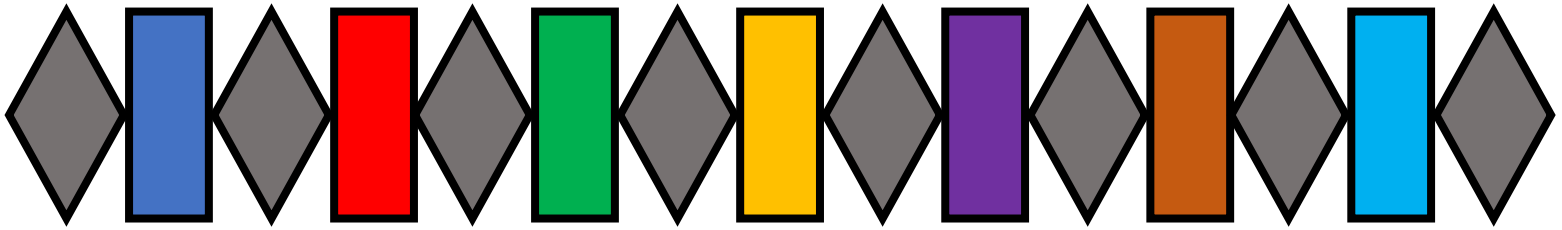
CRISPR: An ancient immune system drives new biotech



Much material adapted from

“Advanced Bioinformatics for Biotechnology”

Video resources

- https://www.youtube.com/watch?v=mXNW_dJotP4
  - Alex Dainis: Stanford genetics blogger

- https://www.youtube.com/watch?v=bXnWIk8FgKc
  - Tessa Montague: Harvard molecular biology PhD student
  - Low-tech pen-&-paper explanations

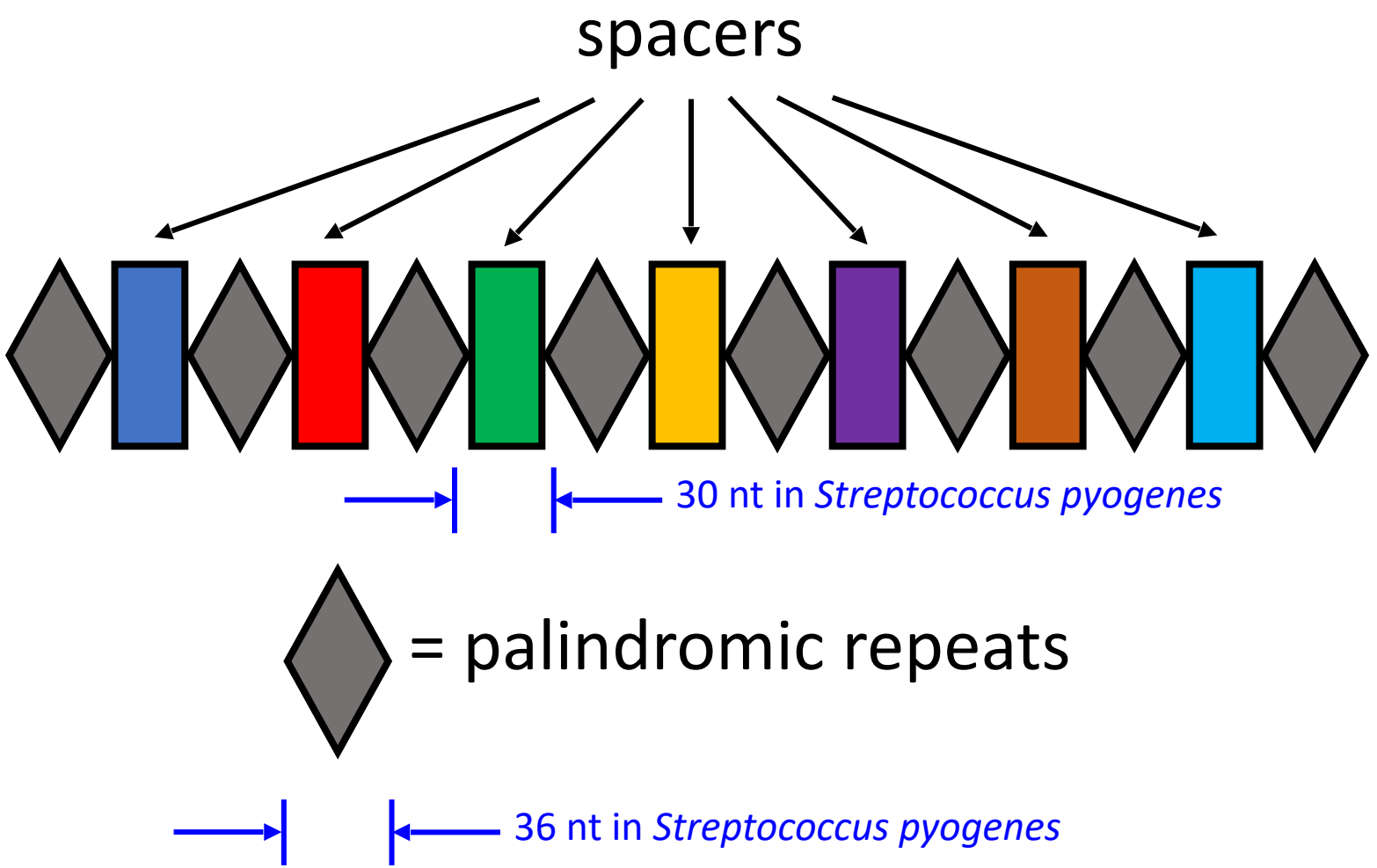
Audio Resource

- https://www.wnycstudios.org/story/antibodies-part-1-crispr
  - Radiolab

The CRISPR timeline

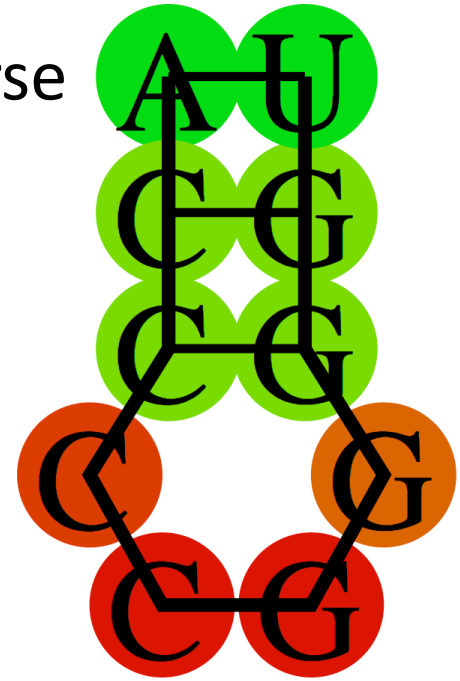
- Billions of years ago: Evolution of CRISPR-Cas system
- 1987: Discovery by humans
- 2002: Name given
- 2007: Role in bacteria identified
- 2012: Genome editing technology
- Tomorrow: today’s knowledge becomes obsolete

Structure of a prokaryotic CRISPR



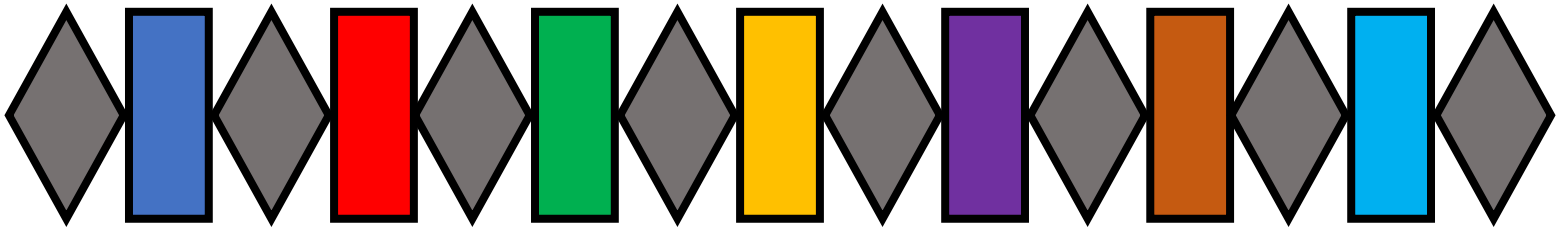
What's a palindrome?

- Civilian: A sentence or phrase that reads the same, forward or backward:
  - Gnu dung
  - Lonely Tylenol
  - Maps, DNA, and spam
- Genetic: A sequence that is its own reverse complement
  - `ACCCGGGT`
  - A palindromic strand can hybridize to itself, forming stem-loop shape



http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi

CRISPR = Clustered Regularly Interspaced Short Palindromic Repeats

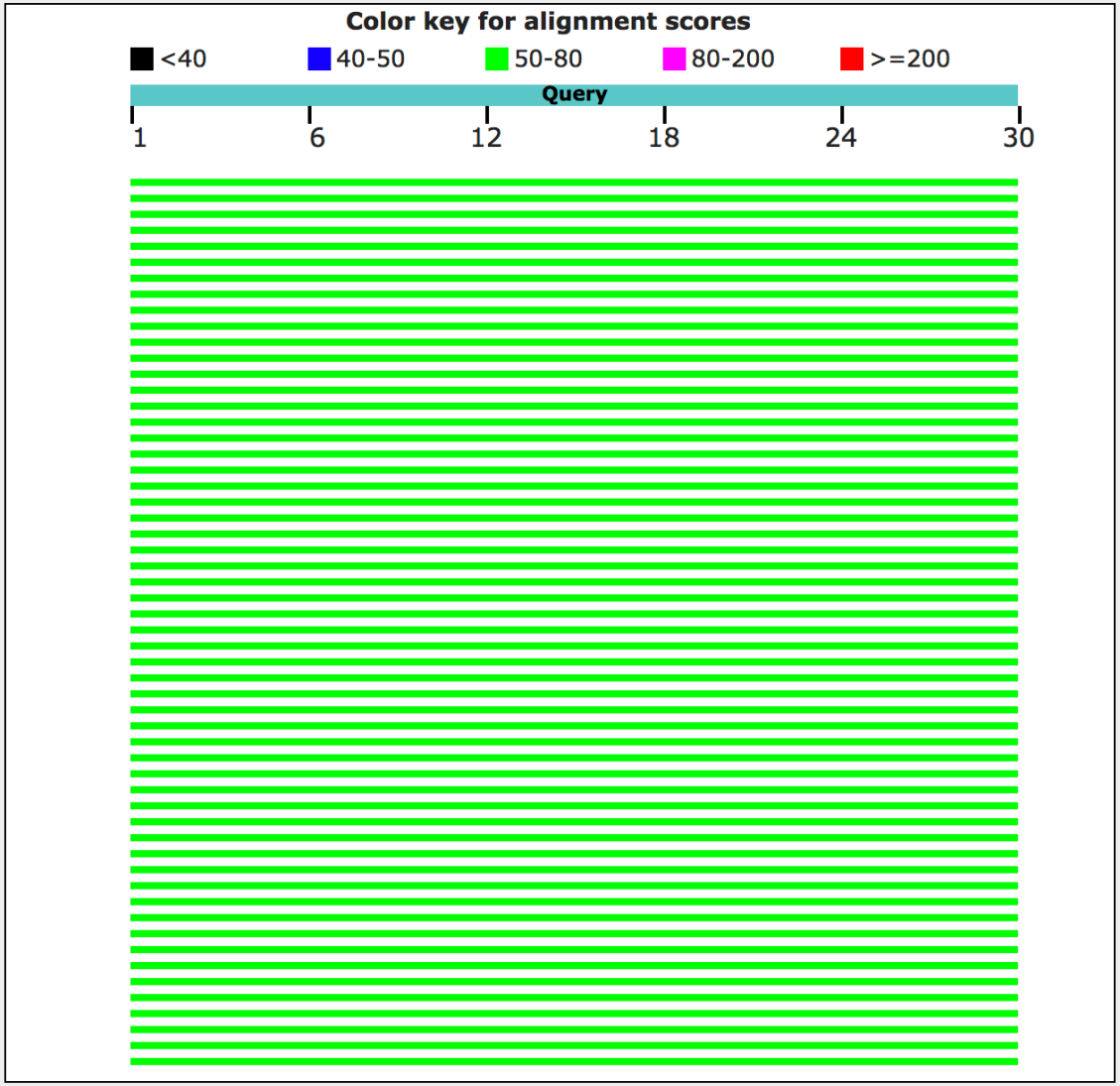


- Named for the repeats
- But the spacers are the magic
- What happens when you BLAST a spacer against GenBank?
- Let's try it!
- TGCGCTGGTTGATTTCTTCTTGCGCTTTTT (*S. pyogenes*)

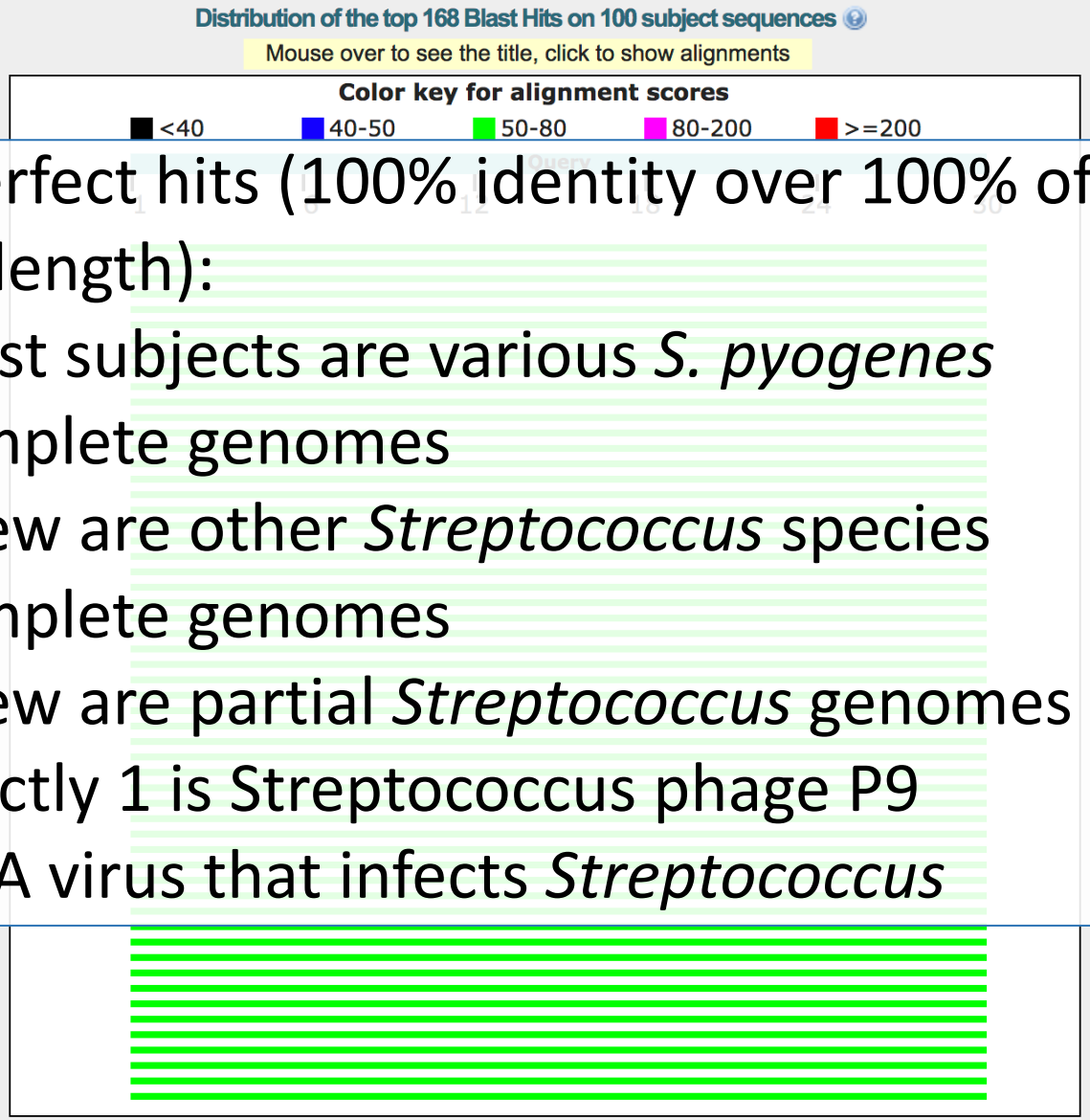
Graphic Summary

Distribution of the top 168 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments



Graphic Summary



~70 perfect hits (100% identity over 100% of query length):

- Most subjects are various *S. pyogenes* complete genomes
- A few are other *Streptococcus* species complete genomes
- A few are partial *Streptococcus* genomes
- Exactly 1 is Streptococcus phage P9
  - A virus that infects *Streptococcus*

The *S. pyogenes* repeat sequence

`GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC`

Looks like a palindrome

Length = 36

Let's align it against a complete *S. pyogenes genome:* *CP031617.1*

Streptococcus pyogenes strain MGAS29409 chromosome, complete genome

Sequence ID: CP031617.1 **Length:** 1894040 **Number of Matches:** 5

Range 1: 1098402 to 1098437 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
67.6 bits(36)	2e-13	36/36(100%)	0/36(0%)	Plus/Minus

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  36
               ||||||||||||||||||||||||||||||||||||
Sbjct  1098437  GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  1098402
```

Range 2: 1098468 to 1098503 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Identities	Gaps	Strand
67.6 bits(36)	2e-13	36/36(100%)	0/36(0%)	Plus/Minus

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  36
               ||||||||||||||||||||||||||||||||||||
Sbjct  1098503  GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  1098468
```

Range 3: 1098534 to 1098569 GenBank Graphics ▼ Next Match ▲ Previous M

Score	Expect	Identities	Gaps	Strand
67.6 bits(36)	2e-13	36/36(100%)	0/36(0%)	Plus/Minus

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  36
               ||||||||||||||||||||||||||||||||||||
Sbjct  1098569  GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  1098534
```

Range 4: 1098600 to 1098635 GenBank Graphics ▼ Next Match ▲ Previous M

Score	Expect	Identities	Gaps	Strand
67.6 bits(36)	2e-13	36/36(100%)	0/36(0%)	Plus/Minus

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  36
               ||||||||||||||||||||||||||||||||||||
Sbjct  1098635  GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  1098600
```

Range 5: 1098339 to 1098371 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Identities	Gaps	Strand
54.7 bits(29)	1e-09	32/33(97%)	1/33(3%)	Plus/Minus

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTC-CCA  32
               |||||||||||||||||||||||||||| |||
Sbjct  1098371  GTTTTAGAGCTATGCTGTTTTGAATGGTCTCCA  1098339
```

The query hits the *S. Pyogenes* genome perfectly, 4 times, and almost perfectly, once

Streptococcus pyogenes strain MGAS29409 chromosome, complete genome

Sequence ID: CP031617.1 **Length:** 1894040 **Number of Matches:** 5

Range 1: 1098402 to 1098437 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
67.6 bits(36)	2e-13	36/36(100%)	0/36(0%)	Plus/Minus

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  36
                ||||||||||||||||||||||||||||||||||||
Sbjct  1098437  GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  1098402
```

66

Range 2: 1098468 to 1098503 GenBank Graphics ▼ Ne

Score	Expect	Identities	Gaps
67.6 bits(36)	2e-13	36/36(100%)	0/36(0%)

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  36
                ||||||||||||||||||||||||||||||||||||
Sbjct  1098503  GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  10
```

66

Range 3: 1098534 to 1098569 GenBank Graphics ▼ Ne

Score	Expect	Identities	Gaps
67.6 bits(36)	2e-13	36/36(100%)	0/36(0%)

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  36
                ||||||||||||||||||||||||||||||||||||
Sbjct  1098569  GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  1098534
```

66

Range 4: 1098600 to 1098635 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Identities	Gaps	Strand
67.6 bits(36)	2e-13	36/36(100%)	0/36(0%)	Plus/Minus

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  36
                ||||||||||||||||||||||||||||||||||||
Sbjct  1098635  GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC  1098600
```

66

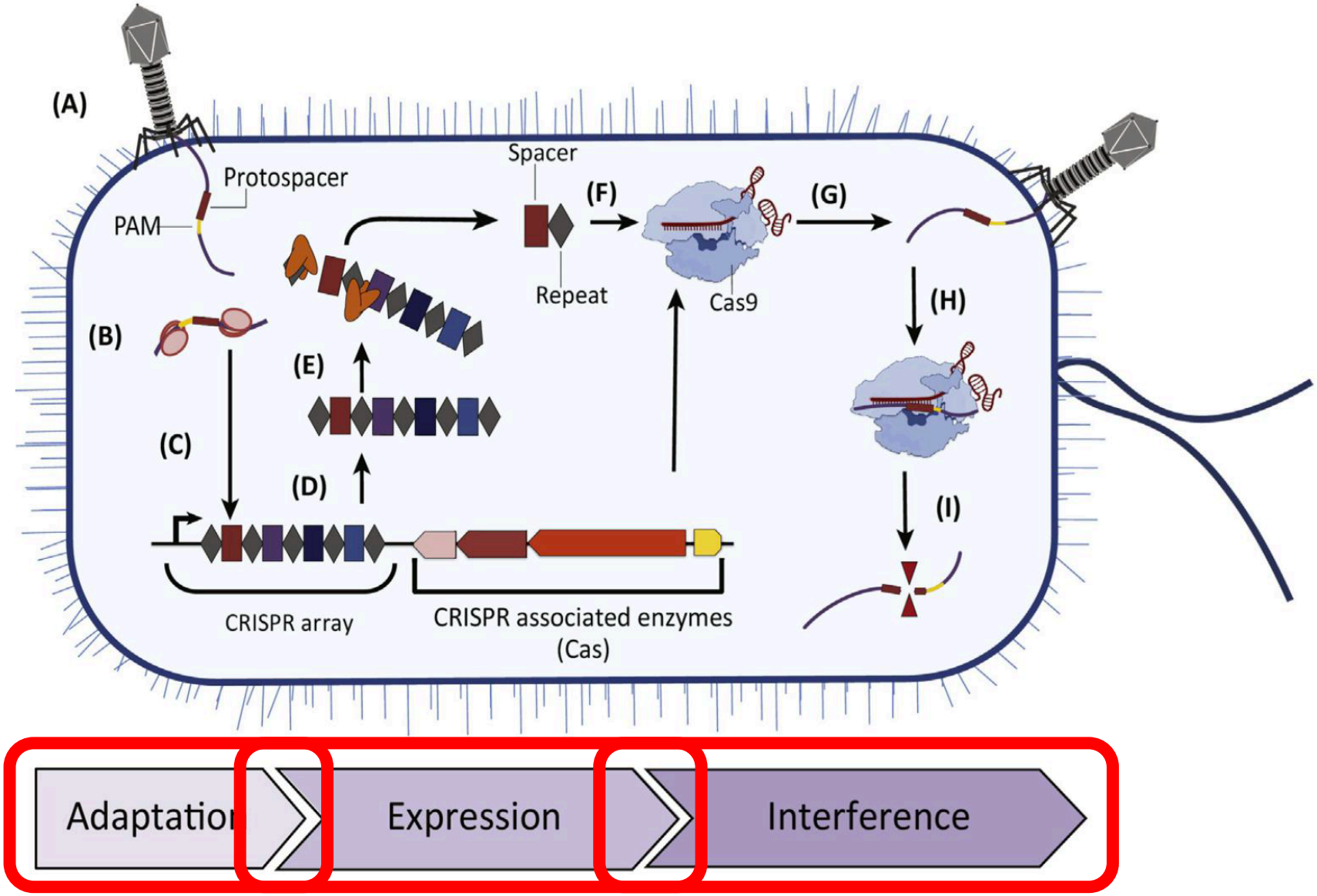
Range 5: 1098339 to 1098371 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Identities	Gaps	Strand
54.7 bits(29)	1e-09	32/33(97%)	1/33(3%)	Plus/Minus

```
Query  1        GTTTTAGAGCTATGCTGTTTTGAATGGTC-CCA  32
                |||||||||||||||||||||||||| |||
Sbjct  1098371  GTTTTAGAGCTATGCTGTTTTGAATGGTCTCCA  1098339
```

Hit coordinates in subjects are all in the same region, evenly spaced

The CRISPR-Cas System: 3 Stages



The CRISPR-Cas System: 3 Stages

Adaption:

- Infection by a virus is survived
- A “wanted poster” describing the virus is added to the CRISPR array

The CRISPR-Cas System: 3 Stages

Expression:

- CRISPR array is transcribed
- Various Cas proteins are expressed
- Cas proteins modify CRISPR transcript

The CRISPR-Cas System: 3 Stages

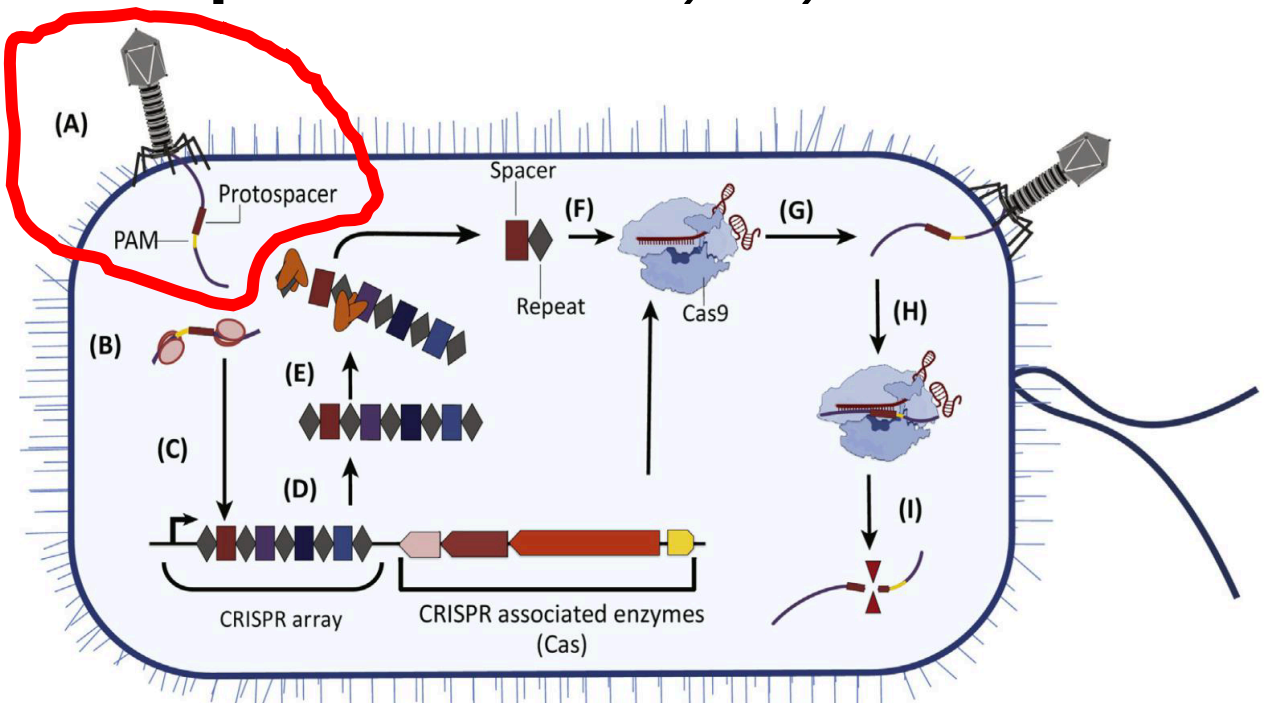
Interference:

- Subsequent infection by same virus
- CRISPR fragment hybridizes to recognized part of viral DNA
- Cas proteins destroy viral DNA

The CRISPR-Cas System: 3 Stages

Let’s break down these steps…

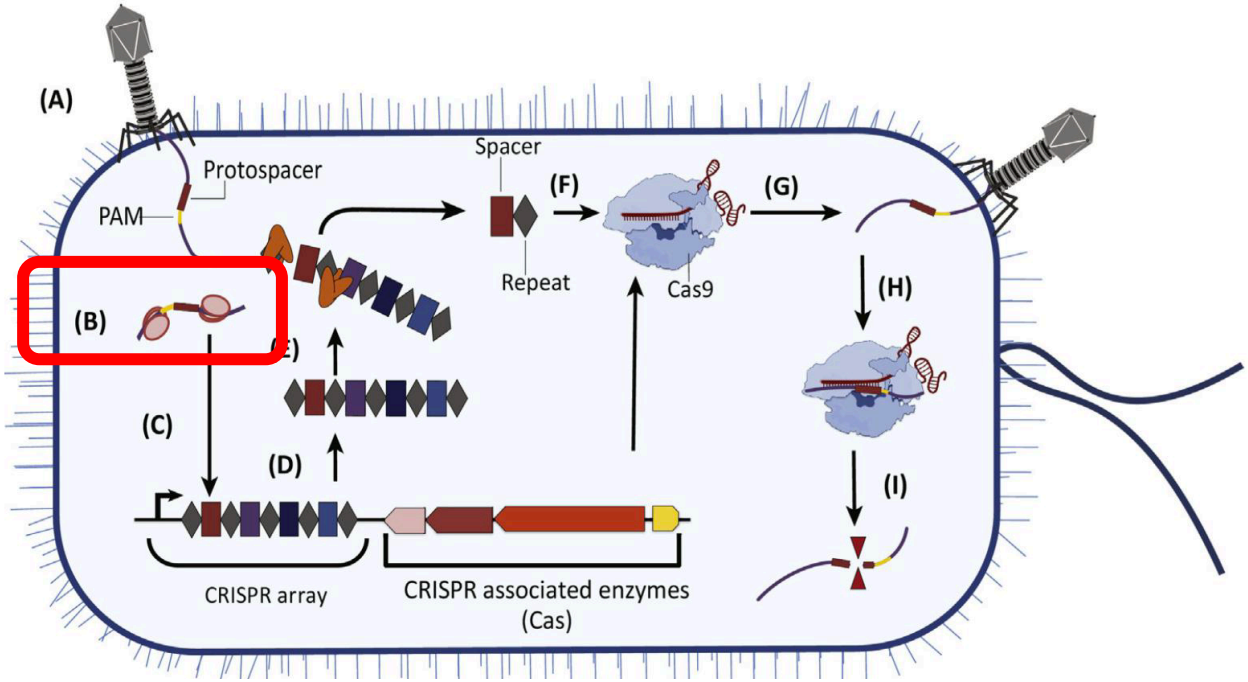
Adaptation: A, B, & C



A: Original infection by virus

- Protospacer is a subsequence of the viral genome
- Protospacer is followed by a "PAM" sequence (Protospacer Adjacent Motif) = NGG in *S. Pyogenes*

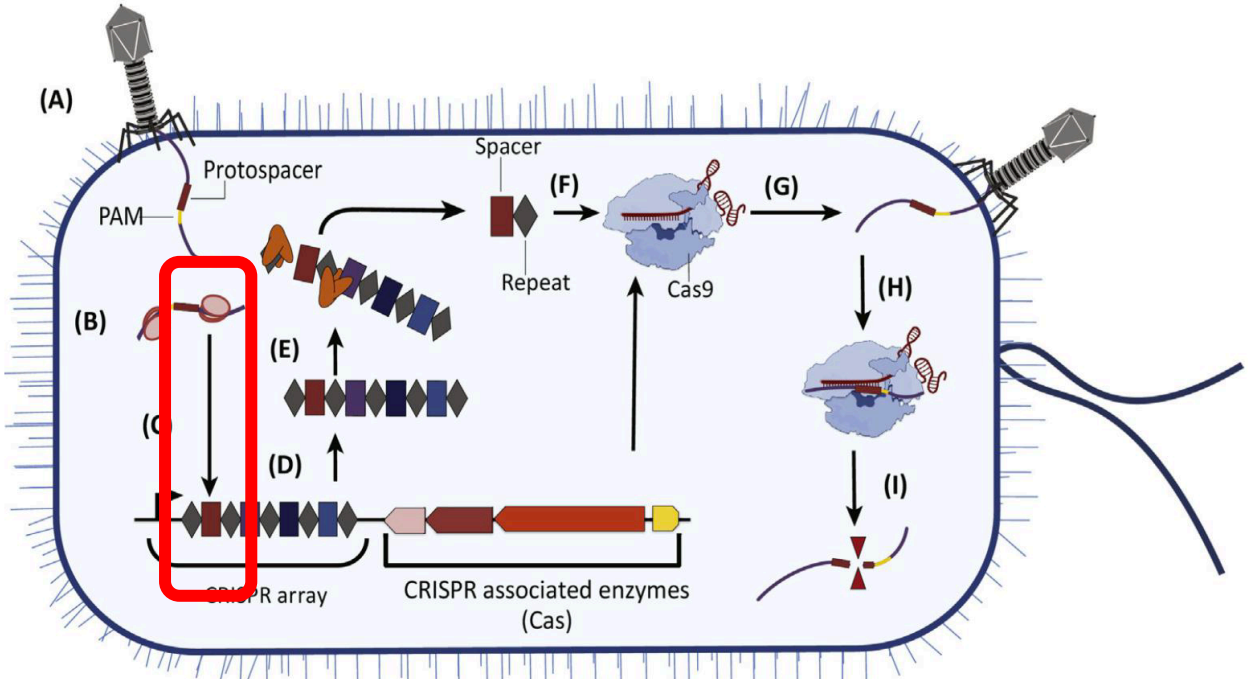
Adaptation: A, B, & C



B: Cas proteins cut viral DNA

- Constant length segment (constant for host species)
  - 30 nt in *S. Pyogenes*
- Segment ends just before (5’ of) PAM

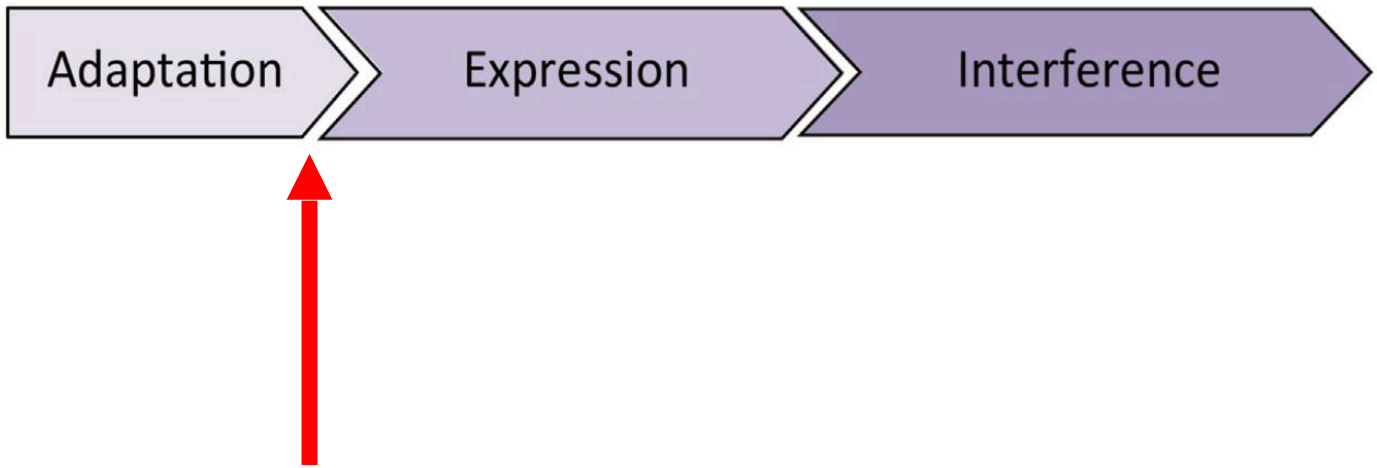
Adaptation: A, B, & C



C: Protospacer is incorporated into the CRISPR array

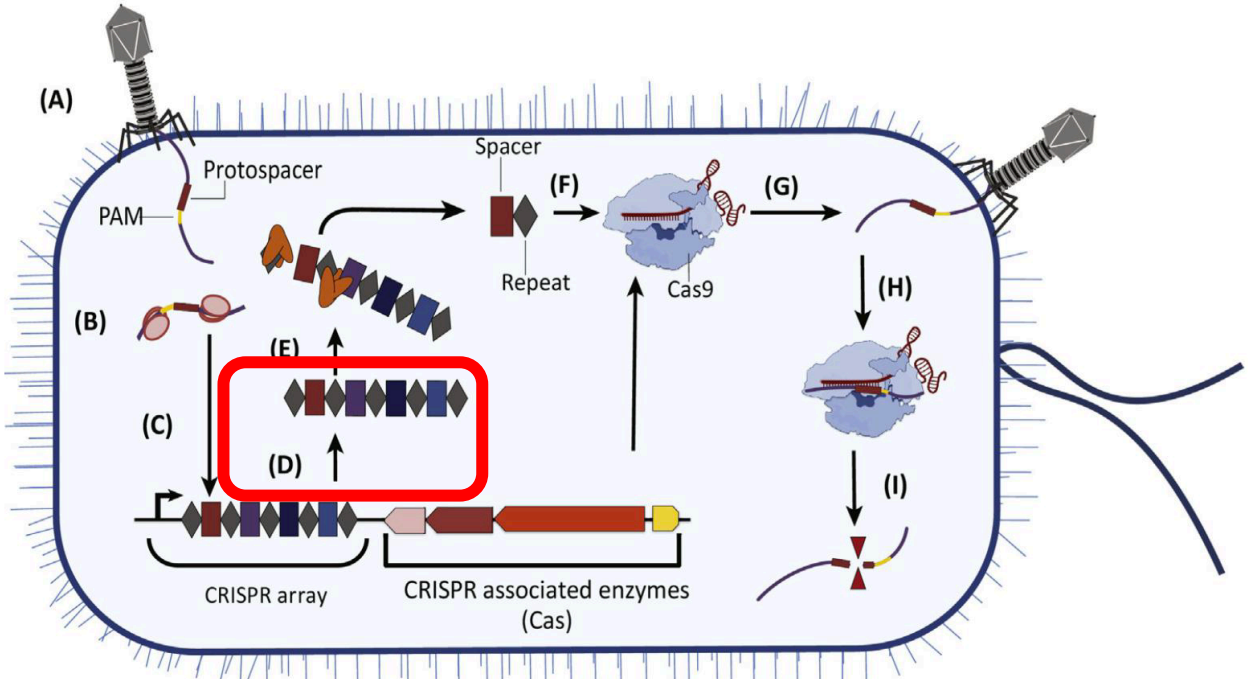
- Protospacer becomes spacer
- Added at 5’ end
- Then a repeat is added at 5’ of the new spacer
- The work is done by various Cas proteins

The CRISPR-Cas System: 3 Stages



Cell is now protected against the virus strain

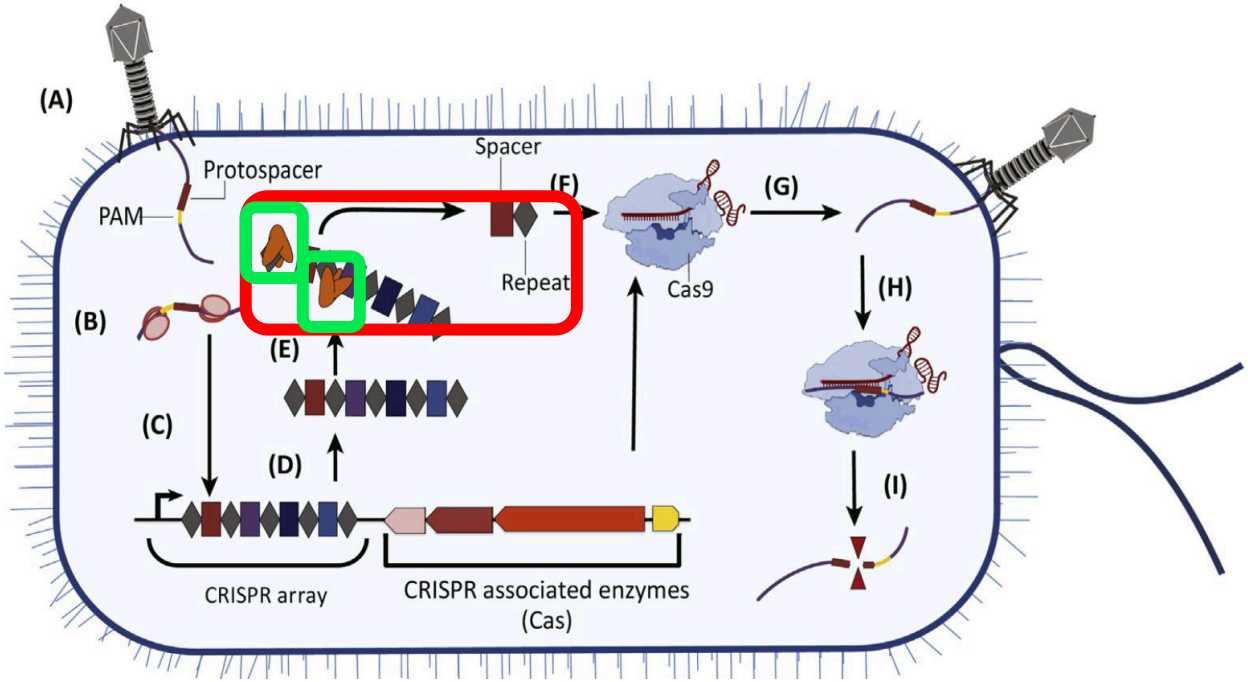
Expression: D, E, F, & G



D: CRISPR array is transcribed

- A single RNA transcript
- It will not be translated

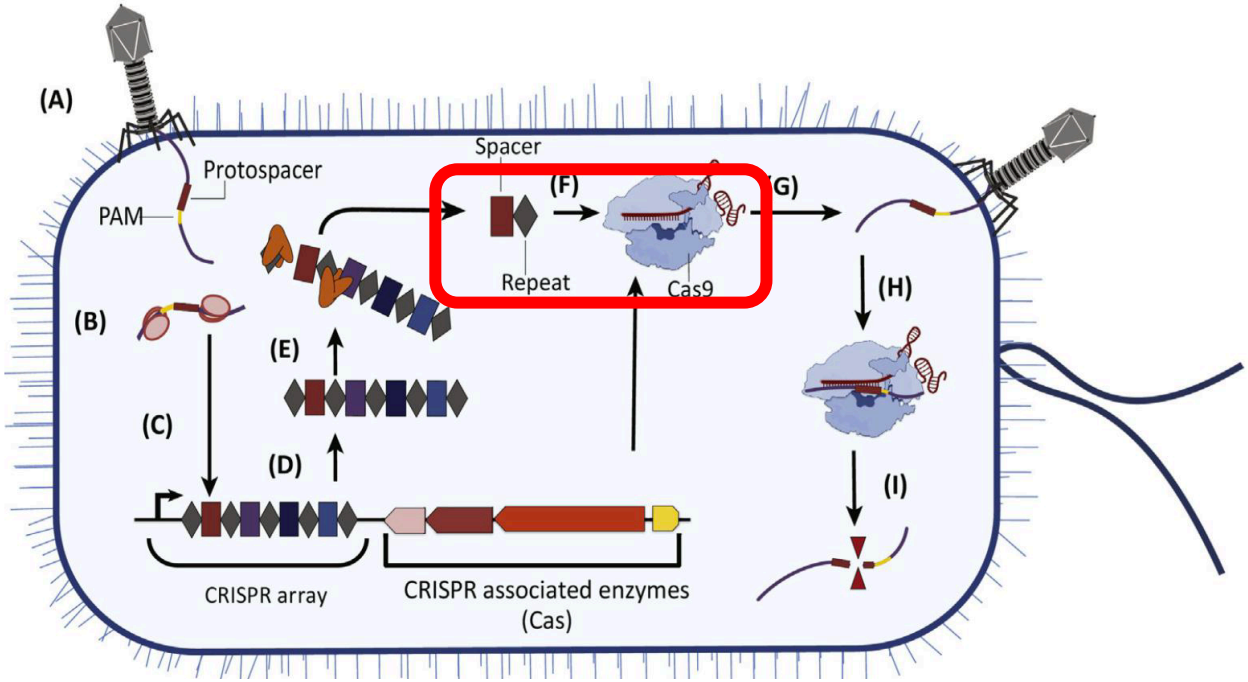
Expression: D, E, F, & G



E: Transcript is cut into spacer-repeat elements

- Elements are called crRNA 
- The work is done by various Cas proteins

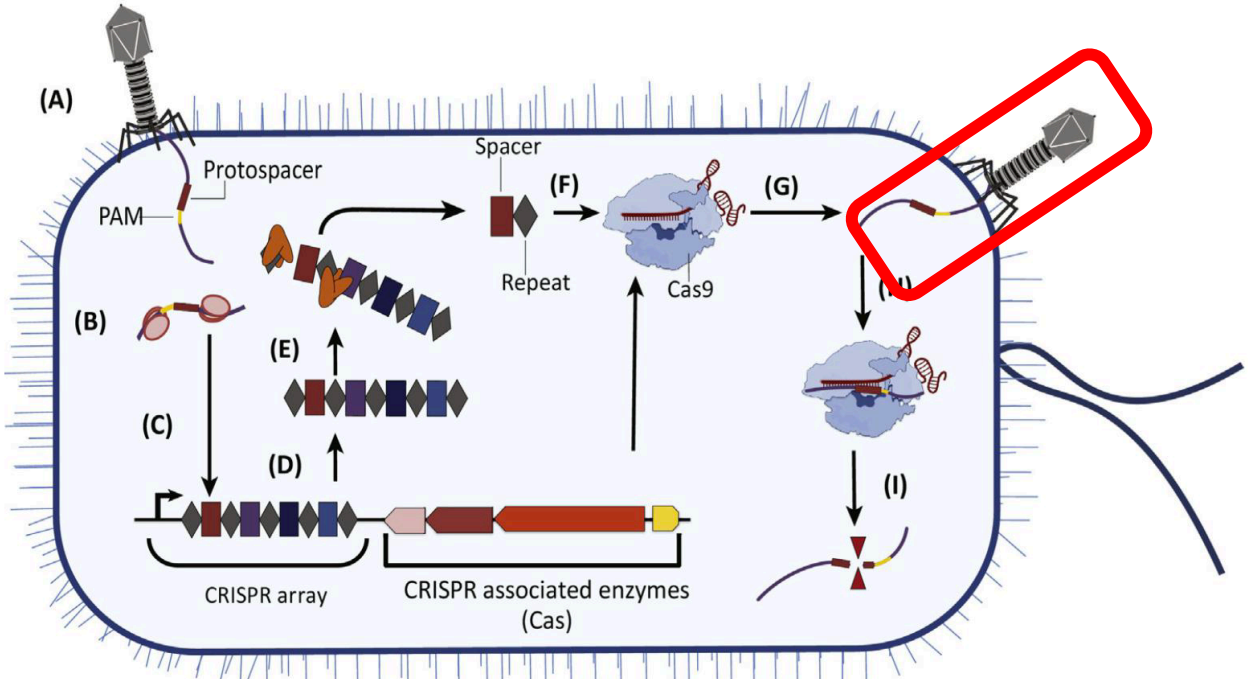
Expression: D, E, F, & G



F: Formation of Cas:crRNA complex

- The Cas part is another Cas protein
- The crRNA is the spacer-repeat from the previous step
- Complex can both recognize and destroy DNA containing the sequence Spacer-PAM
- Doesn't attack cell's own CRISPR array (no PAM)

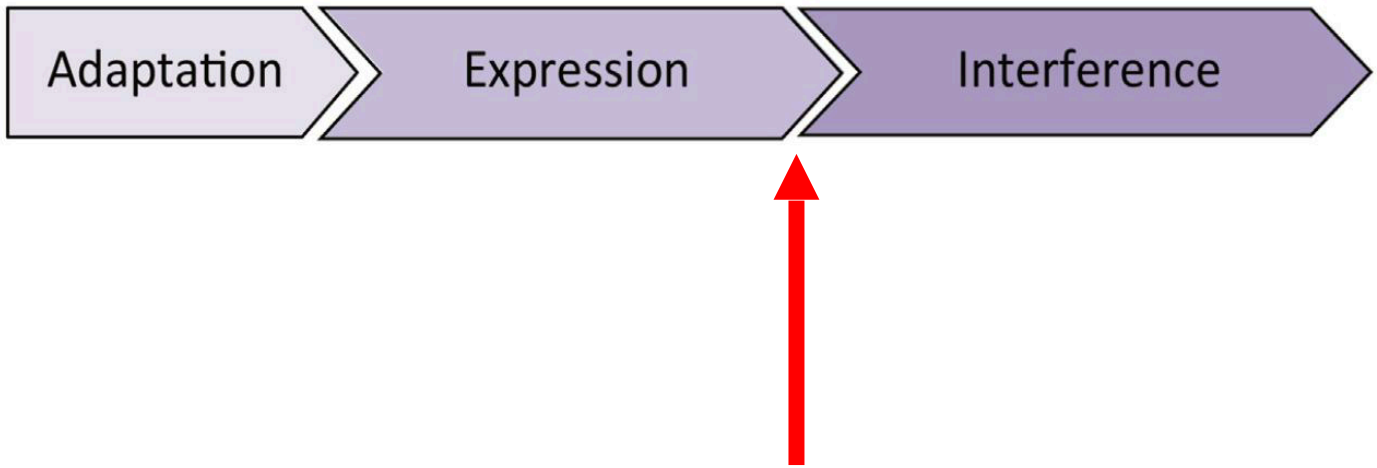
Expression: D, E, F, & G



G: Infection (again)

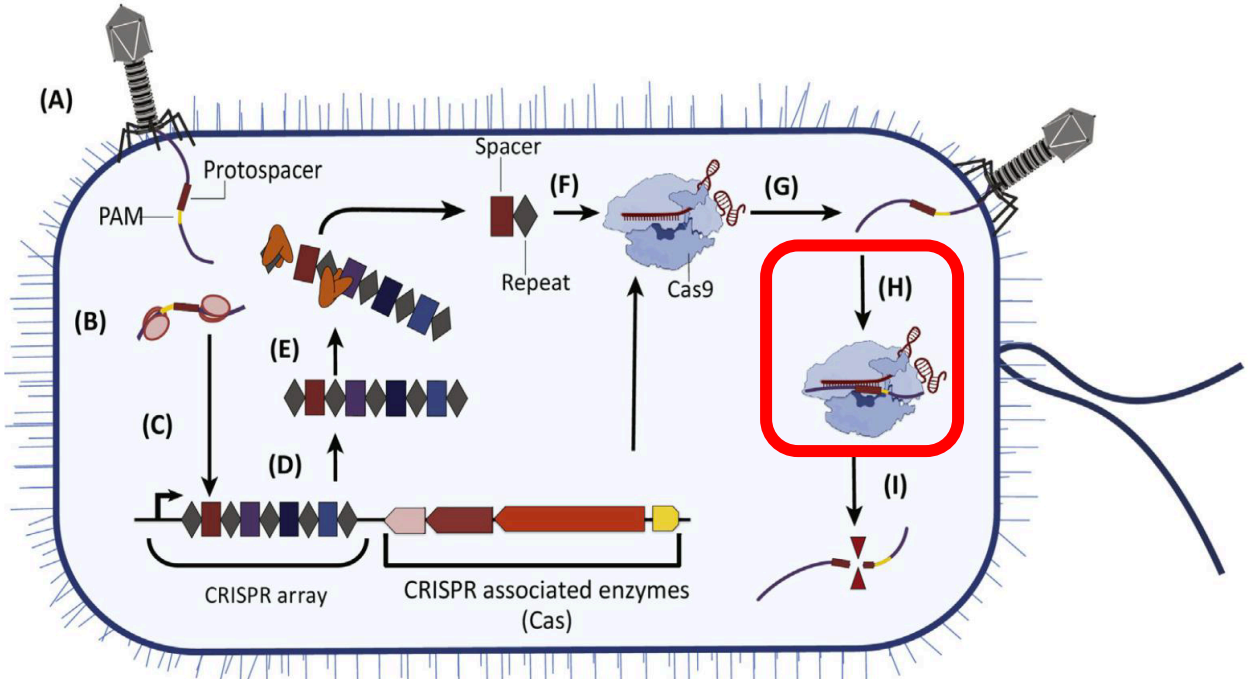
- Virus is same strain as (A)
- Or similar
- Or any strain containing spacer-PAM where spacer is in this cell's CRISPR array

The CRISPR-Cas System: 3 Stages



Invader is about to get a nasty surprise

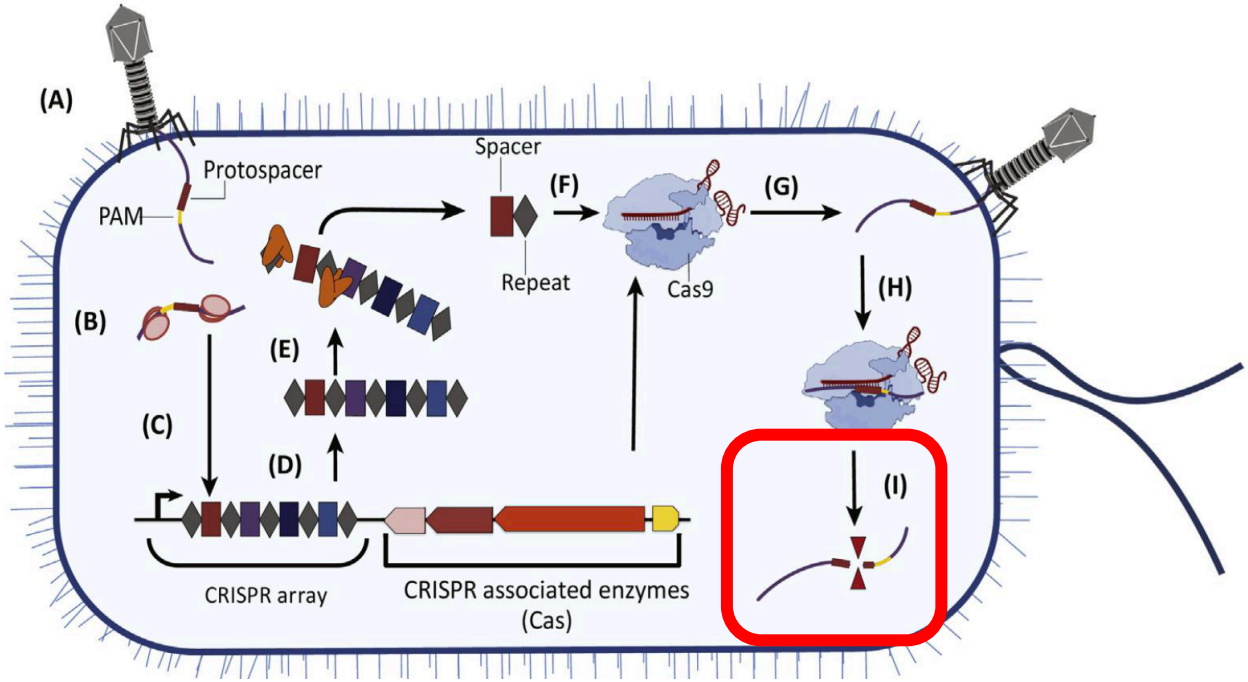
Interference: H & I



H: Cas:crRNA complex binds to invader

- Complex finds spacer-PAM sequence in invader
- Spacer hybridizes to its reverse-complement in invader
- Stem-loop shape of repeat stabilizes components

Interference: H & I



I: Cas:crRNA complex cuts invading DNA

- Invading DNA is no longer viable

The importance of the PAM motif

- Present in the invading virus
- Not in the CRISPR array
- From the cell's point of view:
  - Spacer-PAM means invader
  - Spacer-No-PAM means self
- What if cell's genome (not in the CRISPR array) contains Spacer-PAM in some gene?
  - When cell's CRISPR array acquires the spacer, cell is able to destroy its own chromosome
  - This eventually happens
  - Cell dies → natural selection against such a genome

The big insight

- By 2007, CRISPR (more properly, CRISPR-Cas) was understood to be a bacterial immune system.
- Jennifer Doudna (U.C. Berkeley) realized that the CRISPR-Cas system's ability to target and cut DNA could be leveraged as a tool for gene editing.
  - In prokaryote cells, cut DNA remains cut.
  - Eukaryotes have DNA repair mechanisms, which can be made to insert DNA at the break site.
  - To knock out a gene, insert any single nucleotide ➔ frame shift.
  - To add an entire gene, cut between any 2 genes and insert the new gene.
- Most gene edits are possible using CRISPR-Cas technology.

A nice finale

- ”Within 5 years, Jennifer Doudna will win a Nobel Prize.” – Sami Khuri, 2017
- 2020: Jennifer Doudna wins Nobel Prize in Chemistry