to represent both thymine and uracil. So, when you see an mRNA sequence from a sequence database, it will usually contain T's instead of U's. For that reason, the codon chart for this activity also uses T to represent uracil.

Knowledge Check

1. What amino acid does the codon TGG code for?

2. What amino acid does the codon CAT code for?

3. What amino acid does the codon CGA code for?

4. What is the start codon? Which amino acid does it code for?

5. What are the stop codons? Do they code for an amino acid?

Part 1a: Family 2 Variant - Identify the Gene

In this activity, you will follow the researchers' process to learn how they identified a genetic variant that could be contributing to your daughters' disorder. Using sequencing, the researchers were able to identify rare variants in your daughters' and the other patients' genomes. You will follow a similar process to compare the patients' sequences to the reference genome and identify differences.

See BLAST Written Tutorial: Identifying a Gene

First search for a match for one of the patient's sequences in the entire human genome using NCBI BLAST.

1. Locate the Patient DNA Sequences. Select and copy the sequence for **Family 2 Allele 1**. This sequence is a small portion of an mRNA sequence for our potential gene of interest.

2. Navigate to NCBI Nucleotide BLAST. Paste the sequence for **Family 2 Allele 1** in the box under "Enter Query Sequence."

3. In the section "Choose Source Set", type "Homo sapiens (taxid:9606)" into the box next to "Organism." This tells the software that we are only interested in looking for matching sequences from the human genome.

4. Scroll to the bottom and click "Blast." It may take several seconds for the results to appear. When the results appear, scroll down to the table and find the "Descriptions" tab.

    a. Look at the "Description" column. What do you notice about the names of the sequences?

    b. What gene does this mRNA sequence come from?

Part 1b: Family 2 Variant - Compare to a Reference Sequence

Now that you've identified the gene that the sequence comes from, you need to pick one of the aligned sequences as your reference sequence. A **reference sequence** is a DNA or RNA sequence that is assumed by scientists to be a representative example of the genetic material of a specific species. Reference sequences are typically created by combining sequences from multiple individuals of the same species.

The researchers used "transcript variant 1, mRNA" as their reference sequence, so you should select that as your reference sequence too. You can then compare the patient's sequence to your selected reference sequence and predict how any differences might affect the resulting protein.

5. Click the box next to "select all" at the top of the "Descriptions" tab until all of the boxes are **unselected**. Then select the box next to the "transcript variant 1, mRNA" only.

6. Click on the "Alignments" tab to see how the query sequence you submitted compares with the reference sequence you just selected. In the dropdown menu next to "Alignment View", select the option "Pairwise with dots for identities."

   The **Query** sequence on the top line is the patient sequence (Family 2 Allele 1). The Subject (**Sbjct**) sequence on the bottom line is the reference sequence (transcript variant 1).

   Dots in the **Sbjct** sequence represent nucleotides where the reference sequence matches the query sequence. The numbers at the start and end of each line represent the location of the first nucleotide in the comparison within the whole sequence.

   a. What is the difference between the sequence of Family 2 Allele 1 and the reference sequence? At which nucleotide(s) in the reference sequence does this variant occur?

   b. Use the codon chart to predict what change this variant in Family 2 Allele 1 might cause in the amino acid sequence of the translated protein, if any.
   Hint: This sequence is from the middle of the coding sequence, so you do not need to find a start codon. You can assume the first base of the sequence is the first base of a codon.

Next, check your prediction about whether the variant causes a change in the protein by using the CDS feature. This feature identifies the coding region and translates both the reference and query sequence into a protein sequence. In the protein sequence, amino acids are represented by letters. A stop codon is represented by an asterisk (*).

7. To check your prediction, click the checkbox next to "CDS feature".

   a. What change does this variant in Family 2 Allele 1 cause in the amino acid sequence of the translated protein, if any? At which amino acid in the reference sequence does this change occur? Does this match your prediction?

Next, examine how similar Family 2 Allele 1 is to the reference sequence. BLAST provides an automatic percent identity calculation, which we can use as a starting point for quantitatively comparing the sequences.

8. Locate the percent identity (labeled "Identities") for this nucleotide comparison.

    a. What is the percent identity determined by BLAST?

    b. How was this percent identity value calculated?

In this exercise, you only entered a portion of the Family 2 Allele 1 sequence as the query sequence, since the entire mRNA sequence is 4978 nucleotides long! However, in order to get a more accurate percent identity value, you should calculate percent identity based on the complete sequence.

9. Recalculate the sequence percent identity based on the full length of the gene, mRNA, coding, and protein sequences. You can assume that the rest of the Family 2 Allele 1 sequence is identical to the reference sequence.

    a. Calculate percent identity for each sequence. Round to two decimal places.
        i. Whole gene: 292,344 nucleotides

        ii. mRNA: 4978 nucleotides

        iii. Coding sequence: 2112 nucleotides

        iv. Protein: 703 amino acids

    b. Which of these percent identity values do you think would be most useful for comparing the patient sequences to the reference human genome sequence? Justify your answer.

Part 2: Family 1 Variant

Next, look at one of the alleles from Family 1.

1. Locate the Patient DNA Sequences. Select and copy the sequence for **Family 1 Allele 1**. This sequence is a small portion of an mRNA sequence for our potential gene of interest.

2. Navigate to NCBI Nucleotide BLAST. Paste the sequence for **Family 1 Allele 1** in the box under "Enter Query Sequence."

3. In the next section, "Choose Source Set", type "Homo sapiens (taxid:9606)" into the box next to "Organism."

4. Scroll to the bottom and click "Blast." It may take several seconds for the results to appear. When the results appear, scroll down to the table and find the "Descriptions" tab.

5. Click the box next to "select all" at the top of the "Descriptions" tab until all of the boxes are unselected. Then select the box next to the "transcript variant 1, mRNA" only.

6. Click on the "Alignments" tab to see how the query sequence you submitted compares with the reference sequence you just selected. In the dropdown menu next to "Alignment View", select the option "Pairwise with dots for identities."

    a. What is the difference between the sequence of this allele and the reference sequence? At which nucleotide(s) in the reference sequence does this difference occur?

    b. Use the codon chart to predict what change this variant in Family 2 Allele 1 might cause in the amino acid sequence of the translated protein, if any.
    Hint: This sequence is from the middle of the coding sequence, so you do not need to find a start codon. You can assume the first base of the sequence is the first base of a codon.

7. To check your prediction, click the checkbox next to "CDS feature".

    a. What change does this variant in Family 1 Allele 1 cause in the amino acid sequence of the translated protein, if any? At which amino acid in the reference sequence does this change occur? Does this match your prediction?

b. How is this change similar to or different from the change caused by the variant from Family 2?

8. Locate the percent identity (labeled "Identities") for this nucleotide comparison.

   a. What is the percent identity determined by BLAST?

   b. How was this percent identity value calculated?

9. Recalculate the sequence percent identity based on the full length of the gene, mRNA, coding, and protein sequences. You can assume that the rest of the Family 1 Allele 1 sequence is identical to the reference sequence.

   a. Calculate percent identity for each sequence. Round to two decimal places.
      i. Whole gene: 292,344 nucleotides

      ii. mRNA: 4978 nucleotides

      iii. Coding sequence: 2112 nucleotides
      Hint: A stop codon represents the end of the coding sequence. This variant occurs at nucleotide 1975 in the coding sequence, and it is the first base in the stop codon. So, the coding sequence with this variant is 1977 nucleotides.

      iv. Protein: 703 amino acids
      Hint: Remember that a stop codon ends translation. This variant occurs in place of amino acid 659.

b. Which of these percent identity values do you think would be most useful for comparing the patient sequences to the reference human genome sequence? Justify your answer.

Part 3: Family 4 Variant

In Family 4, one variant appeared at first glance to be a single nucleotide change in the DNA, just like the two variants you've already examined. However, this single nucleotide difference causes a change in the way the mRNA is processed.

After DNA is transcribed into RNA, the RNA molecule goes through a processing step. Portions of the gene sequence, called **introns**, are removed. The remaining portions of the gene sequence, called **exons**, are kept and stitched together to form the final mRNA molecule.

The variant in Family 4 causes a whole exon to get left out of the middle of the mRNA. So even though the DNA only has a change in a single nucleotide, the mRNA is missing 122 nucleotides!

1. Based on this information, calculate percent identity for this variant for each sequence. Round to two decimal places.
    a. Whole gene: 292,344 nucleotides

    b. mRNA: 4978 nucleotides
    Hint: This is length of the reference sequence.

This variant causes an even bigger change in the coding sequence and protein. There are 122 nucleotides missing in the mRNA. Since 122 is not divisible by three, the variant produces a shift in how the coding sequence gets divided into codons.

In the reference sequence, the last two nucleotides of exon 7 (TG) combine with the first nucleotide of exon 8 (G) to produce the codon TGG. During translation, the codon TGG translates to the amino acid tryptophan (abbreviated Trp or W).

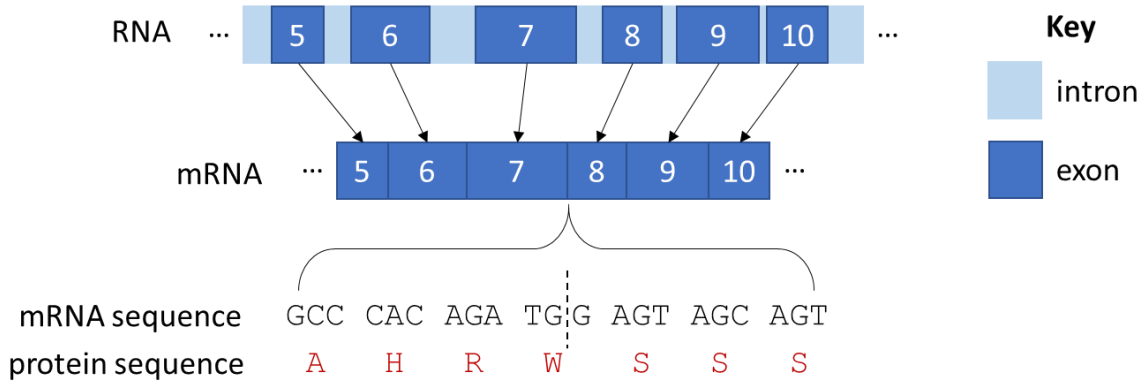

Figure 3. RNA processing without variant. This figure shows a portion of the *ATG7* gene after it has been transcribed into RNA. The RNA then gets processed into mRNA, which contains only exons. In the mRNA sequence without the variant, the last two nucleotides from exon 7 (TG) combine with the first nucleotide of exon 8 (G) to produce the codon TGG.

With the variant, exon 8 is missing, so the last two nucleotides of exon 7 (TG) combine with the first nucleotide of exon 9 (A) to produce the codon TGA. In translation, TGA is a stop codon.

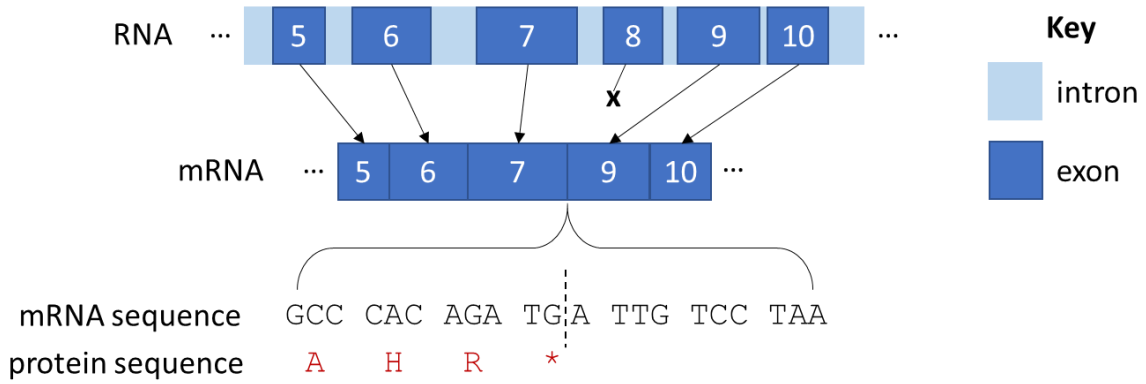

Figure 4. RNA processing with variant. This figure shows a portion of the *ATG7* gene after it has been transcribed into RNA. The RNA then gets processed into mRNA, which contains only exons. In the mRNA sequence with the variant, exon 8 is excluded from the processed mRNA. The last two nucleotides from exon 7 (TG) combine with the first nucleotide of exon 9 (A) to produce the stop codon TGA. In the protein sequence, the stop codon is represented by an asterisk (*).

So, with this variant, translation will end much earlier than usual. This means that this variant has a coding sequence that is only 768 nucleotides long, producing a protein 255 amino acids long.

2. Based on this information, calculate the percent identity for this variant for each sequence. Round to two decimal places.
    a. Coding sequence: 2112 nucleotides

    b. Protein: 703 amino acids

3. Which of the percent identity values for Family 4 do you think would be most useful? Justify your answer.

Part 4. A Silent Variant

While the patients in this study had DNA variants that affected the amino acid sequence of the protein, not all DNA variants will impact the protein sequence. Let's look at an example.

1. Locate the Patient DNA Sequences. Select and copy the sequence called **Unknown Variant**. This sequence is a small portion of a mRNA sequence for our potential gene of interest.

2. Navigate to NCBI Nucleotide BLAST. Paste the sequence **Unknown Variant** in the box under "Enter Query Sequence."

3. In the next section, "Choose Source Set", type "Homo sapiens (taxid:9606)" into the box next to "Organism."

4. Scroll to the bottom and click "Blast." It may take several seconds for the results to appear. When the results appear, scroll down to the table and find the "Descriptions" tab.

5. Click the box next to "select all" at the top of the "Descriptions" tab until all of the boxes are unselected. Then select the box next to the "transcript variant 1, mRNA" only.

6. Click on the "Alignments" tab to see how the query sequence you submitted compares with the reference sequence you just selected. In the dropdown menu next to "Alignment View", select the option "Pairwise with dots for identities."

    a. Is there a difference between the Silent Variant sequence and the reference sequence? At which nucleotide(s) in the reference sequence does this difference occur?

    b. Use the codon chart to predict what change this variant might cause in the amino acid sequence of the translated protein, if any.
    Hint: This sequence is from the middle of the coding sequence, so you do not need to find a start codon. You can assume the first base of the sequence is the first base of a codon.

7. To check your prediction, click the checkbox next to "CDS feature".

    a. What change does this variant cause in the amino acid sequence of the translated protein, if any? At which amino acid in the reference sequence does this change occur? Does this match your prediction?

b. How is this similar to or different from the change caused by the variants from Families 1 and 2?

8. Locate the percent identity (labeled “Identities”) for this nucleotide comparison.

   a. What is the percent identity determined by BLAST?

   b. How was this percent identity value calculated?

9. Recalculate the sequence percent identity based on the full length of the gene, mRNA, coding, and protein sequences. You can assume that the rest of the Silent Variant sequence is identical to the reference sequence.

   a. Calculate percent identity for each sequence. Round to two decimal places.
      i. Whole gene: 292,344 nucleotides

      ii. mRNA: 4978 nucleotides

      iii. Coding sequence: 2112 nucleotides

      iv. Protein: 703 amino acids

   b. Which of these percent identity values do you think would be most useful for comparing this sequence to the reference human genome sequence? Justify your answer.

Part 5: Reflection and Guiding Questions

Activity Reflection

1. What are the possible ways a DNA variant in a gene could affect an amino acid sequence?

2. Since the researchers are studying the variants for all of the families, they need to use a consistent comparison type and percent identity calculation across all of the different variants they identified. What comparison type (whole gene, mRNA, coding sequence, or protein) did you select as the most useful comparison for each patient variant (question 9b from Parts 1 and 2; question 3 from Part 3)? Did you select the same comparison type for each variant?

   If so, can you think of a type of DNA change or variant where that comparison would not be the most useful?

   If not, which comparison would you choose to be consistent across all of the patients? Why?

3. The researchers identified additional variants in the families' DNA in other genes besides *ATG7*. The researchers chose to further investigate the *ATG7* variants in part because those variants are very rare compared to all publicly available human genome sequences. Given what you know from the scenario, why do you think the researchers made this choice?

Guiding Questions Reflection

Revisit the following guiding questions and update your answers to include anything you've learned during this activity.

Given what you know about genome sequencing and genetic variation:

1. What can we learn from comparing genetic information across individuals and species?

Given what you know about (a) how DNA codes for proteins and (b) the connection between protein structure and function:

2. How might a DNA variant affect protein sequence, structure, or function?

Patient DNA Sequences

Note: If BLAST returns an error after copying sequences from this PDF, try instead to copy and paste these sequences from our Patient DNA Sequences Word document.

Part 1. Family 2 Allele 1

>Family 2 Allele 1
TGCACTGTGAGTCATCCAGGACTGGCCGTGATTGCAGGAGCCCTGGCCGTGGAATTGATG

Part 2. Family 1 Allele 1

>Family 1 Allele 1
TATGAATGAGAAGGATTTAACTTCCTAGCCAAGGTGTTTAATTCTTCACATTCCTTCTTA

Part 4. A Silent Variant

>Unknown Variant
TGCACTGTGAGTCGTCCAGGACTAGCCGTGATTGCAGGAGCCCTGGCCGTGGAATTGATG