



# Title: Predicting cell type specific functional consequences of non-coding variation using deep learning

CS230-WINTER 2019

Santosh Kumar  
Sangjukta Kashyap

Stanford University

# Contents

- Team
- Predicting
- Data
- Features
- Models
- Results
- Discussion
- Future
- References

Stanford University

## Team

Santosh Kumar

*Department of Computer Science*

*Stanford University*

[ksantosh@stanford.edu](mailto:ksantosh@stanford.edu)

<https://www.linkedin.com/in/kumsantosh/>

Sangjukta Kashyap

*Department of Computer Science*

*Stanford University*

[sanju321@stanford.edu](mailto:sanju321@stanford.edu)

<https://www.linkedin.com/in/sangjukta-kashyap-99901b5/>

## Mentor

Hoormazd Rezaei

Stanford University

<https://www.linkedin.com/in/hoormazd-rezaei/>

Stanford University

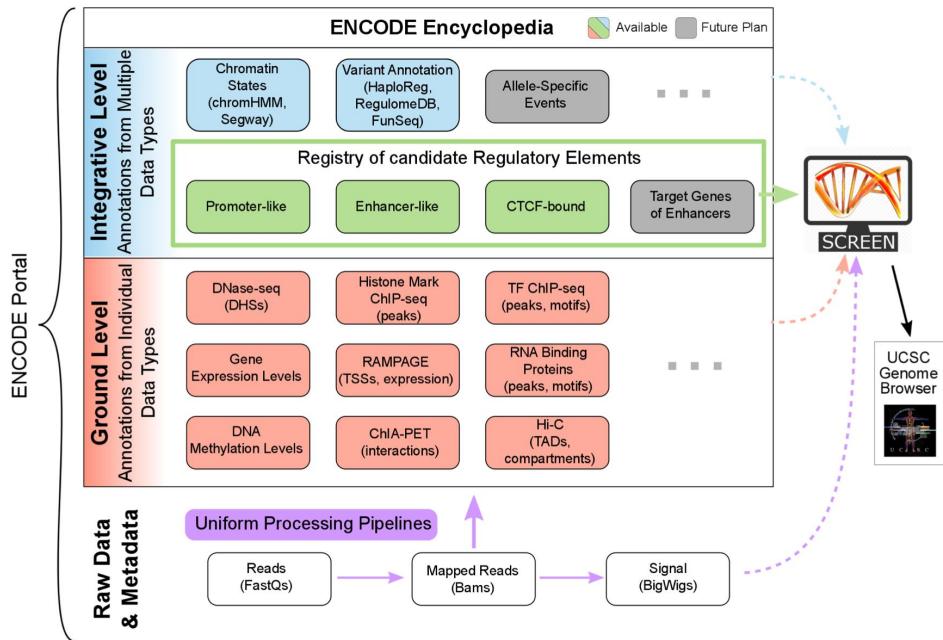
## Predicting

- Predicting the functional consequences of genetic variants in non-coding regions is a challenging problem.
- We used here a deep learning approach, to jointly utilize experimentally confirmed regulatory variants (labeled variants), unlabeled variants genome-wide, and more than a thousand cell/tissue type specific epigenetic annotations to predict functional consequences of non-coding variants.
- Through the application to several experimental datasets, we demonstrate that the proposed method gets good prediction accuracy,

Stanford University

# Data

The dataset we got from Dr. He from Stanford University is ENCODE. The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI).



chr	pos	rs	Label	DNase-E001	H3K27ac-E01	H3K27me3-I	H3K36me3-I	H3K4me1-E	H3K4me3-E	H3K9ac-E12	H3K9me3-E	DNase-E129	H3K27ac-E1	H3K27me3-I	H3K36me3-I	H3K4me1-E	H3K4me3-E	H3K9ac-E12	H3K9me3-E12	
0	chr10	101980135	chr10:10198t	0	0.47	0.4	0.28	4.36	0.42	0.25 ...	0.43	0.34	0.5	0.19	0.13	1.84	0.33	0.11	0.26	0.15
1	chr10	102010516	chr10:10201t	0	0.45	0.43	0.34	3.75	0.42	0.3 ...	0.32	0.37	0.45	0.36	0.33	4.08	0.4	0.31	0.4	0.38
2	chr10	102012645	chr10:10201t	0	0.36	0.29	0.28	1.12	0.29	0.28 ...	0.33	0.28	0.48	0.31	0.23	1.38	0.26	0.3	0.36	0.3
3	chr10	102031373	chr10:10203t	0	0.44	0.33	0.35	0.69	0.56	0.35 ...	0.36	0.58	0.48	0.2	0.15	1.31	0.28	0.2	0.24	1.08
4	chr10	102265445	chr10:10226t	0	0.39	0.07	0.69	0.38	0.12	0.05 ...	0.13	0.1	0.58	0.59	0.45	1.84	0.38	0.37	0.49	0.33

Stanford University

# Features

$127 * 8 = 1016$  Features:

Features example:

DNase-E001 where,

DNase is feature and E001 is Cell type/Tissue type

Labels

0 or 1

No. of training data with Label 0 and 1

- 0 - 22384
- 1 - 693

No. of test data with Label 0 and 1

- 0 - 1451
- 1 - 525

Stanford University

# Models

## Classifier:

1. LogisticRegression
2. DecisionTreeClassifier
3. KNeighborsClassifier
4. BernoulliNB
5. LinearDiscriminantAnalysis
6. GaussianNB
7. RidgeClassifier
8. SGDClassifier
9. Support Vector Machines

## Neural Network

1. Non-regularized NN model
2. NN model with L2 Regularization
3. NN Model with Dropout

Stanford University

# Results

Classifiers	Mean ROC (AUC Value)	Accuracy
LogisticRegression	0.72	95%
DecisionTreeClassifier	0.57	94%
KNeighborsClassifier	0.57	95%
BernoulliNB	0.5	94%
LinearDiscriminantAnalysis	0.69	96%
GaussianNB	0.68	95%
RidgeClassifier	0.55	97%
SGDClassifier	0.72	95%
Support Vector Machines	0.65	95%

NN Model	Accuracy
Non-regularized	96.8%
L2 Regularization	96.3%
Dropout	97.1%

Stanford University

## Discussion

- Dataset is with very less Label 1 and most of them are label 0.
- We tried with different classifiers, we got very good accuracy but average ROC(AUC) Curve
  - LogisticRegression Classifier gives best result
- We also tried with Neural Network
  - Dropout gives best result
- We would like thanks our mentor for this project Mr. Hoormazd Rezaei who guided us during each step
- We would also like to thanks Dr. He who has given us the dataset and project requirements.

Stanford University

## Future

We plan to continue exploring the project with below strategy

1. Work with larger dataset
2. Try convolution network
3. Also try other neural network along with convolution network

Stanford University

# References

1. Ritchie, G. R., Dunham, I., Zeggini, E., Flückeck, P. Functional annotation of noncoding sequence variants. *Nat. Methods* 11, 294–296 (2014).
2. Altshuler, D., Daly, M. J., Lander, E. S. Genetic mapping in human disease. *Science* 322, 881–888 (2008).
3. Kellis, M. et al. Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. USA* 111, 6131–6138 (2014).
4. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* 165, 1519–1529 (2016).
5. Kheradpour, P. et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* 23, 800–811 (2013).
6. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823 (2013).
7. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* 46, 944–950 (2014).
8. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., Goldstein, D. B. Genetic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* 9, e1003709 (2013).
9. Petrovski, S. et al. The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet.* 11, e1005492 (2015).
10. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).
11. Bernstein, B. E. et al. The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* 28, 1045–1048 (2010).
12. Martens, J. H., Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* 98, 1487–1489 (2013).
13. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315 (2014).
14. Ionita-Laza, I., McCallum, K., Xu, B., Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* 48, 214–220 (2016).
15. Quang, D., Chen, Y., Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31, 761–763 (2014).
16. Fu, Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* 15, 480 (2014).
17. Huang, Y. F., Gulko, B., Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* 49, 618–624 (2017).
18. Backenroth, D. et al. FUN-LDA: a latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation. *Am. J. Hum. Genet.* 102, 920–942 (2017).
19. Lu, Q., Powles, R. L., Wang, Q., He, B. J., Zhao, H. Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS Genet.* 12, e1005947 (2016).
20. Zhou, J., Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12, 931–934 (2015).
21. Zou, H., Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* 67, 301–320 (2005).
22. Friedman, J., Hastie, T., Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1 (2010).
23. Prentice, R. L., Pyke, R. Logistic disease incidence models and case-control studies. *Biometrika* 66, 403–411 (1979).
24. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394 (2012).
25. Li, M. J. et al. Predicting regulatory variants with composite statistic. *Bioinformatics* 32, 2729–2736 (2016).
26. Brown, C. D., Mangravite, L. M., Engelhardt, B. E. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* 9, e1003649 (2013).
27. Farh, K. K. H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343 (2015).
28. Maurano, M. T. et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* 47, 1393–1401 (2015).
29. Brown, A. A. et al. Predicting causal variants affecting expression using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.* 49, 1747–1751 (2017).
30. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660 (2015).
31. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427 (2014).
32. Forrest, M. P. et al. Open chromatin profiling in hESC-derived neurons prioritizes functional noncoding psychiatric risk variants and highlights neurodevelopmental loci. *Cell Stem Cell* 21, 305–318 (2017).
33. Duan, J. et al. A rare functional noncoding variant at the GWAS-implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder. *Am. J. Hum. Genet.* 95, 744–753 (2014).
34. Lee, S., Abecasis, G. R., Boehnke, M., Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95, 5–23 (2014).
35. He, Z., Xu, B., Lee, S., Ionita-Laza, I. Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *Am. J. Hum. Genet.* 101, 340–352 (2017).
36. Voight, B. F. et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8, e1002793 (2012).
37. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283 (2013).
38. Liu, D. J. et al. Exome-wide association study of plasma lipids in > 300,000 individuals. *Nat. Genet.* 49, 1758–1766 (2017).
39. Lu, X. et al. Exome chip meta-analysis identifies novel loci and East Asian-specific coding variants that contribute to lipid levels and coronary artery disease. *Nat. Genet.* 49, 1722–1730 (2017).
40. Roussos, P. et al. A role for noncoding variation in schizophrenia. *Cell Rep.* 9, 1417–1429 (2014).
41. Belkin, M., Niyogi, P., Sindhwan, V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* 7, 2399–2434 (2006).
42. Jiang, Y., He, Y., Zhang, H. Variable selection with prior information for generalized linear models via the prior Lasso method. *J. Am. Stat. Assoc.* 111, 355–376 (2016).

Stanford University

Link to youtube video

[https://youtu.be/OCi2CX\\_GVCK](https://youtu.be/OCi2CX_GVCK)

Stanford University