

Capstone Activity

Identification and Bioinformatics Analysis of Cas9 Target Sites

The ability of the CRISPR-Cas9 system to accurately and permanently edit genomes has major implications for the treatment of diseases. Some diseases, such as coronary artery disease, sickle cell disease, and cystic fibrosis, are caused by genetic mutations. A CRISPR-based therapy that can edit the genomic DNA in cells may be able to correct those mutations.

Though this type of therapy is promising, it is not as clear-cut as it may seem on the surface. As with any therapy, there are risks involved that must be analyzed and understood before testing in humans. For example, off-target effects, where a gene or DNA sequence other than the intended target is edited, can have dire effects on an organism. These types of risk can never be completely eliminated, but their probabilities and the conditions in which they may occur must be evaluated.

One of the first steps in designing a CRISPR-based therapy is identifying a gene-editing strategy and selecting a Cas9 target site to cut. In this activity, you will research the genetic basis for a disease and explore a CRISPR-based gene-editing strategy: replacing, inserting, or deleting a sequence. Then you will identify potential Cas9 target sites and use the basic local alignment search tool (BLAST) from the National Center for Bioinformatics Information (NCBI, part of the National Institutes of Health, NIH) to search for similar sequences in the human genome. Using these data, you will analyze your potential Cas9 target sites for risk of off-target effects to identify the most promising candidate for a CRISPR-based therapy.

Part 1. Identify and Catalog Target Sequences

1. Read the background information about the disease you are investigating.

Discuss with your group and answer the disease-specific reading questions.

2. Scan the provided DNA sequence and identify all potential Cas9 target sequences.

Consider the following:

- A Cas9 target site includes a 20-nucleotide protospacer sequence followed downstream by an appropriate PAM sequence (5'-NGG) in the 5' to 3' direction. Therefore, a target sequence is 23 nucleotides long
- A search for PAM sequences first may speed up the process
- Target sequences can be found on either DNA strand, but always in the 5' to 3' direction

3. Select 2–4 candidate target sequences to investigate.

Record each sequence in the table below, using the following naming convention: Gene name abbreviation-your initials-#. For example, GENE9-TRP-1.

| Target Sequence Name | Position # of First Nucleotide | Position # of Last Nucleotide | 23-Nucleotide Target Sequence, 5' to 3' |
|----------------------|--------------------------------|-------------------------------|---|
| | | | |
| | | | |
| | | | |
| | | | |

Part 2. Perform BLAST Search for Off-Target Sequences

A complete or partial Cas9 target sequence can sometimes be found elsewhere in the human genome, so an sgRNA designed against such a site may guide Cas9 to cut off-target sites. You will use the bioinformatics software BLAST to find genes with sequences that completely or partially match the target sites you selected above.

1. Prepare your results table.

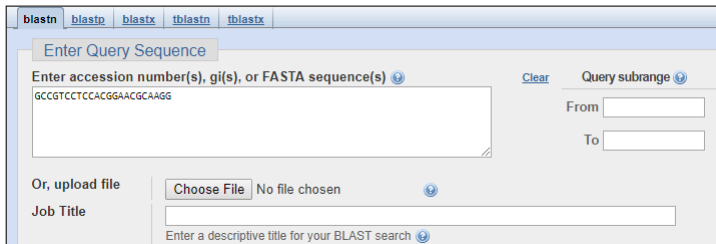
Use the bioinformatics results table provided in the student guide or recreate it in a spreadsheet program.

2. Perform a BLAST search.

The BLAST interface changes frequently. The following instructions and screenshots may deviate slightly from your experience on the site.

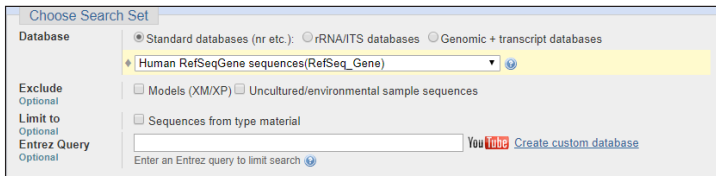
2.1. Visit blast.ncbi.nlm.nih.gov

2.2. Click **Nucleotide BLAST**. Copy your target sequence from your table and paste it under **Enter Query Sequence**.



The screenshot shows the 'Enter Query Sequence' section of the BLAST interface. At the top, there are tabs for 'blastn', 'blastp', 'blastx', 'tblastn', and 'tblastx'. Below the tabs is a text input field containing the sequence 'GCCGTCCTCCACGGAAACGCAAGG'. To the right of the input field is a 'Clear' button and a 'Query subrange' section with 'From' and 'To' input fields. Below the input field is an 'Or, upload file' section with a 'Choose File' button and 'No file chosen' text. At the bottom, there is a 'Job Title' input field with the placeholder text 'Enter a descriptive title for your BLAST search'.

2.3. Under **Choose Search Set > Database**, select **Human RefSeqGene sequences (RefSeq_Gene)**.



The screenshot shows the 'Choose Search Set' section of the BLAST interface. It features a 'Database' section with three radio buttons: 'Standard databases (nr etc.)', 'rRNA/ITS databases', and 'Genomic + transcript databases'. The 'Standard databases (nr etc.)' option is selected, and a dropdown menu below it shows 'Human RefSeqGene sequences(RefSeq_Gene)' selected. To the left, there are 'Exclude' and 'Limit to' sections with checkboxes for 'Models (XM/XP)', 'Uncultured/environmental sample sequences', and 'Sequences from type material'. At the bottom, there is an 'Entrez Query' input field with the placeholder text 'Enter an Entrez query to limit search' and a 'Create custom database' link.

2.4. Select **Show results in a new window** and click **BLAST**.

Search database **Nucleotide collection (nr/nt)** using **Megablast (Optimize for highly similar sequences)**
 Show results in a new window

2.5. The **BLAST** search may take a few minutes to complete, depending on the server usage volume. When complete, a screen similar to that below appears.

The screenshot displays the NCBI BLAST search results interface. At the top, there are navigation links: < Edit Search, Save Search, Search Summary, How to read this report?, BLAST Help Videos, and Back to Traditional Results Page. A message indicates that search parameters were adjusted for a short input sequence. The search parameters are listed on the left: Job Title (Nucleotide Sequence), Query ID (lcl|Query_29909), Database (genomic/9606/RefSeqGene), and Program (BLASTN). The Filter Results section on the right allows for filtering by Organism, Percent Identity, and E value. Below the filters, there are tabs for Descriptions, Graphic Summary, Alignments, and Taxonomy. The main section is titled 'Sequences producing significant alignments' and shows a table with 99 sequences selected. The table columns are: Description, Max Score, Total Score, Query Cover, E value, Per. Ident, and Accession. The top four results are highlighted with checkmarks.

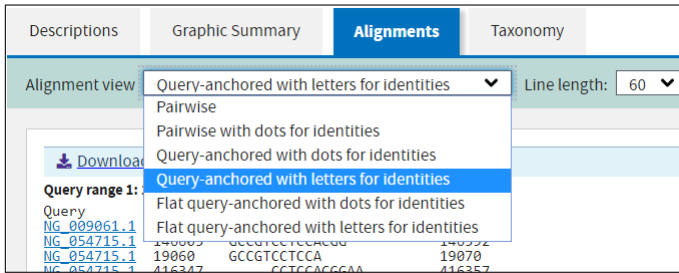
| Description | Max Score | Total Score | Query Cover | E value | Per. Ident | Accession |
|--|-----------|-------------|-------------|---------|------------|-------------|
| <input checked="" type="checkbox"/> Homo sapiens proprotein convertase subtilisin/kexin type 9 (PCSK9), RefSeqGene (LRG_275) on chromosome 1 | 38.2 | 38.2 | 100% | 0.015 | 95.65% | NG_009061.1 |
| <input checked="" type="checkbox"/> Homo sapiens Gse1 coiled-coil protein (GSE1), RefSeqGene on chromosome 16 | 28.2 | 72.8 | 69% | 14 | 100.00% | NG_054715.1 |
| <input checked="" type="checkbox"/> Homo sapiens scaffold attachment factor B2 (SAFB2), RefSeqGene on chromosome 19 | 28.2 | 28.2 | 60% | 14 | 100.00% | NG_050735.1 |
| <input checked="" type="checkbox"/> Homo sapiens tectorin alpha (TECTA), RefSeqGene on chromosome 11 | 28.2 | 28.2 | 60% | 14 | 100.00% | NG_011633.1 |
| <input checked="" type="checkbox"/> Homo sapiens aryl hydrocarbon receptor interacting protein like 1 (AIPL1), RefSeqGene on chromosome 17 | 28.2 | 28.2 | 60% | 14 | 100.00% | NG_008474.1 |

3. Review the **BLAST** search results.

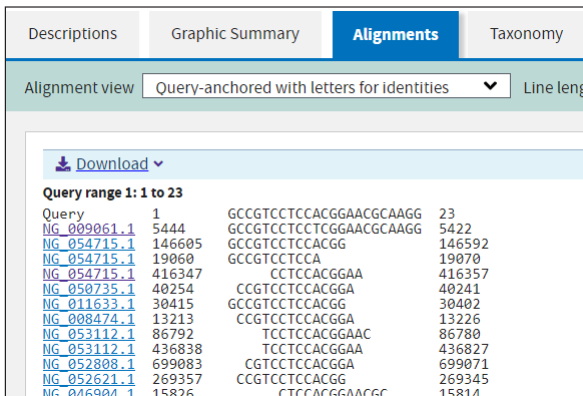
3.1. One of the top results should be an exact match for your query in the gene you are working with. Click the description link and review the information provided. Then return to the **BLAST** results page.

3.2. Click other links in your results list to orient yourself to the information provided. Then return to the **BLAST** results page.

3.3. Select all the sequences (select all), click to open the *Alignments* tab, then select *Alignment view > Query-anchored with letters for identities*.



3.4. Review the sequence information.



- Your sequence (Query) appears at the top of the results for easy reference. The remaining rows display nucleotide sequences from genes that match or partially match your query sequence.
- The links at left are the accession numbers for reference genes. Each gene in the database has its own accession number.
- To the right of each accession number is the start position of the alignment sequence, followed by the sequence and the number of the nucleotide at the end position. Most alignment sequences will not match the full length of the query sequence.

3.5. One of the top results should be an exact match for your query and be located in the gene you are working with. Click the accession number. Ensure the information given for this match is correct (that it matches the gene you are working with). If it does not match exactly, check that your query sequence is correct.

4. Annotate your results.

4.1. In your bioinformatics results table record five alignment results (or as many as possible if there are fewer than five) whose sequences include the PAM sequence 5'-NGG aligned with that of the query sequence, where N is any nucleotide. Use the accession number link to retrieve additional information to fill in the table. There may be multiple results for a single accession number.

4.2. Highlight any exact alignment matches in genes other than your intended target.

4.3. Circle the longest alignment sequence(s) other than your intended target.

5. Repeat steps 2–4 for each of your target sequences.

Focus questions

A. Why might there be multiple results from a single accession number?

B. How might an exact alignment match be an off-target cut site?

C. Does the presence of alignment matches indicate higher or lower risk of off-target effects?

Focus questions

- A. If you were to continue evaluating the candidate target sites for use in a therapy, what are two additional pieces of information or experiments that would help you?**
- B. What health problems could arise from off-target CRISPR-Cas9 activity?**
- C. How would you decide whether the risk of off-target activity for a CRISPR-Cas9 therapy is low enough to be considered safe?**
- D. Should off-target effects be considered for nontherapeutic CRISPR experiments in the laboratory? Explain why or why not.**
- E. Do you think there should be differences between how off-target risk is evaluated for CRISPR-based therapies and for laboratory CRISPR experiments?**

Coronary Artery Disease

Background

Cardiovascular disease is the leading cause of death worldwide, claiming over 17 million lives annually. One type of cardiovascular disease, coronary artery disease (CAD), in which blood vessels near the heart become narrowed due to plaque buildup, claims nearly 8 million lives each year. Lowering levels of low density lipoprotein (LDL) cholesterol has been shown to effectively reduce risk of CAD. LDL receptors (LDLR) in the liver clear LDL from blood plasma. However, levels of LDLR are themselves reduced by proprotein convertase subtilisin/kexin type 9 (PCSK9), a serine protease that binds and degrades the receptors (Figure 6). People with mutations in the *PCSK9* gene commonly have lower levels of LDL cholesterol likely because they have higher levels of the LDL-clearing receptors.

Gene-Editing Therapy Strategy

A goal of gene-editing therapy may be to reduce or eliminate PCSK9 enzyme function. One strategy is to disrupt the gene by making a cut within exon 1 and allowing nonhomologous end joining (NHEJ) to occur. This strategy would reduce levels of functional PCSK9 enzyme in the liver, which would reduce degradation of the LDL receptors to allow more removal of LDL cholesterol from the bloodstream.

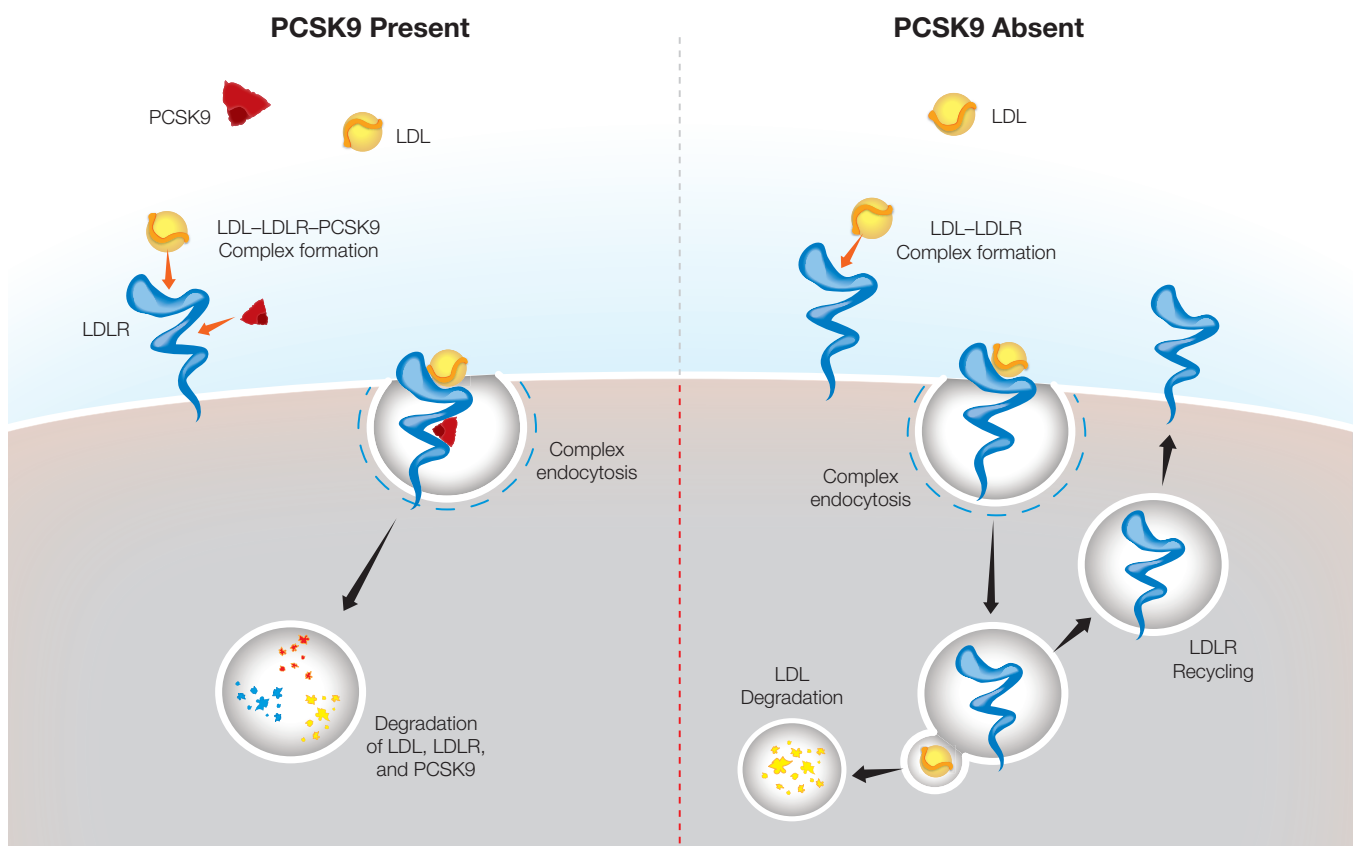


Fig. 6. Degradation of LDL in the presence and absence of PCSK9.

Focus questions

A. Describe how disrupting PCSK9 would impact gene expression. Draw a model that illustrates your description.

B. Describe two potential advantages and two potential disadvantages of administering such a therapy to liver cells only.

C. Is this gene editing strategy an example of replacing, inserting, or deleting a sequence?

PCSK9 gene sequence information

Gene Accession Number: NG_009061.1

Gene Reference: *Homo sapiens proprotein convertase subtilisin/kexin type 9*

Gene Abbreviation: PCSK9

The sequence below is an excerpt of PCSK9 exon 1 nucleotide position 5,387 to 5,446.

```
5387 5' -GGACGAGGACGGCGACTACGAGGAGCTGGTGCTAGCCTTGCCTCCGAGGAGGACGGCCT-3' 5446
      3' -CCTGCTCCTGCCGCTGATGCTCCTCGACCACGATCGGAACGCAAGGCACCTCCTGCCGGA-5'
```

Sickle Cell Disease

Background

Sickle cell disease is an inherited blood disorder in which a person's red blood cells become sickle-shaped, which increases the risk of blood clots. When we scrape a knee or cut a finger, blood clots form externally to create a scab over the wound and promote healing. When blood clots form internally, however, they can block blood vessels and cause pain or even death. Approximately 100,000 people die of complications from sickle cell disease each year.

Sickle cell disease is caused by a single nucleotide polymorphism (SNP) in the hemoglobin B (*HBB*) gene called rs334. People homozygous for adenine at rs334 produce normal hemoglobin while those homozygous for thymine at rs334 produce sickling hemoglobin and have the disease. Heterozygous individuals do not exhibit symptoms of the disease and have increased resistance to malaria, which increases their evolutionary fitness in regions where malaria is common.

Gene-Editing Therapy Strategy

The goal of gene-editing therapy for sickle cell disease is to allow expression of functional non-sickling hemoglobin. Red blood cells, which carry hemoglobin, are formed from hematopoietic stem cells in bone marrow (Figure 5). A potential gene-therapy strategy is to harvest hematopoietic stem cells from a patient, edit the *HBB* gene in those cells, and reintroduce them to the patient. Using a patient's own cells greatly reduces the chance of rejection by the patient's immune system. CRISPR-based gene editing is used to replace thymine with adenine at rs334 by making a cut near the SNP and introducing the correct sequence using homology-directed recombination (HDR).

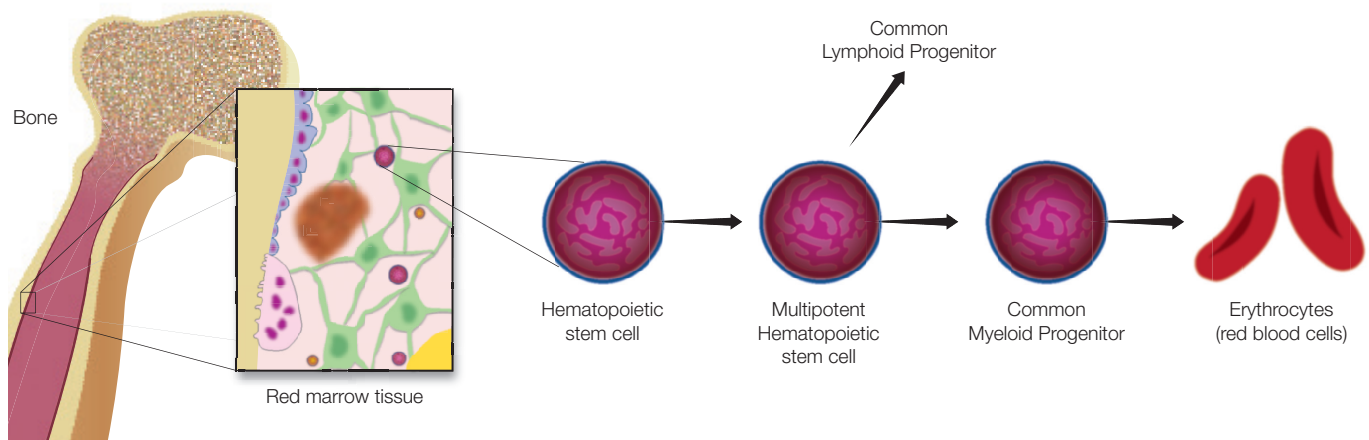


Fig. 5. Differentiation of hematopoietic stem cells into red blood cells.

Focus questions

A. For this gene-editing strategy, why is it useful to edit the DNA of only hematopoietic stem cells?

B. What other cells could be edited to achieve similar results? List two potential advantages and two potential disadvantages of editing these cells instead of hematopoietic cells.

C. Is this gene editing strategy an example of replacing, inserting, or deleting a sequence?

HBB gene sequence information

Gene Accession number: **NG_059281.1**

Gene Reference: **Homo sapiens hemoglobin subunit beta (HBB),
RefSeqGene on chromosome 11**

Gene Abbreviation: **HBB**

The sequence below is an excerpt of *HBB*, nucleotide position 5,053 to 5,106, with rs334 shown bolded and with an asterisk.

*

5053 5' -GGTGCATCTGACTCCTGT**GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT**-3' **5106**
3' -CCACGTAGACTGAGGACACCTCTTCAGACGGCAATGACGGGACACCCCCTTCCA-5'

Cystic Fibrosis

Background

Cystic fibrosis (CF) is an autosomal recessive disease that affects the lungs, pancreas, and small intestine. The disease affects about 70,000 individuals worldwide. It causes buildup of viscous mucus in these organs and frequently leads to severe lung infections. If untreated, most CF patients do not live past their 20s. The disease is caused by mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene on chromosome 7, the most common of which is an in-frame deletion of three base pairs in exon 11 that codes for phenylalanine (F508del). *CFTR* is a protein that transports chloride ions across cell membranes, which is critical to effective clearing of mucus from airways (Figure 7). The F508del mutation impairs the normal production of *CFTR*.

Gene-Editing Therapy Strategy

The goal of cystic fibrosis gene-editing therapy is to correct the *CFTR* gene in lung epithelial stem cells. A therapeutic drug will likely be given by inhalation to target the lungs and the edit will occur *in vivo*. CRISPR technology would be used to create a cut in exon 11 in the vicinity of the *CFTR* F508del mutation and use CRISPR-mediated homologous recombination to replace the mutation with a healthy version of the gene.

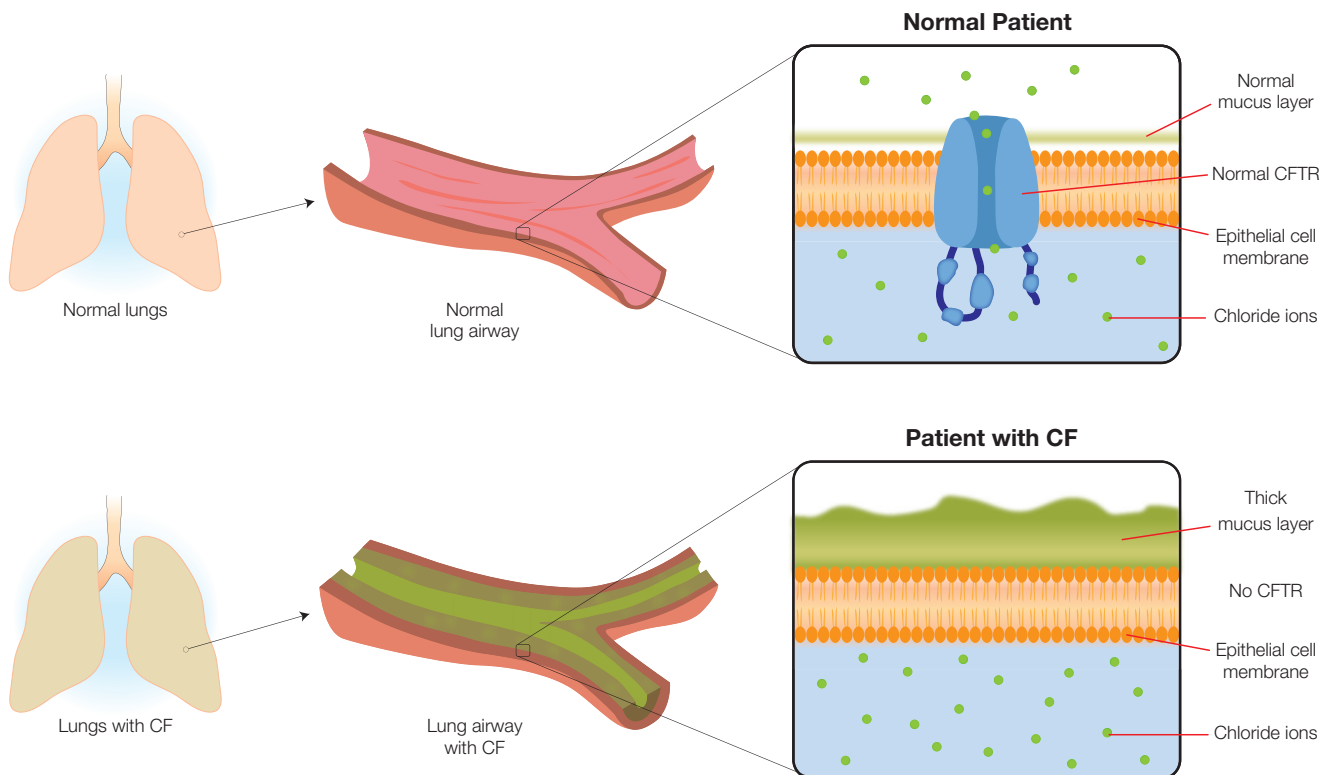


Fig. 7. The mucus layers in lung airways of healthy patients and those with CF.

Focus questions

A. Describe how editing CFTR would impact gene expression. Draw a model that illustrates your description.

B. What are two potential advantages and two potential disadvantages of administering a cystic fibrosis gene editing therapy by inhalation instead of orally, by injection, or another method?

C. Is this gene editing strategy an example of replacing, inserting, or deleting a sequence?

CFTR gene sequence information

Gene Accession Number: NG_016465.4

Gene Reference: *Homo sapiens CF transmembrane conductance regulator (CFTR)*

Gene Abbreviation: CFTR

The sequence below is an excerpt from exon 11 of CFTR, nucleotide position 98,756 to 98,815, with the three-nucleotide mutation shown bolded and with asterisks.

```

                                                                 ***
98756  5' -TTCTGTTCTCAGTTTTCTGGATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGT-3'  98815
      3' -AAGACAAGAGTCAAAGGACCTAATACGGACCGTGGTAATTTCTTTTATAGTAGAAACCA-5'
```


Glossary

Arabinose-inducible promoter — promoter that occurs naturally in bacterial systems and that is used in many expression plasmids to allow regulation of expression of a target gene: expression is induced in the presence of arabinose and repressed in its absence.

Cas9 — CRISPR-associated protein 9 (Cas9), an endonuclease that forms a double-strand break (cuts) in DNA at a specific site within a larger recognition sequence, or target site. It is involved in the natural defense of certain prokaryotes against DNA viruses, and it is heavily utilized in genetic engineering applications to cut DNA at locations specified by a guide RNA (gRNA)

CRISPR — clustered regularly interspaced palindromic repeats (CRISPR) are sequences in the genomes of some prokaryotes that act as a genomic record of previous viral attack. Along with CRISPR-associated (Cas) proteins, bacteria use the sequences to recognize and disarm future invading viruses. Scientists have adapted this system for genetic engineering purposes.

Donor template DNA — engineered sequence of DNA required for homology-directed repair in CRISPR gene editing applications; may include a desired sequence flanked on both sides by “homology arms” that match the sequence upstream and downstream of the cut.

β -galactosidase — encoded by the *lacZ* gene, this enzyme hydrolyzes galactose-containing carbohydrates, including lactose. Conveniently, it also breaks down the colorless compound *X-gal* into two pieces, one of which goes on to form a deep blue pigment.

Guide RNA (gRNA) — non-coding, short RNA sequence that binds to Cas9 and to complementary target DNA sequences, where Cas9 performs its endonuclease activity to cut the target DNA strand.

Guiding region — part of the CRISPR RNA or crRNA in nature, a typically 20-nucleotide region of sgRNA that is complementary to the target DNA sequence and that defines where Cas9 cuts. Scientists can easily customize this sequence for their own targets.

Homology directed repair (HDR) — DNA repair mechanism in which specific proteins patch a double-strand DNA using donor template DNA.

Isopropyl β -d-1-thiogalactopyranoside (IPTG) — a non-metabolizable analog of lactose, which induces transcription of the lac operon

lacZ — part of the *lac* operon in *E. coli*, this gene encodes the enzyme β -galactosidase. For decades, molecular biologists have used the *lacZ* gene as a target site for inserting DNA sequences because the resulting bacterial colony color indicates whether insert was successful.

Non-homologous end joining (NHEJ) — DNA repair mechanism in which specific proteins reconnect the ends of a double-strand DNA break. This process may randomly insert or delete one or more bases that can disrupt gene function or expression.

Protospacer — DNA region targeted for cleavage by the CRISPR system.

Protospacer adjacent motif (PAM) — sequence motif immediately adjacent to the protospacer sequence in the Cas9 recognition sequence that is required for Cas9 function. Cas9 recognizes the PAM sequence 5'-NGG where N can be any nucleotide (A, T, C, or G). When Cas9 binds the PAM, it separates the DNA strands of the adjacent sequence to allow binding of the sgRNA. If the sgRNA is complementary to that sequence, Cas9 cuts the DNA.

Scaffold region — called the trans-activating CRISPR RNA or tracrRNA in nature, a region of sgRNA that forms a multi-hairpin loop structure (scaffold) that binds tightly in a crevice of the Cas9 protein. The sequence of this region is typically the same for all sgRNAs.

Single guide RNA (sgRNA) — engineered form of guide RNA that forms a complex with Cas9; ~100 nucleotide fusion of two regions that occur as separate guide RNAs in nature: the guiding region (crRNA) and the scaffold region (tracrRNA).

X-gal — 5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside, a compound consisting of galactose linked to a substituted indole. Its hydrolysis by β -galactosidase yields an insoluble blue pigment and can be used in bacterial cultures to indicate the presence of active β -galactosidase.