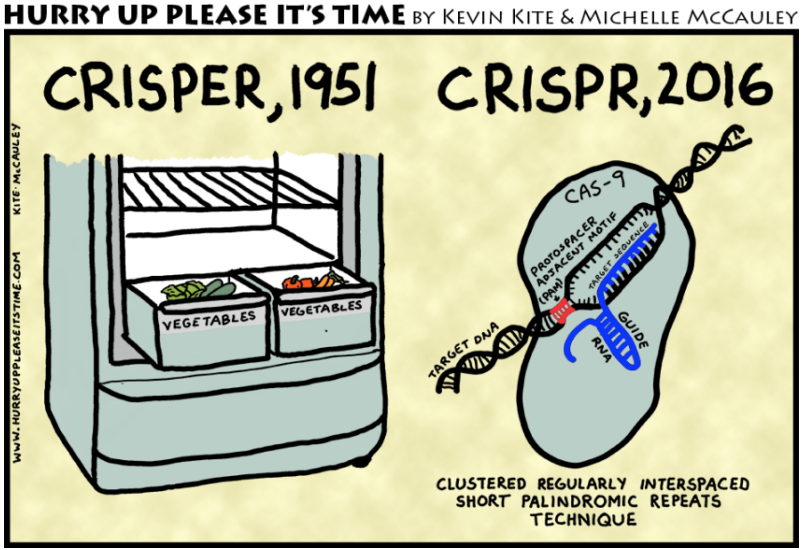
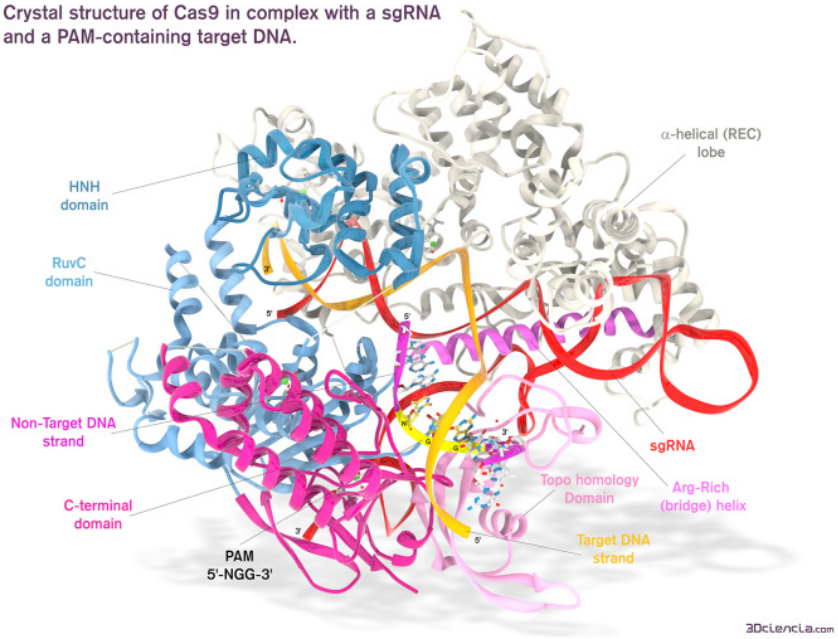
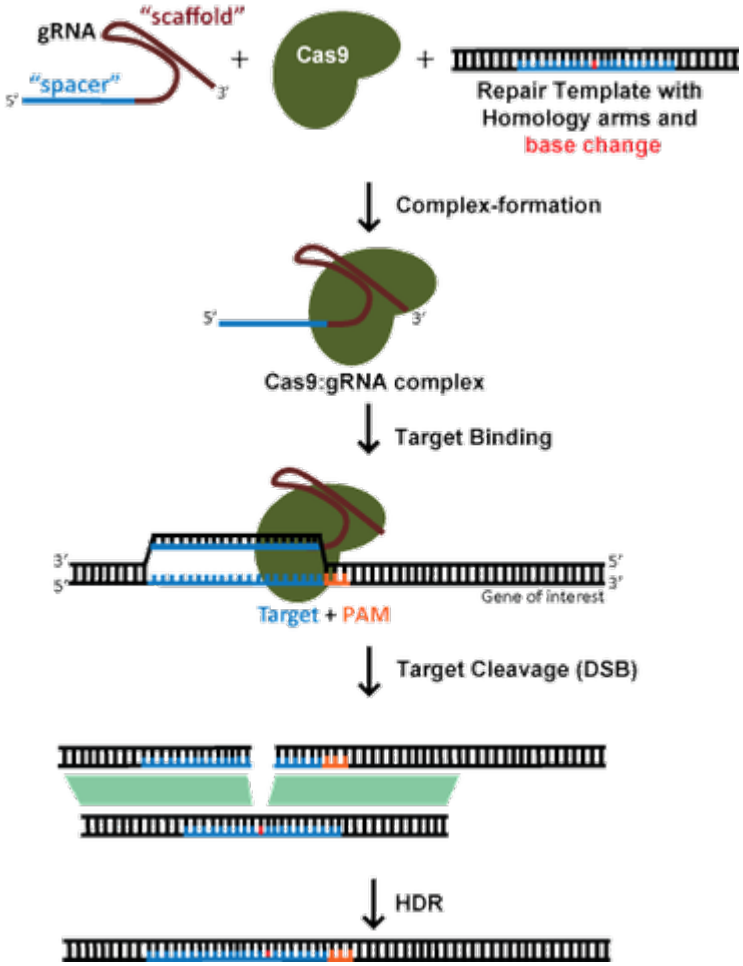
# Project 2.1: Identifying Off-Target CRISPR Sites



# What is CRISPR?

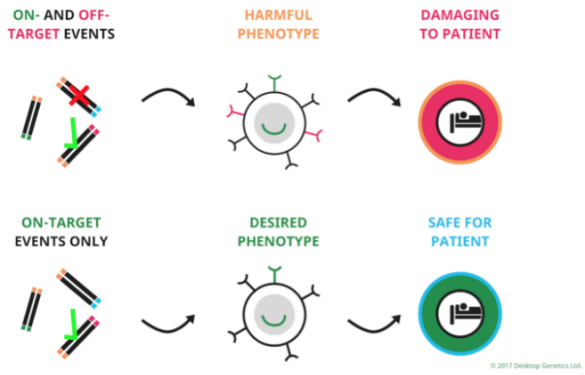
*A new tool to allow precise genome editing*



# Off-Target Effects

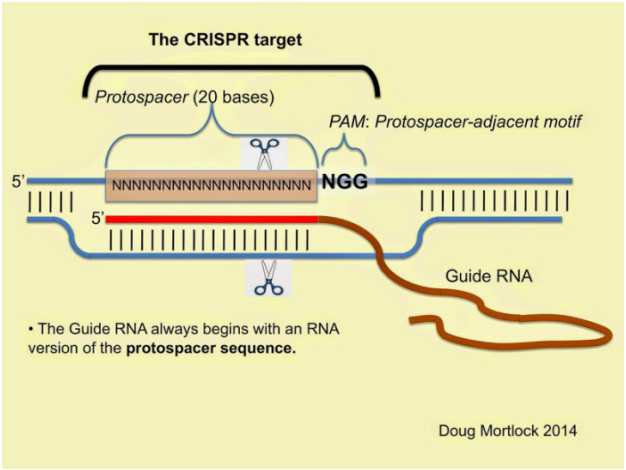
*The major setback for clinical CRISPR/Cas9 use*




human genome
3,000,000,000 base pairs
“NGG” occurs ~ 160,000,000 times

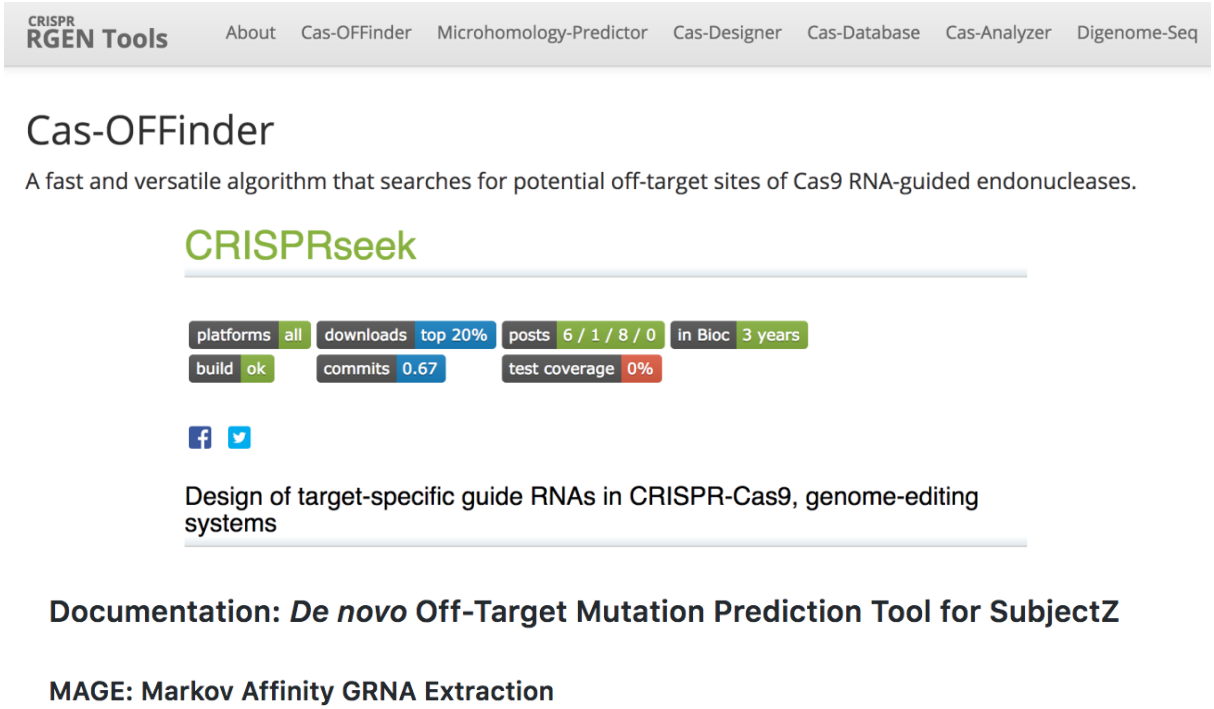
gRNA
20 bases (12 absolute)

… A LOT of potential binding sites!



# Goal: How to predict off-target sites?

*Computational methods to assess gRNAs*



# Principles of CRISPR/Cas9 Specificity

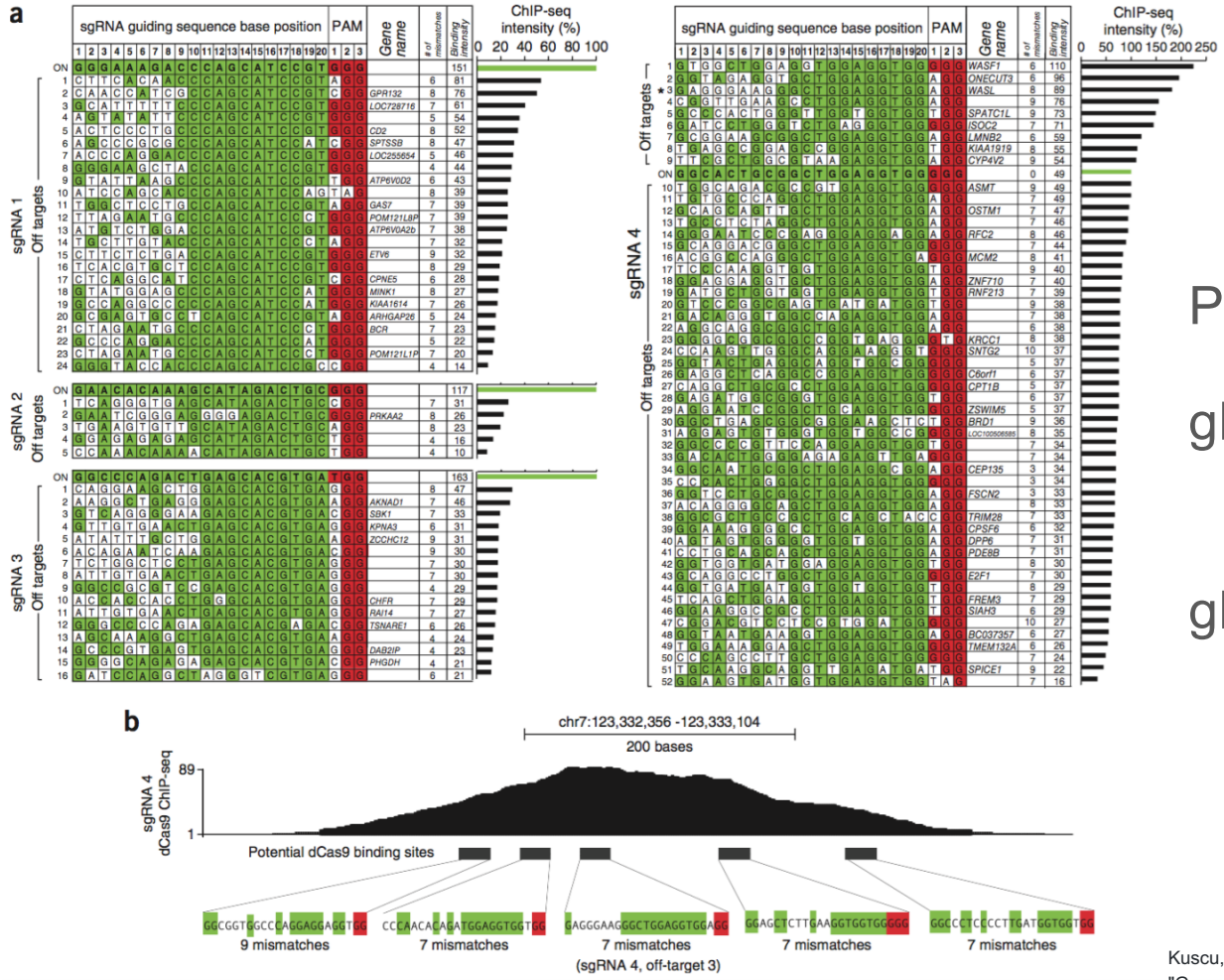
PAM site

gRNA: PAM Proximal vs PAM Distal

Chromatin Structure

Methylation

Cas9 Related Factors




Proximal >> Distal

gRNA tolerates ~7-10 mismatches in Distal

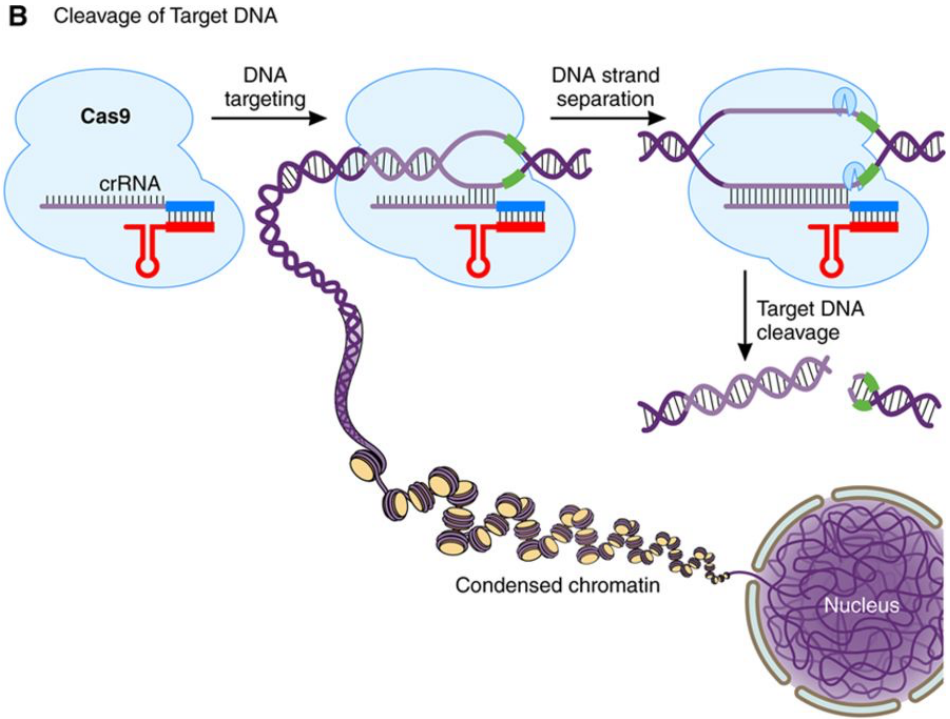
gRNA must be 20 bases

Kuscu, Cem, Sevki Arslan, Ritambhara Singh, Jeremy Thorpe, and Mazhar Adli. "Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease." Nature Biotechnology32.7 (2014): 677-83. Web.

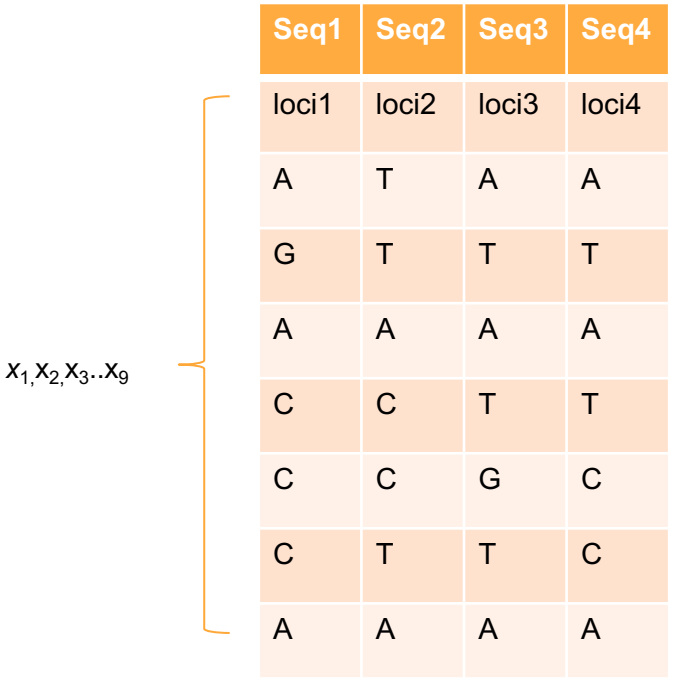
# Chromatin Structure

*Cas9 preferentially accesses open chromatin*

ENCODE contains ChIP-seq data detailing chromatin structures in cells



# Sequence information is multidimensional

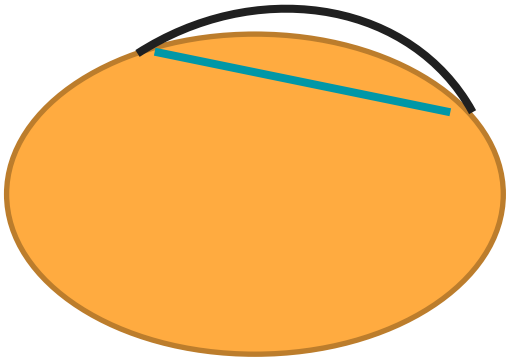


# Multidimensional spaces feature alternative geometries

## Oversimplified example: flight from NYC to Moscow

Euclidean Distance: traditional geometric distance
- Simple: Pythagorean theorem, Norms
- **Straight line**
- **Through the mantle of the earth**

Manifold Distance: distance along the curve or surface
- Found via “walking” over the surface
- **Curved line**
- **Over the surface of the earth**



# There are many multidimensional geometries

Eg. Minkowski geometries

Simply a weighted distance metric on each dimension

Classical example: 4d Euclidean spacetime:

$$-t^2 + x^2 + y^2 + z^2$$

# Minkowski weighting of sequence information defines relative importance of each sample

$x_1, x_2, x_3 .. x_9$

| Seq1 | Seq2 | Seq3 | Seq4 |
|---|---|---|---|
| loci1 | loci2 | loci3 | loci4 |
| A | T | A | A |
| G | T | T | T |
| A | A | A | A |
| C | C | T | T |
| C | C | G | C |
| C | T | T | C |
| A | A | A | A |

$$d(j,y) = \sqrt{\sum a_i (x_{yi} - x_{ji})^2}$$

**We set $a_i$ to weigh the ”seed” region or loci dimensions of the guide RNA more**

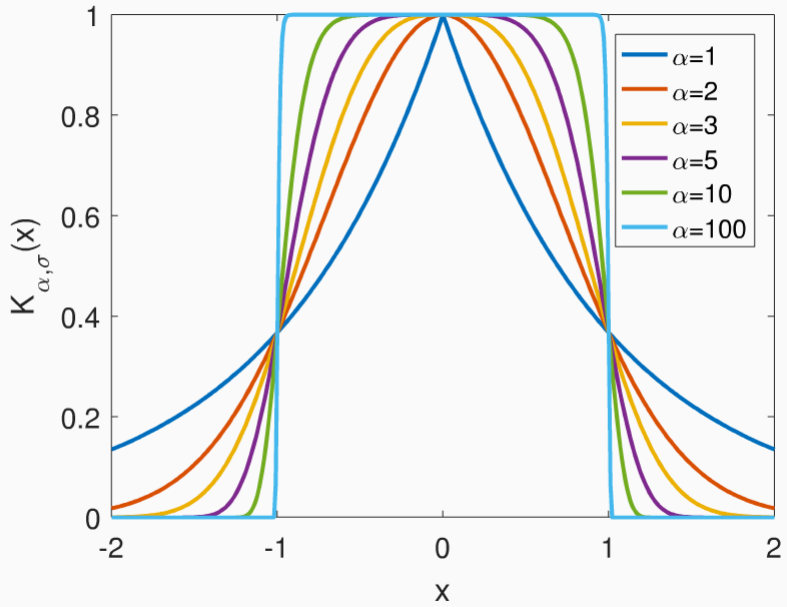
# How do we find manifold distances?

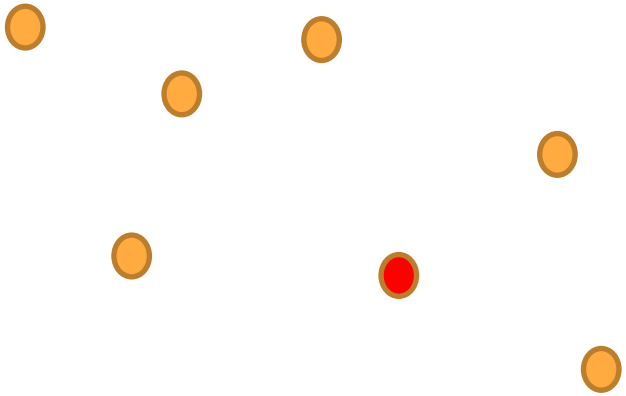
Imagine wandering around blindfold in a room for infinite repetitions.  Eventually you might converge on the walls.  The path along the walls is the manifold distance

First we need to build the network to wander on.

# Defining a kernel over seq space yields connected sequences

$$K_{\alpha,\sigma}(x) = \exp\left(-\left(\frac{|x|}{\sigma}\right)^{\alpha}\right)$$

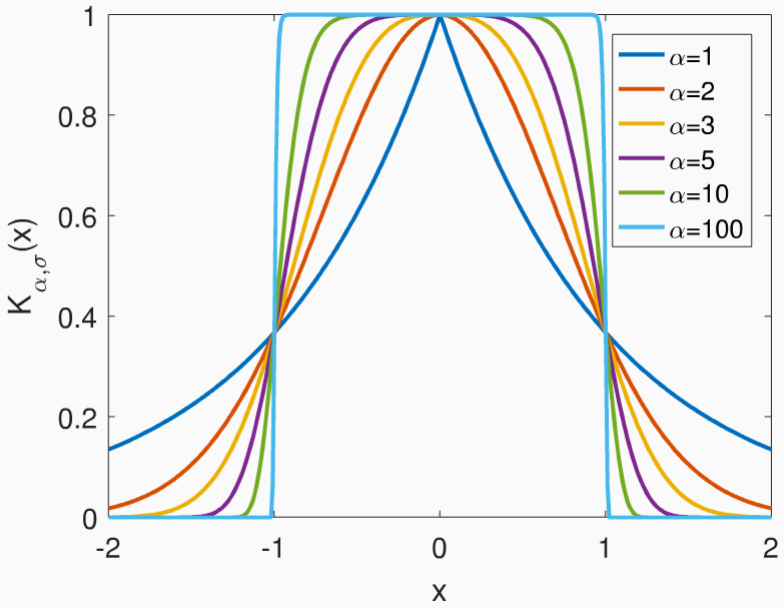


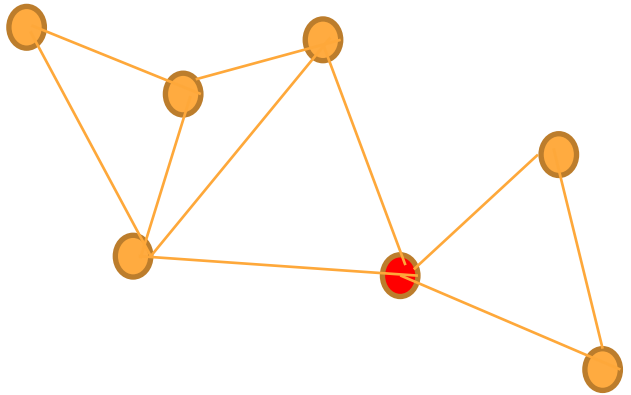



# Defining a kernel over seq space yields connected sequences

$$K_{\alpha,\sigma}(x) = \exp\left(-\left(\frac{|x|}{\sigma}\right)^{\alpha}\right)$$

σ = distance to k-nearest sequence
α = falloff between similar sequences



# How to walk?

Just power the sequence transition matrix (defined by the kernel) *t* times,

- t is optimized using a measure of the uncertainty in the network
- (Von Neumann Entropy)

If you powered it infinite times, you would have the steady-state probabilities of the matrix.  You could find this with the first eigenvector

# Okay, we've walked, what does that mean?

We've now walked onto a manifold and the corresponding probabilities can be viewed as potential distances using a -log transform. (inspired by physics)

These manifold distances are then mapped into probability distributions.

| | seq | diffusion distance | normal distribution probability | exponential distribution |
|---|---|---|---|---|
| 0 | GGGAAAGACCCAGCATCCGT | 0.0 | 0.023202214308488043 | 0.0333287844099327 |
| 45 | GAGAAAAGCCCAGCATCCTT | 0.18585623157998277 | 0.02280492334971664 | 0.02767598829551352 |
| 4 | GAGACAGAGAGAGCATCCGT | 0.21830785661197183 | 0.022655860457788453 | 0.026792274035376654 |
| 47 | CTGACAGACCAAGCATACGT | 0.24358169702627122 | 0.022524006520556004 | 0.02612361575114615 |
| 36 | GAGAATGACGCAGCATCCTC | 0.260319046269238 | 0.02242922319507647 | 0.025690014461274193 |
| 49 | GAGAAACACACAGCATCAGC | 0.2752150578120027 | 0.0223399388060171 | 0.025310171803252447 |
| 14 | CAGAAAGACCCATCATGCCT | 0.2876943009216918 | 0.02226161095239099 | 0.024996282639290462 |
| 43 | AAGAAAAACCCAGCATCTGA | 0.2918024081367915 | 0.02223512833610434 | 0.024893805867133974 |
| 7 | GTGCTAGACCCAGCATCAGT | 0.2971713234994512 | 0.022200000706308187 | 0.024760511274259937 |

# Pipelining to Identify Off-Target Effects

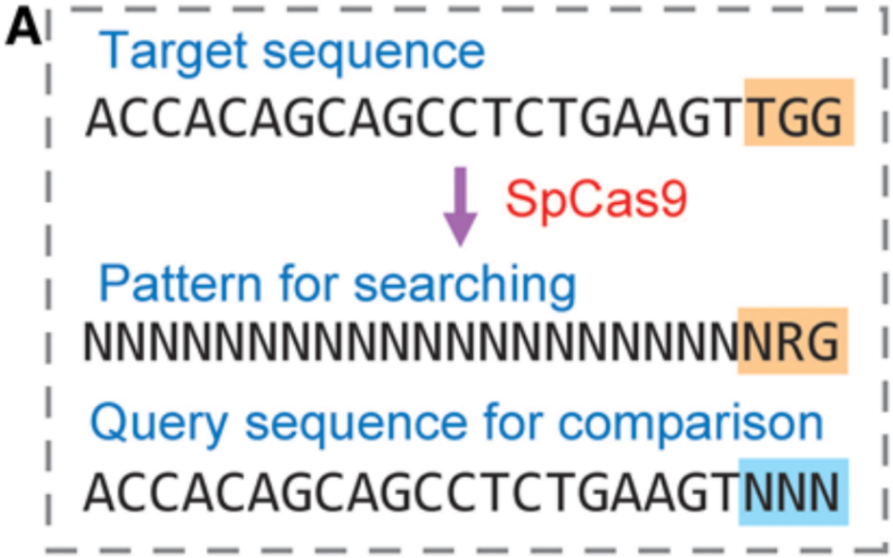
- Many published tools, limited experimental data
- Haeussler et al. 2016 published a compilation of 31 guides and 650 experimentally validated off target sites

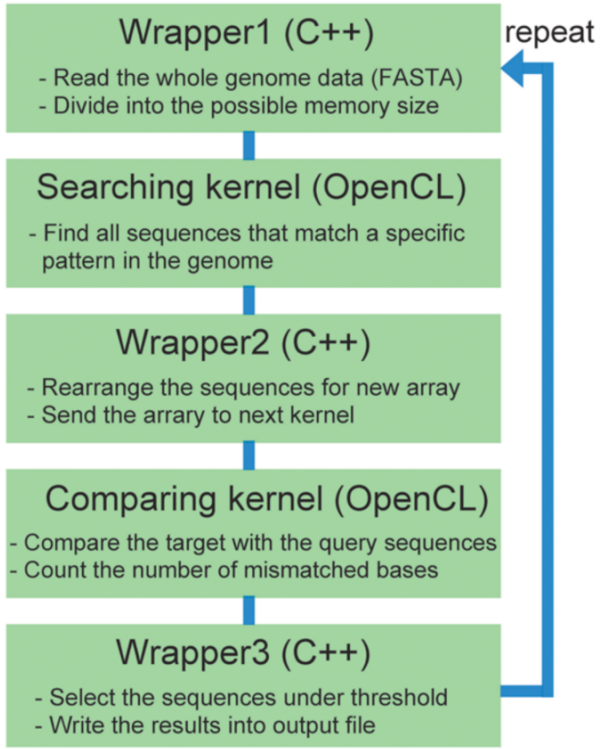
# Pipelining to Identify Off-Target Effects

- Many published tools, limited experimental data
- Haeussler et al. 2016 published a compilation of 31 guides and 650 experimentally validated off target sites

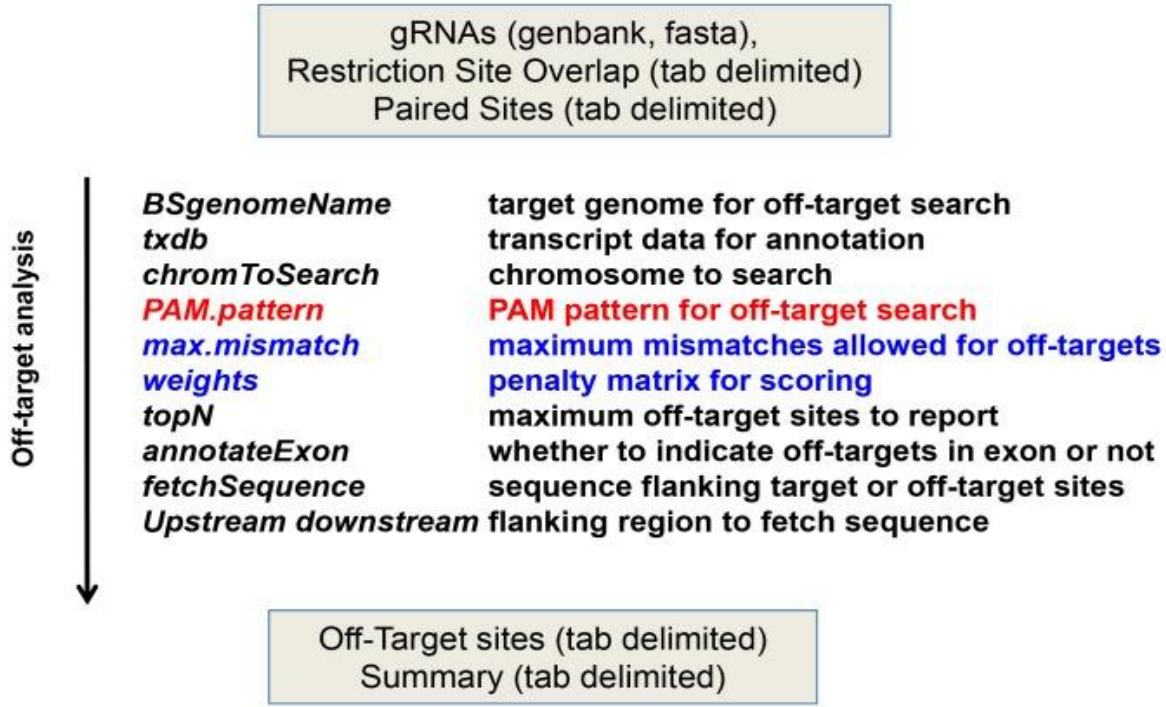
**Pipeline goal: compare the performance of two pipeline tools on this dataset**

# Pipeline: Algorithm 1 - Cas-OFFinder



# Pipeline: Algorithm 2 - CRISPR-SEEK



# Pipelining Results on hg38

- CRISPR-Seek:  516/650
- CasOFFinder: 461/650