CS123A Bioinformatics

Module 5 – Week 16 – Presentation 2

Leonard Wesley

Computer Science Dept

San Jose State Univ

Agenda

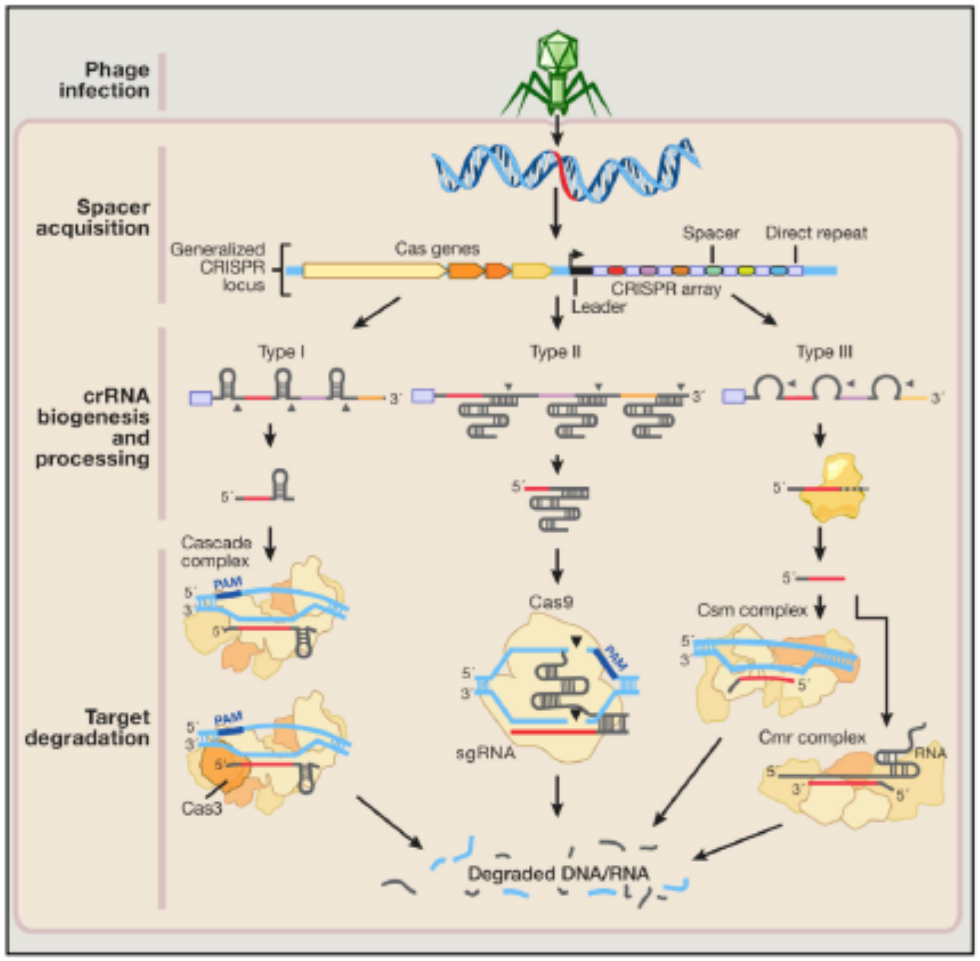
- Introduction To CRISPR (continued)
  - Designing CRISPR-Cas9 Systems

Introduction To CRISPR-Cas 9 System

CRISPR-Cas9 System

- **CRISPR-Cas9** is a unique technology that enables geneticists and medical researchers to edit parts of the genome by removing, adding or altering sections of the DNA sequence.
- It is currently the simplest, most versatile and precise method of genetic manipulation and is therefore causing a buzz in the science world.

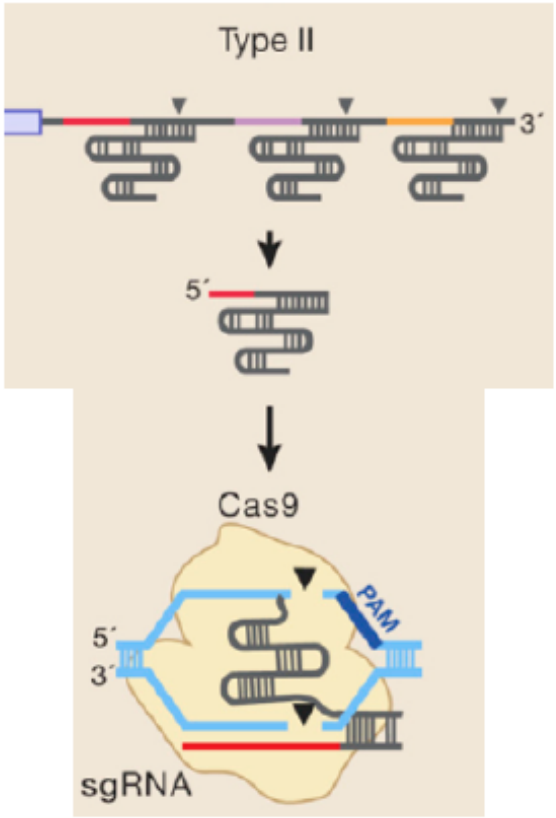
Different CRISPR-Cas Systems



Recall:
The CRISPR array is a **noncoding RNA** transcript that is maturated through distinct pathways that are unique to each **type** of the CRISPR system.

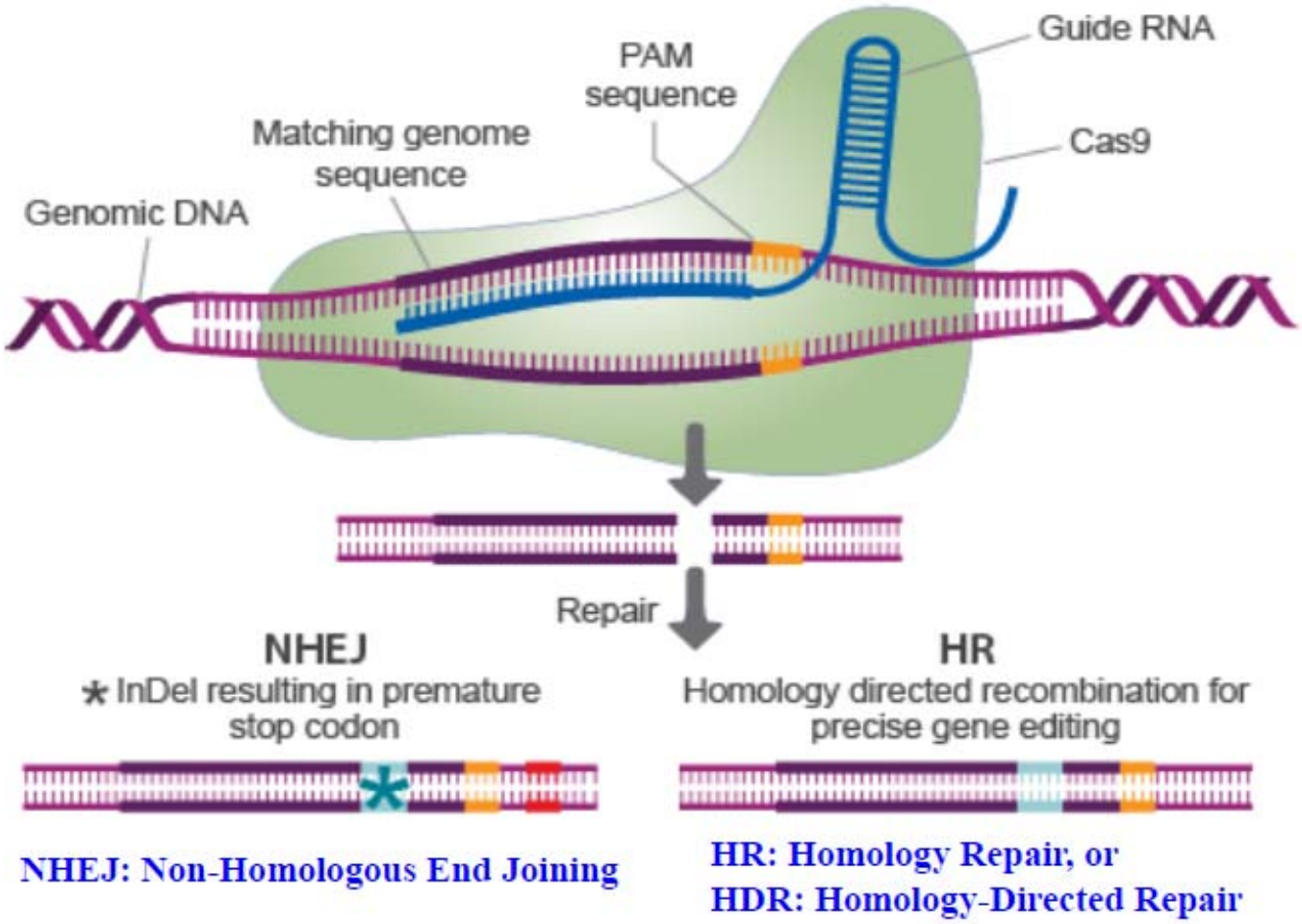
There are three types.

CRISPR-Cas Type II



Recall:
In **type II** CRISPR, an associated ***trans*-activating CRISPR RNA** (tracrRNA) hybridizes with the direct repeats, forming an RNA duplex that is cleaved and processed by endogenous RNase III and other unknown nucleases.

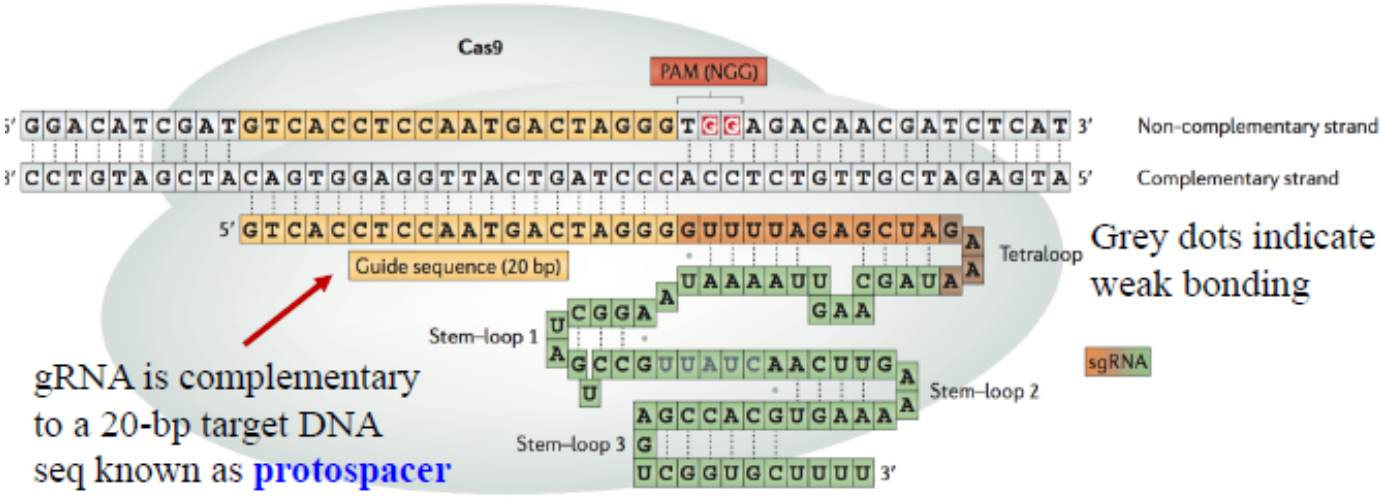
The crRNA-tracrRNA hybrids with Cas9 to form a crRNA-tracrRNA Cas9 complex to mediate interference.


Guide RNA
PAM sequence
Matching genome sequence
Cas9
Genomic DNA
Repair
NHEJ
* InDel resulting in premature stop codon
HR
Homology directed recombination for precise gene editing
NHEJ: Non-Homologous End Joining
HR: Homology Repair, or
HDR: Homology-Directed Repair

CRISPR-Cas9 System

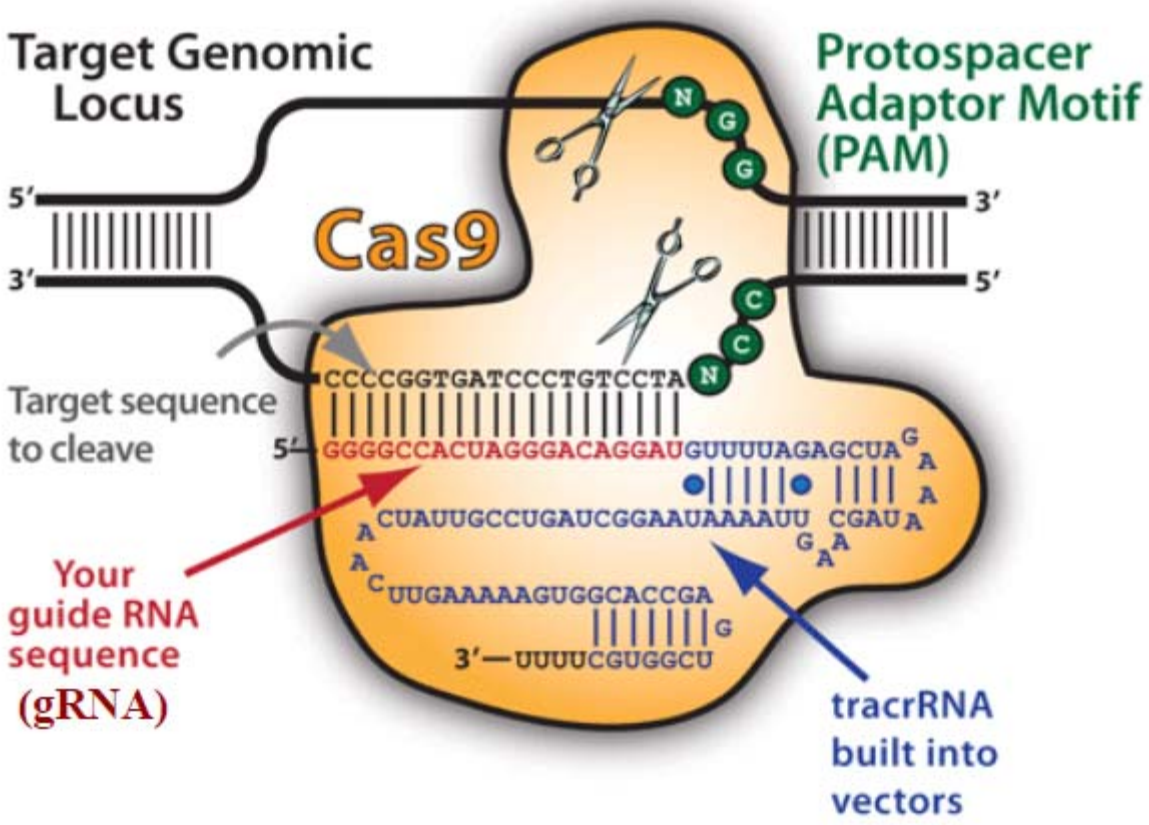
- The CRISPR-Cas9 system consists of two key molecules that introduce a change into the DNA:
  - an **enzyme called Cas9**. This acts as a pair of 'molecular scissors' that can cut the two strands of DNA at a specific location in the genome so that bits of DNA can then be added or removed.
  - a piece of RNA called **guide RNA** (**gRNA**): a small piece of RNA sequence (about 20 bases) located within a longer RNA scaffold: **single guide RNA** (**sgRNA**). The scaffold part binds to DNA and the sequence 'guides' Cas9 to the right part of the genome to make sure to cut at the right point in the genome.

RNA-Guided Engineered Nucleases (RGEN) [I]



An RGEN contains **Cas9** and a single-chain guide RNA (**sgRNA**). The **guide sequence** (**gRNA**) in the sgRNA is complementary to a **protospacer**, which is next to the 5′-NGG-3′ sequence known as: protospacer adjacent motif (**PAM**).

The CRISPR-Cas9 Nuclease Heterocomplex



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

Designing Efficient gRNA's

Recall:

- The sgRNA is functionally equivalent to the crRNA–tracrRNA complex.
- The sgRNA is designed to have a guide sequence domain (gRNA) at the 5′ end, which is complementary to the target sequence.
- The rationally designed sgRNA is then used to guide the Cas9 protein to specific sites in the genome for targeted cleavage.
- Given the critical roles of the gRNA, multiple bioinformatics tools have been developed for the rational design of gRNAs for the CRISPR/Cas9 system.

Computational Challenge Two

Given a DNA sequence, design efficient guide RNAs (gRNAs).

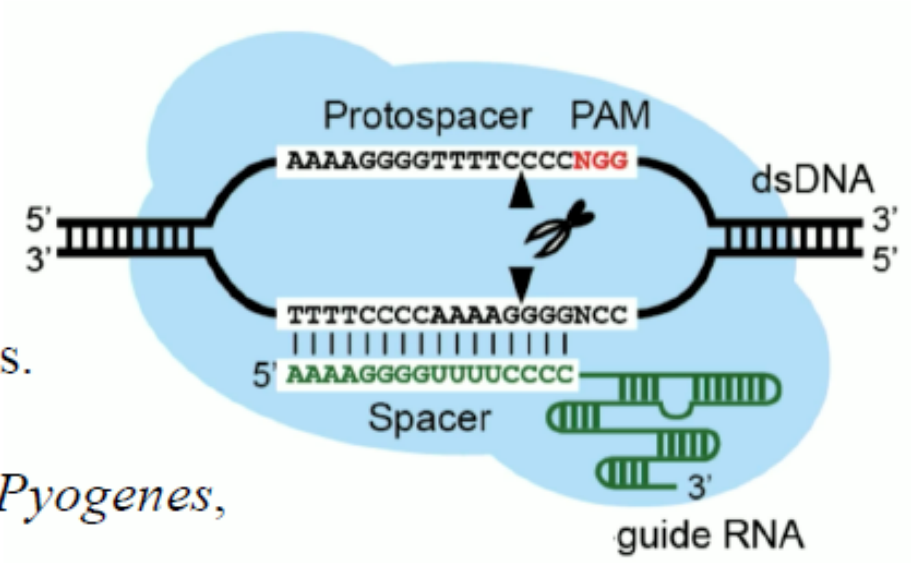
In silico gRNA (and sgRNA) design has become a key issue for successful gene-editing experiments and will allow CRISPR studies to take advantage of various bioinformatics and computational techniques.

The goal is to design *in silico* gRNA (and sgRNA) with **high on-target efficacy** and **reduced off-target effects**.

Challenge Two: First Step

Given a DNA sequence, design efficient guide RNAs (gRNAs).

Finding a gRNA is equivalent to finding a PAM and counting 20 bases backwards.



For *Streptococcus Pyogenes*, the PAM is NGG.

gRNA Design Algorithm

Given a DNA sequence, design efficient gRNA's

1) Find as many gRNA's as possible
   * Might decide to restrict region (exons only, etc..)
   * **General Rules**: Avoid homopolymers and stable secondary structures, and restrict GC content between 30 and 70%
2) Map each gRNA to a Reference Genome:
   * Example: human genome GRCh38 (Genome Reference Consortium 38 – also known as: hg38)
   * Find exact location of gRNA's (exon, UTR, intergenic, etc..)
     - Use Enembl, NCBI, UCSC genome browser, etc..
   * Find Potential Off-Targets
3) Score each gRNA
   * Give a quantitative measure to each gRNA
   * Use as many gRNA features as possible

Desirable Features of gRNA

Some **position-specific sequence** features are associated with CRISPR activities.

- Nucleotides adjacent to the **protospacer adjacent motif** (**PAM**) NGG in the target site are significantly depleted of C's or T's.

Structural characteristics of the sgRNA are important determinants of CRISPR activity.

- On average, non-functional guide sequences have significantly higher potential for self-folding ($\Delta G = -3.1$) than functional guide sequences ($\Delta G = -1.9$).

Desirable Features of gRNA

In general, structural stability of the RNA can be approximated by the **GC content** of the sequence.
Consistent with the free energy calculation, the guide sequence of non-functional sgRNAs has higher GC content on average (0.61) compared with functional sgRNAs (0.57).

High duplex stability is a significant characteristic of non-functional sgRNAs.
On average, non-functional guide sequences were predicted to form more stable RNA/DNA duplexes with the target sequence ($\Delta G = -17.2$) than functional ones ($\Delta G = -15.7$).

Sequence Characteristics of Functional sgRNAs

- In most CRISPR applications, a 20-mer DNA oligo representing the guide sequence is cloned into an expression vector and expressed as the gRNA domain within the sgRNA.

- Thus, the efficiencies of both DNA oligo synthesis and the subsequent transcription process are relevant to CRISPR activity.
- Repetitive bases (i.e., a stretch of contiguous same bases) could potentially be correlated with poor efficiency for DNA oligo synthesis.

Repetitive Bases in sgRNAs

- Repetitive bases are defined as any of the following:
  five contiguous adenines (AAAAA)
  five contiguous cytosines (CCCCC)
  four contiguous guanines (GGGG) or
  four contiguous uracils (UUUU).

- Overall, functional gRNAs were significantly depleted of repetitive bases (5.4 %), when compared with non-functional gRNAs (22.8 %).

Repetitive Bases in sgRNAs

- Among the four bases, four contiguous guanines (GGGG) were especially correlated with poor CRISPR activity.
- GGGG leads to poor yield for oligo synthesis, and also has the propensity to form a special secondary structure called a guanine tetrad, which makes the guide sequence less accessible for target sequence recognition.
  * Consistently, much fewer functional gRNAs were observed to contain the GGGG motif (4.9%) than non-functional ones (17.9%)

Table 1 Significant base counts in functional gRNAs

Mono- or dinucleoside count	Enrichment ratio[a]	*P* value[b]
A	1.39	9.3E–18
U	0.89	1.9E–03
G	0.92	6.2E–03
C	0.95	5.5E–02
GG	0.64	2.3E–11
AG	1.43	1.3E–09
CA	1.38	6.7E–09
AC	1.47	1.2E–08
UU	0.59	7.5E–08
UA	1.84	1.1E–07
GC	0.77	3.2E–06

$$\frac{\#\text{ of A's in gRNA}}{\#\text{ of A's in non gRNA}}$$

[a]The enrichment ratio was determined by comparing the average nucleoside counts of functional gRNAs to that of non-functional gRNAs. [b]The *P* values were calculated with Student's t-test

Targeting of CRISPR-Cas9

- In most cases the guide RNA (gRNA) consists of a specific sequence of 20 bases that are complementary to the target sequence in the gene to be edited. However, not all 20 bases need to match for the guide RNA for binding.
- The problem is that a sequence with, for example, 19 of the 20 complementary bases, may exist elsewhere in the genome. So, there is potential for the guide RNA to bind there instead of or as well as at the target sequence.
- The Cas9 enzyme will then cut at the wrong site and end up introducing a mutation in the wrong location. While this mutation may not matter at all, it could affect a crucial gene or another important part of the genome.

Future of CRISPR-Cas9

- It is likely to be many years before CRISPR-Cas9 is used routinely in humans.
- Much research is still focusing on its use in animal models or isolated human cells, with the aim to eventually use the technology to routinely treat diseases in humans.
- There is a lot of work focusing on eliminating 'off-target' effects, where the CRISPR-Cas9 system cuts at a different gene to the one that was intended to be edited.

CRISPR: In-Lecture Exercise

Extracting the Human HBB Gene

- Download and open the file named “beta_globin_sequence.docx” from the CANVAS -> Files -> Module 3 CRISPR -> Week 10 -> Slides folder.
- Clearly identify the various parts of the sequence: exons, introns, 5’UTR, 3’UTR, etc.
- Copy the gene only, in other words from the underlined A (in bold) on the first page to the last uppercase base (C) on page 2.
- Paste the copied sequence in the window at: https://www.browserling.com/tools/remove-all-whitespace
- Make sure to remove all non-base characters in the sequence just pasted
- Click on “Remove All Spaces!”
- Copy the sequence obtained in a text file that you name: Human_HBB.txt.

Detecting gRNAs in Human HBB with WU-CRISPR:

- Go to http://crispr.wustl.edu/
- Choose "**Option 2: gRNA design for user-provided genomic target sequence:**" and paste the sequence from Human_HBB.txt in the window
- Click on "Go"
- Click on "Retrieve Prediction Results"
- Inspect the results obtained from running WU-CRISPR and answer the following questions:
- 1) How many gRNA's are reported by WU-CRISPR? _____.
  - Click on "Details" of the third hit
- 2) What is the PAM sequence? _____.
  - Return to the page of results
  - Click on "Details" of the top hit
- 3) What is the PAM sequence? _____.
  - Return to the page of results
  - Click on "Return to Homepage"

WU-CRISPR cont.

- We are going to run WU-CRISPR a second time, but with the first option.
- Choose "**Option 1: Search for predesigned gRNA to target human or mouse genes:**" and type "HBB" in the search window
- Click on "Go"
- Inspect the results obtained from running WU-CRISPR and answer the following questions:
- 4) How many gRNA's are reported by WU-CRISPR this time? _____.
- 5) Why do think there is a (big) difference between the number of reported gRNA when using different options?
  - Click on "Details" of the eighth hit (with a Potency Score of 98)
- 6) Explain in your own words the "Off-Target Analysis".
- 7) Copy and paste the accession number NM_001353851 in the NCBI Gene Search Box. What does this gene do? _____

Detecting gRNAs in Human HBB with CHOPCHOP:

- Go to http://chopchop.cbu.uib.no/
- Type HBB in the "Target" window and keep the default species and PAM in the other two windows: "In" and "Using"
- Click on "Find Target Sites!"
- Inspect the results obtained from running CHOPCHOP and answer the following questions:
- 1) How many gRNA's are reported by CHOPCHOP? _____.
  - Consider the gRNA with "Ranking" value of 7.
- 2) What is the PAM sequence of that gRNA? _____.
- 3) How many off-target hits does this gRNA have? _____.
  - Click on the row of gRNA with ranking 7 to go to a new page with more information on the off-target sequences.
- 4) Compare these off-target hits with those reported by WU-CRISPR.

Detecting gRNAs in Human HBB with GPP

- Go to https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design
- Select “human” from the “Select Target Taxon”
- Type HBB in the input window
- Check the “I’m not a robot” box
- Click on “Submit”
- Since the two output files, reported under “Download Results” are “tab-delimited text files, click on “sgRNA Picking Results” and save it in an excel spreadsheet that you name “sgRNA_Picking_Results_HBB.xls”
- 5) How many gRNA’s are reported by GPP? _____.

STAY SAFE