
University of Massachusetts
UMASS Medical School

# CRISPRseek Workshop

Design of target-specific guide RNAs in CRISPR-Cas9 genome-editing systems

August 1st 2014

Lihua Julie Zhu

# Outline

- Background and Motives
- CRISPRseek Functionality
- Dependency
- Installation
- Get help and Reference
- Demo
- Exercise

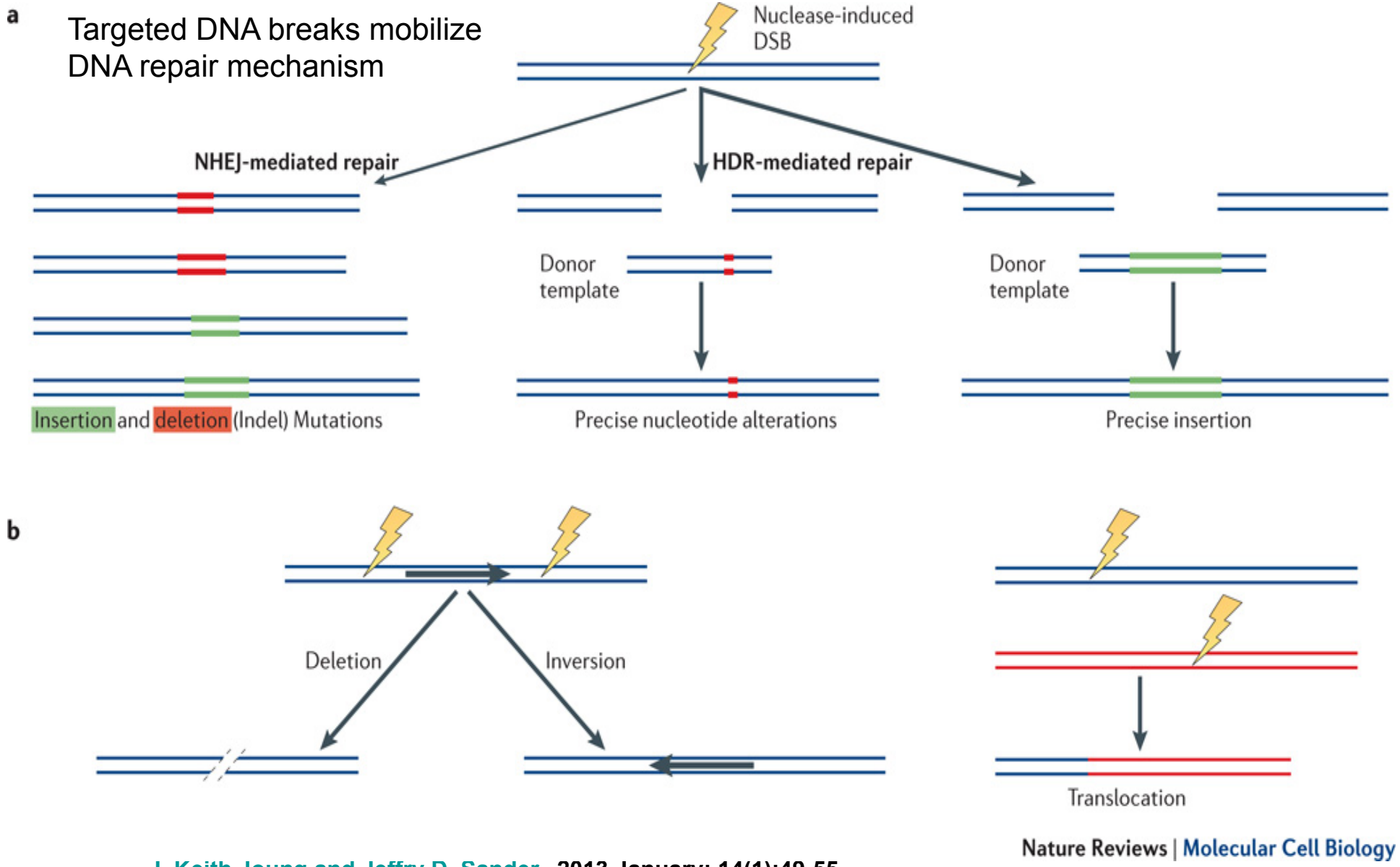
# Genome editing plays an increasingly central role in biomedical research

- Targeted gene mutation
  - Study the function of individual gene
- Create transgenic animals
  - Model disease
- Targeted transgene addition
  - Engineer crops to be disease-resistant
- Gene therapy
  - Erase a genetic mutation
    - Correct faulty genes that cause genetic diseases like Huntington's disease
  - Introduce a mutation to induce the resistance to virus infection, such as HIV infection (CCR5)
    - HIV infection results in the death of immune system cells, particularly CD4+ T-cells leading to AIDS
    - CCR5 is a co-receptor for HIV entry into T-cells and, if CCR5 is not expressed on their surface, HIV infects them with lower efficiency.

# Overview of Genome Editing Methods

- Transgenesis
  - Randomness
- Homologous Recombination
  - More precise
  - Lower efficiency
- Engineered nucleases (molecular scissors)
  - Cut DNA at targeted site in the genome
  - High efficiency
  - Take advantage of endogenous DNA repair machinery

Targeted DNA breaks mobilize DNA repair mechanism

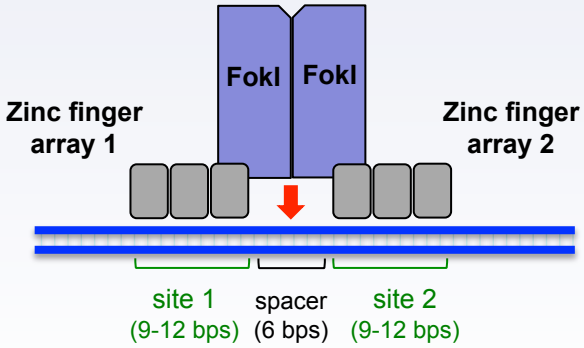



Nature Reviews | Molecular Cell Biology

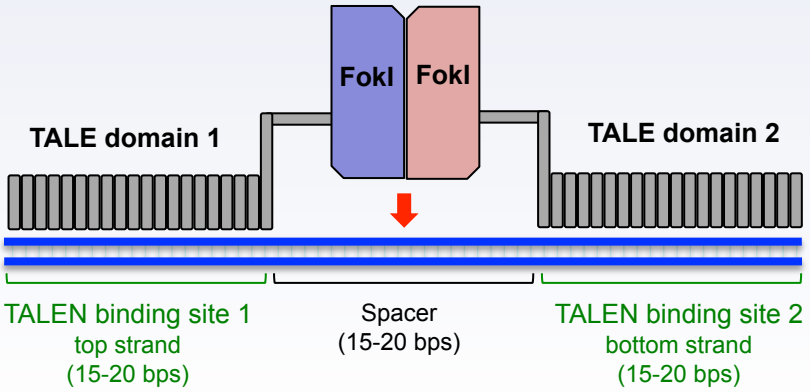
J. Keith Joung and Jeffry D. Sander. 2013 January; 14(1):49-55.

# Earlier Engineered Nucleases

## Zinc Finger Nuclease - ZFN (two subunits)



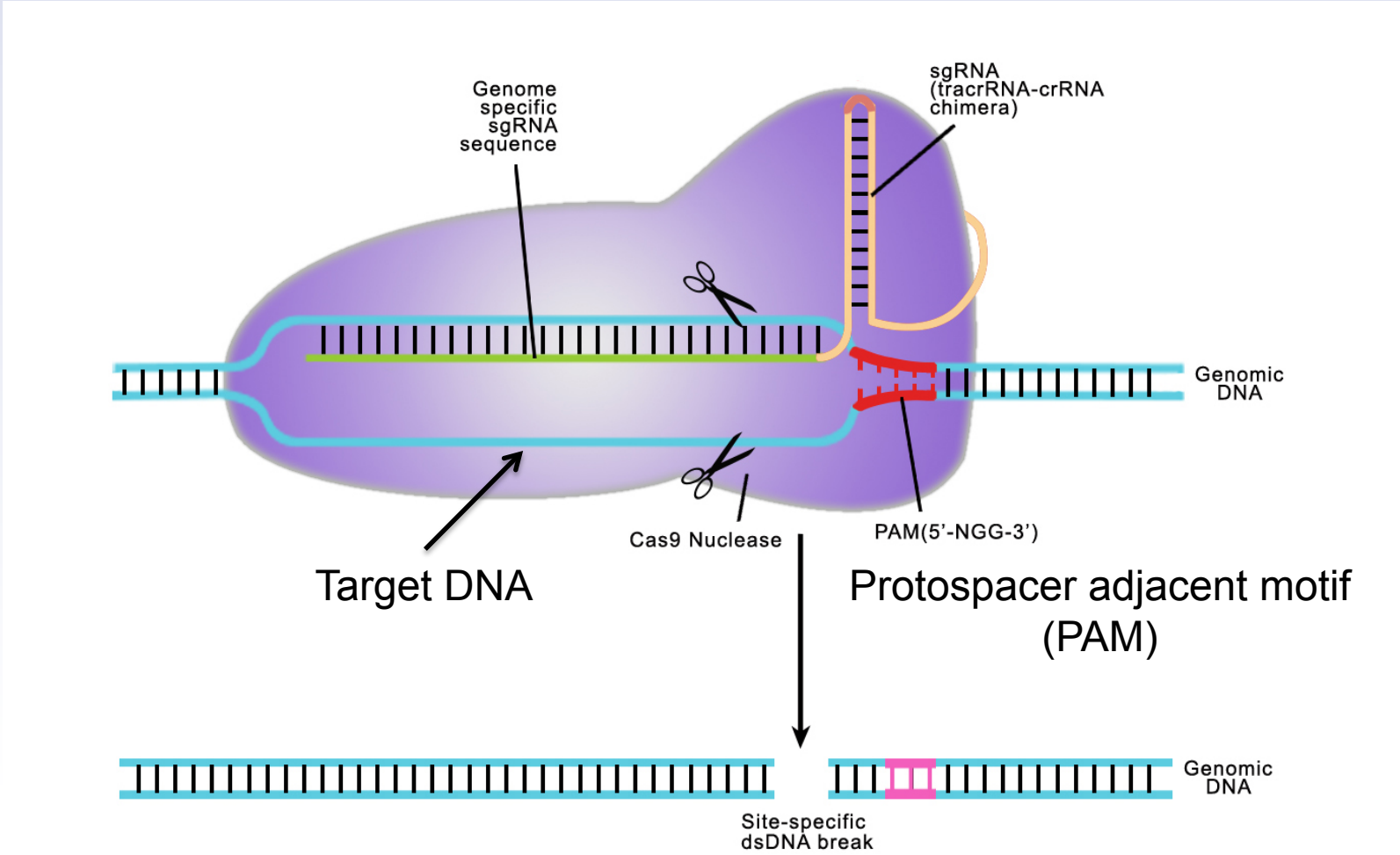

## TALE Nuclease - TALEN (two subunits)



# CRISPR-Cas9 System

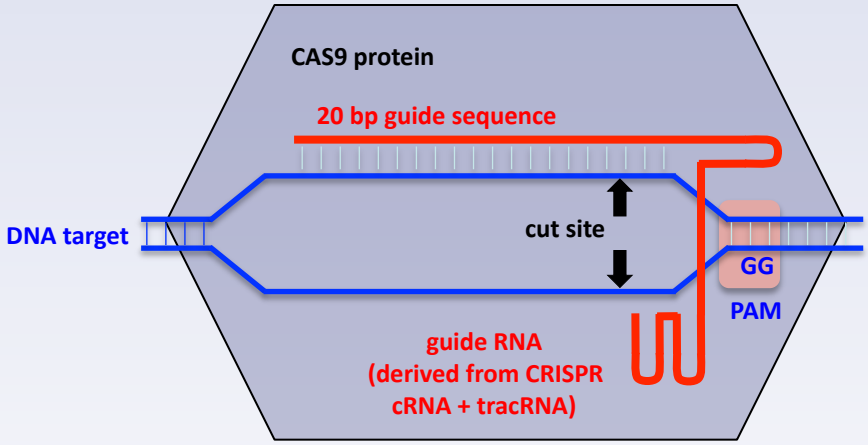
- Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated (Cas) proteins
- An adaptive immune defense system found in archaea and bacteria.
    - It provides resistance to potential invaders, such as plasmids and phages, by recognizing and cleaving foreign nucleic acids
- In 2012, it was shown that CRISPR-cas9 can be easily modified to activate or deactivate target genes with high efficiency and low expense
- CRISPR Therapeutics
- Editas Medicine

# GENOME EDITING WITH CRISPR-CAS9 SYSTEM

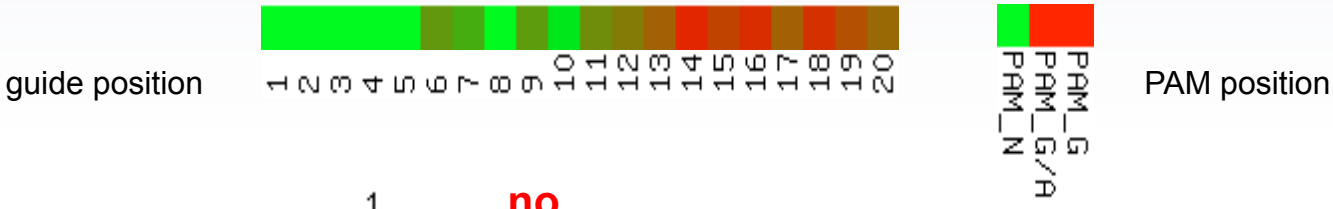



http://www.genecopoeia.com/product/crispr-cas9/

# gRNA design: minimize off-target cleavage

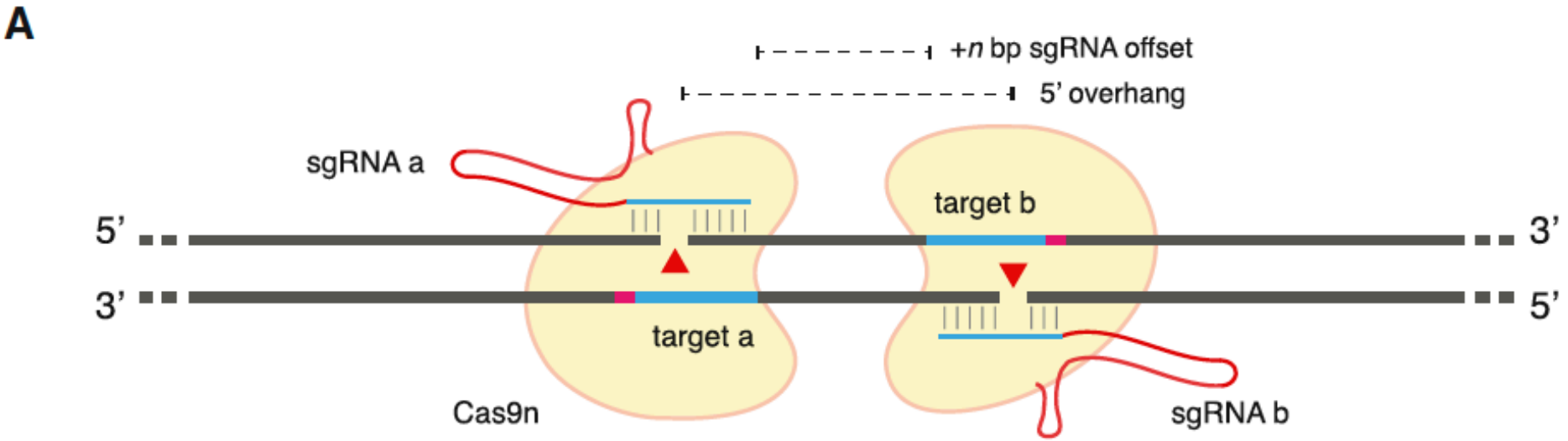



## Effects of single mismatches on cleavage efficiency




P. Hsu, et al., Nature Biotechnology July 2013

# Paired nickases will reduce off-target rates



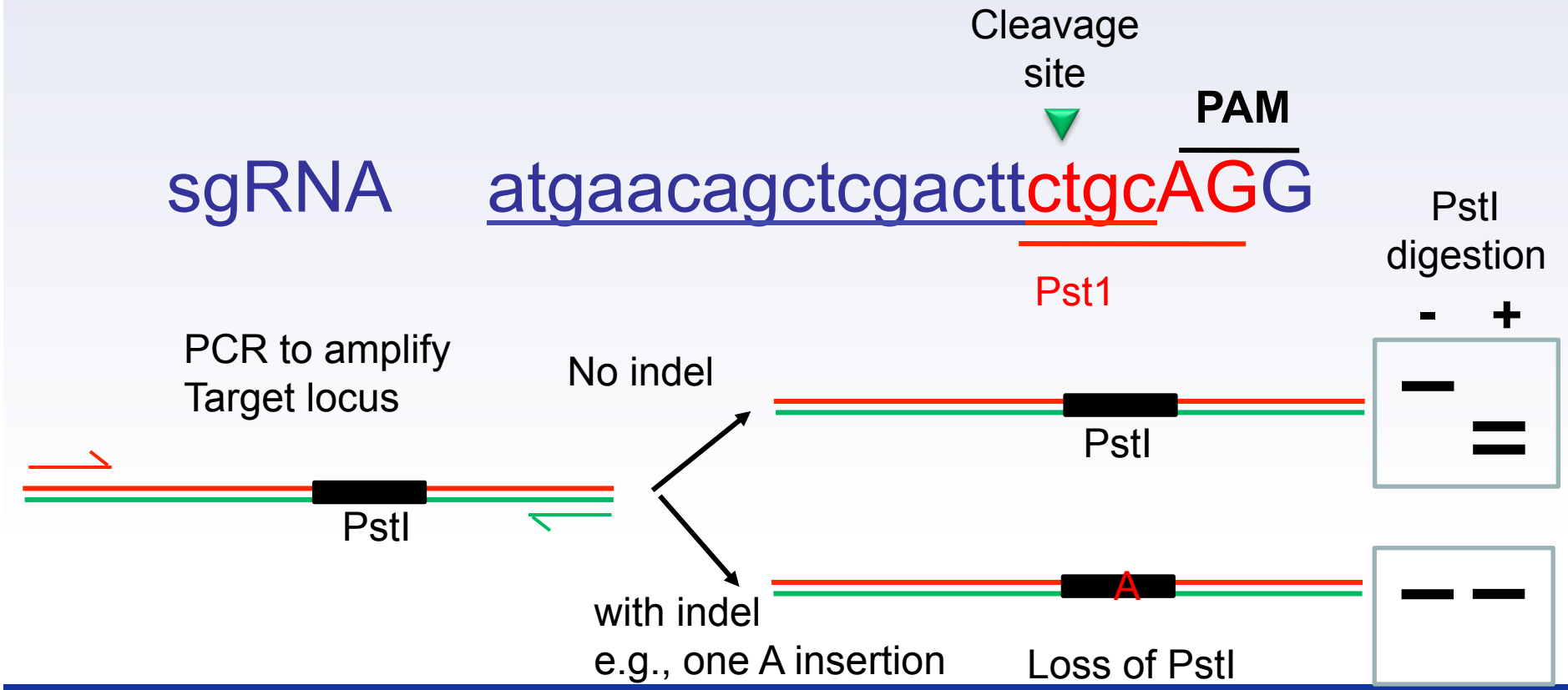

-Nickase only cleaves one strand: low mutation rate
-Requires two nickases to create mutation
-Paired nickases can have high mutation rate with greater specificity

* F. Zhang, et al Cell Aug 28, 2013

# Identify INDELs by restriction enzyme digestion

- For example

*Slide courtesy of Huan Yang*



# CHALLENGES

- Minimize off-target cleavage
- Respond rapidly to CRISPR-cas9 technology
  - Cas9 from different species
  - New off-target analysis data
  - Novel configurations
- Different methods for synthesis and delivery of nucleases to cells
  - Impose different constraints on the gRNAs
- Design nucleases to analyze closely related sequences
  - Cleave one allele but the other
  - Cleave both with similar efficiency
  - Cleave endogenous DNA but not donor DNA

# CRISPRseek

- **Facilitate design of CRISPR-Cas9 genome-editing systems**
  - **Identify target-specific gRNAs for input sequences**
  - **Identify gRNAs common to two related sequences**
  - **Identify gRNAs unique to one of two related sequences**

# FUNCTIONALITY

- Find potential gRNAs in input sequences.
  - Output gRNAs in fasta or genbank format
  - Output gRNAs in paired configuration
  - Output gRNAs annotated with restriction enzyme cut sites
- Off-target searching and scoring in a given genome for each gRNA.
  - Searching for off-targets with user defined maximum mismatch allowed
  - Using position-dependent mismatch penalty scoring system from experimental data
- Filtering gRNAs with minimum off-target cutting
- Retrieve genomic sequences flanking off-target sites and indicating whether the off-target sites are in critical region of the gene such as exon.
- Compare two input sequences to identify gRNAs that specifically target one of the sequences or both.

# DEPENDENCY

- BiocGenerics
  - S4 generic functions needed by many Bioconductor packages.
- BSgenome
  - Supplies infrastructure for efficiently representing, accessing and analyzing whole genome
- Biostrings
  - Implements functions for pattern matching, sequence alignment and string manipulation

# INSTALLATION

Install R 3.1.0

Windows: http://cran.fhcrc.org/bin/windows/base/

OS X: http://cran.fhcrc.org/bin/macosx/

Source (Linux): http://cran.fhcrc.org/sources.html

# INSTALLATION – CNT'D

CRISPRseek can be installed from R as:

```
source("http://bioconductor.org/biocLite.R")
biocLite("CRISPRseek")
```

All the dependent packages will be installed automatically

- BiocGenerics
- Biostring
- BSgenome

The organism-specific package BSgenome.Hsapiens.UCSC.hg19, TxDb.Hsapiens.UCSC.hg19.knownGene need to be installed to run the code snippets in the vignette.

```
biocLite("BSgenome.Hsapiens.UCSC.hg19")
biocLite("TxDb.Hsapiens.UCSC.hg19.knownGene")
```

# MAIN FUNCTIONS

## *offTargetAnalysis*

- Find potential gRNAs for the input sequence
  - Output gRNAs in fasta or genbank format
  - Output gRNAs in paired configuration
  - Output gRNAs with restriction enzyme cut sites
- Searching and scoring off-target sites for each gRNA
- Retrieve genomic sequences flanking off-target sites and indicating whether the off-target sites are in critical region of the gene such as exon
- Output off-target details including mismatch position and cleavage score
- Output gRNAs with topN off-target cutting scores

# MAIN FUNCTIONS – CNT'D

- compare2Sequences
  - Find potential gRNAs for both input sequences
    - Output gRNAs in fasta or genbank format
    - Output gRNAs in paired configuration
    - Output gRNAs with restriction enzyme cut sites
  - Assign relative cleavage score to each gRNA for both input sequences
  - Facilitate identification of gRNAs that specifically target one of the two input sequences
  - Facilitate identification of gRNAs that target both input sequences

# DNA TARGET SITES
# S. PYOGENES

**guide sequence /gRNA in CRISPRseek**

PAM sequence



CCACTGTGTGCACTTCATCCTGG

# DEFAULT PARAMETERS *OFFTARGETANALYSIS*

- ***offTargetAnalysis****(inputFilePath, format = "fasta", findgRNAs = TRUE, exportAllgRNAs = c("all", "fasta", "genbank", "no"), findgRNAsWithREcutOnly = TRUE, REpatternFile, minREpatternSize = 6, overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = TRUE, min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3, gRNA.size = 20, PAM = "NGG", BSgenomeName, chromToSearch = "all", max.mismatch = 4, PAM.pattern = "N[A|G]G$", gRNA.pattern = "", min.score = 0.5, topN = 100, topN.OfftargetTotalScore = 10, annotateExon = TRUE, txdb, outputDir, fetchSequence = TRUE, upstream = 200, downstream = 200, weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583), overwrite = FALSE)*

*Default setting for CRISPR-cas9 system in S. pyogenes*
gRNAf1_rs362331TStart22End44

# Constraint Guide Sequence

- By setting gRNA.pattern to require or exclude specific features within the target site.
  - For example, synthesis of gRNAs in vivo from host U6 promoters is more efficient if the first base is guanine and gRNA synthesis in vitro using T7 promoters is most efficient when the first two bases are GG
    - These features can be specified by setting gRNA.pattern = "^G" and "^GG" respectively
  - Another example is that five consecutive uracils in any position of a gRNA will affect transcription elongation by RNA polymerase III.
    - To avoid premature termination during gRNA synthesis using U6 promoter, we can set gRNA.pattern = "^(?:(?!T{5,}).)+$".
  - In addition, some studies have identified sequence features that broadly correlate with lower nuclease cleavage activity, such as uracil in the last 4 positions of the guide sequence
    - To avoid uracil in these positions, we can specify gRNA.pattern = "[ACG]{4,}.{3}$"
  - Regular expression or IUPAC Extended Genetic Alphabet to represent gRNA pattern. Type help(translatePattern) for a list of IUPAC Extended Genetic Alphabet

# DEFAULT PARAMETERS
# *OFFTARGETANALYSIS*

- ***offTargetAnalysis****(inputFilePath, format = "fasta", findgRNAs = TRUE, exportAllgRNAs = c("all", "fasta", "genbank", "no"), findgRNAsWithREcutOnly = TRUE, REpatternFile, minREpatternSize = 6, overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = TRUE, min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3, gRNA.size = 20, PAM = "NGG", BSgenomeName, chromToSearch = "all", max.mismatch = 4, PAM.pattern = "N[A|G]G$", gRNA.pattern = "", min.score = 0.5, topN = 100, topN.OfftargetTotalScore = 10, annotateExon = TRUE, txdb, outputDir, fetchSequence = TRUE, upstream = 200, downstream = 200, weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583), overwrite = FALSE)*

*Default setting for CRISPR-cas9 system in S. pyogenes*

# DEFAULT PARAMETERS *OFFTARGETANALYSIS*

- ***offTargetAnalysis****(inputFilePath, format = "fasta", findgRNAs = TRUE, exportAllgRNAs = c("all", "fasta", "genbank", "no"), findgRNAsWithREcutOnly = TRUE, REpatternFile, minREpatternSize = 6, overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = TRUE, min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3, gRNA.size = 20, PAM = "NGG", BSgenomeName, chromToSearch = "all", max.mismatch = 4, PAM.pattern = "N[A|G]G$", gRNA.pattern = "", min.score = 0.5, topN = 100, topN.OfftargetTotalScore = 10, annotateExon = TRUE, txdb, outputDir, fetchSequence = TRUE, upstream = 200, downstream = 200, weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583), overwrite = FALSE)*

*Default setting for CRISPR-cas9 system in S. pyogenes*

REpatternFile <- system.file('extdata', 'NEBenzymes.fa', package= 'CRISPRseek')

# DEFAULT PARAMETERS
# *OFFTARGETANALYSIS*

- ***offTargetAnalysis****(inputFilePath, format = "fasta", findgRNAs = TRUE, exportAllgRNAs = c("all", "fasta", "genbank", "no"), findgRNAsWithREcutOnly = TRUE, REpatternFile, minREpatternSize = 6, overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = TRUE, min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3, gRNA.size = 20, PAM = "NGG", BSgenomeName, chromToSearch = "all", max.mismatch = 4, PAM.pattern = "N[A|G]G$", gRNA.pattern = "", min.score = 0.5, topN = 100, topN.OfftargetTotalScore = 10, annotateExon = TRUE, txdb, outputDir, fetchSequence = TRUE, upstream = 200, downstream = 200, weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583), overwrite = FALSE)*

*Default setting for CRISPR-cas9 system in S. pyogenes*
*Position specific penalty score matrix is experimentally determined by Feng Zhang's lab at MIT*
*Alternative weights = c(0, 0, 0, 0, 0, 0.311, 0.329, 0, 0.47, 0.011, 0.335, 0.395, 0.584, 0.829, 0.762, 0.795, 0.67, 0.816, 0.74, 0.656) Hsu et al., 2013*

# DEFAULT PARAMETERS
# *OFFTARGETANALYSIS*

- ***offTargetAnalysis****(inputFilePath, format = "fasta", findgRNAs = TRUE, exportAllgRNAs = c("all", "fasta", "genbank", "no"), findgRNAsWithREcutOnly = TRUE, REpatternFile, minREpatternSize = 6, overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = TRUE, min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3, gRNA.size = 20, PAM = "NGG", BSgenomeName, chromToSearch = "all", max.mismatch = 4, PAM.pattern = "N[A|G]G$", gRNA.pattern = "", min.score = 0.5, topN = 100, topN.OfftargetTotalScore = 10, annotateExon = TRUE, txdb, outputDir, fetchSequence = TRUE, upstream = 200, downstream = 200, weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583), overwrite = FALSE)*

*Default setting for CRISPR-cas9 system in S. pyogenes*
*Position specific penalty score matrix is experimentally determined by* Feng Zhang's lab at MIT
Alternative weights = c(0, 0, 0, 0, 0, 0.311, 0.329, 0, 0.47, 0.011, 0.335, 0.395, 0.584, 0.829, 0.762, 0.795, 0.67, 0.816, 0.74, 0.656)

# Output Files

- gRNAs (genbank, fasta)
- Restriction Site Overlap (tab delimited)
- Paired Sites (tab delimited)
- Off-Target sites (tab delimited)
- Summary (tab delimited)

# DEFAULT PARAMETERS COMPARE2SEQUENCES

- compare2Sequences(inputFile1Path, inputFile2Path, format = "fasta", findgRNAsWithREcutOnly = FALSE, REpatternFile, minREpatternSize = 6, overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = FALSE, min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3, gRNA.size = 20, PAM = "NGG", PAM.pattern = "N[A|G]G$", max.mismatch =4, outputDir, weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583), overwrite = FALSE)

# Caveat

- ***compare2Sequences*** might output no match or more than one match to the alternative input sequence for each gRNAs identified for each input sequence depending on max.mismatch (default is 4 mismatches allowed)
  - Solution
    - Suggest sort the output by gRNAplusPAM and scoreDiff to examine possible multiple off-target sites in the alternative sequence, if you aim to identify gRNAs to target one of the two input sequences only.

# New Features and Future Plan

- Model effect of PAM variants, e.g., NAG vs. NGG
- Model effect of different base changes, e.g., A->G vs. C->G
- Annotate off-targets with gene name
- Speed up off-target search

# REFERENCE AND HELP

- ?CRISPRseek in a R session
- browseVignettes(”CRISPRseek")
- Zhu LJ*, Holmes BR, Aronin N and Brodsky MH*. (2010) [* denotes co-corresponding author] **CRISPRseek: a Bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems**. PloS One (Accepted)
- Email: bioconductor <bioconductor@stat.math.ethz.ch>

# ACKNOWLEDGEMENT


- The Bioconductor core members
  - Hervé Pagès
  - Martin Morgan
- University of Massachusetts
  - Michael Brodsky (PGFE and PMM)
  - Neil Aronin (RNA Therapeutics Institute and Department of Medicine)
- Broad Institute of MIT and Harvard
  - Benjamin Holmes
  - Feng Zhang

# DEMO & EXERCISE

# RSTUDIO IN AMAZON CLOUD

We will be using an Amazon Machine Instance (AMI) on an Ubuntu Linux machine that is pre-configured with RELEASE version of Bioconductor (2.14)

http://ec2-54-83-134-17.compute-1.amazonaws.com
User name: ubuntu
Password: bioc
http://www.bioconductor.org/help/course-materials/2014/BioC2014/CRISPRdemo.Rmd