(Topics in)
Cancer Genomics:

Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

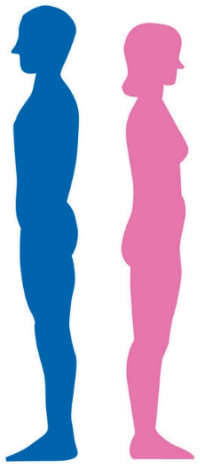
Mark Gerstein
Yale

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

Estimated numbers of **new cases** of invasive cancer in the United States in 2019 by sex and cancer type

Estimated New Cases

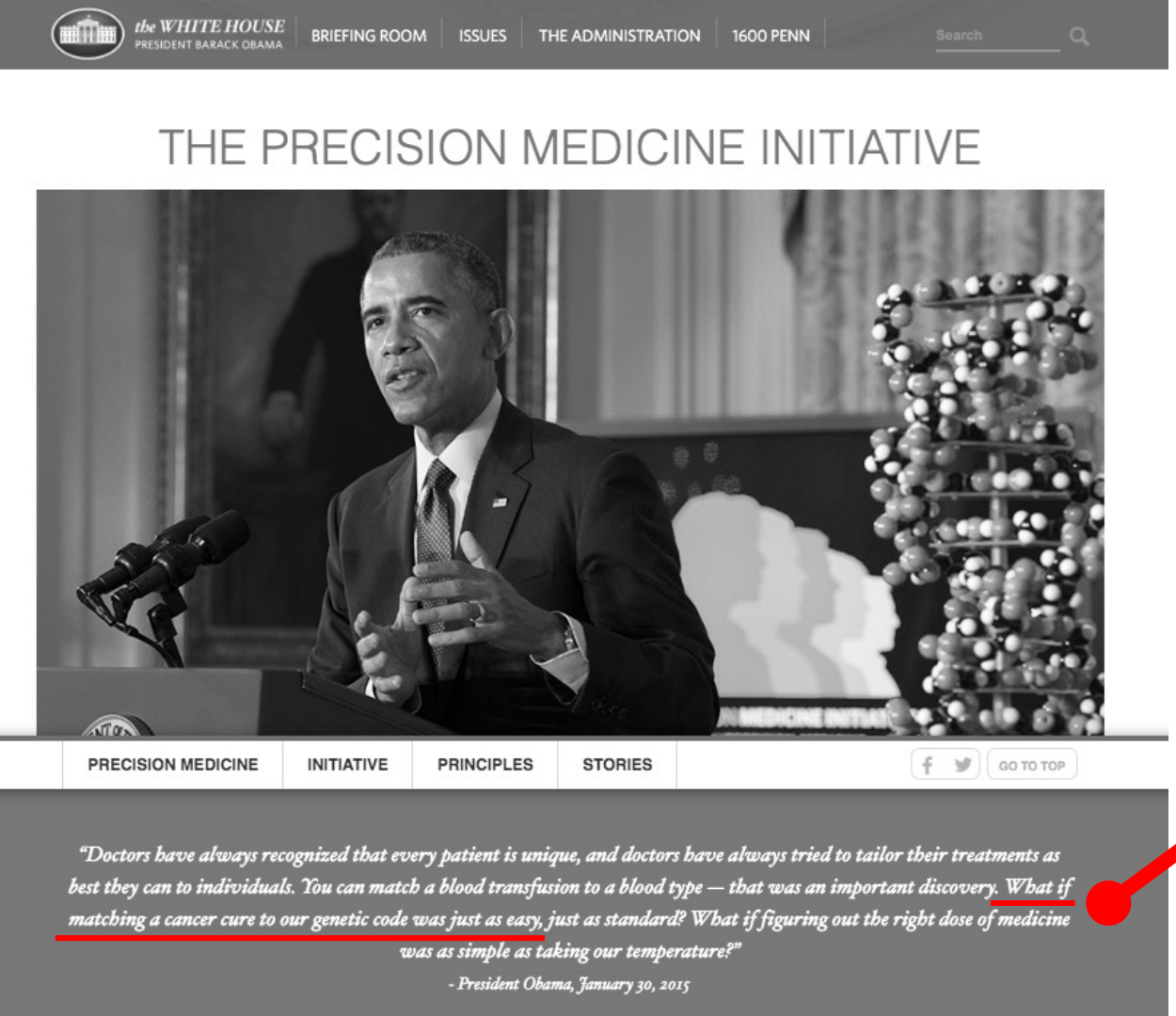
Males				Females		
Prostate	174,650	20%		Breast	268,600	30%
Lung & bronchus	116,440	13%		Lung & bronchus	111,710	13%
Colon & rectum	78,500	9%		Colon & rectum	67,100	8%
Urinary bladder	61,700	7%		Uterine corpus	61,880	7%
Melanoma of the skin	57,220	7%		Melanoma of the skin	39,260	4%
Kidney & renal pelvis	44,120	5%		Thyroid	37,810	4%
Non-Hodgkin lymphoma	41,090	5%		Non-Hodgkin lymphoma	33,110	4%
Oral cavity & pharynx	38,140	4%		Kidney & renal pelvis	29,700	3%
Leukemia	35,920	4%		Pancreas	26,830	3%
Pancreas	29,940	3%		Leukemia	25,860	3%
All Sites	**870,970**	**100%**		**All Sites**	**891,480**	**100%**



1,762,450 new cases per year

~4,800 new cases per day

Segiel et al, Cancer statistics, 2019



Much Interest in Precision Oncology

- Analysis of the exact somatic mutations in a individual
- Highlighting key mutations
- Targeting 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

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

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.

Canonical model of drivers & passengers in cancer

Drivers
directly confer a selective growth advantage to the tumor cell.

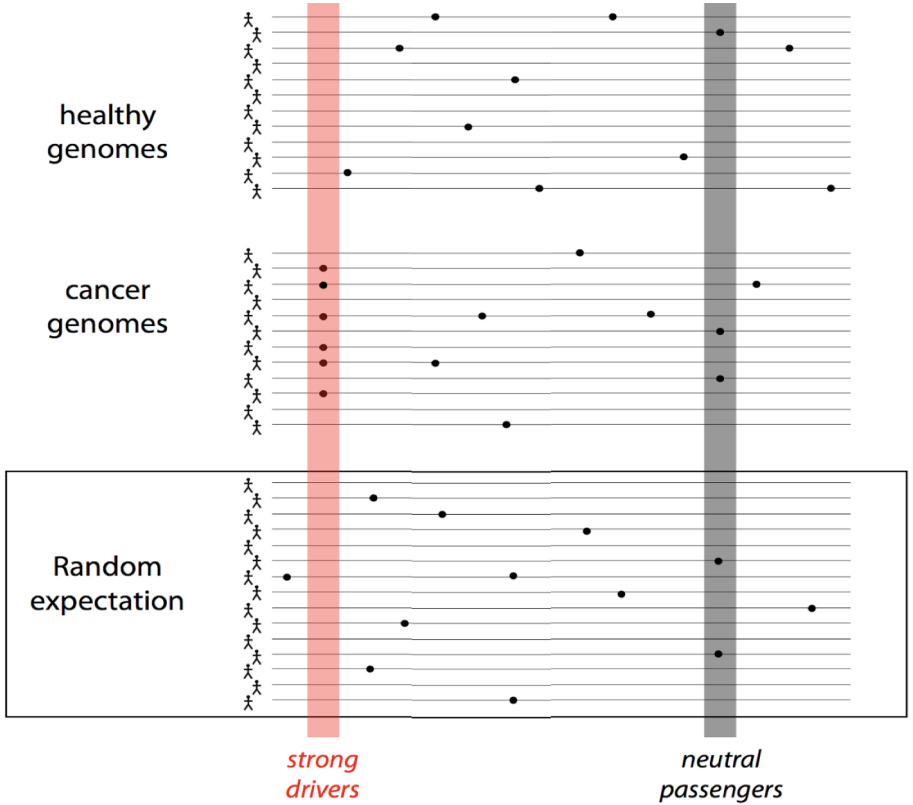
A typical tumor contains 2-8 drivers.

identified through signals of positive selection.

Existing cohorts of ~100s give enough power to identify

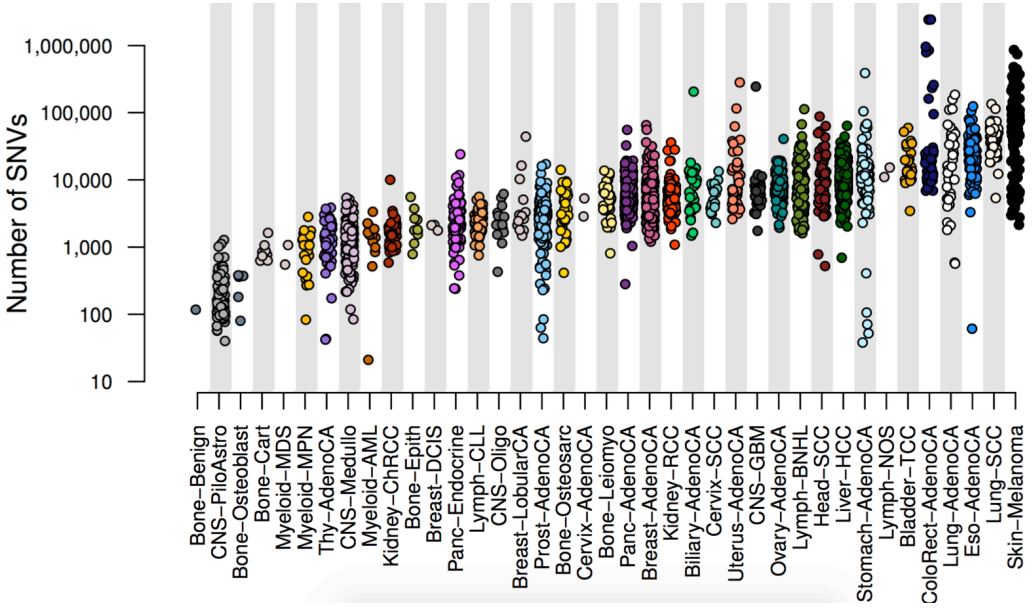
Passengers
Conceptually, a passenger mutation has no direct or indirect effect on tumor progression.

There are 1000s of passengers in a typical cancer genome.



[Vogelstein Science 2013. 339:1546]

PCAWG : most comprehensive resource for cancer whole genome analysis



Adapted from Campbell et. al., bioRxiv ('17)

Project Goals:

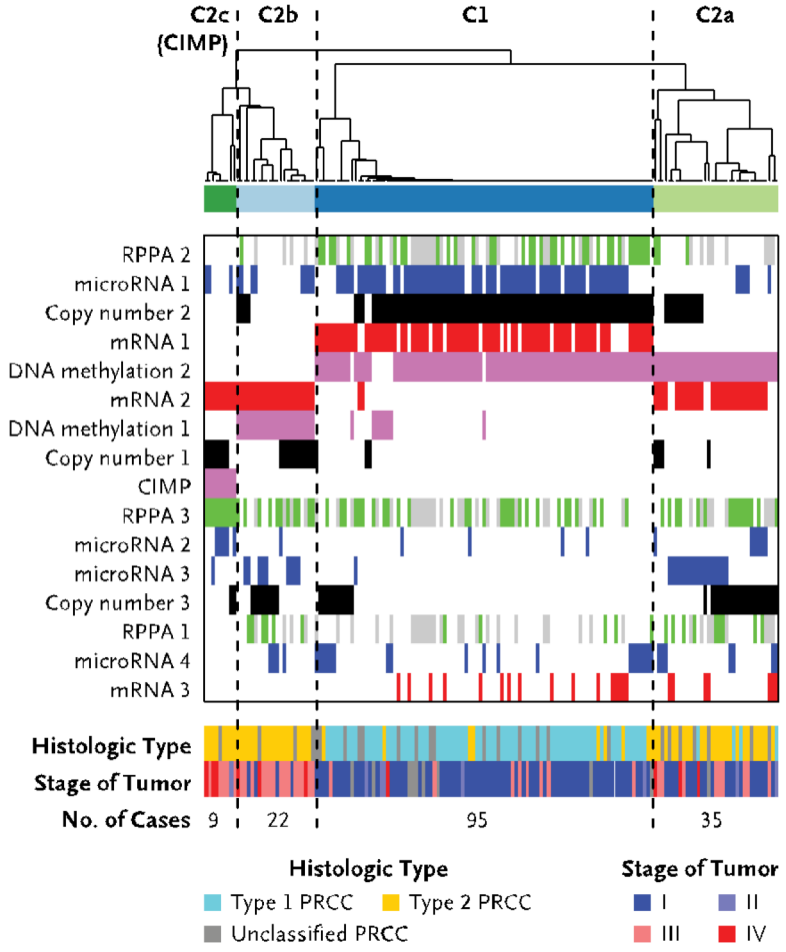
- To understand role of non-coding regions of cancer genomes in disease progression.
- **Union of TCGA-ICGC efforts**

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

PCAWG
PanCancer Analysis
OF WHOLE GENOMES

A case study: pRCC

- Kidney cancer lifetime risk of 1.6% & the papillary type (pRCC) counts for ~10% of all cases
- TCGA sequenced 161 exomes & classified them into subtypes
- 35 WGS of TN pairs



[Cancer Genome Atlas Research Network N Engl J Med. ('16)]

(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

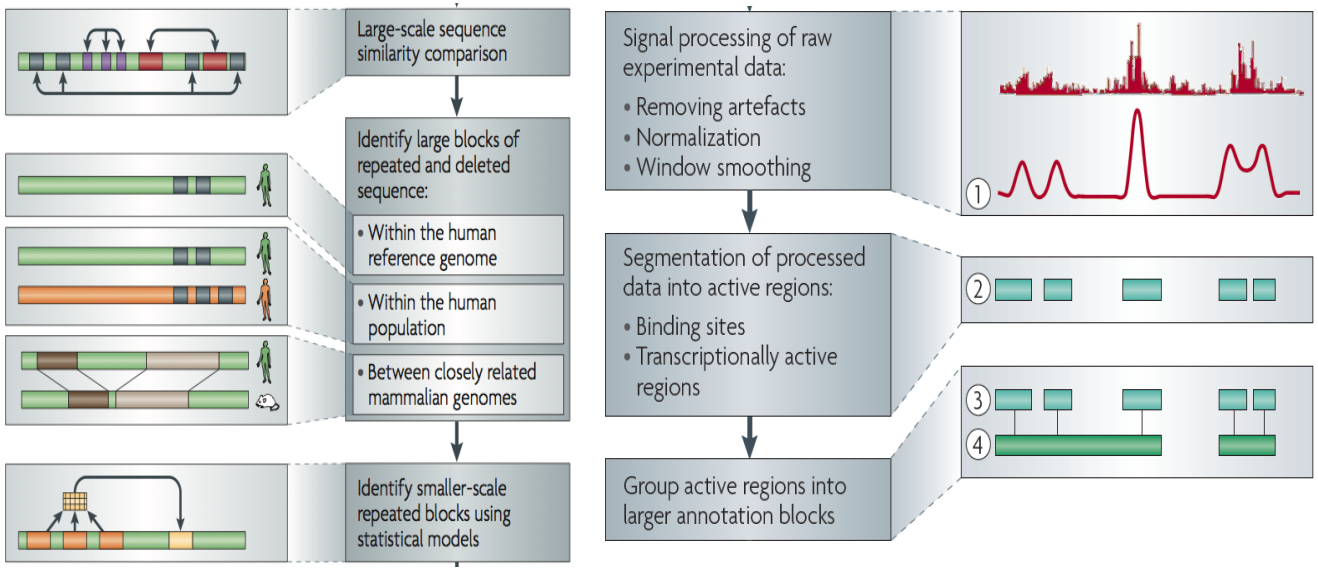
- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

Non-coding Annotations: Overview

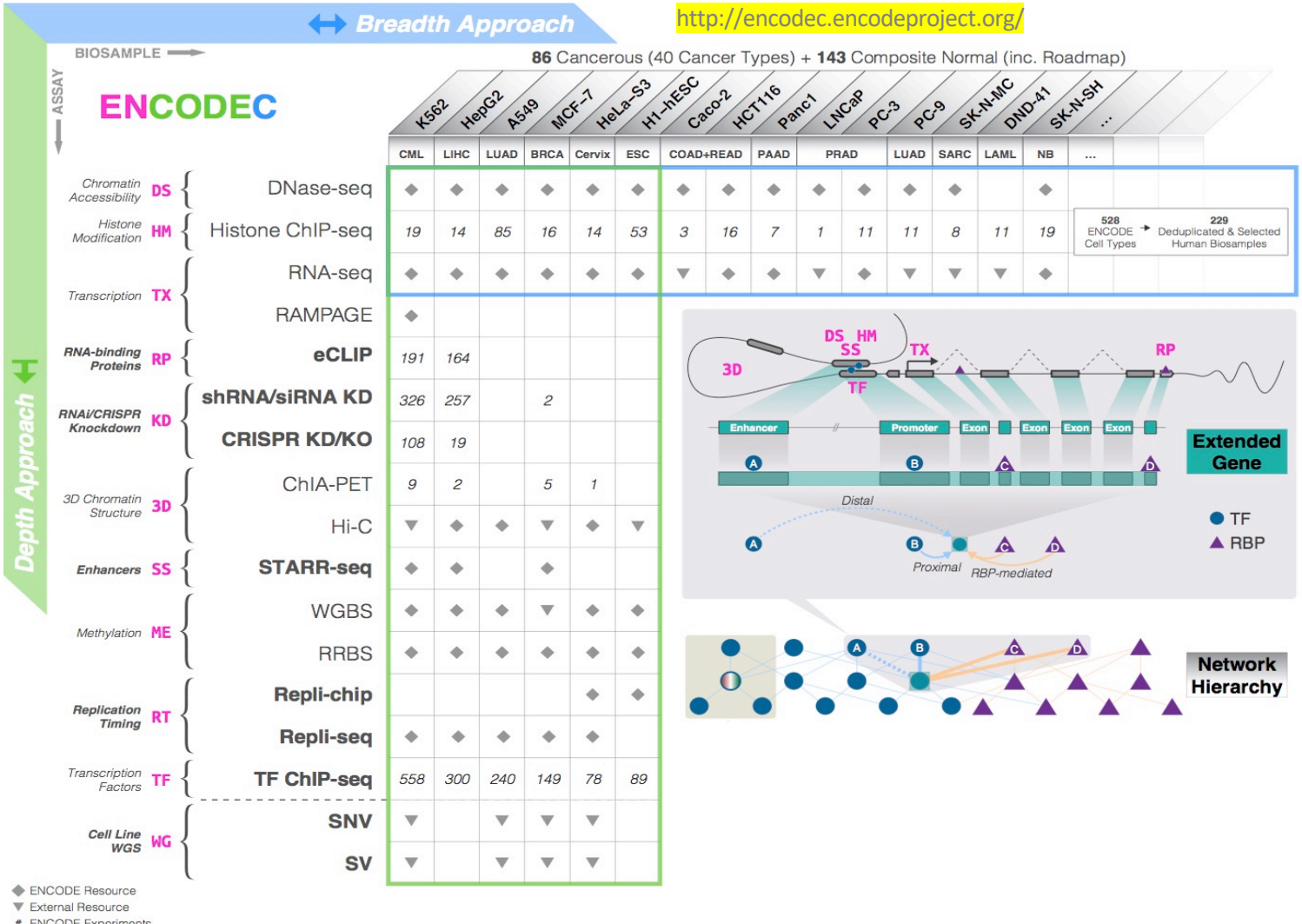
Features are often present on multiple ”scale” (eg elements and connected networks)

Sequence features, incl. **Conservation**

Functional Genomics
Chip-seq (Epigenome & seq. specific TF)
and ncRNA & un-annotated transcription



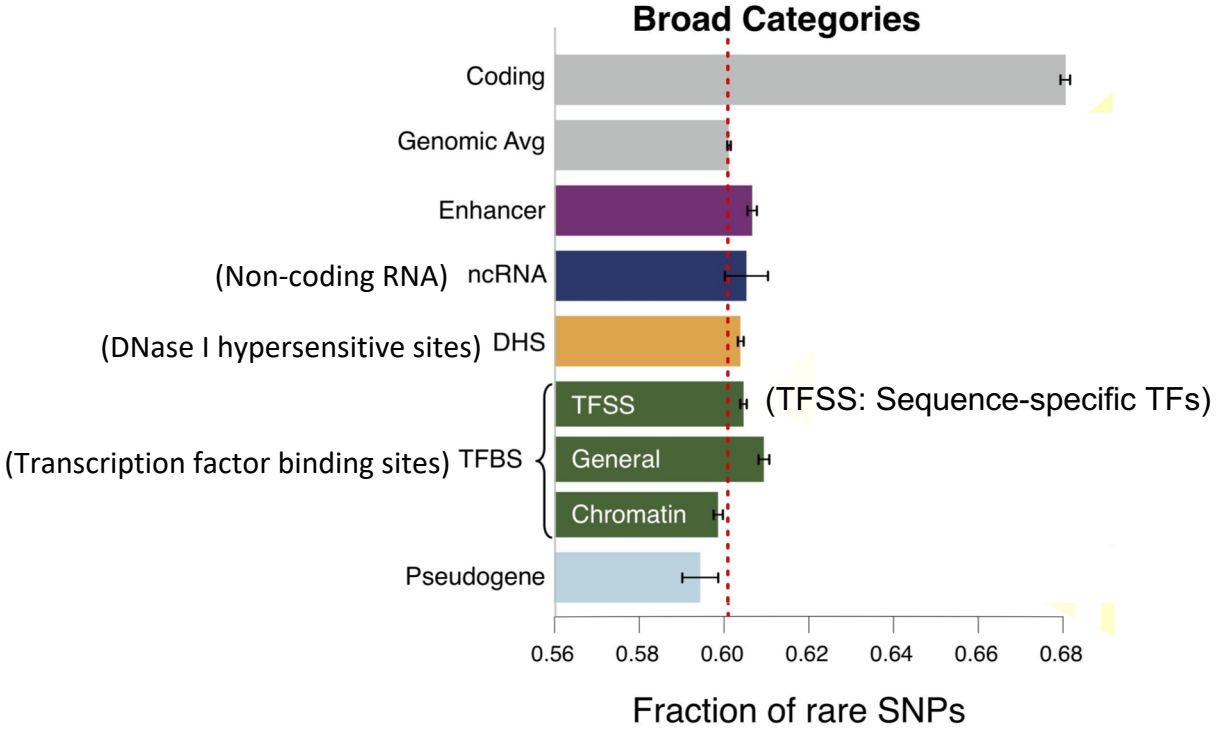
[Alexander et al., *Nat. Rev. Genet.* (’10)]



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

Finding "Conserved" Sites in the Human Population:

Negative selection in non-coding elements based on Production ENCODE & 1000G Phase 1

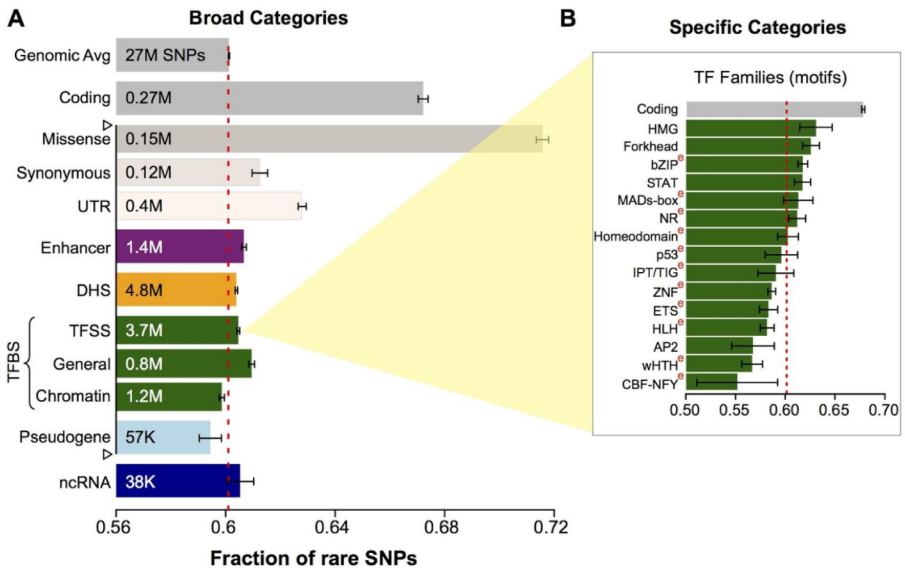


Depletion of Common Variants in the Human Population

Broad categories of regulatory regions under negative selection Related to:

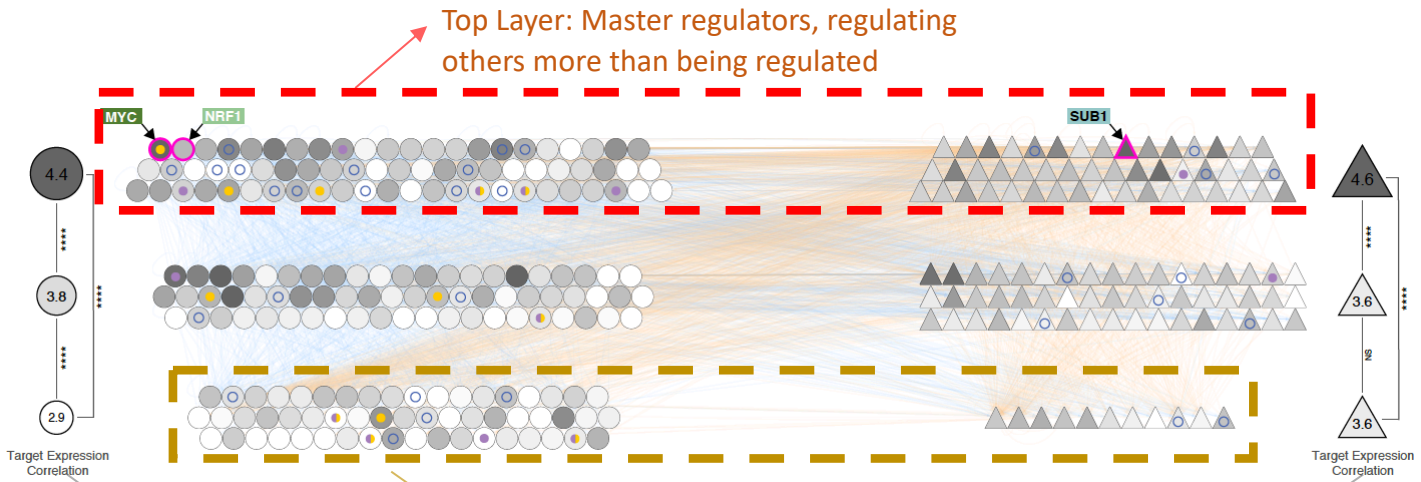
ENCODE, *Nature*, 2012
Ward & Kellis, *Science*, 2012
Mu et al, *NAR*, 2011

Differential selective constraints among specific sub-categories



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

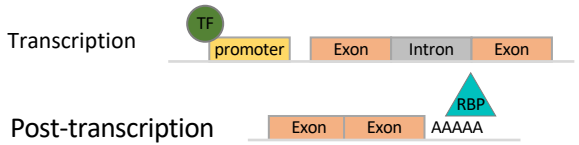
[Khurana et al., *Science* ('13)]



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

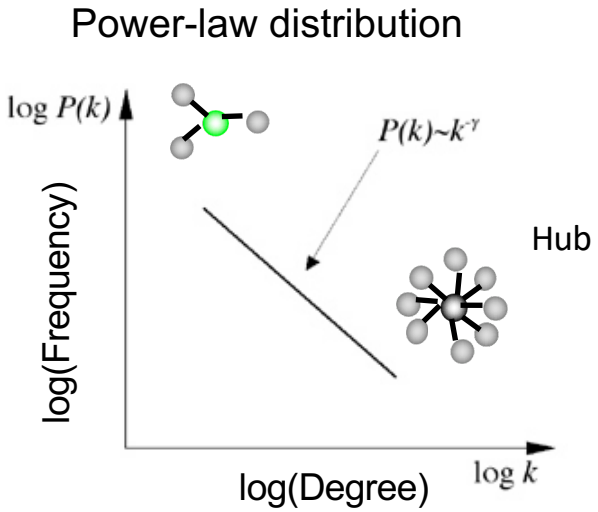
TF-RBP crosstalk

TF-RBP regulate the same gene at different levels



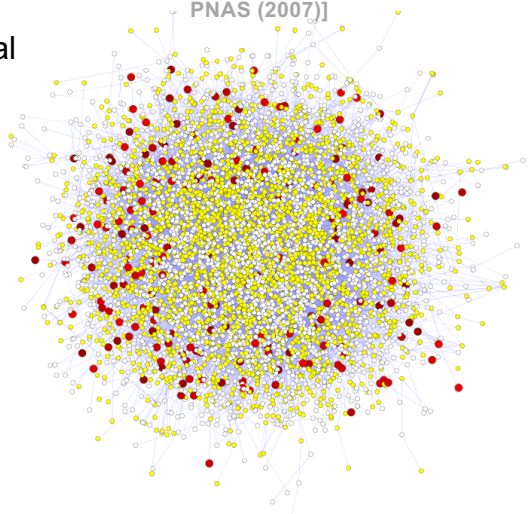
[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

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

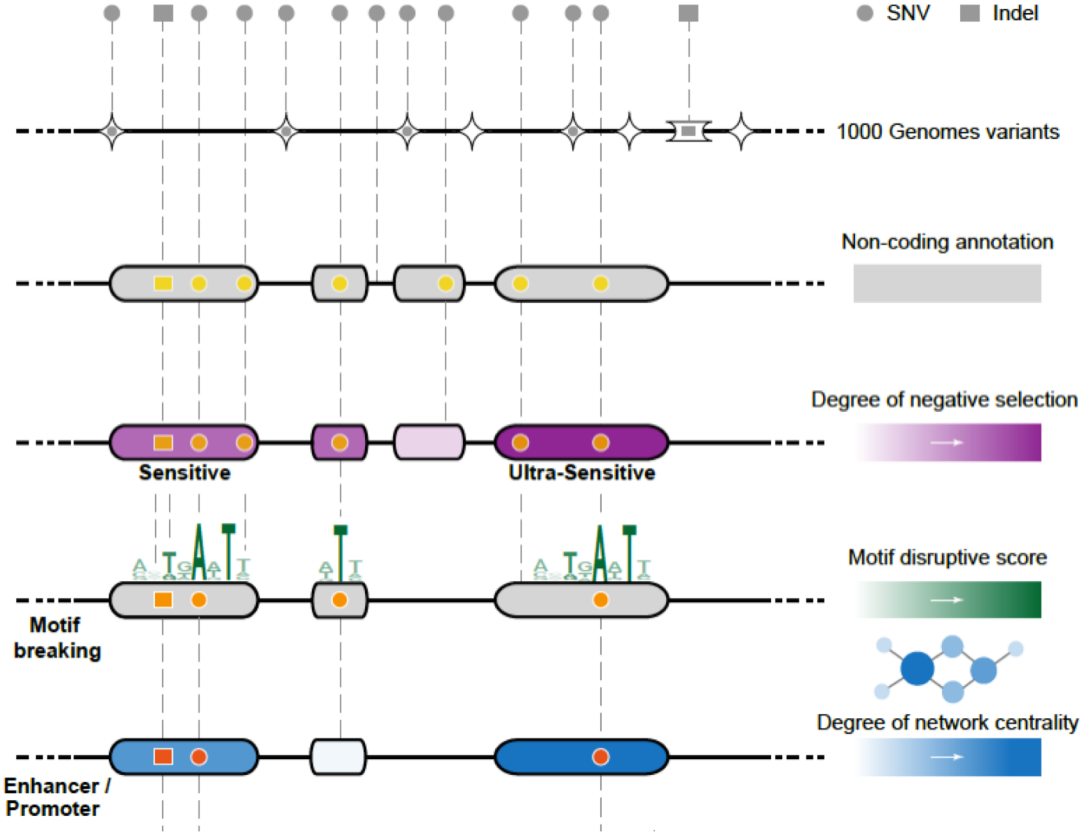
Funseq: a flexible framework to determine functional impact & use this to prioritize variants

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

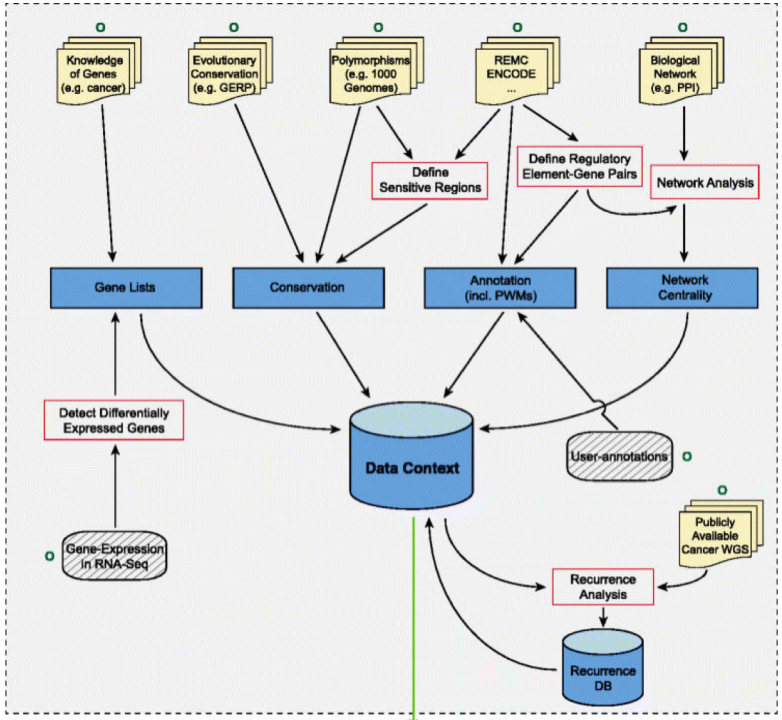
Conservation (GERP, allele freq.)

Mutational impact (motif breaking, Lof)

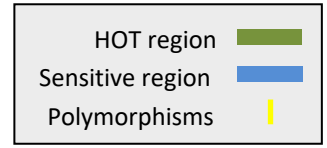
Network (centrality position)



[Fu et al., GenomeBiology ('14), , Khurana et al., Science ('13)]



FunSeq.gersteinlab.org

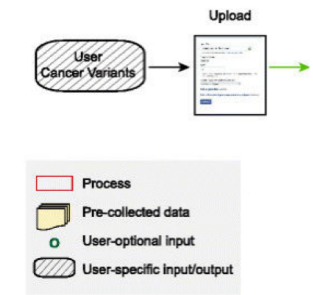


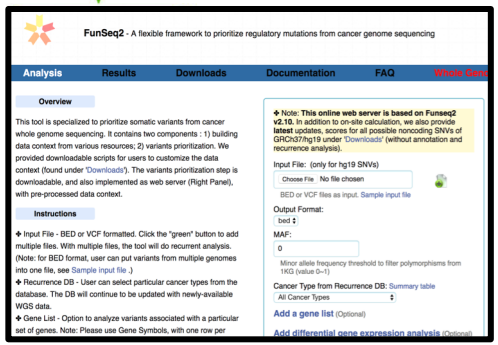
Genome

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

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

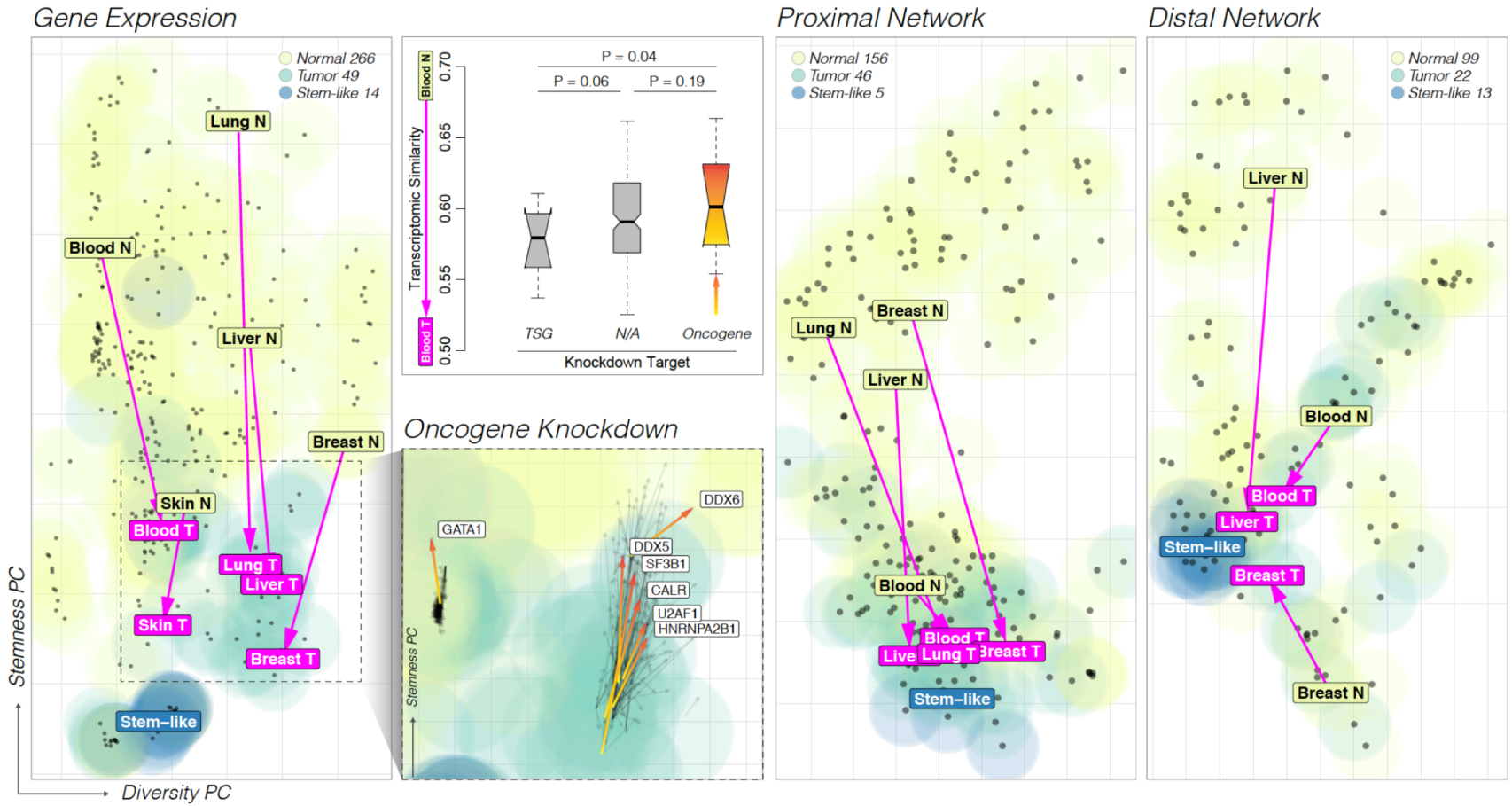
- Practical web server
- Submission of variants & pre-computed large data context from uniformly processing large-scale datasets





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

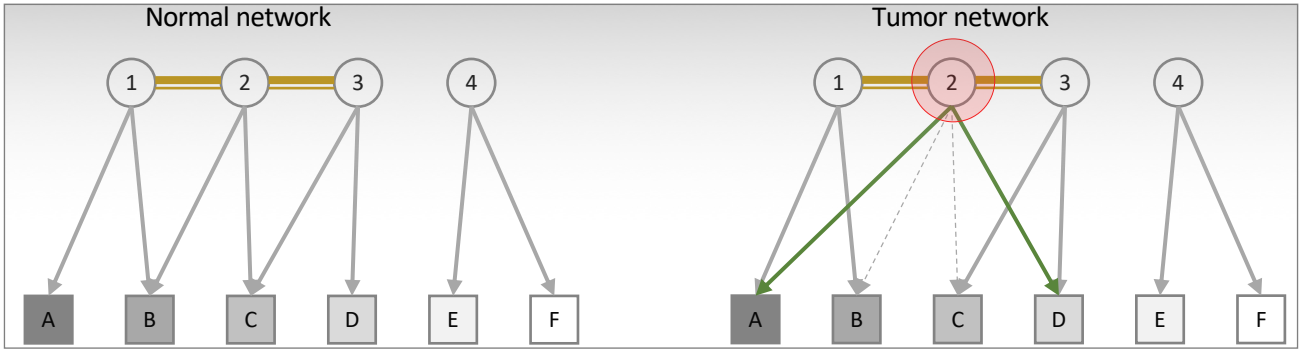
Clustering of ENCODE Biosamples

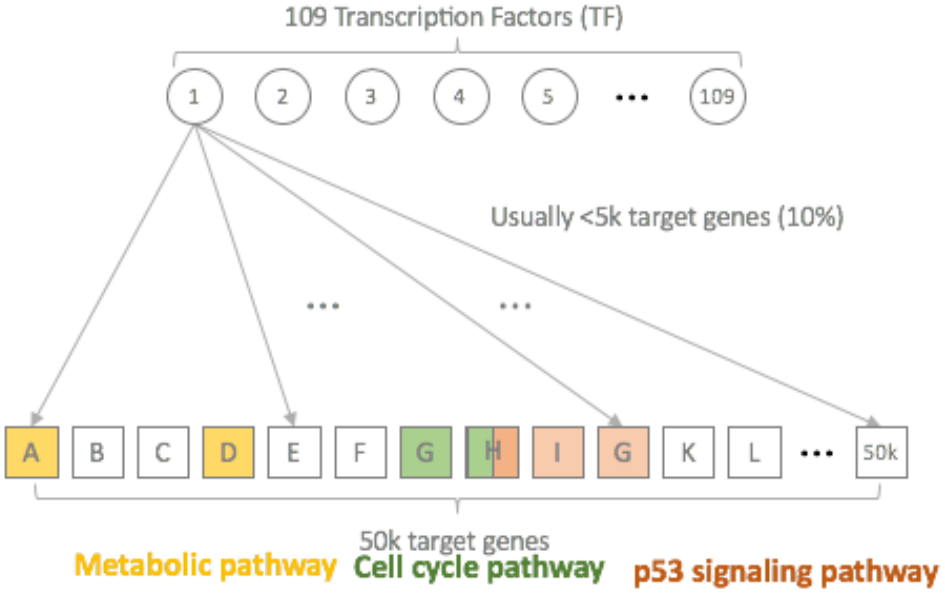


[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

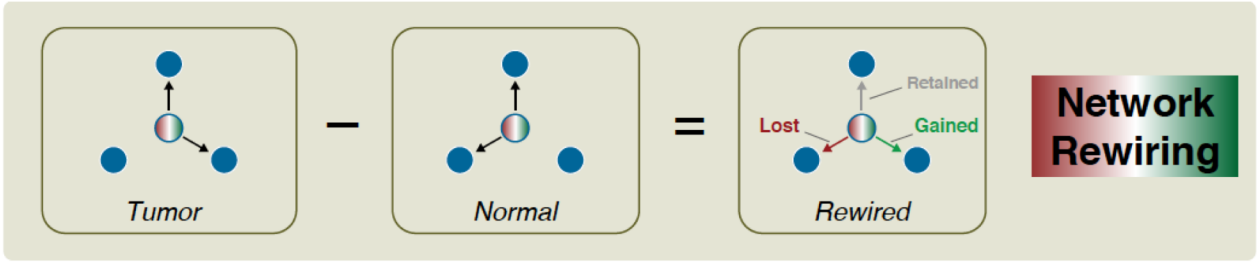
- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers



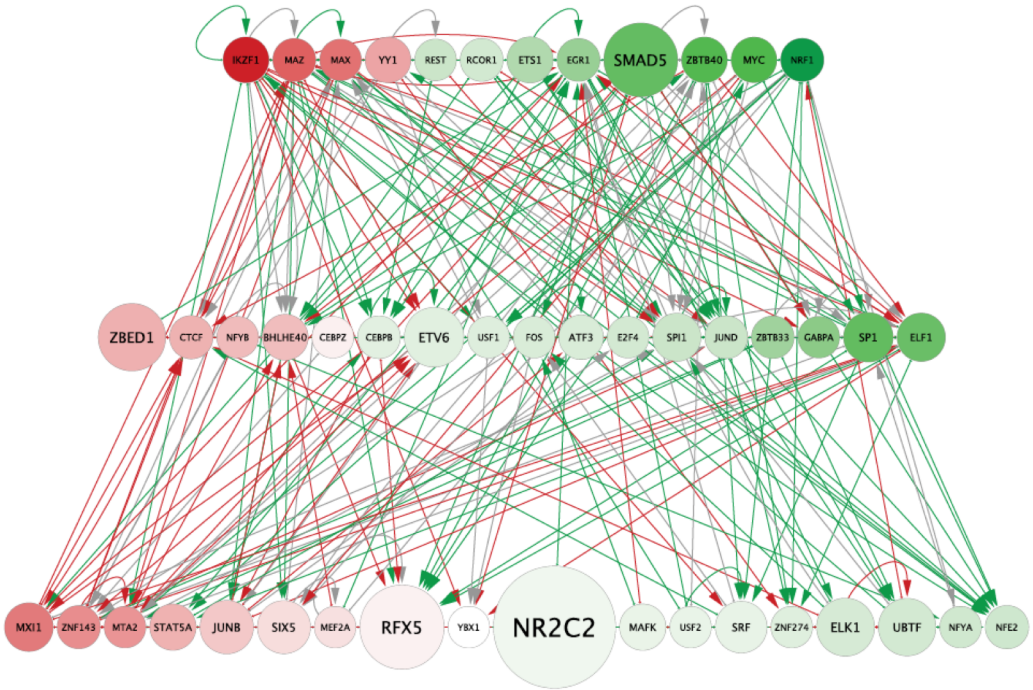


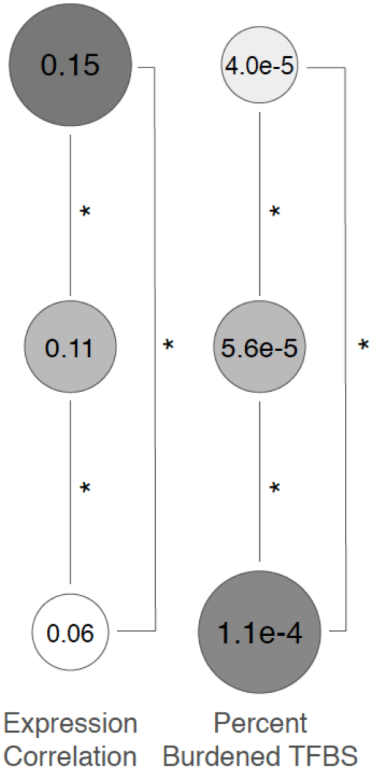
Network rewiring analyses: key cancer-associated regulator identification through network comparisons

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



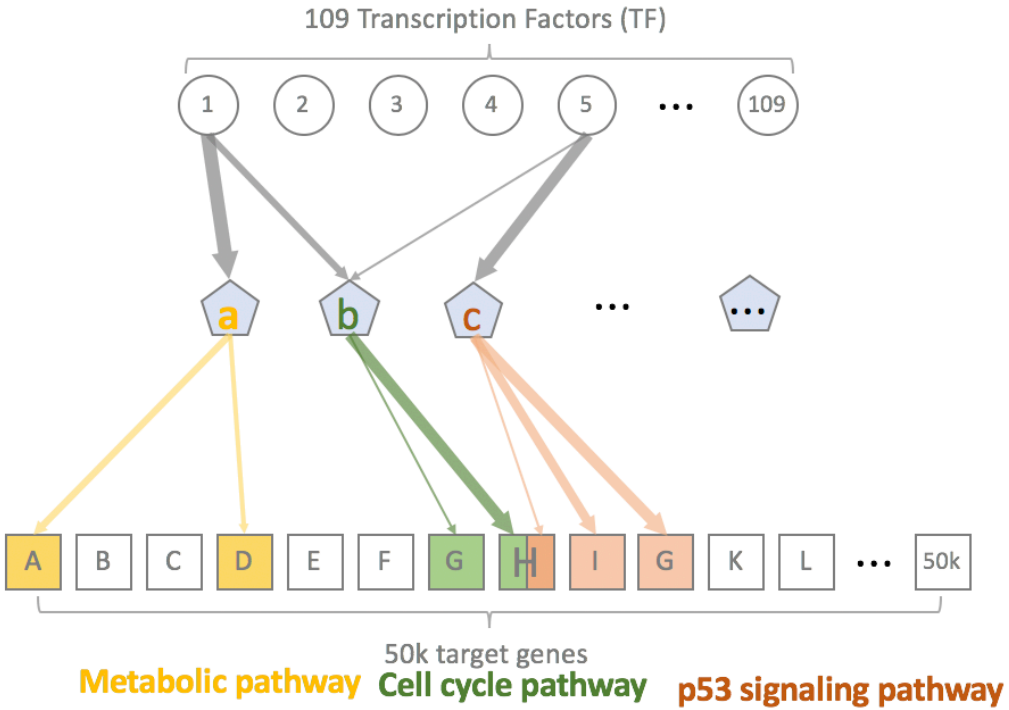
Rewired edges in comparison of GM12878 to K562 109 node TF-TF network (approx. CML)





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

De-noising process by dimension reduction



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

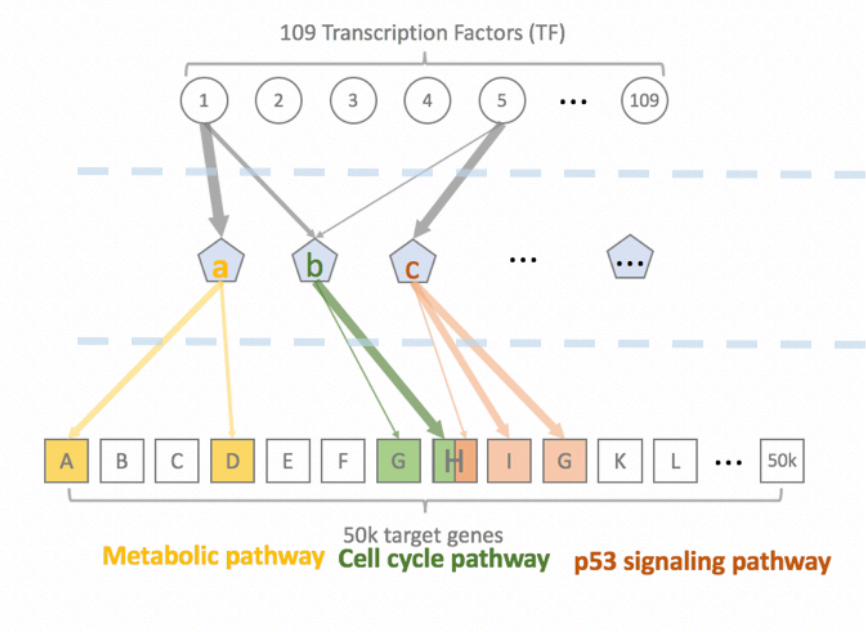
Hidden Layer
(50 biological pathways?)

Challenge: how to define appropriate pathways?

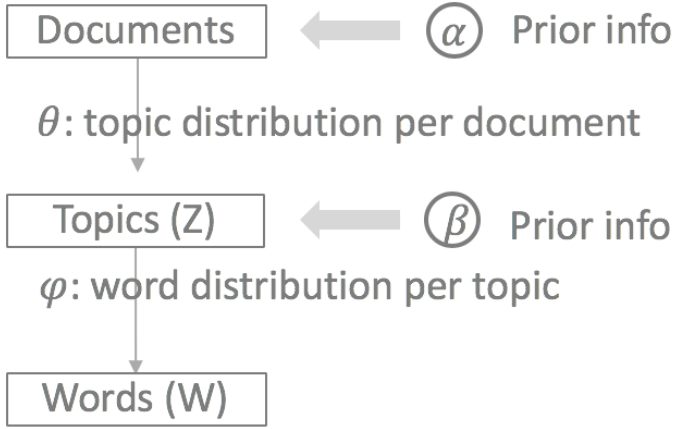
[Zhang et al. (‘20), biorxiv + Nat. Comm. (in press)]

Automatic gene topic identification based on Latent Dirichlet Allocation

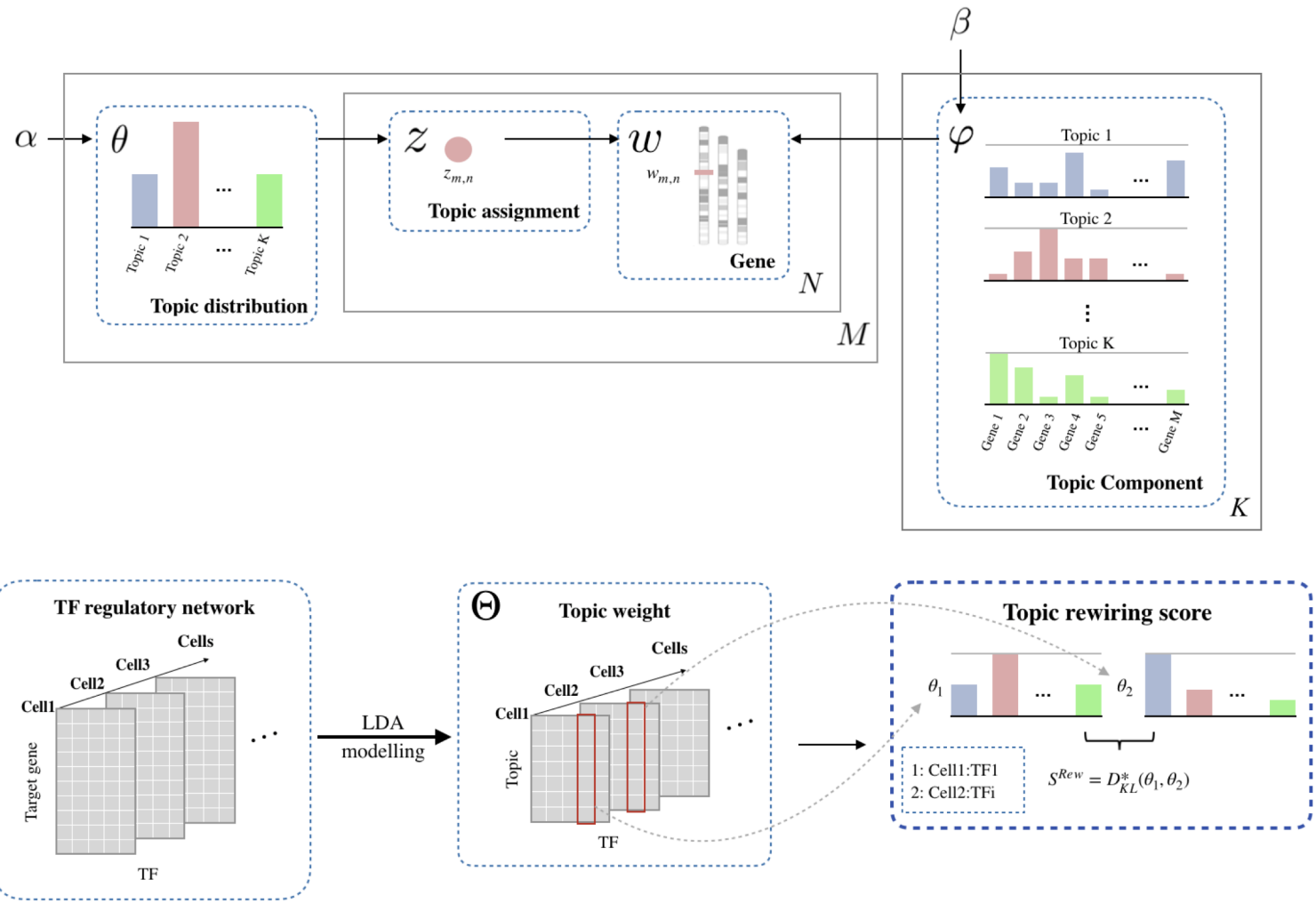
$TF \rightarrow gene$ network

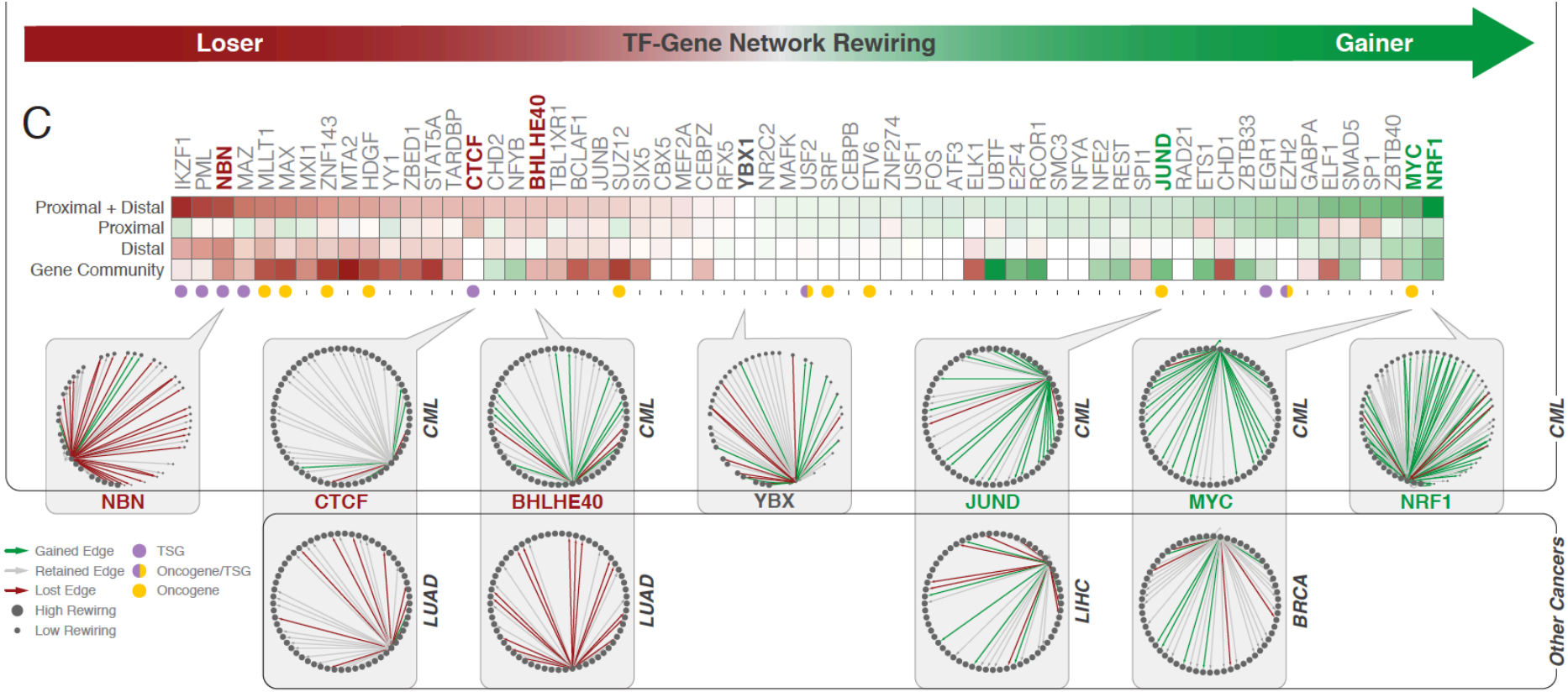


Latent Dirichlet Allocation



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


α
θ
Topic 1
Topic 2
...
Topic K
Topic distribution
z
$z_{m,n}$
Topic assignment
w
$w_{m,n}$
Gene
N
M
β
φ
Topic 1
Topic 2
Topic K
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene M
Topic Component
K
TF regulatory network
Cells
Cell3
Cell2
Cell1
Target gene
TF
LDA
modelling
Θ
Topic weight
Cells
Cell3
Cell2
Cell1
Topic
TF
Topic rewiring score
θ_1
θ_2
1: Cell1:TF1
2: Cell2:TFi
$S^{Rew} = D^*_{KL}(\theta_1, \theta_2)$

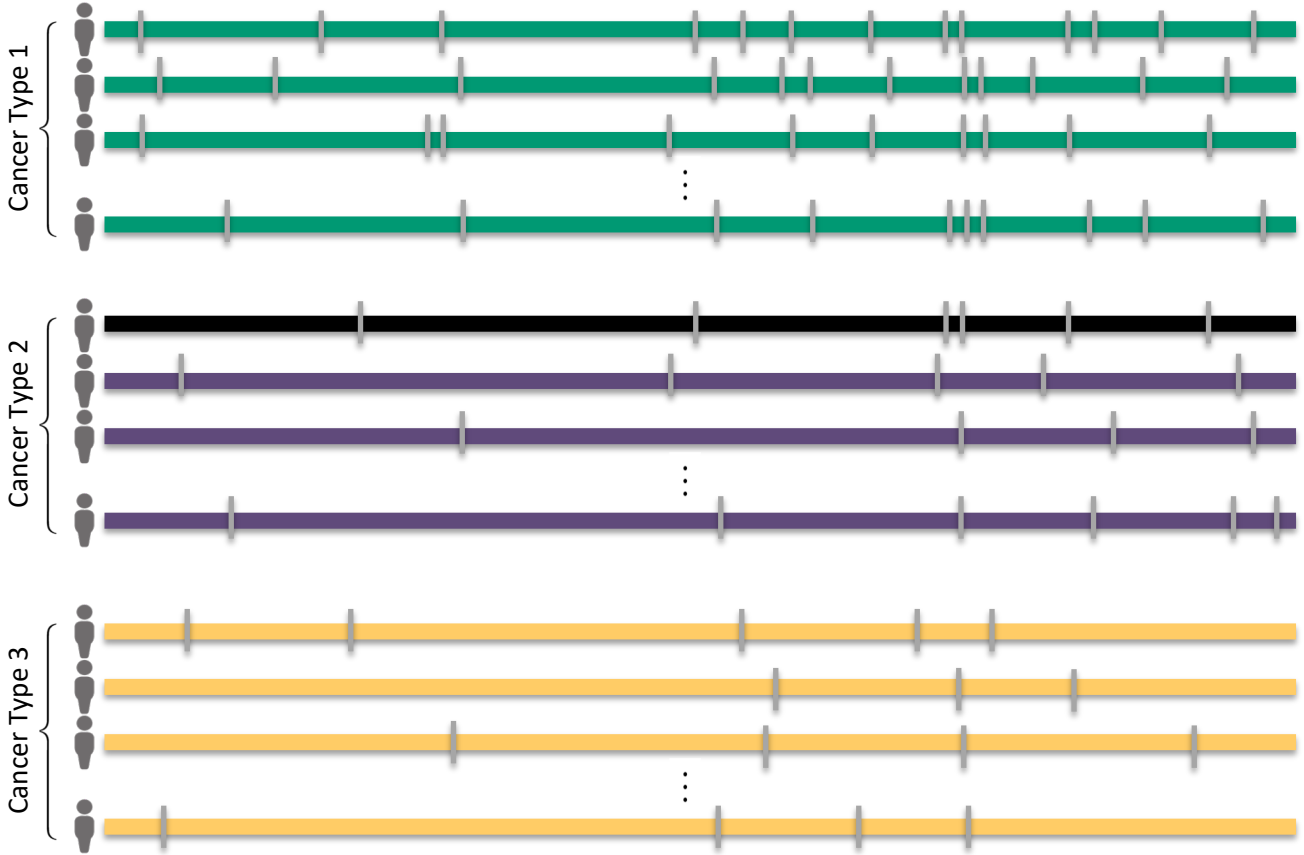


[Zhang et al. (‘20), biorxiv + Nat. Comm. (in press)]

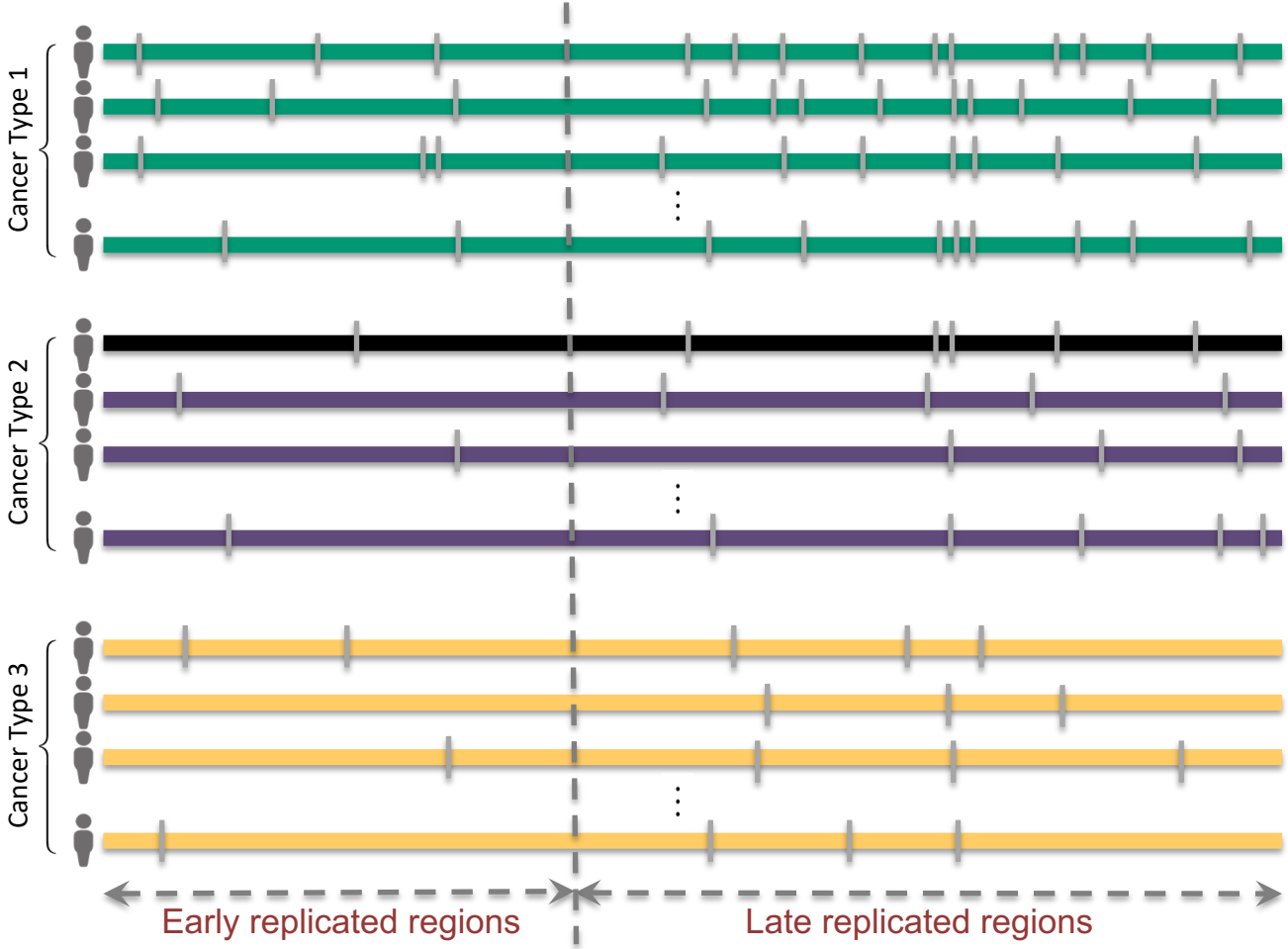
(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

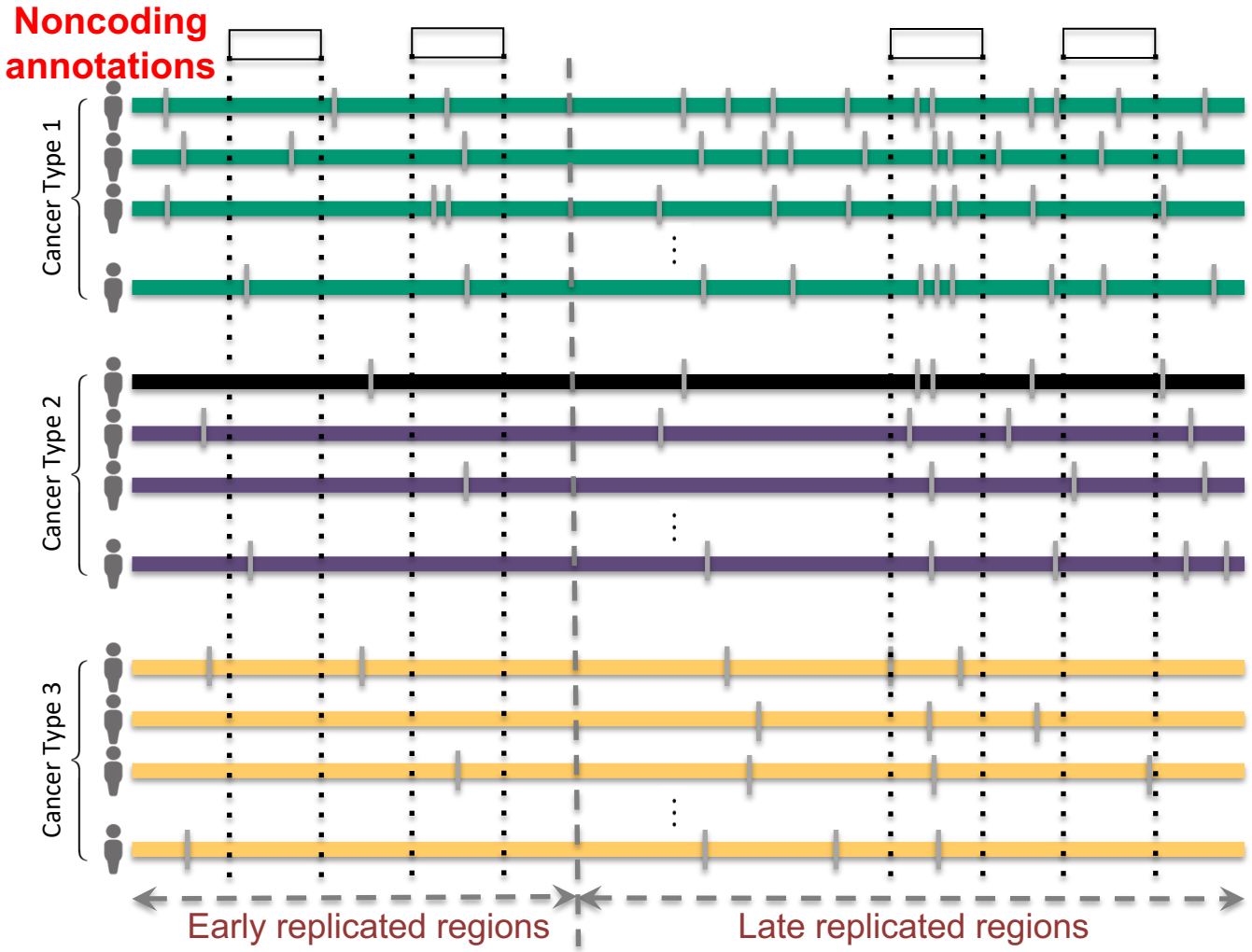
- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

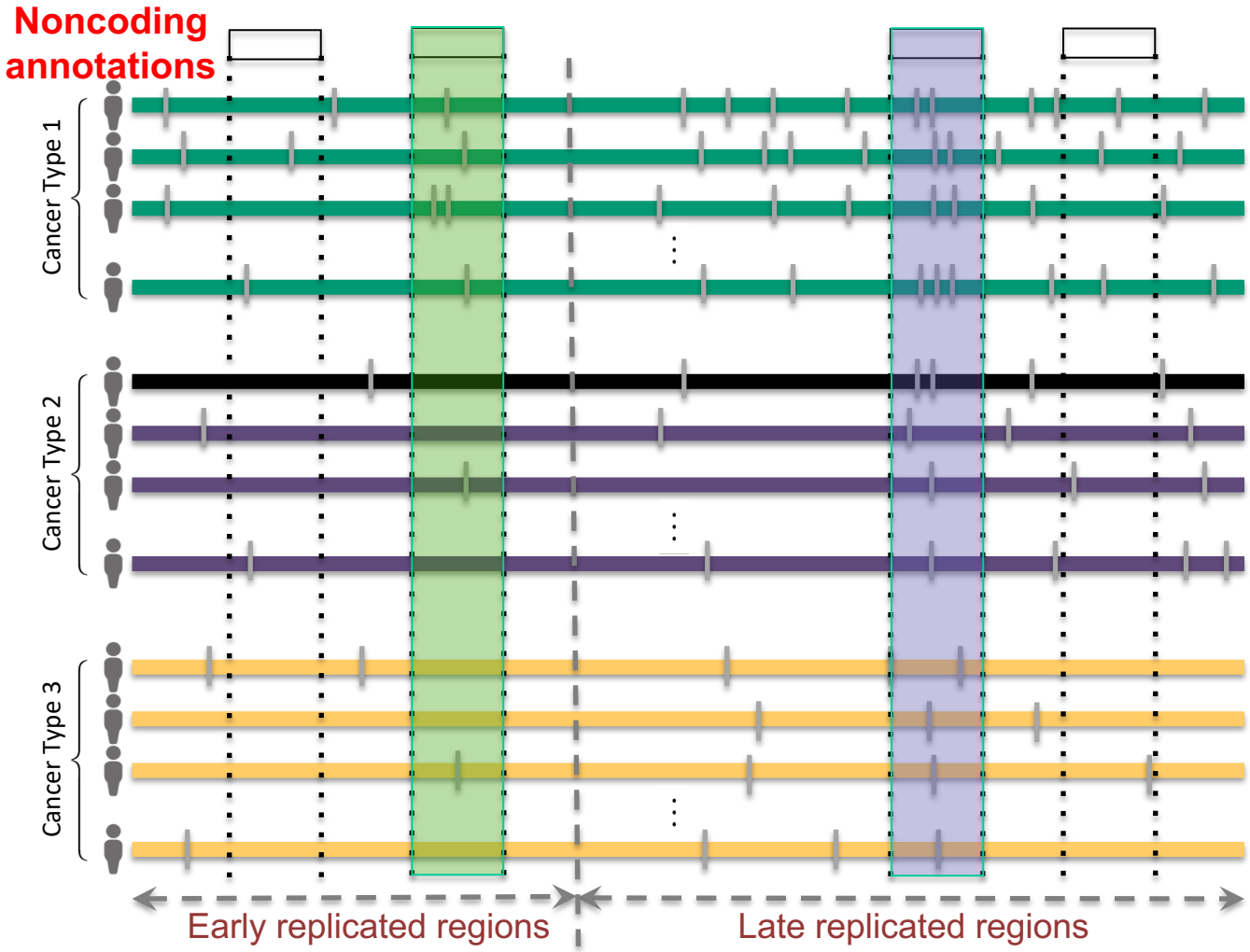
Mutation recurrence



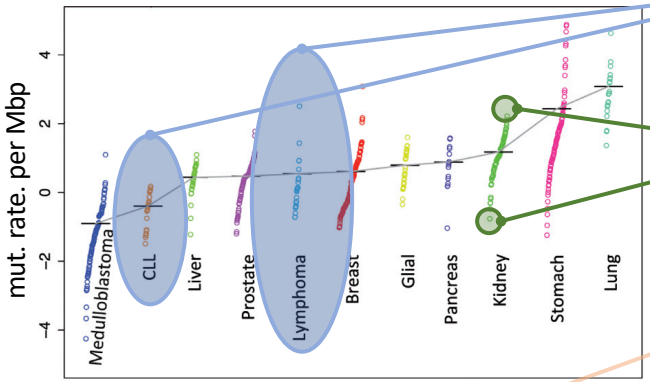
Mutation recurrence

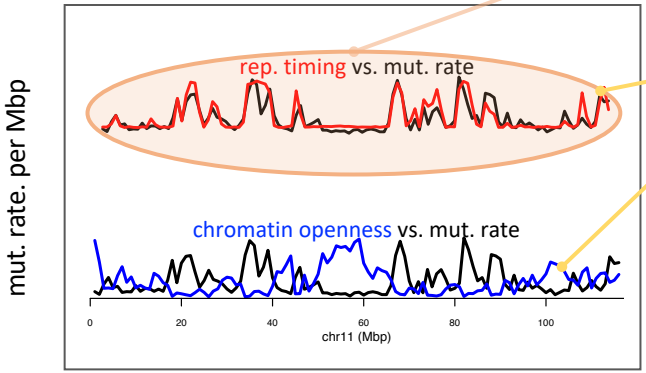



Noncoding annotations
Cancer Type 1
Cancer Type 2
Cancer Type 3
Early replicated regions
Late replicated regions


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)]

Cancer Somatic Mutation Modeling

PARAMETRIC MODELS

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

$x_i : Binomial(n_i, p)$

Model 2a: Varying Mutation Rate with Single Covariate Correction

$x_i : Binomial(n_i, p_i)$

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

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

Model 2b: Varying Mutation Rate with Multiple Covariate Correction

$x_i : Binomial(n_i, p_i)$

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

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

[Lochovsky et al. *NAR* ('15)]

- Suppose there are *k* 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: the covariate rank of the element
- 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

Assume constant background mutation rate in local regions.

Model 3a: Random Permutation of Input Annotations

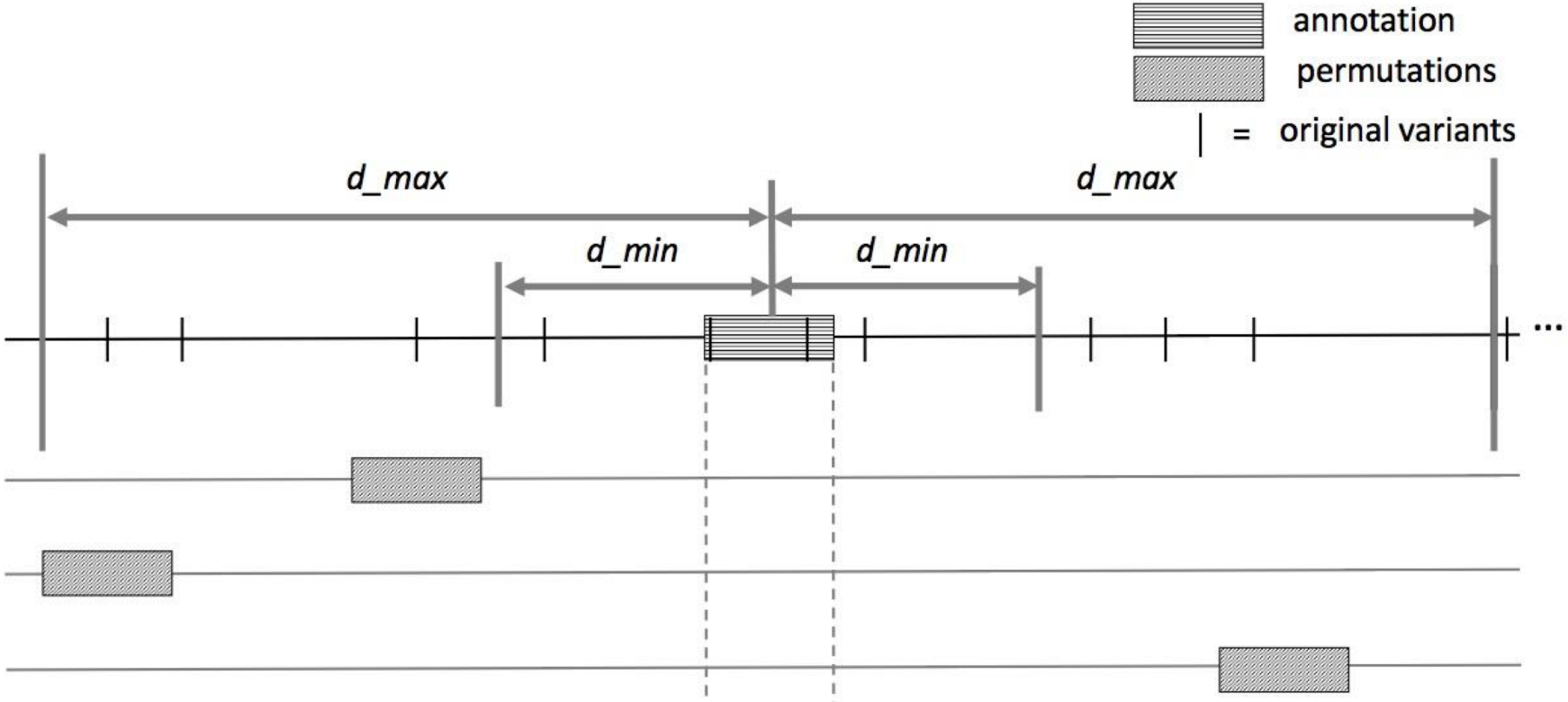
Shuffle annotations within local region to assess background mutation rate.

Model 3b: Random Permutation of Input Variants

Shuffle variants within local region to assess background mutation rate.

[Lochovsky et al. *Bioinformatics* in press]

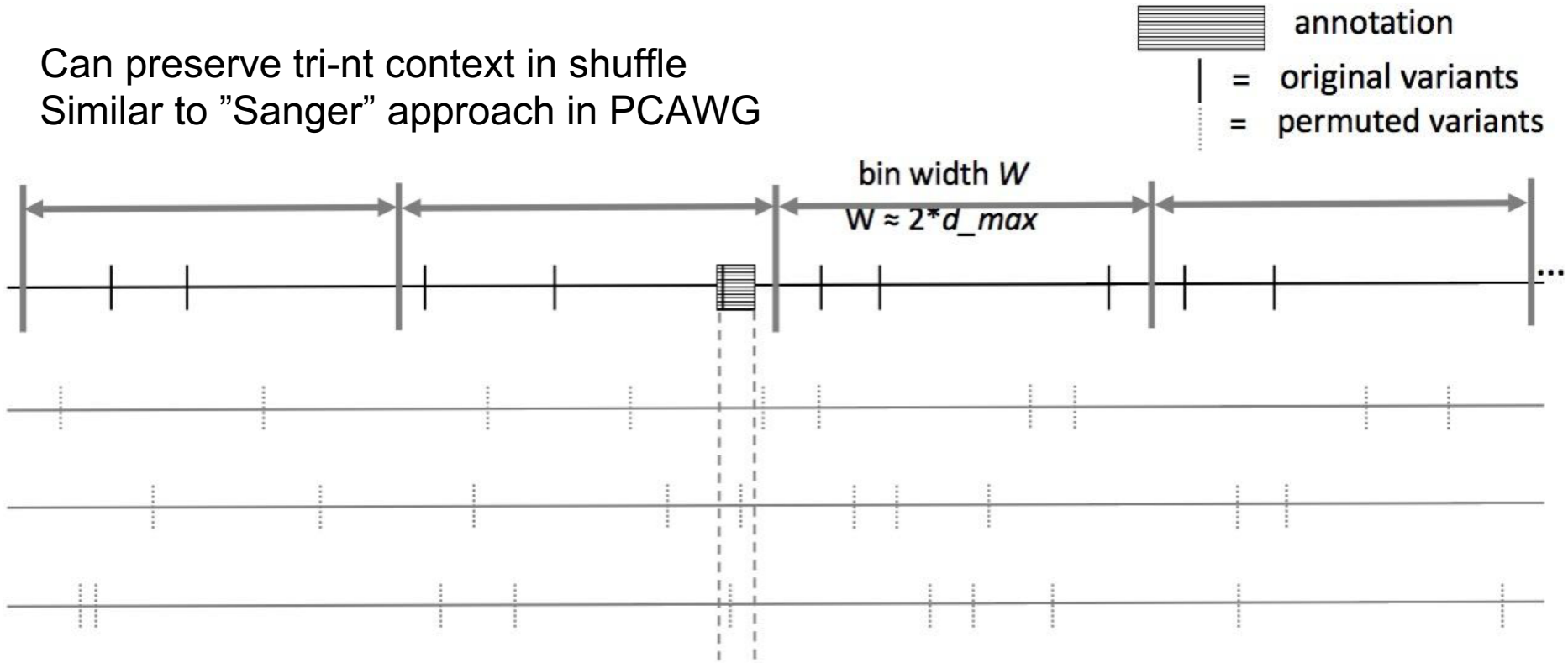
MOAT-a: Annotation-based permutation



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

MOAT-v: Variant-based Permutation

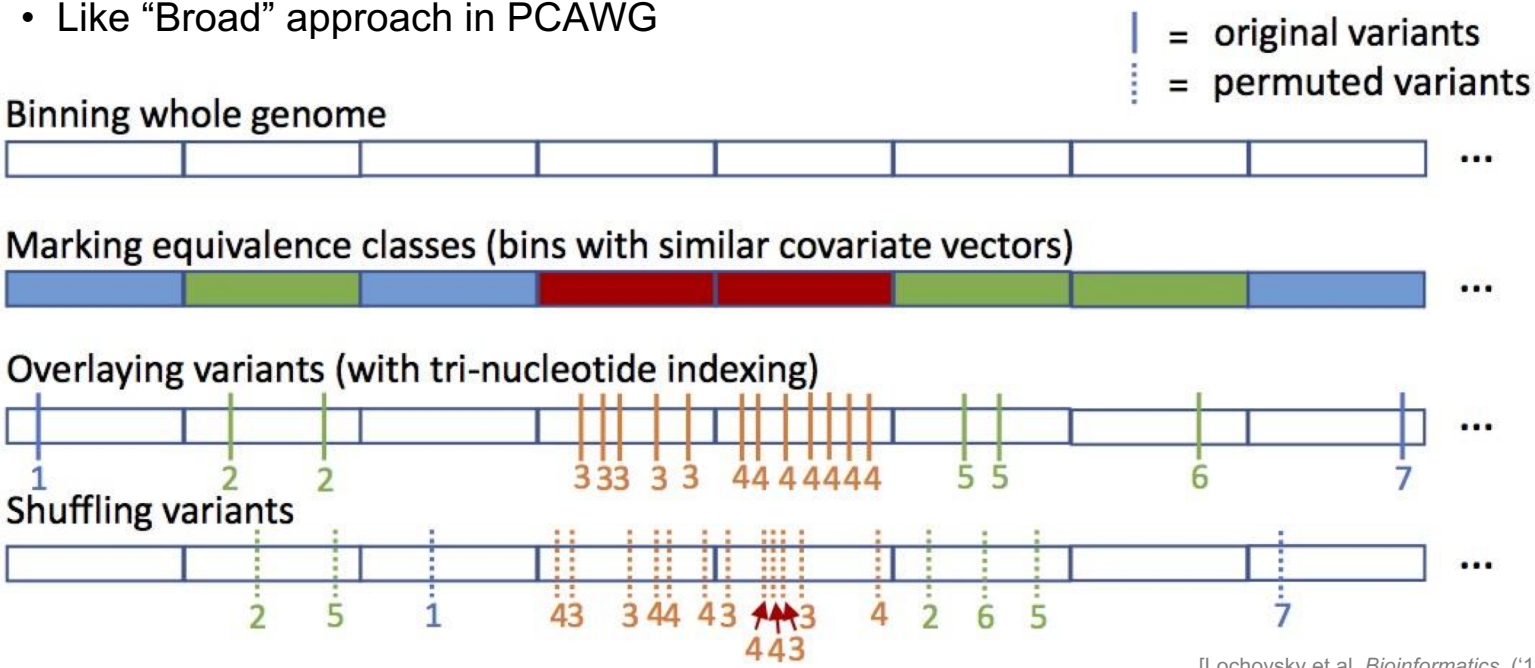
Can preserve tri-nt context in shuffle
Similar to ”Sanger” approach in PCAWG



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

MOAT-s: a variant on MOAT-v

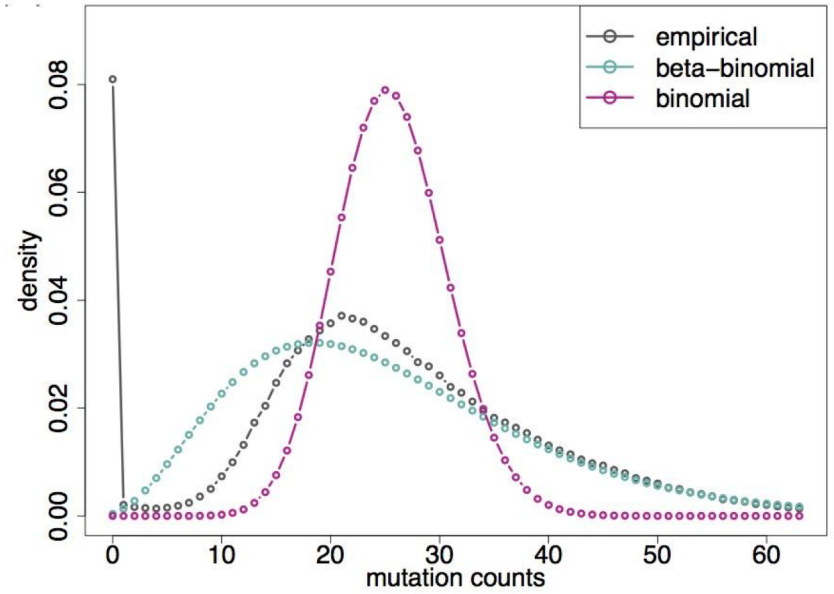
- A somatic variant simulator
  - Given a set of input variants, shuffle to new locations, taking genome structure into account
  - Like “Broad” approach in PCAWG



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

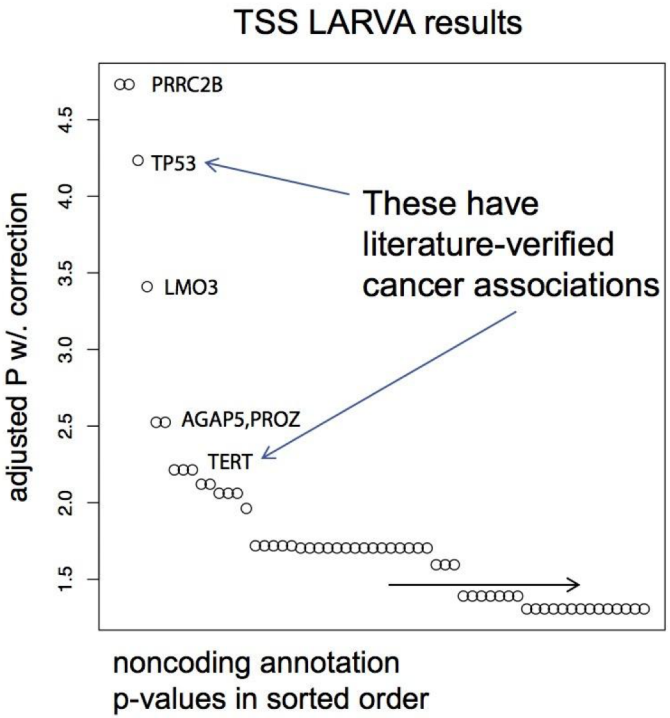
LARVA Model Comparison

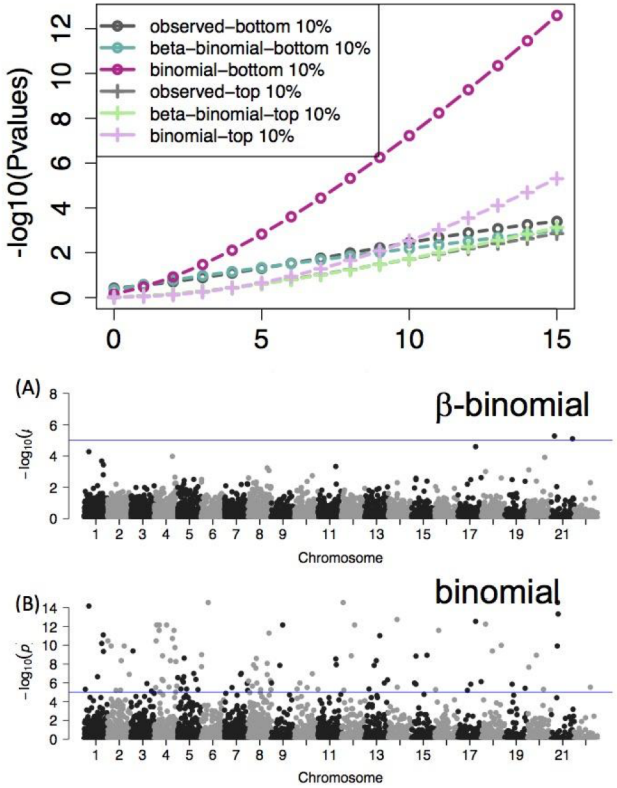
- Comparison of mutation count frequency implied by the binomial model (model 1) and the beta-binomial model (model 2) relative to the empirical distribution
- The beta-binomial distribution is significantly better, especially for accurately modeling the over-dispersion of the empirical distribution



[Lochovsky et al. *NAR* ('15)]

LARVA Results



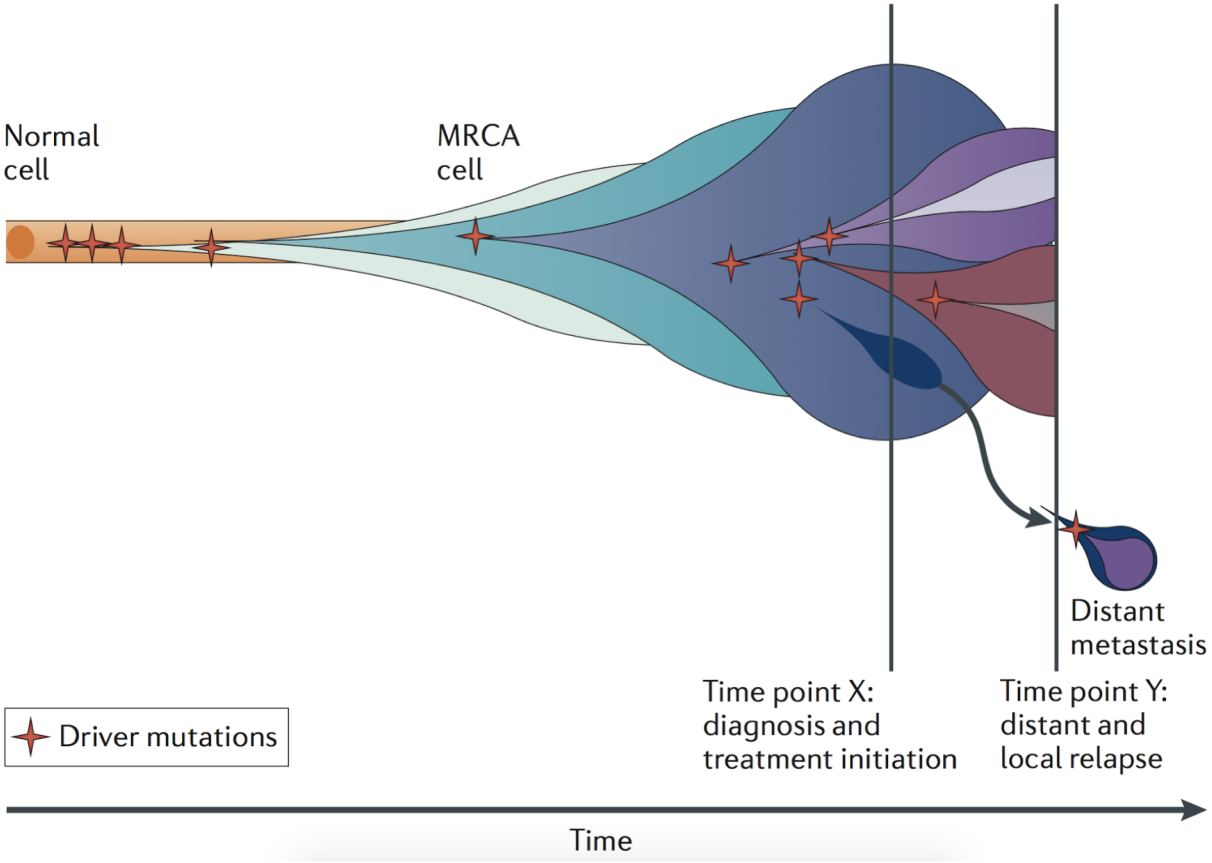


[Lochovsky et al. *NAR* ('15)]

(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

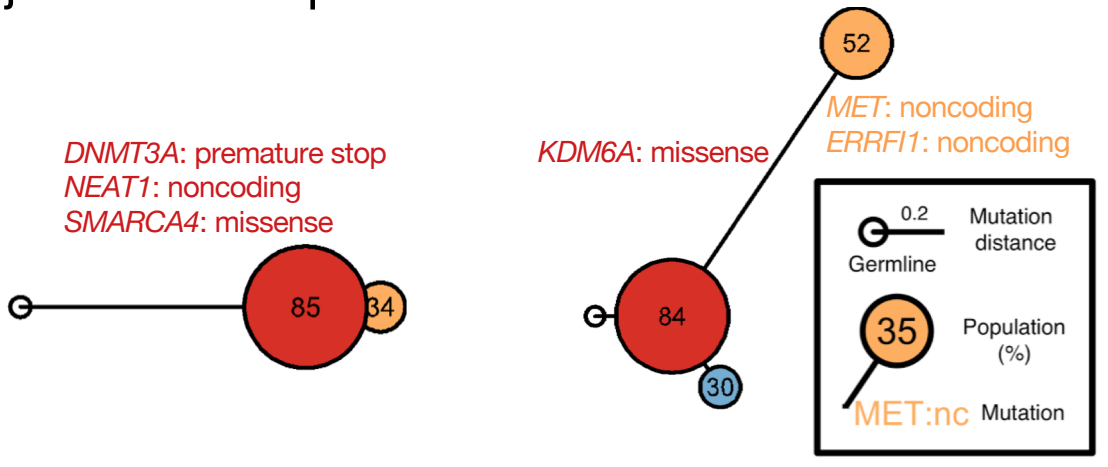
Tumor Evolution: Highlight the Ordering of Key Mutations



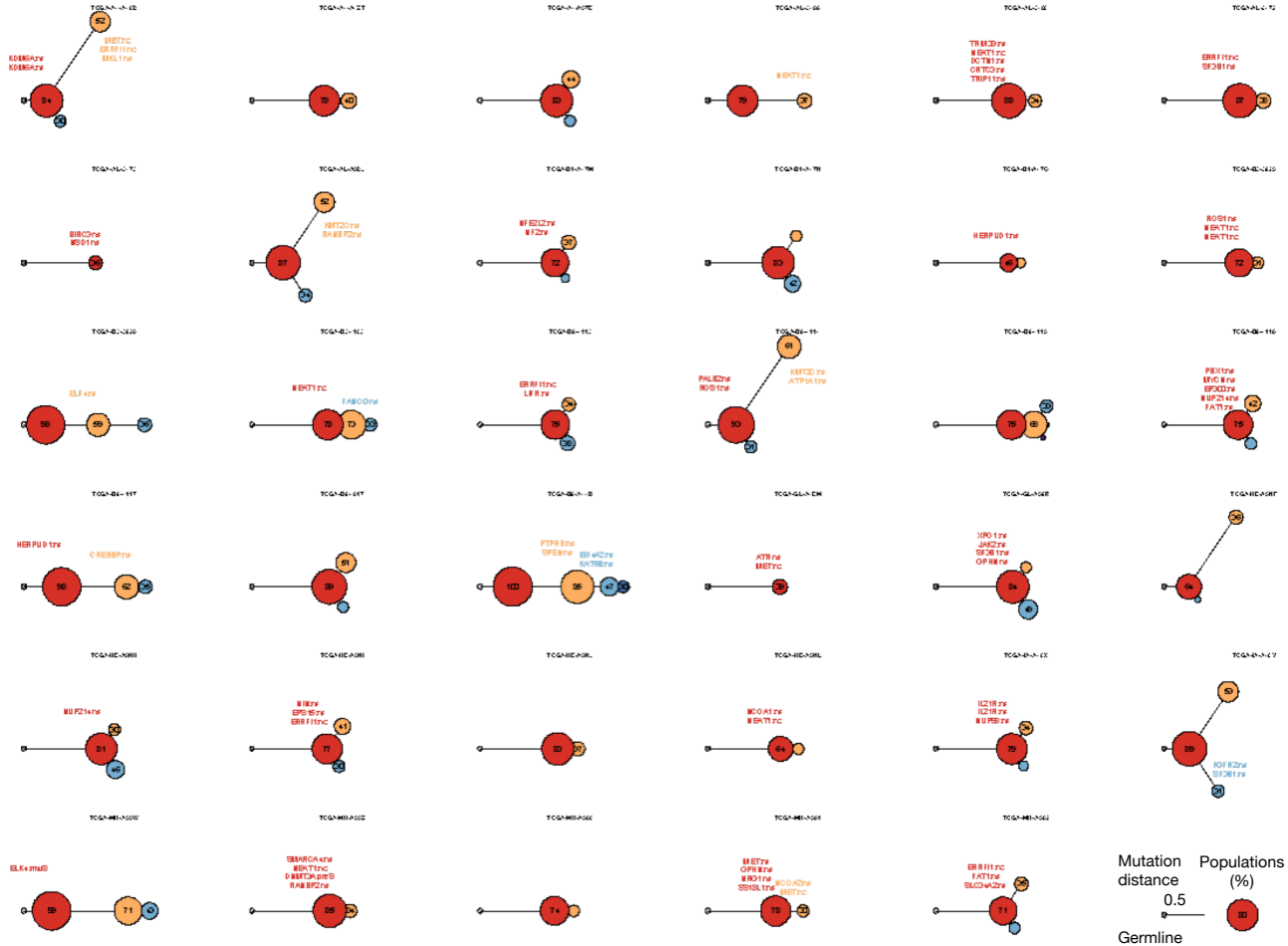
Yates et al, NRG (2012)

Construct evolutionary trees in pRCC

- Infer mutation order and tree structure based on mutation abundance (PhyloWGS, Deshwar et al., 2015)
- Some of the key mutations occur in all the clones while others are just in some parts of the tree



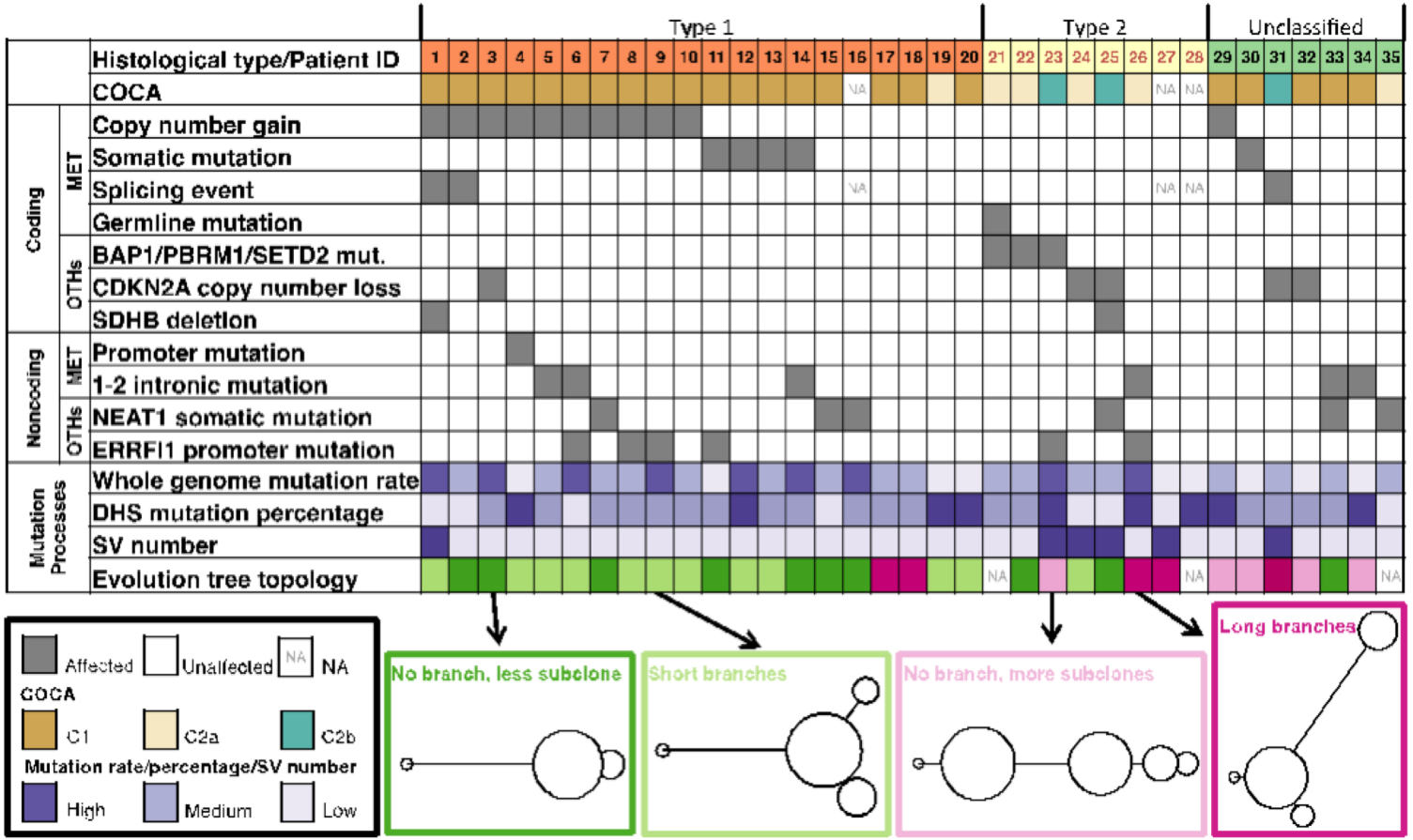
[S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]



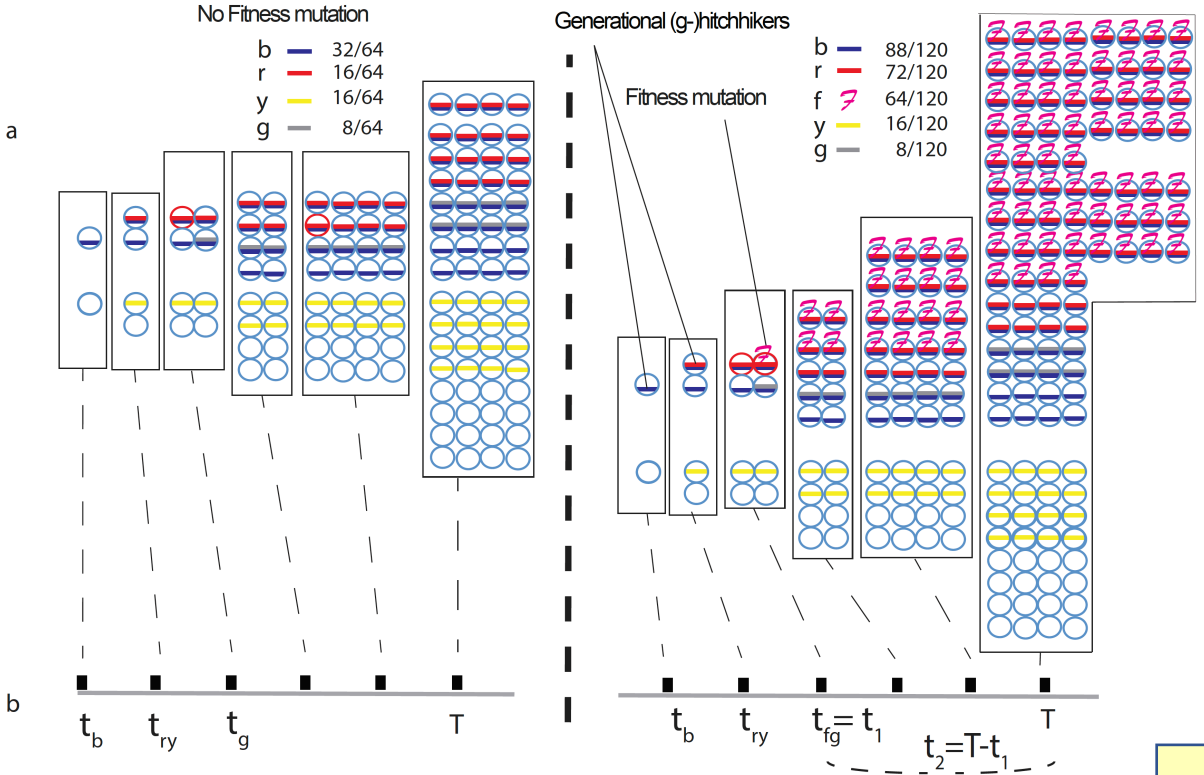
[S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]

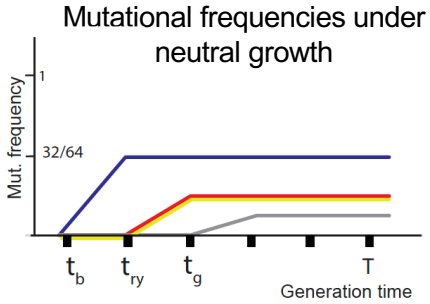
[S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]

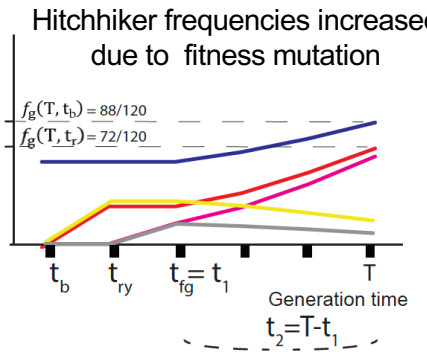
Tree topology correlates with molecular subtypes



Modelling the frequency of “generational” hitchhikers in a fitness population







[Salichos et al. (’20, in press) Nat. Comm.]

$$f_g(T, t_g, t_i - m) = \frac{e^{-r(t_g + t_i - m)} \times (N_{tot} - f_{d(T,t_i)} \times N_{tot}) + \sqrt[k_i]{f_{d(T,t_i)} \times N_{tot}} + f_{d(T,t_i)} \times N_{tot} - \sqrt[k_i]{f_{d(T,t_i)} \times N_{tot}}}{N_{tot}}$$

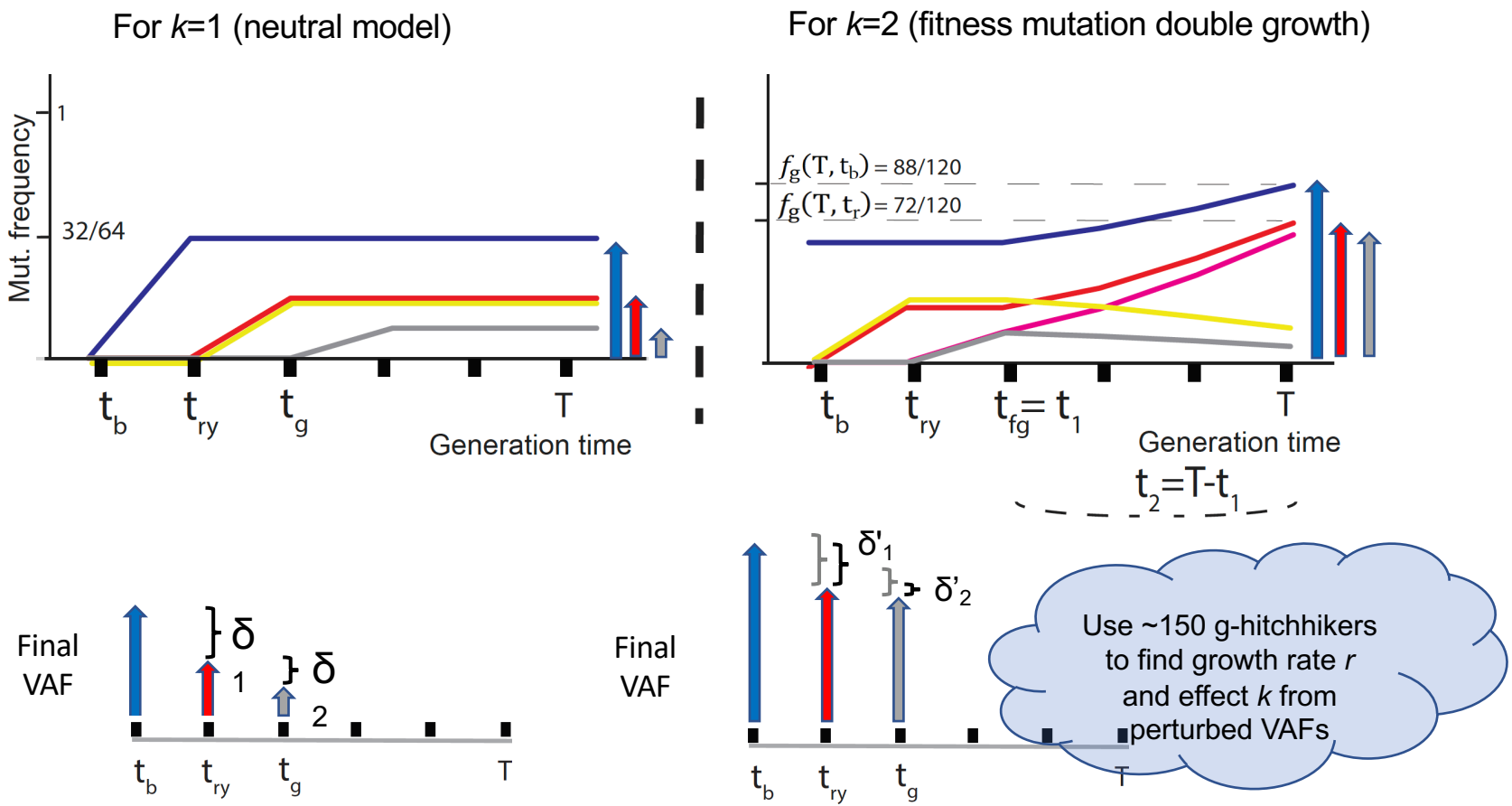
where $f_d(T, t_i)$ is the frequency of the putative driver i occurring at time t_i.

Cell growth without fitness mutation

EXTRA Cell growth WITH fitness mutation

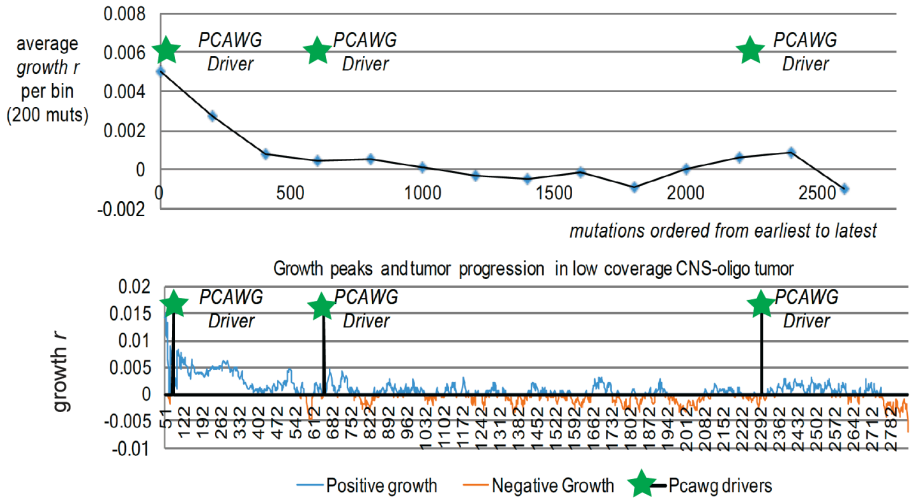
t_g *Generational Time re-optimization parameter*

Modeling scalar effect *k* on growth rate *r* based on VAF perturbations to g-hitchhikers

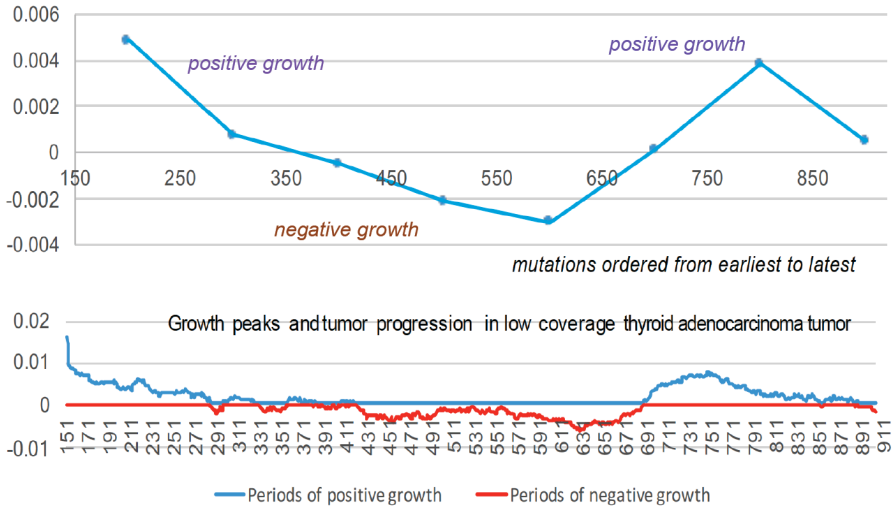


Determining tumor growth in low coverage tumors with known and unknown drivers

Averaged and point growth progression for a low coverage CNS oligo-tumor

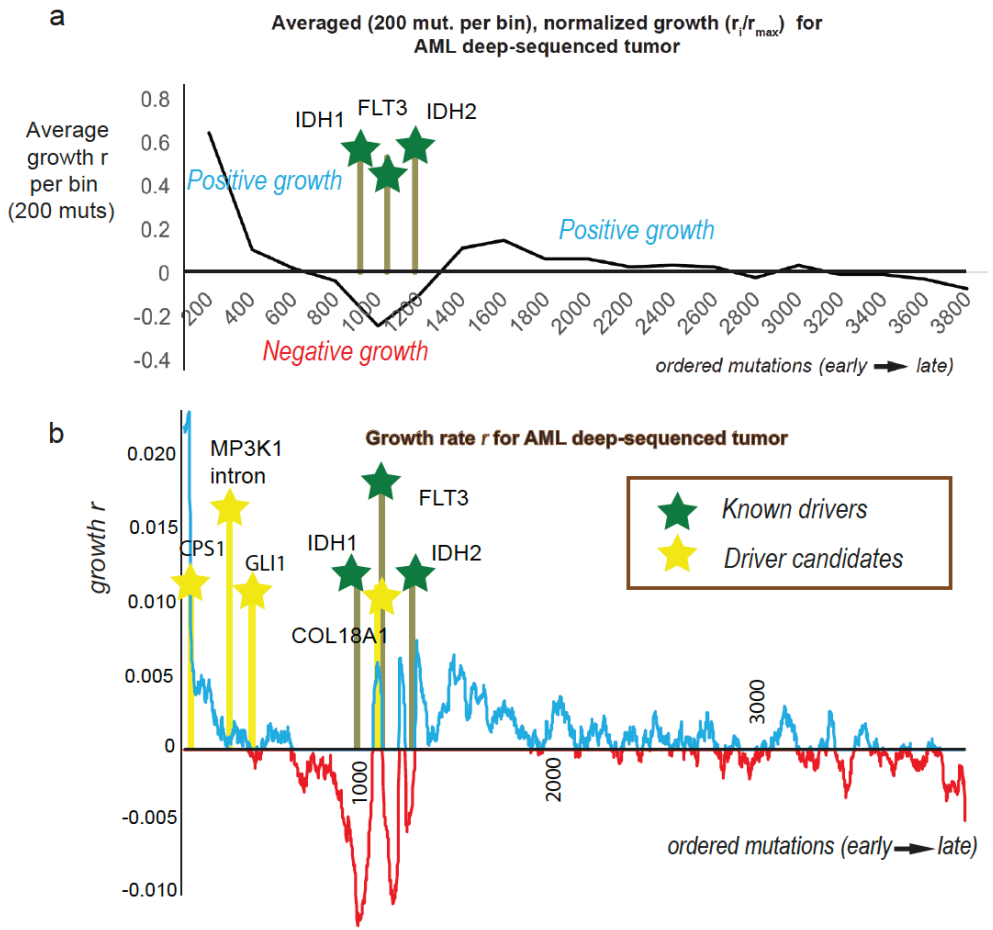


Averaged and point growth progression for a low coverage thyroid adenocarcinoma tumor



[Salichos et al. ('20, in press) Nat. Comm.]

Application to an Ultra-Deep sequenced AML tumor

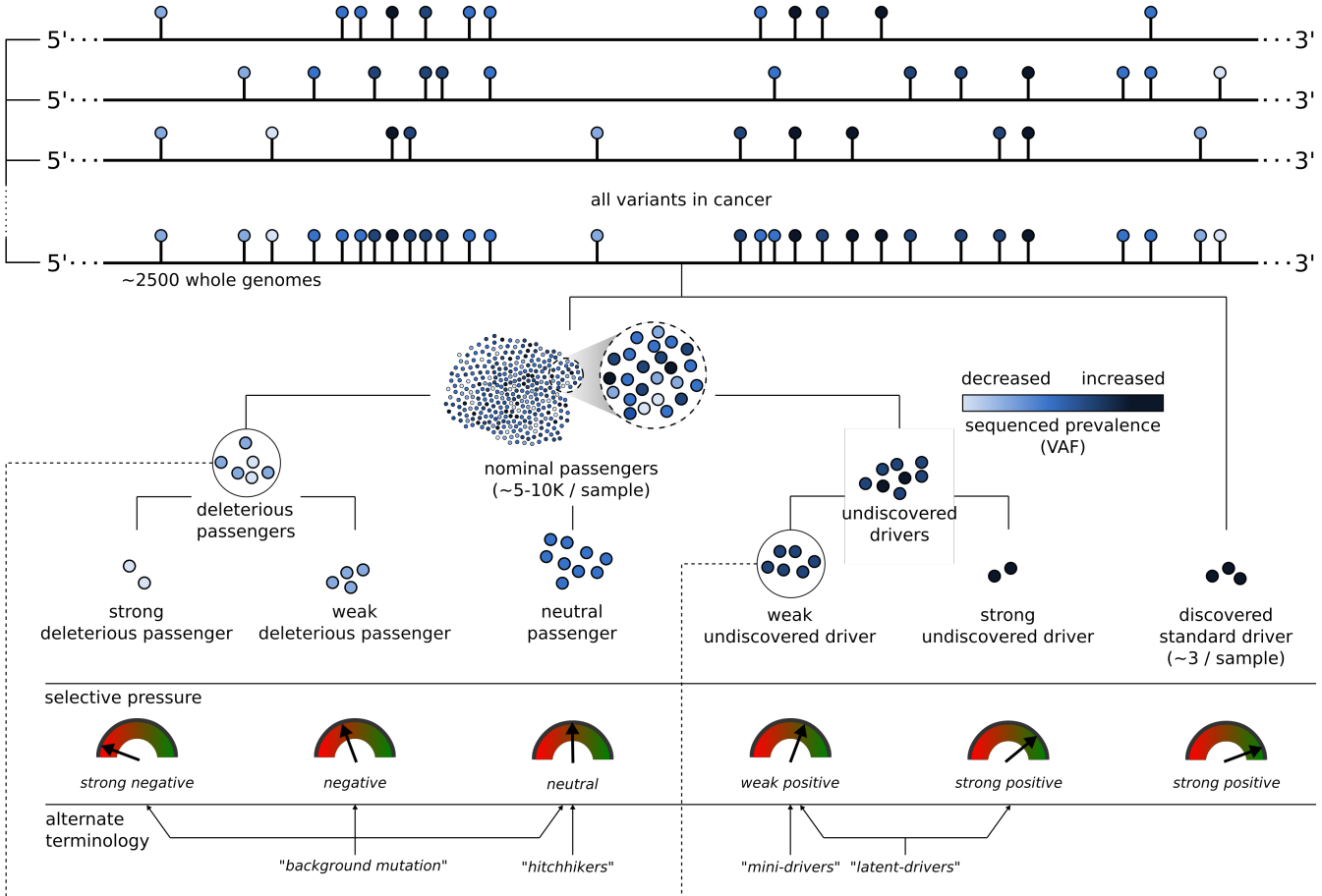


[Salichos et al. ('20, in press) Nat. Comm.]

(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

Conceptual extension of the canonical model of drivers and passengers

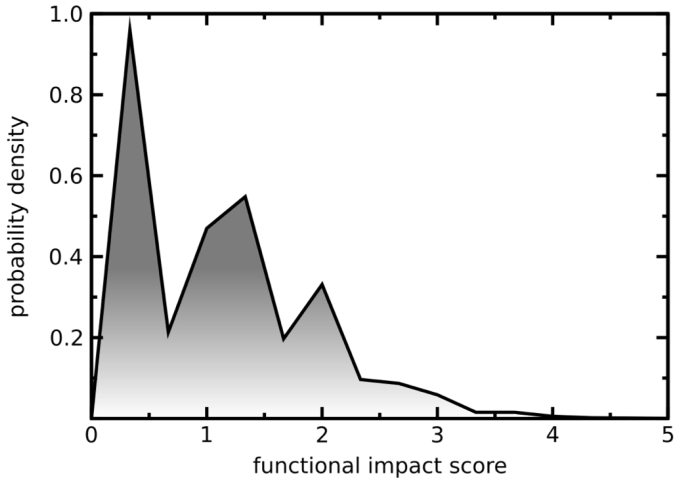


[Kumar, Warrell et al. ('20, in press) *Cell* + biorxiv]

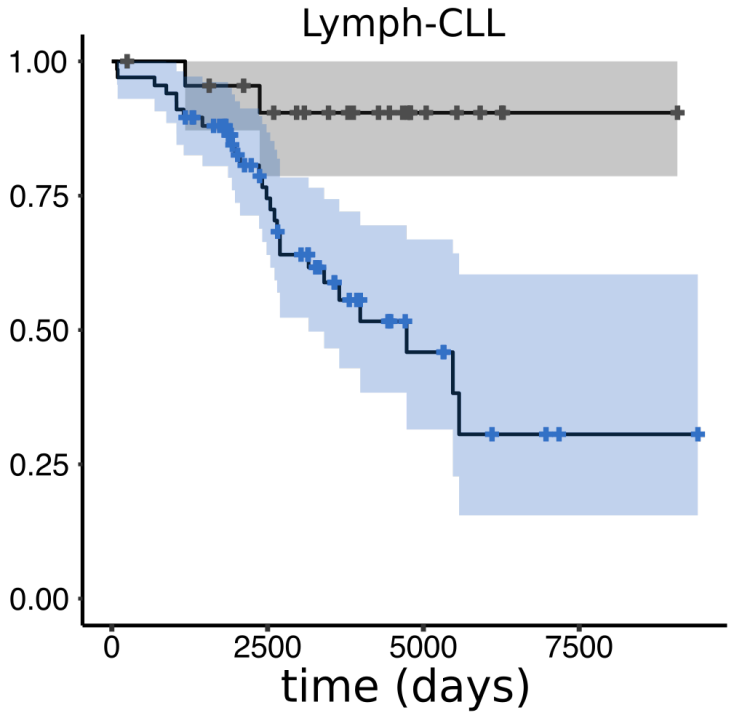
Lectures.GersteinLab.org

50

Overall functional impact distribution of PCAWG mutations

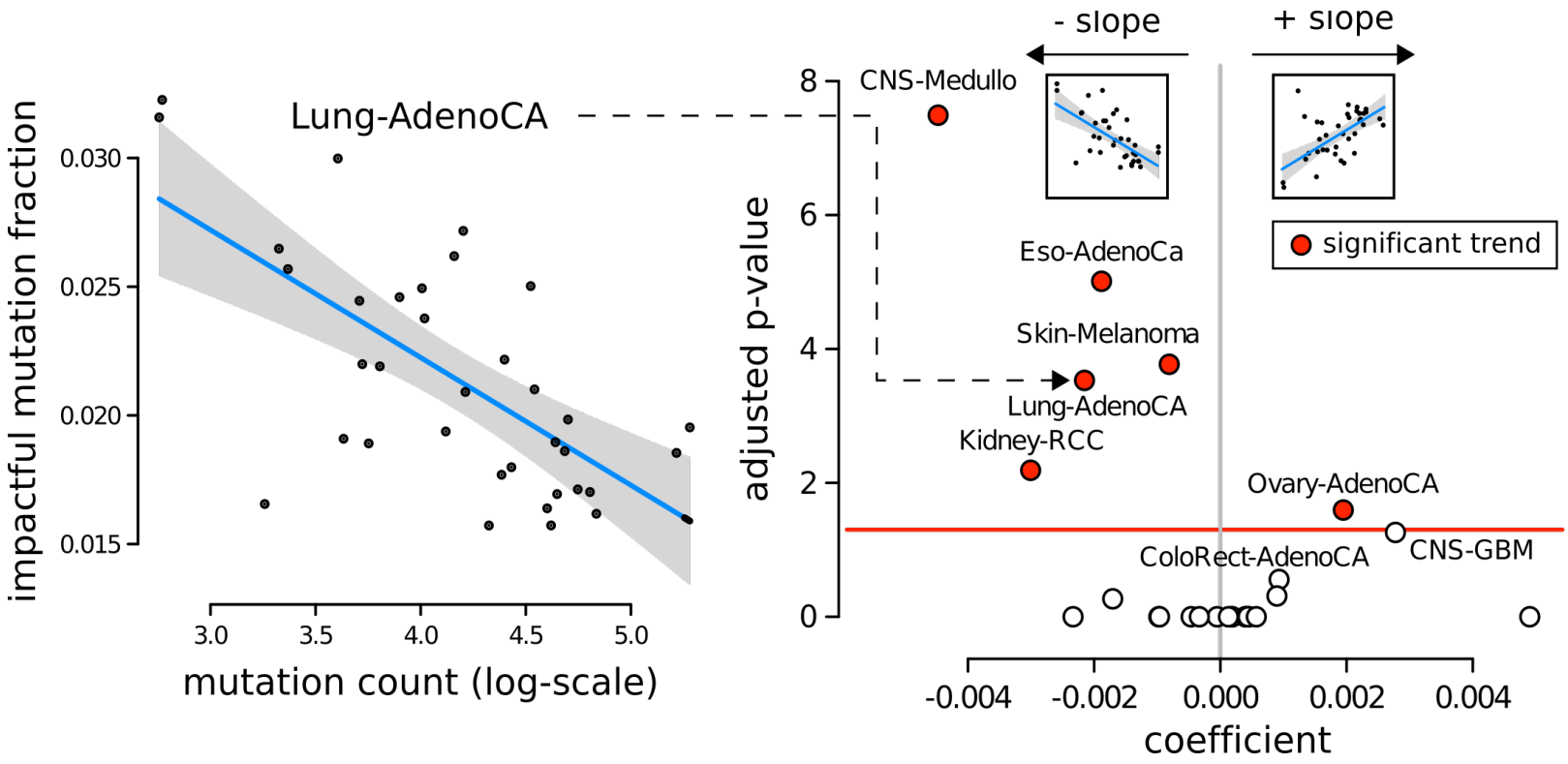


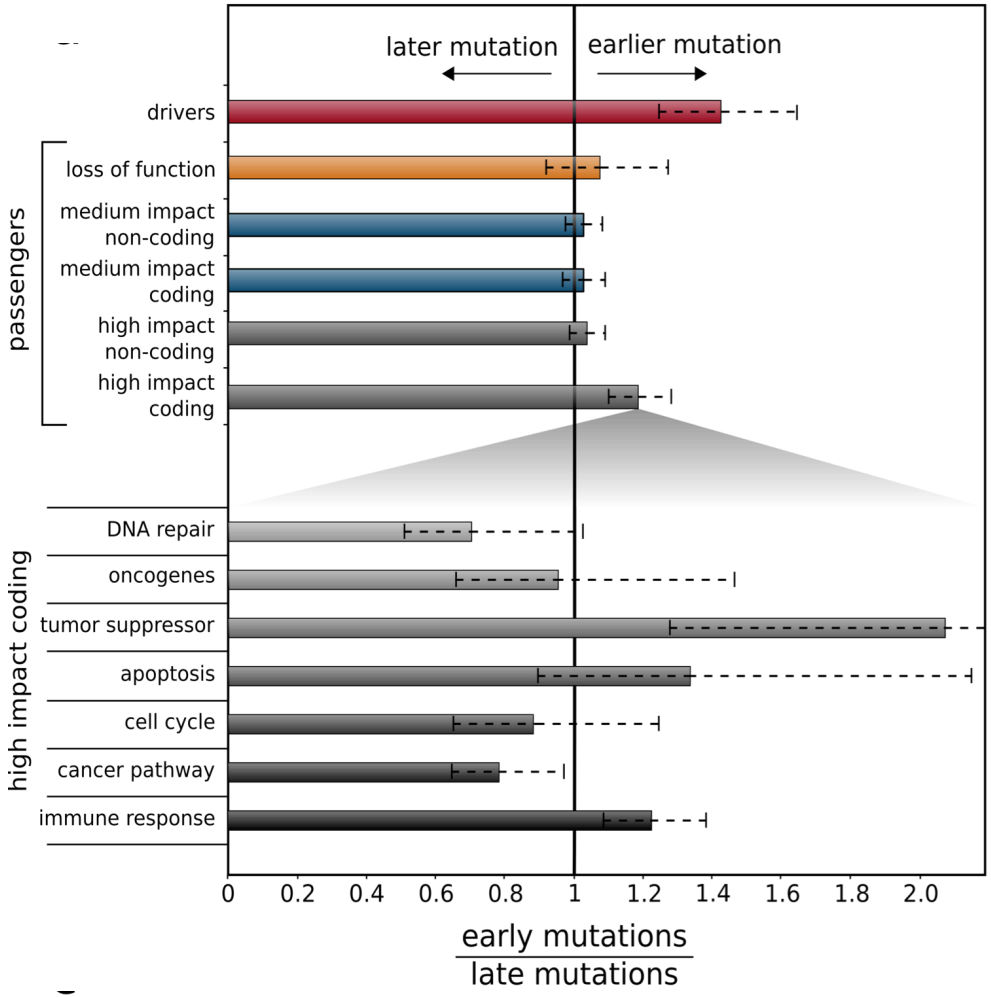
- Funseq molecular functional impact of ~30M variants in >2500 PCAWG samples



Division of PCAWG Lymph-CLL cohort based on average impact of non-driver variants (high v **low**)
[A result of selection?]

In many PCAWG cohorts, the fraction of impactful "passengers" decreases with increase in total mutation burden (A result of selection?)



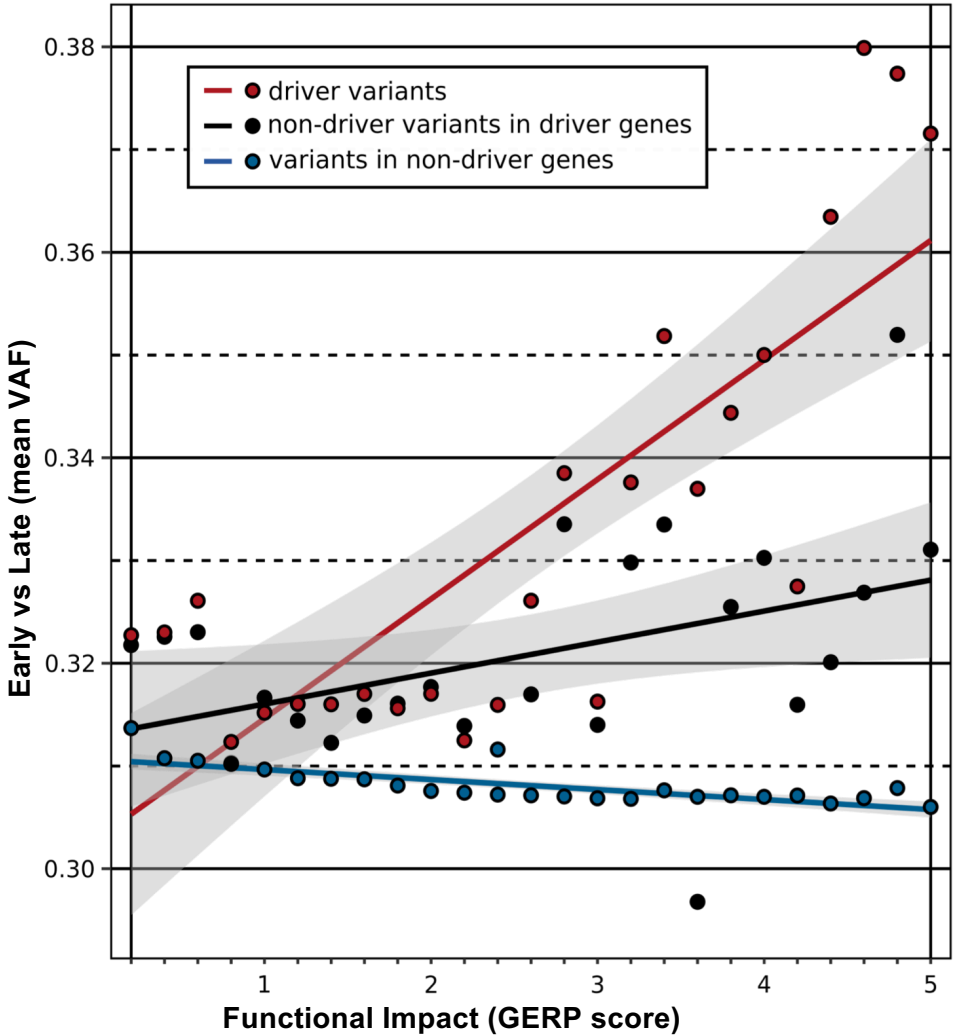


Sub-clonal architecture of mutations in PCAWG

As expected, drivers are enriched in earlier subclones. Overall, no such enrichment among passengers.

High impact passengers are slightly enriched among early subclones (weak drivers?)

Particularly, passengers in tumor suppressor (in contrast to oncogenes, which require specific mutations).



Continuous correlation of functional impact & VAF

Among mutations in driver genes:
higher impact mutation

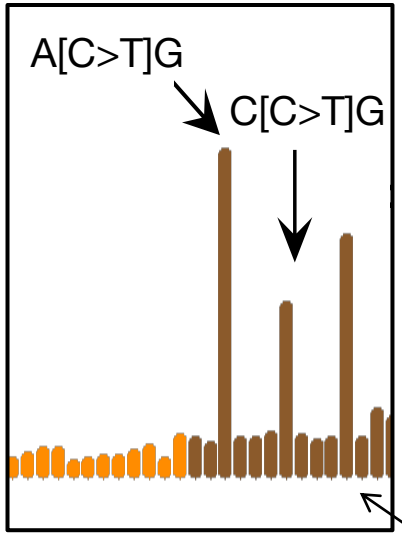
Still true after removing all known driver variants from driver genes.
(Latent drivers?)

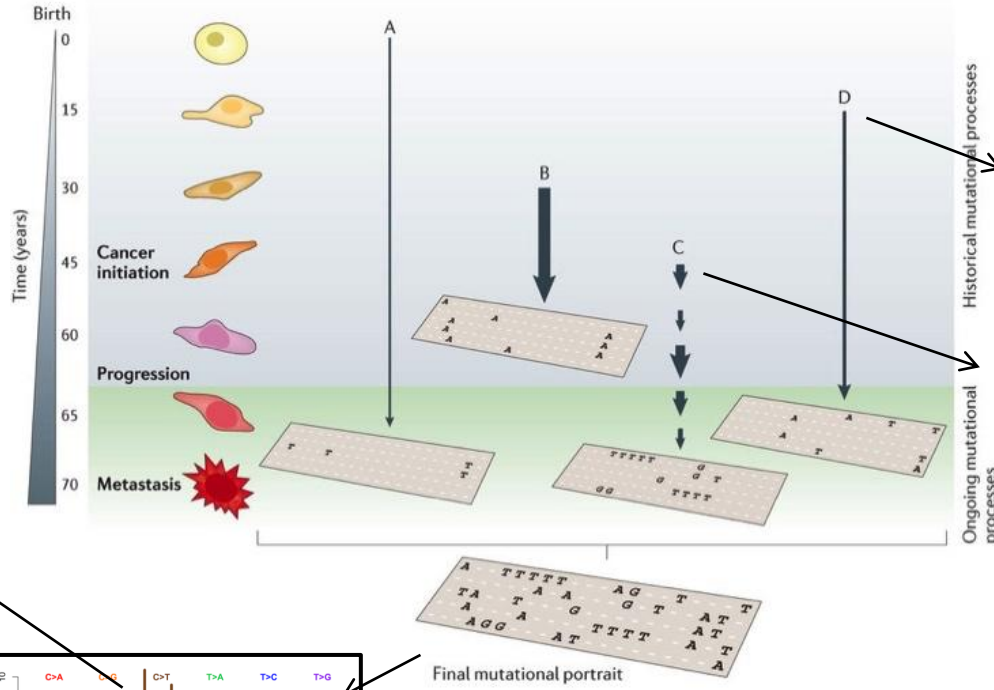
Outside driver genes:
higher impact mutation
(Deleterious passengers?)

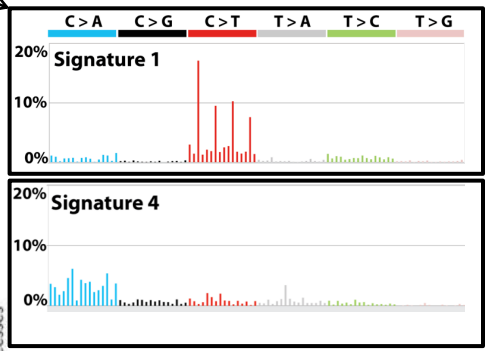
(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

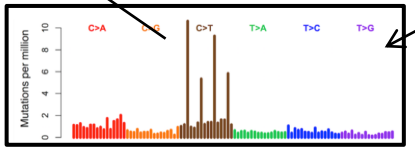
Mutational processes carry context-specific signatures







S: Mutation signature
inferred

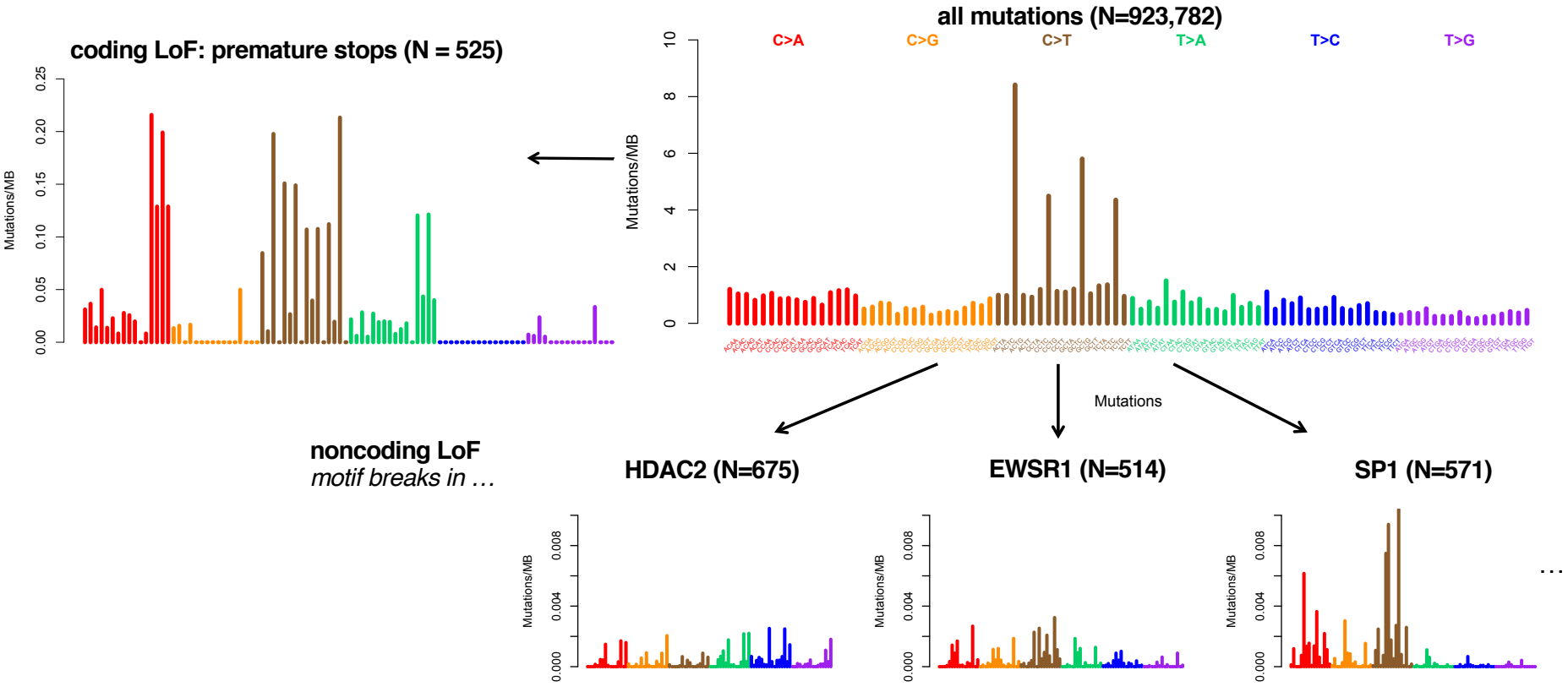


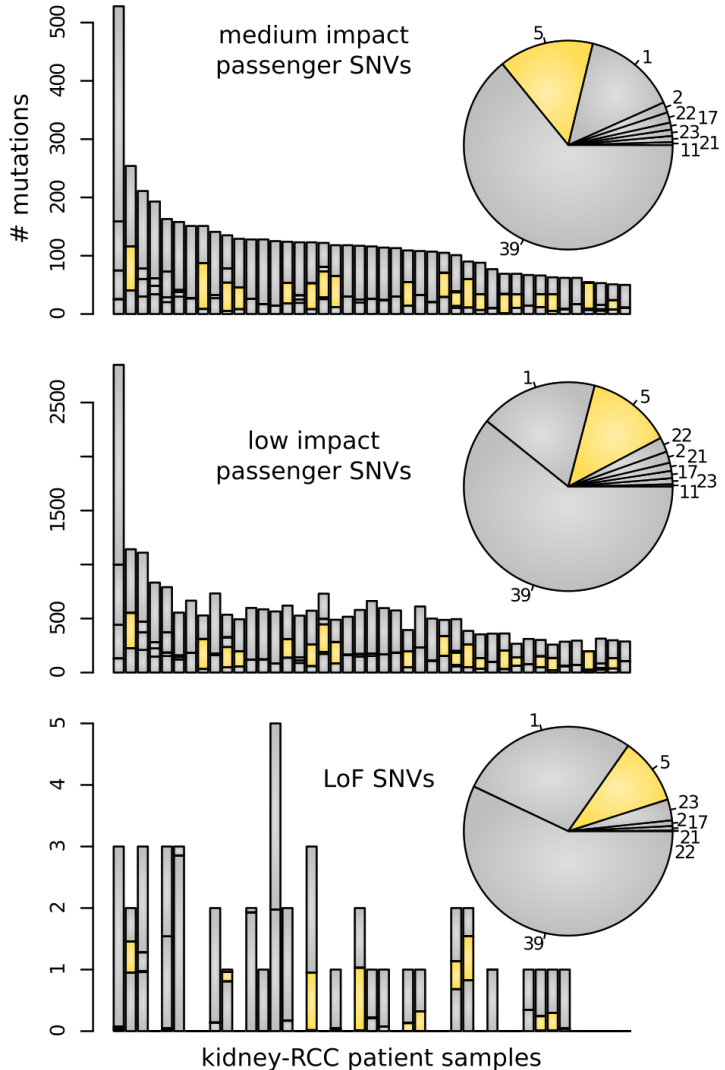
M: Mutation spectrum
observed

$$M = S \times W + \varepsilon$$

[T. Helleday, S. Eshtad & S. Nik-Zainal,
Nature Reviews Genetics ('14), L. Alexandrov et al., Nature ('13)]

Kidney cancer as an example: differential burdening correlates with mutational spectrum



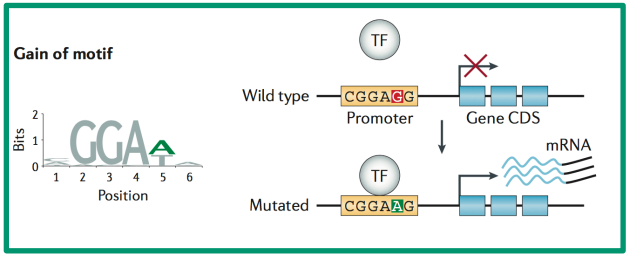


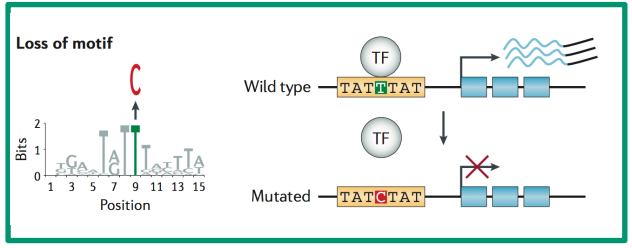
Signatures and molecular impact of passengers: ex of pRCC

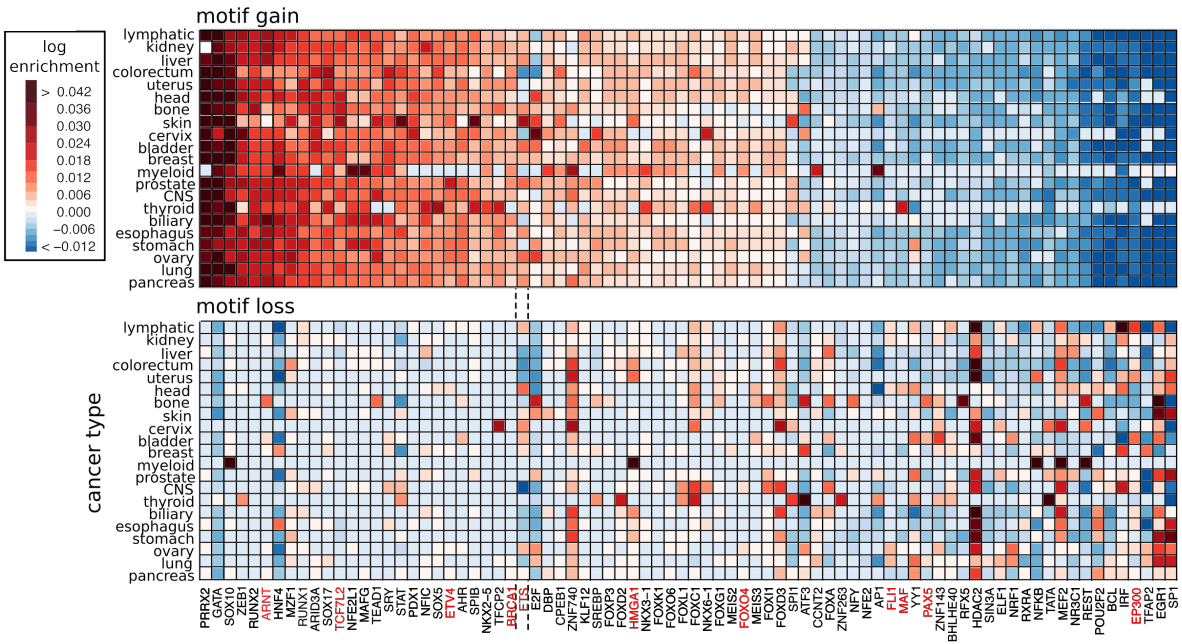
Underlying mutational processes are stochastic but unevenly distributed, which can potentially explain the differential burdening of various genomic elements.

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

Differential Mutational burdening of TF-subnetworks due to SNVs breaking & creating binding sites

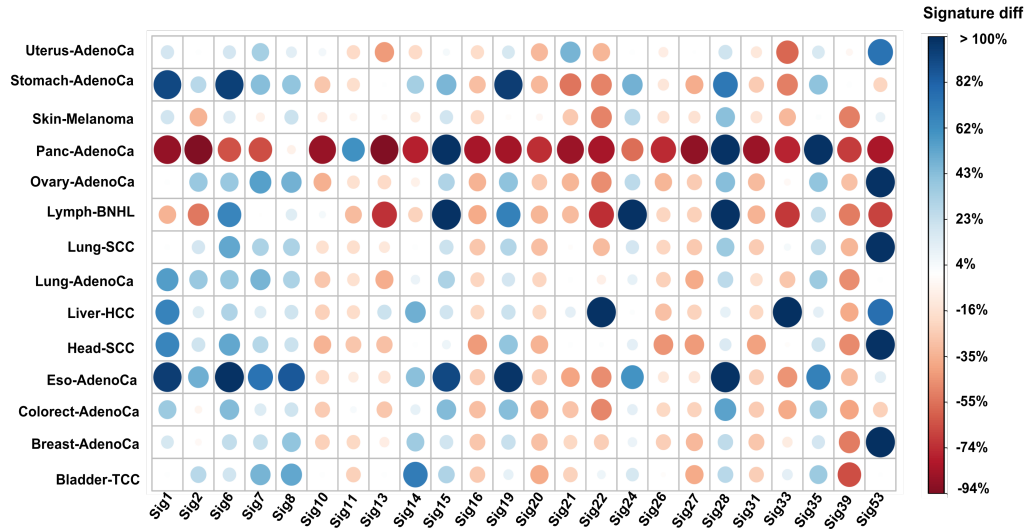






[Kumar, Warrell et al. ('20, in press) *Cell* + biorxiv]

Signature differences between high- and low-impact passengers

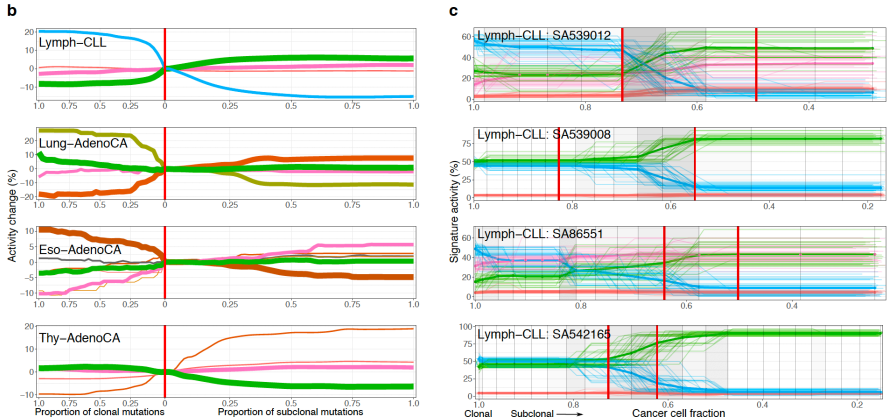


Differing mutational processes could potentially explain the divergence of functional impacts among putative passengers.

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

Mutational processes and fitness

- Mutational process dynamics exhibit common patterns in some cancer types

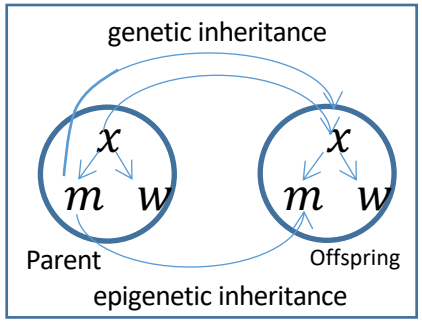


From:

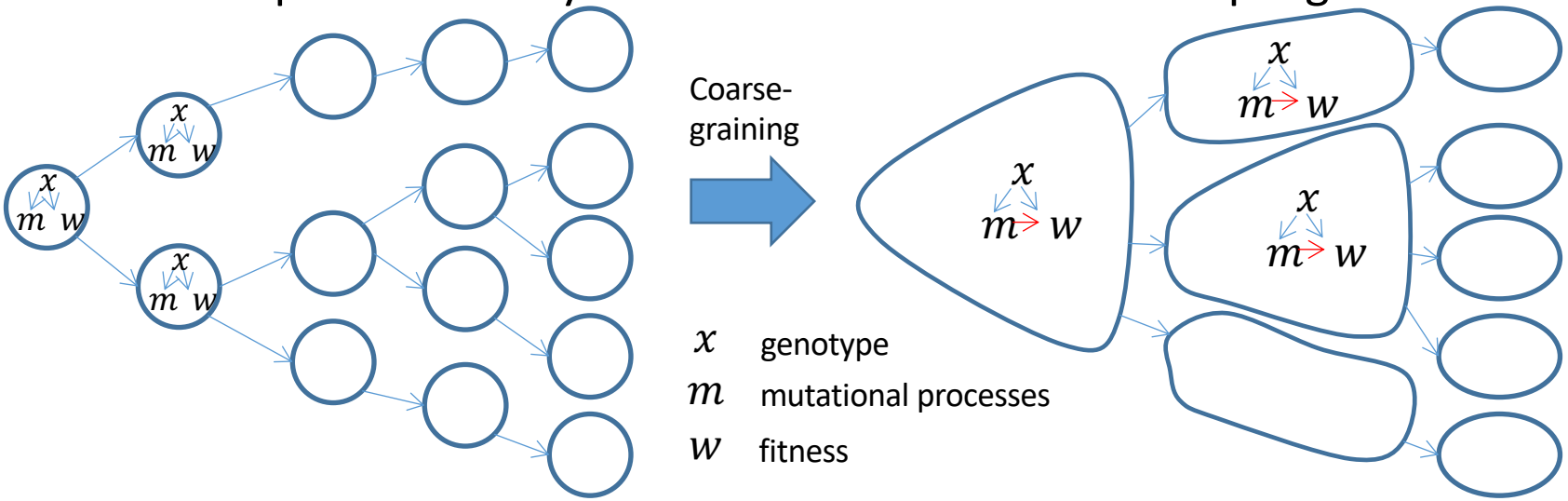
Dentro, S.C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G., Yu, K., Rubanova, Y., Macintyre, G., Vazquez-Garcia, I. and Kleinheinz, K., 2018. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *bioRxiv*.

):

Mutational processes and fitness



- Do mutational processes have effects on fitness?
  - Not necessarily: primarily determine mutations in next generation, rather than number of offspring
- Mutational processes may have fitness effects over multiple generations



- We develop a framework for **cyclic** and **multilevel causation** in evolutionary processes

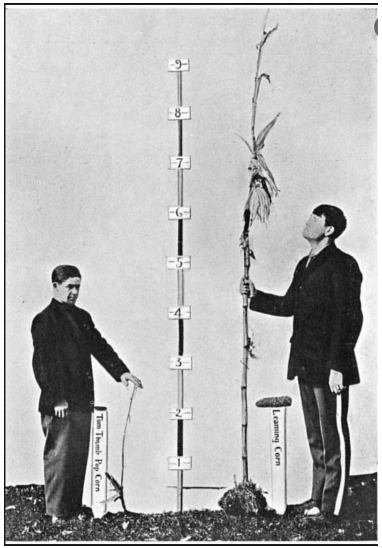
Warrell, J., and Gerstein, M. Cyclic and Multilevel Causality in Evolutionary Processes. *bioRxiv* **(accepted in *Biology and Philosophy*)**

(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

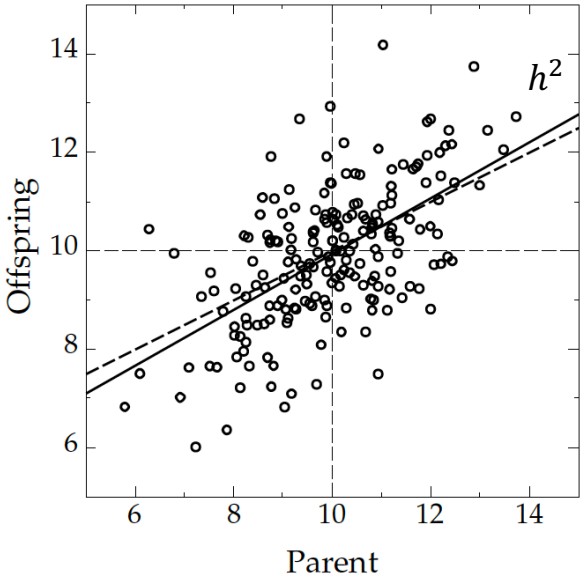
- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

Missing heritability and polygenicity

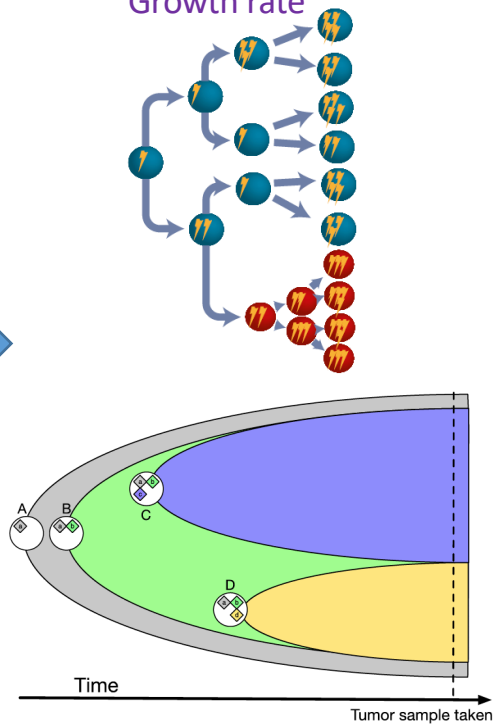
Organismal trait: Height



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

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

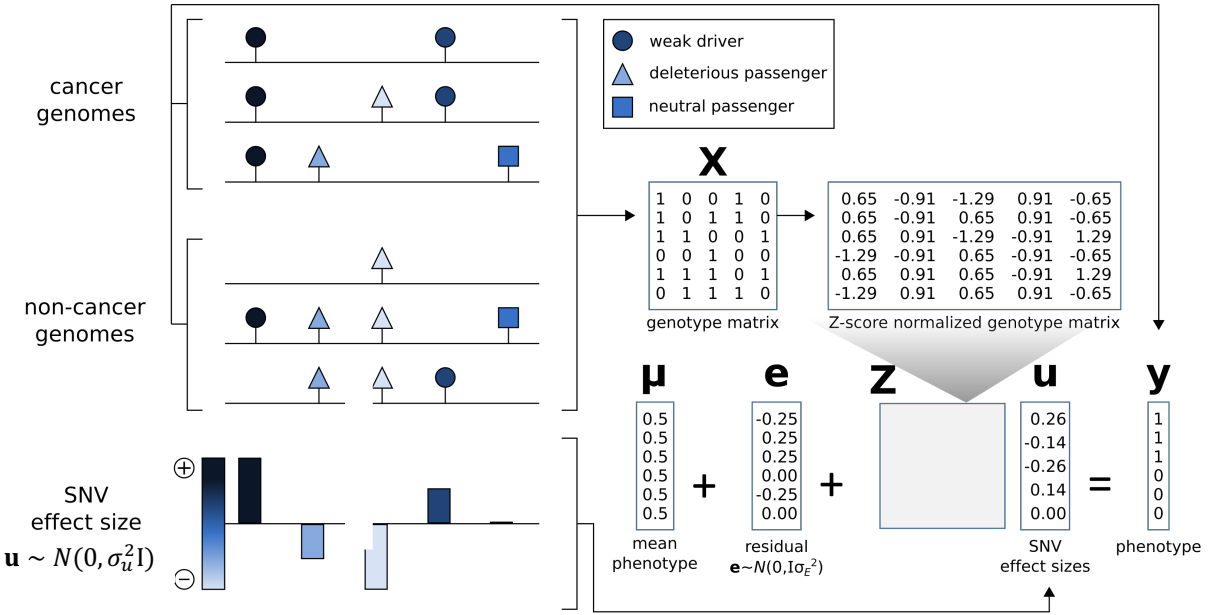
- Model for the effect of an individu SNP on a phenotype

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

- Extension to model the combined effects of multiple SNPs

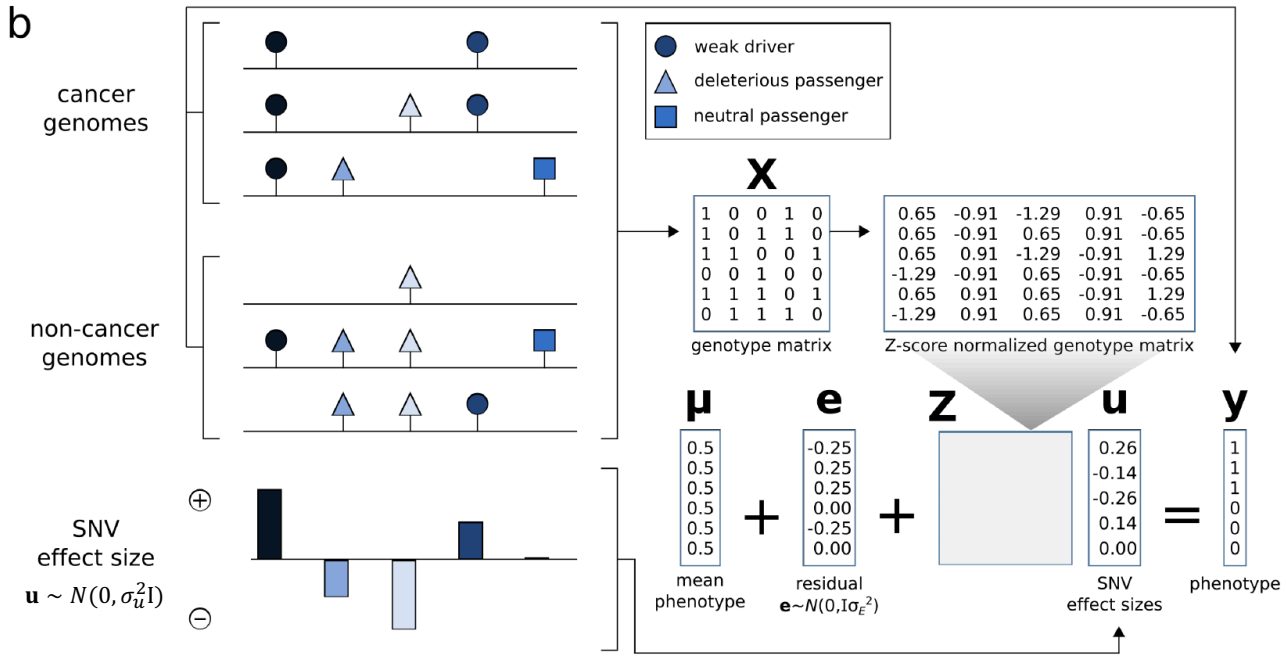
$$y_j = \mu + g_j + e_j \text{ and } g_j = \sum_{i=1}^{m} z_{ij}u_i$$

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



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

Using additive effects to compare different categories of variants



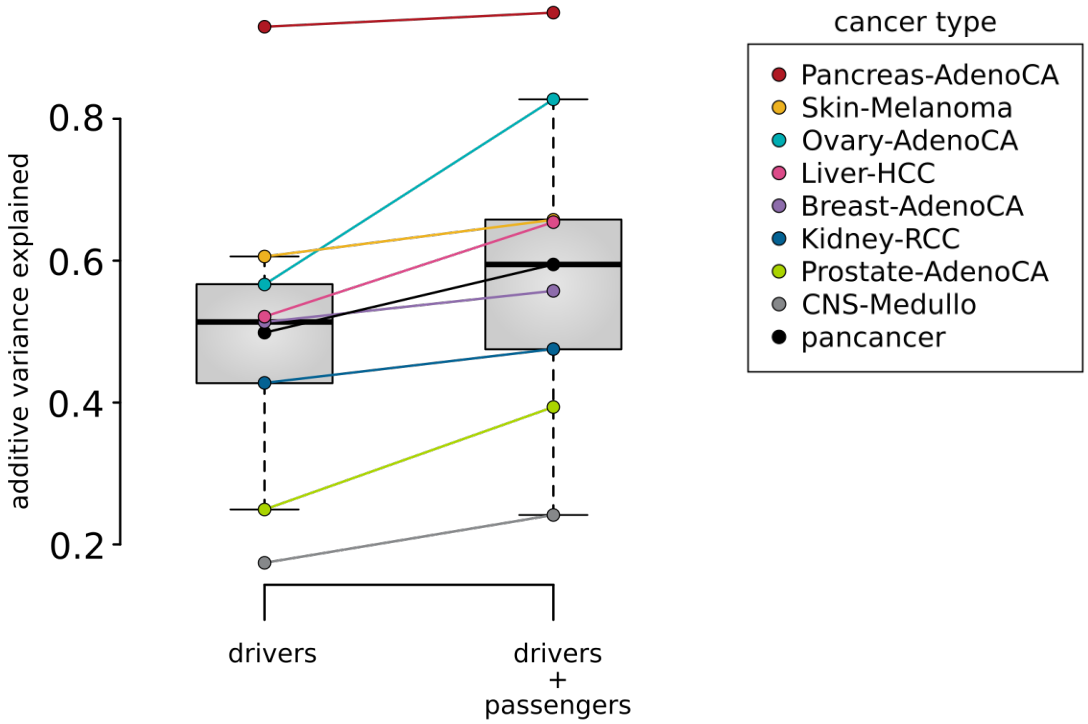
Model: $y_j = \mu + z_j^{\mathrm{drv}} u_1 + \sum_{k \in \{2,3,4\},} z_{ijk} u_{ik} + e_j$

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]

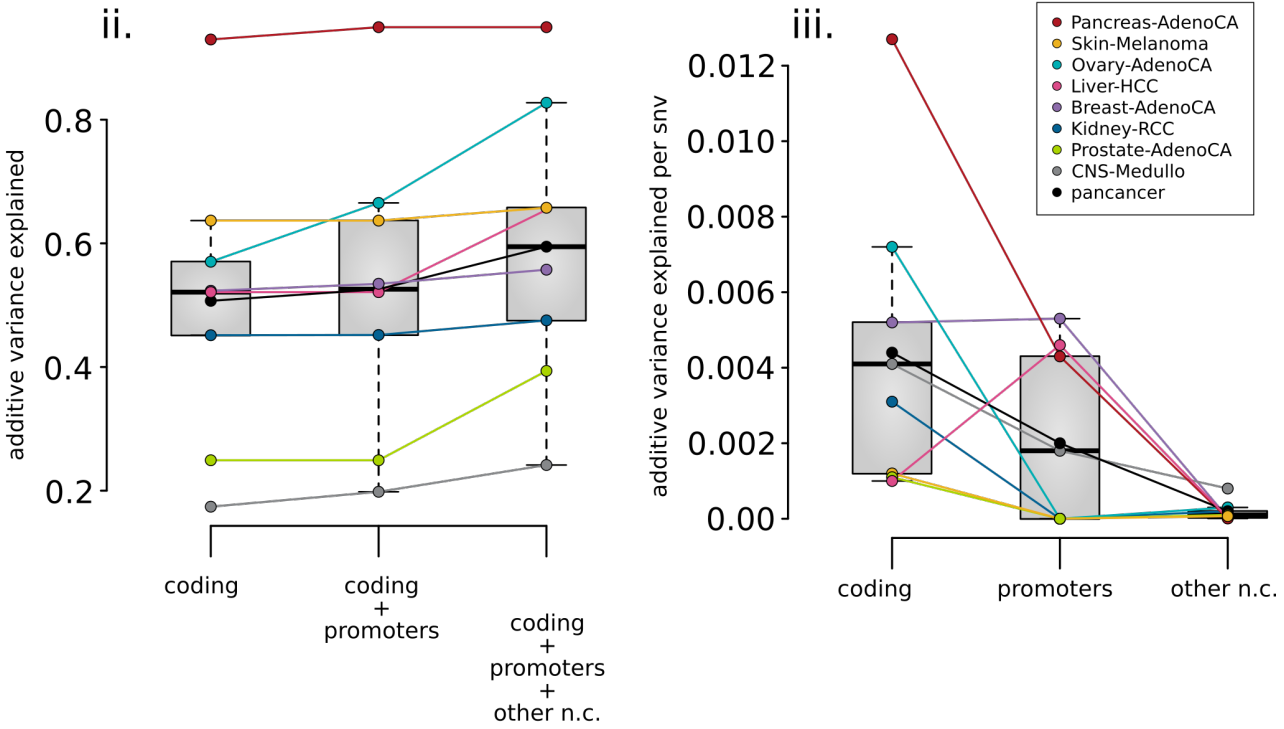
Overall additive variance increase for multiple cancer cohorts in PCAWG with the inclusion of passengers



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]

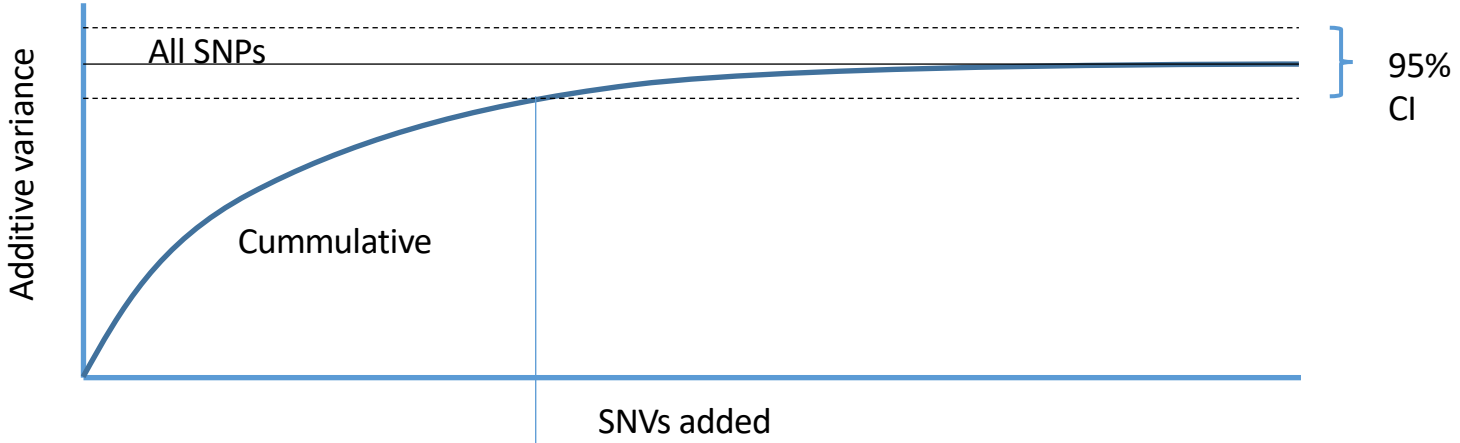
Element level additive variance for multiple cancer cohorts in PCAWG, comparing coding & non-coding



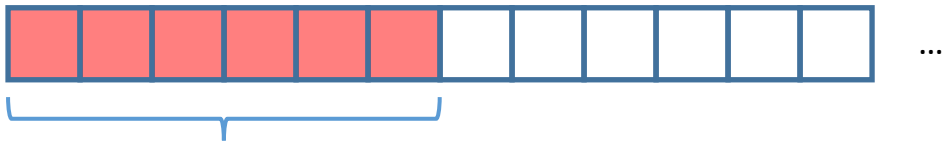
In addition to coding mutations, promoter & other non-coding mutations contributed significant amounts of extra variance (~2% & 7%) .

[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]

(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

(Topics in) Cancer Genomics: Annotating Non-coding Variants, Measuring Regulatory Network Rewiring, Building Background Mutation Models, Analyzing Tumor Evolution & Evaluating the Overall Impact of Passenger Mutations

- **Intro**
  - PMI & Variant Prioritization; driver-passenger model
  - Data source: PCAWG comprehensive WGS on >2.5K + focus on 35 pRCC WGS
- **ENCODEC Annotation**
  - ENCODE cancer resource, with TF & RBP networks
  - Cell-space view of TN pairs
  - FunSeq variant impact measurement integrates conservation & network centrality
- **Network Rewiring**
  - Highlights regulators that change targets greatly
  - LDA approach (from text-mining) finds those that greatly change their gene communities
- **BMR: LARVA/MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Tumor Evolution: Classification + Driver identification**
  - Intro: Mutational timing & tree topology classifies pRCC subtypes
  - Identifying drivers from perturbations in VAF spectra from a single tumor (using many hitchhiking mutations to gain statistical support)
- **Overall Impact of Putative Passengers**
  - Not just high & low impact dichotomy
  - How the fraction of high-impact SNVs scales & relates to survival
  - Differences betw. Impact of early & late passenger mutations (eg in TSGs & oncogenes)
- **Differential Impact of Signatures**
  - Diff. burdening of TF sub-networks naturally results from mutational spectra & signatures differentially affecting binding motifs.
  - High & low impact mutations assoc. w/ diff. signatures
  - How it all relates to selection?
- **Additive Effects Model**
  - To quantify aggregated effect of passengers. Demonstratable effect, particularly for non-coding ones, in addition to known drivers.
  - Recasting as a predictive model to est. number of weak drivers

ENCODEC.gersteinlab.org

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

Evotum

L **Salichos** , W Meyerson , J Warrell

FunSeq.gersteinlab.org

Y **Fu**, E **Khurana**, Z Liu, S Lou, J Bedford, X Mu, K Yip
E Khurana, Y Fu, H Kang, X Mu… M Rubin, C Tyler-Smith

{LARVA,MOAT}.gersteinlab.org

Lochovsky, J **Zhang**, Y Fu, E Khurana

PanCancer.info

S **Kumar**, J Warrell, W Meyerson, P McGillivary,
L Salichos, S Li, A Fundichely, E Khurana, C Chan, M Nielsen,
C Herrman, A Harmanci, L Lochovsky,Y Zhang, X Li, G Getz, J Pedersen,

pRCC

S **Li**, B Shuch


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