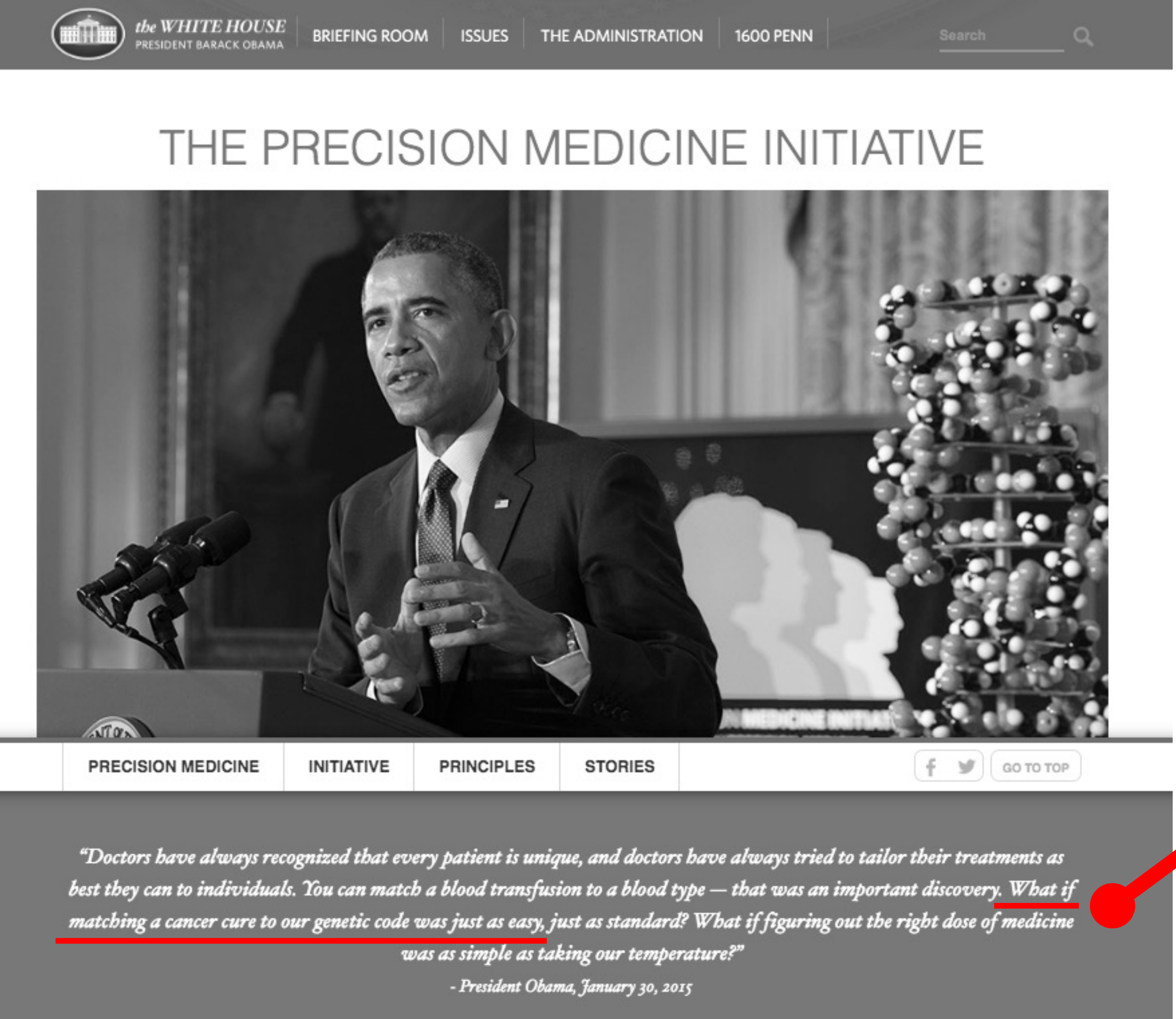
Topics in Precision Oncology: Addressing the role of the non-coding genome in cancer

Mark Gerstein
Yale

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



THE PRECISION MEDICINE INITIATIVE

“Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals. You can match a blood transfusion to a blood type — that was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?”

- President Obama, January 30, 2015

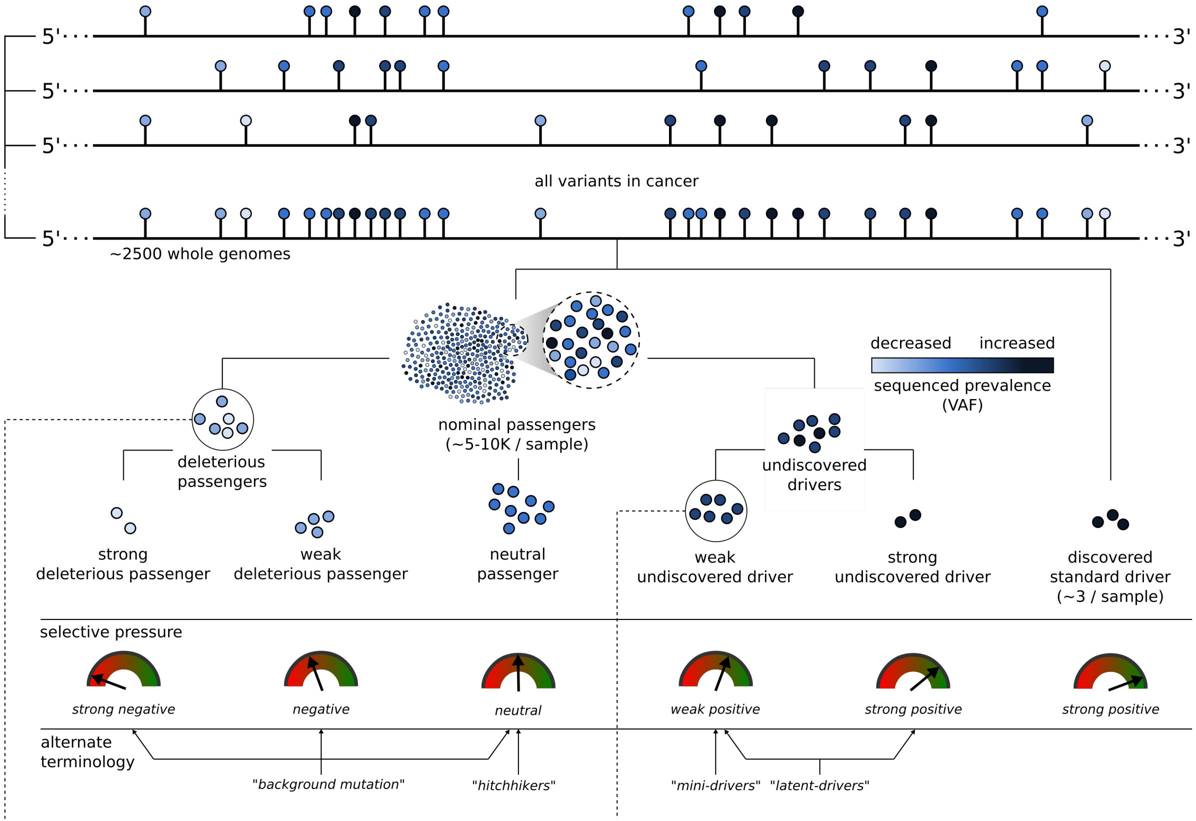
Precision Oncology

- Sub-topic of precision medicine
- Analysis of the exact somatic mutations in a individual, suggesting individualized treatment

What if matching a cancer cure to our genetic code was just as easy

https://obamawhitehouse.archives.gov/blog/2016/02/25/precision-medicine-health-care-tailored-you

Extension of the canonical model of drivers and passengers

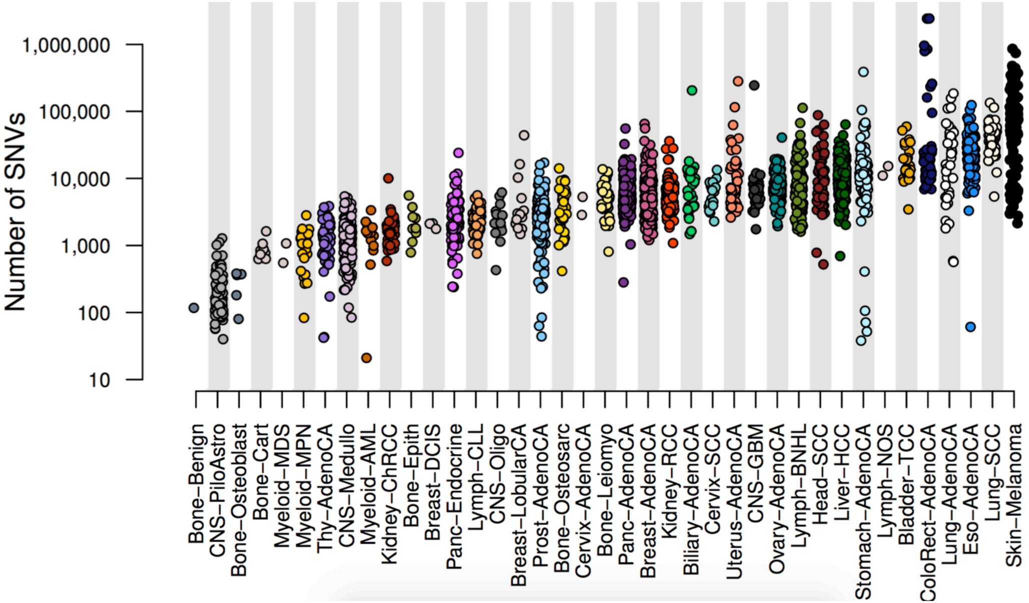


Coding regions are only ~1-2% of the genome yet contain almost all the drivers.

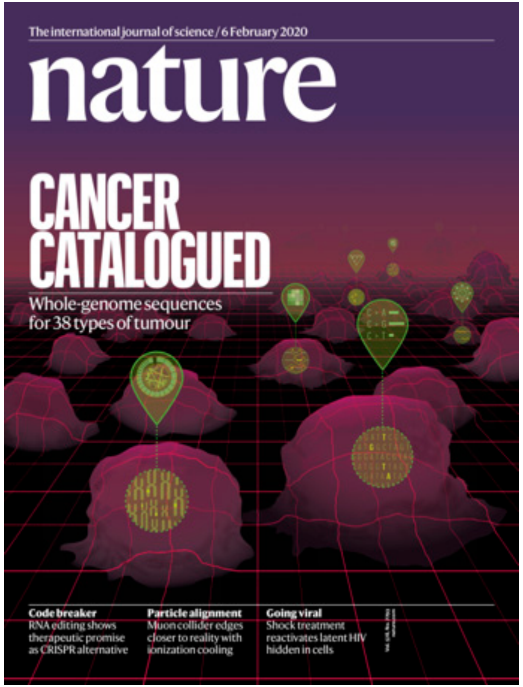
Open Q: what is the role of the non-coding genome in cancer?

[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)


The international journal of science / 6 February 2020
nature
CANCER CATALOGUED
Whole-genome sequences for 38 types of tumour
Code breaker
RNA editing shows therapeutic promise as CRISPR alternative
Particle alignment
Muon collider edges closer to reality with ionization cooling
Going viral
Shock treatment reactivates latent HIV hidden in cells

- **Union of TCGA-ICGC efforts**

- Jointly analyzing ~2800 whole genome tumor/normal pairs
  - ➢ > 580 researchers
  - ➢ ~30M total somatic SNVs

http://encodec.encodeproject.org/

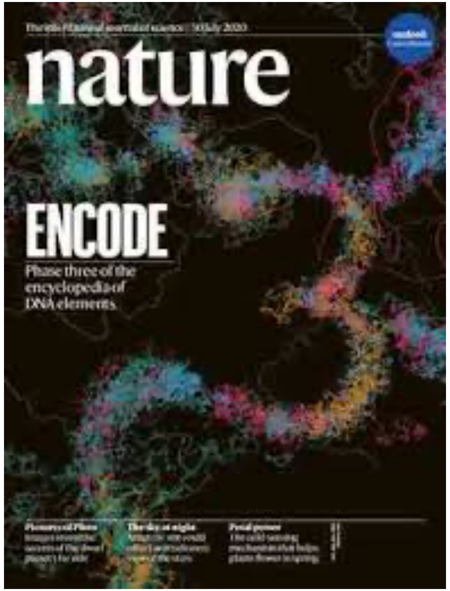
BIOSAMPLE ⟶

ENCODEC

86 Cancerous (40 Cancer Types) + 143 Composite Normal (inc. Roadmap)

			K562	HepG2	A549	MCF-7	HeLa-S3	H1-hESC	Caco-2	HCT116	Panc1	LNCaP	PC-3	PC-9	SK-N-MC	DND-41	SK-N-SH	...
			CML	LIHC	LUAD	BRCA	Cervix	ESC	COAD+READ		PAAD	PRAD		LUAD	SARC	LAML	NB	...
Chromatin Accessibility	DS	DNase-seq	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆		◆	
Histone Modification	HM	Histone ChIP-seq	19	14	85	16	14	53	3	16	7	1	11	11	8	11	19	
Transcription	TX	RNA-seq	◆	◆	◆	◆	◆	◆	▼	◆	◆	▼	◆	▼	▼	▼	◆	
		RAMPAGE	◆															
RNA-binding Proteins	RP	eCLIP	191	164														
RNAi/CRISPR Knockdown	KD	shRNA/siRNA KD	326	257		2												
		CRISPR KD/KO	108	19														
3D Chromatin Structure	3D	ChIA-PET	9	2		5	1											
		Hi-C	▼	◆	◆	▼	◆	▼										
Enhancers	SS	STARR-seq	◆	◆		◆												
Methylation	ME	WGBS	◆	◆	◆	▼	◆	◆										
		RRBS	◆	◆	◆	◆	◆	◆										
Replication Timing	RT	Repli-chip					◆	◆										
		Repli-seq	◆	◆	◆	◆	◆											
Transcription Factors	TF	TF ChIP-seq	558	300	240	149	78	89										
Cell Line WGS	WG	SNV	▼		▼	▼	▼											
		SV	▼		▼	▼	▼											

528 ENCODE Cell Types → 229 Deduplicated & Selected Human Biosamples



The international journal of science | 30 July 2020

nature

ENCODE

Phase three of the encyclopedia of DNA elements

Comprehensive non-coding Annotation

Applicable to cancer genomics

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

Topics in Precision Oncology:
Addressing the role of the non-coding genome in cancer

- **Background**
  - Drivers v passenger
  - Coding v noncoding
  - Pcawg & encode 3

- **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

- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities

Topics in Precision Oncology:
Addressing the role of the non-coding genome in cancer

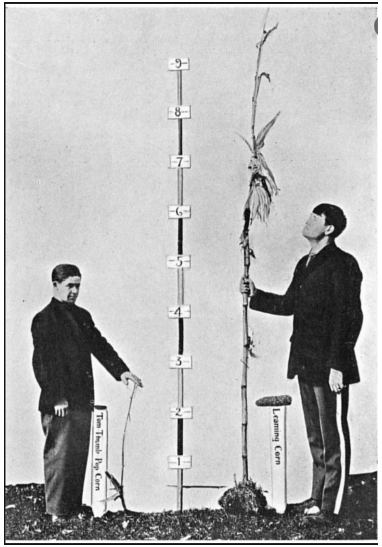
- **Background**
  - Drivers v passenger
  - Coding v noncoding
  - Pcawg & encode 3

- **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

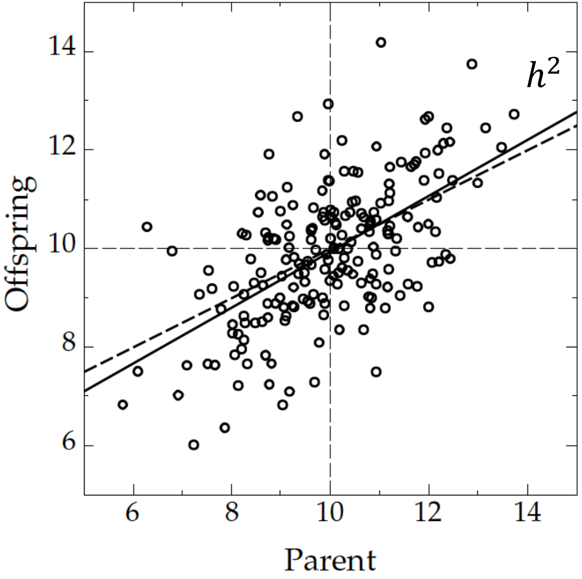
- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities

Relating Germline Missing Heritability to Cancer Studies

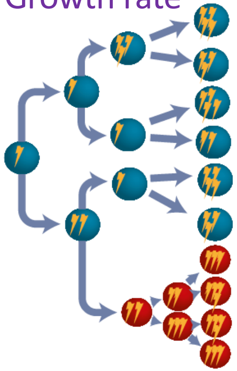
Organismal trait: Height

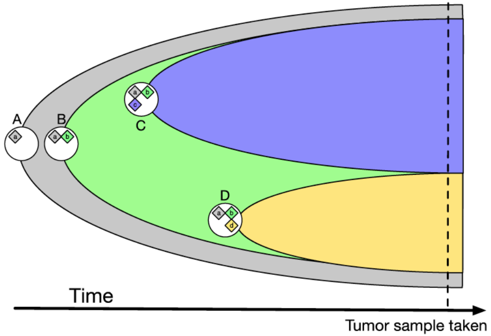


Population level definitions:
Parent-offspring heritability;
Twin-based heritability ...



Subclonal trait in cancer:
Growth rate





SNP-based polygenic & additive model:

$$h^2 = \sigma_u$$

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

Trait

Covariates & fixed effects

Genetic predictors & random effects

Environmental noise

Missing heritability for height & other traits

- Height is a highly polygenic trait:

SNP category	# SNPs	Heritability estimate (h^2)	Year
GWAS SNPs[1]	50	~0.05	2008
Common SNPs[2]	~295K	0.54 (SE 0.1)	2010
Common+rare SNPs[3]	47.1M	0.79 (SE 0.09)	2019
Population estimate (twins)[4]	-	**0.8**	(2012)

SE = standard error

- Many other traits have substantial missing GWAS-based heritability[5]:

[1] Weedon, M.N., Lango, H., Lindgren, C.M., Wallace, C., Evans, D.M., Mangino, M., Freathy, R.M., Perry, J.R., Stevens, S., Hall, A.S. and Samani, N.J., 2008. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics*, *40*(5), p.575.

[2] Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E.,..., Visscher, P., 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, *42*(7), p.565.

[3] Wainschtein, P., Jain, D.P., Yengo, L., Zheng, Z., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., Psaty, B.M., Kooperberg, C., ..., Visscher, P., 2019. Recovery of trait heritability from whole genome sequence data. *bioRxiv*, p.588020.

[4] Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J., 2012. Five years of GWAS discovery. *The American Journal of Human Genetics*, *90*(1), pp.7-24.

[5] Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., and Visscher, P., 2009. Finding the missing heritability of complex diseases. *Nature*, *461*(7265), p.747.

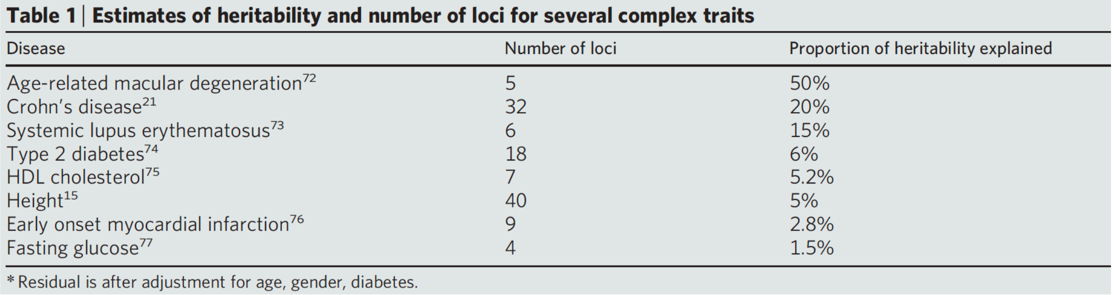


Table 1 | Estimates of heritability and number of loci for several complex traits

Disease	Number of loci	Proportion of heritability explained
Age-related macular degeneration[72]	5	50%
Crohn's disease[21]	32	20%
Systemic lupus erythematosus[73]	6	15%
Type 2 diabetes[74]	18	6%
HDL cholesterol[75]	7	5.2%
Height[15]	40	5%
Early onset myocardial infarction[76]	9	2.8%
Fasting glucose[77]	4	1.5%

* Residual is after adjustment for age, gender, diabetes.

Additive effects model to quantify cumulative effect of nominal passengers in PCAWG

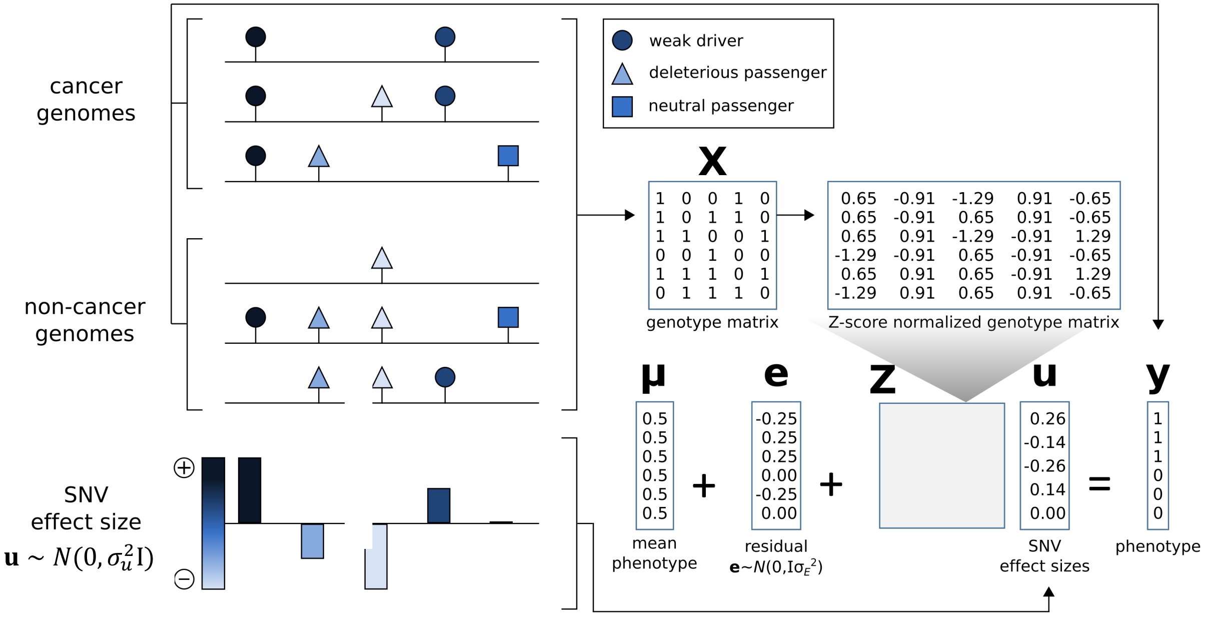
- Model for the effect of an individual SNP on a phenotype

$$y_j = \mu + z_{ij} u_i + e_j$$

- Extension to model the combined effects of multiple SNPs

$$y_j = \mu + g_j + e_j \text{ and } g_j = \sum_{i=1}^{m} z_{ij} u_i$$

$$g_j \sim N(0, \sigma_g^2 = m\sigma_u^2) \qquad \mathbf{u} \sim N(\mathbf{0}, \mathbf{I}\sigma_u^2)$$

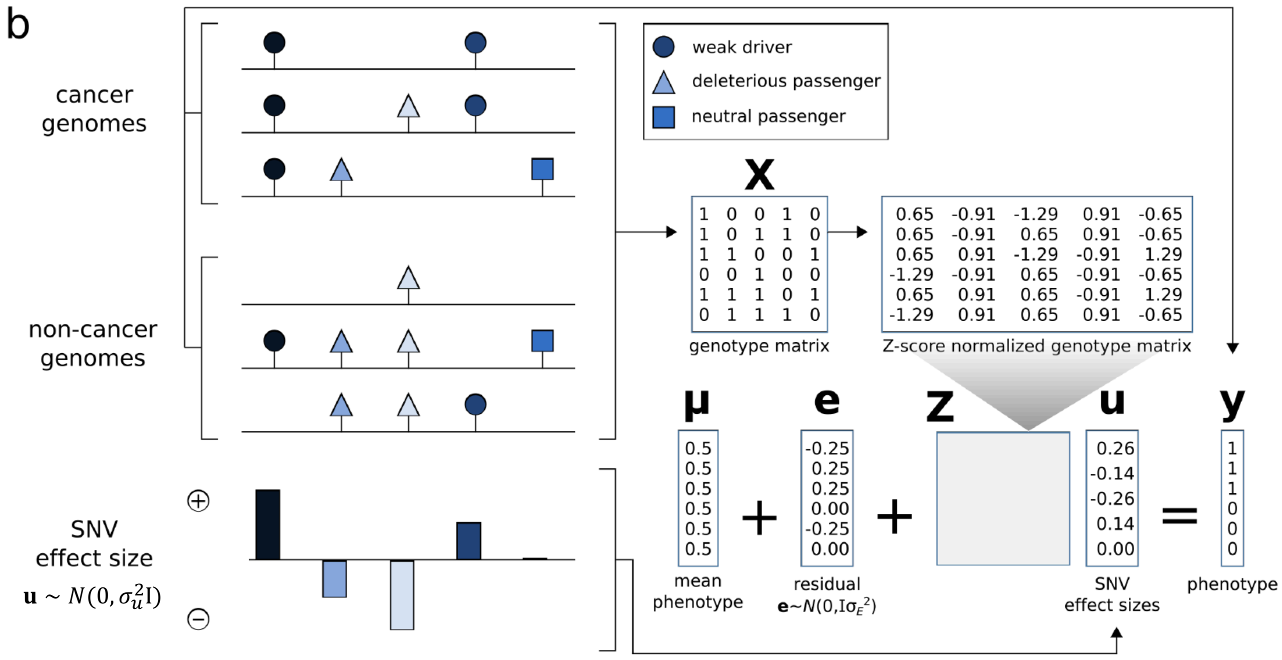


$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0.65 & -0.91 & -1.29 & 0.91 & -0.65 \\ 0.65 & -0.91 & 0.65 & 0.91 & -0.65 \\ 0.65 & 0.91 & -1.29 & -0.91 & 1.29 \\ -1.29 & -0.91 & 0.65 & -0.91 & -0.65 \\ 0.65 & 0.91 & 0.65 & -0.91 & 1.29 \\ -1.29 & 0.91 & 0.65 & 0.91 & -0.65 \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} -0.25 \\ 0.25 \\ 0.25 \\ 0.00 \\ -0.25 \\ 0.00 \end{pmatrix} \quad \mathbf{u} = \begin{pmatrix} 0.26 \\ -0.14 \\ -0.26 \\ 0.14 \\ 0.00 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

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

Using additive effects to compare different categories of variants



$$\mathbf{X} = \begin{array}{ccccc} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{array}$$

$$\begin{array}{ccccc} 0.65 & -0.91 & -1.29 & 0.91 & -0.65 \\ 0.65 & -0.91 & 0.65 & 0.91 & -0.65 \\ 0.65 & 0.91 & -1.29 & -0.91 & 1.29 \\ -1.29 & -0.91 & 0.65 & -0.91 & -0.65 \\ 0.65 & 0.91 & 0.65 & -0.91 & 1.29 \\ -1.29 & 0.91 & 0.65 & 0.91 & -0.65 \end{array}$$

$$\boldsymbol{\mu} = \begin{array}{c} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{array} + \mathbf{e} = \begin{array}{c} -0.25 \\ 0.25 \\ 0.25 \\ 0.00 \\ -0.25 \\ 0.00 \end{array} + \mathbf{Z} \; \mathbf{u} = \begin{array}{c} 0.26 \\ -0.14 \\ -0.26 \\ 0.14 \\ 0.00 \end{array} = \mathbf{y} = \begin{array}{c} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{array}$$

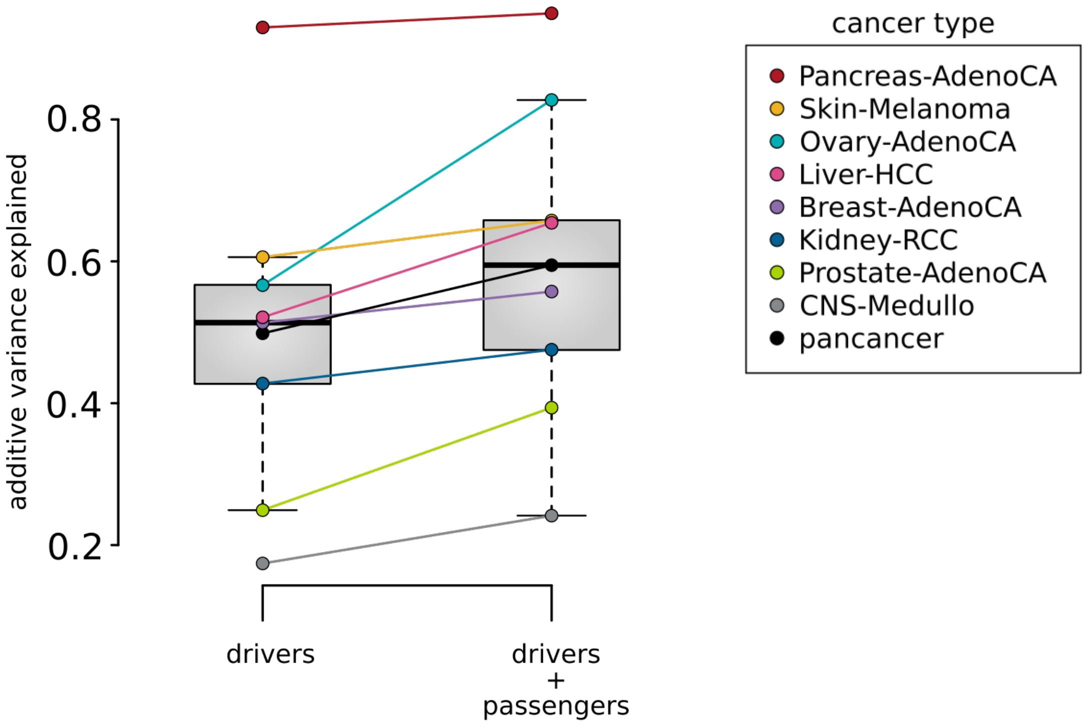
Model: $y_j = \mu + z_j^{\mathrm{drv}} u_1 + \sum_{k \in \{2,3,4\},} z_{ijk} u_{ik} + e_j$

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]

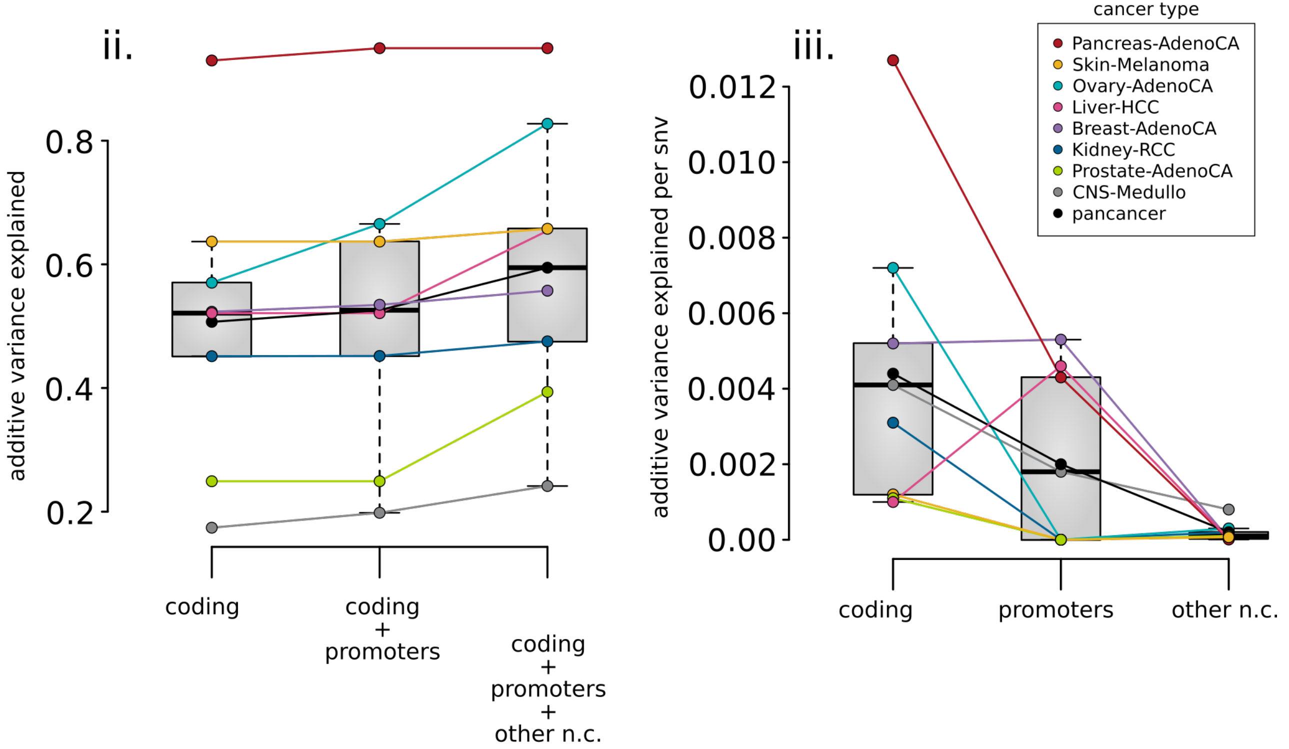
Overall additive variance increase for multiple cancer cohorts in PCAWG with the inclusion of passengers



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]

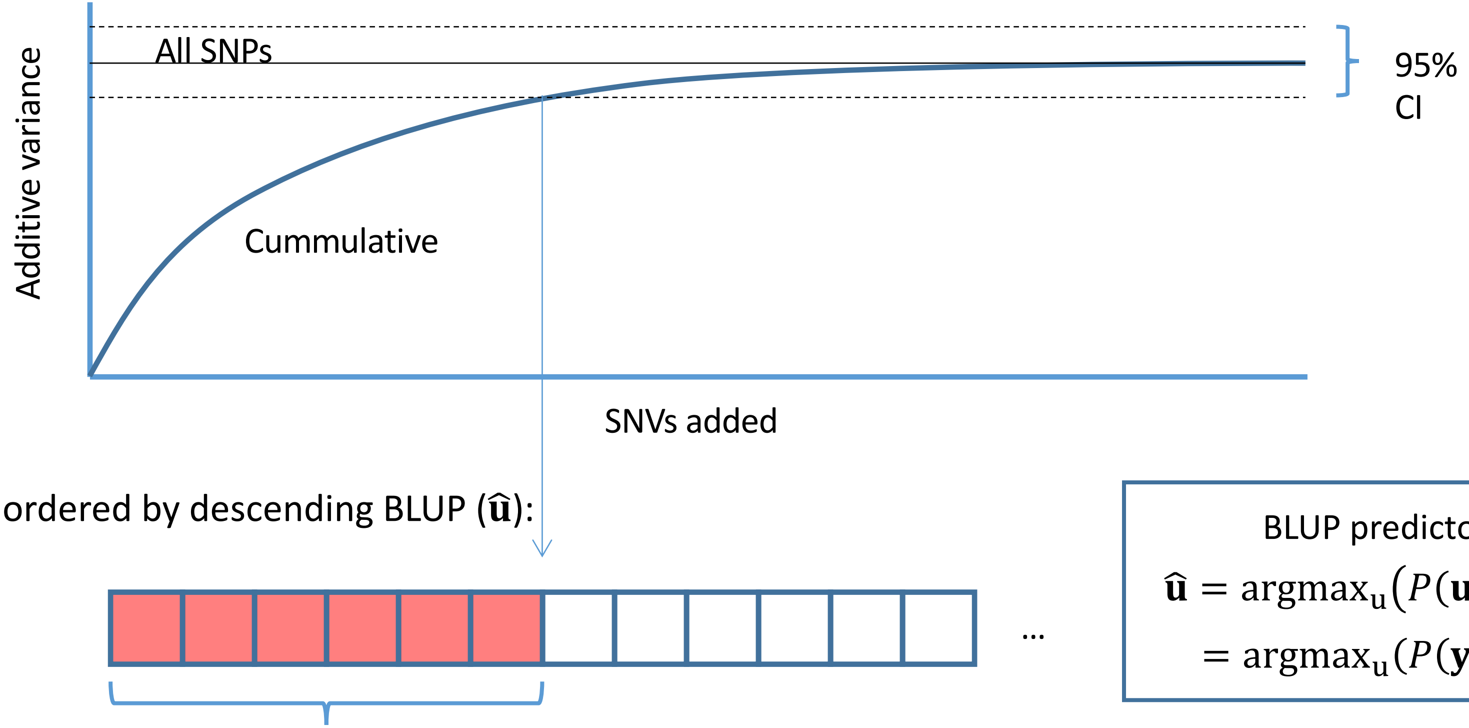
Element level additive variance for multiple cancer cohorts in PCAWG, comparing coding & non-coding



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 ($\hat{\mathbf{u}}$):

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

BLUP predictor:

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

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

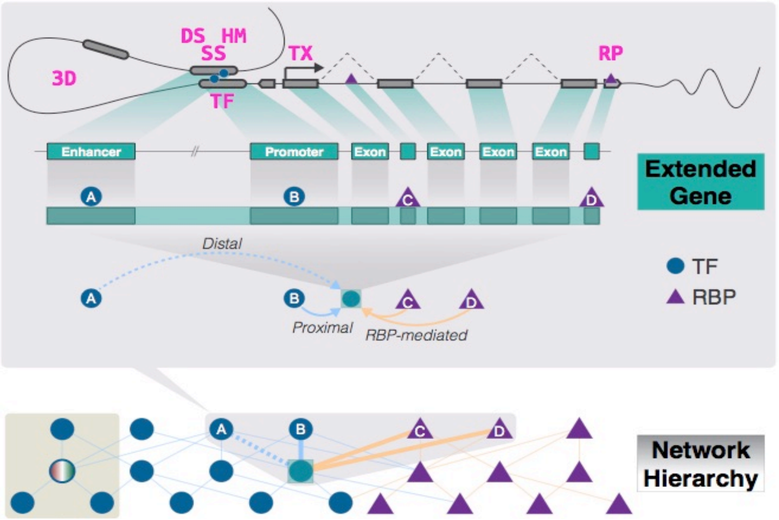
Topics in Precision Oncology:
Addressing the role of the non-coding genome in cancer

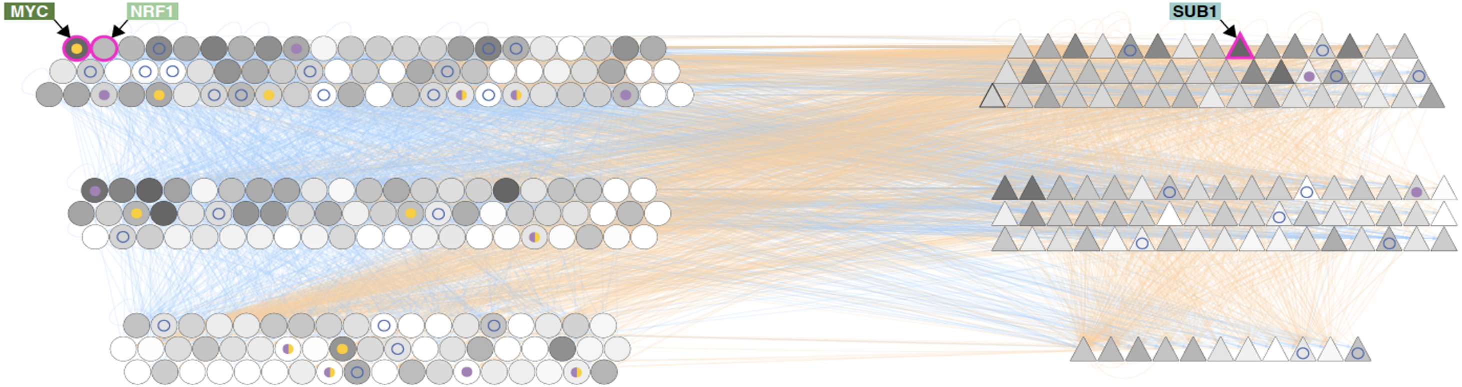
- **Background**
  - Drivers v passenger
  - Coding v noncoding
  - Pcawg & encode 3

- **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

- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities

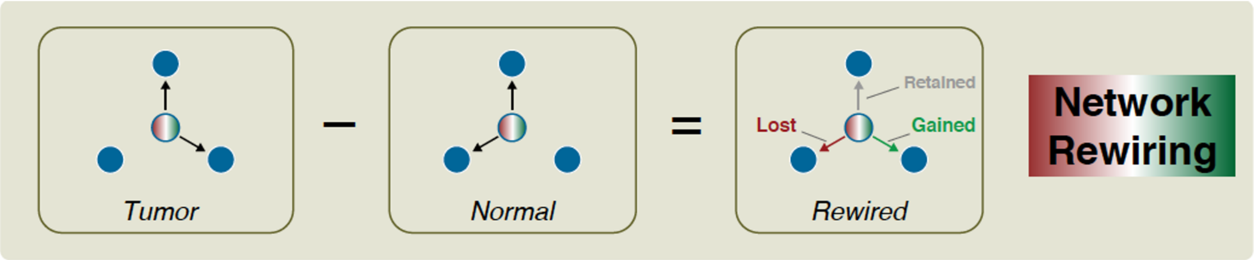
Regulatory Network Construction



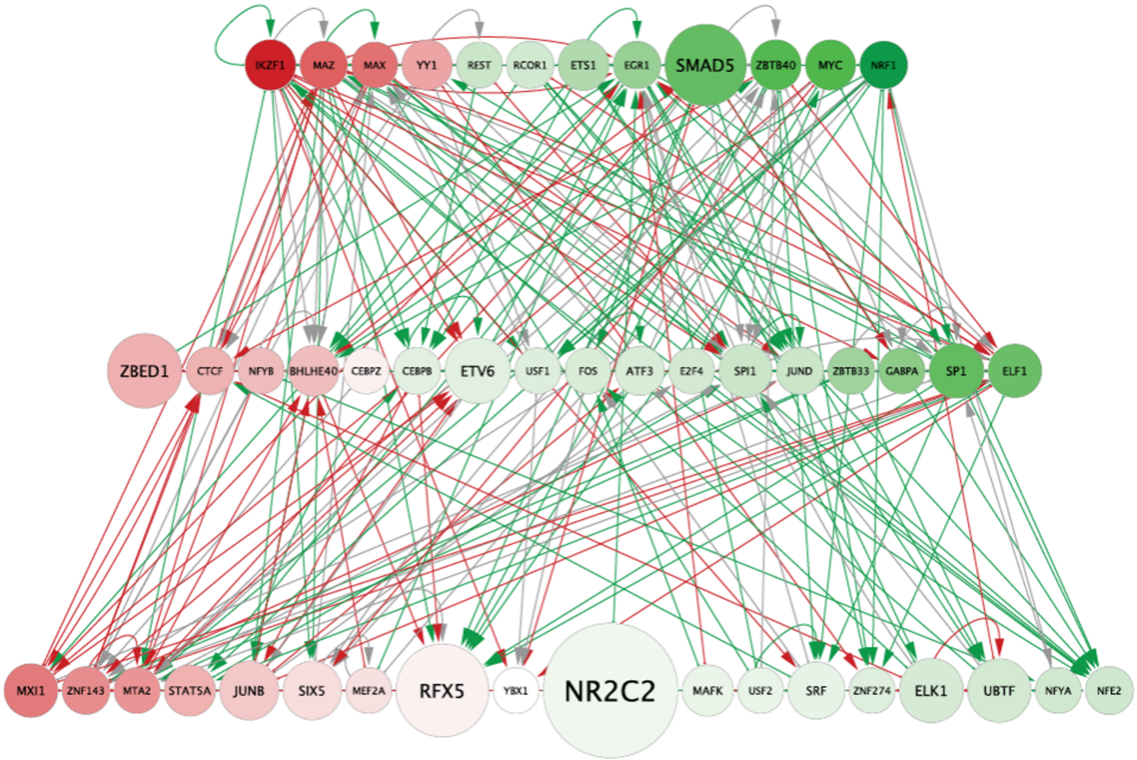


Transcription Factor ***RNA-binding Protein***

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



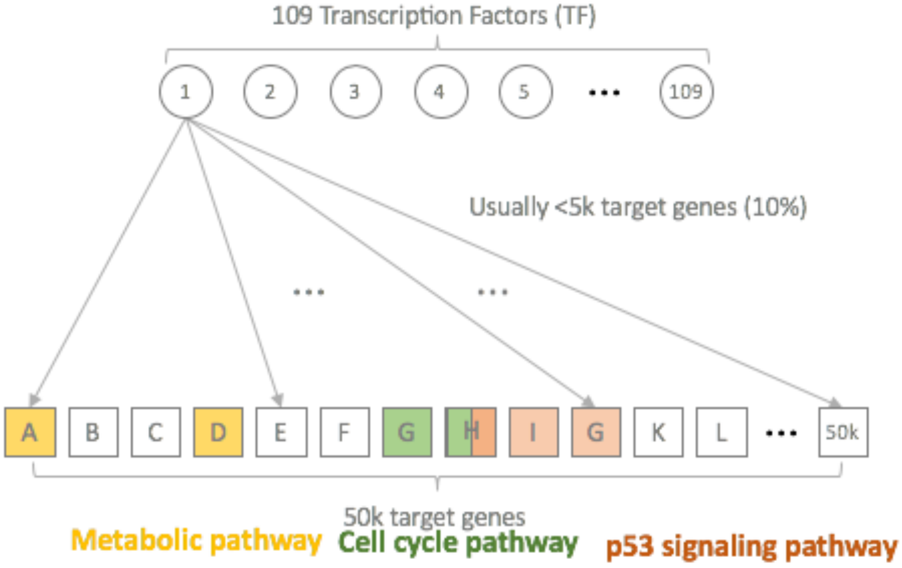
Rewired edges in comparison of GM12878 to K562
109 node TF-TF network (approx. CML)

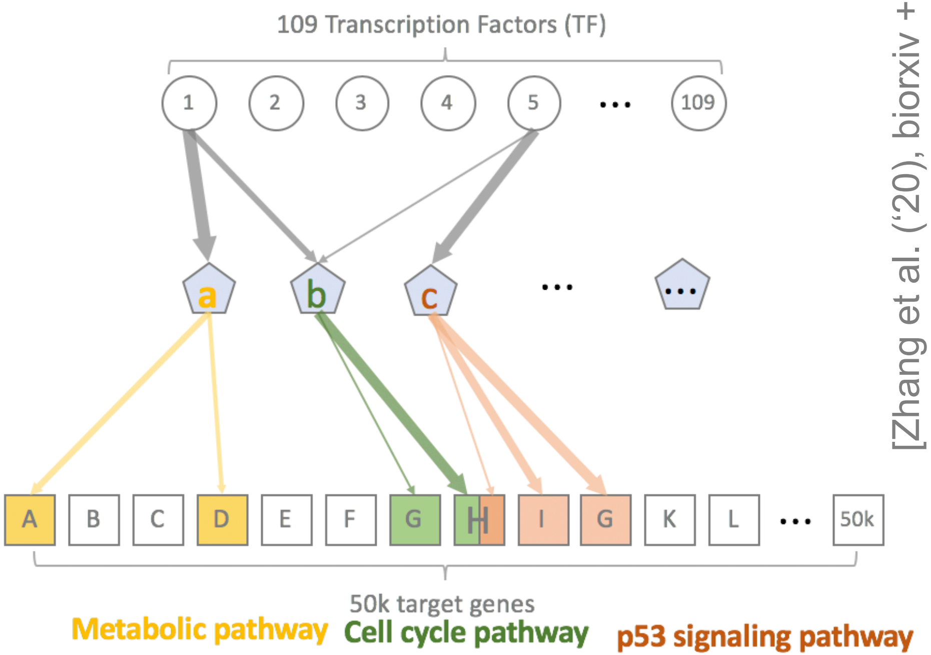


[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

Simplifying Network Rewiring

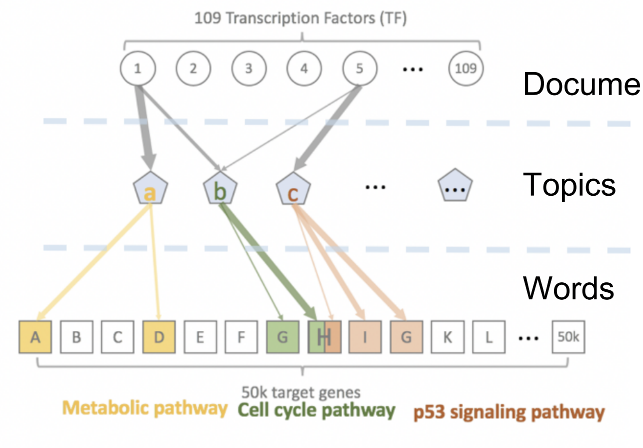
From $TF \rightarrow gene\ (109\times 50{,}000)$
to $TF \rightarrow pathway\ (109\times 50)$

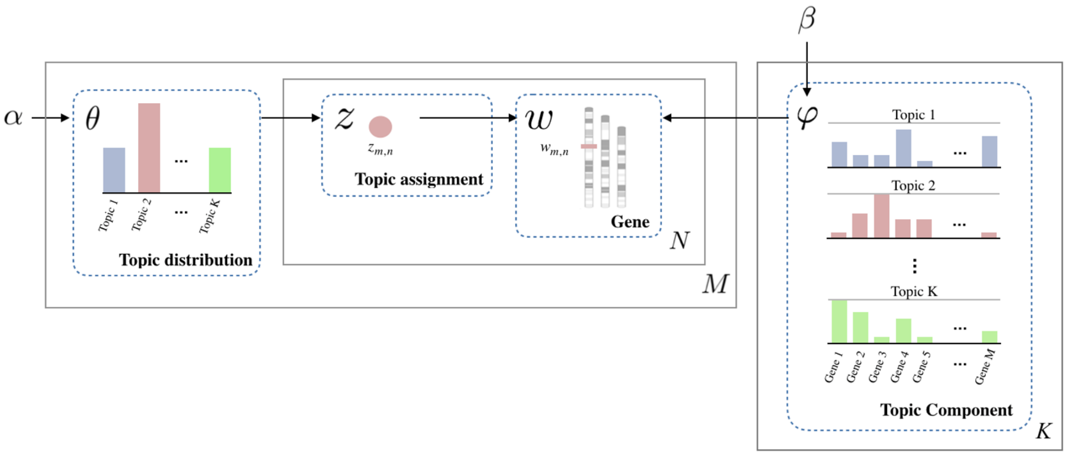


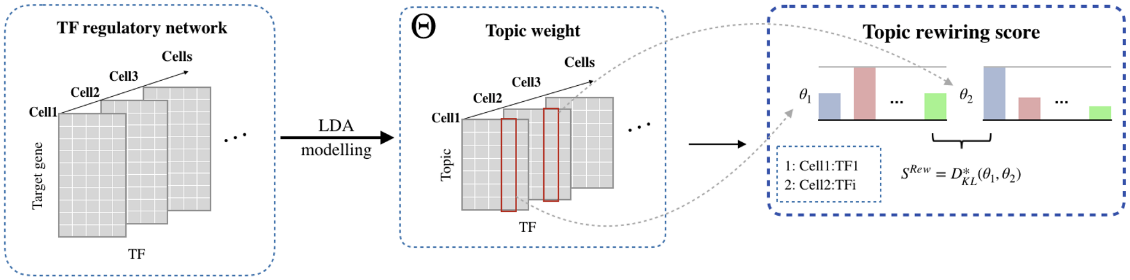


[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

TopicNet: Measuring transcriptional regulatory network change using LDA





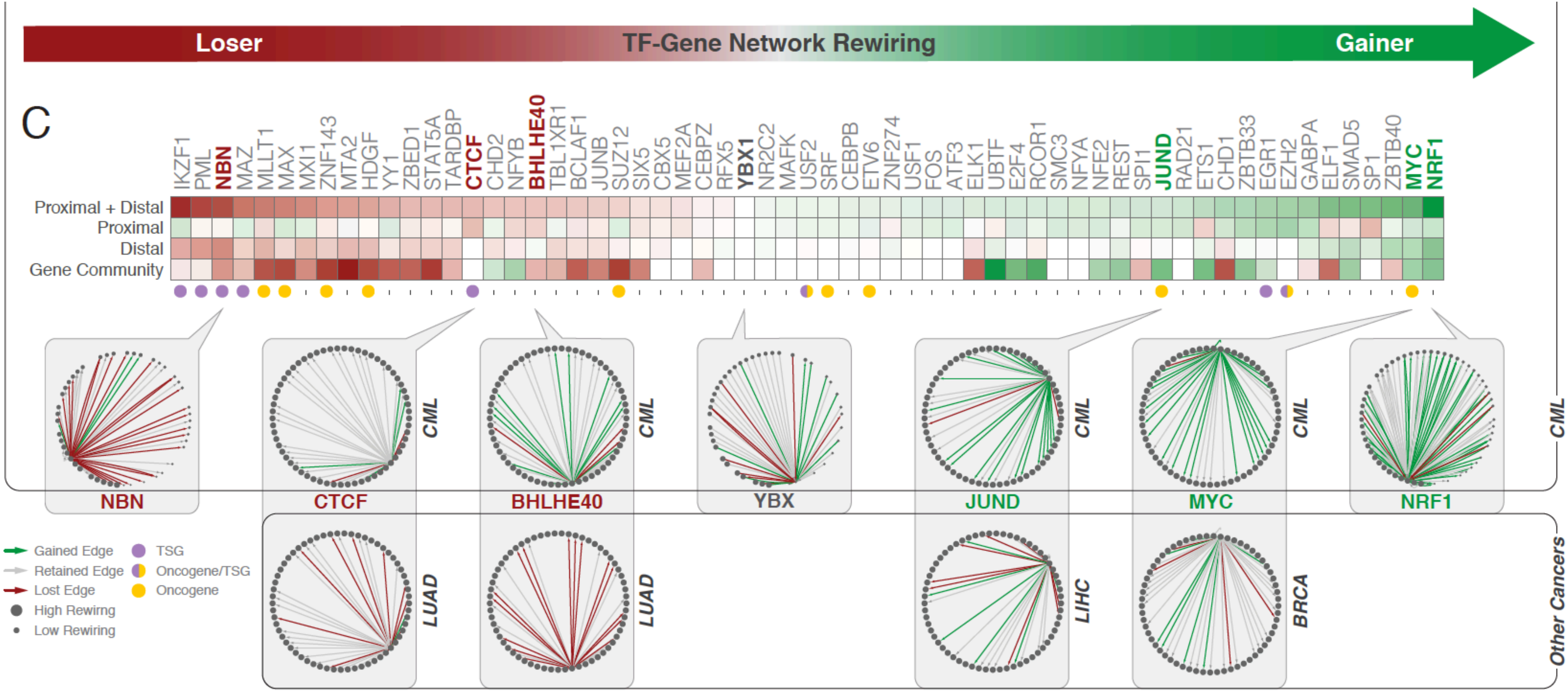


α Prior info β Prior info

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)]

Topics in Precision Oncology:
Addressing the role of the non-coding genome in cancer

- **Background**
  - Drivers v passenger
  - Coding v noncoding
  - Pcawg & encode 3

- **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

- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities

Topics in Precision Oncology:
Addressing the role of the non-coding genome in cancer

- **Background**
  - Drivers v passenger
  - Coding v noncoding
  - Pcawg & encode 3

- **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

- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities

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

ENCODEC.gersteinlab.org

J **Zhang**, D **Lee**, V Dhiman, P Jiang, J Xu, P McGillivray…. S Liu, K White

github.com/gersteinlab/**TopicNet**

S **Lou**, T **Li**, X Kong, J Zhang, J Liu, D Li

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