Center for
BioMolecular
Modeling

Cas9 cuts a *Statistically Unique Sequence* in the human genome.

ENGAGING QUESTION: *How often would you expect to create an Eco RI restriction site (…GAATTC…) if you were to start joining nucleotides together in a random sequence?*

- Provide each student with a pile of nucleotides…with equal numbers of As and Ts and Gs and Cs.
- Instruct your students to join these nucleotides together in any random sequence ….in the 5’ to 3’ direction… to form a 20-nucleotide long sequence.

For example: **`ACGGTATATGCCGATCTCGA`**

- Engage your students in a discussion that focusses on the Eco RI restriction site….GAATTC… in the following way:
    - Phase 1…..Teacher*: Look at your sequence. How many of you started your sequence with a G?*
    - Students: Approximately 25% of students raise their hands.
    - Teacher: *That is because you had a 1 in 4 chance of selecting a G to start your sequence. If that is true, how many of you have a G in the first 4 letters of your sequence?*
    - Students: Most students will raise hands.
    - Phase 2….Teacher: *OK,…let’s look for the second letter in the Eco RI sequence. How many of you have a GA somewhere in your sequence?*
    - Students: Most students will raise their hands.
    - Teacher: *That is because your chance of selecting an A to follow your G was also 1 in 4,….and 4 x4=16. Therefore, statistically we would expect that you would have a GA somewhere in your 20 nucleotide sequence…..*

Phase 3…. From here,….the teacher can continue this discussion, adding one nucleotide at a time to the sequence you are looking for…and multiplying the previous number by 4. In other words…

1x4 = 4	1 in 4 chance of having G
4x4 = 16	1 in 16 chance of having GA
4x4x4 = 64	1 in 64 chance of having GAA
x4x4x4 = 256	1 in 256 chance of having GAAT
4x4x4x4x4 = 1024	1 in 1024 chance of having GAATT
4x4x4x4x4x4 = 4096	1 in 4096 chance of having GAATTC

You wouldn’t expect any student to have a GAA in their first 20 nucleotides. Depending on the time you can devote to this activity, you can have the students construct longer and longer sequences….or combine data from other classes.

So,…if a 6 nucleotide long sequence is predicted to occur once in every 4096 random nucleotides, how many times would you expect to find this sequence in the 3.2×10^9 bp human genome? Answer…. $\frac{3.2 \times 10^9}{4.0 \times 10^3} = \sim 8 \times 10^5 = 800{,}000$

In other words, digesting the human genome with the restriction enzyme Eco RI will make a mess of it…..creating a mixture of 800,000 different DNA fragments,…with an average length of ~4000 bp.

And now,…we are ready to appreciate the unique power of CRISPR Cas9. Cas9 can search through the 3.2 billion bp human genome and find a statistically unique sequence,… and cut it ….at that one site.

The final question to ask your students is then,….***How long (i.e., how many nucleotides) is a statistically unique site in the human genome?***

And the answer is,….16…..because $\underbrace{4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4}_{\text{16 nucleotides}} = 4.3 \times 10^9$ *nucleotides*

And so,… this means that the Cas9 endonuclease can recognize a 16 nucleotide long sequence. And unlike a restriction enzyme, Cas9 is programmed to recognize this sequence by a guide RNA. Cas9 will only cut DNA if it finds a sequence that is exactly complementary to a 16 nucleotide long region of its guide RNA.

Cutting the human genome at a single unique site….such as the beta globin gene…is the first step in editing that gene. Check out the CBM's summer course ***The Science and Ethics of CRISPR-Based Genome Editing*** for more information about the events that follow.