DIVAN: disease-specific non-coding risk variants detection using multi-omics data

STEVE QIN

EMORY UNIVERSITY

EMORY
UNIVERSITY

GWAS Catalog

The NHGRI-EBI Catalog of published gen

Search the catalog

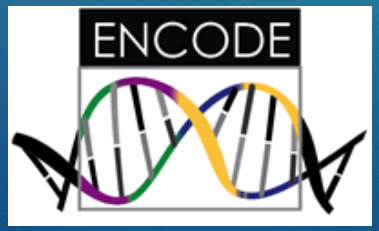
Examples: breast cancer, rs7329174, Yang, 2q37.1, HBS1L

Why these SNPs?

- Understand the biology behind these trait-associated non-coding risk variants
- Identify non-coding risk variants
- Using genomics/epigenomics data
- Methods have been developed:
  - GWAVA, CADD, Eigen, GenoCanyon …

DIVAN: **DI**sease-specific **V**ariant **AN**notation

- A unique model for each disease/ phenotype
- Using trait-associated SNPs identified by GWAS as training data
- Using 1,809 genomic and epigenomic data from ENCODE and REMC as features
  - ChIP-seq (TF, histone), DNase-seq, FAIRE, phastCons, GERP scores



Challenges and our work-arounds

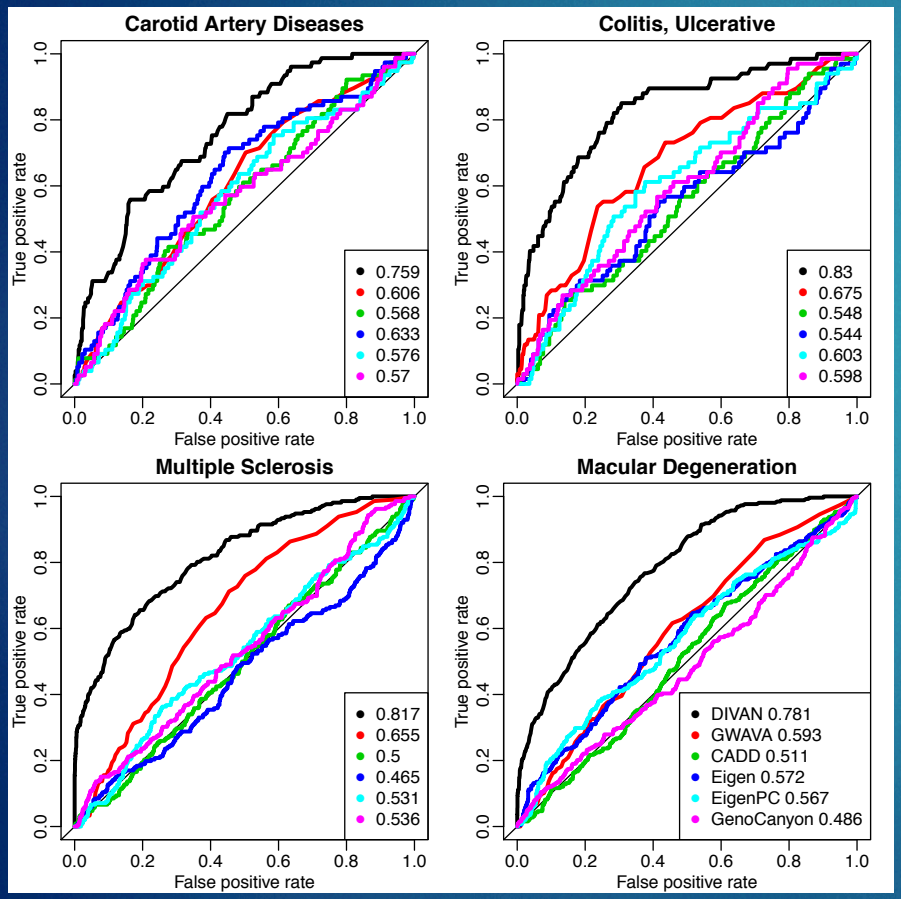
- Challenges
  - Huge collection of multi-omics data, many are correlated
  - Limited size of training data, $p >> n$
  - Diverse data types
- Our workaround
  - Feature selection
  - Ensemble learning
  - Using continuous read count instead of binary indicator of peak overlap as individual features

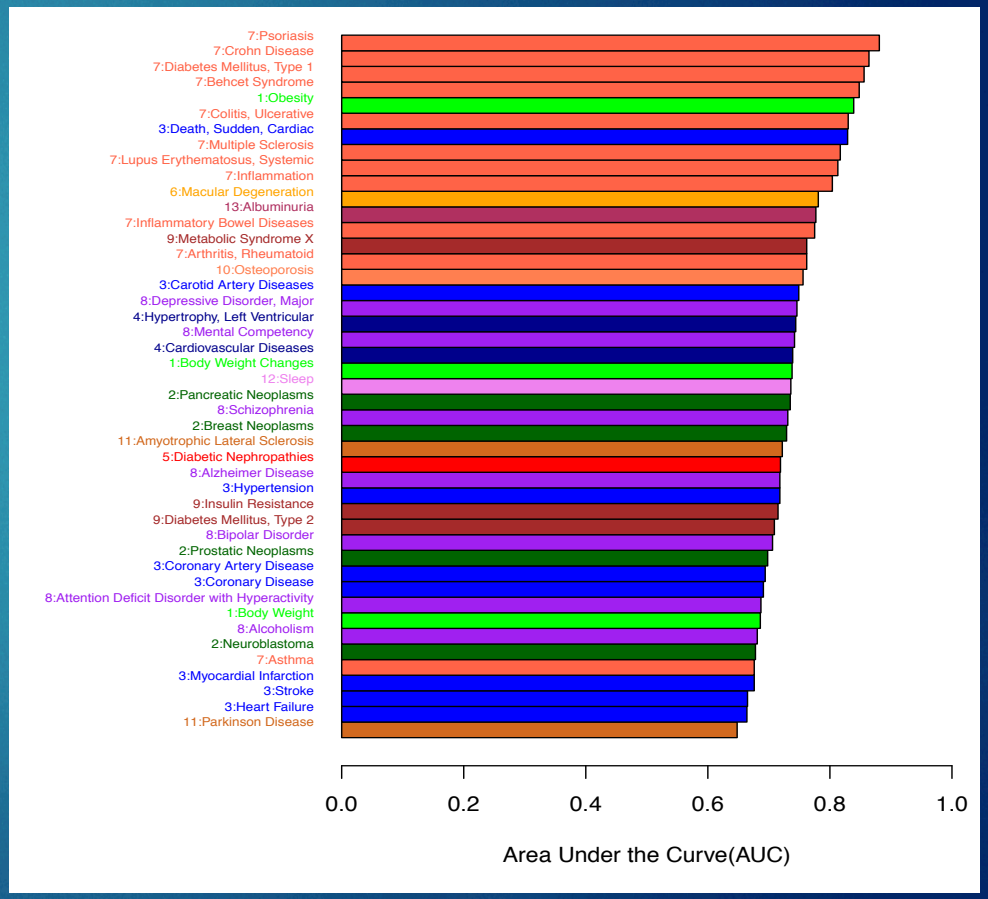
Some results

- Compared to
  - GWAVA, CADD, Eigen, GenoCanyon
- Cross-validation on 45 disease/ phenotype from association result browser
- Independent test on 36 disease/ phenotype from GRASP.

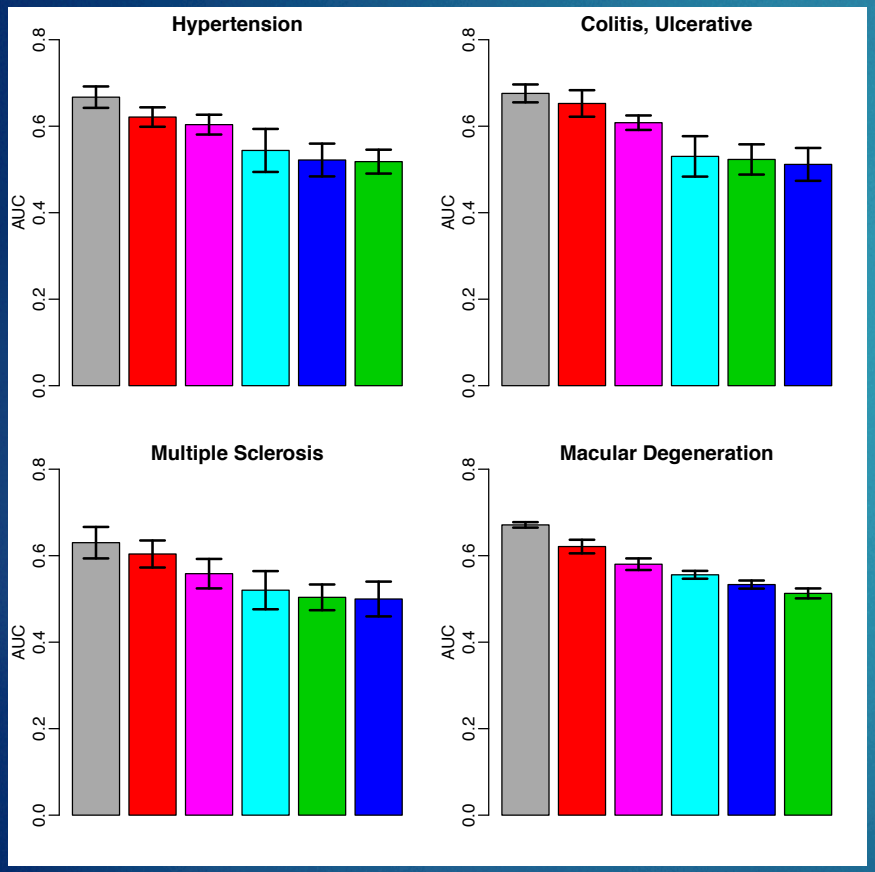
disease	class	#SNP(Association Results Browser)	#SNP(GRASP)
Body Weight	body weight	857	-
Body Weight Changes	body weight	81	26
Obesity	body weight	53	644
Breast Neoplasms	cancer	155	390
Neuroblastoma	cancer	268	53
Pancreatic Neoplasms	cancer	190	66
Prostatic Neoplasms	cancer	215	512
Carotid Artery Diseases	cardiovascular	80	-
Coro-ry Artery Disease	cardiovascular	584	3716
Coro-ry Disease	cardiovascular	176	-
Death, Sudden, Cardiac	cardiovascular	46	-
Heart Failure	cardiovascular	530	22
Hypertension	cardiovascular	201	538
Myocardial Infarction	cardiovascular	584	415
Stroke	cardiovascular	725	21
Cardiovascular Diseases	cariovascular	63	33
Hypertrophy, Left Ventricular	cariovascular	143	13
Diabetic Nephropathies	endocrine	159	-
Macular Degeneration	eye disease	258	5141
Arthritis, Rheumatoid	immune	100	7502
Asthma	immune	252	683
Behcet Syndrome	immune	229	226
Colitis, Ulcerative	immune	67	397
Crohn Disease	immune	59	1172
Diabetes Mellitus, Type 1	immune	147	797
Inflammation	immune	70	76
Inflammatory Bowel Diseases	immune	91	106
Lupus Erythematosus, Systemic	immune	184	215
Multiple Sclerosis	immune	212	473
Psoriasis	immune	106	377
Alcoholism	mental	261	128
Alzheimer Disease	mental	202	961
Attention Deficit Disorder with Hyperactivity	mental	197	295
Bipolar Disorder	mental	268	2028
Depressive Disorder, Major	mental	85	969
Mental Competency	mental	99	-
Schizophrenia	mental	233	1381
Diabetes Mellitus, Type 2	metabolic disease	181	2913
Insulin Resistance	metabolic disease	170	95
Metabolic Syndrome X	metabolic disease	40	25
Osteoporosis	musculoskeletal	67	-
Amyotrophic Lateral Sclerosis	nervous system	197	372

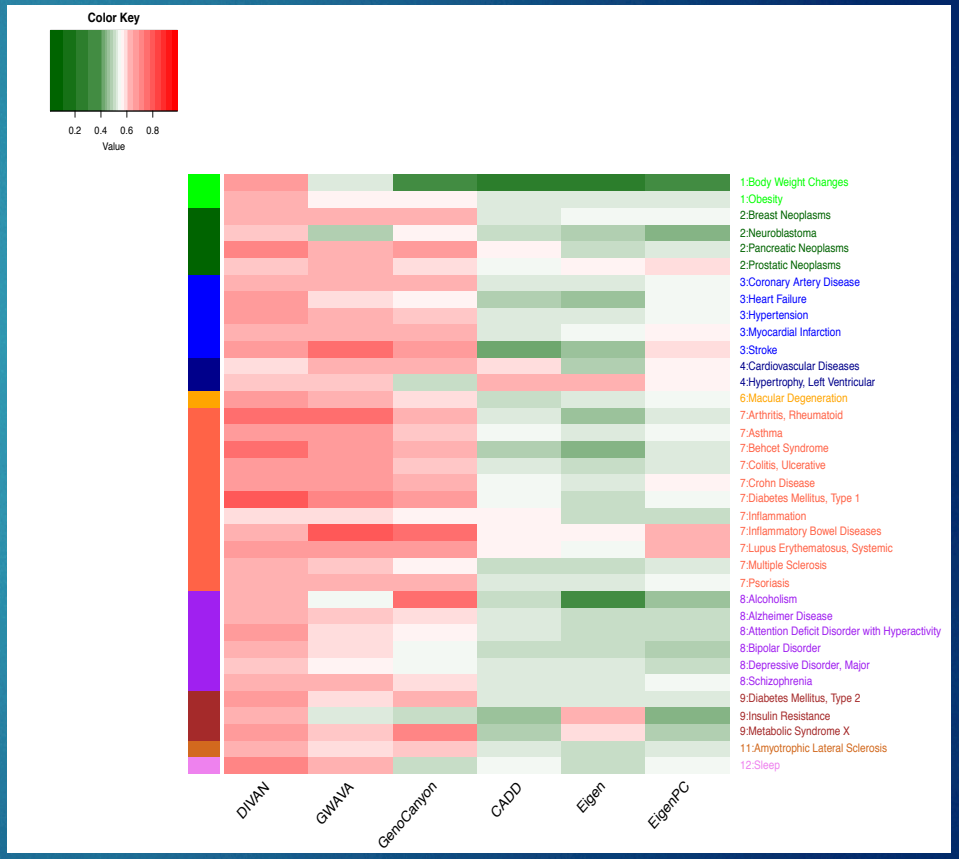
Cross-validation results



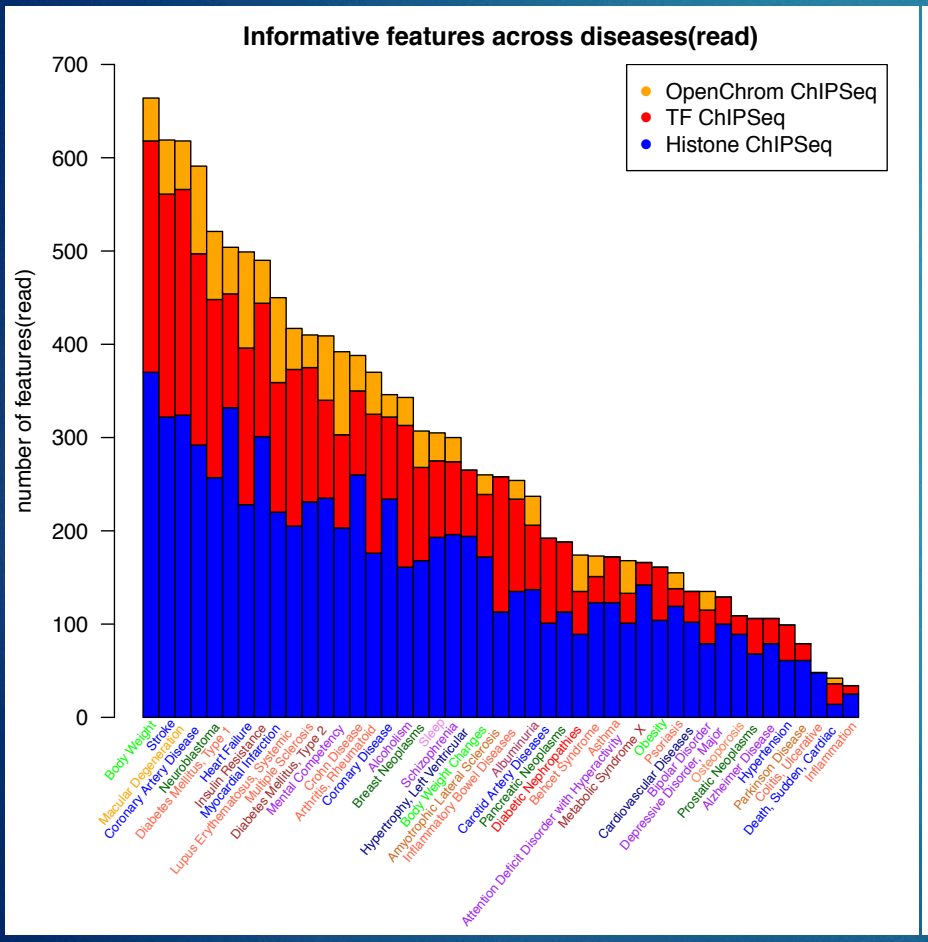


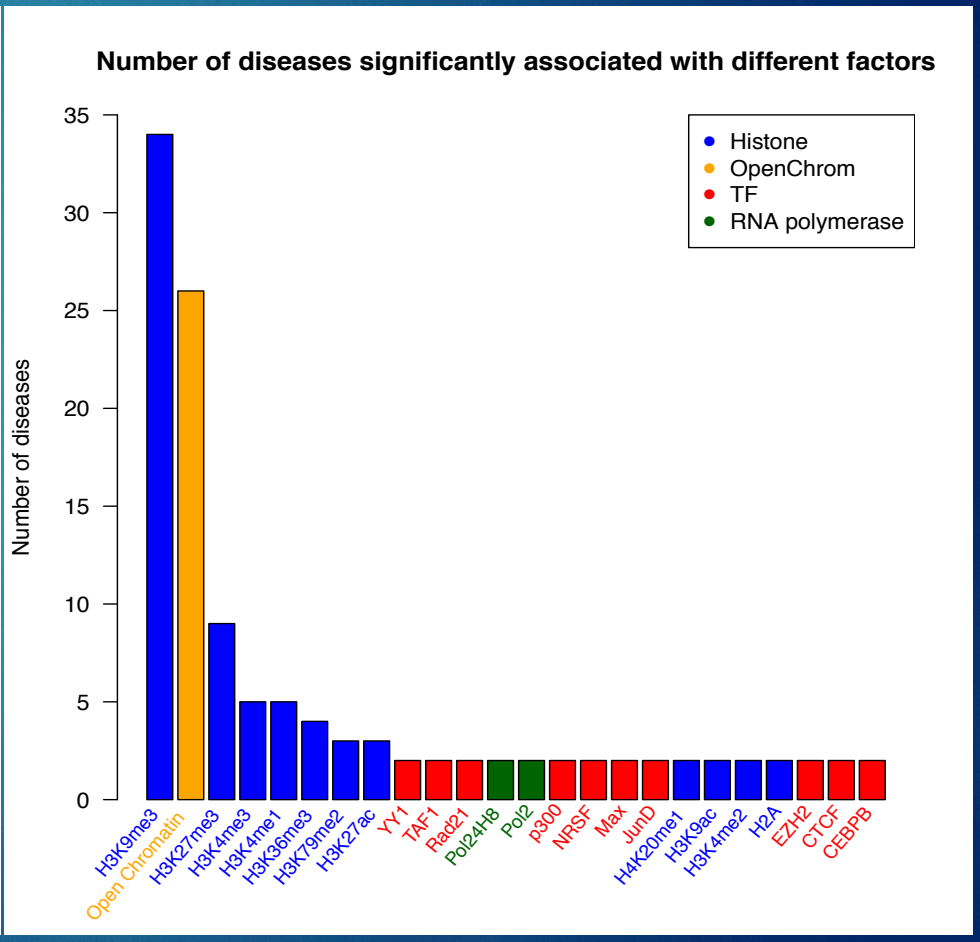
Independent test results





Informative features





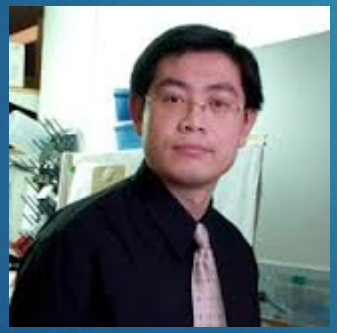
Summary

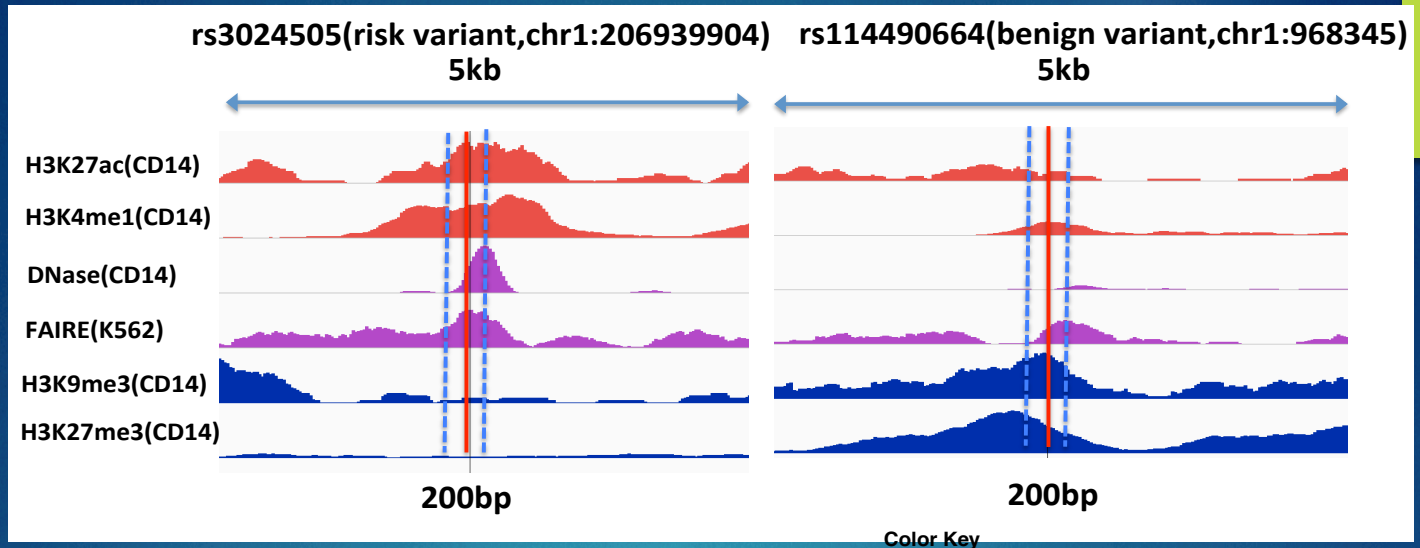
- Disease-specific risk variant identification is feasible
- Need advanced statistical learning techniques
- DIVAN outperforms competitors*,

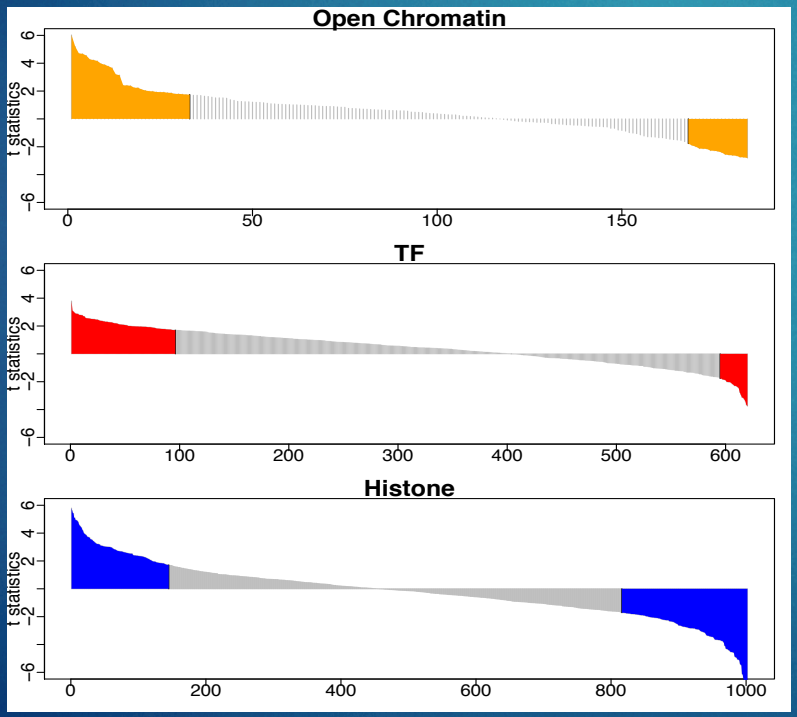
  *on our terms
- Advantages using reads instead of peaks
- Histone marks are the most informative class of features.

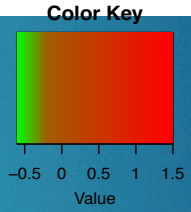
Acknowledgement

- Graduate student
  - Li Chen
- Collaborator
  - Peng Jin




rs3024505(risk variant,chr1:206939904)
5kb
rs114490664(benign variant,chr1:968345)
5kb
H3K27ac(CD14)
H3K4me1(CD14)
DNase(CD14)
FAIRE(K562)
H3K9me3(CD14)
H3K27me3(CD14)
200bp
200bp


Open Chromatin
TF
Histone
t statistics


Color Key
Value