Measuring the Impact of Non-coding Mutations: Repurposing an Additive Effects Model & Developing New Annotations

Mark Gerstein
Yale

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

Dark Matter of the Genome

Non-coding genome is ~99% of the total bases

Not as obvious how to annotate it
or assess the impact of mutations
relative to those in coding regions

[Gravitational lensing by dark matter in Abell 1689 – HST (NASA, ESA)]

Measuring the Impact of Non-coding Mutations: Repurposing an Additive Effects Model & Developing New Annotations

- **Additive-Effects model to measure the Impact of non-coding v coding mutations**
  - Repurposing a formalism from germline genetics for missing heritability to cancer
  - Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  - Notable effect, particularly for non-coding passengers, in addition to known coding drivers.
  - Recasting as a predictive model to est. number of weak drivers

- **New Annotations to help assess the impact non-coding variants**
  - **RADAR Prioritization for RBP sites**
    - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
    - Incorporates new annotation features related to RNA sec. struc & tissue specific effects
  - **uORF Prioritization**
    - Feature integration to find small subset of upstream mutations that potentially alter translation

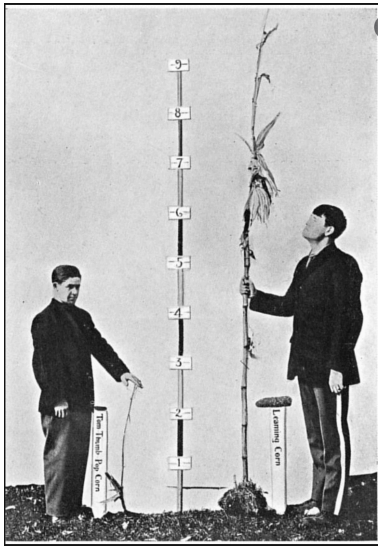
Measuring the Impact of Non-coding Mutations: Repurposing an Additive Effects Model & Developing New Annotations

- **Additive-Effects model to measure the Impact of non-coding v coding mutations**
  - Repurposing a formalism from germline genetics for missing heritability to cancer
  - Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  - Notable effect, particularly for non-coding passengers, in addition to known coding drivers.
  - Recasting as a predictive model to est. number of weak drivers

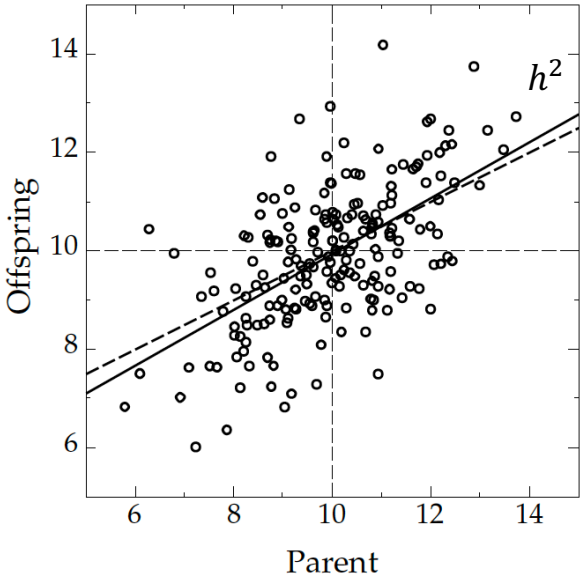
- **New Annotations to help assess the impact non-coding variants**
  - **RADAR Prioritization for RBP sites**
    - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
    - Incorporates new annotation features related to RNA sec. struc & tissue specific effects
  - **uORF Prioritization**
    - Feature integration to find small subset of upstream mutations that potentially alter translation

Relating Germline Missing Heritability to Cancer Studies

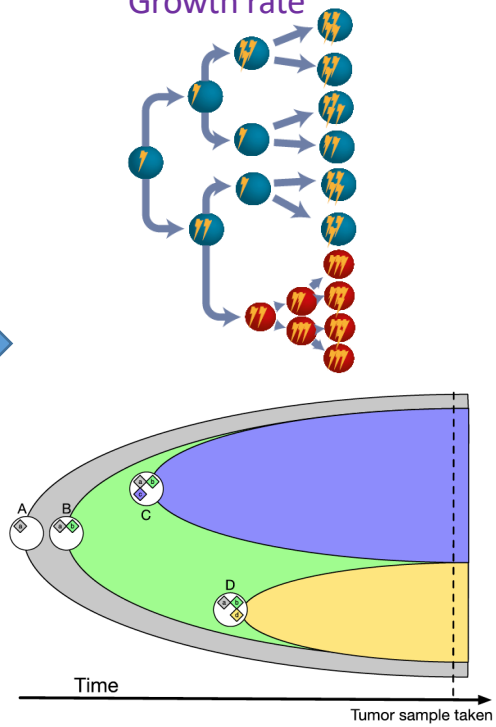
Organismal trait: Height



Population level definitions:
Parent-offspring heritability;
Twin-based heritability …



Subclonal trait in cancer:
Growth rate



SNP-based polygenic & additive model:

$$h^2 = \sigma_u$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

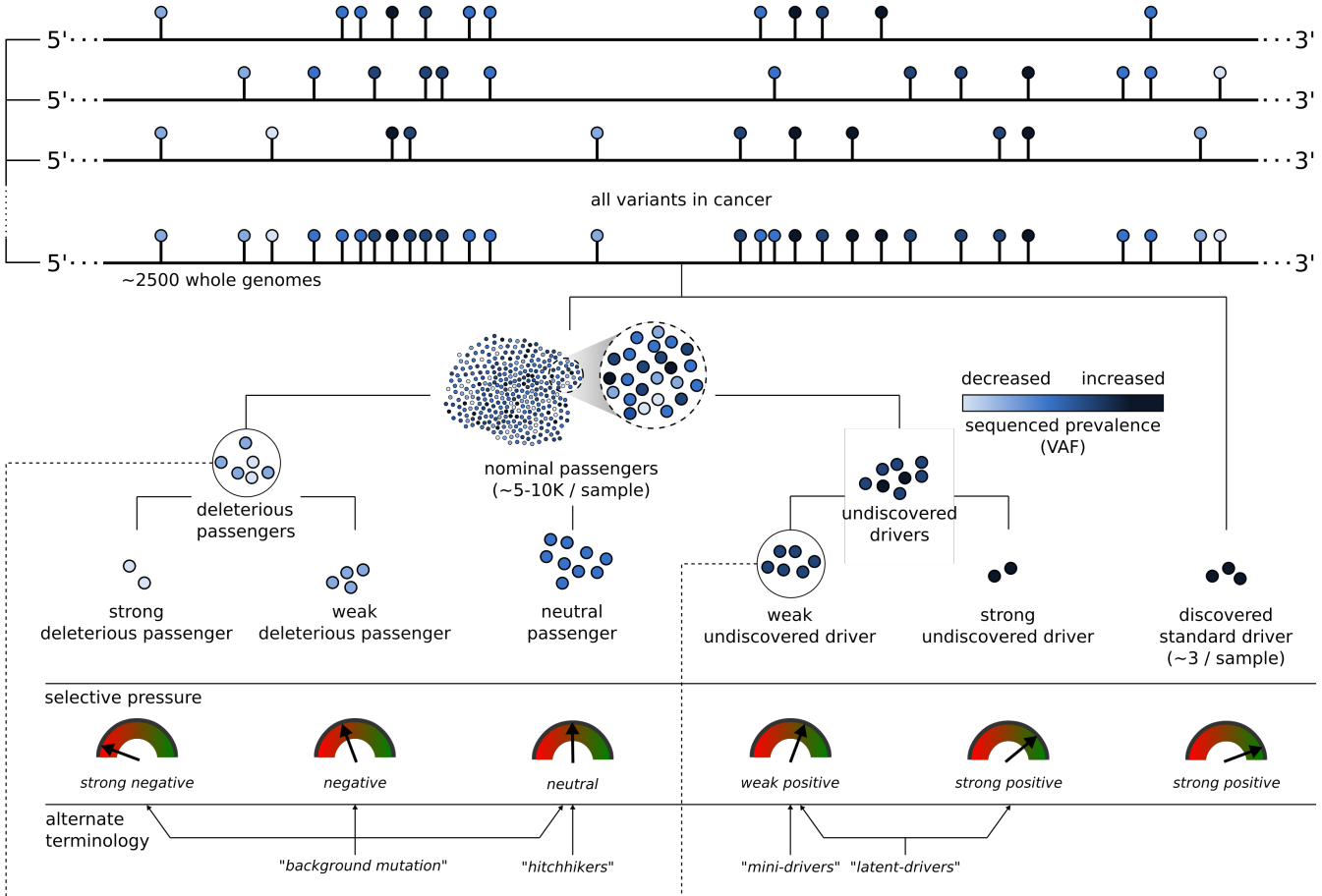
Trait

Covariates & fixed effects

Genetic predictors & random effects

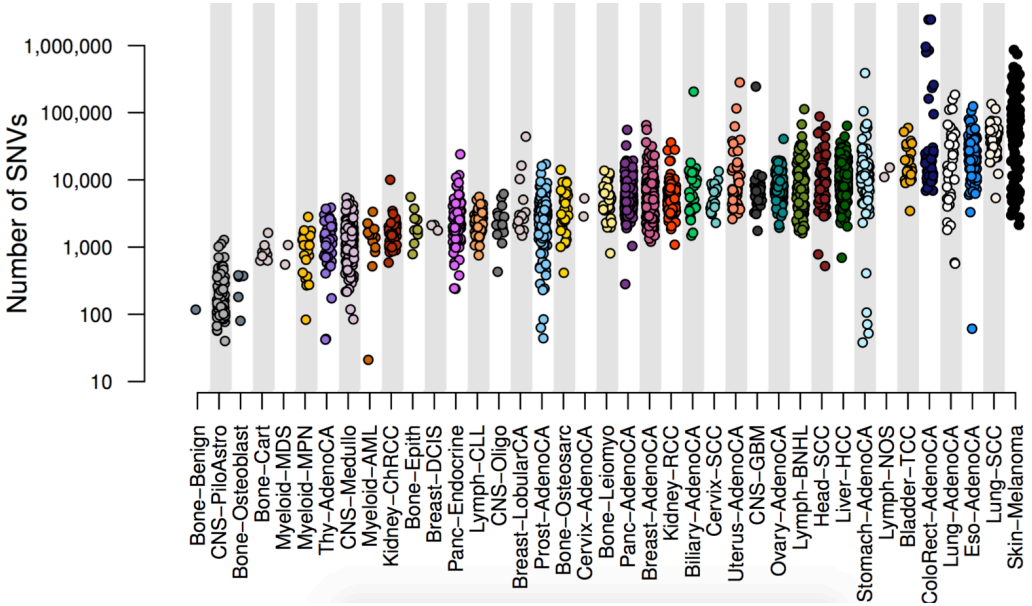
Environmental noise

Conceptual extension of the canonical model of drivers and passengers



[Kumar, Warrell et al. ('20, in press) *Cell* + biorxiv]

PCAWG : most comprehensive resource for cancer whole genome analysis



Adapted from Campbell et. al., bioRxiv ('17).
Now published as Nature 578: 82–93 (2020)

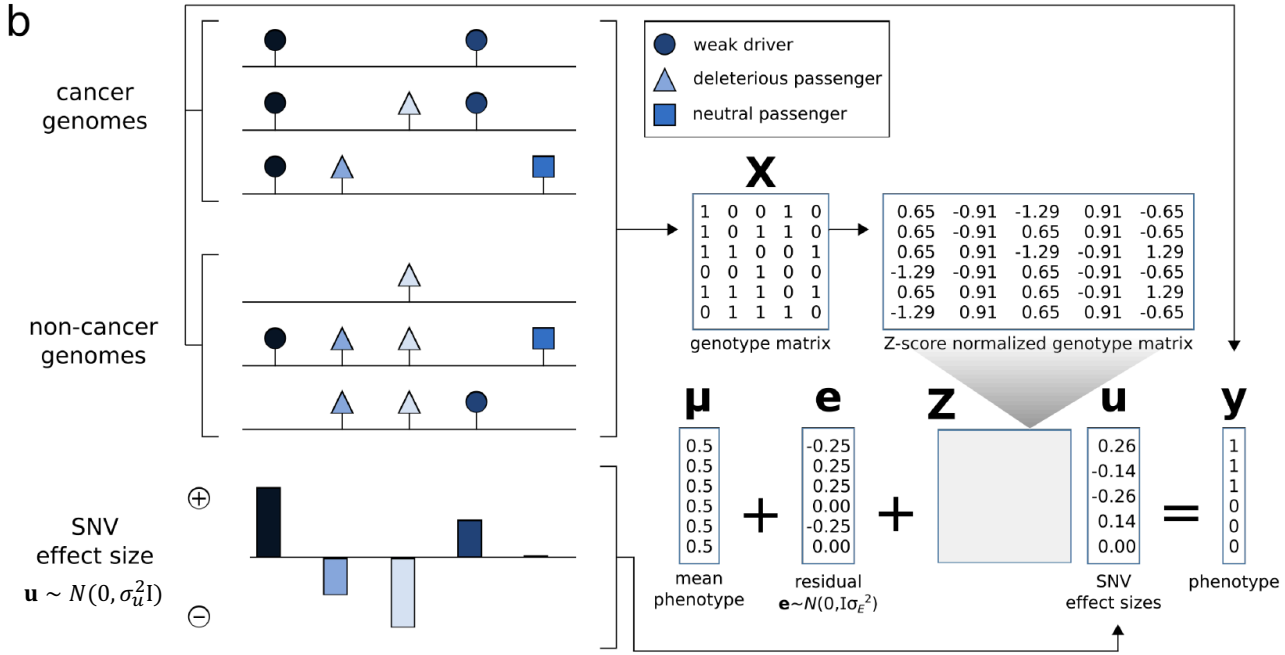
Project Goals:

- To understand role of non-coding regions of cancer genomes in disease progression.
- **Union of TCGA-ICGC efforts**

- Jointly analyzing ~2800 whole genome tumor/normal pairs
  - > 580 researchers
  - 16 thematic working groups
  - ~30M total somatic SNVs

PCAWG
PanCancer Analysis
OF WHOLE GENOMES

Using additive effects to compare different categories of variants



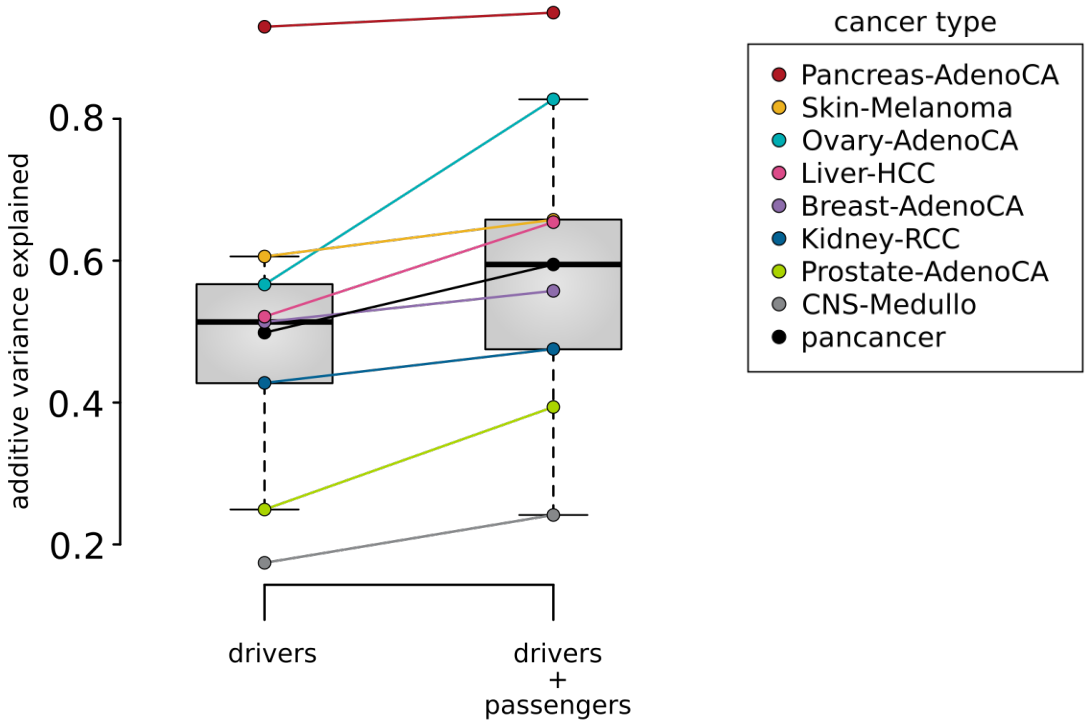
Model: $y_j = \mu + z_j^{\mathrm{drv}} u_1 + \sum_{k \in \{2,3,4\},} z_{ijk} u_{ik} + e_j$

Parameters: $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_E^2)$

Variant categories:
$k = 1$: **coding drivers**
$k = 2$: coding other
$k = 3$: **promoters**
$k = 4$: **other non-coding**

[Kumar, Warrell et al. ('20, in press) *Cell* + biorxiv]

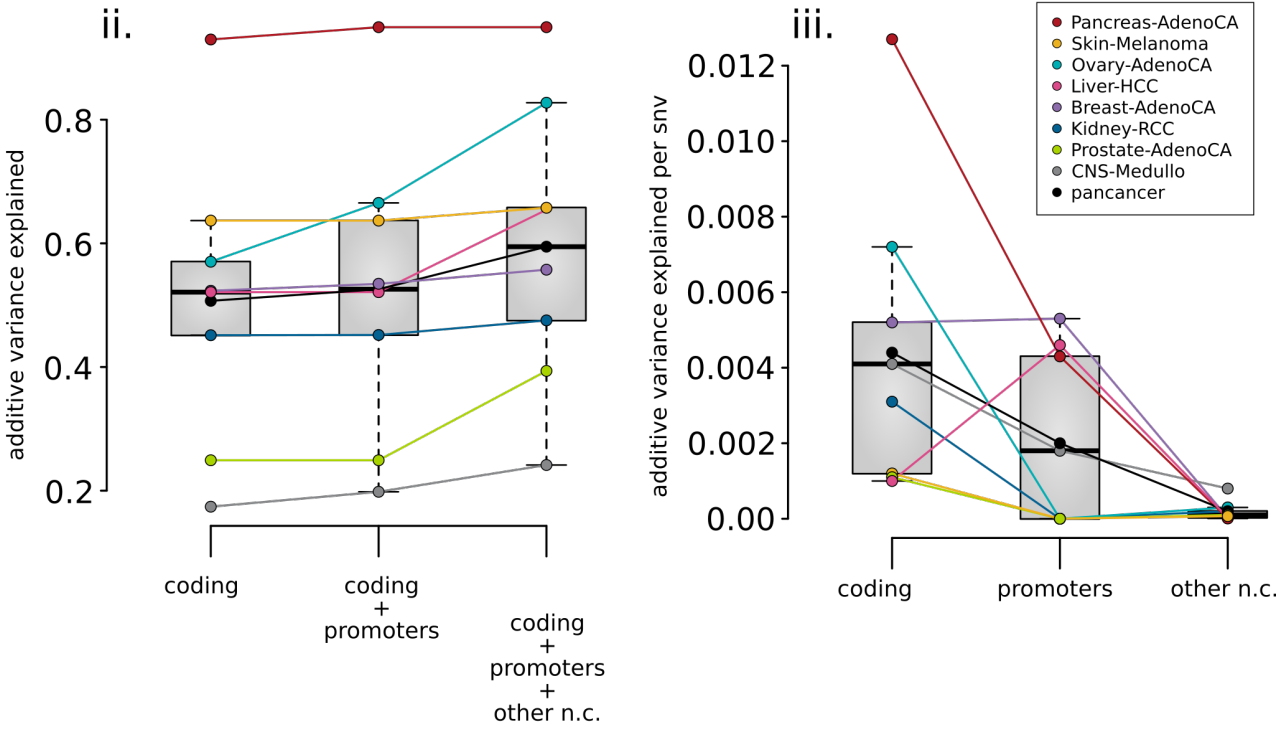
Overall additive variance increase for multiple cancer cohorts in PCAWG with the inclusion of passengers



Increase in the variance from ~50% using drivers alone to ~59% with putative passengers included, averaged across all cohorts.

[Kumar, Warrell et al. ('20, in press) *Cell* + biorxiv]

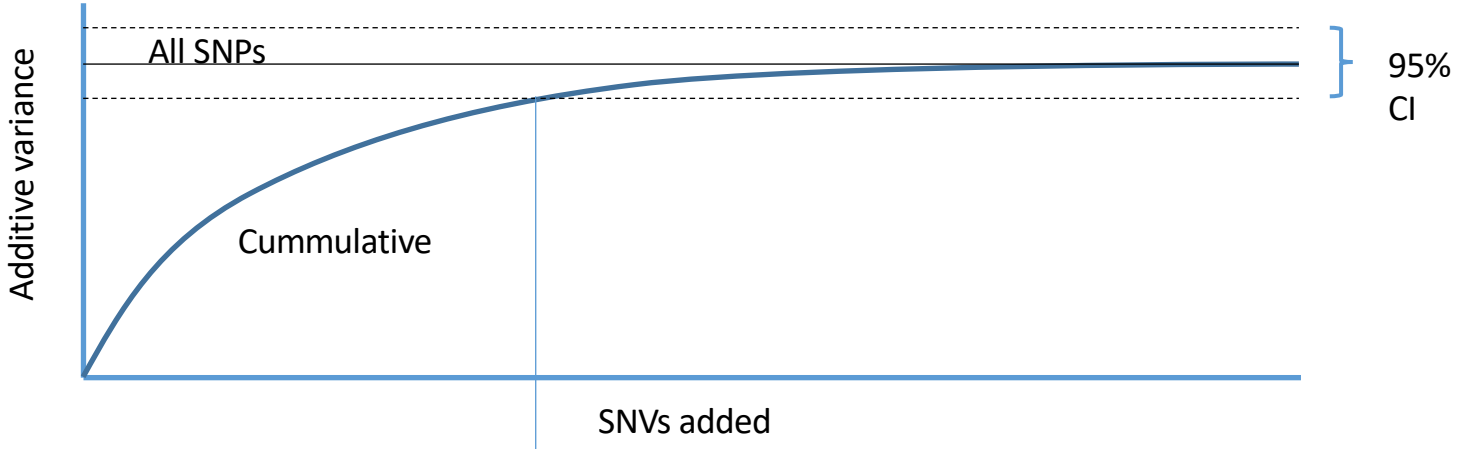
Element level additive variance for multiple cancer cohorts in PCAWG, comparing coding & non-coding



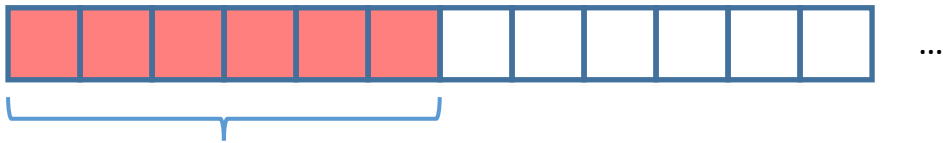
In addition to coding mutations, promoter & other non-coding mutations contributed significant amounts of extra variance (~2% & 7%) .

[Kumar, Warrell et al. ('20, in press) *Cell* + biorxiv]

Recasting the additive effects model in a predictive context: Best Linear Unbiased Predictor (BLUP) analysis



SNVs, ordered by descending BLUP ($\hat{\mathbf{u}}$):



BLUP predictor:

$$\hat{\mathbf{u}} = \text{argmax}_{\mathbf{u}}\left(P(\mathbf{u}|\mathbf{y}, \sigma_u^2)\right)$$
$$= \text{argmax}_{\mathbf{u}}\left(P(\mathbf{y}|\mathbf{u})P(\mathbf{u}|\sigma_u^2)\right)$$

Lower bound on # weak drivers (8.4 pan-cancer average; enriched for PCAWG genes w/ FDR<0.25)

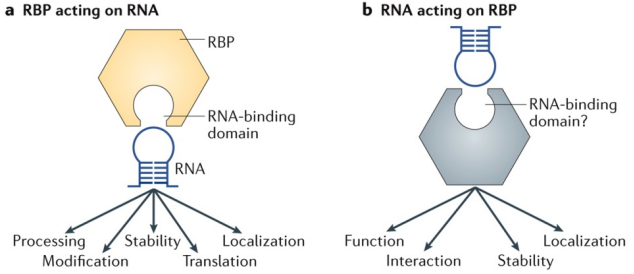
[Kumar, Warrell et al. ('20, in press) *Cell* + biorxiv]

Measuring the Impact of Non-coding Mutations: Repurposing the Additive Effects Model & Developing New Annotations to Assess Variants

- **Additive-Effects model to quantify the Impact of non-coding v coding mutations**
  - Repurposing a formalism in germline genetics for missing heritability to cancer
  - Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  - Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

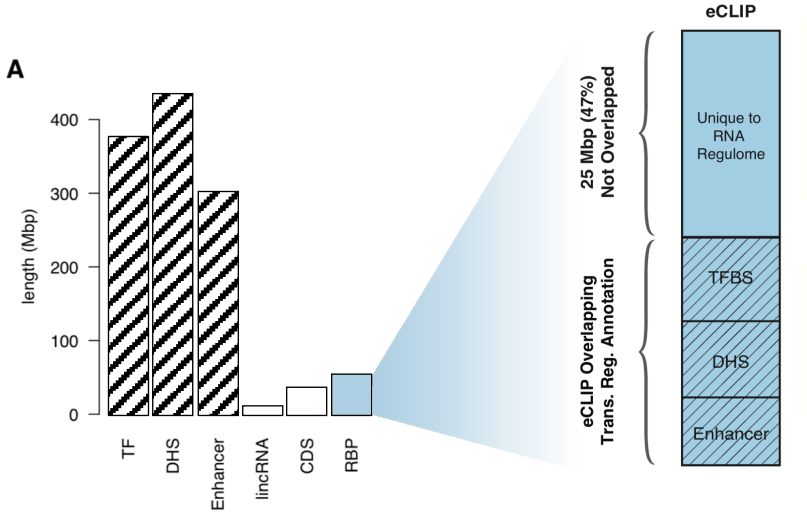
- **New Annotations to help assess the impact non-coding variants**
  - **RADAR Prioritization for RBP sites**
    - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
    - Incorporates new features related to RNA sec. struc & tissue specific effects
  - **uORF Prioritization**
    - Feature integration to find small subset of upstream mutations that potentially alter translation

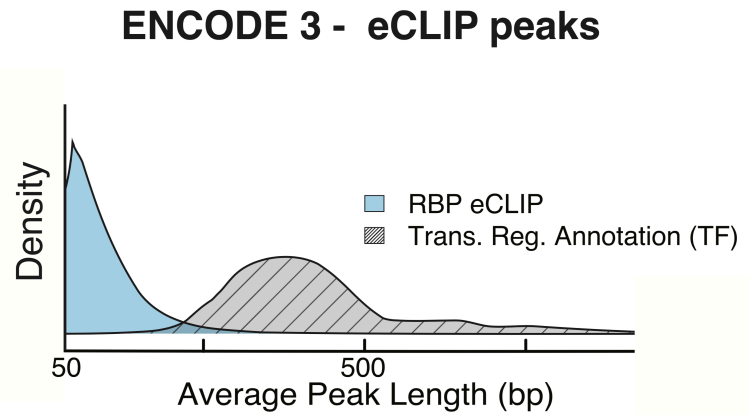
RNA Binding Proteins (RBPs)



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

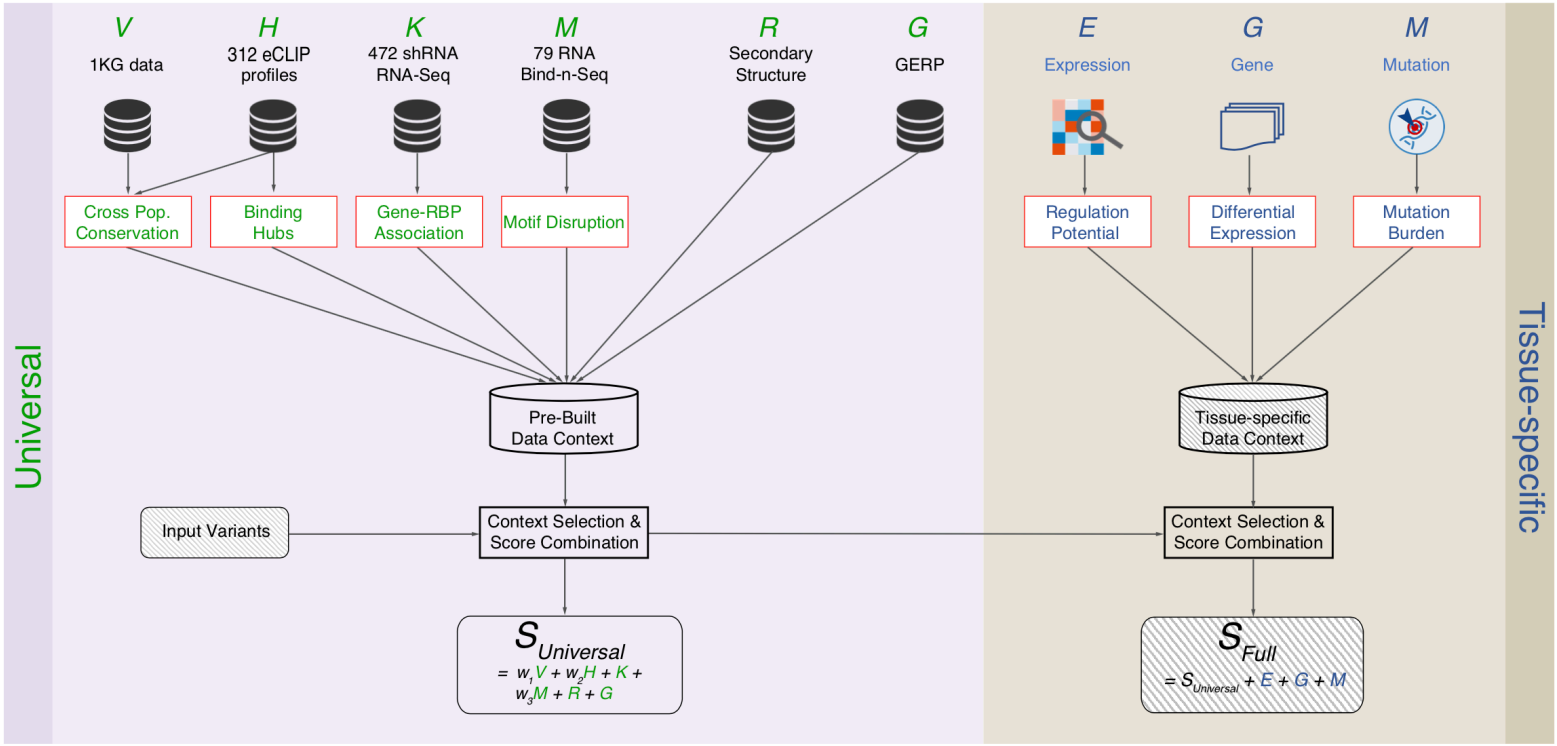
- **Before ENCODE3: >150 expt.**
  in many different cell types
- **ENCODE3 did ~350 focused eCLIP expt.**
  for >110 RBPs on HepG2 & K562
  (Van Nostrand...Yeo. Nat. Meth. '16;
  Van Nostrand...Graveley, Yeo
  (submitted in relation to ENCODE3))

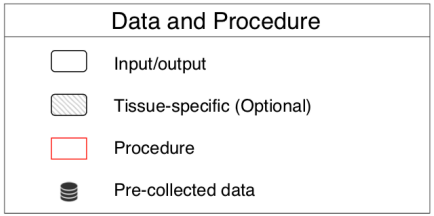


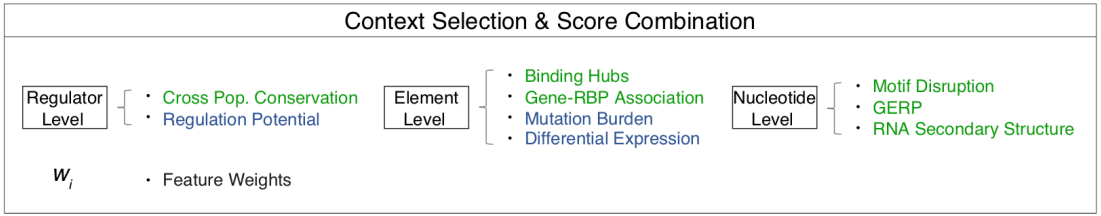


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

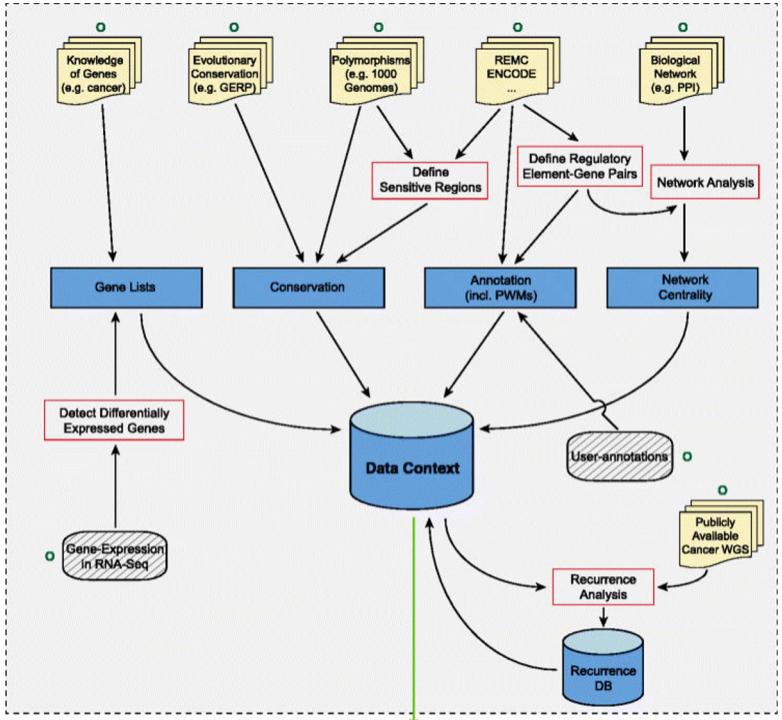
Schematic of RADAR Scoring

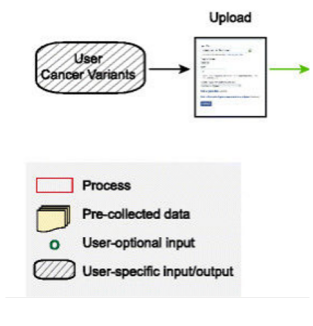


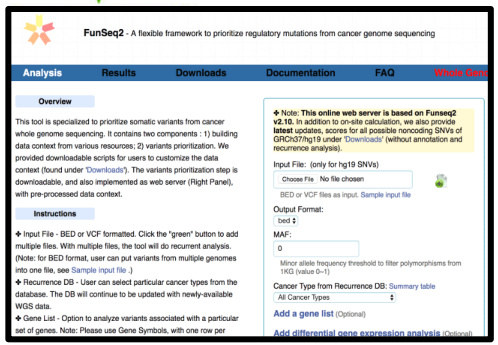




[Zhang*, Liu* et al., *Genome Biology* (in review '18)]







FunSeq.gersteinlab.org

HOT region
Sensitive region
Polymorphisms

Genome

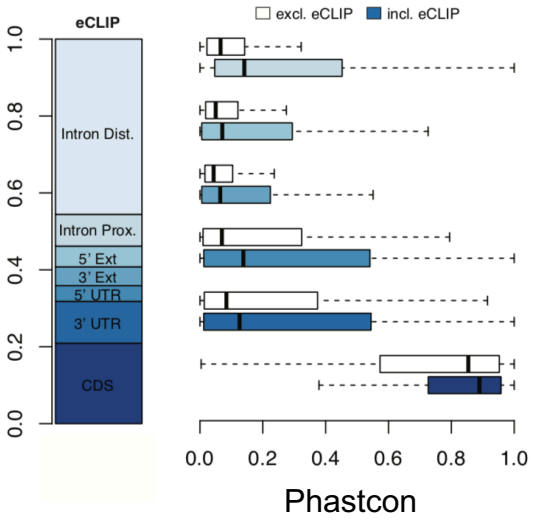
$$w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$$

- Info. theory based method (ie annotation “surprisal”) for weighting consistently many genomic features

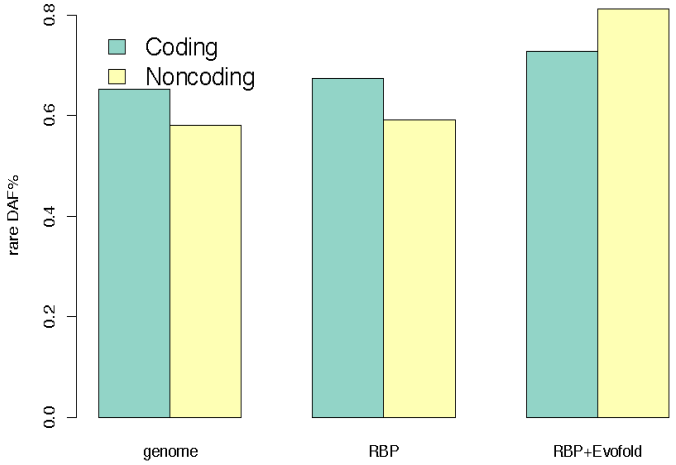
- Practical web server
- Submission of variants & pre-computed large data context from uniformly processing large-scale datasets

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

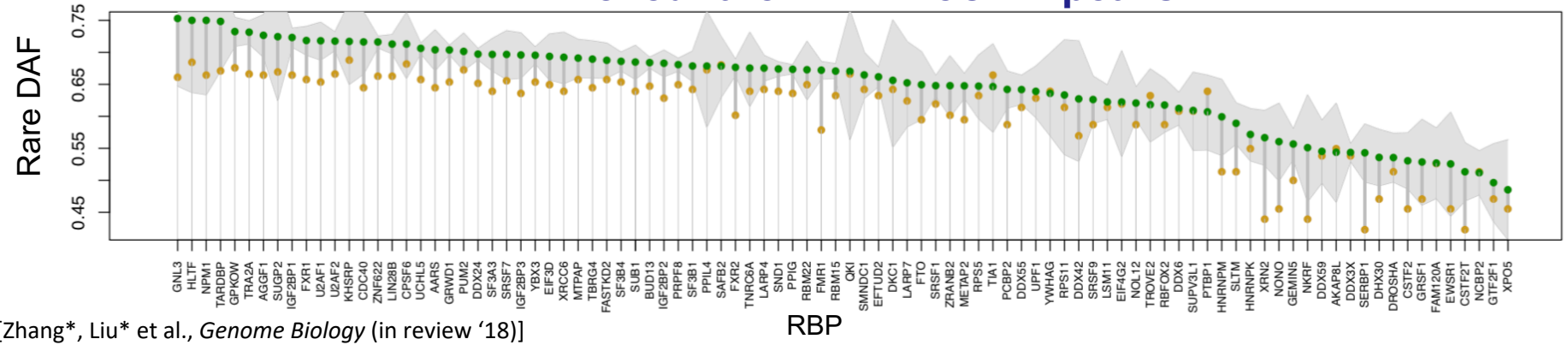
High Phastcon in RBP-overlapped annotations



RNA Structure Cons. from Evofold



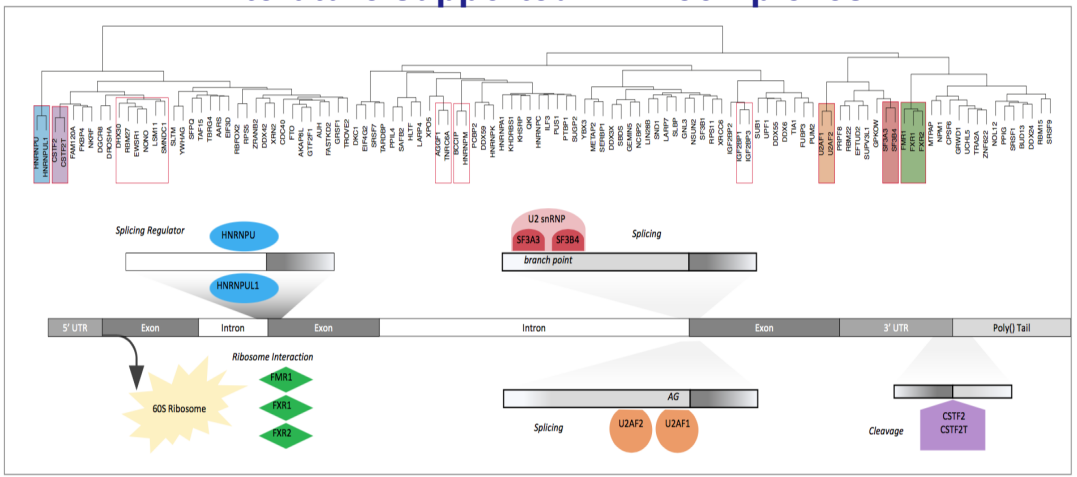
Enriched rare DAF in eCLIP peaks



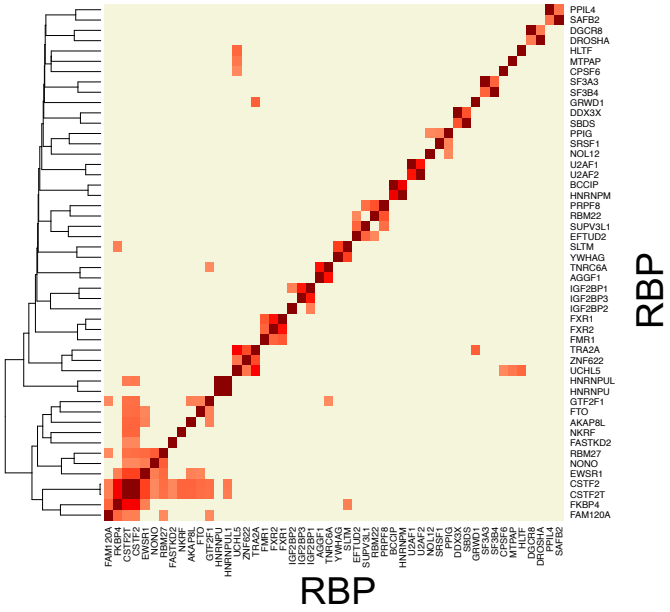
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Co-binding of RBPs form biologically relevant complexes

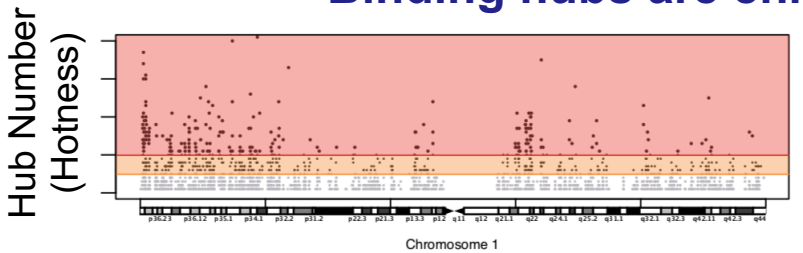
Literature supported RBP complexes

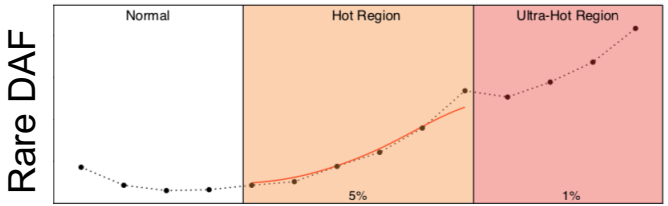


Unique co-binding patterns of RBPs



Binding hubs are enriched for rare variants

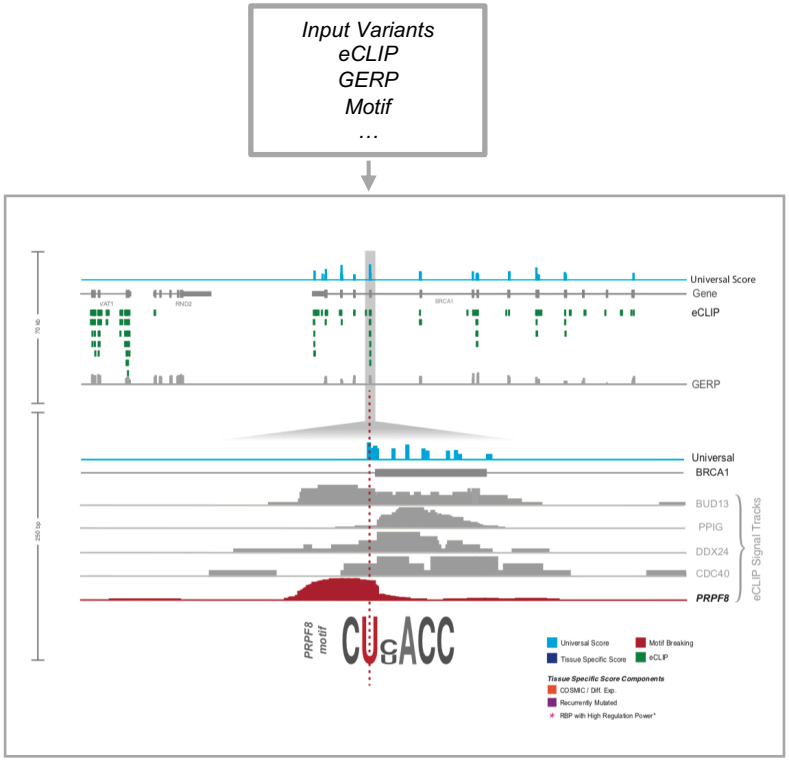




[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Visualization of RADAR Features and Scoring

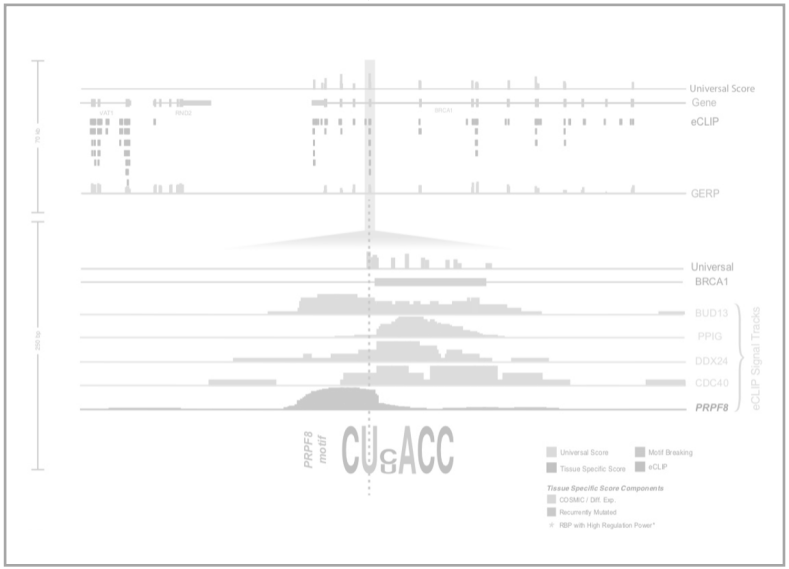
Germline Variants are Score Using a Universal Scoring Scheme



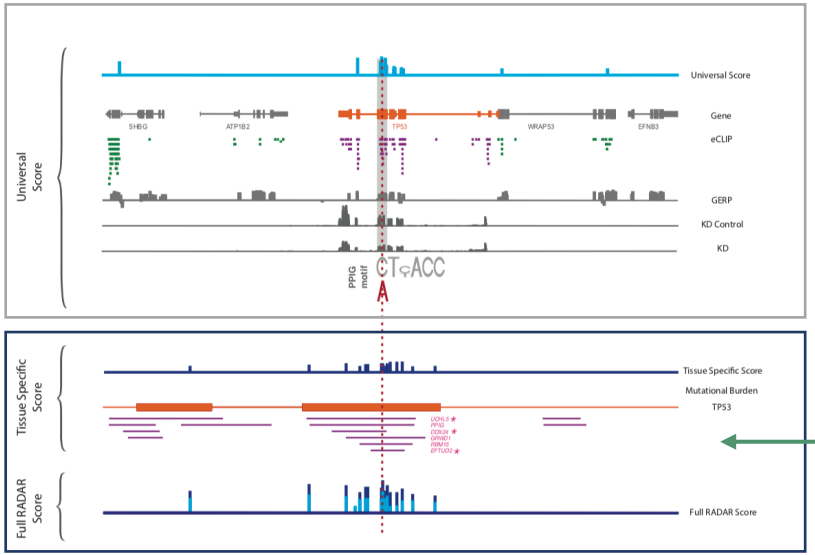
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Visualization of RADAR Features and Scoring

Input Variants
eCLIP
GERP
Motif
…

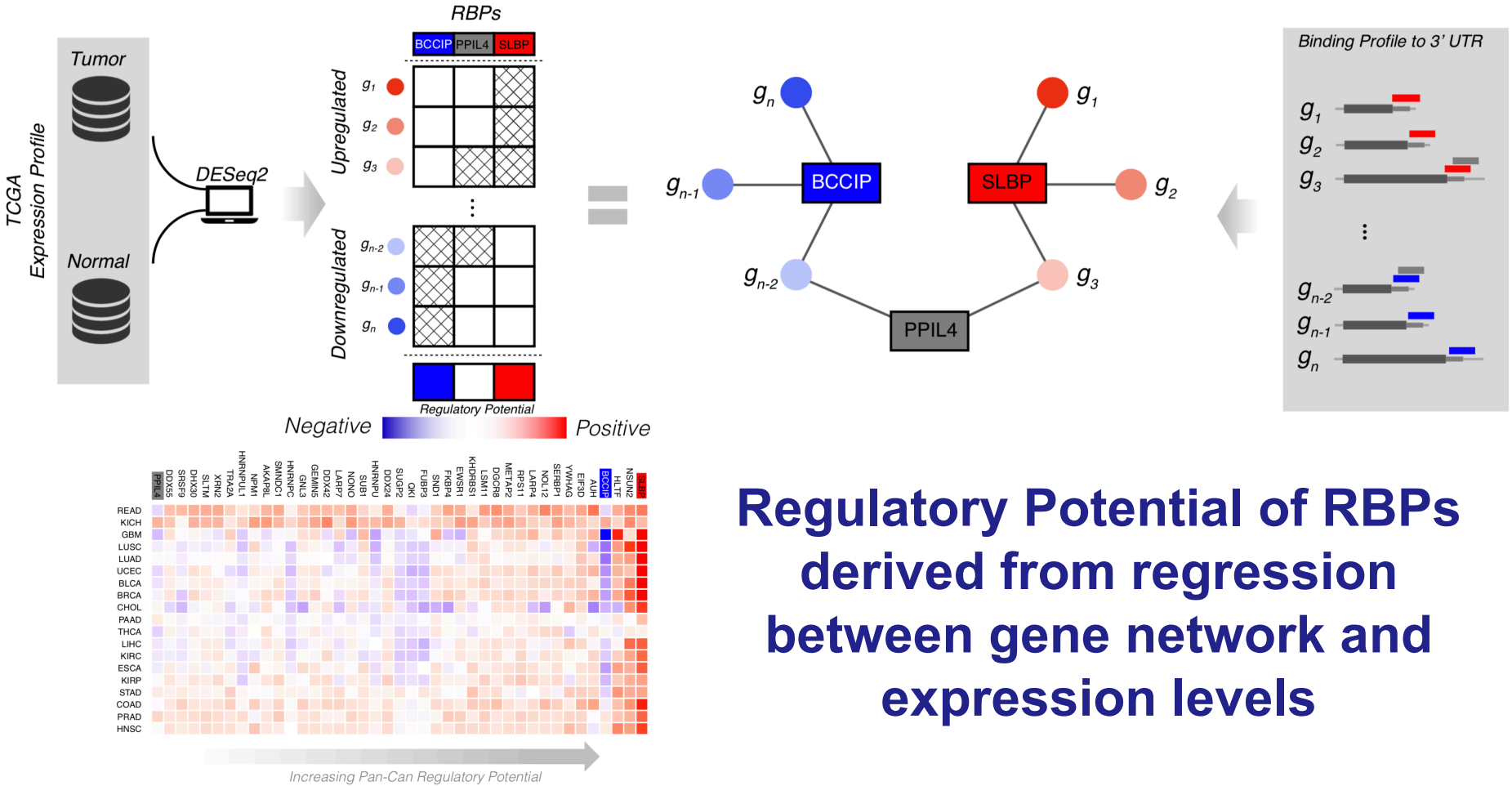


Tissue Specific:
Variants
Expression
Regulatory Potential



Somatic Variant Scored with Universal + Tissue specific context score

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



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

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

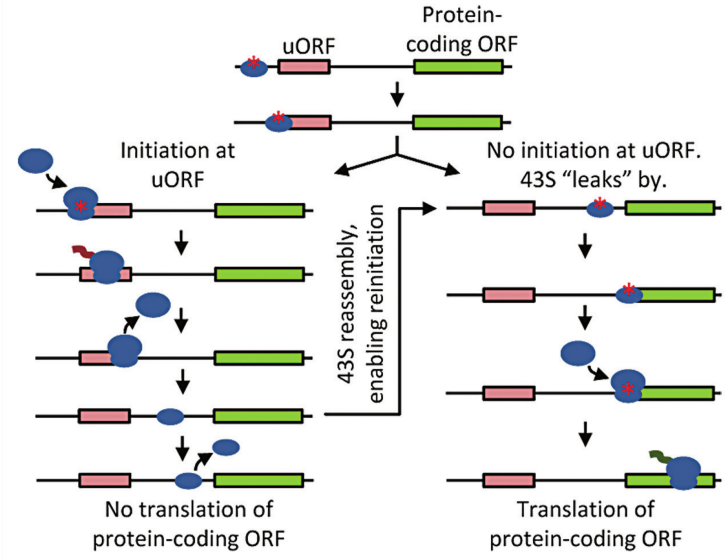
Measuring the Impact of Non-coding Mutations: Repurposing the Additive Effects Model & Developing New Annotations to Assess Variants

- **Additive-Effects model to quantify the Impact of non-coding v coding mutations**
  - Repurposing a formalism in germline genetics for missing heritability to cancer
  - Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  - Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

- **New Annotations to help assess the impact non-coding variants**
  - **RADAR Prioritization for RBP sites**
    - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
    - Incorporates new features related to RNA sec. struc & tissue specific effects
  - **uORF Prioritization**
    - Feature integration to find small subset of upstream mutations that potentially alter translation

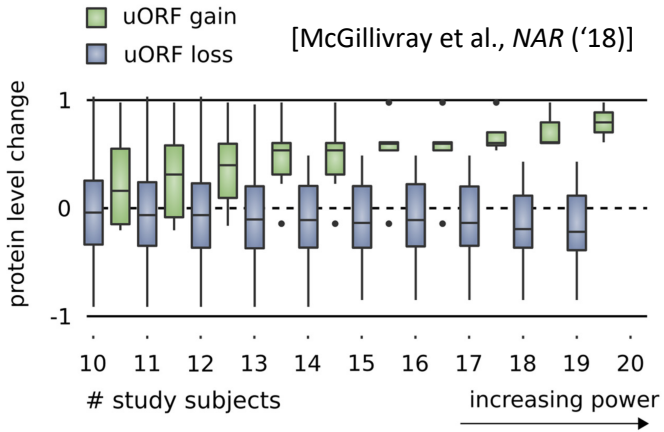
Upstream open reading frames (uORFs) regulate translation are affected by somatic mutation

[Calvo et al., *PNAS* (‘09)]

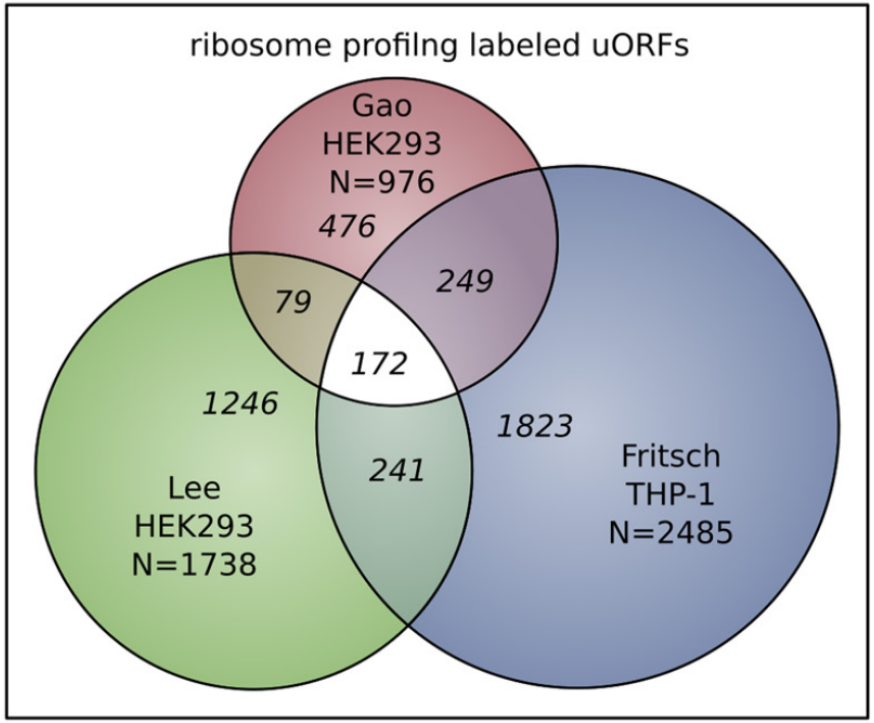


[Ferreira et al., *Bioengineered* (‘14)]

- 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.

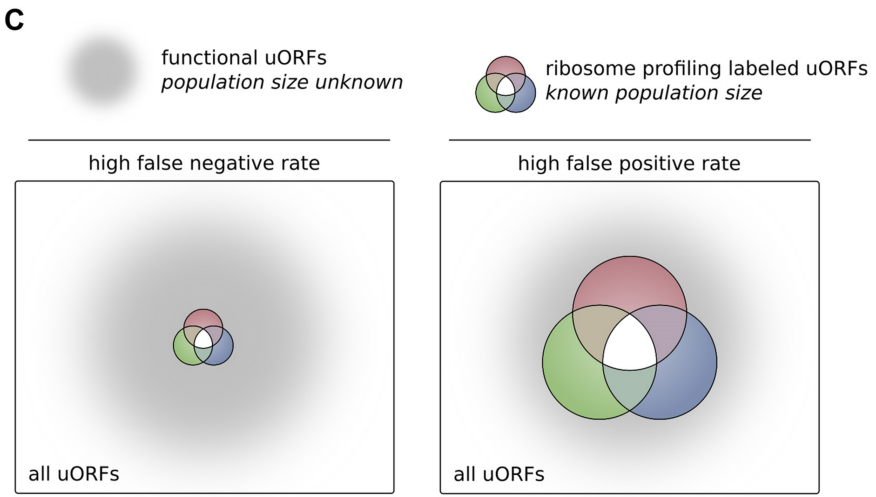


[McGillivray et al., *NAR* (‘18)]



From a “Universe” of 1.3 M pot. uORFs

The population of functional uORFs may be significant

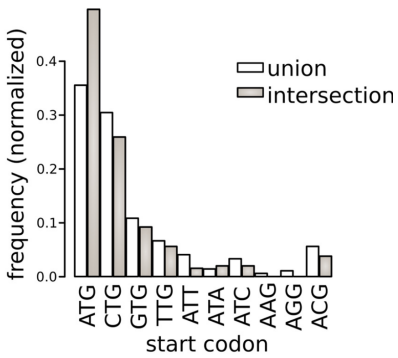


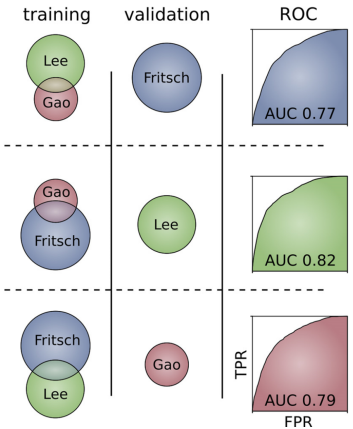
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

[McGillivray et al., *NAR* (‘18)]

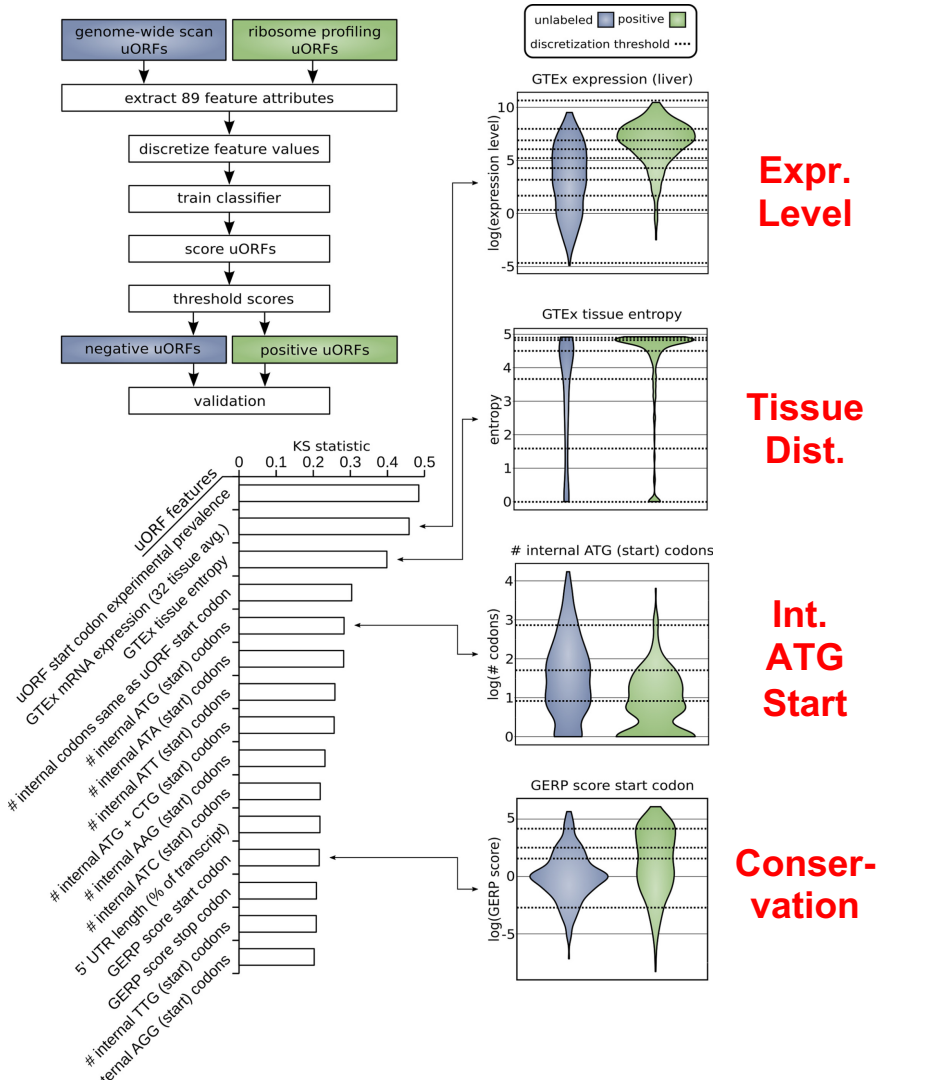
Prediction & validation of functional uORFs using 89 features

- 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).



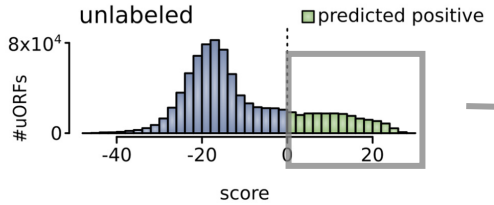


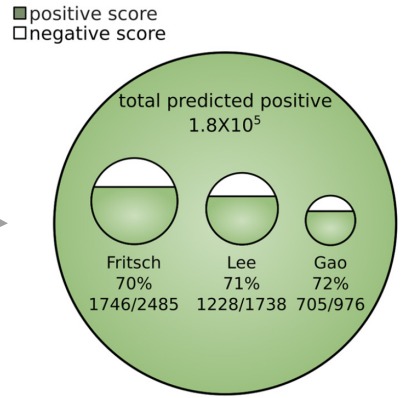
[McGillivray et al., *NAR* ('18)]

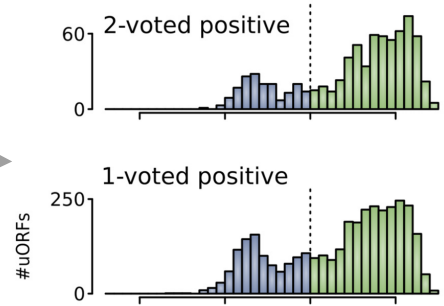


A comprehensive catalog of functional uORFs

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)]

Measuring the Impact of Non-coding Mutations: Repurposing the Additive Effects Model & Developing New Annotations to Assess Variants

- **Additive-Effects model to quantify the Impact of non-coding v coding mutations**
  - Repurposing a formalism in germline genetics for missing heritability to cancer
  - Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  - Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

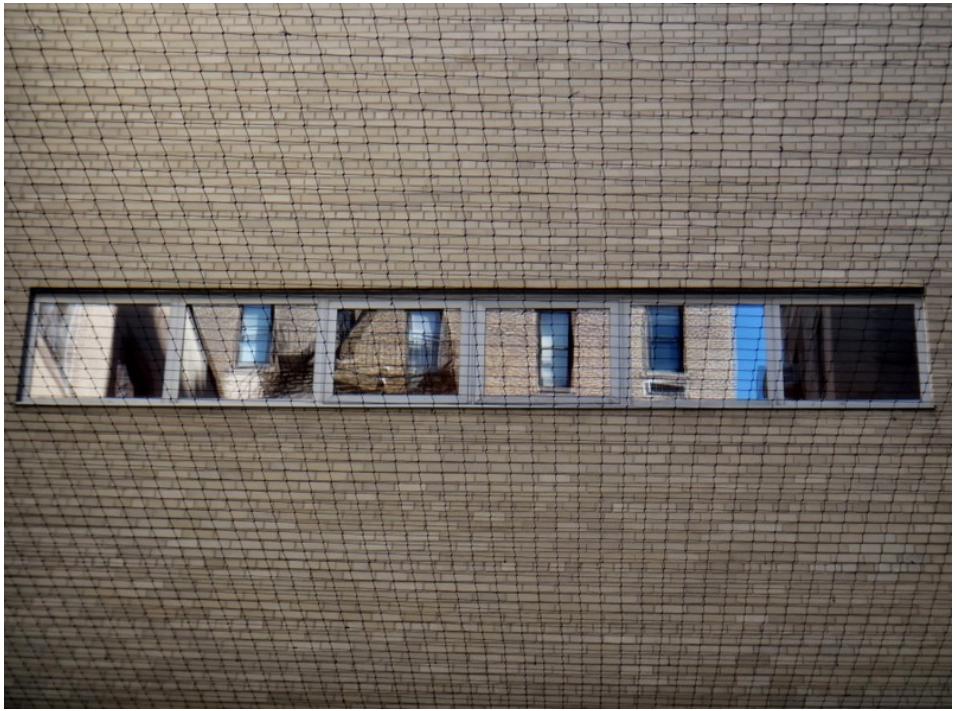
- **New Annotations to help assess the impact non-coding variants**
  - **RADAR Prioritization for RBP sites**
    - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
    - Incorporates new features related to RNA sec. struc & tissue specific effects
  - **uORF Prioritization**
    - Feature integration to find small subset of upstream mutations that potentially alter translation

Measuring the Impact of Non-coding Mutations: Repurposing an Additive Effects Model & Developing New Annotations

- **Additive-Effects model to measure the Impact of non-coding v coding mutations**
  - Repurposing a formalism from germline genetics for missing heritability to cancer
  - Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  - Notable effect, particularly for non-coding passengers, in addition to known coding drivers.
  - Recasting as a predictive model to est. number of weak drivers

- **New Annotations to help assess the impact non-coding variants**
  - **RADAR Prioritization for RBP sites**
    - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
    - Incorporates new annotation features related to RNA sec. struc & tissue specific effects
  - **uORF Prioritization**
    - Feature integration to find small subset of upstream mutations that potentially alter translation

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



PCAWG.gersteinlab.org

S **Kumar**, J **Warrell**, S Li, P McGillivray, W Meyerson, L Salichos, A Harmanci, A Martinez-Fundichely, C Chan, M Nielsen, L Lochovsky, Y Zhang, X Li, S Lou, J Skou Pedersen, C H, G Getz, E Khurana

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


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, 2019.
- 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