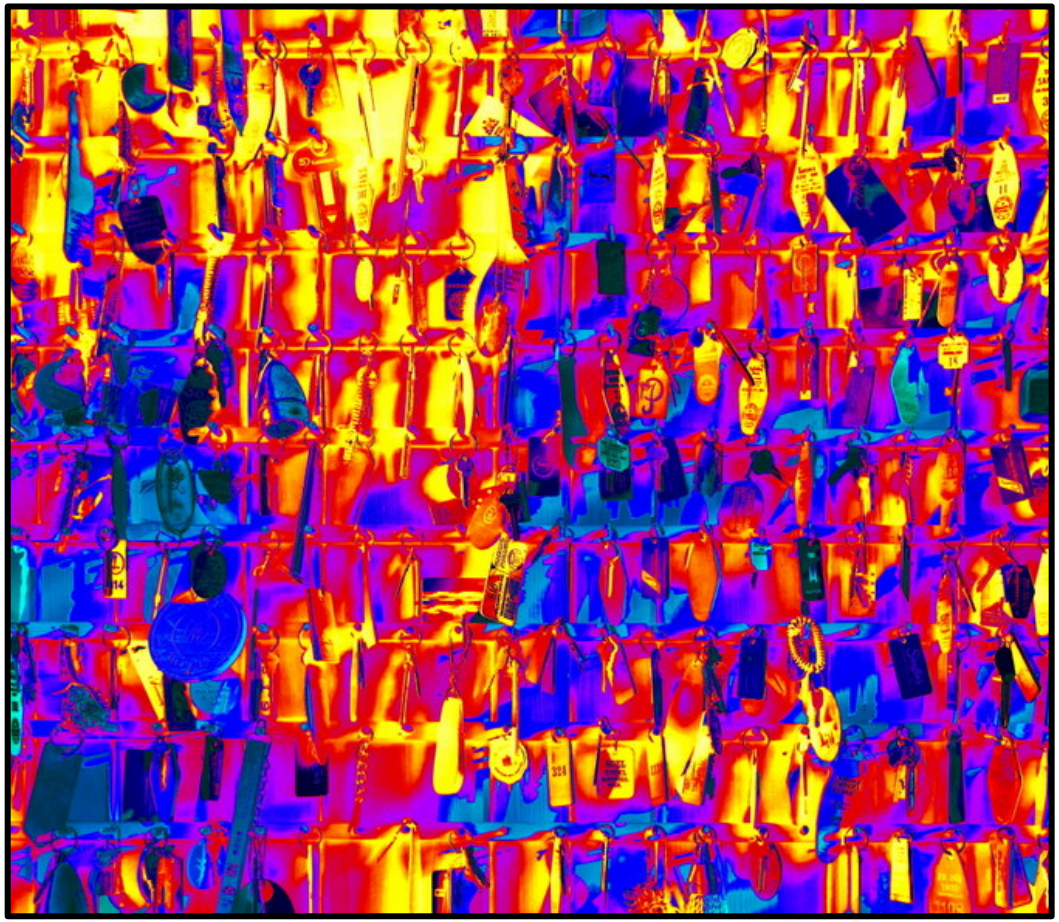


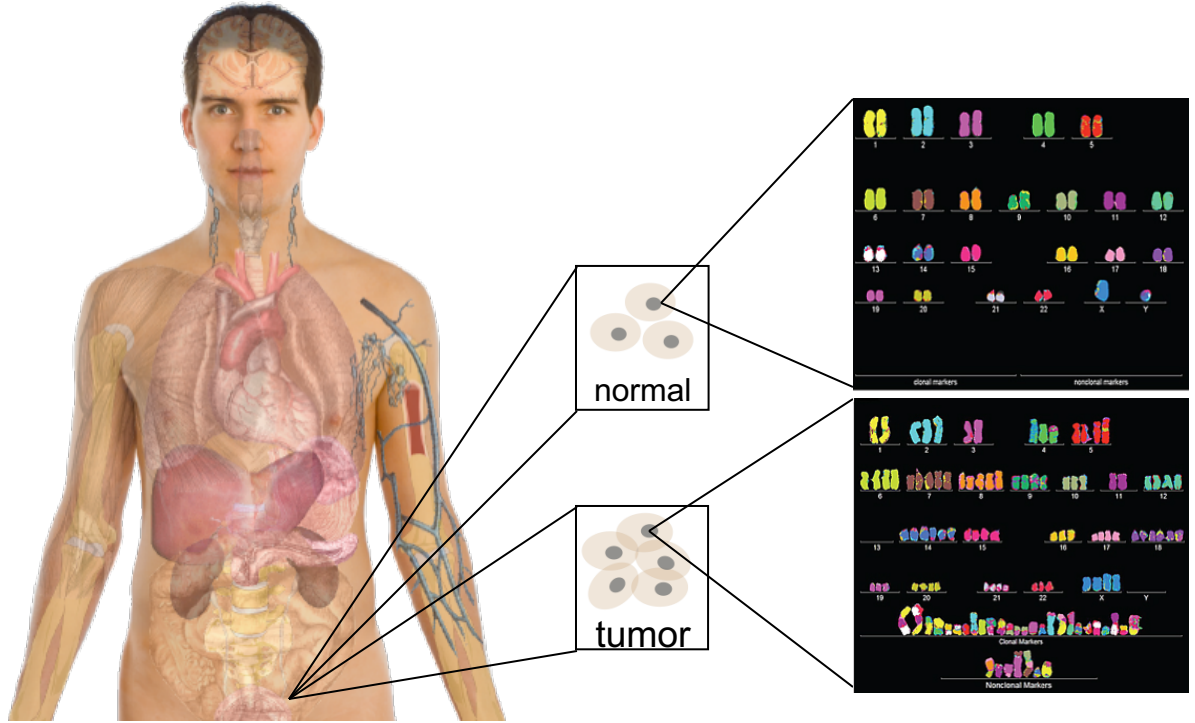
Prioritizing Variants in Personal Genomes: Using functional impact, with particular application to cancer

Mark Gerstein
Yale

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

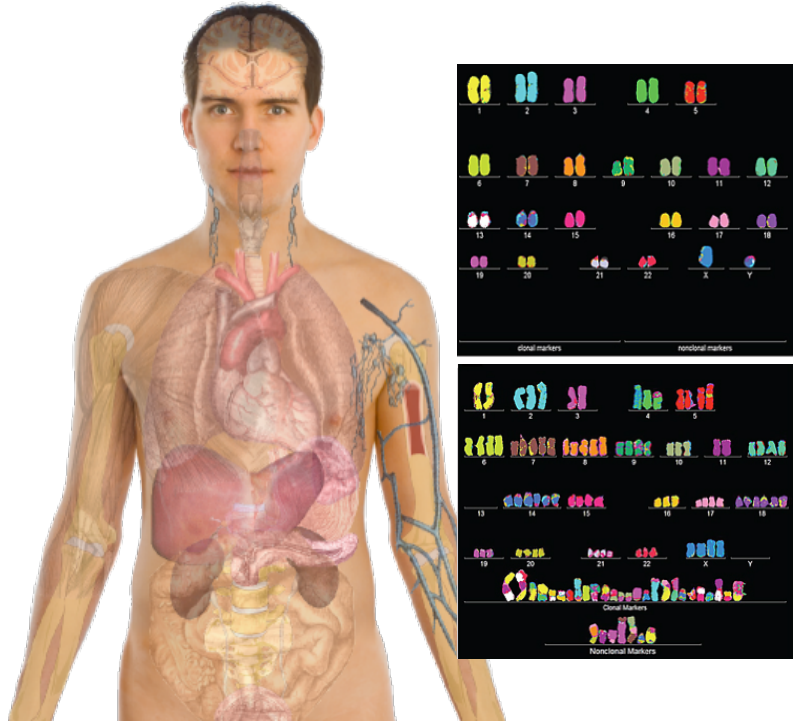
Personal Genomics as a Gateway into Biology

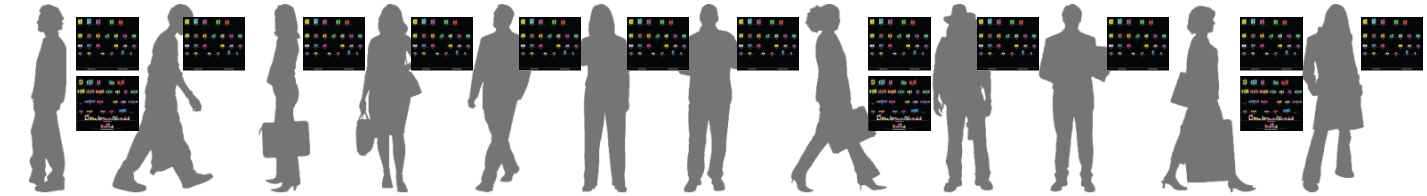
Personal genomes will soon become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.

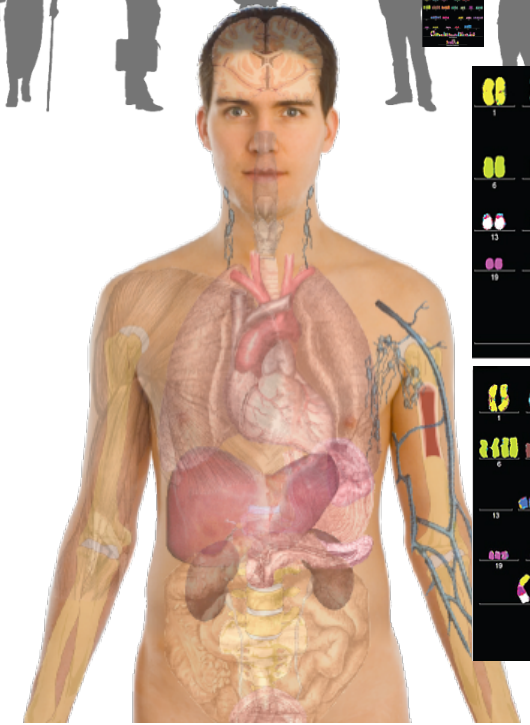


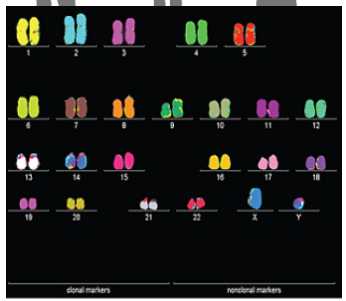
Personal Genomics as a Gateway into Biology

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**

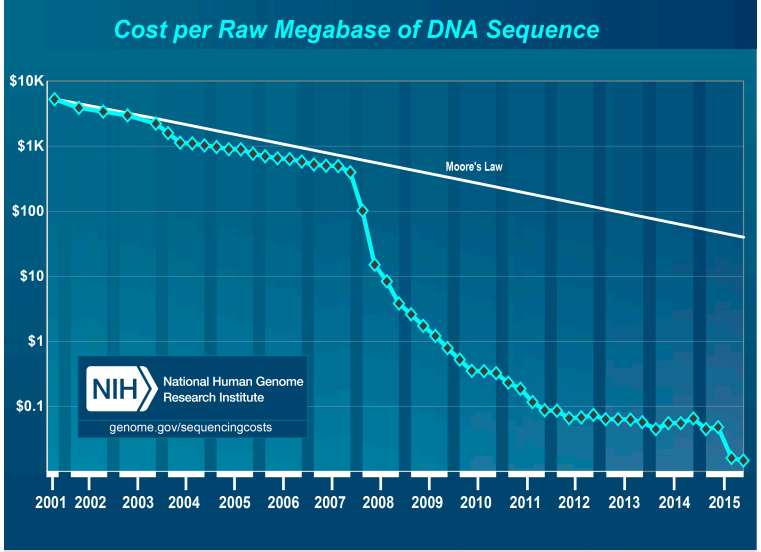
Scaling DBs to the **population**

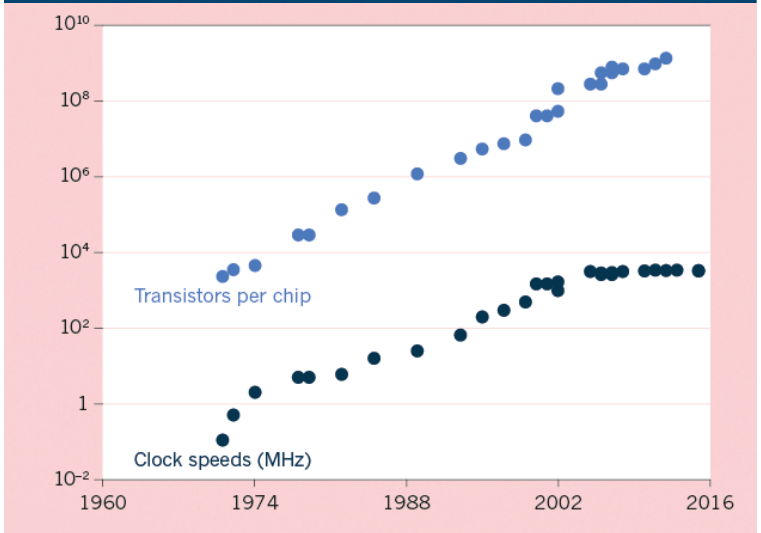
Identifying **key variants** - separating into rare, recurrent, common, &c

The Scaling of Genomic Data Science:

Powered by exponential increases in data & computing

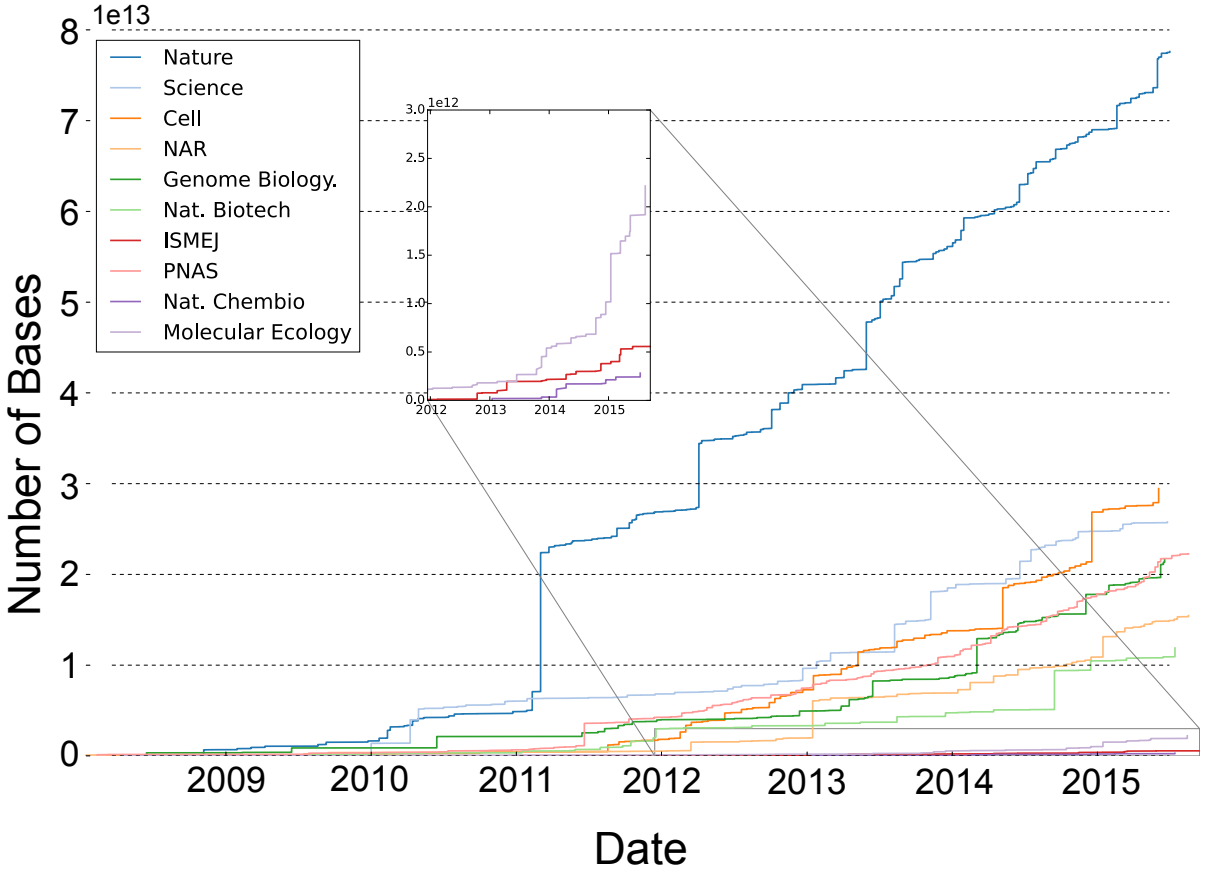
(Moore's Law)





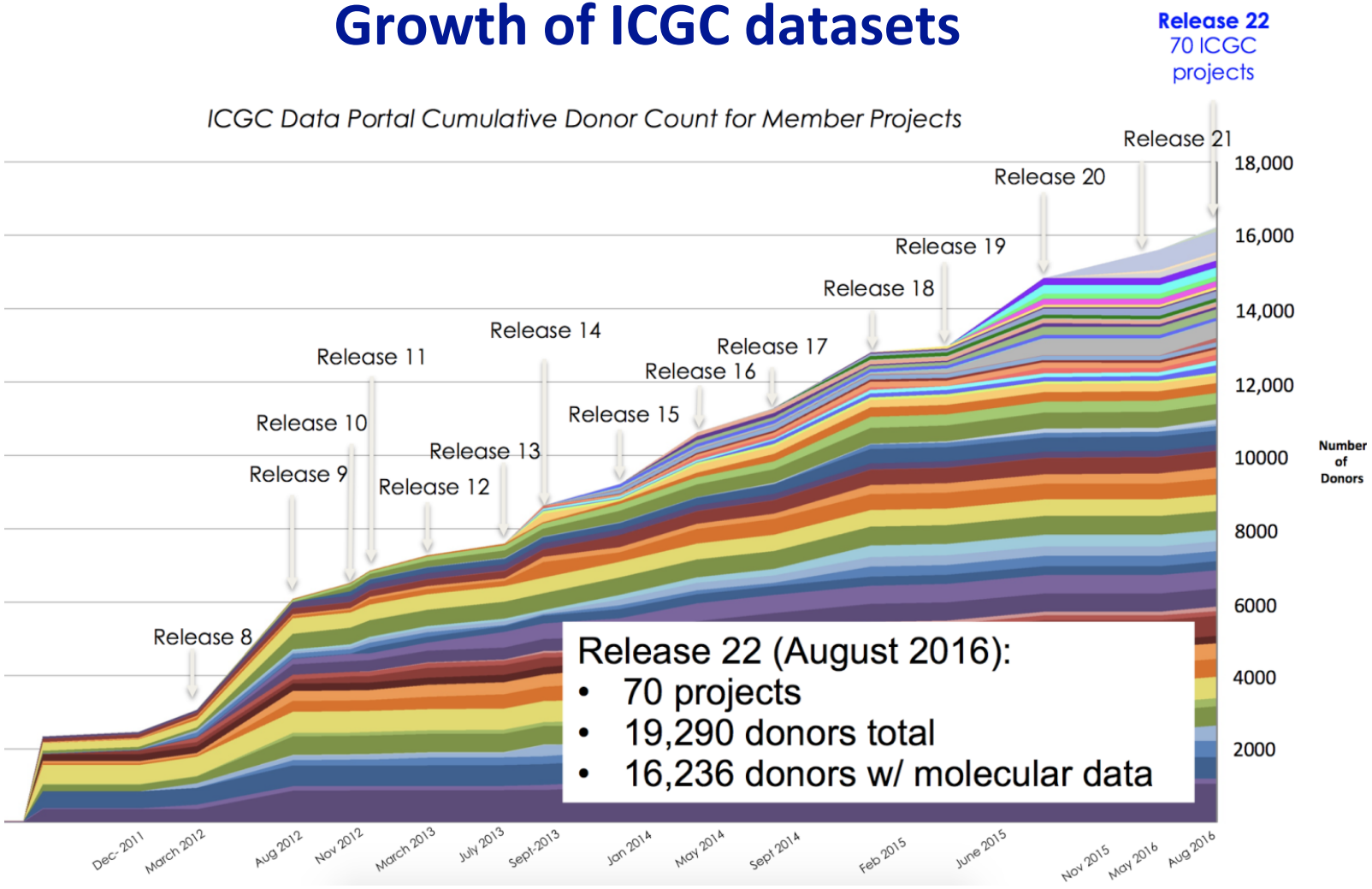
[NHGRI website + Waldrop ('15) Nature]

Exponential Scaling Changes Fields Using Genomic Data

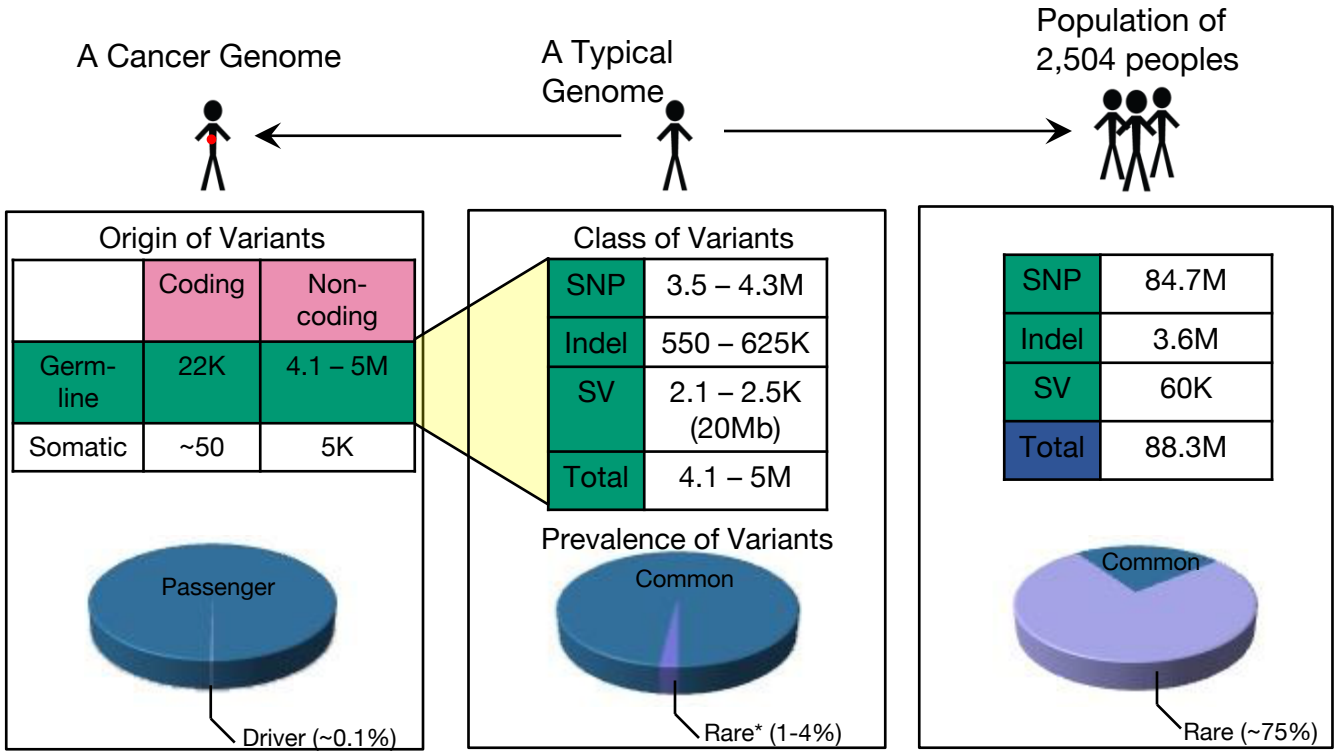


[Muir et al. ('15) GenomeBiol.]

Growth of ICGC datasets



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.

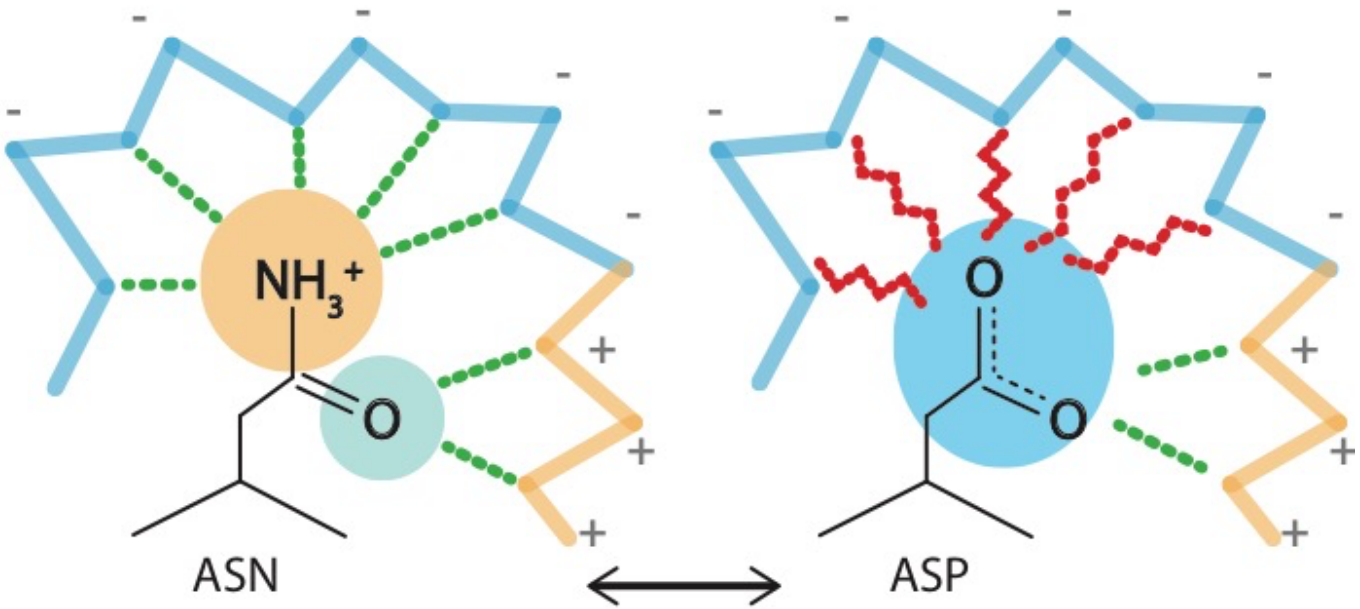
Vogelstein B. Science 2013. 339(6127):1546-1558

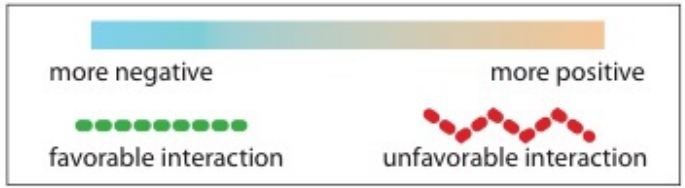
Prioritizing Variants in Personal Genomes: Using functional impact, with particular application to cancer

- 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
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
  - **RADAR:** prioritize variants based on post-transcriptional regulome using ENCODE eCLIP
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation
- (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 15

Prioritizing Variants in Personal Genomes: Using functional impact, 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
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
  - **RADAR:** prioritize variants based on post-transcriptional regulome using ENCODE eCLIP
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation
- (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 15

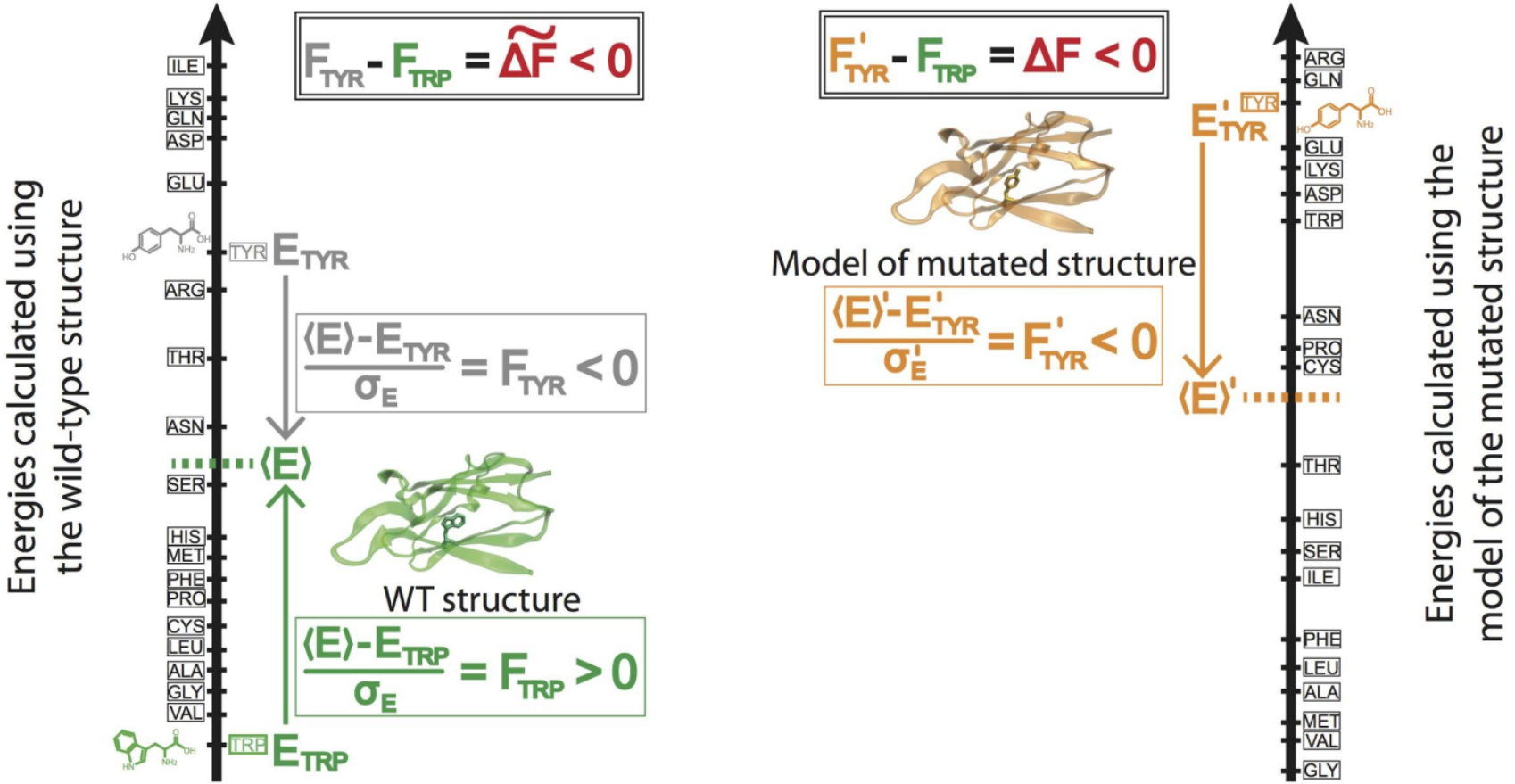




What is localized frustration ?

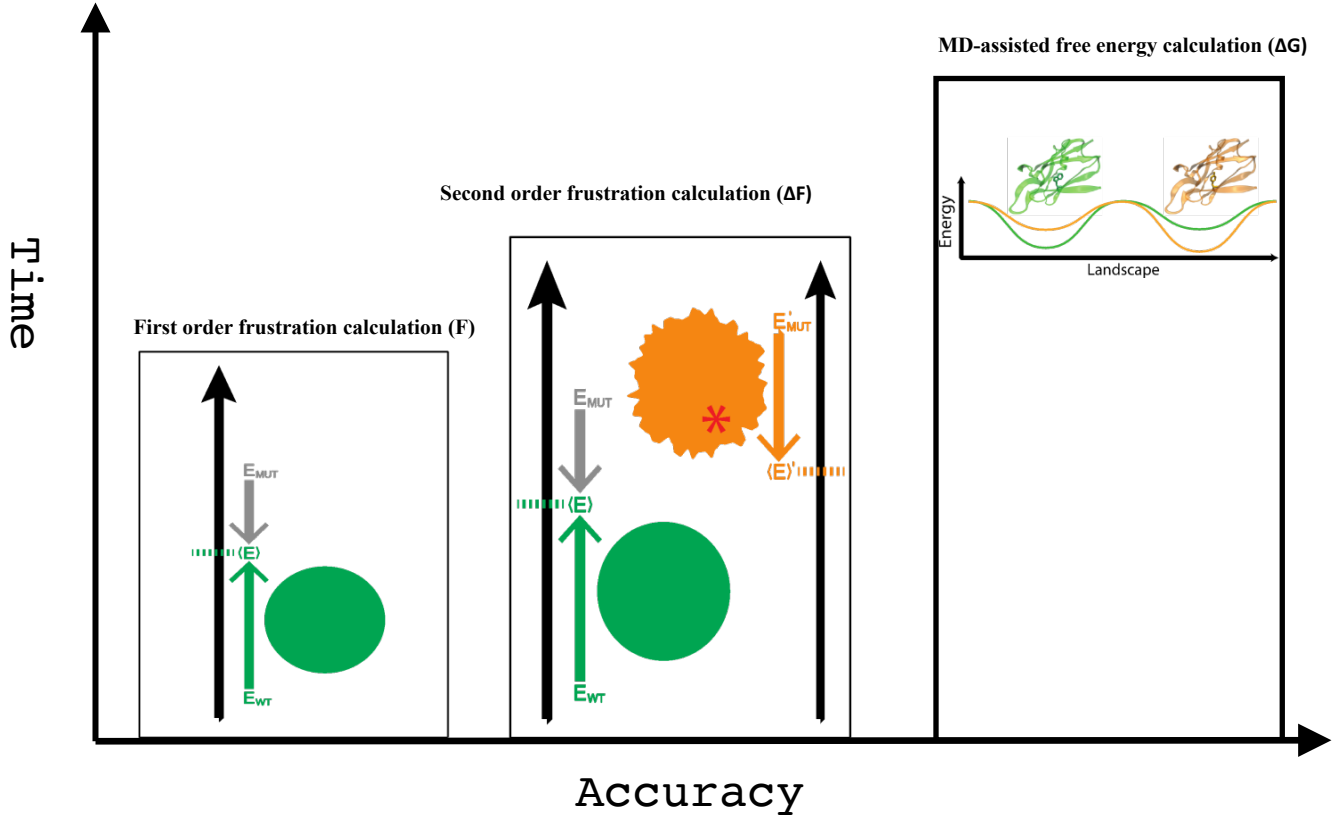
[Ferreiro et al., *PNAS* ('07)]

Workflow for evaluating localized frustration changes (ΔF)

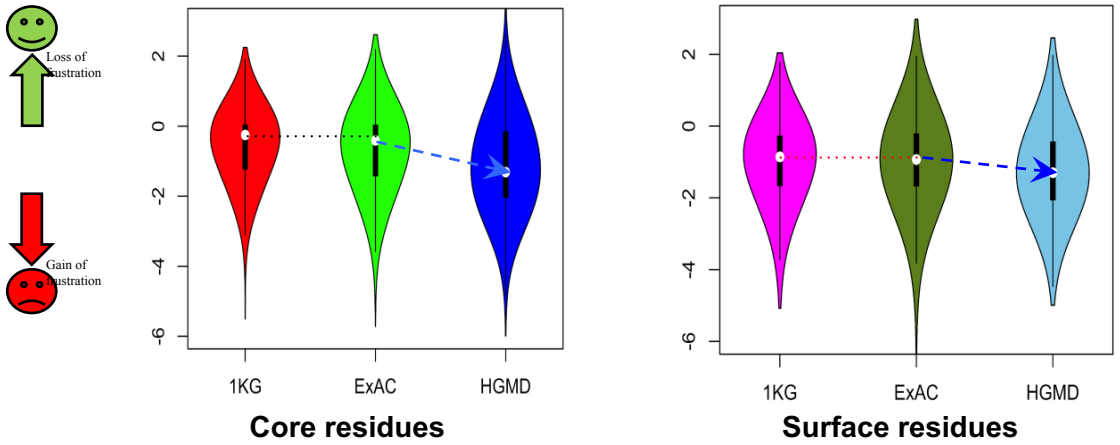


[Kumar et al. *NAR* (2016)]

Complexity of the second order frustration calculation



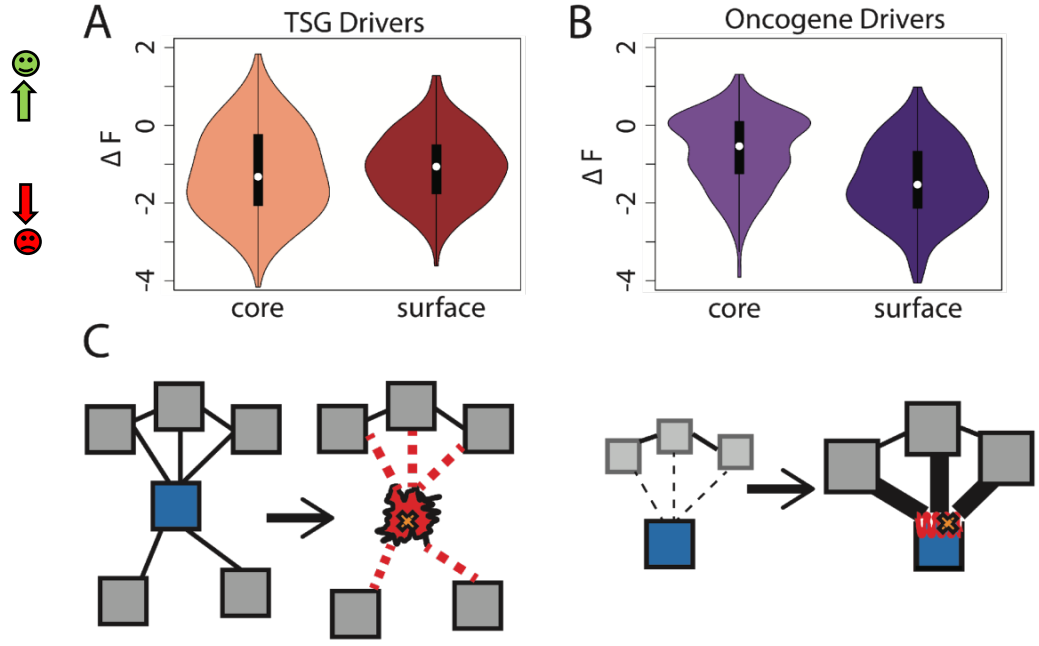
Comparing ΔF values across different SNV categories: disease v normal



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, 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
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
  - **RADAR:** prioritize variants based on post-transcriptional regulome using ENCODE eCLIP
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation
- (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 15

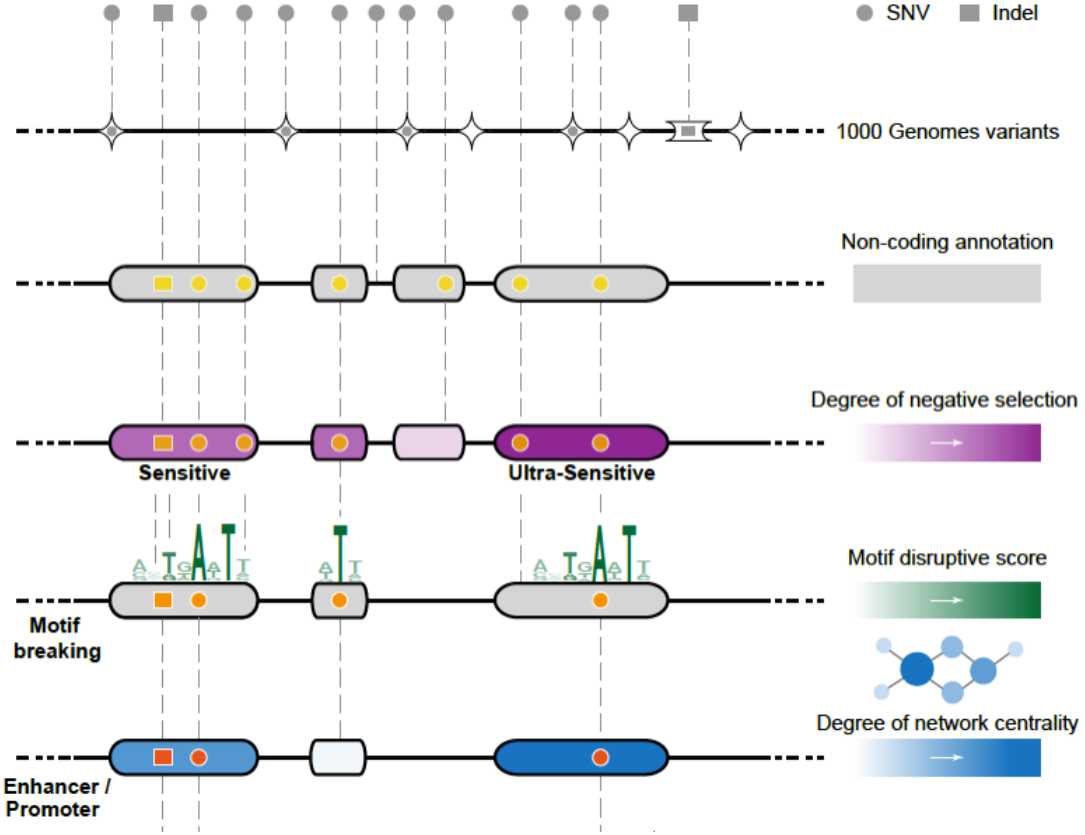
Funseq: a flexible framework to determine functional impact & use this to prioritize variants

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

Conservation (GERP, allele freq.)

Mutational impact (motif breaking, Lof)

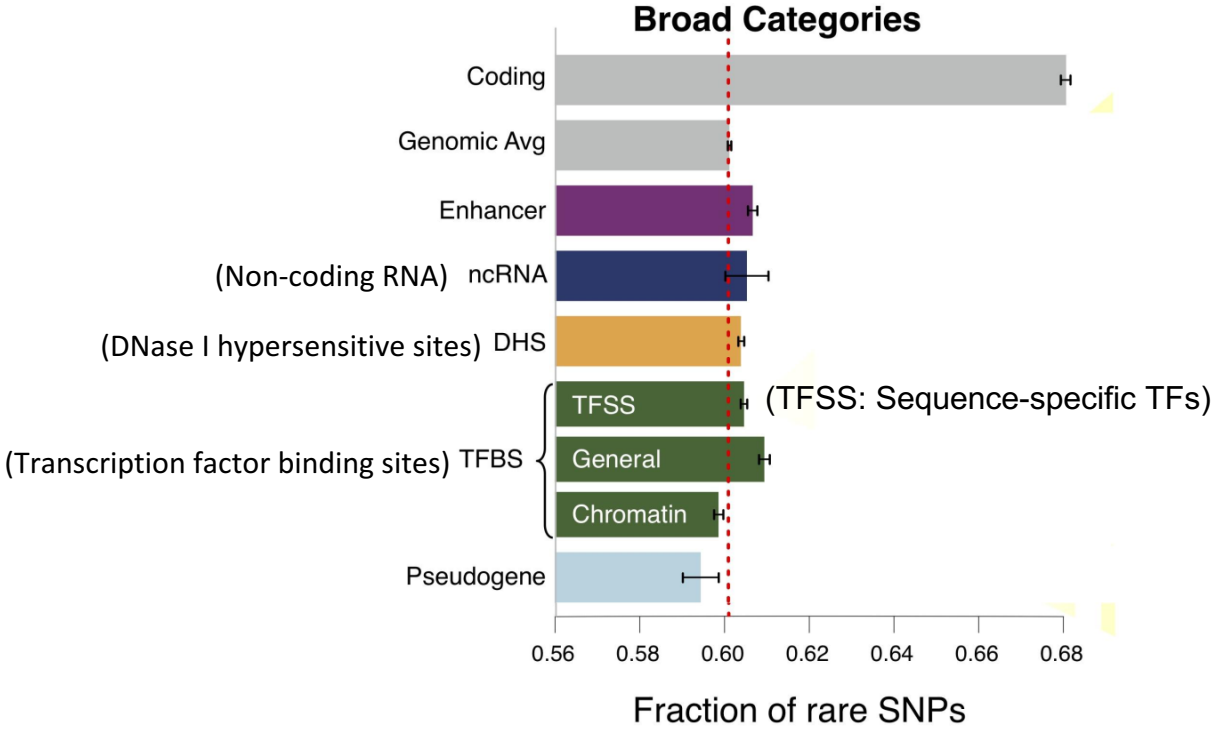
Network (centrality position)



[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

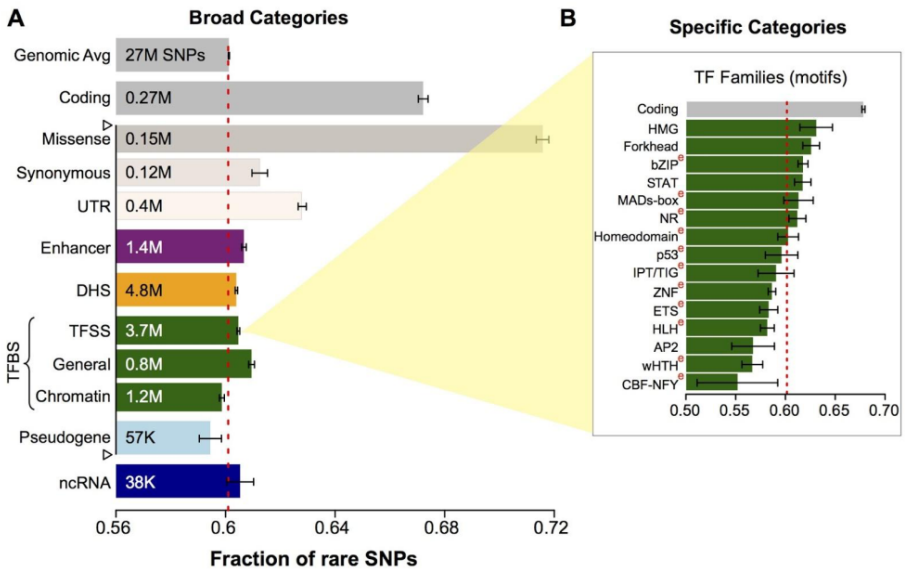


Depletion of Common Variants in the Human Population

Broad categories of regulatory regions under negative selection
Related to:

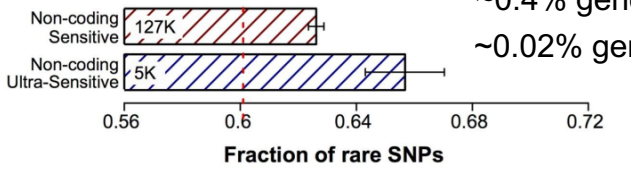
ENCODE, *Nature*, 2012
Ward & Kellis, *Science*, 2012
Mu et al, *NAR*, 2011

Differential selective constraints among specific sub-categories



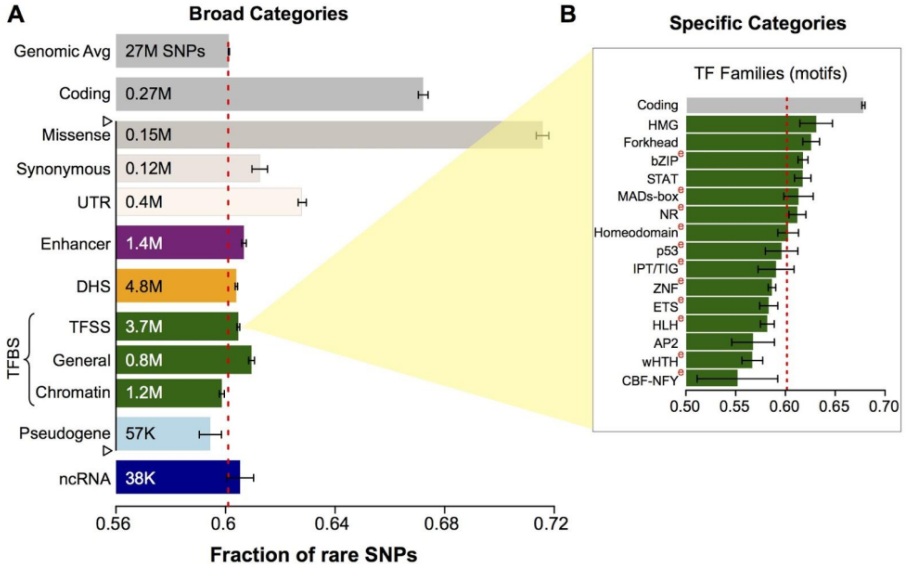
Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

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



~0.4% genomic coverage (~ top 25)

~0.02% genomic coverage (top 5)

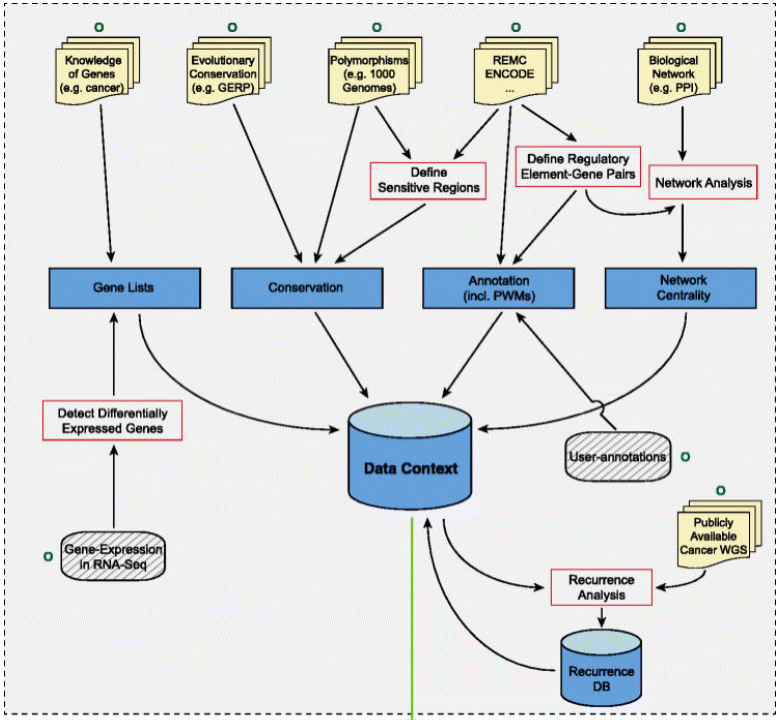


Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

Defining Sensitive non-coding Regions

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

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



FunSeq.gersteinlab.org

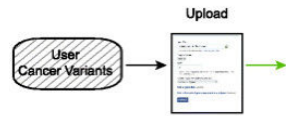
HOT region
Sensitive region
Polymorphisms

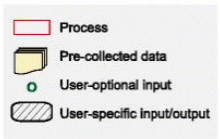
Genome

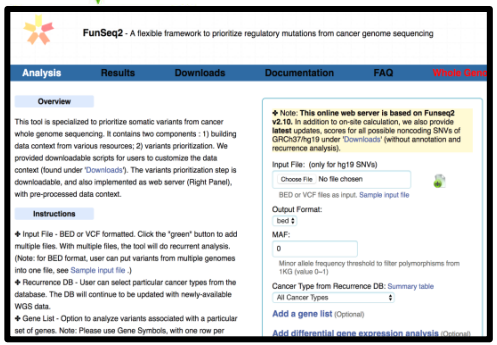
$$w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$$

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

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

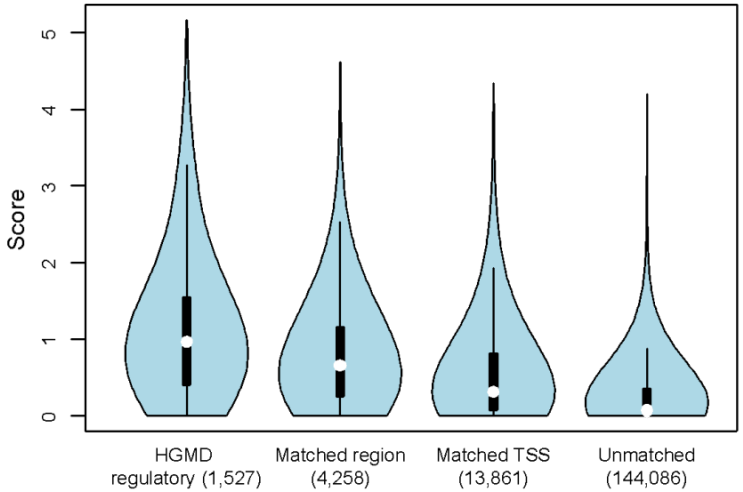


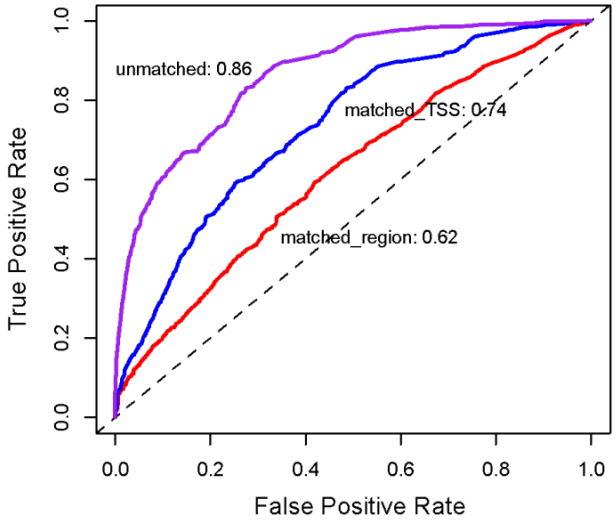




[Fu et al., GenomeBiology ('14)]

Germline pathogenic variants show higher core scores than controls





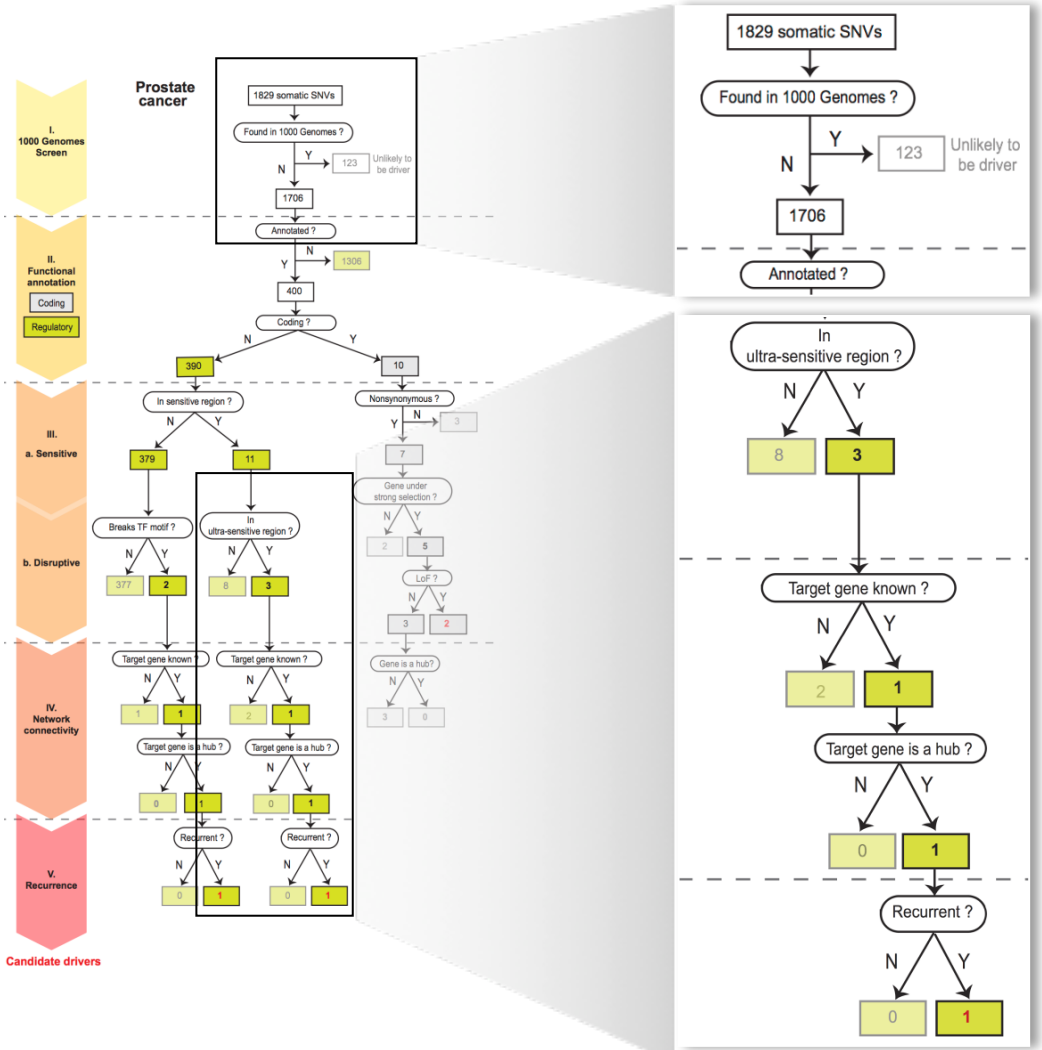
3 controls with natural polymorphisms (allele frequency >= 1%)

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)

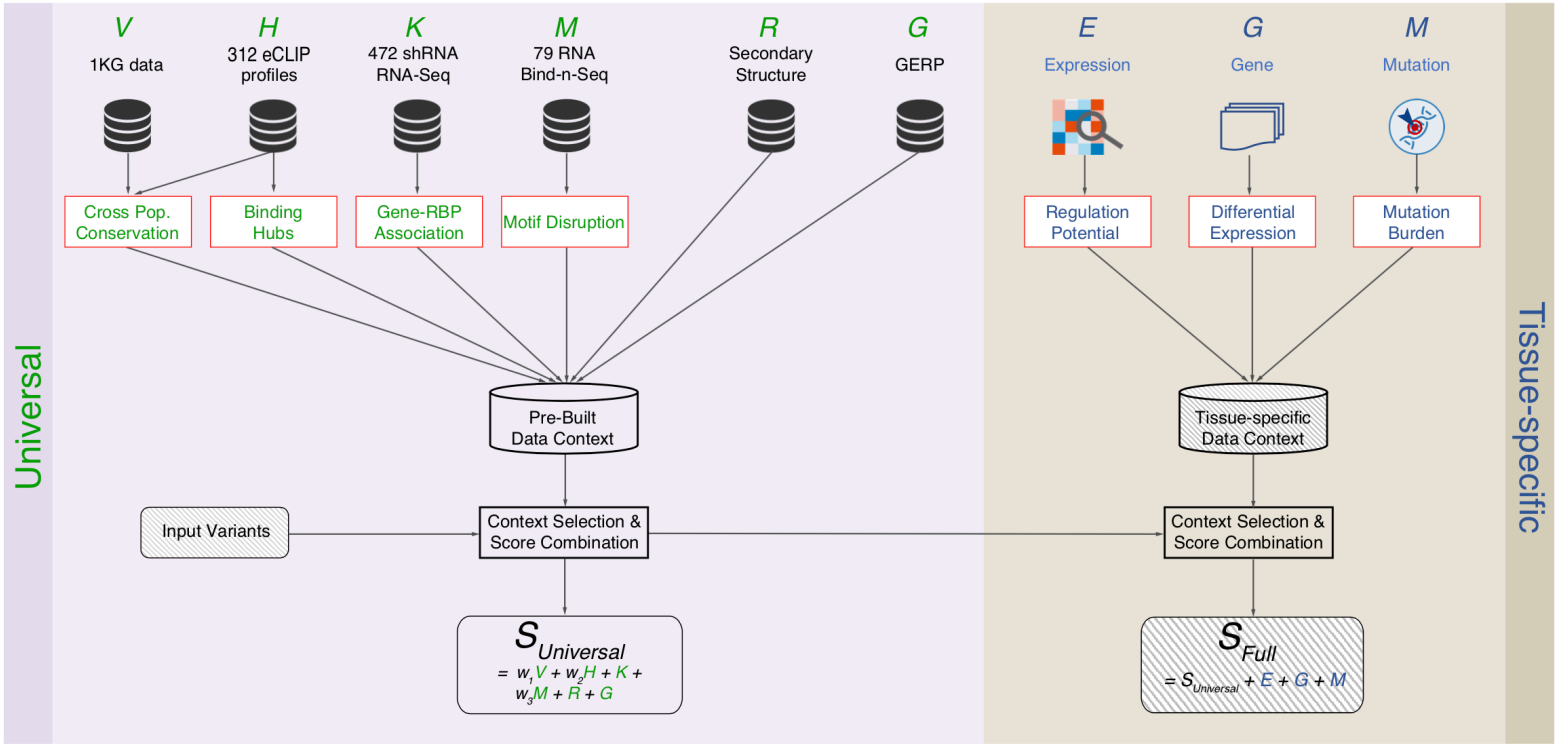


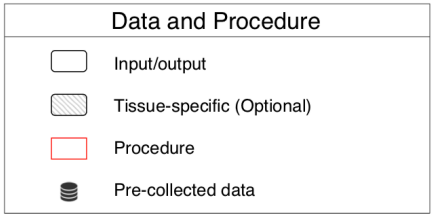
[Khurana et al., *Science* ('13)]

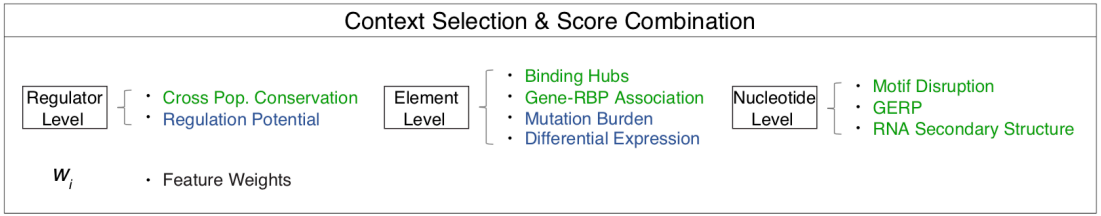
Prioritizing Variants in Personal Genomes: Using functional impact, with particular application to cancer

- 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
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
  - **RADAR:** prioritize variants based on post-transcriptional regulome using ENCODE eCLIP
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation
- (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 15

Schematic of RADAR Scoring

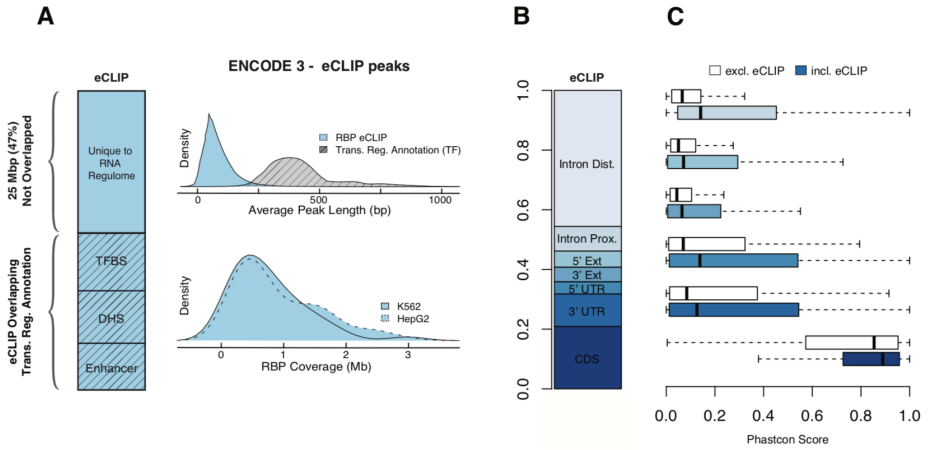




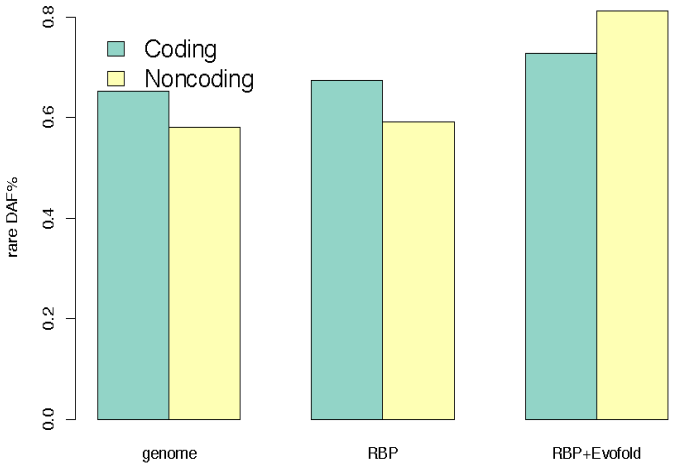


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

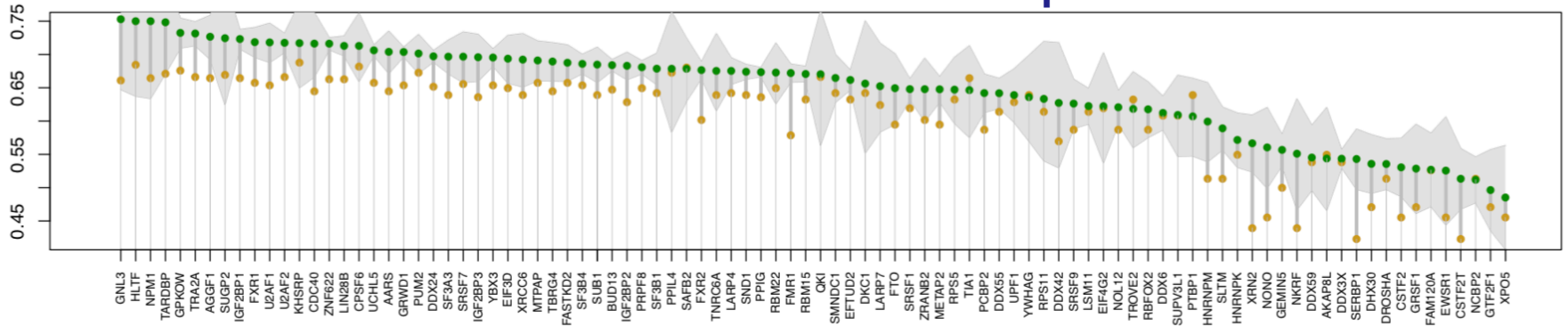
Summary of eCLIP and Phastcon



RNA Structure Cons. from Evofold

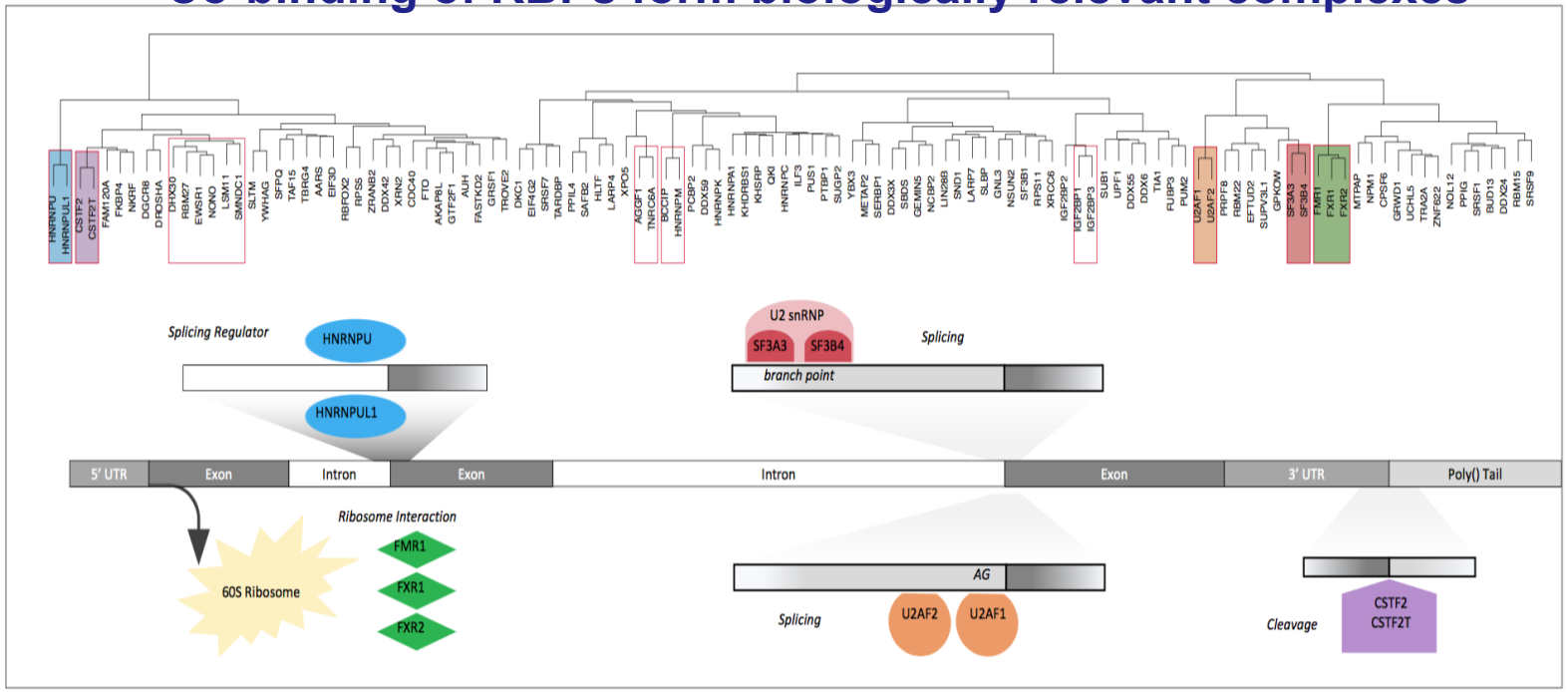


Enriched rare DAF in eCLIP peaks

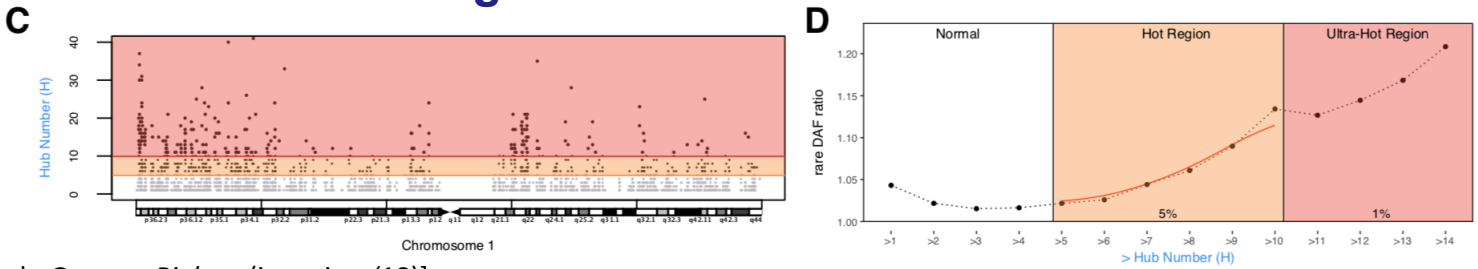


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

Co-binding of RBPs form biologically relevant complexes

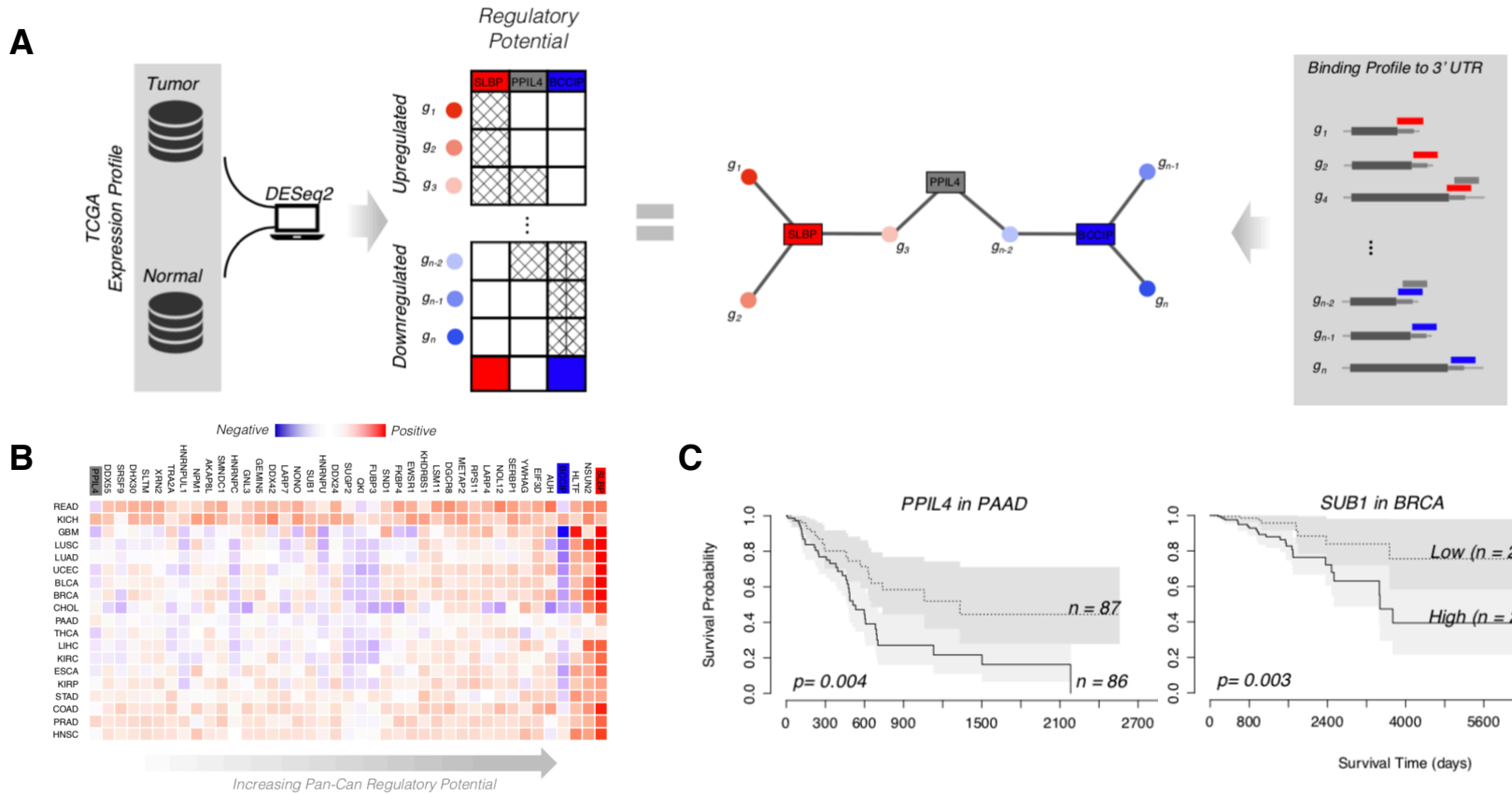


Binding hubs are enriched for rare variants



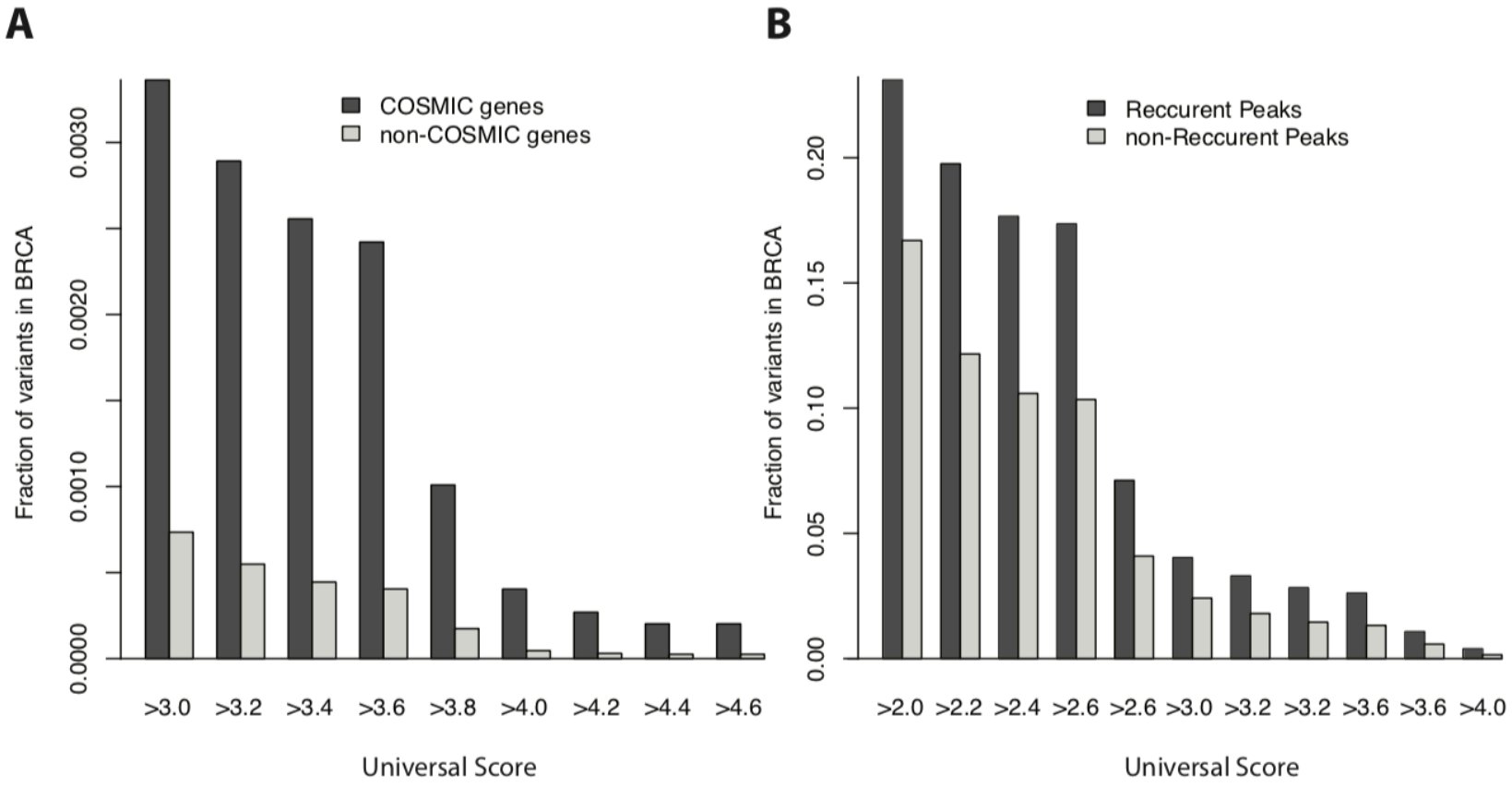
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

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

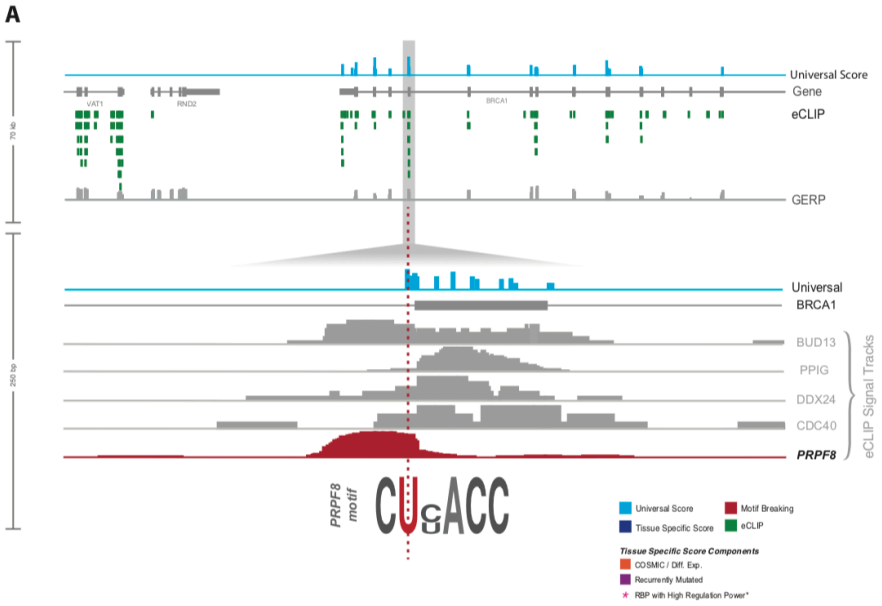


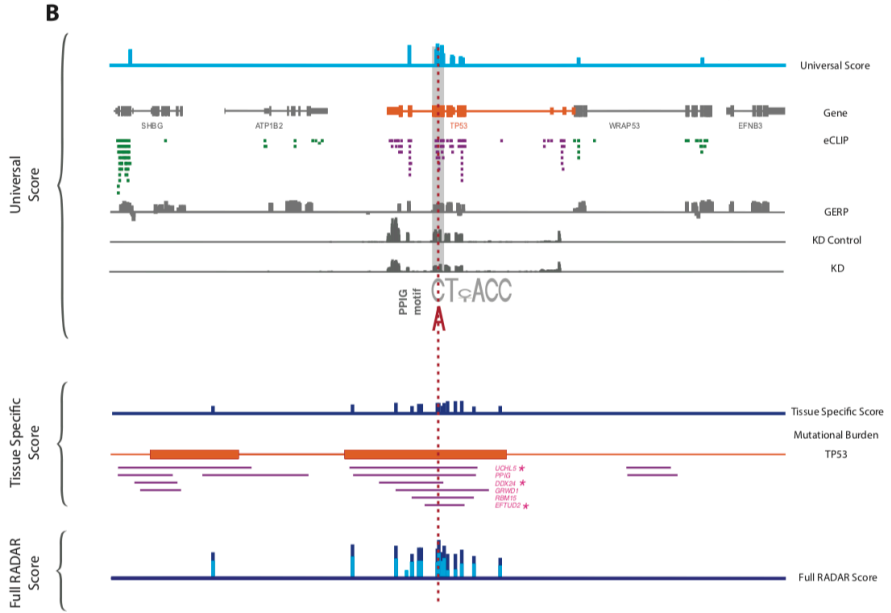
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

RADAR Scores enriched in COSMIC genes and recurrently mutated regions



Visualization of RADAR Features and Scoring





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

RADAR can be run on the command line by following the instructions on the Docs page or through the web here. Running RADAR through the website will print the results after several moments. You can try running RADAR through the web form with a sample file with one variant. Alternatively, you may also input a list of variants into the form as text. If variants are provided in both file and text formats, the variant file will be scored and the text field will be ignored.

More details on the RADAR inputs can be found on the Docs page.

- Variants: a list of variants
  - BED file: a BED file containing the variants
  - Text format: type variants directly into a text box, lines may be tab- or space-delimited
- Cancer type: a TCGA cancer type, only needed if any tissue-specific scores are to be included.
- Tissue-specific scores: which tissue-specific scores should be included along with the universal scores for each variant.

Variants:

Choose File no file selected

E.g. chr1 13506 13507 G A

Cancer type:

Select a cancer

Tissue-specific scores:

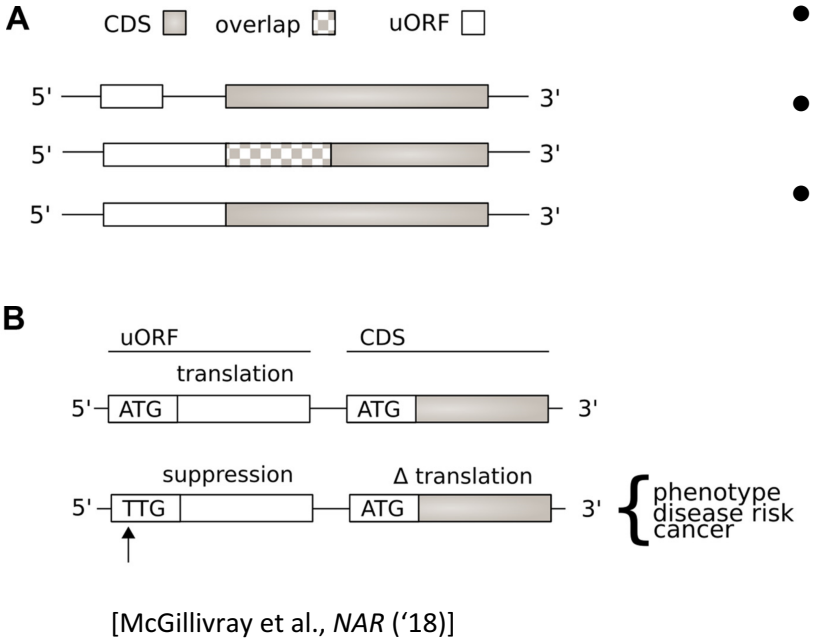
- [ ] Key genes
- [ ] Mutation recurrence
- [ ] RBP-regulation power

Score variants

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

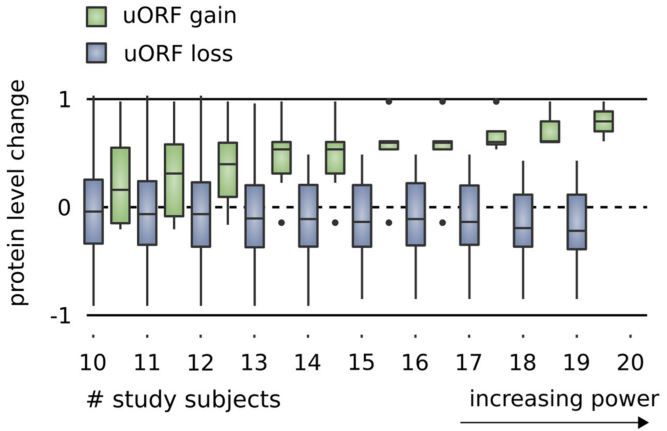
- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data gen. & processing
  - Big-data mining to prioritize key variants as drivers
- Functional impact #1: Coding
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
  - **RADAR:** prioritize variants based on post-transcriptional regulome using ENCODE eCLIP
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation
- (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 15

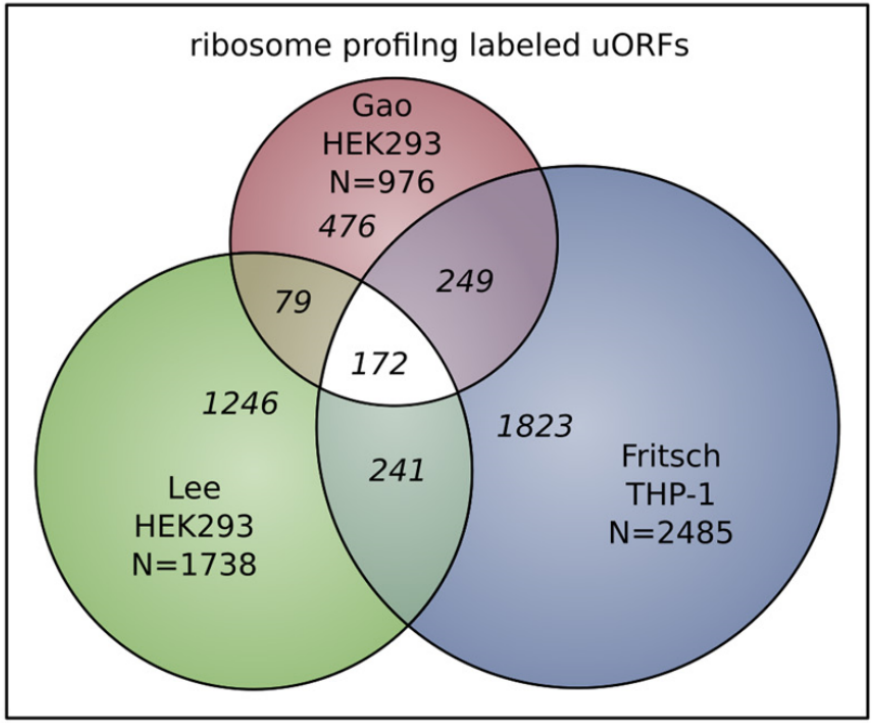
Upstream open reading frames (uORFs) regulate translation are affected by somatic mutation



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

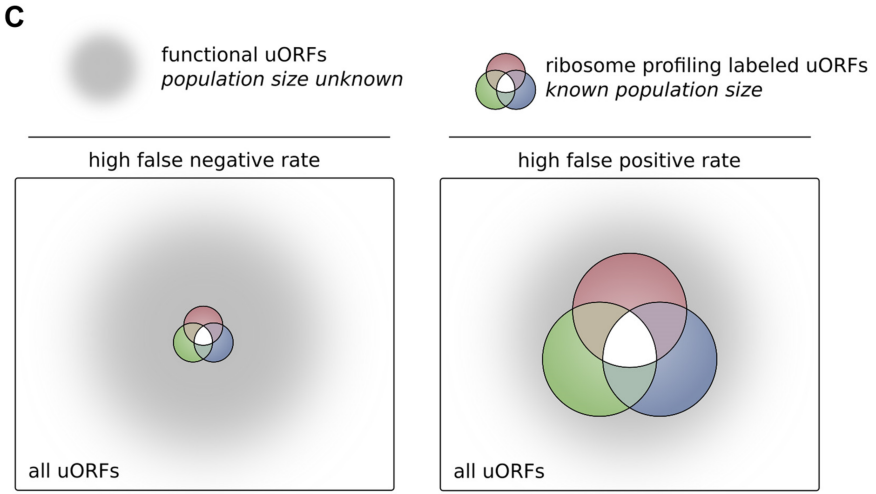
- uORFs regulate the translation of downstream coding regions.
- This regulation may be altered by somatic mutation in cancer.
- In Battle et al. 2014 data uORF gain & loss assoc. protein level change.





From a “Universe” of 1.3 M pot. uORFs

The population of functional uORFs may be significant

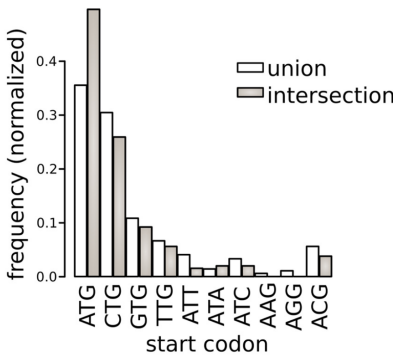


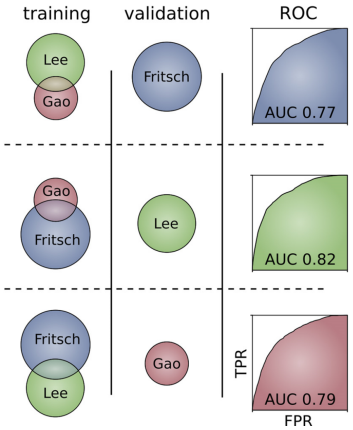
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

[McGillivray et al., *NAR* (‘18)]

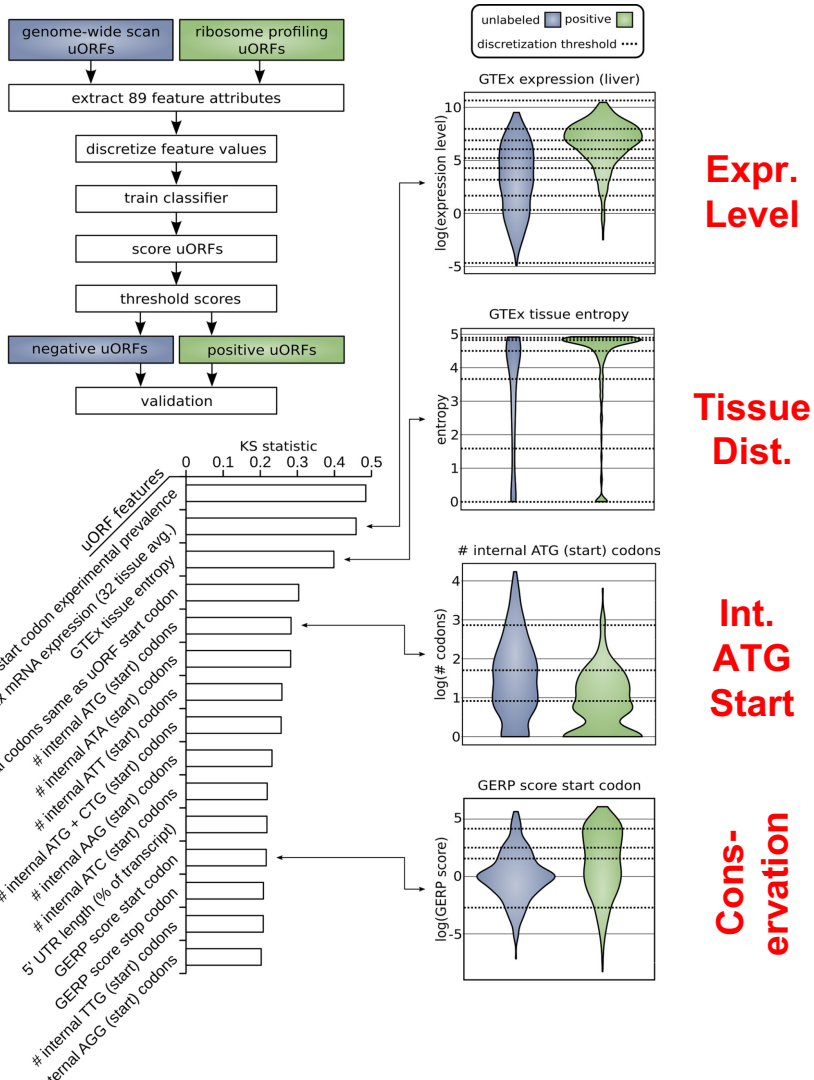
Prediction & validation of functional uORFs using 89 features

- 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).



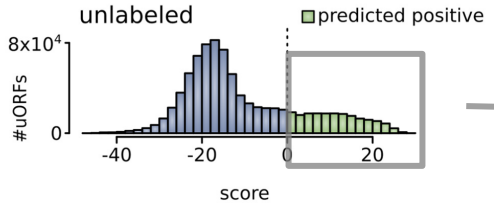


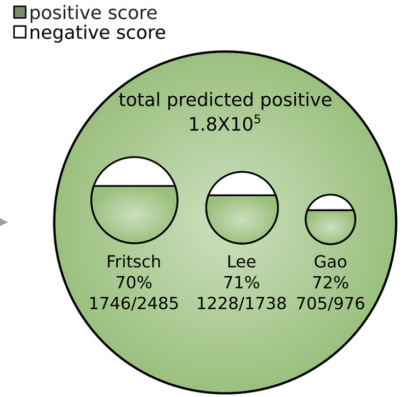
[McGillivray et al., *NAR* ('18)]

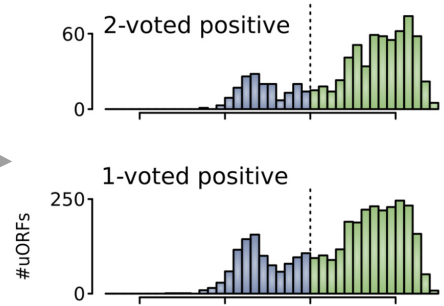


A comprehensive catalog of functional uORFs

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







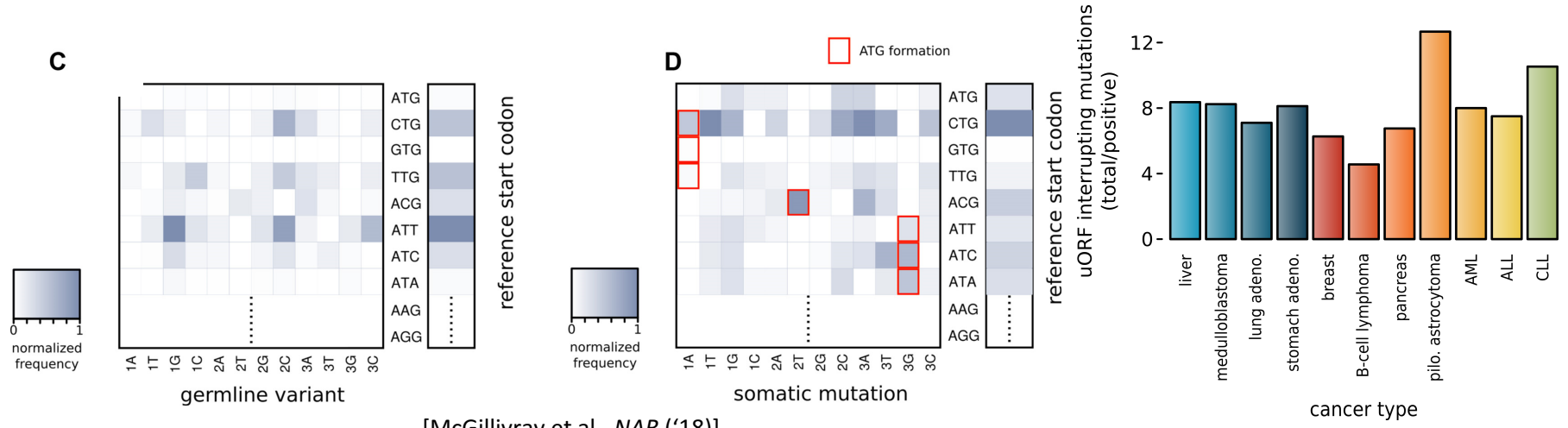
- Predicted functional uORFs may be intersected with disease associated variants.

- **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)]

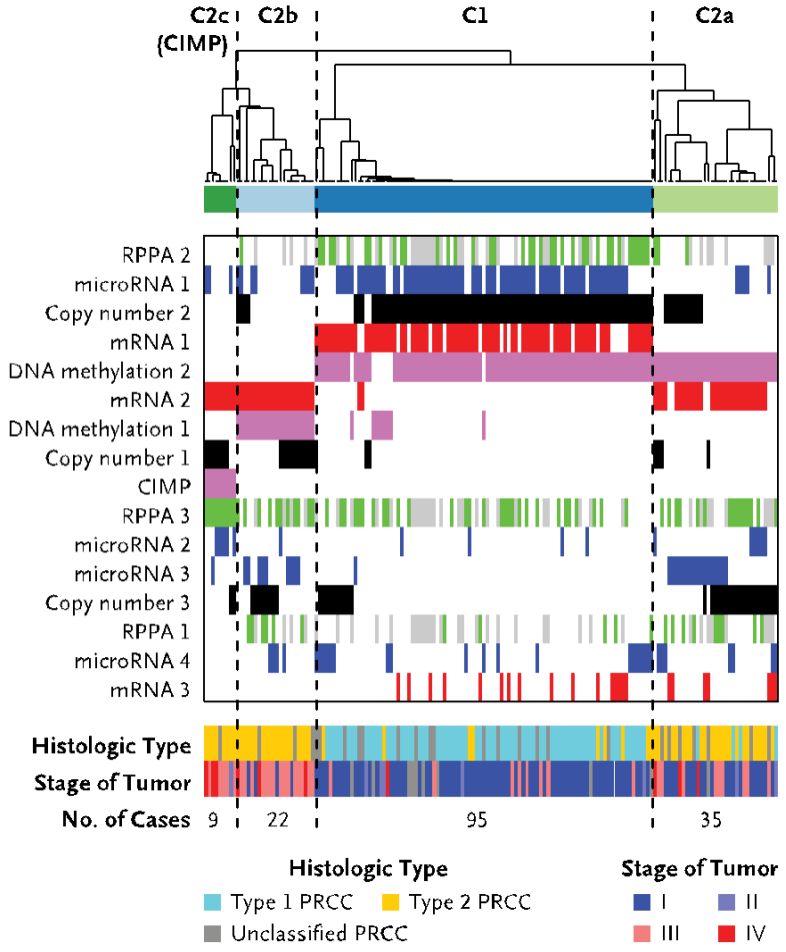
Prioritizing Variants in Personal Genomes: Using functional impact, with particular application to cancer

- 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
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
  - **RADAR:** prioritize variants based on post-transcriptional regulome using ENCODE eCLIP
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation
- (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 15

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....

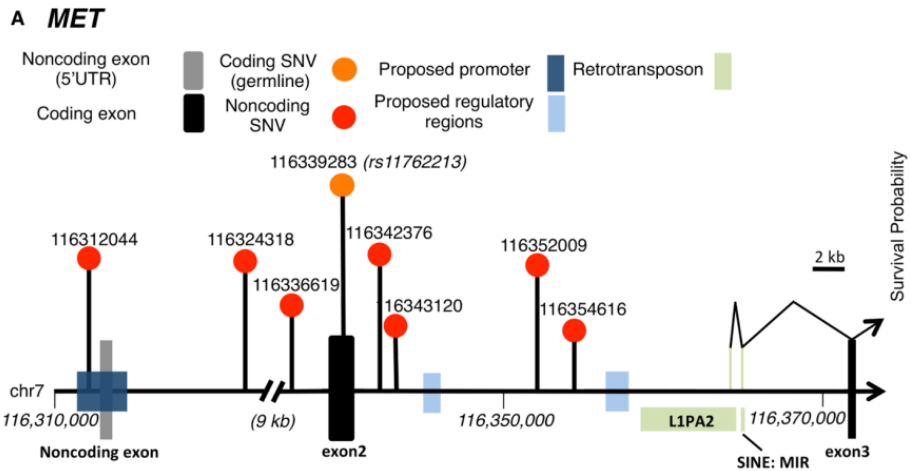
•35 WGS of TN pairs,
perhaps useful? But not that definitive from a recurrence perspective

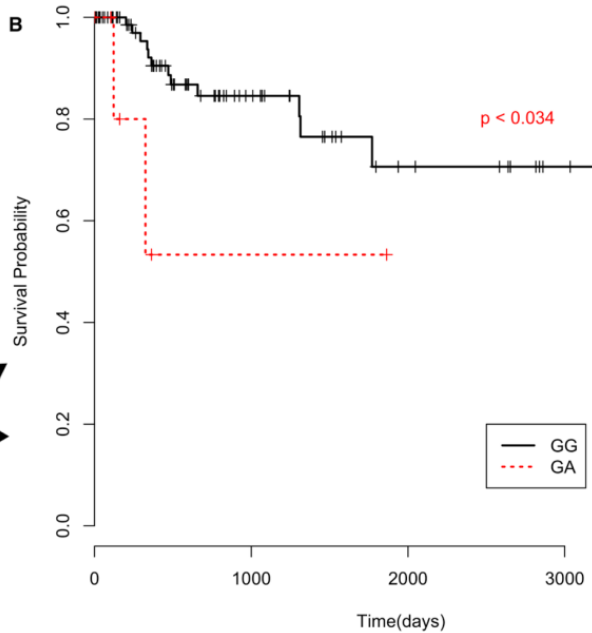


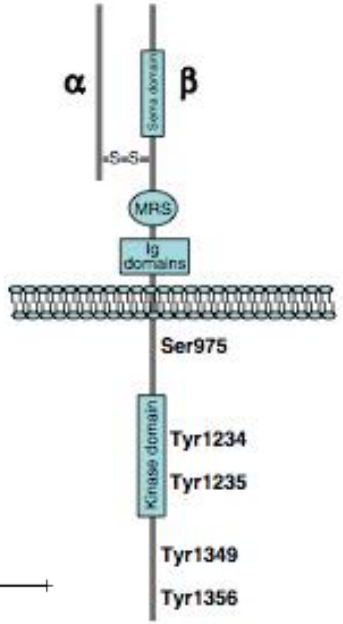
[Cancer Genome Atlas Research Network N Engl J Med. ('16)]

Tyr-kinase MET: Known Facts & New Results

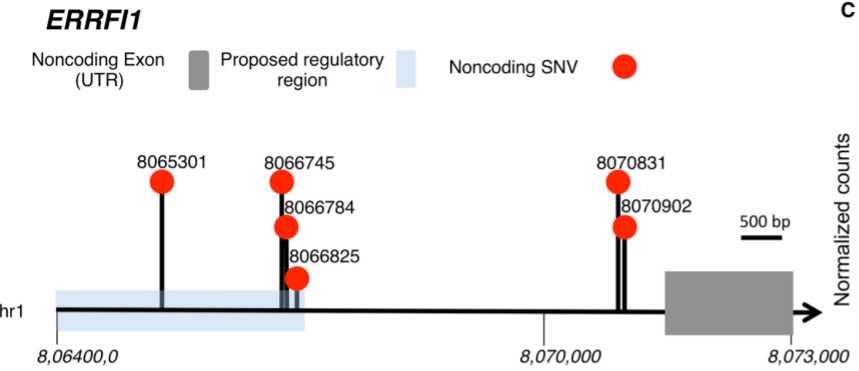
- MET is long known pRCC driver
- In MET, TCGA found somatic SNVs, duplications & an alt. splicing event as drivers (43/161).
- In addition, from 35 WGS we found
  - A noncoding hotspot associated with *MET*
  - Lack of SVs & breakpoints disrupting *MET*
  - Germline SNP (rs11762213) predicts survival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

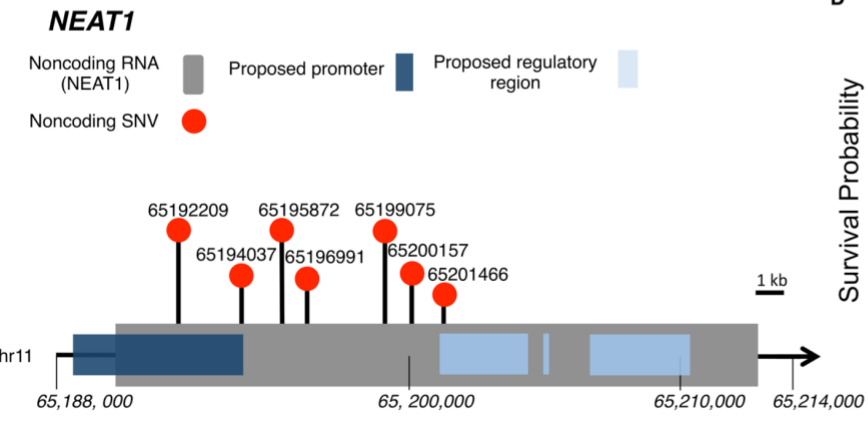


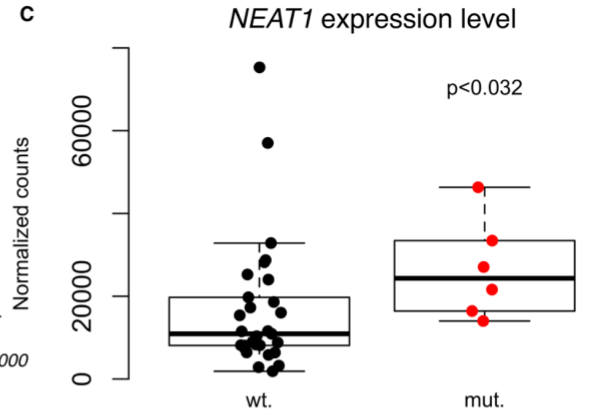


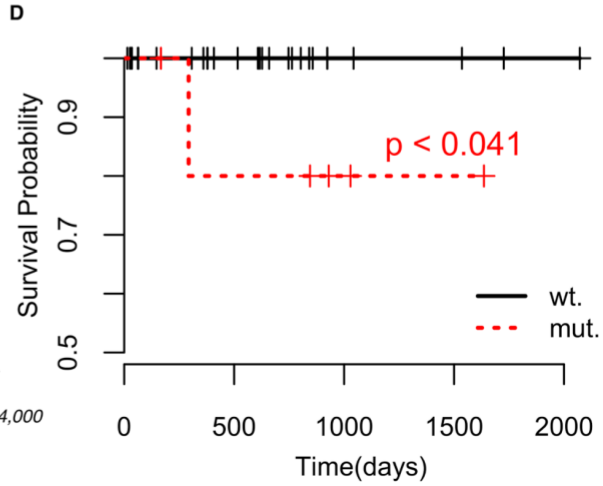


[A. Gentile, L. Trusolino and PM. Comoglio, Cancer and Metastasis Reviews ('08); S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]







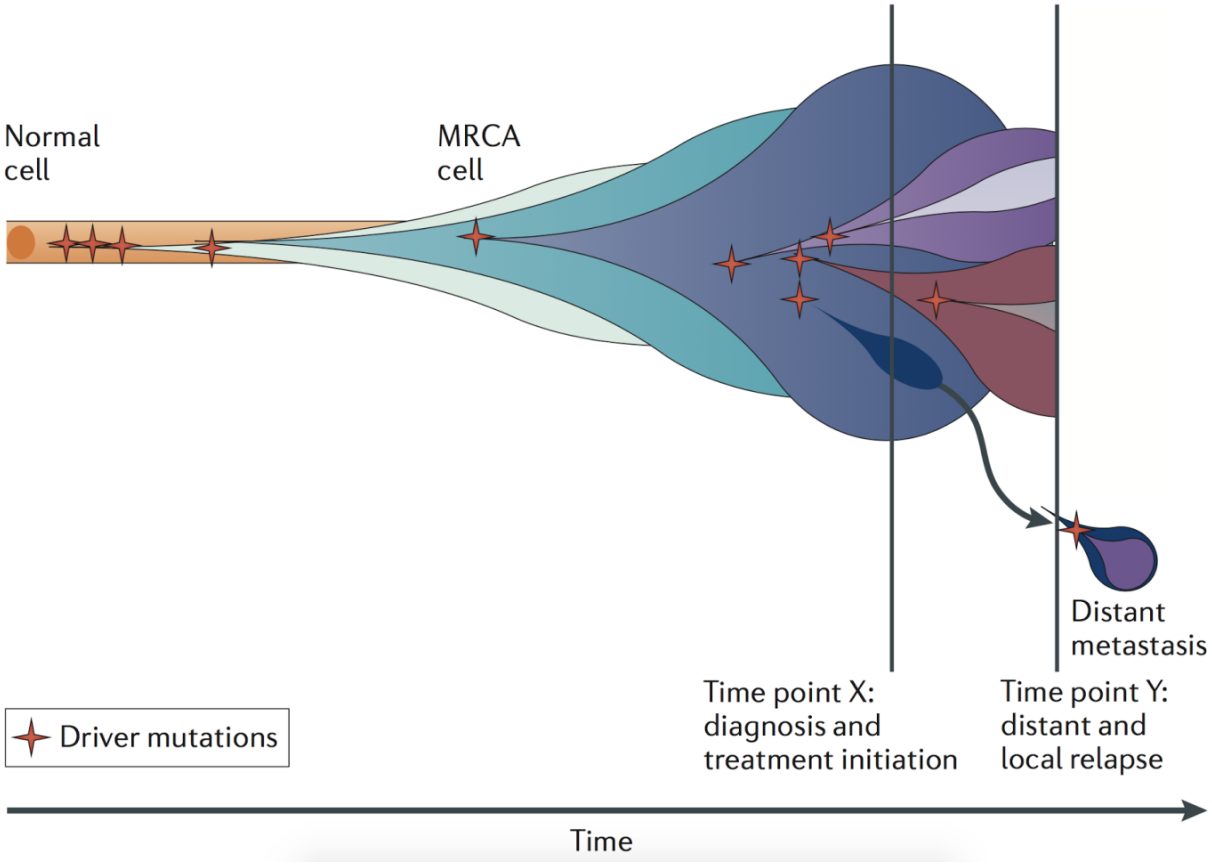


Beyond *MET*: 2 non-coding hotspots in NEAT & ERRFI1,

supported by expr. changes & survival analysis

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

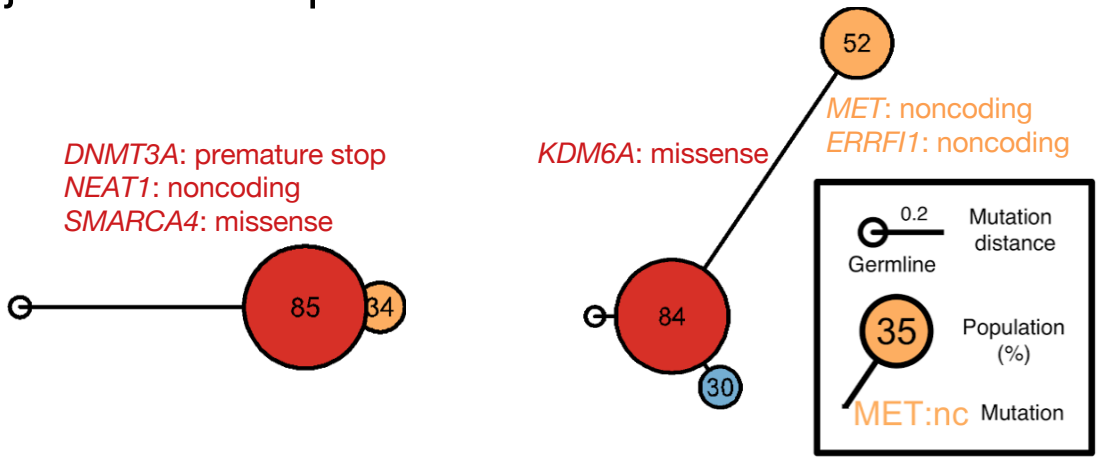
Tumor Evolution: Highlight the Ordering of Key Mutations



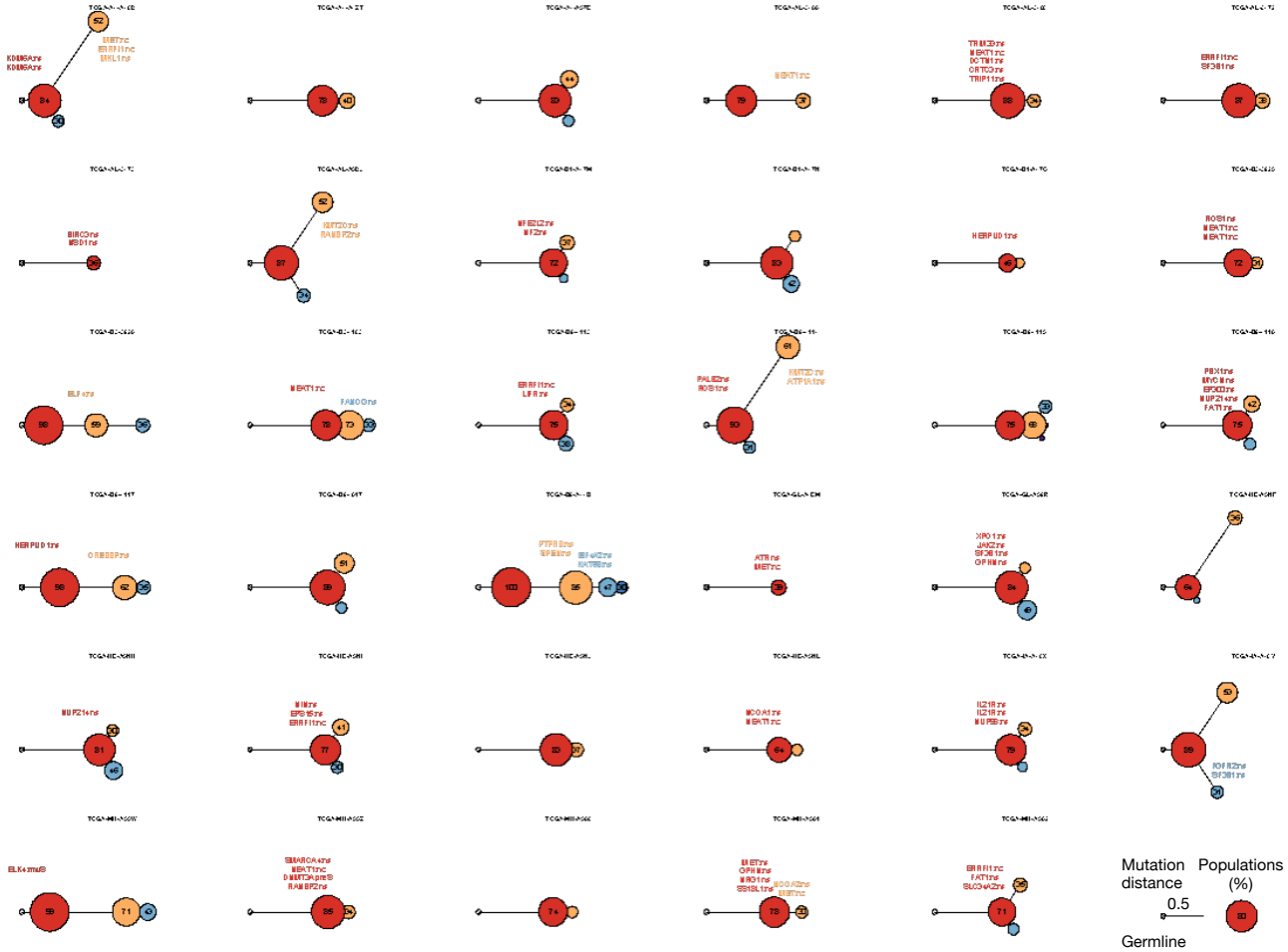
Yates et al, NRG (2012)

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



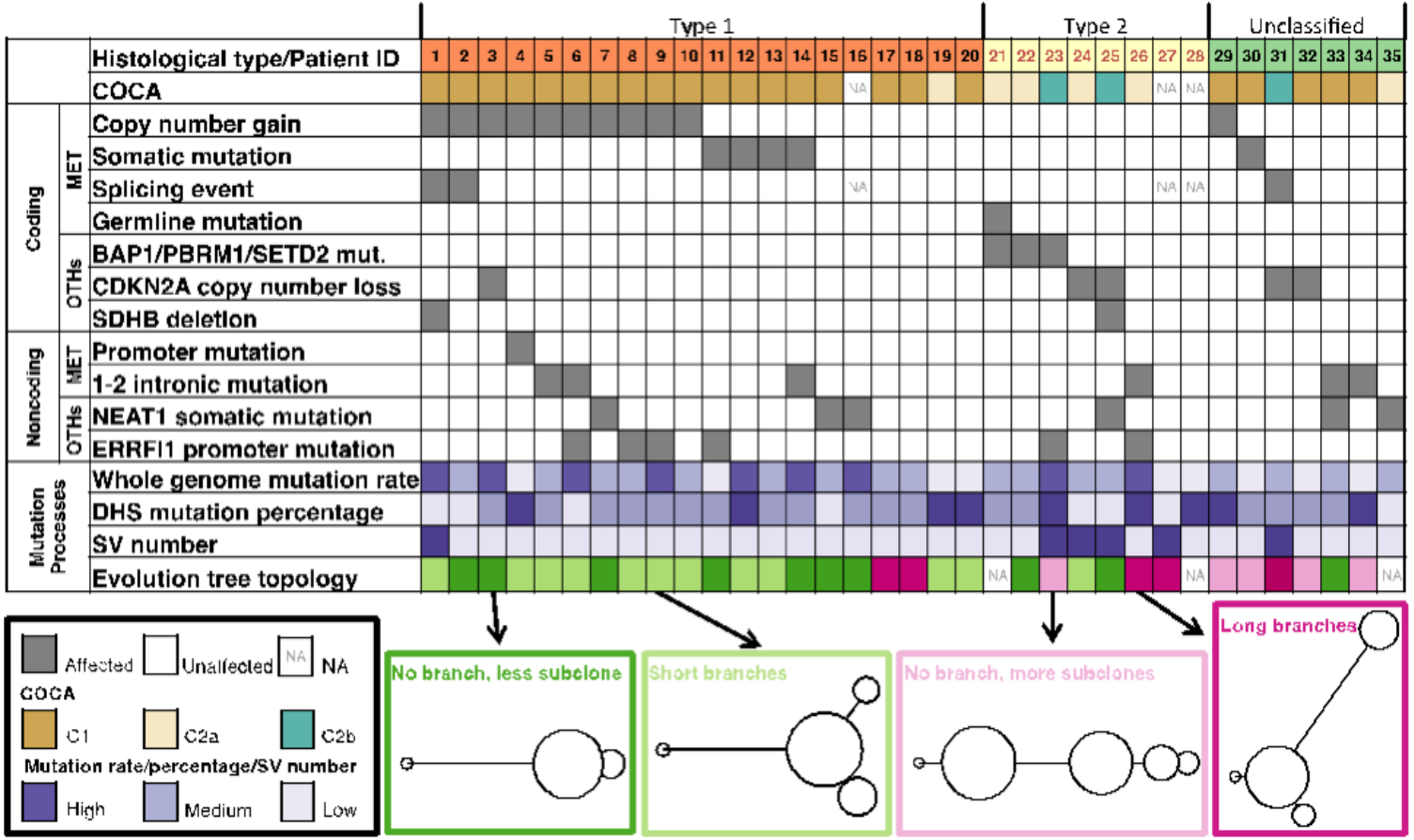
[S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]



[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



Prioritizing Variants in Personal Genomes: Using functional impact, 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
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
  - **RADAR:** prioritize variants based on post-transcriptional regulome using ENCODE eCLIP
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation
- (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 15

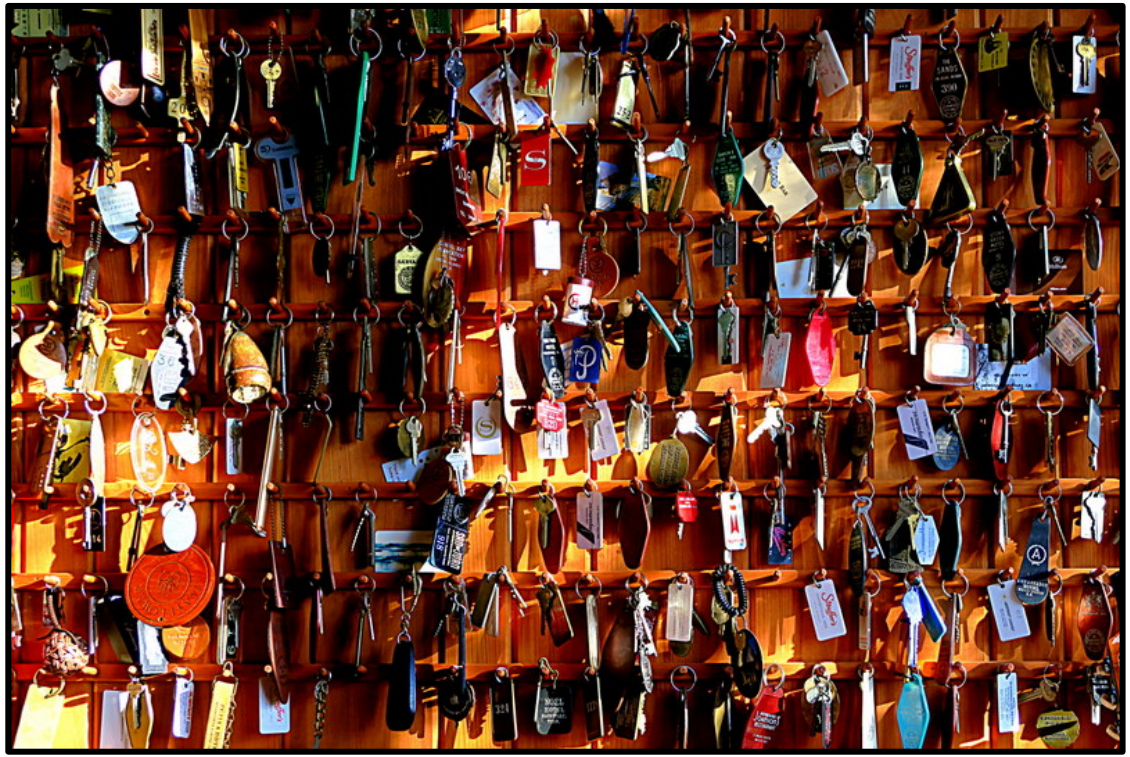
Prioritizing Variants in Personal Genomes: Using functional impact, with particular application to cancer

- 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
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
  - **RADAR:** prioritize variants based on post-transcriptional regulome using ENCODE eCLIP
  - **uORFs:** Feature integration to find small subset of upstream mutations that potentially alter translation
- (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 15

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

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

pRCC - S **Li**, B Shuch



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

github.gersteinlab.org/**uORFs**

P **McGillivray**, R Ault, M Pawashe, R Kitchen, S Balasubramanian

FunSeq.gersteinlab.org

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

RADAR.gersteinlab.org

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


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