Thoughts on Genome Annotation, Prioritizing Variants & the Application of these Concepts in a Disease Context

Mark Gerstein

Yale

Slides freely downloadable from **Lectures.GersteinLab.org** & “tweetable” (via **@MarkGerstein**).
No Conflicts for this Talk. See last slide for more info.

Overall Problem: Finding Key Variants in Personal Genomes

Millions of variants in a personal genome
Thousands, in a cancer genome
Different **contexts** for prioritization

In **rare disease**, only a few
high-impact variants are associated with disease

CAN YOU FIND THE PANDA?

In **cancer**, a few positively selected drivers amongst many passengers

In **common disease**, more variants associated & each has weaker effect,
But one wants to find key ”functional” variant amongst many in LD

Overall Problem: Finding Key Variants in Personal Genomes

CAN YOU FIND THE PANDA?

Millions of variants in a personal genome
Thousands, in a cancer genome
Different **contexts** for prioritization

In **rare disease**, only a few
high-impact variants are associated with disease

In **cancer**, a few positively selected drivers amongst many passengers

In **common disease**, more variants associated & each has weaker effect,
But one wants to find key ”functional” variant amongst many in LD

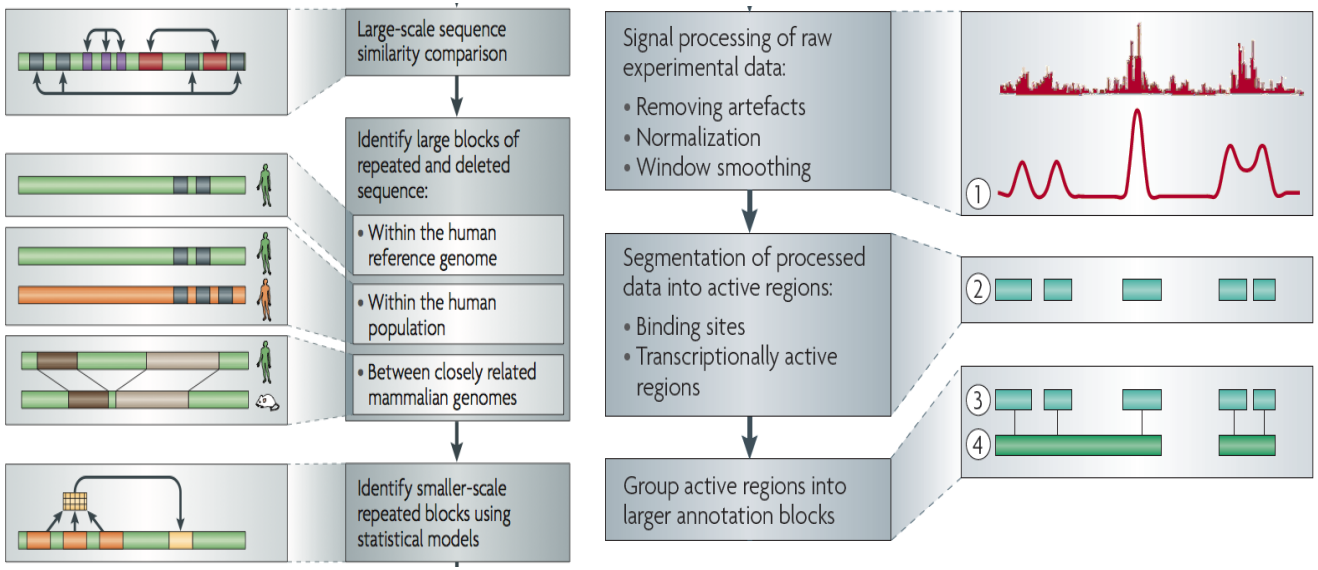
Thus: Need to find & prioritize high impact variants.
Particularly hard for non-coding regions.

Non-coding Annotations: Overview

Features are often present on multiple ”scale” (eg elements and connected networks)

Sequence features, incl. **Conservation**

Functional Genomics
Chip-seq (Epigenome & seq. specific TF)
and ncRNA & un-annotated transcription



What is Annotation? (For Written Texts?)

No. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre

NATURE VOL 409 | 15 FEBRUARY 2001

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

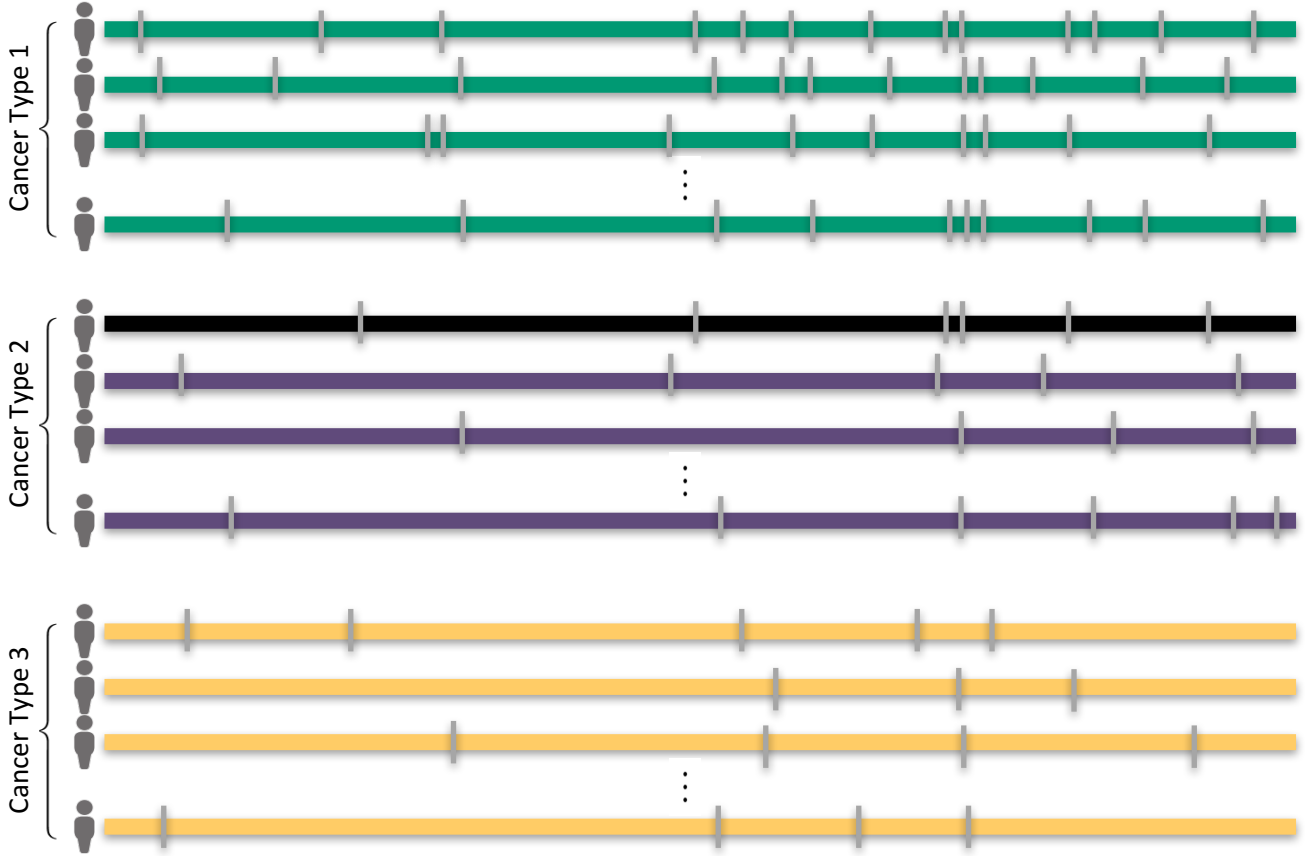
The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

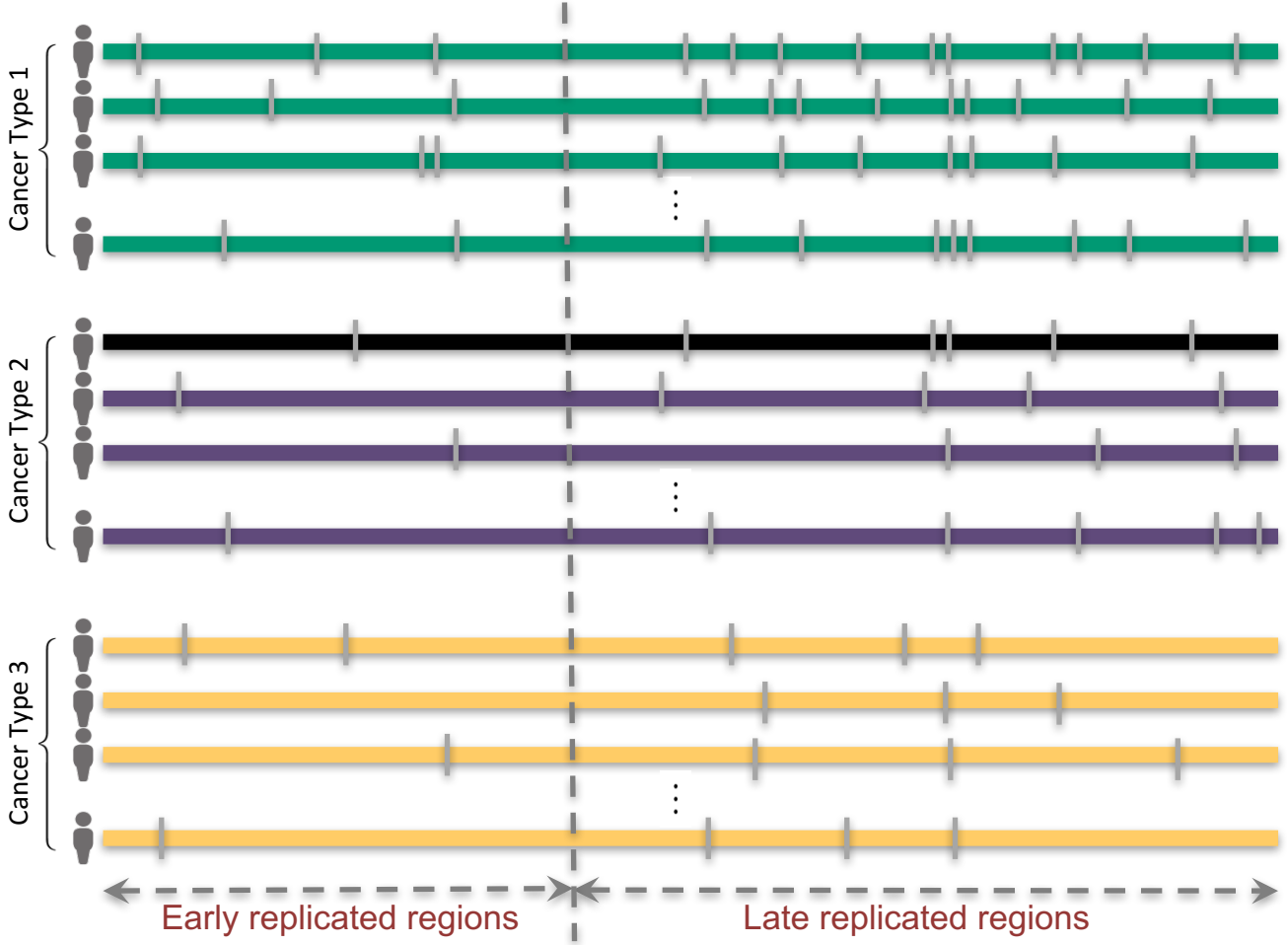
coordinate regulation of the genes in the clusters.

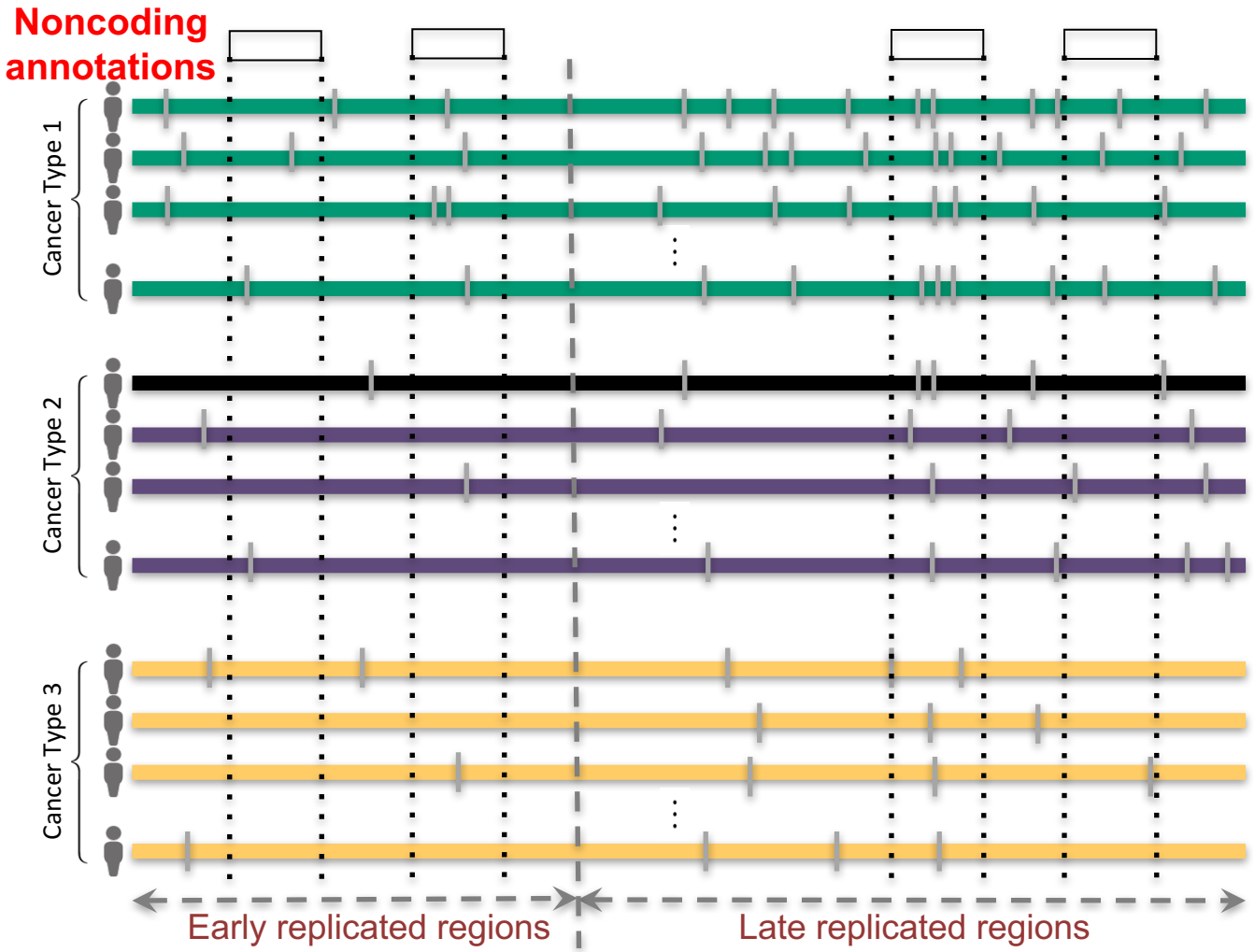
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.
- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.
- The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.
- Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark

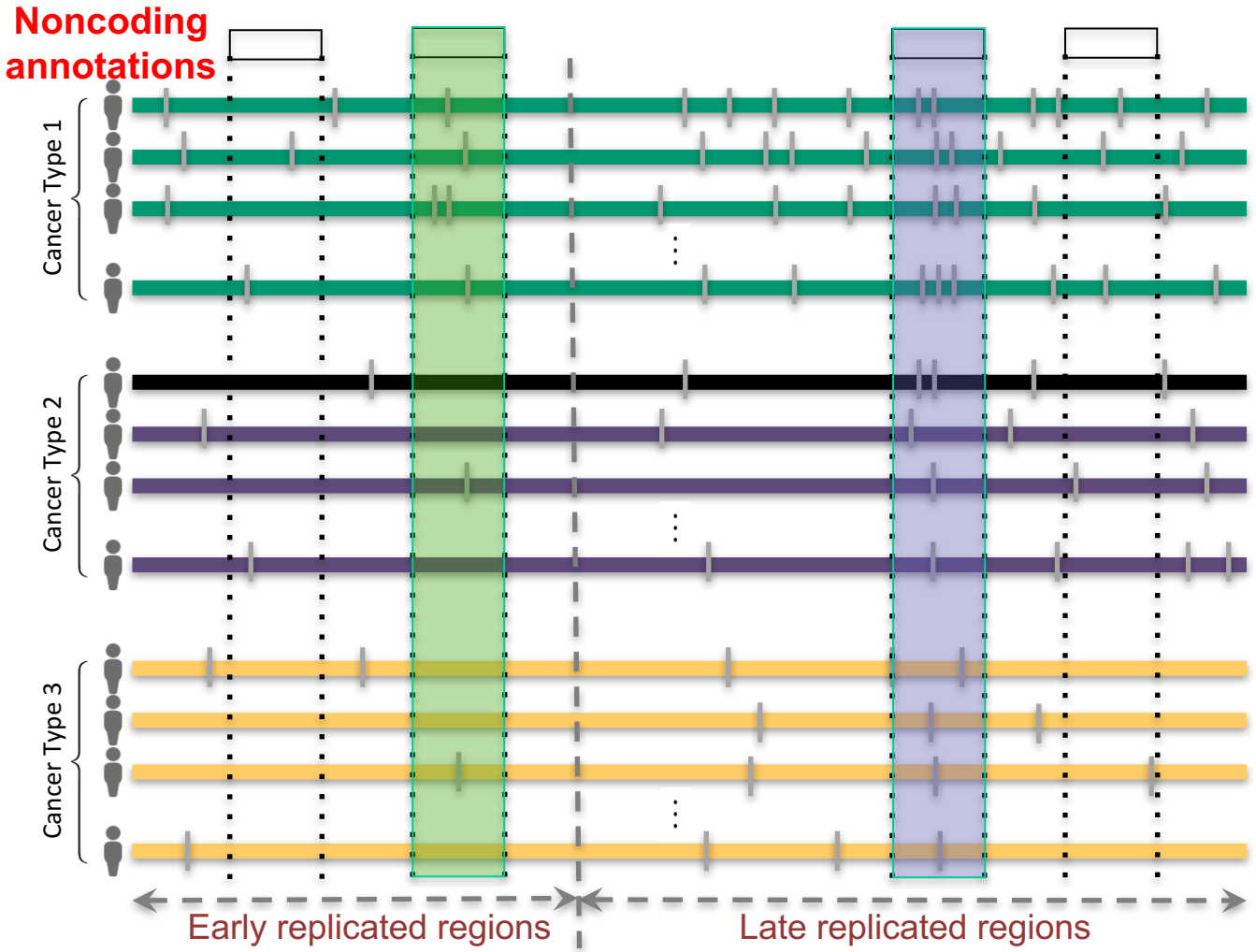
Mutation recurrence



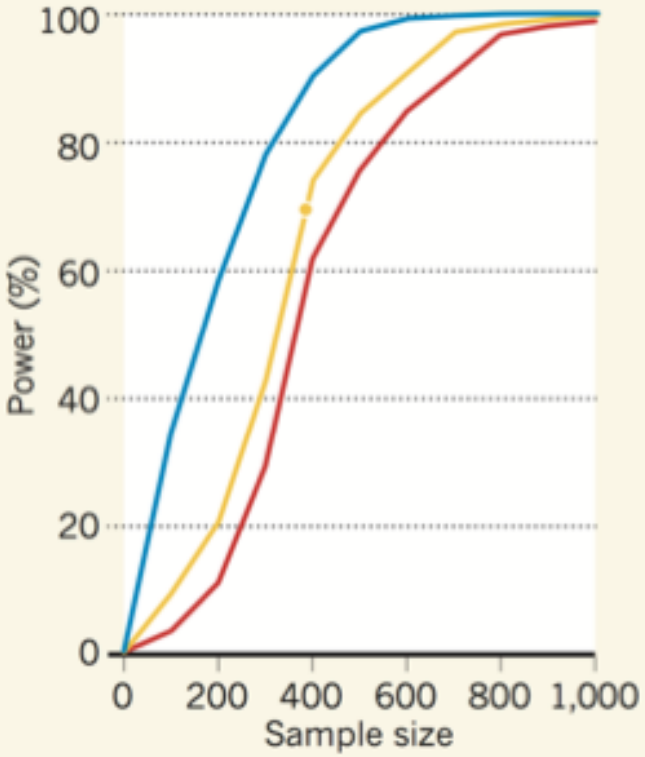
Mutation recurrence

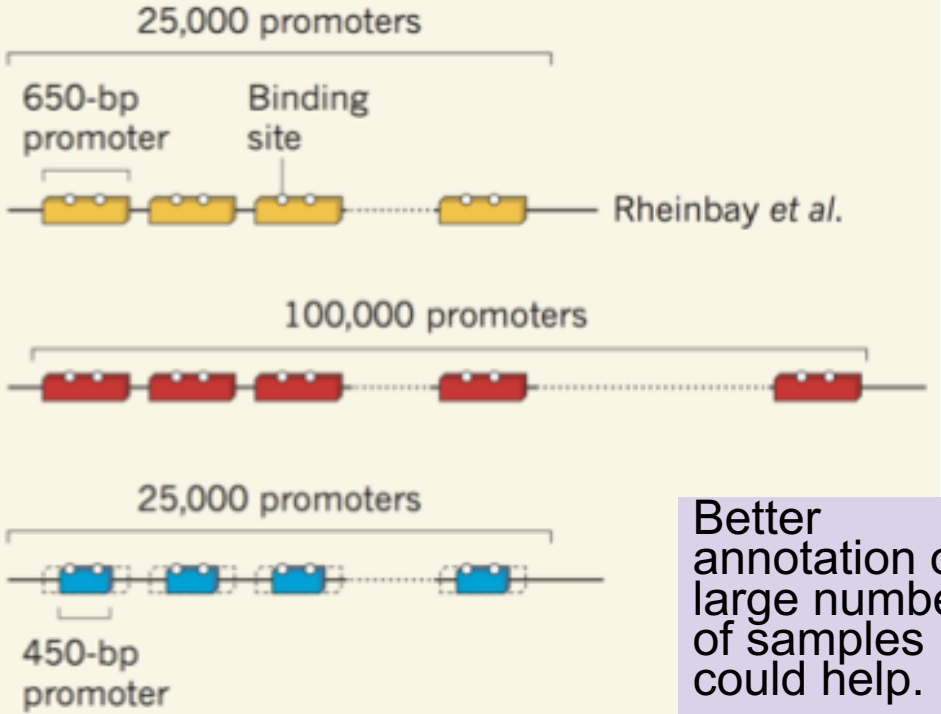



Noncoding annotations
Cancer Type 1
Cancer Type 2
Cancer Type 3
Early replicated regions
Late replicated regions


Noncoding annotations
Cancer Type 1
Cancer Type 2
Cancer Type 3
Early replicated regions
Late replicated regions

Power, as an issue in driver discovery





Better annotation or large number of samples could help.

[Kumar & Gerstein, *Nature* ('17)]

Background on computationally annotation

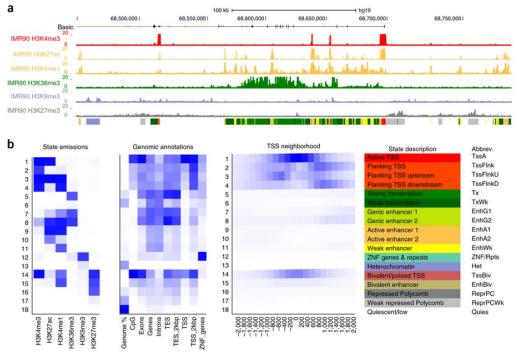
- **Peak calling**:
  - ✓PeakSeq, SPP, MACS2, Hotspot …
  - ✓ENCODE Encyclopedia

- **Genome segmentation**: partition the genome into regions (states) with distinct epigenomic profiles, then assign each state a functional label.
  - ✓ChromHMM: Multivariate Hidden Markov Model
  - ✓Segway: Dynamic Bayesian Network Model

- Supervised regulatory prediction: learn predictive models from labeled dataset of regulatory elements.
  - ✓ CSI-ANN: Time-Delay Neural Network
  - ✓ RFECS: Random Forest
  - ✓ DEEP: Ensemble SVM + Artificial Neural Network
  - ✓ REPTILE: Random Forest
  - ✓ gkm-SVM: Gapped k-mer

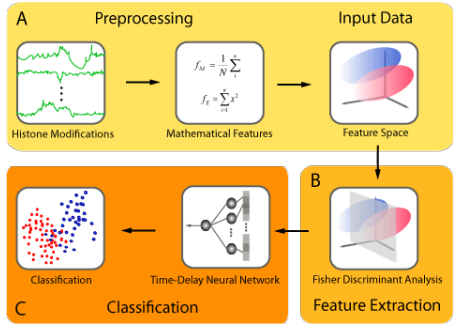
- **Target finding**
  - ✓ Ripple, TargetFinder, JEME, PreSTIGE, IM-PET

ChromHMM



J. Ernst, M. Kellis. *Nat. Protoc., 2017*

CSI-ANN



H.A. Firpi, D. Ucar, K. Tian. Bioinformatics, 2010

High-throughput approaches to dissect enhancer function

	Construct	Tested in	Detection	Advantages	Disadvantages
MPRA/ MPFD/ CRE-seq		Cell lines, Mouse liver, Mouse retina	Barcode RNA-seq	High BC multiplicity Quantitative	Episomal
STARR-seq		Cell lines	Enhancer RNA-seq	Quantitative	Low multiplicity Episomal
TRIP		Mouse ESCs	Barcode RNA-seq	Quantitative Genomic context	Low resolution
FACS-based		Yeast	FACS	Semi-quantitative	Low multiplicity Episomal
FACS-based (FIREWACh)		Mouse ESCs	FACS	Chromosomal	Not quantitative
FACS-based (SIF-seq)		Mouse ESCs	FACS	Chromosomal	Not quantitative

Fumitaka and Ahituv, *Genomics*, 2015

FIREWACh

Murtha, M, et al., *Nat Methods*, 2014

SIF-Seq

Dickel, DE, et al., *Nat Methods*, 2014

MPRA

Patwardhan RP, et al., *Nat Biotech*, 2009
Melnikov A, et al., *Nat Biotech*, 2012
Kheradpour P, et al, *Genome Res*, 2013
Birnbaum RY, et al, *PLoS Genet*, 2014

CRE-Seq

Kwasnieski JC, et al, *PNAS*, 2012
White MA, et al, *PNAS*, 2013
Kwasnieski JC, et al, *Genome Res*, 2014

MPFD

Patwardhan RP, et al., *Nat Biotech*, 2012

STARR-Seq

Arnold, CD, et al., *Science*, 2013
Arnold, CD, et al., *Nat Genet*, 2014
Shlyueva, D, et al, *Mol Cell*, 2014

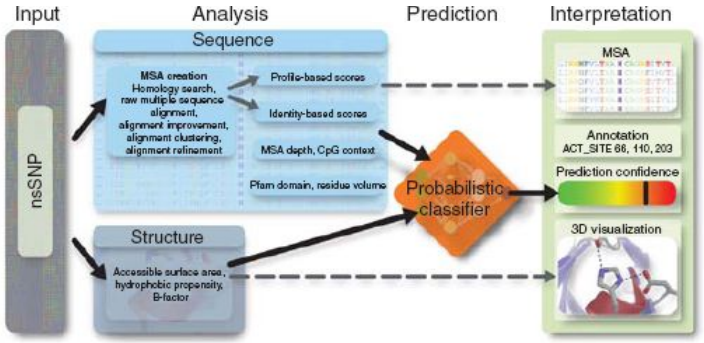
TRIP

Akhtar, W, et al., *Cell*, 2013

Genetic variant annotation: coding and noncoding

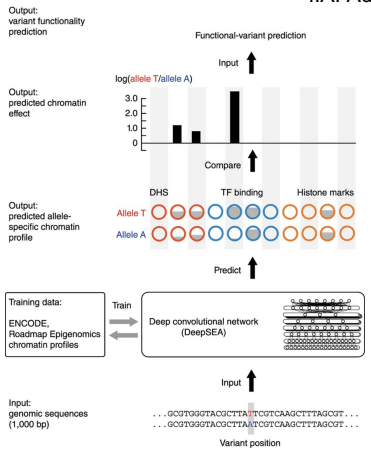
- Tools developed specifically for coding variants:
  - ✓PolyPhen-2
  - ✓SnpEff
  - ✓ SIFT
  - ✓...
- Tools developed specifically for noncoding variants:
  - ✓RegulomeDB
  - ✓HaploReg
  - ✓DeepSEA
  - ✓GWAVA
  - ✓...
- Tools for both coding and noncoding variants:
  - ✓CADD
  - ✓ANNOVAR
  - ✓VEP
  - ✓FATHMM-MKL
  - ✓ ….

Polyphen-2



I.A. Adzhubei, *et al. Nat. Methods, 2010*

DeepSEA



J. Zhou, O.G. Troyanskaya, Nat. Methods, 2015

Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

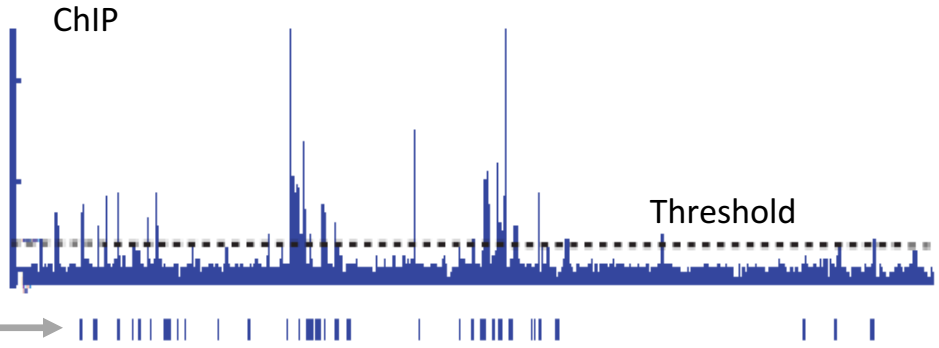
- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

Summarizing the Signal: "Traditional" ChipSeq Peak Calling

- Generate & threshold the signal profile to identify candidate target regions
  - Simulation (PeakSeq),
  - Local window based Poisson (MACS),
  - Fold change statistics (SPP)



Potential Targets

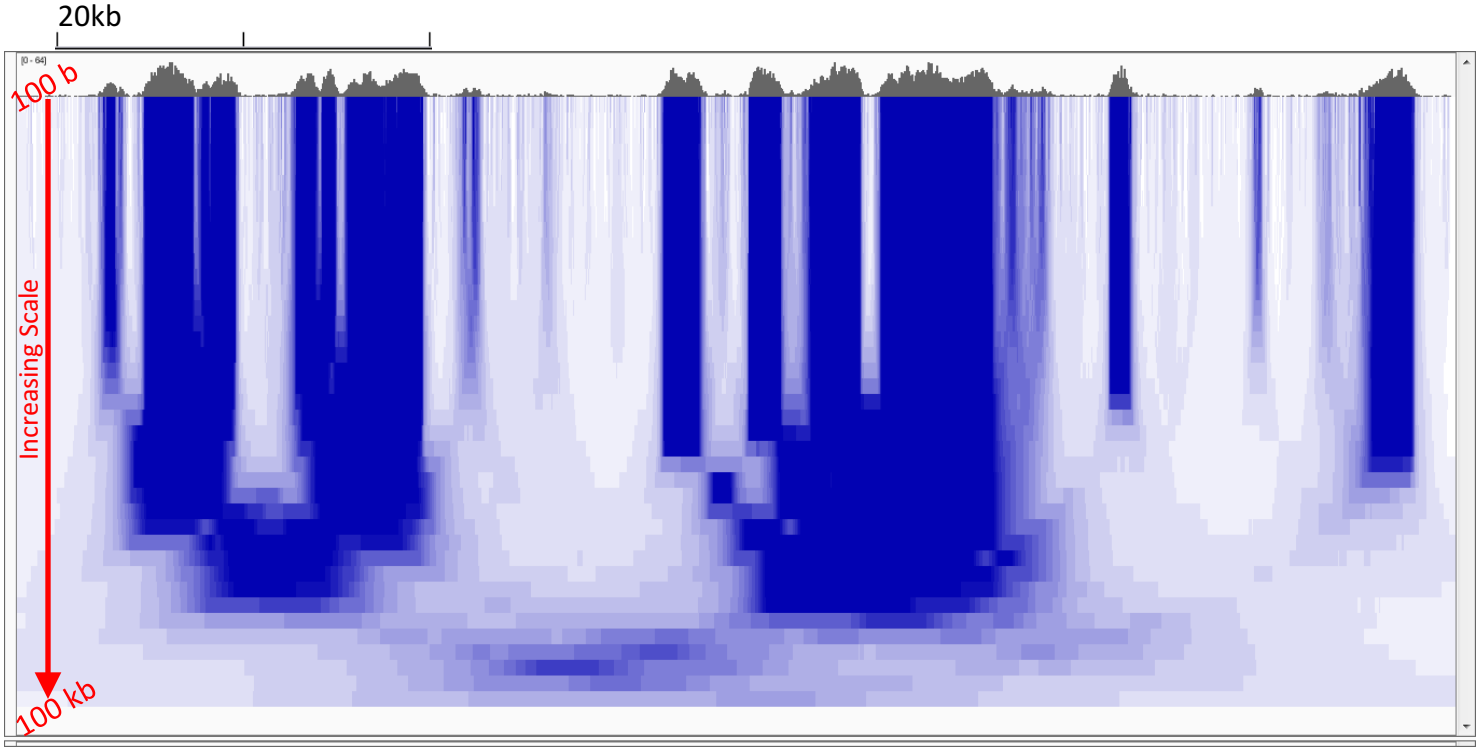
- Score against the control

Significantly Enriched targets

Now an update: "PeakSeq 2" => MUSIC

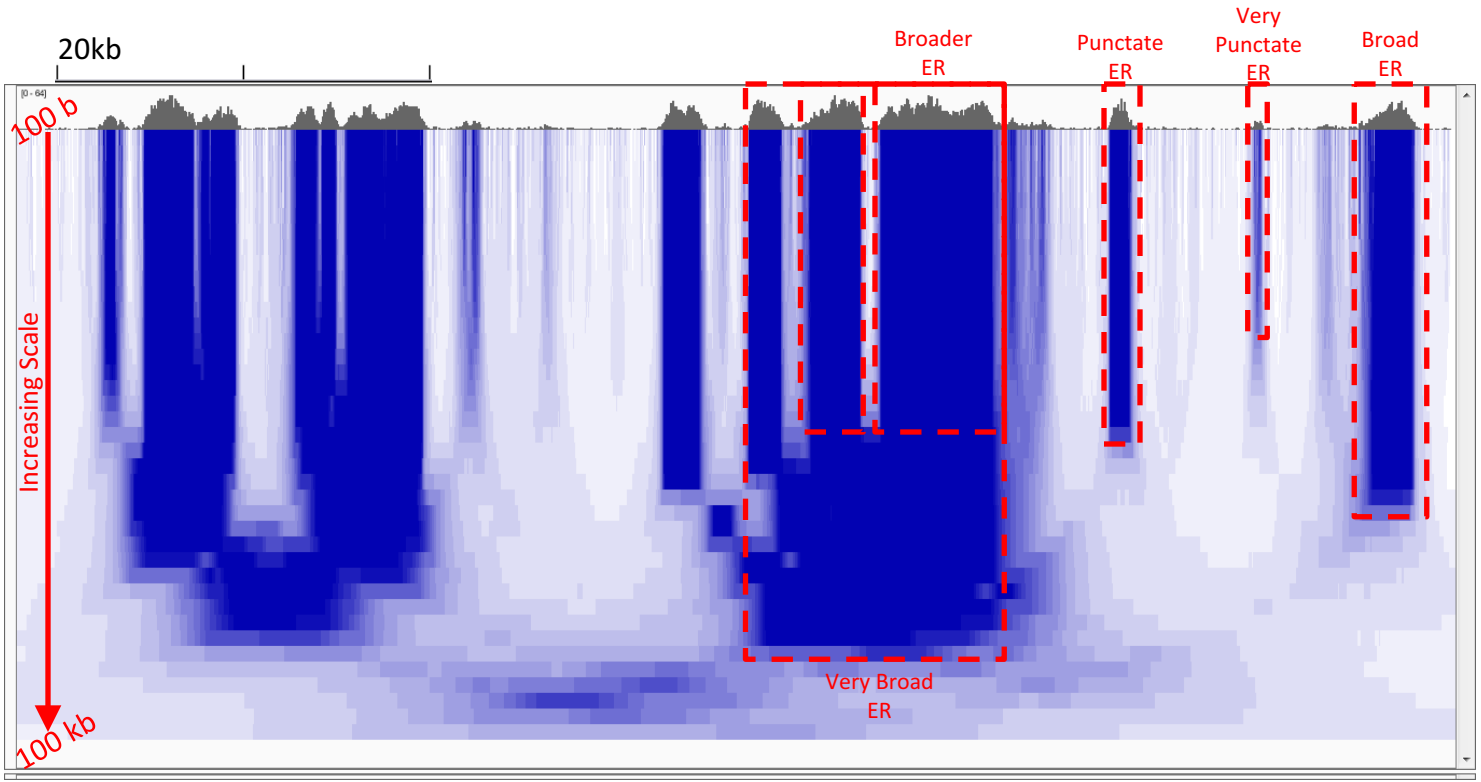
[Rozowsky et al. ('09) *Nat Biotech*]

Multiscale Decomposition



[Harmanci *et al, Genome Biol.* ('14)]

Multiscale Decomposition



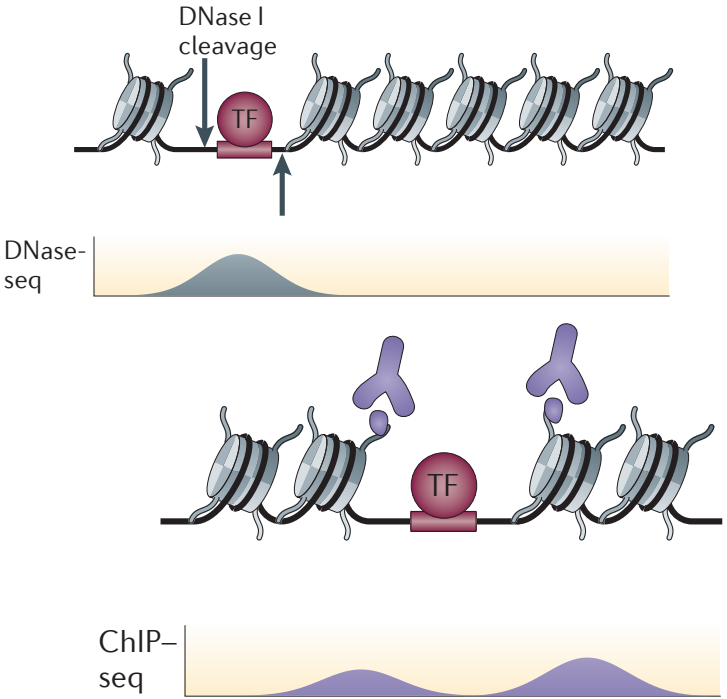
[Harmanci *et al, Genome Biol.* ('14)]

Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

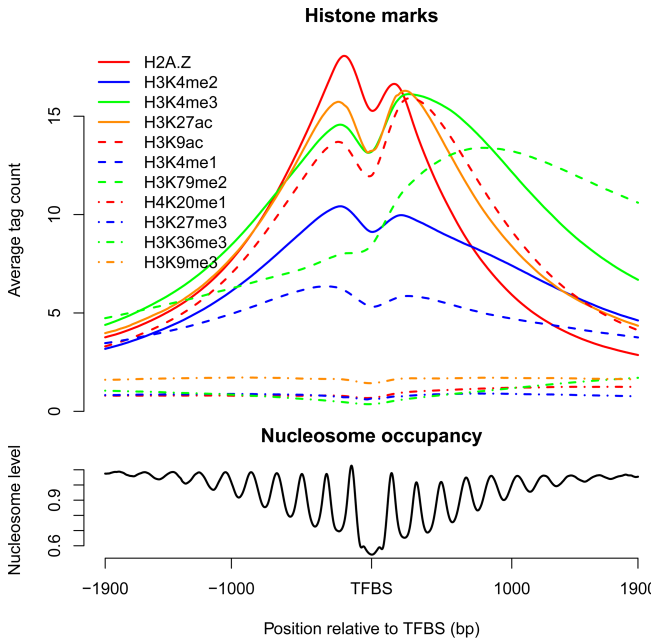
- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

Unique shape associated histone signals flanking active enhancers identified through STARR-seq

Arnold, et al., Science

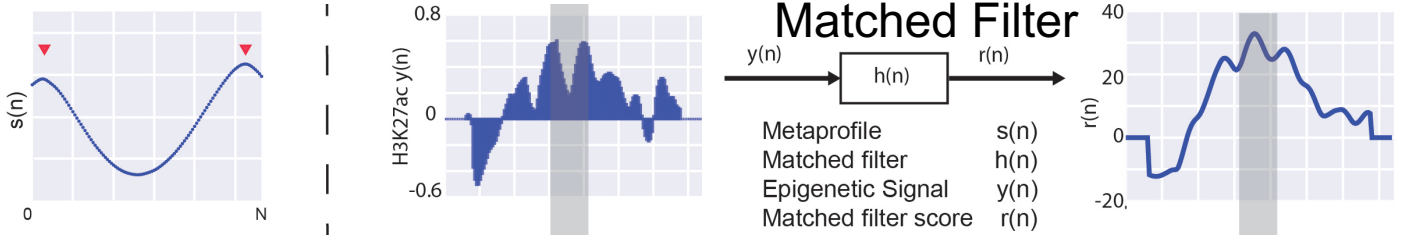


Shlyueva, et al., Nat Rev Genet

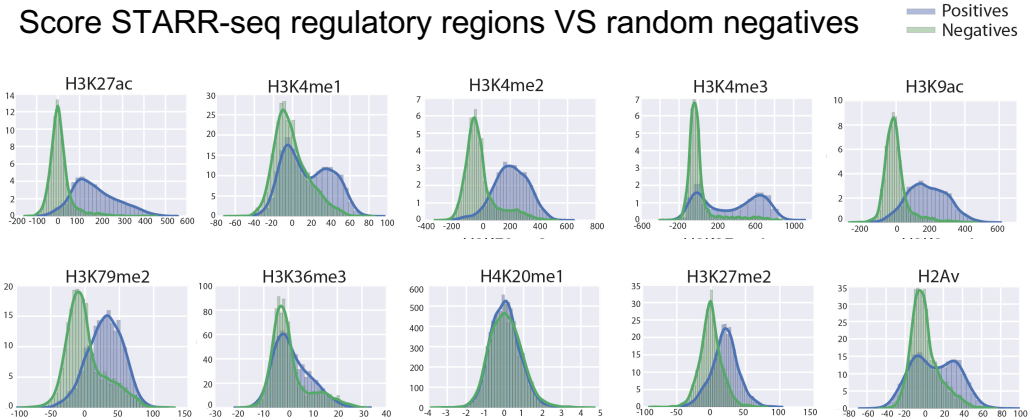


Nie, et al., PloS one

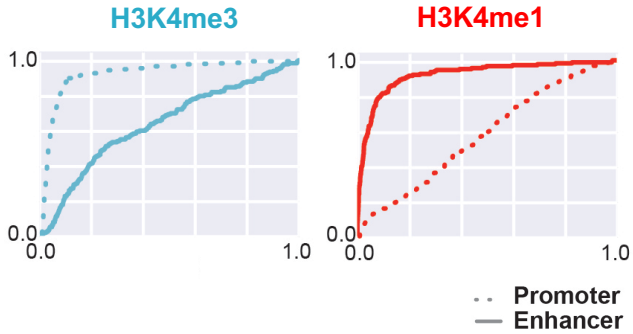
Matched Filter recognize shape patterns



Score STARR-seq regulatory regions VS random negatives



Evaluate using ROC curve

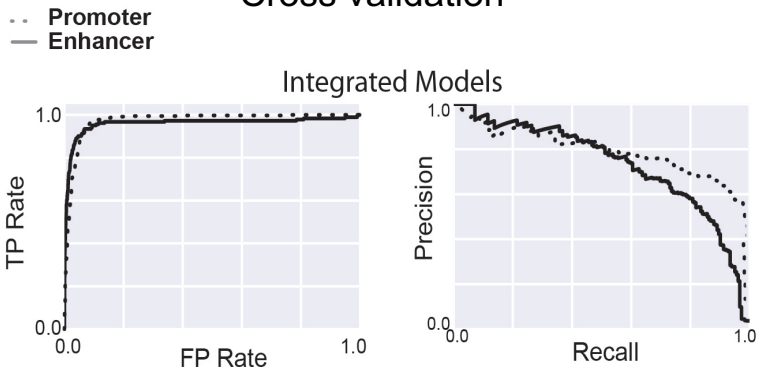


[biorxiv.org/content/early/2018/08/05/385237]

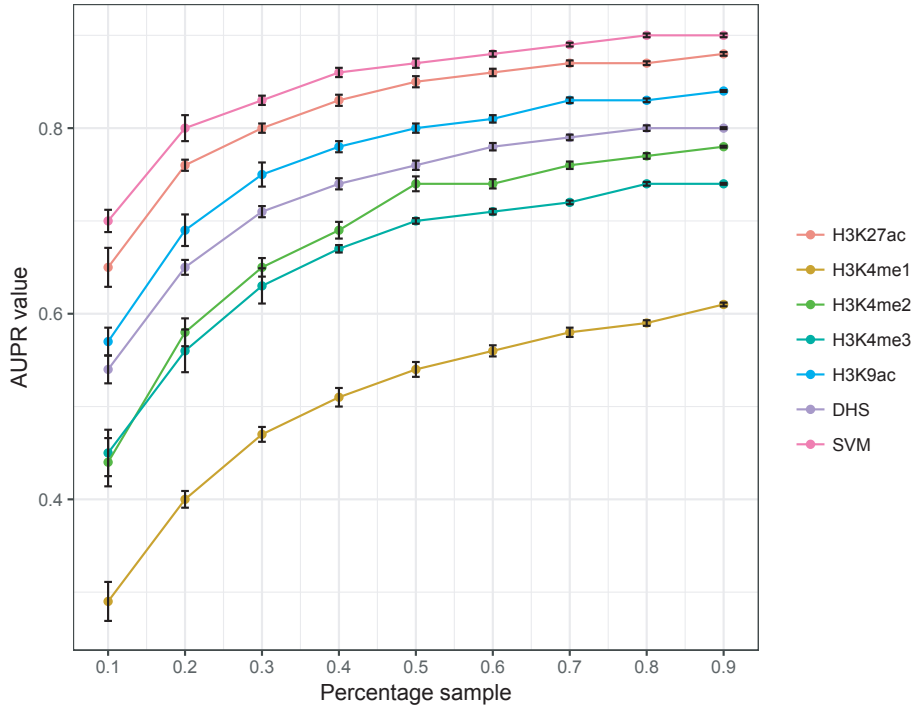
Integrate matched filter scores of multiple features

Model	AUROC	AUPR
Random Forest	0.96 (0.95)	0.91 (0.79)
Ridge Regression	0.95 (0.94)	0.90 (0.77)
Linear SVM	0.96 (0.95)	0.91 (0.78)
Naive Bayes	0.95 (0.93)	0.89 (0.72)

Cross validation

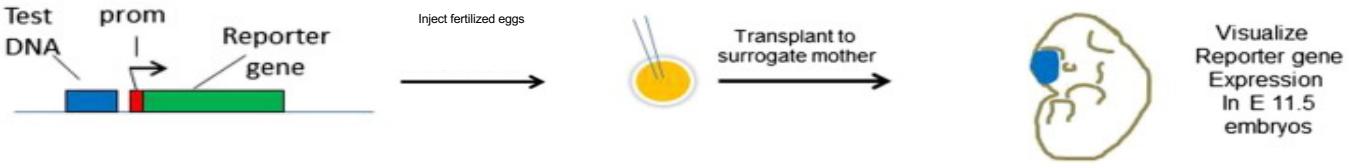


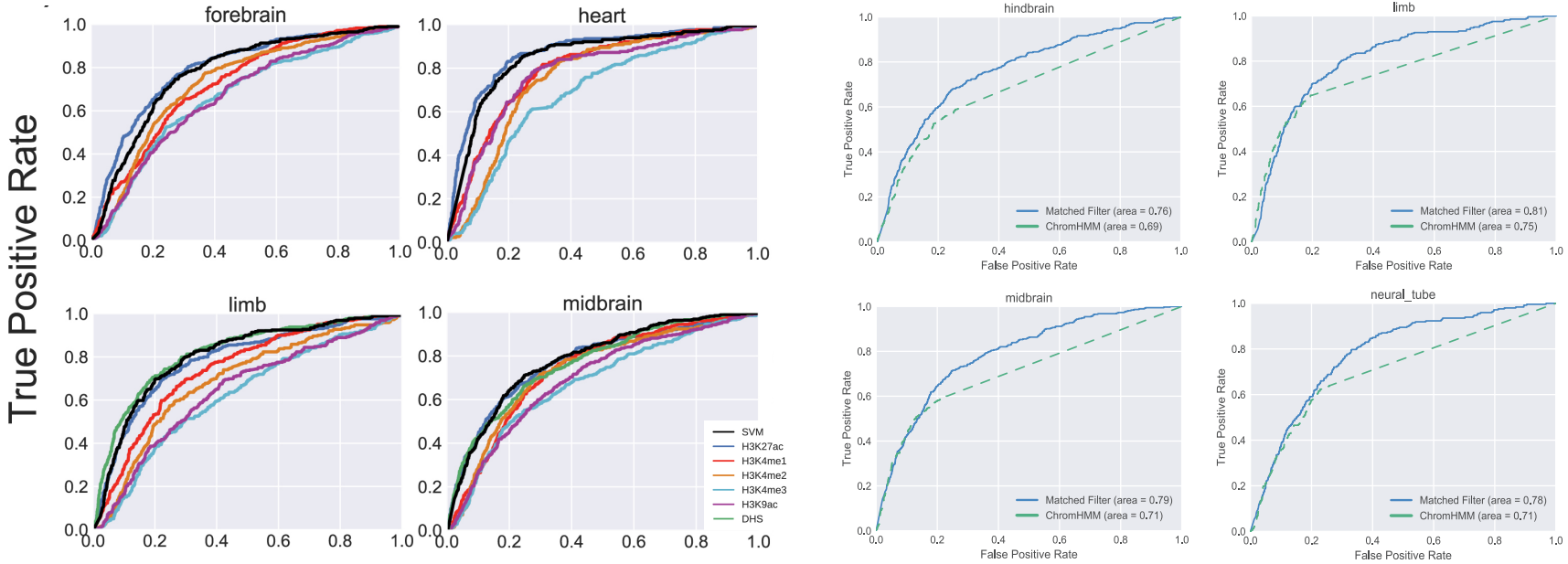
Large scale STARR-seq experiment data helps to improve the performance of integrated model



[biorxiv.org/content/early/2018/08/05/385237]

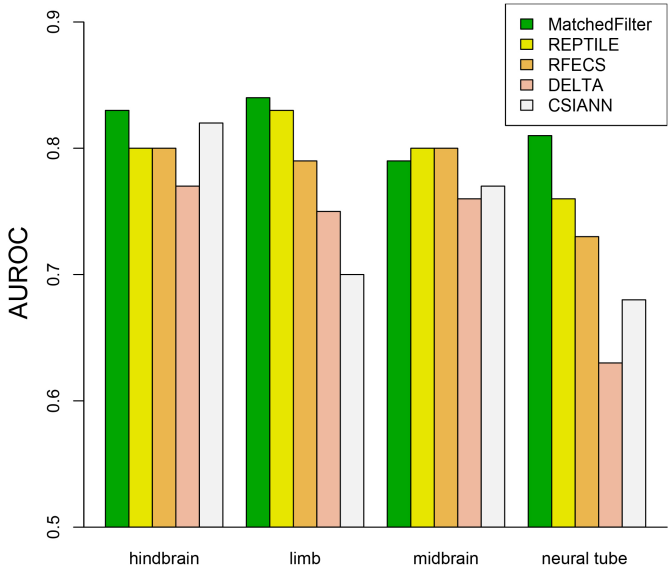
Validation with transgenic mouse enhancer assay

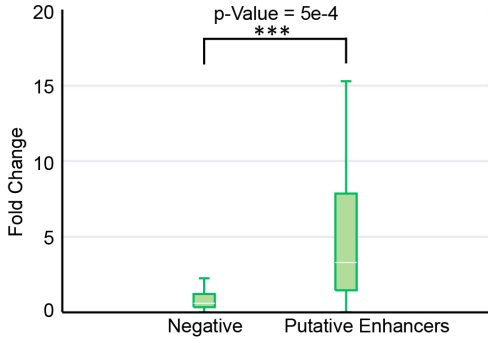




[biorxiv.org/content/early/2018/08/05/385237]

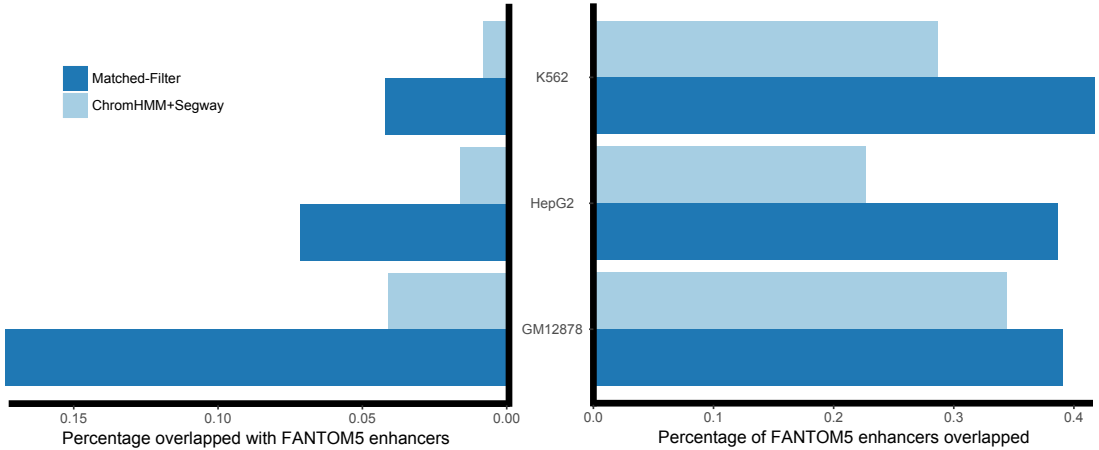
Matched-Filter can be applied across different organisms





Validation using transduction-based reporter assay (H1-hESC, HOS, A549 and TZMBL)

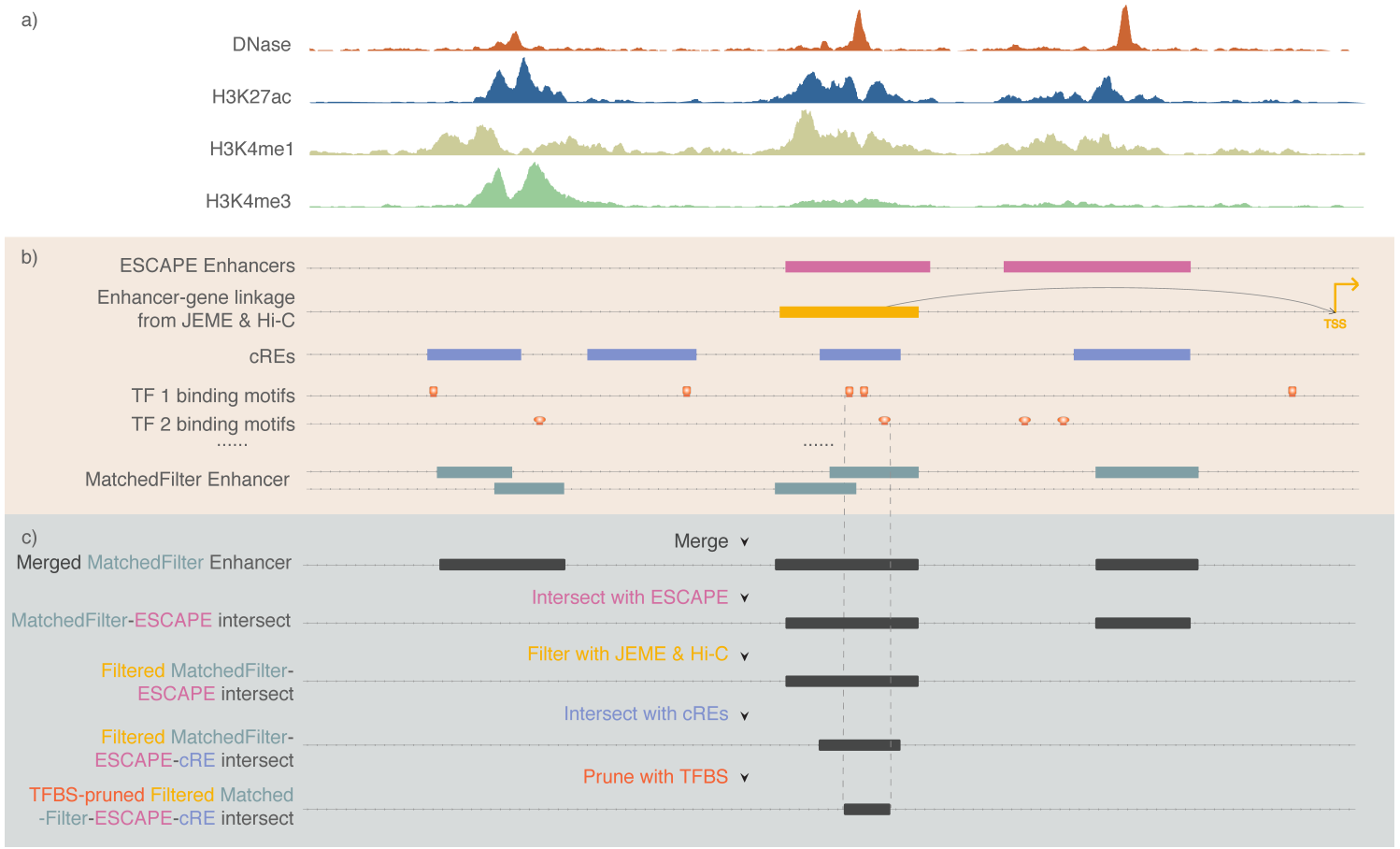
Compare overlap with FANTOM5 enhancers



Compare Matched-Filter performance with other state-of-the-art methods

[biorxiv.org/content/early/2018/08/05/385237]

Constructing a high-confidence set of cell-specific enhancers



Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

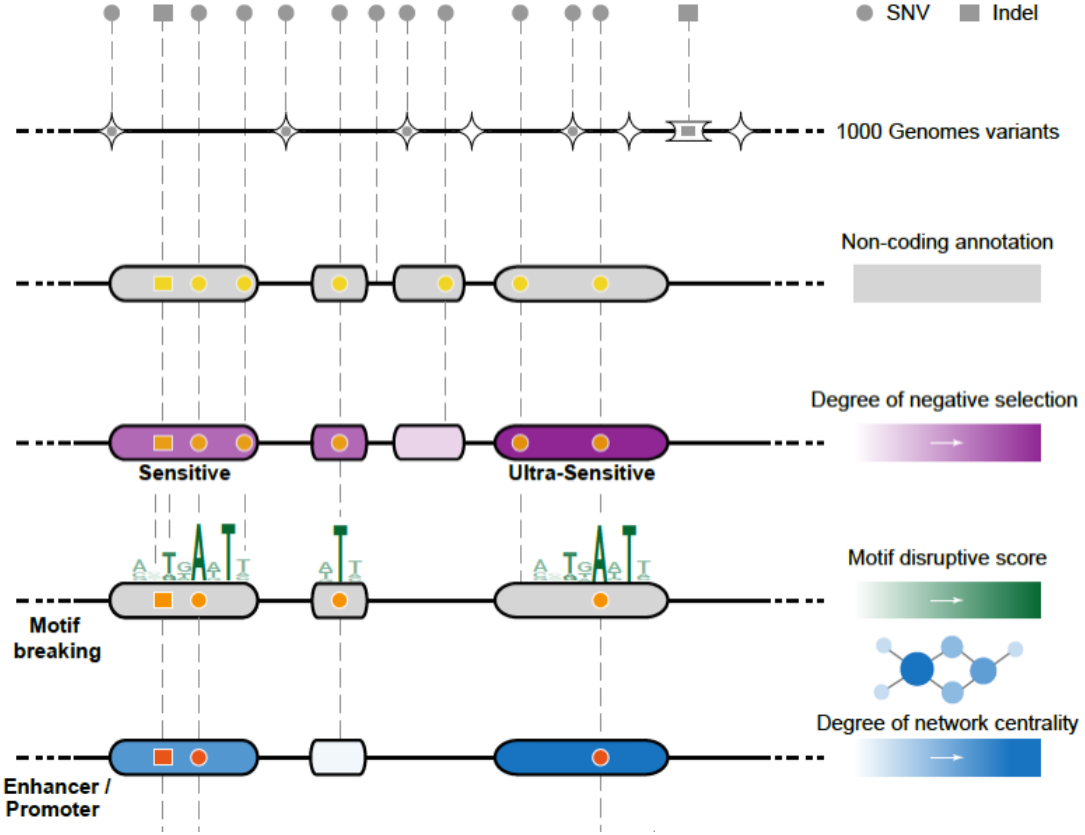
Funseq: a flexible framework to determine functional impact & use this to prioritize variants

Annotation (tf binding sites open chromatin, ncRNAs) & Chromatin Dynamics

Conservation (GERP, allele freq.)

Mutational impact (motif breaking, Lof)

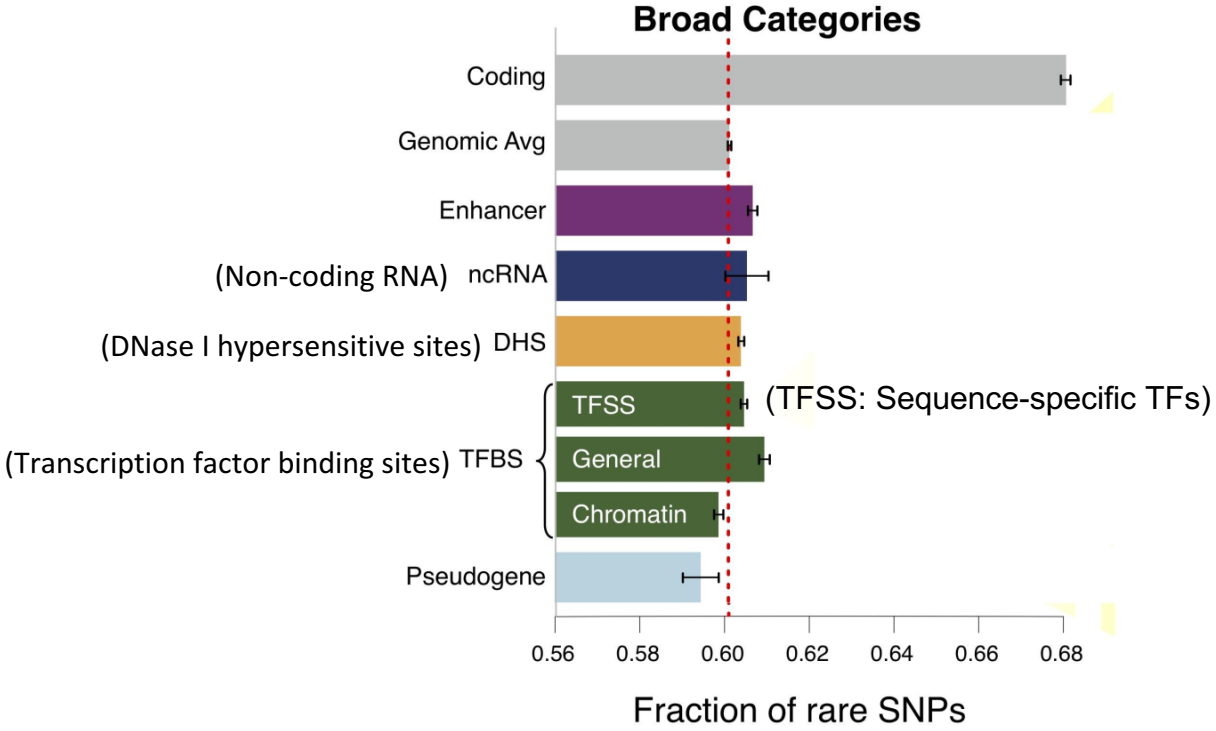
Network (centrality position)



[Fu et al., GenomeBiology ('14), , Khurana et al., Science ('13)]

Finding "Conserved" Sites in the Human Population:

Negative selection in non-coding elements based on Production ENCODE & 1000G Phase 1

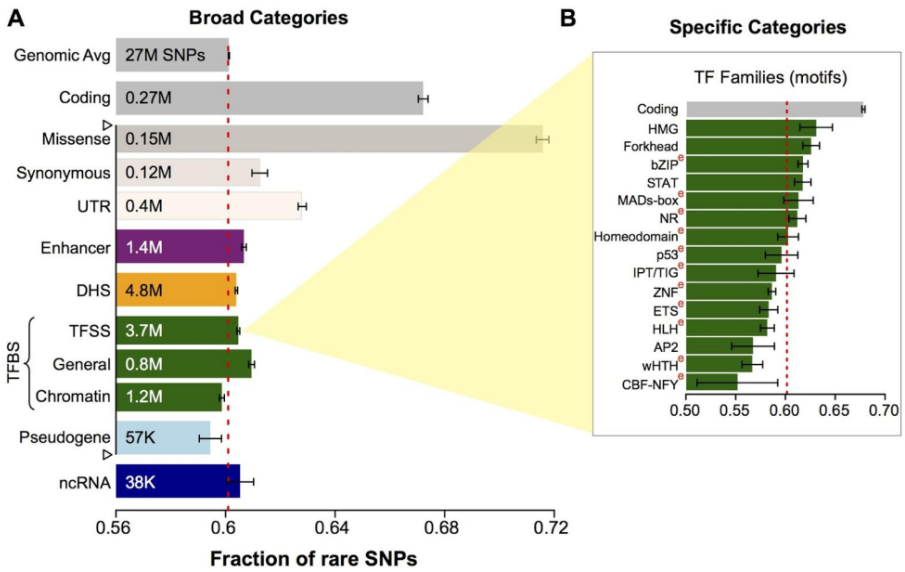


Depletion of Common Variants in the Human Population

Broad categories of regulatory regions under negative selection Related to:

ENCODE, *Nature*, 2012
Ward & Kellis, *Science*, 2012
Mu et al, *NAR*, 2011

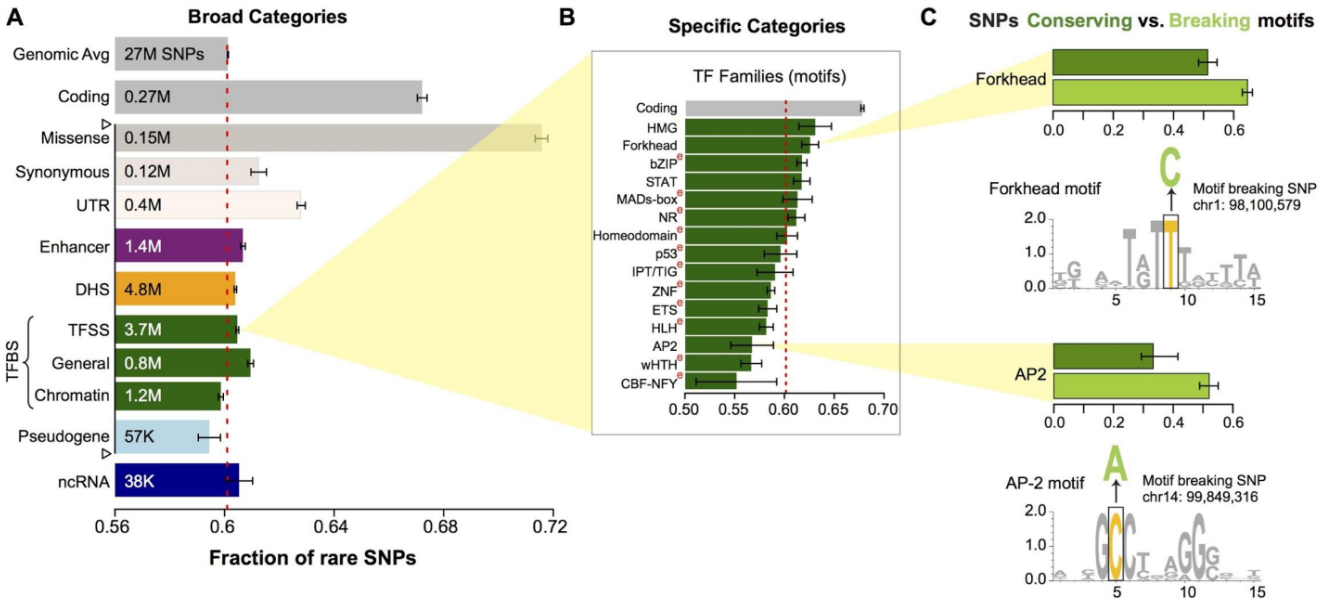
Differential selective constraints among specific sub-categories



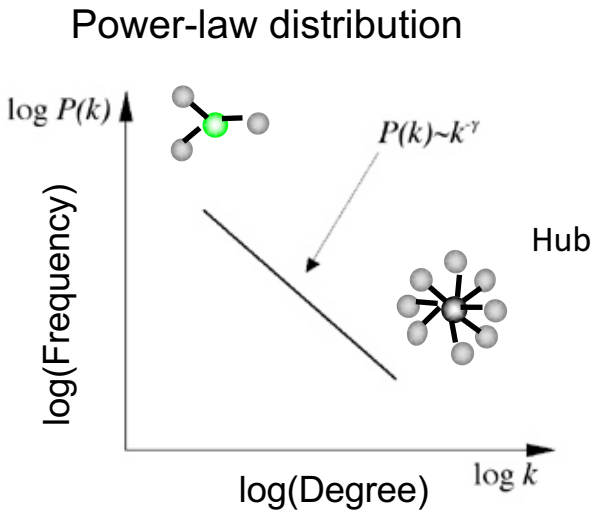
Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

[Khurana et al., *Science* ('13)]

SNPs which break TF motifs are under stronger selection



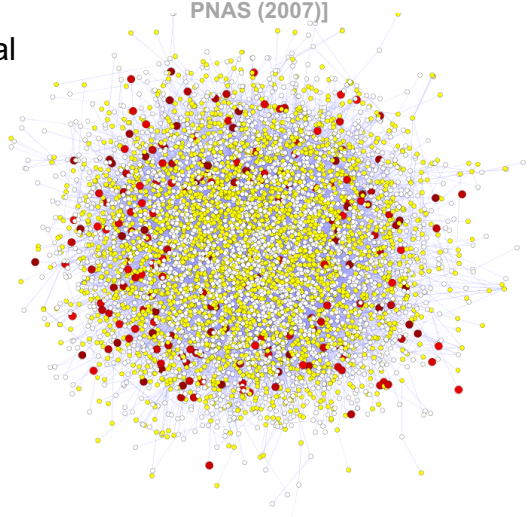
[Khurana et al., *Science* ('13)]



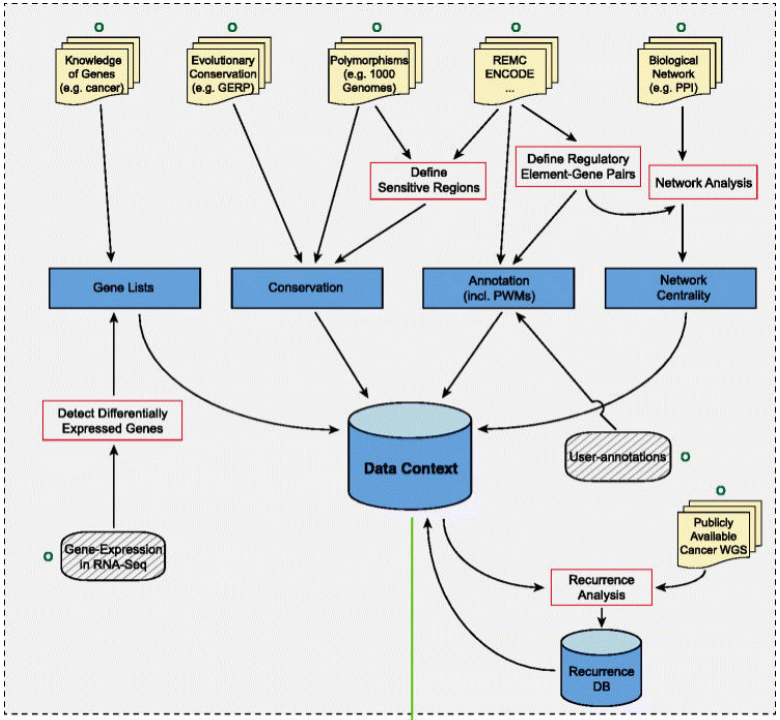
Hubs Under Constraint: A Finding from the Network Biology Community

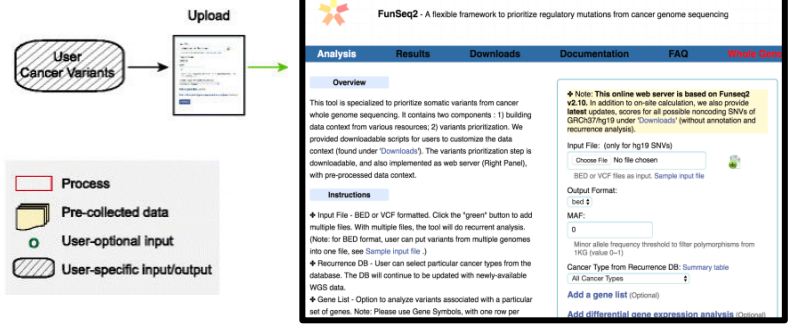
- High likelihood of positive selection
- Lower likelihood of positive selection
- Not under positive selection
- No data about positive selection

[Nielsen et al. *PLoS Biol.* (2005), HPRD, Kim et al. PNAS (2007)]



- More Connectivity, More Constraint: Genes & proteins that have a more central position in the network tend to evolve more slowly and are more likely to be essential.
- This phenomenon is observed in **many organisms & different kinds of networks**
  - **yeast PPI** - Fraser et al ('02) Science, ('03) BMC Evo. Bio.
  - **Ecoli PPI** - Butland et al ('04) Nature
  - **Worm/fly PPI** - Hahn et al ('05) MBE
  - **miRNA net** - Cheng et al ('09) BMC Genomics





FunSeq.gersteinlab.org

HOT region
Sensitive region
Polymorphisms

Genome

$$w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$$

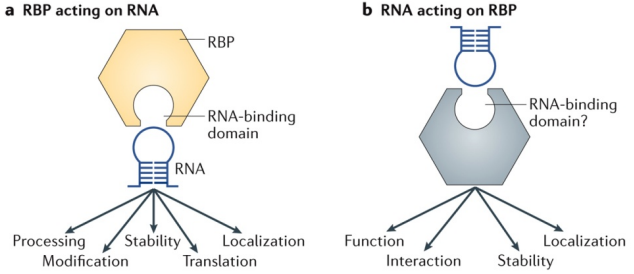
- Info. theory based method (ie annotation "surprisal") for weighting consistently many genomic features
- Practical web server
- Submission of variants & pre-computed large data context from uniformly processing large-scale datasets

[Fu et al., GenomeBiology ('14)]

Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

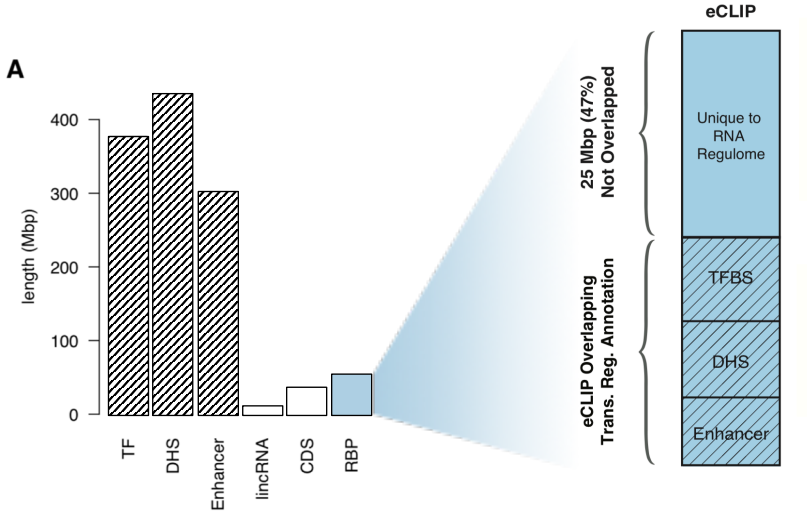
- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

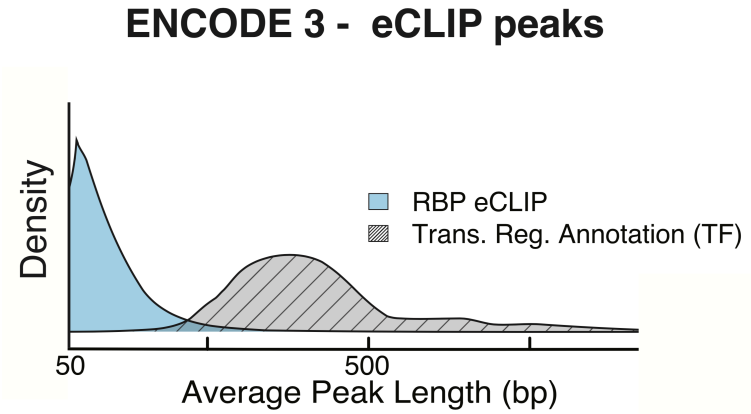
RNA Binding Proteins (RBPs)



Nat Rev Mol Cell Biol. 2018 May;19(5):327-341. doi: 10.1038/nrm.2017.130. Epub 2018 Jan 17.

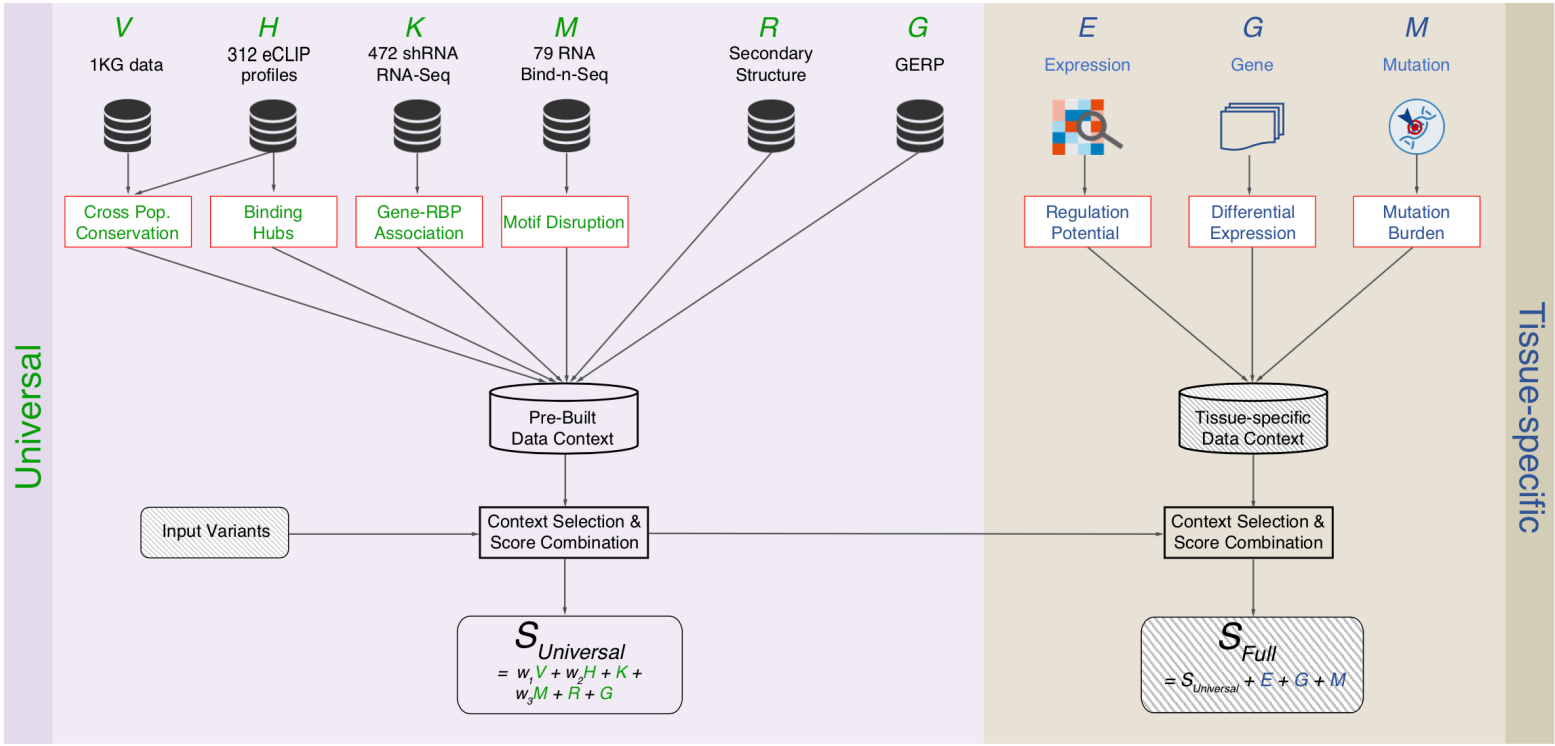
- **Before ENCODE3: >150 expt.**
  in many different cell types
- **ENCODE3 did ~350 focused eCLIP expt.**
  for >110 RBPs on HepG2 & K562
  (Van Nostrand...Yeo. Nat. Meth. '16;
  Van Nostrand...Graveley, Yeo
  (submitted in relation to ENCODE3))

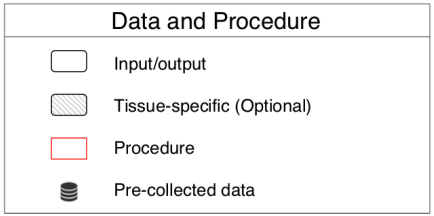


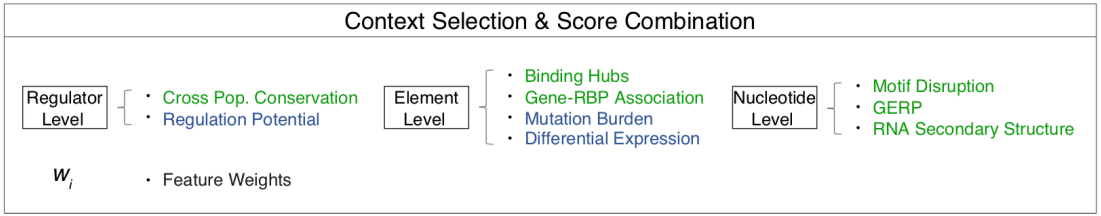


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

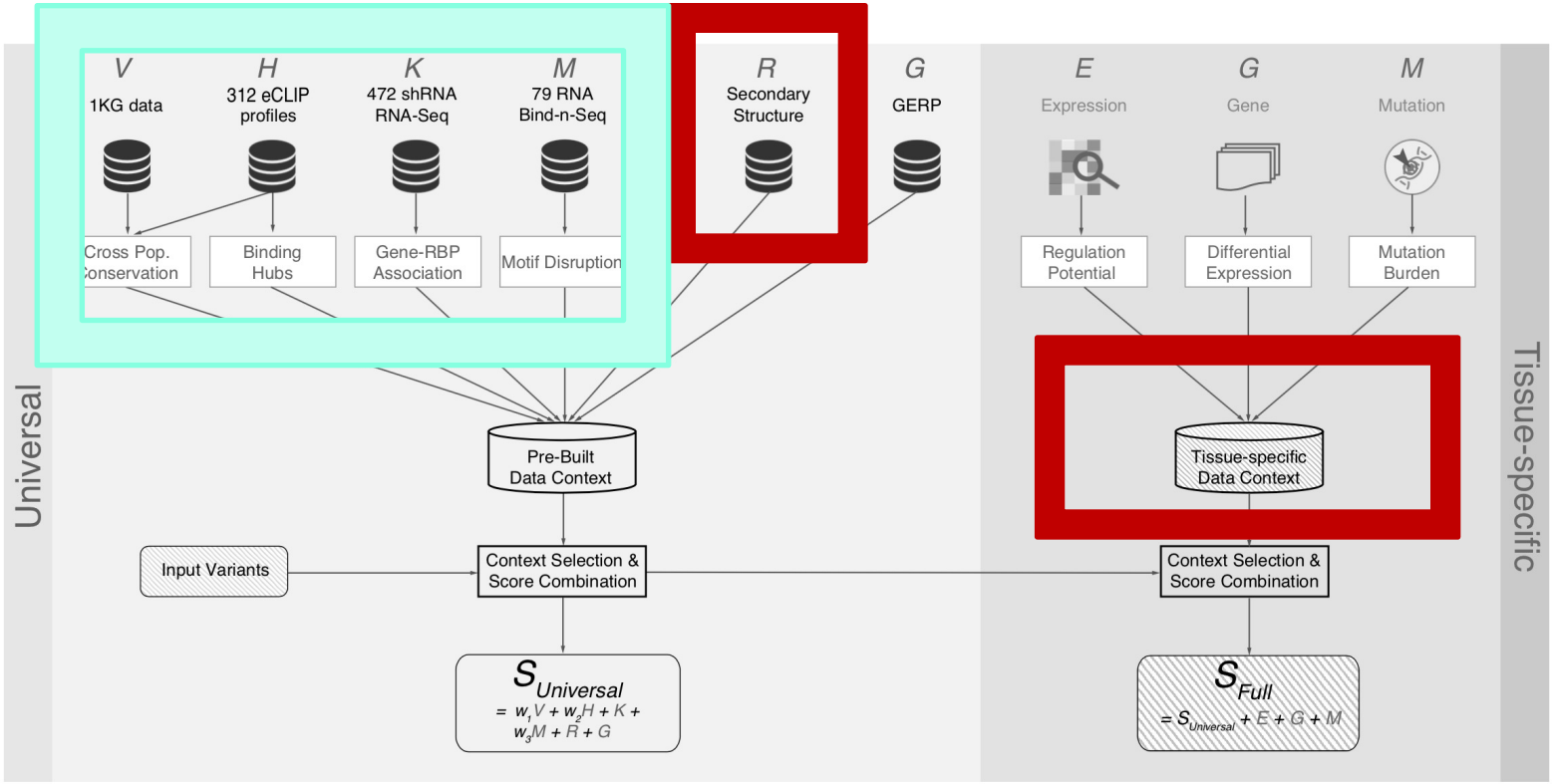
Schematic of RADAR Scoring

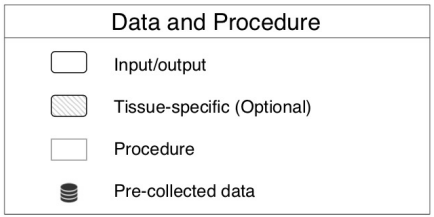


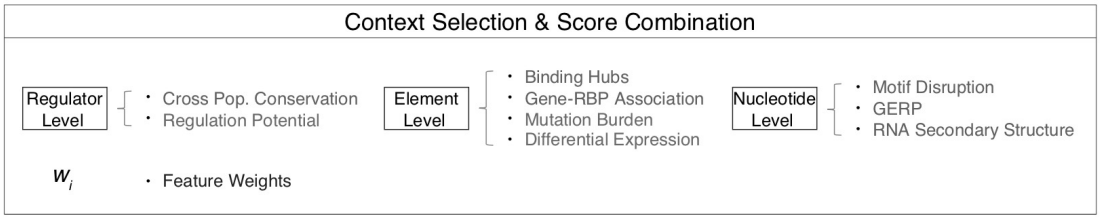




[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

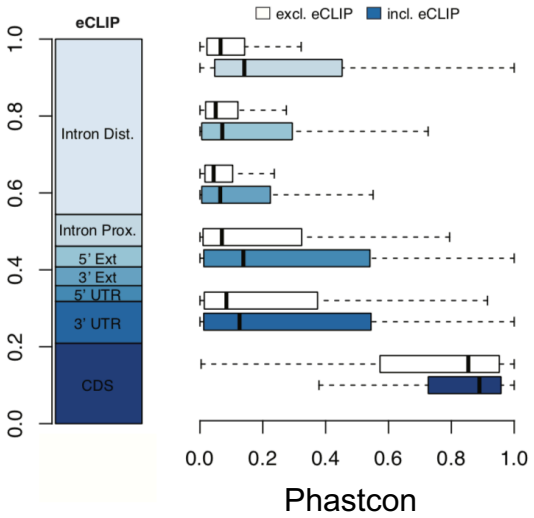




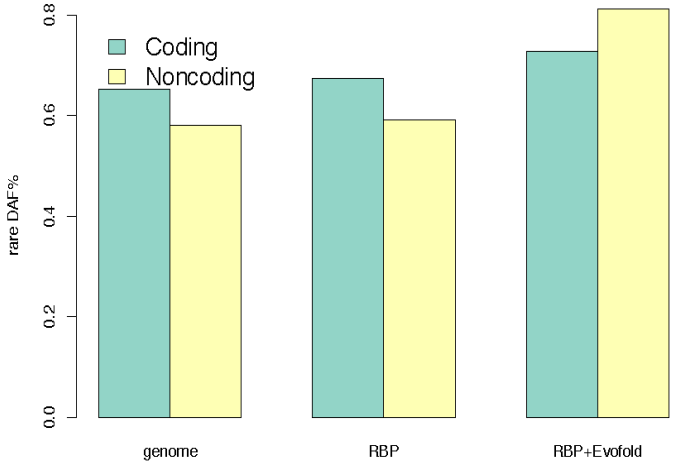


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

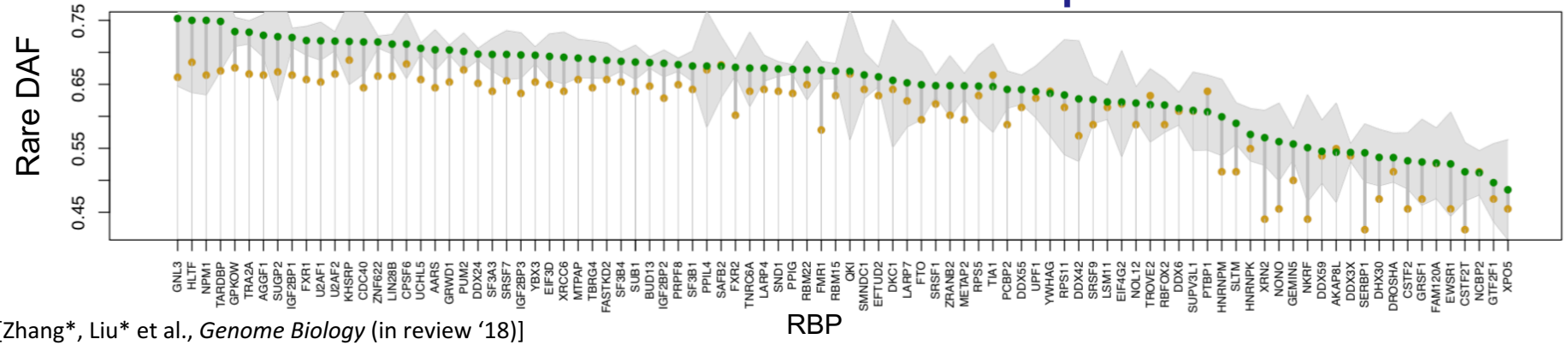
High Phastcon in RBP-overlapped annotations



RNA Structure Cons. from Evofold



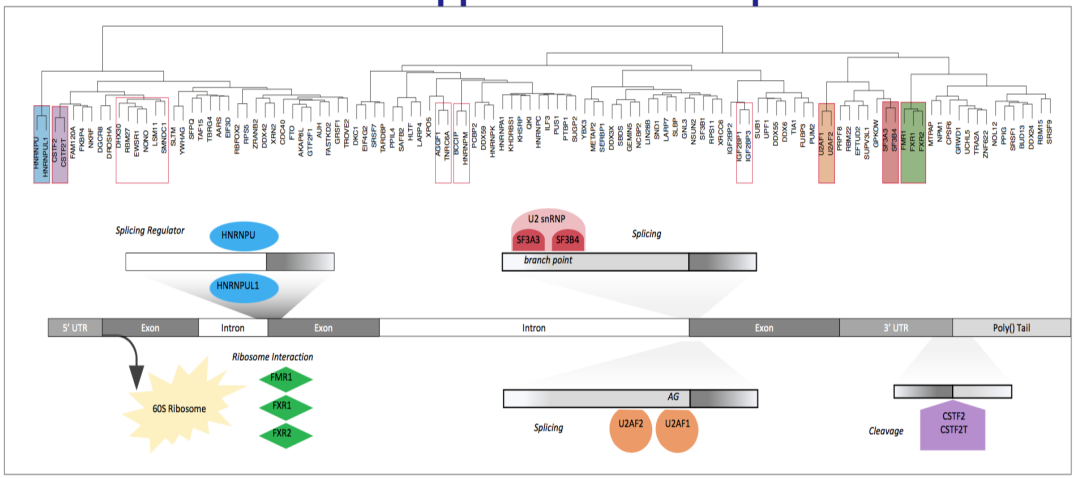
Enriched rare DAF in eCLIP peaks



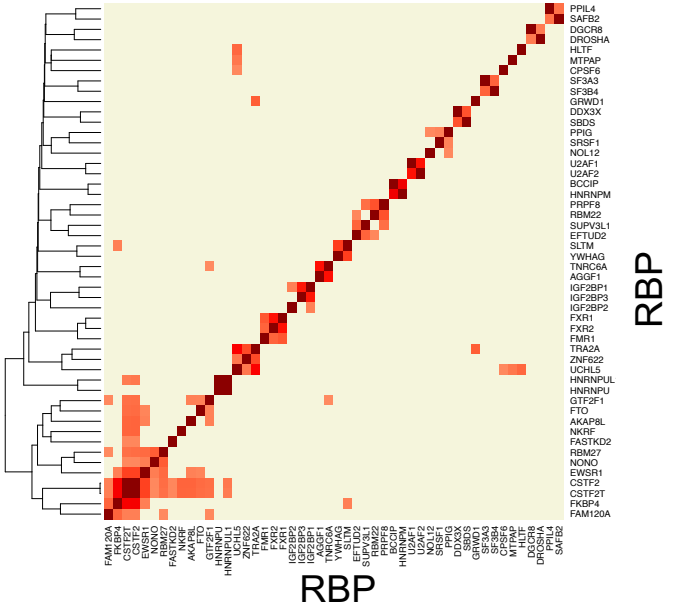
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Co-binding of RBPs form biologically relevant complexes

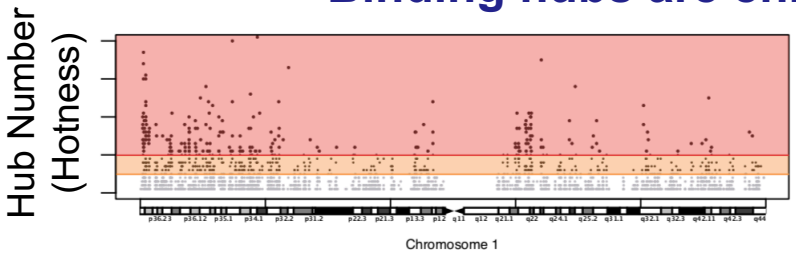
Literature supported RBP complexes

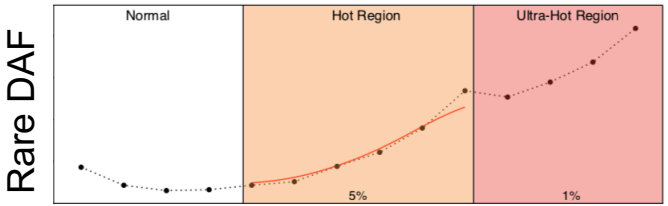


Unique co-binding patterns of RBPs



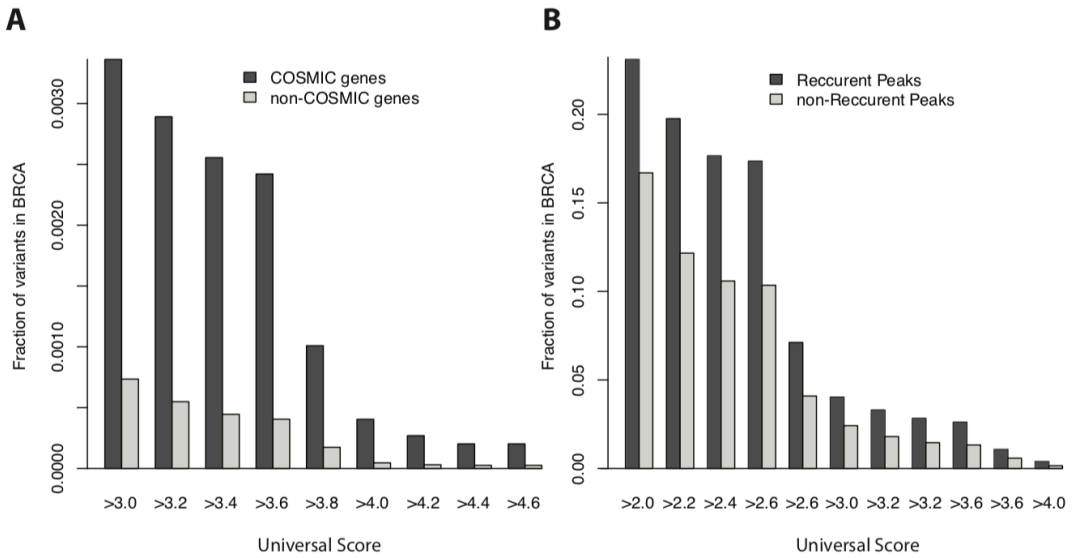
Binding hubs are enriched for rare variants





[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Validation for Somatic Variants:
RADAR Scores enriched in COSMIC genes & recurrently mutated regions
+ higher for tissue matched context

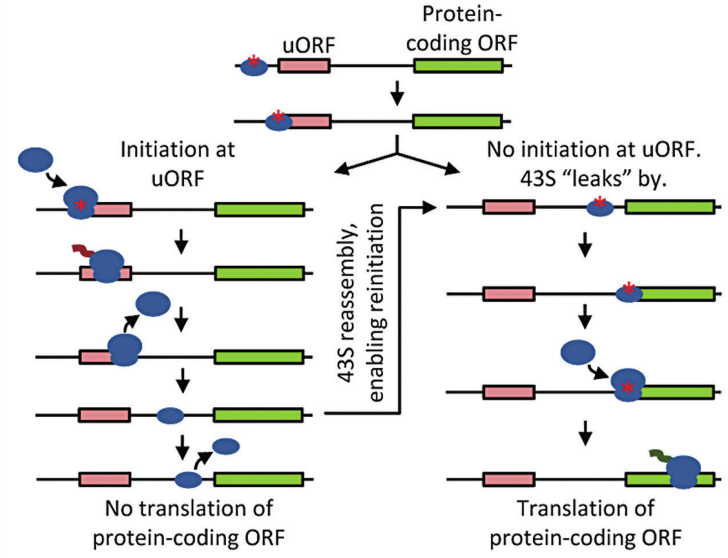


Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

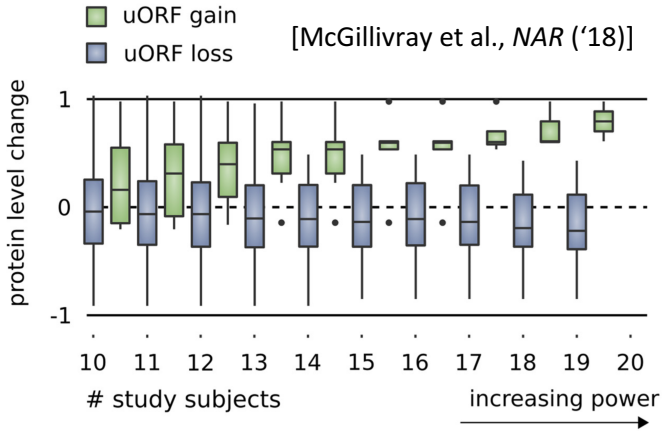
Upstream open reading frames (uORFs) regulate translation are affected by somatic mutation

[Calvo et al., *PNAS* (‘09)]

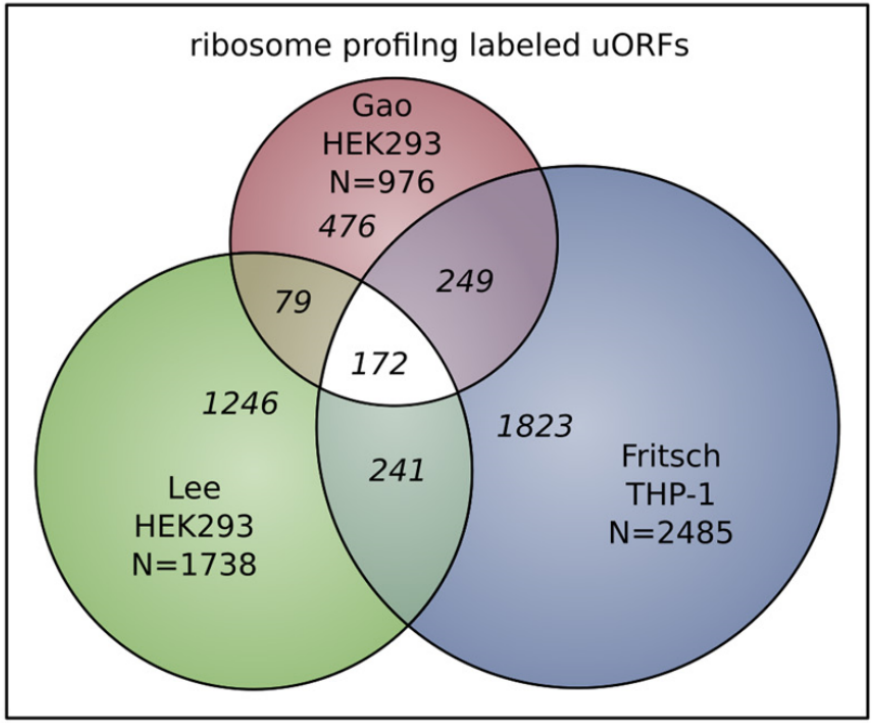


[Ferreira et al., *Bioengineered* (‘14)]

- uORFs regulate the translation of downstream coding regions.
- This regulation may be altered by somatic mutation in cancer.
- In Battle et al. 2014 data uORF gain & loss assoc. protein level change.

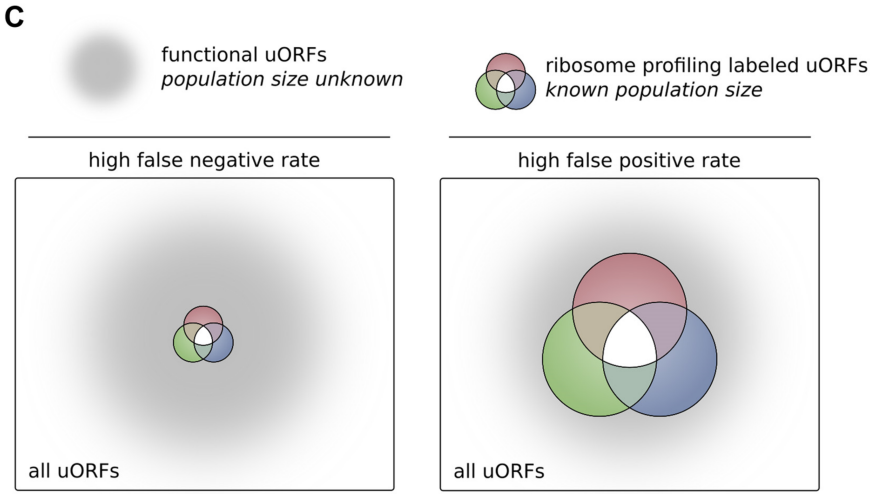


[McGillivray et al., *NAR* (‘18)]



From a “Universe” of 1.3 M pot. uORFs

The population of functional uORFs may be significant

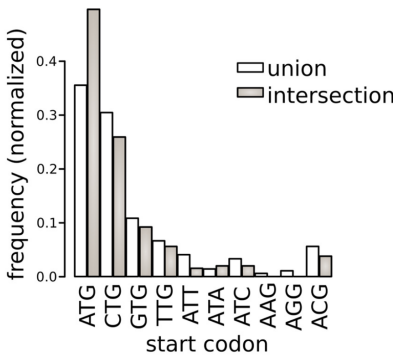


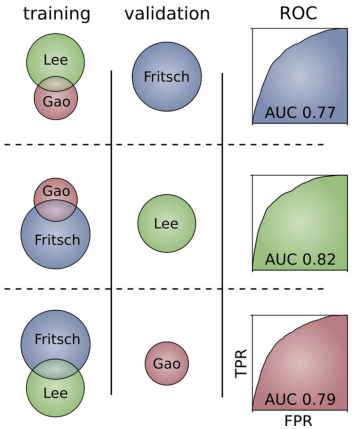
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

[McGillivray et al., *NAR* (‘18)]

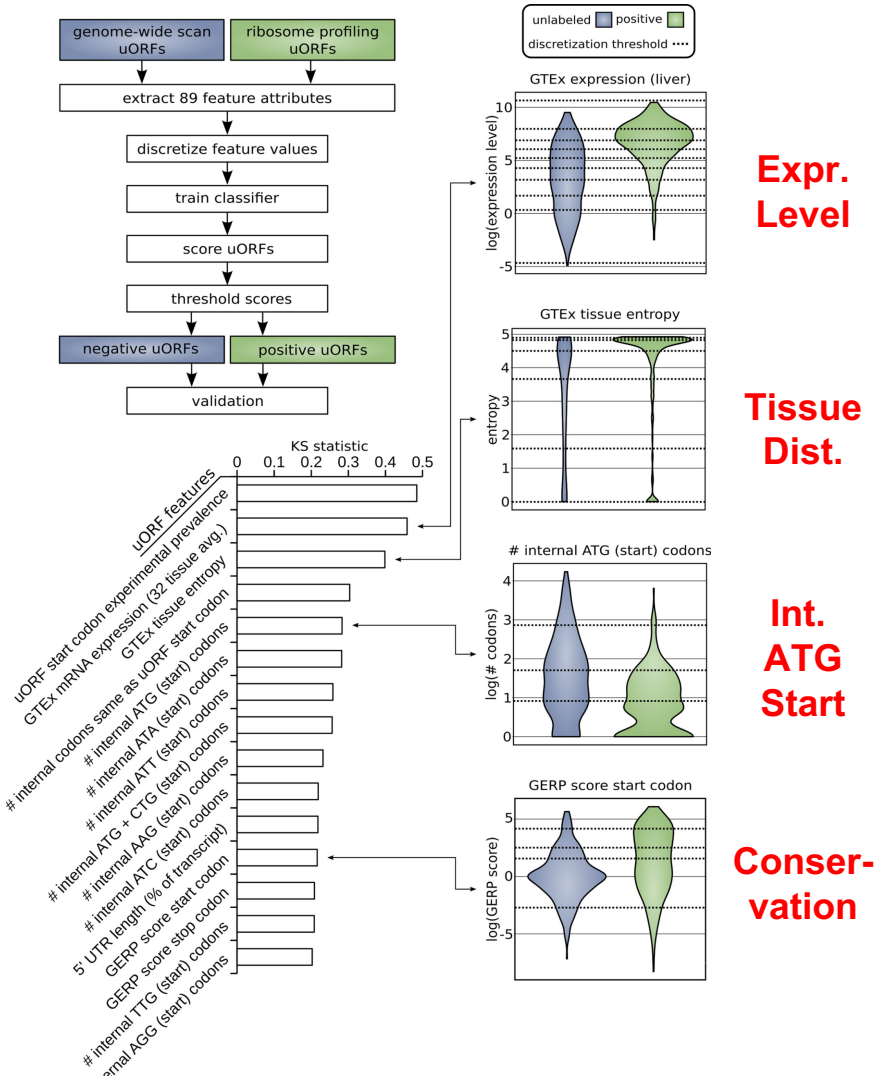
Prediction & validation of functional uORFs using 89 features

- All near-cognate start codons predicted.
- Cross-validation on independent ribosome profiling datasets and validation using in vivo protein levels and ribosome occupancy in humans (Battle et al. 2014).



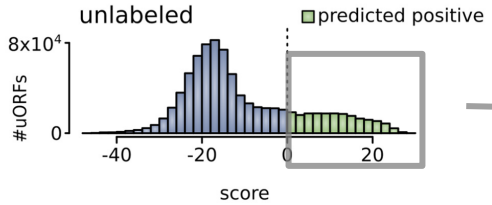


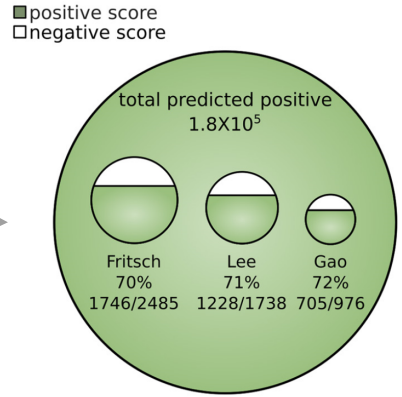
[McGillivray et al., *NAR* ('18)]

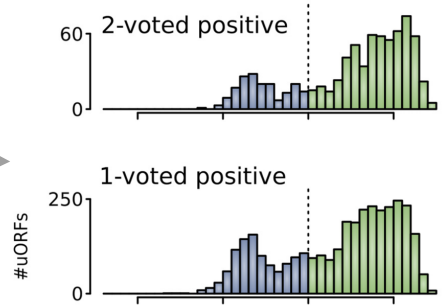


A comprehensive catalog of functional uORFs

Universe of **1.3M** uORFs scored via Simple Bayes algo.







- Predicted functional uORFs may be intersected with disease associated variants.

- **180K**: Large predicted positive set likely to affect translation
- Calibration on gold standards, suggests getting **~70%** of known

[McGillivray et al., *NAR* ('18)]

Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

Cancer Somatic Mutation Modeling

PARAMETRIC MODELS

Model 1: Constant Background Mutation Rate (Model from Previous Work)

$x_i : Binomial(n_i, p)$

Model 2a: Varying Mutation Rate with Single Covariate Correction

$x_i : Binomial(n_i, p_i)$

$p_i : Beta(\mu|R_i, \sigma|R_i)$

$\mu|R_i, \sigma|R_i$: constant within the same covariate rank

Model 2b: Varying Mutation Rate with Multiple Covariate Correction

$x_i : Binomial(n_i, p_i)$

$p_i : Beta(\mu|\boldsymbol{R_i}, \sigma|\boldsymbol{R_i})$

$\mu|\boldsymbol{R_i}, \sigma|\boldsymbol{R_i}$: constant within the same covariate rank

[Lochovsky et al. *NAR* ('15)]

- Suppose there are *k* genome elements. For element *i*, define:
  - n_i: total number of nucleotides
  - x_i: the number of mutations within the element
  - p: the mutation rate
  - R_i: the covariate rank of the element
- Non-parametric model is useful when covariate data is missing for the studied annotations
  - Also sidesteps issue of properly identifying and modeling every relevant covariate (possibly hundreds)

NON-PARAMETRIC MODELS

Assume constant background mutation rate in local regions.

Model 3a: Random Permutation of Input Annotations

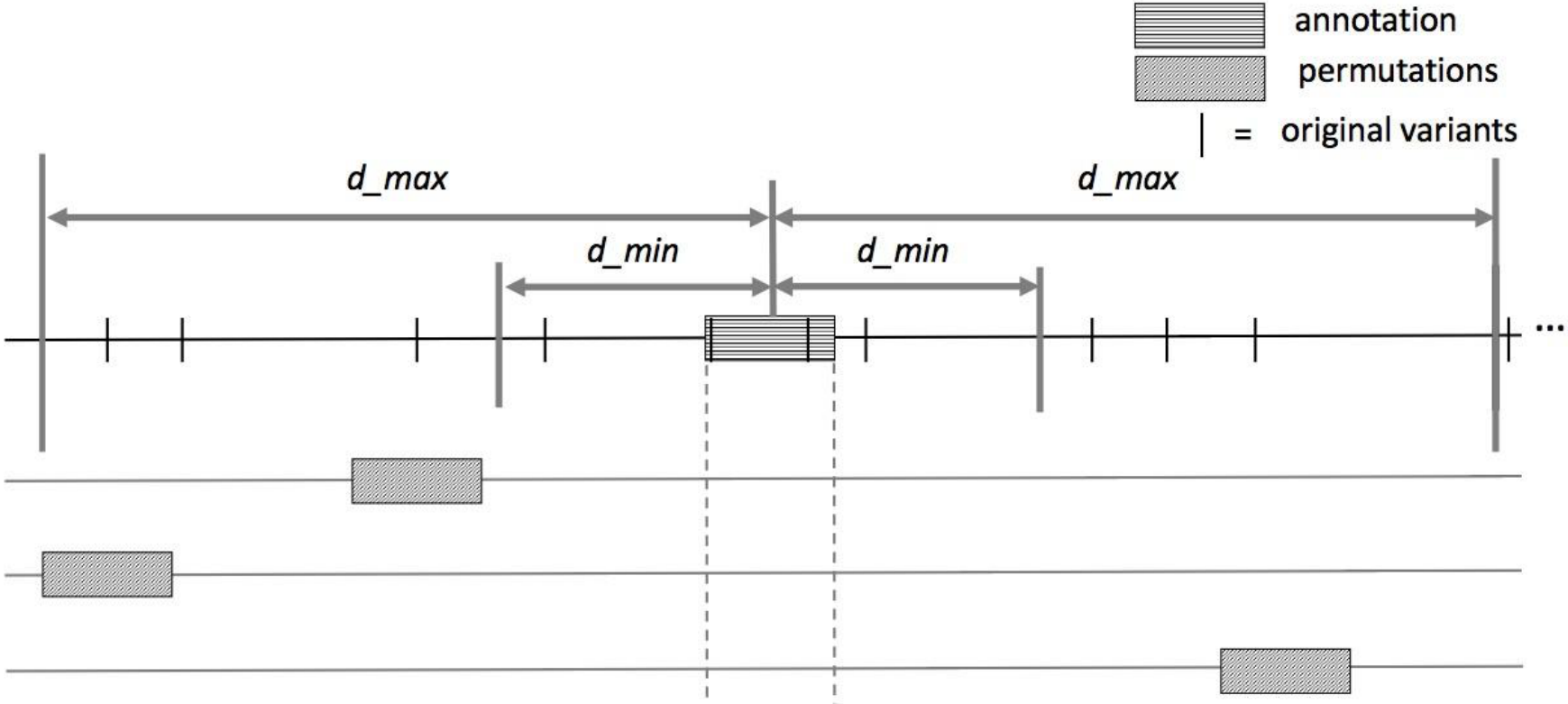
Shuffle annotations within local region to assess background mutation rate.

Model 3b: Random Permutation of Input Variants

Shuffle variants within local region to assess background mutation rate.

[Lochovsky et al. *Bioinformatics* in press]

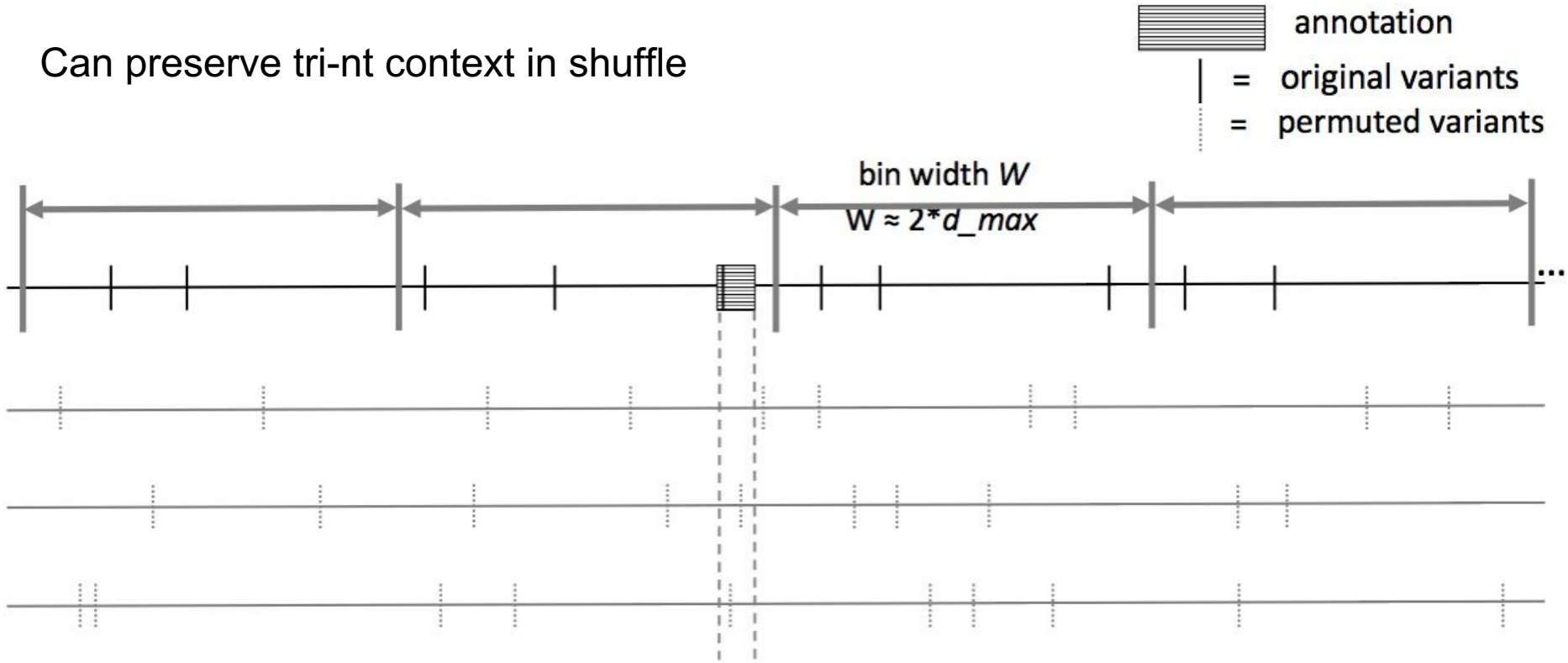
MOAT-a: Annotation-based permutation



[Lochovsky et al. *Bioinformatics* in press]

MOAT-v: Variant-based Permutation

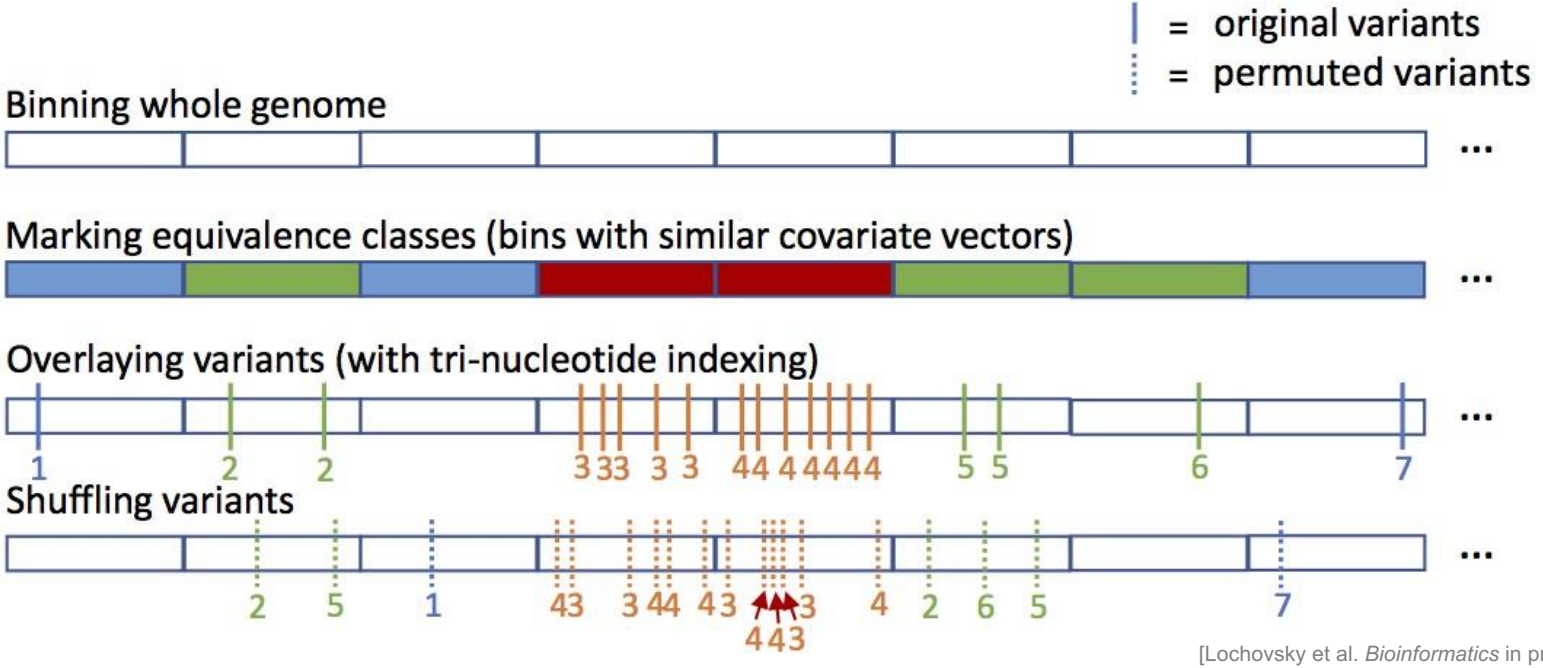
Can preserve tri-nt context in shuffle



[Lochovsky et al. *Bioinformatics* in press]

MOAT-s: a variant on MOAT-v

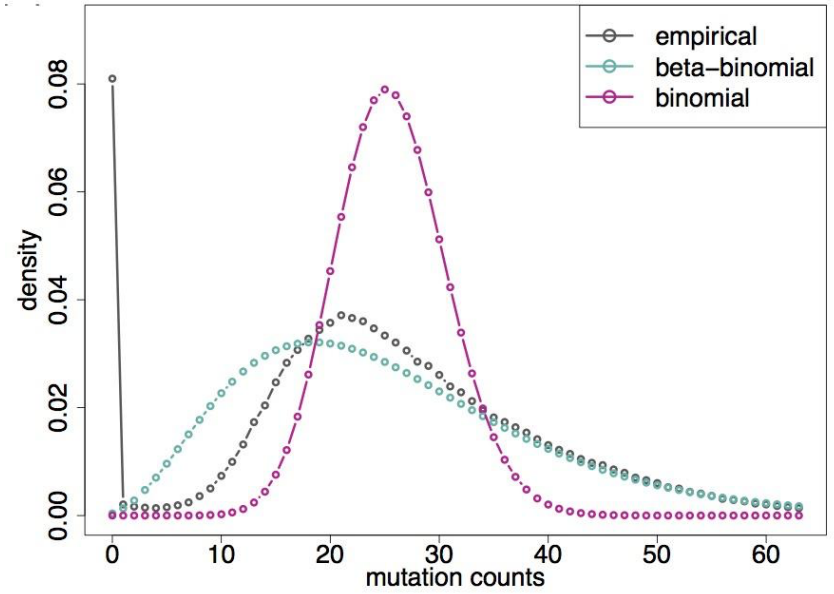
- A somatic variant simulator
  - Given a set of input variants, shuffle to new locations, taking genome structure into account



[Lochovsky et al. *Bioinformatics* in press]

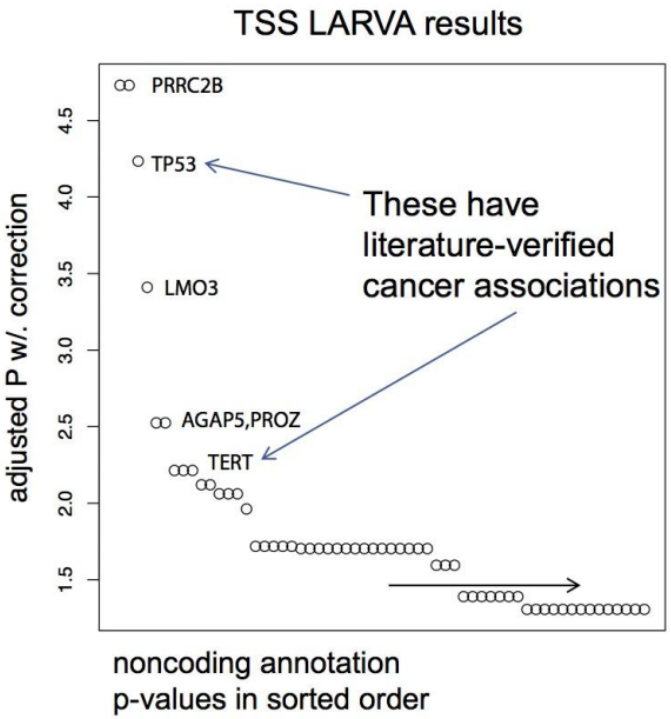
LARVA Model Comparison

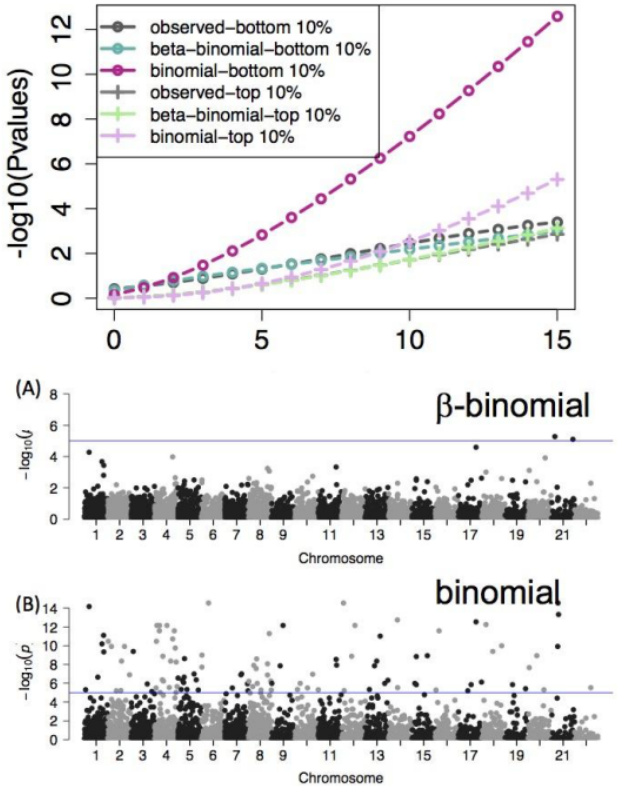
- Comparison of mutation count frequency implied by the binomial model (model 1) and the beta-binomial model (model 2) relative to the empirical distribution
- The beta-binomial distribution is significantly better, especially for accurately modeling the over-dispersion of the empirical distribution



[Lochovsky et al. *NAR* ('15)]

LARVA Results





[Lochovsky et al. *NAR* ('15)]

MOAT: recapitulates LARVA with GPU-driven runtime scalability

Gene Name	Documented role with cancer	Pubmed ID
SLC3A1	Cysteine transporter SLC3A1 promotes breast cancer tumorigenesis	28382174
ADRA2B	reduce cancer cell proliferation, invasion, and migration	25026350
SIL1	subtype-specific proteins in breast cancer	23386393
TCF24	NA	NA
AGAP5	significant mutation hotspots in cancer	25261935
TMPRSS13	Type II transmembrane serine proteases in cancer and viral infections	19581128
ERO1L	Overexpression of ERO1L is Associated with Poor Prognosis of Gastric Cancer	26987398

.
.
.

MOAT's high mutation burden elements recapitulate LARVA's results & published noncoding cancer-associated elements.

Computational efficiency of MOAT's NVIDIA™ CUDA™ version, with respect to the number of permutations, is dramatically enhanced compared to CPU version.

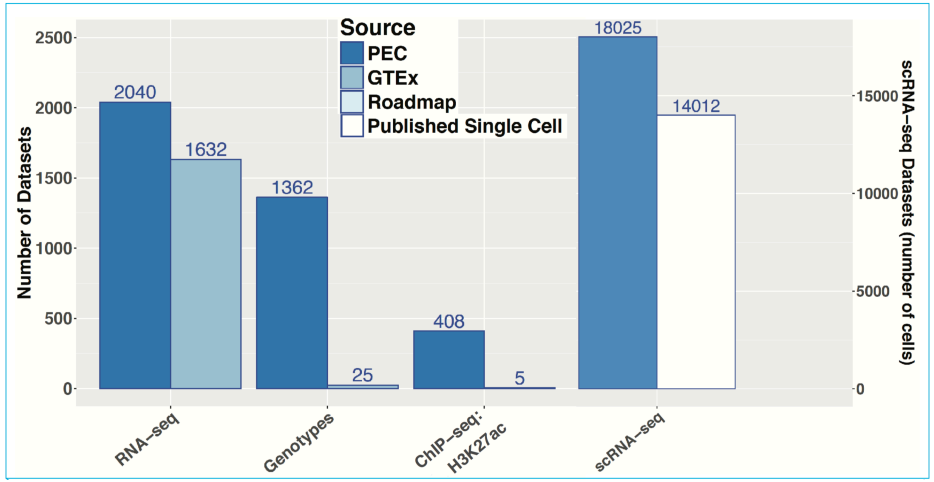
Number of permutations	Fold speedup of CUDA version
1k	14x
10k	100x
100k	256x

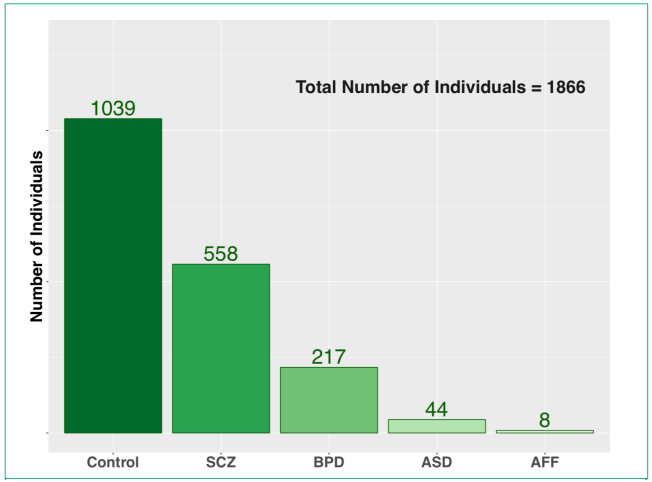
[Lochovsky et al. *Bioinformatics* in press]

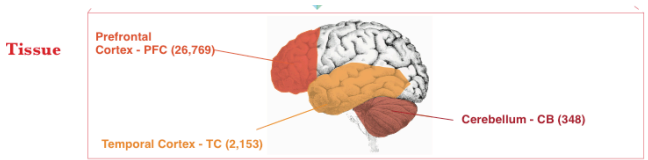
Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

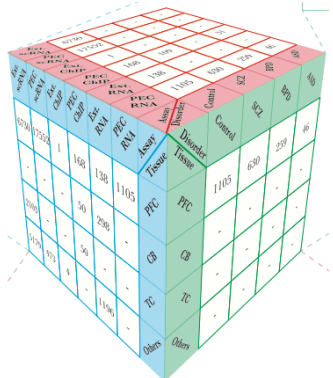
- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

Multi-omics of PsychENCODE & other consortia to understand functional genomics in brain disorders



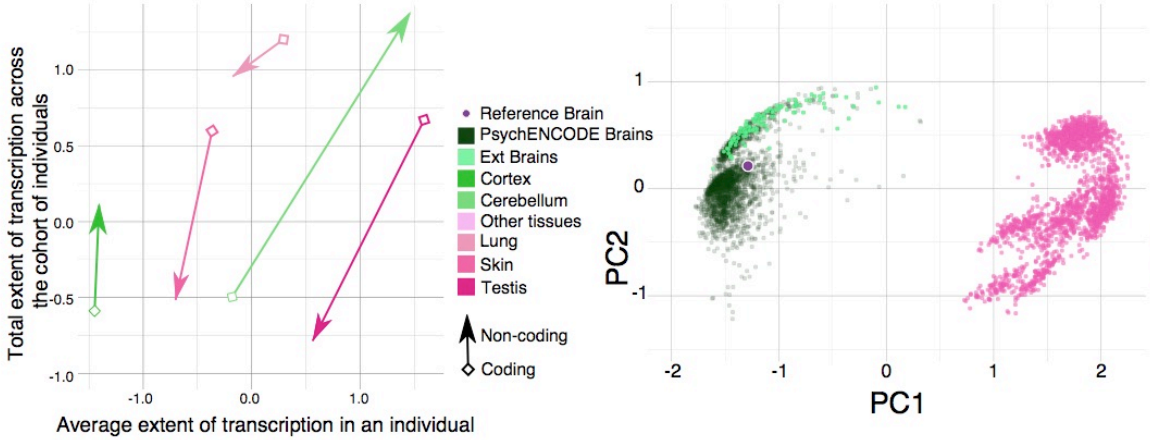




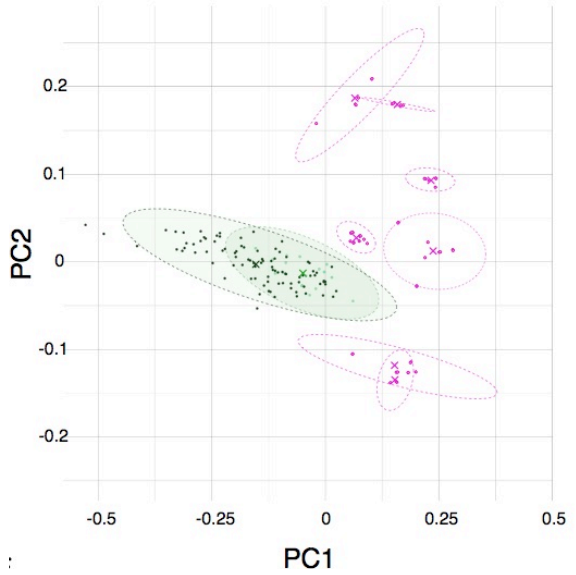


- Brain transcriptome and epigenome
- Single cell deconvolution
- QTL
- Regulatory network
- Deep Neural Network for predicting disorders

Brain specific transcriptome and epigenome from comparative analysis

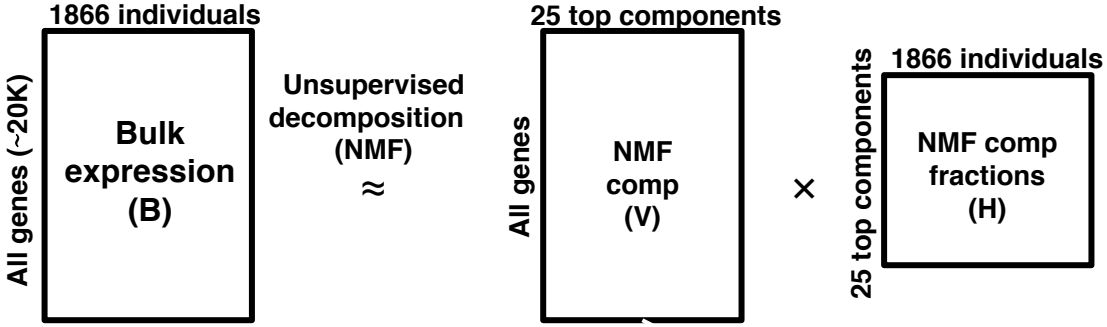


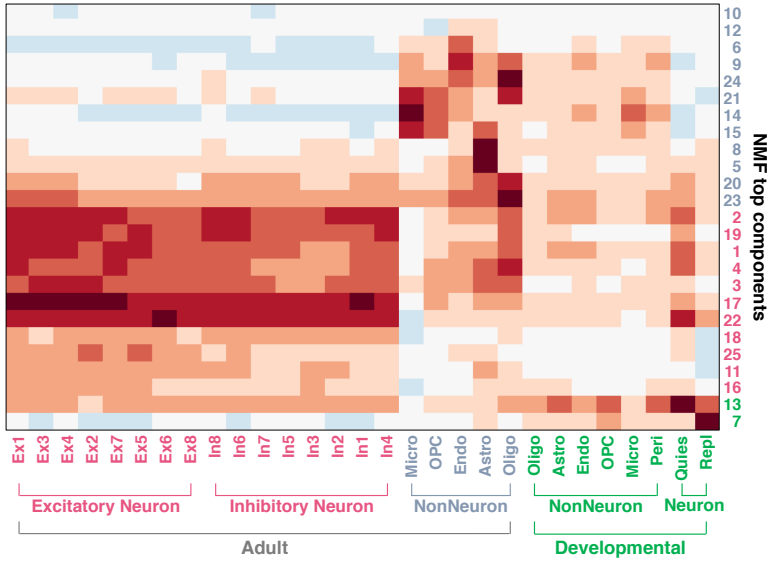
Transcriptome

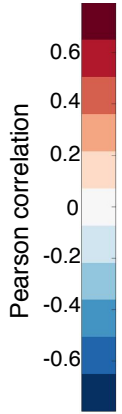


Epigenome
(~79,000 brain enhancers)

Single cell deconvolution Step 1: unsupervised learning to see cell types

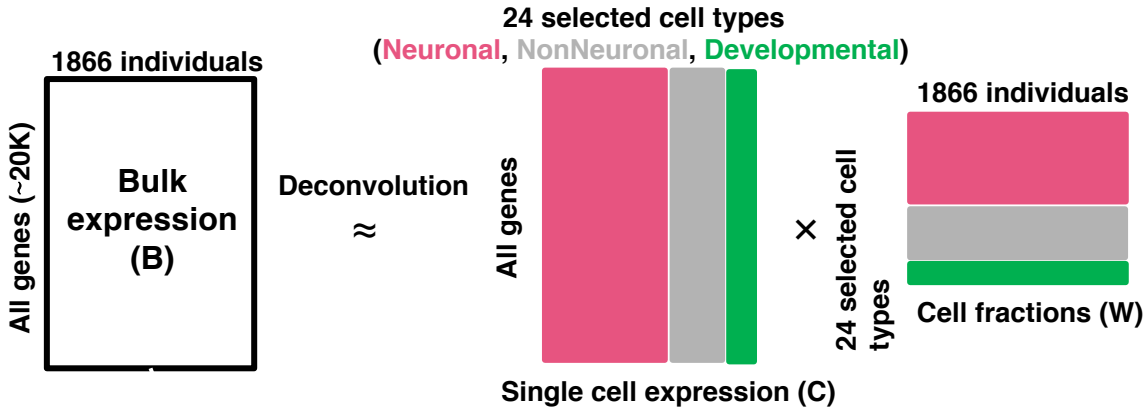


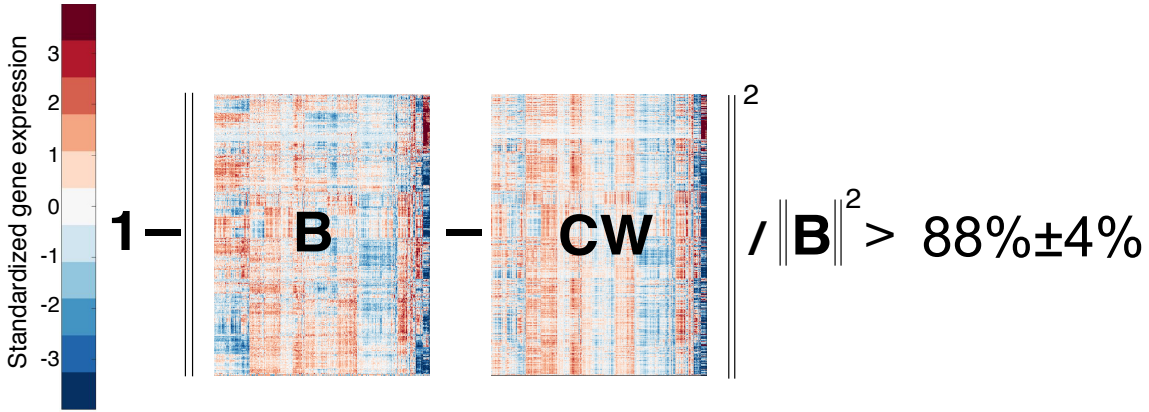




Single cell signatures
- ~14,000 cells (Lake et al., Science, 2016&2018)
- ~400 cells (Darmanis et al., PNAS, 2015)
- ~18,000 cells (PsychENCODE, submitted)

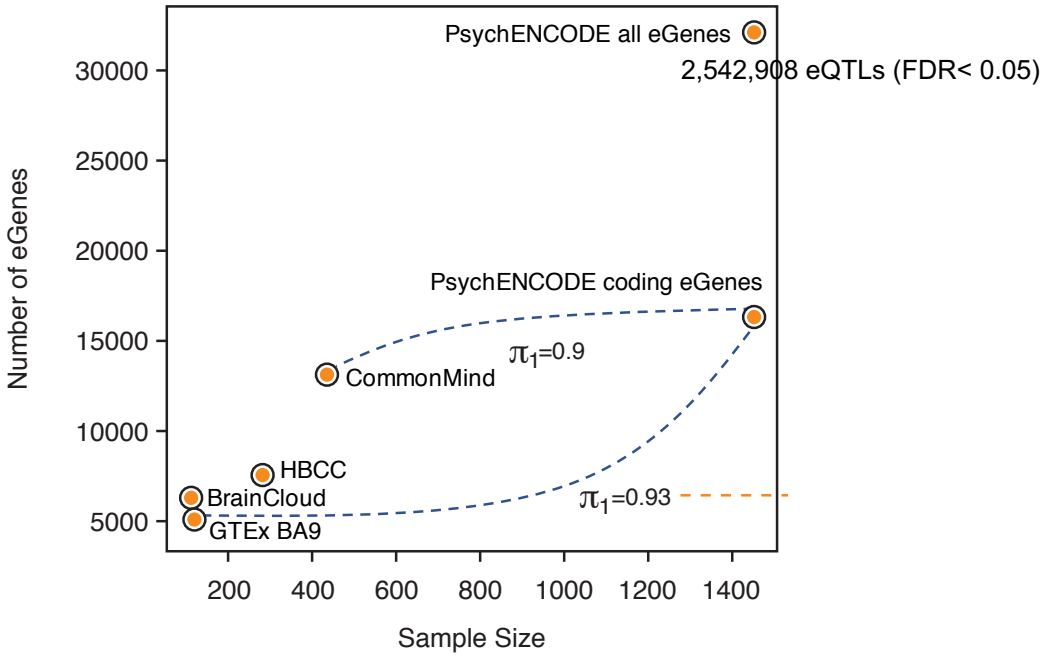
Single cell deconvolution Step 2: supervised learning to estimate cell fractions

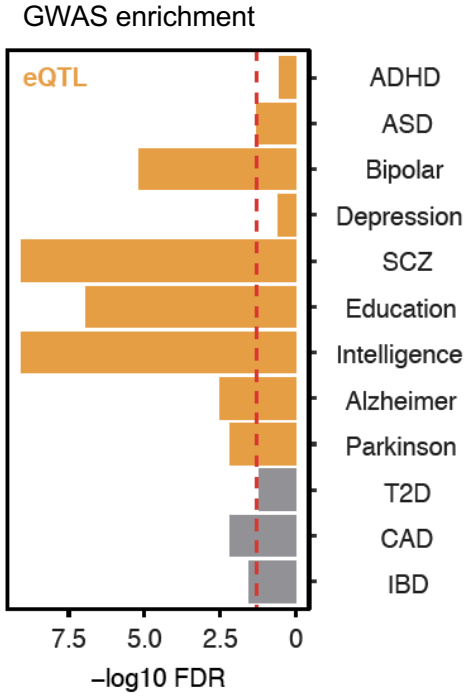




Individual and cross-population reconstruction accuracy via deconvolution

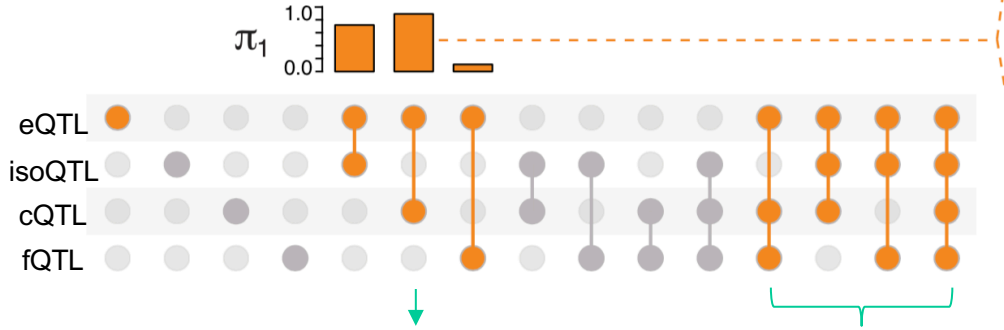
Larger Brain eQTL sets than previous studies



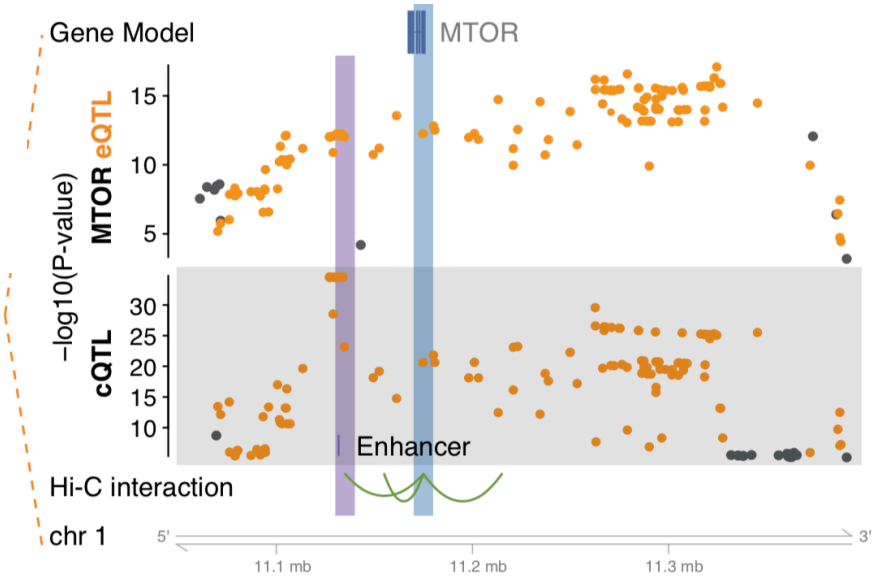


multi-QTLs

	Numbers of QTLs	eGenes Enhancers Cell types	SNPs
eQTL	2,542,908	32,944	1,341,182
isoQTL	2,628,259	19,790	1,052,939
cQTL*	8,464	8,484	7,983
fQTL	4,199	9	1,672

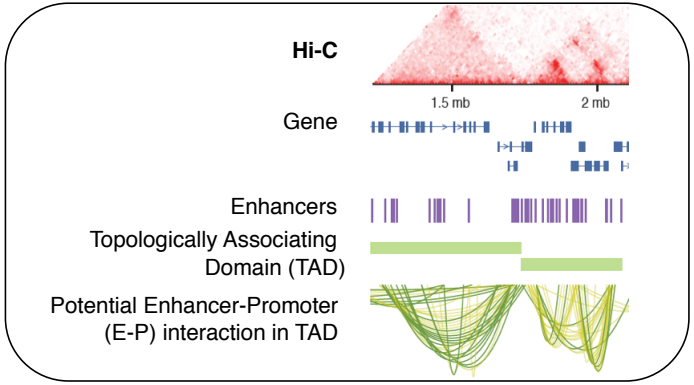


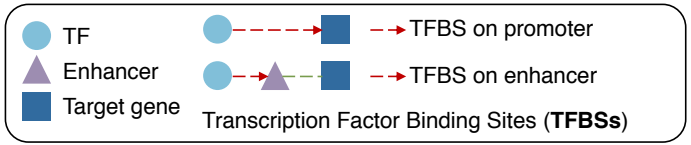
eQTLs and cQTLs significantly overlap

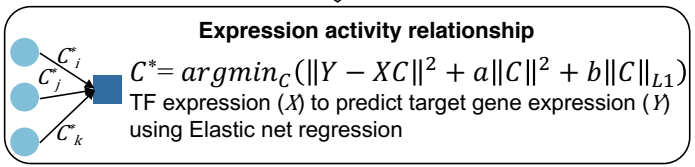


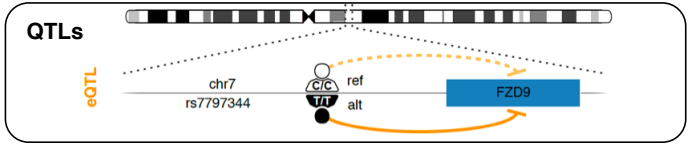
1391 SNPs (multi-QTLs) in at least three types among eQTLs, isoQTLs, cQTLs, fQTLs

Gene regulatory network inference

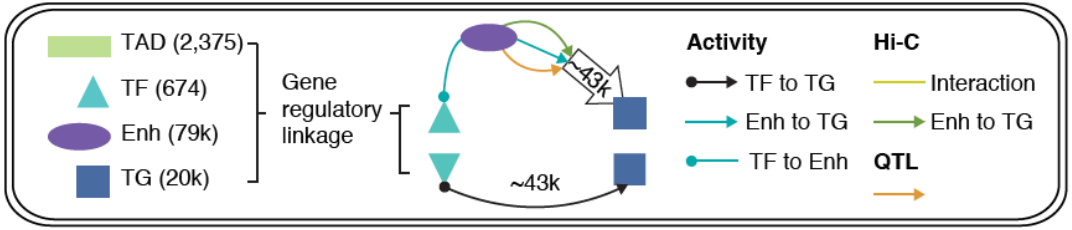


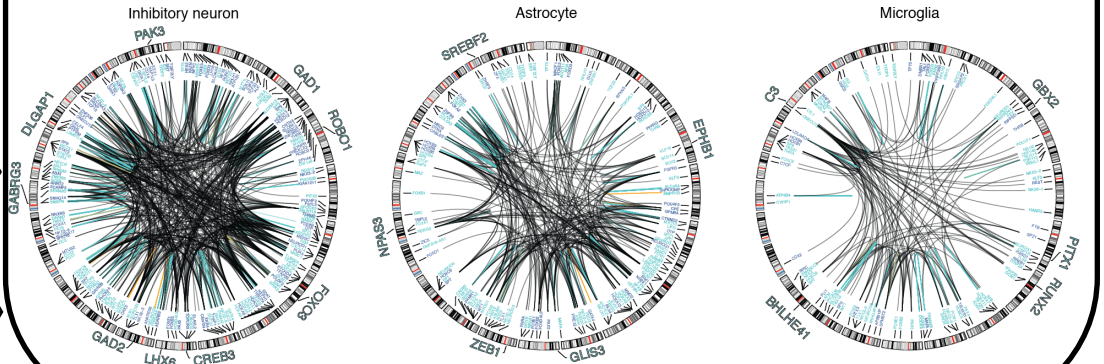






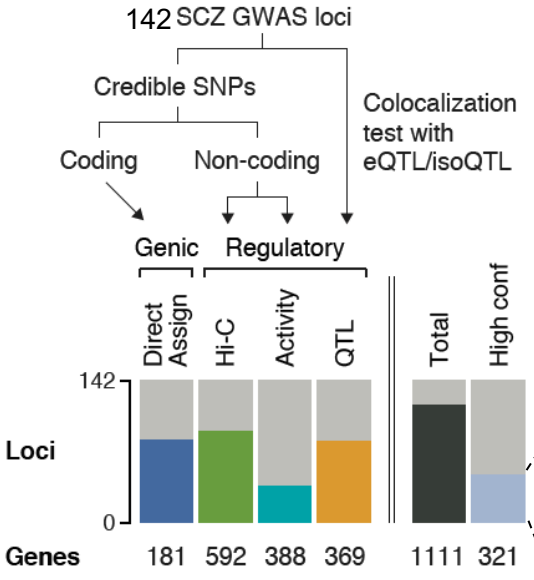
Imputed gene regulatory network linking TFs, enhancers and genes plus QTLs

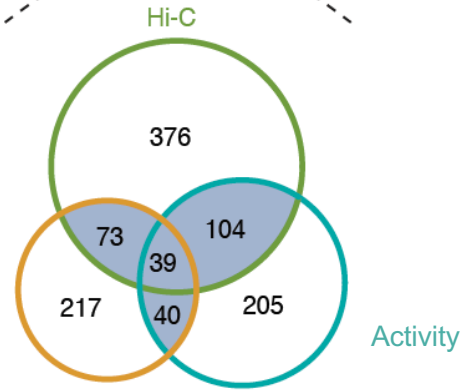




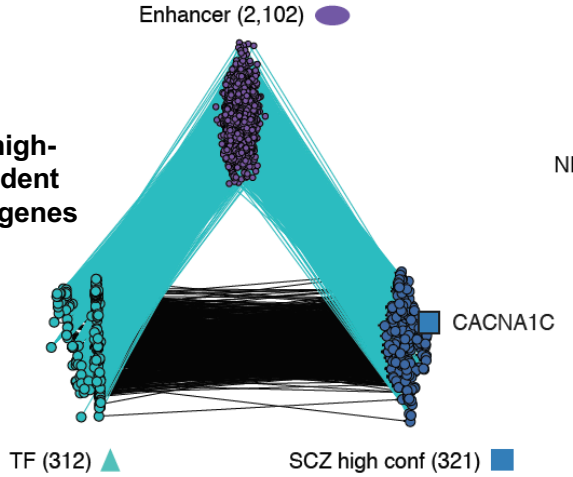
e.g., subnetworks targeting single cell marker genes

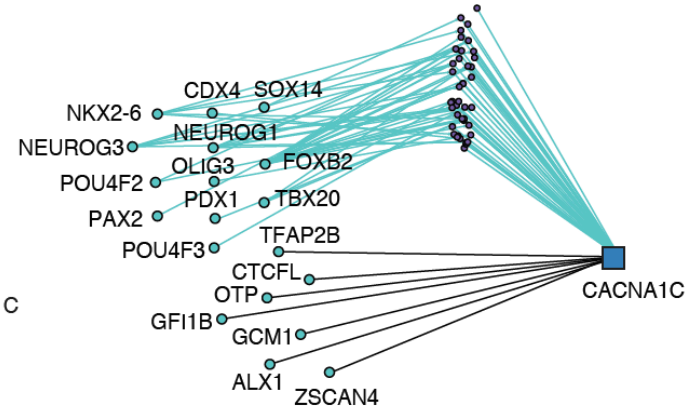
Linking GWAS SNPs to new disease genes using gene regulatory network





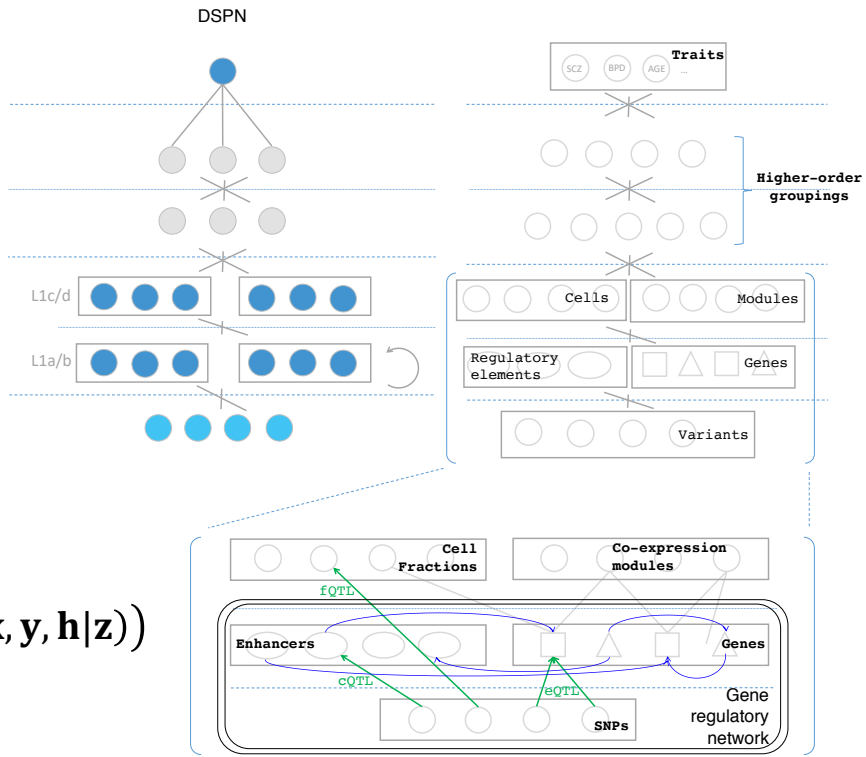
321 high-confident SCZ genes





Deep Structured Phenotype Network (DSPN)

Gene regulatory network builds DSPN skeleton



Boltzmann machine

y: phenotypes

h: hidden units (e.g., circuits)

x: intermediate phenotypes (e.g., genes, enhancers)

z: genotypes (e.g., SNPs)

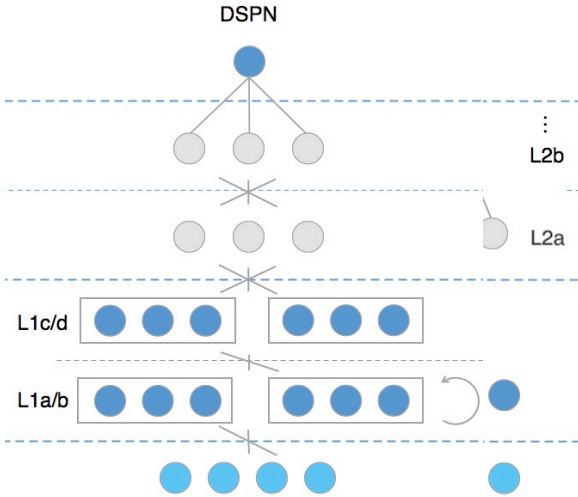
Energy model:

$p(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) \propto \exp\left(-E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z})\right)$

W: weights (e.g., regulatory network)

$$E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{W}_1\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W}_2\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W}_3\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W}_4\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W}_5\mathbf{y} - \boldsymbol{Bias}$$

DSPN improves brain disease prediction by adding deep layers

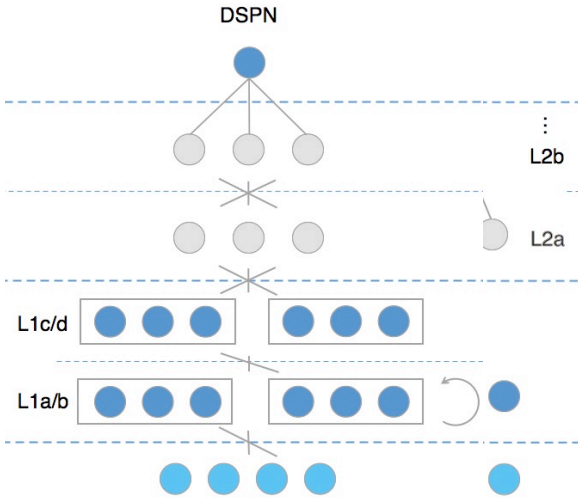


Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	59.0%	**73.6%**
Bipolar Disorder	**56.7%**	63.3%	71.1%	67.2%	**76.7%**
Autism Spectrum Disorder	**50.0%**	51.7%	67.2%	62.5%	**68.3%**

X 6.0

Accuracy = chance to correctly predict disease/health

DSPN improves brain disease prediction by adding deep layers

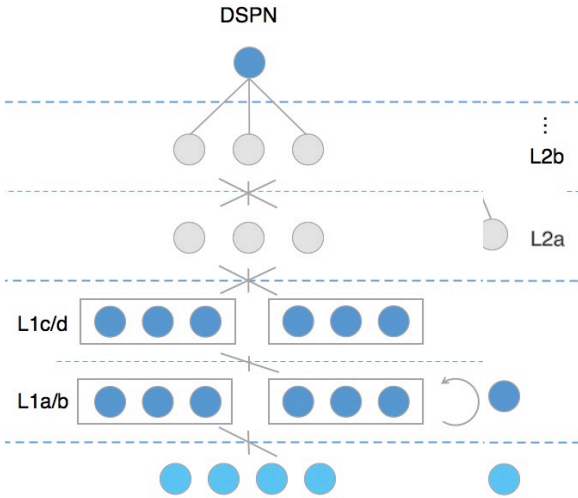


Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	54.6%	63.0%	70.0%	59.0%	73.6%
Bipolar Disorder	56.7%	63.3%	71.1%	67.2%	76.7%
Autism Spectrum Disorder	50.0%	51.7%	67.2%	62.5%	68.3%

X 2.5

Accuracy = chance to correctly predict disease/health

DSPN improves brain disease prediction by adding deep layers

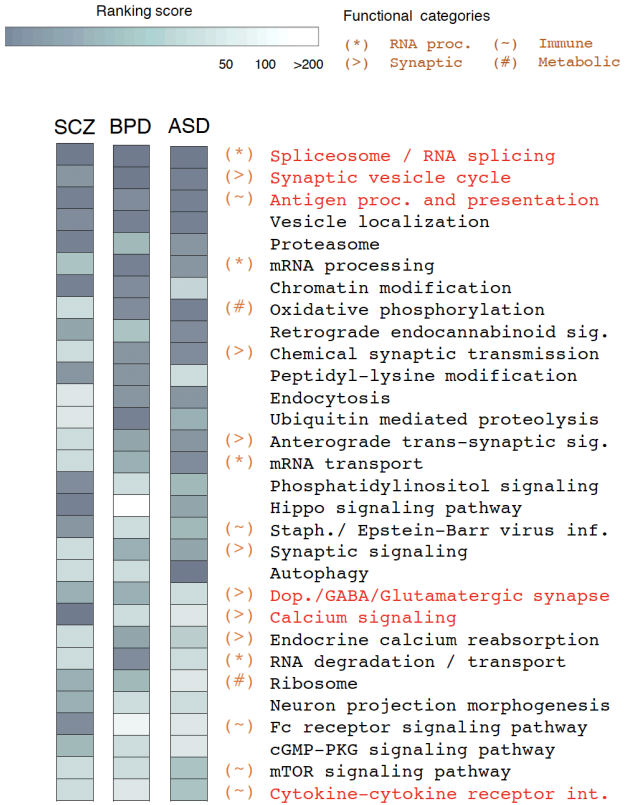


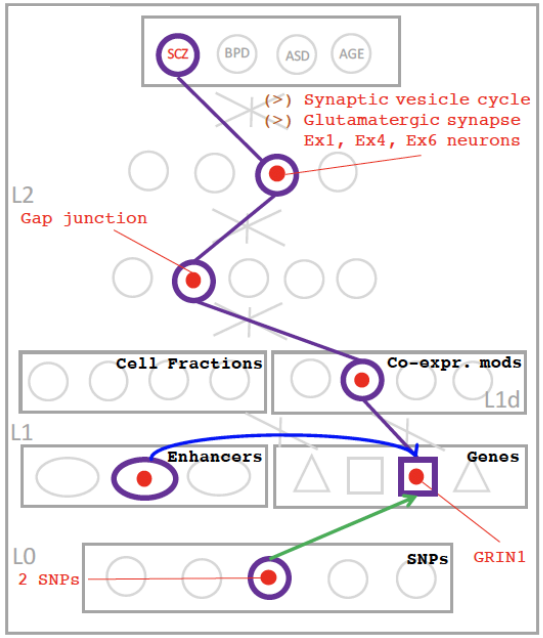
Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	**59.0%**	73.6%
Bipolar Disorder	**56.7%**	63.3%	71.1%	**67.2%**	76.7%
Autism Spectrum Disorder	**50.0%**	51.7%	67.2%	**62.5%**	68.3%

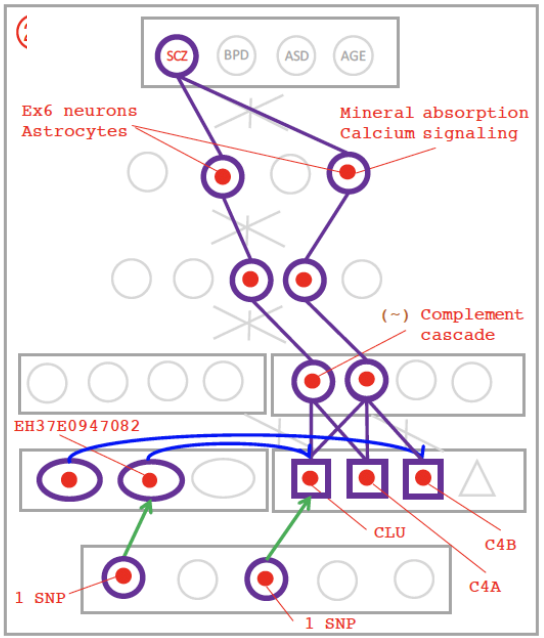
X 3.1

Accuracy = chance to correctly predict disease/health

DSPN discovers molecular pathways from genotype to phenotype







Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

Thoughts on Genome Annotation, Prioritizing Variants & Application of these concepts in a disease context

- **Background**
  - Types of variants: Germline, Somatic, &c
  - Types of annotations: peaks, segmentations, model predictions
  - Genomic covariates
- **Music**
  - Multi-scale peak calling
- **Matched Filter**
  - Integrating cross-assay signal-track patterns associated with enhancer
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR**
  - Adapts FunSeq approach to RNA
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORFs**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **PsychENCODE (Application)**
  - Population-level analysis of functional genomics data related to mental disease
  - Single-cell deconvolution explaining across-population variation
  - Large QTL resource (~2.5M eQTLs)
  - Regulatory network construction using QTLs, Hi-C, & activity relationships. Used to link GWAS SNPs to genes.
  - Embedding the reg. network in a deep-learning model (DSPN) to predict psychiatric disease phenotype from genotype and transcriptome data.

PsychENCODE Acknowledgment

- Geetha Senthil
- Lora Bingaman
- David Panchision
- Alexander Arguello
- Thomas Lehner

Daifeng Wang, Shuang Liu, Jonathan Warrell, Hyejung Won, Xu Shi, Fabio Navarro, Declan Clarke, Mengting Gu, Prashant Emani, Yucheng T. Yang, Min Xu, Michael Gandal, Shaoke Lou, Jing Zhang, Jonathan J. Park, Chengfei Yan, Suhn Kyong Rhie, Kasidet Manakongtreecheep, Holly Zhou, Aparna Nathan, Mette Peters, Eugenio Mattei, Dominic Fitzgerald, Tonya Brunetti, Jill Moore, Yan Jiang, Kiran Girdhar, Gabriel Hoffman, Selim Kalayci, Zeynep Hulya Gumus, Greg Crawford, **PsychENCODE Consortium**, Panos Roussos, Schahram Akbarian, Andrew E. Jaffe, Kevin White, Zhiping Weng, Nenad Sestan, **Daniel H. Geschwind, James A. Knowles**

Dedicated to **Pamela Sklar**

The PsychENCODE Consortium: **Allison E Ashley-Koch, Duke University; Gregory E Crawford, Duke University; Melanie E Garrett, Duke University; Lingyun Song, Duke University; Alexias Safi, Duke University; Graham D Johnson, Duke University; Gregory A Wray, Duke University; Timothy E Reddy, Duke University; Fernando S Goes, Johns Hopkins University; Peter Zandi, Johns Hopkins University; Julien Bryois, Karolinska Institutet; Andrew E Jaffe, Lieber Institute for Brain Development; Amanda J Price, Lieber Institute for Brain Development; Nikolay A Ivanov, Lieber Institute for Brain Development; Leonardo Collado-Torres, Lieber Institute for Brain Development; Thomas M Hyde, Lieber Institute for Brain Development; Emily E Burke, Lieber Institute for Brain Development; Joel E Kleiman, Lieber Institute for Brain Development; Ran Tao, Lieber Institute for Brain Development; Joo Heon Shin, Lieber Institute for Brain Development; Schahram Akbarian, Icahn School of Medicine at Mount Sinai; Kiran Girdhar, Icahn School of Medicine at Mount Sinai; Yan Jiang, Icahn School of Medicine at Mount Sinai; Marija Kundakovic, Icahn School of Medicine at Mount Sinai; Leanne Brown, Icahn School of Medicine at Mount Sinai; Bibi S Kassim, Icahn School of Medicine at Mount Sinai; Royce B Park, Icahn School of Medicine at Mount Sinai; Jennifer R Wiseman, Icahn School of Medicine at Mount Sinai; Elizabeth Zharovsky, Icahn School of Medicine at Mount Sinai; Rivka Jacobov, Icahn School of Medicine at Mount Sinai; Olivia Devillers, Icahn School of Medicine at Mount Sinai; Elie Flatow, Icahn School of Medicine at Mount Sinai; Gabriel E Hoffman, Icahn School of Medicine at Mount Sinai; Barbara K Lipska, Human Brain Collection Core, National Institutes of Health, Bethesda, MD; David A Lewis, University of Pittsburgh; Vahram Haroutunian, Icahn School of Medicine at Mount Sinai and James J Peters VA Medical Center; Chang-Gyu Hahn, University of Pennsylvania; Alexander W Charney, Mount Sinai; Stella Dracheva, Mount Sinai; Alexey Kozlenkov, Mount Sinai; Judson Belmont, Icahn School of Medicine at Mount Sinai; Diane DelValle, Icahn School of Medicine at Mount Sinai; Nancy Francoeur, Icahn School of Medicine at Mount Sinai; Evi Hadjimichael, Icahn School of Medicine at Mount Sinai; Dalila Pinto, Icahn School of Medicine at Mount Sinai; Harm van Bakel, Icahn School of Medicine at Mount Sinai; Panos Roussos, Mount Sinai; John F Fullard, Mount Sinai; Jaroslav Bendl, Mount Sinai; Mads E Hauberg, Mount Sinai; Lara M Mangravite, Sage Bionetworks; Mette A Peters, Sage Bionetworks; Yooree Chae, Sage Bionetworks; Junmin Peng, St. Jude Children's Hospital; Mingming Niu, St. Jude Children's Hospital; Xusheng Wang, St. Jude Children's Hospital; Maree J Webster, Stanley Medical Research Institute; Thomas G Beach, Banner Sun Health Research Institute; Chao Chen, Central South University; Yi Jiang, Central South University; Rujia Dai, Central South University; Annie W Shieh, SUNY Upstate Medical University; Chunyu Liu, SUNY Upstate Medical University; Kay S. Grennan, SUNY Upstate Medical University; Yan Xia, SUNY Upstate Medical University/Central South University; Ramu Vadukapuram, SUNY Upstate Medical University; Yongjun Wang, Central South University; Dominic Fitzgerald, The University of Chicago; Lijun Cheng, The University of Chicago; Miguel Brown, The University of Chicago; Mimi Brown, The University of Chicago; Tonya Brunetti, The University of Chicago; Thomas Goodman, The University of Chicago; Majd Alsayed, The University of Chicago; Michael J Gandal, University of California, Los Angeles; Daniel H Geschwind, University of California, Los Angeles; Hyejung Won, University of California, Los Angeles; Damon Polioudakis, University of California, Los Angeles; Brie Wamsley, University of California, Los Angeles; Jiani Yin, University of California, Los Angeles; Tarik Hadzic, University of California, Los Angeles; Luis De La Torre Ubieta, UCLA; Vivek Swarup, University of California, Los Angeles; Stephan J Sanders, University of California, San Francisco; Matthew W State, University of California, San Francisco; Donna M Werling, University of California, San Francisco; Joon-Yong An, University of California, San Francisco; Brooke Sheppard, University of California, San Francisco; A Jeremy Willsey, University of California, San Francisco; Kevin P White, The University of Chicago; Mohana Ray, The University of Chicago; Gina Giase, SUNY Upstate Medical University; Amira Kefi, University of Illinois at Chicago; Eugenio Mattei, University of Massachusetts Medical School; Michael Purcaro, University of Massachusetts Medical School; Zhiping Weng, University of Massachusetts Medical School; Jill Moore, University of Massachusetts Medical School; Henry Pratt, University of Massachusetts Medical School; Jack Huey, University of Massachusetts Medical School; Tyler Borrman, University of Massachusetts Medical School; Patrick F Sullivan, University of North Carolina - Chapel Hill; Paola Giusti-Rodriguez, University of North Carolina - Chapel Hill; Yunjung Kim, University of North Carolina - Chapel Hill; Patrick Sullivan, University of North Carolina - Chapel Hill; Jin Szatkiewicz, University of North Carolina - Chapel Hill; Suhn Kyong Rhie, University of Southern California; Christoper Armoskus, University of Southern California; Adrian Camarena, University of Southern California; Peggy J Farnham, University of Southern California; Valeria N Spitsyna, University of Southern California; Heather Witt, University of Southern California; Shannon Schreiner, University of Southern California; Oleg V Evgrafov, SUNY Downstate Medical Center; James A Knowles, SUNY Downstate Medical Center; Mark Gerstein, Yale University; Shuang Liu, Yale University; Daifeng Wang, Stony Brook University; Fabio C. P. Navarro, Yale University; Jonathan Warrell, Yale University; Declan Clarke, Yale University; Prashant S. Emani, Yale University; Mengting Gu, Yale University; Xu Shi, Yale University; Min Xu, Yale University; Yucheng T. Yang, Yale University; Robert R. Kitchen, Yale University; Gamze Gürsoy, Yale University; Jing Zhang, Yale University; Becky C Carlyle, Yale University; Angus C Nairn, Yale University; Mingfeng Li, Yale University; Sirisha Pochareddy, Yale University; Nenad Sestan, Yale University; Mario Skarica, Yale University; Zhen Li, Yale University; Andre M.M. Sousa, Yale University; Gabriel Santpere, Yale University; Jinmyung Choi, Yale University; Ying Zhu, Yale University; Tianliuyun Gao, Yale University; Daniel J Miller, Yale University; Adriana Cherskov, Yale University; Mo Yang, Yale University; Anahita Amiri, Yale University; Gianfilippo Coppola, Yale University; Jessica Mariani, Yale University; Soraya Scuderi, Yale University; Anna Szekely, Yale University; Flora M Vaccarino, Yale University; Feinan Wu, Yale University; Sherman Weissman, Yale University; Tanmoy Roychowdhury, Mayo Clinic Rochester; Alexej Abyzov, Mayo Clinic Rochester;.**

Acknowledgments! Also, Hiring: See **Jobs**.gersteinlab.org

MUSIC.gersteinlab.org
A **Harmanci**, J Rozowsky

github.com/gersteinlab/**MatchedFilter**
A **Sethi**, M **Gu**, E Gumusgoz, L Chan, KK Yan, J Rozowsky,
I Barozzi, V Afzal, J Akiyama, I Plajzer-Frick, C Yan, C Pickle,
M Kato, T Garvin, Q Pham, A Harrington, B Mannion, E Lee,
Y Fukuda-Yuzawa, A Visel, D E. Dickel, K Yip, R Sutton, LA Pennacchio

FunSeq.gersteinlab.org
Y **Fu**, E **Khurana**, Z Liu, S Lou, J Bedford, X Mu, K Yip

RADAR.gersteinlab.org
J **Zhang**, J **Liu**, D Lee, J-J Feng, L Lochovsky, S Lou, M Rutenberg-Schoenberg

github.gersteinlab.org/**uORFs**
P **McGillivray**, R Ault, M Pawashe, R Kitchen, S Balasubramanian

MOAT.gersteinlab.org
L **Lochovsky**, J Zhang

LARVA.gersteinlab.org
L **Lochovsky**, J Zhang, Y Fu, E Khurana


STOP

Info about this talk

No Conflicts

Unless explicitly listed here. There are no conflicts of interest relevant to the material in this talk

General PERMISSIONS

- This Presentation is copyright Mark Gerstein, Yale University, 2017.
- Please read permissions statement at

  sites.gersteinlab.org/Permissions
- Basically, feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or website link). Paper references in the talk were mostly from Papers.GersteinLab.org.

PHOTOS & IMAGES

For thoughts on the source and permissions of many of the photos and clipped images in this presentation see streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: flickr.com/photos/mbgmbg/tags/kwpotppt