Genomics & Data Science:
Approaches to identifying key variants through functional impact & recurrence

Mark Gerstein
Yale

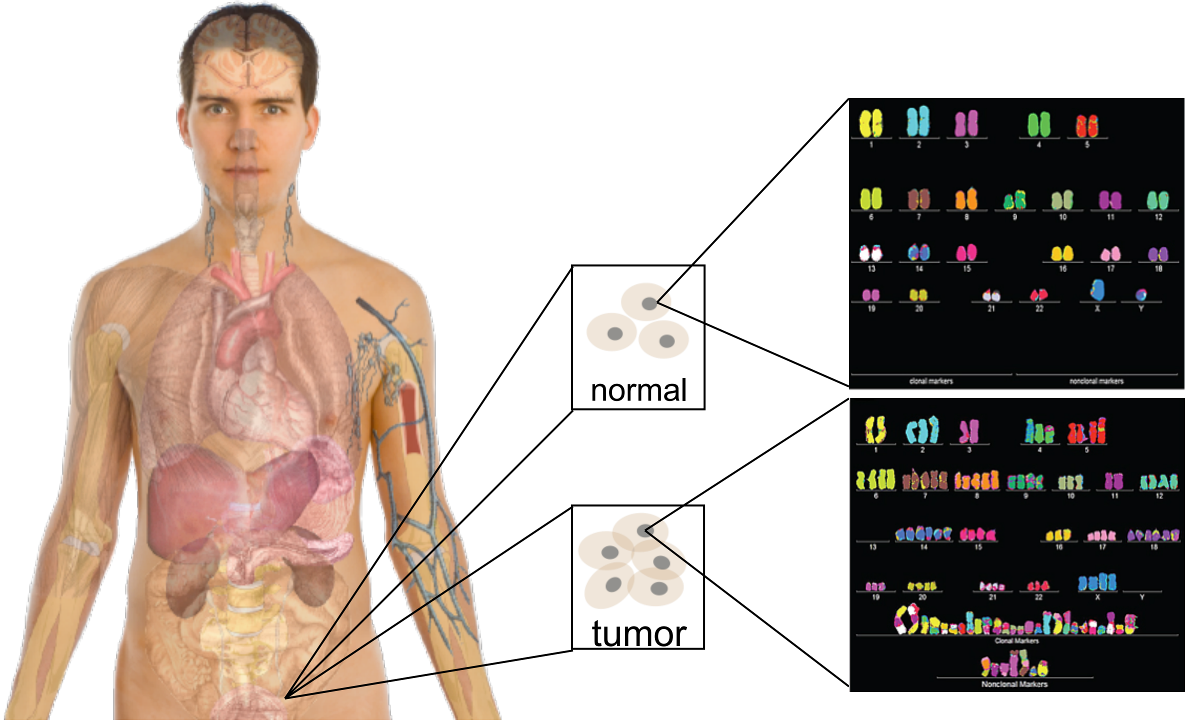
Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @MarkGerstein).

No Conflicts for this Talk

See last slide for more info.

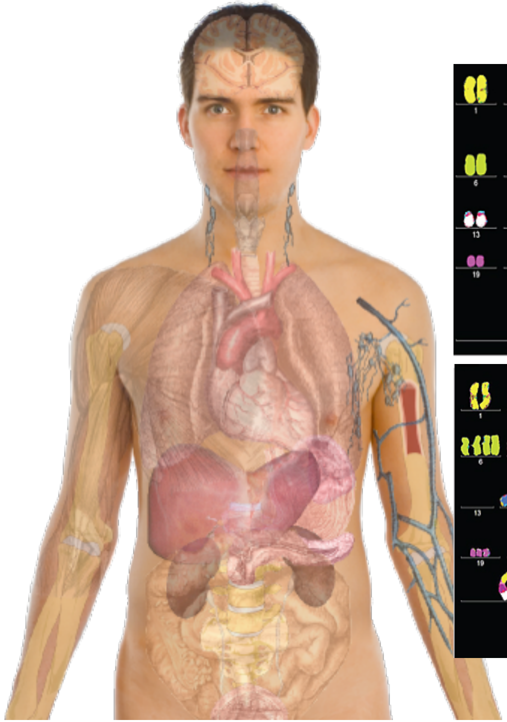
Personal Genomics as a Gateway into Biology

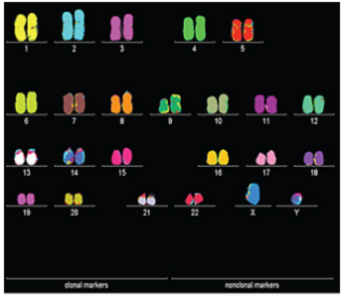
Personal genomes will soon become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.

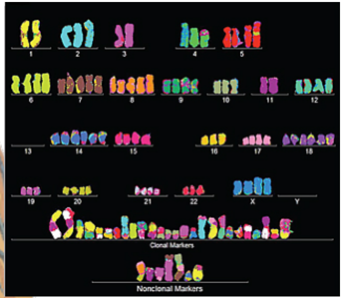


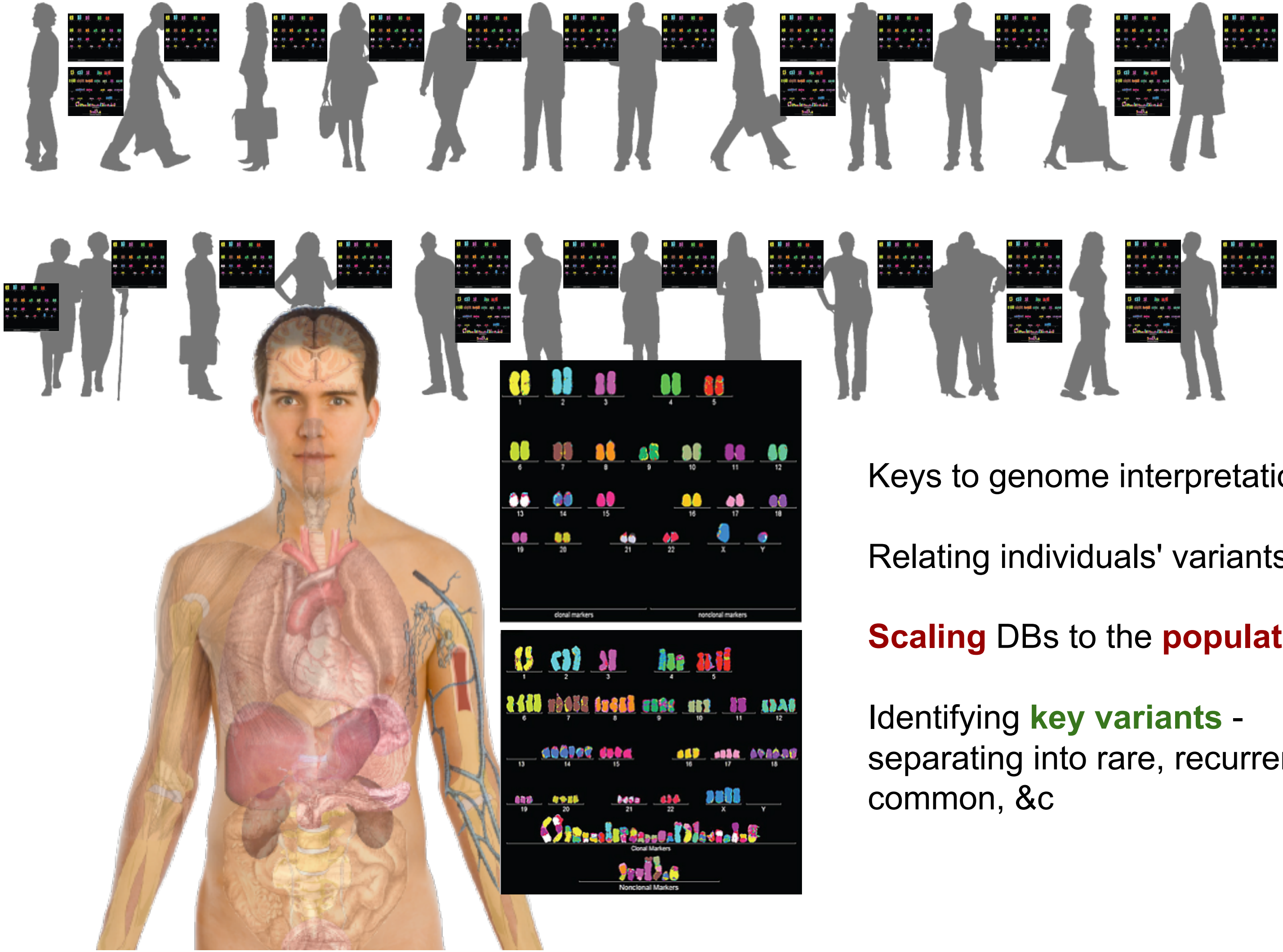
Personal Genomics as a Gateway into Biology

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.









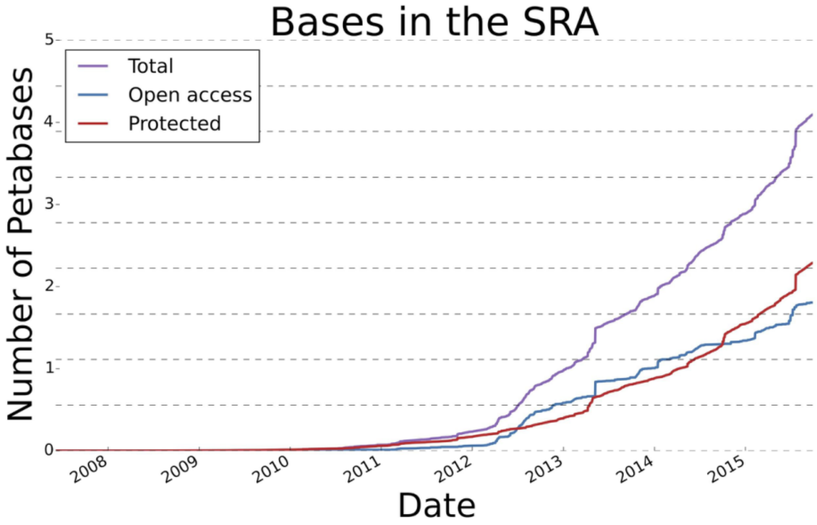
Keys to genome interpretation

Relating individuals' variants to **DBs**

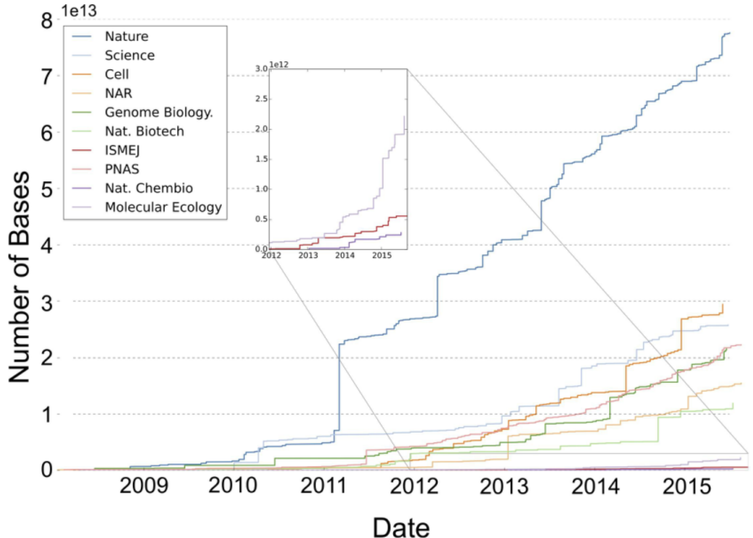
Scaling DBs to the **population**

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

DB Growth: explosion of data scale & a diversity of uses



- The type of sequence data deposited has changed as well.
  - Protected data represents an increasing fraction of all submitted sequences.

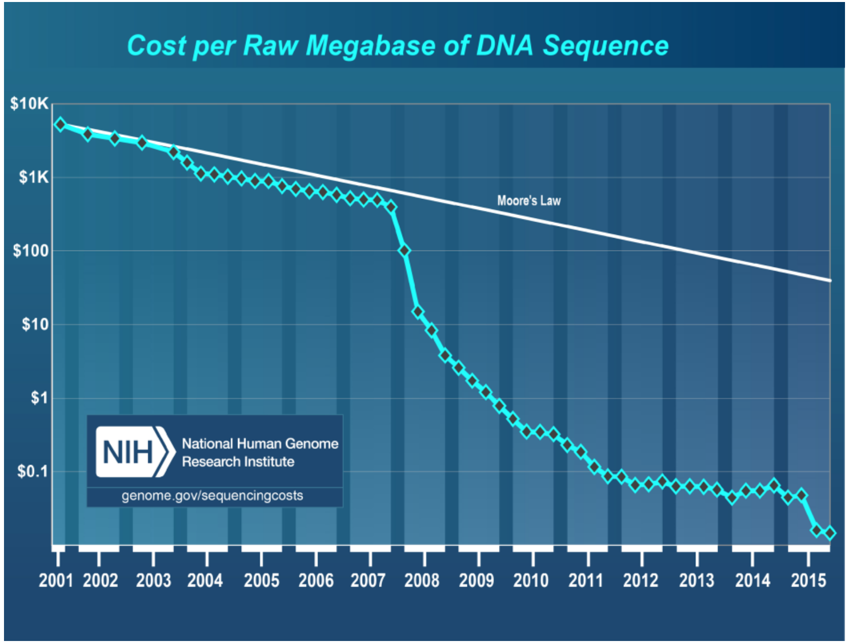


[Muir et al. ('15) GenomeBiol.]

Sequencing Data Explosion: Faster than Moore’s Law?

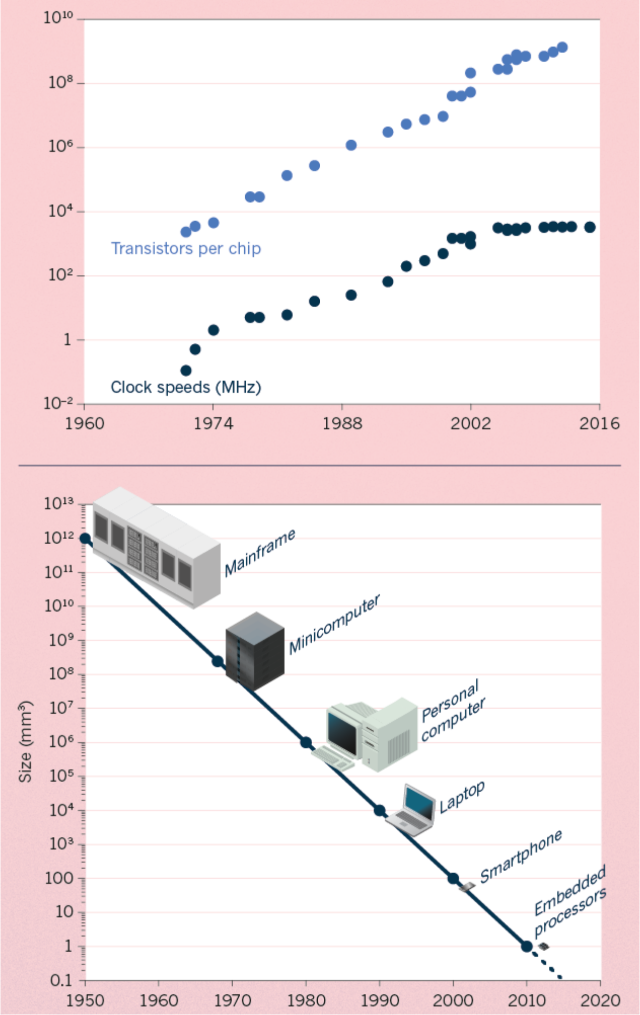
- In the early 2000’s, improvements in Sanger sequencing produced a scaling pattern similar to Moore’s law.

- The advent of NGS was a shift to a new technology with dramatic decrease in cost).



Moore's Law: Exponential Scaling of Computer Technology

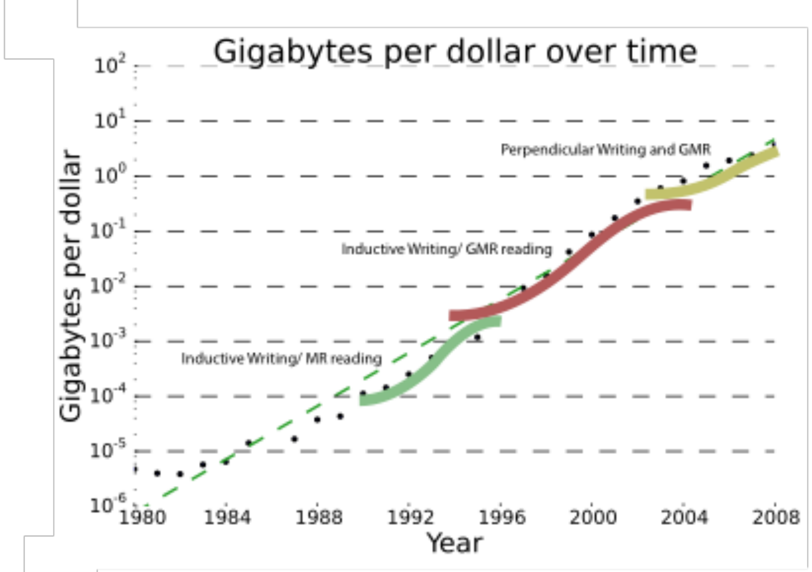
- Exponential increase in the number of transistors per chip.
- Led to improvements in speed and miniaturization.
- Drove widespread adoption and novel applications of computer technology.

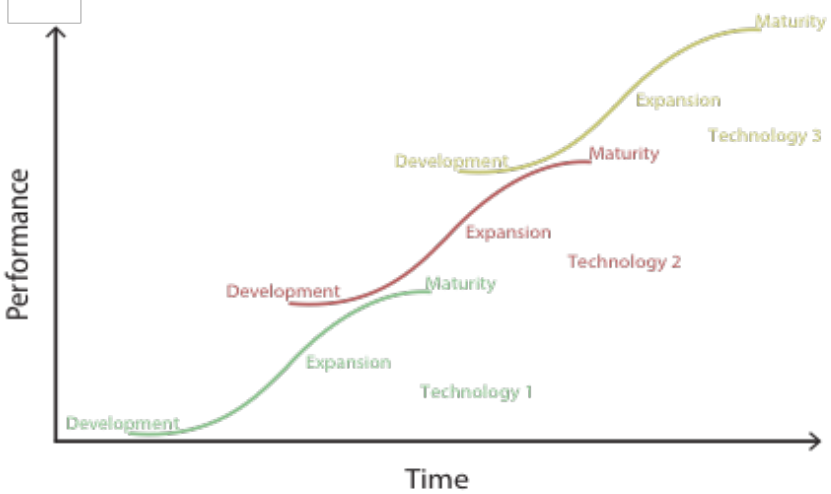


[Waldrop ('15) Nature]

Kryder's Law and S-curves underlying exponential growth

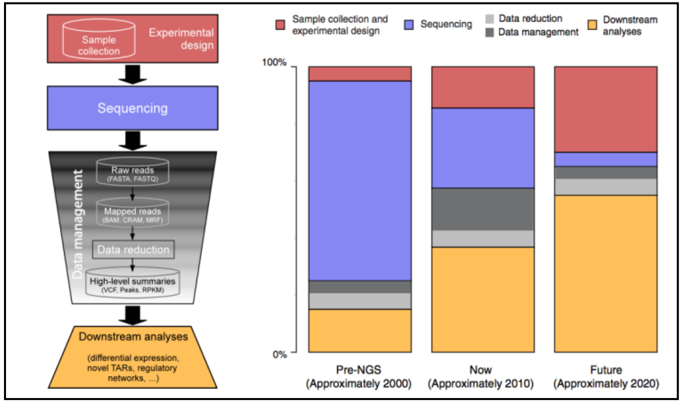
- Moore's & Kryder's Laws
    - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Kryder's law is a superposition of S-curves for different technologies





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

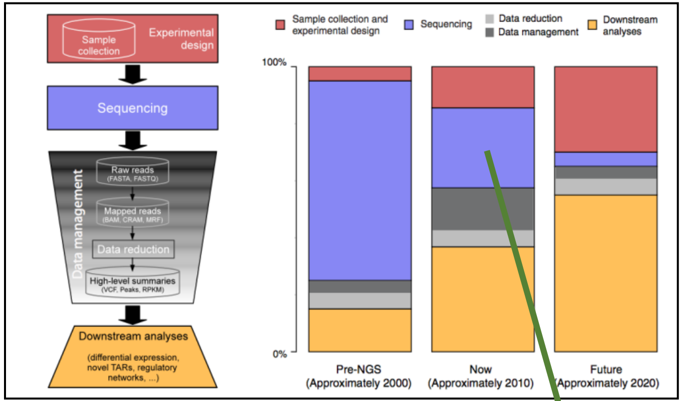
The changing costs of a sequencing pipeline



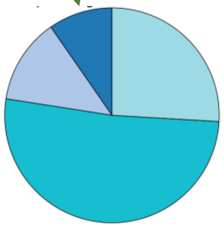
From ‘00 to ~’20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

The changing costs of a sequencing pipeline

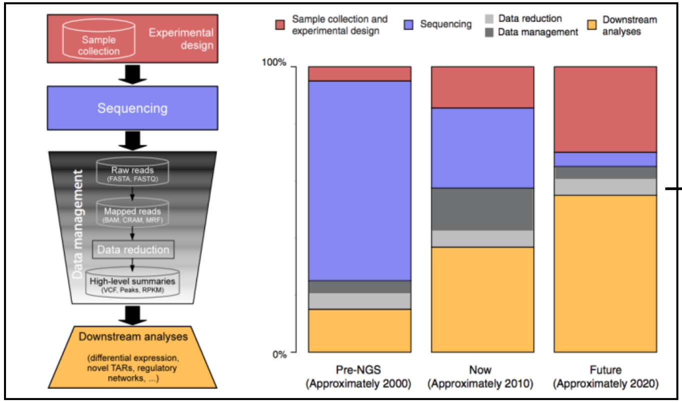


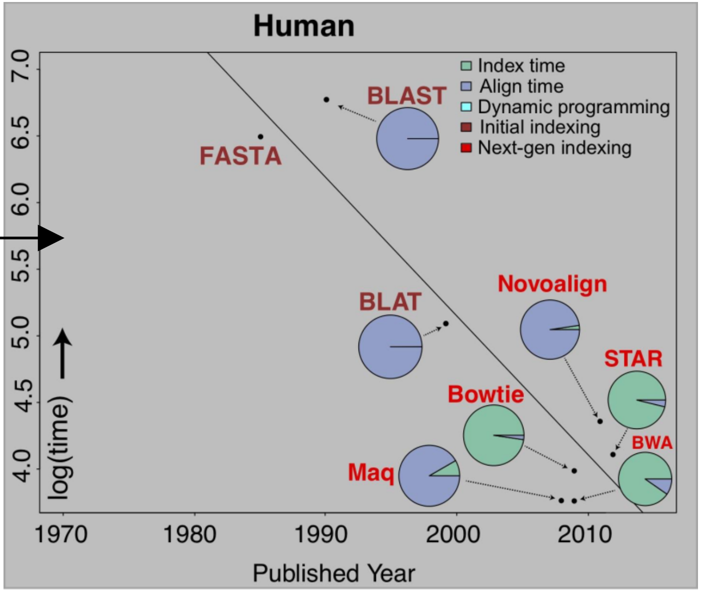
From '00 to ~'20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



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

The changing costs of a sequencing pipeline



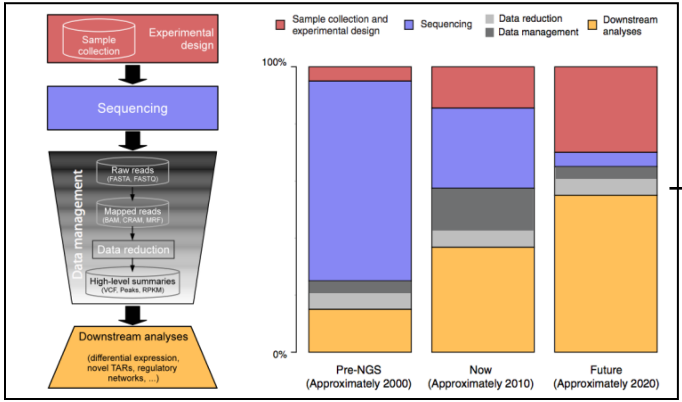


From ‘00 to ~’20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

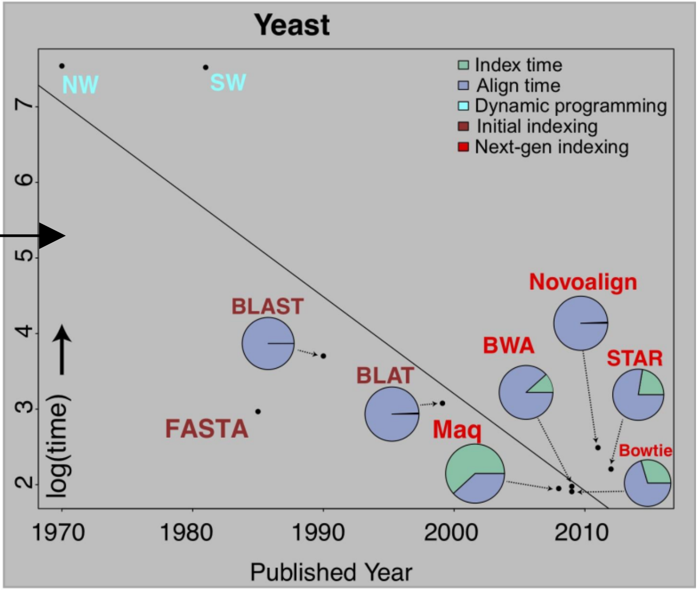
Alignment algorithms scaling to keep pace with data generation

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

The changing costs of a sequencing pipeline



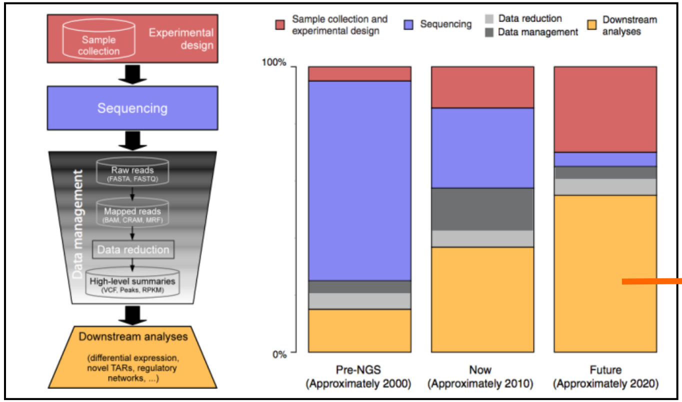
From ‘00 to ~’20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

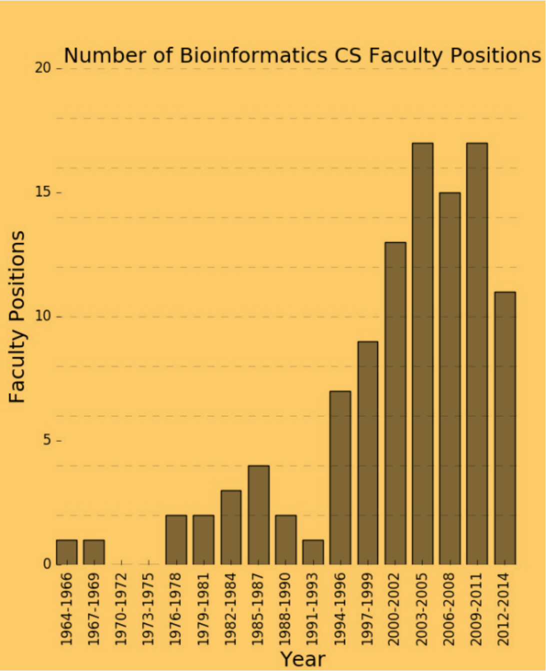


Alignment algorithms scaling to keep pace with data generation

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

The changing costs of a sequencing pipeline

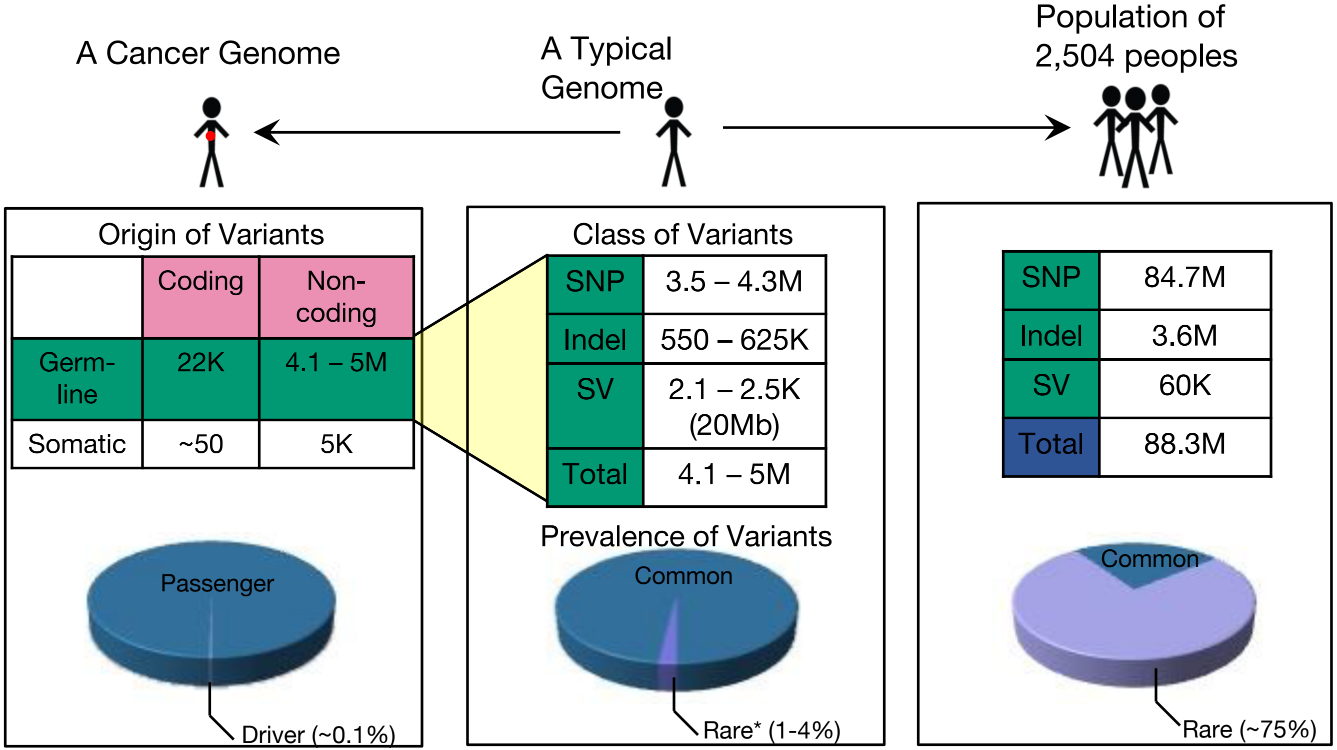




From ‘00 to ~’20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

Human Genetic Variation



Origin of Variants

	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K

Class of Variants

SNP	3.5 – 4.3M
Indel	550 – 625K
SV	2.1 – 2.5K (20Mb)
Total	4.1 – 5M

SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M

* 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

Genomics & Data Science:
Approaches to identifying key variants through functional impact & recurrence

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data generation & processing
  - Mining the data to prioritize variants for key drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - LoF annotation as a complex problem + finding deleterious LoFs
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with an entropy based weighting scheme.
  - Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence: Statistics for driver identification
  - **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 than but speeded up w/ GPUs

Genomics & Data Science:
Approaches to identifying key variants through functional impact & recurrence

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data generation & processing
  - Mining the data to prioritize variants for key drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - LoF annotation as a complex problem + finding deleterious LoFs
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with an entropy based weighting scheme.
  - Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence:
  Statistics for driver identification
  - **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 than but speeded up w/ GPUs

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

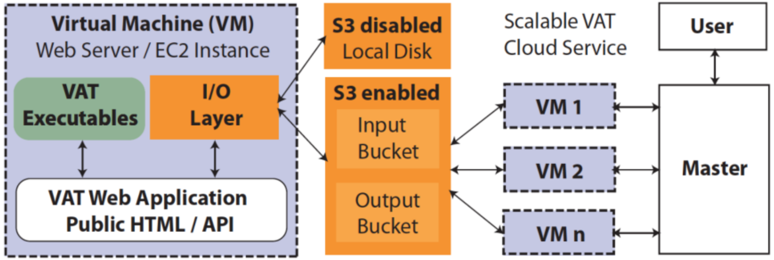
VCF Input

Output:

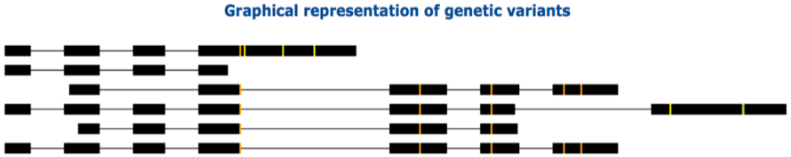
- Annotated VCFs
- Graphical representations of functional impact on transcripts

Access:

- Webserver
- AWS cloud instance
- Source freely available



CLOUD APPLICATION



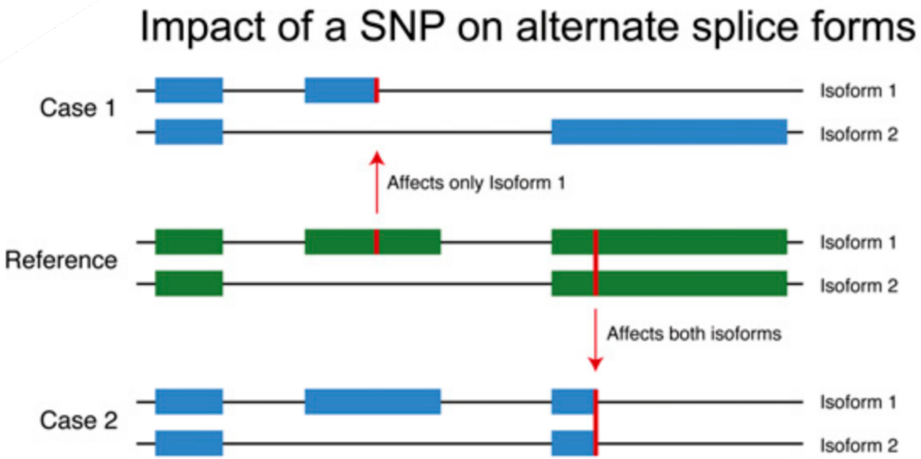
vat.gersteinlab.org

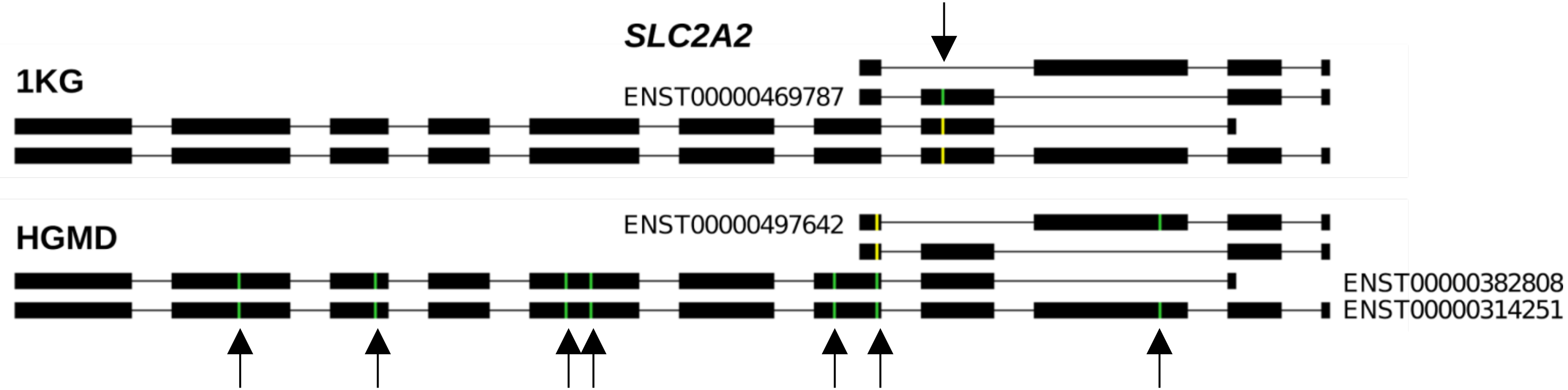
Habegger L., Balasubramanian S.*, et al. Bioinformatics, 2012*

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





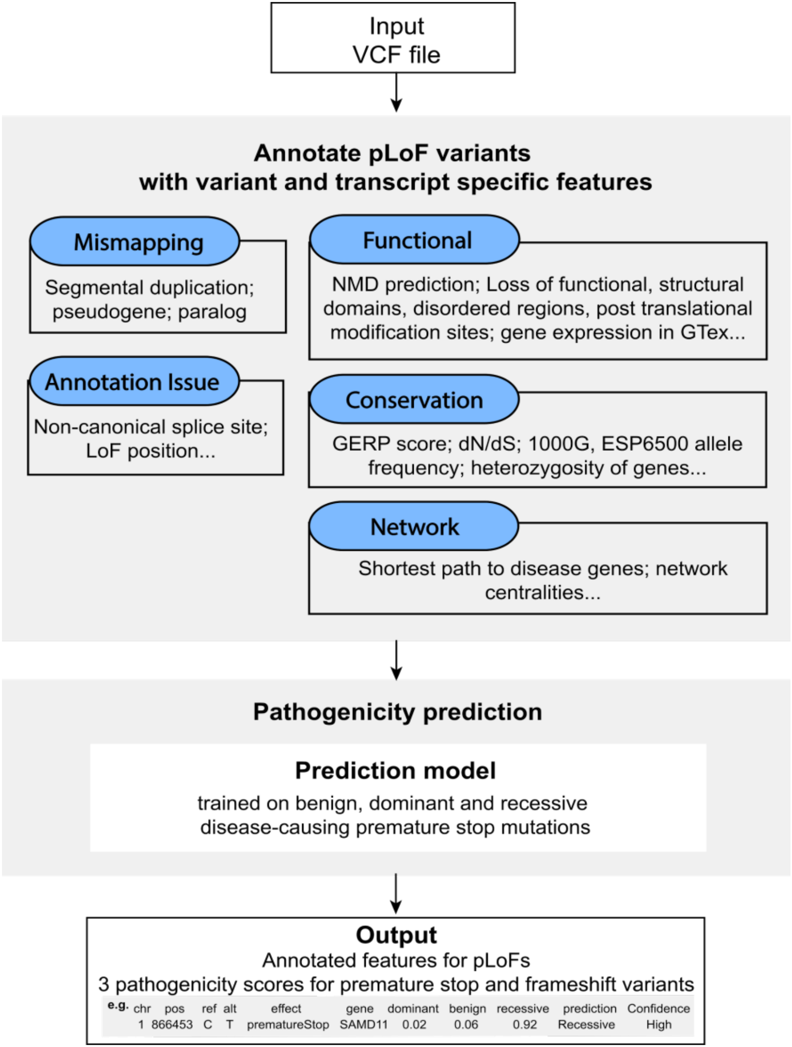
Annotation of Loss-of-Function Transcripts (ALoFT)

Runs on top of VAT

Output:

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

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



Input VCF file

Annotate pLoF variants with variant and transcript specific features

Mismapping: Segmental duplication; pseudogene; paralog

Functional: NMD prediction; Loss of functional, structural domains, disordered regions, post translational modification sites; gene expression in GTex...

Annotation Issue: Non-canonical splice site; LoF position...

Conservation: GERP score; dN/dS; 1000G, ESP6500 allele frequency; heterozygosity of genes...

Network: Shortest path to disease genes; network centralities...

Pathogenicity prediction

Prediction model: trained on benign, dominant and recessive disease-causing premature stop mutations

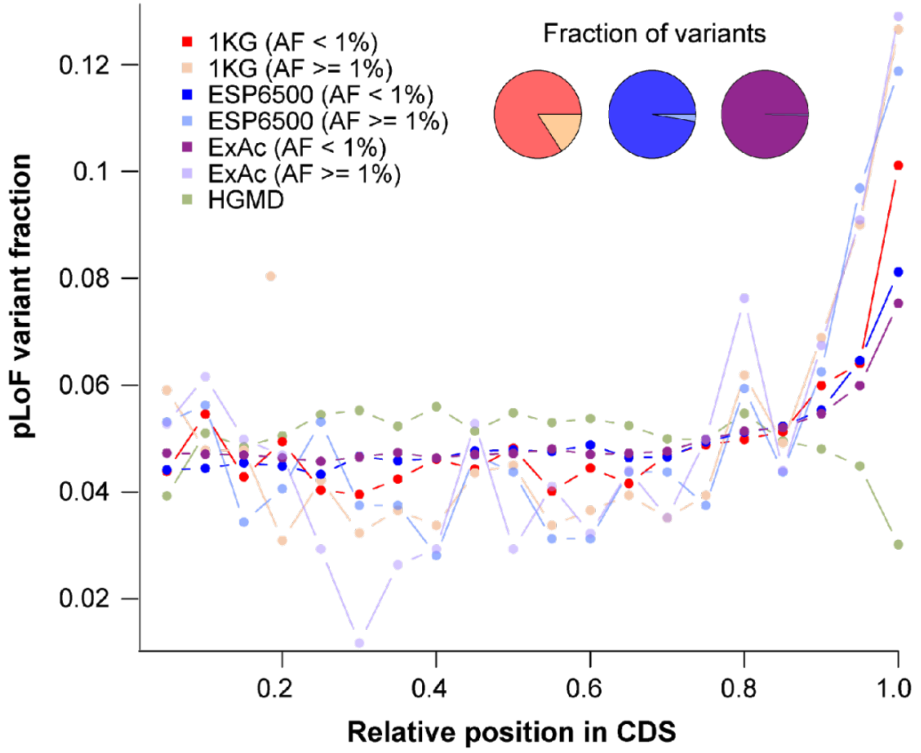
Output

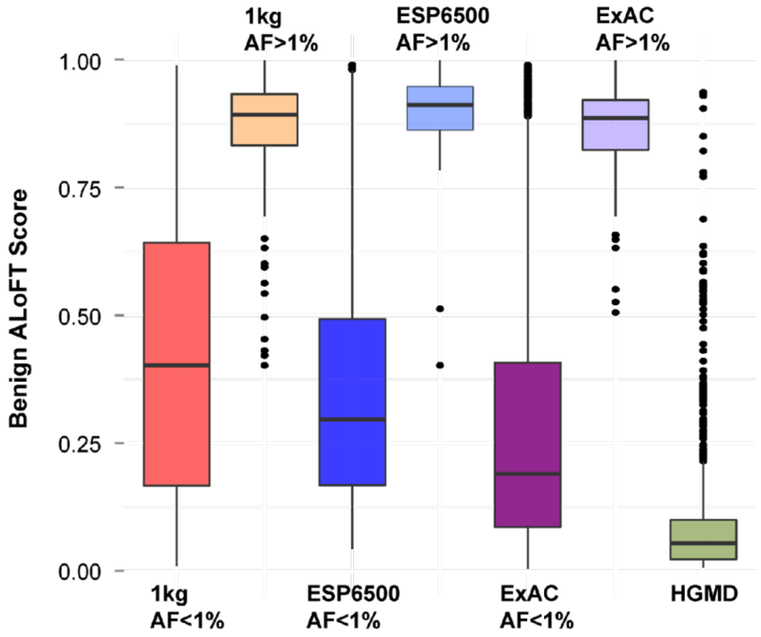
Annotated features for pLoFs

3 pathogenicity scores for premature stop and frameshift variants

e.g. chr	pos	ref	alt	effect	gene	dominant	benign	recessive	prediction	Confidence
1	866453	C	T	prematureStop	SAMD11	0.02	0.06	0.92	Recessive	High

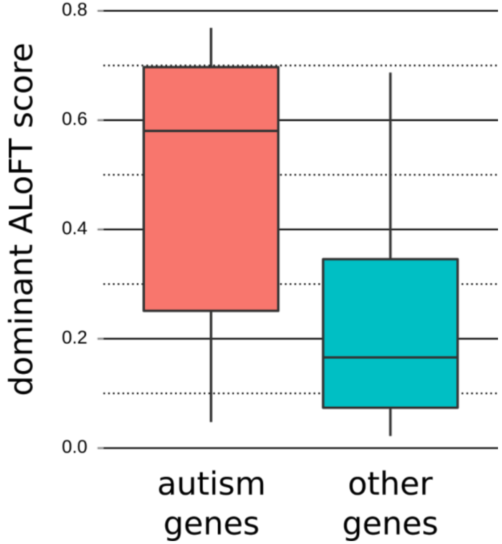
LoF distribution varies as expected by mutation set (from healthy people v from disease)

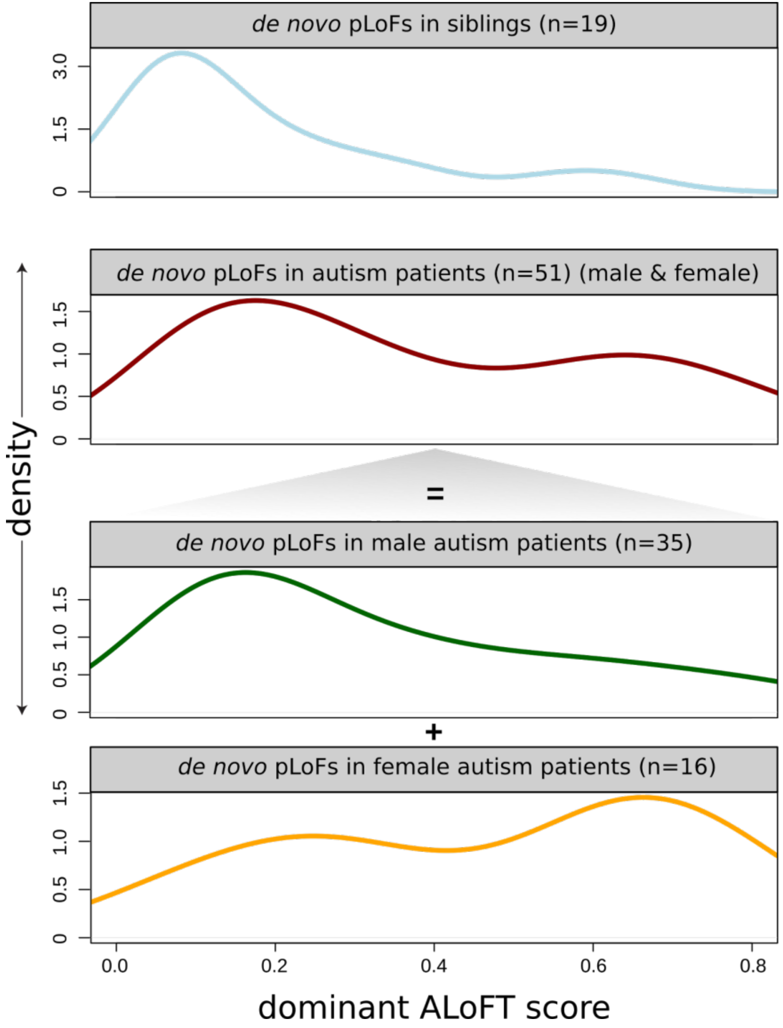




Balasubramanian S.*, Fu Y.* et al., *NComms.*, ’17

Application to LoF mutations in autism spectrum disorder





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

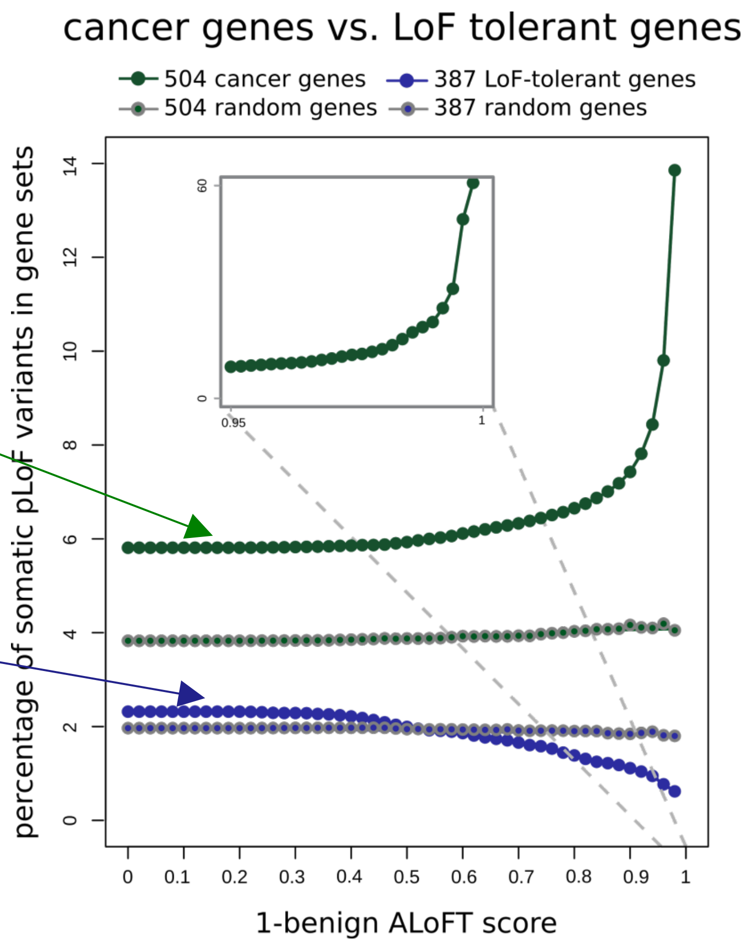
ALoFT identifies deleterious somatic LoF variants

Cancer genes:

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

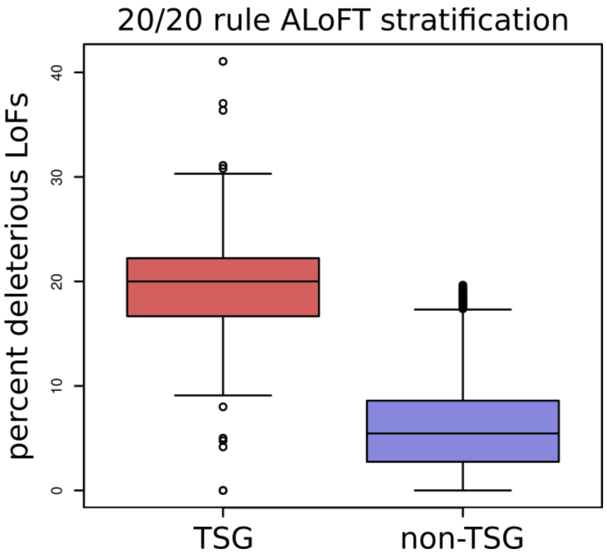
LoF tolerant genes:

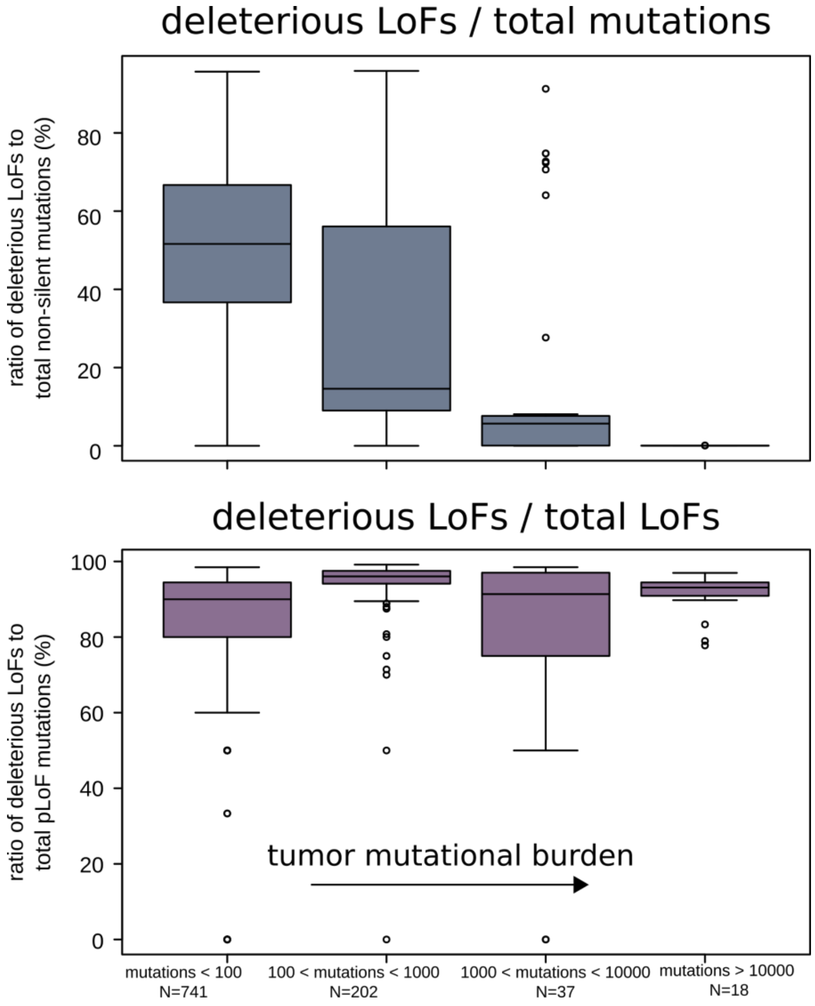
- LoF in the 1KG cohort.
- *Depleted in deleterious LoFs.*



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

ALoFT refines cancer mutation characterization





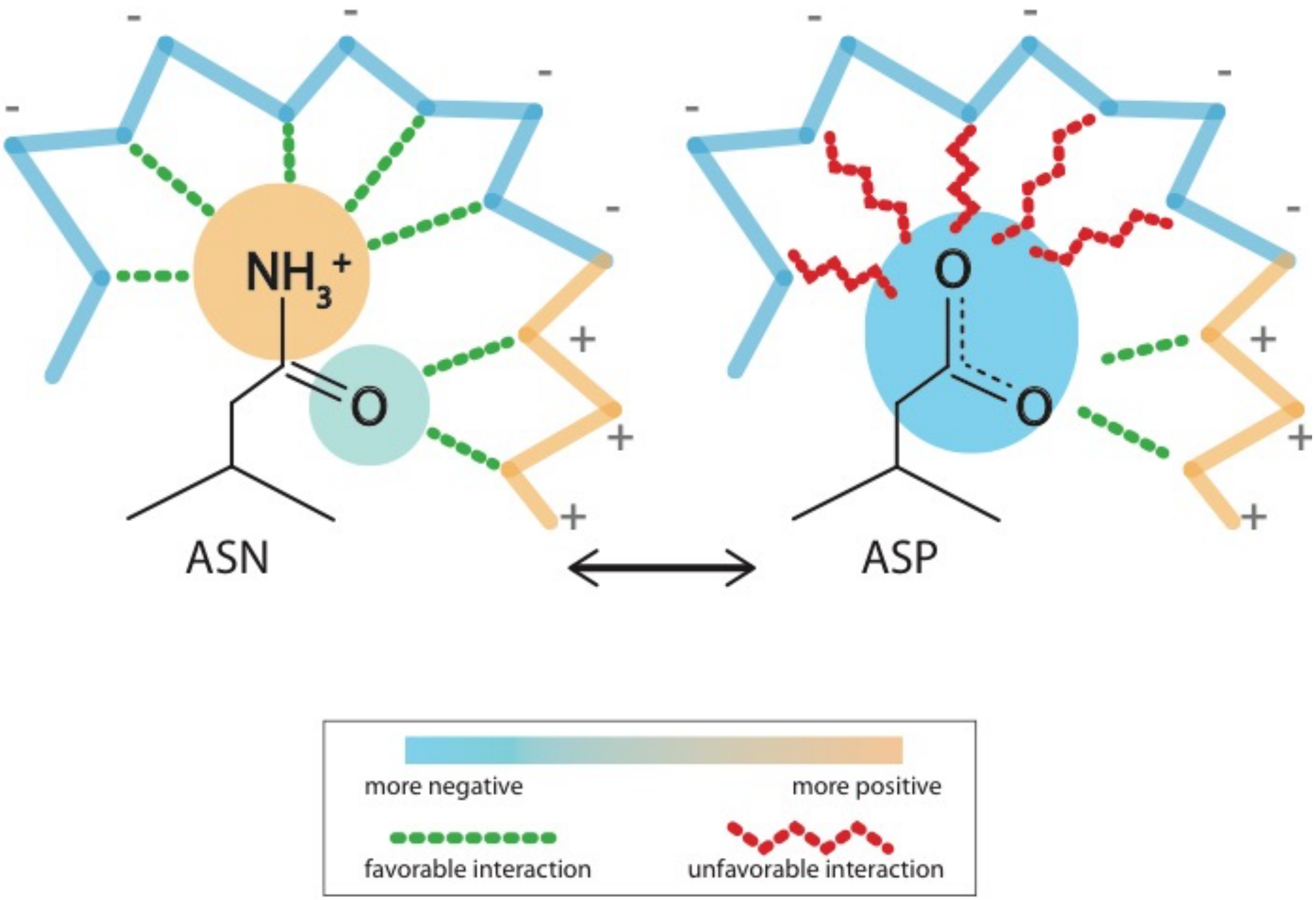
Vogelstein et al. '13: if >20% of mutations in gene inactivating → tumor suppressor gene (TSG).

ALoFT further refines 20/20 rule predictions.

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

Genomics & Data Science:
Approaches to identifying key variants through functional impact & recurrence

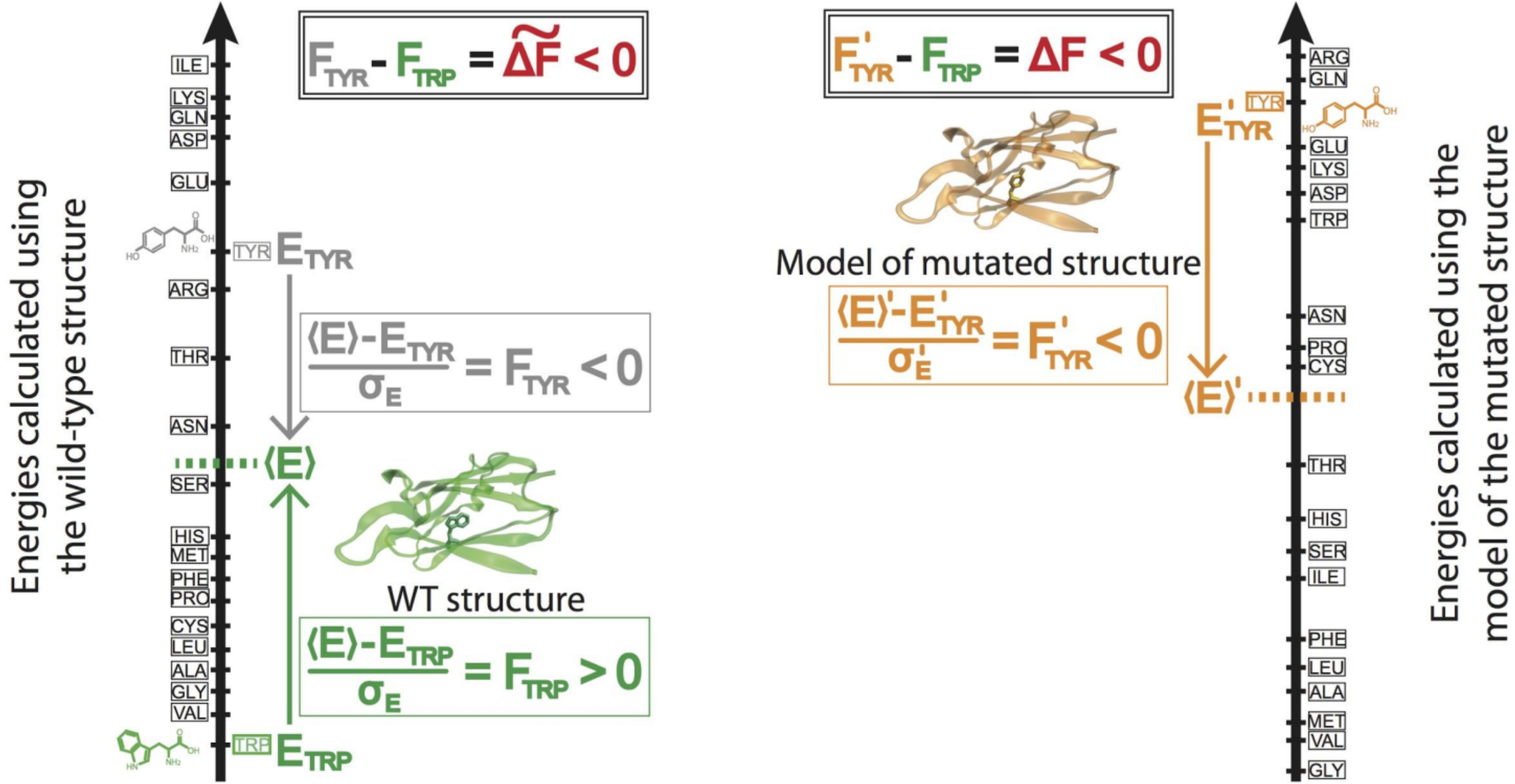
- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data generation & processing
  - Mining the data to prioritize variants for key drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - LoF annotation as a complex problem + finding deleterious LoFs
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with an entropy based weighting scheme.
  - Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence:
  Statistics for driver identification
  - **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 than but speeded up w/ GPUs



What is localized frustration ?

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

Workflow for evaluating localized frustration changes (ΔF)



$$F_{TYR} - F_{TRP} = \widetilde{\Delta F} < 0$$

$$\frac{\langle E\rangle - E_{TYR}}{\sigma_E} = F_{TYR} < 0$$

$$\frac{\langle E\rangle - E_{TRP}}{\sigma_E} = F_{TRP} > 0$$

WT structure

Energies calculated using the wild-type structure

$$F'_{TYR} - F_{TRP} = \Delta F < 0$$

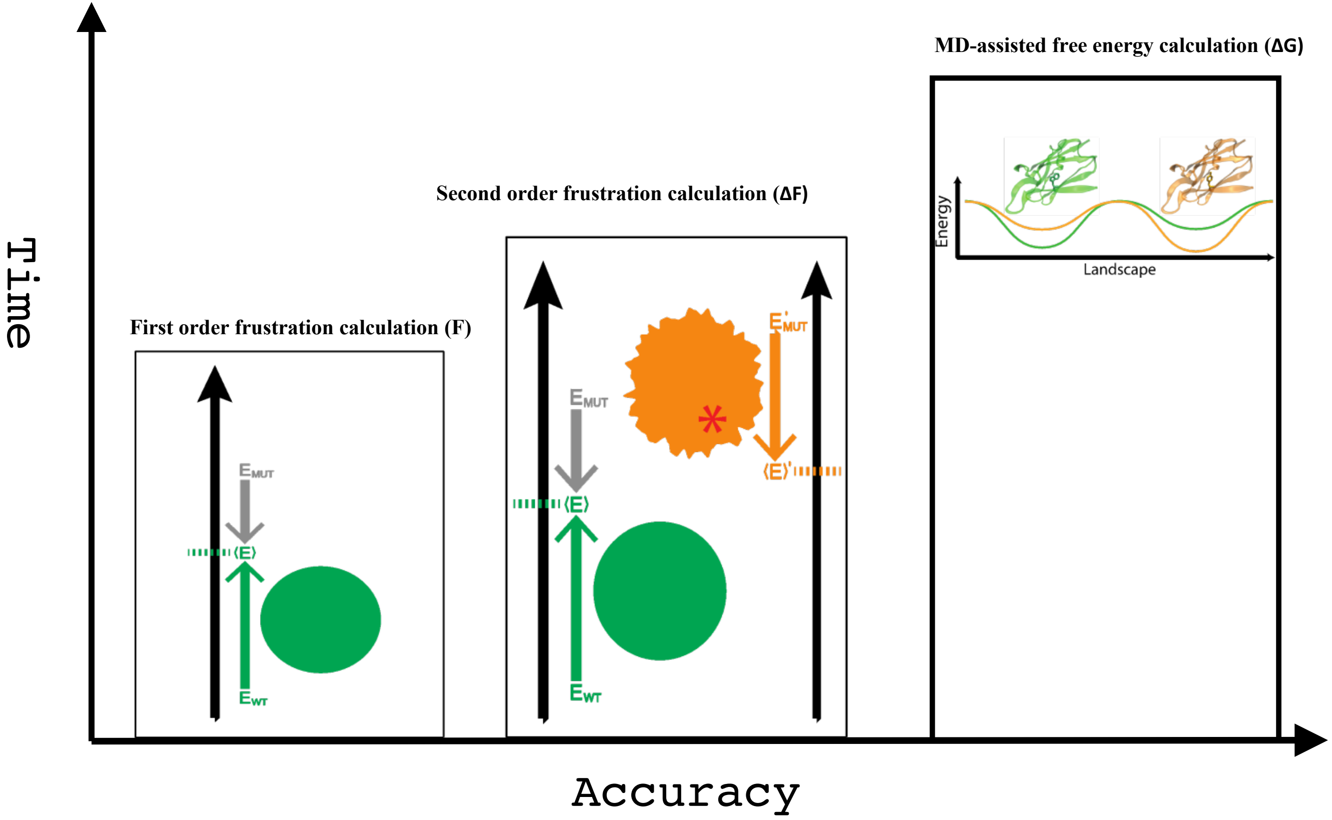
Model of mutated structure

$$\frac{\langle E\rangle' - E'_{TYR}}{\sigma'_E} = F'_{TYR} < 0$$

Energies calculated using the model of the mutated structure

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

Complexity of the second order frustration calculation

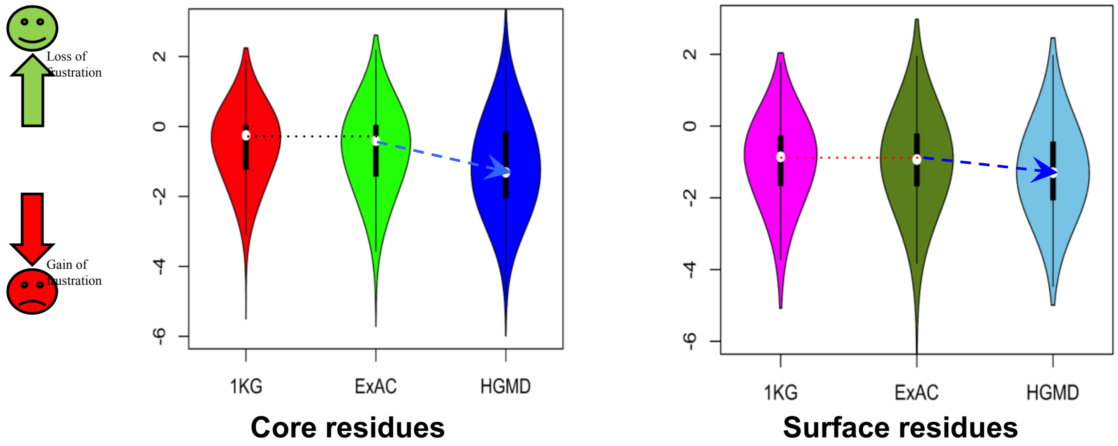


First order frustration calculation (F)

Second order frustration calculation (ΔF)

MD-assisted free energy calculation (ΔG)

Comparing ΔF values across different SNV categories: disease v normal



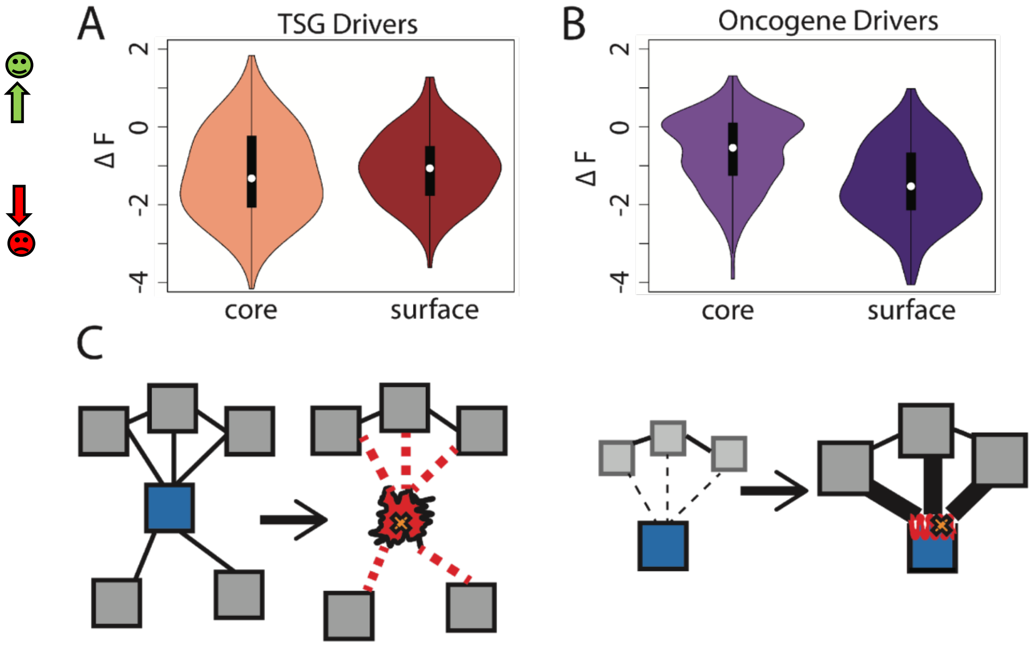
Core residues

Surface residues

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)]

Genomics & Data Science:
Approaches to identifying key variants through functional impact & recurrence

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data generation & processing
  - Mining the data to prioritize variants for key drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - LoF annotation as a complex problem + finding deleterious LoFs
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with an entropy based weighting scheme.
  - Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence:
  Statistics for driver identification
  - **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 than but speeded up w/ GPUs

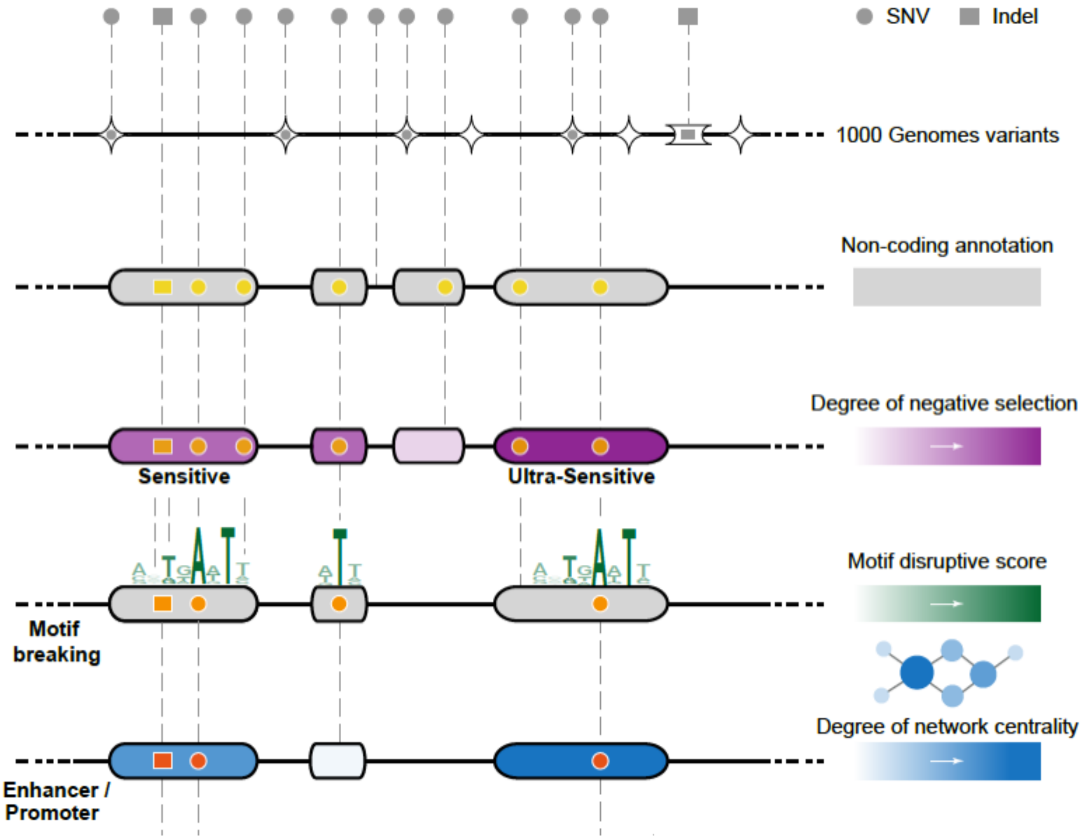
Funseq: a flexible framework to determine functional impact & use this to prioritize variants

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

Conservation (GERP, allele freq.)

Mutational impact (motif breaking, Lof)

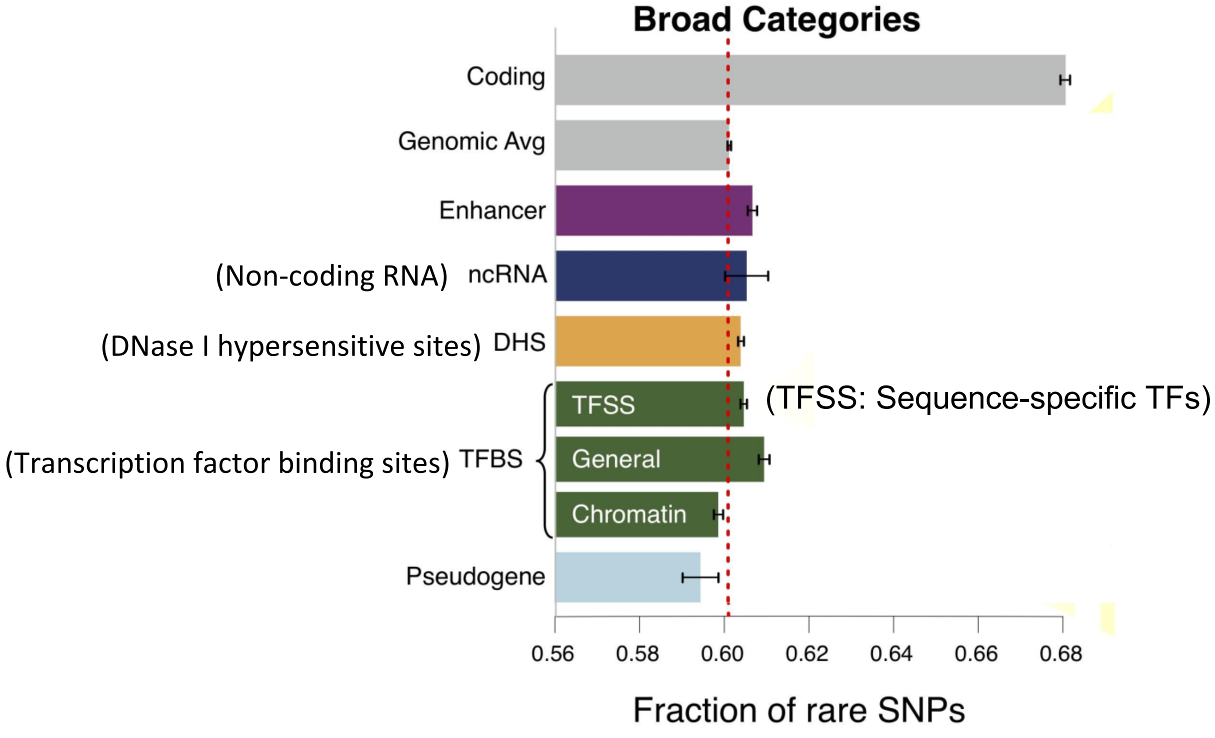
Network (centrality position)



[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

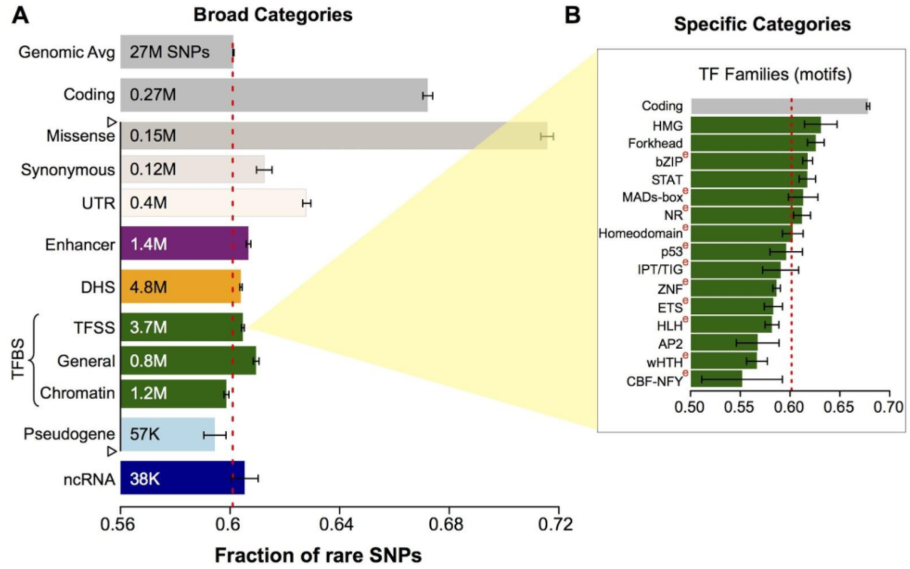


Depletion of Common Variants in the Human Population

Broad categories of regulatory regions under negative selection Related to:

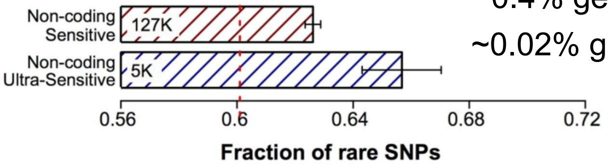
ENCODE, *Nature*, 2012
Ward & Kellis, *Science*, 2012
Mu et al, *NAR*, 2011

Differential selective constraints among specific sub-categories



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

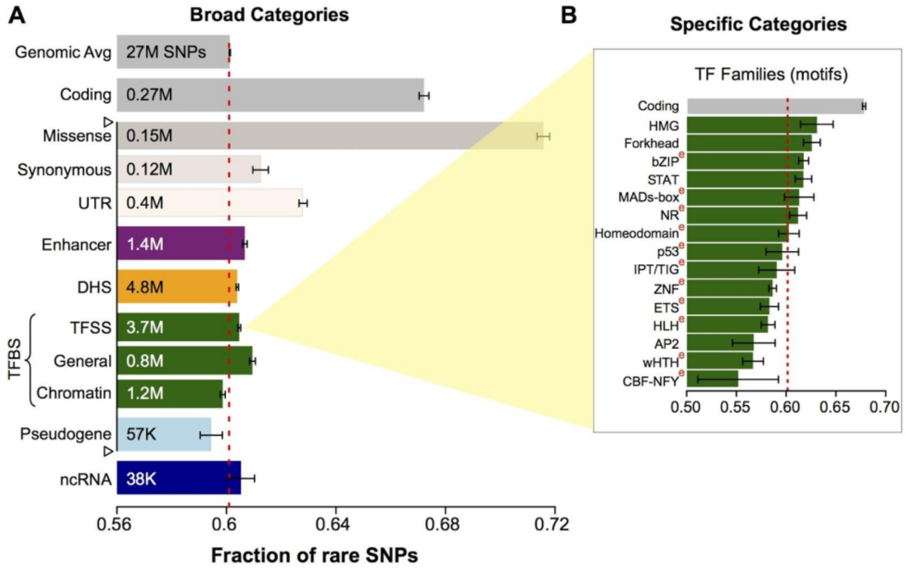
[Khurana et al., *Science* ('13)]



~0.4% genomic coverage (~ top 25)

~0.02% genomic coverage (top 5)

Defining Sensitive non-coding Regions

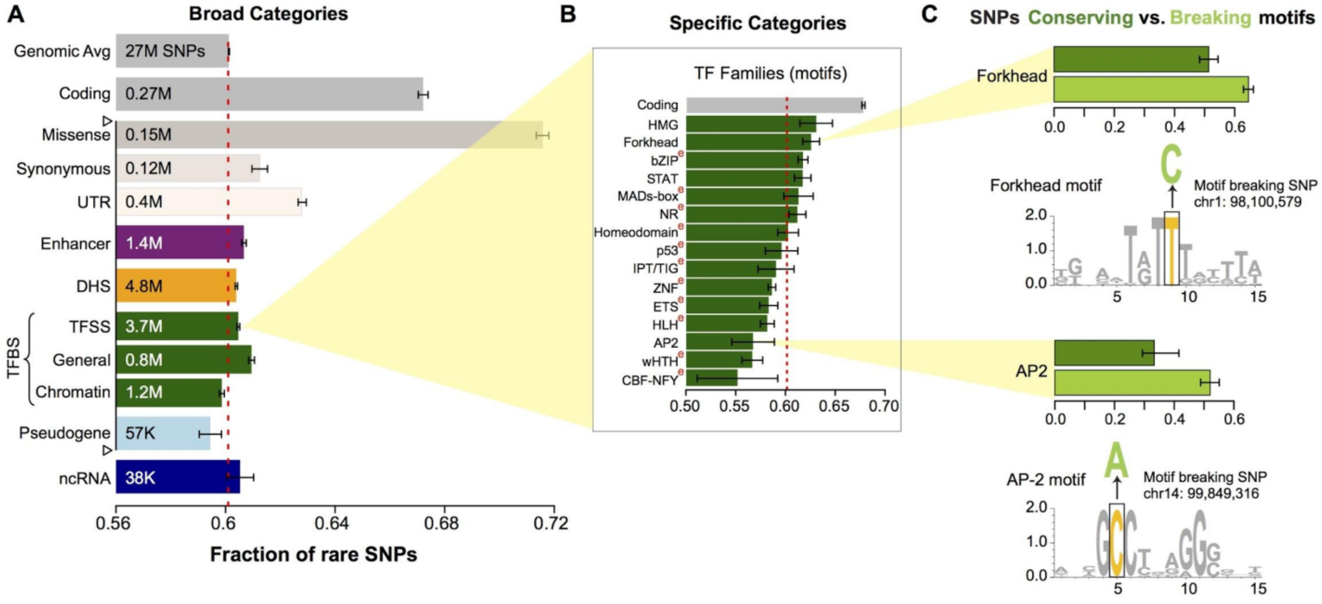


Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

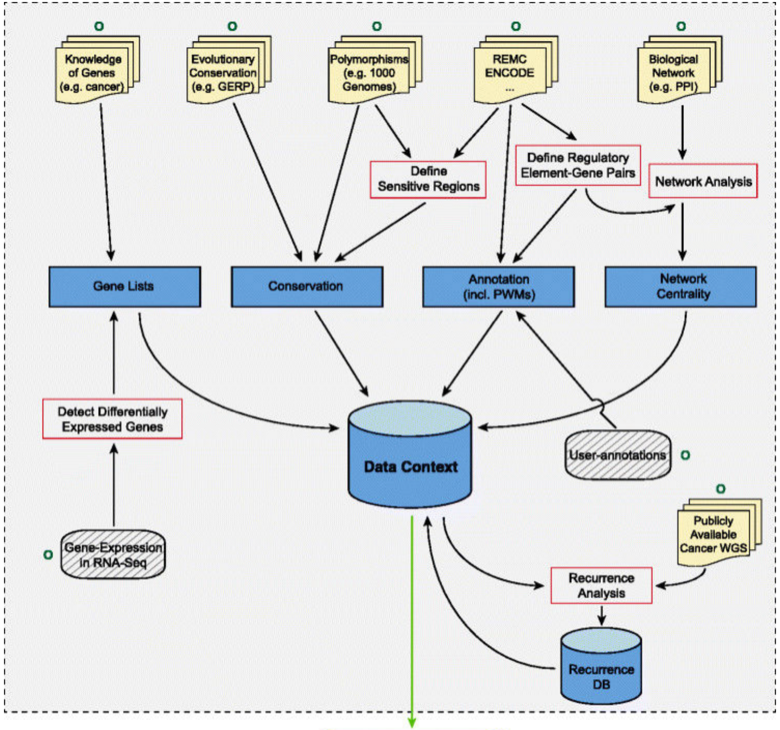
Start 677 high-resolution non-coding categories; Rank & find those under strongest selection

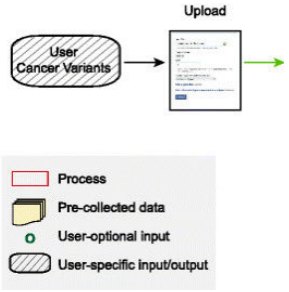
[Khurana et al., *Science* ('13)]

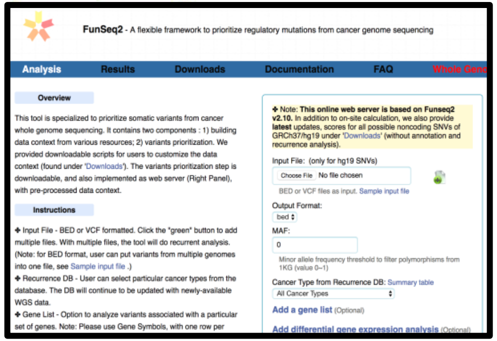
SNPs which break TF motifs are under stronger selection



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


Knowledge of Genes (e.g. cancer)
Evolutionary Conservation (e.g. GERP)
Polymorphisms (e.g. 1000 Genomes)
REMC ENCODE ...
Biological Network (e.g. PPI)
Define Sensitive Regions
Define Regulatory Element-Gene Pairs
Network Analysis
Gene Lists
Conservation
Annotation (incl. PWMs)
Network Centrality
Detect Differentially Expressed Genes
Data Context
User-annotations
Publicly Available Cancer WGS
Gene-Expression in RNA-Seq
Recurrence Analysis
Recurrence DB





FunSeq2 - A flexible framework to prioritize regulatory mutations from cancer genome sequencing

Overview

This tool is specialized to prioritize somatic variants from cancer whole genome sequencing. It contains two components : 1) building data context from various resources; 2) variants prioritization. We provided downloadable scripts for users to customize the data context (found under 'Downloads'). The variants prioritization step is downloadable, and also implemented as web server (Right Panel), with pre-processed data context.

Instructions

✚ Input File - BED or VCF formatted. Click the "green" button to add multiple files. With multiple files, the tool will do recurrent analysis. (Note: for BED format, user can put variants from multiple genomes into one file, see Sample input file .)

✚ Recurrence DB - User can select particular cancer types from the database. The DB will continue to be updated with newly-available WGS data.

✚ Gene List - Option to analyze variants associated with a particular set of genes. Note: Please use Gene Symbols, with one row per

✚ Note: **This online web server is based on Funseq2 v2.10.** In addition to on-site calculation, we also provide **latest** updates, scores for all possible noncoding SNVs of GRCh37/hg19 under 'Downloads' (without annotation and recurrence analysis).

Input File: (only for hg19 SNVs)

Choose File No file chosen

BED or VCF files as input. Sample input file

Output Format:

bed

MAF:

0

Minor allele frequency threshold to filter polymorphisms from 1KG (value 0–1)

Cancer Type from Recurrence DB: Summary table

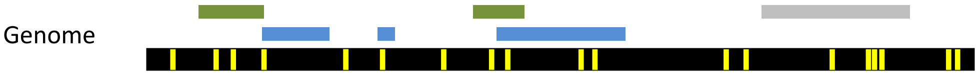
All Cancer Types

Add a gene list (Optional)

Add differential gene expression analysis (Optional)

FunSeq.gersteinlab.org

HOT region 🟩
Sensitive region 🟦
Polymorphisms 🟨



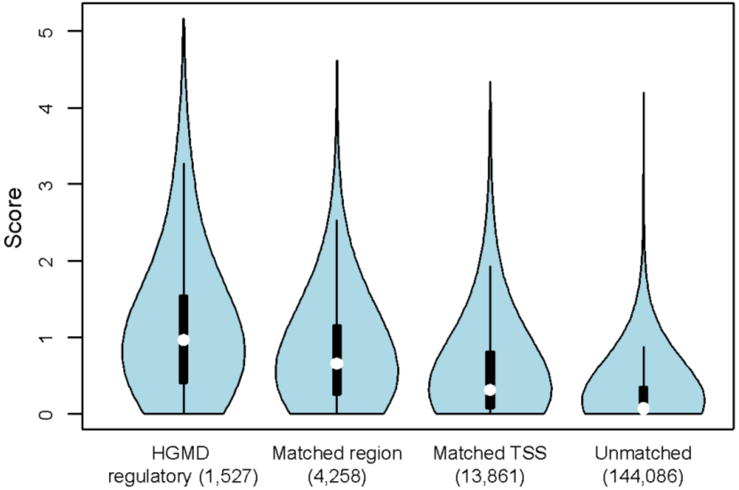
$$w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$$

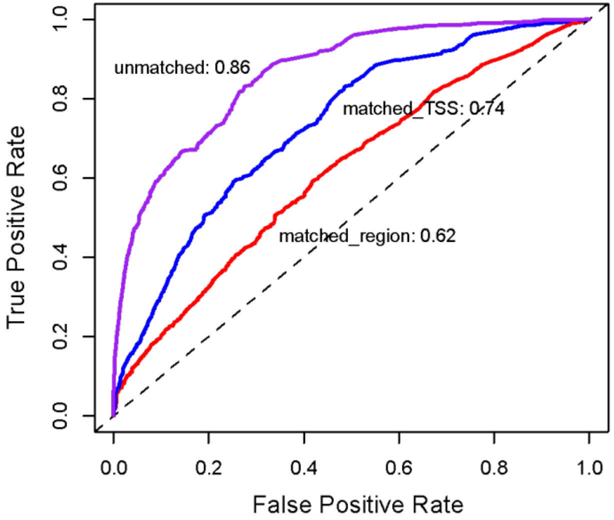
- Entropy based method for weighting consistently many genomic features

- 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





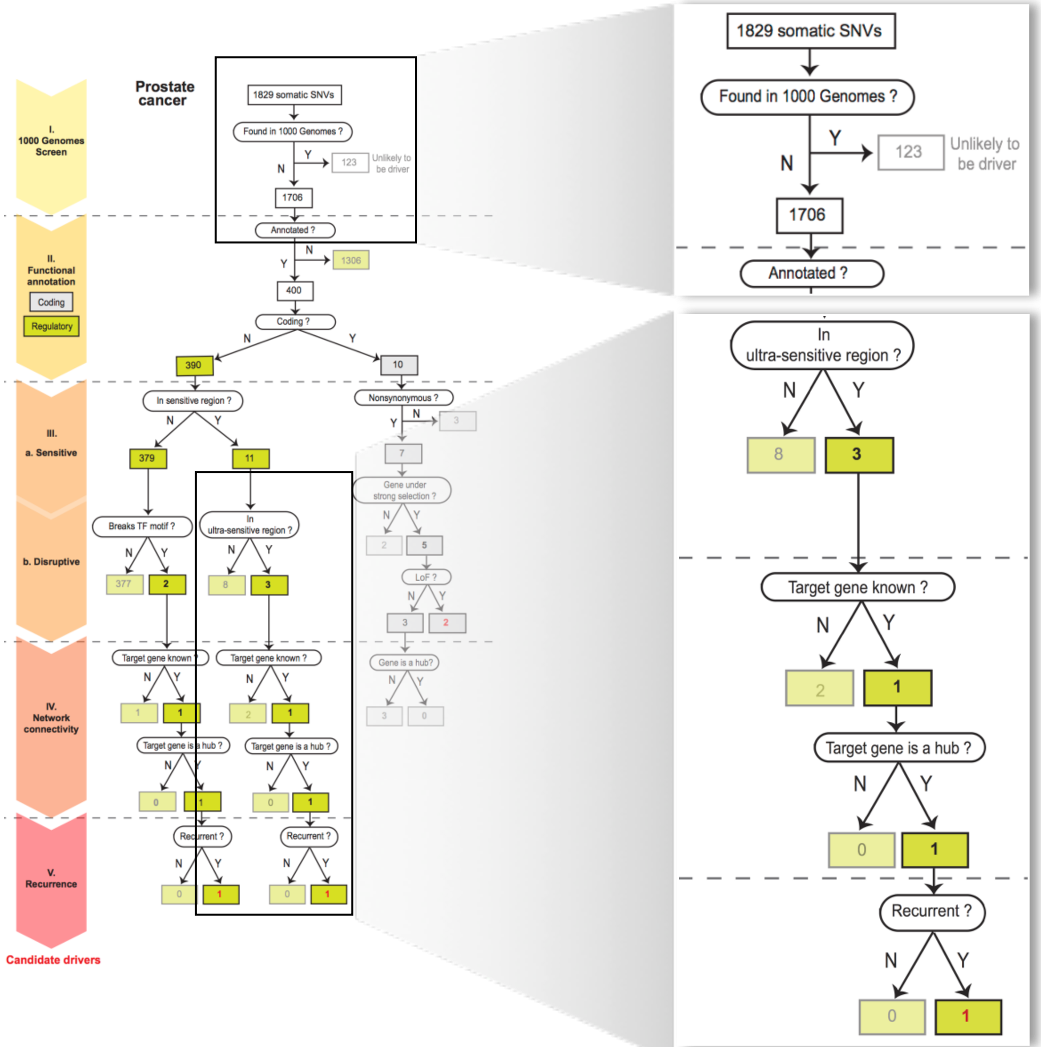
3 controls with natural polymorphisms (allele frequency >= 1%)

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)

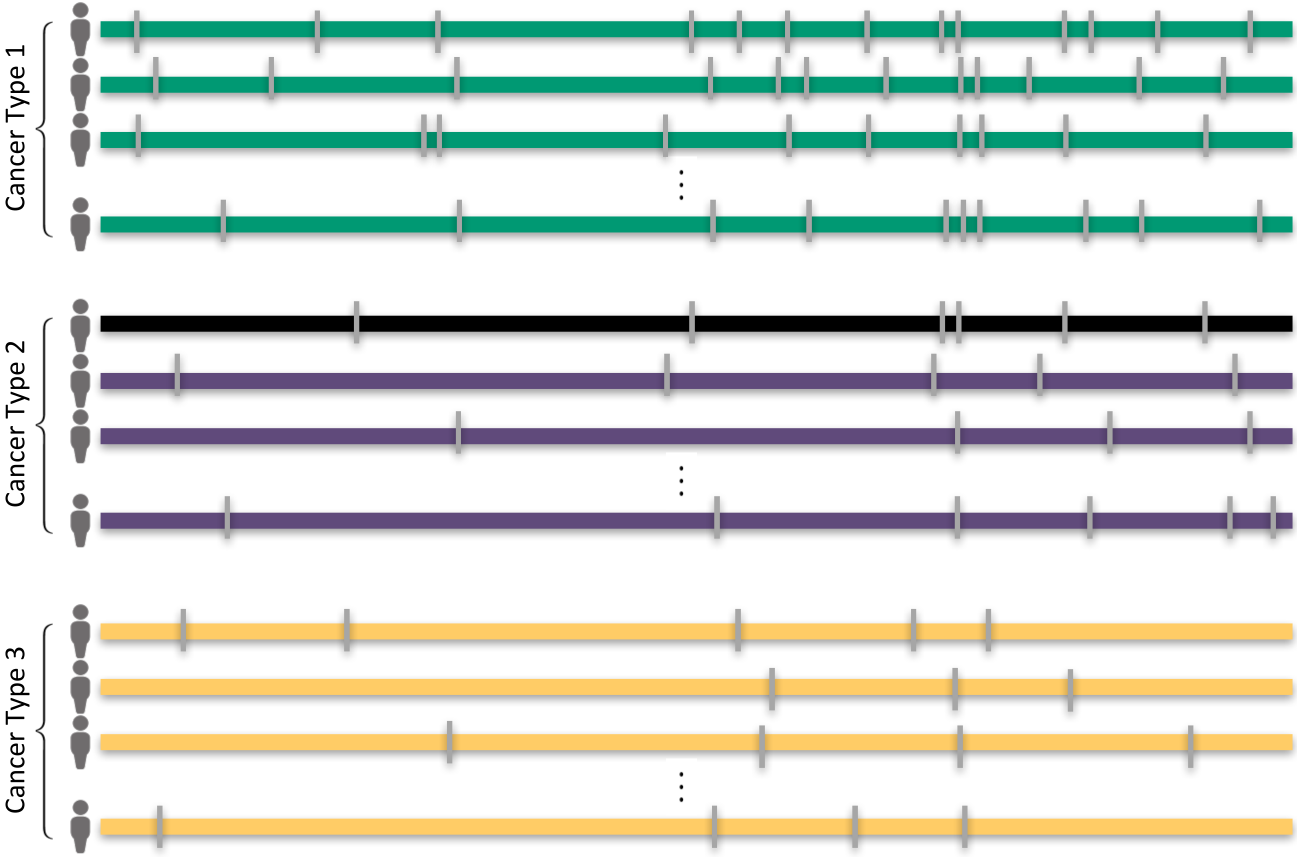


[Khurana et al., *Science* ('13)]

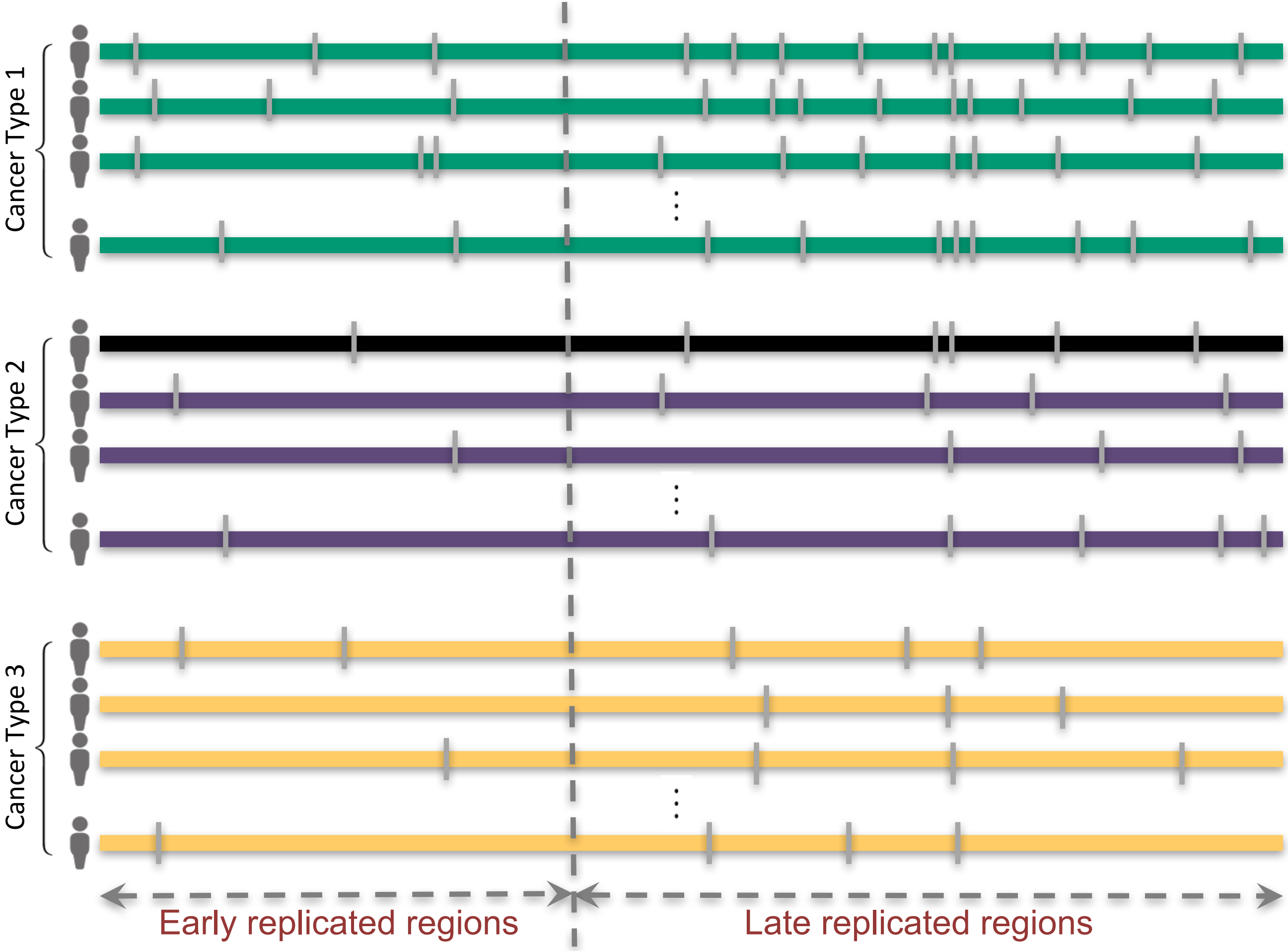
Genomics & Data Science:
Approaches to identifying key variants through functional impact & recurrence

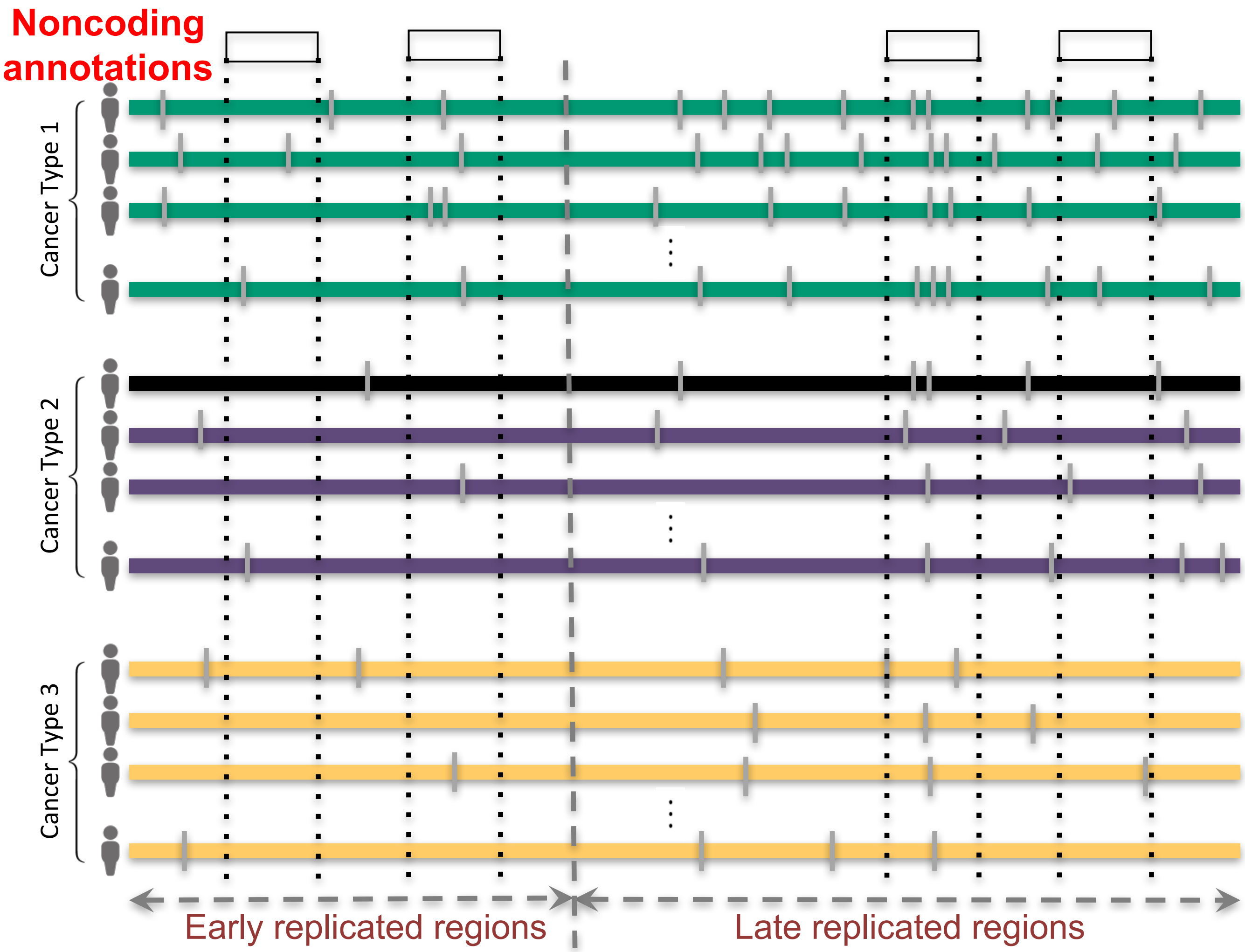
- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data generation & processing
  - Mining the data to prioritize variants for key drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - LoF annotation as a complex problem + finding deleterious LoFs
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with an entropy based weighting scheme.
  - Prioritizing rare variants with “sensitive sites” (human conserved)
- Recurrence:
  Statistics for driver identification
  - **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 than but speeded up w/ GPUs

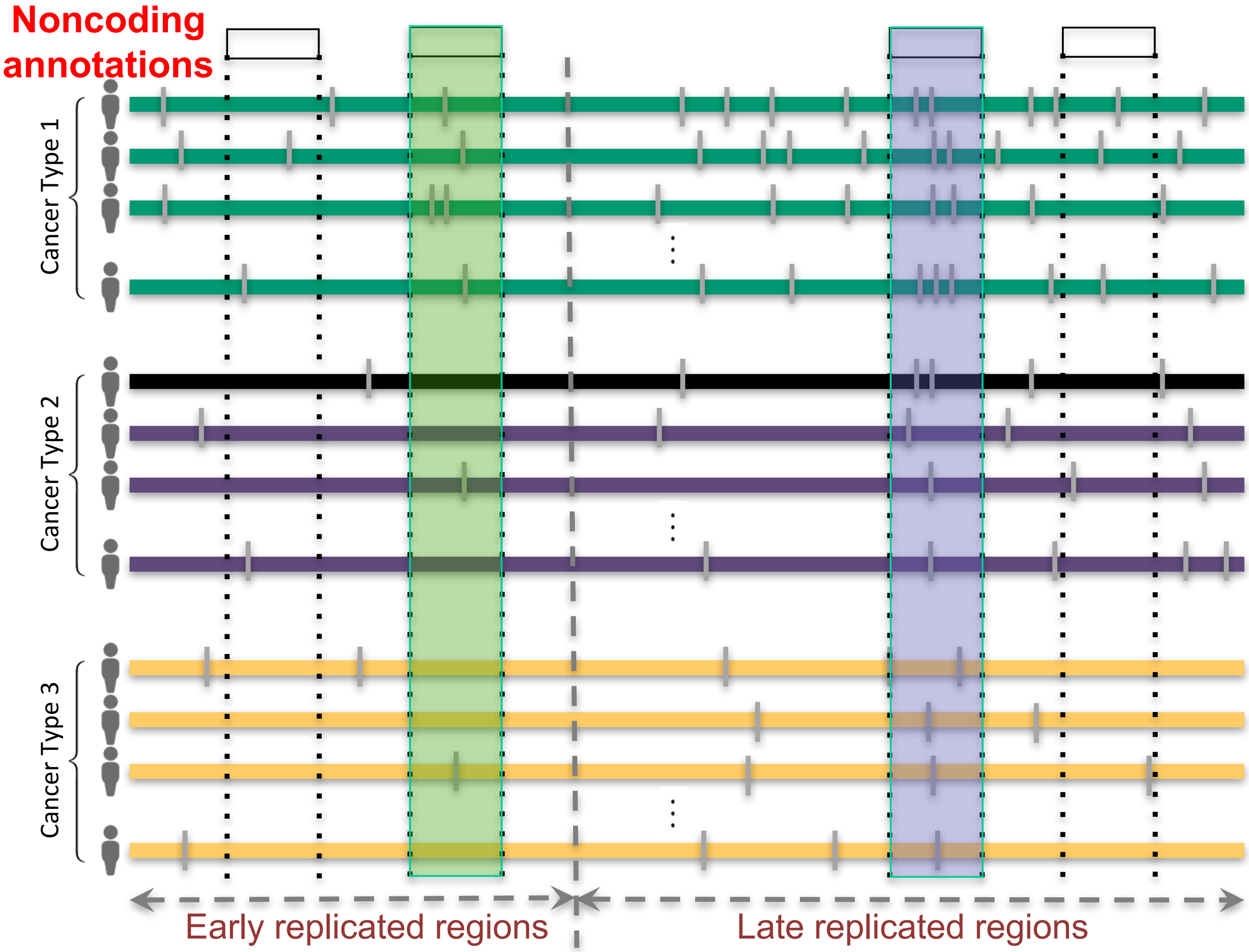
Mutation recurrence



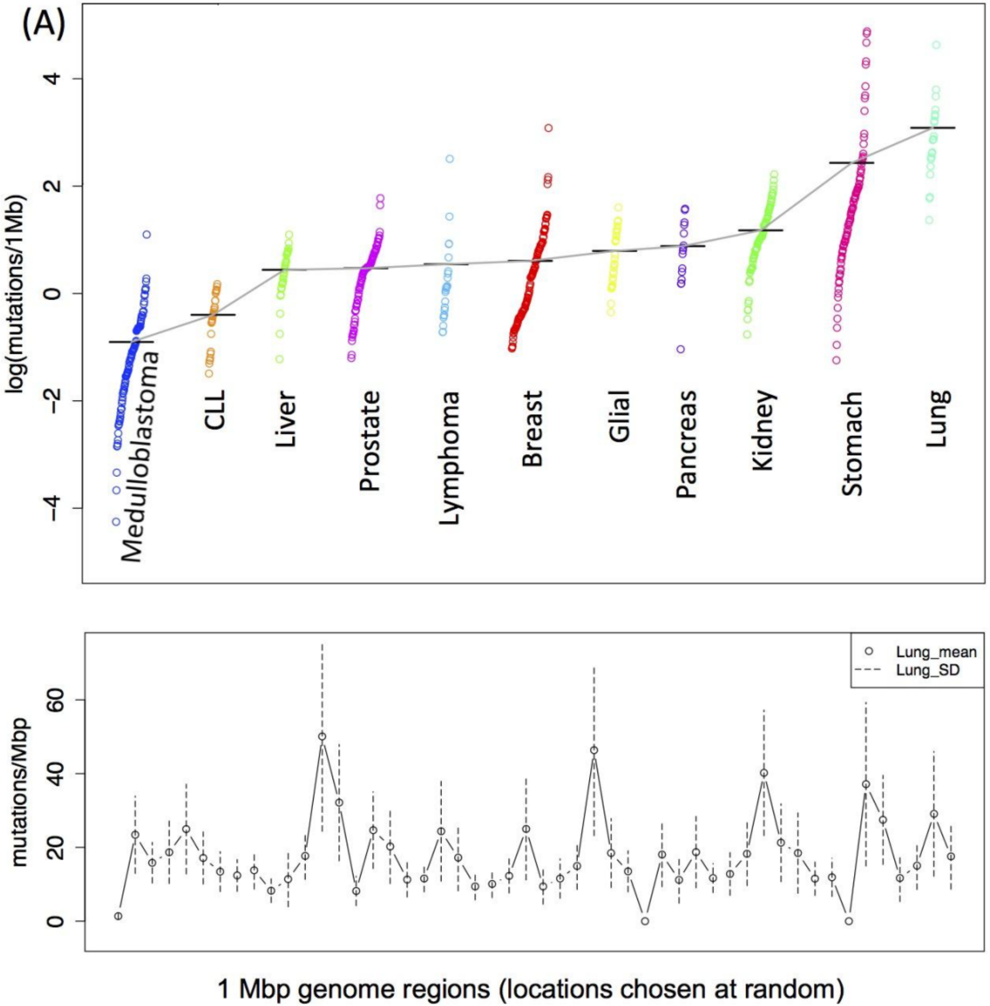
Mutation recurrence



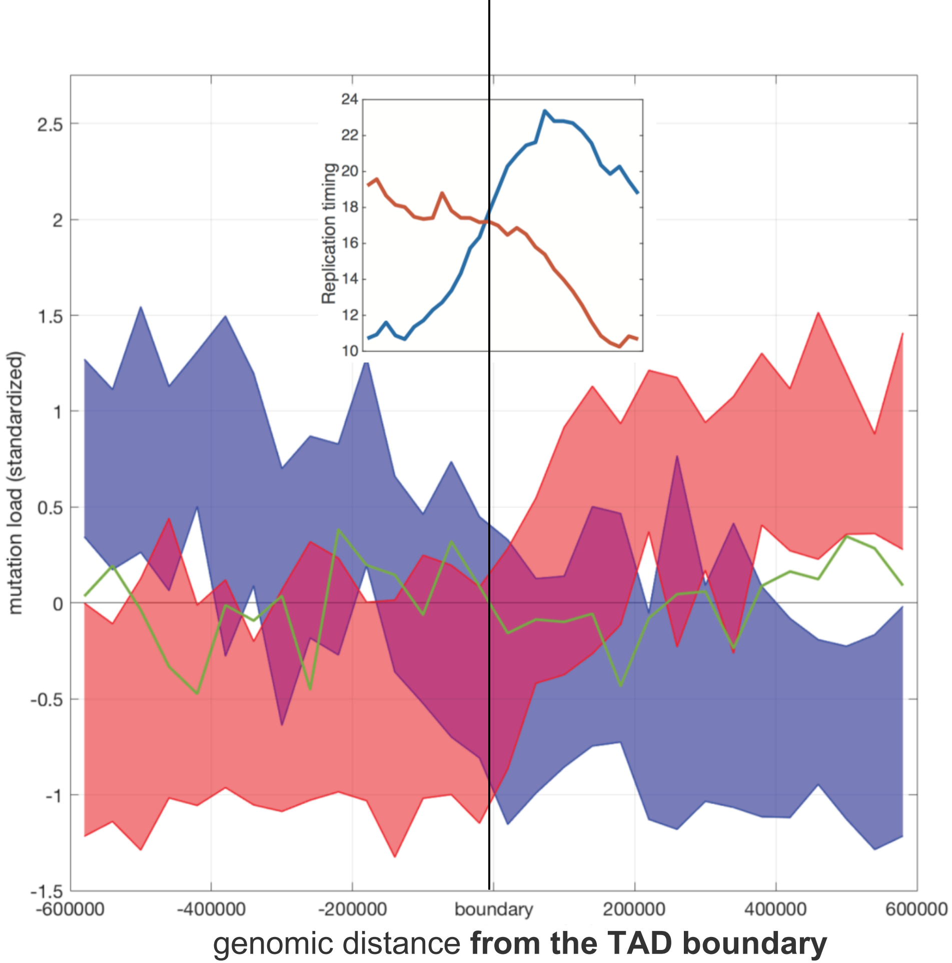

Noncoding annotations
Cancer Type 1
Cancer Type 2
Cancer Type 3
Early replicated regions
Late replicated regions


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

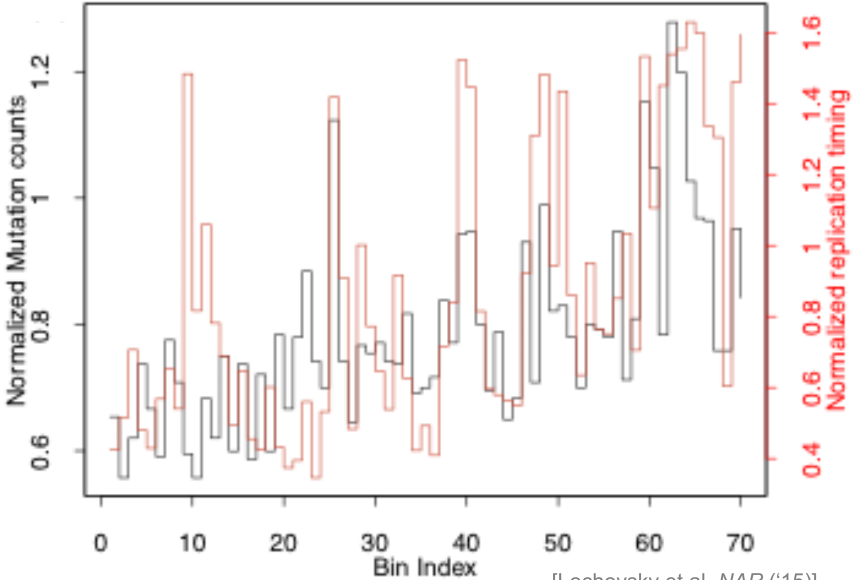
Cancer Somatic Mutational Heterogeneity, across cancer types, samples & regions



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



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



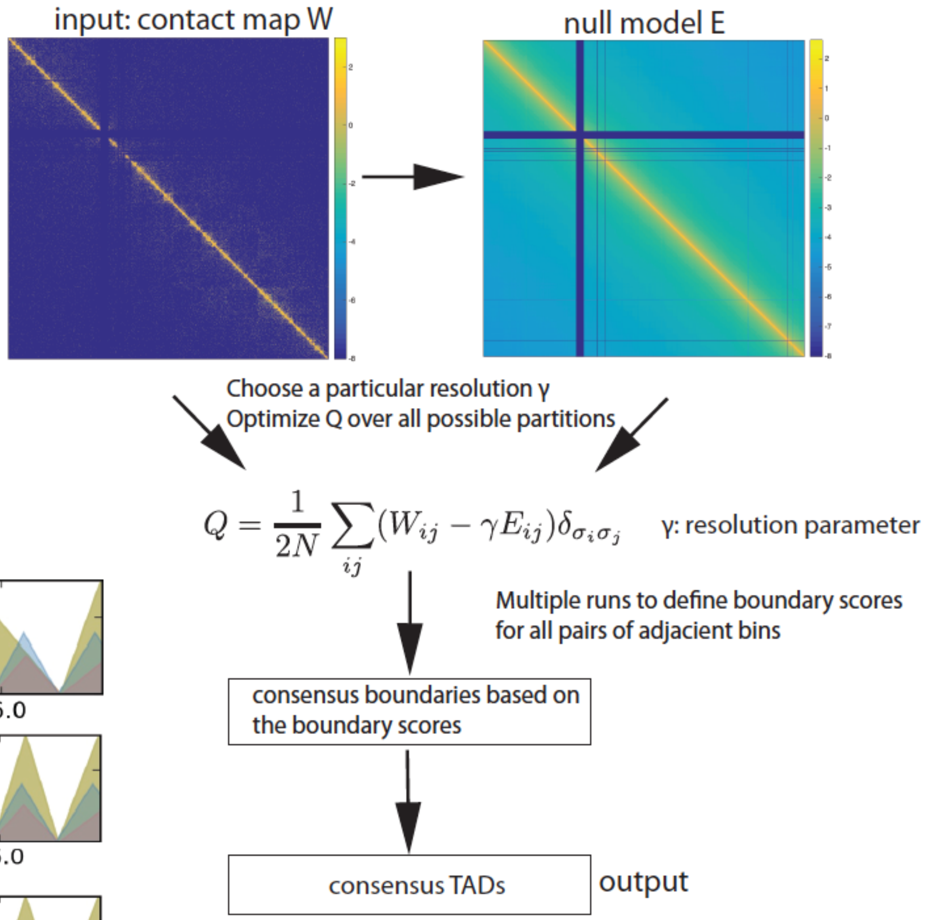
[Lochovsky et al. *NAR* ('15)]

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

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

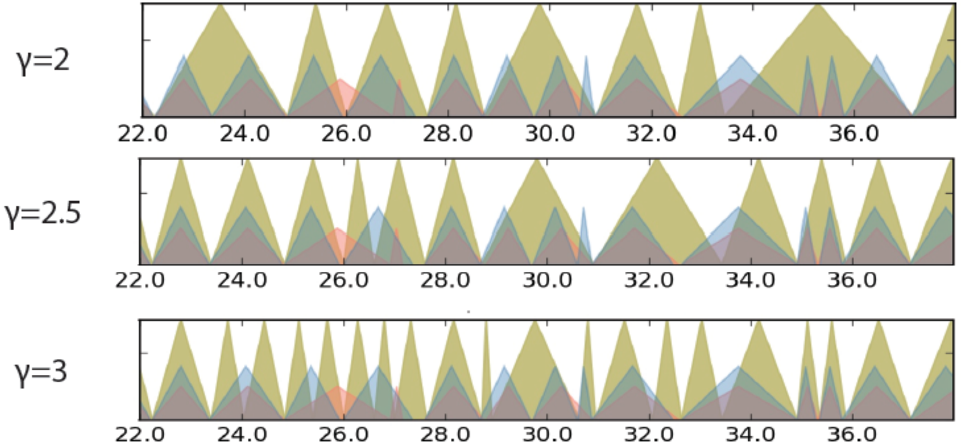
γ: resolution parameter

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

consensus boundaries based on the boundary scores

consensus TADs

output



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

Genomics & Data Science:
Approaches to identifying key variants through functional impact & recurrence

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data generation & processing
  - Mining the data to prioritize variants for key drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - LoF annotation as a complex problem + finding deleterious LoFs
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with an entropy based weighting scheme.
  - Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence:
  Statistics for driver identification
  - **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 than but speeded up w/ GPUs

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

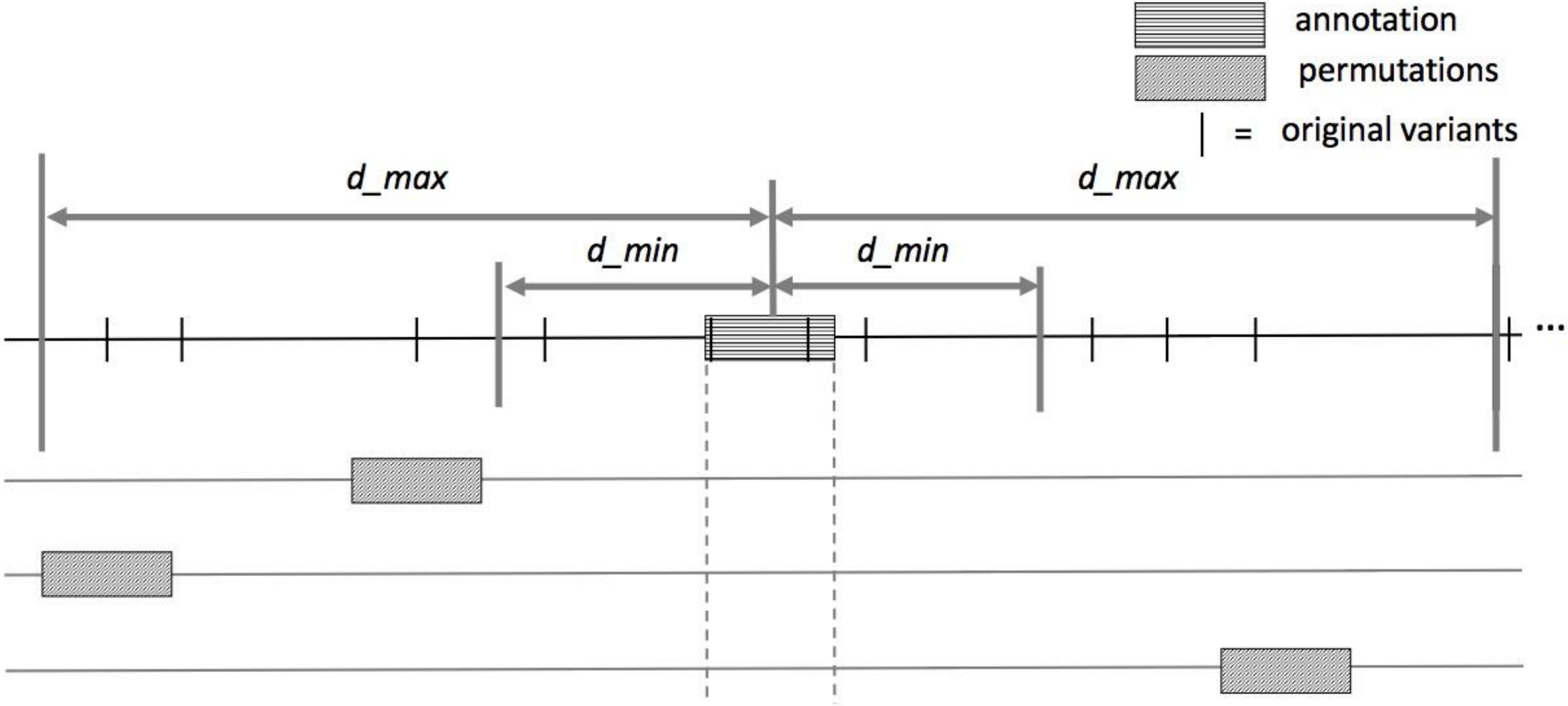
Shuffle annotations within local region to assess background mutation rate.

Model 3b: Random Permutation of Input Variants

Shuffle variants within local region to assess background mutation rate.

[Lochovsky et al. *Bioinformatics* in press]

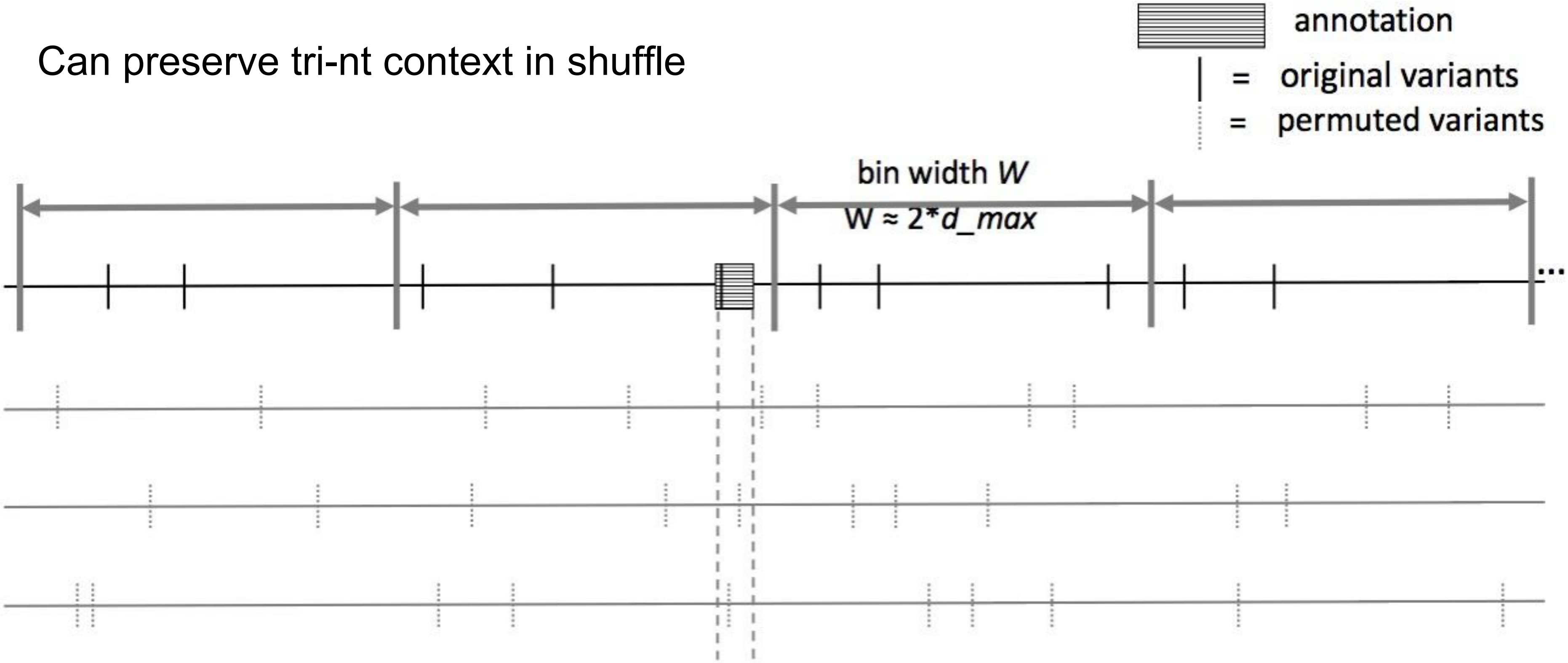
MOAT-a: Annotation-based permutation



[Lochovsky et al. *Bioinformatics* in

MOAT-v: Variant-based Permutation

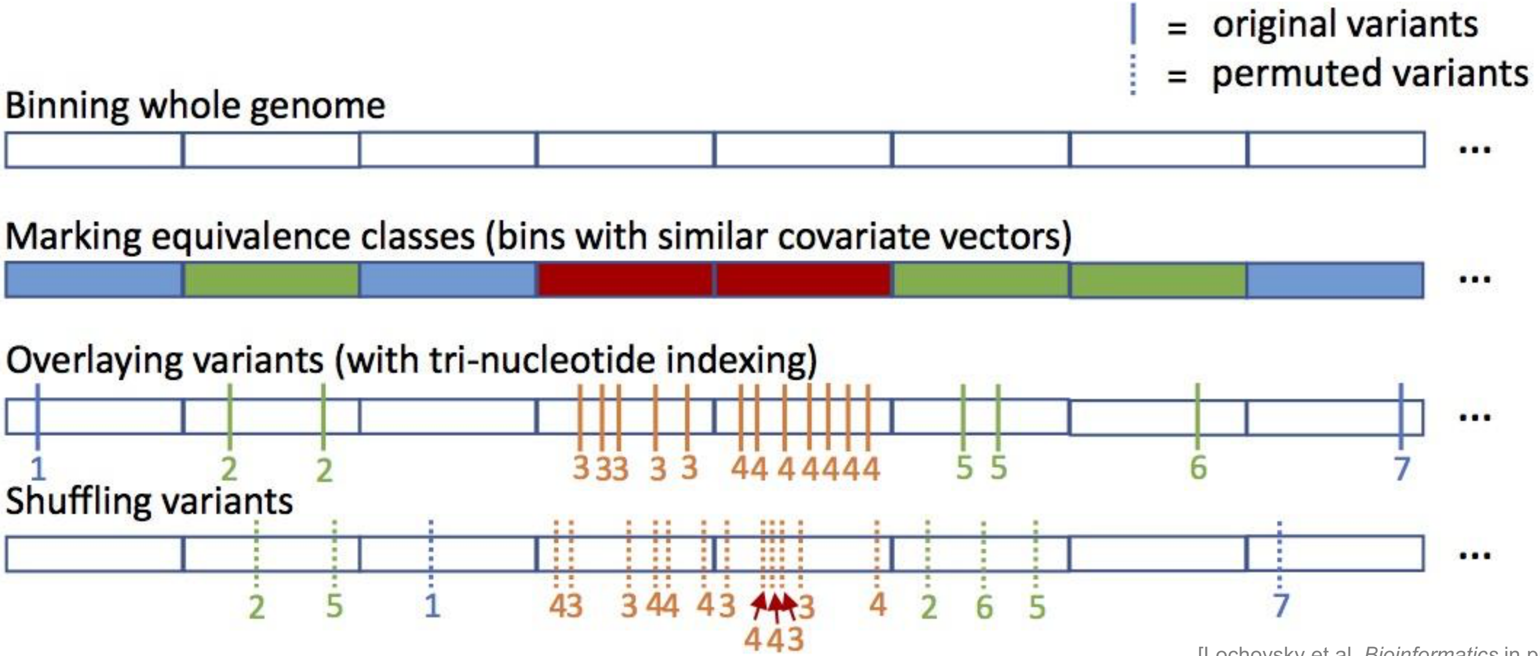
Can preserve tri-nt context in shuffle



[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



| = original variants

⋮ = permuted variants

Binning whole genome

Marking equivalence classes (bins with similar covariate vectors)

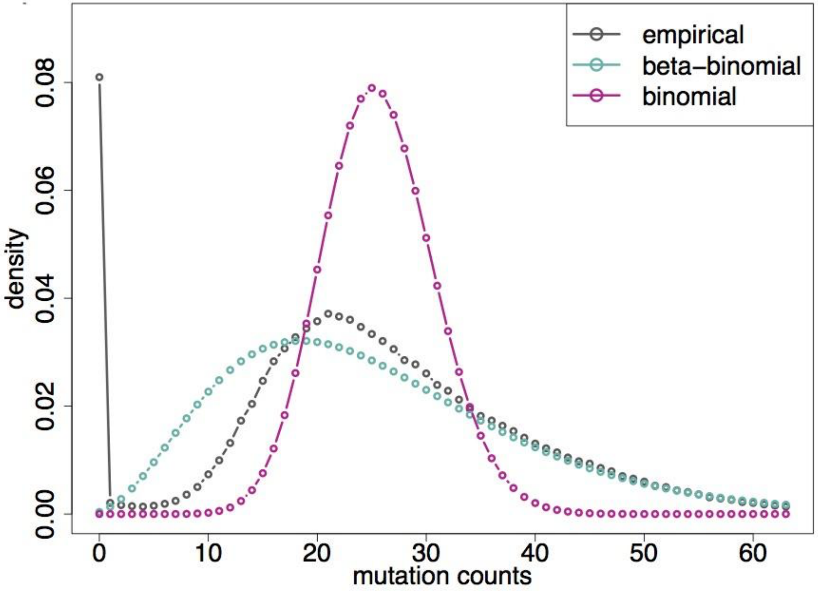
Overlaying variants (with tri-nucleotide indexing)

Shuffling variants

[Lochovsky et al. *Bioinformatics* in press]

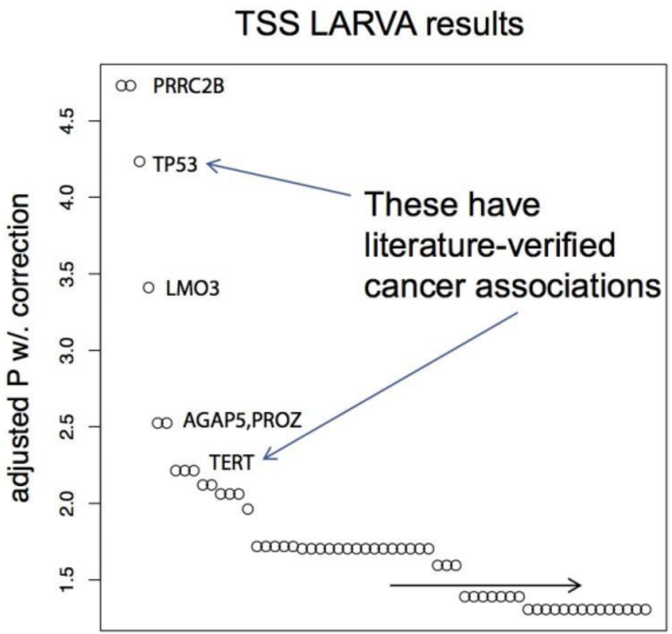
LARVA Model Comparison

- 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



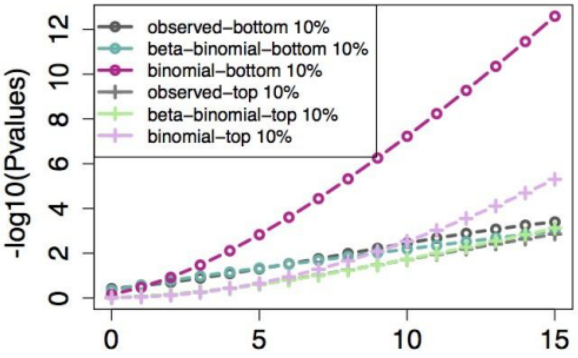
[Lochovsky et al. *NAR* ('15)]

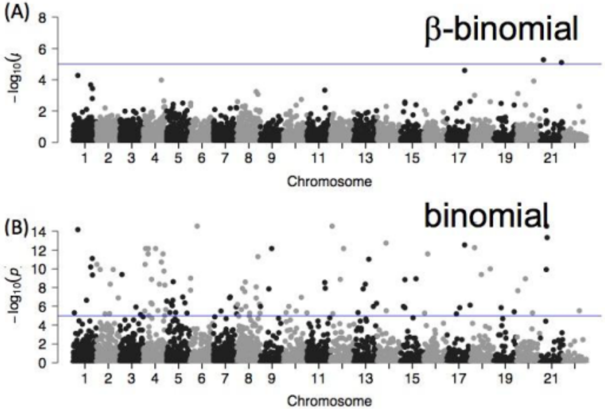
LARVA Results



These have literature-verified cancer associations

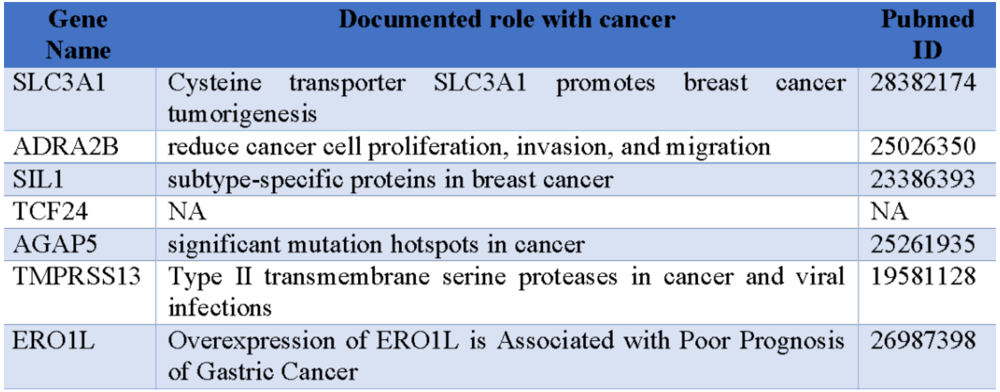
noncoding annotation p-values in sorted order





[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]

Genomics & Data Science:
Approaches to identifying key variants through functional impact & recurrence

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data generation & processing
  - Mining the data to prioritize variants for key drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - LoF annotation as a complex problem + finding deleterious LoFs
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with an entropy based weighting scheme.
  - Prioritizing rare variants with "sensitive sites" (human conserved)
- Recurrence:
  Statistics for driver identification
  - **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 than but speeded up w/ GPUs

Genomics & Data Science: Approaches to identifying key variants through functional impact & recurrence

- Introduction
  - An individual's disease variants as the public's gateway into genomics & biology
  - **The exponential scaling** of data generation & processing
  - Mining the data to prioritize variants for key drivers
- Functional impact #1: Coding
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - LoF annotation as a complex problem + finding deleterious LoFs
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
- Functional impact #2: Non-coding
  - **FunSeq** integrates evidence, with an entropy based weighting scheme.
  - Prioritizing rare variants with “sensitive sites” (human conserved)
- Recurrence:
  Statistics for driver identification
  - **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 than but speeded up w/ GPUs

Acknowledgments

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, XJ Mu, KY Yip

Hiring Postdocs.See **Jobs**.gersteinlab.org

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

MOAT.gersteinlab.org

L **Lochovsky**, J **Zhang**

Info about this talk

General PERMISSIONS

- This Presentation is copyright Mark Gerstein, Yale University, 2016.
- Please read permissions statement at
  gersteinlab.org/misc/permissions.html .
- Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org). 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