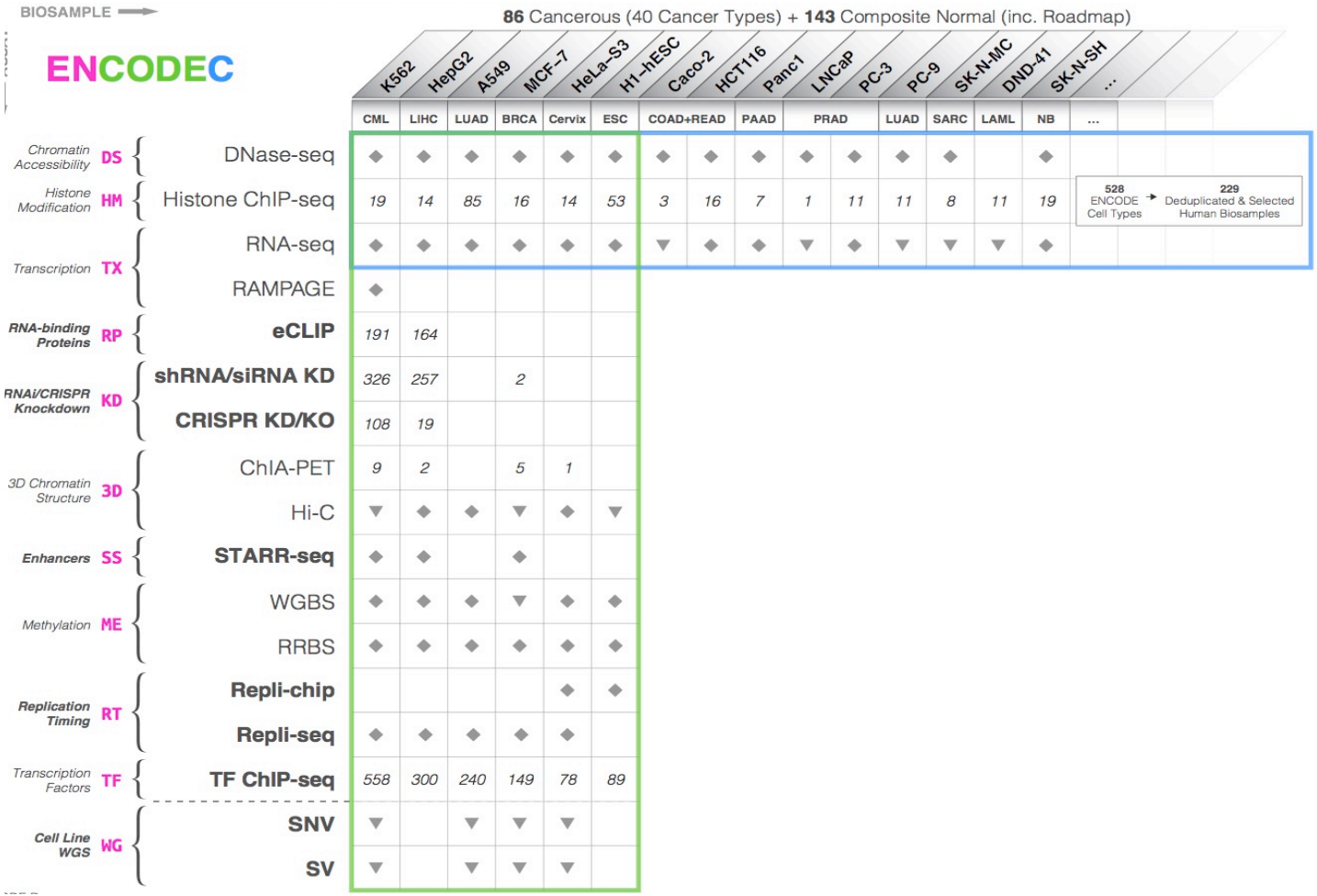
Using ENCODE Data for Cancer Genomics

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

http://encodec.encodeproject.org/



[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

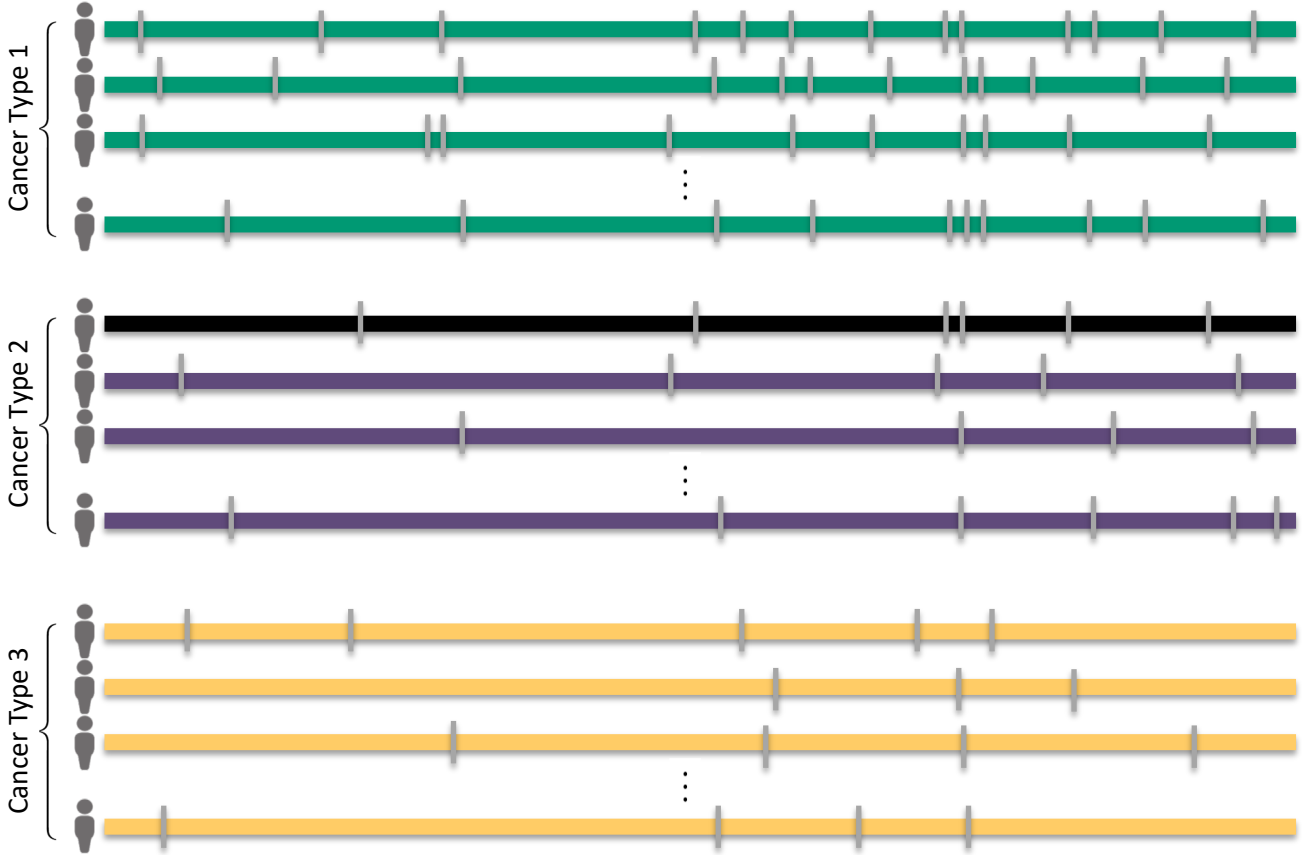
Using ENCODE Data for Cancer Genomics

- **BMR Correction: LARVA/MOAT/NIMBUS**
  - Parametric models explicitly modeling genomic covariates
  - Many ENCODE covariates useful in accurately estimating background mutation rate
- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data in certain cell lines highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
- **RADAR Variant Prioritization**
  - Prioritizes germline & somatic variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

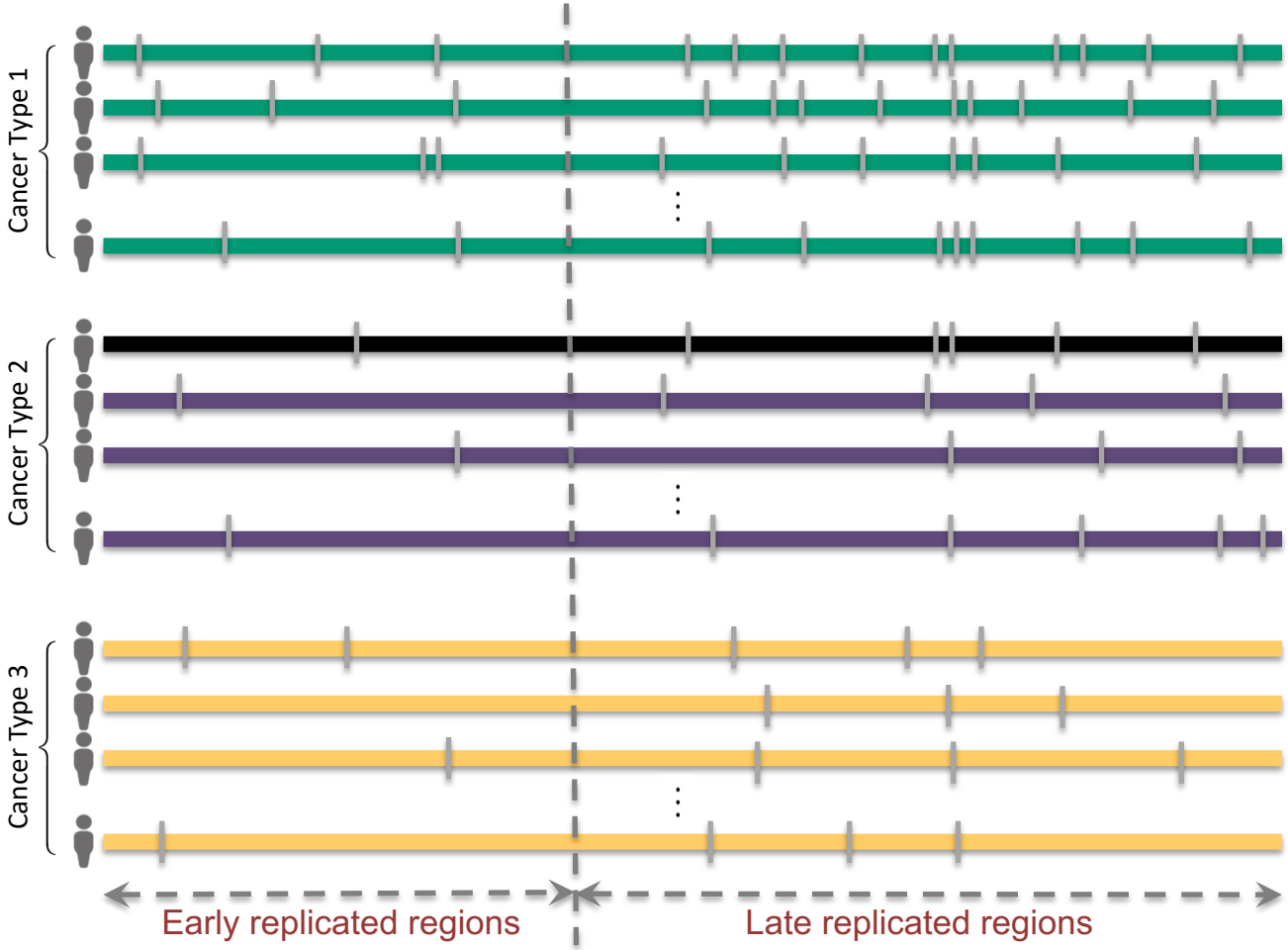
Using ENCODE Data for Cancer Genomics

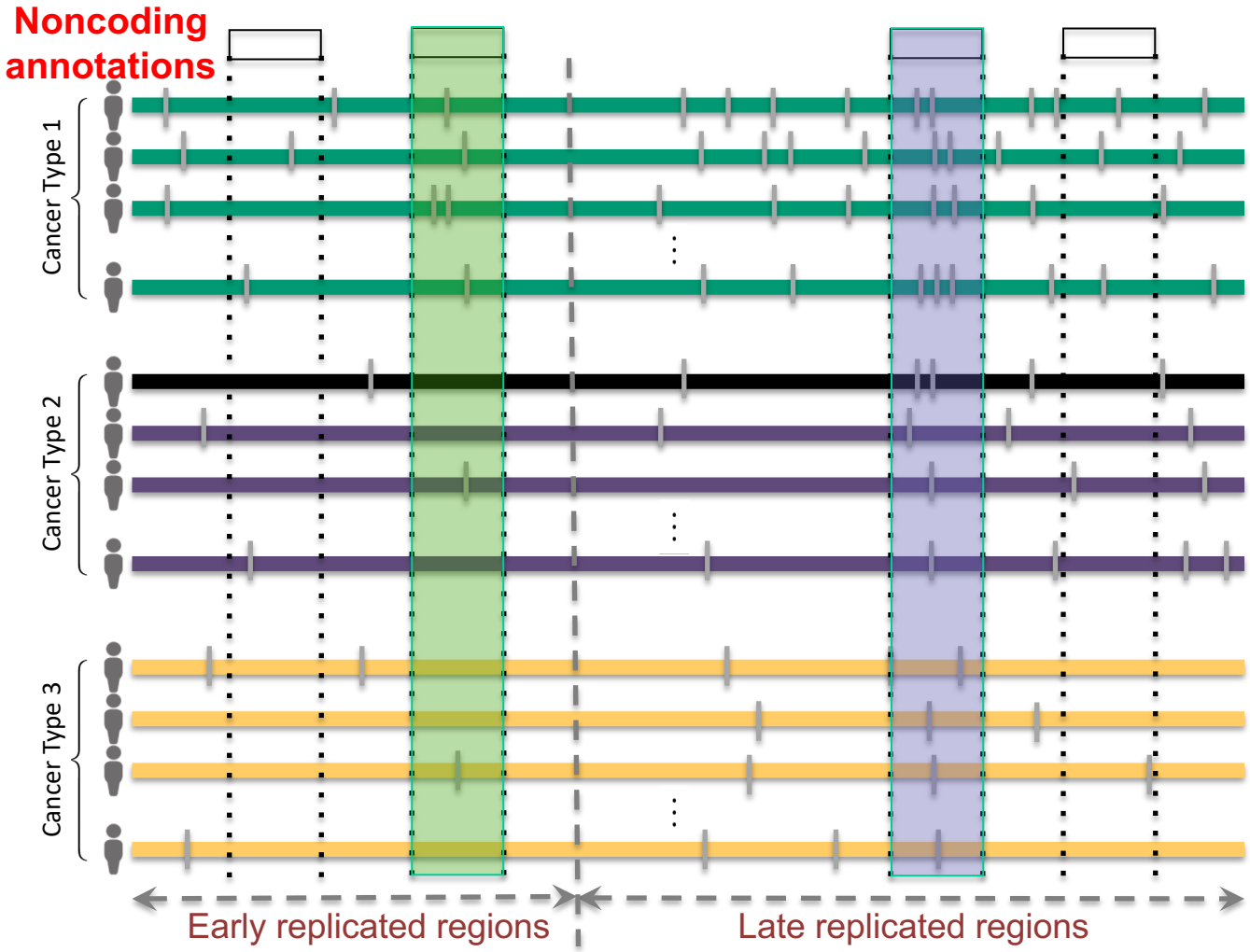
- **BMR Correction: LARVA/MOAT/NIMBUS**
  - Parametric models explicitly modeling genomic covariates
  - Many ENCODE covariates useful in accurately estimating background mutation rate
- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data in certain cell lines highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
- **RADAR Variant Prioritization**
  - Prioritizes germline & somatic variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

Mutation recurrence

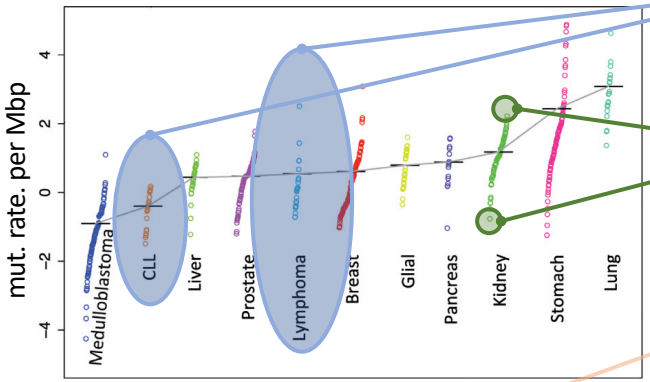


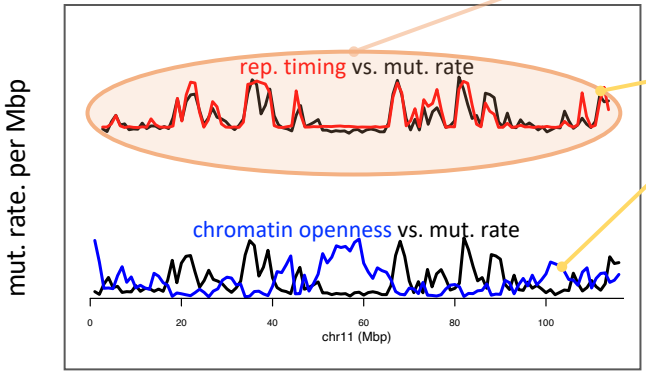
Mutation recurrence




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

violation of the constant mutation rate assumption





mutation rate changes across tumor

within one tumor type

mutation rate changes across patients

within one tumor type
within one patient

mutation rate changes across regions

within one tumor type
within one patient

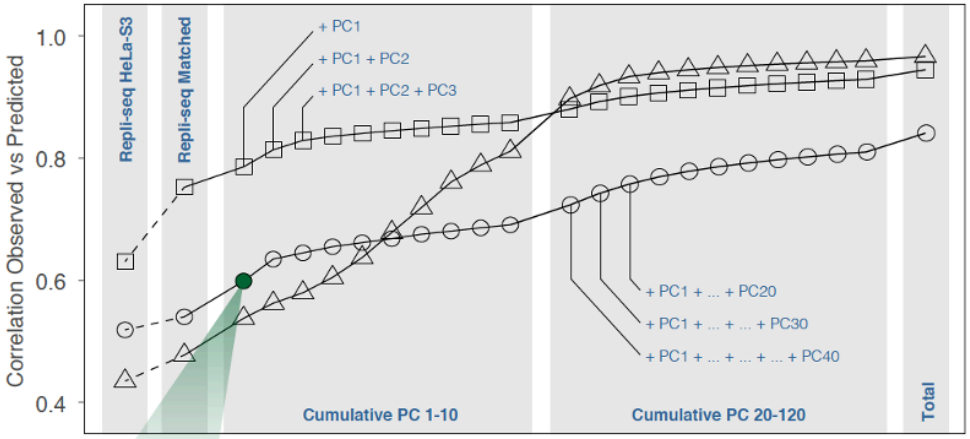
mutation rate changes with many covariates

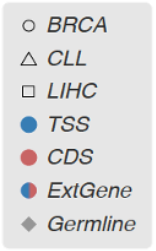
inappropriate models

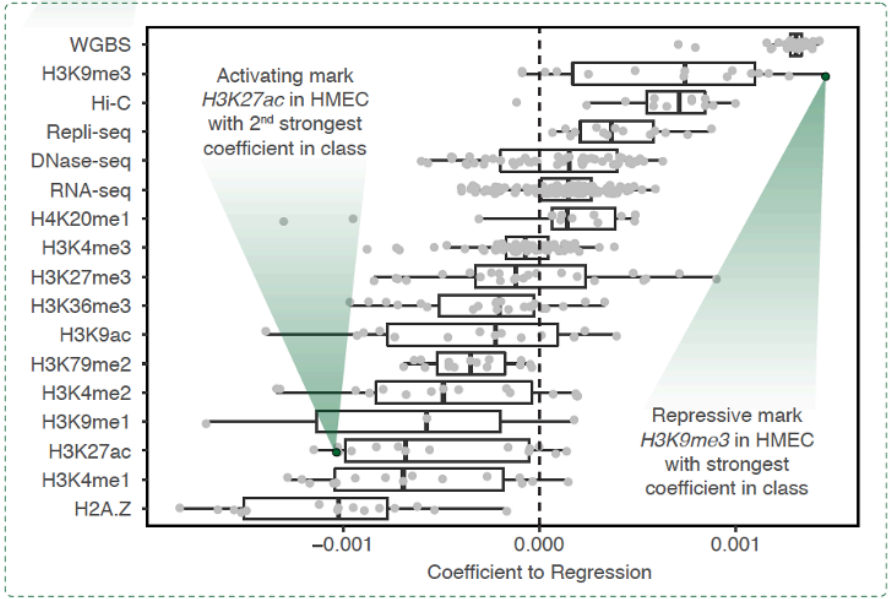
Bad data fitting

Inaccurate burden test results

[Lochovsky et al. *NAR* ('15); Zhang et al. biorxiv + BMC Bioinfo ('20), in press]







Accurately modeling background mutation rate with full spectrum of ENCODE data

[Zhang et al. *Nat. Comm.* ('20);
Zhang et al. biorxiv + BMC Bioinfo ('20), in press]

Cancer Somatic Mutation Modeling

PARAMETRIC MODELS (LARVA/NIMBUS)

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

$x_i : Binomial(n_i, p)$

Model 2: Varying Mutation Rate with Covariate Correction (Beta Binom.)

$x_i : Binomial(n_i, p_i)$

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

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

Model 3: Varying Mutation Rate with Covariate Correction (Neg. Binom.)

$x_i|p_i \sim Pois(p_i)$

$p_i \sim gamma(\mu_i, \theta_i)$

$\log(\mu_i) \sim \beta_0 + \beta_1 v_1 + \cdots + \beta_k v_k$

[Zhang et al. biorxiv + BMC Bioinfo. ('20); Lochovsky et al. *NAR* ('15)]

- Suppose there are *L* 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 & v_k : covariates
- 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 (MOAT)

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.

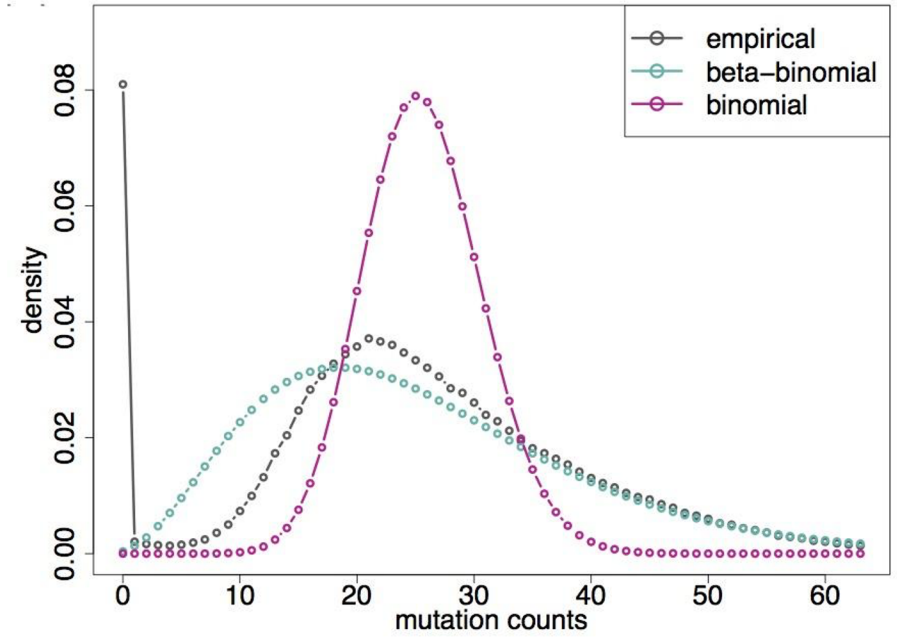
Model 3b: Random Permutation of Input Variants

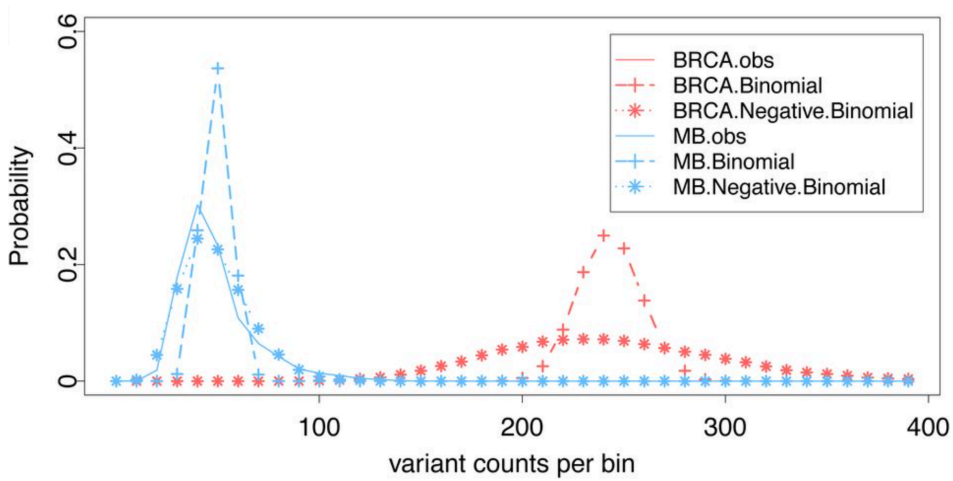
Shuffle variants within local region to assess background mutation rate.

[Lochovsky et al. *Bioinformatics* ('17)]

LARVA/NIMBUS Model Comparison

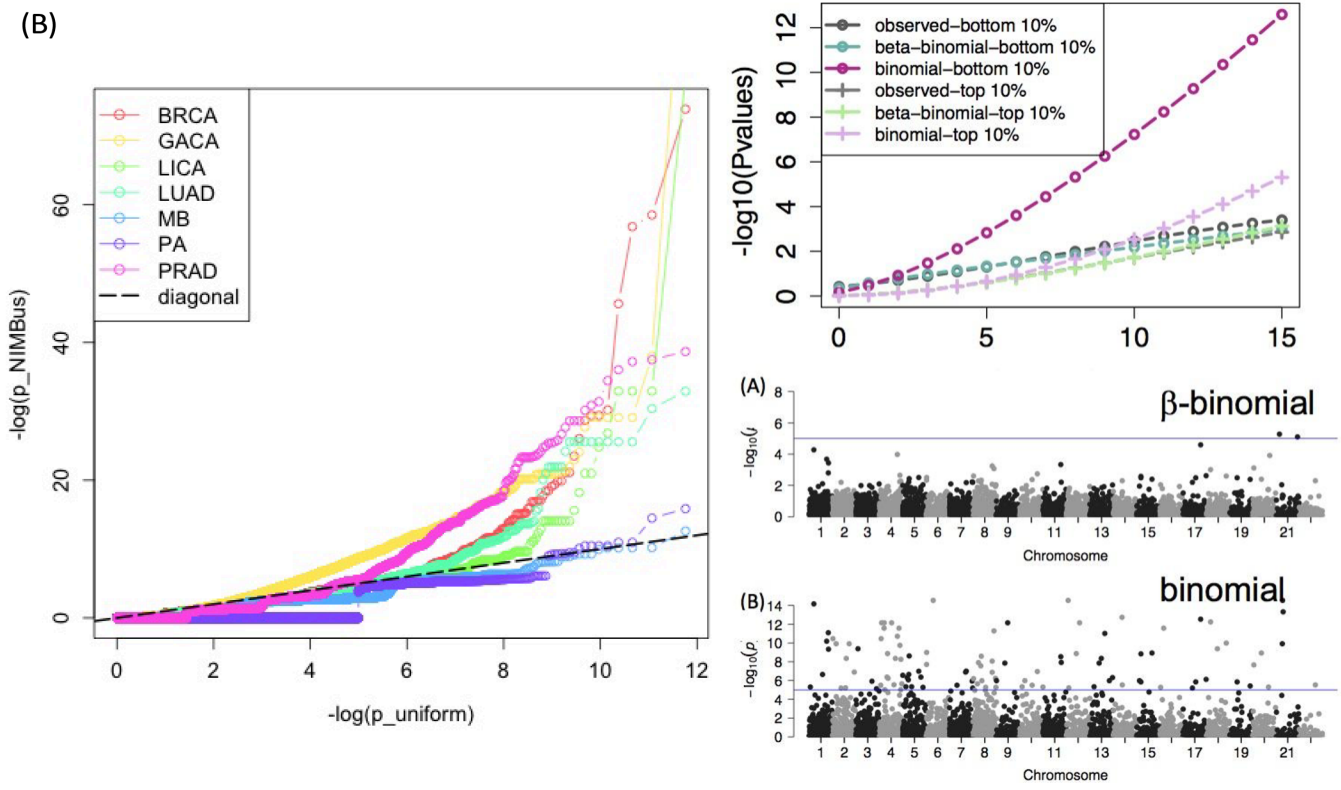
- Comparison of mutation count frequency implied by the binomial model (model 1) and the beta-binomial model (model 2) relative to the empirical distribution
- The beta-binomial/negative binomial distribution is significantly better, especially for accurately modeling the over-dispersion of the empirical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





[Zhang et al. biorxiv + BMC Bioinfo. ('20); Lochovsky et al. *NAR* ('15)]

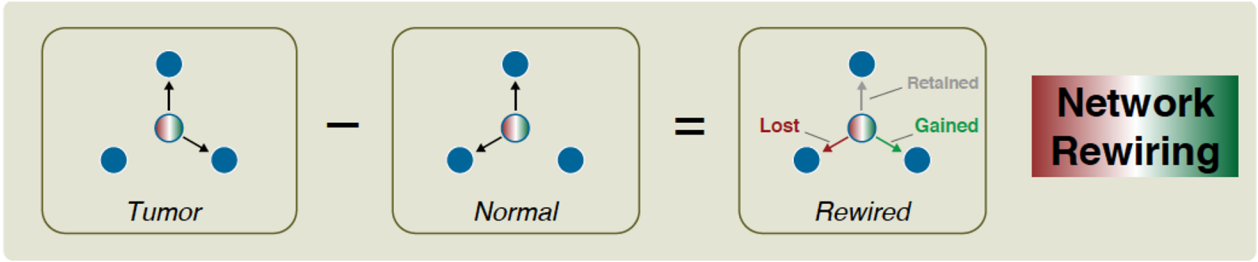
LARVA/NIMBUS Results: Reducing P-value inflation



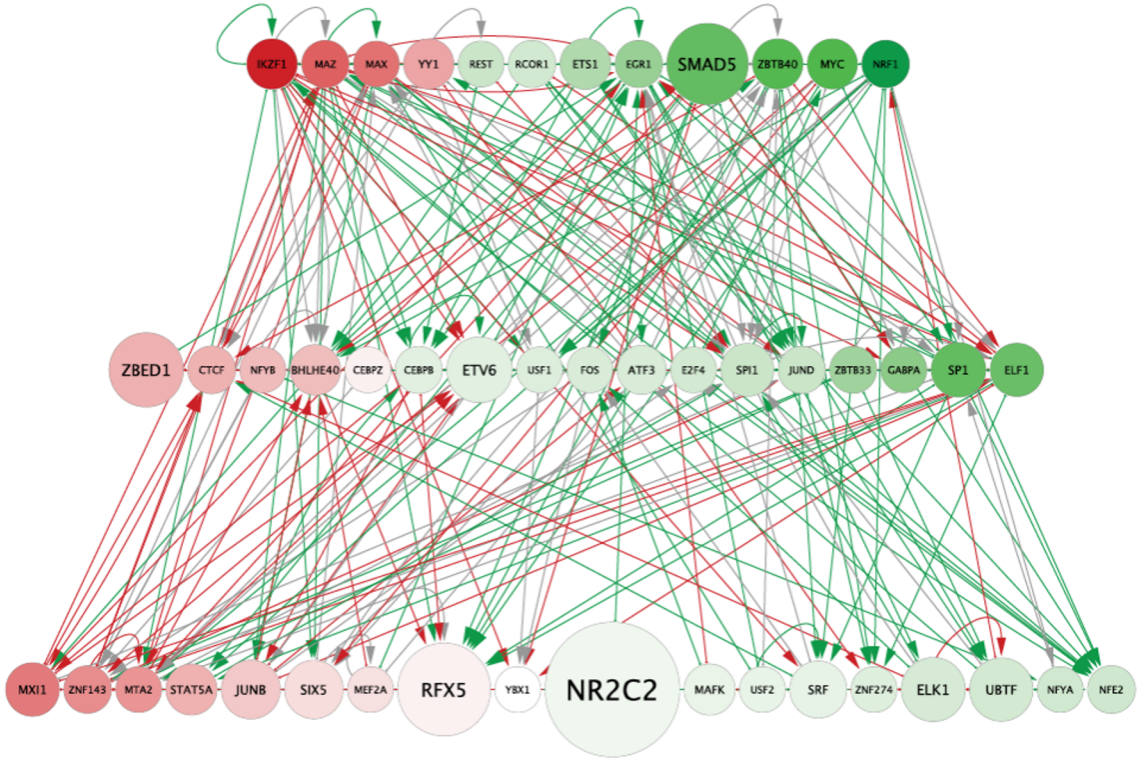
[Zhang et al. biorxiv + BMC Bioinfo. ('20); Lochovsky et al. *NAR* ('15)]

Using ENCODE Data for Cancer Genomics

- **BMR Correction:**
  LARVA/MOAT/NIMBUS
  - Parametric models explicitly modeling genomic covariates
  - Many ENCODE covariates useful in accurately estimating background mutation rate
- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data in certain cell lines highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
- **RADAR Variant Prioritization**
  - Prioritizes germline & somatic variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1



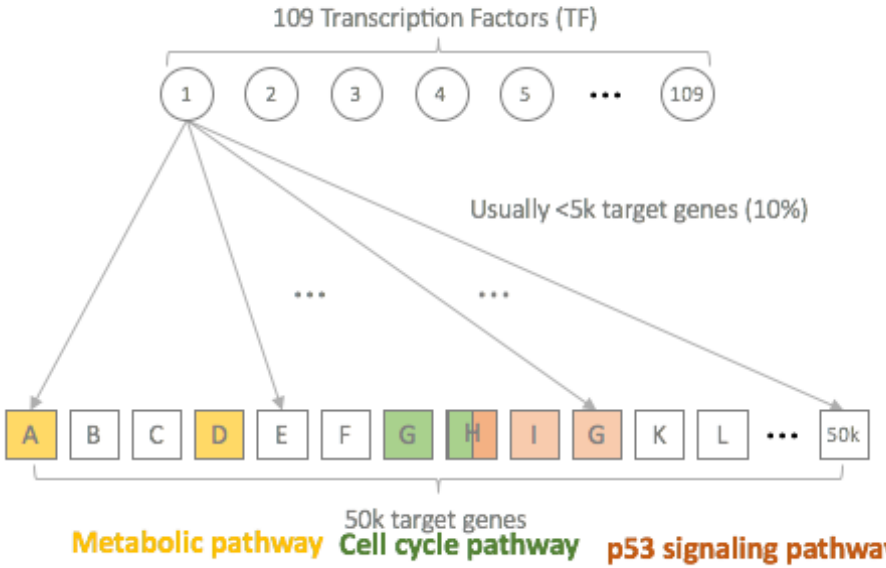
Rewired edges in comparison of GM12878 to K562 109 node TF-TF network (approx. CML)

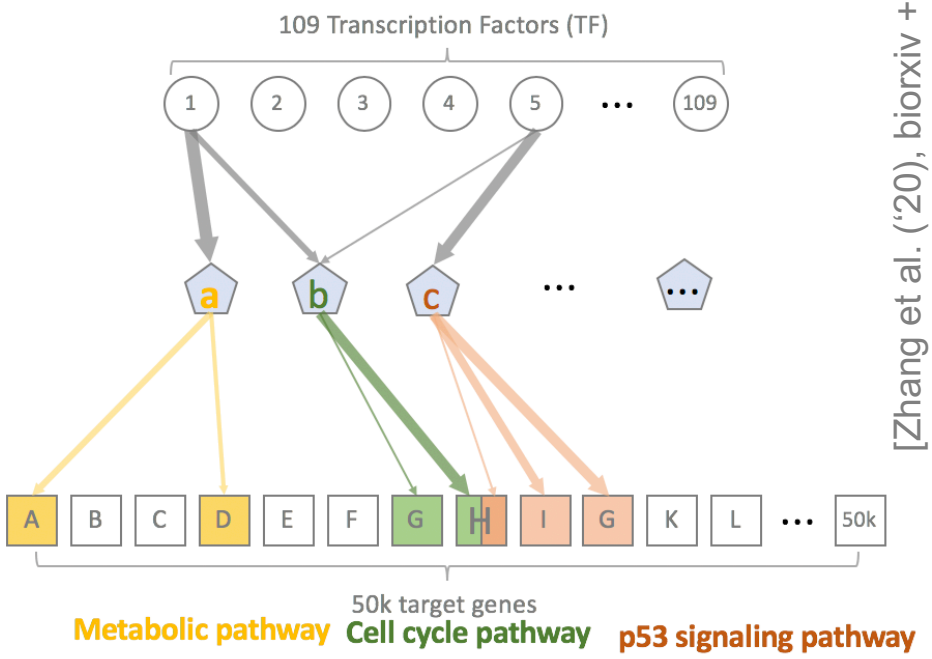


[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

Simplifying Network Rewiring

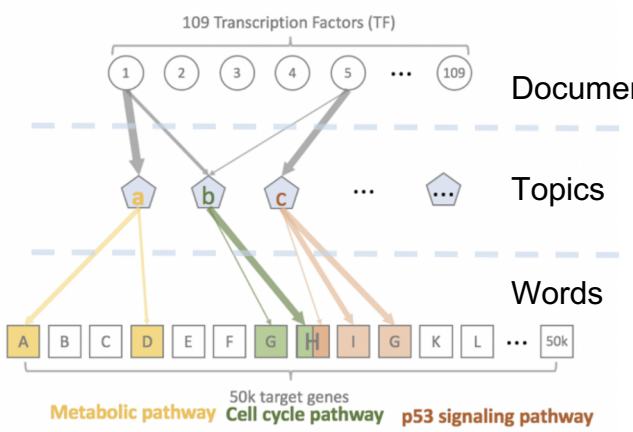
From $TF \rightarrow gene$ $(109\times50{,}000)$
to $TF \rightarrow pathway$ (109×50)

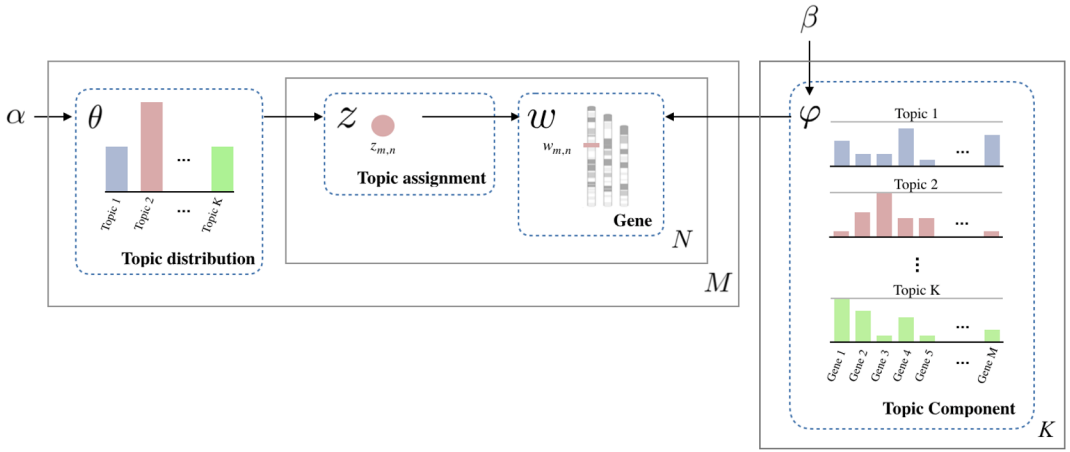


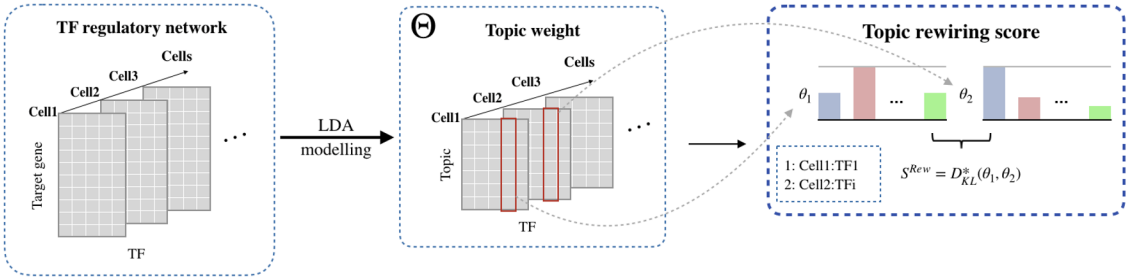


[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

TopicNet: Measuring transcriptional regulatory network change using LDA







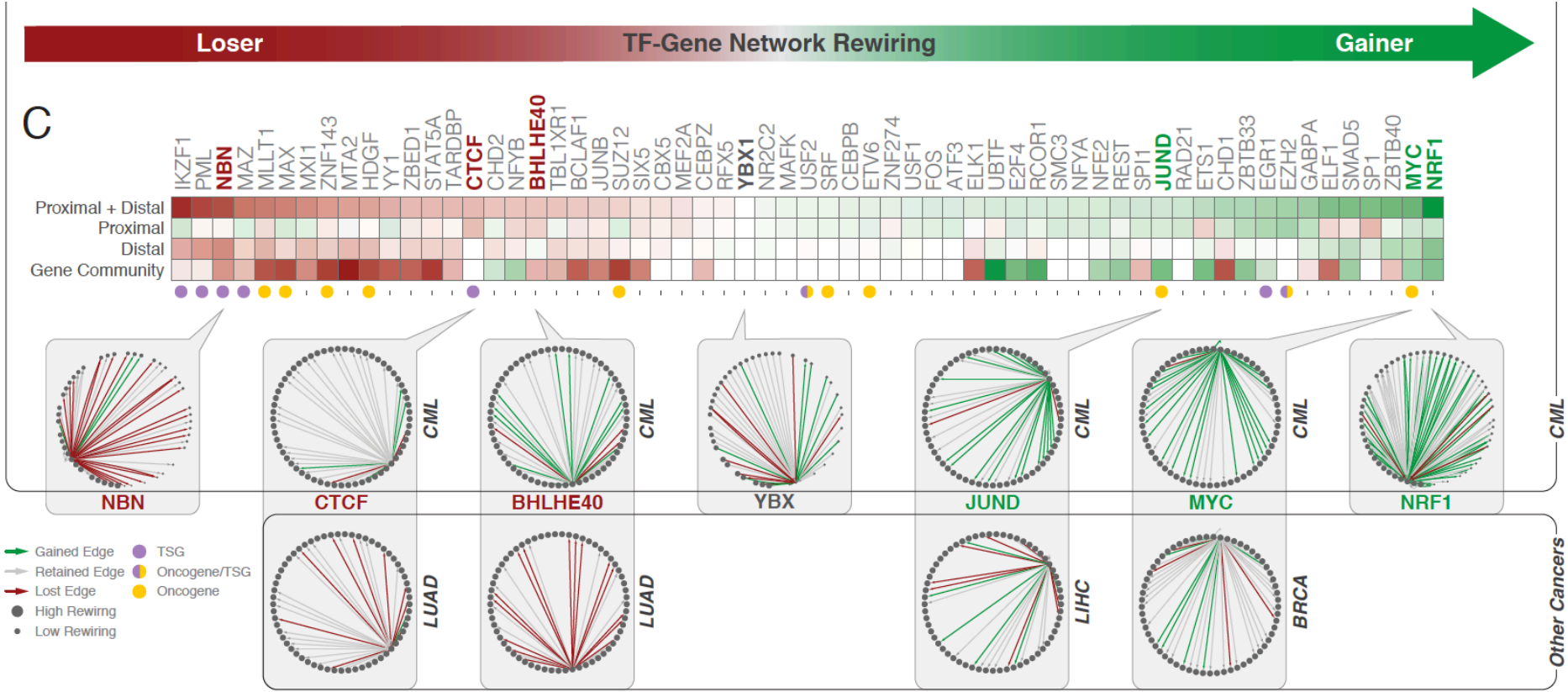
α Prior info

β Prior info

θ: topic distribution per document

φ: word distribution per topic

[Lou et al. bioxriv + Bioinformatics ('20)]

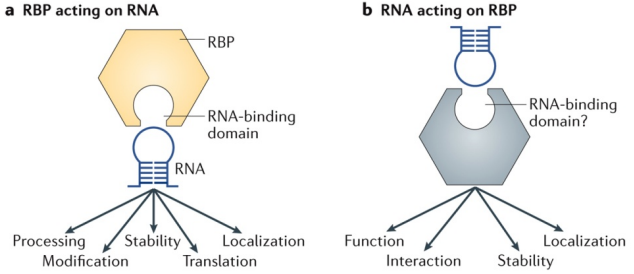


[Zhang et al. (‘20), biorxiv + Nat. Comm. (in press)]

Using ENCODE Data for Cancer Genomics

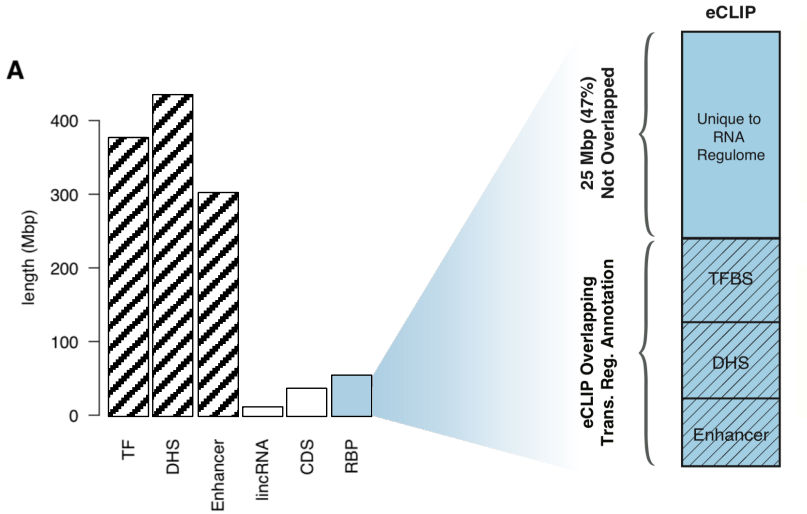
- **BMR Correction:**
  LARVA/MOAT/NIMBUS
  - Parametric models explicitly modeling genomic covariates
  - Many ENCODE covariates useful in accurately estimating background mutation rate
- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data in certain cell lines highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
- **RADAR Variant Prioritization**
  - Prioritizes germline & somatic variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

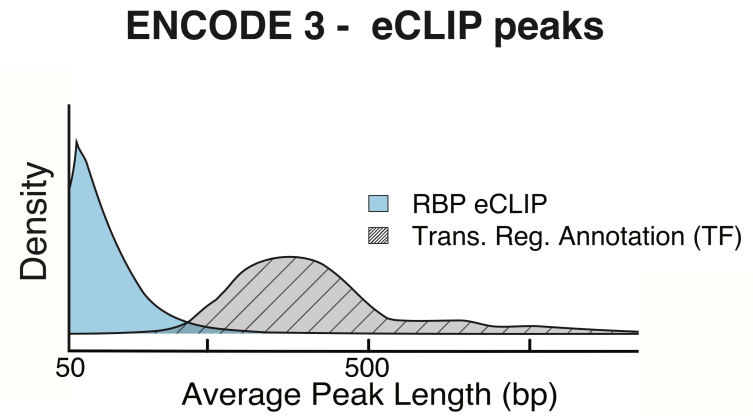
RNA Binding Proteins (RBPs)



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

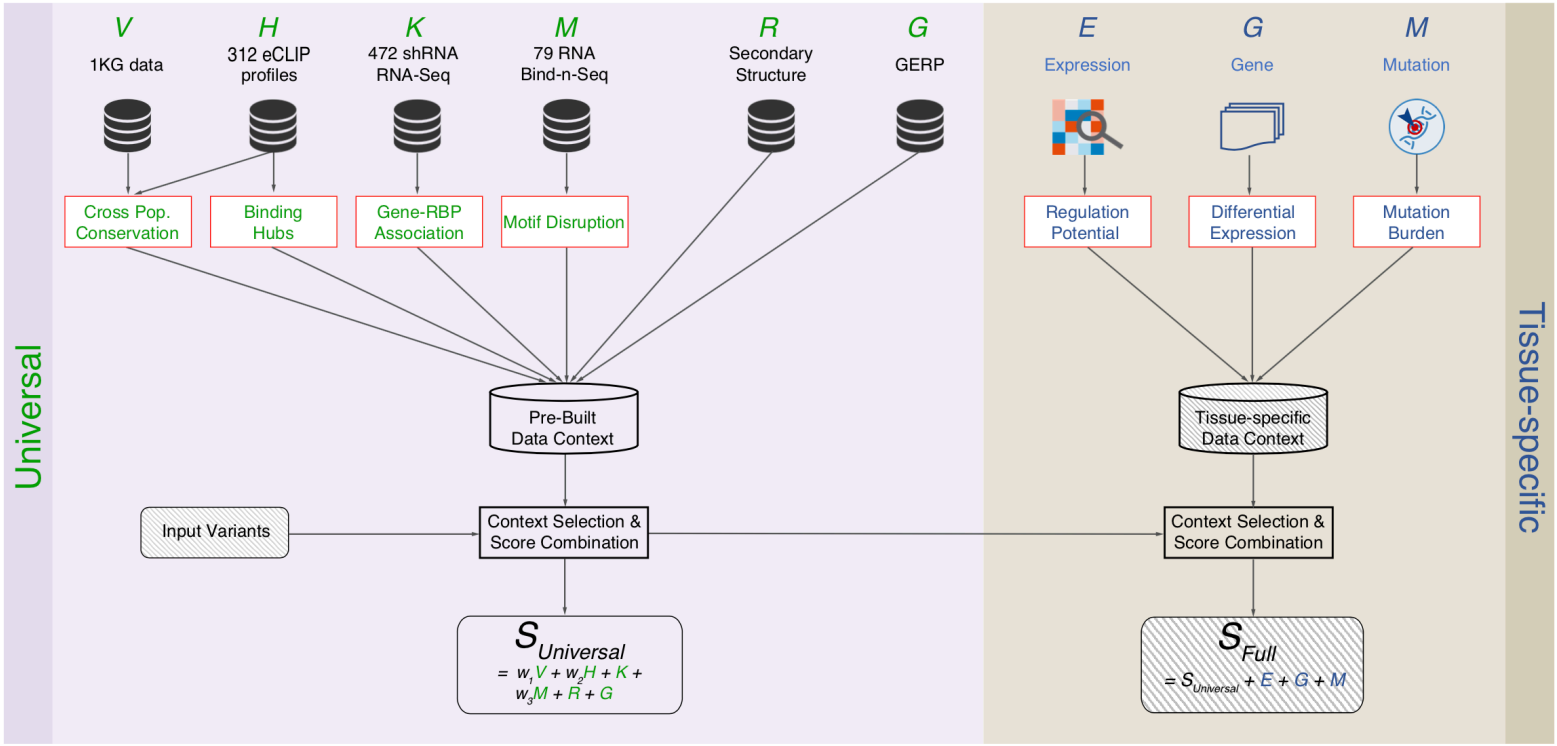
- **Before ENCODE3: >150 expt.**
  in many different cell types
- **ENCODE3 did ~350 focused eCLIP expt.**
  for >110 RBPs on HepG2 & K562
  (Van Nostrand...Yeo. Nat. Meth. '16;
  Van Nostrand...Graveley, Yeo
  (submitted in relation to ENCODE3))

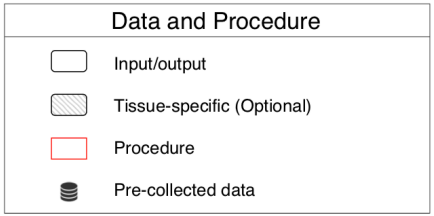


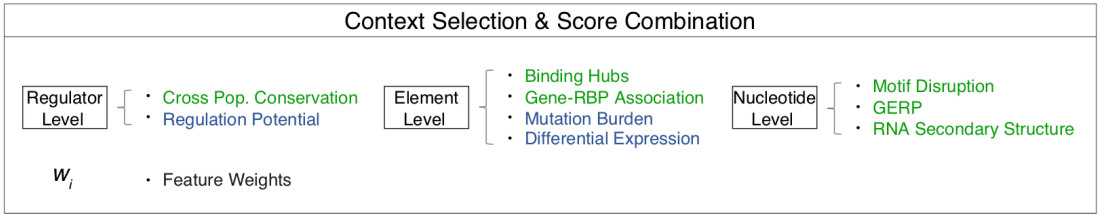


[Zhang*, Liu* et al., *Genome Biology* '20]

Schematic of RADAR Scoring

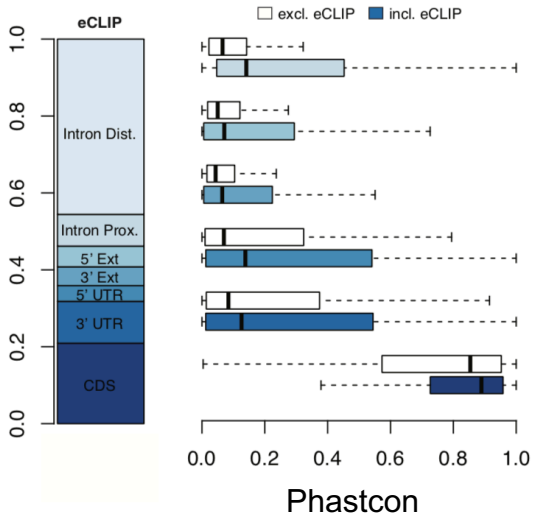




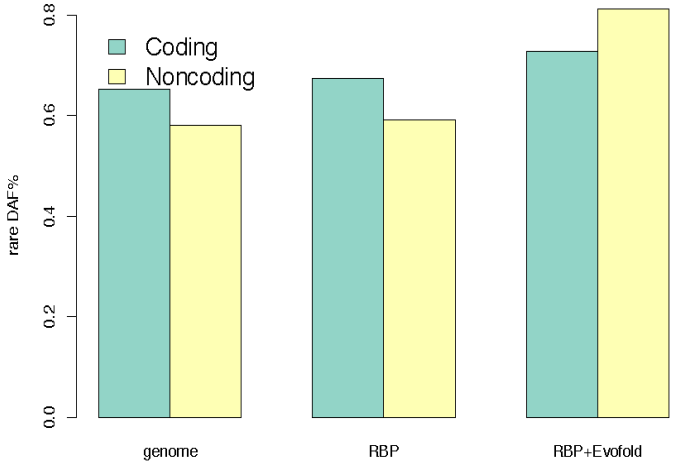


[Zhang*, Liu* et al., *Genome Biology* '20]

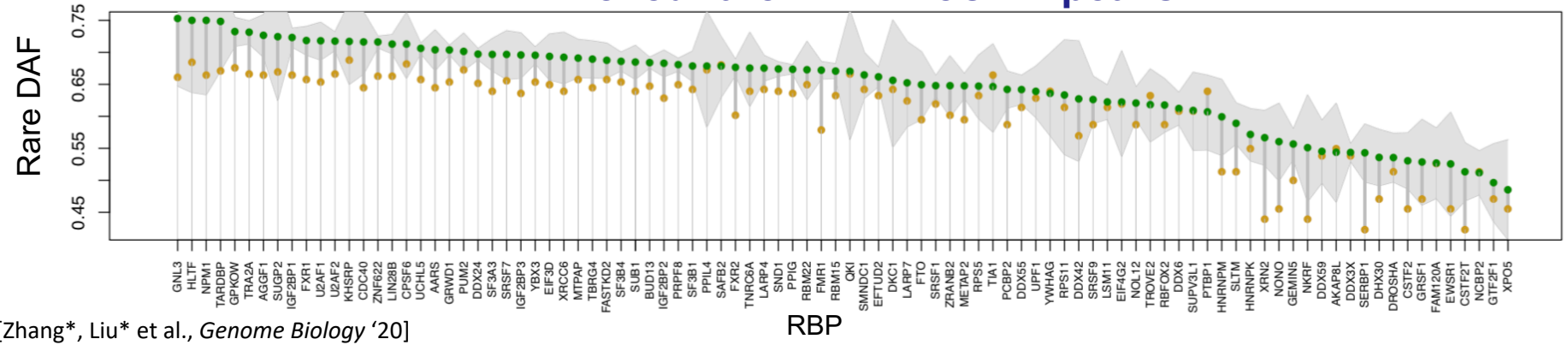
High Phastcon in RBP-overlapped annotations



RNA Structure Cons. from Evofold



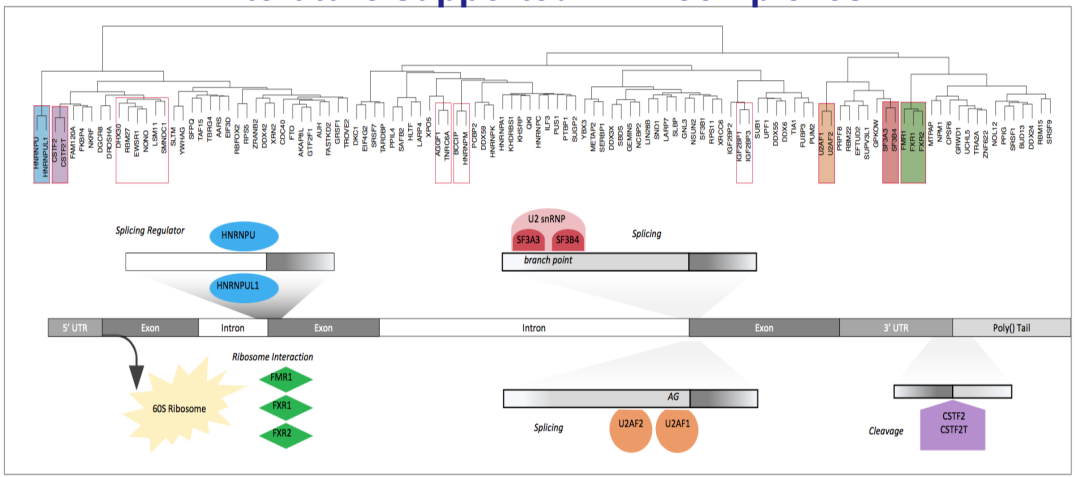
Enriched rare DAF in eCLIP peaks



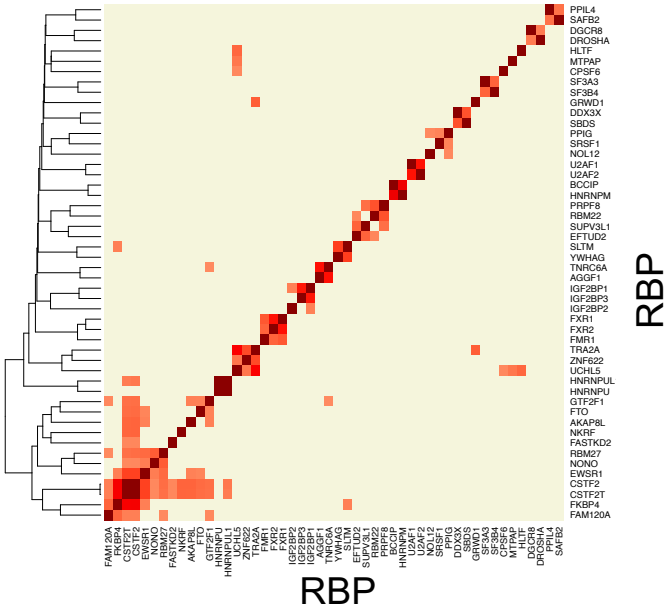
[Zhang*, Liu* et al., *Genome Biology* '20]

Co-binding of RBPs form biologically relevant complexes

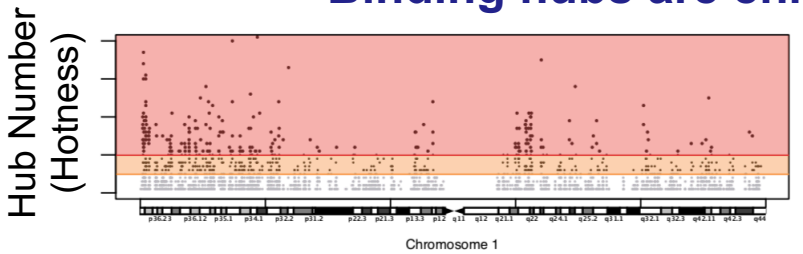
Literature supported RBP complexes

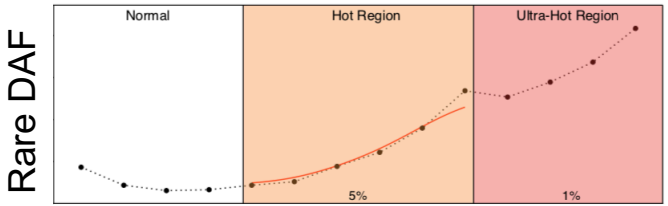


Unique co-binding patterns of RBPs



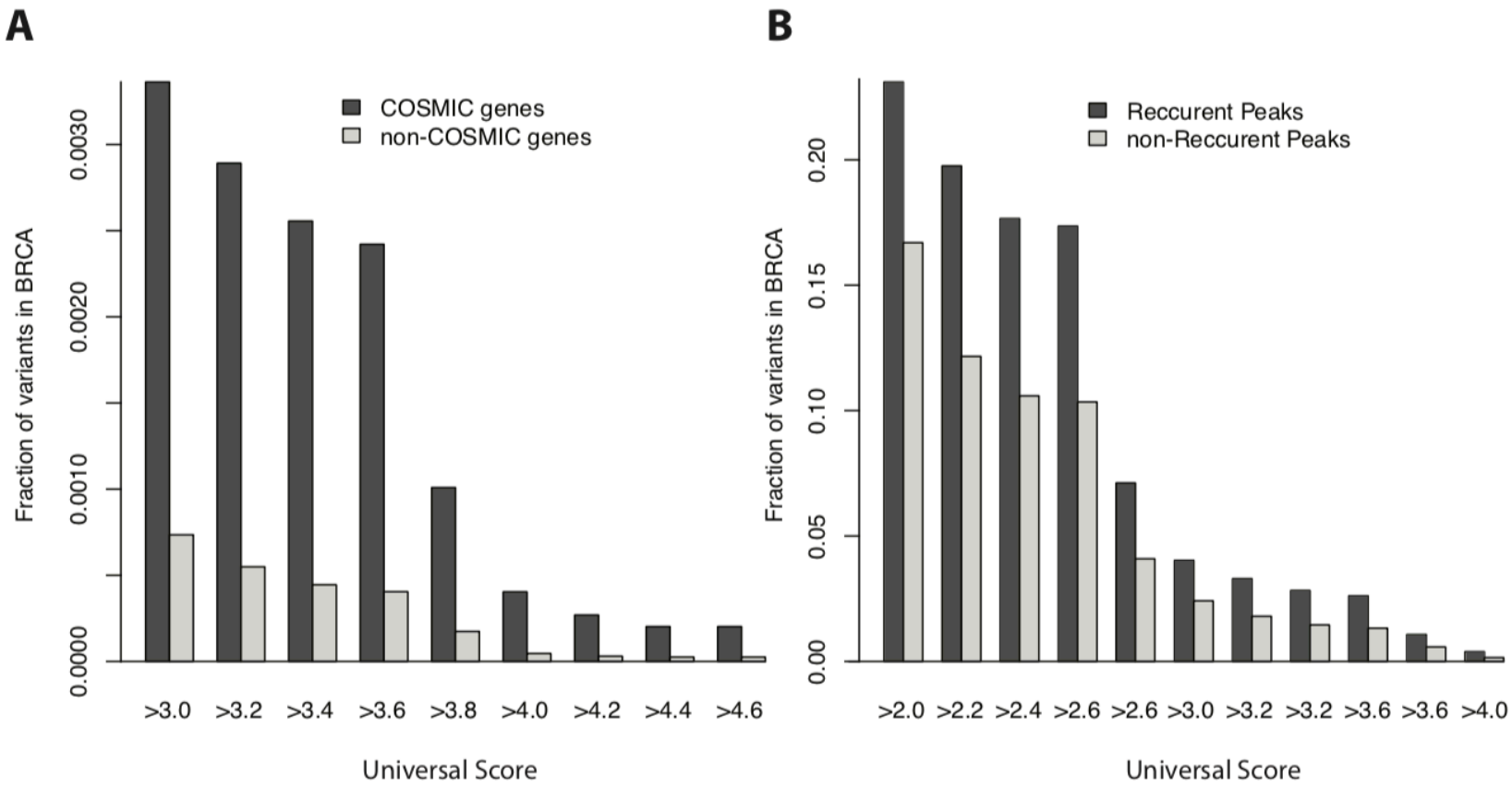
Binding hubs are enriched for rare variants





[Zhang*, Liu* et al., *Genome Biology* '20]

RADAR Scores enriched in COSMIC genes and recurrently mutated regions

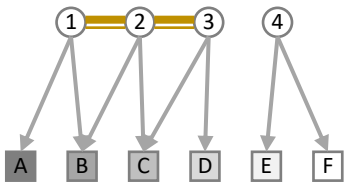


[Zhang*, Liu* et al., *Genome Biology* '20]

Using ENCODE Data for Cancer Genomics

- **BMR Correction:**
  LARVA/MOAT/NIMBUS
  - Parametric models explicitly modeling genomic covariates
  - Many ENCODE covariates useful in accurately estimating background mutation rate
- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data in certain cell lines highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
- **RADAR Variant Prioritization**
  - Prioritizes germline & somatic variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

Normal Network



Co-regulation

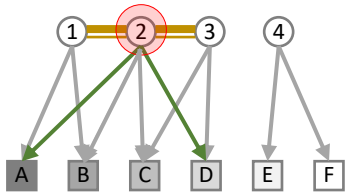
TF/RBP to gene

TF/RBP

gene

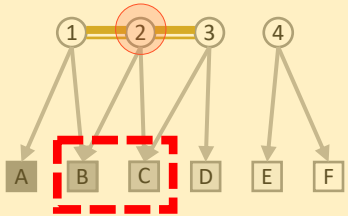
High expression — low expression

Disease Network：
dotted line = lost edge



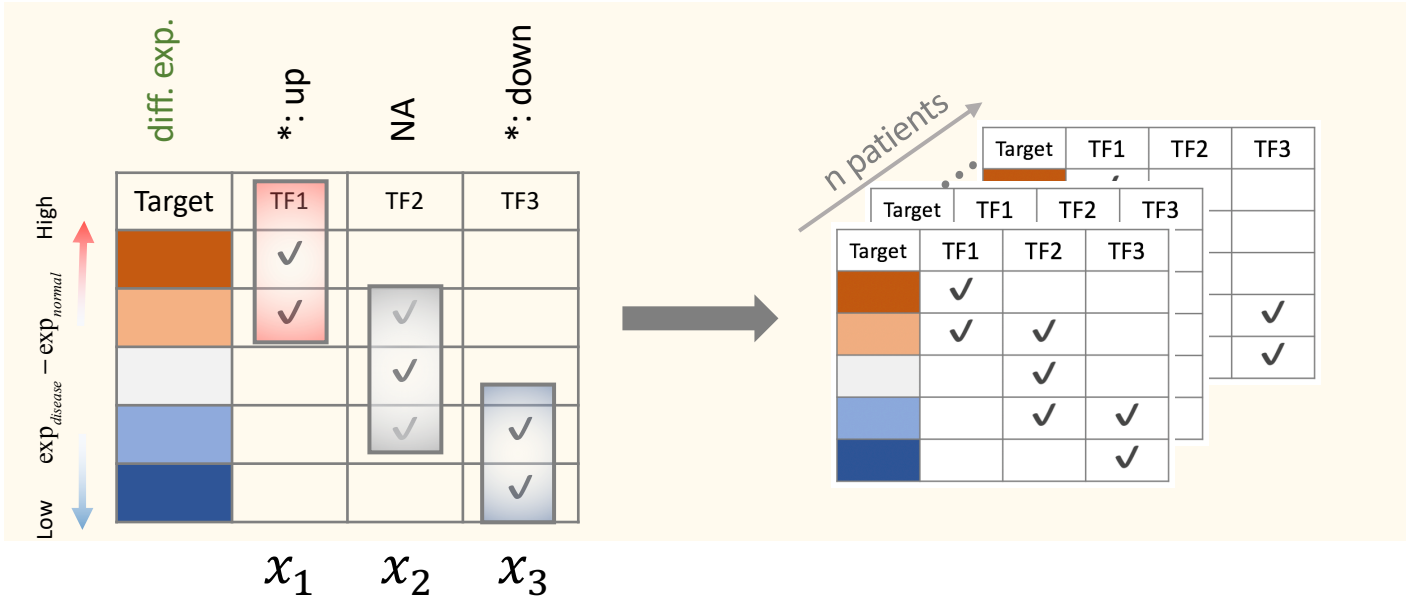
Principles

Direct target gain/loss



Target gene expression changes

[Zhang et al. ('19), biorxiv.org]

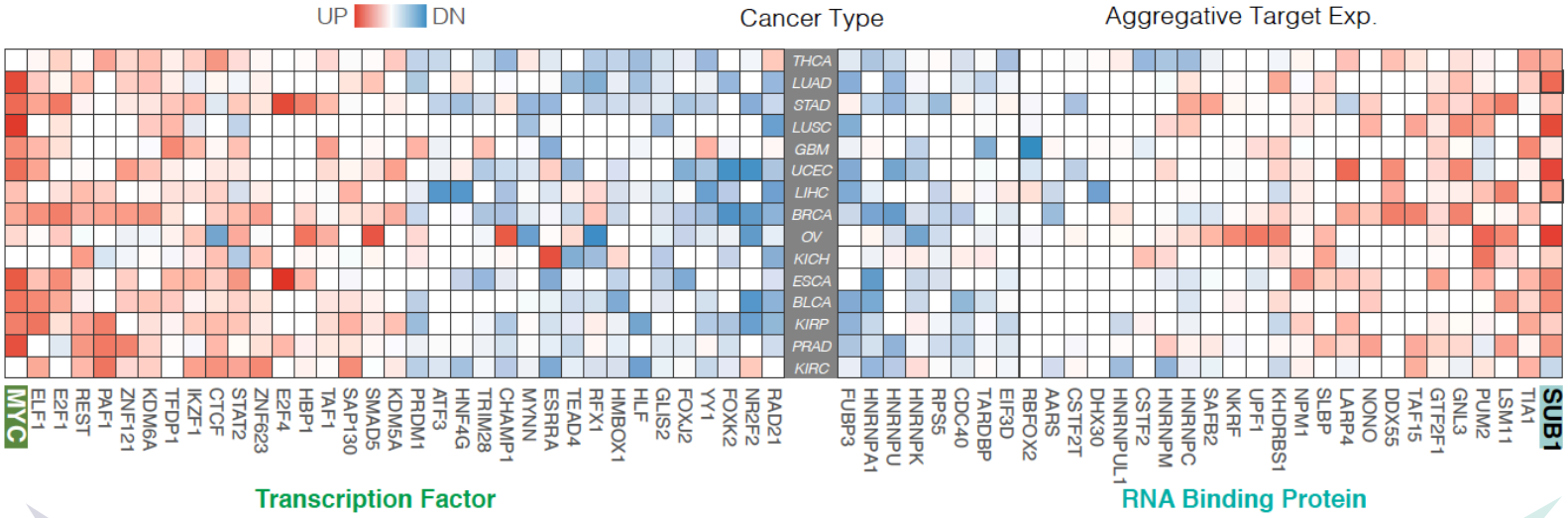


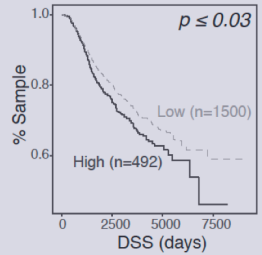
2198 ChIP-seq

459 eCLIP

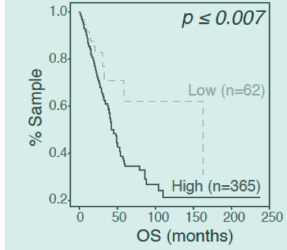
$$\underbrace{y = \left(\exp_{disease} - \exp_{normal}\right)}_{\text{differential expression}} \sim \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}_{\text{Network for Regulator 1 to k}}$$

[Zhang et al. ('19), biorxiv.org]

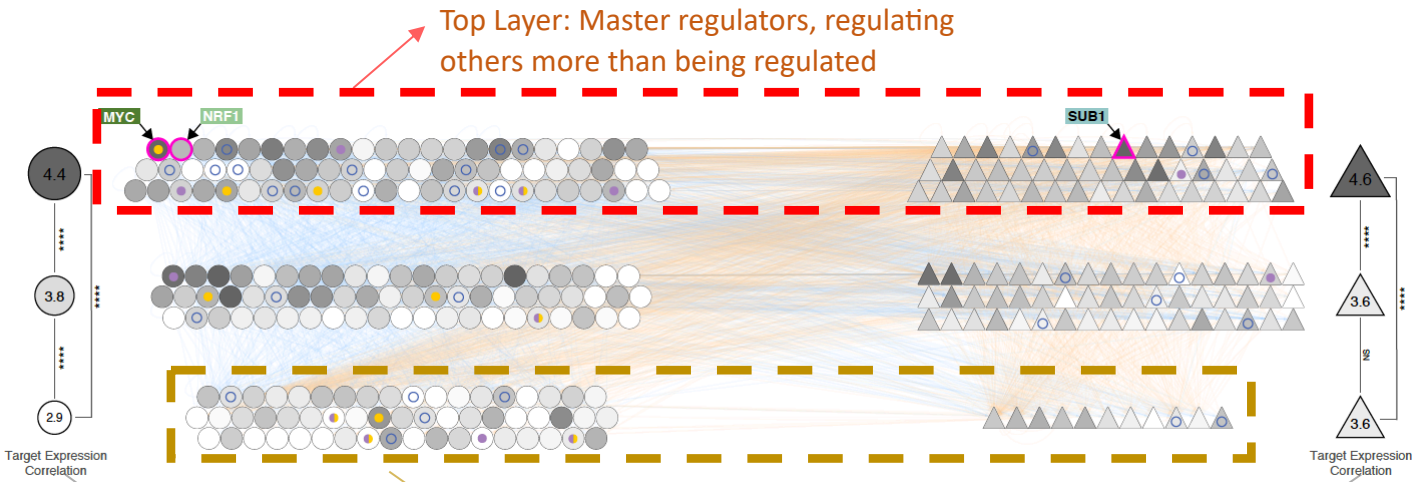




Aggregated t-statistic in regression over TCGA samples



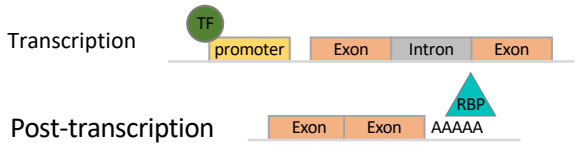
[Zhang et al. ('19), biorxiv.org]



How much power each regulator has in driving tumor-normal differential expressions

TF-RBP crosstalk

TF-RBP regulate the same gene at different levels



[Zhang et al. ('19), biorxiv.org]

Using ENCODE Data for Cancer Genomics

- **BMR Correction:**
  LARVA/MOAT/NIMBUS
  - Parametric models explicitly modeling genomic covariates
  - Many ENCODE covariates useful in accurately estimating background mutation rate
- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data in certain cell lines highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
- **RADAR Variant Prioritization**
  - Prioritizes germline & somatic variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

Using ENCODE Data for Cancer Genomics

- **BMR Correction: LARVA/MOAT/NIMBUS**
  - Parametric models explicitly modeling genomic covariates
  - Many ENCODE covariates useful in accurately estimating background mutation rate
- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data in certain cell lines highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
- **RADAR Variant Prioritization**
  - Prioritizes germline & somatic variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

ENCODEC.gersteinlab.org
J **Zhang**, D **Lee**, V **Dhiman**, P **Jiang**, J **Xu**,
P McGillivray, H Yang…. S Liu, K White

NIMBus.gersteinlab.org
J **Zhang**, J **Liu**, P McGillivray, C Yi, L Lochovsky, D Lee

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

{LARVA,MOAT}.gersteinlab.org
Lochovsky, J **Zhang**, Y Fu, E Khurana

github.com/gersteinlab/**TopicNet**
S **Lou**, T **Li**, X **Kong**, J Zhang, J Liu, D Lee


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