Predicting functional elements in noncoding DNA

Meer Mustafa, Kathryn Wendorf, Neelang Parghi, Alexander Lucaci

Presentation outline

Introduction

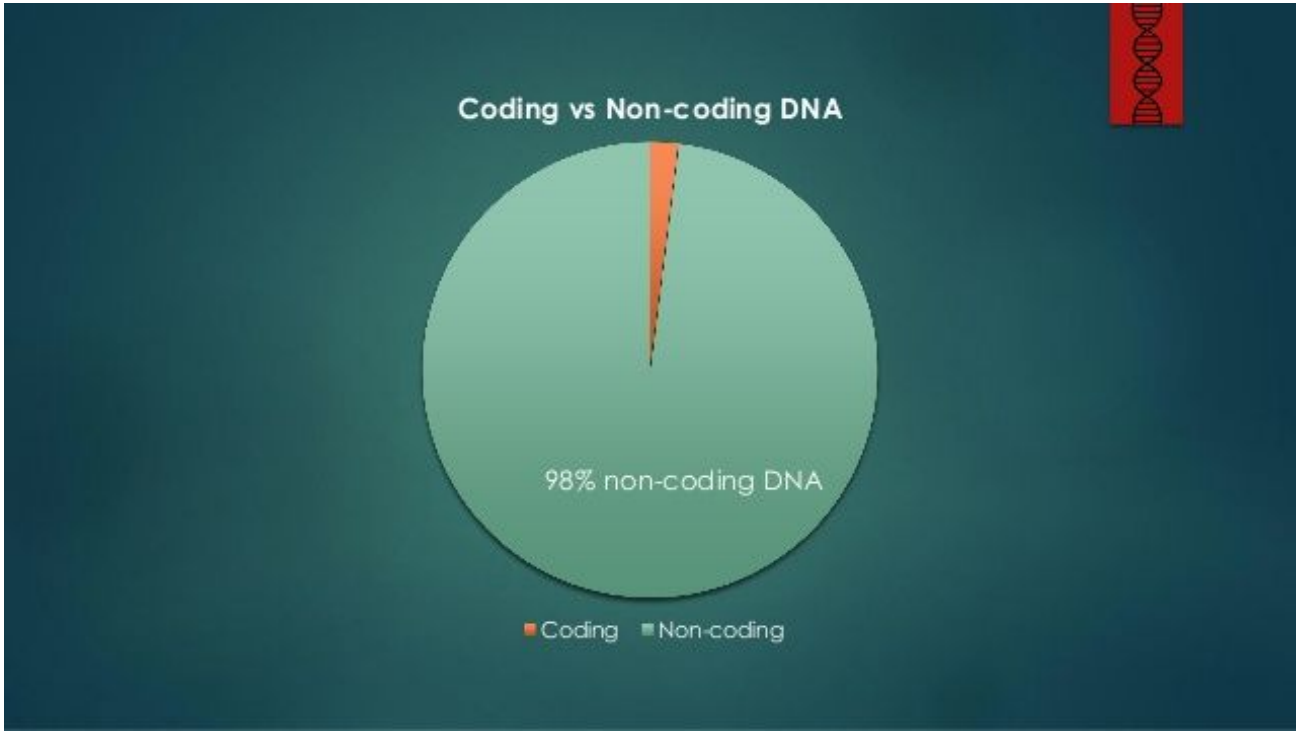
Question

Data types

Methods

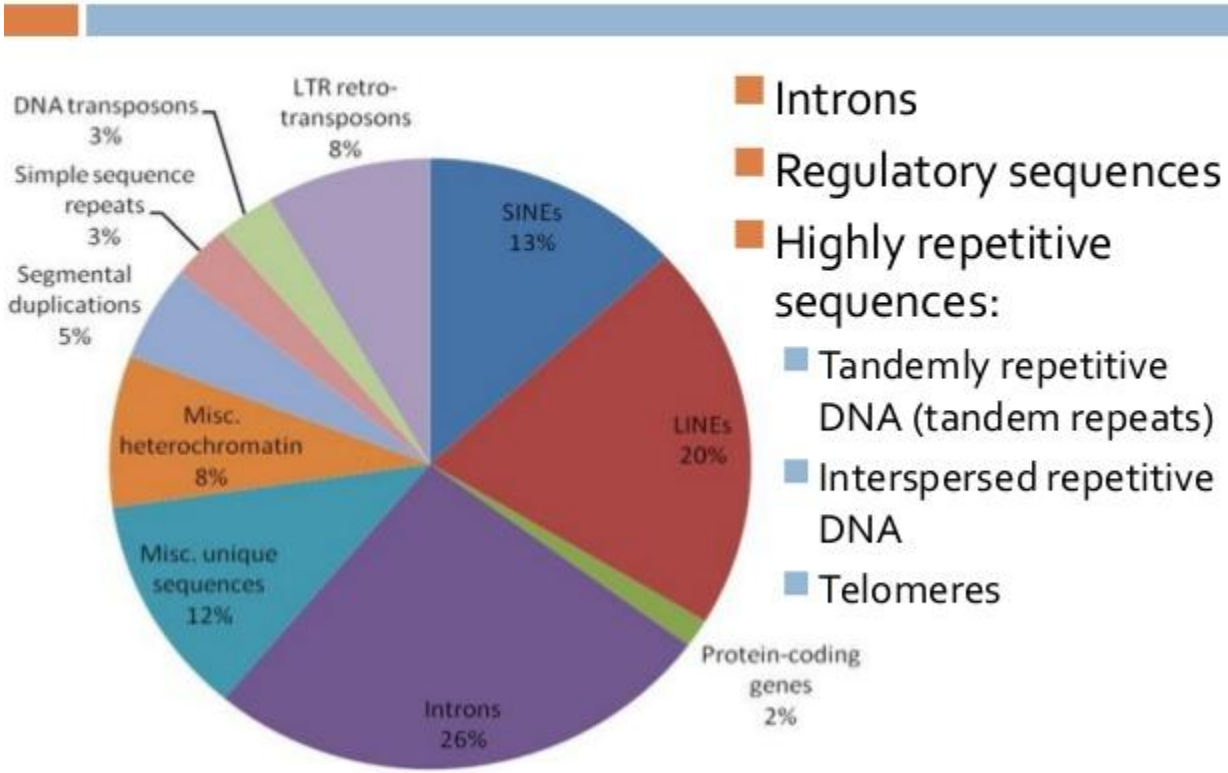
Results

Future directions



https://www.slideshare.net/planetcassandra/nyc-2013-analyzing-the-human-genomedna-with-cassandra

Classes of Noncoding Sequences



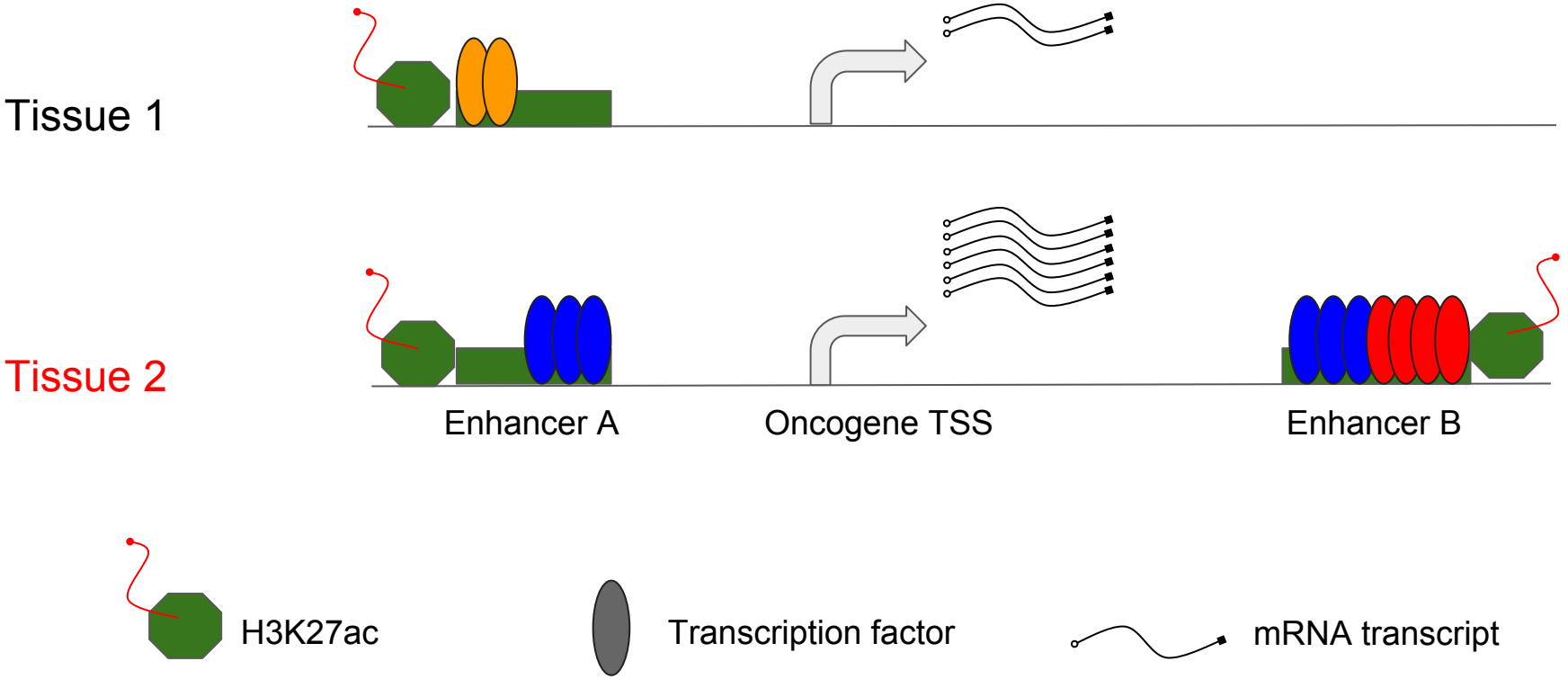
https://www.slideshare.net/thelawofscience/organization-eukaryotic-genome

Why is the non-coding genome important?

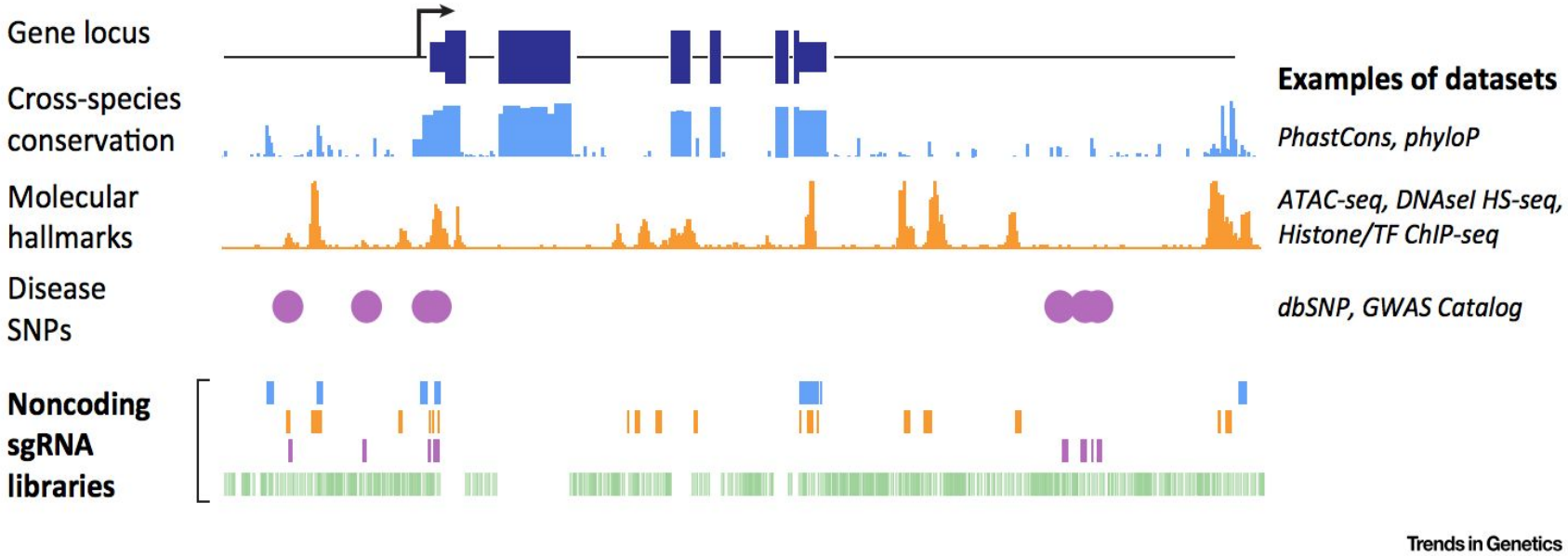
- ~98% of human genome does not code for proteins
- Non-coding genome affects gene regulation and disease
- Mapping of chromatin state and chromosome conformation has been used to identify regulatory elements.

Main problem: no overarching framework to translate the non-coding genomic sequence into functional elements.

Noncoding variation and regulation

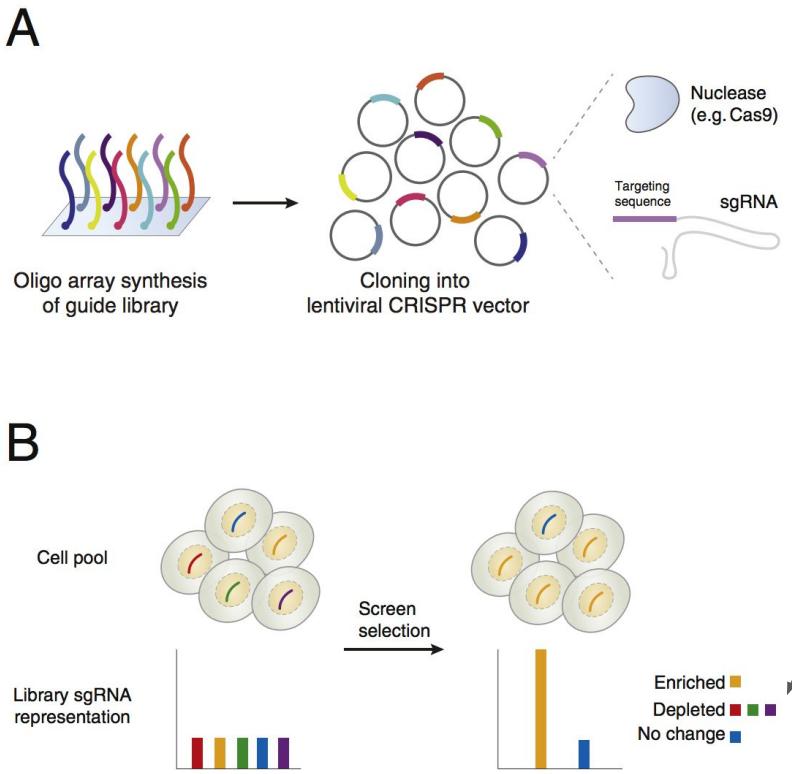


What are CRISPR screens?



Wright & Sanjana 2016

How do they work?



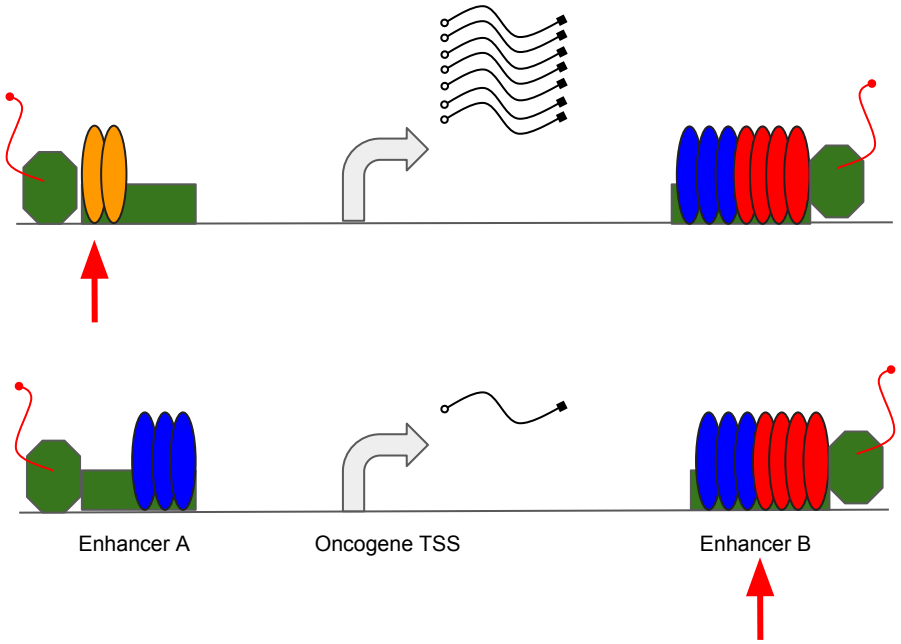
Guides that drop out were essential for survival

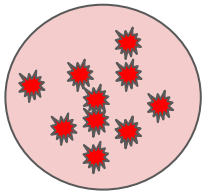
Sanjana et al 2016

Noncoding variation and regulation

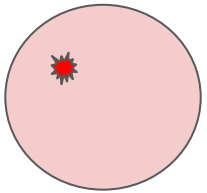
sgRNA #1

sgRNA #2





Cell growth



Cell death

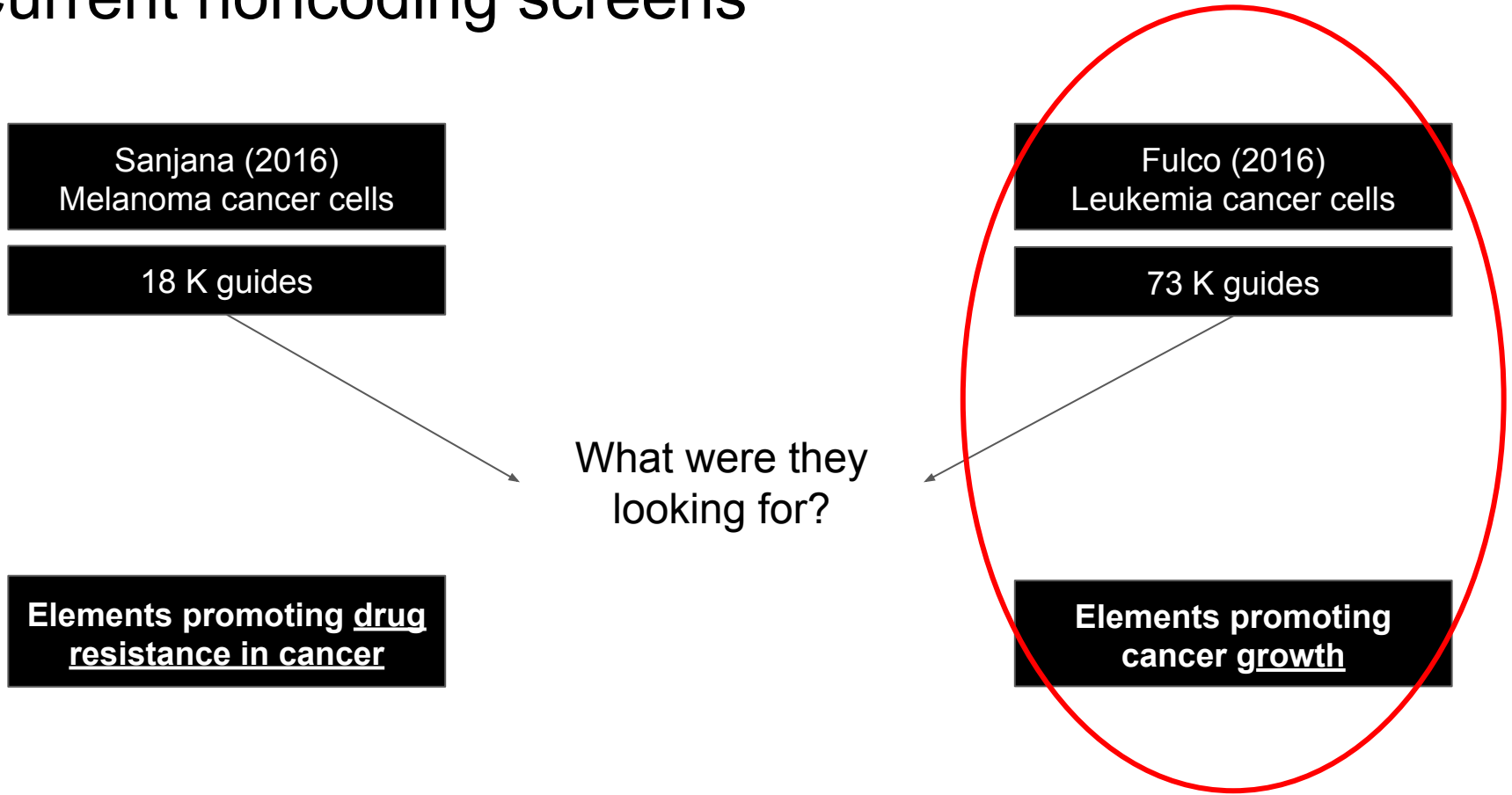
Project background

Sanjana lab generated CRISPR screen looking into BRAF inhibitor resistance in melanoma.

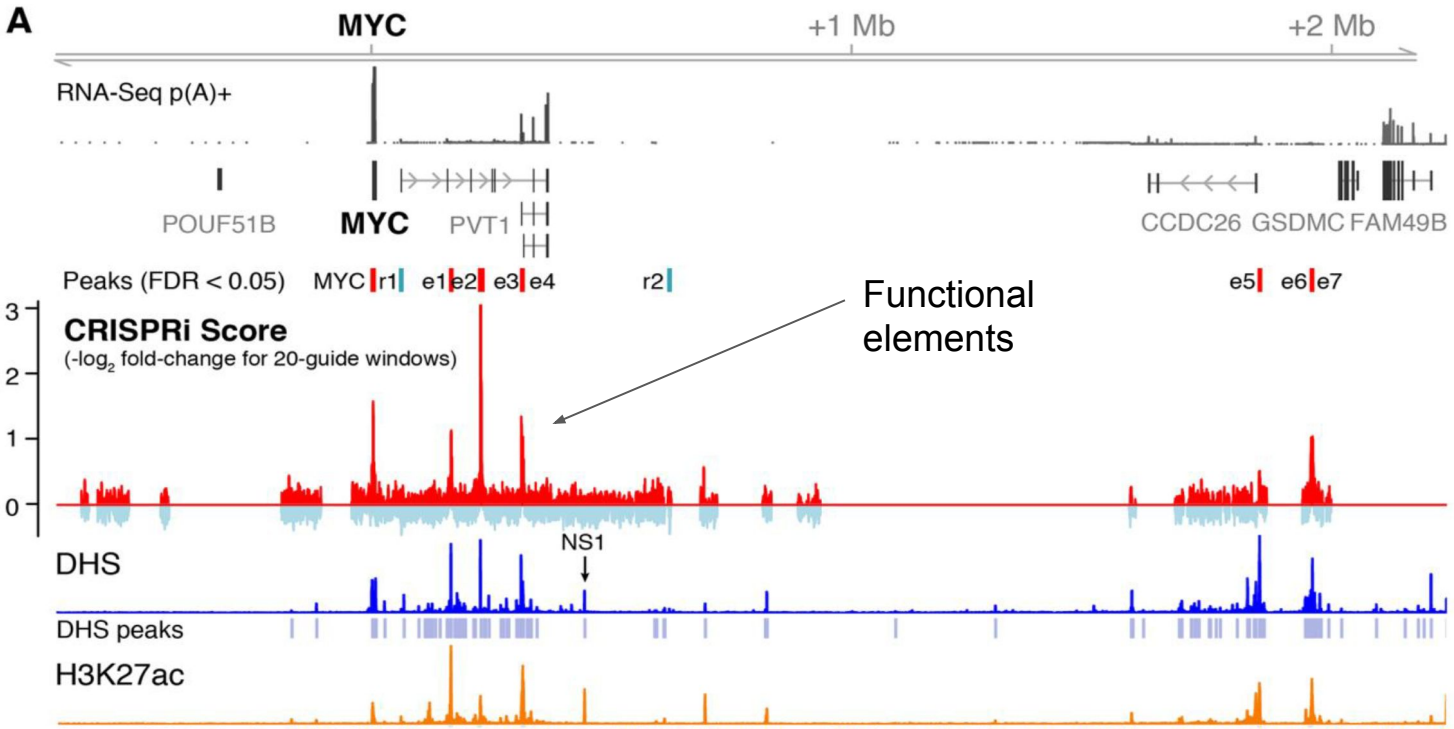
- Demonstrated that noncoding mutations are involved in gene regulation and chemotherapeutic resistance.

- Found that noncoding loci that modulate drug resistance also harbor predictive hallmarks of functional elements (e.g. enhancers, repressors)

Current noncoding screens

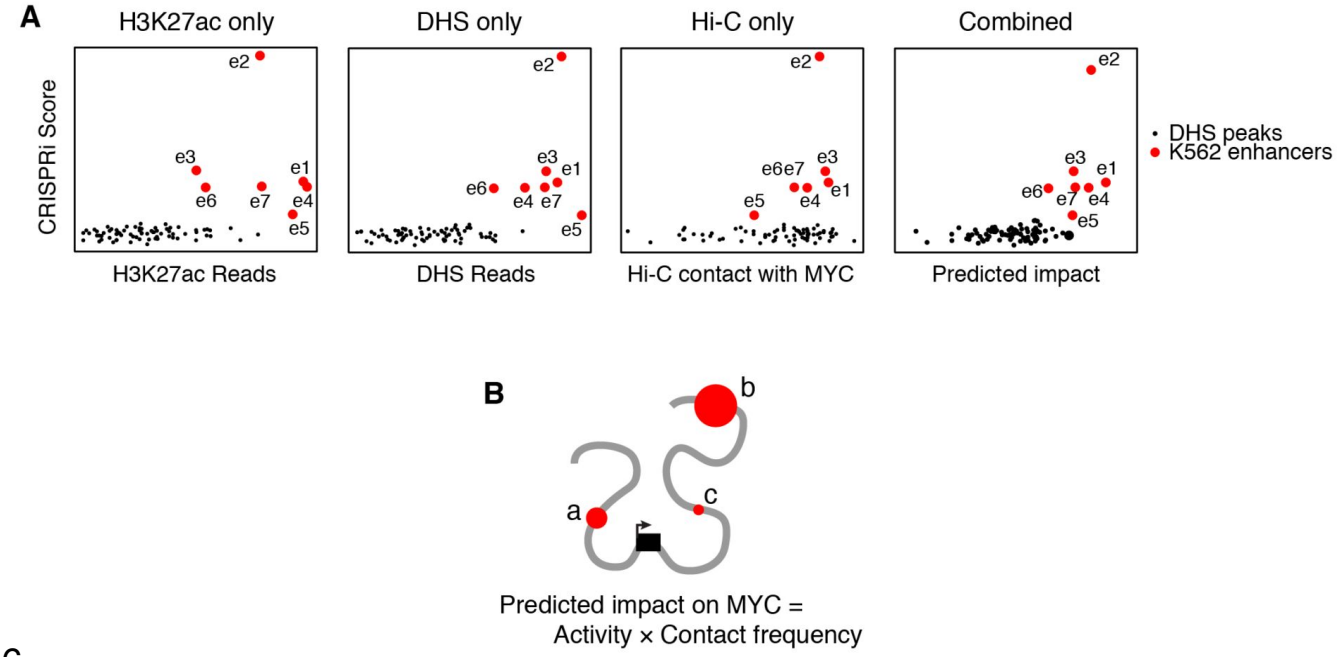


A noncoding screen to find functional elements



What's been done thus far with functional validation?

A very simple heuristic...

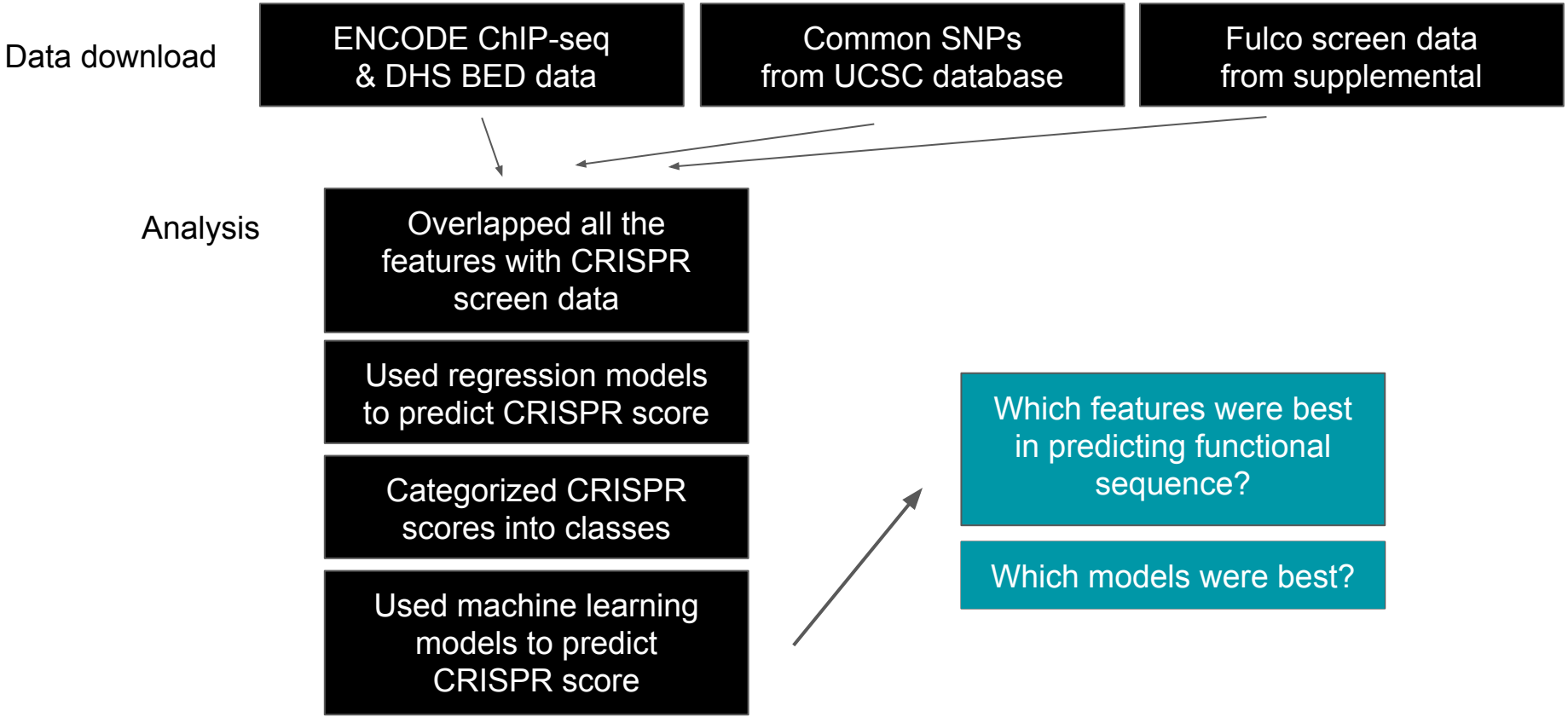


Fulco et al 2016

An open-ended question:

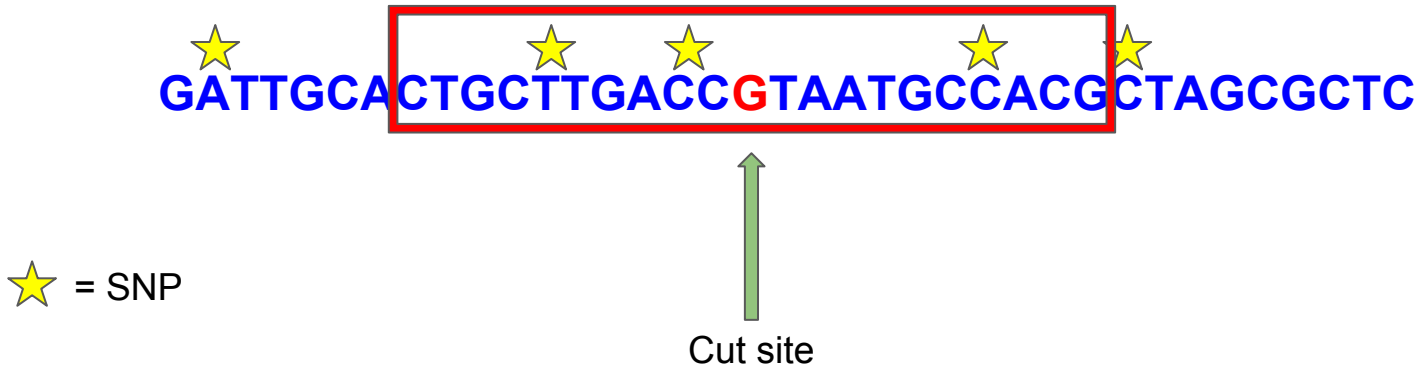
Given new biological data -- DNA-binding proteins, open chromatin, histone modifications, chromosome conformation capture, raw DNA sequence -- can we say whether it’s functional or not?

Computational workflow

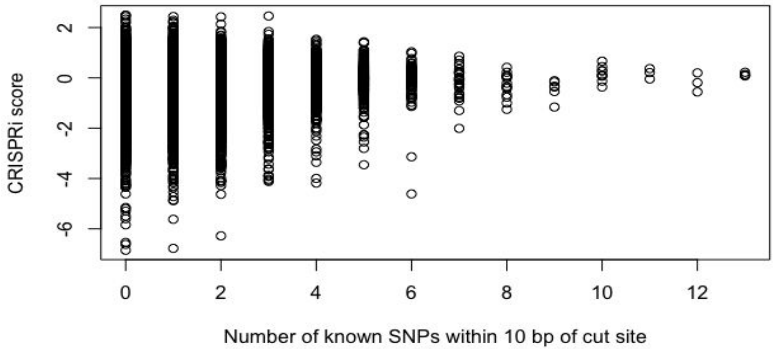


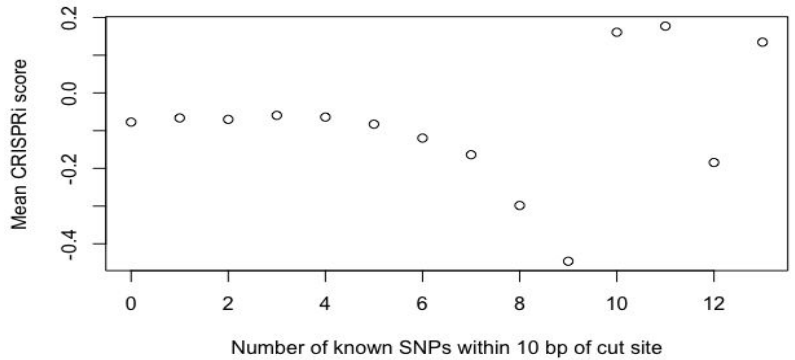
Is functionality correlated with human-to-human conservation?

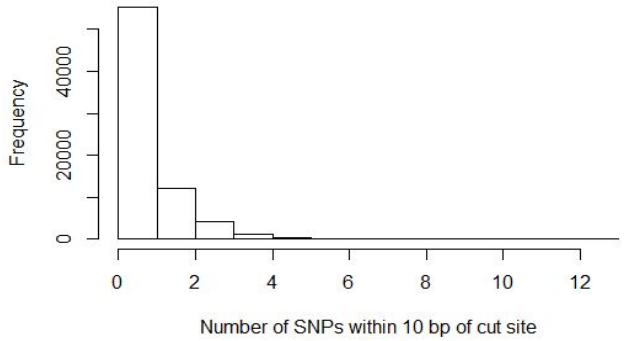
- A list of known SNPs and their allele frequencies in the region of our CRISPRi scores was queried from dbSNP (common SNPs list)
- For each CRISPR cut site, we counted how many SNPs fell within a distance of the site, and investigated how this number relates to CRISPRi score
- For example, the SNP number for this cut site would be 3.



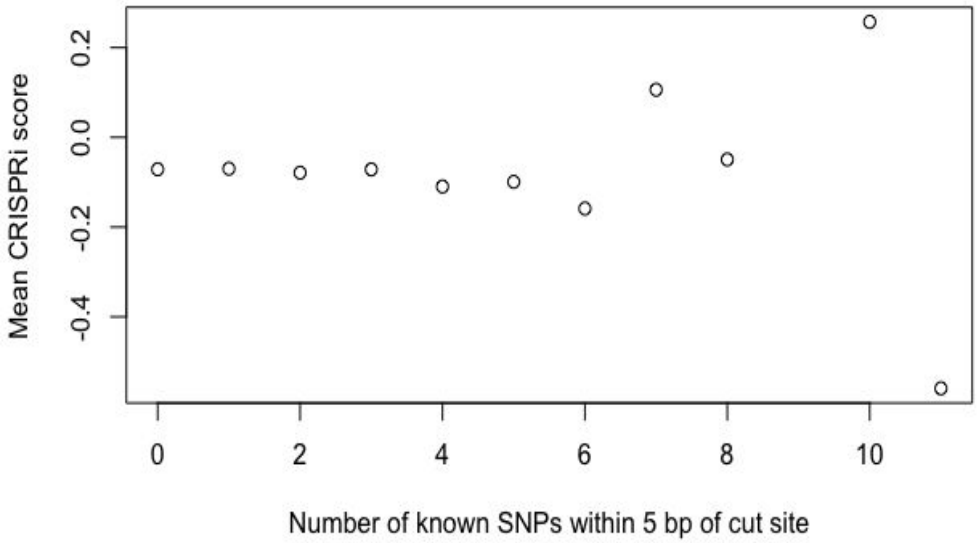
10 bp windows

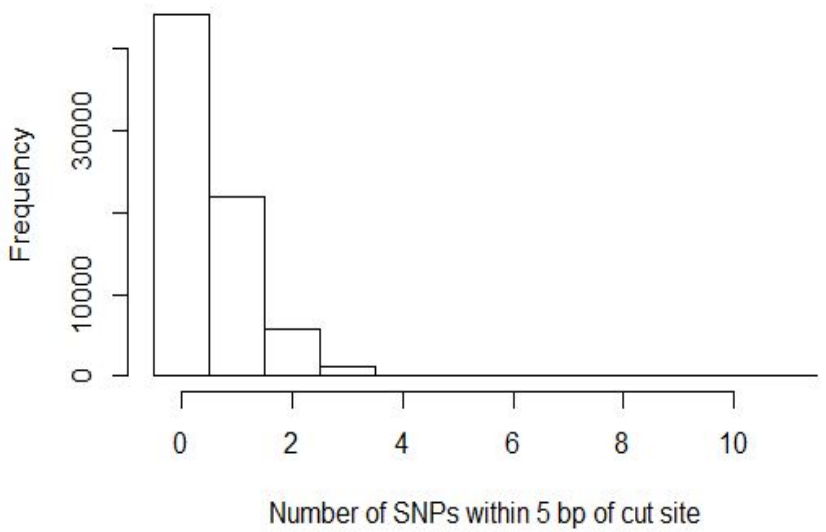




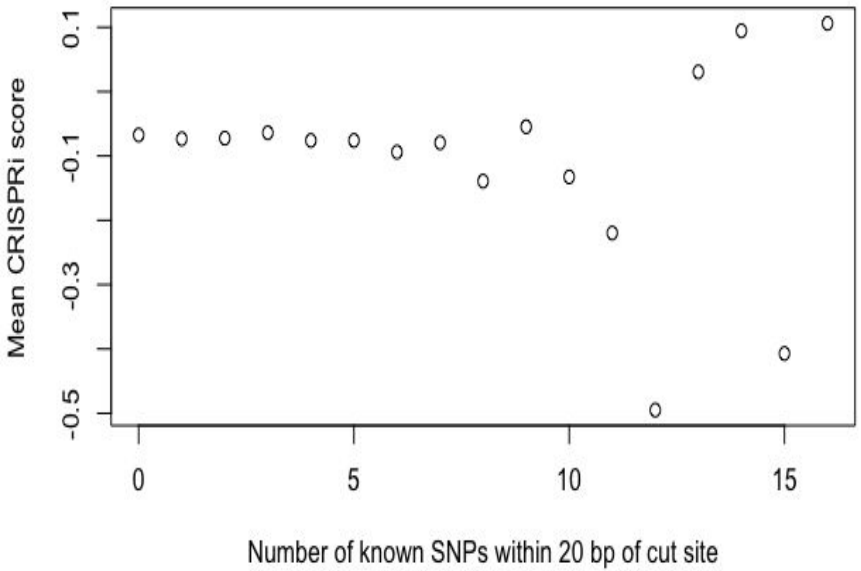


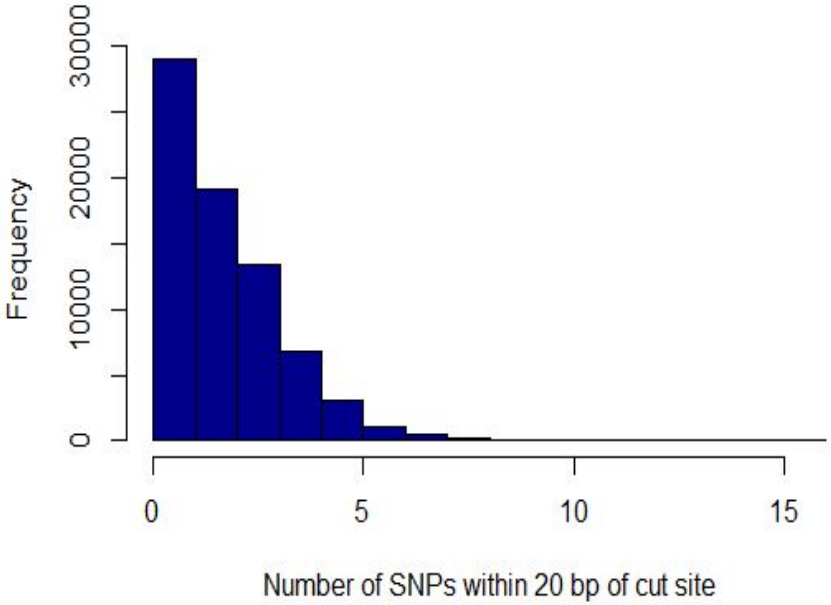
Similar for 5 bp...





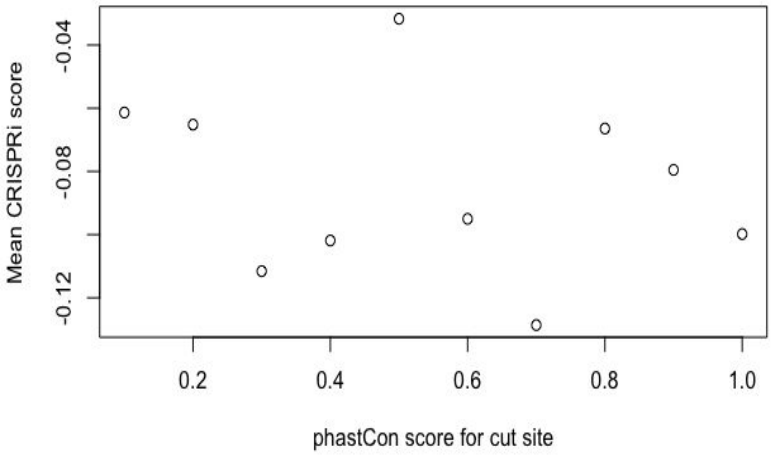
Not as good for 20 bp...

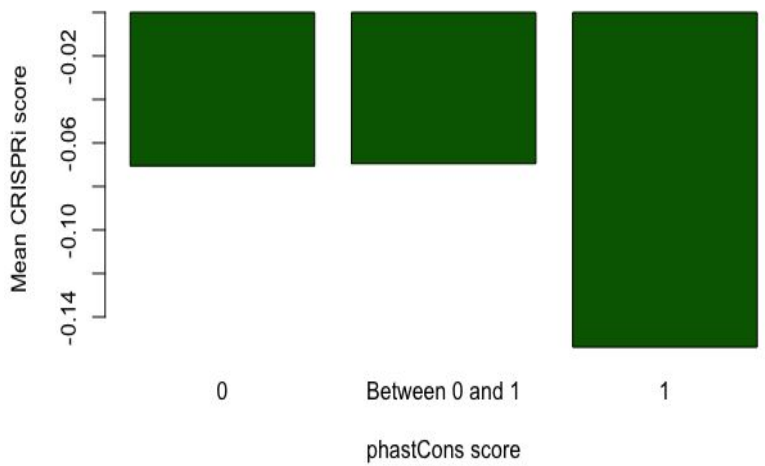




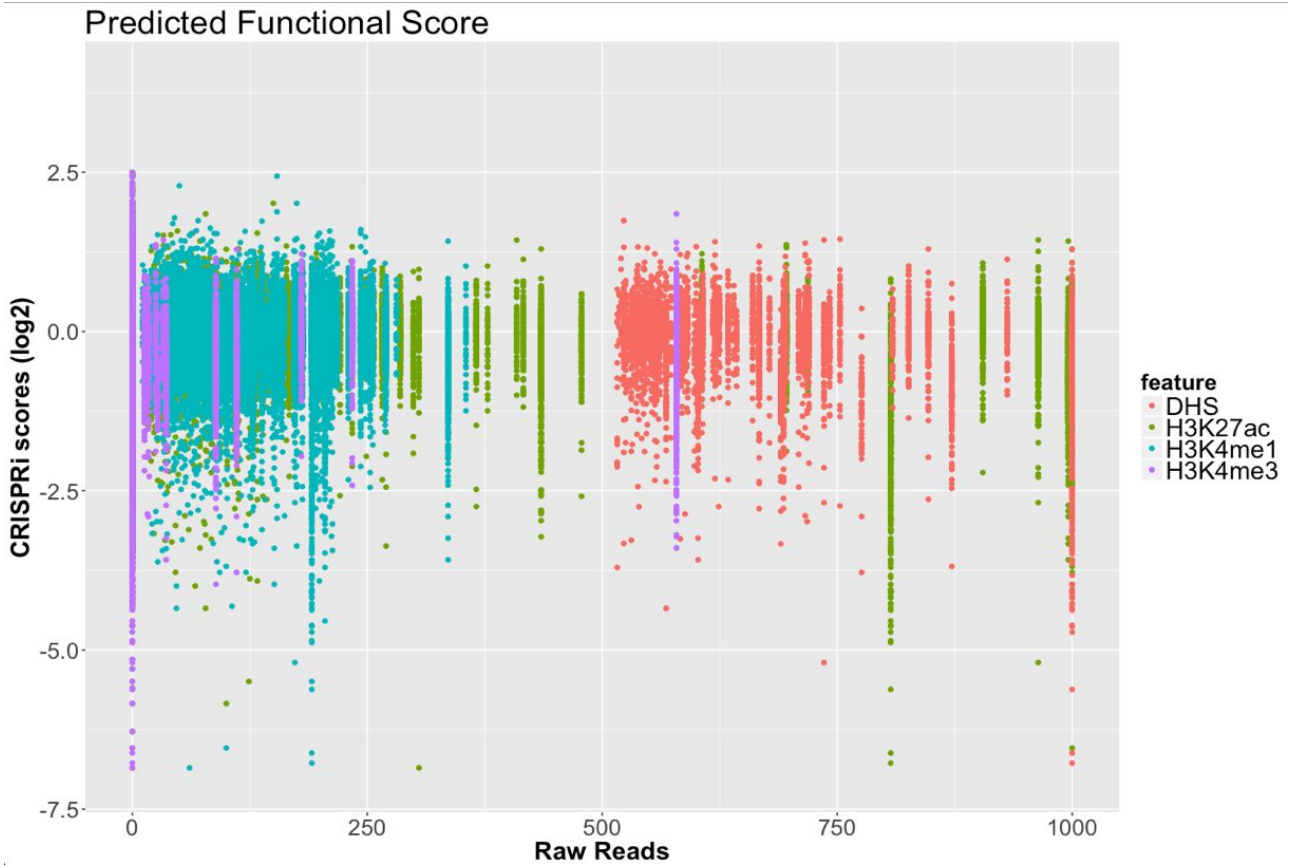
Conservation across vertebrate species

PhastCons: a score from 0 to 1 representing how conserved the sequence is





Other feature data



Actually normalized reads

Methods

Used three different prediction algorithms: SVM, Random Forest, KNN

Applied 10-fold cross validation to each:
the training and testing samples
entire data set.

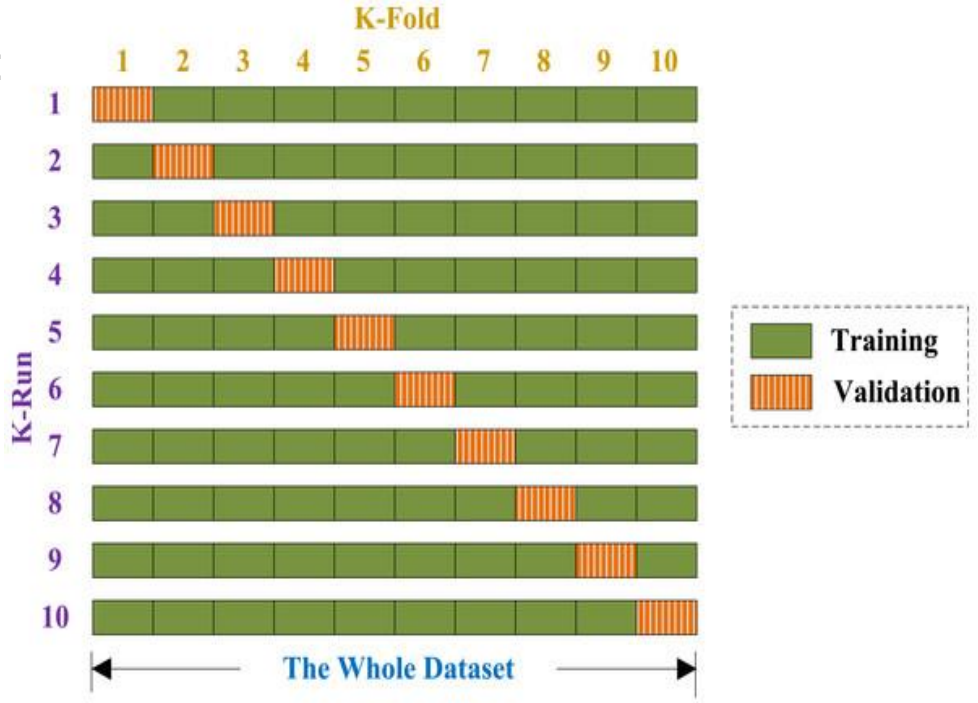


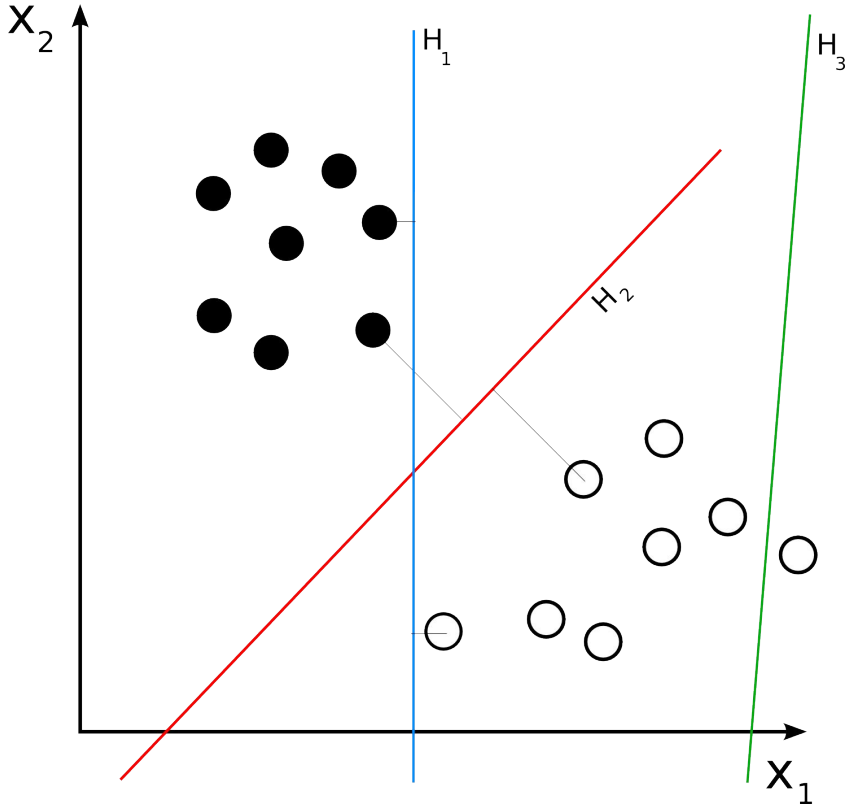
Image from: “Detection of Alzheimer’s disease by displacement field machine learning.” Y. Zhang and S. Wang. *PeerJ*.

SVM - support vector machines

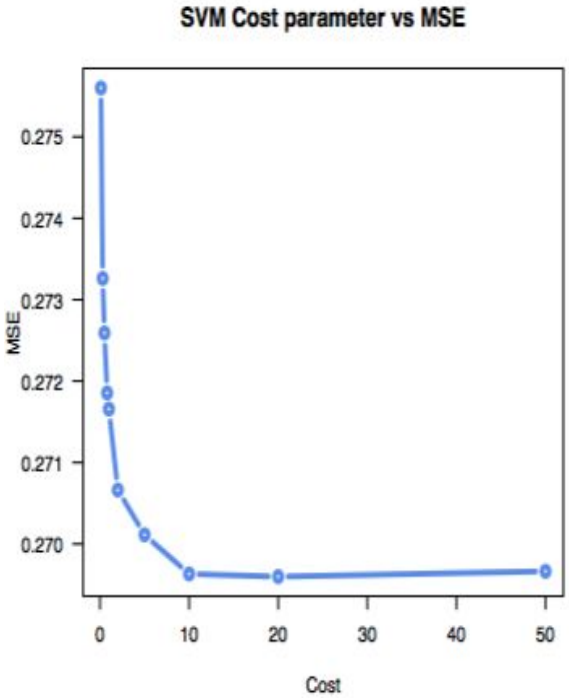
Idea: A classifier builds a model from a labeled training set and returns an optimal hyperplane that puts new examples into one of the categories non-probabilistically.

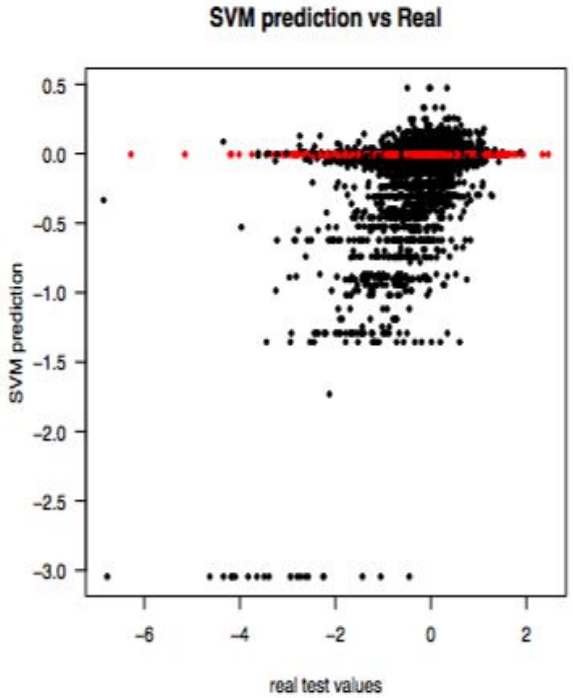
Goal: Maximize the hyperplane margin.

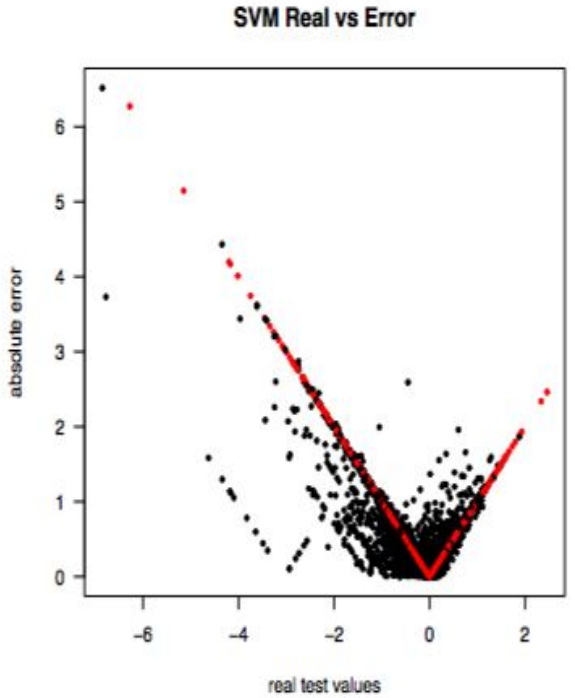
This represents the decision boundary.



SVM - support vector machine results







red points are where there are 0 values for all features i.e. no presence of any data type at the genomic site

Random Forest

Uses a series of decision trees -> A forest!

Create random bootstrapped subsets of data and create decision trees from them.

What class does each tree predict? The these outputs is our prediction.

Bootstrap aggregating: “bagging.”

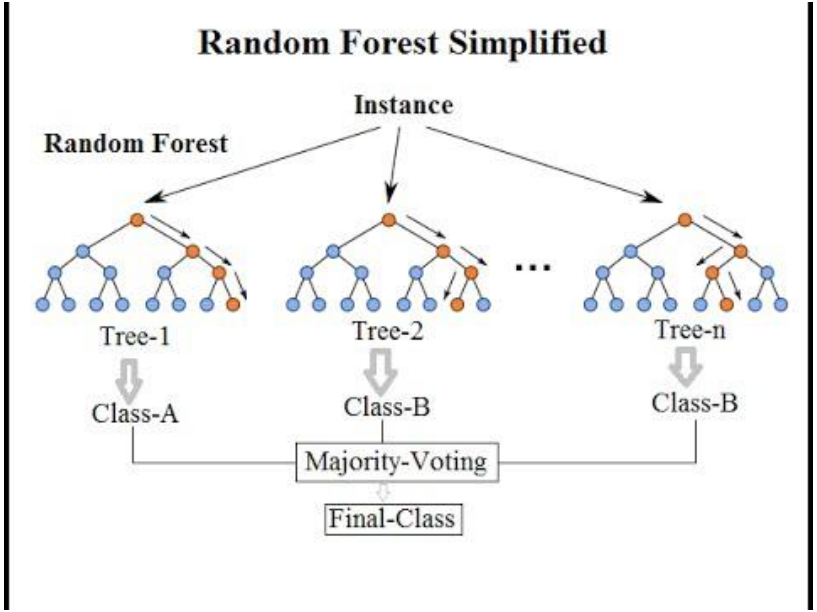
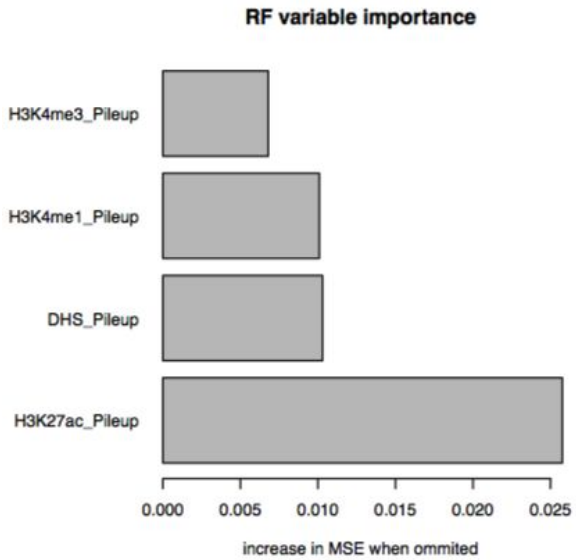
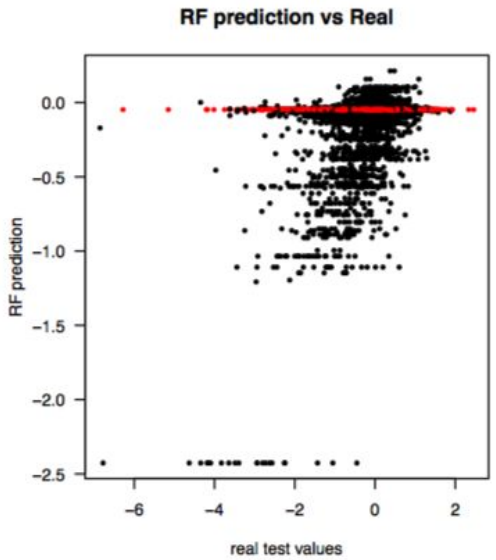
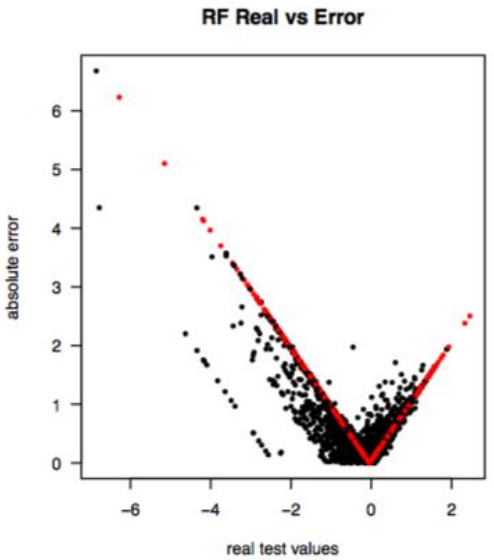


Image from “Random Forest based Classification:” https://www.youtube.com/watch?v=ajTc5y3OqSQ

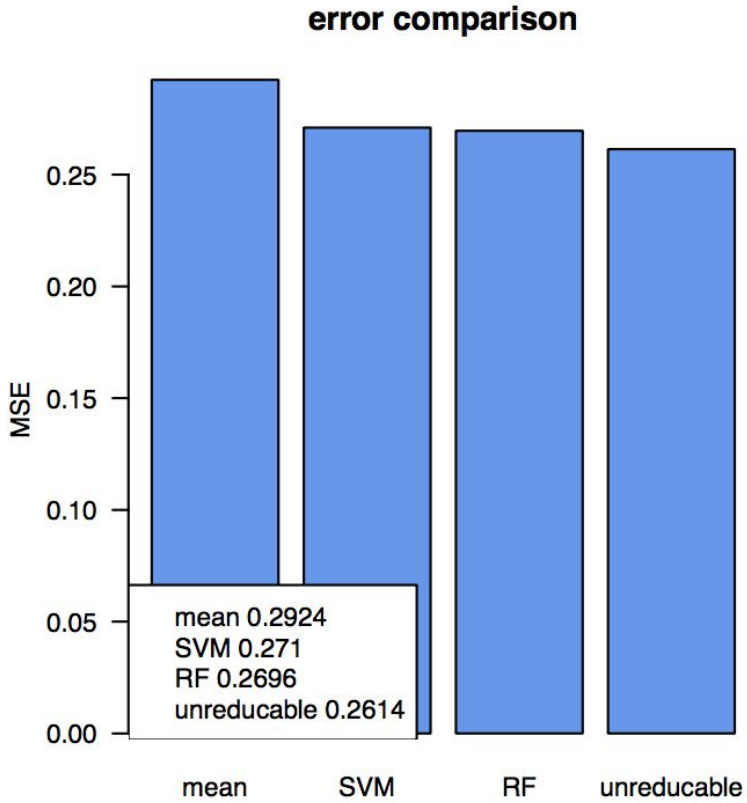
Random Forest - results

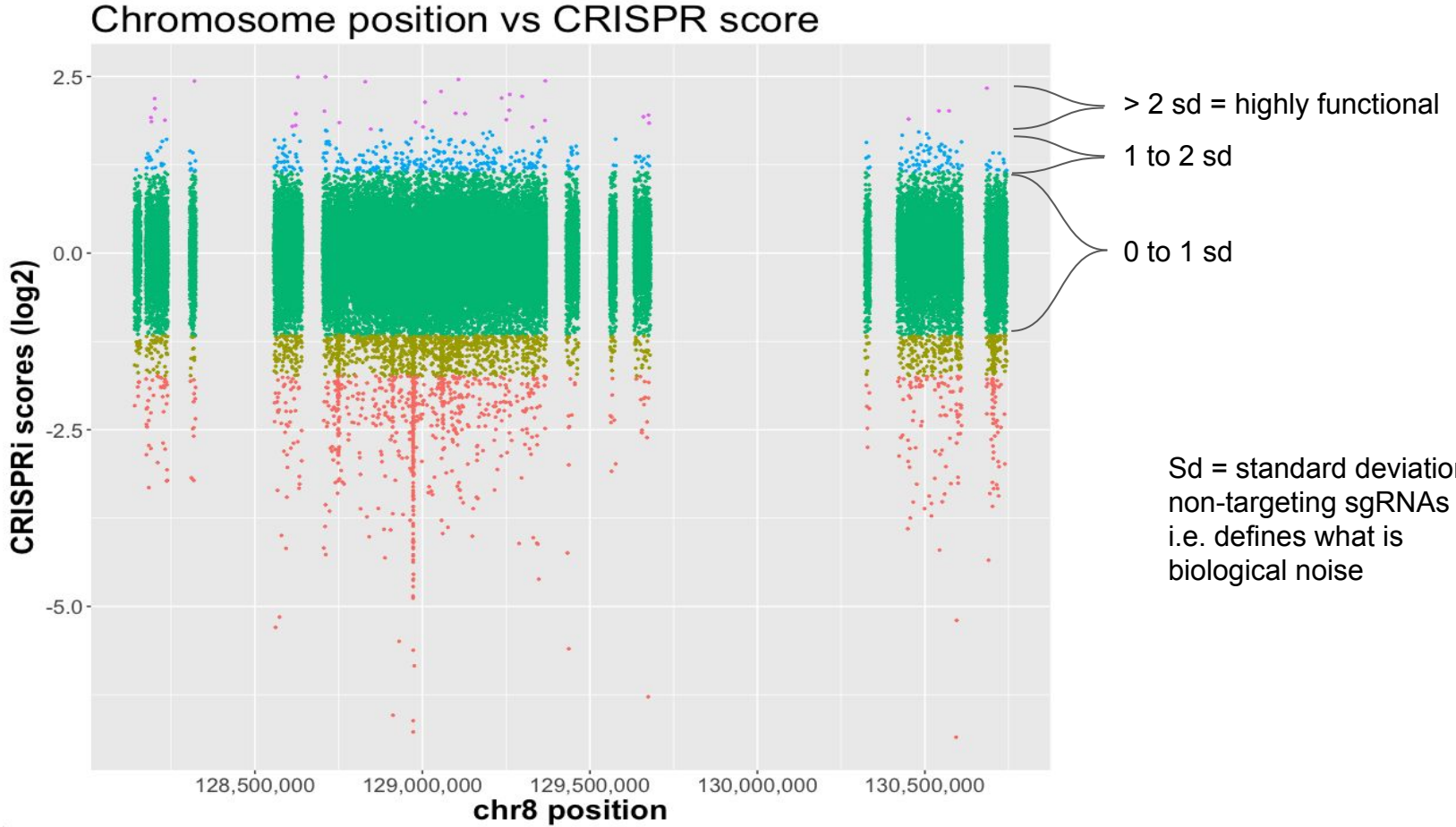




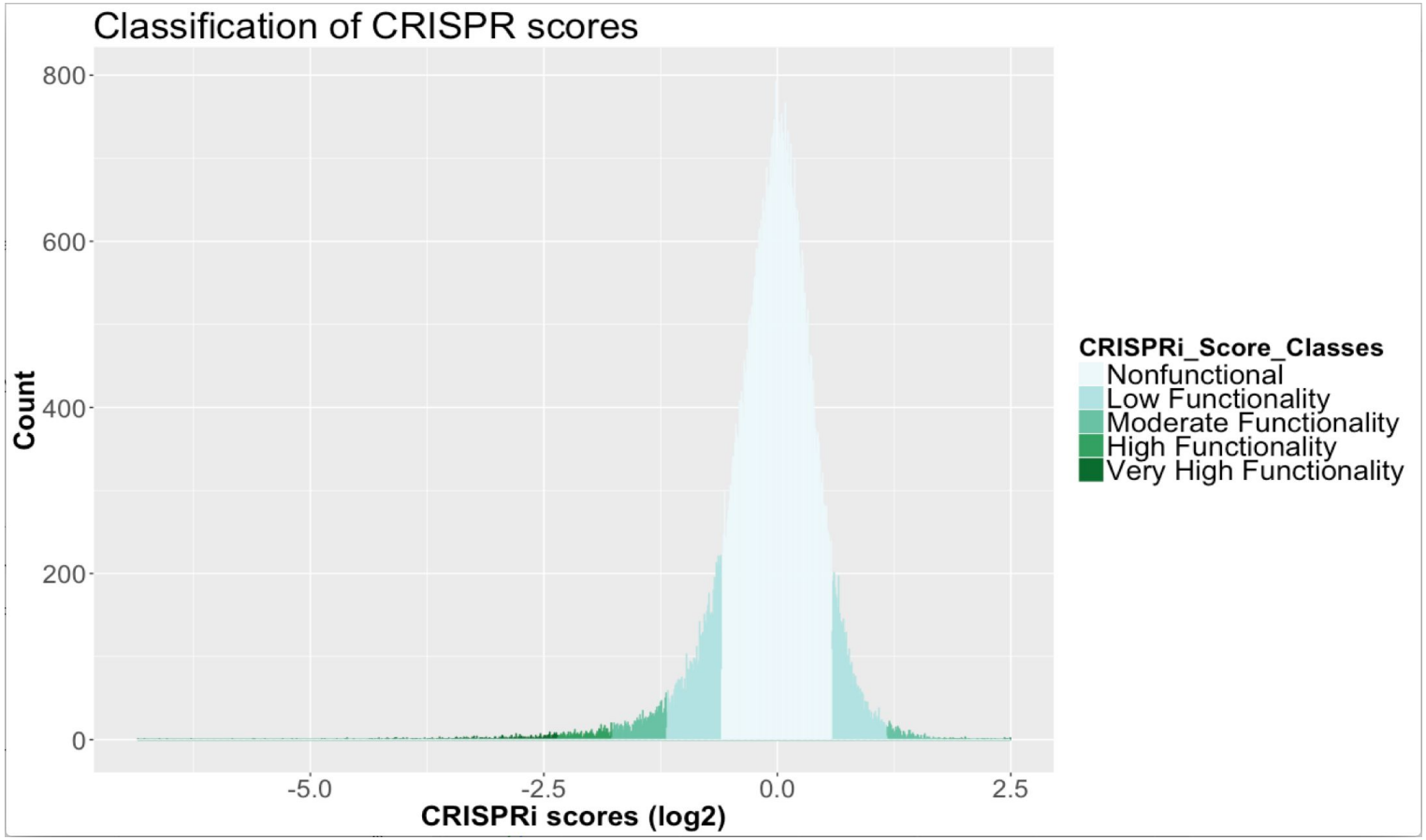


Comparing SVM and RF





Sd = standard deviation of non-targeting sgRNAs
i.e. defines what is biological noise


Classification of CRISPR scores
800
600
400
200
0
Count
-5.0
-2.5
0.0
2.5
CRISPRi scores (log2)
CRISPRi_Score_Classes
Nonfunctional
Low Functionality
Moderate Functionality
High Functionality
Very High Functionality

K-nearest neighbors

Makes predictions based on an instance's nearest neighbors.

Very simple: no need for assumptions, just look at the neighbors. Can be robust noise.

Interpreting model can be difficult: can't say, "As this happens to X, that happens to Y."

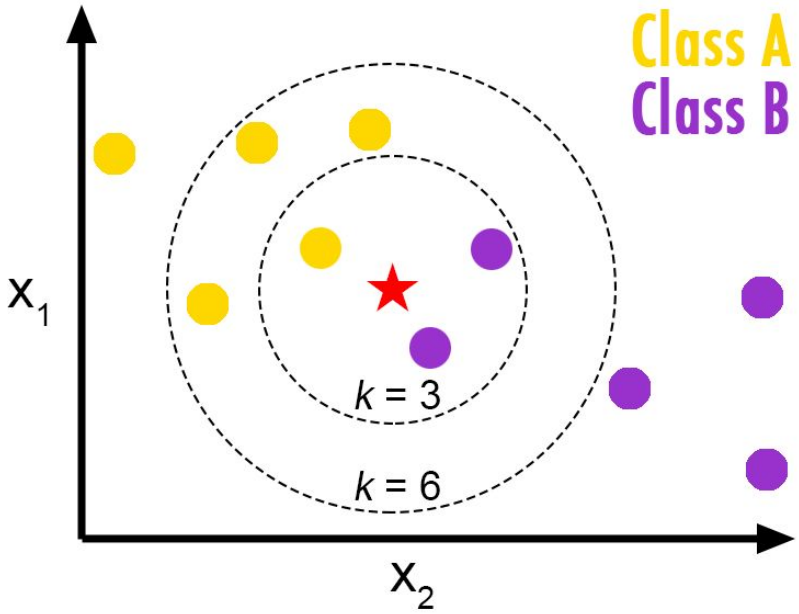
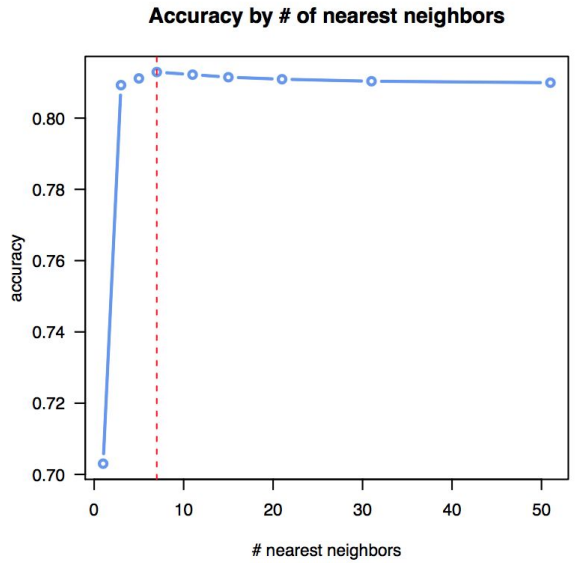


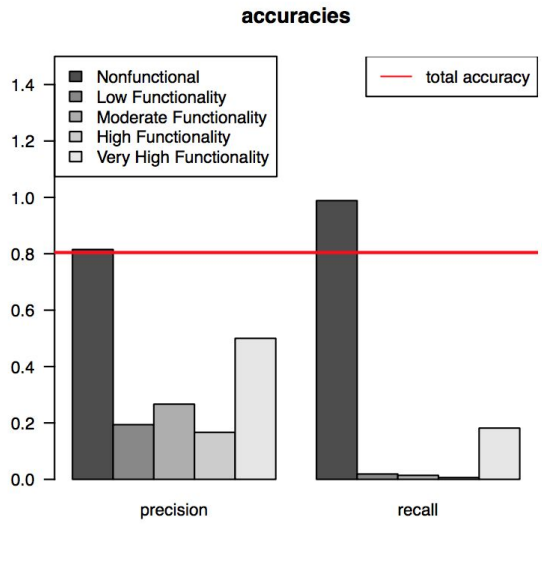
Image from: http://en.proft.me/2017/01/22/classification-using-k-nearest-neighbors-r/

K-nearest neighbors - results

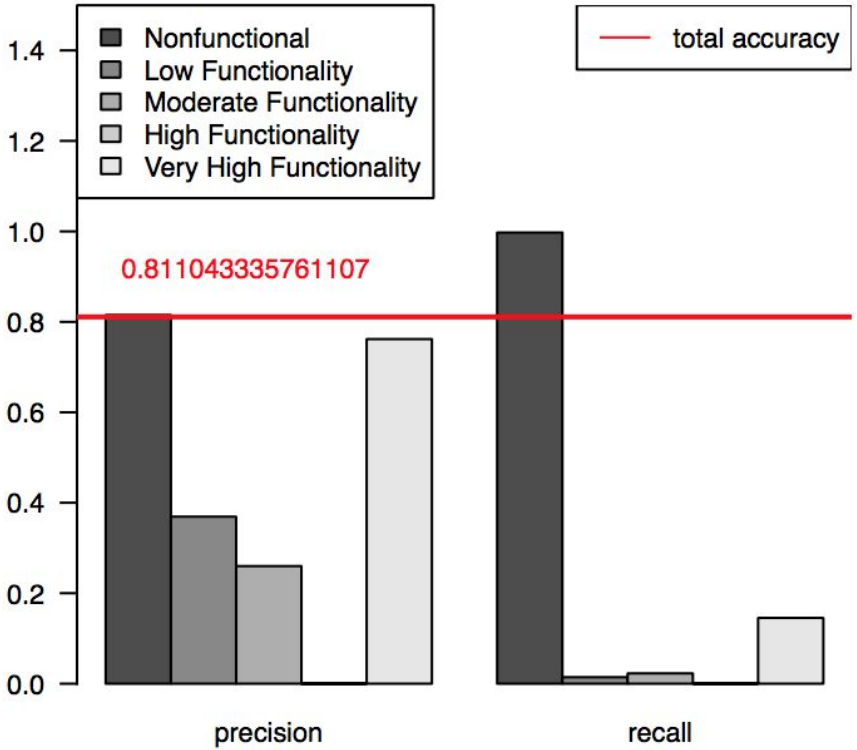
Divided the scores into five different classes

Highest accuracy: 80.44883% with k = 7

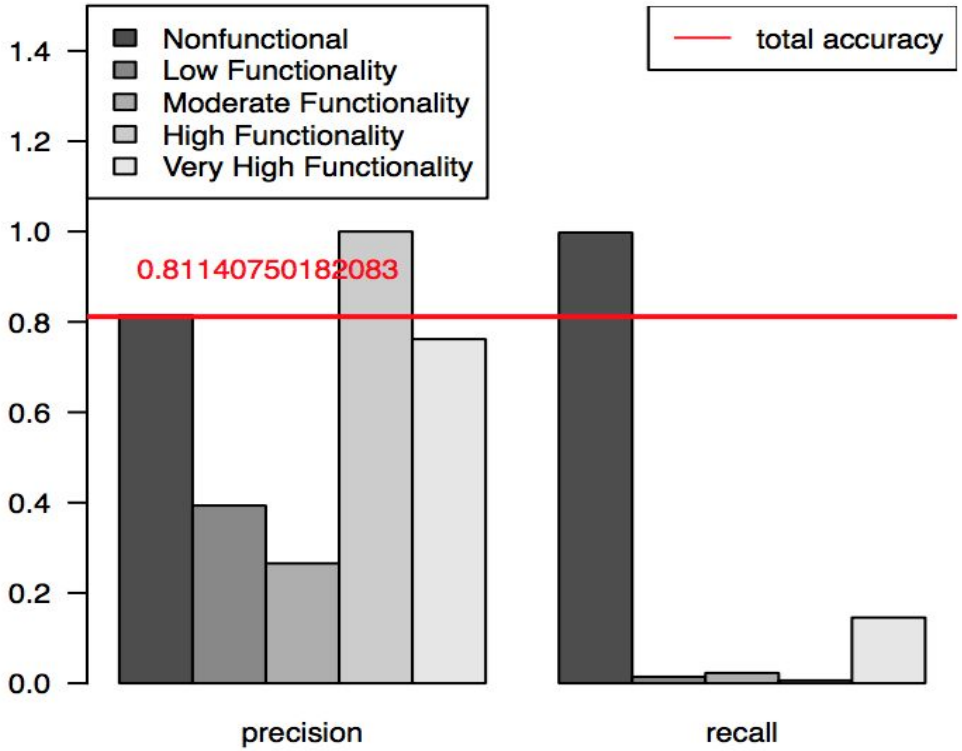


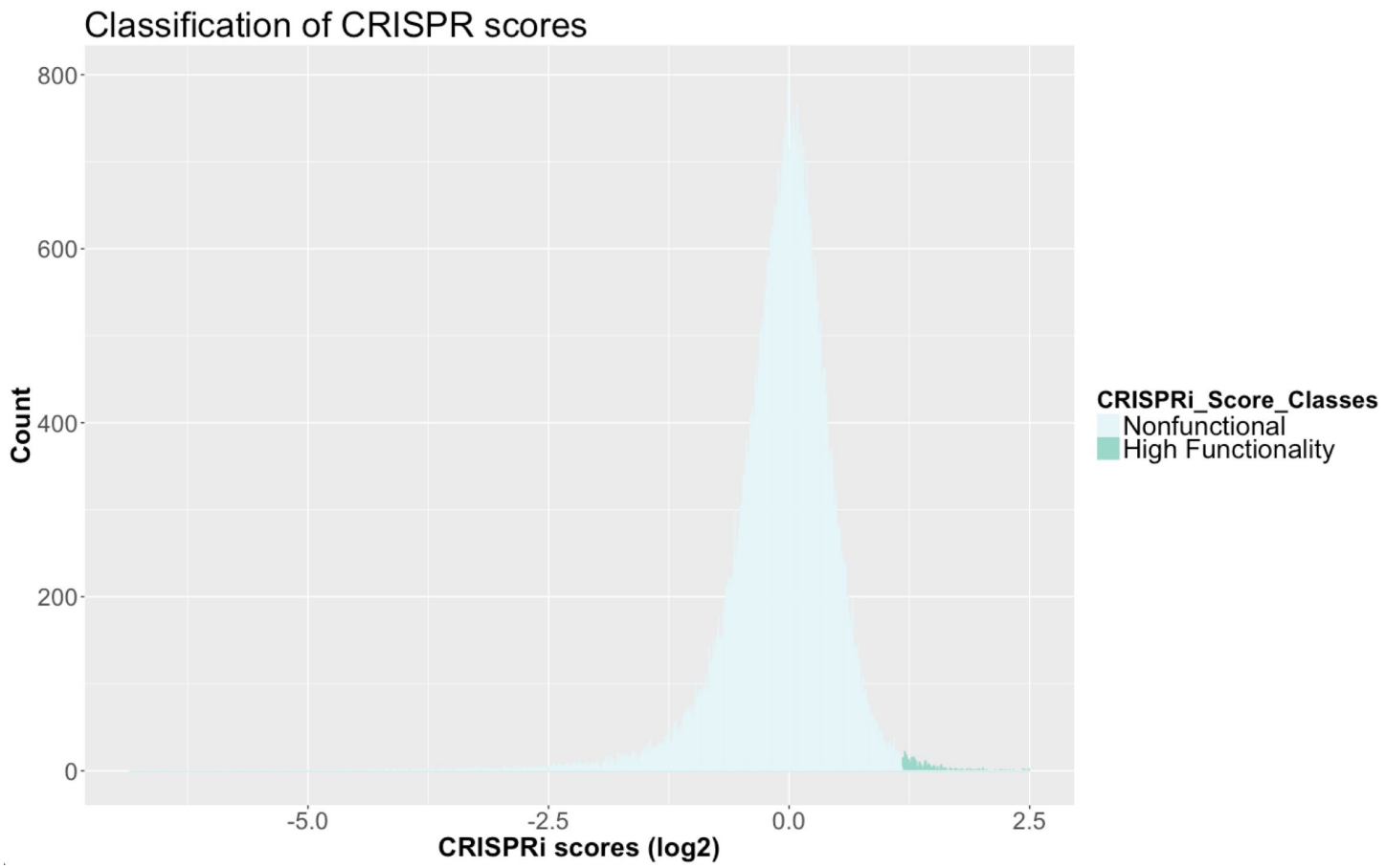


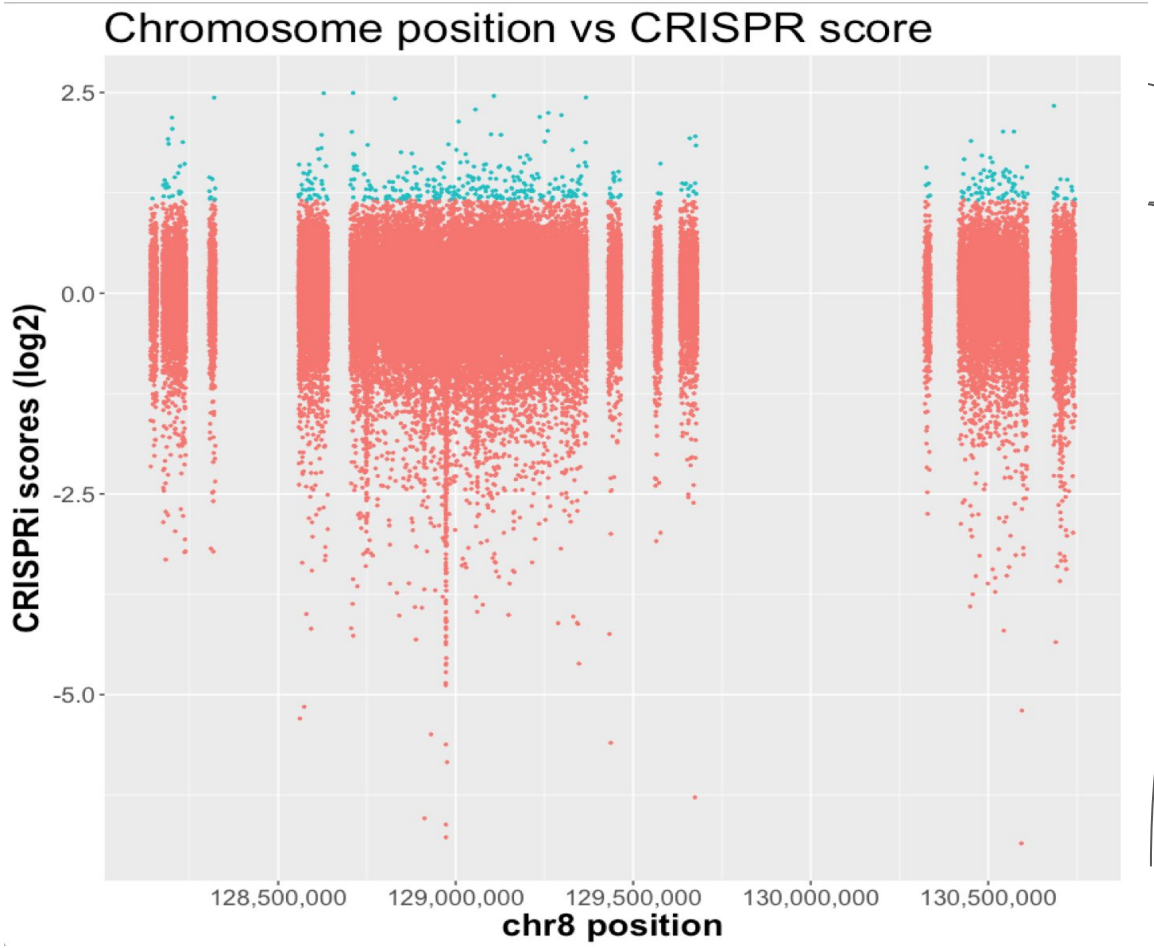
SVM results - 5 classes



Random Forest results - 5 classes




Classification of CRISPR scores
800
600
400
200
0
Count
-5.0
-2.5
0.0
2.5
CRISPRi scores (log2)
CRISPRi_Score_Classes
Nonfunctional
High Functionality



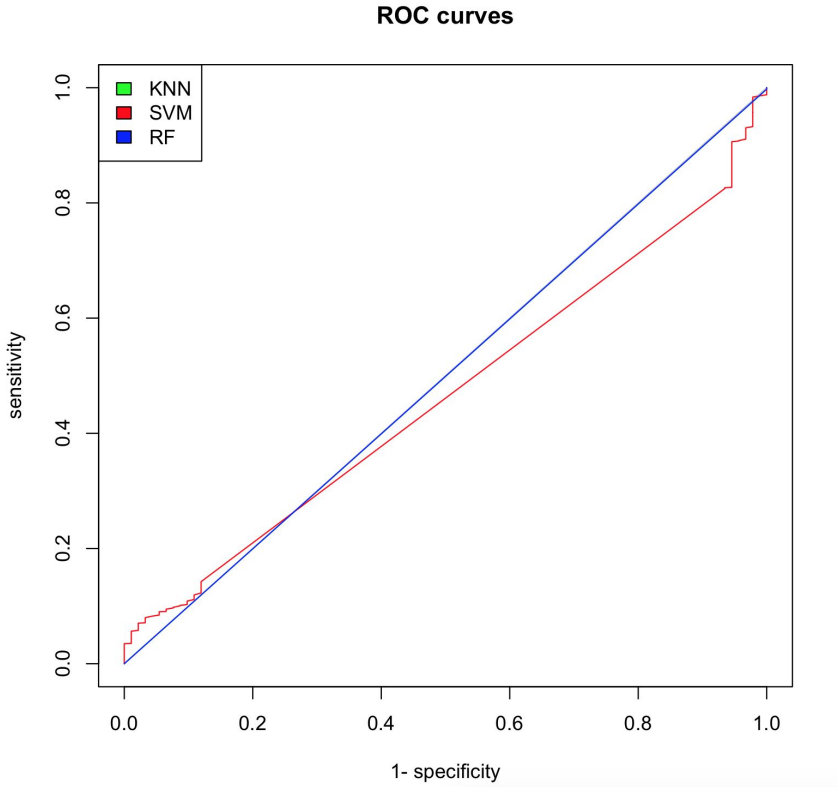
> 2 sd = highly functional

0 to 2 sd

Sd = standard deviation of non-targeting sgRNAs
i.e. defines the range of biological noise

Problems with two-class data

- Extremely unbalanced classes
- ROC plot of 50/50 balance between the two classes



Results

Sequence conservation - sequences conserved across humans have a slightly higher CRISPR score, while sequences conserved across vertebrates have a slightly lower CRISPR score

SVM - noticeable reduction in error from naive mean

Random Forest - slight improvement over SVM

KNN - didn’t perform as well as SVM or RF

All models were more accurate with 2 class

Future work

Find and analyze **cancer-associated** SNPs

Integrate other high-resolution data types into the model:

- Transcription factor binding information (ChIP-seq)
- Intensity for Hi-C contact
- Frequency of Hi-C contact
- ChIA-PET (promoter-enhancer interactions)

Understanding tissue-specific effects when comparing one screen to another

Much later… after we know which features are the most predictive, what do they mean in a biological sense?

Acknowledgements

Neville Sanjana

Brian Parker

Rich Bonneau

NEW YORK UNIVERSITY

NEW YORK GENOME CENTER®

Appendix

Testing different predictive models & resamplings

10-fold cross validation:

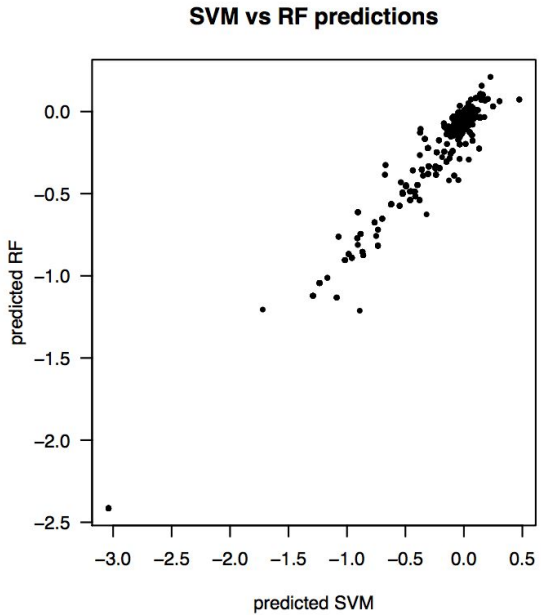
	Raw CRISPR score (Mean Squared Error)	5 classes (Mean Misclassification Error)	2 classes (Mean Misclassification Error)
SVM	0.20	0.20	
Random Forest	0.189	0.188	0.00541
KNN		0.231	0.0125

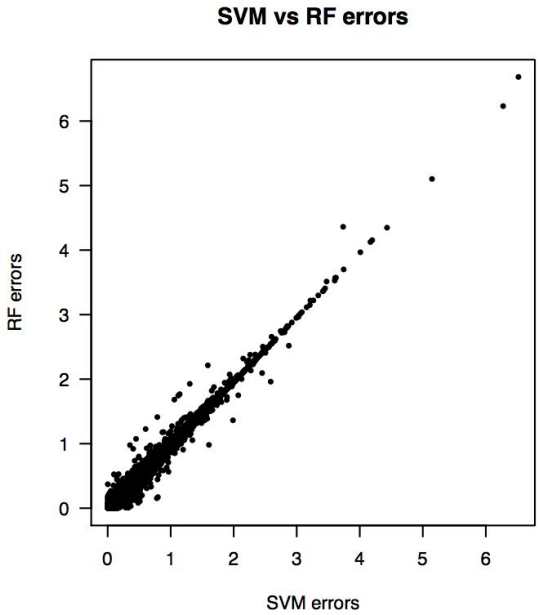
Bootstrapping (resampling the dataset using replacement):

SVM			
Random Forest		0.188	0.00534
KNN		0.228	0.0141

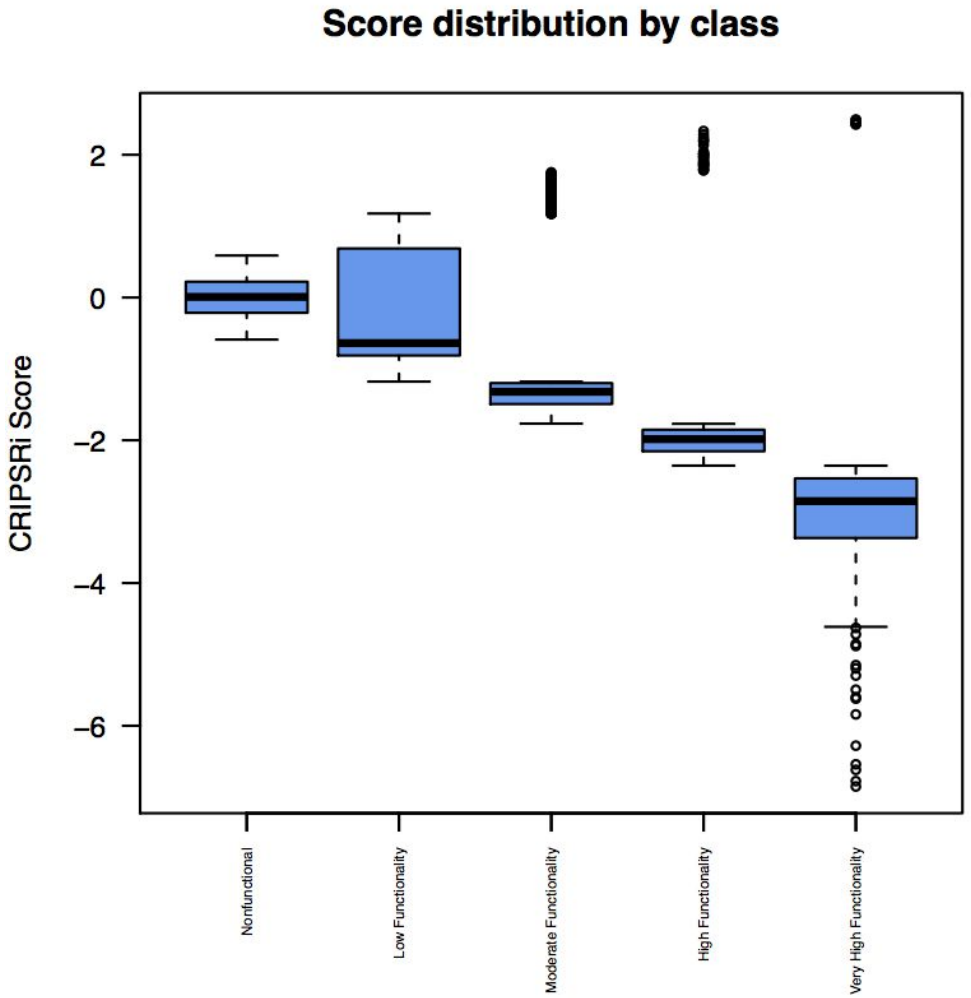
SVM vs. RF (without classes)

The errors show a high correlation, meaning that if SVM makes a mistake - RF will also make a similar mistake on that entry.

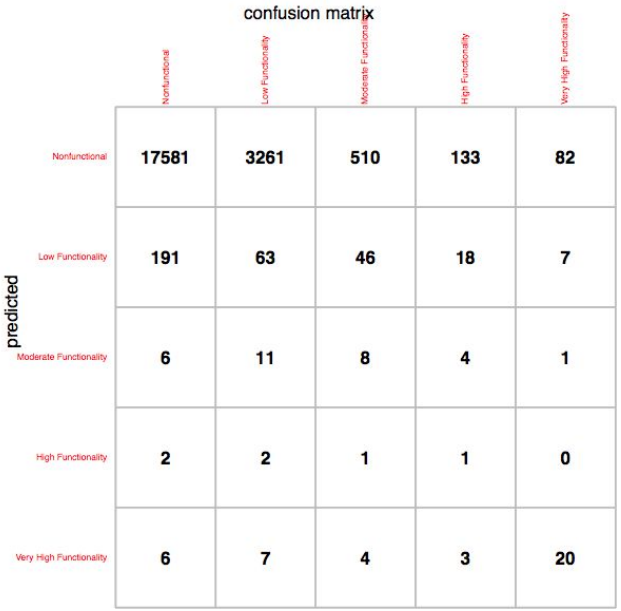


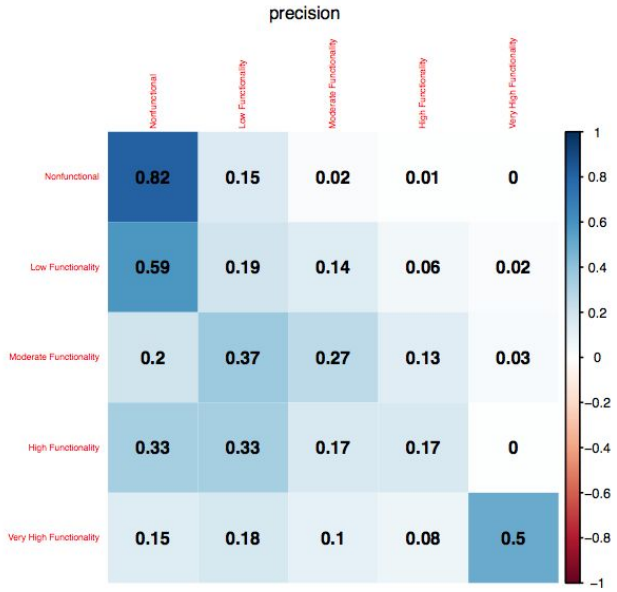


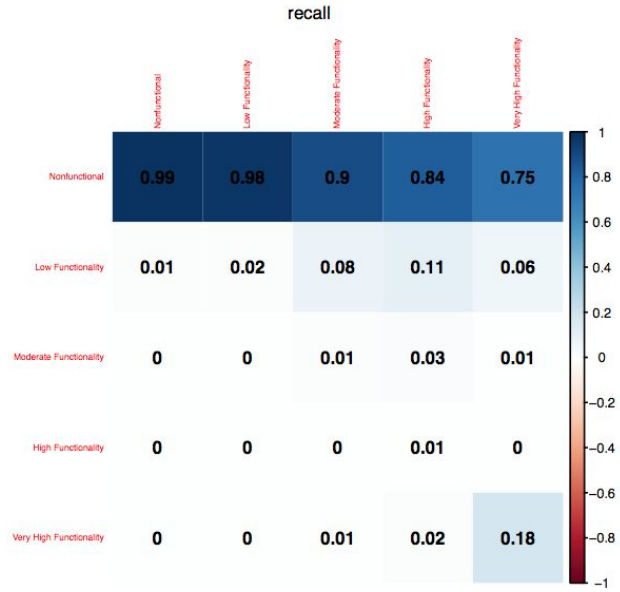
Distribution of 5 classes by score



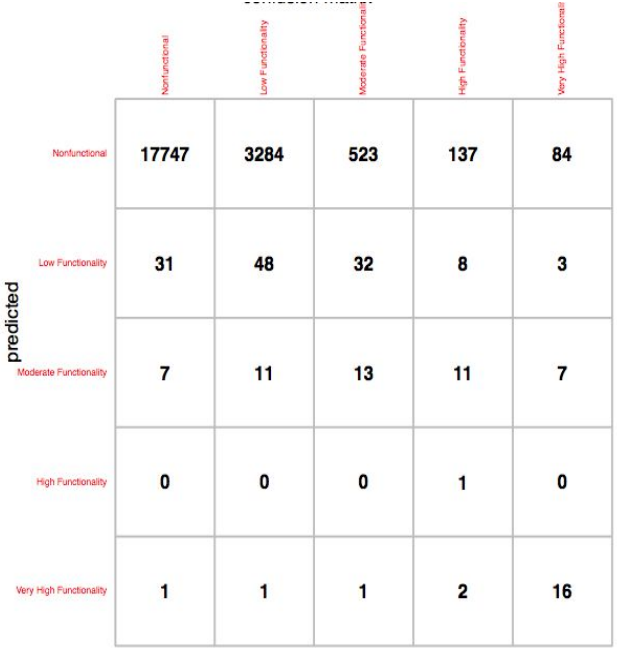
Confusion matrix and precision/recall for kNN for 5 classes

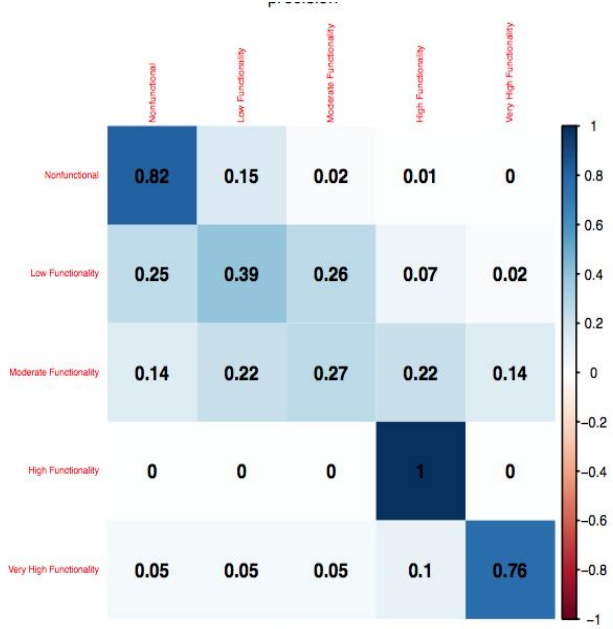


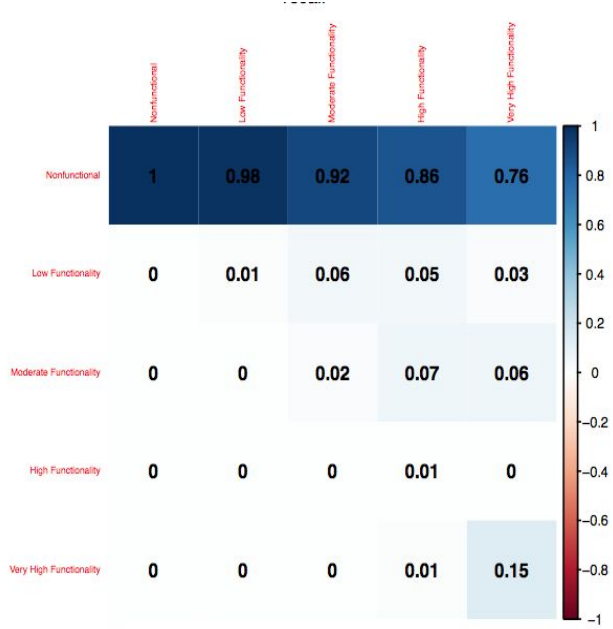




Confusion matrix and precision/recall for SVM for 5 classes







Confusion matrix and precision/recall for RF for 5 classes