
University of Massachusetts Medical School

# CRISPRseek and GUIDEseq packages for CRISPR-Cas9 Genome Editing Studies

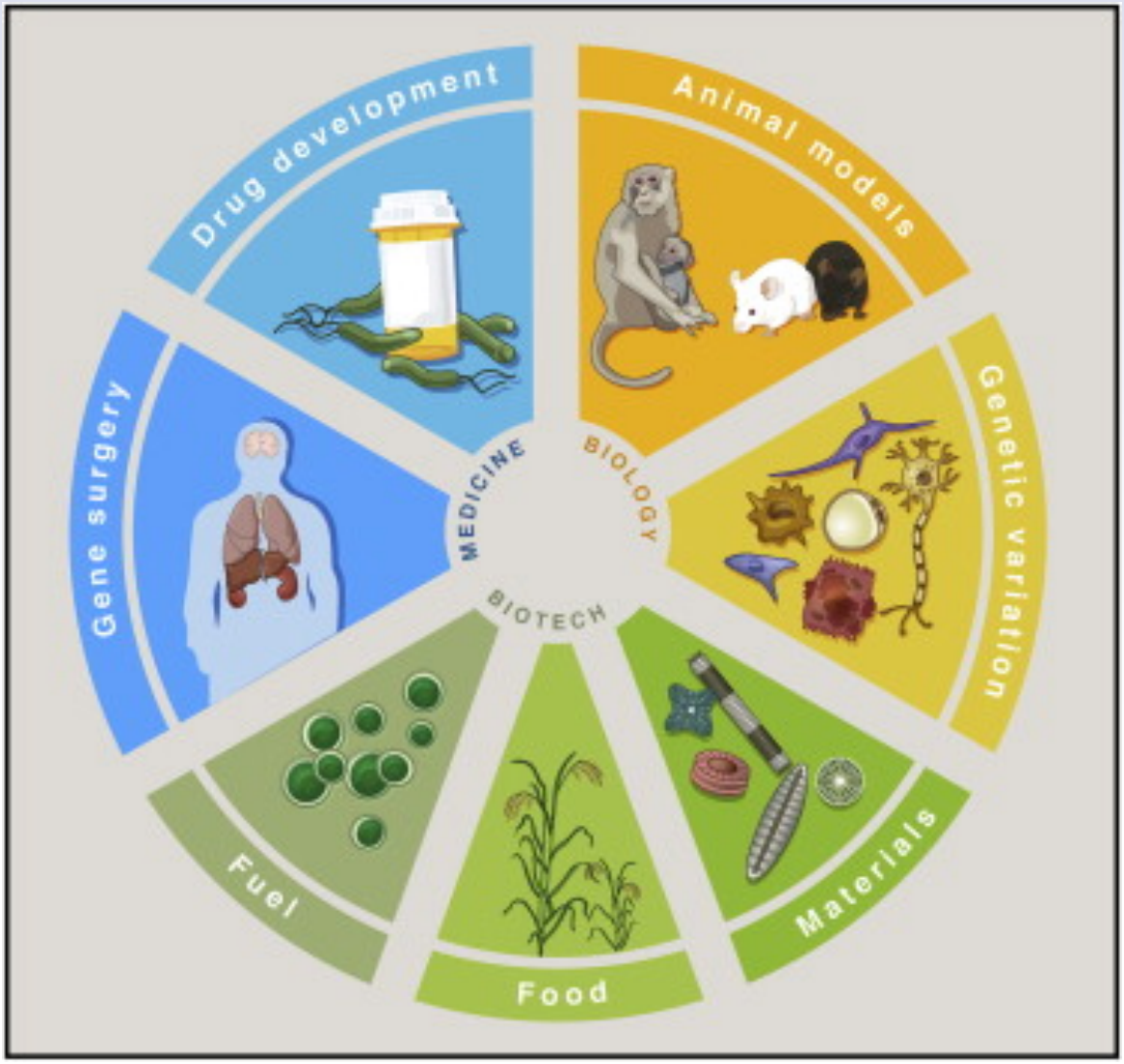
July 27th 2017

Lihua Julie Zhu, PhD

Research Professor and Head of Bioinformatics Core

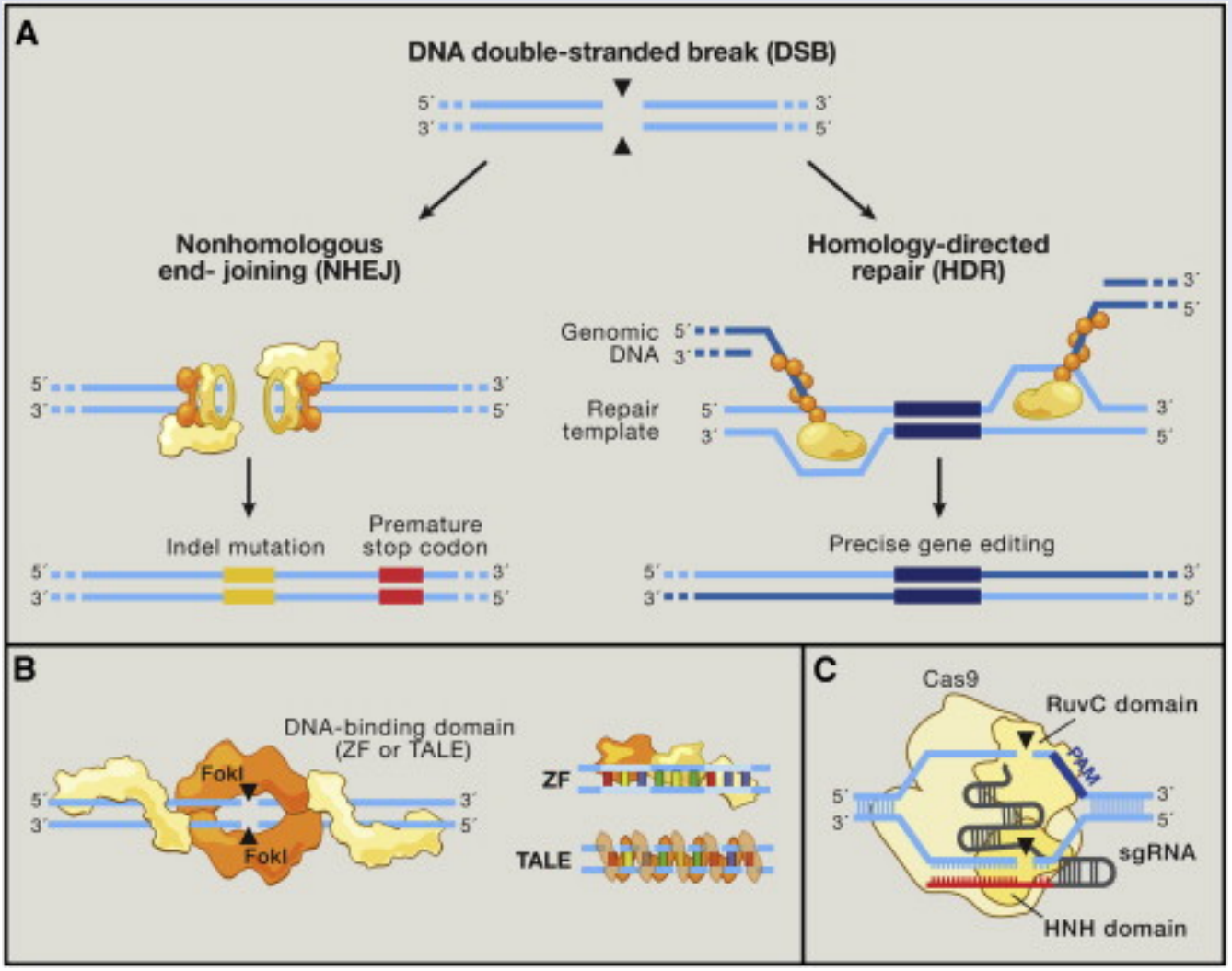
MCCB, Umass Medical School

# Genome Editing Applications



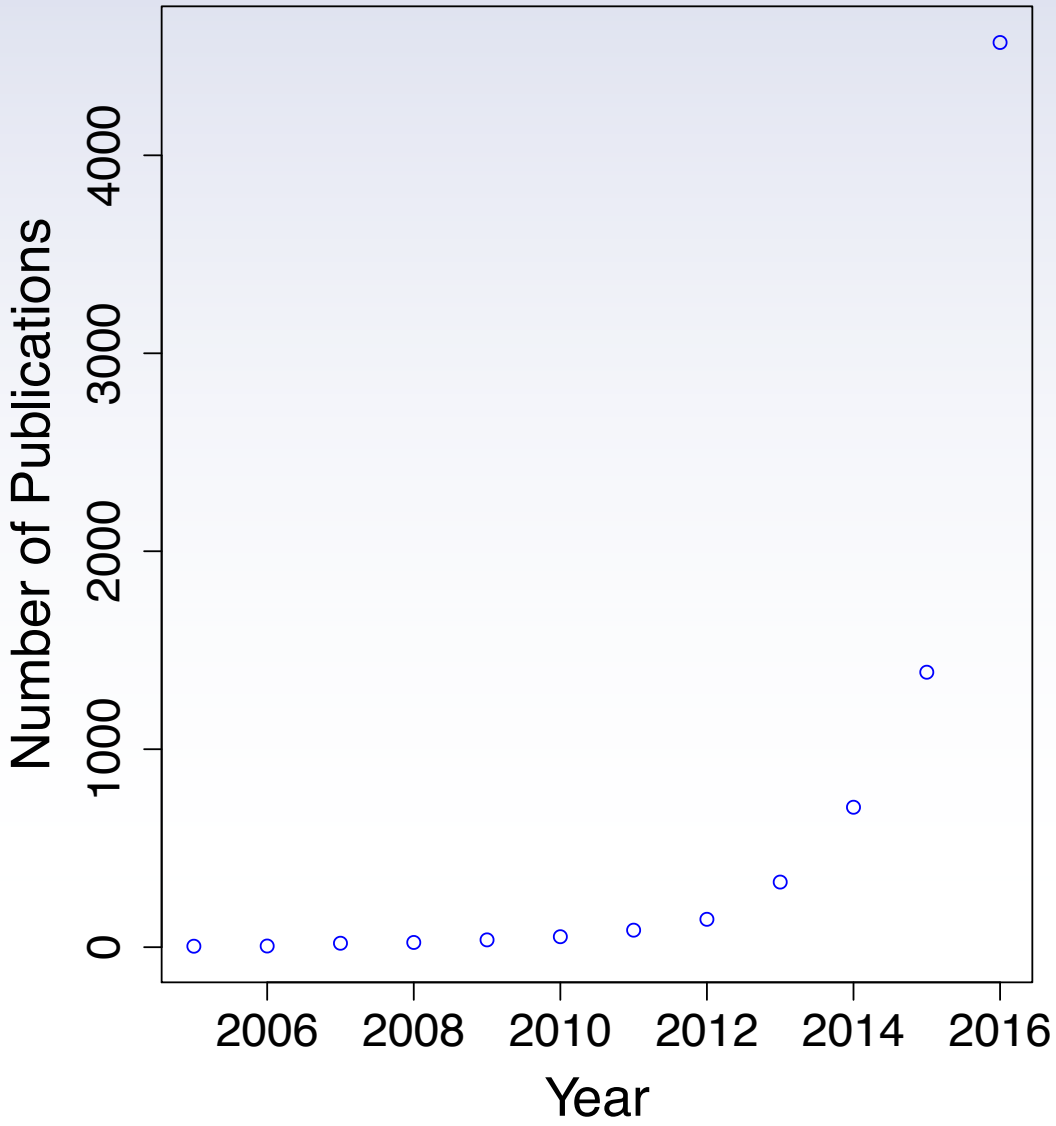

Adapted from Hsu,PD et al., Cell. Volume 157, Issue 6, 5 June 2014, Pages 1262-1278

# Genome Editing



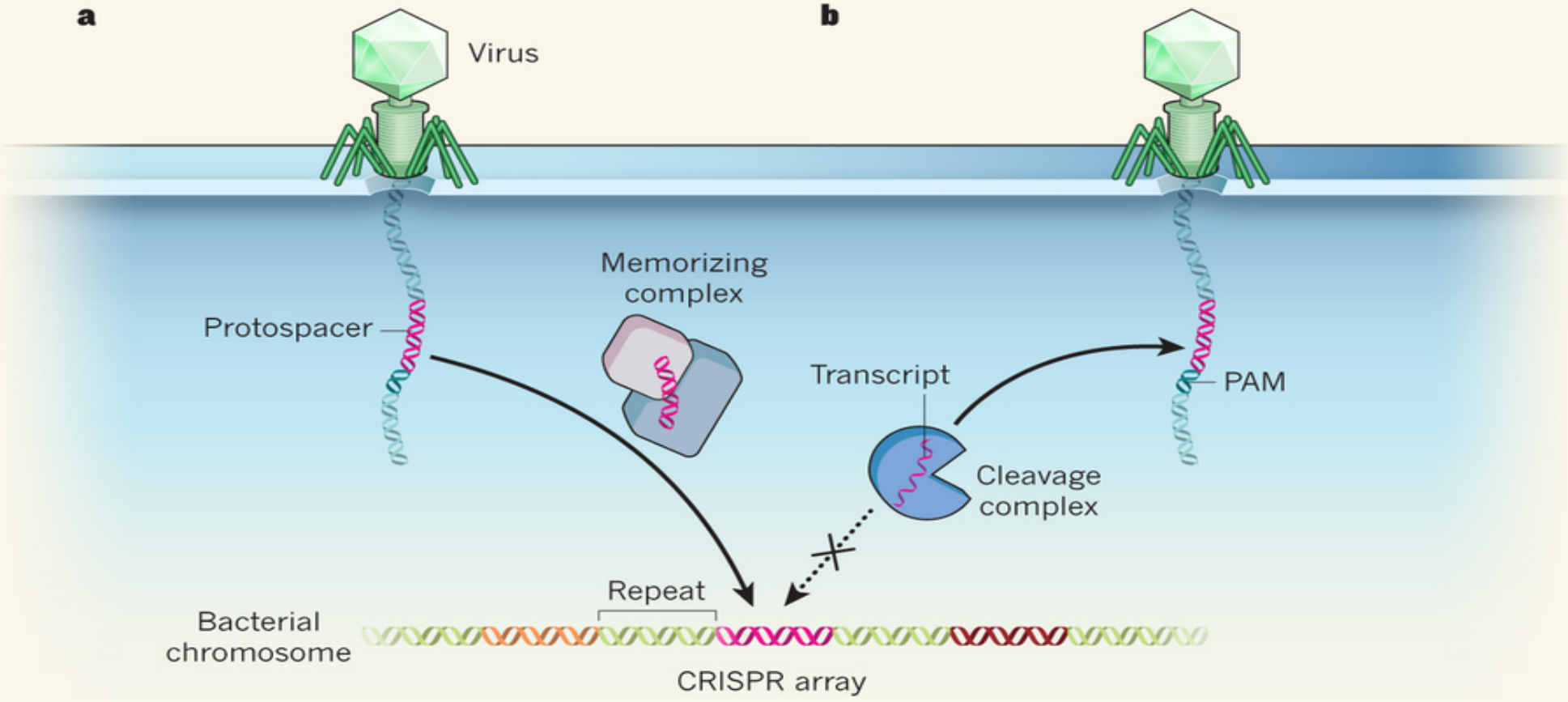

Adapted from Hsu,PD et al., Cell. Volume 157, Issue 6, 5 June 2014, Pages 1262-1278

# CRISPR-Cas9 Publications



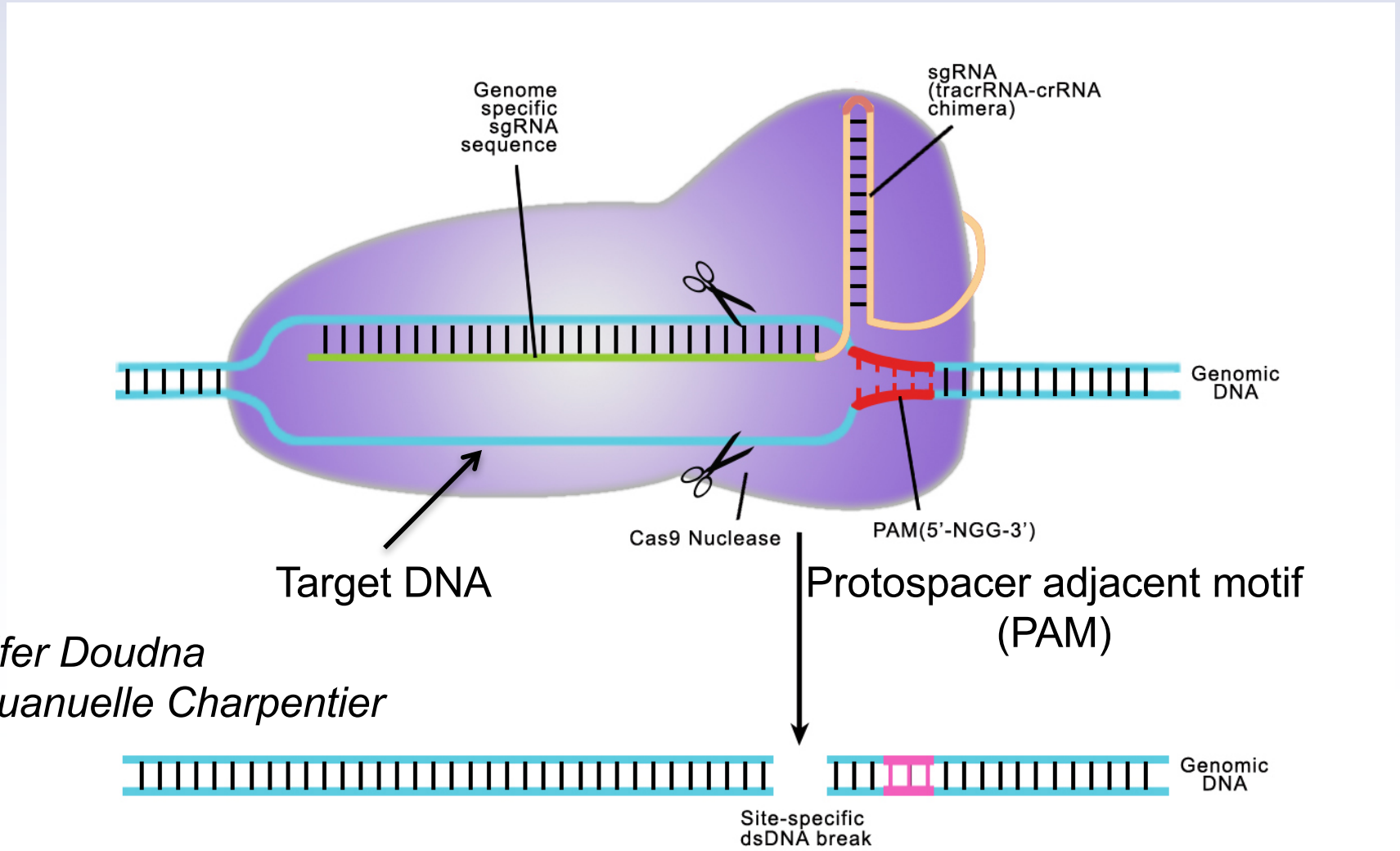
# CRISPR-Cas9 System

An adaptive immune defense system found in bacteria




*Adapted from Nature* **519**, 166–167 (12 March 2015) Clustered Regularly Interspaced Short Palindromic Repeats

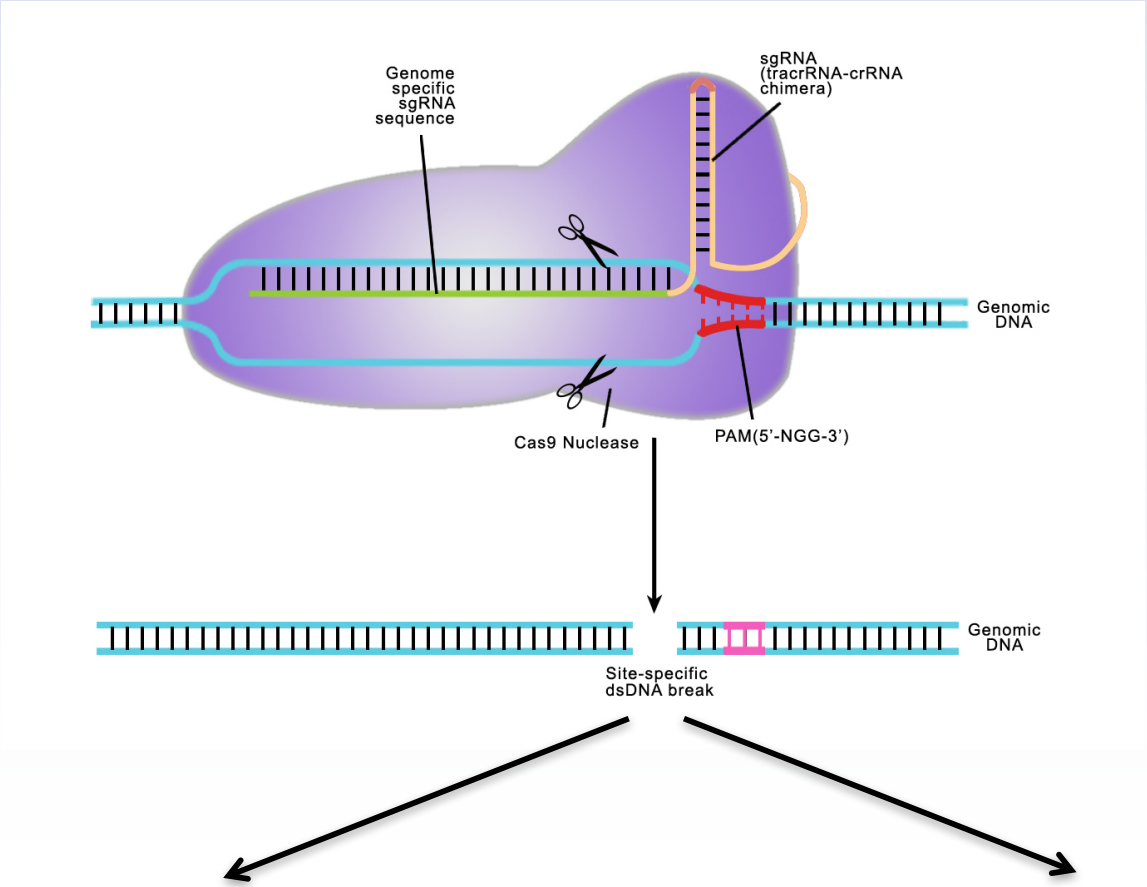
# Genome Editing with CRISPR-Cas9 System




*Jennifer Doudna*
*Emmuanuelle Charpentier*

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

# Genome Editing with CRISPR-Cas9 System




Gene disruption (NHEJ)

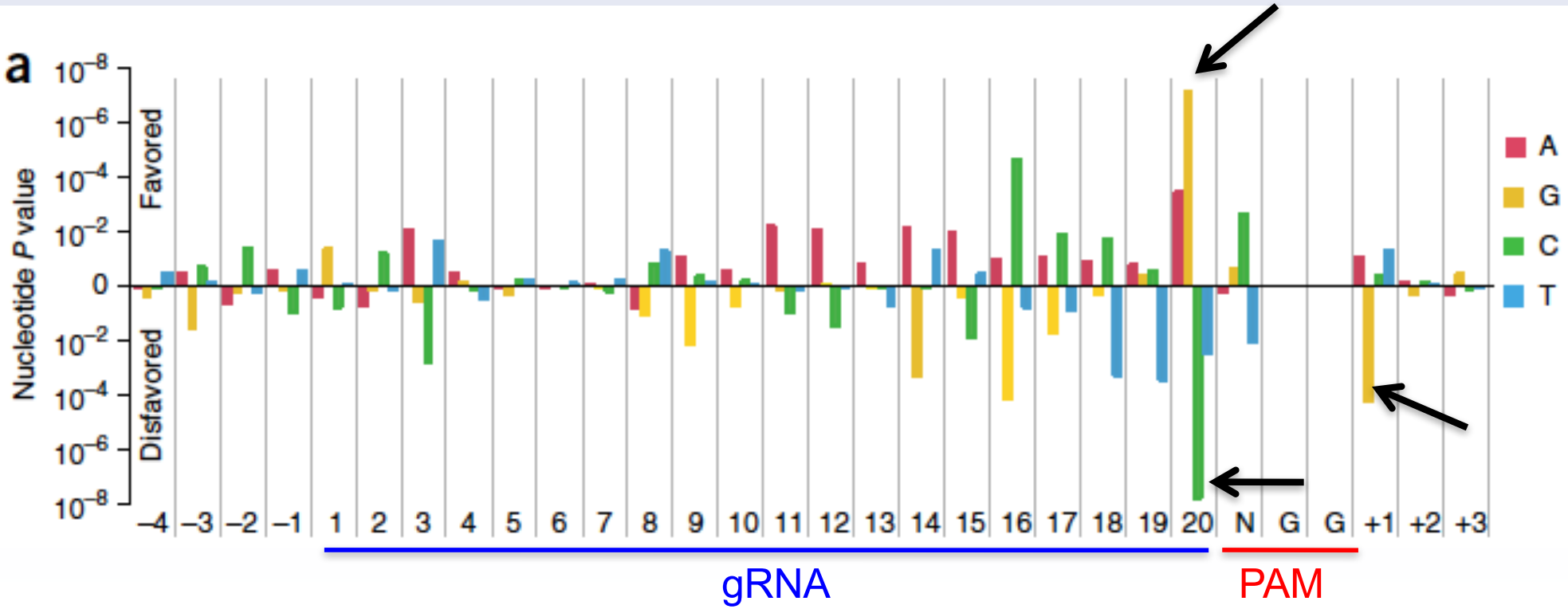
Targeted modification (HDR)

## CRISPRseek

1. High on-target cleavage (high efficacy/efficiency)

1. Low off-target cleavage (high specificity)

# Features that Predict On-target Efficacy of CRISPR-Cas9 System from S. pyogenes (SpCas9)

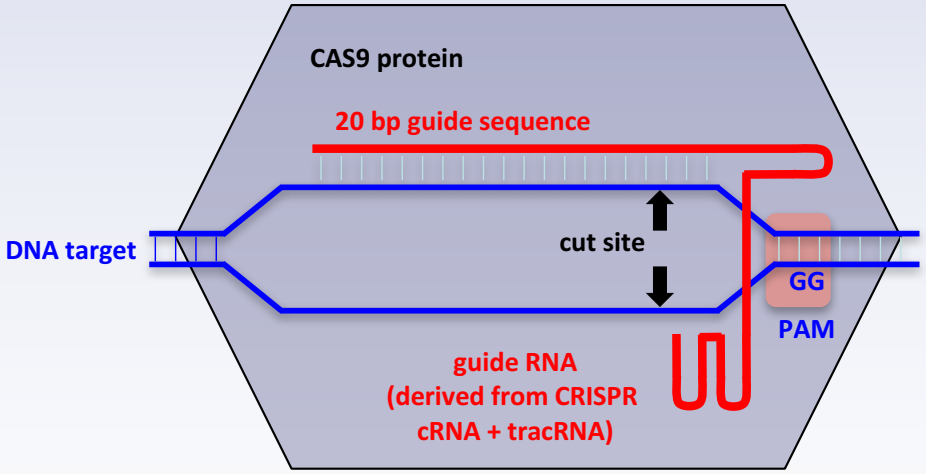



-Assayed 1841 gRNAs tiling across all possible target sites of a panel of 6 mouse genes and 3 human genes

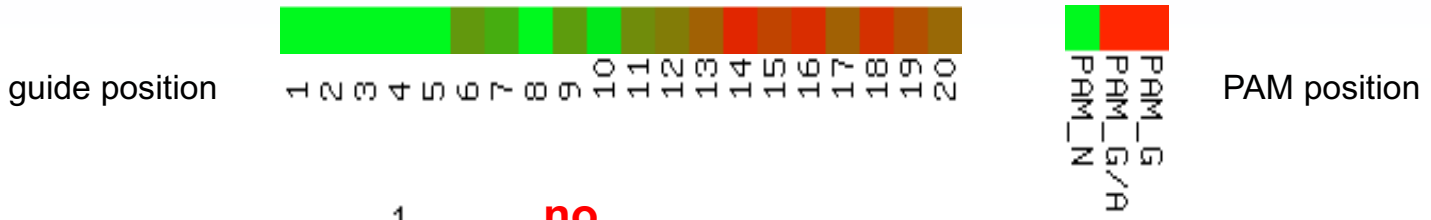
-Construct a predictive model of efficacy with 72 features

Doench, et al Nbt Aug 21, 2014

# Mismatch Number and Positions Affect Off-target Cleavage in SpCas9

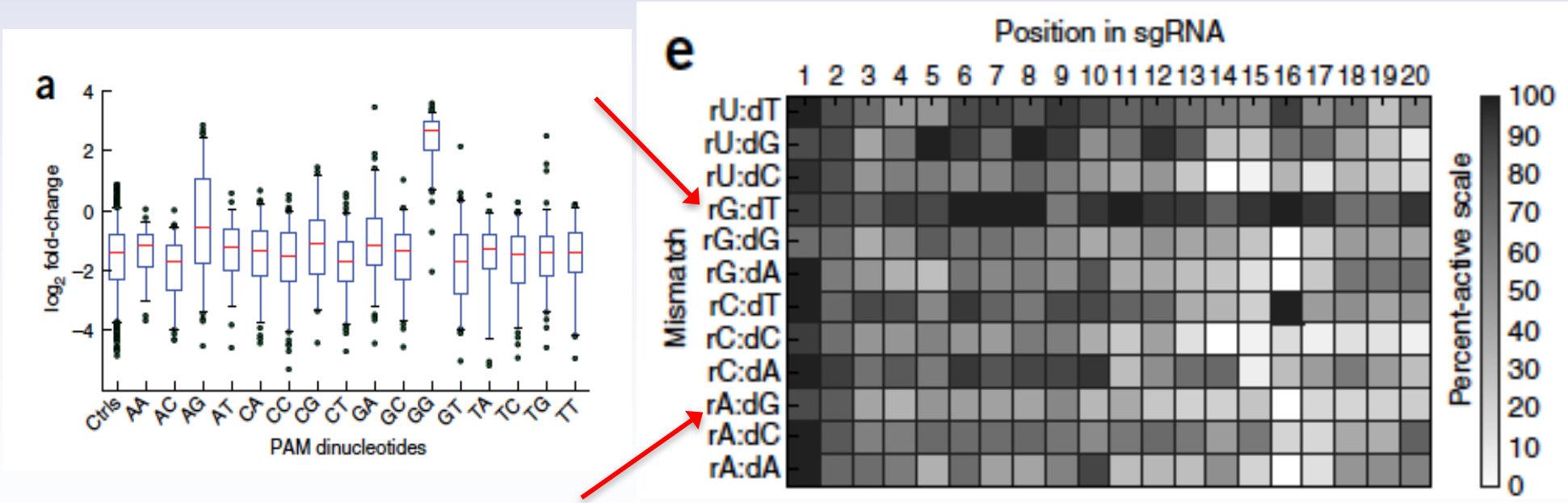



Tested >700 gRNA variants for 15 target sites within EMX1 gene in human cell line
Transfected the cells with gRNAs containing all possible single mismatches and compared to gRNAs without mismatch




P. Hsu, et al., Nbt, July 2013

# Mismatch Type Also Affects Off-target Cleavage in SpCas9




Heat-map of the percent-active values for all sgRNA-DNA interactions including all possible one-nucleotide mismatches

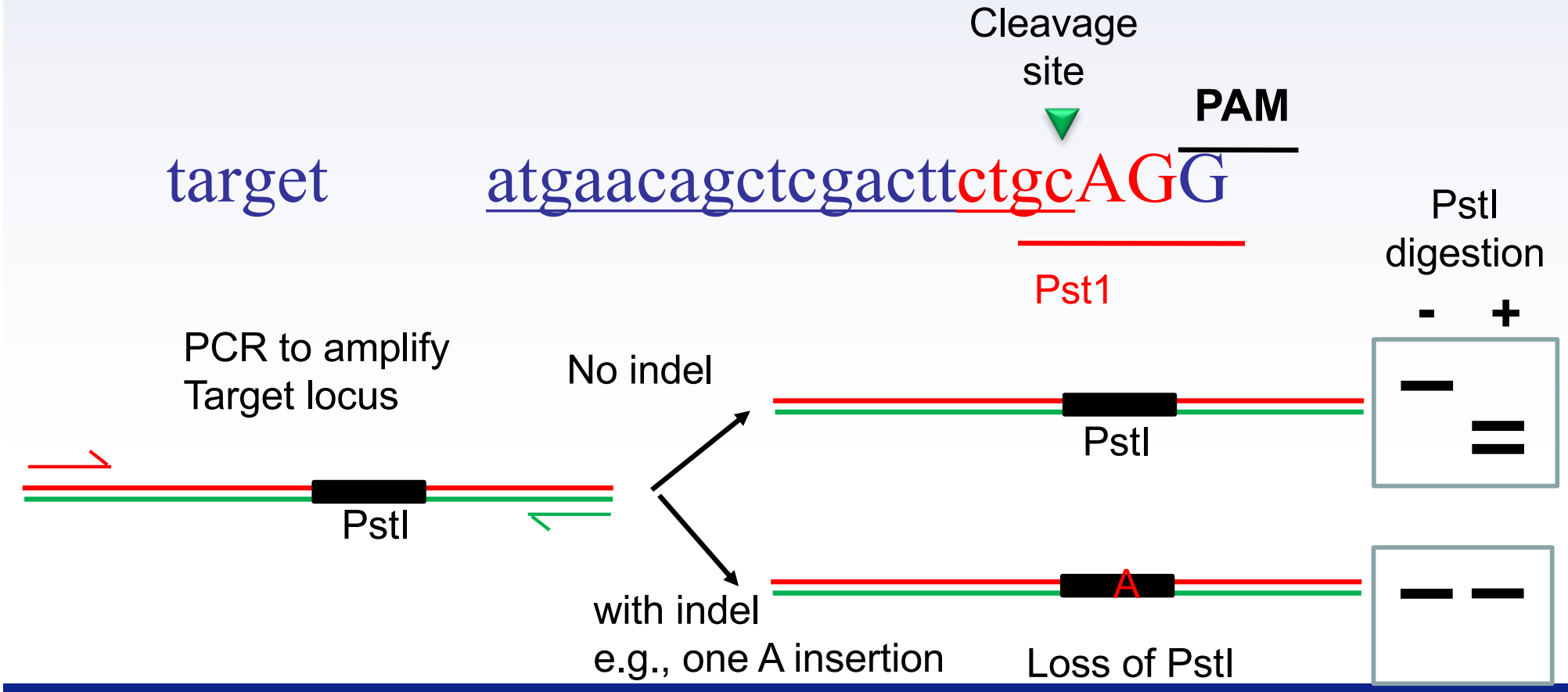
Assayed 65 perfect-match gRNAs targeting human CD33 gene and their 4,290 gRNAs with one mismatch

JG. Doench, et al., Nbt, Jan 2016

# Identify INDELs by Restriction Enzyme Digestion

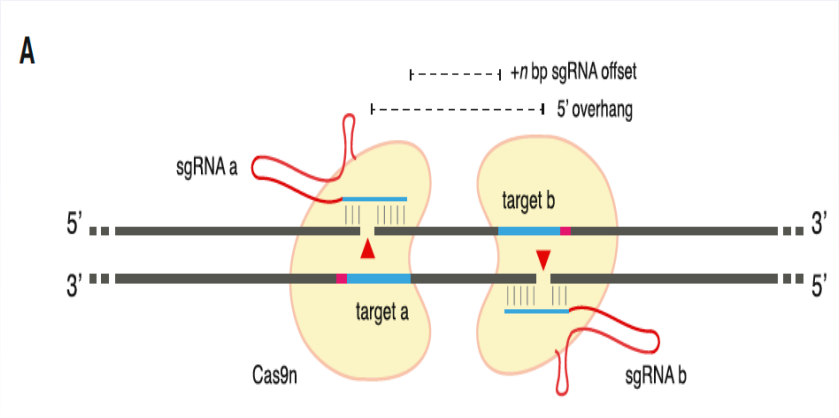
- For example

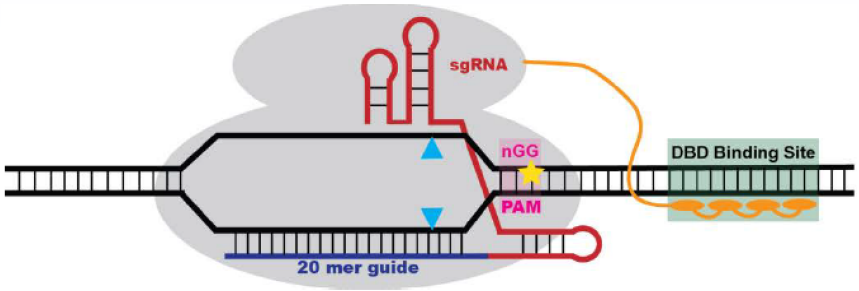
*Slide courtesy of Huan Yang*



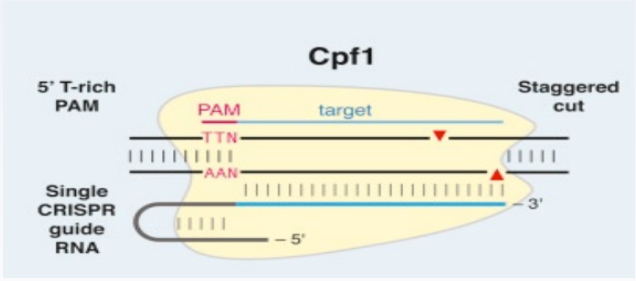
# Evolving CRISPR-Cas9 Technology

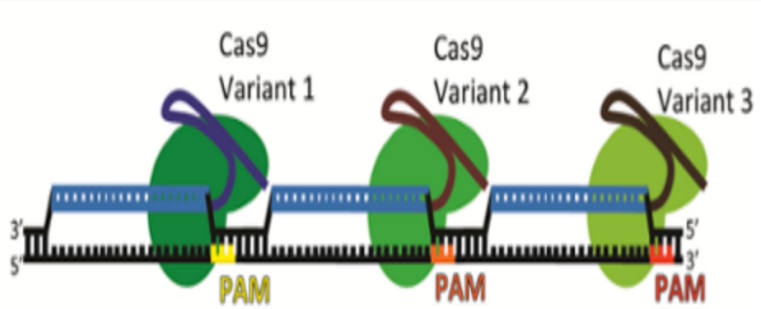
- Alternative configuration
  - Paired nickases
  - dCas9-fokI dimers
  - Cas9-ZFP fusion




- Cas9 variants
  - Different bacteria species
  - SpCas9 mutant
- New Nucleases




- Zhang et. al. Cell. Aug 2013
- Tsai et al. nbt. Feb 2014
- *Bolukbasi* et al. Nat Methods. Dec 2015
- Kleinstiver et al. Nature 2015

# Design Goals For CRISPRseek

- Identify gRNAs with high on-target and low off-target cleavage
- Respond rapidly to CRISPR-Cas9 technology
  - Cas9 from different species
  - Novel configurations (paired nickases and dCas9-FokI dimers)
  - Alternative scoring model from newly published data
- Accommodate different methods for synthesis and delivery of nucleases to cells
  - Impose different constraints on the gRNAs
- Monitor cleavage
  - Restriction site
- Design gRNAs to analyze closely related sequences
  - Cleave one allele but not the other or both

# Main Functions of CRISPRseek

## *offTargetAnalysis workflow*

- gRNA searching and off-target analysis and annotation for one or a set of input sequences

## *compare2Sequences workflow*

- Identify gRNAs that specifically target one of the two input sequences or both


chr1 (q25.1)
33
31.3
1p31.1
1q12
32.1
Window Position
Human Feb. 2009 (GRCh37/hg19) chr1:173,735,071-173,735,193 (123 bp)
Scale
50 bases
hg19
chr1:
173,735,080
173,735,090
173,735,100
173,735,110
173,735,120
173,735,130
173,735,140
173,735,150
173,735,160
CRISPRseek
Alcds2_gR252f
Alcds2_gR241r
Alcds2_gR240r
Alcds2_gR244f
Alcds2_gR233f
Alcds2_gR213r
Alcds2_gR217f
Alcds2_gR196f
Alcds2_gR195f
Alcds2_gR216f
Alcds2_gR194f
Alcds2_gR215f
Alcds2_gR191f
Alcds2_gR212f
Alcds2_gR171r
Alcds2_gR192r
Alcds2_gR217f
Alcds2_gR196f
Alcds2_gR195f
Alcds2_gR216f
Alcds2_gR215f
Alcds2_gR194f
Alcds2_gR212f
Alcds2_gR191f

# Summary of gRNAs

| gRNAsPlusPAM | toViewInUCSC | target efficacy | Top10 offtarget Score | RE site |
|---|---|---|---|---|
| CTACTGTGTGCACTTCATCCTGG | chr4:3215834-3215856 | 0.0768 | 20 | |
| TGAAGTGCACACAGTAGATGAGG | chr4:3215828-3215850 | 0.2668 | 21.9 | |
| GAAGTGCACACAGTAGATGAGGG | chr4:3215827-3215849 | 0.3687 | 11.9 | |
| CACACAGTAGATGAGGGAGCAGG | chr4:3215821-3215843 | 0.2049 | 31.5 | BsaXI Cac8I |
| AGTAGATGAGGGAGCAGGCGTGG | chr4:3215816-3215838 | 0.6173 | 12.3 | BsaXI Cac8I |
| GTAGATGAGGGAGCAGGCGTGGG | chr4:3215815-3215837 | 0.6881 | 15.8 | |

# Detailed Information of Off-targets

| gRNAPlusPAM | OffTargetSequence | inExon | inIntron | entrez_id | gene | score | n.mismatch | mismatch | alignment | NGG | strand | chrom | chromStart | chromEnd |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CCAGTTTGTGGATCCTGC | CCAGTTTGTGGATCCTGCTC | TRUE | | 2623 | GATA1 | 100 | 0 | | ..... | 1 | + | chrX | 48649564 | 48649586 |
| CCAGTTTGTGGATCCTGC | ACAATTTCTGGATCCTGCTCCAG | | TRUE | | | 2.6 | 3 | 20,17,13 | A..A...C..... | 0 | - | chrX | 70453482 | 70453504 |
| CCAGTTTGTGGATCCTGC | GCAGTTTATGGATCCTGCTGGAG | | | | | 1.5 | 3 | 20,13,1 | G.....A.....( | 0 | + | chr11 | 128187317 | 128187339 |
| CCAGTTTGTGGATCCTGC | TGAGTTTGTGGTTCCTGCTCTGG | | TRUE | | | 1.4 | 3 | 20,19,9 | TG.....T..... | 1 | - | chr4 | 115807036 | 115807058 |
| CCAGTTTGTGGATCCTGC | ACAGATTGTGGATCCTCCTCTGG | | TRUE | | | 1.3 | 3 | 20,16,4 | A...A.....C.. | 1 | + | chr5 | 58007599 | 58007621 |
| CCAGTTTGTGGATCCTGC | TCACTTTGTGGACCCTGCTCTGG | | | | | 1.2 | 3 | 20,17,8 | T..C.....C..... | 1 | + | chr11 | 127110216 | 127110238 |
| CCAGTTTGTGGATCCTGC | CCAGGTTGTTGAGCCTGCTCAAG | | TRUE | | | 1 | 3 | 16,11,8 | ....G....T..G..... | 0 | + | chr6 | 124177937 | 124177959 |
| CCAGTTTGTGGATCCTGC | ACAGTATGTGAATCCTGCTCCGG | | | | | 0.9 | 3 | 20,15,10 | A....A....A..... | 1 | - | chr8 | 54555407 | 54555429 |
| CCAGTTTGTGGATCCTGC | ACAGTGTGTGAATCCTGCTCTGG | | | | | 0.9 | 3 | 20,15,10 | A....G....A..... | 1 | - | chr4 | 136302357 | 136302379 |
| CCAGTTTGTGGATCCTGC | TCAGTGTGTGGTTCCTGCTCCAG | | | | | 0.9 | 3 | 20,15,9 | T....G.....T..... | 0 | - | chr8 | 58413423 | 58413445 |
| CCAGTTTGTGGATCCTGC | CCACTTTGGGGTTCCTGCTCCGG | | TRUE | | | 0.8 | 3 | 17,12,9 | ...C....G..T..... | 1 | - | chr15 | 89911561 | 89911583 |
| CCAGTTTGTGGATCCTGC | GTAGTTTGTGGATCCTGTTCTAG | | TRUE | | | 0.7 | 3 | 20,19,3 | GT.....T. | 0 | + | chrX | 10161855 | 10161877 |
| CCAGTTTGTGGATCCTGC | CCAGTTTTTGGACCCTGCTGGAG | | TRUE | | | 0.5 | 3 | 13,8,1 | .....T....C.....( | 0 | + | chr3 | 76589098 | 76589120 |

# Improved Off-target Scores Can Account For Differences In Allele Specificity

```
Target allele        2331(T)     GACCCACGCCTGCTCCCTCATCTACTGTGTGCACTTCATCCTGGA
Non-target allele    2331(C)                          C
```

**gRNA: 16T** CCTCATCTACTGTGTGCACTTCA

**gRNA: 16C** CCTCATCCACTGTGTGCACTTCA

| name | gRNAPlusPAM | targetSeqName | guideAlignment2OffTarget | scoreForSeq1 | scoreForSeq2 | scoreDiff |
|---|---|---|---|---|---|---|
| **rs363099C_gR38f** | **AGGGTTTCTCCGCTCAGCCTTGG** | **rs363099C** | .........T......... | **100** | **92.1** | **7.9** |
| rs363099T_gR45r | TATCTGAGAAAGAACATCCAAGG | rs363099T | ..... | 100 | 100 | 0 |
| rs363099T_gR19f | ACAGCACGGAAAAGTTTGGAGGG | rs363099T | ..... | 100 | 100 | 0 |
| rs363099T_gR18f | AACAGCACGGAAAAGTTTGGAGG | rs363099T | ..... | 100 | 100 | 0 |
| rs363099C_gR45r | TATCTGAGAAAGAACATCCAAGG | rs363099C | ..... | 100 | 100 | 0 |
| rs363099C_gR19f | ACAGCACGGAAAAGTTTGGAGGG | rs363099C | ..... | 100 | 100 | 0 |
| rs363099C_gR18f | AACAGCACGGAAAAGTTTGGAGG | rs363099C | ..... | 100 | 100 | 0 |
| **rs363099T_gR38f** | **AGGGTTTCTTCGCTCAGCCTTGG** | **rs363099T** | .........C......... | **92.1** | **100** | **-7.9** |

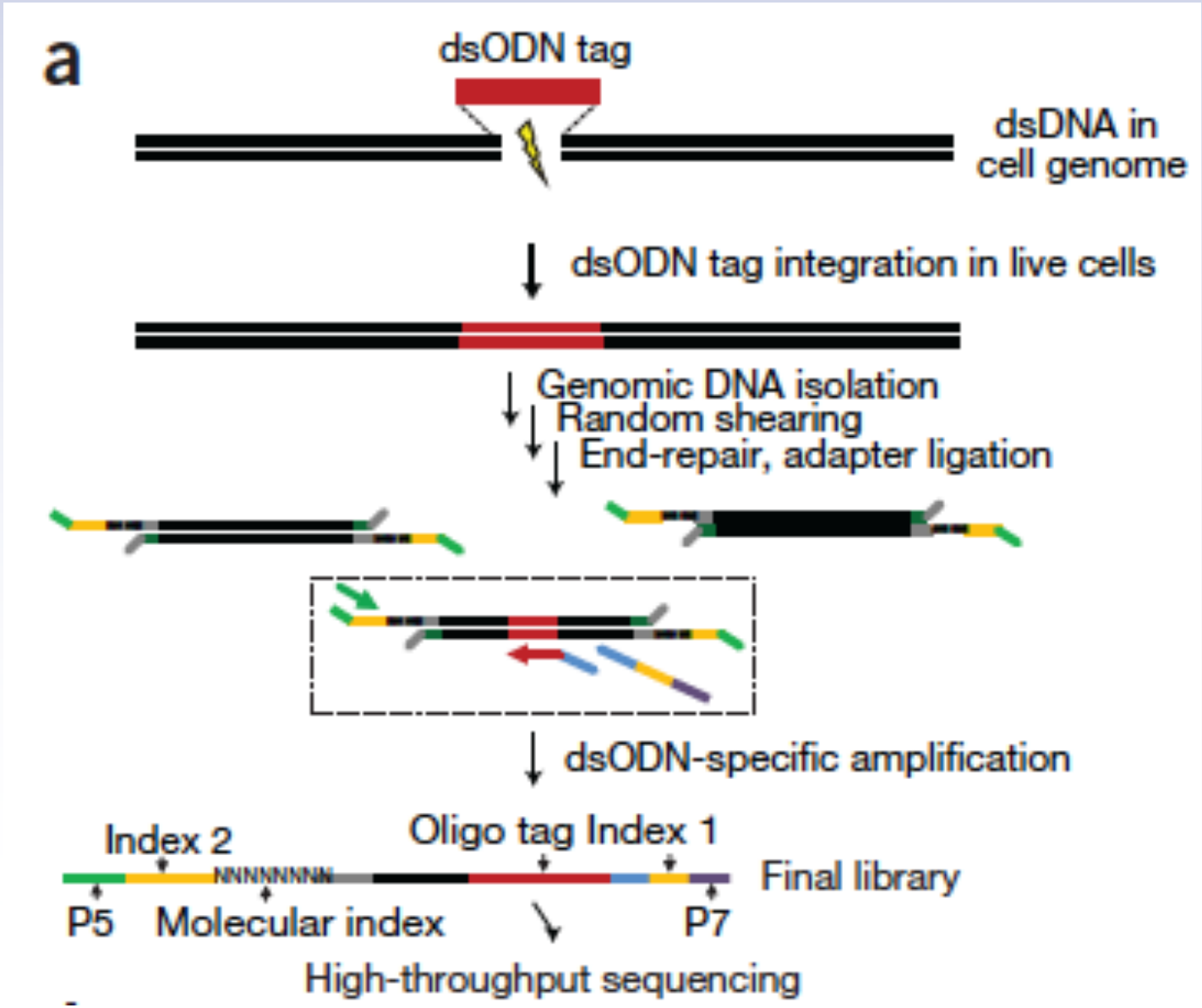
| | | | | | | |
|---|---|---|---|---|---|---|
| **rs363099C_gR38f** | **AGGGTTTCTCCGCTCAGCCTTGG** | **rs363099C** | .........T......... | **1** | **0.941176** | **0.0588** |
| rs363099T_gR45r | TATCTGAGAAAGAACATCCAAGG | rs363099T | ..... | 1 | 1 | 0 |
| rs363099T_gR19f | ACAGCACGGAAAAGTTTGGAGGG | rs363099T | ..... | 1 | 1 | 0 |
| rs363099T_gR18f | AACAGCACGGAAAAGTTTGGAGG | rs363099T | ..... | 1 | 1 | 0 |
| rs363099C_gR45r | TATCTGAGAAAGAACATCCAAGG | rs363099C | ..... | 1 | 1 | 0 |
| rs363099C_gR19f | ACAGCACGGAAAAGTTTGGAGGG | rs363099C | ..... | 1 | 1 | 0 |
| rs363099C_gR18f | AACAGCACGGAAAAGTTTGGAGG | rs363099C | ..... | 1 | 1 | 0 |
| **rs363099T_gR38f** | **AGGGTTTCTTCGCTCAGCCTTGG** | **rs363099T** | .........C......... | **0.533333** | **1** | **-0.4667** |

# Genome-wide Unbiased Identification of DSBs enabled by sequencing

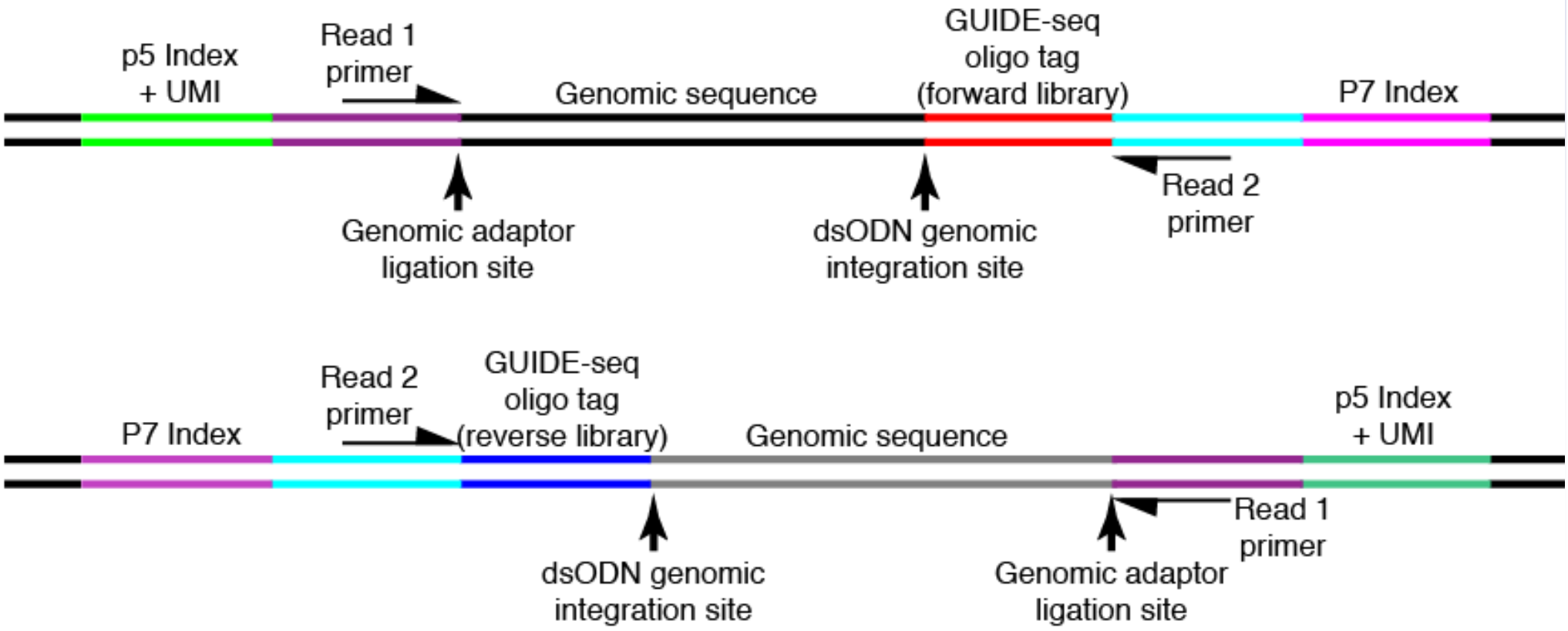
## GUIDE-seq



Tsai *et al*. Nat. Biotech, 33(2), 187–197 2015.

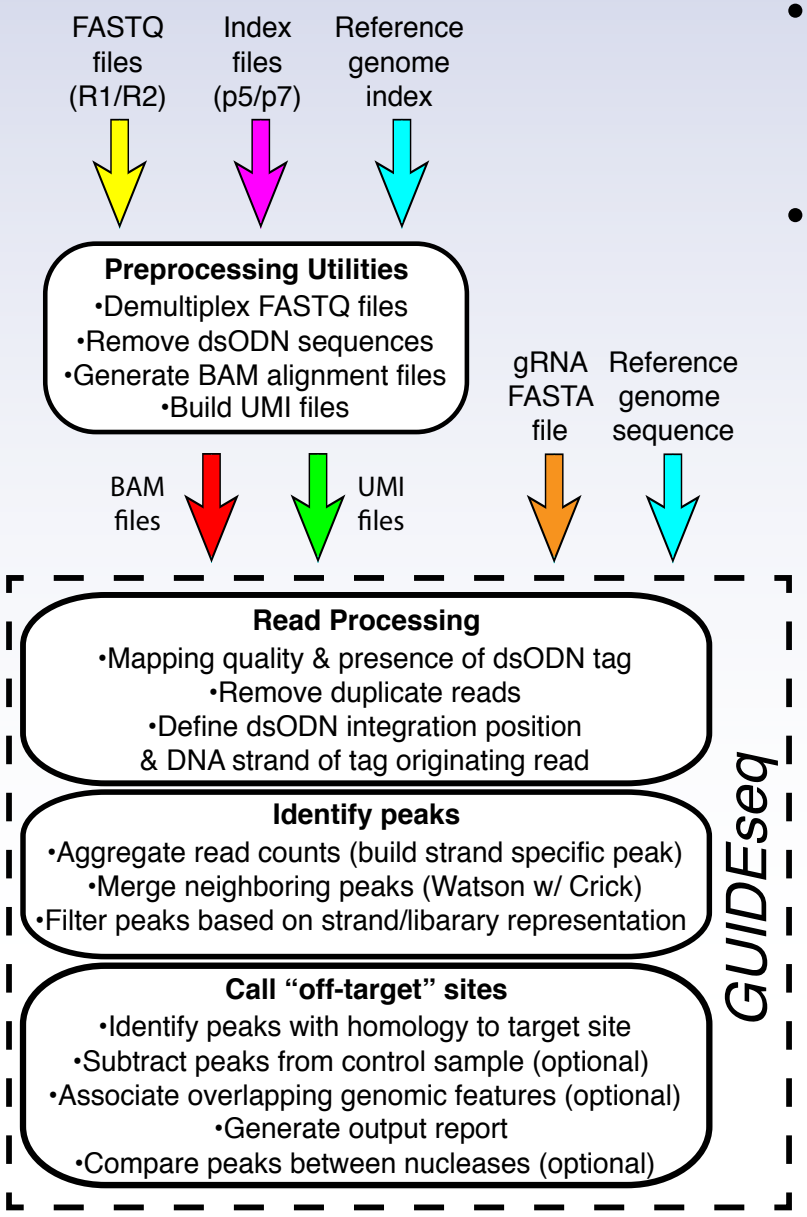



Tsai SQ 2015 Nature Method



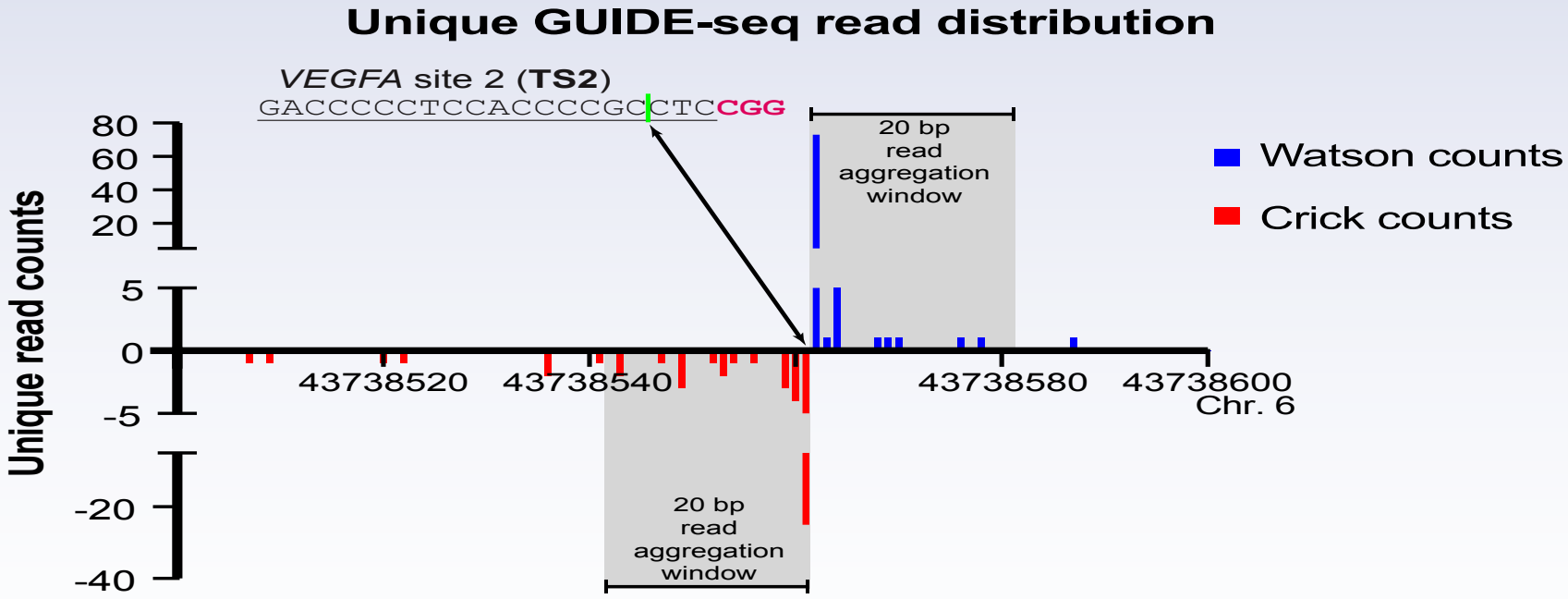

Adapted from Zhu et al., 2017 BMC Genomics **18**(1)

**Figure 1**




- Preprocessing scripts at http://mccb.umassmed.edu/GUIDE-seq/
- Additional File 1: Preprocessing steps to generate alignment and umi files as input for GUIDEseq (Zhu LJ, Lawrence M, Gupta A, Pages H, Kucukural A, Garber M and Wolfe SA. 2017. GUIDEseq: a bioconductor package to analyze GUIDE-Seq datasets for CRISPR-Cas nucleases. *BMC Genomics*, **18**(1))

Adapted from Zhu et al., 2017
BMC Genomics **18**(1)




Adapted from Zhu et al., 2017 BMC Genomics **18**(1)

- Workflow function ***GUIDESeqAnalysis***
  - More than 60 parameters
  - SpCas9
    - *BSgenomeName*
    - *gRNA.file*
    - *Alignment.inputfile*
    - *umi.inputfile*
    - *outputDir*

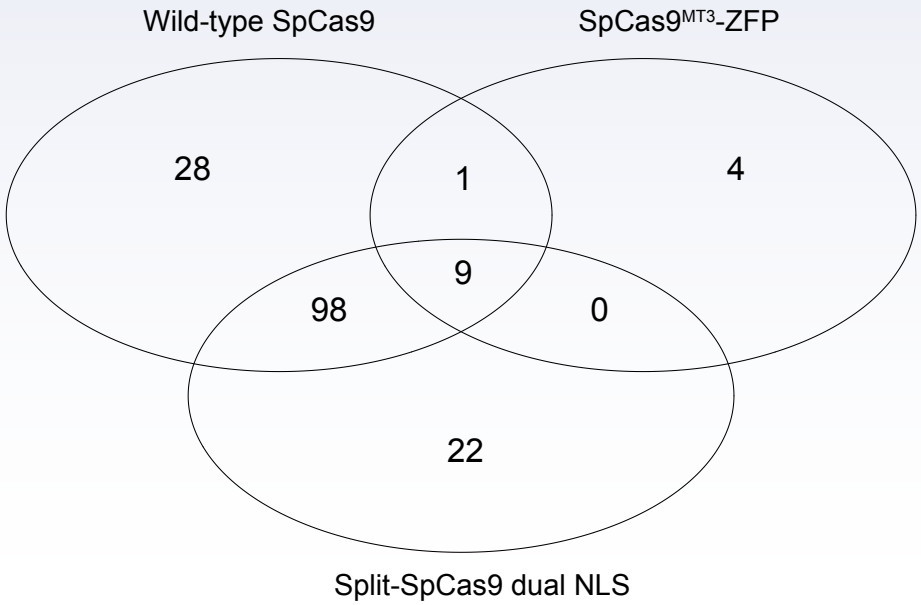
# USE CASES

1. Analysis of SpCas9 GUIDE-seq data
2. Analysis of NmCas9 GUIDE-seq data
3. Analysis of Cpf1 GUIDE-seq data
4. Annotate off-targets
5. Merge off-targets from multiple experiments to facilitate comparisons among different nuclease configurations or variants

# Compare Multiple Experiments

- Function *combineOfftargets*

**Figure 4**



# Preprocessing GUIDE-seq Data

- Scripts can be downloaded at http://mccb.umassmed.edu/GUIDE-seq/
- Additional File 1: Preprocessing steps to genereate alignment and umi files as input for GUIDEseq (Zhu LJ, Lawrence M, Gupta A, Pages H, Kucukural A, Garber M and Wolfe SA (2017). GUIDEseq: a bioconductor package to analyze GUIDE-Seq datasets for CRISPR-Cas nucleases. *BMC Genomics*, **18**(1))
- Bin barcode
  - Assign reads to different samples using p5 and p7 indexes
- Remove adaptors (dsODN)
- Extract UMI for each read
- Map to genome

# Reference and Help

- http://pgfe.umassmed.edu/bioinformatics/workshop/
- http://www.bioconductor.org/help/course-materials/2014/BioC2014/CRISPRseek-forBioc2014.pdf
- Zhu LJ*, Holmes BR, Aronin N and Brodsky MH*. (2014) CRISPRseek: a Bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. PloS One Sept 23rd 2014
- Zhu LJ (2015). Overview of guide RNA design tools for CRISPR-Cas9 genome editing technology. Front. Biol., 10(4)
- Zhu LJ*, Lawrence M, Gupta A, Pages H, Kucukural A, Garber M and Wolfe SA (2017). GUIDEseq: a bioconductor package to analyze GUIDE-Seq datasets for CRISPR-Cas nucleases. *BMC Genomics*, **18**(1)

# Future Directions

- More Precise gRNA efficacy prediction
- More Accurate off-target cleavage prediction
- Expanding dataset
  - GUIDEseq

# Acknowledgement


- **University of Massachusetts Medical School**
  - Neil Aronin (CRISPRseek)
  - Michael Brodsky (CRISPRseek)
  - Manuel Garber (GUIDEseq)
  - Ankit Gupta (GUIDEseq)
  - Alper Kucukural (GUIDEseq)
  - Scot Wolfe (GUIDEseq & MotifStack)
- **Broad Institute of MIT and Harvard**
  - Feng Zhang (CRISPRseek)
    - Benjamin Holmes
- **Genentech**
  - Michael Lawrence (GUIDEseq)
- **The Bioconductor core members**
  - Hervé Pagès (CRISPRseek & GUIDEseq)
  - Martin Morgan (CRISPRseek)