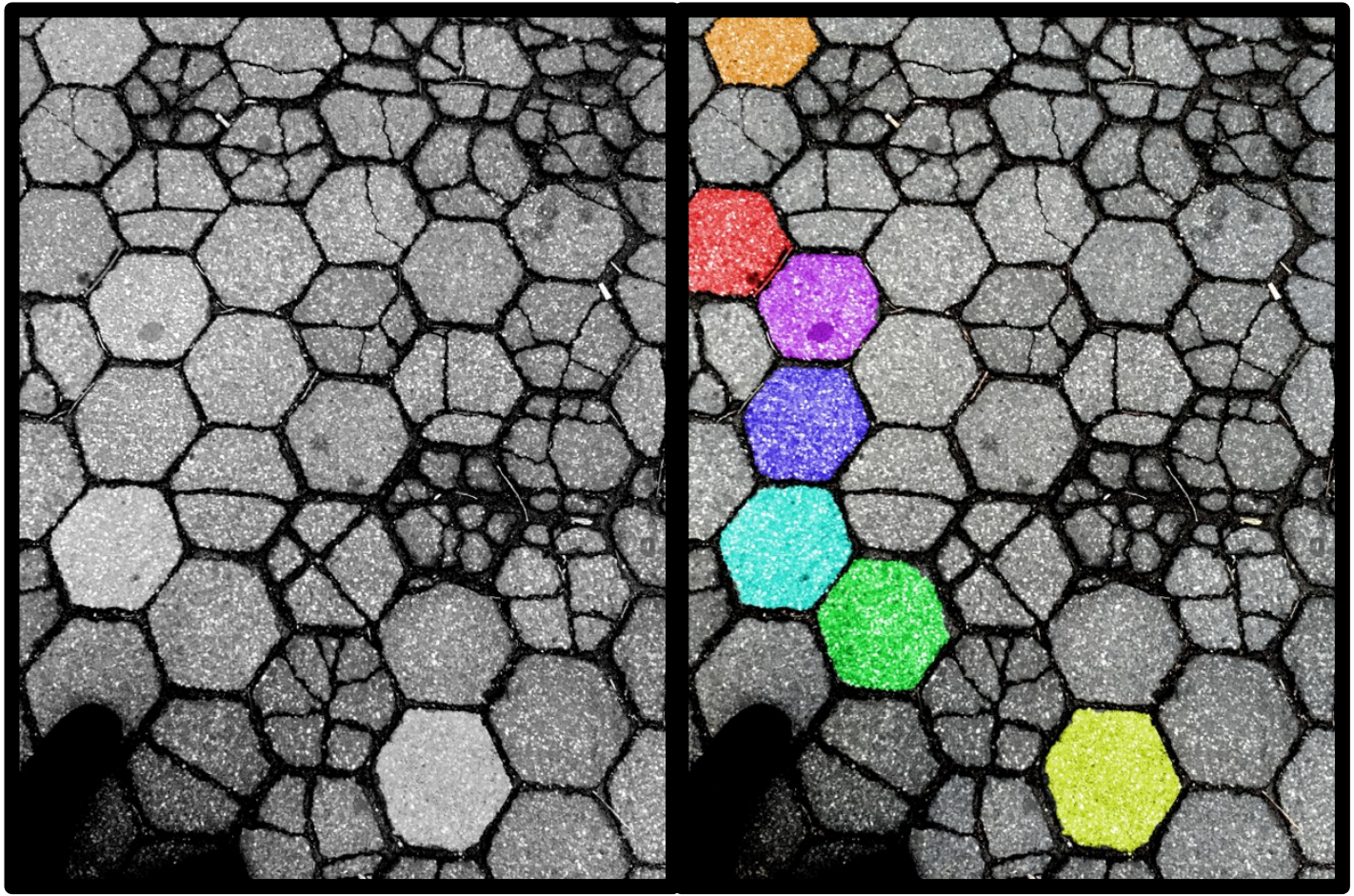
RADAR: Annotation & prioritization of variants in the post-transcriptional regulome for RBPs



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.

RADAR: Annotation & prioritization of variants in the post-transcriptional regulome for RBPs

- Background on prioritizing non-coding variants:
  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

- Next step in prioritizing variants associated with RNA:
  uORFs - Feature integration to find small subset of upstream mutations that potentially alter translation

RADAR: Annotation & prioritization of variants in the post-transcriptional regulome for RBPs

- Background on prioritizing non-coding variants:
  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

- Next step in prioritizing variants associated with RNA:
  uORFs - Feature integration to find small subset of upstream mutations that potentially alter translation

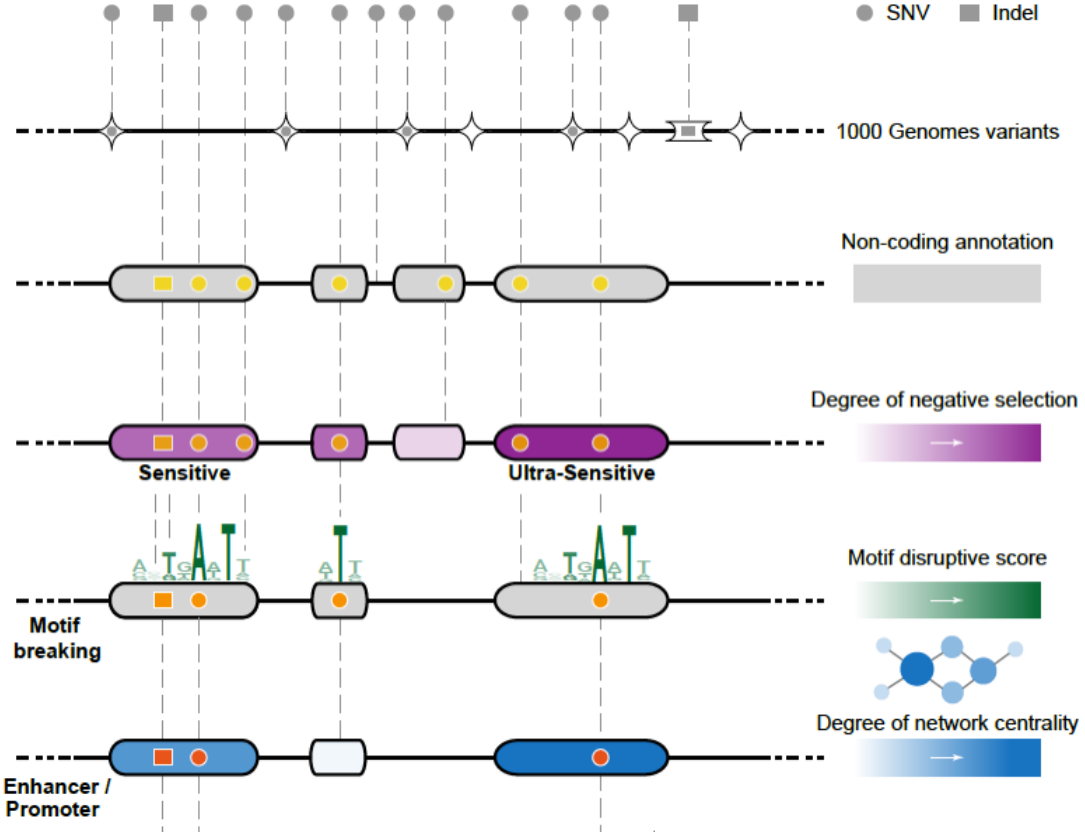
Funseq: a flexible framework to determine functional impact & use this to prioritize variants

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

Conservation (GERP, allele freq.)

Mutational impact (motif breaking, Lof)

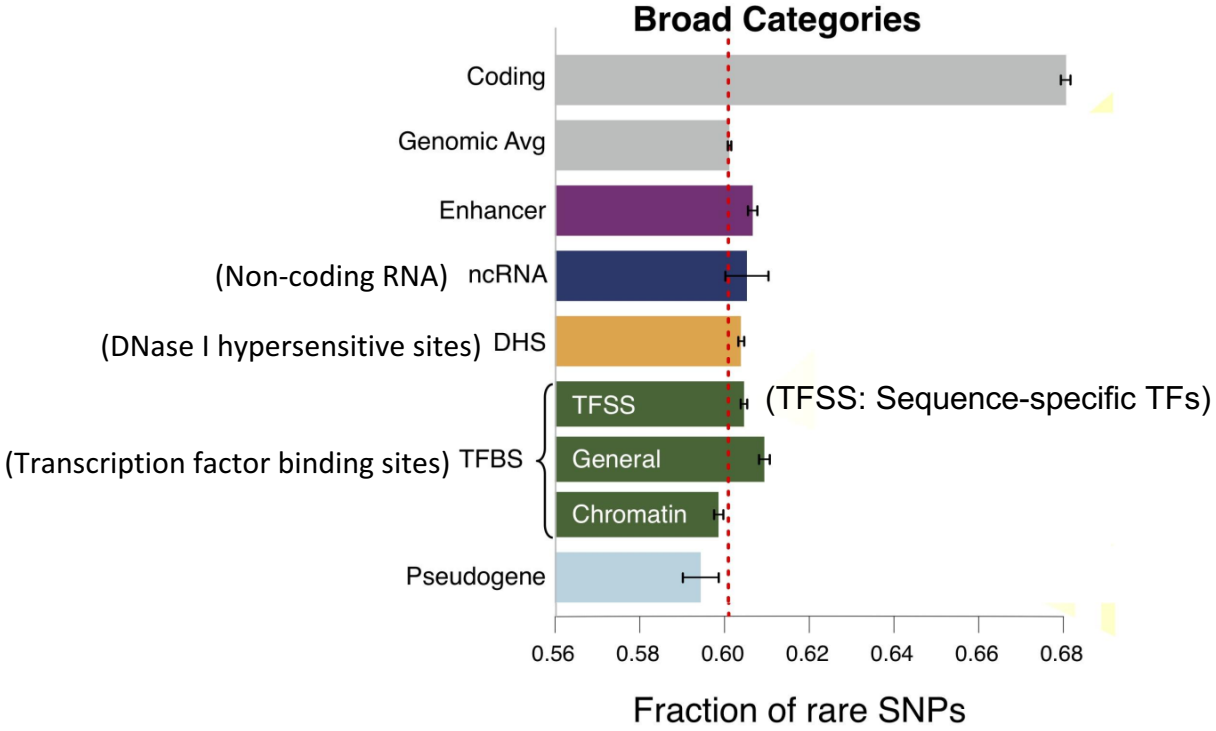
Network (centrality position)



[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

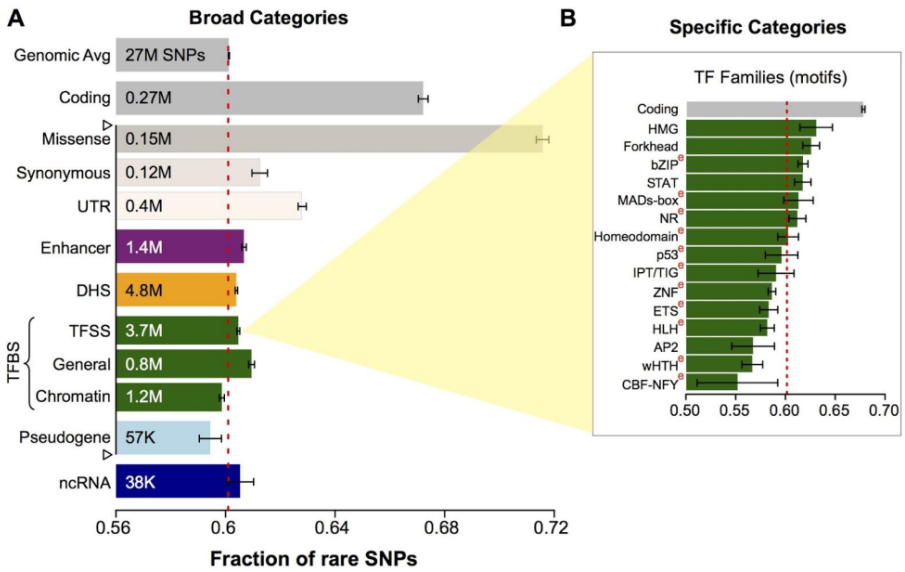


Depletion of Common Variants in the Human Population

Broad categories of regulatory regions under negative selection Related to:

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

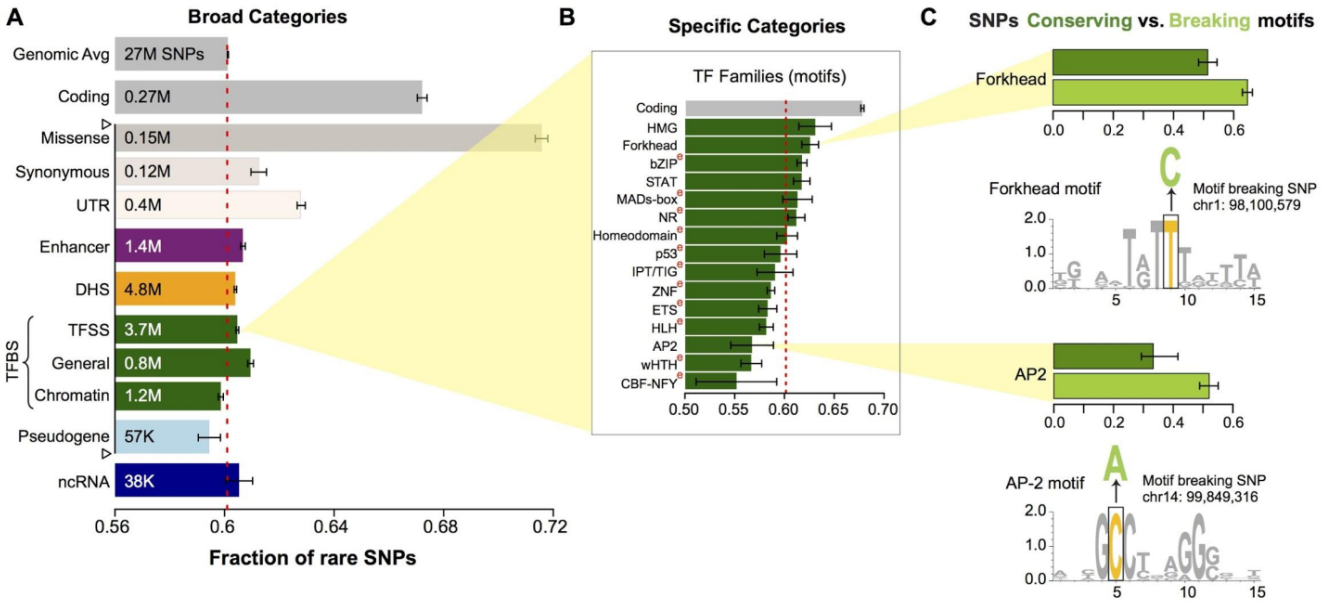
Differential selective constraints among specific sub-categories



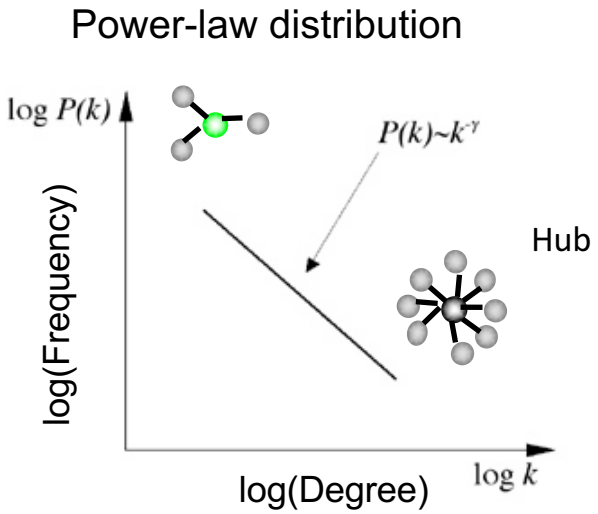
Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

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

SNPs which break TF motifs are under stronger selection



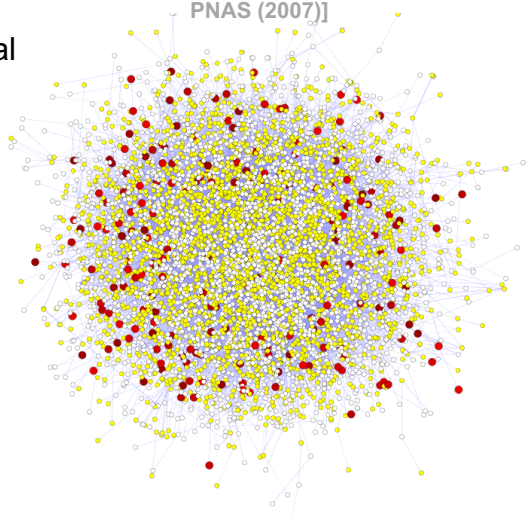
[Khurana et al., *Science* ('13)]



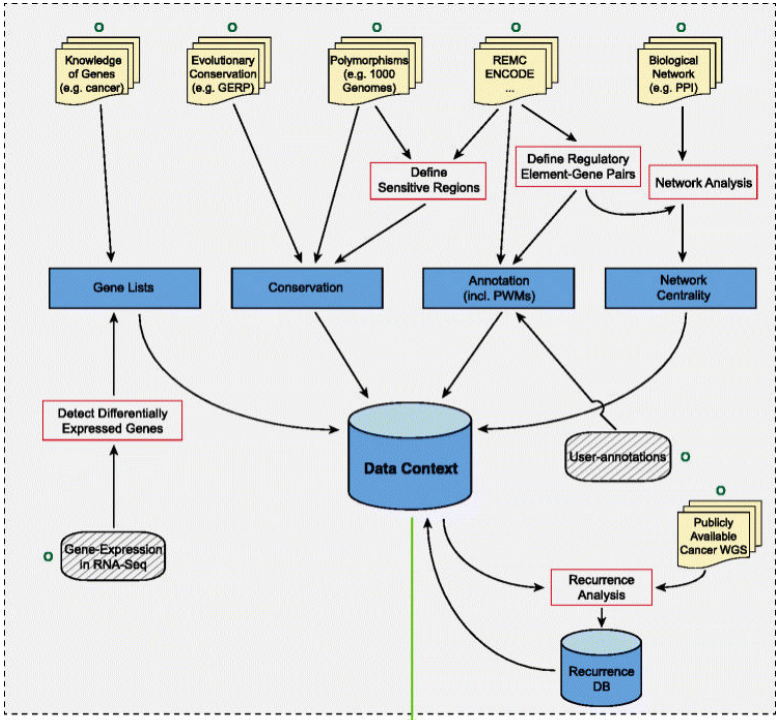
Hubs Under Constraint: A Finding from the Network Biology Community

- 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

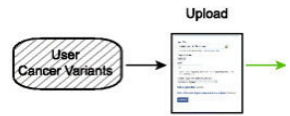
HOT region
Sensitive region
Polymorphisms

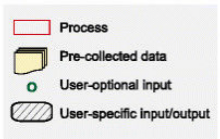
Genome

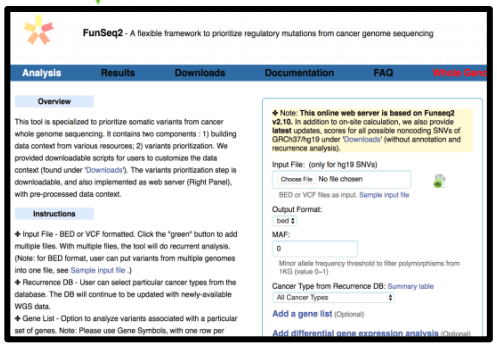
$$w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$$

- 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)]

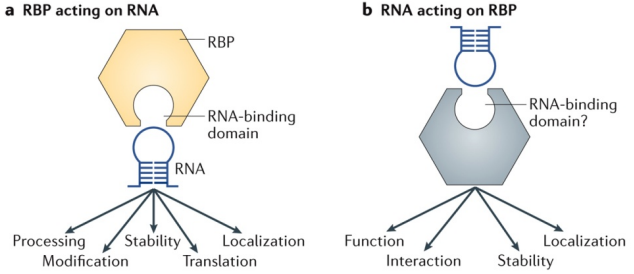
RADAR: Annotation & prioritization of variants in the post-transcriptional regulome for RBPs

- Background on prioritizing non-coding variants:
  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

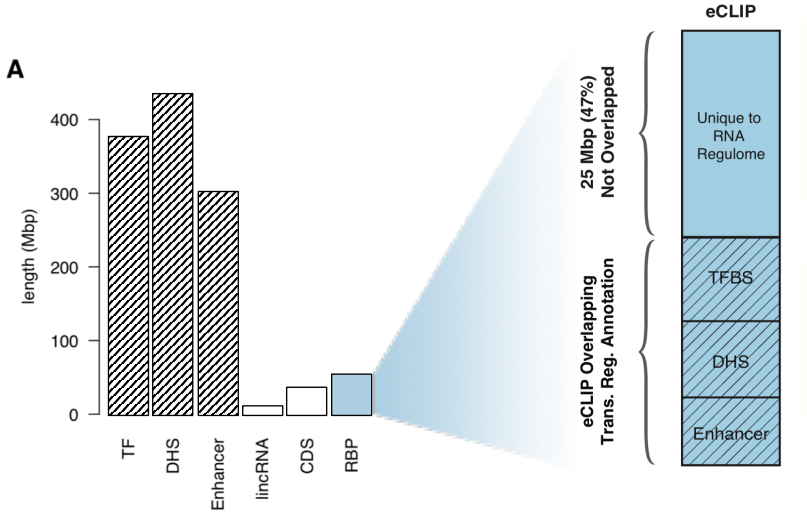
- Next step in prioritizing variants associated with RNA:
  uORFs - Feature integration to find small subset of upstream mutations that potentially alter translation

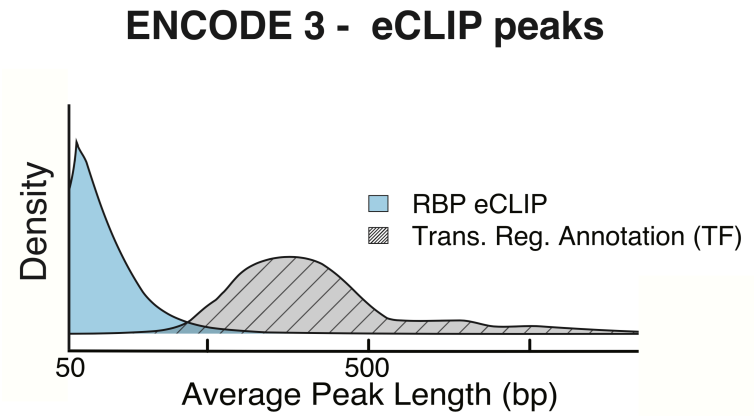
RNA Binding Proteins (RBPs)



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

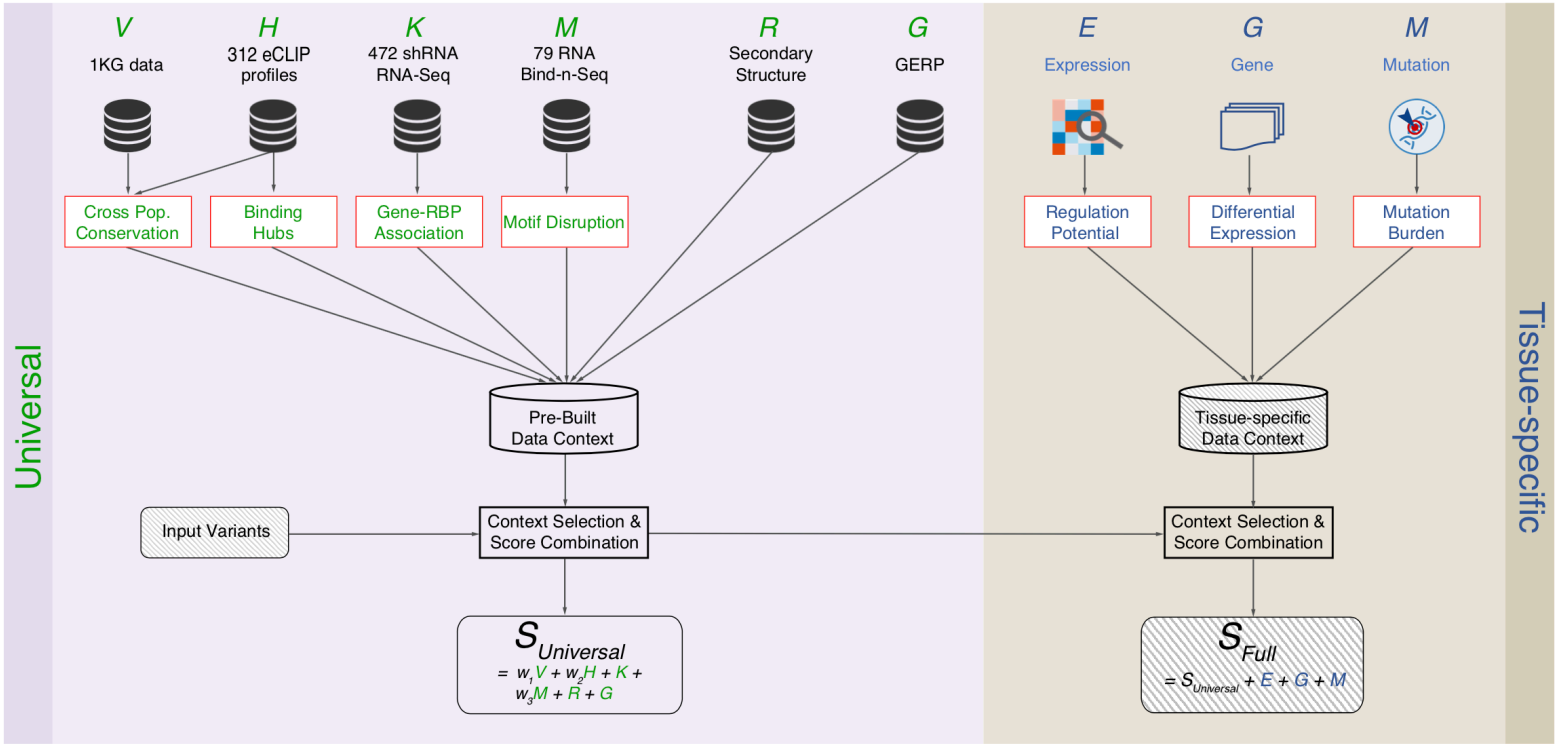
- **Before ENCODE3: >150 expt.**
  in many different cell types
- **ENCODE3 did ~350 focused eCLIP expt.**
  for >110 RBPs on HepG2 & K562
  (Van Nostrand...Yeo. Nat. Meth. '16;
  Van Nostrand...Graveley, Yeo
  (submitted in relation to ENCODE3))

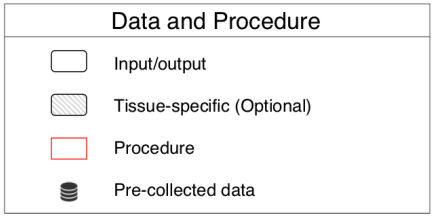


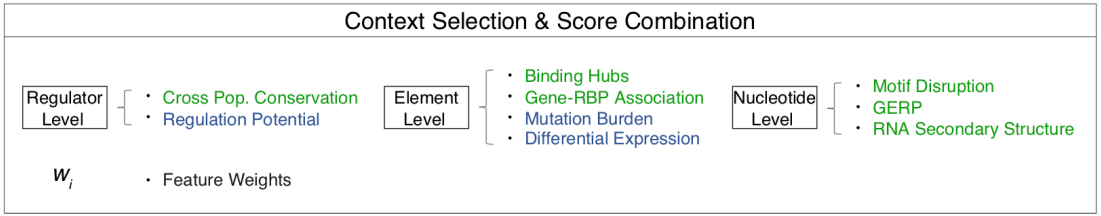


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

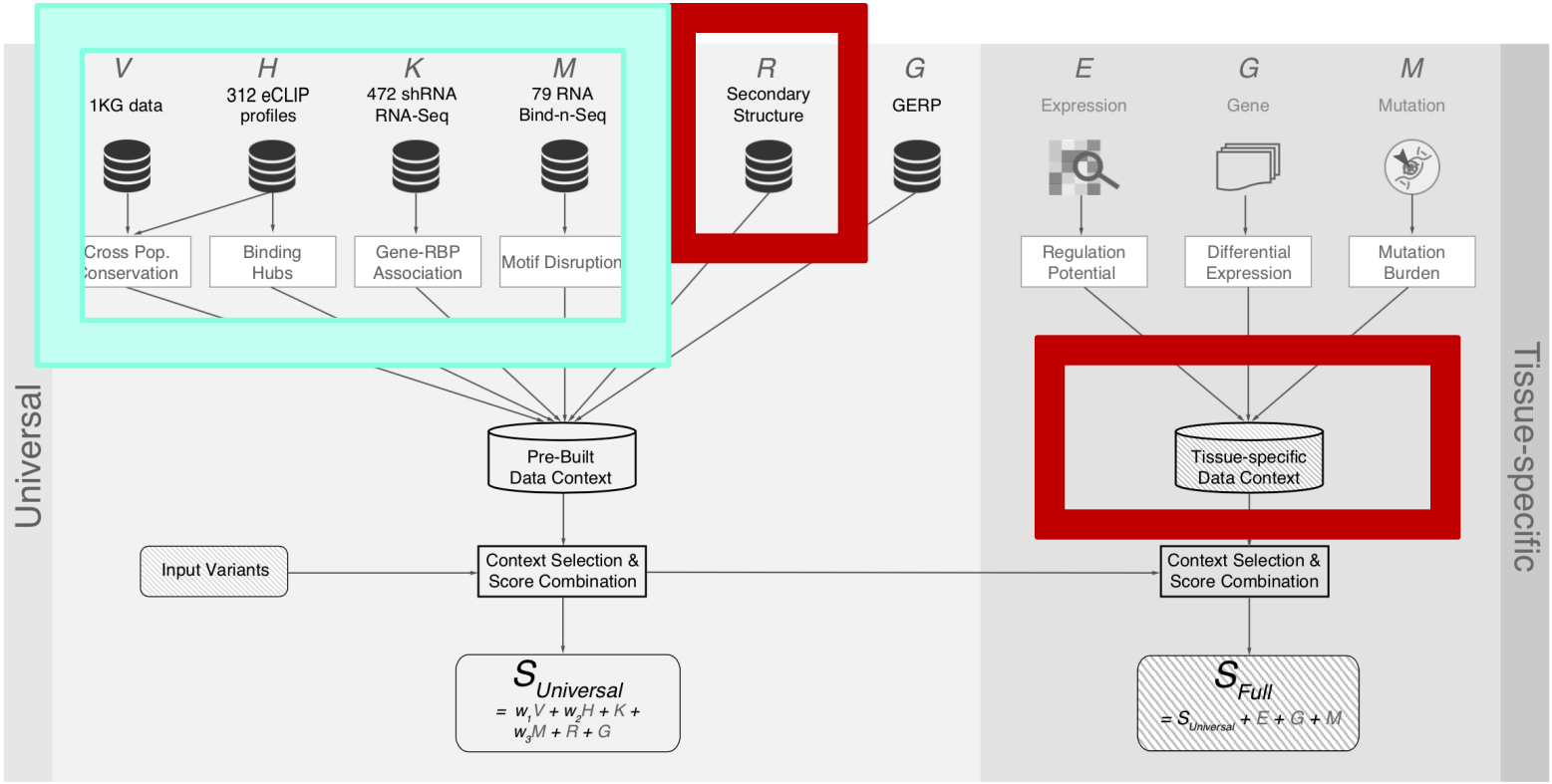
Schematic of RADAR Scoring

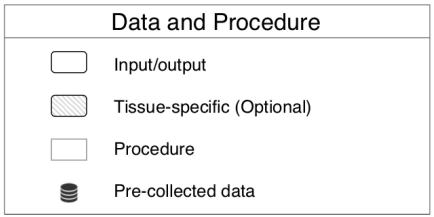


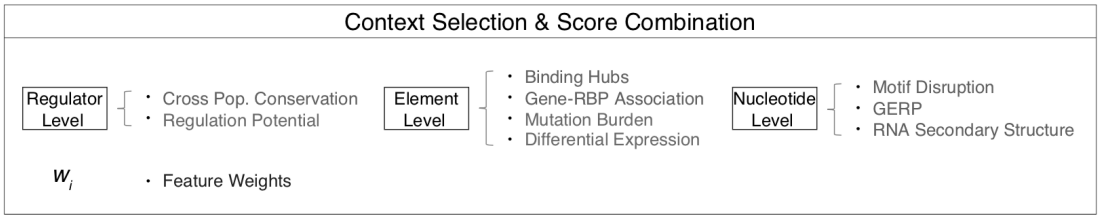




[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

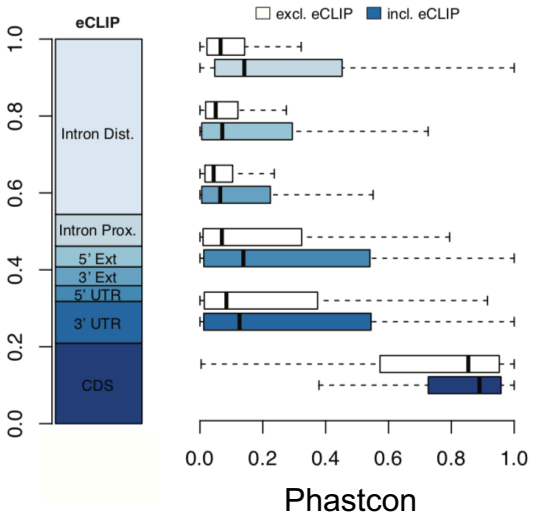




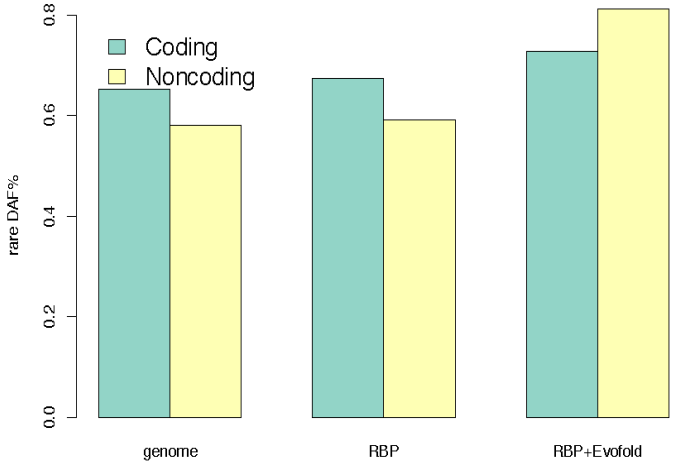


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

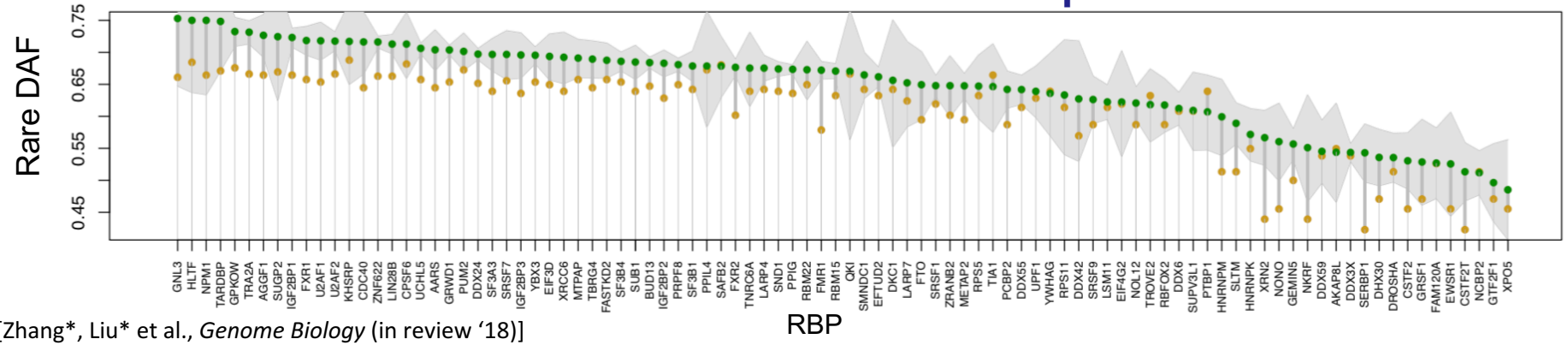
High Phastcon in RBP-overlapped annotations



RNA Structure Cons. from Evofold



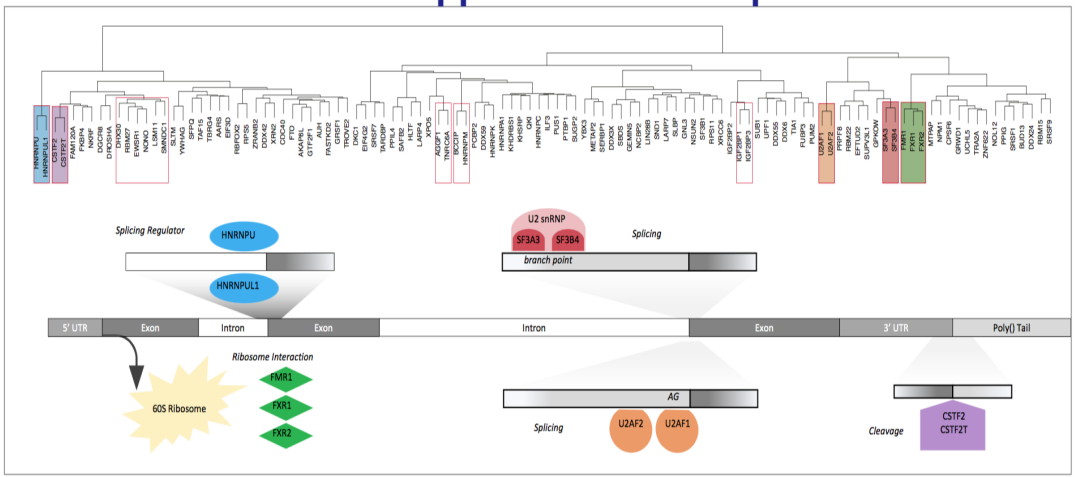
Enriched rare DAF in eCLIP peaks



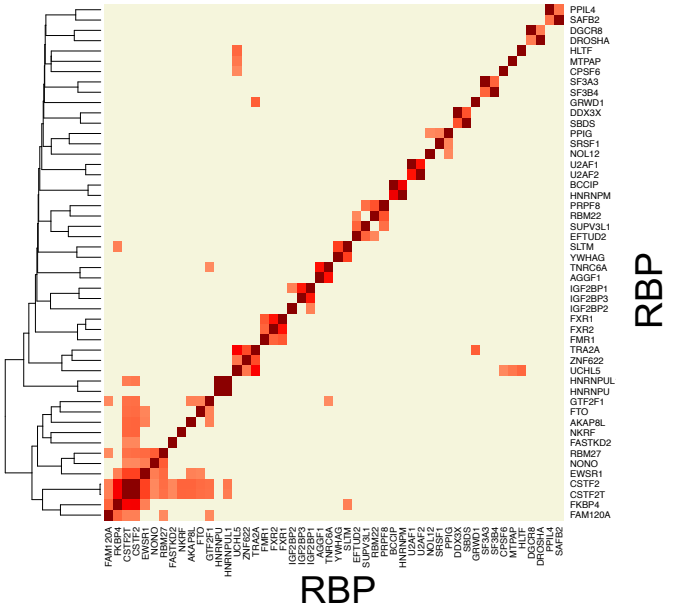
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Co-binding of RBPs form biologically relevant complexes

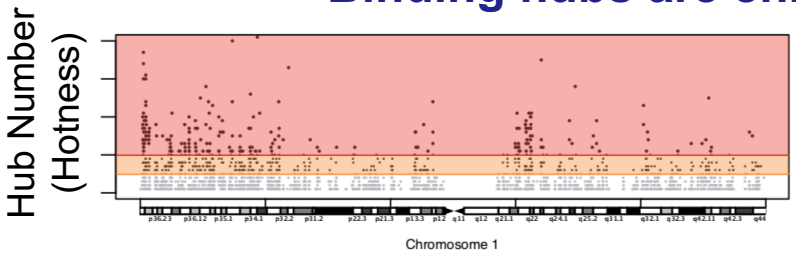
Literature supported RBP complexes

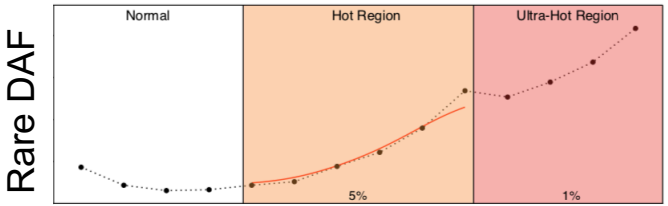


Unique co-binding patterns of RBPs

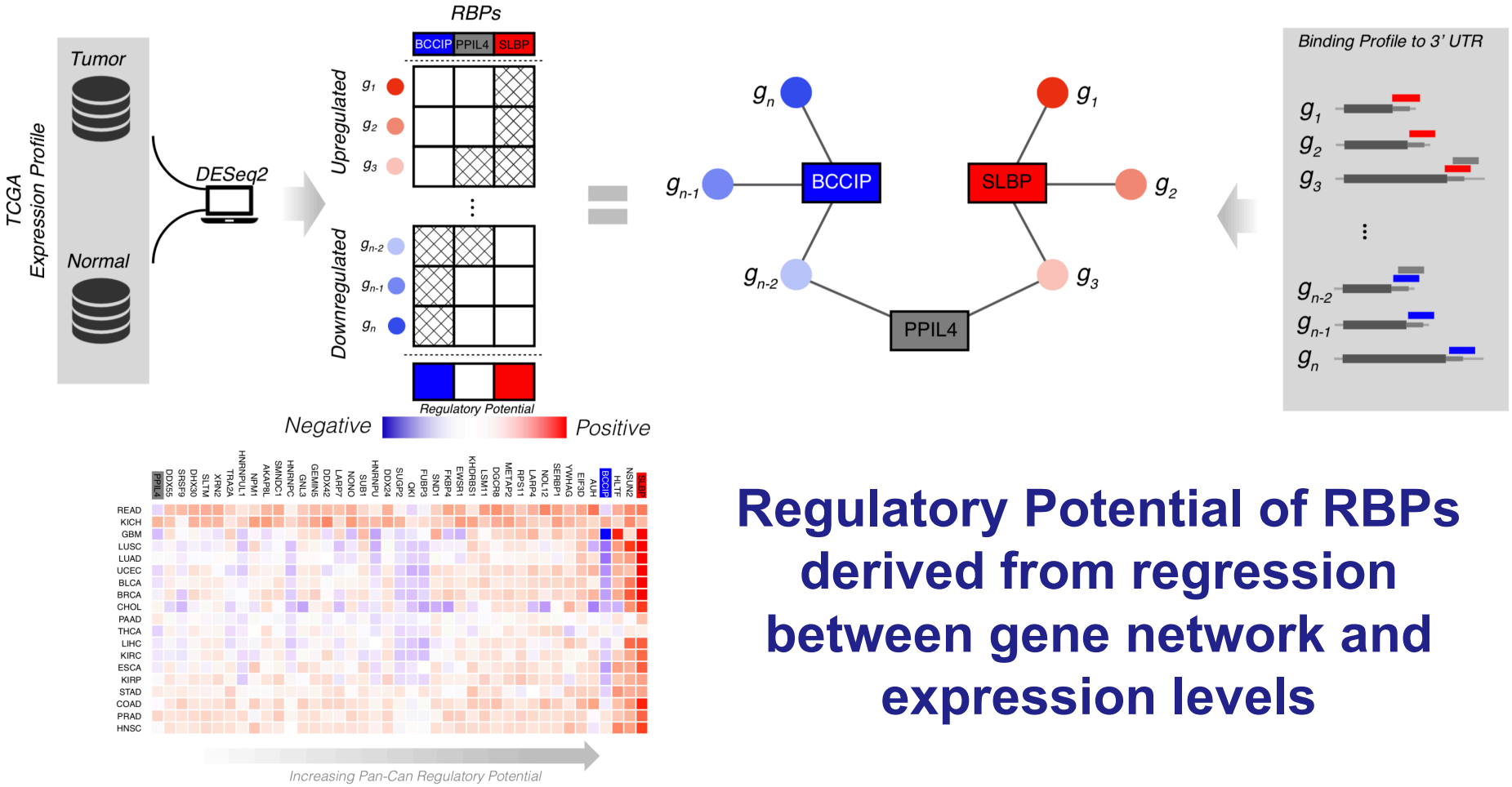


Binding hubs are enriched for rare variants





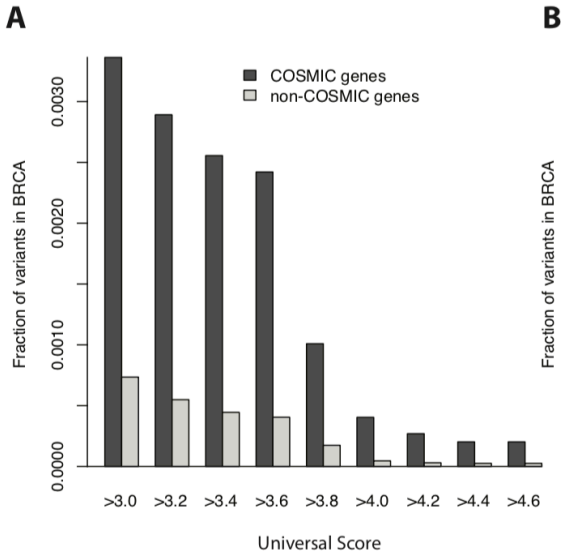
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

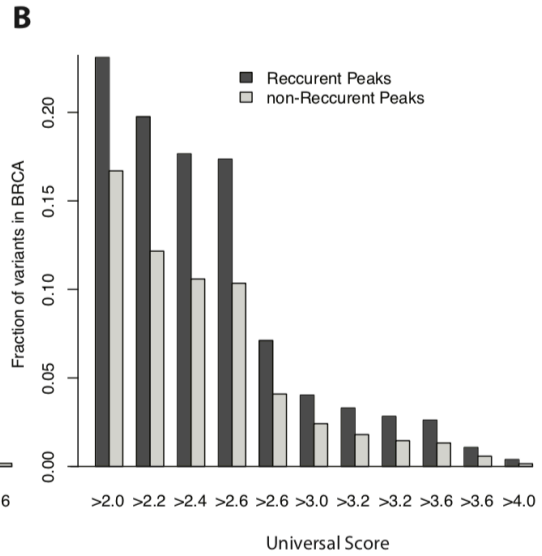


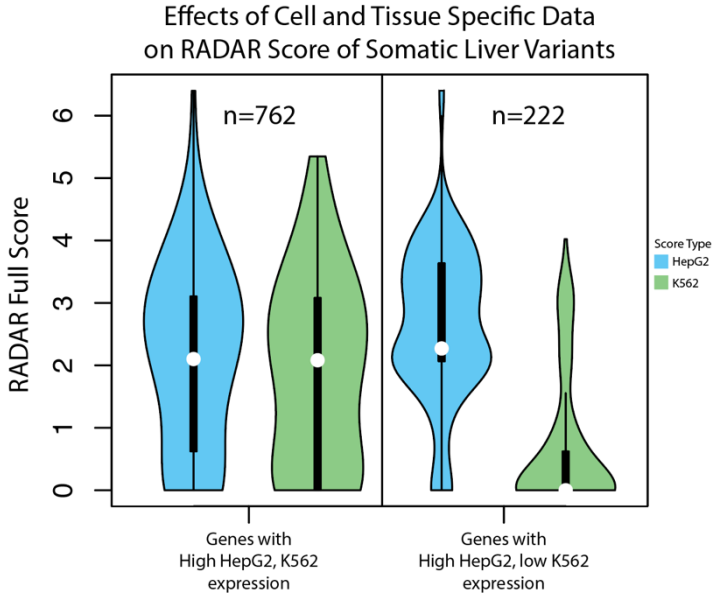
Regulatory Potential of RBPs derived from regression between gene network and expression levels

[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

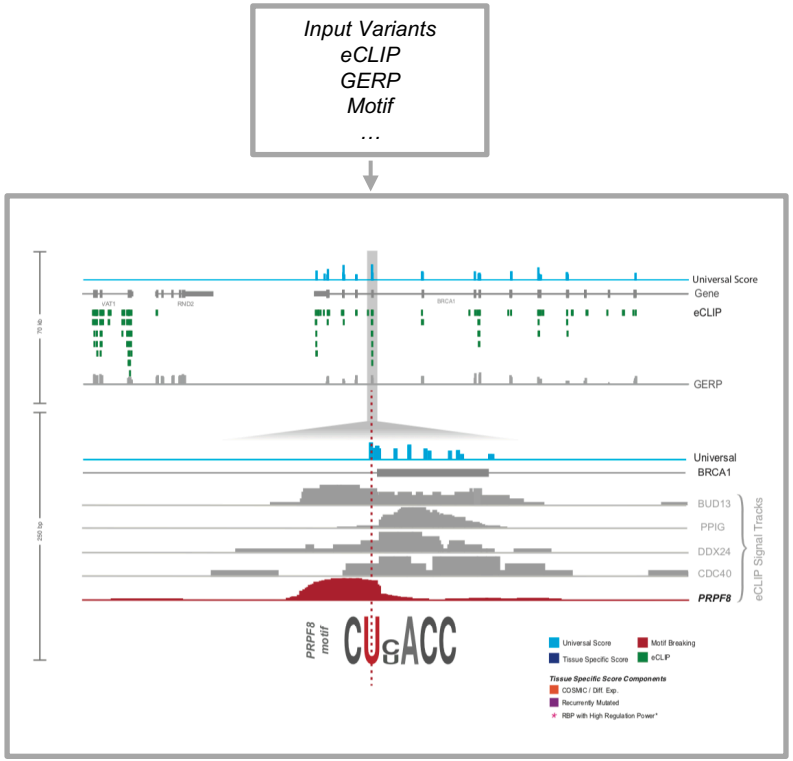






Visualization of RADAR Features and Scoring

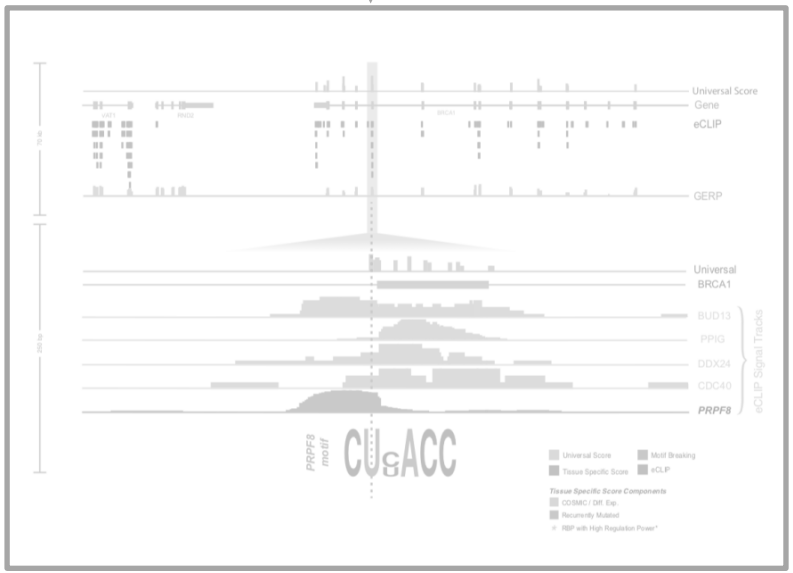
Germline Variants are Score Using a Universal Scoring Scheme



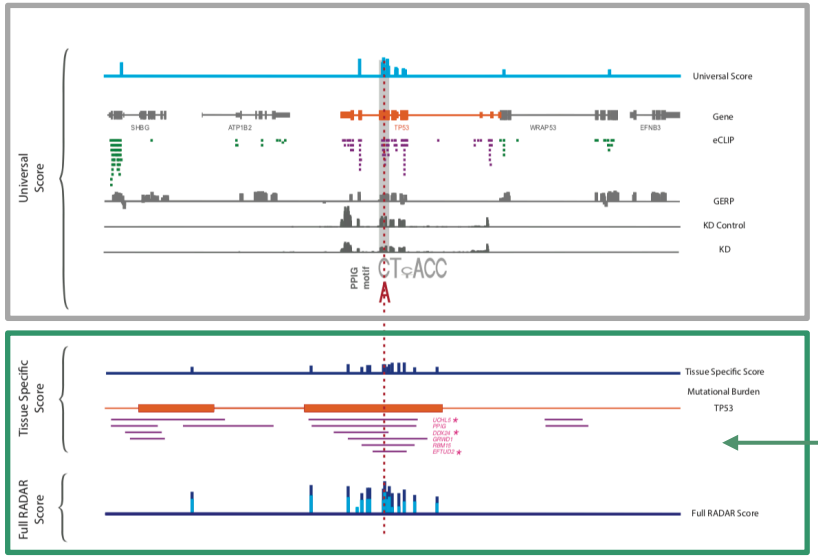
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Visualization of RADAR Features and Scoring

Input Variants
eCLIP
GERP
Motif
…



Tissue Specific:
Variants
Expression
Regulatory Potential



Somatic Variant Scored with Universal + Tissue specific context score

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

RADAR: Annotation & prioritization of variants in the post-transcriptional regulome for RBPs

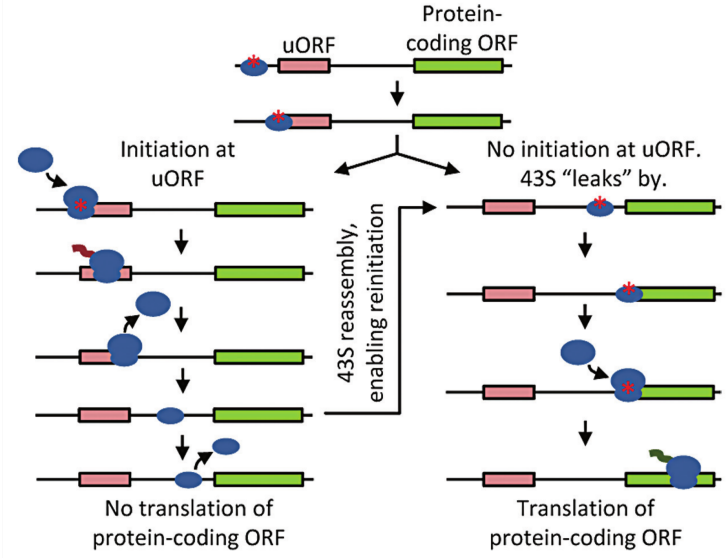
- Background on prioritizing non-coding variants:
  FunSeq integrates evidence, with a “surprisal” based weighting scheme. Prioritizing variants within “sensitive sites” (human conserved)

- **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

- Next step in prioritizing variants associated with RNA:
  uORFs - Feature integration to find small subset of upstream mutations that potentially alter translation

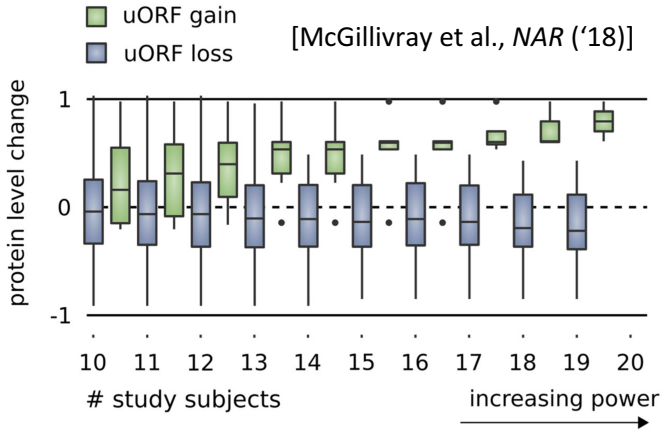
Upstream open reading frames (uORFs) regulate translation are affected by somatic mutation

[Calvo et al., *PNAS* (‘09)]

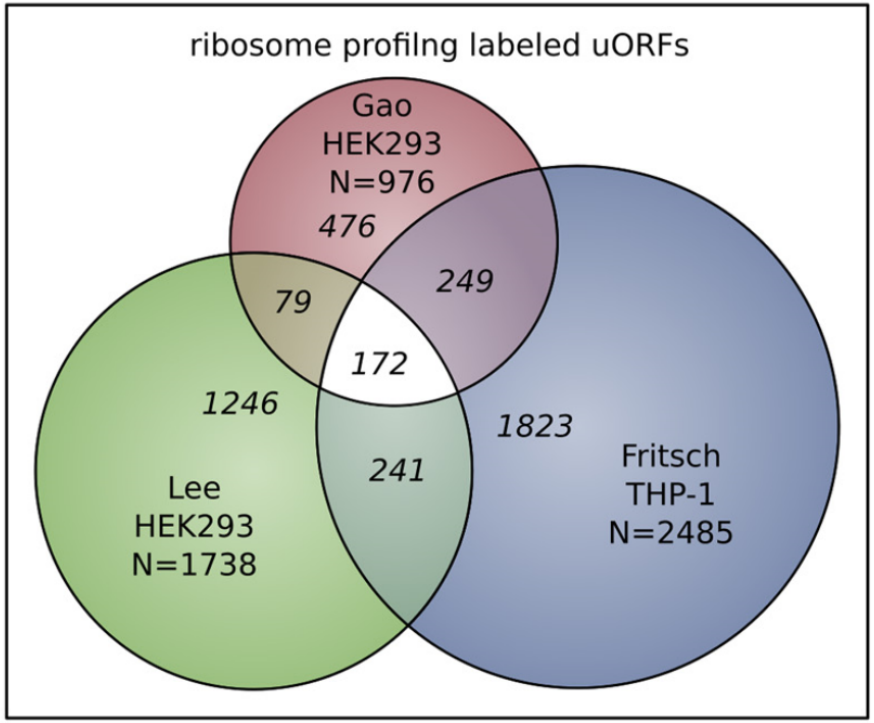


[Ferreira et al., *Bioengineered* (‘14)]

- 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.

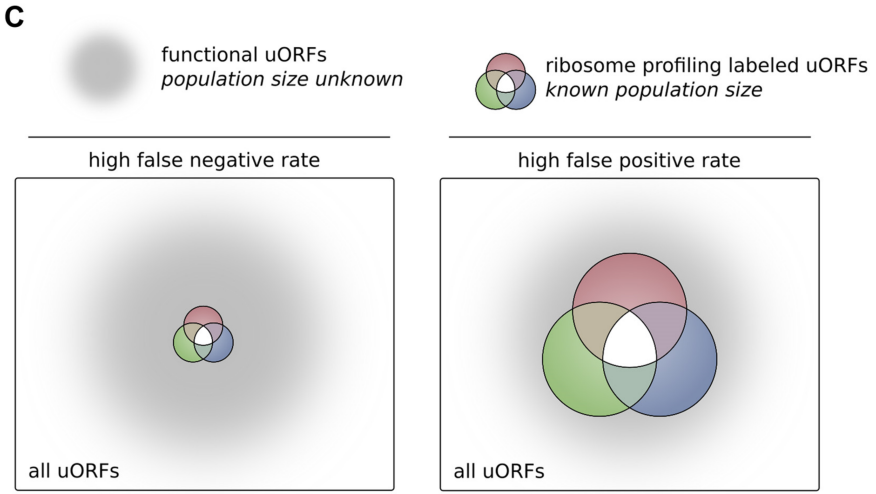


[McGillivray et al., *NAR* (‘18)]



From a “Universe” of 1.3 M pot. uORFs

The population of functional uORFs may be significant

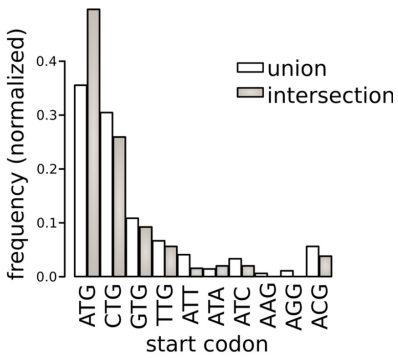


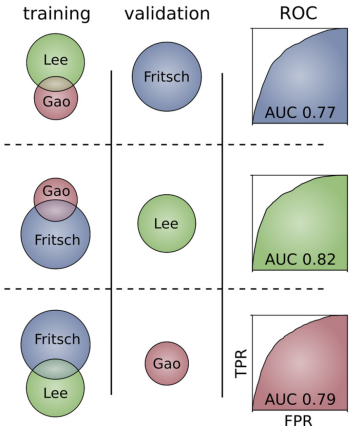
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

[McGillivray et al., *NAR* (‘18)]

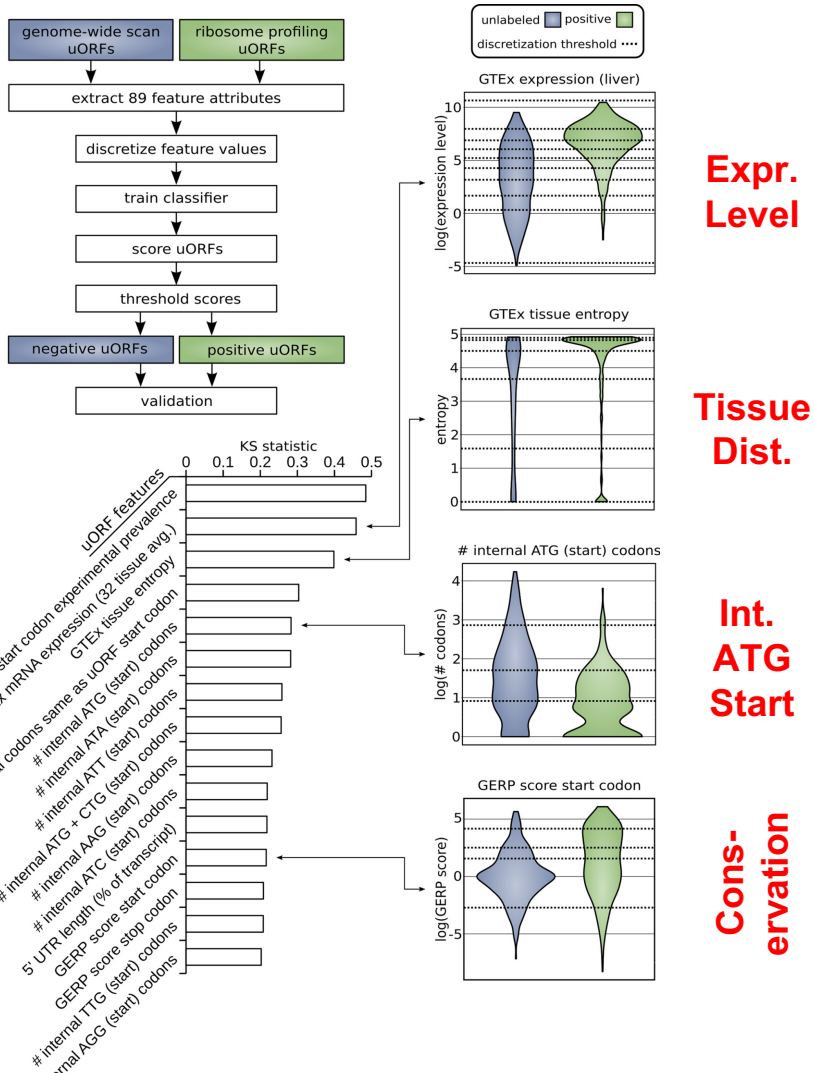
Prediction & validation of functional uORFs using 89 features

- 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)]



Expr. Level

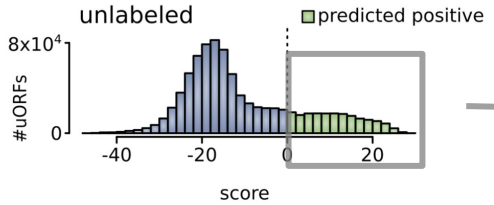
Tissue Dist.

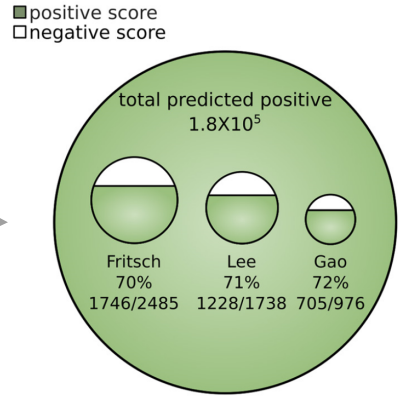
Int. ATG Start

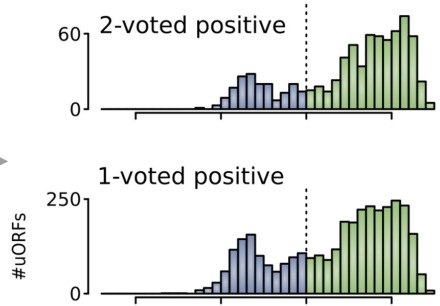
Cons- ervation

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)]

RADAR: Annotation & prioritization of variants in the post-transcriptional regulome for RBPs

- Background on prioritizing non-coding variants:
  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

- Next step in prioritizing variants associated with RNA:
  uORFs - Feature integration to find small subset of upstream mutations that potentially alter translation

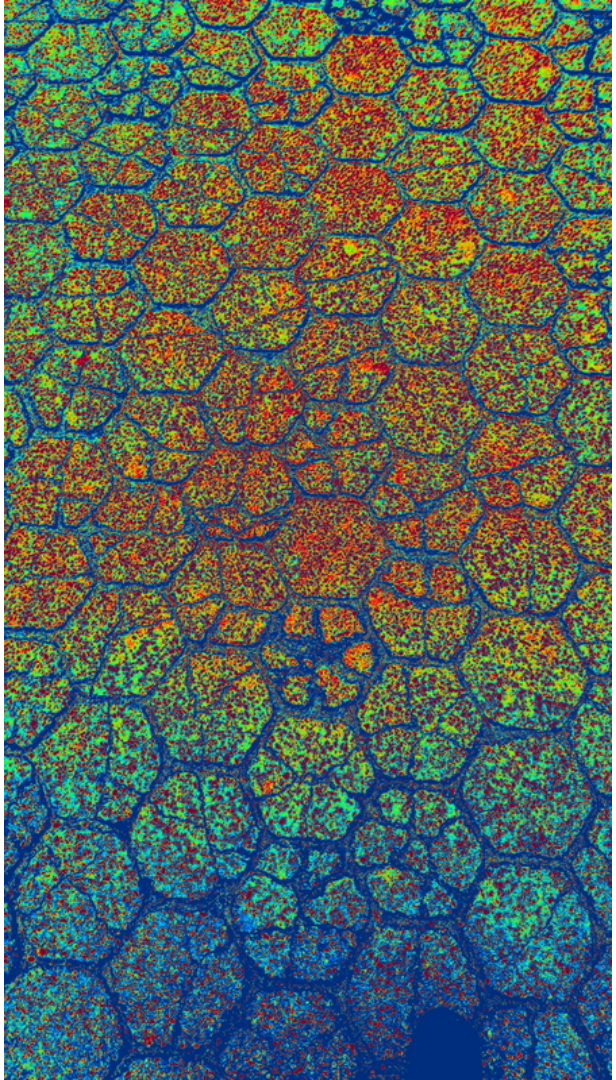
RADAR: Annotation & prioritization of variants in the post-transcriptional regulome for RBPs

- Background on prioritizing non-coding variants:
  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

- Next step in prioritizing variants associated with RNA:
  uORFs - Feature integration to find small subset of upstream mutations that potentially alter translation

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



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,
L Lochovsky, J-J Feng, S Lou,
M Rutenberg-Schoenberg

github.gersteinlab.org/**uORFs**

P **McGillivray**, R Ault,
M Pawashe, R Kitchen, S Balasubramanian


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