Evaluation of "near-coding" elements: uORFs & RBP sites

M Gerstein

(Gencode call)

Coding v Non-coding

- Coding
  - Easily interpretable, particularly related to structure
  - Available in large quantities
  - Exomes have the current potential for great scale (Scale of EXAC, >60K exomes [Lek et al. ‘16])
- Non-coding
  - Not as interpretable & hard to connect to genes
- “Near coding”
  - Bits of non-coding, close to genes & readily linked to them
  - EX: Splice sites, promotors, uORF

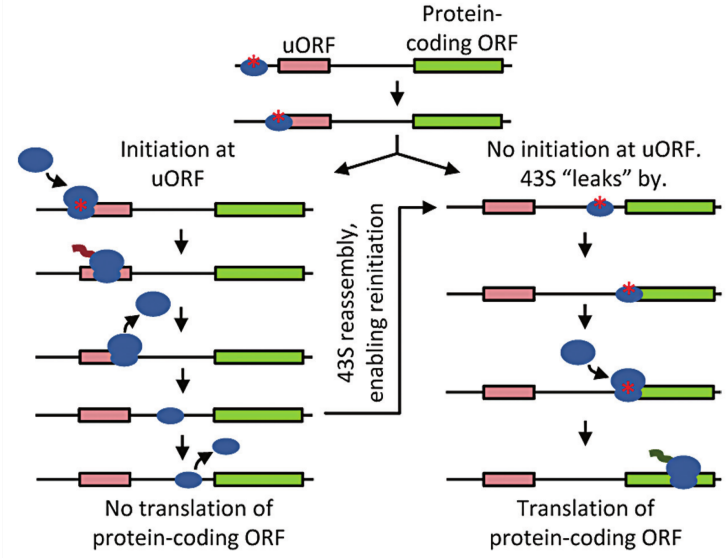
Evaluation of “near-coding” elements

- **uORFs:**
  Feature integration to find small subset of upstream mutations that potentially alter translation

- **RADAR:**
  prioritize variants based on post-transcriptional regulome using ENCODE eCLIP

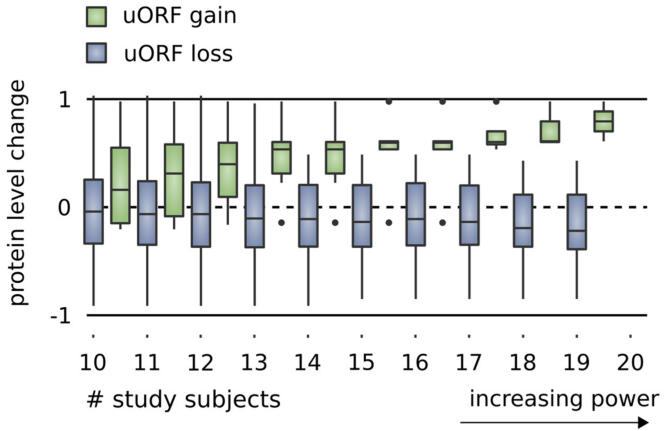
Upstream open reading frames (uORFs) regulate translation are affected by mutation

[Calvo et al., *PNAS* (‘09)]

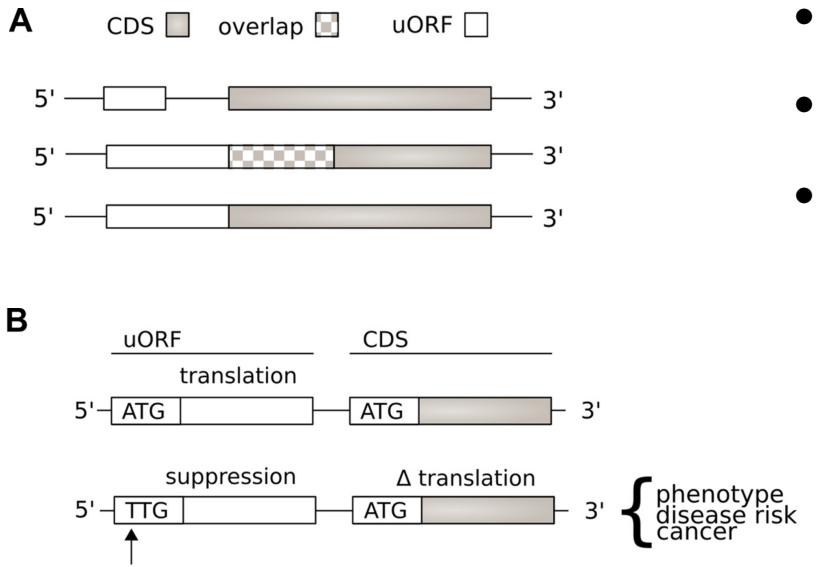


[Ferreira et al., *Bioengineered* (‘14)]

- 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.

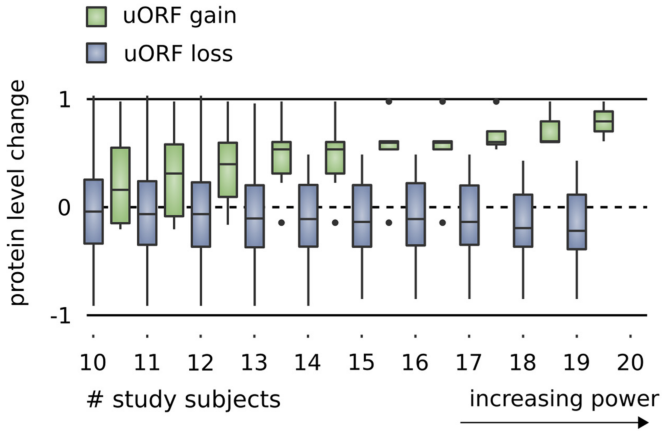


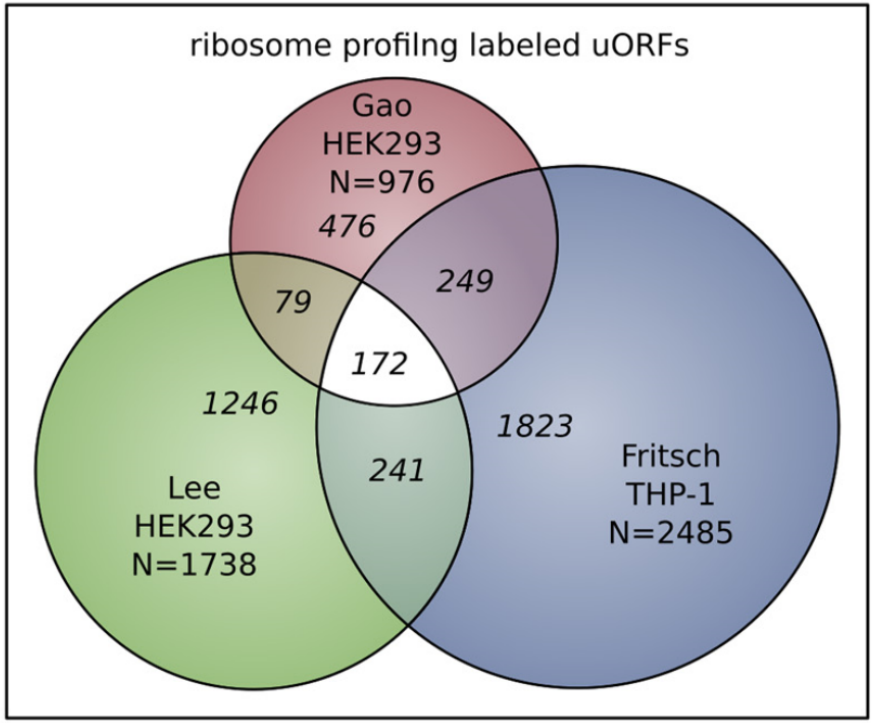
Upstream open reading frames (uORFs) regulate translation are affected by mutation



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

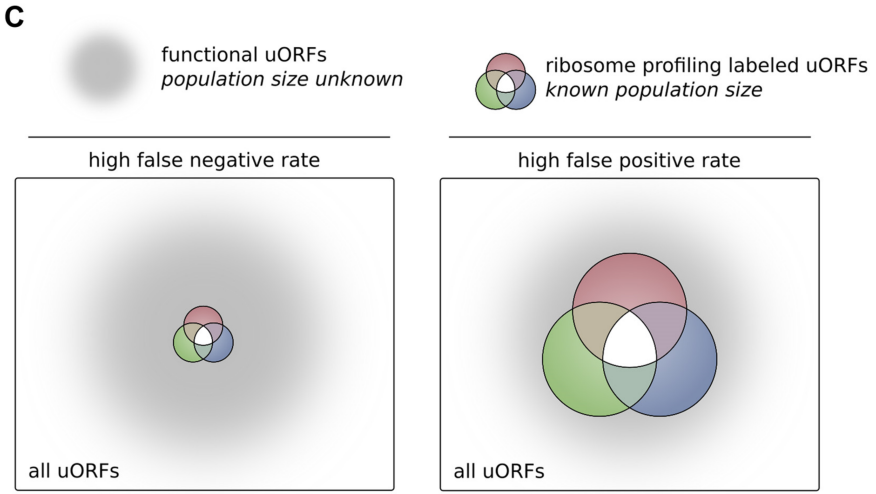
- uORFs regulate the translation of downstream coding regions.
- This regulation may be altered by somatic mutation in cancer.
- In Battle et al. 2014 data uORF gain & loss assoc. protein level change.





From a “Universe” of 1.3 M pot. uORFs

The population of functional uORFs may be significant

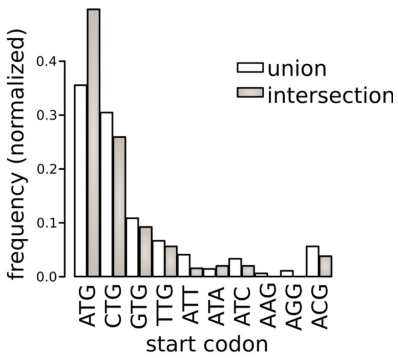


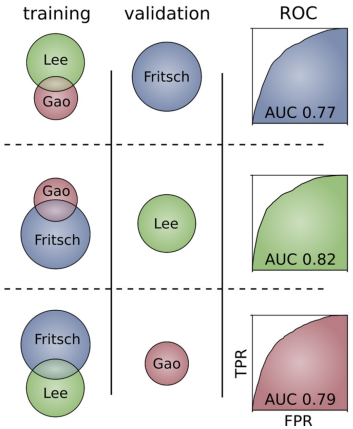
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

[McGillivray et al., *NAR* (‘18)]

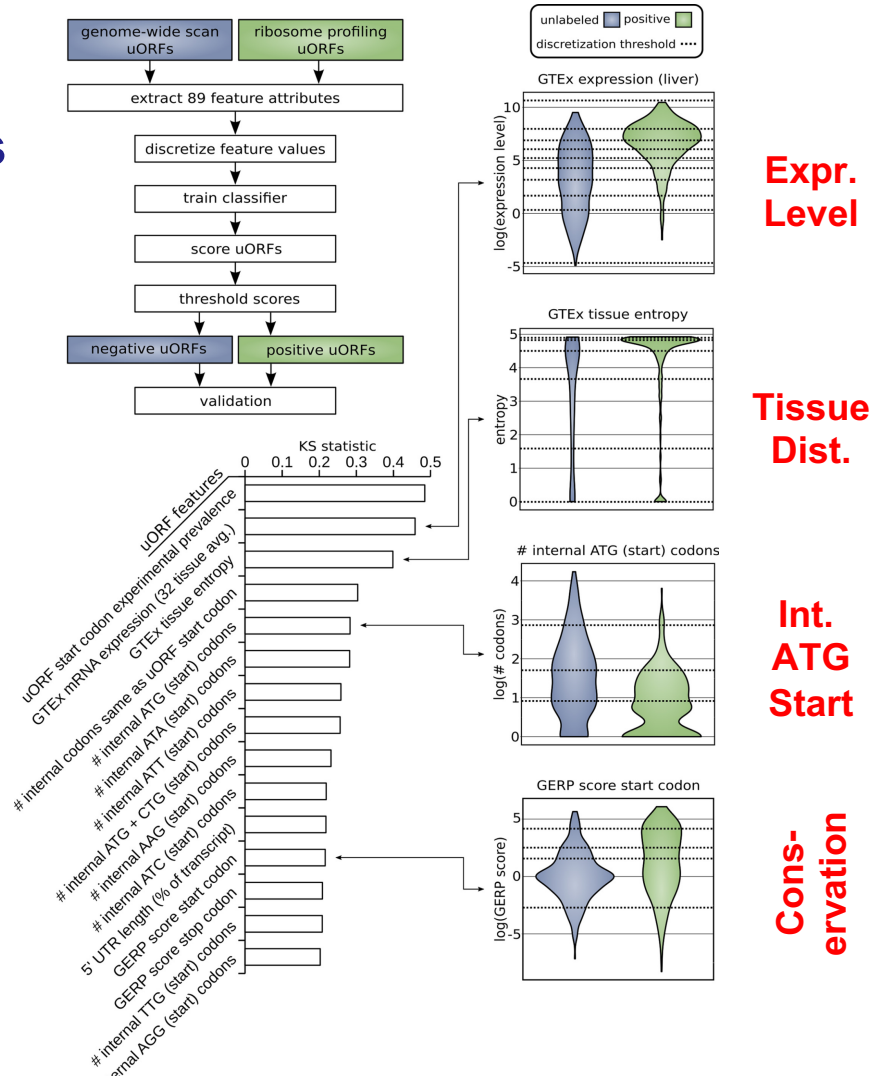
Prediction & validation of functional uORFs using 89 features

- 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).



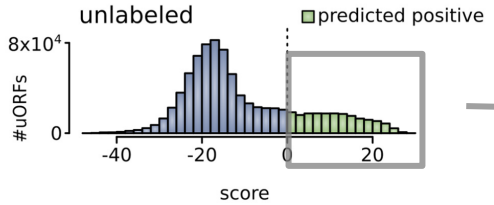


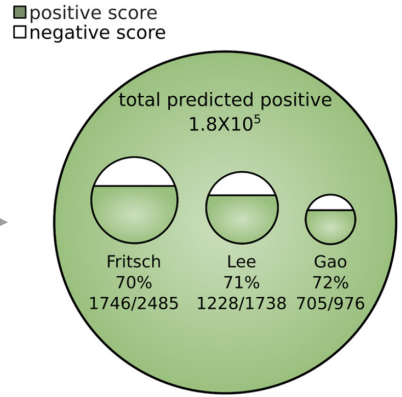
[McGillivray et al., *NAR* ('18)]

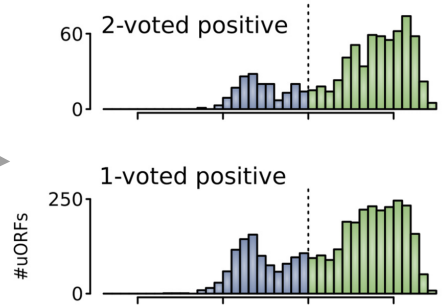


A comprehensive catalog of functional uORFs

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







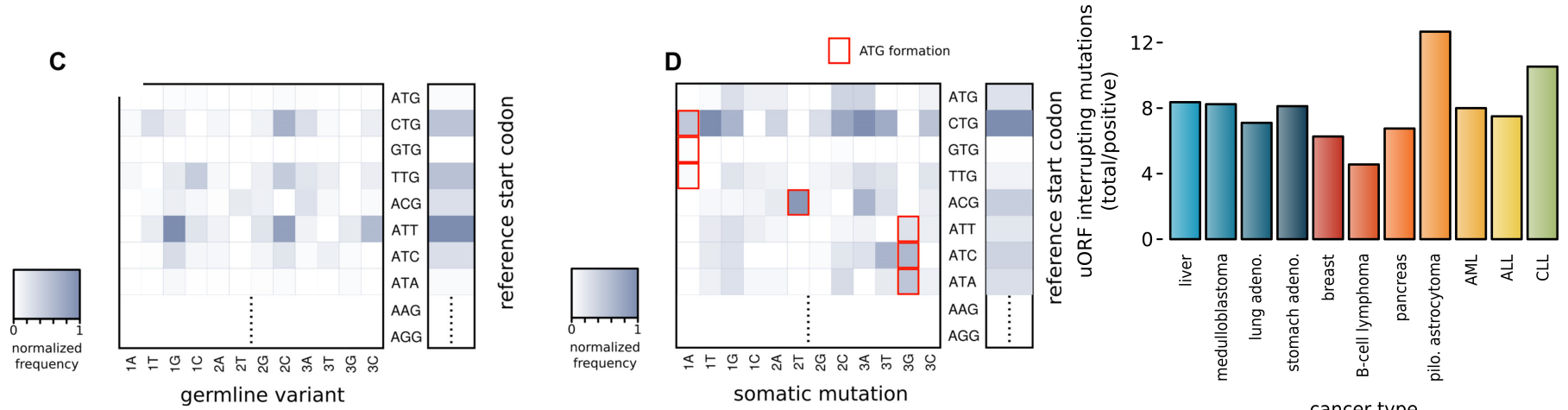
- Predicted functional uORFs may be intersected with disease associated variants.

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

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

Somatic alteration of uORFs disproportionately affects certain cancers and molecular pathways

- uORF gain and loss occurs in cancer (incl. in cancer associated genes, e.g., MYC, BCL2, etc.).
- Alteration of translation may contribute to cancer.
- These changes are concentrated in certain cancers and pathways.
- Mutations leading to uORFs diff in somatic vs. germline.



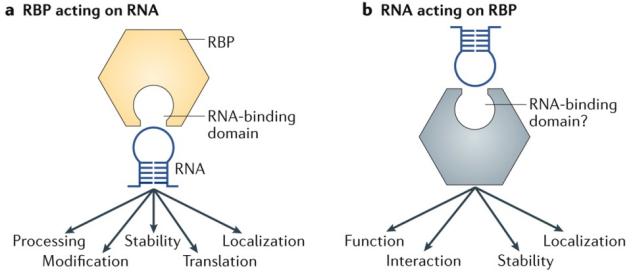
[McGillivray et al., *NAR* ('18)]

Evaluation of “near-coding” elements

- **uORFs:**
  Feature integration to find small subset of upstream mutations that potentially alter translation

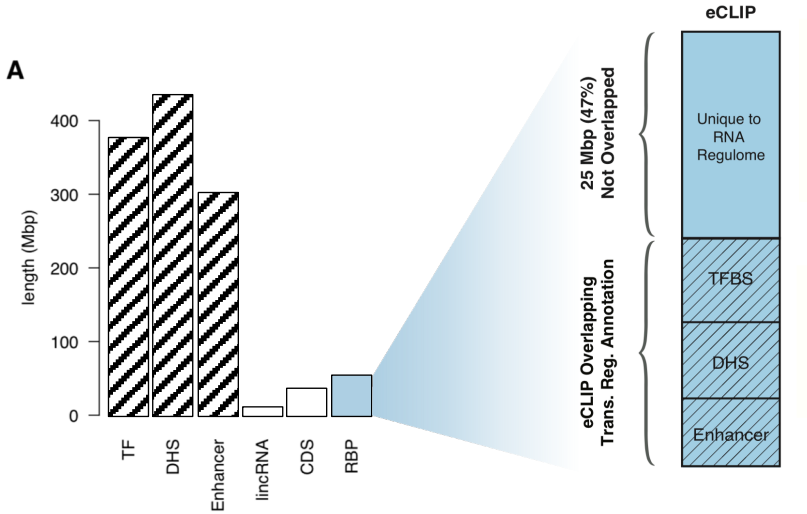
- **RADAR:**
  prioritize variants based on post-transcriptional regulome using ENCODE eCLIP

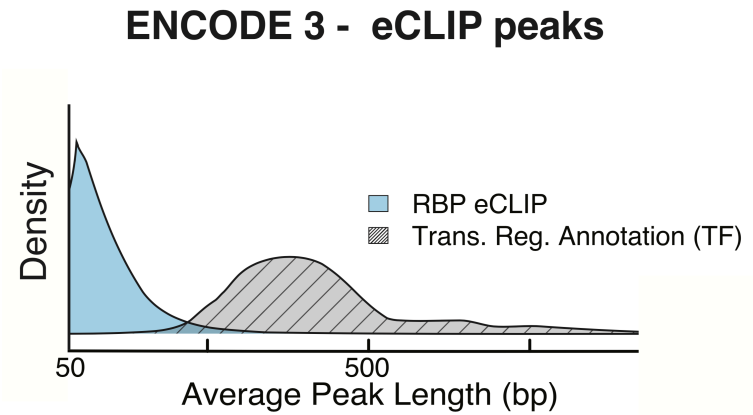
RNA Binding Proteins (RBPs)



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

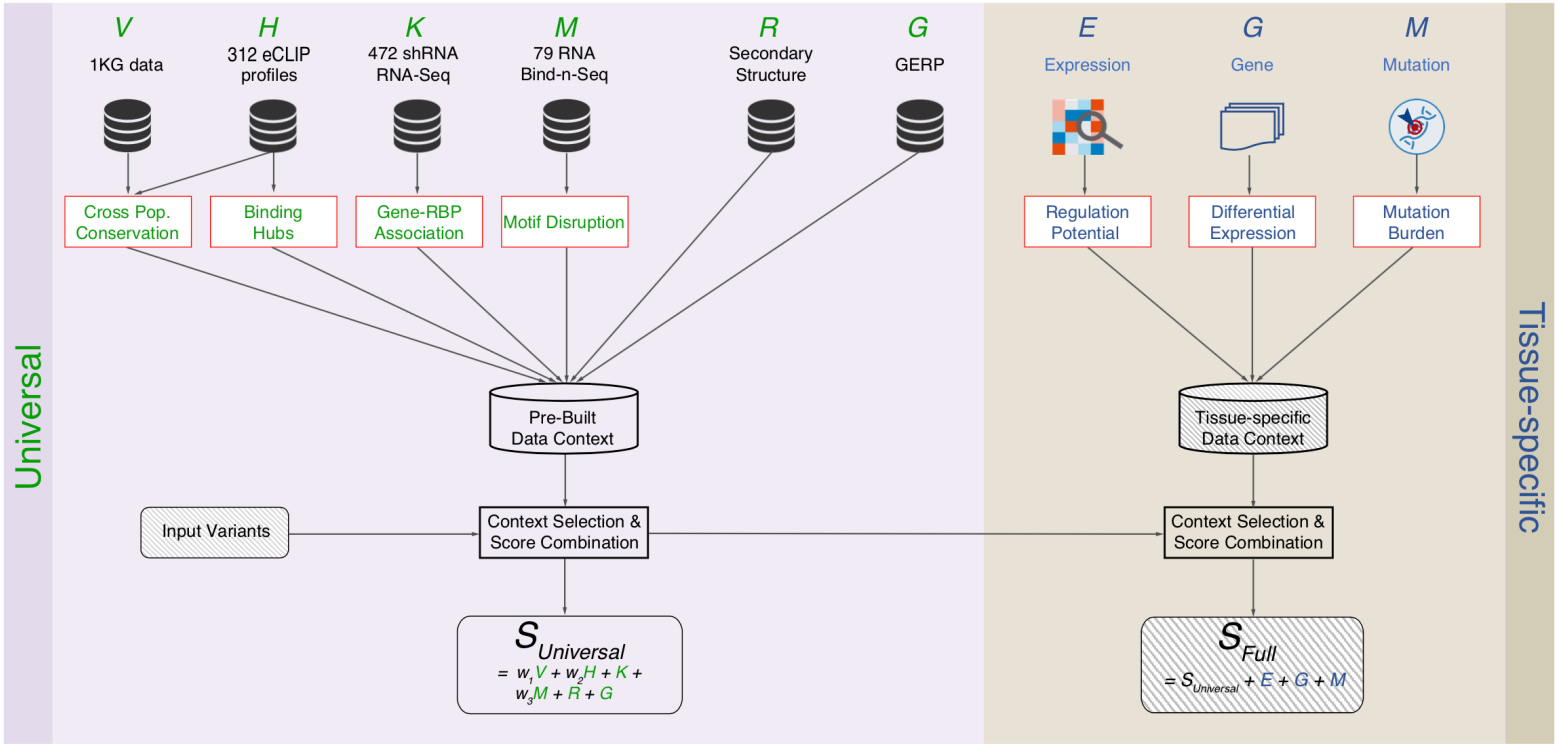
- **Before ENCODE3: >150 expt.**
  in many different cell types
- **ENCODE3 did ~350 focused eCLIP expt.**
  for >110 RBPs on HepG2 & K562
  (Van Nostrand...Yeo. Nat. Meth. '16;
  Van Nostrand...Graveley, Yeo
  (submitted in relation to ENCODE3))

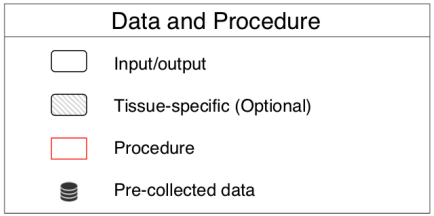


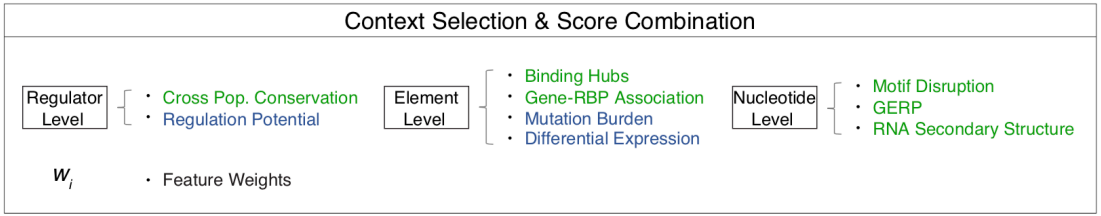


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

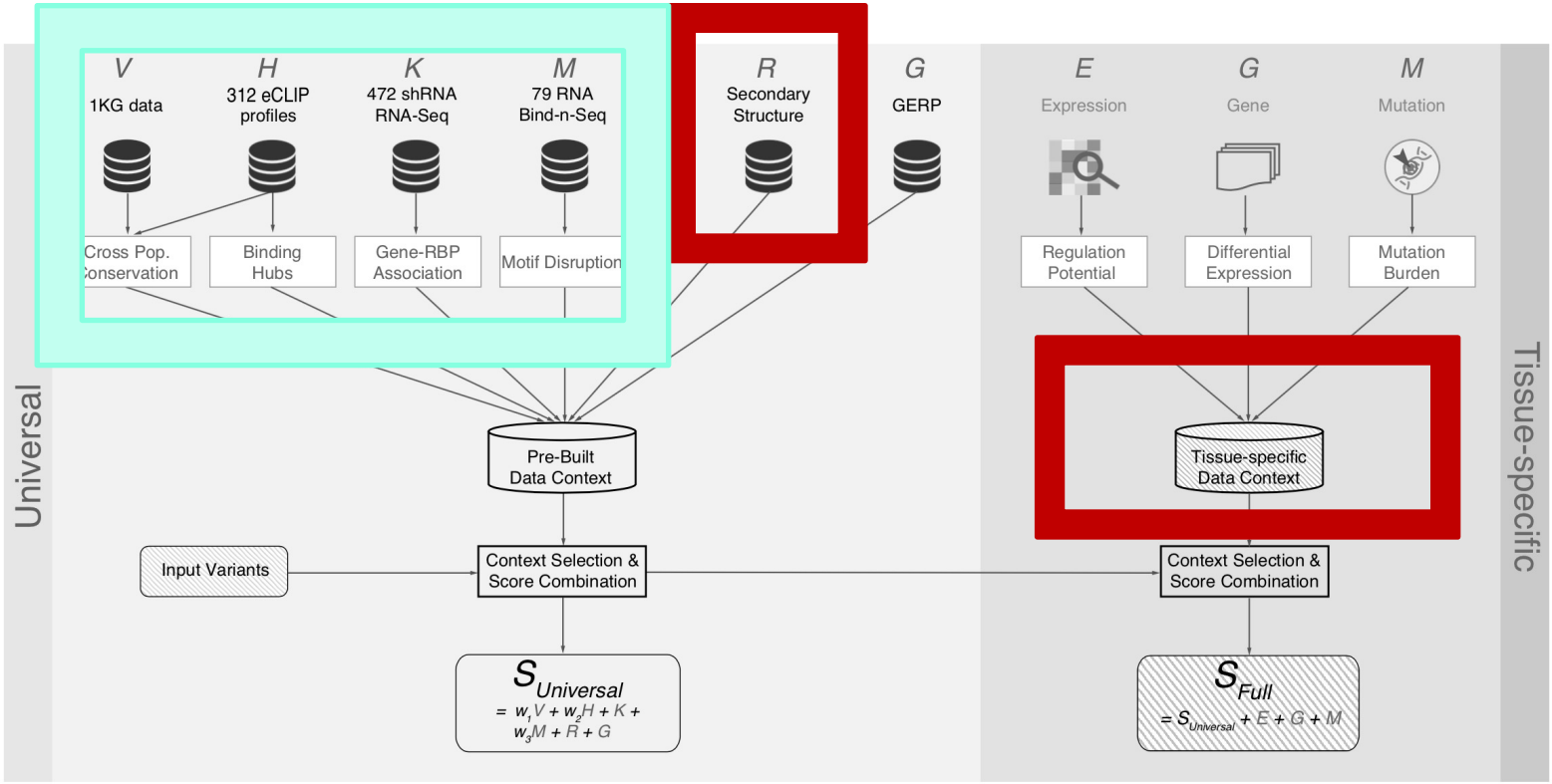
Schematic of RADAR Scoring

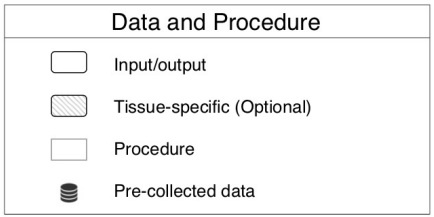


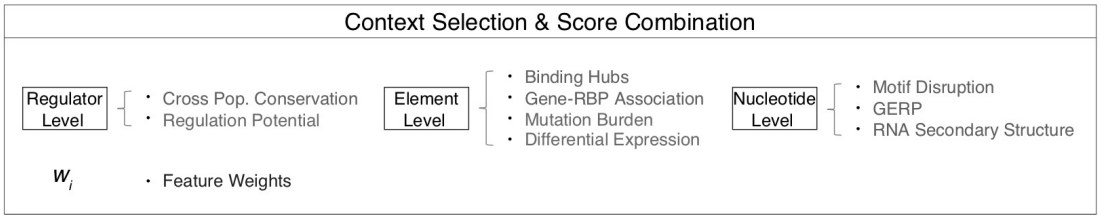




[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

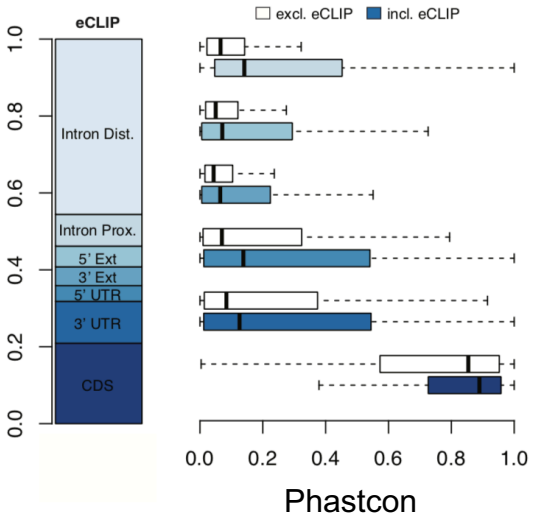




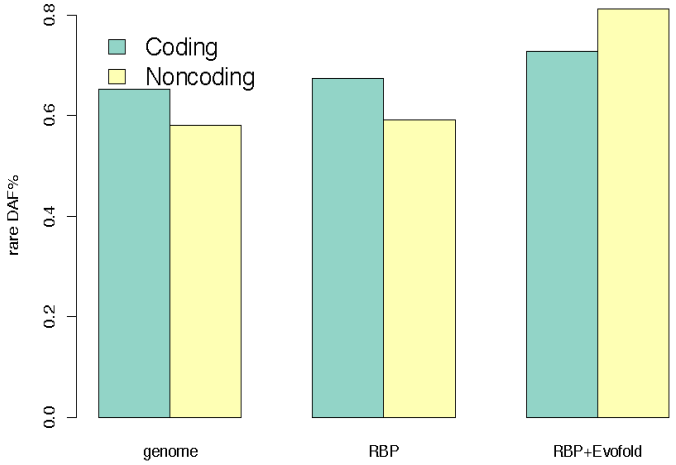


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

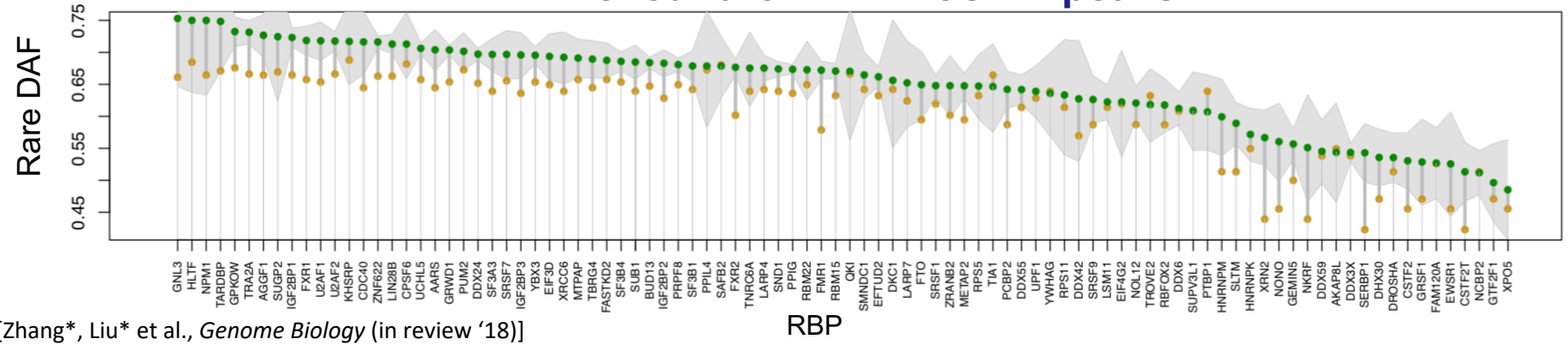
High Phastcon in RBP-overlapped annotations



RNA Structure Cons. from Evofold



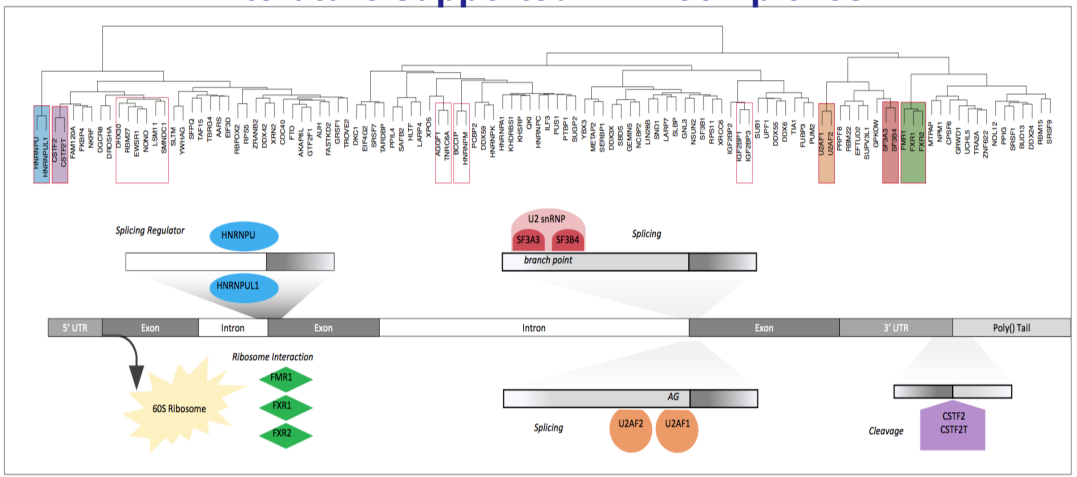
Enriched rare DAF in eCLIP peaks



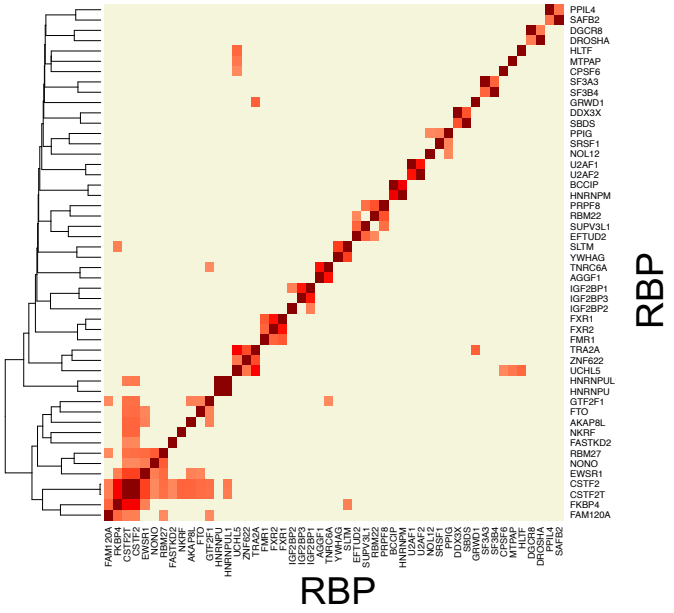
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Co-binding of RBPs form biologically relevant complexes

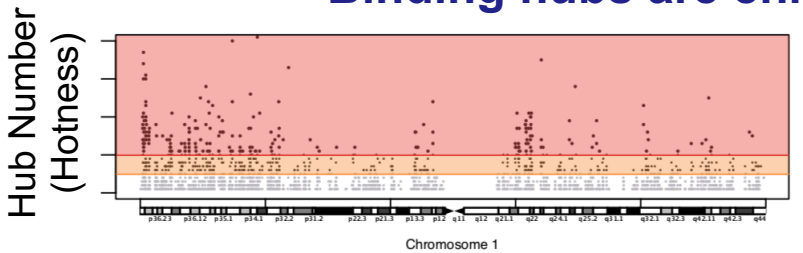
Literature supported RBP complexes

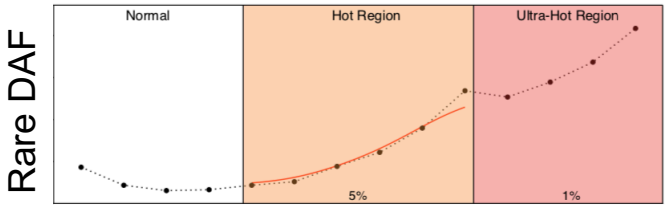


Unique co-binding patterns of RBPs

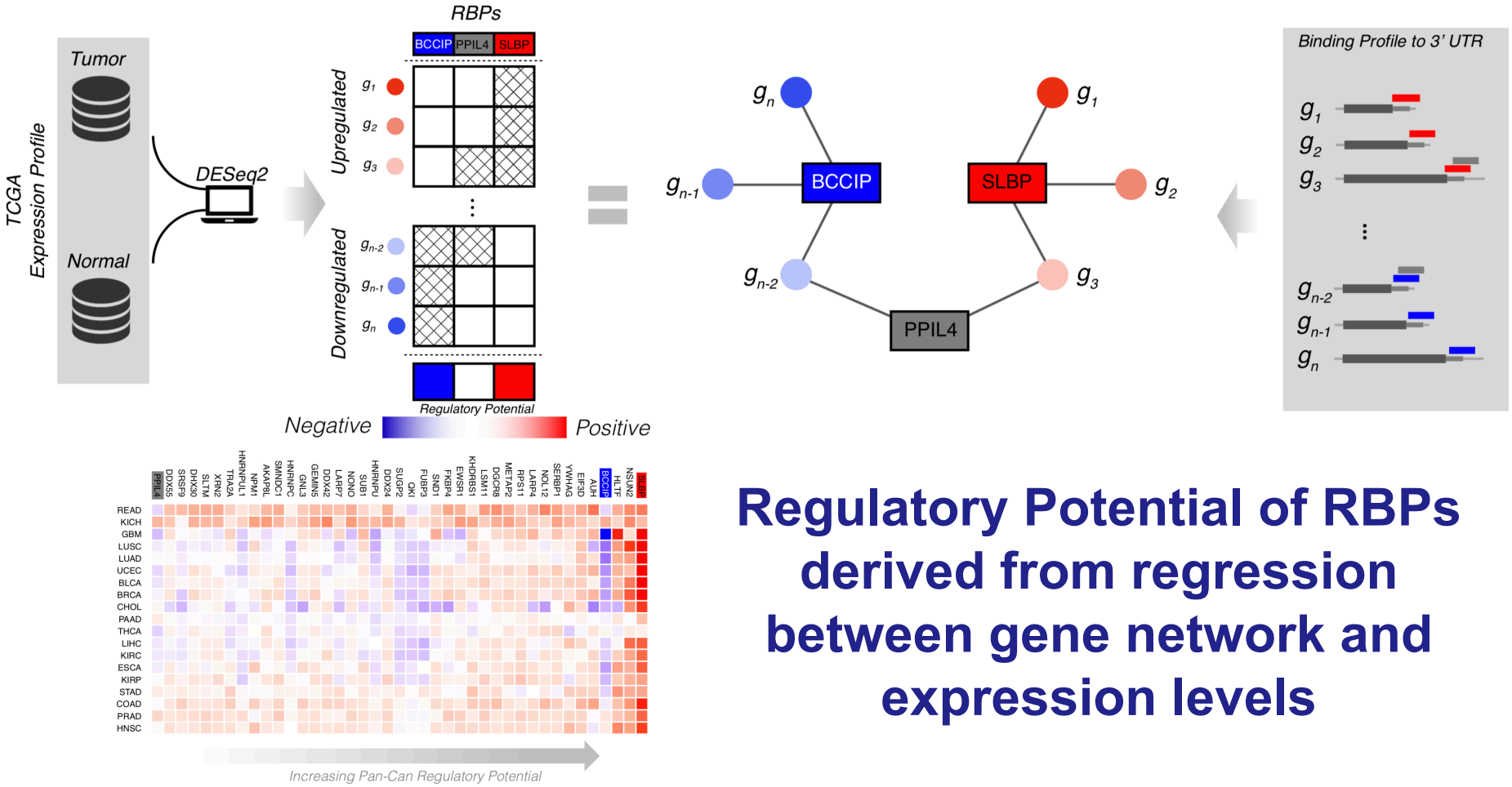


Binding hubs are enriched for rare variants





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

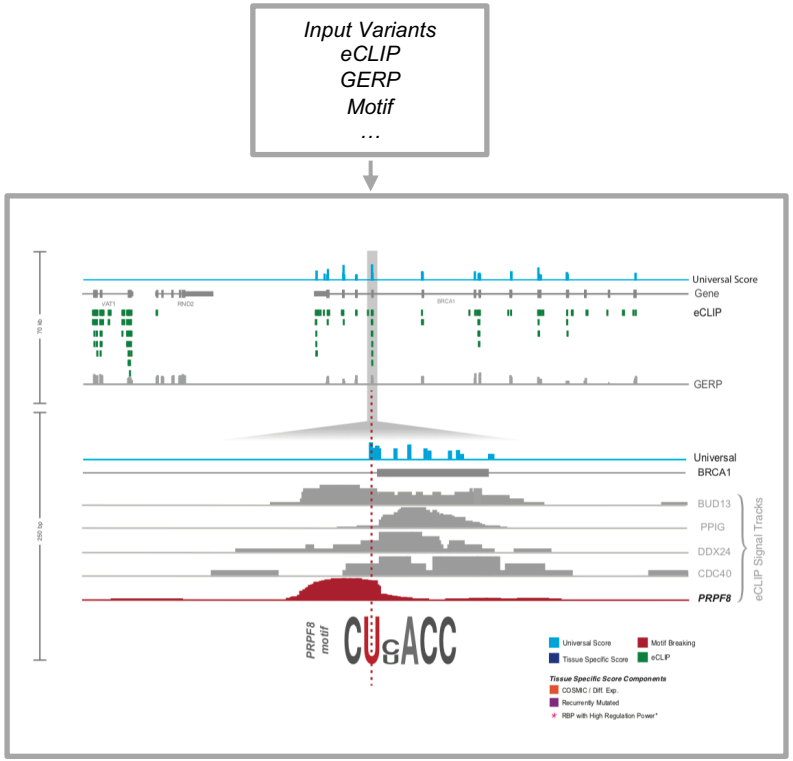


Regulatory Potential of RBPs derived from regression between gene network and expression levels

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

Visualization of RADAR Features and Scoring

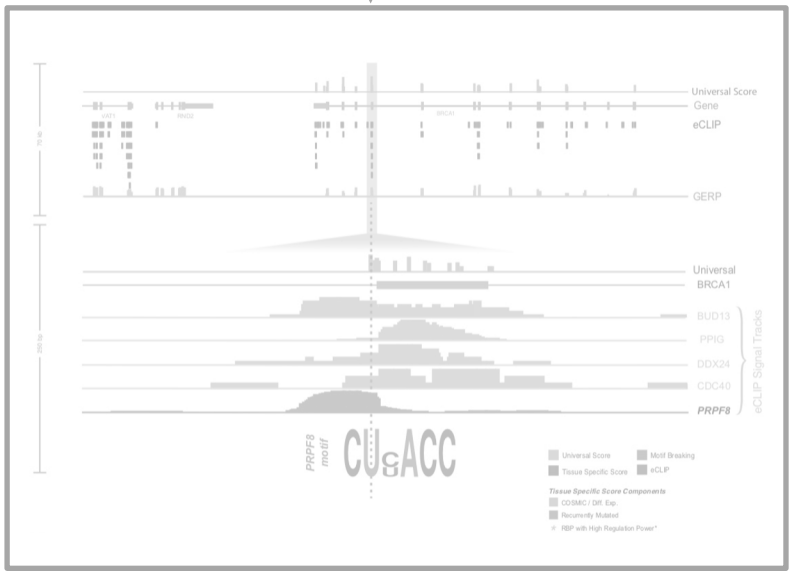
Germline Variants are Score Using a Universal Scoring Scheme



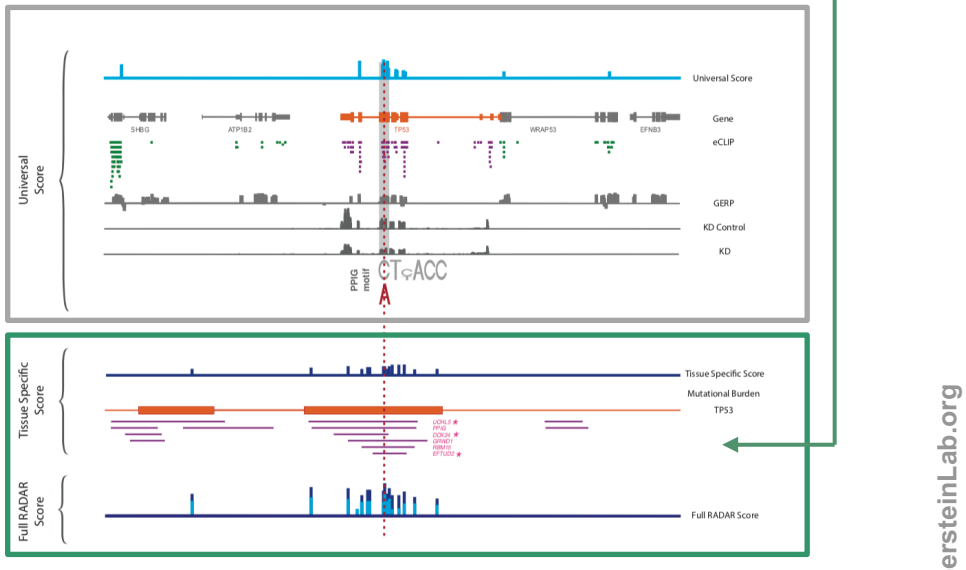
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

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 can be run on the command line by following the instructions on the Docs page or through the web here. Running RADAR through the website will print the results after several moments. You can try running RADAR through the web form with a sample file with one variant. Alternatively, you may also input a list of variants into the form as text. If variants are provided in both file and text formats, the variant file will be scored and the text field will be ignored.

More details on the RADAR inputs can be found on the Docs page.

- Variants: a list of variants
  - BED file: a BED file containing the variants
  - Text format: type variants directly into a text box, lines may be tab- or space-delimited
- Cancer type: a TCGA cancer type, only needed if any tissue-specific scores are to be included.
- Tissue-specific scores: which tissue-specific scores should be included along with the universal scores for each variant.

Variants:

Choose File no file selected

E.g. chr1 13506 13507 G A

Cancer type:

Select a cancer

Tissue-specific scores:

- [ ] Key genes
- [ ] Mutation recurrence
- [ ] RBP-regulation power

Score variants

Evaluation of “near-coding” elements

- **uORFs:**
  Feature integration to find small subset of upstream mutations that potentially alter translation

- **RADAR:**
  prioritize variants based on post-transcriptional regulome using ENCODE eCLIP

Acknowledgements

github.gersteinlab.org/**uORFs**

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

RADAR.gersteinlab.org

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