CS123A Bioinformatics

Module 5 – Week 16 – Presentation 1

Leonard Wesley

Computer Science Dept

San Jose State Univ

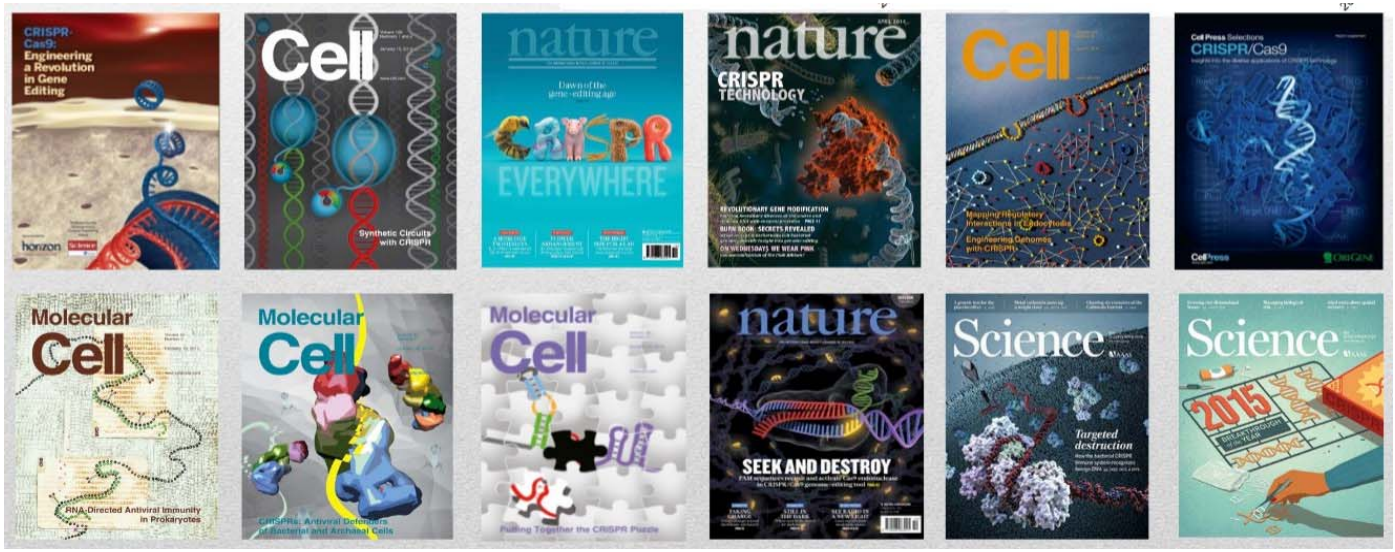
Agenda

- Introduction To CRISPR
  - Background & History
  - Bacteria & The Immune System Of Bacteria
  - What Is CRISPR & How It Works

Introduction To CRISPR

The CRISPR Craze

- CRISPR: *Clustered Regularly Interspaced Short Palindromic Repeats*



- 100's of publications; 1000's of citations
- Supplanting other technologies (TALEN, ZFN)
- Scientific democratization (easy and inexpensive)

Industrial Interest In CRISPR

CRISPR THERAPEUTICS

editas MEDICINE

Caribou Biosciences

- Commercial products (agriculture, biofuels, food, drug development, etc.)
- Large venture capital investments
- All scales (start-ups, mid-size, large, non-profits)

2016 Canada Gairdner International Awards

Emmanuelle Charpentier (Germany/Sweden), Jennifer Doudna (USA), and Feng Zhang (USA) for development of CRISPR-CAS as a genome editing tool for eukaryotic cells

Rodolphe Barrangou (USA) and Philippe Horvath (France) for establishing and characterizing CRISPR-Cas bacterial immune defence system

BROAD INSTITUTE

320 Charles Street

Cambridge, MA 02141

105 Broadway

Cambridge, MA 02142

75 Ames Street

Cambridge, MA 02142

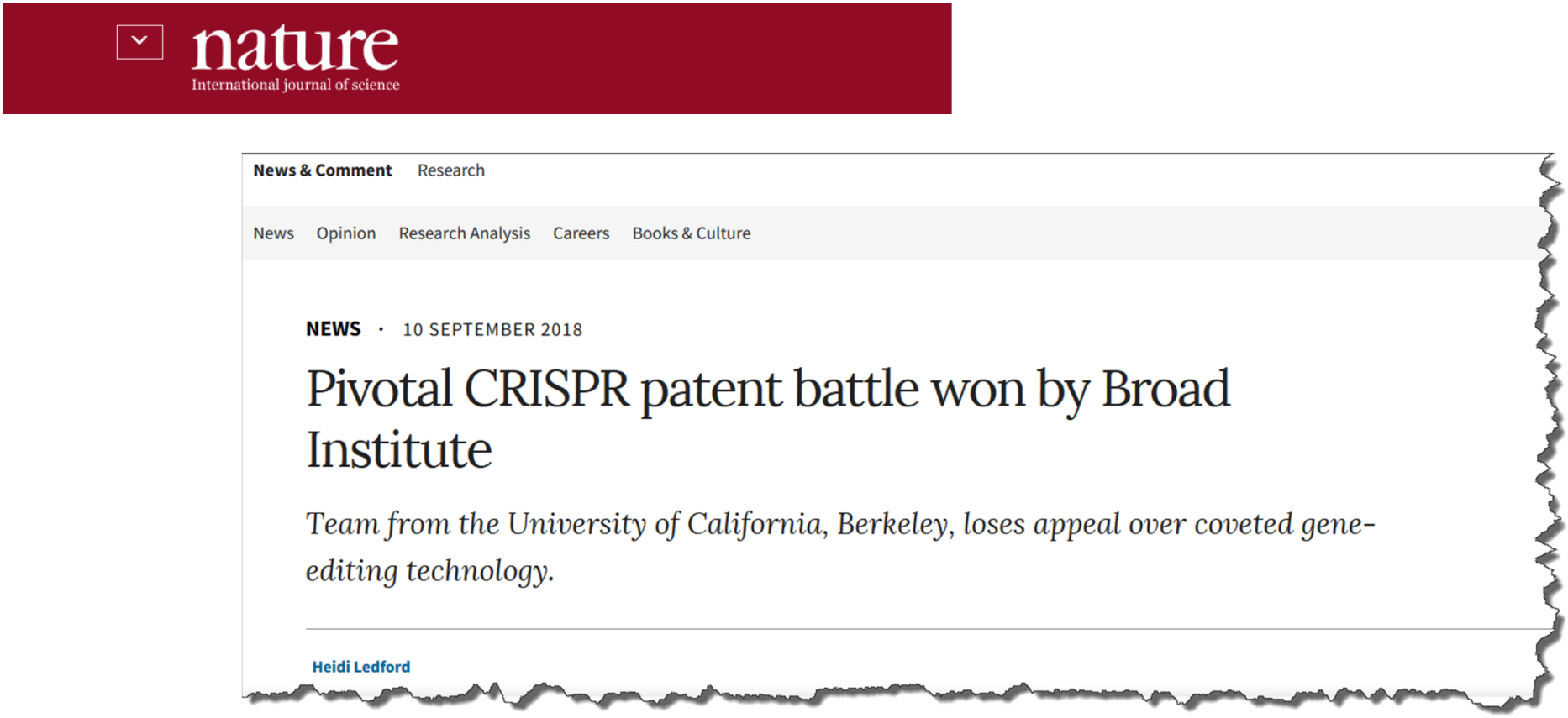
415 Main Street

Cambridge, MA 02142

Major CRISPR R&D Centers

CRISPR and Berkeley: The discovery of the century

The CRISPR Patent Battle Is Over



CRISPR: In The News

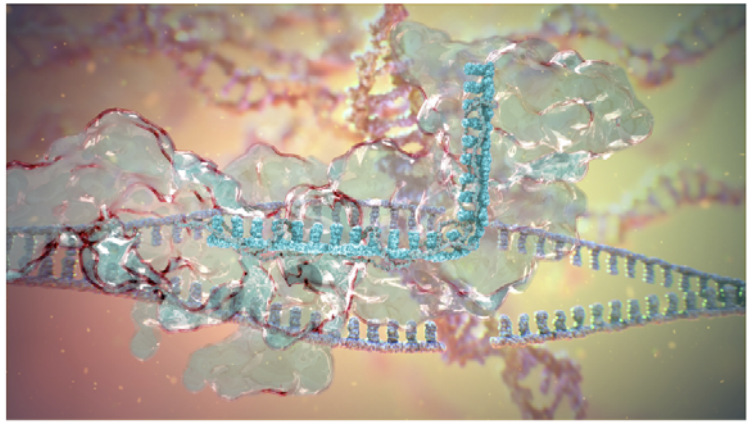
NEWS / 10.22.18

Machine learning tool predicts peptides' potential as immune activators

A deep neural network algorithm called BOTA uses bacterial genomes to identify unrecognized bacterial antigens. READ MORE

Intellia's gene-editing ATTR treatment cuts abnormal proteins in monkeys

by Amirah Al Idrus | Oct 19, 2018 10:55am



So far, one dose of Intellia's gene-editing treatment kept circulating TTR protein levels down in monkeys for more than six months. (AstraZeneca)

New research presents a promising new avenue for research into treating genetic conditions during fetal development.

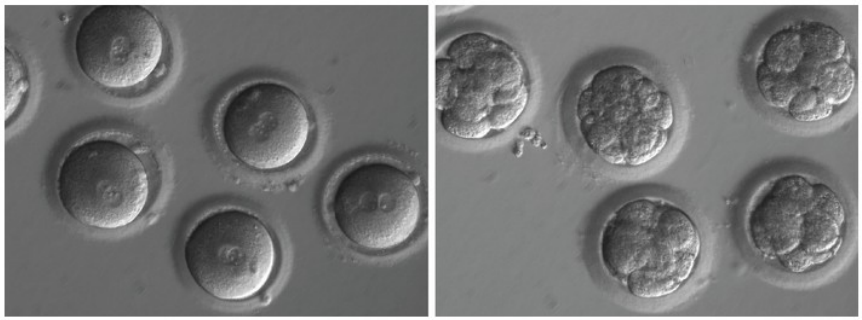
Credit: © Ilhedgehogll / Fotolia

CRISPR In The News *(Cont.)*

In Breakthrough, Scientists Edit a Dangerous Mutation From Genes in Human Embryos

查看简体中文版 | 查看繁體中文版 | Leer en español

By **PAM BELLUCK** AUG. 2, 2017

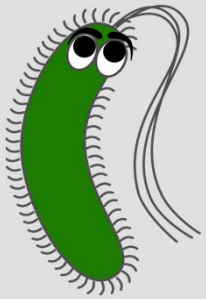


Newly fertilized eggs before gene editing, left, and embryos after gene editing and a few rounds of cell division. A study published on Wednesday announced that edited human embryos can repair common and serious disease-causing gene mutations. Shoukhrat Mitalipov

Gene Editing Spurs Hope for Transplanting Pig Organs Into Humans

By **GINA KOLATA** AUG. 10, 2017

Piglets whose genes were edited to remove retroviruses, which could help clear the way for pig organs to be transplanted to humans. eGenesis

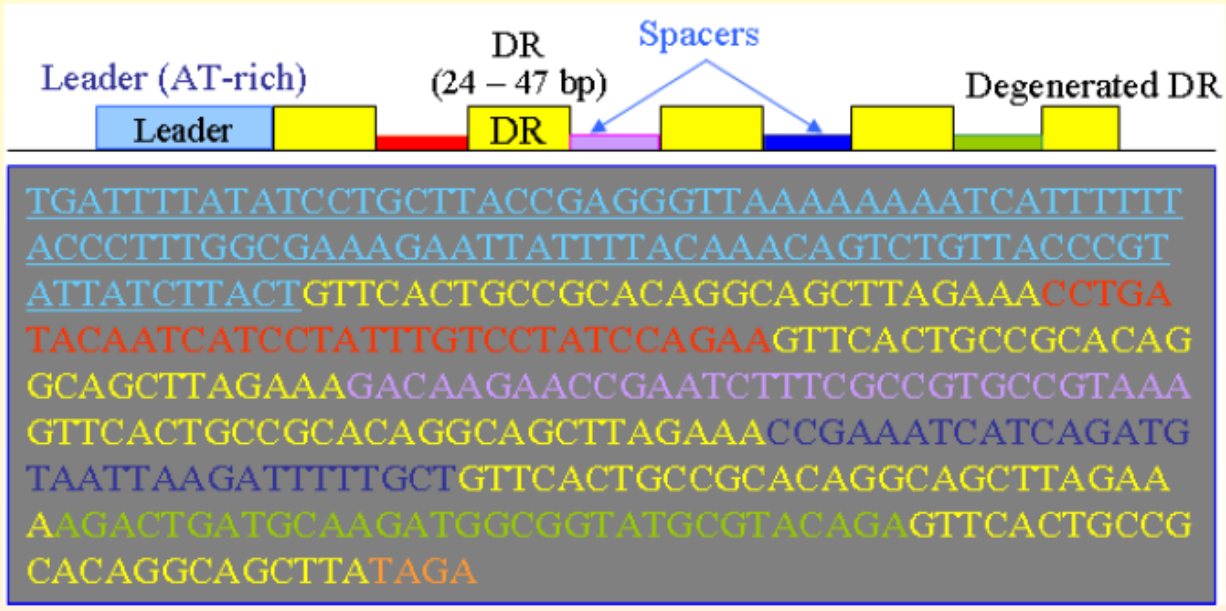


Clustered
Regularly
Interspaced
Short
Palindromic
Repeats

What Is CRISPR?

Clustered Regularly Interspaced Short Palindromic Repeats

CRISPR ARRAY

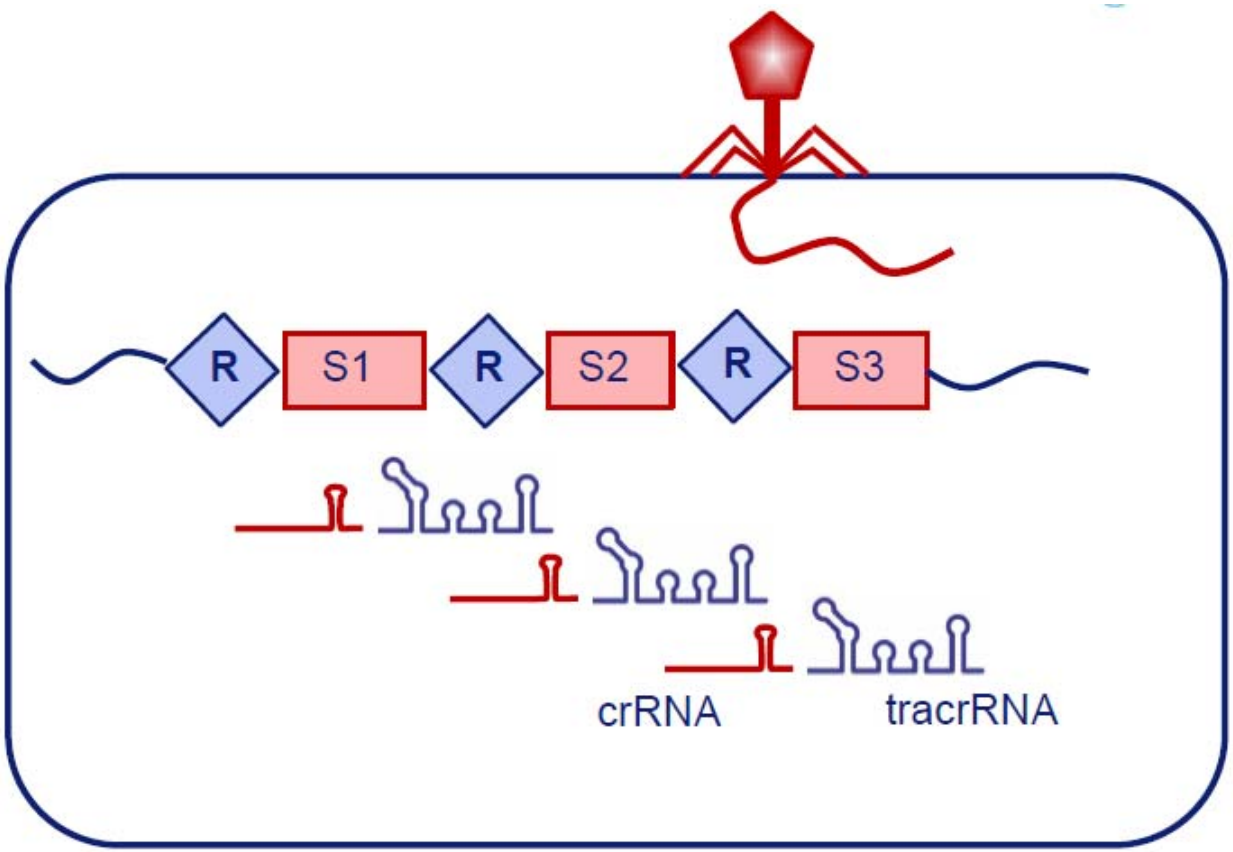


from CRISPRs web server: http://crispr.i2bc.paris-saclay.fr/index.php?page=FAQs

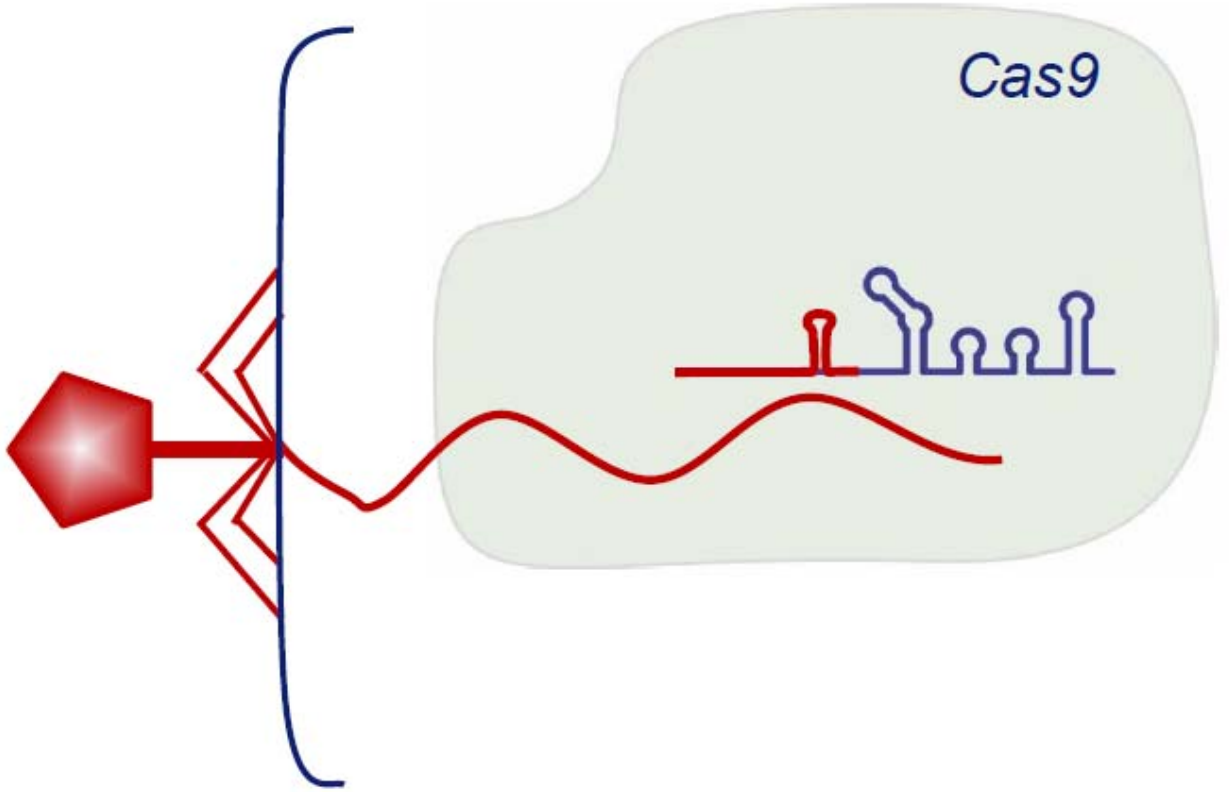
Palindromic

GTTCACTGCCGCACAGGCAGCTTAGAAA (direct repeat)
TTTCTAAGCTGCCTGTGCGGCAGTGAAC (reverse complement)

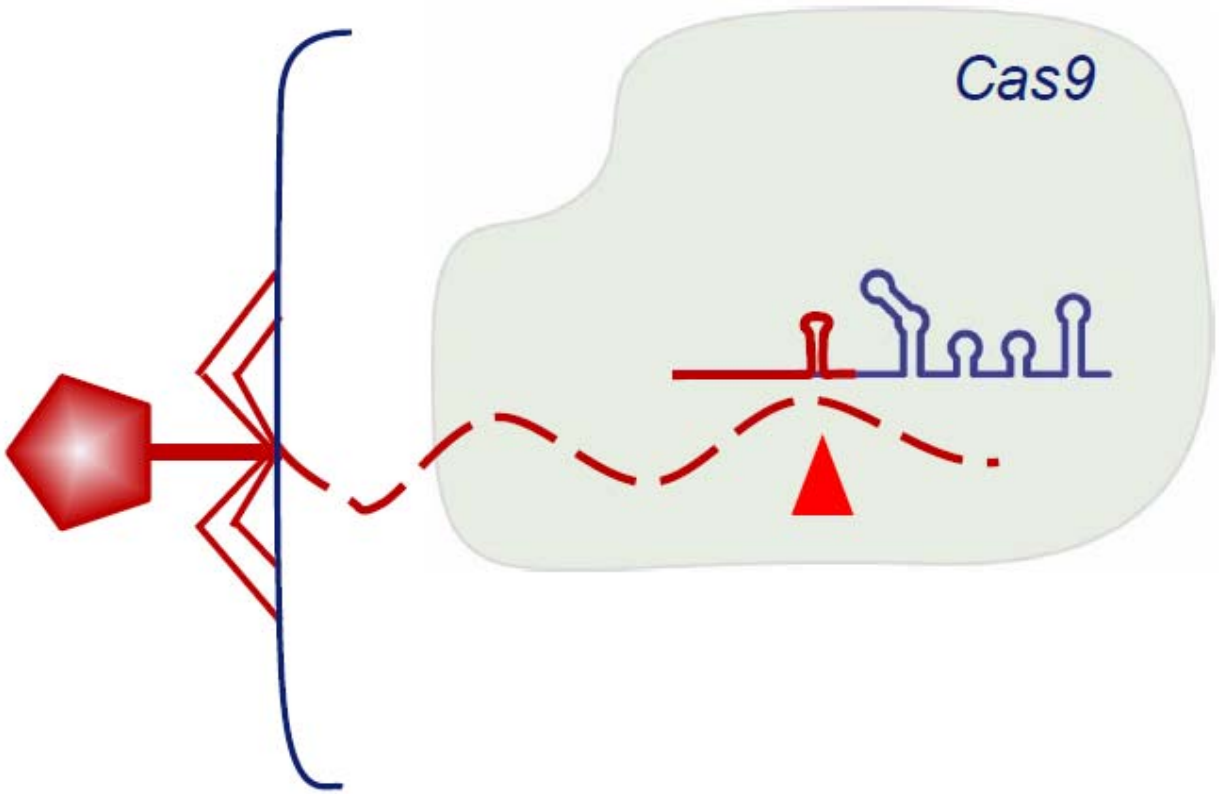
How CRISPR Works



How CRISPR Works *(Cont.)*



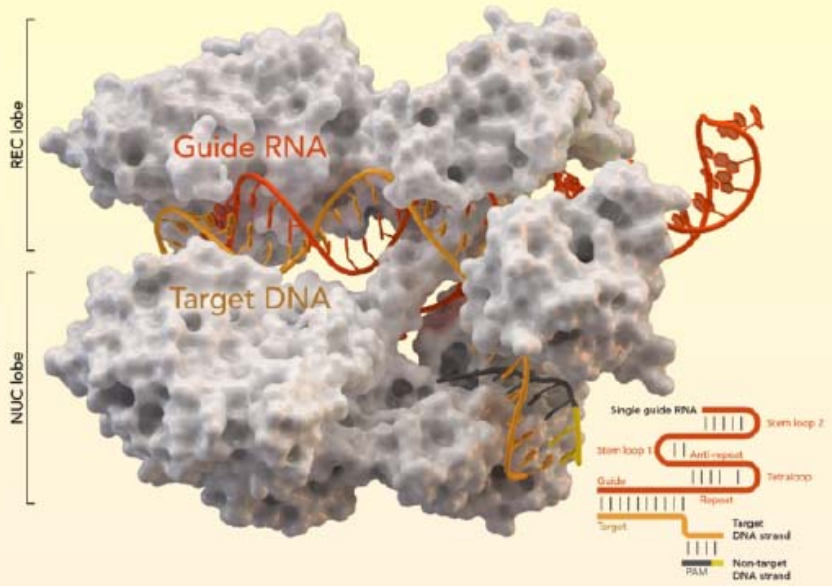
How CRISPR Works *(Cont.)*



How CRISPR Works: Bacterial Immunity

- Adaptation
  - short pieces of DNA homologous to virus sequences are integrated into CRISPR loci
  - repeat is duplicated
- Expression
  - long transcript of a CRISPR locus is generated
  - transcript is processed into short CRISPR RNAs (crRNAs)
- Interference
  - crRNAs guide Cas proteins to the target sequences that match the spacers
  - foreign DNA or RNA is targeted and cleaved

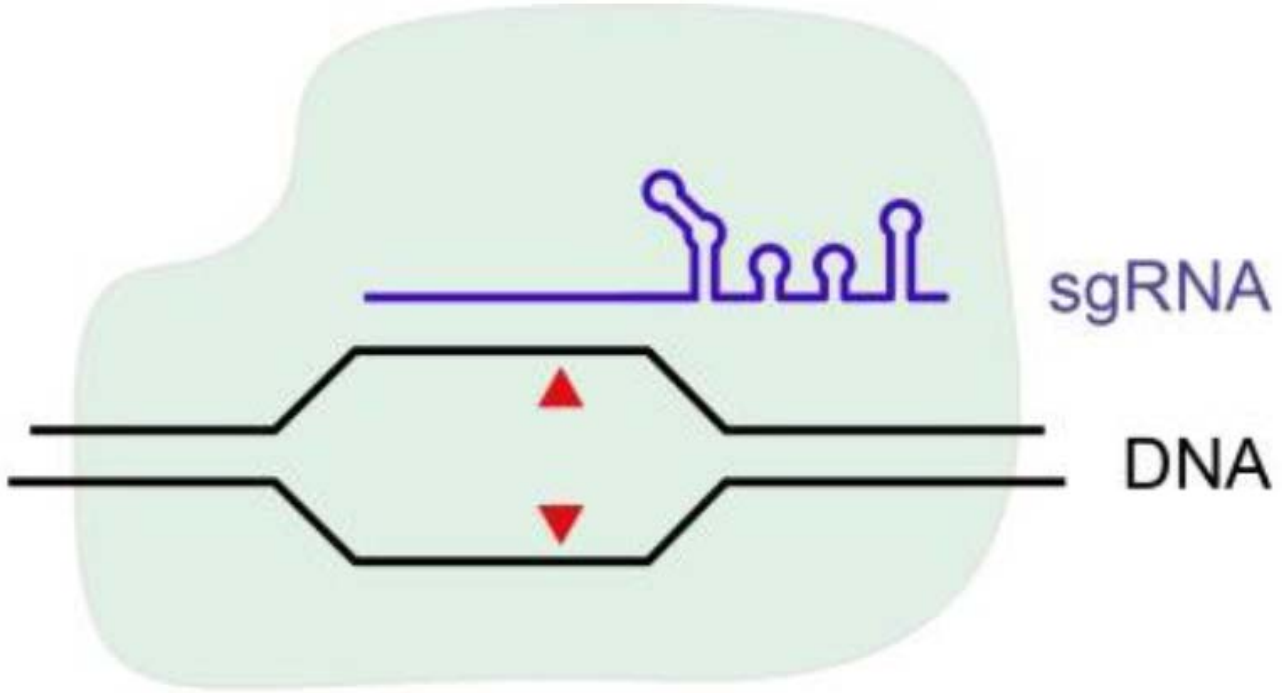
How CRISPR Works: Gene Editing



by Thomas Splettstoesser (www.scistyle.com) [CC BY-SA 4.0] via Wikimedia Commons

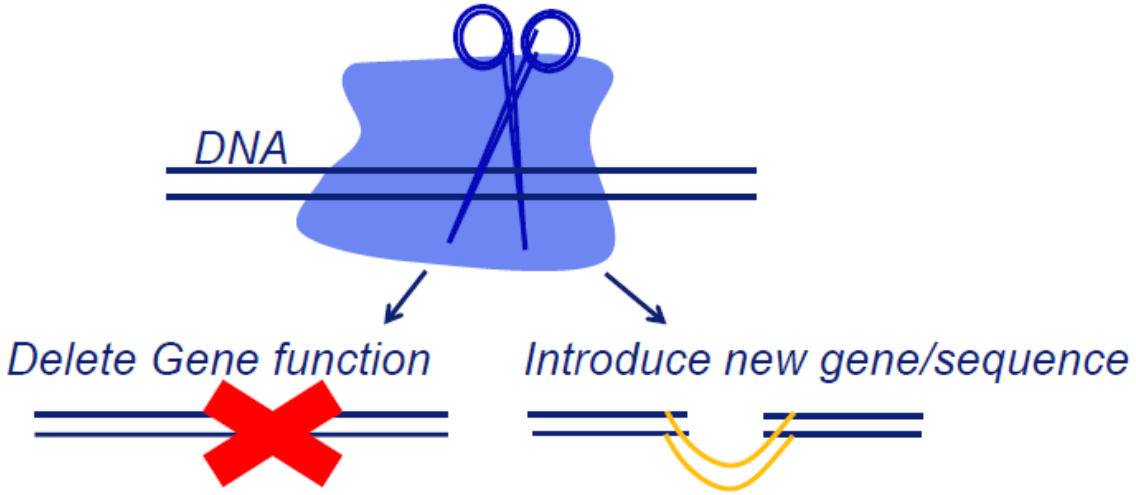
- location for editing is specified by guide RNA
- sequence must be followed by PAM sequence
- Cas protein cuts the DNA
- cut is "repaired" using cell's own machinery

CRISPR/Cas9: A Potential Tool For Gene Editing/Biomed R&D



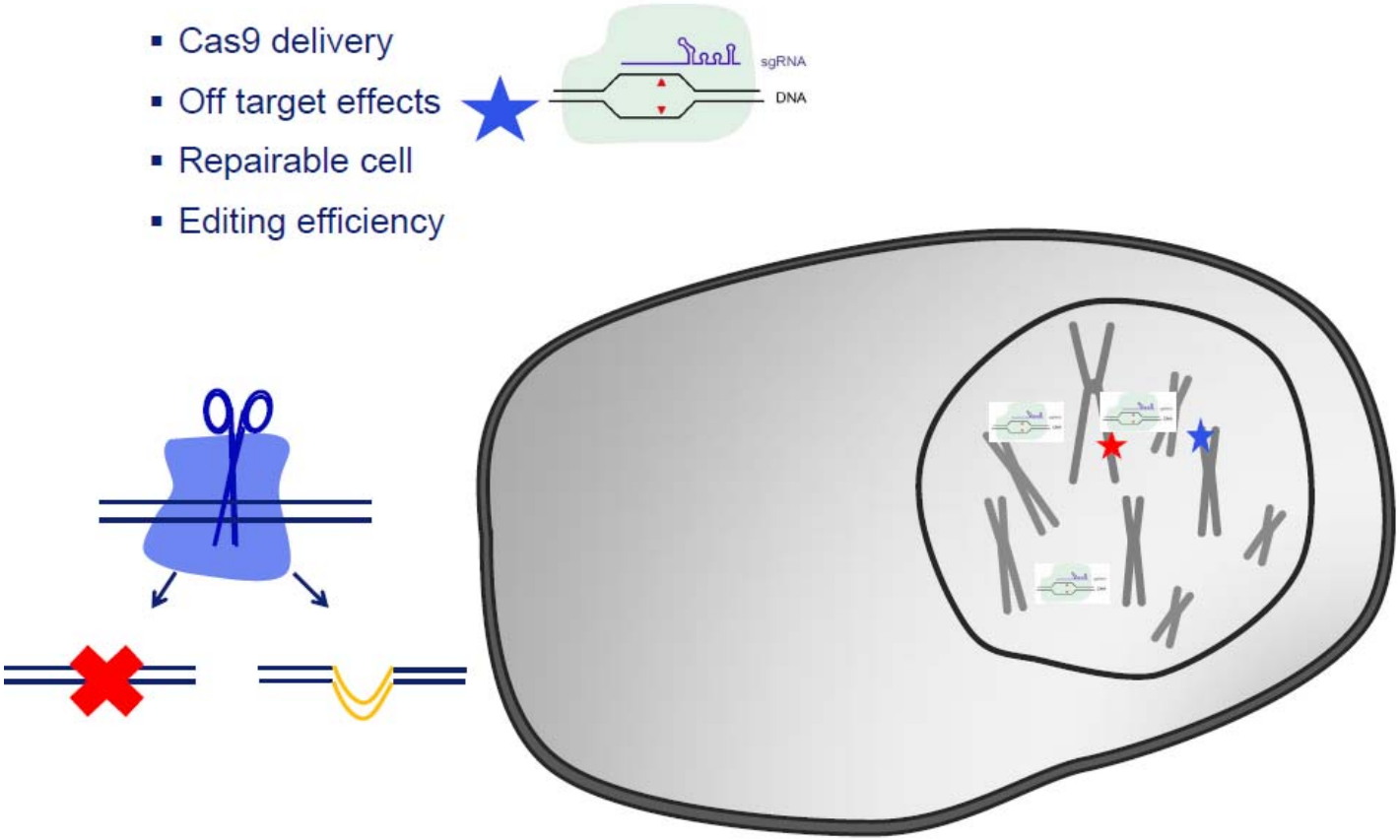
CRISPR/Cas9: Gene Editing *(Cont.)*

- Generation of:
  - Mutations
  - (large) deletions
  - Integrations (reporters, tags)
  - Activation/repression of transcription



What to think of when you design your experiment

- Cas9 delivery
- Off target effects
- Repairable cell
- Editing efficiency



PAM

- Proto-spacer Adjacent Motif
- 2-6 base pair DNA sequence immediately following target DNA
- Cas proteins will not successfully bind to target if it is not followed by the appropriate PAM sequence
- Selection of spacer from viral DNA is determined by location of PAM
- Canonical PAM is 5'-NGG-3'
- PAMs differ between CRISPR-Cas variants

  Bioinformatic problem: Find good CRISPR-Cas variant for gene target

Types Of CRISPRs

Type II : precisely target DNA with Cas9 or Cpf1

Type I: hit and destroy DNA with Cas3

Type III: cleave either DNA or RNA with Cas10

Different PAM Sequences

Species/Variant of Cas9	PAM Sequence
Streptococcus pyogenes (SP); SpCas9	NGG
SpCas9 D1135E variant	NGG (reduced NAG binding)
SpCas9 VRER variant	NGCG
SpCas9 EQR variant	NGAG
SpCas9 VQR variant	NGAN or NGNG
Staphylococcus aureus (SA); SaCas9	NNGRRT or NNGRR(N)
Neisseria meningitidis (NM)	NNNNGATT
Streptococcus thermophilus (ST)	NNAGAAW
Treponema denticola (TD)	NAAAAC
Cpf1 (from various species)	TTN
Additional Cas9s from various species	PAM sequence may not be characterized

Cas

- **Cas** stands for **C**RISPR-**as**sociated.

- Cas is a collection of genes responsible for the multiple stages of CRISPR acquisition/adaptation and interference.

- Editing requires only two components:
  - a Cas nuclease and
  - a programmable guide RNA

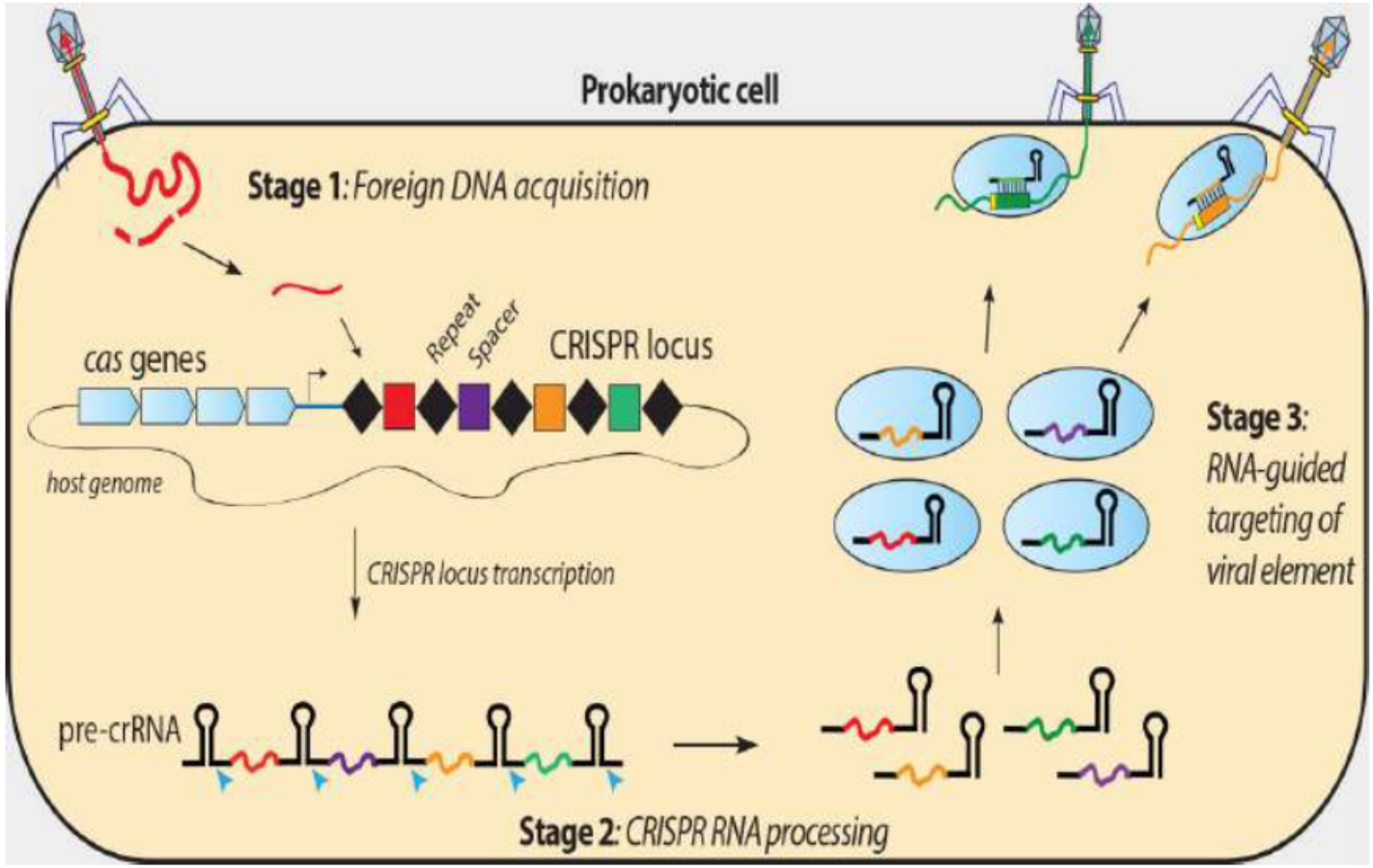
CRISPR-Cas Systems

- CRISPR-Cas systems provide endogenous adaptive immunity in approximately:
  - – 40% of bacterial genomes and
  - – 70% of sequenced archaeal species,

and act against invading genetic elements in a conserved sequence of events:

- – **adaptation**
- – **expression**
- – **interference**

Three Stages of CRISPR-Cas



Some Computational Aspects

Given a DNA Sequence Locate CRISPR Arrays

- SOME TOOLS:
  - **"CRISPRCasFinder** is a web service offering fundamental tools for CRISPR detection, including the shortest ones, allowing an accurate definition of the DR consensus boundaries and extraction of the related spacers."
  - **"CRISPRDetect** is a tool to discover and explore the CRISPR noncoding RNAs in sequence data. It is a bioinformatics tool to find CRISPR arrays."

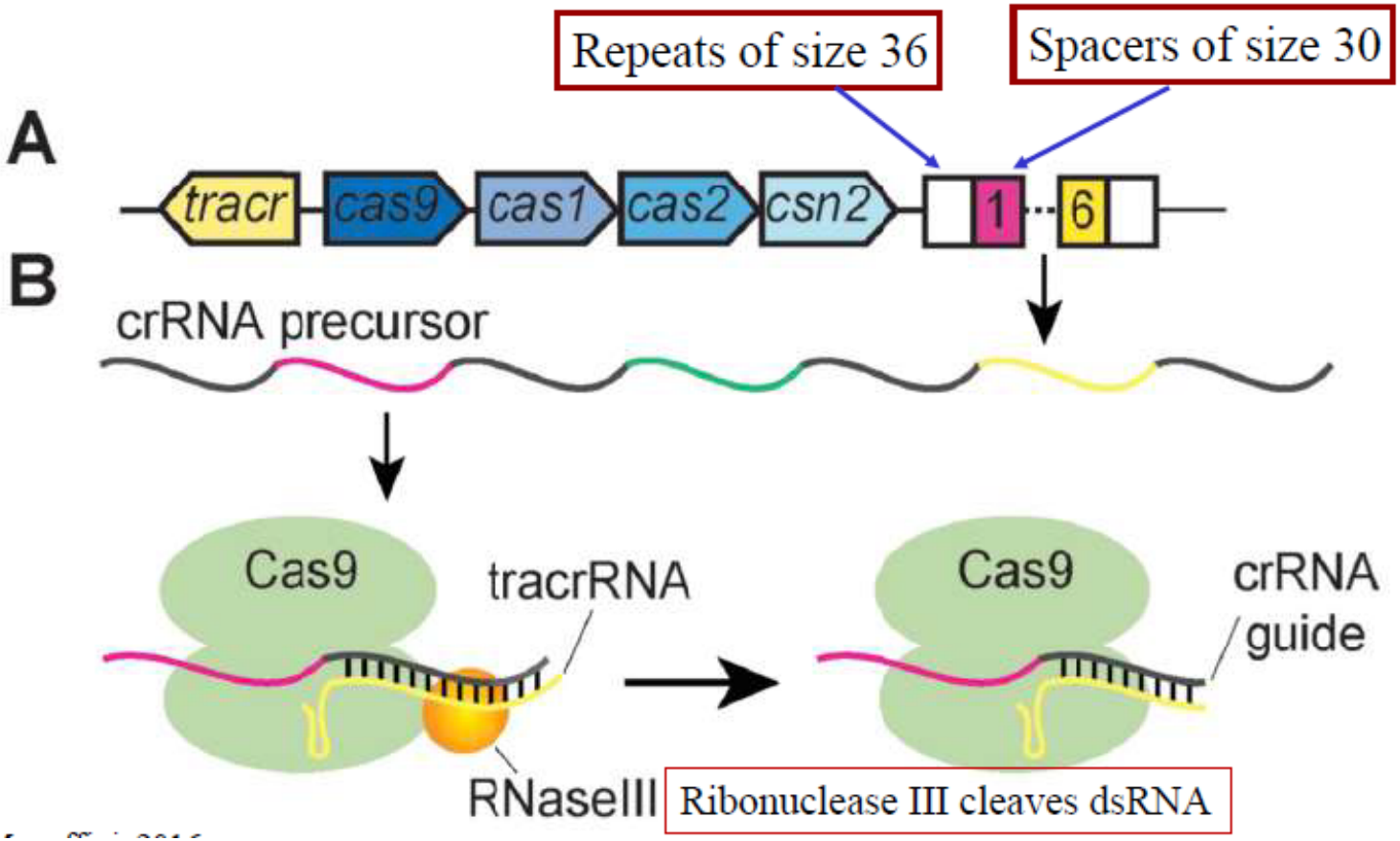
CRISPR ARRAY: Streptococcus pyrogenes

Table 1. CRISPR and prophage content of current *S. pyogenes* genomes.

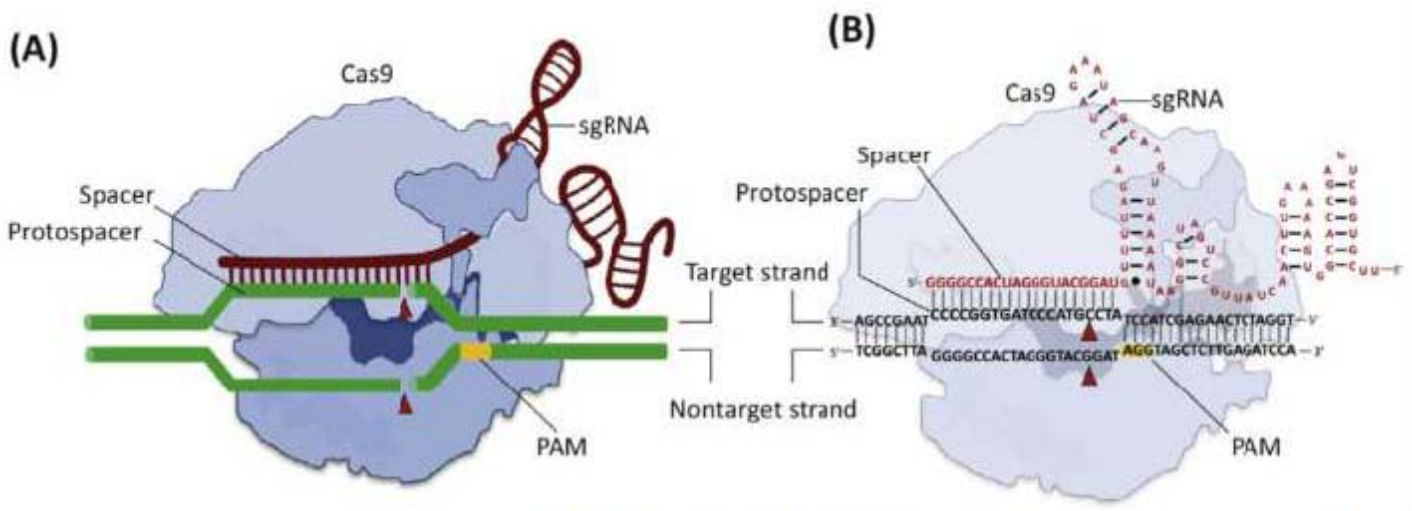
Strain[a]	Serotype	CRISPR 2A[b]	CRISPR 1C[b]	Prophages in the genome[c]	Prophages targeted by CRISPR		Ref.
SF370	M1	6	3	370.1 – 4	10270.1, 2 315.2, 3, 4, 5 SPsP2, 3, 4, 5 10750.1, 2, 3 10394.3, 4, 5	9429.2 8232.2, 3, 5 6180.2 NZ131.3	(Ferretti, et al., 2001)

(a) Strains are ordered from highest to lowest total number of CRISPR spacers.
(b) Number of spacers are indicated.
(c) Prophages and prophage remnants are indicated.

CRISPR ARRAY: Streptococcus pyrogenes



CRISPR-Cas9: DNA Cleavage



Streptococcus pyogenes sgRNA and CRISPR-Cas9 protein in complex with a dsDNA target. The spacer sequence hybridizes to the target strand of the targeted DNA and the Cas9 nuclease generates a DS break (DSB) (red arrows) three nucleotides upstream of the 5'-NGG-3' protospacer adjacent motif (PAM) sequence.
After the Cas9:sgRNA disassociates from the target sequence.

CRISPR-Cas9: Computational Aspects – Finding CRISPR Arrays

GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAA
ACTGCGCTGGTTGATTTCTTCTTGCGCTTTTTGT
TTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC
TTATATGAACATAACTCAATTTGTAAAAAAGTTT
TAGAGCTATGCTGTTTTGAATGGTCCCAAAACAG
GAATATCCGCAATAATTAATTGCGCTCTGTTTTA
GAGCTATGCTGTTTTGAATGGTCCCAAAACAGTG
CCGAGGAAAAATTAGGTGCGCTTGGCGTTTTAGA
GCTATGCTGTTTTGAATGGTCCCAAAACTAAATT
TGTTTAGCAGGTAAACCGTGCTTTGTTTTAGAGC
TATGCTGTTTTGAATGGTCCCAAAACTTCAGCAC
ACTGAGACTTGTTGAGTTCCATGTTTTAGAGCTA
TGCTGTTTTGAATGGTCTCCATTC

>NC_002737.2:860819-861250 Streptococcus pyogenes M1 GAS
GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC
TGCGCTGGTTGATTTCTTCTTGCGCTTTTTGTTTTA
GAGCTATGCTGTTTTGAATGGTCCCAAAACTTATAT
GAACATAACTCAATTTGTAAAAAAGTTTTAGAGCTA
TGCTGTTTTGAATGGTCCCAAAACAGGAATATCCGC
AATAATTAATTGCGCTCTGTTTTAGAGCTATGCTGT
TTTGAATGGTCCCAAAACAGTGCCGAGGAAAAATTA
GGTGCGCTTGGCGTTTTAGAGCTATGCTGTTTTGAA
TGGTCCCAAAACTAAATTTGTTTAGCAGGTAAACCG
TGCTTTGTTTTAGAGCTATGCTGTTTTGAATGGTCC
CAAAACTTCAGCACACTGAGACTTGTTGAGTTCCAT
GTTTTAGAGCTATGCTGTTTTGAATGGTCTCCATTC

Direct Repeat consensus :
GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC

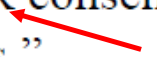
Finding CRISPR Arrays

Computational Challenge One

Given a DNA sequence, where are the CRISPR-arrays?

Some available tools:

1) "CRISPRCasFinder is a web service offering fundamental tools for CRISPR detection, including the shortest ones, allowing an accurate definition of the DR consensus boundaries and extraction of the related spacers."



Direct Repeats

2) "CRISPRDetect is a tool to discover and explore the CRISPR noncoding RNAs in sequence data.
It is a bioinformatics tool to find CRISPR arrays."

Streptococcus pyogenes and CRISPRCasFinder

- We are going to retrieve the whole genome of strain SF370 of *Streptococcus pyogenes* from NCBI:
  - Go to NCBI: http://www.ncbi.nlm.nih.gov/
  - Choose "Nucleotide" from the drop-down window (that shows "All Databases")
  - Enter "NC_002737.2" (which is the accession number of strain SF370 of
  - *Streptococcus pyogenes*) in the search window
  - Click on the blue search button
  - From the next page, click on "FASTA" to get the whole genome in FASTA format
  - Go to "Send to:" (right hand side of page) and save the file, that you name "S_pyogenes.txt" on your computer.

Streptococcus pyogenes and CRISPRCasFinder (2)

- We are now ready to find the CRISPR arrays in strain SF370 of *Streptococcus pyogenes.*
  - Go to https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index
  - Either use the "Browse" button to upload "S_pyogenes.txt" or copy the sequence from "S_pyogenes.txt" and paste it in the "Sequence(s)" window
- Click on "Run CRISPRCasFinder" at the bottom of the page".

- CRISPR arrays having evidence-levels **3** and **4** are considered as highly likely candidates, whereas evidence-levels **1** and **2** indicate potentially invalid CRISPR arrays. So, we only consider "NC_002737_1" from the new page.
- Click on "NC_002737_2_Streptococcus_pyogenes_M1_GAS_complete_sequence_1"

Streptococcus pyogenes and CRISPRCasFinder (3)

- 1) How many spacers are there in the NC_002737_1 CRISPR array? _____.
- 2) Record in the following table, the starting and ending locations of the CRISPR array reported by the package.
- **CRISPR Array Starting Position Ending Position NC_002737_2**
- 3) What is the consensus sequence of the repeats of the CRISPR array?
- 4) What percentage is reported for the conservation of direct repeats (Conservation DR)? _____. Did you expect such a high/low percentage and why? _____ (Yes/No)
- 5) What percentage is reported for the conservation of spacers (Conservation Spacer)? _____. Did you expect such a high/low percentage and why? _____ (Yes/No)

Double Check Our Work

We would like to check and see from which viruses these spacers come. Recall that a bacteriophage, also known informally as a phage, is a virus that infects and replicates within bacteria and archaea.

- 6) Take the first spacer from the previous question and BLAST it (use blastn) at NCBI. Did the blast produce a significant match between the first spacer and some virus? Explain.

- 8) Repeat question 6 with a few of the remaining spacers.

CRISPR Continued Next Class