TEACHING THE GENOME GENERATION

Sequence Comparison and Identity

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

Contents

Sequence Comparison and Identity Module Overview

Introduction & Theme

This module aims to introduce students to the concepts of sequence comparison and identity. Comparing DNA and protein sequences is a core component of bioinformatics analyses, and sequence comparison techniques can be applied in a variety of fields, from human health to evolution. Within the context of sequence comparisons, the activities in this module reinforce math skills, the fundamentals of the central dogma, and the connection between protein structure and function.

The first set of activities utilizes the five genes from the Teaching the Genome Generation™ (TtGG) laboratory curriculum to give teachers and students an entry point into sequence comparison and identity using genes they may already be familiar with. The TtGG laboratory curriculum was designed to explore human genetic variation, with a focus on variants in five different genes: *ACE*, *ACTN3*, *OXTR*, *CYP2C19*, and *TAS2R38*. Each of these genes are associated with an interesting phenotype and has a different type of molecular variant.

The second part of the module centers around a real research study. In the study, researchers sequenced the genomes of two sisters who had a previously undiagnosed rare neurodevelopmental disorder. After identifying a potentially causal variant in the gene **Autophagy related 7** (*ATG7*), the researchers searched for other families with similar variants. Ultimately, the researchers identified 12 individuals from 5 families with similar systems, who all had rare recessive variants in *ATG7*.

The activities in the second part of the module guide students through parts of the process that the researchers used to identify the *ATG7* variants and establish that the variants have an impact on ATG7 protein structure and function.

Note: Per common practice regarding gene and protein nomenclature, gene symbols throughout these lessons are displayed in italics, while protein names and symbols are not italicized.

Reference: Collier, J.J., et al. (2021). Developmental Consequences of Defective ATG7-Mediated Autophagy in Humans. *N Engl J Med, 384*:2406-2417. https://dx.doi.org/10.1056/NEJMoa1915722

Learning Outcomes

Essential Question

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

Enduring Understandings

- Changes in the genetic code can affect changes in protein structure, which can change protein function.
- Bioinformatics tools enable us to understand biological data.
- Bioinformatics and genetic information can inform new ideas in a variety of fields, from evolution to human health.

Skills

Data analysis and interpretation; computation (math skills); asking questions and formulating hypotheses; problem solving; pattern recognition; constructing explanations

Lesson & Activity Descriptions

Introduction to Sequence Comparison

This brief reading and knowledge check introduce students to the concept of sequence comparison and the formula for percent identity calculations.

Sequence Comparison with the TtGG Genes

This suite of five activities introduces applications of sequence comparison through the TtGG genes. There is one version of the activity for each gene, and all of the activities cover the same concepts. Students practice applying sequence comparison and percent identity calculations between individuals, across species, and across gene families. These activities do not require the use of online databases.

Sequencing for Rare Disease Diagnosis

Sequencing for Rare Disease Diagnosis: Scenario Introduction

This short reading and associated questions introduce the research study for the second half of the module, with a focus on one of the families who participated in the study.

Identifying a Variant Using BLAST

In this activity, students use BLAST to identify variants in *ATG7* from several of the patient families, including missense and nonsense variants. Students then assess the impact of the variant on the gene, mRNA, coding, and protein sequences using percent identity calculations. At the end of activity, students reflect on the different impacts that variants can have on mRNA and protein sequences.

Connecting Protein Structure and Function using PolyPhen-2, UniProt, and BLAST

In the first part of this activity, students use the online tool PolyPhen-2 to predict the effect of a specific patient variant on the function of Autophagy related 7 protein (ATG7). Students then brainstorm what types of information PolyPhen-2 could be using to make this prediction.

In the second part, students learn about ATG7 protein structure and function using the UniProt database and simplified structural models. Students use this information, plus information about the location of the variant, to assess how this patient variant could impact the protein's structure and function.

In the third part, students use protein BLAST to compare the human ATG7 protein sequence to the fruit fly, chimpanzee, mouse, chicken, and yeast ATG7 sequences. Students calculate percent identity for the whole protein sequence and for a small region of the sequence around the patient variant. Students use that data to discover that the patient variant occurs within an evolutionarily conserved sequence and consider how that information could be used to predict whether variant affects protein function.

Finally, students reflect on their original hypothesis regarding the type of information PolyPhen-2 uses to make variant effect predictions.

Final Reflection and Bioethics

For the final reflection, students reflect on the guiding questions for the module, summarizing their learnings. Then, in the bioethics exploration, students address ethics questions focused on the goals and outcomes of rare disease research and the inequities in patient access to researchers, healthcare, and a diagnosis.

Lessons & Activities

Introduction to Sequence Comparison

Note: There are two options included in the module for introducing sequence comparison to your students. The first is this reading and associated knowledge check, which you can assign to students to complete either in class or in advance of class. The second is presenting the Introduction to Sequence Comparison Teacher Slides. You are welcome to choose either option or use both.

Sequence comparison is a technique to determine how similar two or more nucleotide or protein sequences are to each other. As you complete the activity, you will discover different applications of sequence comparison used by scientists.

Sequence Alignment

In sequence comparison, the first step is aligning the sequences. The goal of sequence alignment is to line up the nucleotides or amino acids of each sequence such that there are as many matches as possible. Researchers typically use computer programs, like the online Basic Local Alignment Search Tool (BLAST), to align sequences.

BLAST displays aligned sequences one above the other. The **Query** sequence is typically the sequence of interest and the Subject (**Sbjct**) sequence is typically the sequence you are comparing to.

```
Query  1      CACTGCCCGAGGCTGACCGAGAGCGAGGTGCCATCATGGGCATCCAGGGTGAGATCCAGA  60
Sbjct  18688  ............................................................  18747
```

The numbers at the beginning and end of each sequence represent the position of the first and last nucleotide or amino acid within the whole sequence. Dots in the Subject sequence line indicate positions at which the nucleotides or amino acids **match** the Query sequence. In the above example, all the nucleotides in the query and subject sequences match.

Sequence Comparison

After the sequences are aligned, the next step is to identify differences between the sequences. There are two main kinds of differences: mismatches and gaps.

Mismatches

Mismatches are positions where the nucleotides or amino acids in the sequences are not the same, or do not match. Mismatches are represented as letters in the Subject sequence. In the example below, there are two mismatches: the Query sequence has a cytosine (C) at one position where the Subject sequence has a thymine (T), and at another position where the Subject sequence has a guanine (G).

```
Query  1      CACCGCCCGAGGCTCACCGAGAGCGAGGTGCCATCATGGGCATCCAGGGTGAGATCCAGA  60
Sbjct  18688  ...T.........G............................................  18747
```

Gaps

Gaps are positions where one sequence has one or more nucleotides or amino acids that the other sequence does not have. Gaps are represented as dashes (-). In the example below, there are two gaps: the Query sequence has an extra thymine (T) relative to the Subject sequence, and the Subject sequence has an extra cytosine (C) relative to the Query sequence.

```
Query  1      CACTGCCCGATGGCTGACCGAGAGCGAGGTGCCAT-ATGGGCATCCAGGGTGAGATCCAG  59
Sbjct  18688  ..........-........................C.......................  18746
```

Percent Identity

When comparing sequences, **percent identity** provides a measure of how similar two sequences are. The formula for percent identity uses the total number of nucleotide or amino acid **positions** in the sequence comparison and the number of nucleotide or amino acid positions that are different, or **divergent**, between the sequences:

$$Percent\ Identity = \frac{\#\ positions - \#\ divergent\ positions}{\#\ positions}\ x\ 100\%$$

In the example below, there are two divergent positions when comparing the two sequences.

```
Query  1      CACTGCCCGATGGCTGACCGAGAGCAAGGTGCCATCATGGGCATCCAGGGTGAGATCCAG  60
Sbjct  18688  ..........-.............G................................  18746
```

In BLAST, different lengths of sequences can be compared. When the alignment is complete, BLAST will display the alignment such that each full line within a sequence comparison has 60 total positions, regardless of the number of nucleotides in each individual sequence. In the sequence comparison above, Query has 60 nucleotides and Subject has 59 nucleotides and one gap. In total there are 60 positions being compared. Two of those positions are divergent (one gap and one mismatch).

The percent identity for this comparison could be calculated as follows:

$$Percent\ Identity = \frac{60 - 2}{60}\ x\ 100\% = \frac{58}{60}\ x\ 100\% = 96.67\%$$

Knowledge Check

1. What is the difference between a mismatch and a gap?

2. What is the percent identity for the following comparison of 60 nucleotide positions?

```
Query  1     GGAGCCCTGGGCCGTGGAATTGATGGTATCTGTTTTCCAGCATGCAGAAGGGGGCTATGC  60
Sbjct  1794  ..........-..........................G......C.................  1852
```

3. You have two sequences you want to compare with BLAST. Your Query sequence is 100 nucleotides long, and your Subject sequence is 70 nucleotides long. When you compare the two sequences, you find that all 70 nucleotides from the Subject sequence exactly match the first 70 nucleotides from the Query sequence. Using 100 nucleotides as the total number of positions, what is the percent identity between these two sequences?

   Note: Students may question why we do not use 70 nucleotides as the total number of positions, since the first 70 nucleotides of the two sequences have 100% identity and the Subject sequence is only 70 nucleotides long. This is an important question! There are multiple ways to calculate percent identity, and that is one other valid way. However, for the purpose of these activities, we have defined percent identity to use the total number of nucleotide positions in the comparison.

4. The sequence comparison below shows two aligned amino acid sequences, with the amino acids abbreviated as single letters. What is the percent identity for this comparison of 60 amino acid positions?

```
Query  1     DPGSSKLQFAPFSSALNVGFWHELTQKKLNEYRLDETPKVIKGYYYNGDPSGFPARLTLE  60
Sbjct  7     ...L...........D..................A..D.........SA.L.......  66
```

Sequence Comparison with the TtGG Genes

Did you know that humans are, on average, 99.9% genetically identical? Only 0.1% of all our DNA bases are different, but those differences are what influence our traits and help make us each who we are.

The Teaching the Genome Generation™ (TtGG) curriculum was designed to explore human genetic variation, with a focus on variants in five different genes: *ACE*, *ACTN3*, *OXTR*, *CYP2C19*, and *TAS2R38*. Each of these genes are associated with an interesting phenotype and has a different type of molecular variant. In these activities, you can explore one (or more!) of the TtGG genes through the lens of sequence comparison.

Note: This activity does not introduce percent identity. Before beginning, be sure that either your students complete the Introduction to Sequence Comparison reading or you introduce sequence comparison and percent identity using the Introduction to Sequence Comparison Teacher Slides.

These activities can be used as an extension to the TtGG laboratories, but they have been designed such that they can also be used independently of the laboratory protocols. If you would like more background information on these five genes and their respective variants, either for yourself or for your students, check out our TtGG Gene Info Sheets.

If you or your students would like a refresher on the concept of a DNA variant prior to beginning this activity, check out our Minute to Understanding video "What are DNA variants?"

Sequence Comparison with *ACE*

The angiotensin I converting enzyme (*ACE*) gene codes for the protein angiotensin-converting enzyme (ACE), which functions as a protease that cuts other proteins. ACE plays a central role in the system that controls blood pressure by regulating the volume of fluids in the body. Variants in the human *ACE* gene are associated with differences in athletic endurance performance. In this activity, you'll compare ACE DNA sequences from different individuals and different organisms.

When comparing DNA sequences, **percent identity** provides a measure of how similar two sequences are. The formula for percent identity uses the total number of nucleotide **positions** in the sequence comparison and the number of nucleotide positions that are different, or **divergent**, between the sequences:

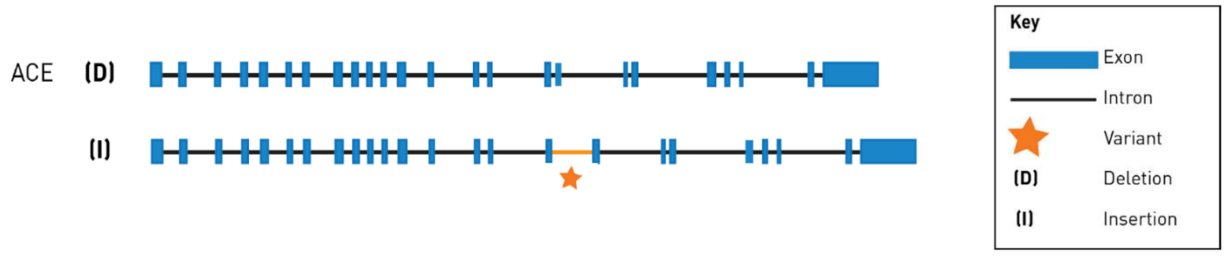
$$Percent\ Identity = \frac{\#\ positions - \#\ divergent\ positions}{\#\ positions}\ x\ 100\%$$

Part 1. Compare to a Reference Sequence

One common type of sequence comparison is comparing an individual's DNA sequence to a reference sequence. A **reference sequence** is a DNA 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 the DNA sequences of multiple individuals from the same species.

Comparing an individual person's DNA to a reference sequence allows us to identify variants, or differences, between that person's DNA sequence and the reference.

Below are box-line diagrams representing the two common alleles of the *ACE* gene. Box-line diagrams are a common visual representation of a gene structure where boxes indicate the parts of the gene that code for proteins (also called exons), and the black lines are the regions in between called introns.



The insertion (**I**) allele of the *ACE* gene has an insertion of 287 base pairs within intron 16. The deletion (**D**) allele of the of the *ACE* gene does not contain the 287 base pair insertion.

The entire human *ACE* gene sequence, including the 287 base pair insertion, is 21,597 nucleotides.

1. You want to compare an individual person's DNA sequence for the *ACE* gene to the human reference sequence for *ACE* to identify which *ACE* gene variant that person carries. We'll call this person **Jean**.

   Jean's *ACE* DNA sequence contains the insertion, while the reference sequence contains the deletion. What is the percent identity of this gene for this comparison? You can assume that all of the nucleotides are the same between the two sequences besides the 287 nucleotide insertion. Round your answer to two decimal places.

Sometimes researchers know which gene a DNA sequence is from, but other times they don't. Comparing sequences and calculating percent identity can help them figure out which gene their sequence is from. Since humans are 99.9% genetically identical to each other on average, a very high percent identity provides more confidence that the sequence is from a given gene.

2. If you didn't know which gene **Jean**'s DNA sequence was from, would you feel confident that this sequence was from *ACE* given the percent identity you calculated in question 1? Why or why not?

3. What if the percent identity was 95%? 75%? Justify your answer.

Note: Questions 2 and 3 could serve as a prompt for small group or whole group discussion.

Part 2. Compare Across Species

Imagine that you are a researcher studying athletic performance, and you want to learn more about the *ACE* gene. You plan to use a model organism to conduct some experiments to better understand the impact of the *ACE* gene on physical endurance. Model organisms, such as mice and fruit flies, are often used as a representation of human biology because they are easier to study in controlled environments and share much of the same physiology as humans.

You discover that mice also have an *ACE* gene. How similar is the mouse *ACE* gene to the human *ACE* gene?

Start by looking at a small section of the coding region of the *ACE* gene. The sequence comparison below compares a portion of the ACE gene for two sequences:

- The **Query** sequence is the **mouse** reference sequence for *ACE*.
- The **Subject** ("Sbjct") sequence is the **human** reference for *ACE*.

When comparing across species, it can be helpful to record which sequence corresponds to which organism. In the figure below, write in which sequence is **mouse** and which is **human**.

```
_______ Query  1     TTGTATGAGTCCATTTGGCAGAACTTTACTGACTCAAAGCTGCGAAGGATCATCGGATCT  60
_______ Sbjct        C.......AC.G..C...........C..G...C.GC.......C...........G..
```

1. How many nucleotides are different between the two sequences? What types of differences are they?

2. In this comparison, there are 60 positions total. What is the percent identity of this section of the *ACE* gene?

Now, look at a different section of the *ACE* gene and see how it compares between mouse (**Query**) and human (**Sbjct**).

```
_______ Query  1     GGCTCTACAACATCCGTAACCATCACAGCCTCCGCCGGCCCCACCGTGGGCCCCAGTTTG  60
_______ Sbjct        ......T..G......---...C.G........A.....A.TC..AC...........C.
```

3. How many nucleotides are different between the two sequences? What types of differences are they?

4. What is the percent identity of this section of the *ACE* gene?

The percent identity for the comparison of the whole protein coding sequences of human *ACE* and mouse *ACE* is 83%. The percent identity of mouse and human genes is 85% on average, but it varies from 60% to 99% for individual genes.

5. Given an 83% identity for *ACE* and knowing that mice are often used as a model organism for human biology, would you feel confident using mice as a model to study the impact of *ACE* on endurance performance?

6. What other information might you want to know about the mouse *ACE* gene or protein to help you make your decision?

Comparing DNA sequences across species can also provide information about how species evolved over time. In general, the greater the percent identity of DNA sequences between two species, the more recently they have shared a common ancestor. It takes time for genetic differences to accumulate, so organisms with fewer genetic differences are typically more closely related.

7. Which species would you expect to have a more similar *ACE* DNA sequence to humans: mice or chimpanzees? Justify your answer.

Check your hypothesis by comparing a portion of the chimpanzee *ACE* reference sequence (**Query**) and the human *ACE* reference sequence (**Sbjct**).

```
Query  1    CTGTATGAACCGGTCTGGCAGAACTTCACGGACCCGCAGCTGCGCAGGATCATCGGAGCT  60

Sbjct       ............A............................................
```

8. How many nucleotides are different between the two sequences? What types of differences are they?

9. What is the percent identity of this section of the *ACE* gene?

10. Compare your percent identity calculations for questions 2 and 4 (mouse) with your calculation for question 9 (chimpanzee). Do they support your hypothesis about whether the mouse or chimpanzee *ACE* DNA sequence is more similar to the human sequence? Explain your reasoning.

11. If not, why do you think that might be?

Note: Questions 10 and 11 could serve as a prompt for small group or whole group discussion.

Part 3. Compare Within Species

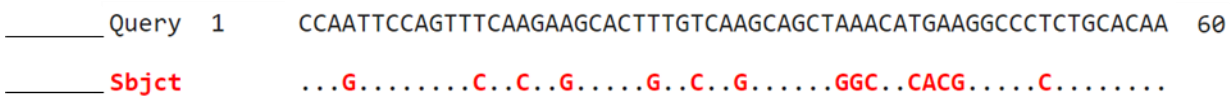
Another type of sequence comparison is comparing the DNA sequences of different genes within the same species. Comparing different genes within the same species can help scientists identify gene families.

Gene families are groups of genes with similar functions. Comparing sequences helps identify gene families because a gene's sequence determines its associated protein's structure, which determines protein function. This is especially useful in organisms where a full genome sequence is not known. By comparing new gene sequences to known genes, scientists can determine if the new gene serves a similar function to a known gene.

Comparing genes within species can also provide us information about evolution. Sometimes, as species evolve, genes get duplicated. Over time, these gene duplicates accumulate changes and become different enough that they serve different, yet related, functions.

Let's look at an example. *ACE* is in the same gene family as Angiotensin Converting Enzyme 2 (*ACE2*). How similar are these two sequences?

Start by comparing a small section of the human *ACE2* reference sequence (**Query**) to the human *ACE* reference sequence (**Sbjct**). When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from ***ACE2*** and which is from ***ACE***.



1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity for this comparison?

3. Given this percent identity and what you know about the function of ACE, what might you predict the function of ACE2 in the body to be?

   Note: This question could serve as a prompt for small group or whole group discussion. If you would like more information about the function of ACE2, check out the following resources:
   - *ACE2 UniProt Entry*
   - *ACE2 GeneCards Entry*
   - *ACE2 OMIM Entry*

Sequence Comparison with *ACTN3*

The alpha actinin 3 (*ACTN3*) gene codes for the alpha actinin 3 (ACTN3) protein, which is a key structural protein in muscle. Variants in the human *ACTN3* gene are associated with differences in athletic performance. In this activity, you'll compare *ACTN3* DNA sequences from different individuals and different organisms.

When DNA comparing sequences, **percent identity** provides a measure of how similar two sequences are. The formula for percent identity uses the total number of nucleotide **positions** in the sequence comparison and the number of nucleotide positions that are different, or **divergent**, between the sequences:

$$Percent\ Identity = \frac{\#\ positions - \#\ divergent\ positions}{\#\ positions}\ x\ 100\%$$

Part 1. Compare to a Reference Sequence

One common type of sequence comparison is comparing an individual's DNA sequence to a reference sequence. A **reference sequence** is a DNA 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 the DNA sequences of multiple individuals from the same species.

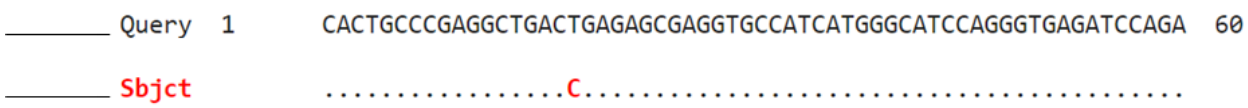
Comparing an individual person's DNA to a reference sequence allows us to identify variants, or differences, between that person's DNA sequence and the reference.

In this example, we will compare an individual person's DNA sequencing data for the *ACTN3* gene to the human reference sequence for *ACTN3* to identify which *ACTN3* gene variant that person carries. We'll call this person **Jean**.

The sequence comparison below looks at a portion of the *ACTN3* gene for two sequences:

- The **Query** sequence is **Jean**'s *ACTN3* sequence data.
- The **Subject** ("Sbjct") sequence is the human **reference** sequence for *ACTN3*.

When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from **Jean** and which is the **reference** sequence.



1. How many nucleotides are different between the two sequences within this portion of the ACTN3 gene? What is/are the variant nucleotide(s) in the individual's DNA sequence and the reference DNA sequence?

2. In this comparison, there are 60 positions total. What is the percent identity for this comparison?

3. The entire human *ACTN3* gene sequence is 16,489 nucleotides. What is the percent identity for the entire *ACTN3* sequence? You can assume that all the other nucleotides are the same between the two sequences. Round your answer to two decimal places.

Sometimes researchers know which gene a DNA sequence is from, but other times they don't. Comparing sequences and calculating percent identity can help them figure out which gene their sequence is from. Since humans are 99.9% genetically identical to each other on average, a very high percent identity provides more confidence that the sequence is from a given gene.

4. If you didn't know which gene **Jean**'s DNA sequence was from, would you feel confident that this sequence was from *ACTN3* given the percent identity you calculated in question 3? Why or why not?

5. What if the percent identity was 99%? 95%? 75%? Justify your answer.

Note: Questions 4 and 5 could serve as a prompt for small group or whole group discussion.

Part 2. Compare Across Species

Imagine that you are a researcher studying athletic performance, and you want to learn more about the *ACTN3* gene. You plan to use a model organism to conduct some experiments to better understand the impact of the *ACTN3* gene on sprinting performance. Model organisms, such as mice and fruit flies, are often used as a representation of human biology because they are easier to study in controlled environments and share much of the same physiology as humans.

You discover that mice also have an *ACTN3* gene. How similar is the mouse *ACTN3* gene to human *ACTN3*?

Start by looking at a small section of the *ACTN3* gene, comparing the mouse reference sequence (**Query**) to the human reference sequence (**Sbjct**). When comparing across species, it can be helpful to record which sequence corresponds to which organism. In the figure below, write in which sequence is **mouse** and which is **human**.

```
_______ Query  1      CGTTGCCCGAGGCTGATCGAGAGCGAGGCGCCATCCTGGGCATCCAAGGAGAGATCCAGA  60

_______ Sbjct         .AC............C...........T......A..........G..T..........
```

1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity of this section of the *ACTN3* gene?

Now, look at a different section of the *ACTN3* gene and see how it compares between mouse (**Query**) and human (**Sbjct**).

```
_______ Query  1      TGATTCTTCCTGT---TC--CCCCAGAGCCTGCTAACAGCACACGAACAGTTCAAGGCAA  60

_______ Sbjct         ...CA.......CCTG..GT..............G.....G.....T.............
```

3. How many nucleotides are different between the two sequences? What types of differences are they?

4. What is the percent identity of this section of the *ACTN3* gene?

The percent identity for the comparison of the whole protein coding sequences of human *ACTN3* and mouse *ACTN3* is 89%. The percent identity of mouse and human genes is 85% on average, but it varies from 60% to 99% for individual genes.

5. Given an 89% identity for *ACTN3* and knowing that mice are often used as a model organism to study human biology, would you feel confident using mice as a model to study the impact of *ACTN3* on sprinting performance?

6. What other information might you want to know about the mouse *ACTN3* gene or protein to help you make your decision?

Comparing DNA sequences across species can also provide information about how species evolved over time. In general, the greater the percent identity of DNA sequences between two species, the more recently they have shared a common ancestor. It takes time for genetic differences to accumulate, so organisms with fewer genetic differences are typically more closely related.

7. Which species would you expect to have a more similar *ACTN3* DNA sequence to humans: mice or chimpanzees? Justify your answer.

Check your hypothesis by comparing a portion of the chimpanzee *ACTN3* reference sequence (**Query**) and the human *ACTN3* reference sequence (**Sbjct**).

```
Query  1    CATTGCCCGAGGCTGACCGAGAGCGAGGTGCCATCATGGGCATCCAGGGTGAGATCCAGA  60

Sbjct       ..C.........................................................
```

8. How many nucleotides are different between the two sequences? What types of differences are they?

9. What is the percent identity of this section of the *ACTN3* gene?

10. Compare your percent identity calculations for questions 2 and 4 (mouse) with your calculation for question 9 (chimpanzee). Do they support your hypothesis about whether the mouse or chimpanzee *ACTN3* DNA sequence is more similar to the human sequence? Explain your reasoning.

11. If not, why do you think that might be?

Note: Questions 10 and 11 could serve as a prompt for small group or whole group discussion.

Part 3. Compare Within Species

Another type of sequence comparison is comparing the DNA sequences of different genes within the same species. Comparing different genes within the same species can help scientists identify gene families.

Gene families are groups of genes with similar functions. Comparing sequences helps identify gene families because a gene's sequence determines its associated protein's structure, which determines protein function. This is especially useful in organisms where a full genome sequence is not known. By comparing new gene sequences to known genes, scientists can determine if the new gene serves a similar function to a known gene.

Comparing genes within species can also provide us information about evolution. Sometimes, as species evolve, genes get duplicated. Over time, these gene duplicates accumulate changes and become different enough that they serve different, yet related, functions.

Let's look at an example. *ACTN3* is in the same gene family as alpha actinin 2 (*ACTN2*). How similar are these two sequences?

Start by comparing a small section of the human *ACTN2* reference sequence (**Query**) to the human *ACTN3* reference sequence (**Sbjct**). When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from ***ACTN2*** and which is from ***ACTN3***.

```
_______ Query  1     CACTGCCCGAGGCTGACCGAGAGCGAGGTGCCATCATGGGCATCCAGGGTGAGATCCAGA 60

_______ Sbjct        .G...........G...G.......GCAGT.........C.......AAC...G.GG...
```

1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity for this comparison?

Overall, human *ACTN2* has a 74% identity with human *ACTN3*.

3. Given this percent identity and what you know about the function of ACTN3, what might you predict the function of ACTN2 in the body to be?

   Note: This question could serve as a prompt for small group or whole group discussion. If you would like more information about the function of ACTN2, check out the following resources:
   - *ACTN2 UniProt Entry*
   - *ACTN2 GeneCards Entry*
   - *ACTN2 OMIM Entry*

Sequence Comparison with *OXTR*

The oxytocin receptor (*OXTR*) gene codes for the oxytocin receptor (OXTR) protein, which functions as a receptor for the hormone and neurotransmitter oxytocin. Variants in the human *OXTR* gene are associated with differences in social ability, emotional sensitivity, and response to emotional stress. In this activity, you'll compare *OXTR* DNA sequences from different individuals and different organisms.

When comparing DNA sequences, **percent identity** provides a measure of how similar two sequences are. The formula for percent identity uses the total number of nucleotide **positions** in the sequence comparison and the number of nucleotide positions that are different, or **divergent**, between the sequences:

$$Percent\ Identity = \frac{\#\ positions - \#\ divergent\ positions}{\#\ positions}\ x\ 100\%$$

Part 1. Compare to a Reference Sequence

One common type of sequence comparison is comparing an individual's DNA sequence to a reference sequence. A **reference sequence** is a DNA 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 the DNA sequences of multiple individuals from the same species.

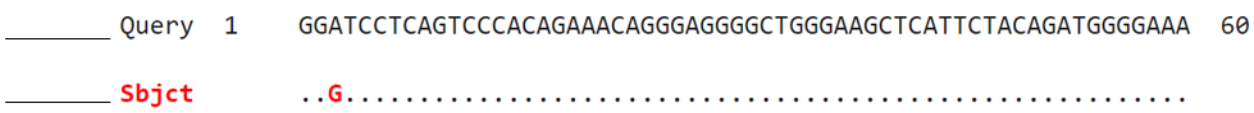
Comparing an individual's DNA to a reference sequence allows us to identify variants, or differences, between the individual's DNA sequence and the reference.

In this example, we will compare an individual person's DNA sequencing data for the *OXTR* gene to the human reference sequence for *OXTR* to identify which *OXTR gene* variant that person carries. We'll call this person **Jean**.

The sequence comparison below looks at a portion of the *OXTR* gene for two sequences:

- The **Query** sequence is **Jean**'s *OXTR* sequence data.
- The **Subject** ("Sbjct") sequence is the human **reference** sequence for *OXTR.*

When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from **Jean** and which is the **reference** sequence.



1. How many nucleotides are different between the two sequences within this portion of the *OXTR* gene? What is/are the variant nucleotide(s) in the individual's DNA sequence and the reference DNA sequence?

2. In this comparison, there are 60 positions total. What is the percent identity for this comparison?

3. The entire human *OXTR* gene sequence is 19,206 nucleotides. What is the percent identity for the entire *OXTR* sequence? You can assume that all the other nucleotides are the same between the two sequences. Round your answer to two decimal places.

Sometimes researchers know which gene a DNA sequence is from, but other times they don't. Comparing sequences and calculating percent identity can help them figure out which gene their sequence is from. Since humans are 99.9% genetically identical to each other on average, a very high percent identity provides more confidence that the sequence is from a given gene.

4. If you didn't know which gene **Jean**'s DNA sequence was from, would you feel confident that this sequence was from *OXTR* given the percent identity you calculated in question 3? Why or why not?

5. What if the percent identity was 99%? 95%? 75%? Justify your answer.

Note: Questions 4 and 5 could serve as a prompt for small group or whole group discussion.

Part 2. Compare Across Species

Imagine that you are a researcher investigating whether genetics can impact social behavior, and you want to learn more about the *OXTR* gene. You plan to use a model organism to conduct some experiments to better understand the impact of the *OXTR* gene on social behaviors. Model organisms, such as mice and fruit flies, are often used as a representation of human biology and behavior because they are easier to study in controlled environments and share much of the same physiology as humans.

You discover that mice also have an *OXTR* gene. How similar is the mouse *OXTR* gene to human *OXTR*?

Start by looking at a small section of the coding region of the *OXTR* gene, comparing the mouse reference sequence (**Query**) to the human reference sequence (**Sbjct**). When comparing across species, it can be helpful to record which sequence corresponds to which organism. In the figure below, write in which sequence is **mouse** and which is **human**.

Note: The OXTR sequence comparison in Part 1 looked at a variant in an intron. The mouse and human OXTR intron sequences are too different to do a useful sequence comparison, so this section instead uses a sequence from an exon of the OXTR gene. If you have covered introns/exons and conserved sequences in evolution, you could use this as a discussion point with your students.

```
_____ Query  1     GTCAGTAGTGTCAAGCTTATCTCCAAGGCCAAAATCCGCACAGTGAAGATGACCTTCATC  60

_____ Sbjct        .....C..C........C..............G........G..C........T......
```

1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity of this section of the *OXTR* gene?

Now, look at a different section of the *OXTR* gene and see how it between across mouse (**Query**) and human (**Sbjct**).

```
_____ Query  1     TGCTATGGTCTCATCAGCTTCAAGATCTGGCAGAATCTGCGACTCAAGAC-G--GCAGCC  60

_____ Sbjct        .....C..C..T.......................CT....G........C.CT.....G
```

3. How many nucleotides are different between the two sequences? What types of differences are they?

4. What is the percent identity of this section of the *OXTR* gene?

The percent identity for the comparison of the whole protein coding sequences of human *OXTR* and mouse *OXTR* is 92%. The percent identity of mouse and human genes is 85% on average, but it varies from 60% to 99% for individual genes.

Note: If you have discussed the difference between introns/exons with your students above, you may want to point out to students that all of the percent identity values above refer only to comparisons between the coding sequences. This point could be used to frame a class discussion around why protein coding regions have a higher percent identity than non-coding regions.

5. Given an 92% identity for *OXTR* and knowing that mice are often used as a model organism to study human biology, would you feel confident using mice as a model to study *OXTR*? Justify your answer.

6. What other information might you want to know about the mouse *OXTR* gene or protein to help you make your decision?

Comparing DNA sequences across species can also provide information about how species evolved over time. In general, the greater the percent identity of DNA sequences between two species, the more recently they have shared a common ancestor. It takes time for genetic differences to accumulate, so organisms with fewer genetic differences are typically more closely related.

7. Which species would you expect to have a more similar *OXTR* DNA sequence to humans: mice or chimpanzees? Justify your answer.

Check your hypothesis by comparing a portion of the chimpanzee *OXTR* reference sequence (**Query**) and the human *OXTR* reference sequence (**Sbjct**).

```
Query  1     GTCAGCAGCGTCAAGCTCATCTCCAAGGCCAAGATCCGCACGGTCAAGATGACTTTCATC  60

Sbjct        ............................................................
```

8. How many nucleotides are different between the two sequences? What types of differences are they?

9. What is the percent identity of this section of the *OXTR* gene?

10. Compare your percent identity calculations for questions 2 and 4 (mouse) with your calculation for question 9 (chimpanzee). Do they support your hypothesis about whether the mouse or chimpanzee *OXTR* DNA sequence is more similar to the human sequence? Explain your reasoning.

11. If not, why do you think that might be?

Note: Questions 10 and 11 could serve as a prompt for small group or whole group discussion.

Part 3. Compare Within Species

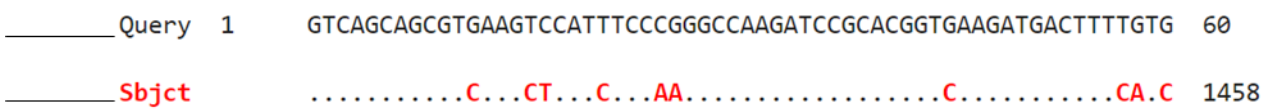
Another type of sequence comparison is comparing the DNA sequences of different genes within the same species. Comparing different genes within the same species can help scientists identify gene families.

Gene families are groups of genes with similar functions. Comparing sequences helps identify gene families because a gene's sequence determines its associated protein's structure, which determines protein function. This is especially useful in organisms where a full genome sequence is not known. By comparing new gene sequences to known genes, scientists can to determine if the new gene serves a similar function to a known gene.

Comparing genes within species can also provide us information about evolution. Sometimes, as species evolve, genes get duplicated. Over time, these gene duplicates accumulate changes and become different enough that they serve different, yet related, functions.

Let's look at an example. *OXTR* is in the same gene family as arginine vasopressin receptor 1A (*AVPR1A*). How similar are these two sequences?

Start by comparing a small section of the human *AVPR1A* reference sequence (**Query**) to the human *OXTR* reference sequence (**Sbjct**). When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from ***AVPR1A*** and which is from ***OXTR***.



1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity for this comparison?

Overall, human *AVPR1A* has a 57% identity with human *OXTR*.

3. Given this percent identity and what you know about the function of OXTR, what might you predict the function of AVPR1A in the body to be?

   Note: This question could serve as a prompt for small group or whole group discussion. If you would like more information about the function of AVPR1A, check out the following resources:
   - *AVPR1A UniProt Entry*
   - *AVPR1A GeneCards Entry*
   - *AVPR1A OMIM Entry*

Sequence Comparison with *CYP2C19*

The Cytochrome P450, family 2, subfamily C, polypeptide 19 (*CYP2C19*) gene codes for the Cytochrome P450, family 2, subfamily C, polypeptide 19 (CYP2C19) protein, which is an enzyme that catalyzes many reactions involved in drug metabolism, synthesis of cholesterol, steroids, and other lipids. Variants in the human *CYP2C19* gene are associated with differences in drug metabolism. In this activity, you'll compare *CYP2C19* DNA sequences from different individuals and different organisms.

When comparing DNA sequences, **percent identity** provides a measure of how similar two sequences are. The formula for percent identity uses the total number of nucleotide **positions** in the sequence comparison and the number of nucleotide positions that are different, or **divergent**, between the sequences:

$$Percent\ Identity = \frac{\#\ positions - \#\ divergent\ positions}{\#\ positions}\ x\ 100\%$$

Part 1. Compare to a Reference Sequence

One common type of sequence comparison is comparing an individual's DNA sequence to a reference sequence. A **reference sequence** is a DNA 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 the DNA sequences of multiple individuals from the same species.

Comparing an individual's DNA to a reference sequence allows us to identify variants, or differences, between the individual's DNA sequence and the reference.

In this example, we will compare an individual person's DNA sequencing data for the *CYP2C19* gene to the human reference sequence for *CYP2C19* to identify which *CYP2C19* gene variant that person carries. We'll call this person **Jean**.

The sequence comparison below looks at a portion of the *CYP2C19* gene for two sequences:

- The **Query** sequence is **Jean**'s *CYP2C19* sequence data.
- The **Subject** ("Sbjct") sequence is the human **reference** sequence for *CYP2C19.*

When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from **Jean** and which is the **reference** sequence.

```
_______ Query  1    ATTTTCCCACTATCATTGATTATTTCCCAGGAACCCATAACAAATTACTTAAAAACCTTG  60

_______ Sbjct       ............................G.............................
```

1. How many nucleotides are different between the two sequences within this portion of the *CYP2C19* gene? What is/are the variant nucleotide(s) in the individual's DNA sequence and the reference DNA sequence?

2. In this comparison, there are 60 positions total. What is the percent identity for this comparison?

3. The entire human *CYP2C19* gene sequence is 90,209 nucleotides. What is the percent identity for the entire *CYP2C19* sequence? You can assume that all the other nucleotides are the same between the two sequences. Round your answer to two decimal places.

Sometimes researchers know which gene a DNA sequence is from, but other times they don't. Comparing sequences and calculating percent identity can help them figure out which gene their sequence is from. Since humans are 99.9% genetically identical to each other on average, a very high percent identity provides more confidence that the sequence is from a given gene.

4. If you didn't know which gene **Jean**'s DNA sequence was from, would you feel confident that this sequence was from *CYP2C19* given the percent identity you calculated in question 3? Why or why not?

5. What if the percent identity was 99%? 95%? 75%? Justify your answer.

Note: Questions 4 and 5 could serve as a prompt for small group or whole group discussion.

Part 2. Compare Across Species

Imagine that you are a researcher studying drug metabolism, and you want to learn more about the *CYP2C19* gene. You plan to use a model organism to conduct some experiments to better understand the impact of the *CYP2C19* gene on drug metabolism. Model organisms, such as mice and fruit flies, are often used as a representation of human biology because they are easier to study in controlled environments and share much of the same physiology as humans.

When you try to find the sequence for *CYP2C19* in mice, you are surprised to find that mice do not have a *CYP2C19* gene! However, mice do have several genes that share sequence similarity with human *CYP2C19*, like cytochrome P450, family 2, subfamily C, polypeptide 65 (*CYP2C65*). These two genes are part of the same gene family, or group of genes with similar functions.

You decide to compare a small section of the mouse reference sequence for *CYP2C65* (**Query**) to the human reference sequence for *CYP2C19* (**Sbjct**). When comparing across species, it can be helpful to record which sequence corresponds to which organism. In the figure below, write in which sequence is **mouse** and which is **human**.

```
_____ Query  1     ATTTTCCTGCTGTCATTGATTATCTACCAGGAAGACACAGAAAATTACATAAAAATTTTG  60

_____ Sbjct        .......CA..A...........T.C..G....CC..T.AC.......T......CC...
```

1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity for this comparison?

Now, look at a different section of mouse *CYP2C65* (**Query**) and human *CYP2C19* (**Sbjct**).

```
_____ Query  1     ACAATCCTCGGGACTTTATTGATTGTTTCCTGATCAAAATGGAACAGGAAAAGCACAACC  60

_____ Sbjct        ....C.....................C.................GA..........A....
```

3. How many nucleotides are different between the two sequences? What types of differences are they?

4. What is the percent identity for this comparison?

While the start of the sequences for human *CYP2C19* and mouse *CYP2C65* are similar, the mouse sequence (1941 nucleotides) is much shorter than the human sequence (4131 nucleotides). The percent identity for the comparison of the whole protein coding sequences of human *CYP2C19* and mouse *CYP2C65* is about 30%. For genes shared between mice and humans, percent identity is 85% on average.

5. Given a 30% identity for human *CYP2C19* and mouse *CYP2C65*, would you feel confident using mice as a model for your study on drug metabolism? Justify your answer.

6. What is one other system or method you might be able to use to study human *CYP2C19* and drug metabolism?

Comparing DNA sequences across species can also provide information about how species evolved over time. In general, the greater the percent identity of DNA sequences between two species, the more recently they have shared a common ancestor. It takes time for genetic differences to accumulate, so organisms with fewer genetic differences are typically more closely related.

7. Which species would you expect to have a more similar DNA sequence to human *CYP2C19*: mice or chimpanzees? Justify your answer.

Unlike mice, chimpanzees have a *CYP2C19* gene. Check your hypothesis by comparing a portion of the chimpanzee *CYP2C19* reference sequence (**Query**) and the human *CYP2C19* reference sequence (**Sbjct**).

```
Query  1     ATTTTCCCACTATCATTGATTATTTCCCGGGAACCCATAACAAATTACTTAAAAACCTTG  60

Sbjct        ............................................................
```

8. How many nucleotides are different between the two sequences? What types of differences are they?

9. What is the percent identity of this section of the *CYP2C19* gene?

10. Compare your percent identity calculations for questions 2 and 4 (mouse) with your calculation for question 9 (chimpanzee). Do they support your hypothesis about whether the mouse *CYP2C65* or chimpanzee *CYP2C19* DNA sequence is more similar to the human *CYP2C19* sequence? Explain your reasoning.

11. If not, why do you think that might be?

Note: Questions 10 and 11 could serve as a prompt for small group or whole group discussion.

Part 3. Compare Within Species

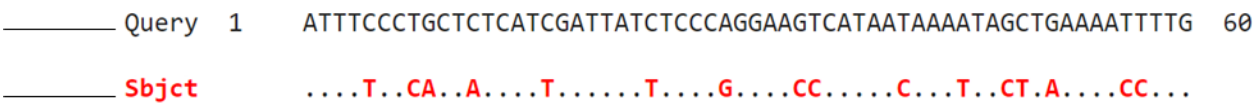
Another type of sequence comparison is comparing the DNA sequences of different genes within the same species. Comparing different genes within the same species can help scientists identify gene families.

Gene families are groups of genes with similar functions. Comparing sequences helps identify gene families because a gene's sequence determines its associated protein's structure, which determines protein function. This is especially useful in organisms where a full genome sequence is not known. By comparing new gene sequences to known genes, scientists can determine if the new gene serves a similar function to a known gene.

Comparing genes within species can also provide us information about evolution. Sometimes, as species evolve, genes get duplicated. Over time, these gene duplicates accumulate changes and become different enough that they serve different, yet related, functions.

Let's look at an example. *CYP2C19* is in the same gene family as cytochrome P450 family 2 subfamily C member 8 (*CYP2C18*). How similar are these two sequences?

Start by comparing a small section of the human *CYP2C18* reference sequence (**Query**) to the human *CYP2C19* reference sequence (**Sbjct**). When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from ***CYP2C18*** and which is from ***CYP2C19***.



1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity for this comparison?

Overall, human *CYP2C18* has an 84% identity with human *CYP2C19*.

3. Given this percent identity and what you know about the function of CYP2C19, what might you predict the function of CYP2C18 in the body to be?

   Note: This question could serve as a prompt for small group or whole group discussion. If you would like more information about the function of CYP2C18, check out the following resources:
   - *CYP2C18 UniProt Entry*
   - *CYP2C18 GeneCards Entry*
   - *CYP2C18 OMIM Entry*

Sequence Comparison with *TAS2R38*

The Taste 2 Receptor Member 38 (*TAS2R38*) gene produces the Taste 2 Receptor Member 38 (TAS2R38) protein, which functions as a receptor to perceive a wide range of bitter compounds. Variants in the human *TAS2R38* gene are associated with differences in ability to taste specific bitter compounds. In this activity, you'll compare *TAS2R38* DNA sequences from different individuals and different organisms.

When comparing DNA sequences, **percent identity** provides a measure of how similar two sequences are. The formula for percent identity uses the total number of nucleotide **positions** in the sequence comparison and the number of nucleotide positions that are different, or **divergent**, between the sequences:

$$Percent\ Identity = \frac{\#\ positions - \#\ divergent\ positions}{\#\ positions}\ x\ 100\%$$

Part 1. Compare to a Reference Sequence

One common type of sequence comparison is comparing an individual's DNA sequence to a reference sequence. A **reference sequence** is a DNA 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 the DNA sequences of multiple individuals from the same species.

Comparing an individual's DNA to a reference sequence allows us to identify variants, or differences, between the individual's DNA sequence and the reference.

In this example, we will compare an individual person's DNA sequencing data for the *TAS2R38* gene to the human reference sequence for *TAS2R38* to identify which *TAS2R38* gene variant that person carries. We'll call this person **Jean**.

The sequence comparison below looks at a portion of the *TAS2R38* gene for two sequences:

- The **Query** sequence is **Jean**'s *TAS2R38* sequence data.
- The **Subject** ("Sbjct") sequence is the human **reference** sequence for *TAS2R38.*

When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from **Jean** and which is the **reference** sequence.

```
_____Query  1    GTGAATTTTTGGGATGTAGTGAAGAGGCAGCCACTGAGCAACAGTGATTGTGTGCTGCTG  60

_____Sbjct       ............................G............................
```

1. How many nucleotides are different between the two sequences within this portion of the *TAS2R38* gene? What is/are the variant nucleotide(s) in the individual's DNA sequence and the reference DNA sequence?

2. In this comparison, there are 60 positions total. What is the percent identity for this comparison?

3. The entire human *TAS2R38* gene sequence is 1,002 nucleotides. What is the percent identity for the entire *TAS2R38* sequence? You can assume that all the other nucleotides are the same between the two sequences. Round your answer to two decimal places.

Sometimes researchers know which gene a DNA sequence is from, but other times they don't. Comparing sequences and calculating percent identity can help them figure out which gene their sequence is from. Since humans are 99.9% genetically identical to each other on average, a very high percent identity provides more confidence that the sequence is from a given gene.

4. If you didn't know which gene **Jean**'s DNA sequence was from, would you feel confident that this sequence was from *TAS2R38* given the percent identity you calculated in question 3? Why or why not?

5. What if the percent identity was 99%? 95%? 75%? Justify your answer.

Note: Questions 4 and 5 could serve as a prompt for small group or whole group discussion.

Part 2. Compare Across Species

Imagine that you are a researcher studying taste perception, and you want to learn more about the *TAS2R38* gene. You plan to use a model organism to conduct some experiments to better understand the impact of the *CYP2C19* gene on the ability to taste bitter compounds. Model organisms, such as mice and fruit flies, are often used as a representation of human biology and behavior because they are easier to study in controlled environments and share much of the same physiology as humans.

Although you're interested in human taste perception, you plan to use a model organism. Model organisms, such as mice, are often used as a representation of human biology and behavior.

When you try to find the sequence for *TAS2R38* in mice, you are surprised to find that mice do not have a *TAS2R38* gene! Mice instead have a very similar gene called Taste 2 Receptor Member 138 (*TAS2R138)*. These two genes are part of the same gene family, or group of genes with similar functions.

You decide to compare a small section of the mouse reference sequence for *TAS2R138* (**Query**) to the human reference sequence for *TAS2R38* (**Sbjct**). When comparing across species, it can be helpful to record which sequence corresponds to which organism. In the figure below, write in which sequence is **mouse** and which is **human**.

```
_______ Query  1    GTAAATGTTTGGGATGTGGTAAAAAAGCAGCCCTTGAACAACTGTGACATCGCACTGCTG  60

_______ Sbjct       ..G...T..........A..G..G.G....G.AC...G....A....TTGT.TG......
```

1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity for this comparison?

Now, look at a different section of mouse *TAS2R138* (**Query**) and human *TAS2R38* (**Sbjct**).

```
_______Query  1    -AGATTTCTTCAGATGCTTCTAGTTGTTCTTCTTCTCTCCTGCATCTGCACTGCCCTTTG  60

_______Sbjct       A....C..-C........C..G.G.A..A.....TG.................T...C..
```

3. How many nucleotides are different between the two sequences? What types of differences are they?

4. What is the percent identity for this comparison?

The percent identity for the comparison of the whole protein coding sequences of human *TAS2R38* and mouse *TAS2R138* is 76%. The percent identity of mouse and human genes is 85% on average, but it varies from 60% to 99% for individual genes.

5. Given a 76% identity for human *TAS2R38* and mouse *TAS2R138*, would you feel confident using mice as a model for your research on taste receptors? Justify your answer.

6. What is one other system or method you might be able to use to study human *TAS2R38* and taste perception?

Comparing DNA sequences across species can also provide information about how species evolved over time. In general, the greater the percent identity of DNA sequences between two species, the more recently they have shared a common ancestor. It takes time for genetic differences to accumulate, so organisms with fewer genetic differences are typically more closely related.

7. Which species would you expect to have a more similar DNA sequence to human *TAS2R38*: mice or chimpanzees? Justify your answer.

Check your hypothesis by comparing a portion of the chimpanzee *TAS2R38* reference sequence (**Query**) and the human *TAS2R38* reference sequence (**Sbjct**).

```
Query  1    GTGAATTTTTGGGATGTAGTGAAGAGGCAGCCACTGAGCAACAGTGATTGTGTGCTGCTG  60

Sbjct       ..............................G.............................
```

8. How many nucleotides are different between the two sequences? What types of differences are they?

9. What is the percent identity of this section of the *TAS2R38* gene?

10. Compare your percent identity calculations for questions 2 and 4 (mouse) with your calculation for question 9 (chimpanzee). Do they support your hypothesis about whether the mouse *TAS2R138* or chimpanzee *TAS2R38* DNA sequence is more similar to the human *TAS2R38* sequence? Explain your reasoning.

11. If not, why do you think that might be?

Note: Questions 10 and 11 could serve as a prompt for small group or whole group discussion.

Part 3. Compare Within Species

Another type of sequence comparison is comparing the DNA sequences of different genes within the same species. Comparing different genes within the same species can help scientists identify gene families.

Gene families are groups of genes with similar functions. Comparing sequences helps identify gene families because a gene's sequence determines its associated protein's structure, which determines protein function. This is especially useful in organisms where a full genome sequence is not known. By comparing new gene sequences to known genes, scientists can determine if the new gene serves a similar function to a known gene.

Comparing genes within species can also provide us information about evolution. Sometimes, as species evolve, genes get duplicated. Over time, these gene duplicates accumulate changes and become different enough that they serve different, yet related, functions.

Let's look at an example. *TAS2R38* is in the same gene family as taste 2 receptor member 3 (*TAS2R3*). How similar are these two sequences?

Start by comparing a small section of the human *TAS2R3* reference sequence (**Query**) to the human *TAS2R38* reference sequence (**Sbjct**). When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from ***TAS2R3*** and which is from ***TAS2R38***.

```
_______ Query  1    TGGCTTGCCACCTGTCTTGGTGTCCTCTACTGCCTGAAAATCGCCAGTTTCTCTCACCCC  60

_______ Sbjct       ........TG....C..CA.CC.G..T......TCC..GC..AT.C...........A..
```

1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity for this comparison?

Overall, human *TAS2R3* has an 41% identity with human *TAS2R38*.

3. Given this percent identity and what you know about the function of TAS2R38, what might you predict the function of TAS2R3 in the body to be?

   Note: This question could serve as a prompt for small group or whole group discussion. If you would like more information about the function of TAS2R3, check out the following resources:
   - *TAS2R3 UniProt Entry*
   - *TAS2R3 GeneCards Entry*
   - *TAS2R3 OMIM Entry*

Sequencing for Rare Disease Diagnosis

Sequencing for Rare Disease Diagnosis: Scenario Introduction

Scenario

You are a parent of two daughters, both of whom have a rare, undiagnosed disorder. Your daughters have similar symptoms, including learning difficulties and ataxia, an inability to control body movements.

You want to find a medical diagnosis for your daughters so that you can connect with other patients and their families and learn more about how to best support your daughters. However, you've been to many doctors and none of them have been able to identify a known disorder that matches all of your daughters' symptoms.

Finally, a doctor connects your family to a group of researchers who study human disease. After hearing about your family's efforts to find a diagnosis and care for your daughters, the researchers suggest that they might be able to learn more by sequencing your daughters' DNA. After sequencing, they'll compare your daughters' DNA sequences to a reference sequence to look for any differences, or variants.

Reflection

1. How are these patients and their parents' experiences similar to or different from your own experience at the doctor?

2. What questions would you have for the researchers as the parent? As one of the daughters?

Note: These questions could be implemented as a think-pair-share or small group discussion.

If your students are new to these topics, you may want to provide students with additional background on genome sequencing and/or rare diseases. The following videos could serve as a starting point:

- *Would you have your genome sequenced? | Dr Saskia Sanderson | TEDxGoodenoughCollege*
  - *Watch until 3:31*
- *Impact of Genetic Sequencing for Rare Disease Diagnosis | Illumina*

Guiding Questions

During these activities, you will be learning more about what the researchers did in their study and what they discovered. As you go, you'll be learning more about the questions below. Before you begin, write out your thoughts in response to these questions. Don't worry if you don't feel confident in your answers! You can revisit these after each activity so that you can see what you've learned.

Given what you know about genome sequencing and genetic variation:

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

Note: Students are not expected to be able to fully answer these questions at this stage.

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?

Note: Students are not expected to be able to fully answer these questions at this stage.

Identifying a Variant using BLAST

Note: This activity does not introduce percent identity. If you plan to use this activity on its own, make sure that either your students complete the Introduction to Sequence Comparison reading or you introduce sequence comparison and percent identity using the Introduction to Sequence Comparison Teacher Slides.

Students should also complete Sequencing for Rare Disease Diagnosis: Scenario Introduction before beginning this activity.

If you or your students would like a refresher on the concept of a DNA variant prior to beginning this activity, check out our Minute to Understanding video "What are DNA variants?"

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.

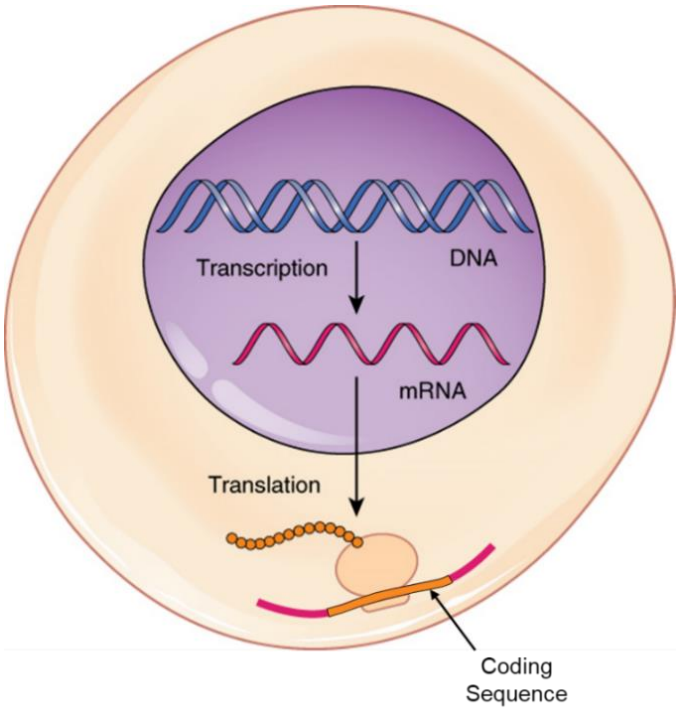


Figure 1. **Central Dogma**[1].

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. All of the sequences you will need are in this Patient DNA Sequences document.

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

Note: This part of the activity includes information on RNA processing. If your students are not familiar with RNA processing, you can skip this part of the activity.

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).

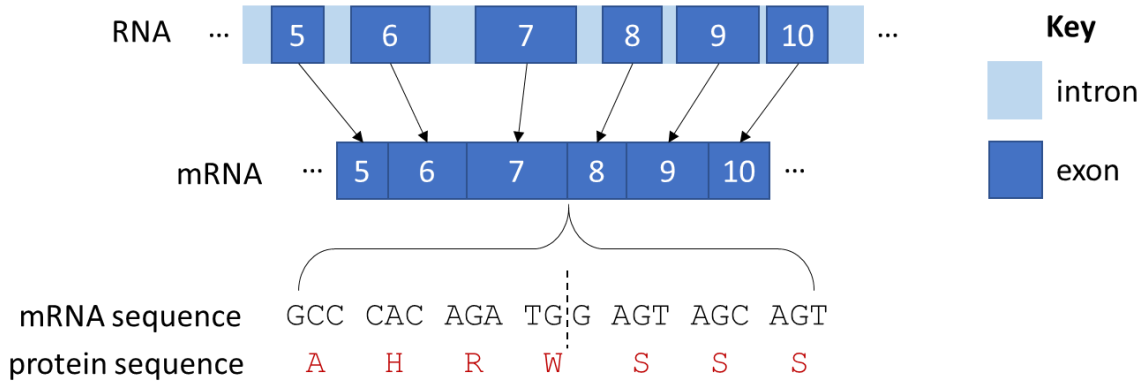


Figure 3. RNA processing without variant. This figure shows a portion of the *ATG7* gene after it has been transcribed into RNA. The RNA then gets processed into mRNA, which contains only exons. In the mRNA sequence without the variant, the last two nucleotides from exon 7 (TG) combine with the first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

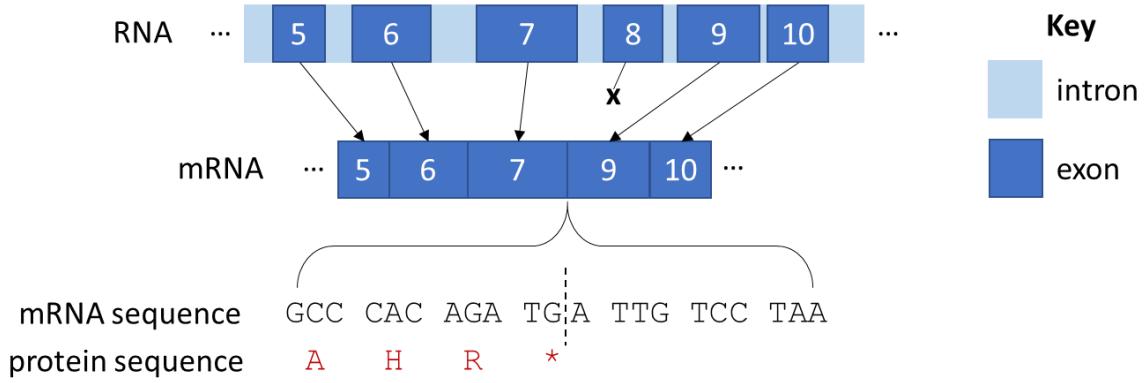


Figure 4. RNA processing with variant. This figure shows a portion of the *ATG7* gene after it has been transcribed into RNA. The RNA then gets processed into mRNA, which contains only exons. In the mRNA sequence with the variant, exon 8 is excluded from the processed mRNA. The last two nucleotides from 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?

Connecting Protein Structure and Function using PolyPhen-2, UniProt, and BLAST

Note: This activity does not introduce percent identity. If you plan to use this activity on its own, make sure that either your students complete the Introduction to Sequence Comparison reading or you introduce sequence comparison and percent identity using the Introduction to Sequence Comparison Teacher Slides.

Students should also complete Sequencing for Rare Disease Diagnosis: Scenario Introduction before beginning this activity.

If you or your students would like a refresher on the concept of a DNA variant prior to beginning this activity, check out our Minute to Understanding video "What are DNA variants?"

Introduction

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, they eventually identified ten people with similar symptoms and similar variants from four other families. The researchers sequenced those patients' genomes too.

Ultimately, they identified variants in all of the patients' genomes in a gene called Autophagy related 7 (*ATG7*), which codes for the Autophagy related 7 protein (ATG7). Through DNA sequence comparison, the researchers found that the patients' DNA variants result in changes in the amino acid sequence of the patients' ATG7 proteins.

In this activity, you will follow the researchers' process to learn how they investigated whether a change in the ATG7 amino acid sequence can impact ATG7 protein function. You will specifically examine a variant from **Family 2**, which results in a change from arginine (R) to histidine (H) at amino acid #576 in the protein sequence.

Make a Prediction

Given what you know about the relationship between protein structure and function:

1. Do you think the patients' ATG7 variants will impact protein function? Why or why not?

2. What would you want to know about the ATG7 protein and the patients' variants to better predict whether the variants will impact function?

Part 1. Functional Effect Prediction

To investigate the potential impact of the *ATG7* variants on protein function, we'll use a bioinformatics tool called PolyPhen-2. This tool makes predictions about how a single amino acid change will affect protein function. Let's use this tool to investigate **Family 2 Allele 1**.

See PolyPhen-2 Written Tutorial
Or PolyPhen-2 Tutorial Series videos: Introduction, Submitting a Query using a Protein Identifier, Accessing Your Results, and Interpreting a PolyPhen-2 Report

1. Navigate to PolyPhen-2.

2. In the field titled "Protein or SNP identifier" enter `ATG7_HUMAN`. Be sure to type in all caps.

3. The DNA variant in **Family 2 Allele 1** results in a change from Arginine (R) to Histidine (H) at amino acid #576 in the protein sequence.
    a. In the field titled "Position", enter `576`.
    b. Next to "AA_1", select `R` for Arginine.
    c. Next to "AA_2", select `H` for Histidine.
    d. In the field titled "Query description", enter `Family 2 Allele 1`.

4. Click the button labeled "Submit Query".

5. When the next page loads, wait for ~1 minute, then click the button labeled "Refresh". There should now be a link labeled "View" in the Results column. Click on that link.
    Troubleshooting notes:
    If the link does not appear, wait another minute, then click "Refresh" again.
    If you get an error, return to Step 1 and try entering the information again.
    Note: PolyPhen-2 is a real research tool. There may be researchers using the tool at the same time as your class, which can affect how quickly your students are able to obtain results. If the tool is not working or if it is running too slowly, you can instead have your students look at the PolyPhen-2 results in our Example Result Slides.

6. The results page should open either in a new tab or in the same window. When the page appears, find the "Results" section and locate the "Prediction/Confidence" report.
    a. What effect does PolyPhen-2 predict this variant will have on ATG7 protein function?

    b. Given what you know about the connection between structure and function, what information about the protein might PolyPhen-2 be using to make this prediction?

Note: This question could serve as a prompt for small group or whole group discussion. If your students are familiar with the levels of protein structure, you may wish to prompt students to think about how changing an amino acid could affect the different levels of protein structure. You can use this discussion as a transition to introduce Part 2.

Part 2a. ATG7 Function

Next, we will look at a database called UniProt. UniProt contains information about the sequence, structure, and function of different proteins. We will be using UniProt to learn more about how the function of ATG7 is related to its structure. First, we will learn more about the function of ATG7.

See UniProt Written Tutorial: Navigating a UniProtKB Entry – Protein Function
Or UniProt Tutorial Videos: Searching for a Protein, Navigating a UniProtKB Protein Entry and Finding Information on Protein Function

1. Navigate to the entry for Human ATG7 on UniProt.

2. The first section of the main entry page is titled "Function" and includes a description of the protein's function. Read over the questions below, then read through the description. Take notes under each question as you read.

    a. What are some science words you recognize?

    b. What are some science words that are unfamiliar to you?

    c. What type of protein is ATG7? Is it a structural protein, a transport protein, a hormone, a contractile protein, a defense protein, an enzyme, or a storage protein?

    d. ATG7 is involved in a process called autophagy. What do you think autophagy means?
    Hint: What other words appear near "autophagy" and "autophagic" in the description?

Note: A small group or whole class discussion after this section is recommended, to help students interpret the information and clarify their ideas. You may wish to define autophagy for the students if they do not arrive at a definition themselves. Humans Can Survive Without Key Autophagy Gene, *the popular science article about this research study published in The Scientist, contains an infographic and definition of autophagy that might be useful.*

Part 2b. Structural Changes

Next, we will use UniProt to learn more about the 3-dimensional (3D) structure of ATG7.

> *See UniProt Written Tutorial: Navigating a UniProtKB Entry – Protein Function*
> *Or UniProt Tutorial Videos: Searching for a Protein, Navigating a UniProtKB Protein Entry and Finding Information on Protein Function*

1. Navigate to the entry for Human ATG7 on UniProt.

2. Locate the "Function" section of the entry and then find the heading "Features". Notice the length of the protein is shown in grey from amino acid 1 to 703. Underneath the grey bar, colored shapes identify sections of the protein that are important for its function, also known as the protein's features. These features are also listed in the table below the grey bar.

3. Are there any protein features near the variant site (amino acid #576)? If so, which feature(s)?

The active site is the portion of an enzyme's structure that is directly responsible for interacting with other proteins or molecules.

4. How do you think a variant near the active site of ATG7 affect the protein's ability to complete its enzymatic function?

In addition to looking at the location of the variant within the amino acid sequence, we can also look at the protein's 3D structure. Remember that proteins aren't just flat sequences of amino acids—those amino acid strands fold up into a specific shape to make a functioning protein! We can represent a protein's 3D structure using connected spheres to represent all of the amino acids of the protein. **Figure 1** is a representation of the 3D structure of ATG7 protein from yeast.

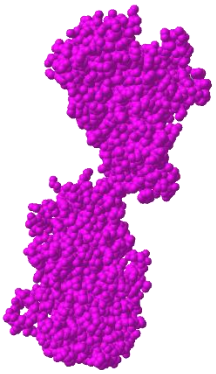


Figure 1. ATG7 3D Protein Structure[2]

Did you know? We are looking at the yeast ATG7 protein because determining the structure of protein is very difficult, and scientists have not yet determined the exact structure of human ATG7 protein. However, the yeast and human ATG7 proteins are similar enough in sequence that we can still get useful information from the 3D structure of the yeast protein.

Scientists studied ATG7 protein in model organisms, like yeast, and discovered that two ATG7 proteins need to come together and attach to each other to work properly. **Figure 2** shows two ATG7 proteins (one pink and one blue) joining together to form one functioning structure.

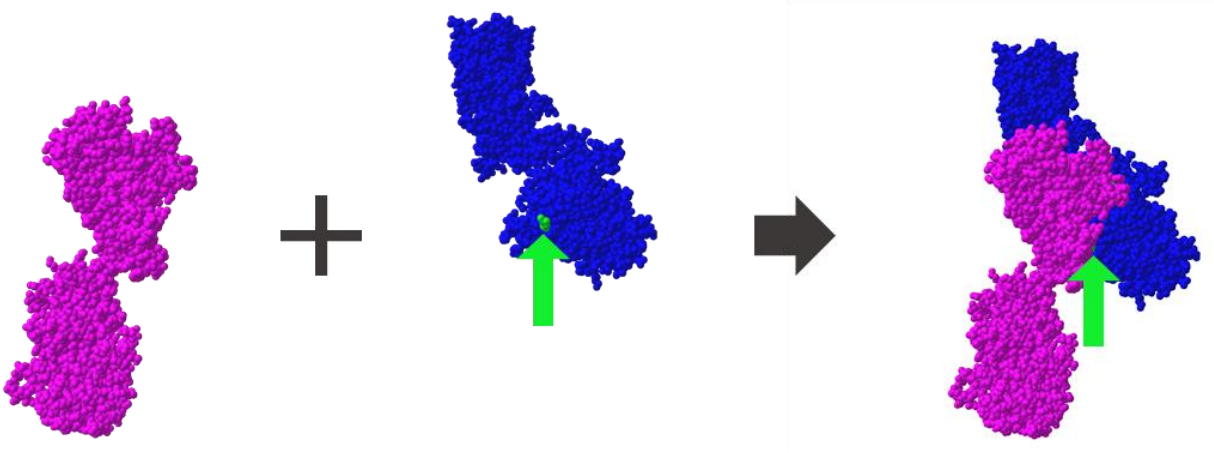


Figure 2. ATG7 Protein Interactions.[2] Two copies of the ATG7 protein interact with each other to form one functioning structure. Amino acid 576, the amino acid changed by the Family 2 Allele 1 variant, is highlighted in green.

All amino acids have unique chemical structures, leading to different protein sizes, shapes, and ways of interacting with other molecules depending on which amino acids are present. Remember that ATG7 amino acid 576 (the highlighted location in Figure 2) is typically an Arginine, but in proteins expressed from Family 2 Allele 1, that amino acid will instead be a Histidine.

5. Do you think this amino acid change will affect the ability of two ATG7 proteins to join together? Justify you answer using **Figure 2** and/or your knowledge of amino acid properties.

6. Researchers found that the two patients in Family 2 had a higher percentage of individual ATG7 proteins and a lower percentage of paired ATG7 proteins when compared to controls without the ATG7 variant. Does this evidence support your prediction in question 5? Why or why not?

[2] Structure from Kaiser SE, et al. (2012). Nat Struct Mol Biol, 19:1242-9. Data obtained through the Molecular Modeling Database (MMDB) and modeled using iCn3D.

Part 2c. Amino Acid Properties

When two ATG7 proteins join together, the R group of the arginine at amino acid #576 of one of the two proteins interacts with the R group of the glutamic acid at amino acid #589 on the other protein. Use Figure 3, which shows the structure of arginine and glutamic acid, to answer the following question.

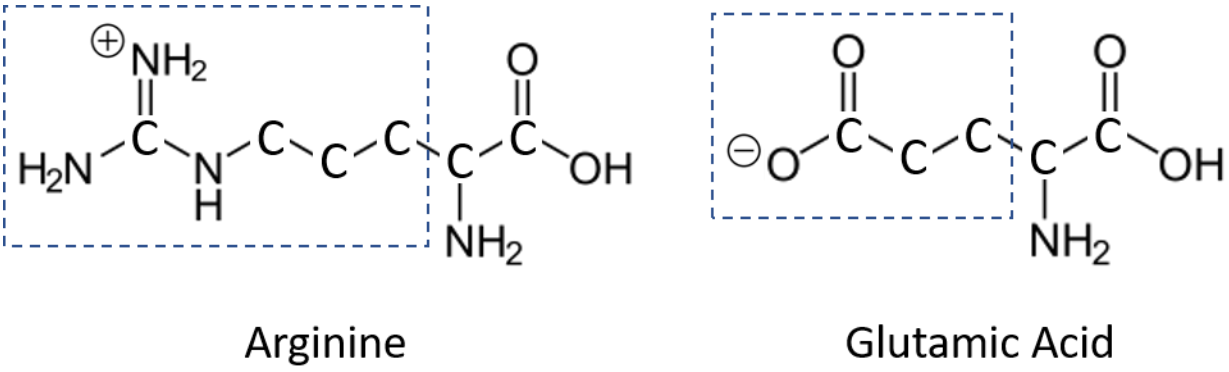


Figure 3. Structure of Arginine and Glutamic Acid. The R group, also known as the side chain, of each amino acid is highlighted by a blue box.

1. What type(s) of interaction(s) could be occurring between the R group of the arginine and the R group of the glutamic acid?

The variant in Family 2 Allele 1 results in an amino acid change from arginine to histidine at amino acid #576. Use Figure 4, which shows the structures of arginine and histidine, to answer the following questions.

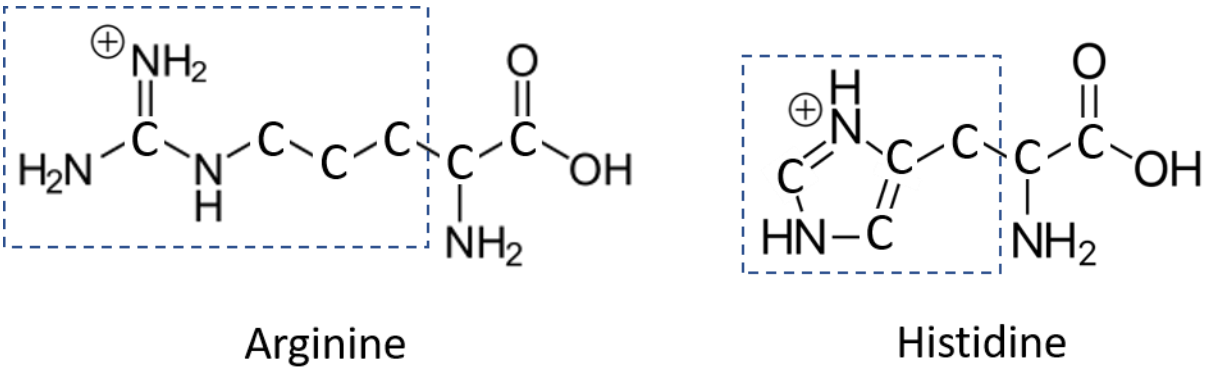


Figure 4. Structure of Arginine and Histidine. The R group, also known as the side chain, of each amino acid is highlighted by a blue box.

2. What similarities and differences do you notice in the structure of these two amino acids? You may want to think about size, shape, and charge.

3. How would you predict the interactions of amino acid #576 with the glutamic acid at amino acid #589 might change when the arginine at #576 is replaced by histidine?

Part 3. Conserved Sequences

One way to identify regions of a protein that are important for protein function is to look for conserved sequences. A conserved sequence is a protein or DNA sequence that is identical or highly similar across multiple species. If a region of a protein is conserved across many species, we can often assume that the structure of that region is important for proper protein function.

To check if the variant in Family 2 Allele 1 occurs in a conserved sequence, you will compare the human ATG7 protein reference sequence to ATG7 protein sequences from fruit fly, chimpanzee, yeast, chicken, and mouse. All of the sequences you will need are in this ATG7 Protein Sequences document.

Before you begin, let's make some predictions about how similar each of these species' ATG7 protein sequences is to the human ATG7 sequence.

1. Which species (chimpanzee, mouse, chicken, fruit fly, or yeast) do you think is most closely related to humans? Which species is least closely related to humans?

2. Given your answer to the previous question, which species do you expect to have the greatest percent identity for ATG7 protein sequence compared to human? What percent identity do you expect for this comparison?

3. Which species do you expect to have the lowest percent identity for ATG7 protein sequence compared to human? What percent identity do you expect for this comparison?

Now, let's check if the variant occurs in a conserved sequence by using BLAST.

See BLAST Tutorial Video: Comparing two or more protein sequences

4. Locate the ATG7 protein sequences.

5. Navigate to Protein BLAST and select the option to "align two or more sequences."

6. In the box labeled "Enter Query Sequence," paste the "Human ATG7 Protein" sequence. Copy the entire text of the sequence as it is written in the sequences file including the description line beginning with ">."

7. In the box labeled "Enter Subject Sequence," paste the chimpanzee, mouse, chicken, fruit fly, and yeast sequences. You can copy and paste all five sequences simultaneously.

8. Scroll to the bottom and click "BLAST." It may take several seconds for the results to appear.

9. Once the results load, scroll down to the table and navigate to the "Alignments" tab. In the "Alignment view" dropdown menu, select "Pairwise with dots for identities" to compare the five sequences to the human protein sequence.

10. Scroll down the page and look at each comparison. Within each comparison, the **Query** sequence is the human sequence, and the subject (**Sbjct**) sequence is either the chimpanzee, mouse, chicken, fruit fly, or yeast sequence.
    a. What do you notice about the similarities and differences between each subject sequence and the human sequence?

11. Now locate the percent identity score for each comparison. Enter the values into the second column of Table 1 (**% Identity for Whole Protein**). Then answer the questions below.

Table 1.

Comparison to Human ATG7	**% Identity for Whole Protein**	**% Identity for Variant Region**
Chimpanzee		
Mouse		
Chicken		
Fruit Fly		
Yeast		

    b. What is the range of percent identity across all of the comparisons?

    c. Which comparison has the highest percent identity? Which has the lowest?

    d. Does this data match your predictions? Why do you think those two organisms have the highest and lowest percent identities?

Next, to see if this variant is located in a region that is conserved, we'll look at the percent identity of a 50 amino acid region around the site of the variant.

12. Copy the sequence below and search for it on your BLAST results page. On your browser select "Edit > Find" or use keyboard shortcuts Ctrl+F or Cmd+F. The location of the variant is bolded for your reference. Searching for the sequence should highlight our region of interest in all five comparisons.

    CYFCNDVVAPGDSTRDRTLDQQCTVS**R**PGLAVIAGALAVELMVSVLQHPE

    If this does not work in your browser, look for the region between amino acids 550 and 599 of the **Query** sequence. The sequence you identify should match the above sequence, starting with C and ending with E.

13. In the human reference sequence (the **Query** sequence), the amino acid at the variant site (#576) is an arginine (**R**). In the other five species, is the equivalent amino acid also an arginine? Or does it vary between the different species?

14. For each comparison, calculate the percent identity of **the 50 amino acid region only**.
    e. Calculate the percent identity of this region for:
        i. Chimpanzee

        ii. Mouse

        iii. Chicken

        iv. Fruit Fly

        v. Yeast

    f. Enter the values above into the third column of Table 1. Are these values a high or low level of percent identity?

g. How does the percent identity of the region around the variant compare to the percent identity for the whole protein?

15. Is the variant in **Family 2 Allele 1** in a conserved region of the protein? Use the data you collected to justify your answer.

Part 4. Reflection

Activity Reflection

1. At the start of this activity, you predicted whether the patients' variants would impact ATG7 protein function. After investigating the effect of one variant from Family 2, do you still agree with your initial prediction? Why or why not?

2. In Part 1 of this activity, you used PolyPhen-2 to predict the effect of the variant from Family 2 on ATG7 function. Based on what you learned from the rest of the activity, what information about the protein do you think PolyPhen-2 is using to make this prediction? Is this the same or different from what you predicted at the end of Part 1?

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?

Final Reflection and Bioethics

Guiding Questions Final Reflection

Revisit the following guiding questions and update your answers to include everything you've learned from completing the activities in this module.

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?

Bioethics

Outcomes of Rare Disease Research

At the beginning of this module, you were asked to imagine that you were the parent of two daughters with a rare disorder. This scenario was based on a true story of a family searching for a medical diagnosis for their daughters so that they could connect with other patients and families. However, after many doctors' visits, the family still had not been able to receive a diagnosis.

They decided to participate in a research study and have the daughters' genomes sequenced. Using the sequencing data, researchers were able to identify variants in the gene Autophagy related 7 (*ATG7*) as the genetic cause of the daughters' disorder. They also connected the family with ten other individuals who had similar symptoms and genetic variants and who ranged in age from 5 years old to 71 years old.

1. Once again, imagine you are the parents in this family. How do you think you would feel after learning for the first time that there were other people, both kids and adults, who had the same disorder as your daughters? What would be going through your mind?

2. How do you think you would feel as one of the daughters? What would be going through your mind?

After the research study, the family finally received a medical diagnosis for their daughters: *Spinocerebellar ataxia, autosomal recessive 31 (SCAR31)*.

Although researchers identified genetic variants and established a diagnosis of SCAR31, there is still no treatment available for the disorder. The researchers know that the daughters' *ATG7* variants affect their Autophagy related 7 (ATG7) protein structure, which in turn affects ATG7 protein function in their cells. However, the researchers still do not fully understand how the loss of ATG7 protein function produces the daughters' specific symptoms.

3. Is there value in knowing the genetic cause of a disorder when there is no treatment available yet? How might having a diagnosis impact the patients and their families even without an available treatment?

As part of the study, the researchers performed **gene editing** on the patients' cells in the laboratory, replacing the patients' *ATG7* gene with a functioning copy of the *ATG7* gene. They found that the gene edited cells had normal ATG7 protein function.

These promising results in the lab are an important first step in the lengthy process for therapy research and development. After this discovery phase, all potential treatments must undergo preclinical testing in human cells or animal models in the laboratory to make sure they are safe and effective. Treatments must then go through **clinical trials**, where they are evaluated for safety and efficacy in groups of patients, before they can be made widely available. The entire treatment research and development process typically takes several years to complete.

4. If you were a patient or a family member, how would you want the researchers to proceed?

5. Would you want to participate in a clinical trial for gene editing? Why or why not?

6. Considering the time and resources needed to develop new therapies, what might be some barriers to future research on a treatment for this disorder?

Note: For more on this topic, check out the lesson "Identifying & Understanding Rare Genetic Conditions: Meet Tess Bigelow" from the Personal Genetics Education Project (pgEd).

Access to Researchers and Therapies

After identifying *ATG7* variants in the two daughters of the first family, the researchers used GeneMatcher to identify additional patients with mutations in *ATG7* for their study.

GeneMatcher is a freely accessible site designed to help connect doctors, researchers, and patients who are interested in the same genes. The site enabled the researchers to connect with other patients with *ATG7* variants from the UK, France, Germany, Switzerland, and Saudi Arabia.

1. Most of the families included in this study are from Europe. What are some possible explanations for this?

Even if a therapy was developed for treating SCAR31, patients would first need to obtain a diagnosis before they could be treated. From the story of the first family in this research study, we know that finding a diagnosis for a rare disorder can be a long, difficult process. Rare diseases are often overlooked as a possibility, because doctors are typically trained to consider more common diagnoses first. It can be challenging to find a doctor who is willing to consider a rare disease diagnosis.

Additionally, a patient's ability to access doctors, researchers, genome sequencing, and therapies can vary significantly depending on factors such as health insurance status, socioeconomic status, age, race, and geographic location.

2. How might limited access to health care impact this already challenging endeavor of obtaining a diagnosis? Who is likely to be most impacted?

Note: For more on this topic, check out the lesson "When New Treatments Come with Big Hopes and a Big Price Tag" from the Personal Genetics Education Project (pgEd).

References:

Collier, J.J., et al. (2021). Developmental Consequences of Defective ATG7-Mediated Autophagy in Humans. *N Engl J Med, 384*:2406-2417. https://dx.doi.org/10.1056/NEJMoa1915722

Melchor, A. (2021). Humans Can Survive Without Key Autophagy Gene. *The Scientist*. https://www.the-scientist.com/news-opinion/humans-can-survive-without-key-autophagy-gene-68986

Implementation Strategies

This module was built to be flexible and adaptable to different classroom environments. These strategies are provided as examples for how you might implement these activities in the classroom. If you have any questions about implementing this module, reach out to the TtGG team at ttgg@jax.org.

Strategy 1: Full Module

Day 1

Pre-work:

- Complete the Introduction to Sequence Comparison & Identity reading and knowledge check
- Review the TtGG Gene Info Sheet(s) for instructor's gene(s) of choice

In class:

- Review the Introduction to Sequence Comparison Teacher Slides and address any questions
- Sequence Comparison with *ACE*, *ACTN3*, *OXTR*, *CYP2C19*, and/or *TAS2R38*
    - Individually, each student working on the same gene
    - In groups, each group working on the same gene
    - In groups as jigsaw, each group working on a different gene

Day 2

Pre-work:

- Complete the Sequencing for Rare Disease Diagnosis: Scenario Introduction reading and reflection, and answer the guiding questions
- Complete the Identifying a Variant using BLAST introduction, including the Central Dogma and Codon Chart readings and knowledge checks

In class:

- Discuss student pre-work answers to the scenario reflection and guiding questions
    - In pairs/groups
    - As a whole class
- Complete the Identifying a Variant using BLAST activity Parts 1-4
    - Individually
    - In pairs/groups
- Complete the Identifying a Variant using BLAST activity Part 5
    - Individually (in class or as post-work)
    - In pairs/groups
    - As a whole class

Day 3

Pre-work:

- Complete the Connecting Protein Structure & Function introduction, including making predictions

In class:

- Discuss student pre-work predictions in pairs/groups
- Complete the Connecting Protein Structure & Function activity Part 1
    - Individually
    - In pairs/groups
- Discuss the PolyPhen-2 Model (Part 1, question 6b) as a whole class, documenting student suggestions to revisit at the end of the activity

- Complete the Connecting Protein Structure & Function activity Part 2a
  - Individually
  - In pairs/groups
- Discuss the function of ATG7 as a whole class
- Complete the Connect Protein Structure & Function activity Parts 2b-c

Day 4
In class:

- Complete the Connect Protein Structure & Function activity Part 3
  - Individually
  - In pairs/groups, each group completing entire activity
  - In groups as jigsaw, each group working on different organism
- Complete the Connect Protein Structure & Function activity Part 4
  - Individually
  - In pairs/groups
- Revisit the PolyPhen2 model predictions (Part 1, question 6b; Part 4, question 2) as a whole class
- Complete the Final Reflection & Bioethics activity
  - Individually
  - In pairs/groups
  - As a whole class

Strategy 2: Sequence Comparison with the TtGG Genes Only

Pre-work:

- Complete the Introduction to Sequence Comparison & Identity reading and knowledge check
- Review the TtGG Gene Info Sheet(s) for instructor's gene(s) of choice
- Answer guiding question #1 only from Sequencing for Rare Disease Diagnosis Introduction (students do not need to read the scenario or answer the scenario reflection questions)

In-class:

- Review the Introduction to Sequence Comparison Teacher Slides and address any questions
- Sequence Comparison with *ACE*, *ACTN3*, *OXTR*, *CYP2C19*, and/or *TAS2R38*
  - Individually, each student working on the same gene
  - In groups, each group working on the same gene
  - In groups as jigsaw, each group working on a different gene

Post-work:

- Revisit guiding question #1 from the Sequencing for Rare Disease Diagnosis Introduction
  - Students revise and submit their answer
  - Group discussion as a class

Strategy 3: Sequencing for Rare Disease Diagnosis Only

Day 1
Pre-work:

- Complete the Introduction to Sequence Comparison & Identity reading and knowledge check
- Complete the Central Dogma and Codon Chart readings and knowledge checks from the Identifying a Variant using BLAST introduction

In class:

- Complete the Sequencing for Rare Disease Diagnosis: Scenario Introduction reading and reflection, and answer the guiding questions

    - Individually
    - In pairs/groups
- Review the Introduction to Sequence Comparison Teacher Slides and address any questions
- Complete the Identifying a Variant using BLAST Background reading and activity Parts 1-4
    - Individually
    - In pairs/groups

Post-work:

- Complete the Identifying a Variant using BLAST activity Part 5

Day 2

Pre-work:

- Complete the Connecting Protein Structure & Function introduction, including making predictions

In class:

- Discuss student post-work reflection and pre-work predictions in pairs/groups
- Complete the Connecting Protein Structure & Function activity Part 1
    - Individually
    - In pairs/groups
- Discuss the PolyPhen-2 Model (Part 1, question 6b) as a whole class, documenting student suggestions to revisit at the end of the activity
- Complete the Connecting Protein Structure & Function activity Part 2a
    - Individually
    - In pairs/groups
- Discuss the function of ATG7 as a whole class
- Complete the Connect Protein Structure & Function activity Parts 2b-c

Day 3

In class:

- Complete the Connect Protein Structure & Function activity Part 3
    - Individually
    - In pairs/groups, each group completing entire activity
    - In groups as jigsaw, each group working on different organism
- Complete the Connect Protein Structure & Function activity Part 4
    - Individually
    - In pairs/groups
- Revisit the PolyPhen2 model predictions (Part 1, question 6b; Part 4, question 2) as a whole class
- Complete the Final Reflection & Bioethics activity
    - Individually
    - In pairs/groups
    - As a whole class

NGSS Alignments

Activity or Lesson	Disciplinary Core Ideas	Cross Cutting Concepts	Science & Engineering Practices
Introduction to Sequence Comparison & Percent Identity		Patterns	Using Mathematics and Computational Thinking
Sequence Comparison with the TtGG Genes	**LS1.A** Structure and Function **LS3.A** Inheritance of Traits **LS4.A** Evidence of Common Ancestry and Diversity	Structure and Function Patterns	Using Mathematics and Computational Thinking Engaging in Argument from Evidence
Sequencing for Rare Disease Diagnosis: Scenario Introduction			Asking Questions and Defining Problems Obtaining, Evaluating, and Communicating Information
Identifying a Variant Using BLAST	**LS1.A** Structure and Function **LS3.A** Inheritance of Traits **LS3.B** Variation of Traits	Structure and Function Patterns	Analyzing and Interpreting Data Using Mathematics and Computational Thinking Engaging in Argument from Evidence
Connecting Protein Structure and Function using PolyPhen-2, UniProt, and BLAST	**LS1.A** Structure and Function **LS3.B** Variation of Traits **LS4.A** Evidence of Common Ancestry and Diversity	Structure and Function Patterns Cause and Effect	Analyzing and Interpreting Data Using Mathematics and Computational Thinking Constructing Explanations and Designing Solutions Engaging in Argument from Evidence Obtaining, Evaluating, and Communicating Information
Sequencing for Rare Disease Diagnosis: Final Reflection & Bioethics			Engaging in Argument from Evidence Obtaining, Evaluating, and Communicating Information

Supporting Materials

Answer Key

An answer key is available to teachers upon request by emailing the TtGG team at ttgg@jax.org.

Database Links

- NCBI BLAST
  - Nucleotide BLAST
  - Protein BLAST
- PolyPhen-2
- UniProt

Note: If any of these databases are not working when you implement the module in the classroom, you can provide a screenshot of the activity results to students. Screenshots of the database results for each activity can be found in the Example Result Slides.

Documents and Slides

- TtGG Gene Info Sheets
- Introduction to Sequence Comparison Teacher Slides
- Example Result Slides

Tutorials

- BLAST
  - Written Tutorials
  - Video Tutorials
    - Comparing two or more protein sequences
- PolyPhen-2
  - Written Tutorials
  - Video Tutorials
    - Introduction
    - Submitting a Query using a Protein Identifier
    - Submitting a Query using a SNP Identifier
    - Submitting a Query using a Protein Sequence
    - Accessing Your Results
    - Interpreting a PolyPhen-2 Report
- UniProt
  - Written Tutorials
  - Video Tutorials
    - Introduction
    - Searching for a Protein
    - Navigating a UniProtKB Protein Entry
    - Finding Information on Protein Function

Minute to Understanding Video

- What are DNA variants?

Related Ethics Lessons from the Personal Genetics Education Project (pgEd)

- Snapshots
  - Identifying & Understanding Rare Genetic Conditions: Meet Tess Bigelow
  - When New Treatments Come with Big Hopes and a Big Price Tag
  - Privacy Protections for Genetic Information: Meet GINA
  - Introduction to Genetics and Medicine
- Full lessons
  - Introduction to Personal Genetics
  - Personalized Medicine
  - Genome Editing and CRISPR

Related Ethics Materials from ELSIhub

- The Disability Rights Critique of Technologies that Eliminate Human Genetic Variation
- Paying for Cures: The Ethics and Economics of Gene Therapies for Rare Diseases

Feedback

Did you use any of the Sequence Comparison and Identity content in your class(es)? If so, we'd love to hear how it went. Use the following form to provide the TtGG team with feedback.

Feedback Form