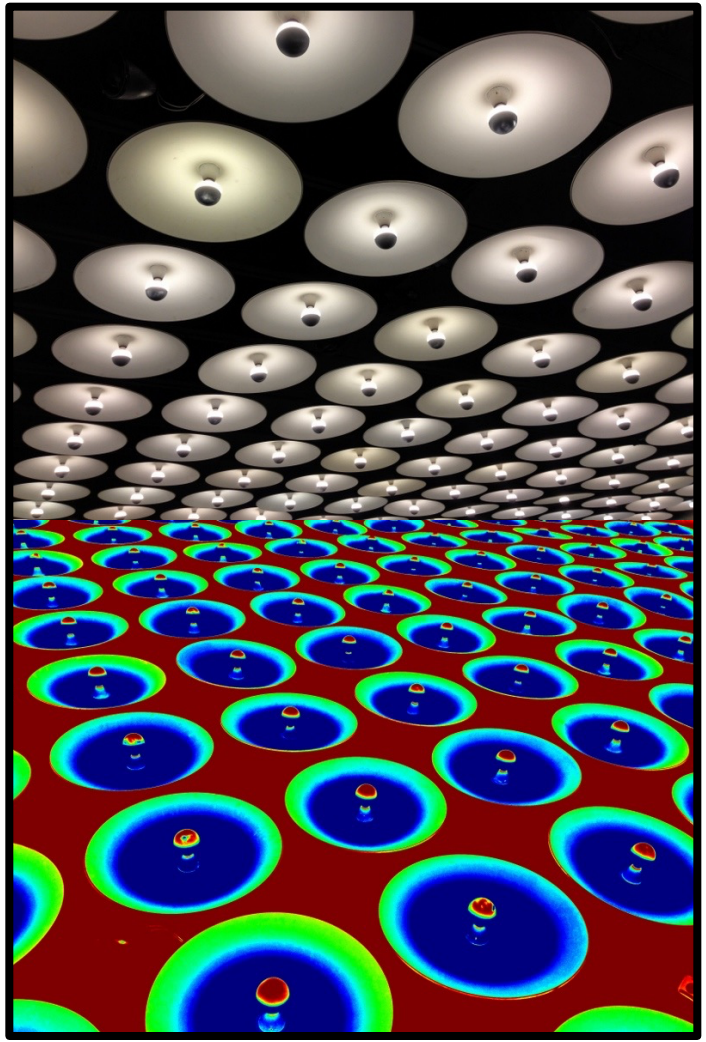


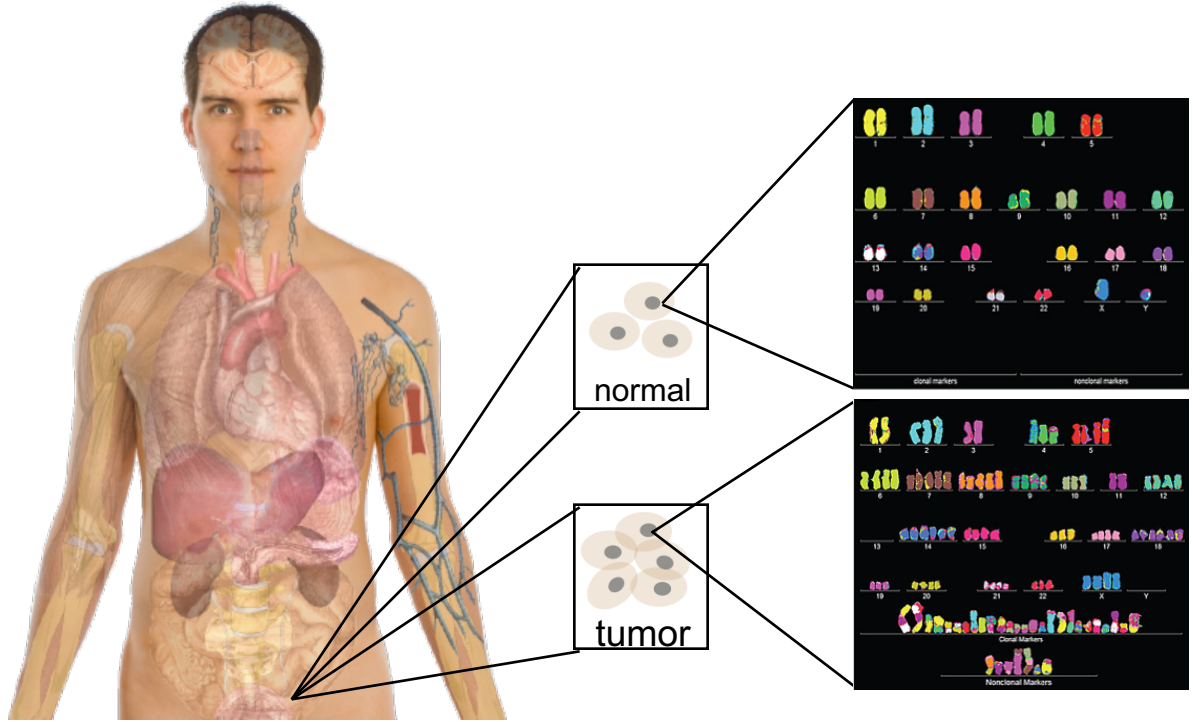
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

Mark Gerstein
Yale

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

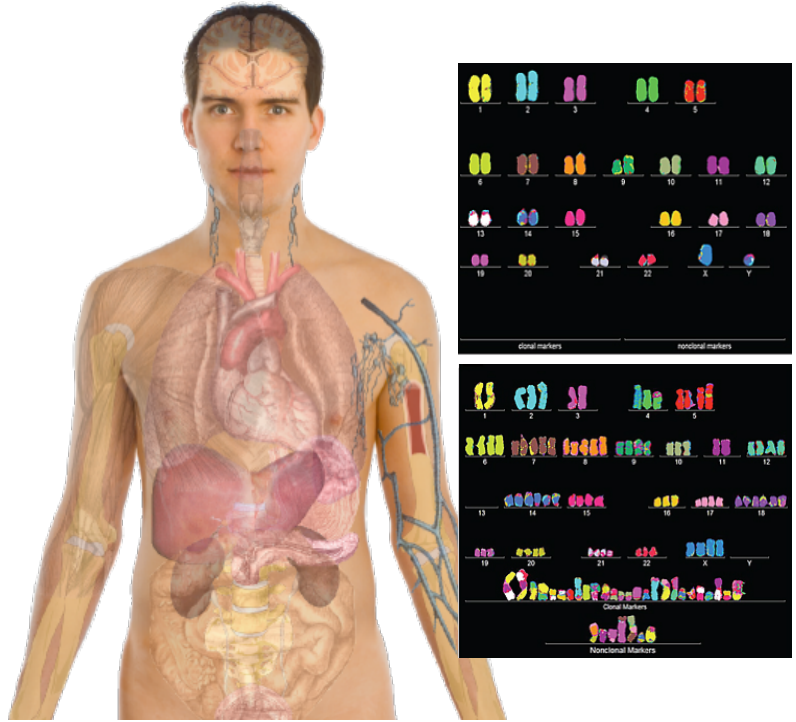
Personal Genomics as a Gateway into Biology

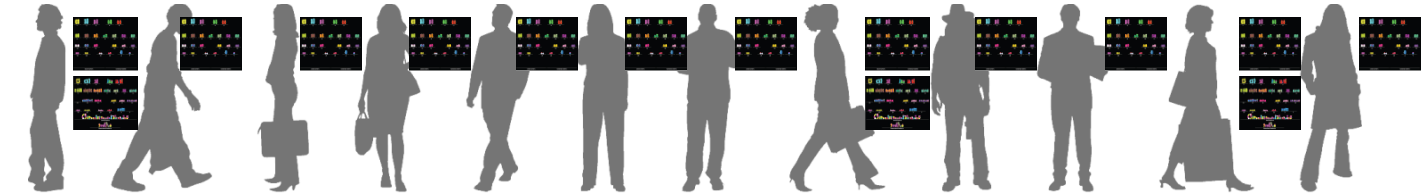
Personal genomes will soon become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.

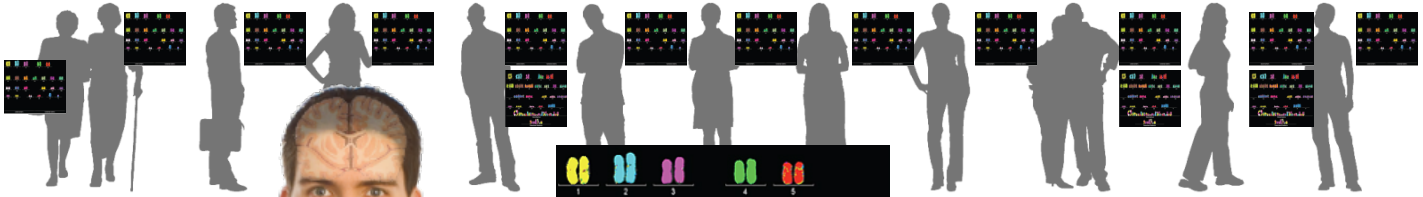


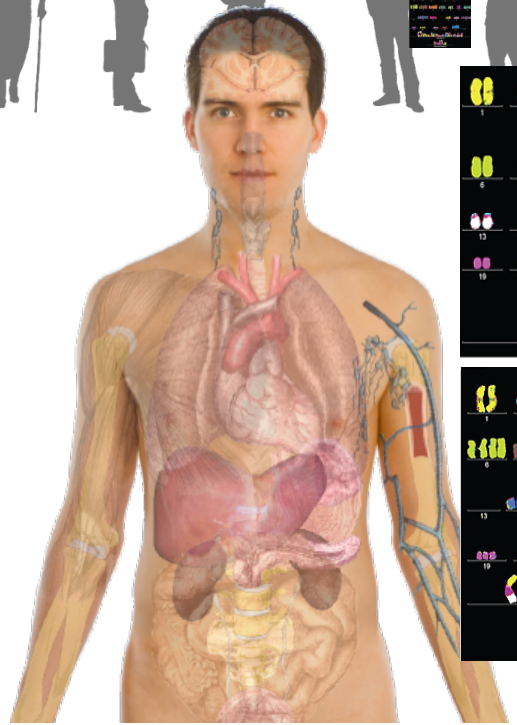
Personal Genomics as a Gateway into Biology

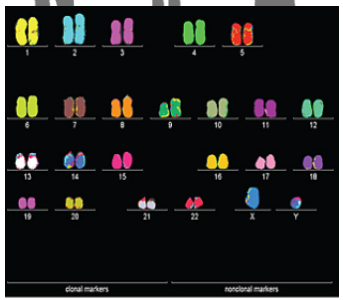
Personal genomes will soon become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.











Keys to genome interpretation

Relating individuals' variants to **DBs**

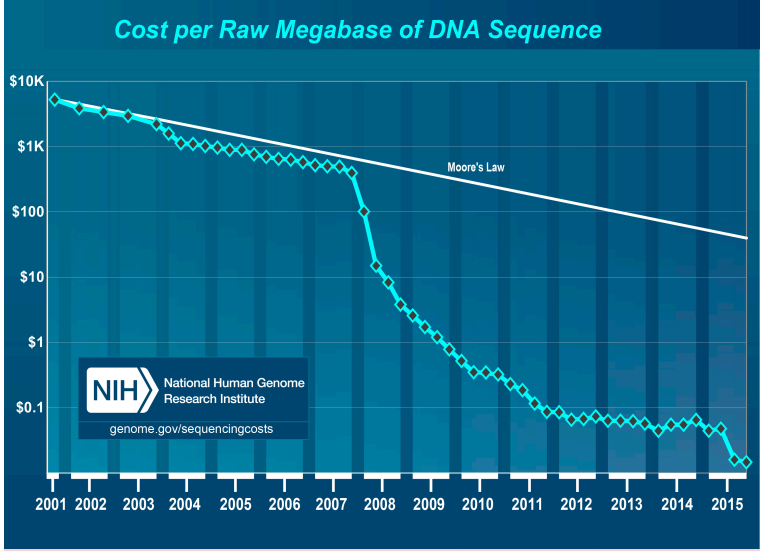
Scaling DBs to the **population**

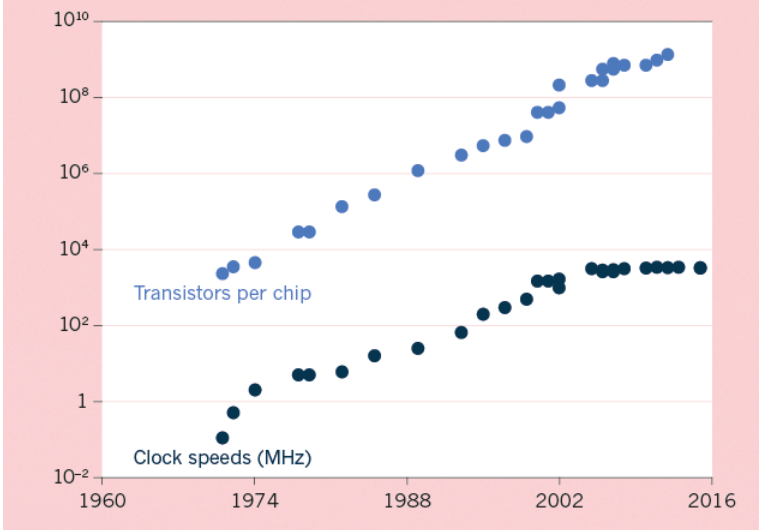
Identifying **key variants** - separating into rare, recurrent, common, &c

The Scaling of Genomic Data Science:

Powered by exponential increases in data & computing

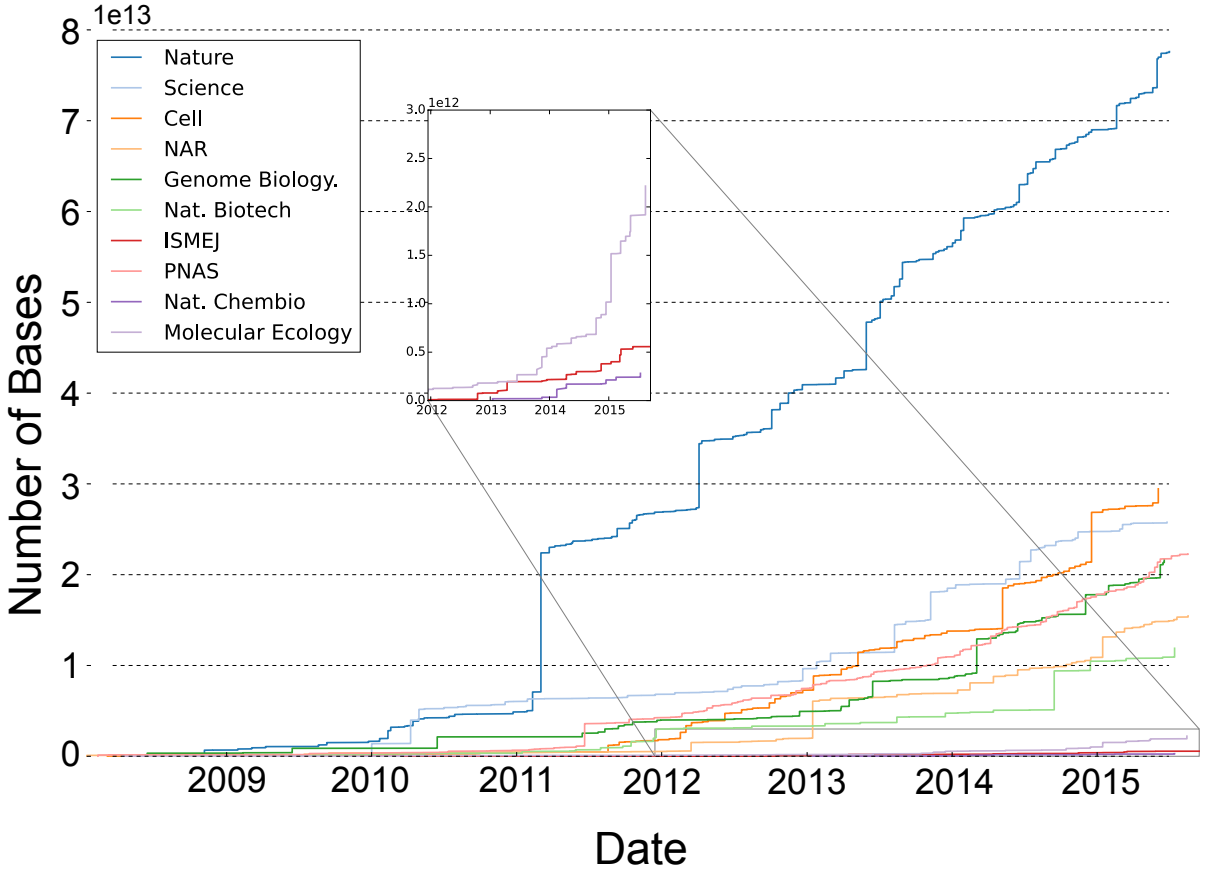
(Moore's Law)





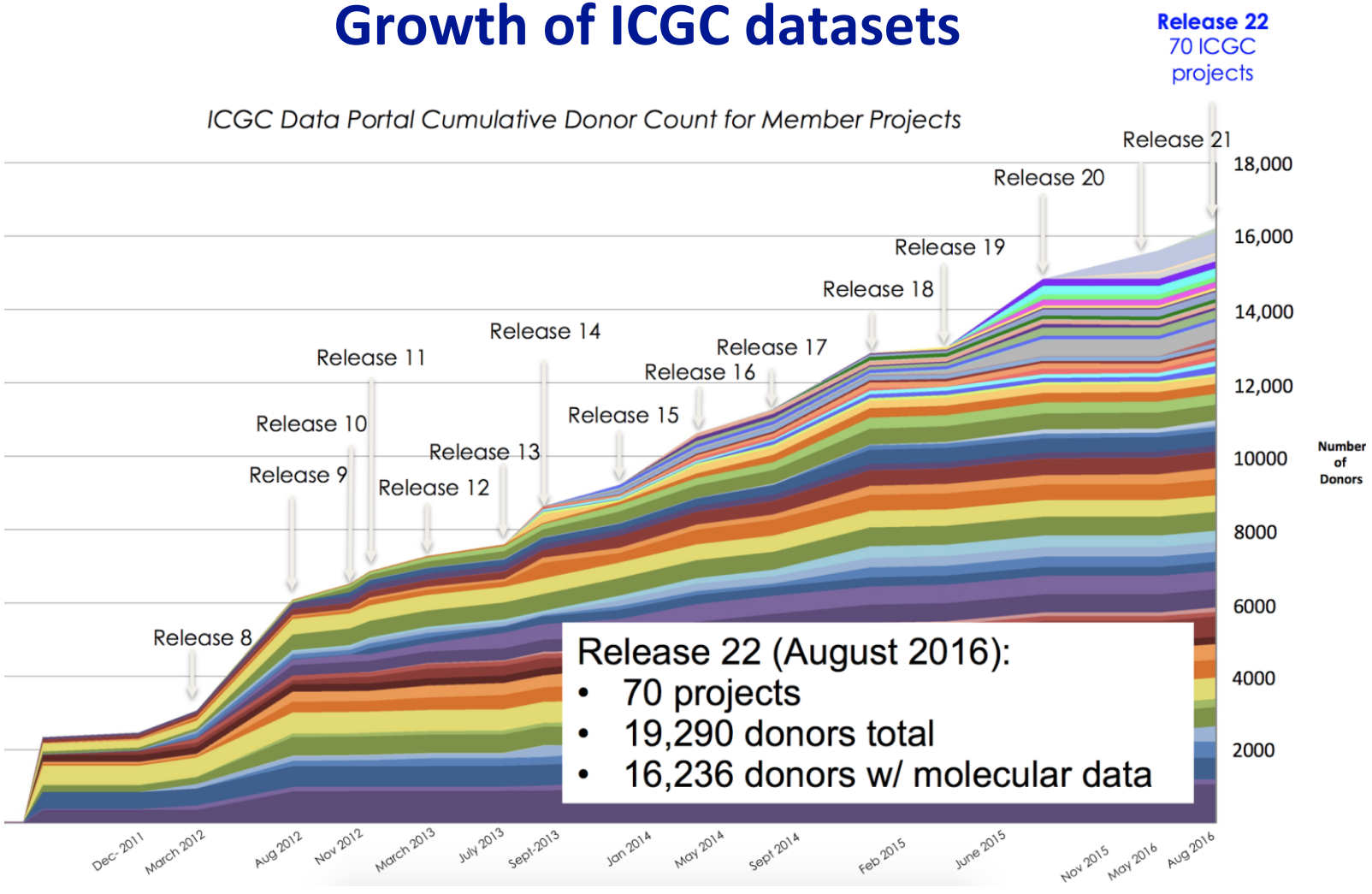
[NHGRI website + Waldrop ('15) Nature]

Exponential Scaling Changes Fields Using Genomic Data



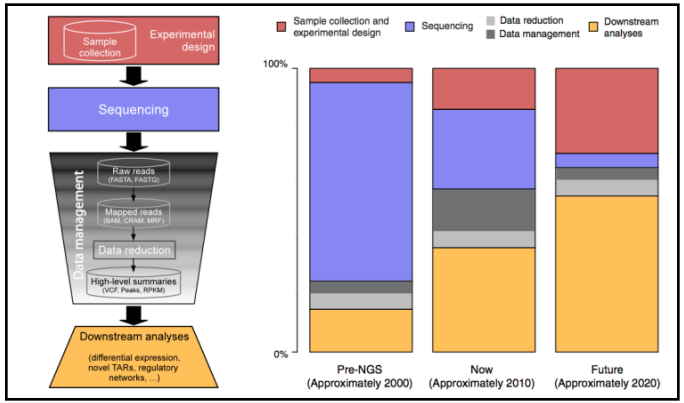
[Muir et al. ('15) GenomeBiol.]

Growth of ICGC datasets



Hardeep Nahal , 12th Scientific ICGC Workshop (Sept 2016)

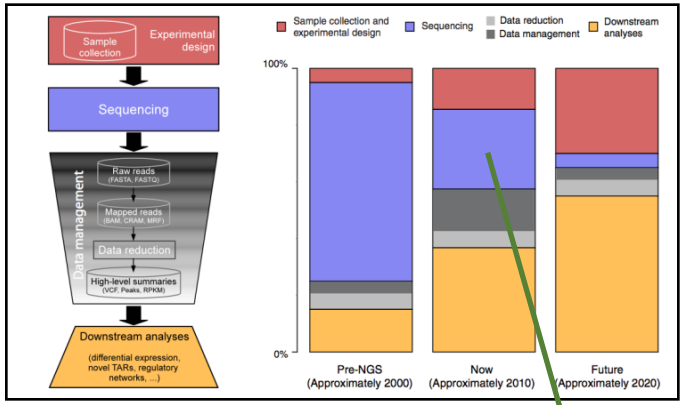
The changing costs of a sequencing pipeline



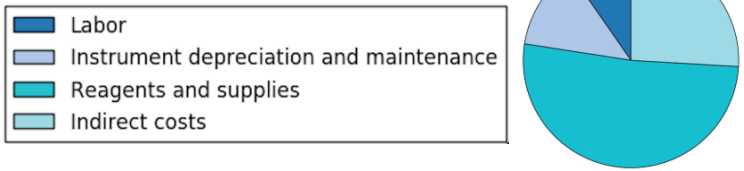
From ‘00 to ~’20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline

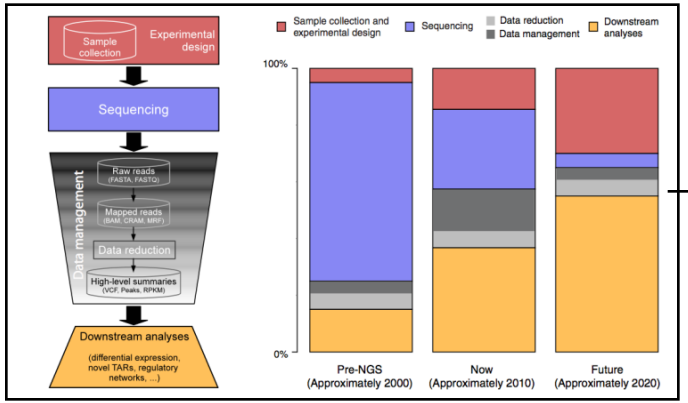


From ‘00 to ~’20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

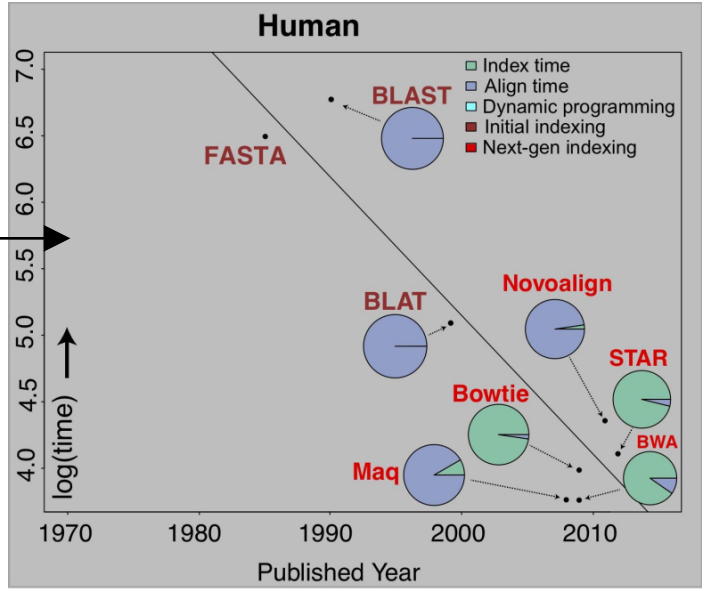


[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline



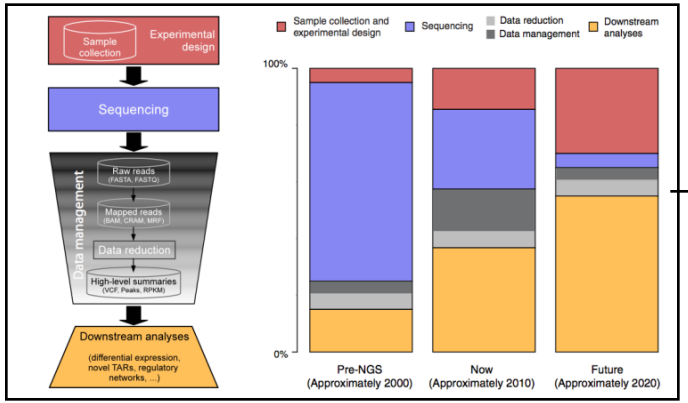
From '00 to ~'20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



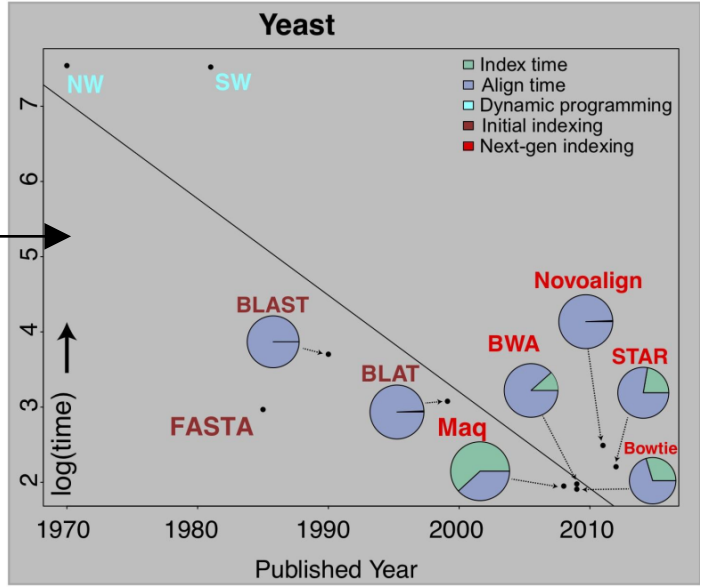
Alignment algorithms scaling to keep pace with data generation

[Sboner et al. ('11), Muir et al. ('15) Genome Biology]

The changing costs of a sequencing pipeline



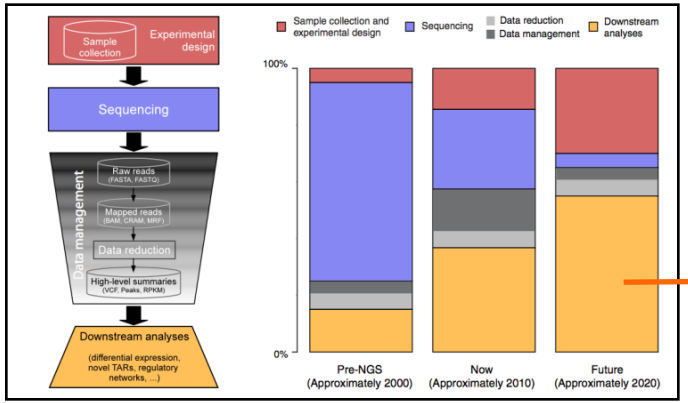
From ‘00 to ~’20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



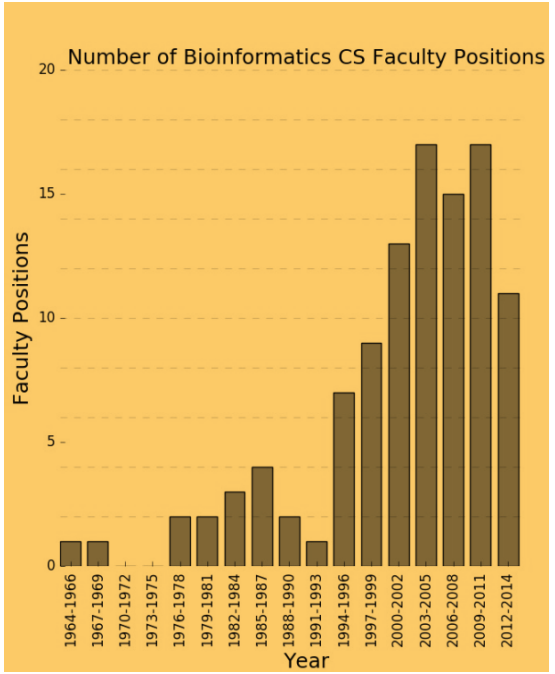
Alignment algorithms scaling to keep pace with data generation

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline

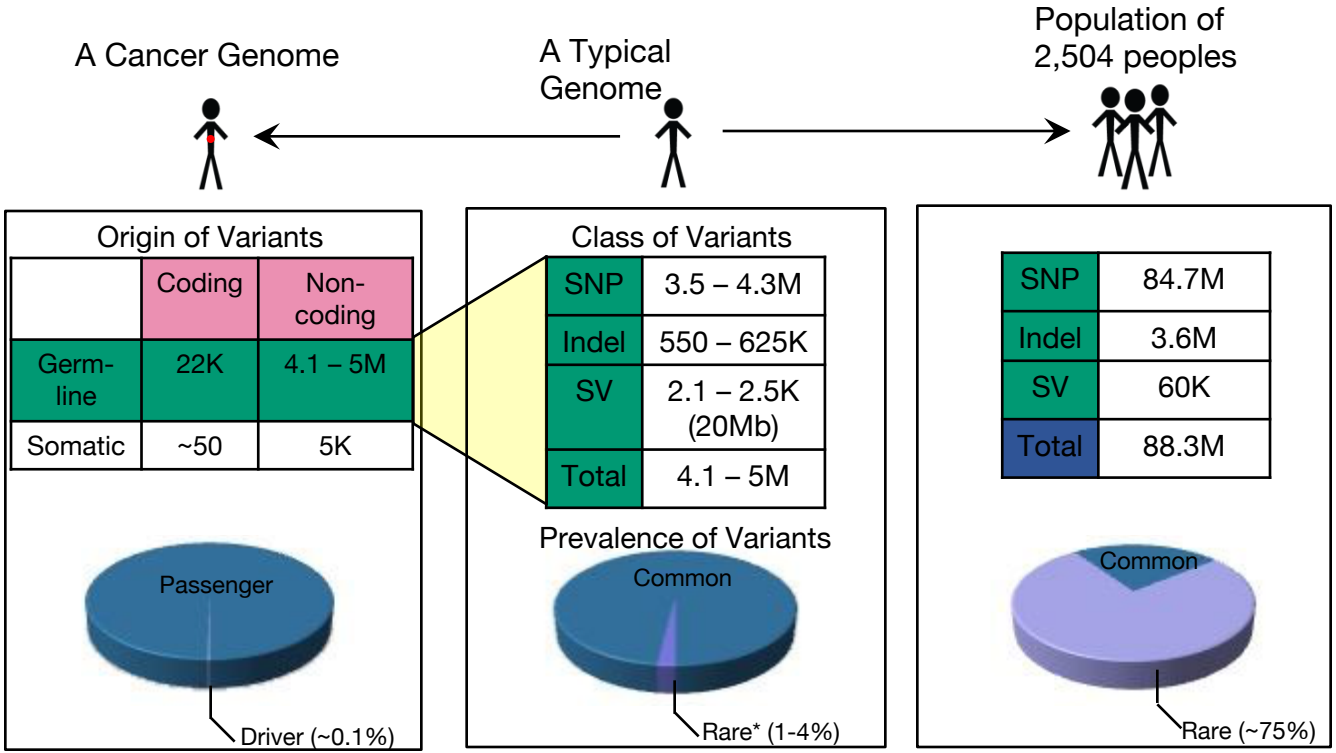


From '00 to ~'20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



[Sboner et al. ('11), Muir et al. ('15) Genome Biology]

Human Genetic Variation



* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Finding Key Variants

Germline

CAN YOU FIND THE PANDA?

- **Common variants**
  - Can be most readily associated with phenotype (ie disease) via GWAS
  - Usually their functional effect is weaker
  - Many are non-coding
  - Issue of LD in identifying the actual causal variant.
- **Rare variants**
  - Associations are usually underpowered due to low frequencies but often have larger functional impact
  - Can be collapsed in the same element to gain statistical power (burden tests).

McCarthy, M. et al. Nat. Rev. Genet. 2008. 9, 356-369, Zuk, O. et al. PNSA. 2014. Vol. 11, no. 4, MacArthur DG et al. Nature 2014. 508:469-476

Finding Key Variants

Somatic

CAN YOU FIND THE PANDA?

- **Overall**
  - Often these can be thought of as very rare variants
- **Drivers**
  - Driver mutation is a mutation that directly or indirectly confers a selective growth advantage to the cell in which it occurs.
  - A typical tumor contains 2-8 drivers; the remaining mutations are passengers.
- **Passengers**
  - Conceptually, a passenger mutation has no direct or indirect effect on the selective growth advantage of the cell in which it occurred.

Vogelstein B. Science 2013. 339(6127):1546-1558

Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

Variant Annotation Tool (VAT), developed for 1000G FIG

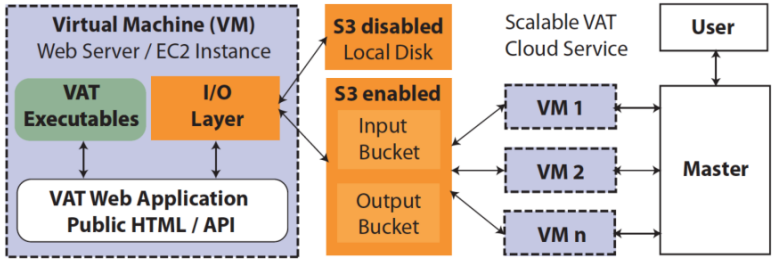
VCF Input

Output:

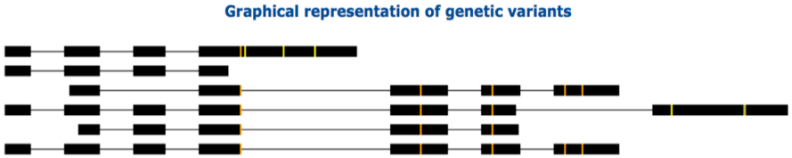
- Annotated VCFs
- Graphical representations of functional impact on transcripts

Access:

- Webserver
- AWS cloud instance
- Source freely available



CLOUD APPLICATION



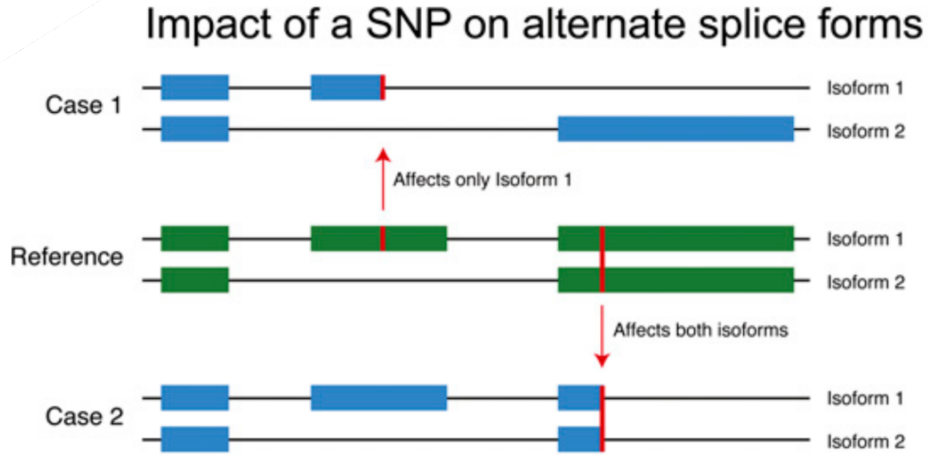
vat.gersteinlab.org

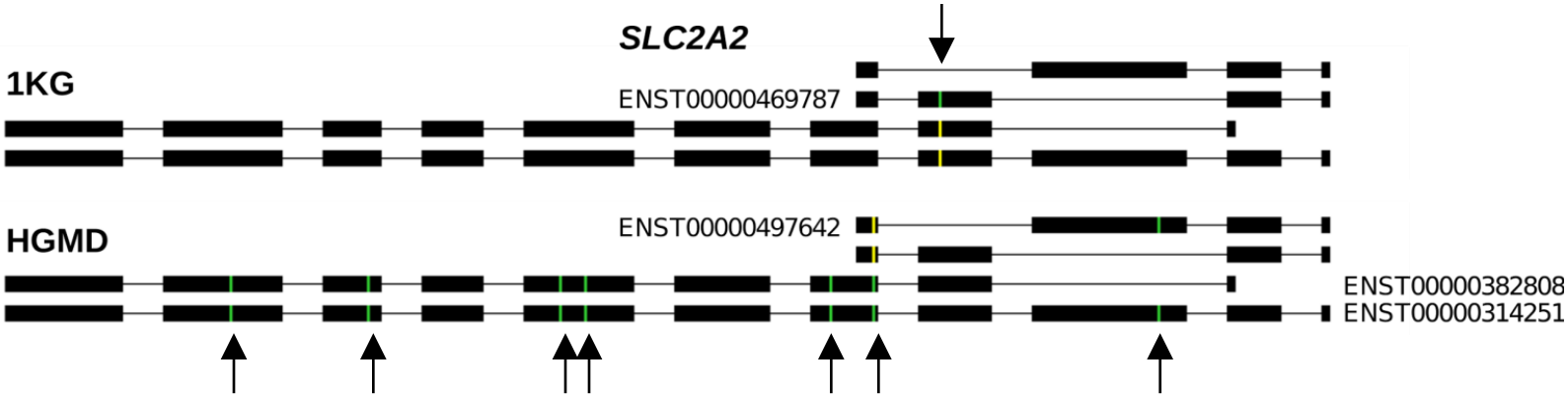
Habegger L., Balasubramanian S.*, et al. Bioinformatics, 2012*

Complexities in LOF annotation

Transcript isoforms,
distance to stop,
functional domains,
protein folding,
etc.

Balasubramanian S. et al., *Genes Dev.,* '11
Balasubramanian S.*, Fu Y.* et al., *NComms.*, '17



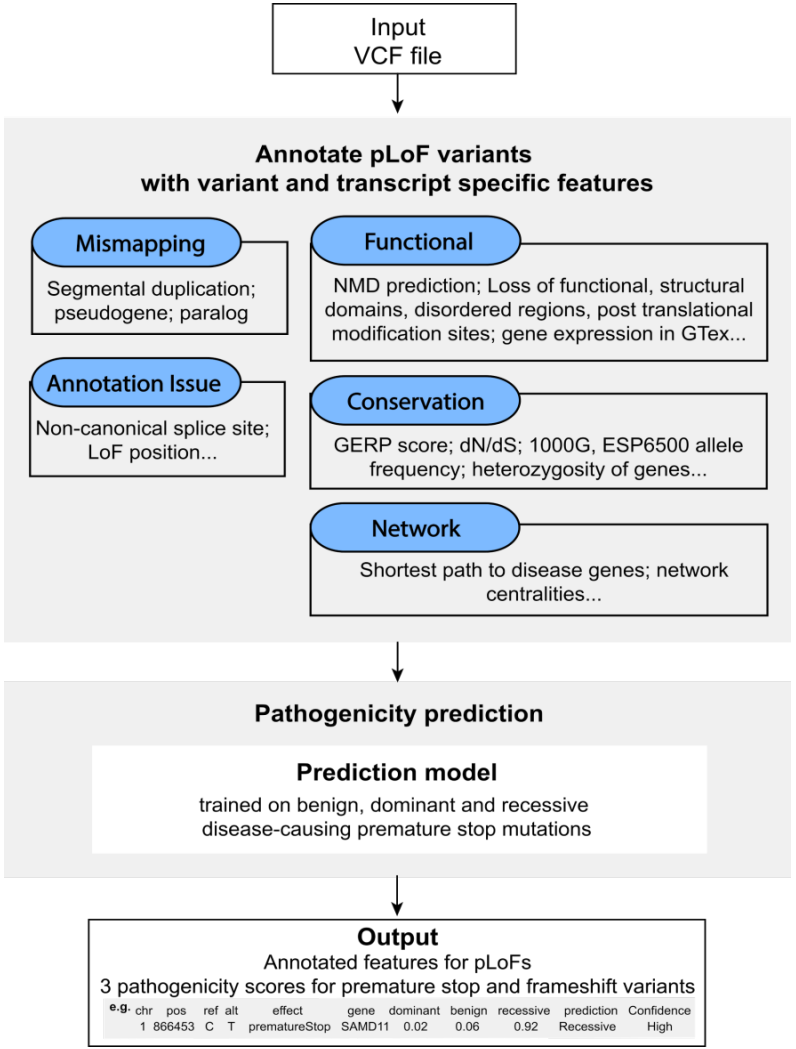


Annotation of Loss-of-Function Transcripts (ALoFT)

Runs on top of VAT

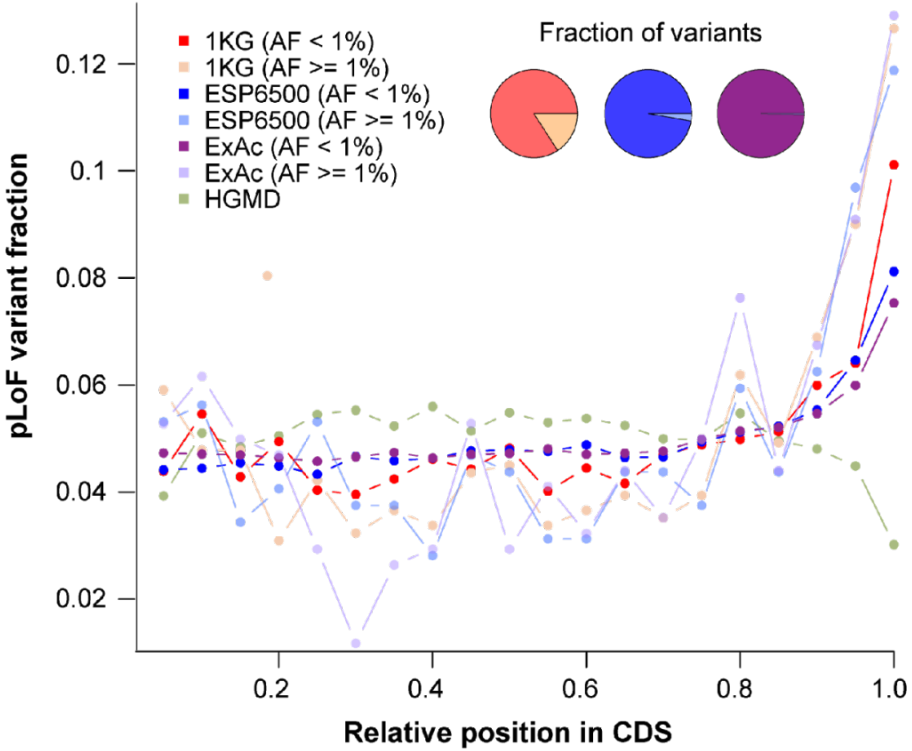
Output:

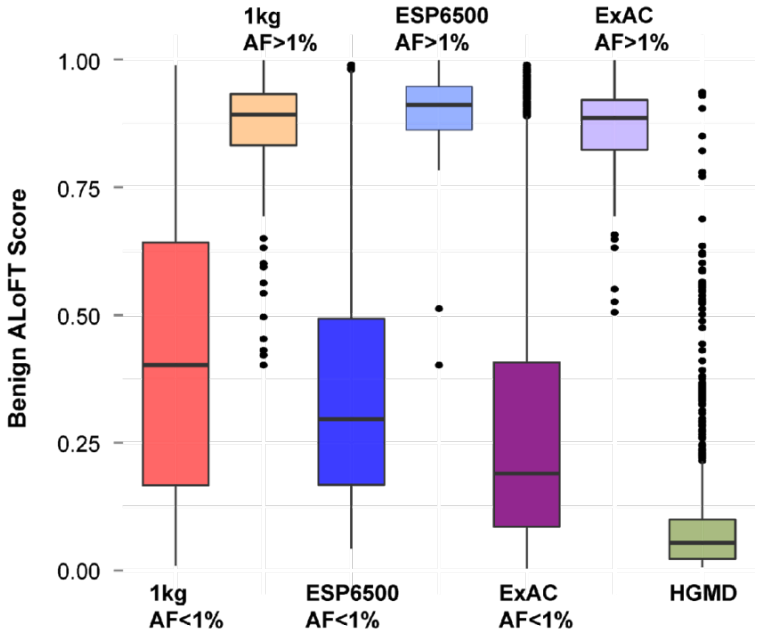
- Impact score: benign or deleterious.
- Decorated VCF.



Balasubramanian S.*, Fu Y.* et al., *NComms.*, '17

LoF distribution varies as expected by mutation set (from healthy people v from disease)





Balasubramanian S.*, Fu Y.* et al., *NComms.*, '17

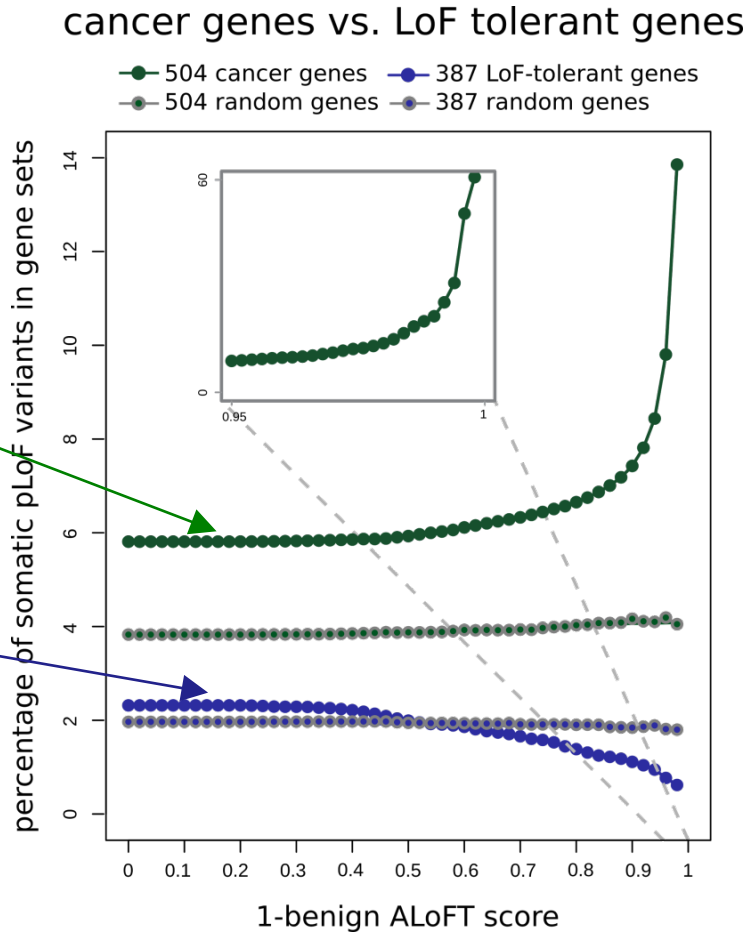
ALoFT identifies deleterious somatic LoF variants

Cancer genes:

- COSMIC consensus.
- *Enriched in deleterious LoFs.*

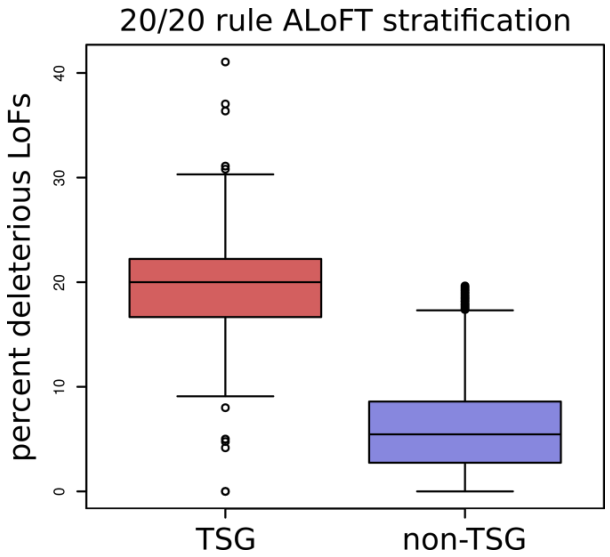
LoF tolerant genes:

- LoF in the 1KG cohort.
- *Depleted in deleterious LoFs.*



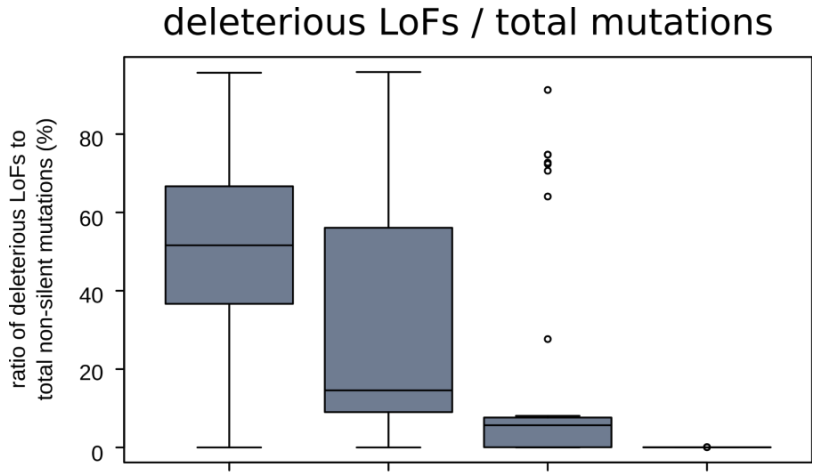
Balasubramanian S.*, Fu Y.* et al., *NComms.*, '17

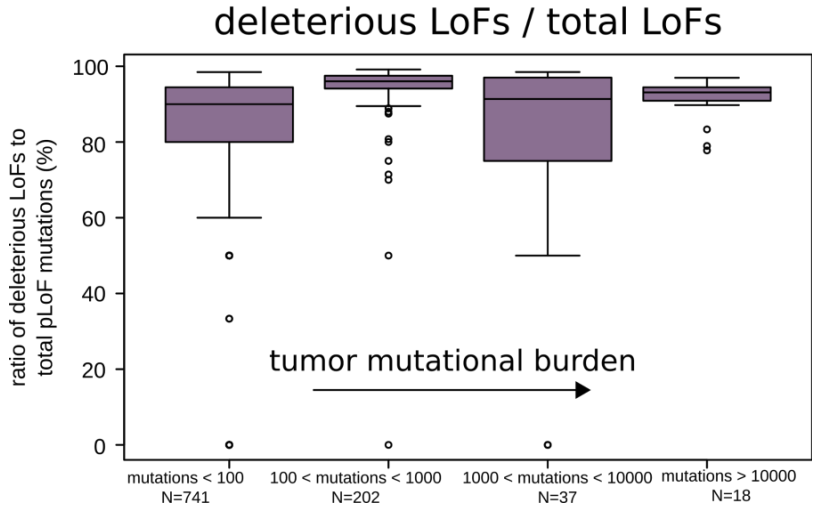
ALoFT refines cancer mutation characterization



Vogelstein et al. '13: if >20% of mutations in gene inactivating → tumor suppressor gene (TSG).

ALoFT further refines 20/20 rule predictions.

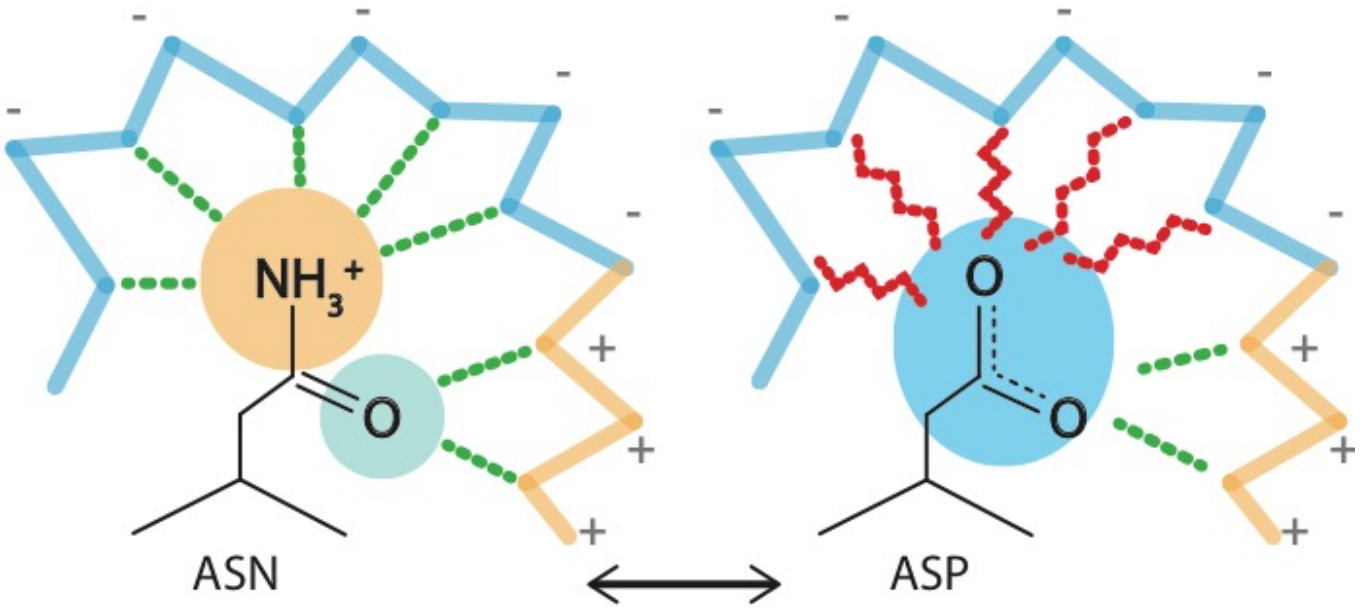


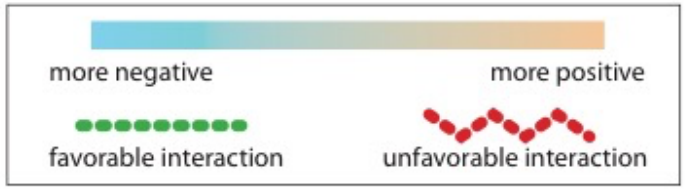


Balasubramanian S.*, Fu Y.* et al., *NComms.*, '17

Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

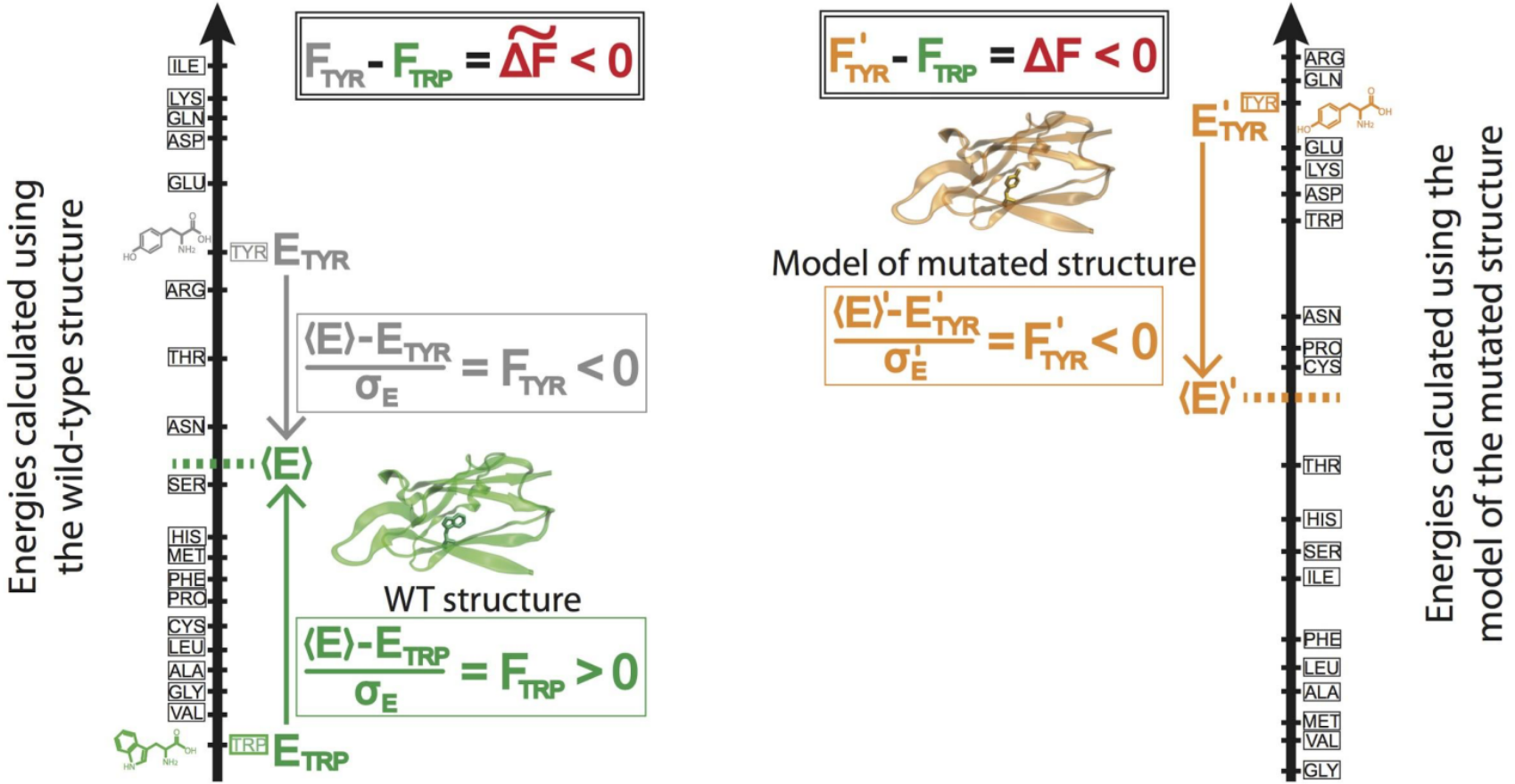




What is localized frustration ?

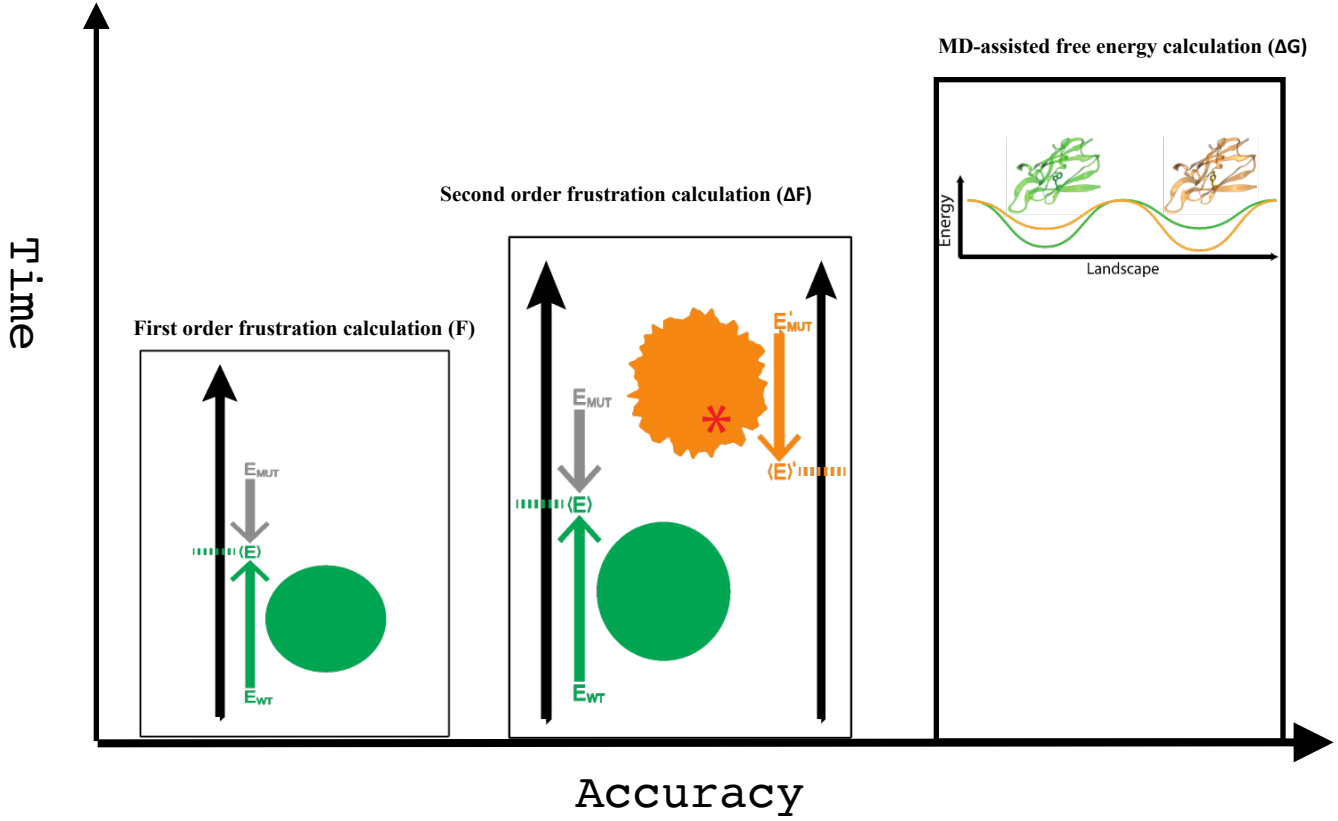
[Ferreiro et al., *PNAS* ('07)]

Workflow for evaluating localized frustration changes (ΔF)

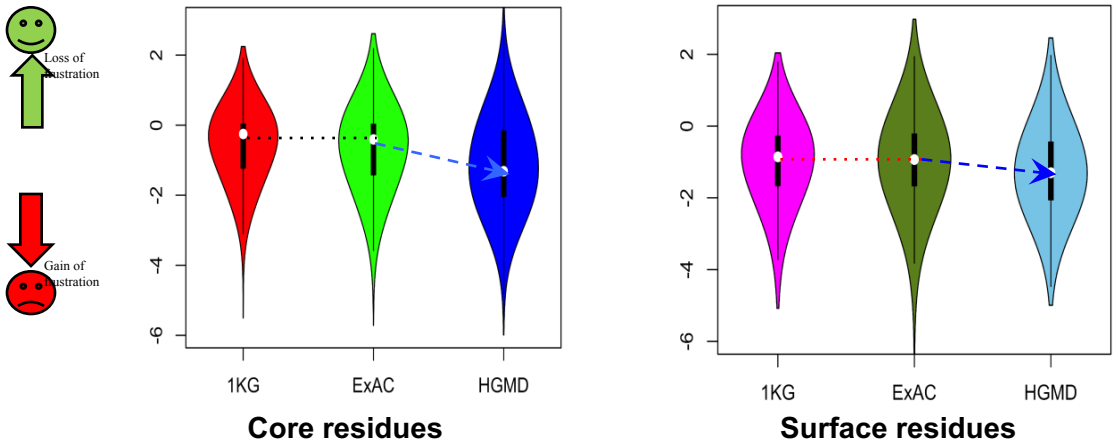


[Kumar et al. *NAR* (2016)]

Complexity of the second order frustration calculation



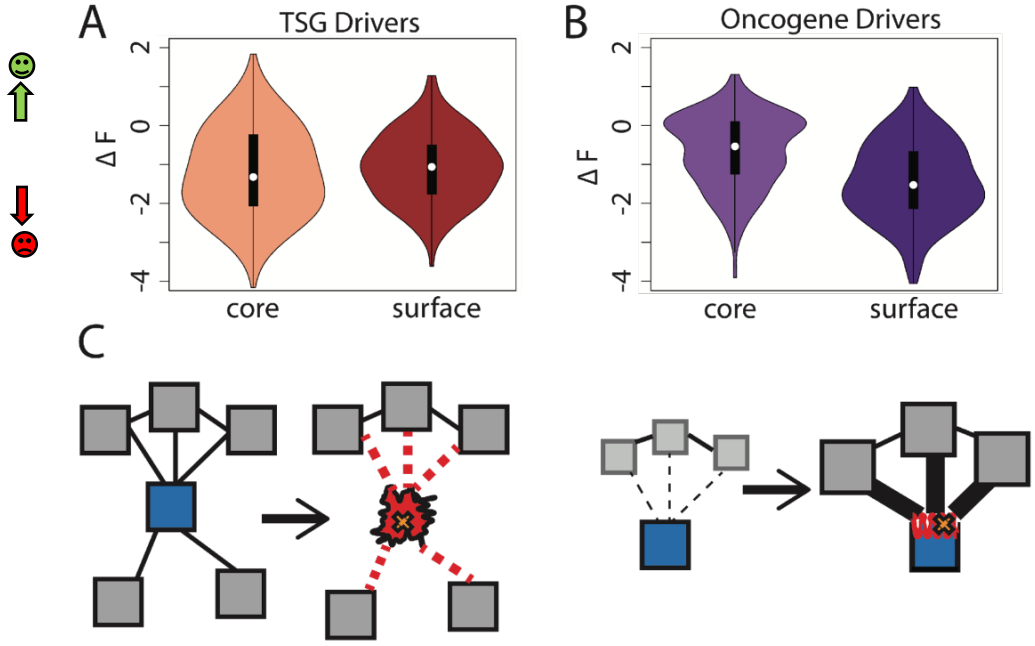
Comparing ΔF values across different SNV categories: disease v normal



Normal mutations (1000G) tend to unfavorably frustrate (less frustrated) surface more than core, but for disease mutations (HGMD) no trend & greater changes

[Kumar et al, *NAR* (2016)]

Comparison between ΔF distributions: TSGs v. oncogenes



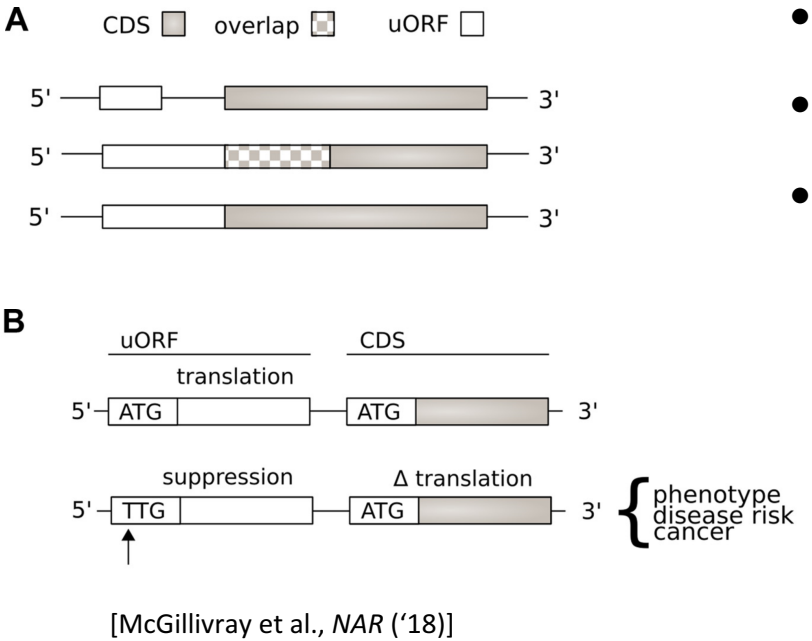
SNVs in TSGs change frustration more in core than the surface, whereas those associated with oncogenes manifest the opposite pattern. This is consistent with differences in LOF v GOF mechanisms.

[Kumar et al, *NAR* (2016)]

Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

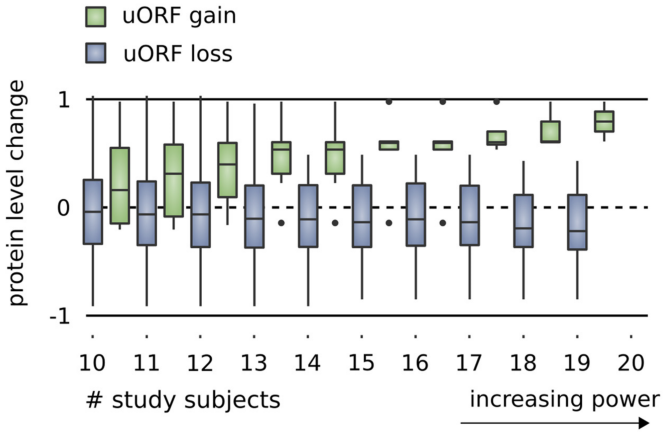
- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

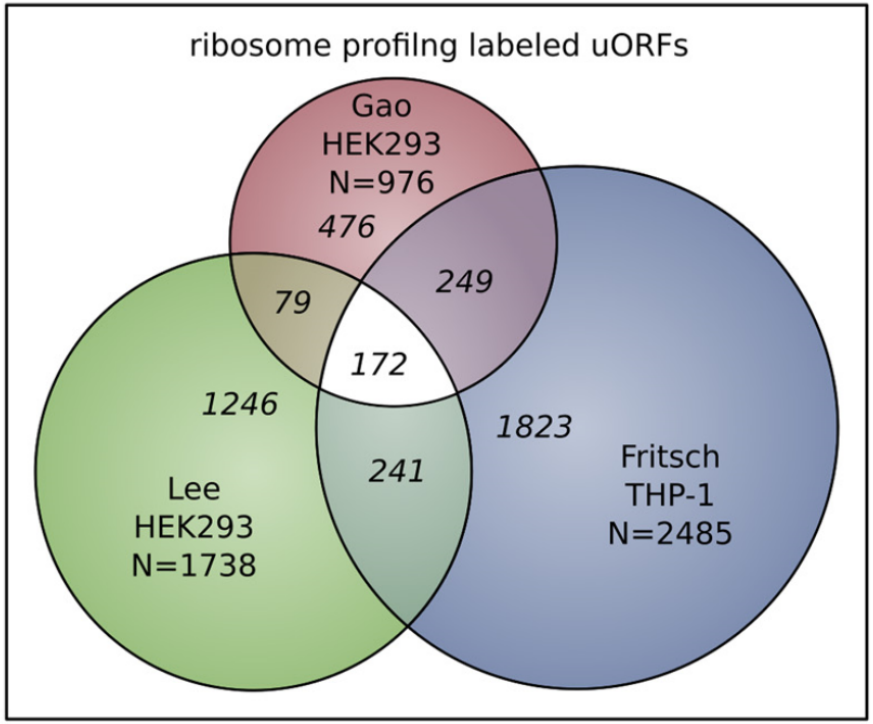
Upstream open reading frames (uORFs) regulate translation are affected by somatic mutation



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

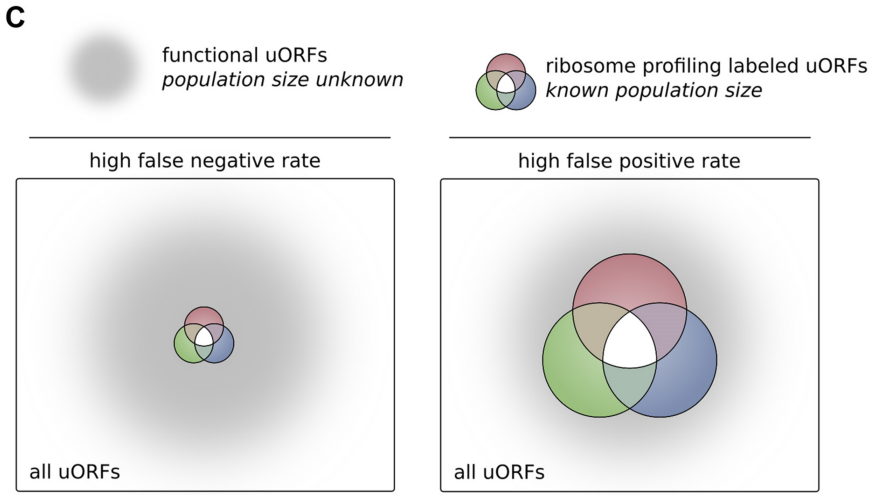
- uORFs regulate the translation of downstream coding regions.
- This regulation may be altered by somatic mutation in cancer.
- In Battle et al. 2014 data uORF gain & loss assoc. protein level change.





From a “Universe” of 1.3 M pot. uORFs

The population of functional uORFs may be significant

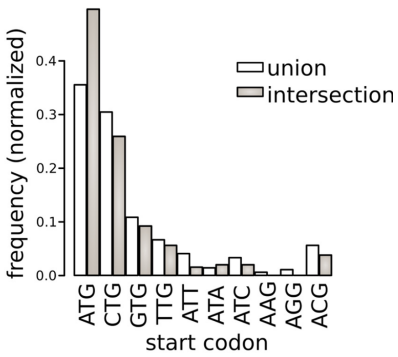


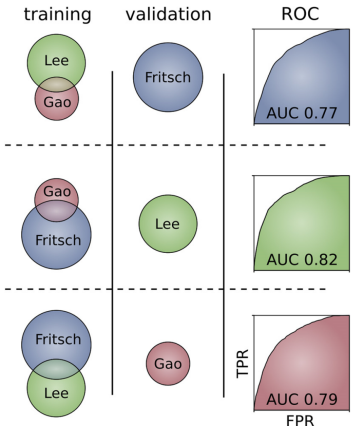
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

[McGillivray et al., *NAR* (‘18)]

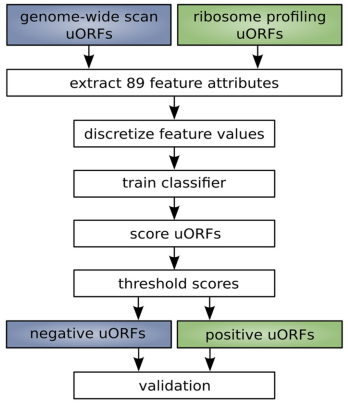
Prediction & validation of functional uORFs using 89 features

- All near-cognate start codons predicted.
- Cross-validation on independent ribosome profiling datasets and validation using in vivo protein levels and ribosome occupancy in humans (Battle et al. 2014).



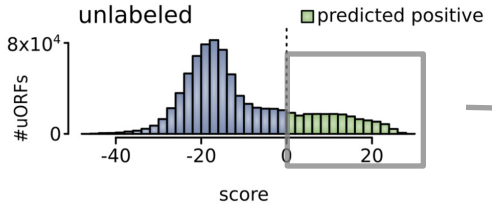


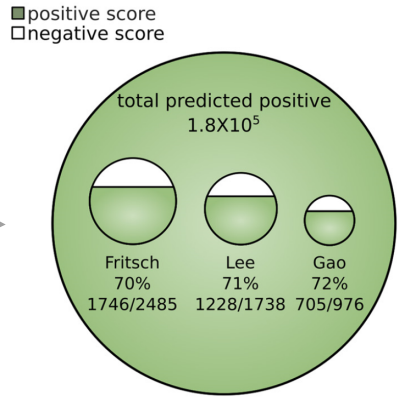
[McGillivray et al., *NAR* ('18)]

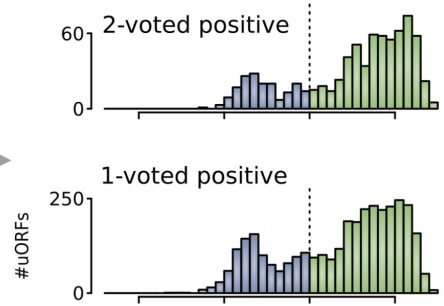


A comprehensive catalog of functional uORFs

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







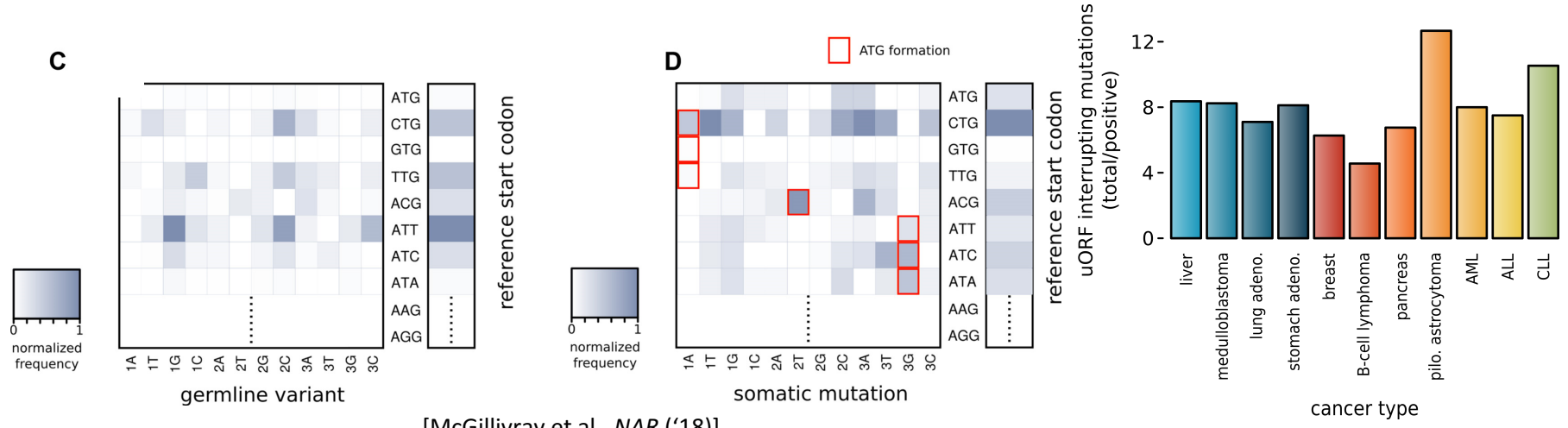
- Predicted functional uORFs may be intersected with disease associated variants.

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

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

Somatic alteration of uORFs disproportionately affects certain cancers and molecular pathways

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

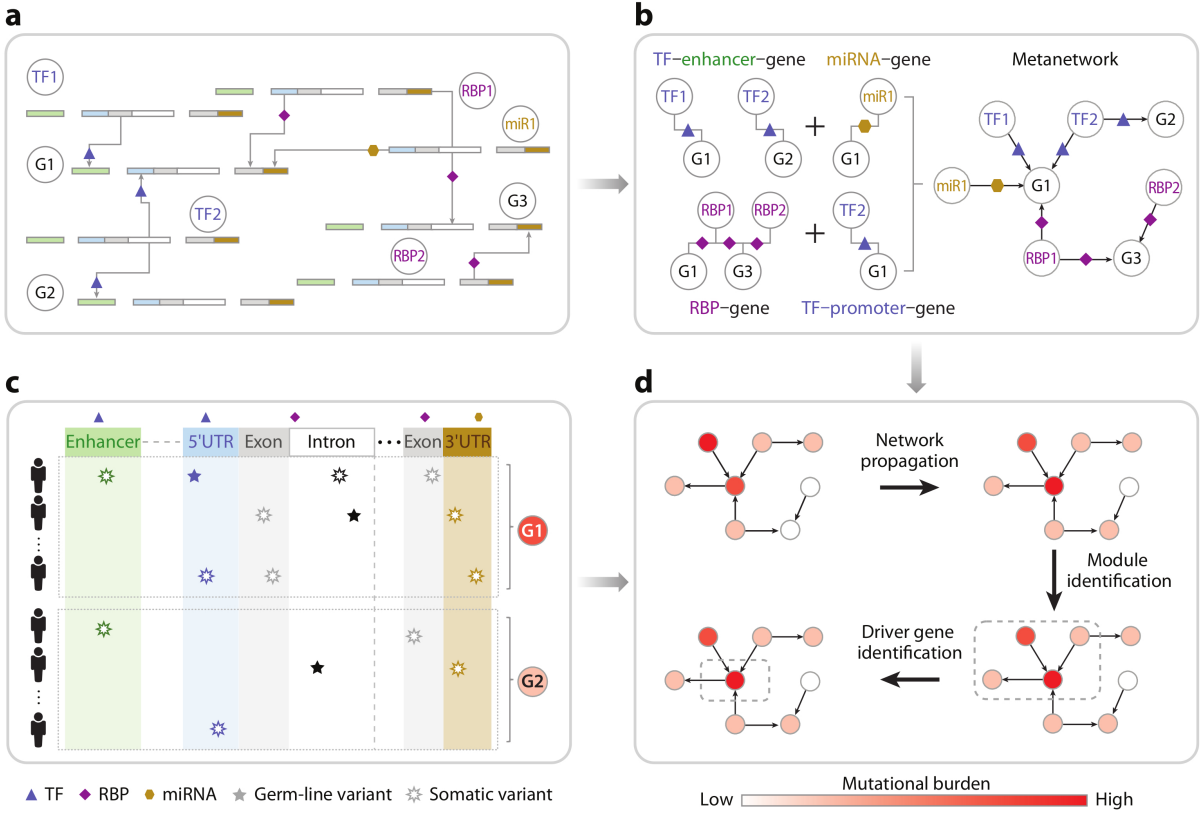


[McGillivray et al., *NAR* ('18)]

Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

Coding and non-coding elements may synergistically contribute to cancer



[McGillivray et al., *Ann. Rev. Biomedical Data Science* ('18), in press.]

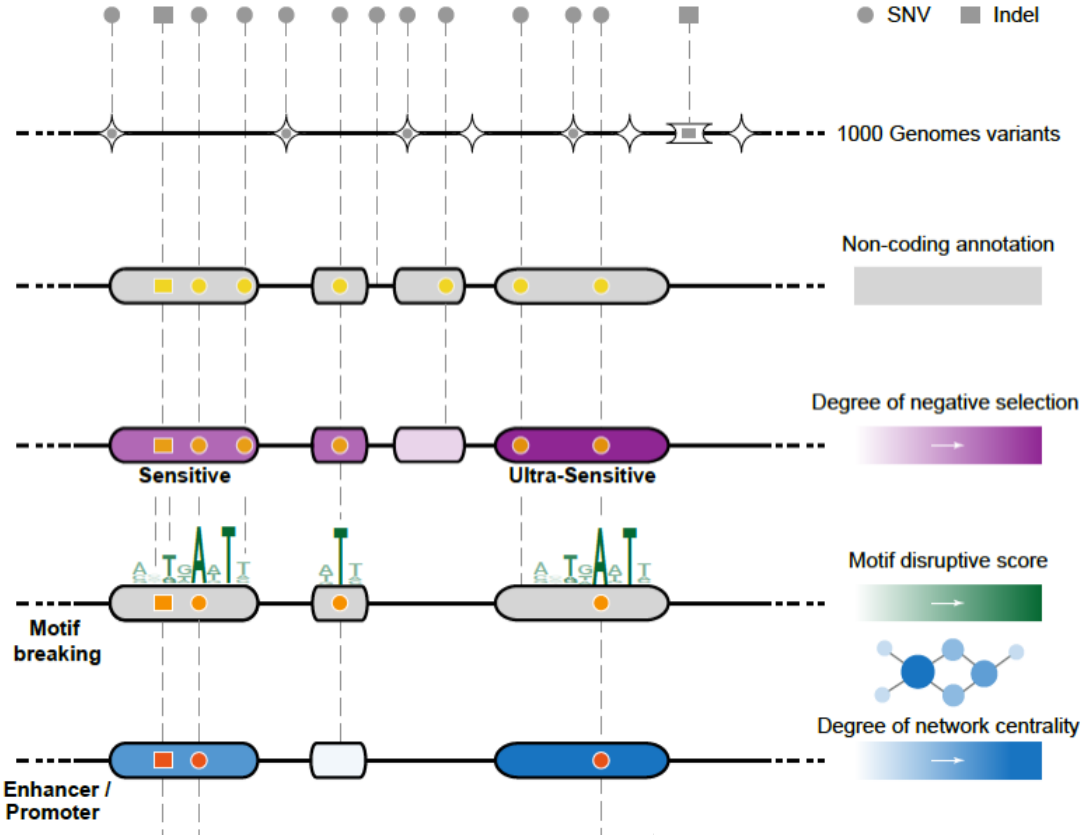
Funseq: a flexible framework to determine functional impact & use this to prioritize variants

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

Conservation (GERP, allele freq.)

Mutational impact (motif breaking, Lof)

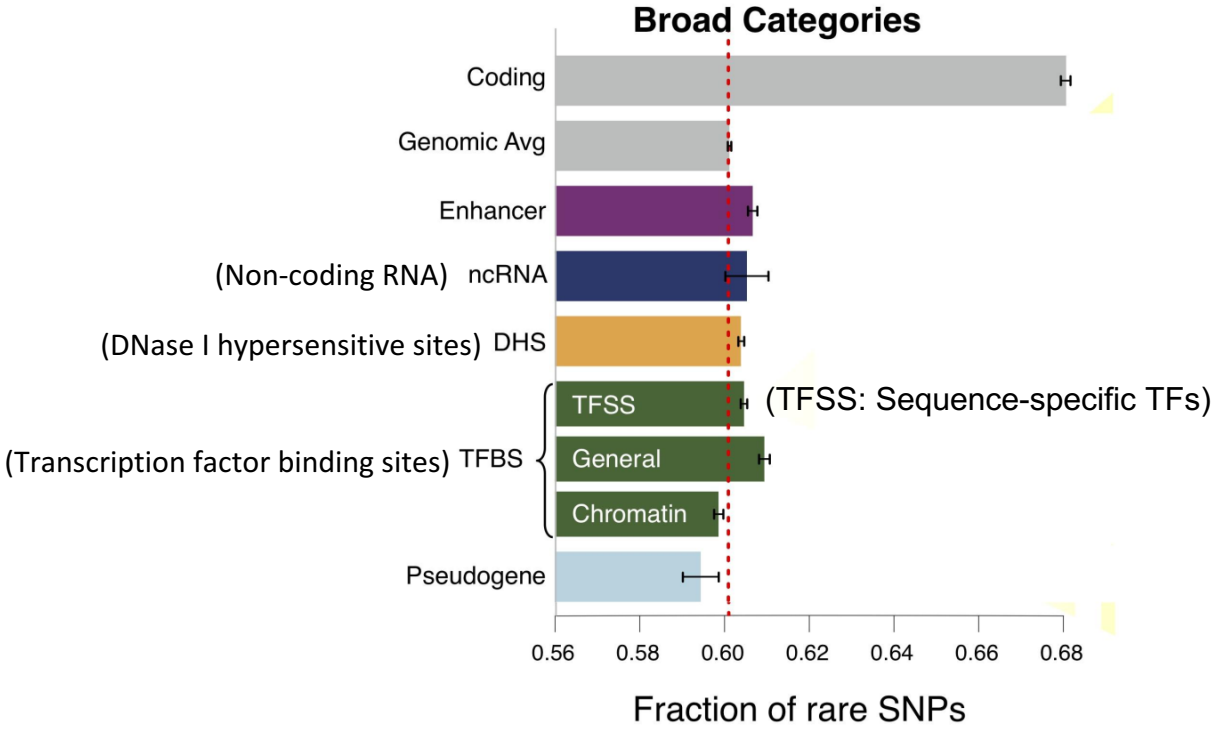
Network (centrality position)



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

Finding "Conserved" Sites in the Human Population:

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

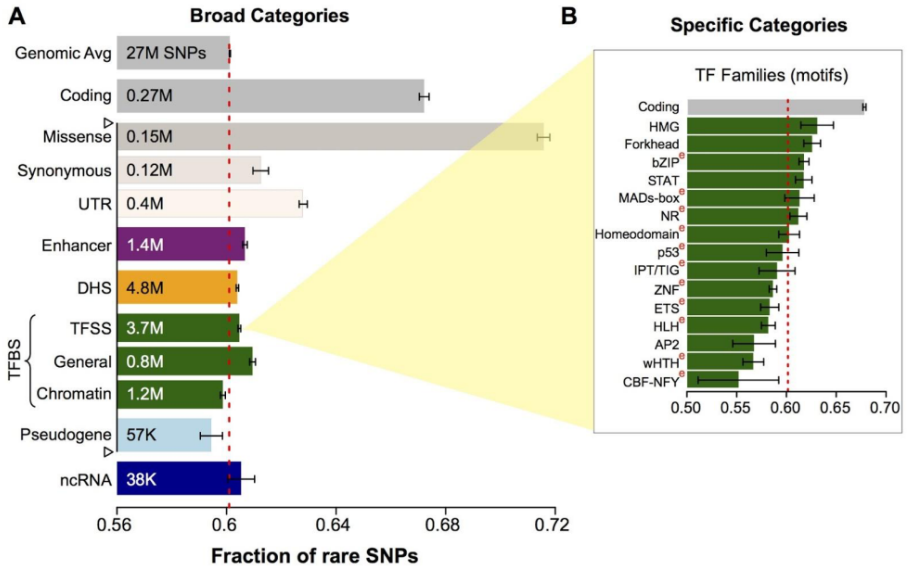


Depletion of Common Variants in the Human Population

Broad categories of regulatory regions under negative selection Related to:

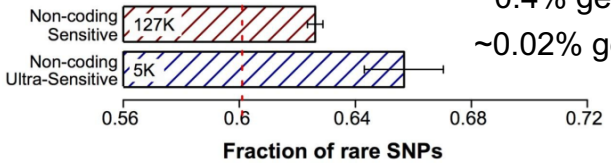
ENCODE, *Nature*, 2012
Ward & Kellis, *Science*, 2012
Mu et al, *NAR*, 2011

Differential selective constraints among specific sub-categories



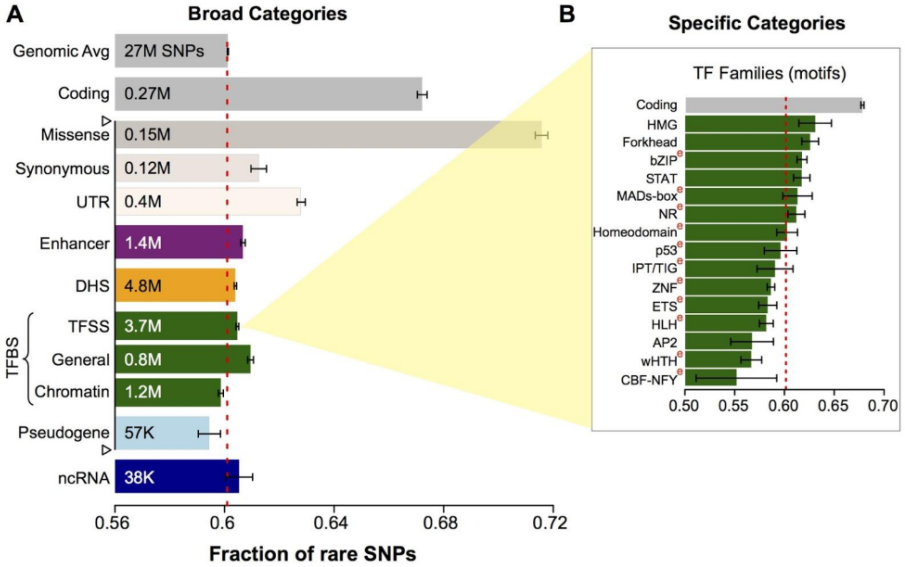
Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

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



~0.4% genomic coverage (~ top 25)

~0.02% genomic coverage (top 5)



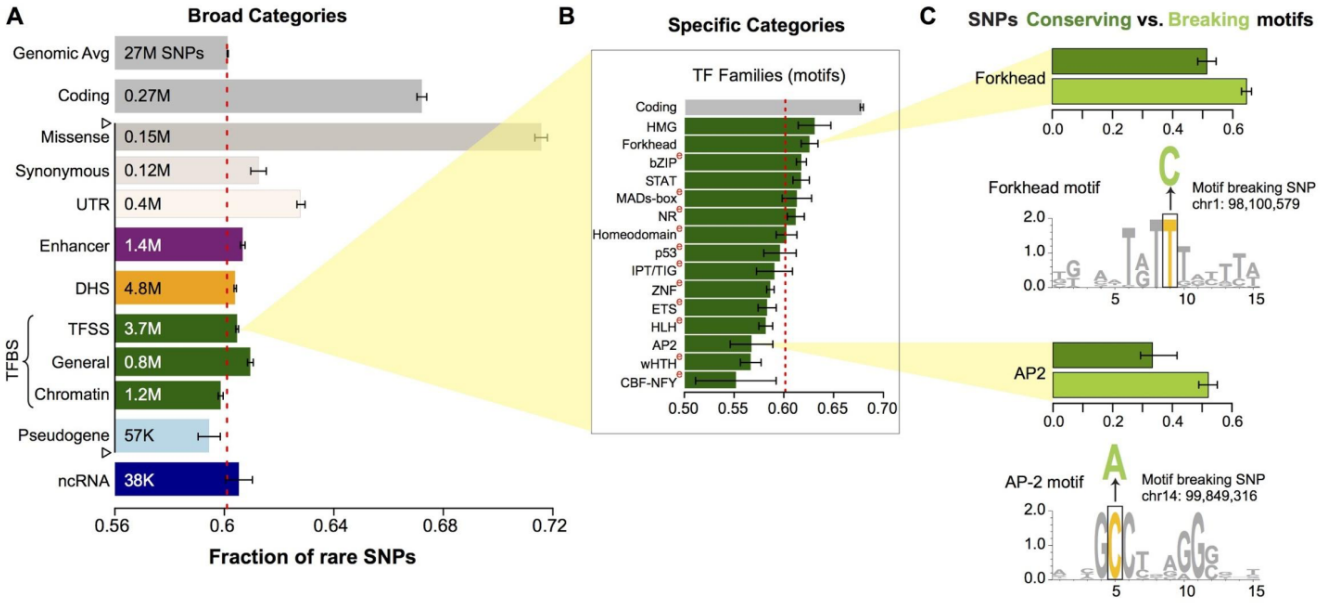
Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

Defining Sensitive non-coding Regions

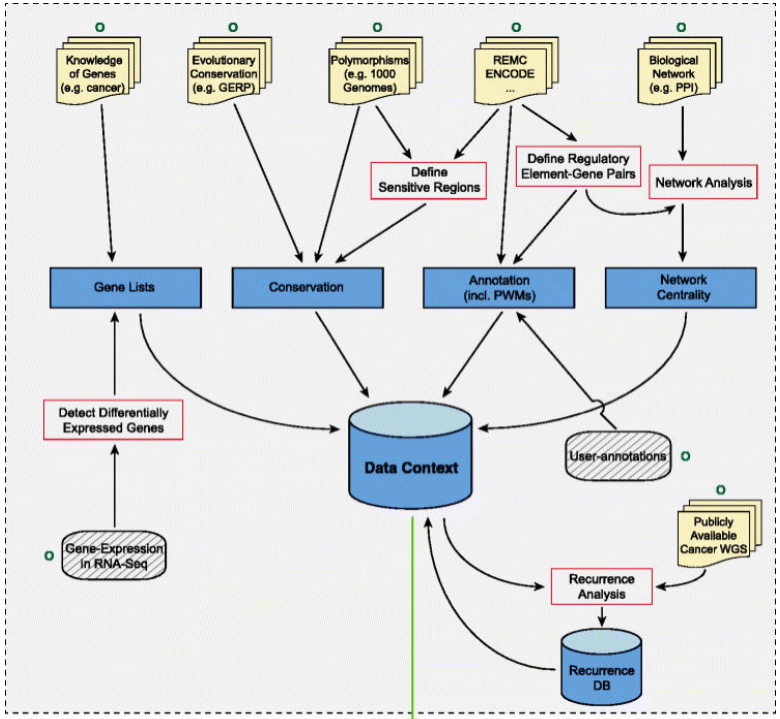
Start 677 high-resolution non-coding categories; Rank & find those under strongest selection

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

SNPs which break TF motifs are under stronger selection



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

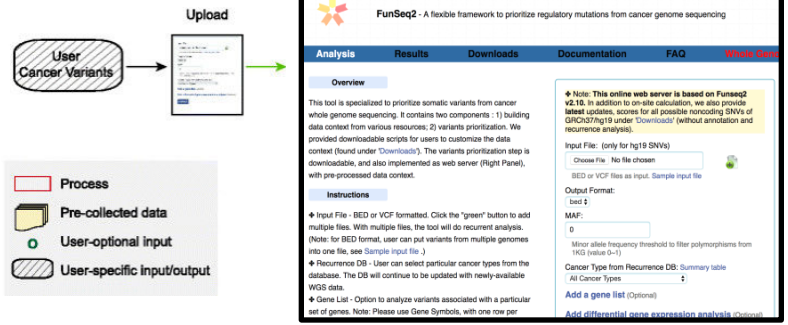


FunSeq.gersteinlab.org

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

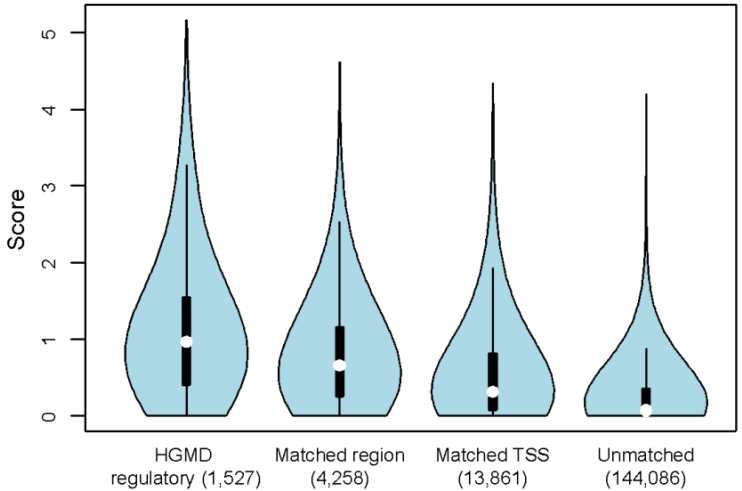
- Info. theory based method (ie annotation "surprisal") for weighting consistently many genomic features

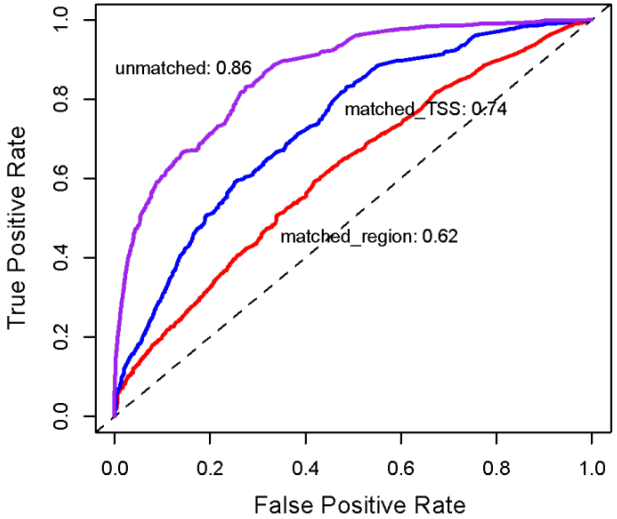
- Practical web server
- Submission of variants & pre-computed large data context from uniformly processing large-scale datasets



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

Germline pathogenic variants show higher core scores than controls





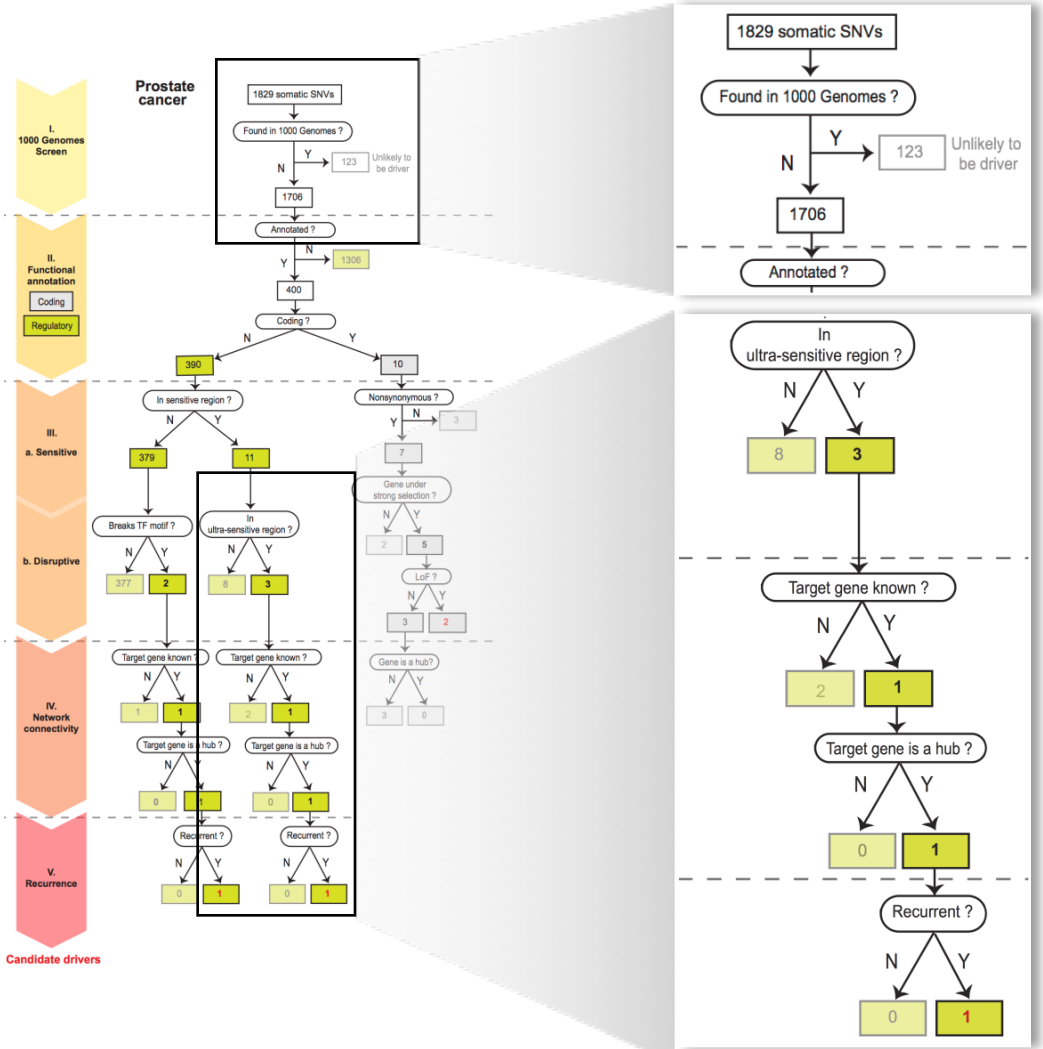
3 controls with natural polymorphisms (allele frequency >= 1%)

1. Matched region: 1kb around HGMD variants
2. Matched TSS: matched for distance to TSS
3. Unmatched: randomly selected

Ritchie et al., Nature Methods, 2014

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

Flowchart for 1 Prostate Cancer Genome (from Berger et al. '11)

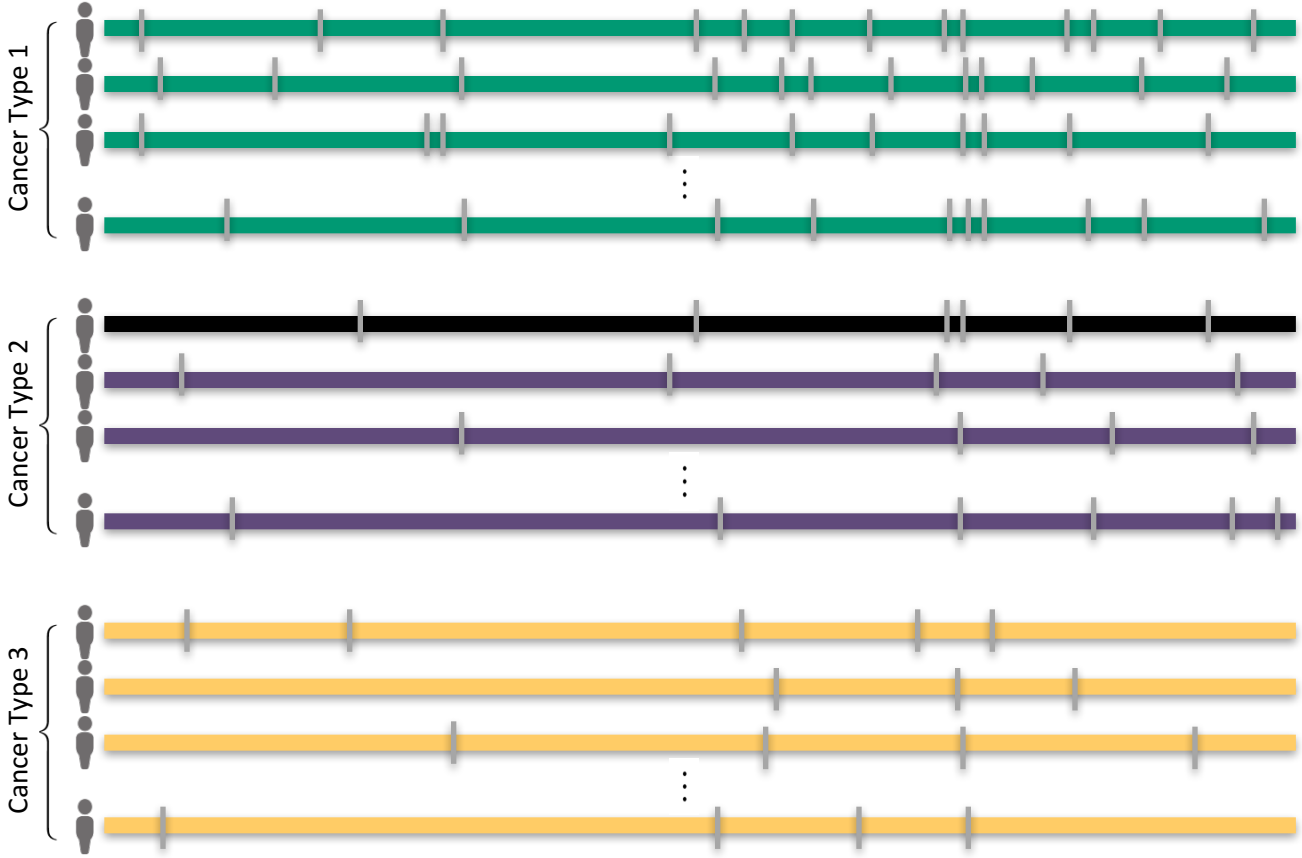


[Khurana et al., *Science* ('13)]

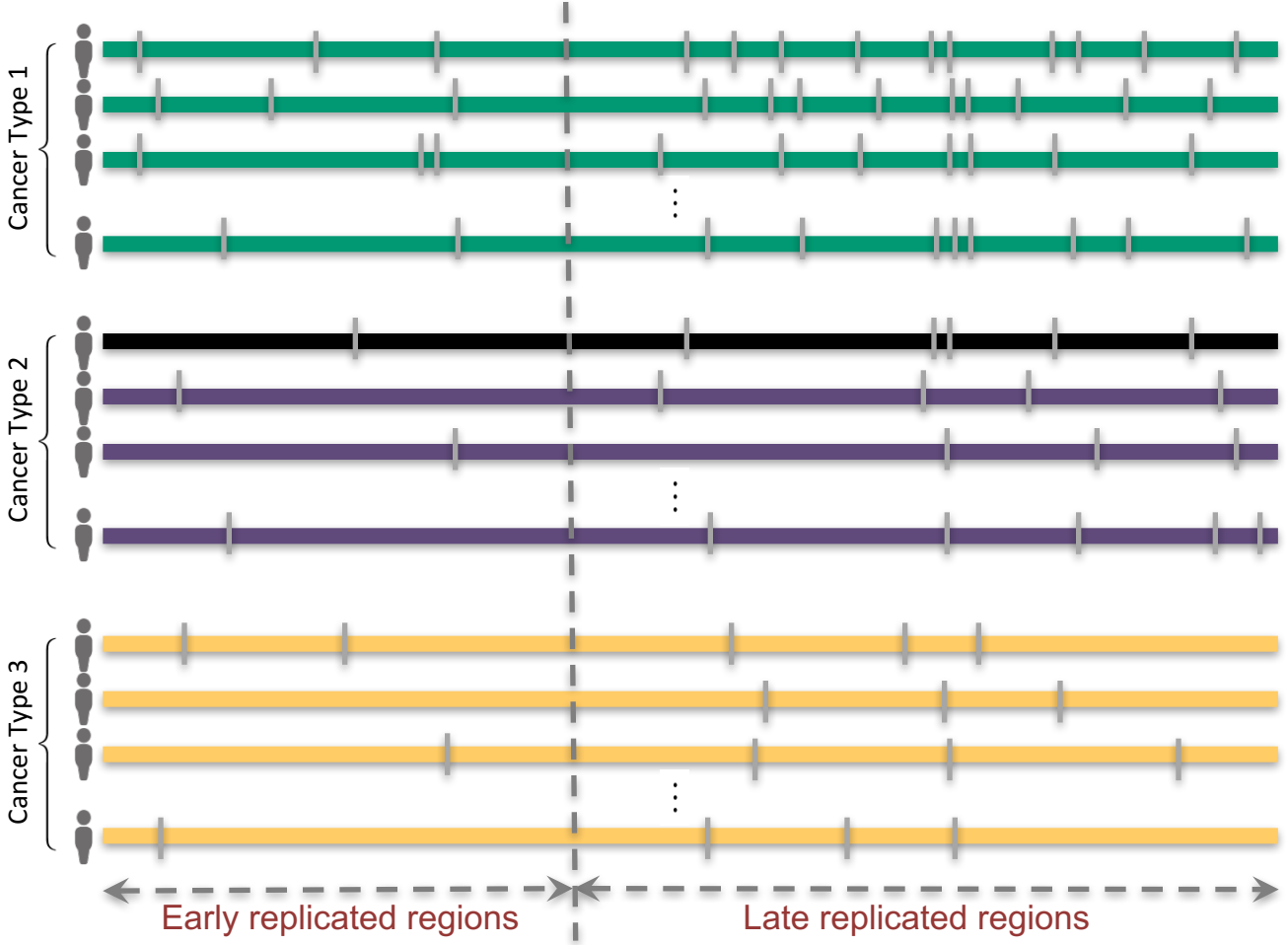
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

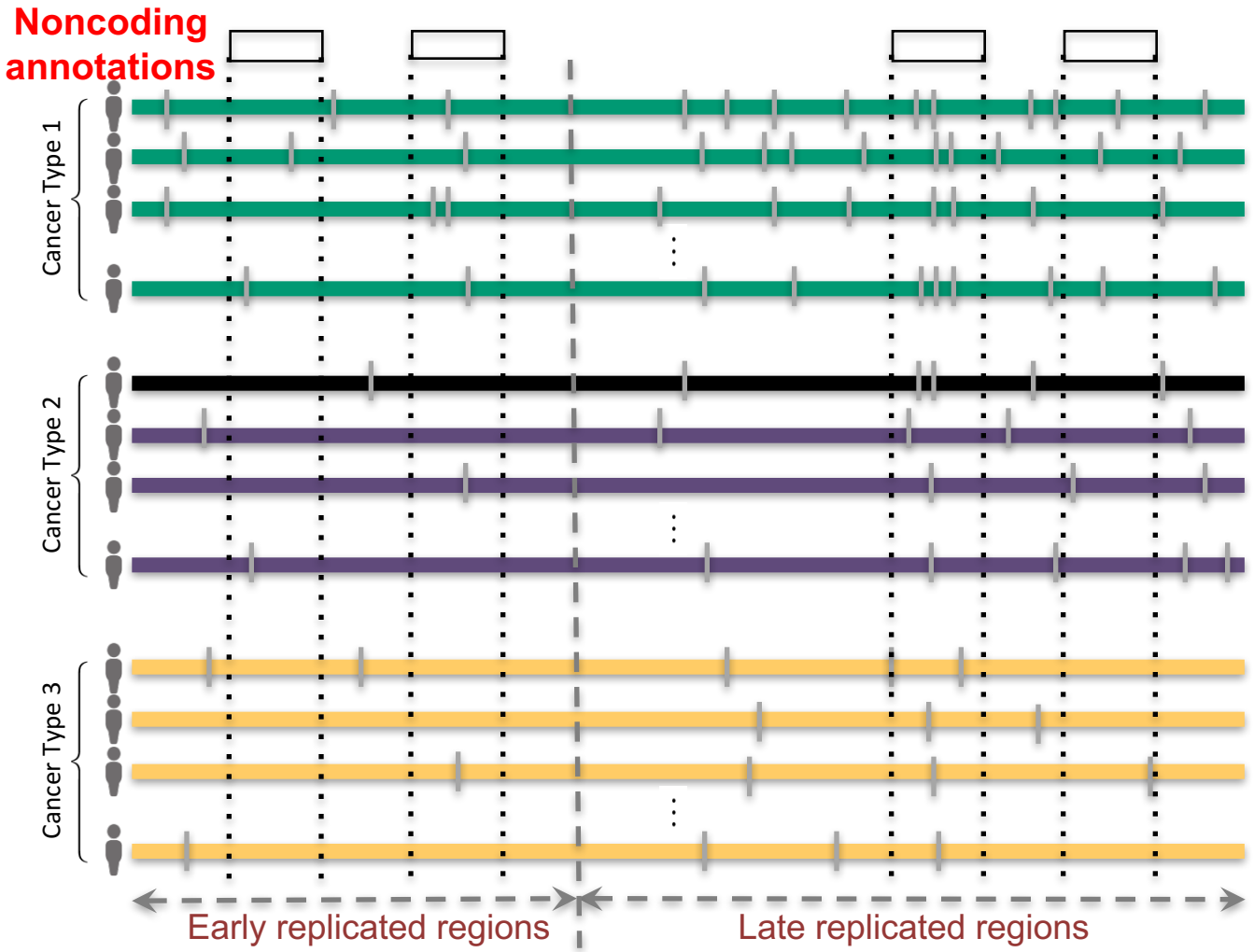
- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

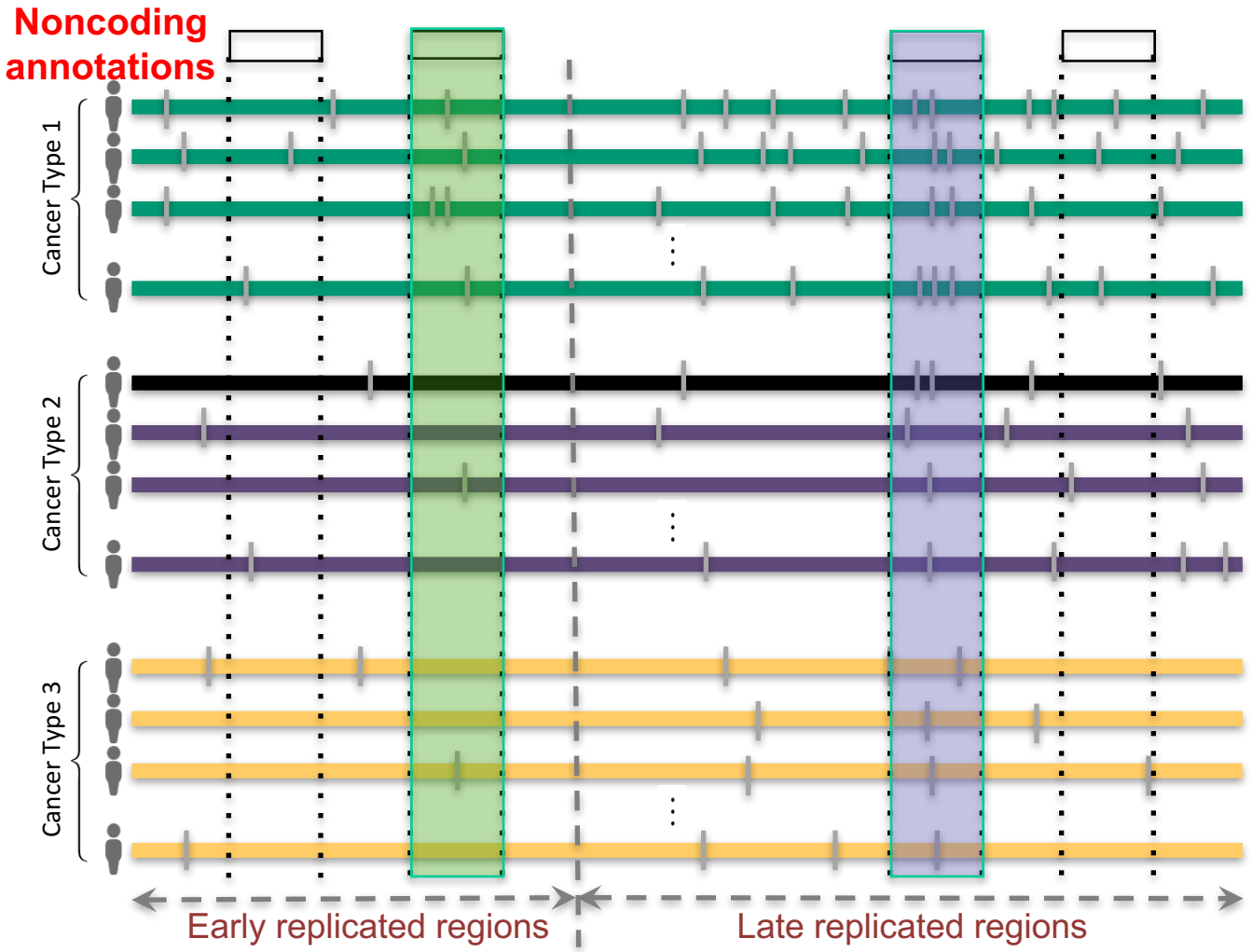
Mutation recurrence



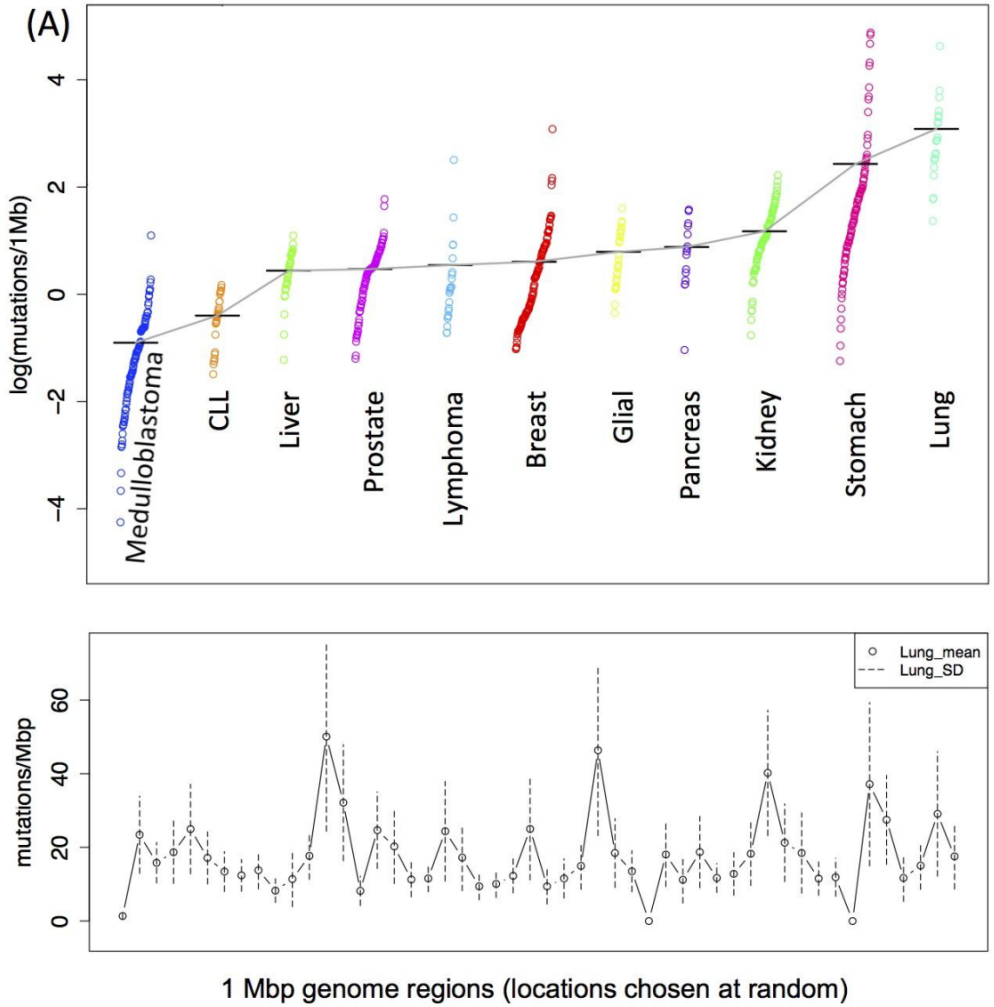
Mutation recurrence



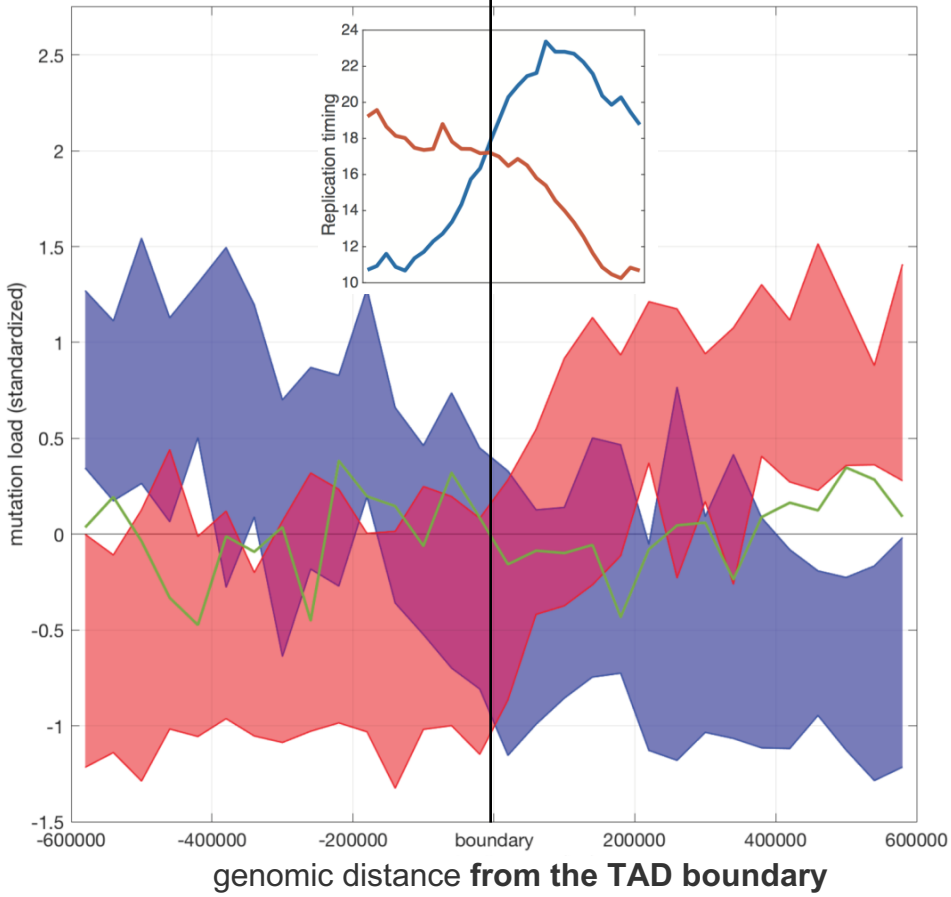

Noncoding annotations
Cancer Type 1
Cancer Type 2
Cancer Type 3
Early replicated regions
Late replicated regions


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

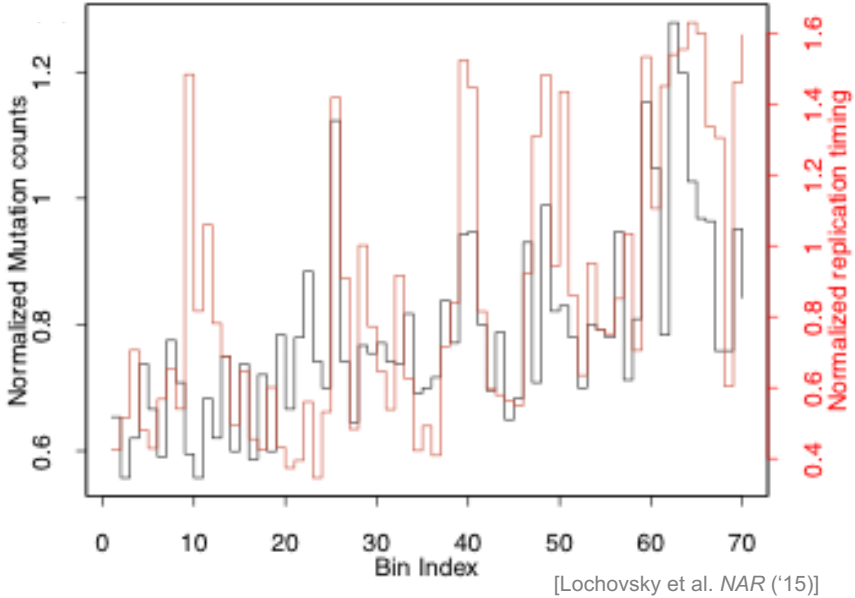
Cancer Somatic Mutational Heterogeneity, across cancer types, samples & regions



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



[Yan et al., *PLOS Comp. Bio. ('17)*; S. Li et al., PLOS Genetics ('17)]]



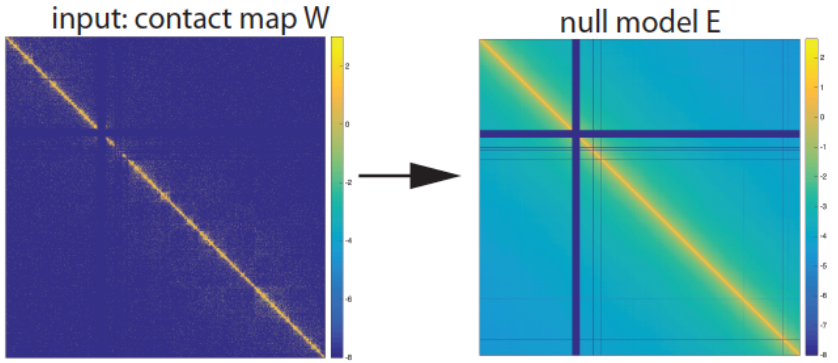
[Lochovsky et al. *NAR* ('15)]

Chromatin remodeling failure leads to more mutations in early-replicating regions

Variation in somatic mutations is closely associated with chromatin structure (TADs) & replication timing

mrTADFinder:

Identifying TADs at multiple resolutions by maximizing modularity vs appropriate null



Choose a particular resolution γ
Optimize Q over all possible partitions

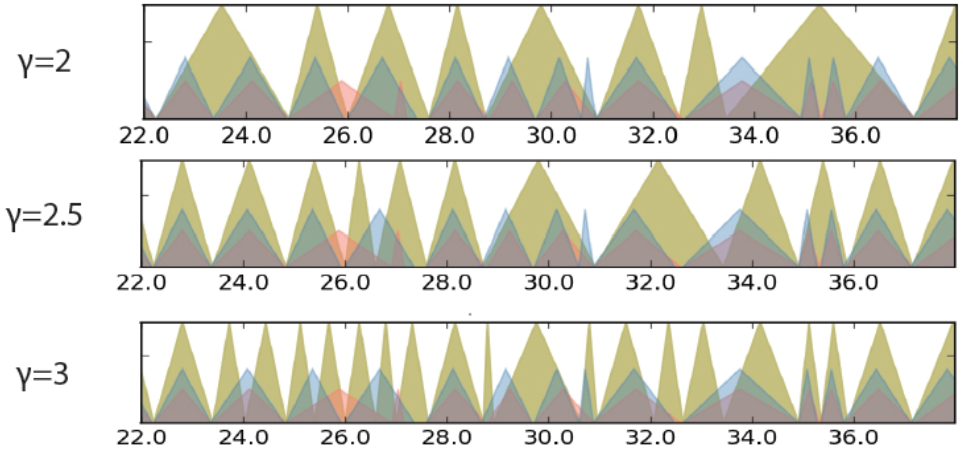
$$Q = \frac{1}{2N}\sum_{ij}(W_{ij} - \gamma E_{ij})\delta_{\sigma_i\sigma_j}$$

γ: resolution parameter

Multiple runs to define boundary scores for all pairs of adjacient bins

consensus boundaries based on the boundary scores

consensus TADs — output



[Yan et al., *PLOS Comp. Bio. ('17)*]

Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

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

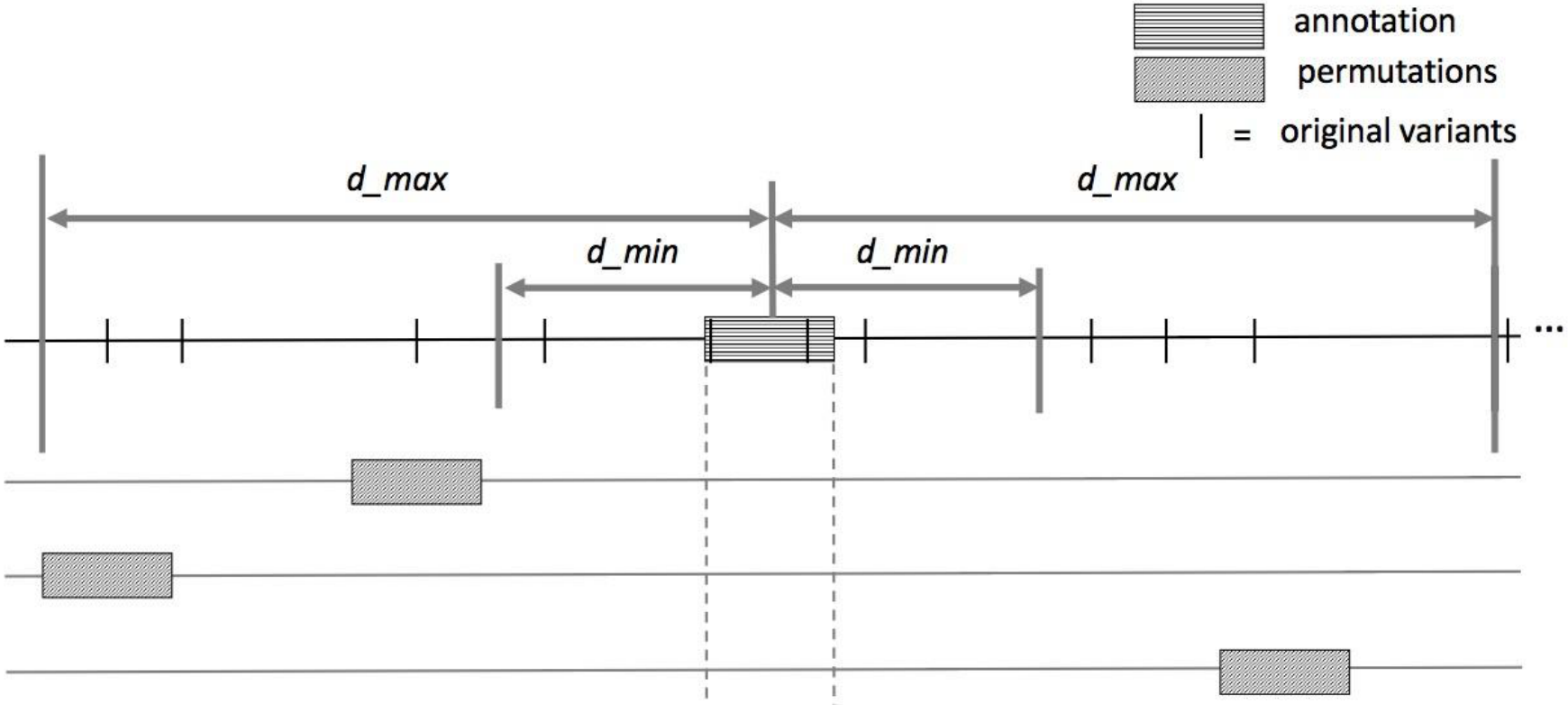
Shuffle annotations within local region to assess background mutation rate.

Model 3b: Random Permutation of Input Variants

Shuffle variants within local region to assess background mutation rate.

[Lochovsky et al. *Bioinformatics* in press]

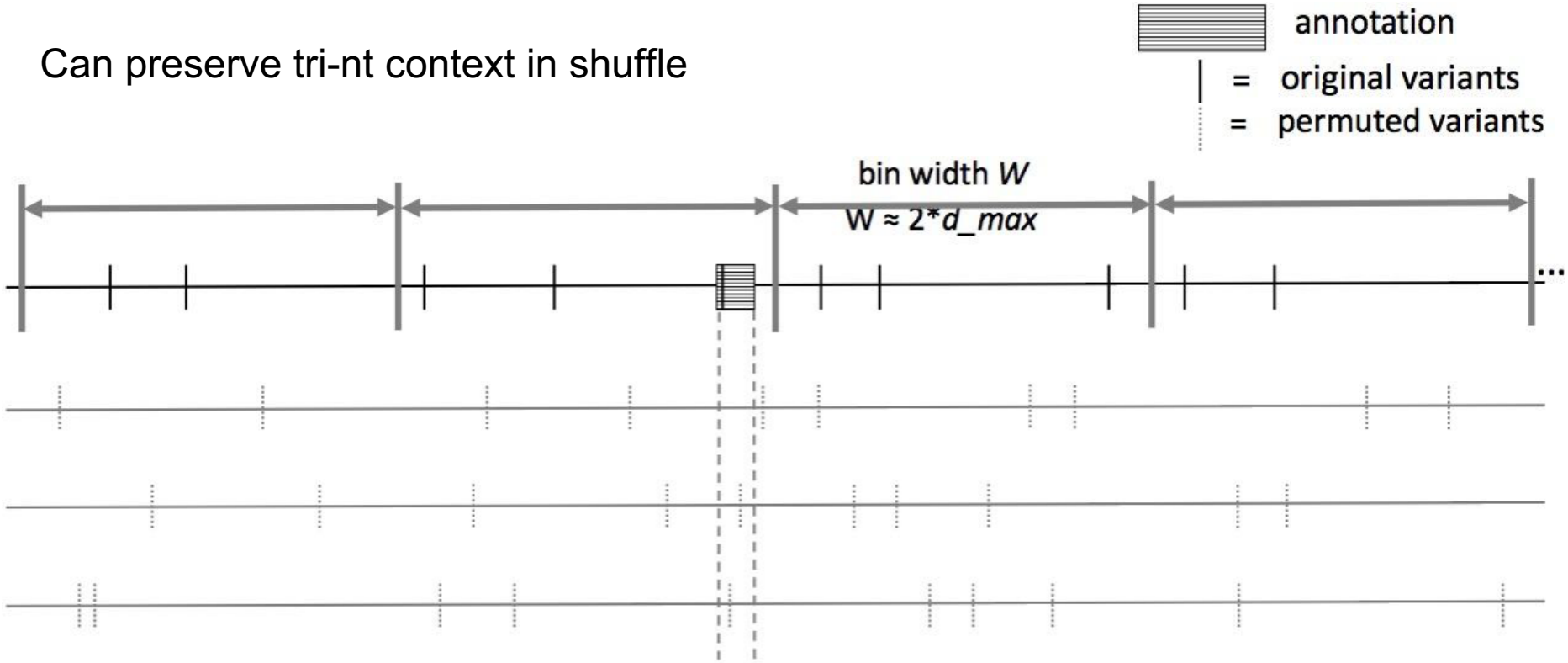
MOAT-a: Annotation-based permutation



[Lochovsky et al. *Bioinformatics* in press]

MOAT-v: Variant-based Permutation

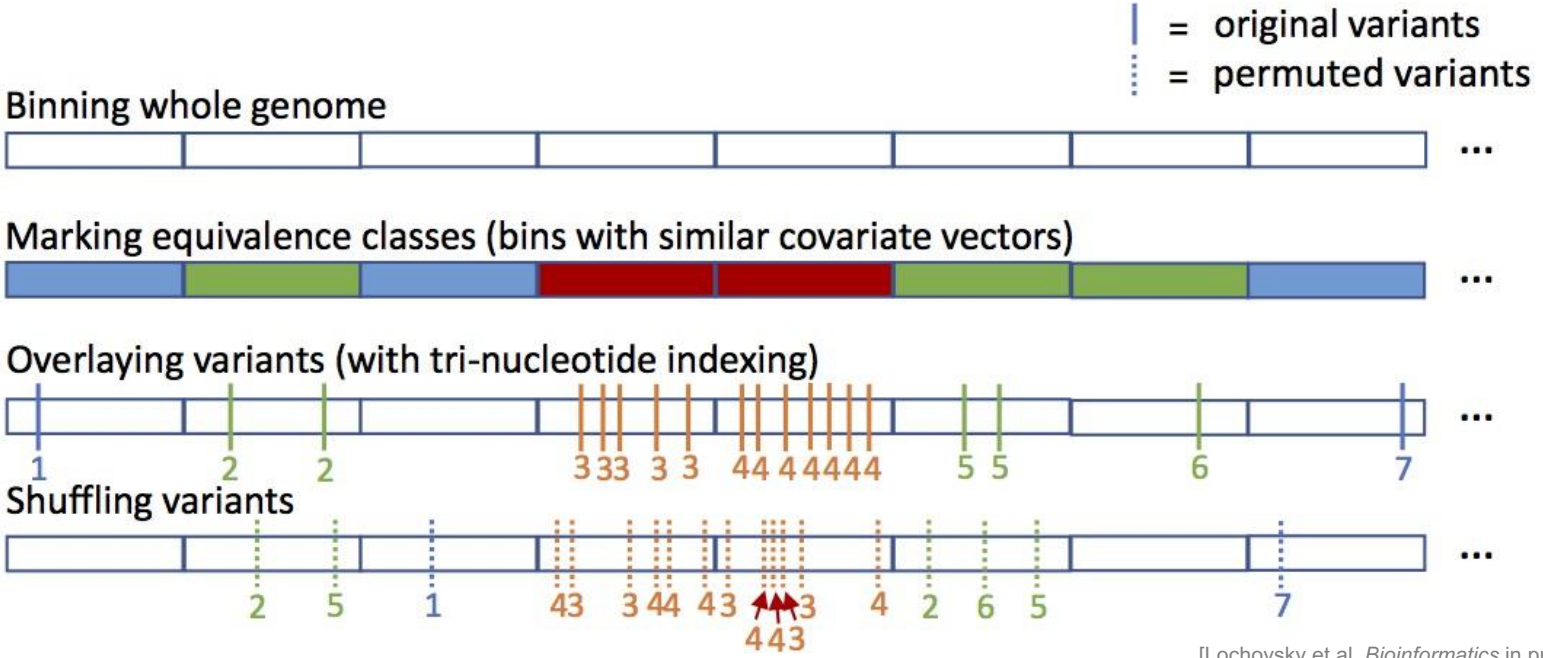
Can preserve tri-nt context in shuffle



[Lochovsky et al. *Bioinformatics* in press]

MOAT-s: a variant on MOAT-v

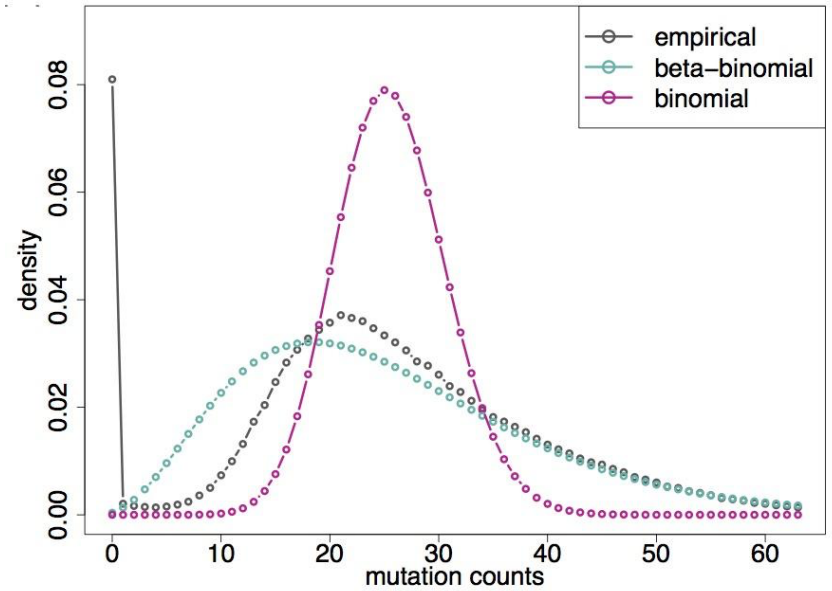
- A somatic variant simulator
  - Given a set of input variants, shuffle to new locations, taking genome structure into account



[Lochovsky et al. *Bioinformatics* in press]

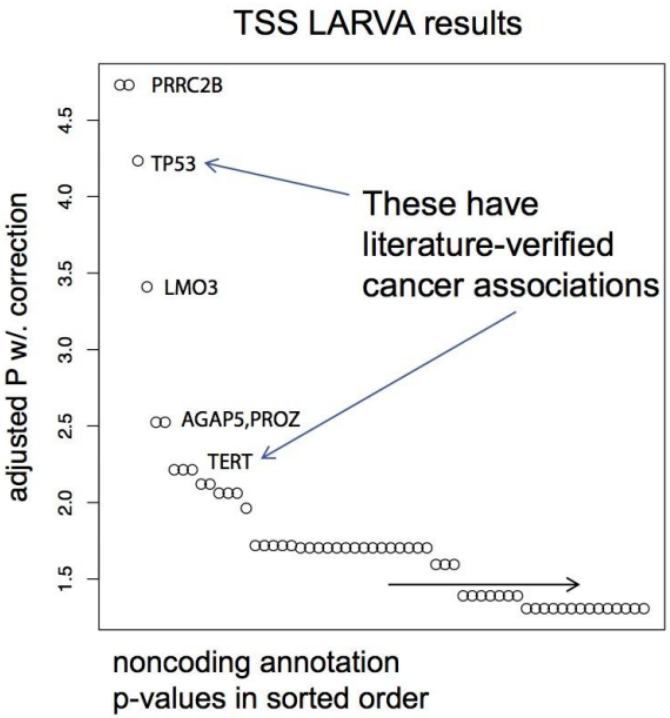
LARVA Model Comparison

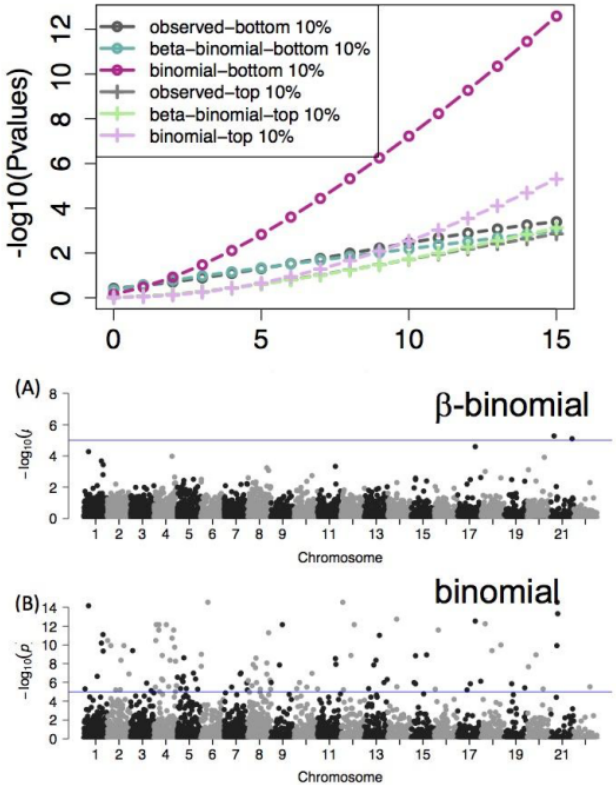
- Comparison of mutation count frequency implied by the binomial model (model 1) and the beta-binomial model (model 2) relative to the empirical distribution
- The beta-binomial distribution is significantly better, especially for accurately modeling the over-dispersion of the empirical distribution



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

LARVA Results





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

MOAT: recapitulates LARVA with GPU-driven runtime scalability

Gene Name	Documented role with cancer	Pubmed ID
SLC3A1	Cysteine transporter SLC3A1 promotes breast cancer tumorigenesis	28382174
ADRA2B	reduce cancer cell proliferation, invasion, and migration	25026350
SIL1	subtype-specific proteins in breast cancer	23386393
TCF24	NA	NA
AGAP5	significant mutation hotspots in cancer	25261935
TMPRSS13	Type II transmembrane serine proteases in cancer and viral infections	19581128
ERO1L	Overexpression of ERO1L is Associated with Poor Prognosis of Gastric Cancer	26987398

.
.
.

MOAT's high mutation burden elements recapitulate LARVA's results & published noncoding cancer-associated elements.

Computational efficiency of MOAT's NVIDIA™ CUDA™ version, with respect to the number of permutations, is dramatically enhanced compared to CPU version.

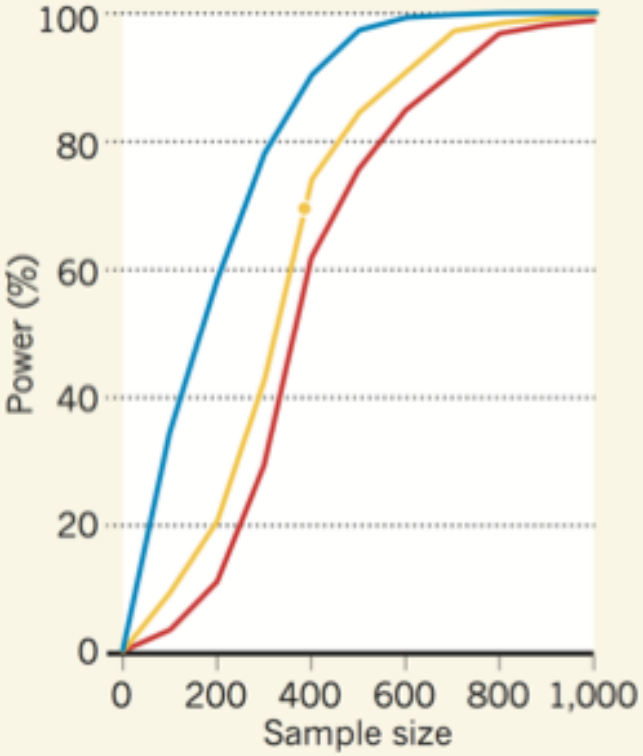
Number of permutations	Fold speedup of CUDA version
1k	14x
10k	100x
100k	256x

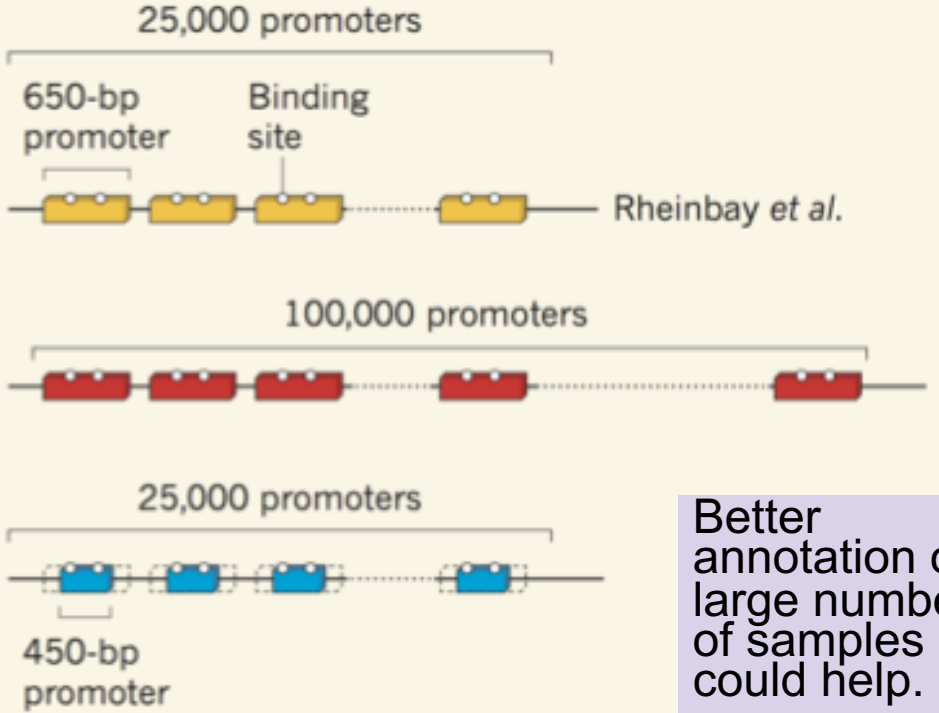
[Lochovsky et al. *Bioinformatics* in press]

Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

Power, as an issue in driver discovery





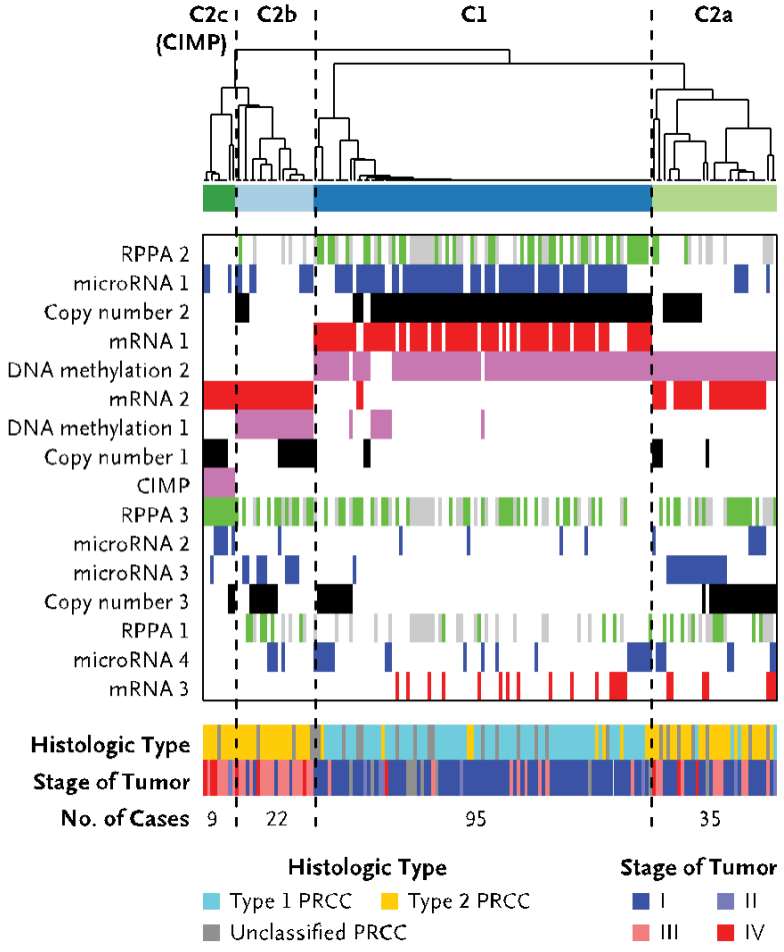
Better annotation or large number of samples could help.

[Kumar & Gerstein, *Nature* ('17)]

An (underpowered) case study: pRCC

- Kidney cancer lifetime risk of 1.6% & the papillary type (pRCC) counts for ~10% of all cases
- TCGA project sequenced 161 pRCC exomes & classified them into subtypes
  - Yet, cannot pin down the cause for a significant portion of cases....

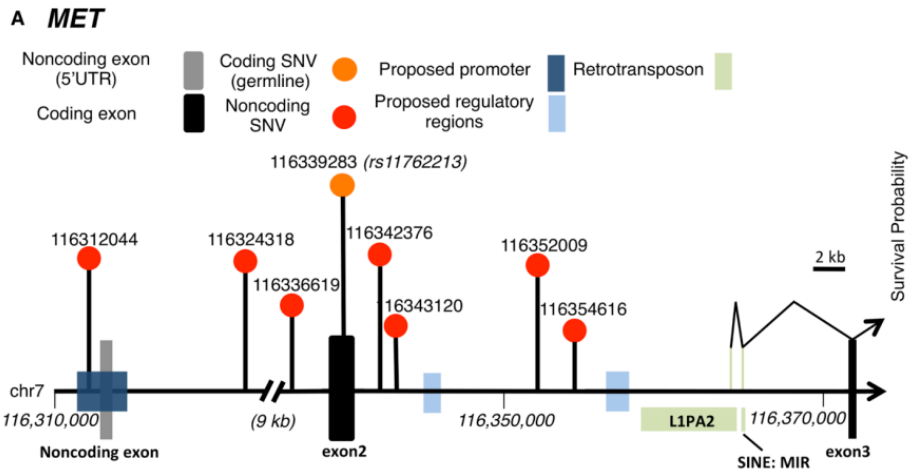
•35 WGS of TN pairs,
perhaps useful? But not that definitive from a recurrence perspective

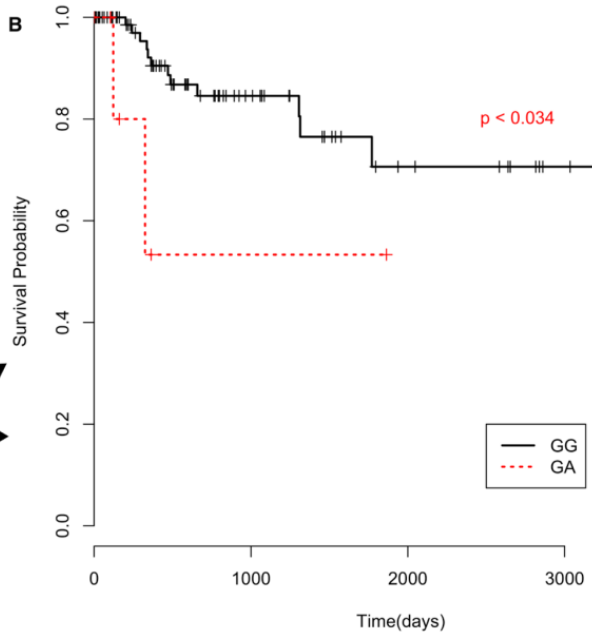


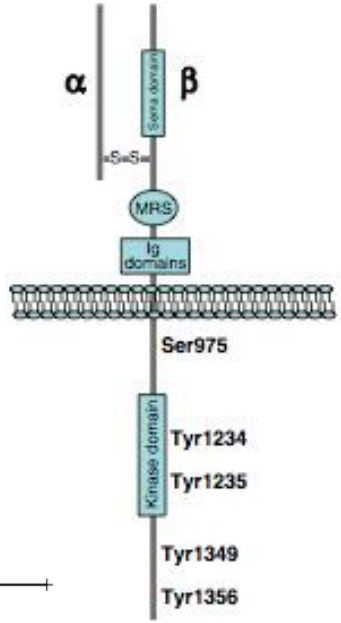
[Cancer Genome Atlas Research Network N Engl J Med. (‘16)]

Tyr-kinase MET: Known Facts & New Results

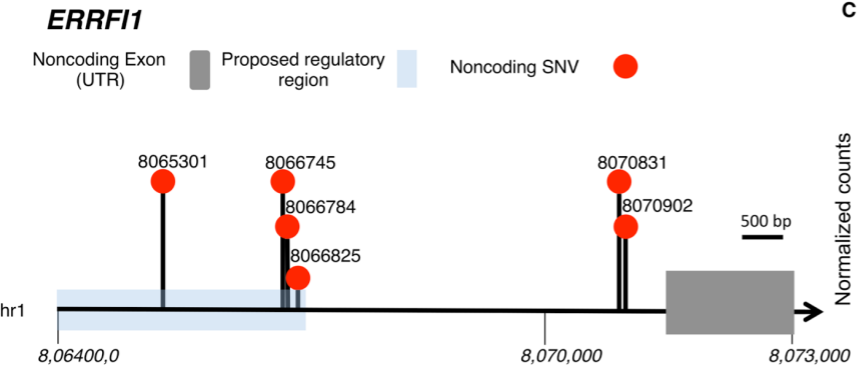
- MET is long known pRCC driver
- In MET, TCGA found somatic SNVs, duplications & an alt. splicing event as drivers (43/161).
- In addition, from 35 WGS we found
  - A noncoding hotspot associated with *MET*
  - Lack of SVs & breakpoints disrupting *MET*
  - Germline SNP (rs11762213) predicts survival in type 2 patients

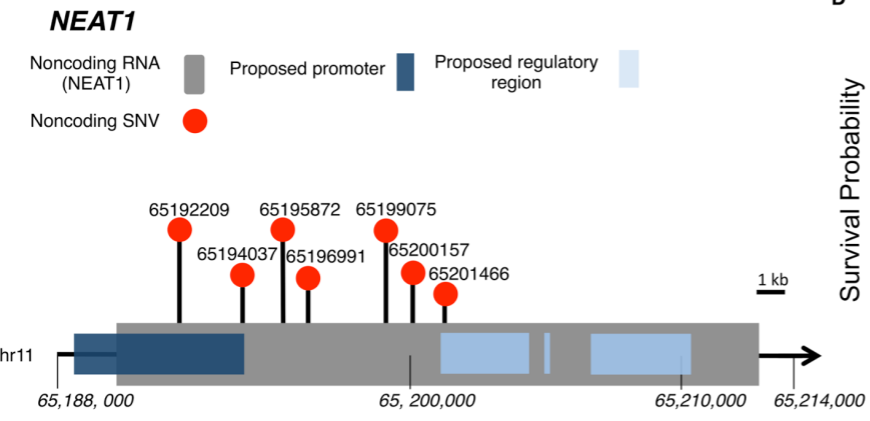


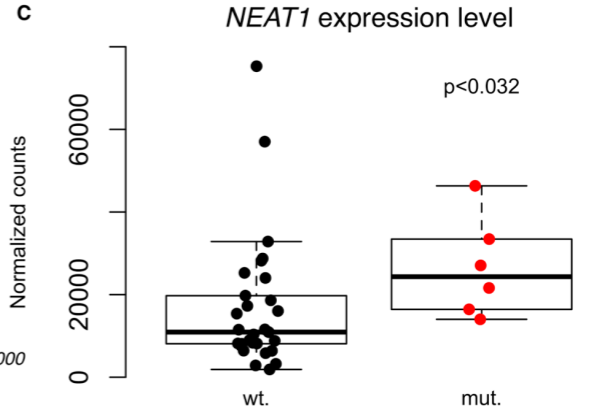


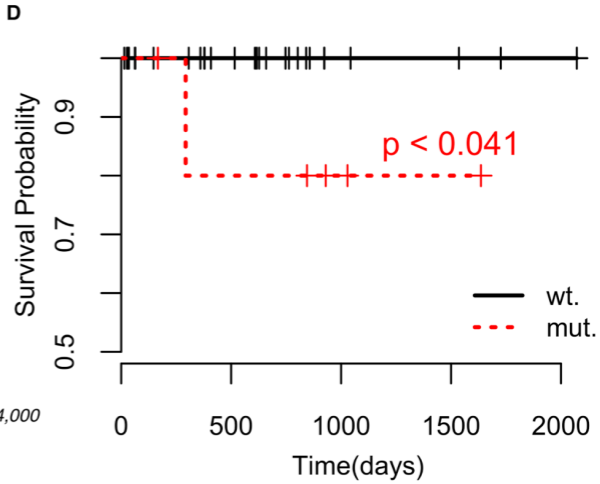


[A. Gentile, L. Trusolino and PM. Comoglio, Cancer and Metastasis Reviews ('08); S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]







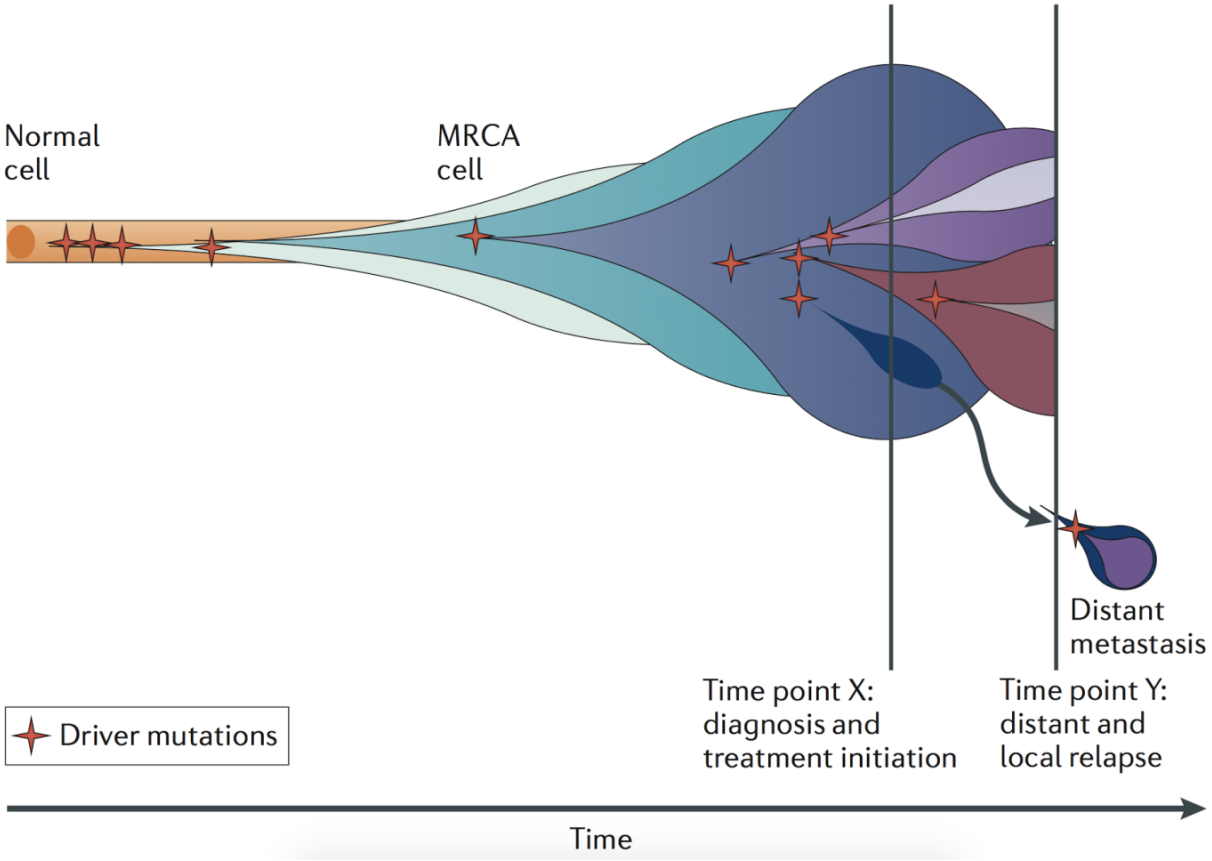


Beyond *MET*: 2 non-coding hotspots in NEAT & ERRFI1,

supported by expr. changes & survival analysis

[Li et al. PLOS Genetics ('17)]

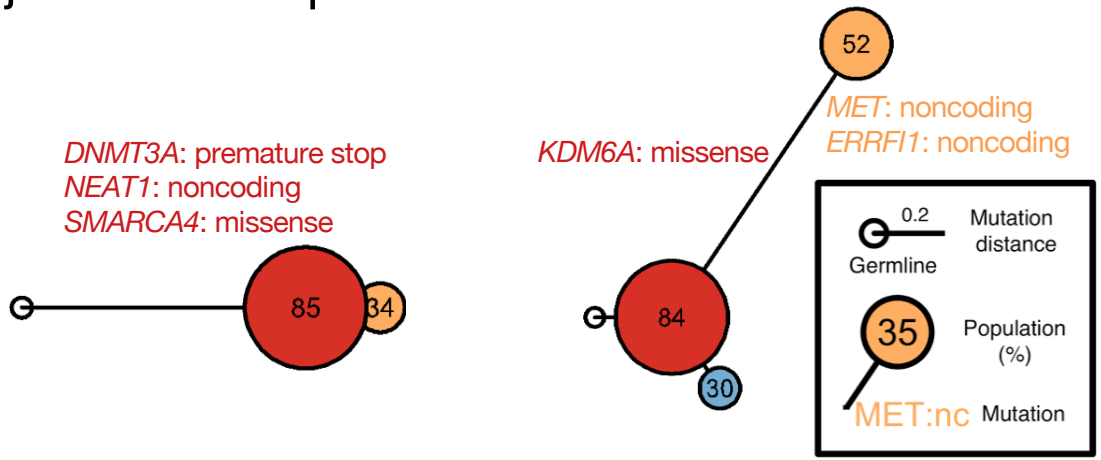
Tumor Evolution: Highlight the Ordering of Key Mutations



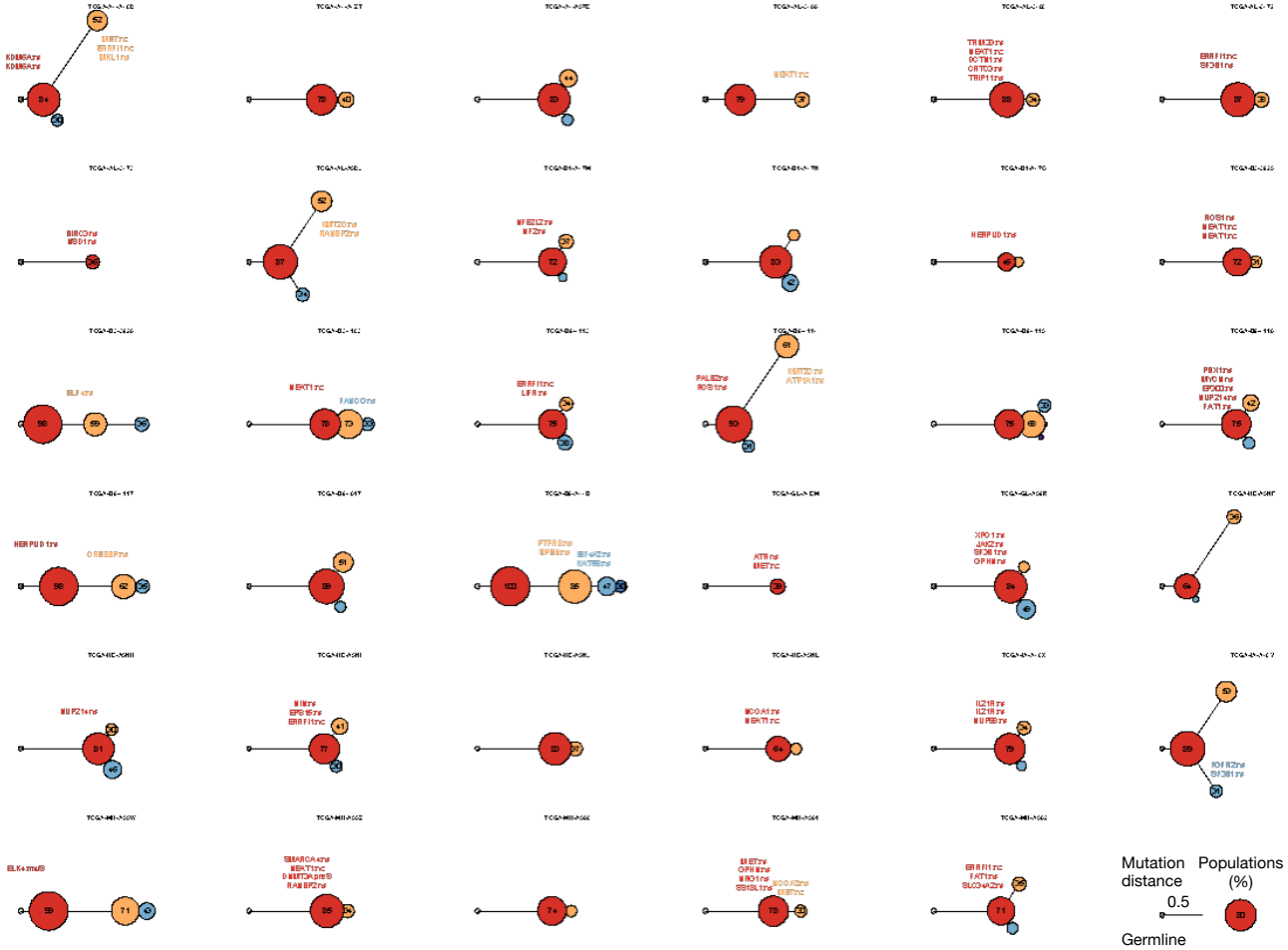
Yates et al, NRG (2012)

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



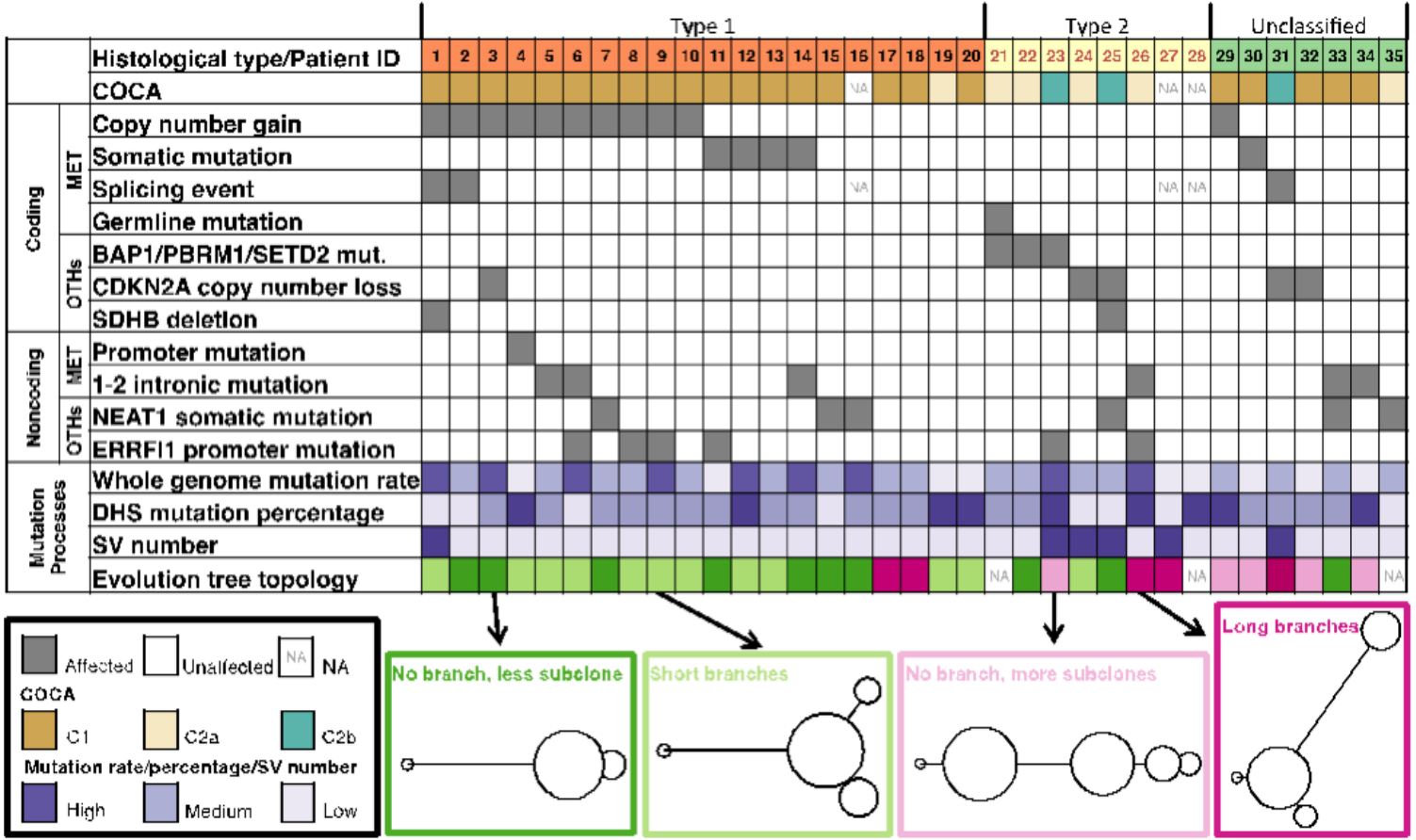
[S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]



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

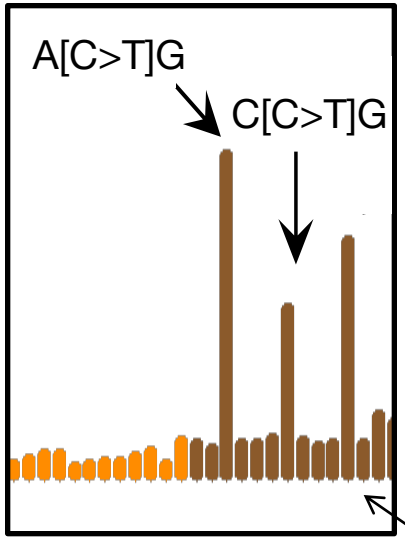
[S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]

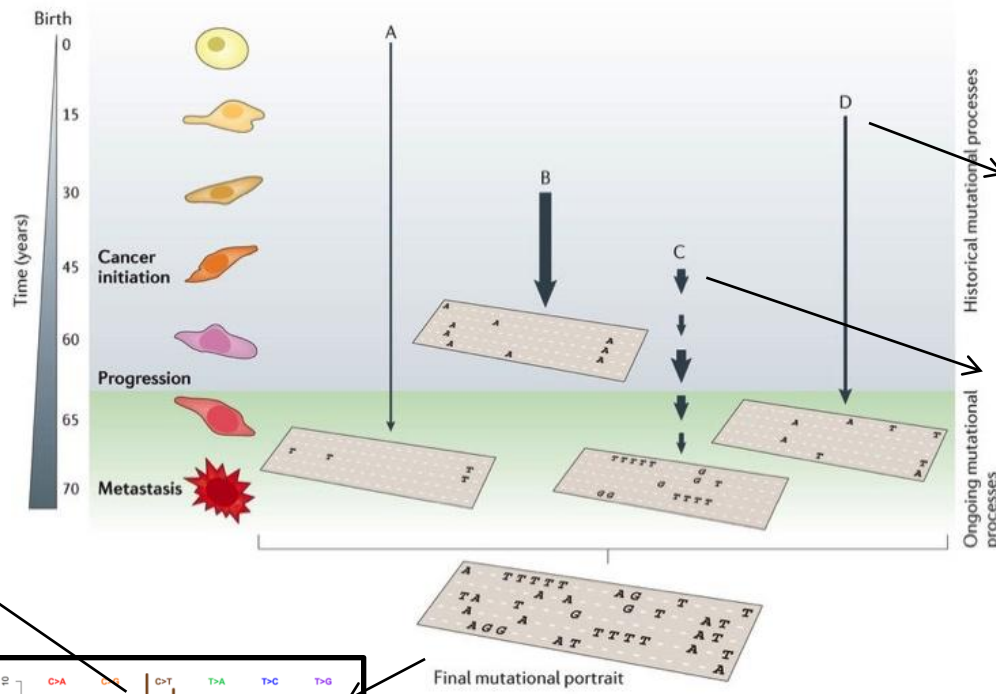
Tree topology correlates with molecular subtypes

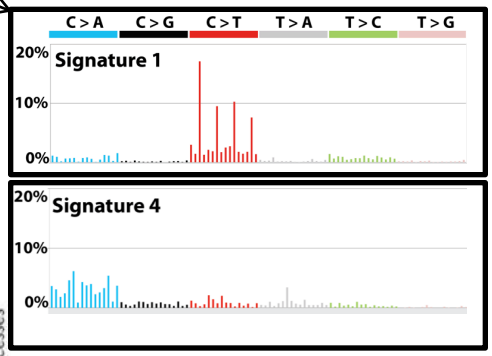


[Li et al., PLOS Genetics ('17)]

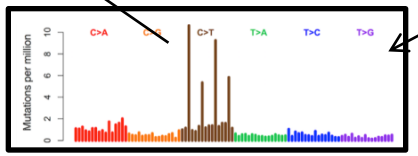
Mutational processes carry context-specific signatures







S: Mutation signature
inferred



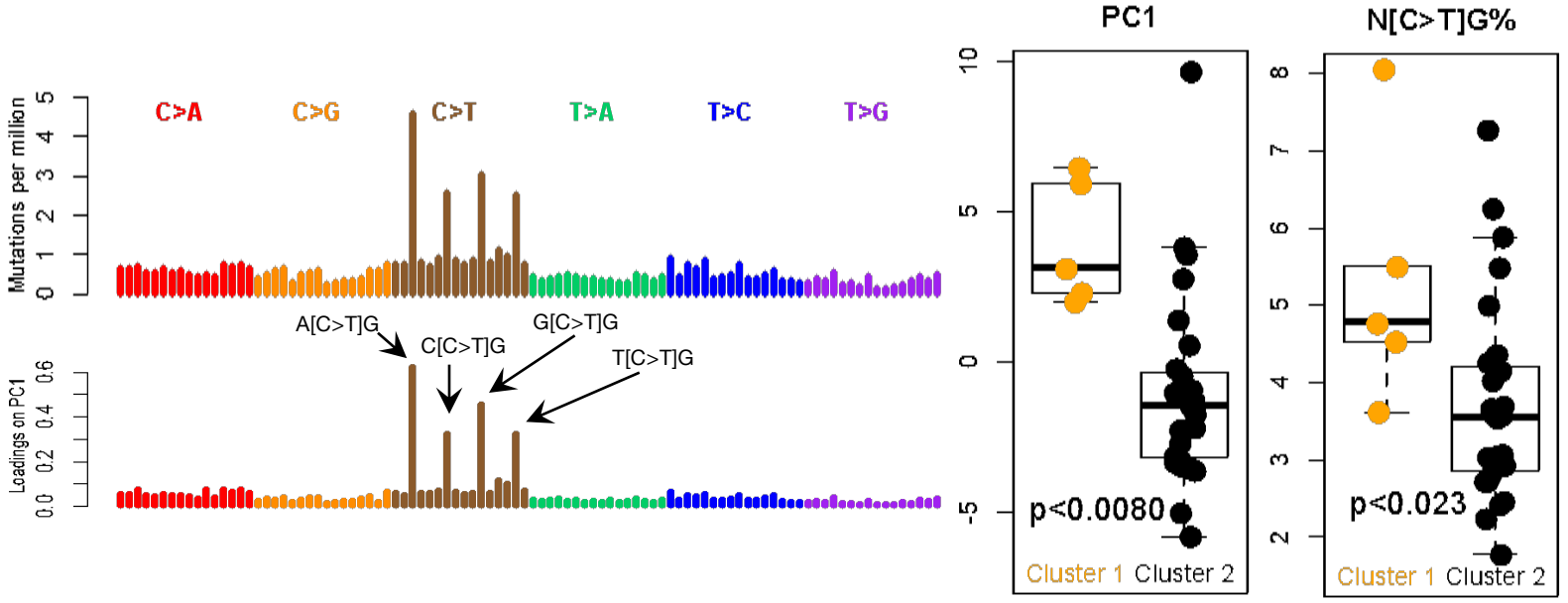
M: Mutation spectrum
observed

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

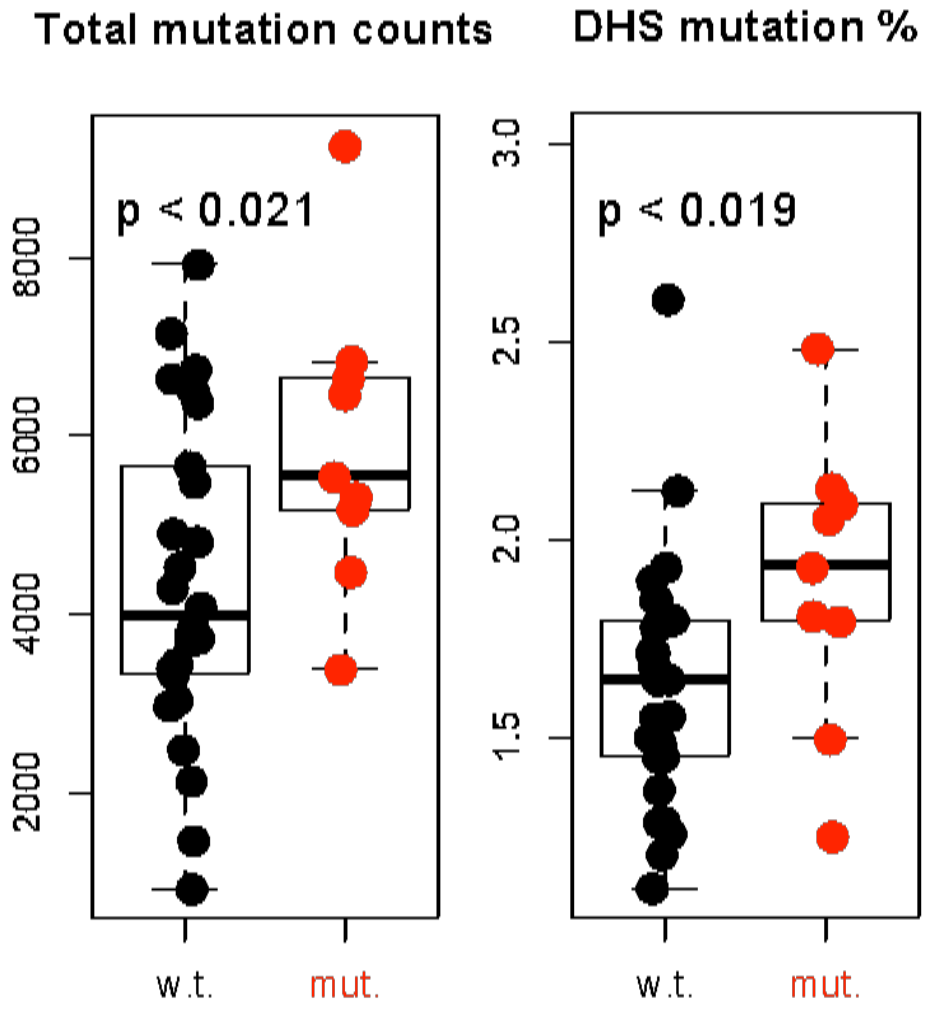
[T. Helleday, S. Eshtad & S. Nik-Zainal,
Nature Reviews Genetics (’14), L. Alexandrov et al., Nature (‘13)]

CpGs drive inter-patient variation in pRCC mutational spectra

- The loadings on PC1 are mostly [C>T]G
- Confirmed by higher C>T% in CpGs in the hypermethylated group (cluster1)



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



Key mutation affects mutational landscape which, in turn, affects overall burden in pRCC

- Chromatin remodeling defect ("mut") leads to more mutations in open chromatin (raw number & fraction) in those pRCC cases w/ the mutation

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

Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
- Recurrence: Statistics for driver identification
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
- Recurrence #2: (Low-power) application to **pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways

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

github.com/gersteinlab/**Frustration**
S **Kumar**, D Clarke

github.com/gersteinlab/**MrTADfinder**
KK **Yan**, S Lou

VAT.gersteinlab.org
L **Habegger**, S Balasubramanian, DZ Chen, E Khurana, A Sboner, A Harmanci, J Rozowsky, D Clarke, M Snyder

ALoFT.gersteinlab.org
S **Balasubramanian**, Y **Fu**, M Pawashe, P McGillivray, M Jin, J Liu, K Karczewski, D MacArthur

FunSeq.gersteinlab.org
Y **Fu**, E **Khurana**, Z Liu, S Lou, J Bedford, X Mu, K Yip

pRCC - S **Li**, B Shuch

MOAT.gersteinlab.org - L **Lochovsky**, J **Zhang**

CostSeq2 - P **Muir**, S Li, S Lou, D Wang, DJ Spakowicz, L Salichos, J Zhang, GM Weinstock, F Isaacs, J Rozowsky

LARVA.gersteinlab.org
L **Lochovsky**, J **Zhang**, Y Fu, E Khurana

github.gersteinlab.org/**uORFs**
P **McGillivray**, R Ault, M Pawashe, R Kitchen, S Balasubramanian


STOP

Info about this talk

No Conflicts

Unless explicitly listed here. There are no conflicts of interest relevant to the material in this talk

General PERMISSIONS

- This Presentation is copyright Mark Gerstein, Yale University, 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