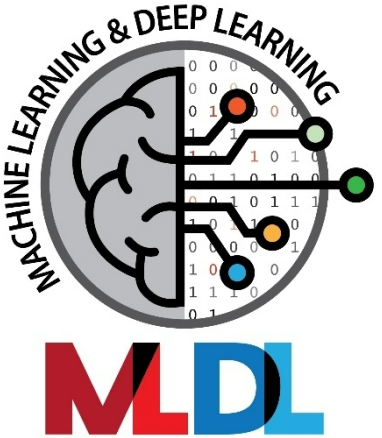
# Data Science for Detection of Genome Editing

Stephen Verzi, Raga Krishnakumar,
Drew Levin, Dan Krofcheck,
Callie Boskin, Kelly Williams




Sandia Machine Learning and Deep Learning Workshop
July 19-22, 2021

This work was funded by an LDRD.

SAND2021-XXXX PE





SANDIA BIOSCIENCE

# Abstract

Detection of genome editing is a challenging and broad area of research, and this research investigates whether data science techniques are an appropriate solution methodology.

We have gathered example data, for both edit and non-edit (or control) situations, and we have developed a data processing and analysis pipeline which includes genomic noise counting as well as machine learning (Random Forest and Deep Neural Network) and anomaly detection models.

We will present results on genomic noise characterization as well as edit detection.

SANDIA BIOSCIENCE

# Outline

- Problem
- Data
- Solution
- Results
- Next Steps

SANDIA BIOSCIENCE

Sandia National Laboratories

# Outline

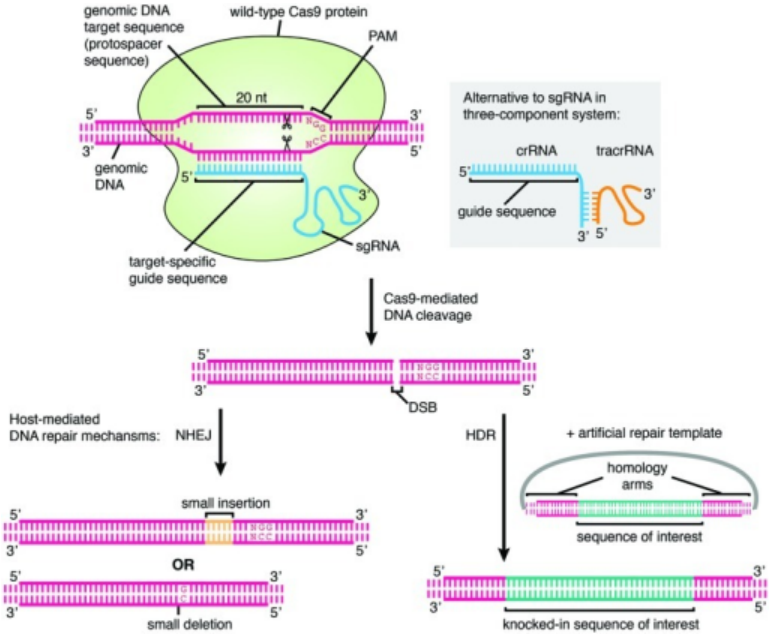
- Problem
- Data
- Solution
- Results
- Next Steps

SANDIA BIOSCIENCE

Sandia National Laboratories

# Our Problem

- Can we detect homology directed repair (HDR) & non-homologous end joining (NEHJ) genome editing from normal mutation and/or machine error in deep sequence data?
- CRISPR Cas9
  - NHEJ
    - (small) insertion
    - deletion
  - HDR
    - (large) insertion



# Outline

- Problem
- Data
- Solution
- Results
- Next Steps

SANDIA BIOSCIENCE

Sandia National Laboratories

# Datasets Under Study

| | | | |
|---|---|---|---|
| Paper | wang2018 | doudna2014 (lin2014) | cho2014 |
| Pubmed | 29843790 | 25497837 | 24253446 |
| Project | PRJNA454824 | PRJNA269153 | PRJDB1494 |
| Mutation Types | HDR (ssODN) vs NHEJ | HDR (ssODN) vs NHEJ | NHEJ only |
| Cas Variety | 5 (3 Cas9, 2 Cpf1) | 1 (Cas9) | 1 (Cas9) |
| Target Sites | 90 | 2 (EMX1, DYRK1) | 2 (CCR5, C4BPB) |
| Delivery | Cas/sgRNA-Plasmid | Cas/sgRNA-RNP | Cas-plasmid + sgRNA RNA |
| Cells | HEK293T | HEK293T | K562, HeLa |
| SRA Total | 48.7 Gbyte | 7.1 Gbyte | 23.3 Gbyte |
| Sequencing | Amplicon | Amplicon | Amplicon + Deep |
| Note | 1144 samples | Nocodazole S-phase synch | Many off-targets too |

# Datasets Under Study

| | soyk2017 | Iyer2018 |
|---|---|---|
| Paper | | |
| Pubmed | 27918538 | 29985941 |
| Project | PRJNA327675 | ERP024425 |
| Mutation Types | NHEJ | NHEJ |
| Cas Variety | 1 (Cas9) | 1 (Cas9) |
| Target Sites | 2 (SP5G – a1 & a2) | 2 (Tyr – 2F & 2R) |
| Delivery | Cas/sgRNA | Cas/gRNA |
| Cells | Solanum lycopersicum (Tomato) | Embryo (Mouse) |
| SRA Total | 33.8 Gbyte | >980 Gbyte |
| Sequencing | Amplicon | Amplicon + Deep |
| Note | Edited SP5G promotes day-neutrality | Tyr knockout changes coat color |

# Datasets Under Study

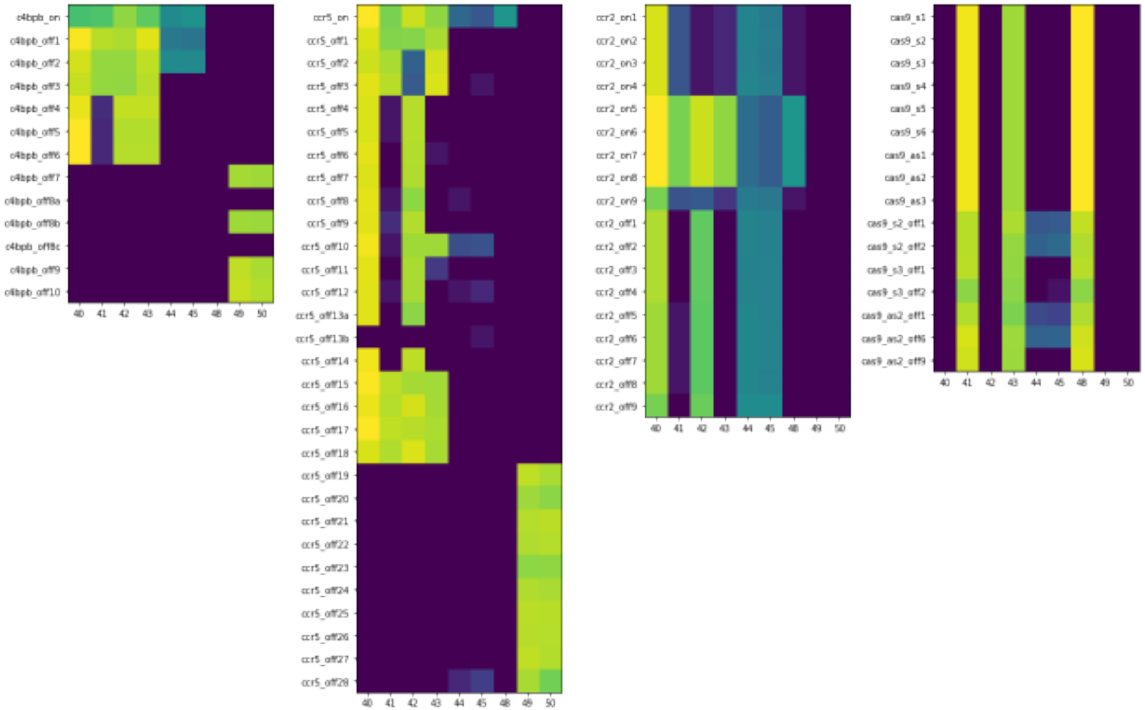
| Paper | guo2018 | veres2014 | church2014 (yang2014) |
|---|---|---|---|
| Pubmed | 30340517 | 24996167 | 25425480 |
| Project | PRJNA475379 | PRJNA240654 | PRJNA259786 |
| Mutation Types | NHEJ | NHEJ | NHEJ |
| Cas Variety | 1 (Cas9) | 2 (Cas9, Talen) | 1 (Cas9) |
| Target Sites | 71 | 2 (SORT1, LINC00116) | 1 (TAZ) |
| Delivery | Cas/sgRNA | Cas/sgRNA | Cas/sgRNA |
| Cells | Mouse & HEK293T | HEK293T, K562 | PGP1-hiPSC |
| SRA Total | 46.2 Gbyte | 0.87 Tbyte | 0.14 Tbyte |
| Sequencing | Amplicon | WGS | WGS |
| Note | 99 samples | 10 samples | 6 samples + many off-targets |

# Datasets Under Study

| | | |
|---|---|---|
| Paper | chakrabarti2019 | van Overbeek2016 |
| Pubmed | 30554945 | 27499295 |
| Project | E-MAT-7095, E-MAT-7091 | PRJNA326019 |
| Mutation Types | NHEJ | NHEJ, MMEJ |
| Cas Variety | 1 (Cas9) | 1 (Cas9) |
| Target Sites | 1491/1248/649 | 223 |
| Delivery | Cas/sgRNA | Cas/sgRNA |
| Cells | HepG2 | HCT116, HEK293T, K562 |
| SRA Total | 80 x ~10 Gbyte per pool | 6195 x ~1 Mbyte per file |
| Sequencing | AMPLICON | AMPLICON |
| Note | 450 genes, many sgRNAs | many samples and sites |

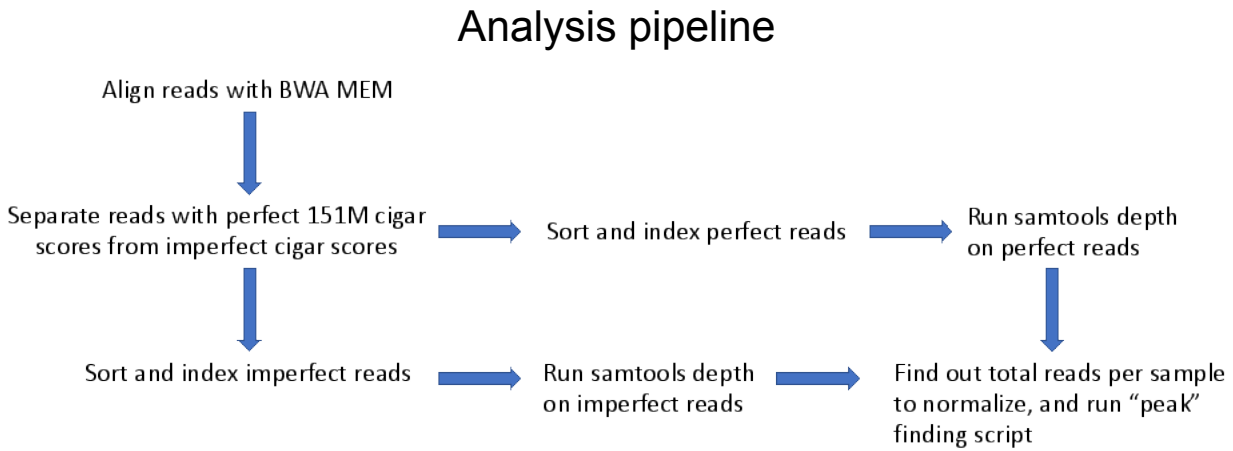
# Dataset Visualization

- Cho et al., 2014 datasets x target sites



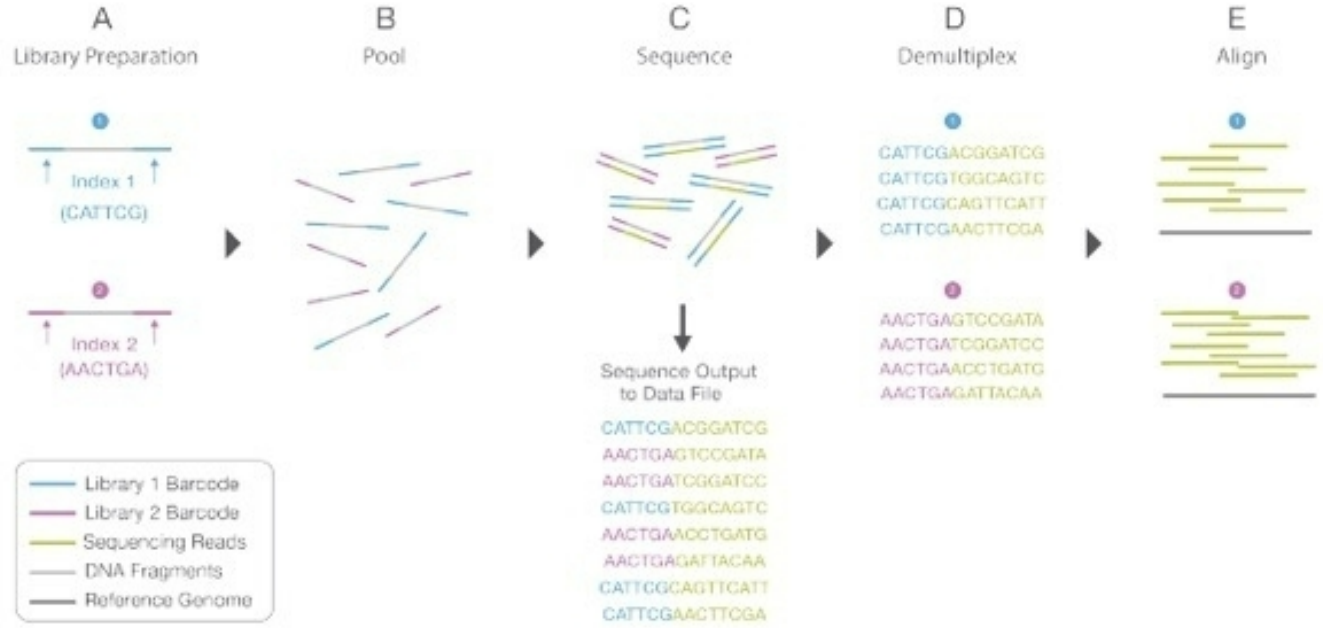
# Application of Our Methodology

- Using data from Iyer et al., 2018



# Data Preprocessing

- Sequencing and alignment



# Data Preprocessing

- Handling imperfect alignments

**Non perfect matching after alignment**

SAM Flag 2048 = supplemental alignment (alignment at a secondary location, eg. chimeric reads)

SM/HM
MS/MH
SMS/HMH

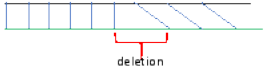

Hard clipping

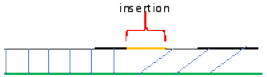
Often supplemental alignments are hard clipped


Soft clipping

M – match or mismatch
S – soft clipped
H – hard clipped
I – insertion
D - deletion

MDM
MIM




NM = number of mutations introduced during alignment


- "noise" for the genome counter

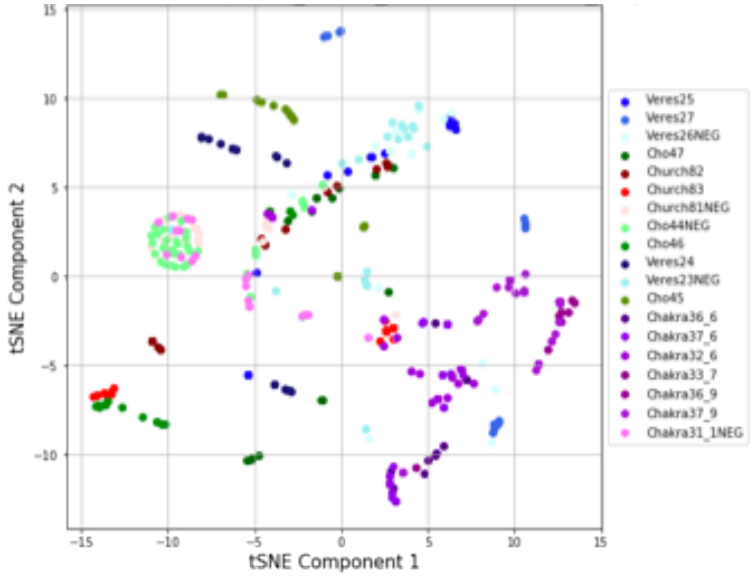
# Genome Counter Features and Analysis

Genome counter

- Non-matches
- Total reads
- Matches
- Deletions
- Insertions
- Clips
  - soft
  - hard
- Nucleotides: A,C,G,T
- others

tSNE (t-distributed Stochastic Neighbor Embedding) visualization

DBSCAN clustering



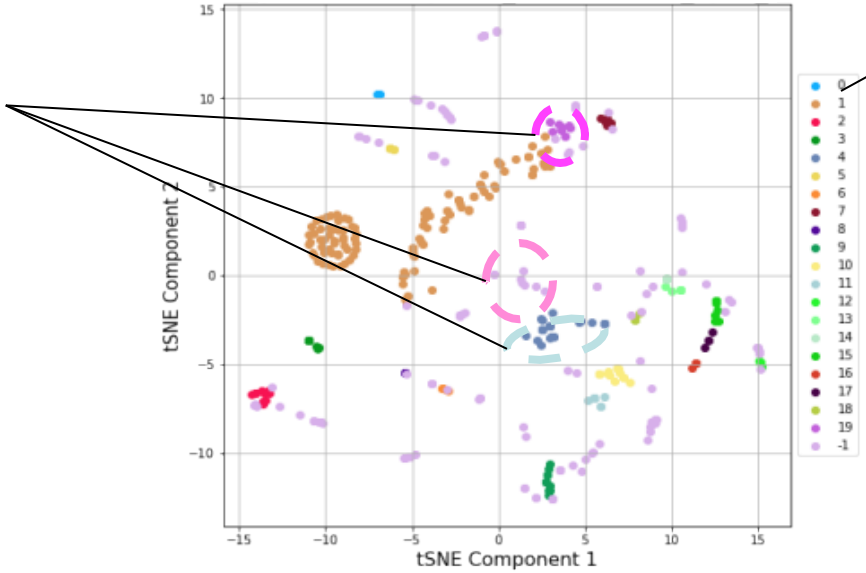

SANDIA BIOSCIENCE

# Cluster Analysis

- Analysis of edit (positive) versus no edit (negative)

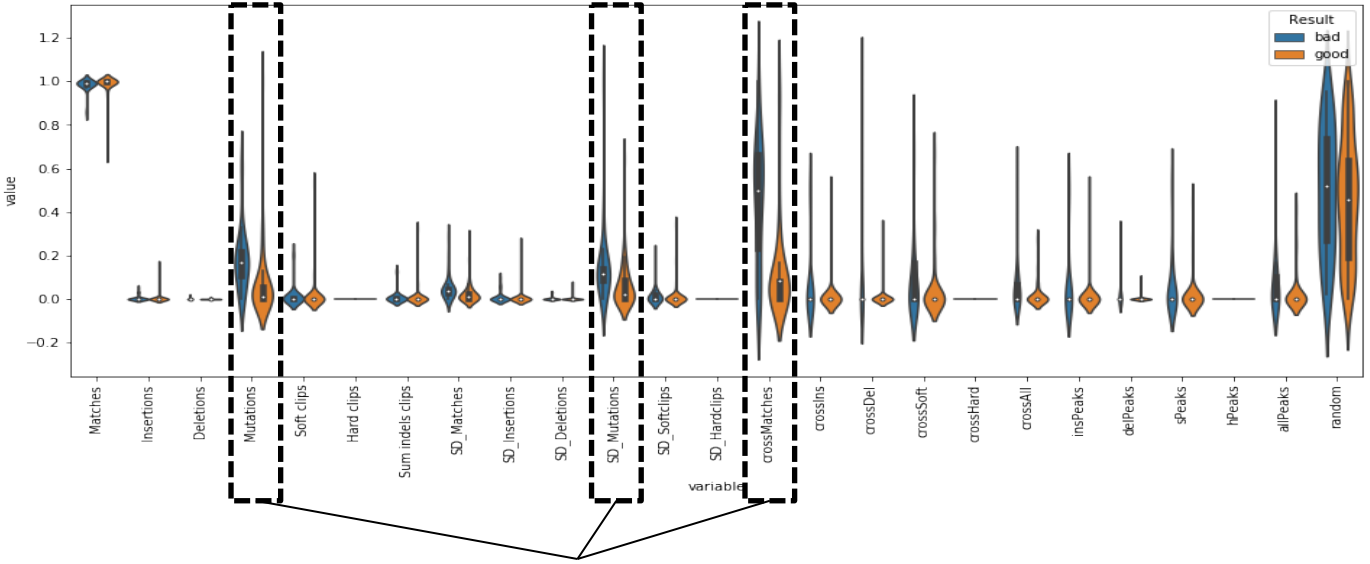
Most negatives are in cluster 1

Non-cluster 1 negatives are either in cluster 4, 19 or singletons (not in any cluster)



# Cluster Analysis

- Non-edited samples that are correctly clustered (good) versus mis-clustered (bad)




Features that are mis-leading: Mutations and crossMatches

# Outline

- Problem
- Data
- Solution
- Results
- Next Steps

SANDIA BIOSCIENCE

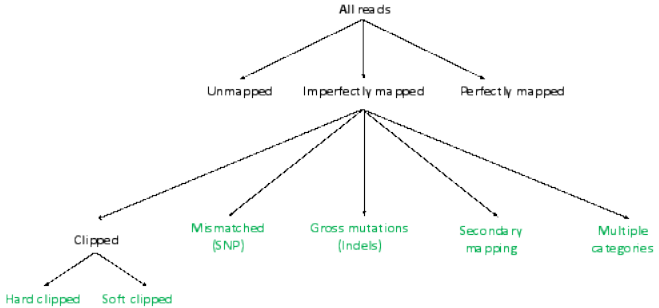
Sandia National Laboratories

# Our Solution

We construct a data processing pipeline to facilitate distinguishing amongst:

- Normal DNA sequence variations,
- Sequencer machine error,
- Penetration of editing: failure of editing in many cells in a tissue, and
- Occasional NHEJ (with different outcomes in different cells in a tissue), even when attempting HDR.

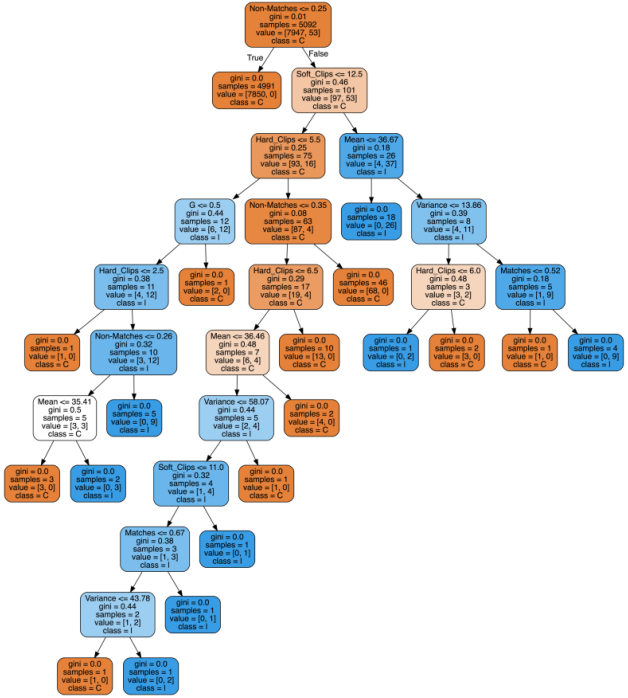
## Data Processing Pipeline




As part of the data pipeline, we focus upon and investigate imperfectly mapped reads.

# Machine Learning and Anomaly Detection Algorithms

## Random forest and XGBoost




Use decision trees to distinguish edit from no edit

# Feature Importance

## Random forest and XGBoost

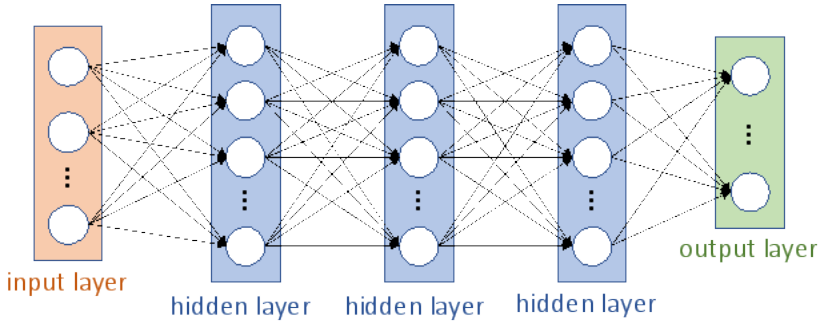
| | Pre-smote | Post-smote | XGB f imp | RF f imp |
|---|---|---|---|---|
| Matches | -0.334572 | -0.570796 | 0.01247 | 0.10471 |
| random | -0.005705 | -0.004884 | 0.00411 | 0.00918 |
| crossMatches | 0.004259 | -0.014383 | 0.02273 | 0.10219 |
| delPeaks | 0.050335 | 0.181153 | 0.05403 | 0.00837 |
| crossSoft | 0.057115 | 0.191615 | 0.01571 | 0.03634 |
| crossDel | 0.058821 | 0.229701 | 0.0232 | 0.00709 |
| insPeaks | 0.065569 | 0.325902 | 0.02968 | 0.01515 |
| crossIns | 0.065626 | 0.319157 | 0.00691 | 0.00869 |
| allPeaks | 0.078111 | 0.256226 | 0.01316 | 0.0611 |
| sPeaks | 0.081481 | 0.287515 | 0.01505 | 0.03051 |
| crossAll | 0.140162 | 0.424111 | 0.01956 | 0.12038 |
| Soft clips | 0.197254 | 0.469467 | 0.028 | 0.04279 |
| SD_Softclips | 0.211422 | 0.539551 | 0.01245 | 0.0655 |
| SD_Matches | 0.232588 | 0.573208 | 0.01184 | 0.13165 |
| SD_Insertions | 0.250619 | 0.46122 | 0.00523 | 0.01011 |
| hPeaks | 0.252463 | 0.479862 | 0.0111 | 0.00703 |
| Insertions | 0.267338 | 0.451866 | 0.04347 | 0.01114 |
| crossHard | 0.27584 | 0.494565 | 0.02723 | 0.01036 |
| SD_Deletions | 0.293409 | 0.424185 | 0.11218 | 0.01295 |
| Deletions | 0.323052 | 0.371885 | 0.02046 | 0.01471 |
| Sum indels cl | 0.357617 | 0.57898 | 0.45137 | 0.14595 |
| SD_Hardclips | 0.389561 | 0.509483 | 0.03067 | 0.02356 |
| Hard clips | 0.400161 | 0.508797 | 0.02939 | 0.02053 |

Augment genome counter with statistically derived features and Smote

# Machine Learning and Anomaly Detection Algorithms

Deep Convolutional Neural Network (CNN)

- 1D CNNs in each hidden layer
- Trained over
  - All edit sites (positive)
  - Each control file (negative)
    - Cho
    - Church
    - Veres
    - 1000 genomes




- Use genome counter features and deep convolutional filters

# Machine Learning and Anomaly Detection Algorithms

Anomaly (Edit) Detection Using Deep Spiking

- Bio-inspired simplification of CRISPR edit precision prediction [Chakrabarti, 2019]
  - Using spiking adaptive median-filtering [Verzi, 2018]

$$o_{ij} = \begin{cases} \hat{\rho}_{ij}^{1}, & \text{if } \exists m, \hat{\rho}_{ij}^{m} > \theta^{m} \\ x_{ij}, & \text{otherwise} \end{cases}$$

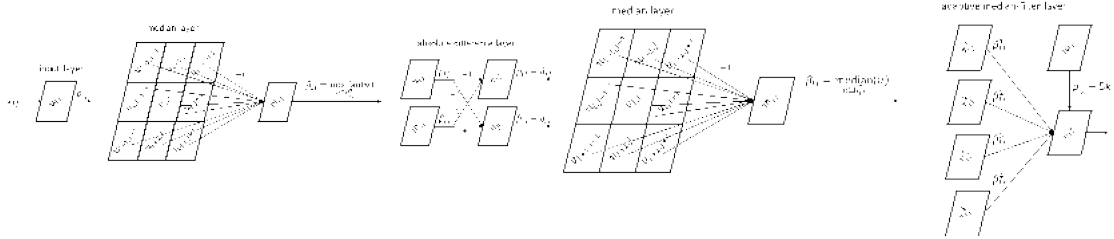
$$\hat{\rho}_{ij}^{m} = \underset{x \in \Omega_{ij}^{m}}{\text{median}}\{x\}$$

$$\Omega_{ij}^{m} = \{x_{lr} \mid i - m \le l \le i + m, j - m \le r \le j + m\}$$

$$\theta^{m} = s \cdot \underset{x \in \Omega_{ij}^{m}}{\text{median}}\left\{\hat{\rho}_{ij}^{m} - x\right\} + \delta^{m}$$

$$\delta^{m} = \frac{\left((2m+1)^{2}-1\right)}{2}$$

Find tightly grouped noise signals from genome counter



SANDIA BIOSCIENCE

Sandia National Laboratories

# Outline

- Problem
- Data
- Solution
- Results
- Next Steps

SANDIA BIOSCIENCE

Sandia National Laboratories

# Results

- Random forest and XGBoost

| | pre-SMOTE | post-SMOTE |
|---|---|---|
| Matches | -0.254619 | -0.292882 |
| random | 0.012406 | 0.014812 |
| SD_Softclips | 0.097822 | 0.097472 |
| SD_Insertions | 0.142293 | 0.161023 |
| Soft clips | 0.149396 | 0.165603 |
| Insertions | 0.150198 | 0.172695 |
| Deletions | 0.209405 | 0.248075 |
| Hard clips | 0.235041 | 0.277336 |
| SD_Deletions | 0.235192 | 0.277489 |
| SD_Hardclips | 0.24263 | 0.285641 |
| SD_Matches | 0.244244 | 0.276878 |
| Sum indels clips | 0.257533 | 0.295573 |
| delPeaks | 0.280862 | 0.330604 |
| insPeaks | 0.292268 | 0.327164 |
| crossIns | 0.295258 | 0.330344 |
| crossDel | 0.381553 | 0.440137 |
| crossHard | 0.384375 | 0.443334 |
| hPeaks | 0.404201 | 0.465554 |
| sPeaks | 0.406744 | 0.442842 |
| allPeaks | 0.43573 | 0.475255 |
| crossSoft | 0.43661 | 0.462243 |
| crossAll | 0.545992 | 0.597058 |
| Class | 1 | 1 |

- Overall high specificity and accuracy
- Variability in precision and recall depending on data
- XGB > RF for recall

Random Forest vs XGBoost (hyperparameter optimization)

| | Xgboost | RF |
|---|---|---|
| 1 | 0.9756 | 0.9982 |
| 2 | 0.9774 | 0.9982 |
| 3 | 0.9737 | 0.9826 |
| 4 | 0.9765 | 0.9982 |
| 5 | 0.9759 | 0.9982 |
| 6 | 0.9765 | 0.9466 |
| 7 | 0.9744 | 0.9982 |
| 8 | 0.9755 | 0.8698 |
| 9 | 0.9756 | 0.997 |
| 10 | 0.9788 | 0.9436 |

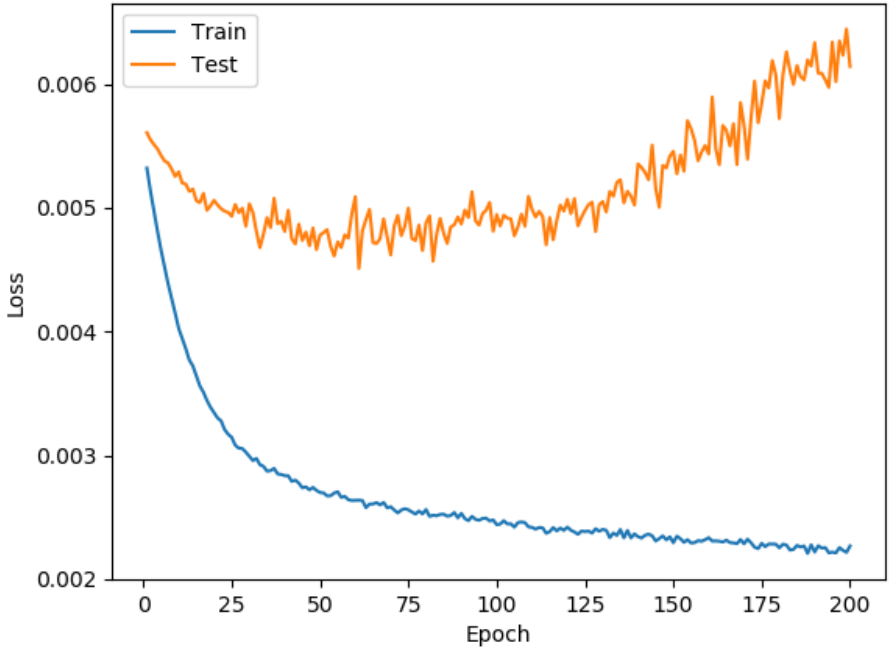
Random Forest vs XGBoost (test data set metrics)

| Test set | Algorithm | Accuracy | Precision | Recall | Specificity | F1 |
|---|---|---|---|---|---|---|
| Data set 1 | RF | 0.995 | 0.5 | 0.1 | 0.9995 | 0.1666 |
| Data set 1 | XGB | 0.995 | 0.4444 | 0.4 | 0.9983 | 0.421 |
| Data set 2 | RF | 0.97 | 0.96226 | 0.51 | 0.99875 | 0.66667 |
| Data set 2 | XGB | 0.985 | 0.93023 | 1 | 0.98125 | 0.96386 |

# Results

- Deep Convolutional Neural Network (CNN)

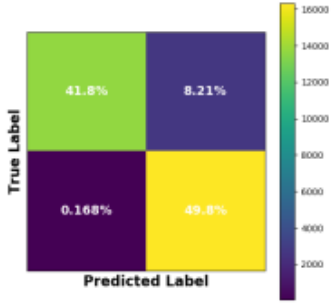
Type 0 – all negatives
Type 1 – non-sgRNA positives
Type 2 – sgRNA positives

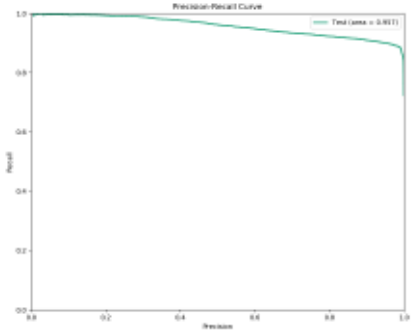


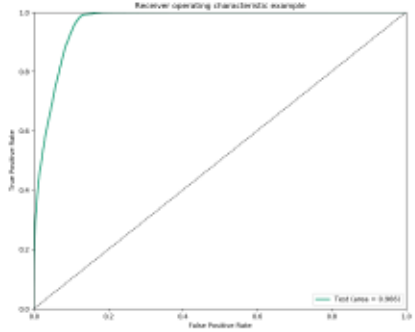
# Results

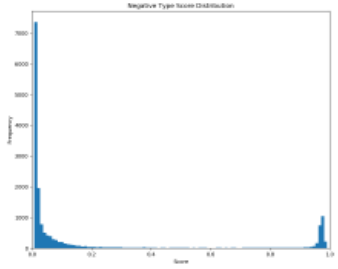
- Deep Convolutional Neural Network (CNN)
  - Balanced training data

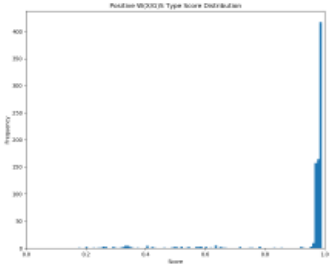
Type 0 – all negatives
Type 1 – non-sgRNA positives
Type 2 – sgRNA positives

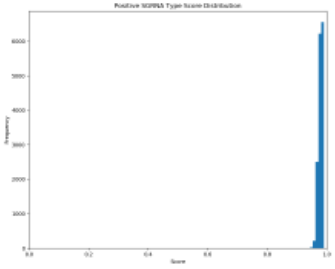


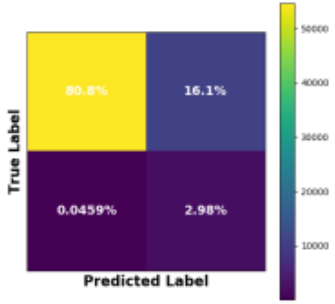



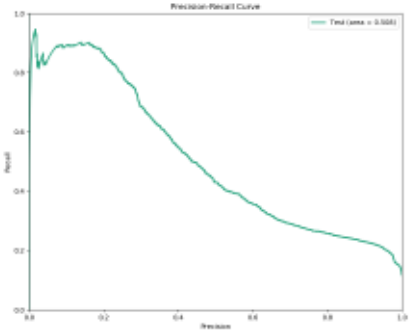



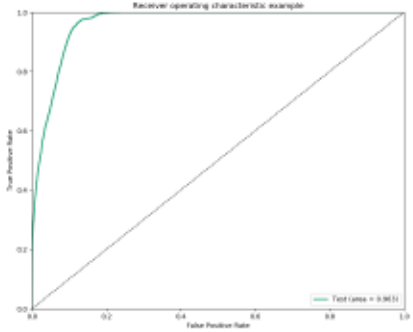
# Results

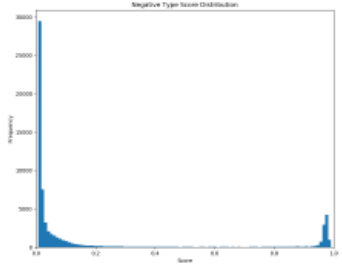
- Deep Convolutional Neural Network (CNN)
  - Imbalanced training data

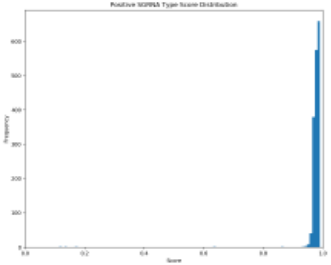
Type 0 – all negatives
Type 1 – non-sgRNA positives
Type 2 – sgRNA positives





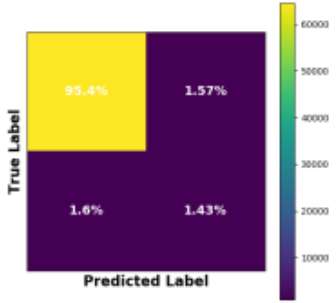


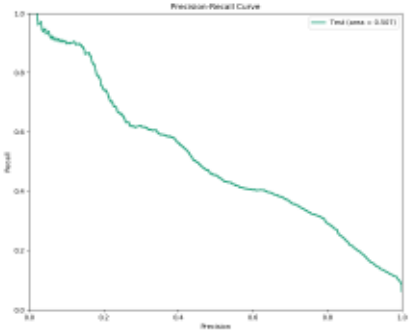


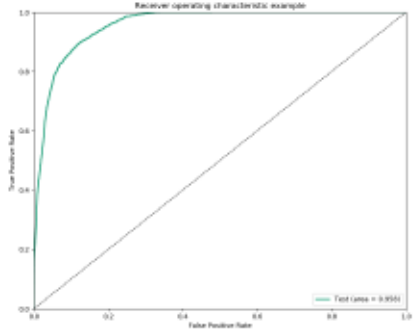
# Results

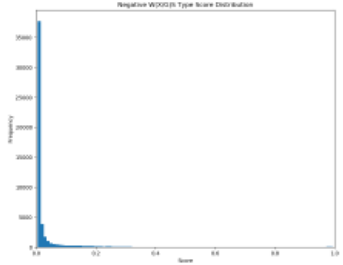
- Deep Convolutional Neural Network (CNN)
  - Imbalanced training data

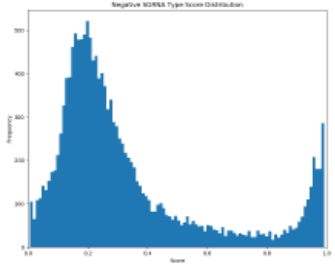
Type 0 – non-sgRNA negatives
Type 1 – sgRNA negatives
Type 2 – non-sgRNA positives
Type 3 – sgRNA positives



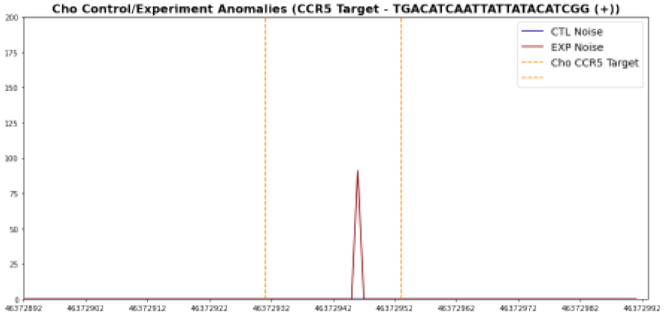



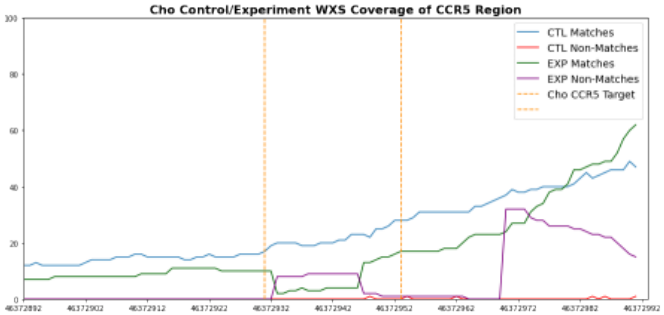


# Results

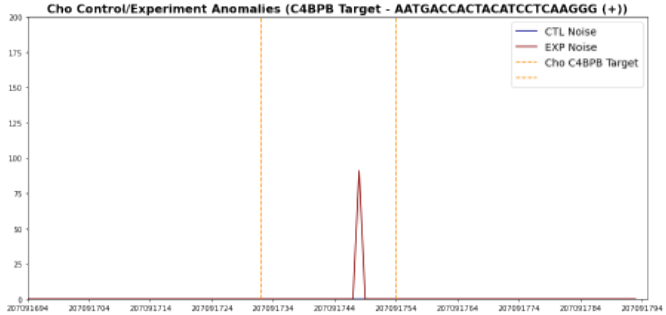
- Anomaly (Edit) Detection Using Deep Spiking

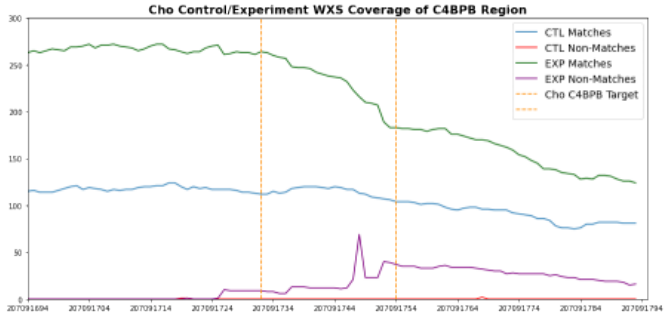
CCR5 (chromosome 3)




C4BPB (chromosome 1)



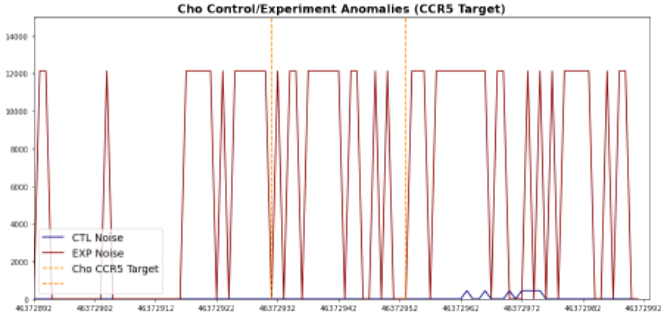



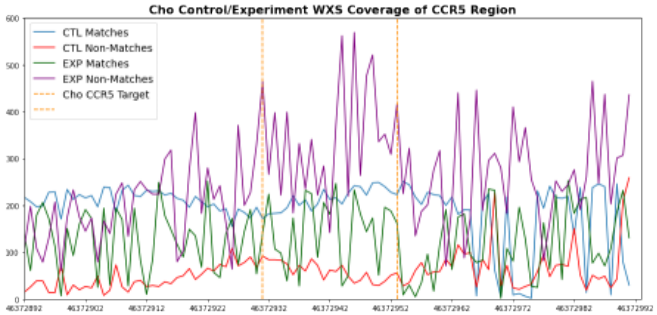

WXS – trimmed

# Results

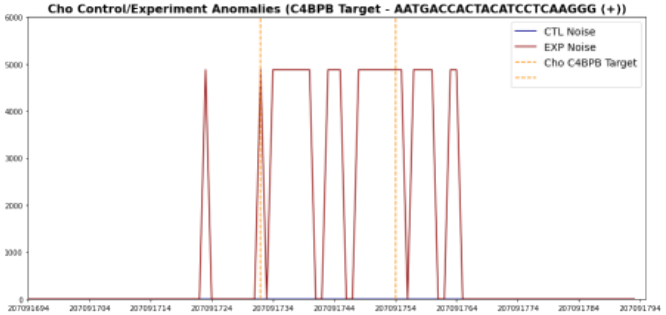
- Anomaly (Edit) Detection Using Deep Spiking

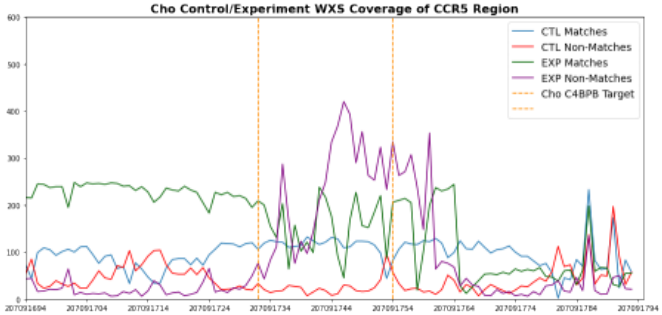
CCR5 (chromosome 3)




C4BPB (chromosome 1)




WXS – trimmed

# Outline

- Problem
- Data
- Solution
- Results
- Next Steps

SANDIA BIOSCIENCE

Sandia National Laboratories

# Next Steps

Path Forward

- Acquire more datasets for verification, validation and publishable results – ongoing
- Data science/machine learning algorithm training/testing, improvement & application
- Publish
  - Manuscript and TA for precise CRISPR edit detection with deep spiking anomaly detector
  - Manuscript for ML algorithms applied to CRISPR edit detection

SANDIA BIOSCIENCE

# Thank You!

- Questions?

# Backups

SANDIA BIOSCIENCE

Sandia National Laboratories