Epigenetic Data Boosts the Accuracy of Recombination Hotspot Prediction by Machine Learning Models

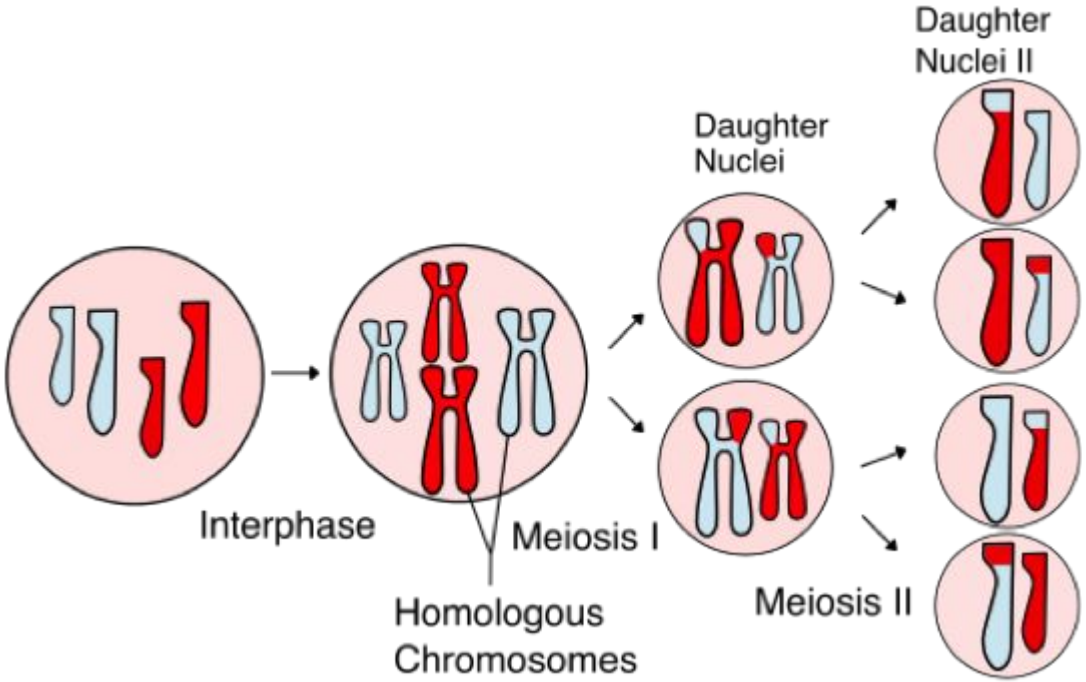
Lawrence Wong
Collaborator: Joy Linyue Fan

Motivation

Genetic recombination plays a key role in:

- Driving force behind evolution and genetic diversity
- Implicated in genetic disease
  - Translocation events and other mispairings
  - Non-homologous recombinations predicate cancer
- GWAS and Linkage Disequilibrium (LD) studies

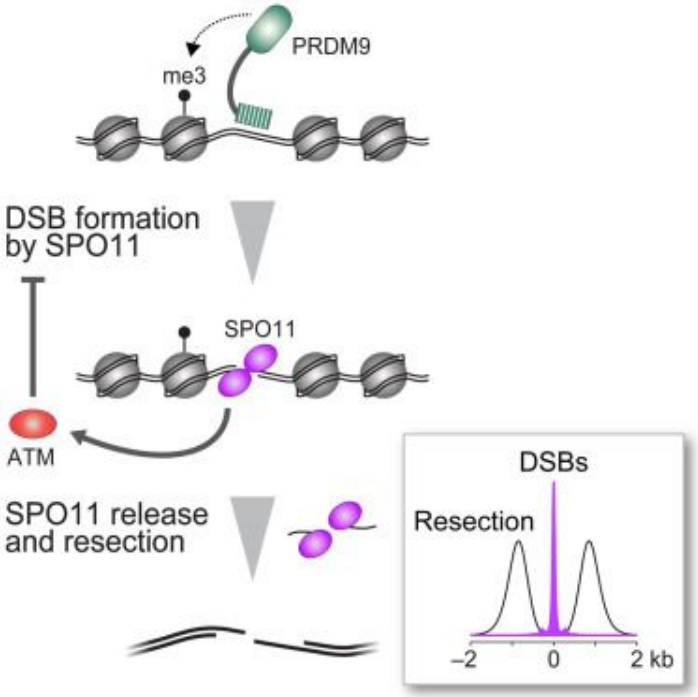
Recombination Hotspots (and Coldspots)



(*Location matters in recombination* 2018)

- **Hotspots** are regions in a genome that exhibit elevated rates of recombination.
- Identified through microarray and linkage disequilibrium studies.
- In humans, crossovers occur within 2 kbps regions that are spaced 50-100 kbps apart.

Spo11 and PRDM9's Role in Recombination



(Lange et al., 2016)

- **PRDM9** binds to motifs and recruits recombination machinery. Known to have an affinity for a identified degenerate 13-mer motifs.

- **Spo11** makes double strand breaks (DSB). It has preferences for GC rich regions.

Existing Hotspot Prediction Models

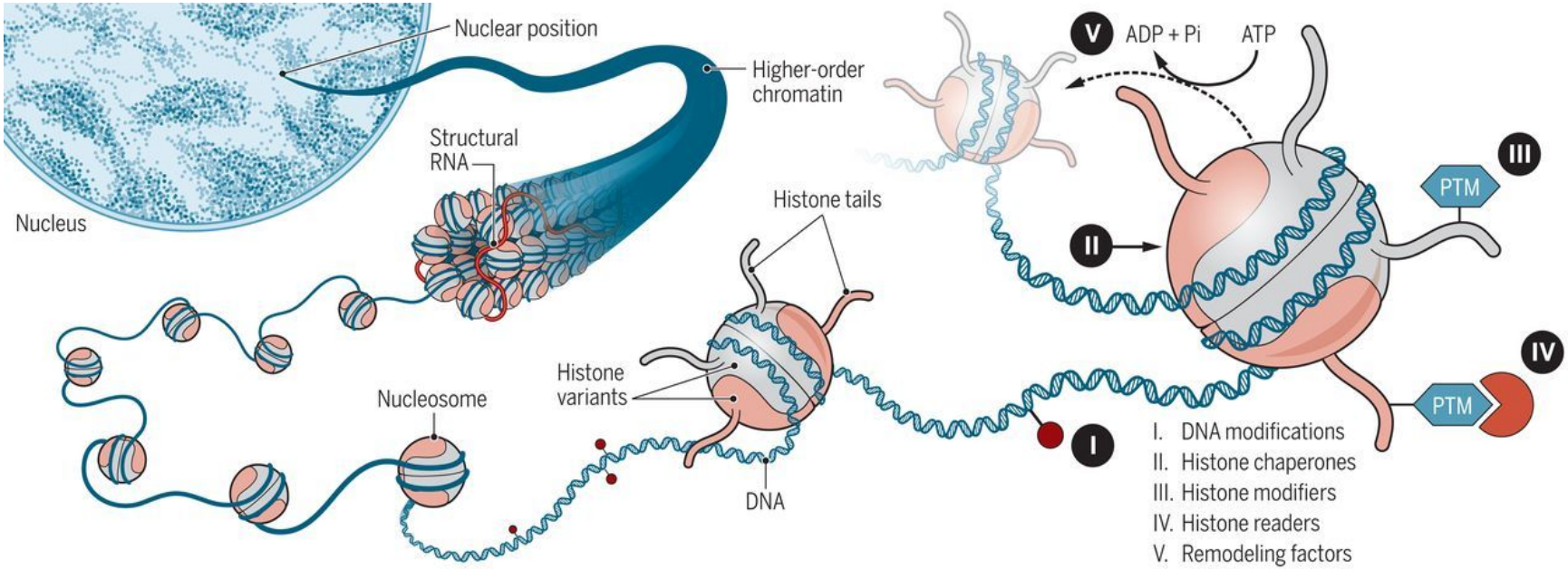
- Formulated the problem as a supervised learning problem of binary classification.
  - Random Forest (RF), Support Vector Machines (SVM), Neural Networks
- Most uses sequence based features (Maruf & Shatabda, 2018)

1. **RF-DYMHC** (Jiang et al., 2007) - Random Forest model uses gapped dinucleotide compositions.

2. **IDQD** (Liu et al., 2012) - QDA model uses with k-mer frequencies.

3. **iRSpot-PseDNC** (Chen et al., 2013) - SVM model uses pseudo dinucleotide composition created from local structural properties of DNA.

Epigenetics & Chromatin Structure



(Yadav et al., 2018)

Goals

1. Find the effect of epigenetic data on hotspot predictions
    a. Baseline Models
    b. Neural Networks

2. Discover motifs for PRDM9
    a. Gibbs sampling on hotspot sequences
    b. Gibbs sampling on predicted hotspot sequences

Epigenetic Data and Hotspot Predictions

Datasets

Roadmap Epigenomics Project

- H3K4me3 chromatin mark annotations
- H3K36me3 chromatin mark annotations
- DNAse I Hypersensitivity data

NCBI SNP Database

- SNP data

Human Genome Project

- GRCh37 reference genome

HapMap Phase I, II

- Finescale genetic map of locations with high recombination rates

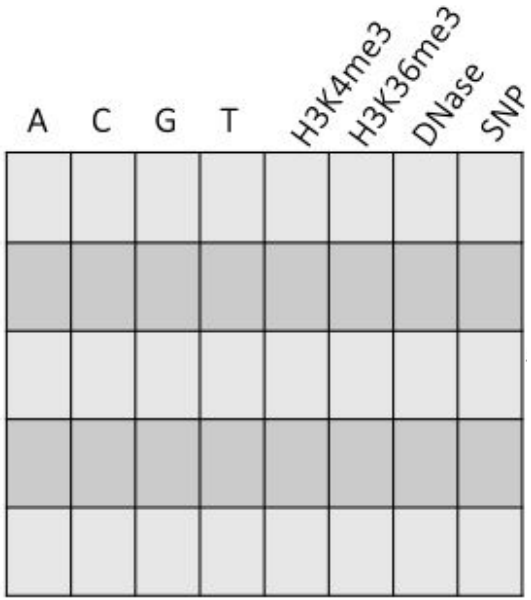
Data Preprocessing

1. Convert recombination rates into 0s and 1s.

2. Map recombination intervals with corresponding nucleotides in genome and one-hot encode nucleotides.

3. Map recombination intervals with chromatin marks, DNAse I hypersensitivity, and SNPs.

4. Divide intervals with all variables into sections of 2 kbps.



Models

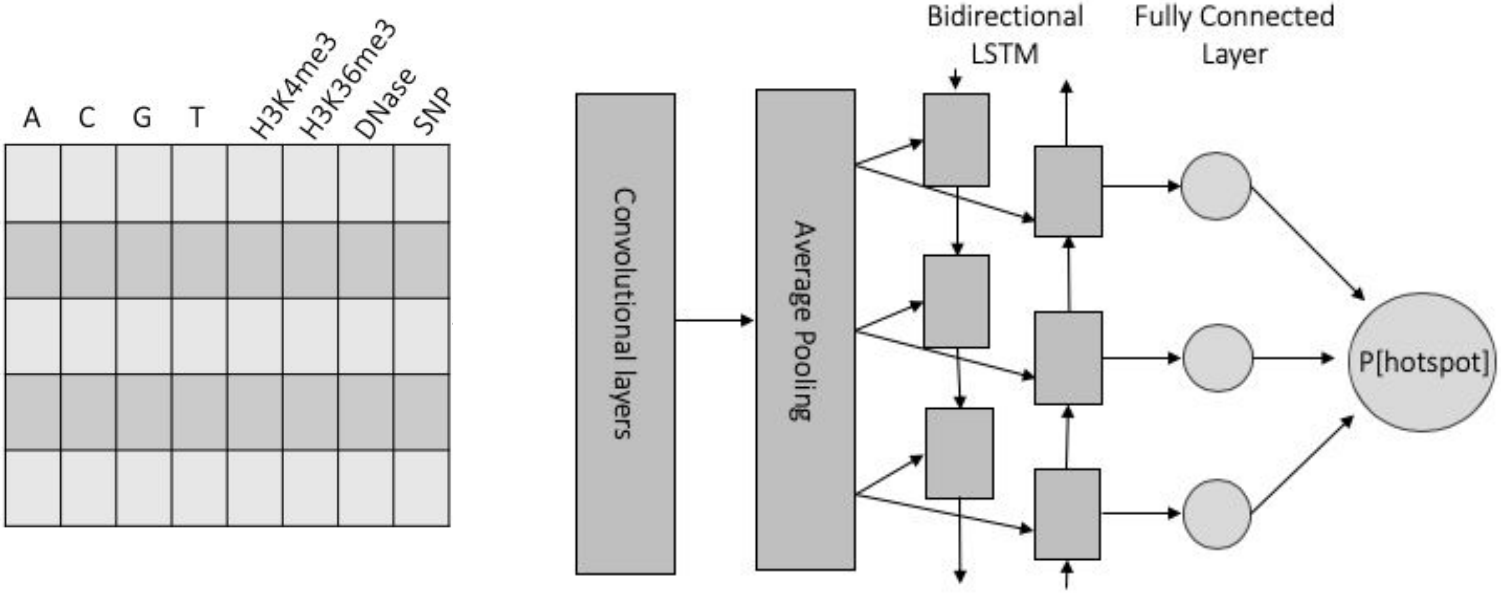
Baseline Models

- Logistic Regression w/ L2 penalty
- Random Forest Classifier
- Naive Bayes Classifier

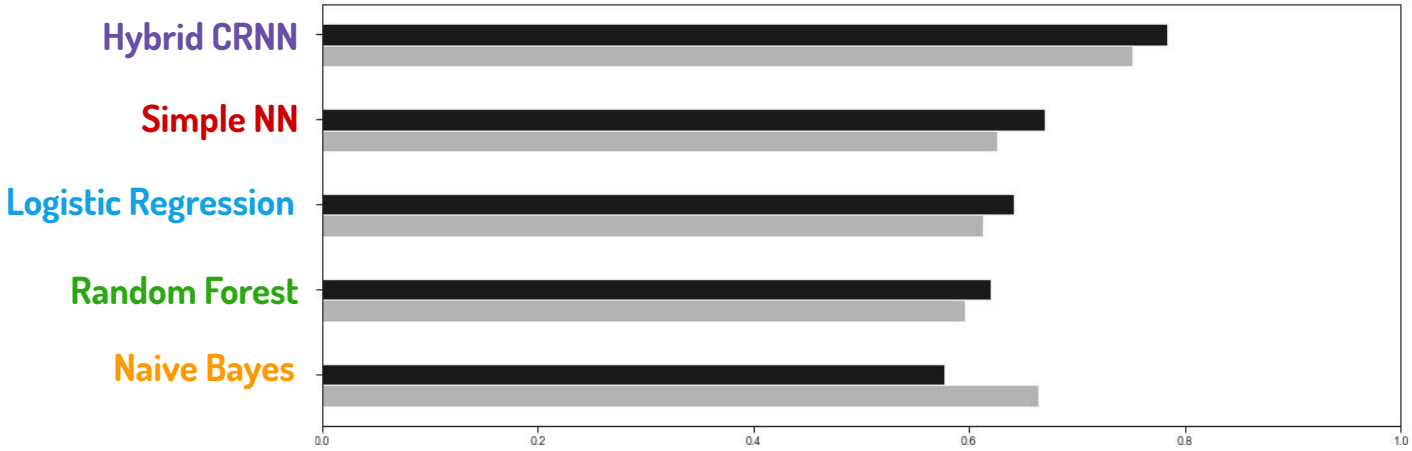
Neural Network Models

- Simple Neural Network
- Hybrid Convolutional/Recurrent Neural Network

Hybrid Convolutional/Recurrent NN Architecture

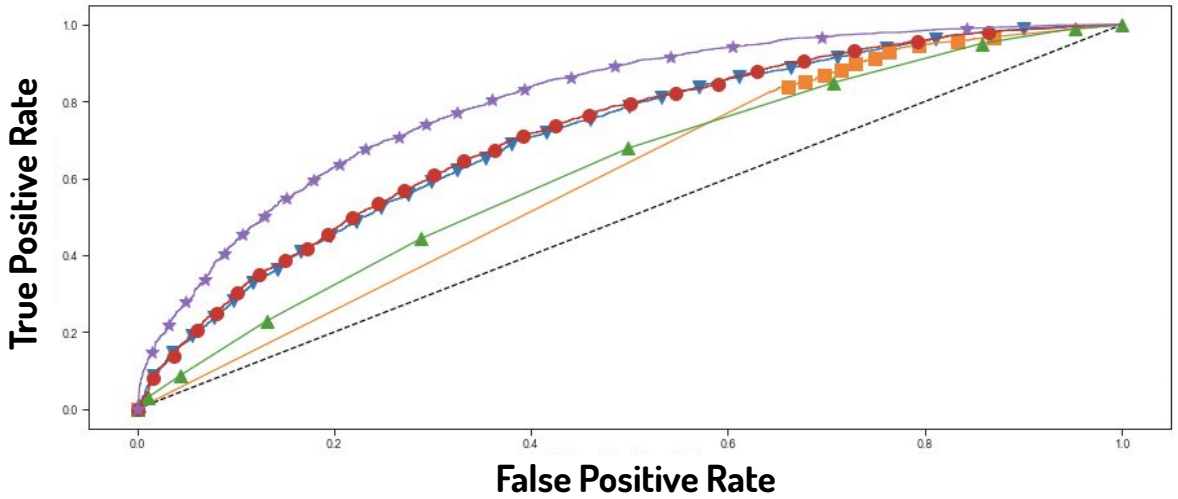


Methods



Full 8 parameters

Sequence only



Hybrid CRNN

Simple NN

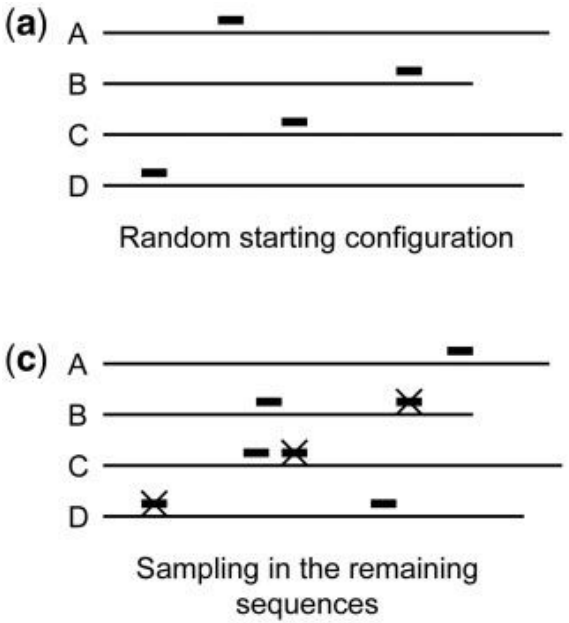
Logistic Regression

Random Forest

Naive Bayes

Motif Discovery via Gibbs Sampling

Gibbs Sampling

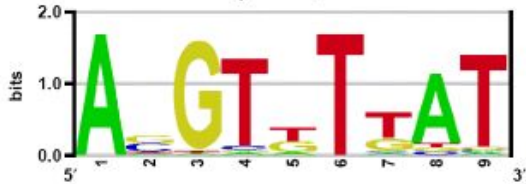


	1	2	3	4	5	6	7	8	9
A	.97	.10	.02	.03	.10	.01	.05	.85	.03
C	.01	.40	.01	.04	.05	.01	.05	.05	.03
G	.01	.40	.95	.03	.40	.01	.3	.05	.03
T	.01	.10	.02	.90	.45	.97	.6	.05	.91

PWM

WebLogo, enoLOGOS, etc.

?



(Lange et al., 2016)

Methods

Discovered Motifs

Known PRDM9 Motif

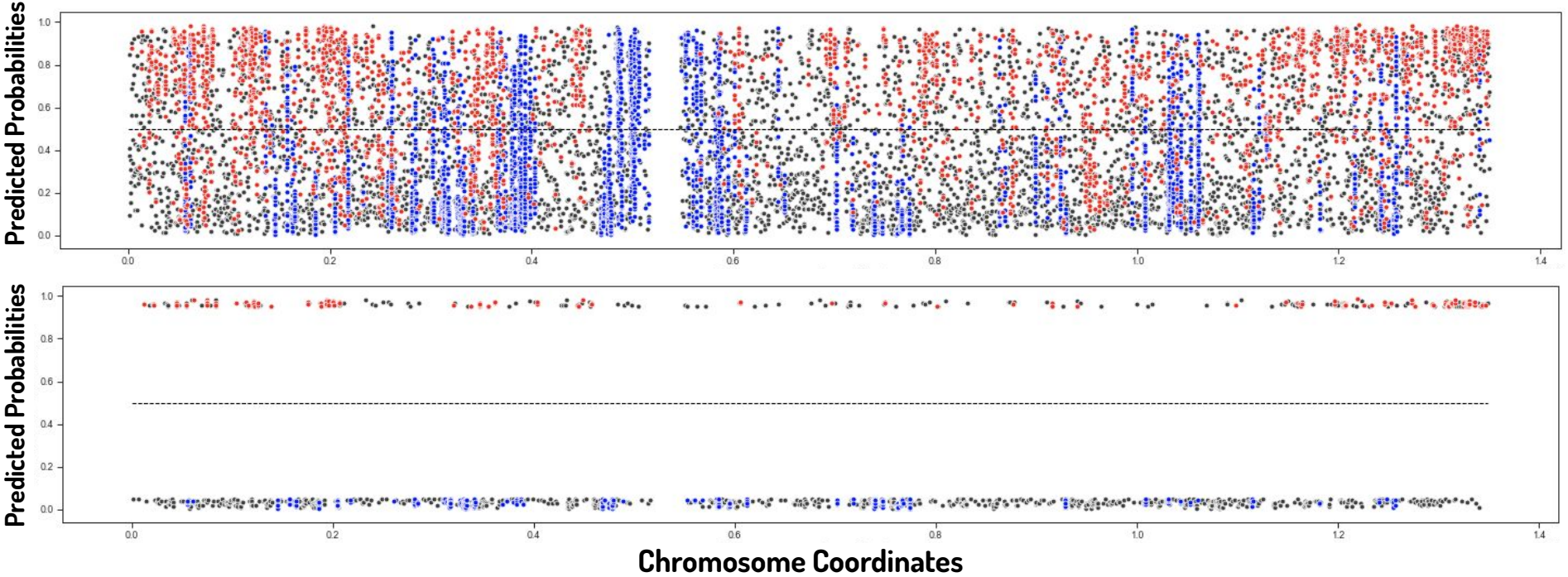
CCNCCNTNNCCNC

TABLE I

10 MOST FREQUENTLY FOUND MOTIFS DISCOVERED BY GIBBS SAMPLING

Motifs	Percent Identity	Frequency
CCCCCCCCCCCC	0.920	8
GCCCCCCCCCCC	0.846	4
CCCTCCCCCCCC	0.846	4
CCCCCTCCCCCC	0.920	4
CCCCCCCCCCGCC	0.846	4
CCCCTCCCCCCC	0.846	4
CCCCGGCCCCCC	0.846	3
CCCCCGGCCCCC	0.920	3
CCCCCCCCCTCCC	0.846	3
CCCGCCCCCCCC	0.846	3

Chromosome 11 Hotspot Predictions



Discussion

- Inclusion of epigenetic data increases the accuracy of hotspot predictions depending on the model.
- The hybrid CRNN outperforms other baseline machine learning models.
- Gibbs sampling on the predicted hotspot sequences from the hybrid CRNN model found motifs contradicting what we know.

Future Work

- Expand testing/training to other chromosomes.
- Experimental validation of new hotspots.
- Include other epigenetic data.
- Add features generated from the DNA sequences.

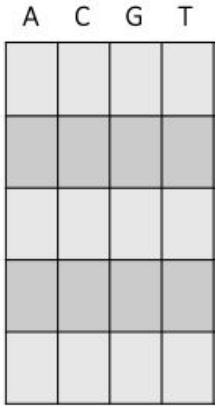
Acknowledgements

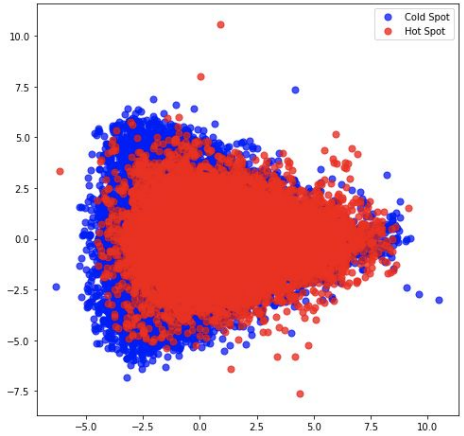
- Joy Linyue Fan (Co-Author)
- Wouter Meuleman (Project Mentor)
- Samuel Kim (6.047 TA)
- Manolis Kellis (6.047 Professor)

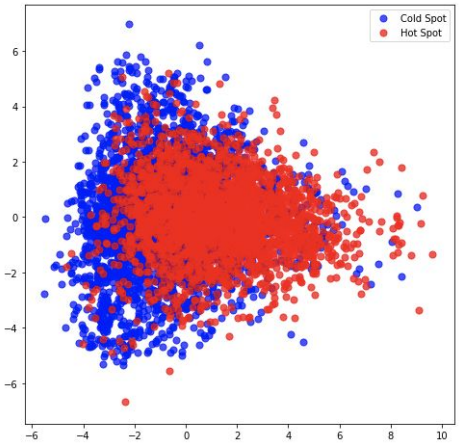
Questions / Feedbacks?

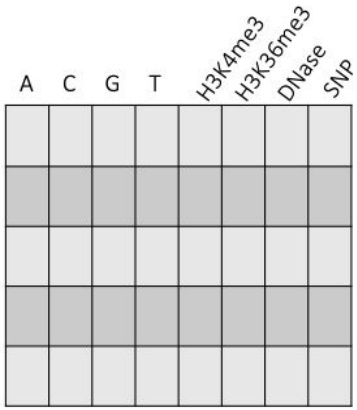
Works Cited

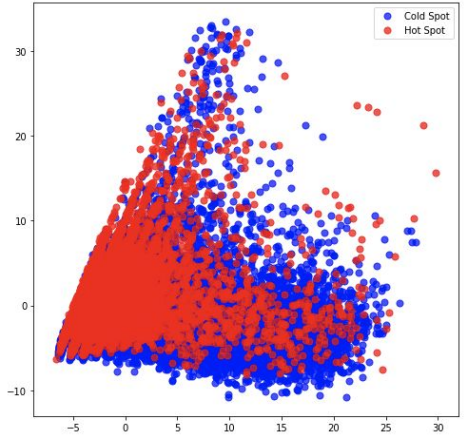
- *Finding sequence motifs in prokaryotic genomes - a brief ...* (n.d.). Retrieved March 9, 2022, from https://www.researchgate.net/publication/26317798_Finding_sequence_motifs_in_prokaryotic_genomes_-_A_brief_practical_guide_for_a_microbiologist
- Lange, J., Yamada, S., Tischfield, S. E., Pan, J., Kim, S., Zhu, X., Socci, N. D., Jasin, M., & Keeney, S. (2016, October 13). *The landscape of mouse meiotic double-strand break formation, processing, and Repair*. Cell. Retrieved March 6, 2022, from https://www.sciencedirect.com/science/article/pii/S0092867416313174
- *Location matters in recombination*. Fred Hutch. (2018, September 17). Retrieved March 6, 2022, from https://www.fredhutch.org/en/news/spotlight/2018/09/bs_nambiar_moleculcarcell.html
- Yadav, T., https://orcid.org/0000-0001-6557-7204, J.-P. Q., , G. A. https://orcid.org/0000-0001-5570-0723, Tejas YadavInstitut Curie, 75248 P. C. 05, Jean-Pierre Quivy https://orcid.org/0000-0001-6557-7204Institut Curie, 75248 P. C. 05, & Geneviève Almouzni* https://orcid.org/0000-0001-5570-0723 Institut Curie, 75248 P. C. 05. (2018, September 28). *Chromatin plasticity: A versatile landscape that underlies cell fate and identity*. Science. Retrieved March 6, 2022, from https://www.science.org/doi/10.1126/science.aat8950
- Maruf, M. A. A., & Shatabda, S. (2018, June 20). *IRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's pseudo components*. Genomics. Retrieved March 12, 2022, from https://www.sciencedirect.com/science/article/pii/S0888754318302143
- Jiang, P., Wu, H., Wei, J., Sang, F., Sun, X., & Lu, Z. (2007). RF-DYMHC: Detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features. *Nucleic Acids Research*, *35*(Web Server). https://doi.org/10.1093/nar/gkm217
- Liu, G., Liu, J., Cui, X., & Cai, L. (2012). Sequence-dependent prediction of recombination hotspots in saccharomyces cerevisiae. *Journal of Theoretical Biology*, *293*, 49–54. https://doi.org/10.1016/j.jtbi.2011.10.004
- Chen, W., Feng, P.-M., Lin, H., & Chou, K.-C. (2013). IRSpot-PSEDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Research*, *41*(6). https://doi.org/10.1093/nar/gks1450

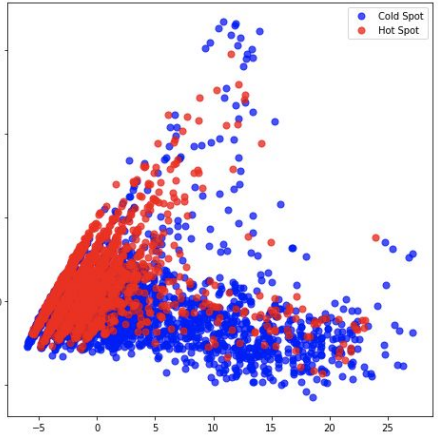

A C G T


Cold Spot
Hot Spot


Cold Spot
Hot Spot


A C G T
H3K4me3
H3K36me3
DNase
SNP


Cold Spot
Hot Spot


Cold Spot
Hot Spot

Methods