

Cas9 Cuts a *Statistically Unique Sequence* in the Human Genome

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 nucleotide 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**

- Focus on the *Eco RI* restriction site....GAATTC....
 - **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: *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: *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: *Your chance of selecting an A to follow your G was also 1 in 4 and $4 \times 4 = 16$. Statistically we would expect you to have a GA somewhere in your 20 nucleotide sequence.*
 - **Phase 3** Teacher: *What is your chance of having Eco RI somewhere in your sequence? Starting with the first Eco RI nucleotide, G multiply each of nucleotides by 4.*

$1 \times 4 = 4$	1 in 4 chance of having G
$4 \times 4 = 16$	1 in 16 chance of having GA sequence
$4 \times 4 \times 4 = 64$	1 in 64 chance of having GAA sequence
$4 \times 4 \times 4 \times 4 = 256$	1 in 256 chance of having GAAT sequence
$4 \times 4 \times 4 \times 4 \times 4 = 1024$	1 in 1024 chance of having GAATT sequence
$4 \times 4 \times 4 \times 4 \times 4 \times 4 = 4096$	1 in 4096 chance of having GAATTC sequence

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.

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: $\frac{3.2 \times 10^9}{4.0 \times 10^3} = \sim 8 \times 10^5 = 800,000$

What would happen when Eco RI digests the human genome? You would have a mess of a mixture of 800,000 different DNA fragments with an average length of ~ 4000 bp.

Now we are ready to appreciate the unique power of CRISPR/Cas9. Cas9 can search through the 3.2 billion bp human genome and a statistically unique sequence and cut it -- at that site.

The next question to ask your students is: ***How many nucleotides long is a statistically unique site in the human genome?***

It's 16. since $4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 = 4.3 \times 10^9$ nucleotides

16 nucleotides

The Cas9 endonuclease can recognize a 16 nucleotide long sequence, 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. If the scientists are knocking out the gene, scientists will rely on the cell's *error-prone* DNA repair system to repair the cut, since there is a high probability the cell won't be able to successfully repair the gene. If the goal is to repair the gene, the cell's homologous DNA repair system will be used to insert a corrective gene into the cut site.

We hope to see you at one of our professional development courses in the future!

<https://cbm.msos.edu/summerCourses/index.php>