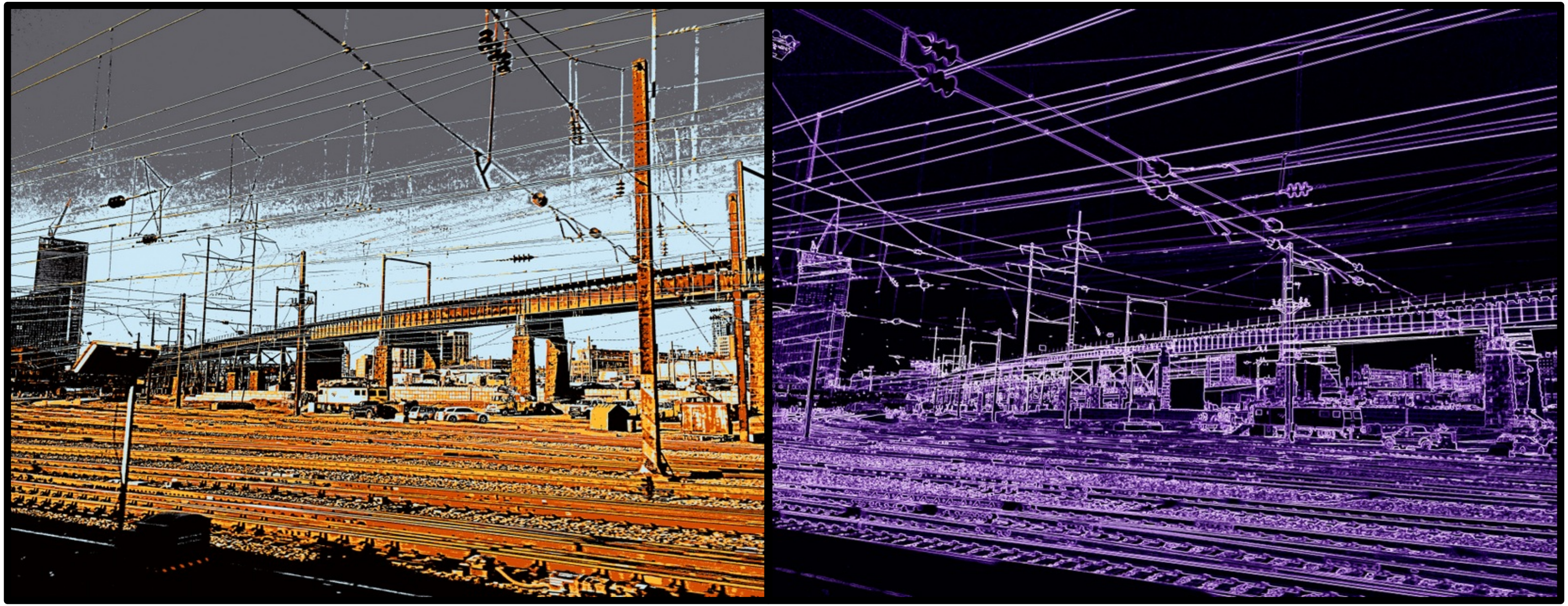
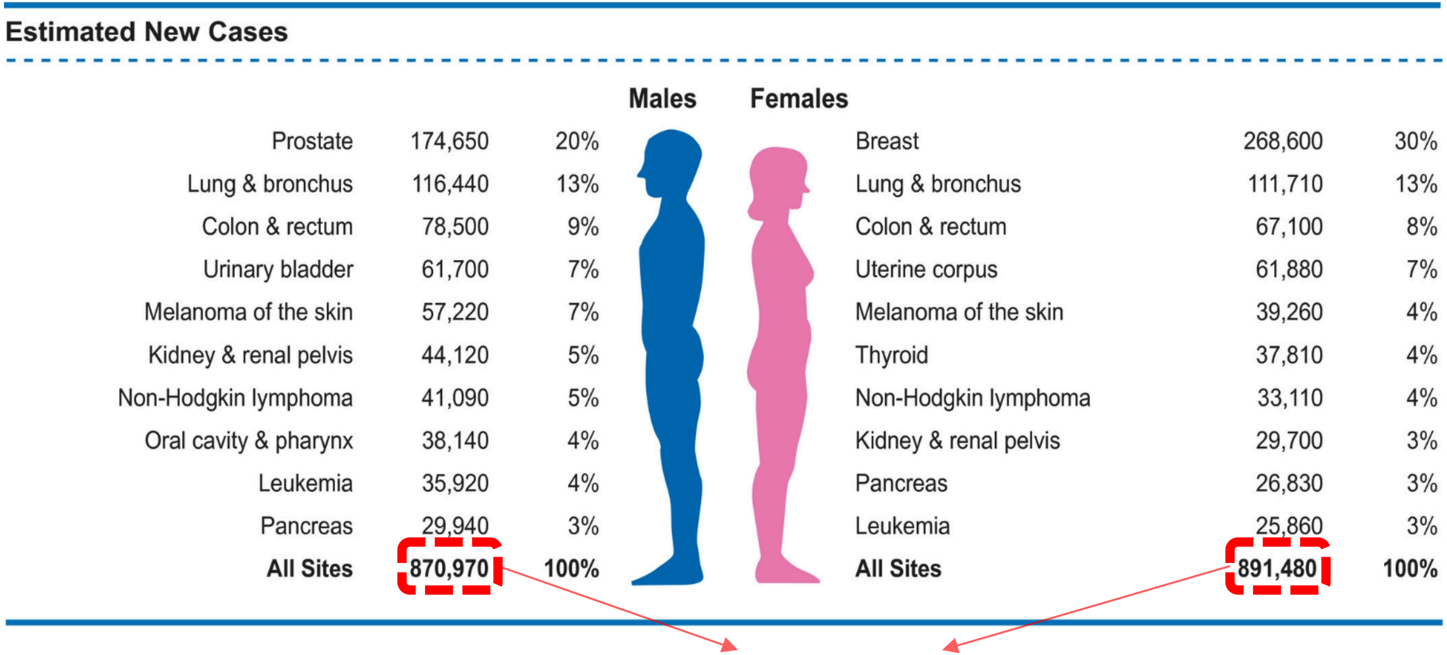
# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation & the Application of all of these to Cancer



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

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



**Estimated New Cases**

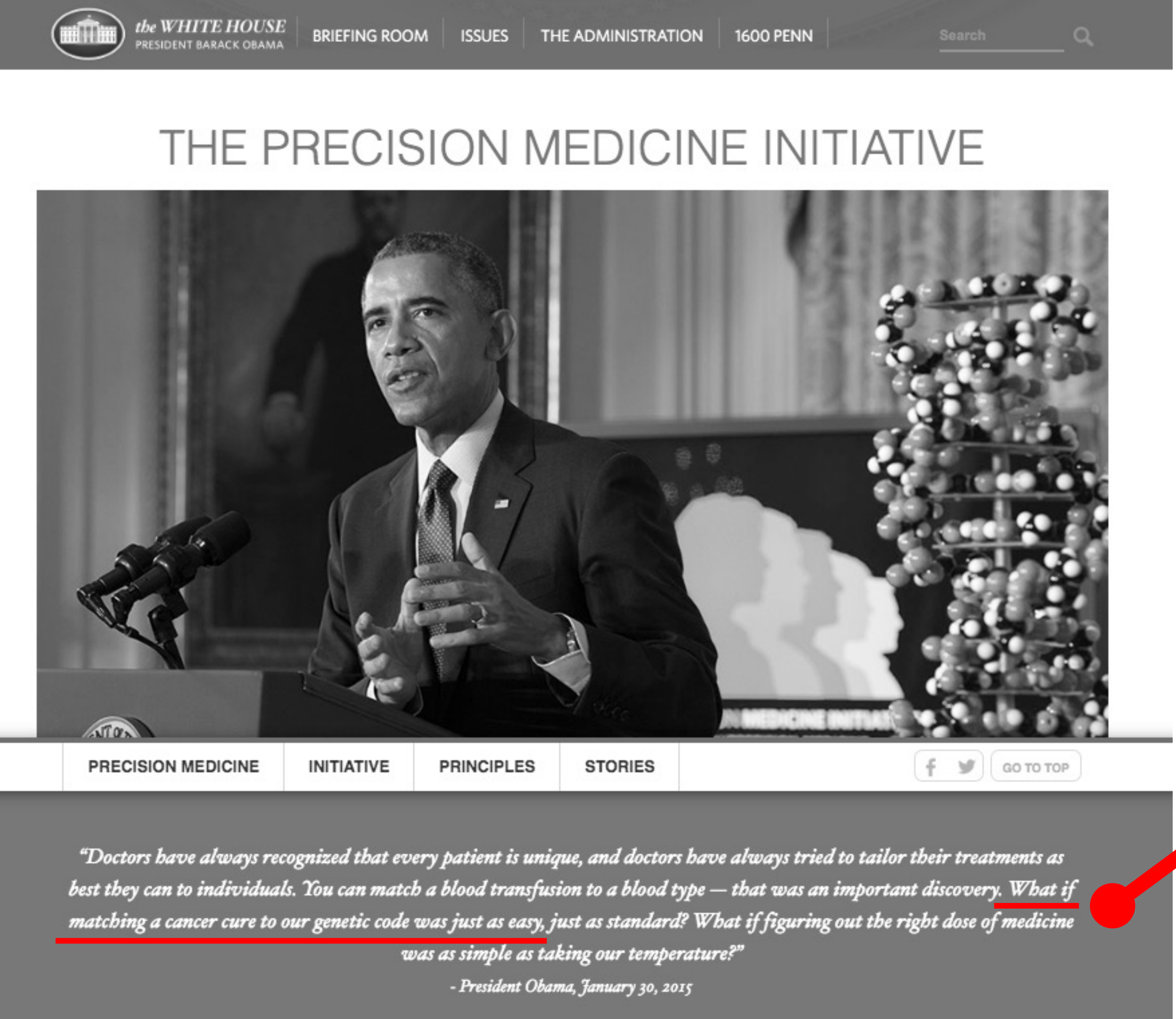
| Males | | | Females | | |
|---|---|---|---|---|---|
| Prostate | 174,650 | 20% | Breast | 268,600 | 30% |
| Lung & bronchus | 116,440 | 13% | Lung & bronchus | 111,710 | 13% |
| Colon & rectum | 78,500 | 9% | Colon & rectum | 67,100 | 8% |
| Urinary bladder | 61,700 | 7% | Uterine corpus | 61,880 |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

1,762,450 new cases per year

~4,800 new cases per day

*Segiel et al, Cancer statistics, 2019*

# Much Interest in Precision Oncology



THE PRECISION MEDICINE INITIATIVE

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

*- President Obama, January 30, 2015*

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

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

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

# Overall Problem:
# Finding Key Variants in Personal Genomes

CAN YOU FIND THE PANDA?

**Millions** of variants in a personal genome
**Thousands**, in a cancer genome
Different **contexts** for prioritization

In **rare disease**, only a few
high-impact variants are associated with disease

In **cancer**, a few positively selected drivers amongst many passengers

In **common disease**, more variants associated & each has weaker effect,
But one wants to find key ”functional” variant amongst many in LD

# Overall Problem: Finding Key Variants in Personal Genomes

**Millions** of variants in a personal genome
**Thousands**, in a cancer genome
Different **contexts** for prioritization

In **rare disease**, only a few
high-impact variants are associated with disease

CAN YOU FIND THE PANDA?

In **cancer**, a few positively selected drivers amongst many passengers

In **common disease**, more variants associated & each has weaker effect,
But one wants to find key ”functional” variant amongst many in LD

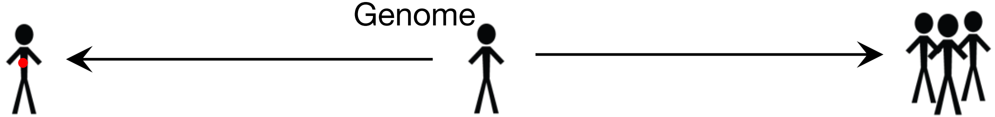
**Thus: Need to find & prioritize high impact variants.**
**Particularly hard for non-coding regions.**

# Human Genetic Variation

A Cancer Genome

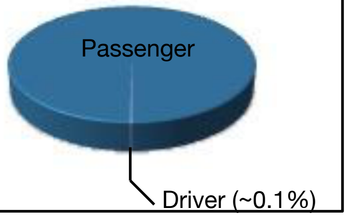
A Typical Genome

Population of 2,504 peoples



Origin of Variants

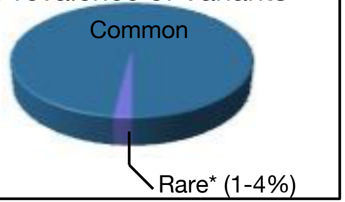
| | Coding | Non-coding |
|---|---|---|
| Germ-line | 22K | 4.1 – 5M |
| Somatic | ~50 | 5K |



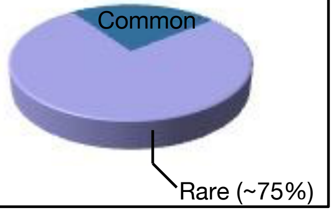
Class of Variants

| | |
|---|---|
| SNP | 3.5 – 4.3M |
| Indel | 550 – 625K |
| SV | 2.1 – 2.5K<br>(20Mb) |
| Total | 4.1 – 5M |

Prevalence of Variants



| | |
|---|---|
| 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.

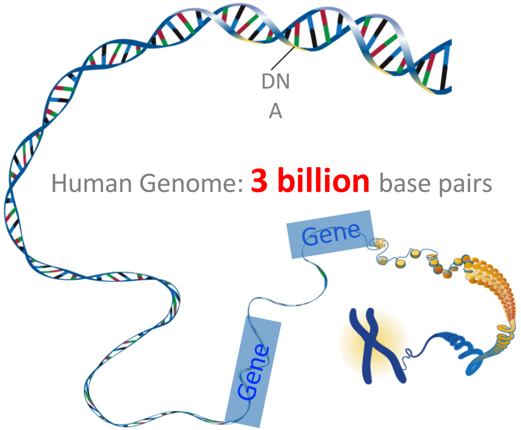
**The 1000 Genomes Project Consortium**, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

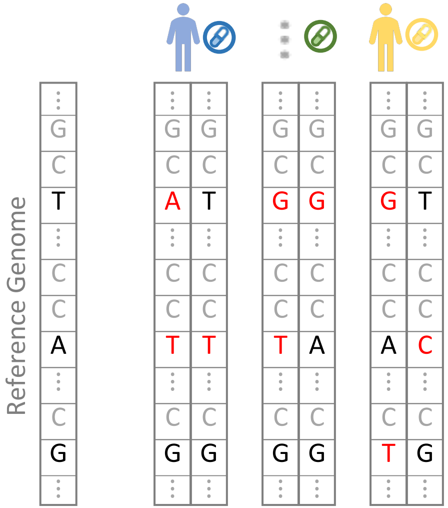
- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a "surprisal" based weighting scheme.
  - Prioritizing variants within "sensitive sites" (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1



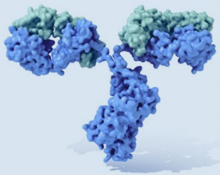
Human Genome: **3 billion** base pairs



| Reference Genome | | | | | | |
|---|---|---|---|---|---|---|
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| G | G | G | G | G | G | G |
| C | C | C | C | C | C | C |
| T | A | T | G | G | G | T |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| C | C | C | C | C | C | C |
| C | C | C | C | C | C | C |
| A | T | T | T | A | A | C |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| C | C | C | C | C | C | C |
| G | G | G | G | G | T | G |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

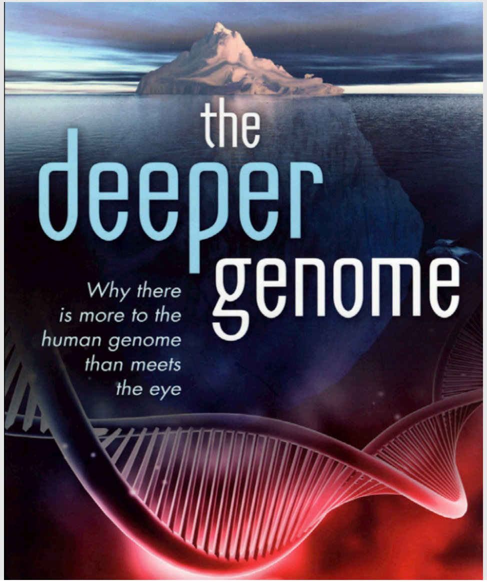
*Image adapted from NHGRI*

Protein Coding Regions:
Part of the genome we can “see”
< 2% of the genome



**The Noncoding Regions:** Dark Matter in the Genome

- >98% of the genome
- Host ~90% of disease risk loci
- contains extensive regulatory information

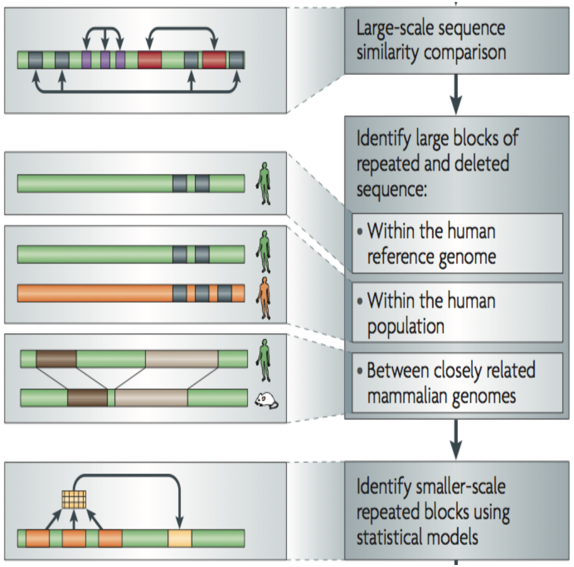


*Greenbaum & Gerstein, Cell 15’*

# Non-coding Annotations: Overview

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

Sequence features, incl. **Conservation**



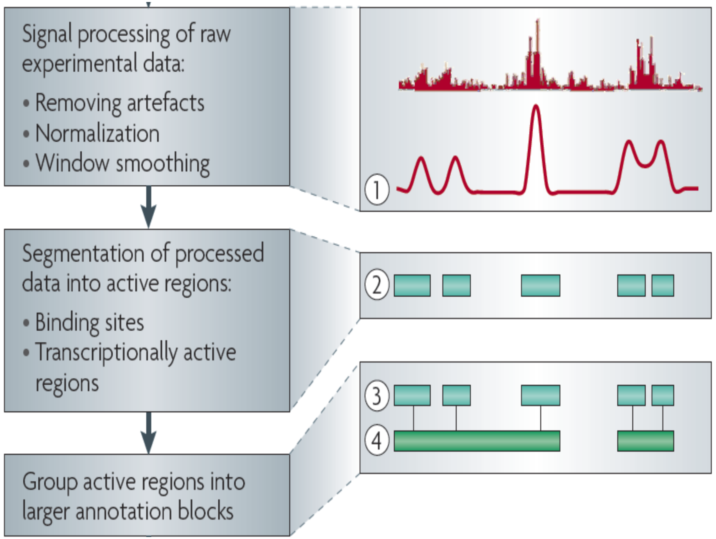
Large-scale sequence similarity comparison

Identify large blocks of repeated and deleted sequence:

- Within the human reference genome
- Within the human population
- Between closely related mammalian genomes

Identify smaller-scale repeated blocks using statistical models

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



Signal processing of raw experimental data:

- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:

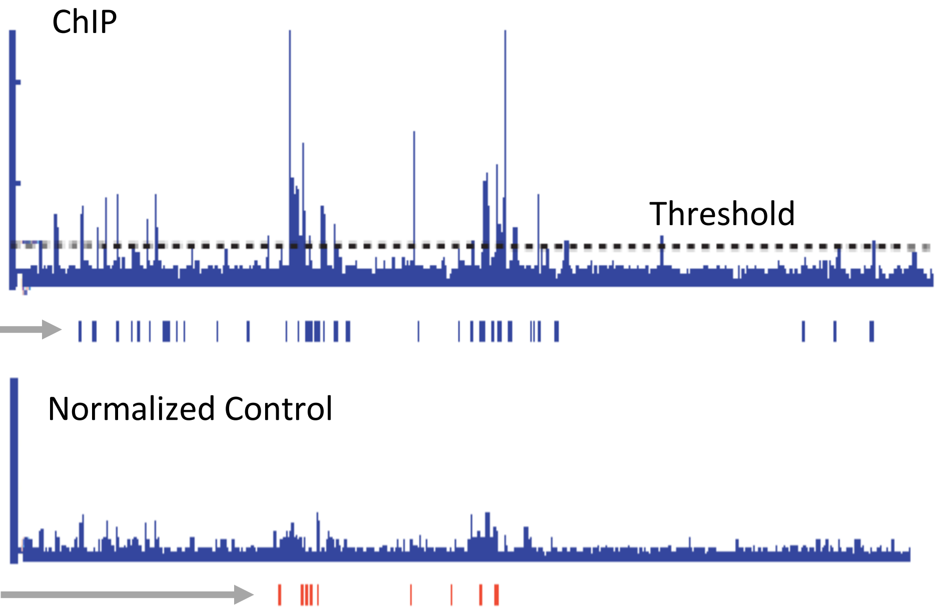
- Binding sites
- Transcriptionally active regions

Group active regions into larger annotation blocks

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

# Summarizing the Signal: "Traditional" ChipSeq Peak Calling

- Generate & threshold the signal profile to identify candidate target regions
  - Simulation (PeakSeq),
  - Local window based Poisson (MACS),
  - Fold change statistics (SPP)



Potential Targets

- Score against the control

Significantly Enriched targets

Now an update: "PeakSeq 2" => MUSIC

[Rozowsky et al. ('09) *Nat Biotech*]

# Background on computationally annotation

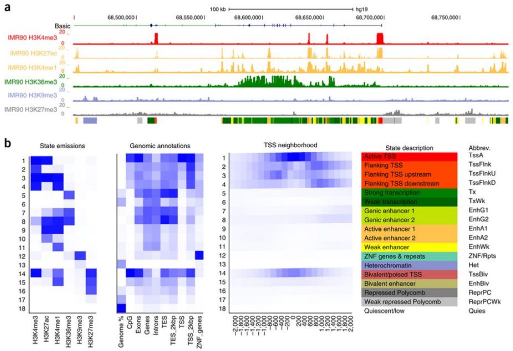
- **Peak calling**:
  - ✓PeakSeq, SPP, MACS2, Hotspot …
  - ✓ENCODE Encyclopedia

- **Genome segmentation:** partition the genome into regions (states) with distinct epigenomic profiles, then assign each state a functional label.
  - ✓ChromHMM: Multivariate Hidden Markov Model
  - ✓Segway: Dynamic Bayesian Network Model

- Supervised regulatory prediction: learn predictive models from labeled dataset of regulatory elements.
  - ✓ CSI-ANN: Time-Delay Neural Network
  - ✓ RFECS: Random Forest
  - ✓ DEEP: Ensemble SVM + Artificial Neural Network
  - ✓ REPTILE: Random Forest
  - ✓ gkm-SVM: Gapped k-mer

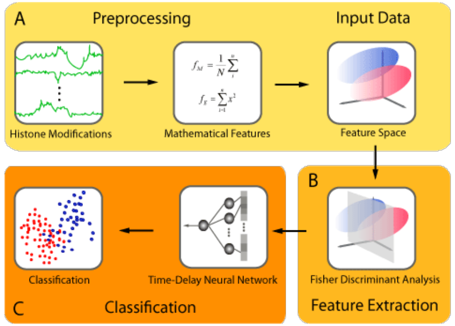
- **Target finding**
  - ✓ Ripple, TargetFinder, JEME, PreSTIGE, IM-PET

ChromHMM



J. Ernst, M. Kellis. *Nat. Protoc., 2017*

CSI-ANN

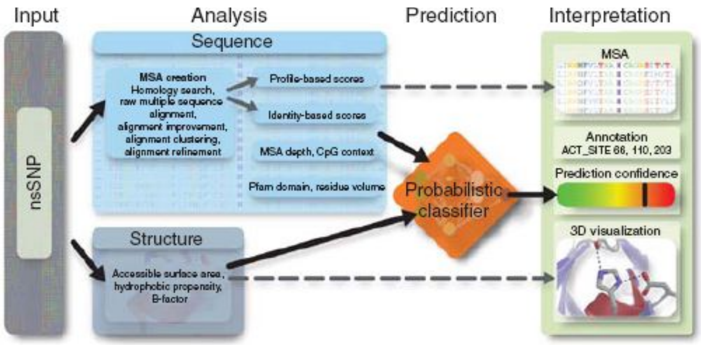


H.A. Firpi, D. Ucar, K. Tian. Bioinformatics, 2010

# Genetic variant annotation: coding and noncoding

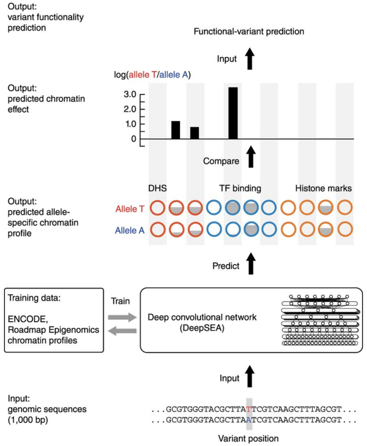
- Tools developed specifically for coding variants:
  - ✓PolyPhen-2
  - ✓SnpEff
  - ✓ SIFT
  - ✓...
- Tools developed specifically for noncoding variants:
  - ✓RegulomeDB
  - ✓HaploReg
  - ✓DeepSEA
  - ✓GWAVA
  - ✓...
- Tools for both coding and noncoding variants:
  - ✓CADD
  - ✓ANNOVAR
  - ✓VEP
  - ✓FATHMM-MKL
  - ✓ ….

Polyphen-2



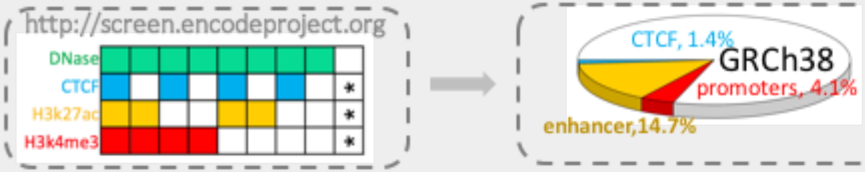
I.A. Adzhubei, *et al. Nat. Methods*, *2010*

DeepSEA



J. Zhou, O.G. Troyanskaya, Nat. Methods, 2015

# Major takeaway from annotation experience for disease studies: ***<span>less is more</span>***



- 1.97M elements
- anchored on DHS sites
- 20.1% of the genome

Individual $\xrightarrow{\text{Genotype}}$ Cohorts

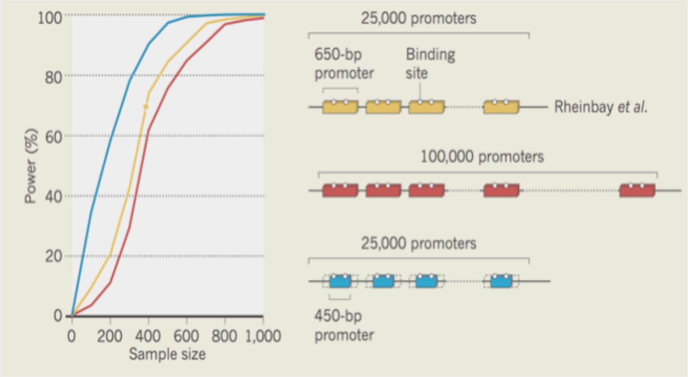
| | |
|---|---|
| SNP | 3.5 – 4.3M |
| Indel | 550 – 625K |
| SV | 2.1 – 2.5K |
| Total | 4.1 – 5M |

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

V.S.

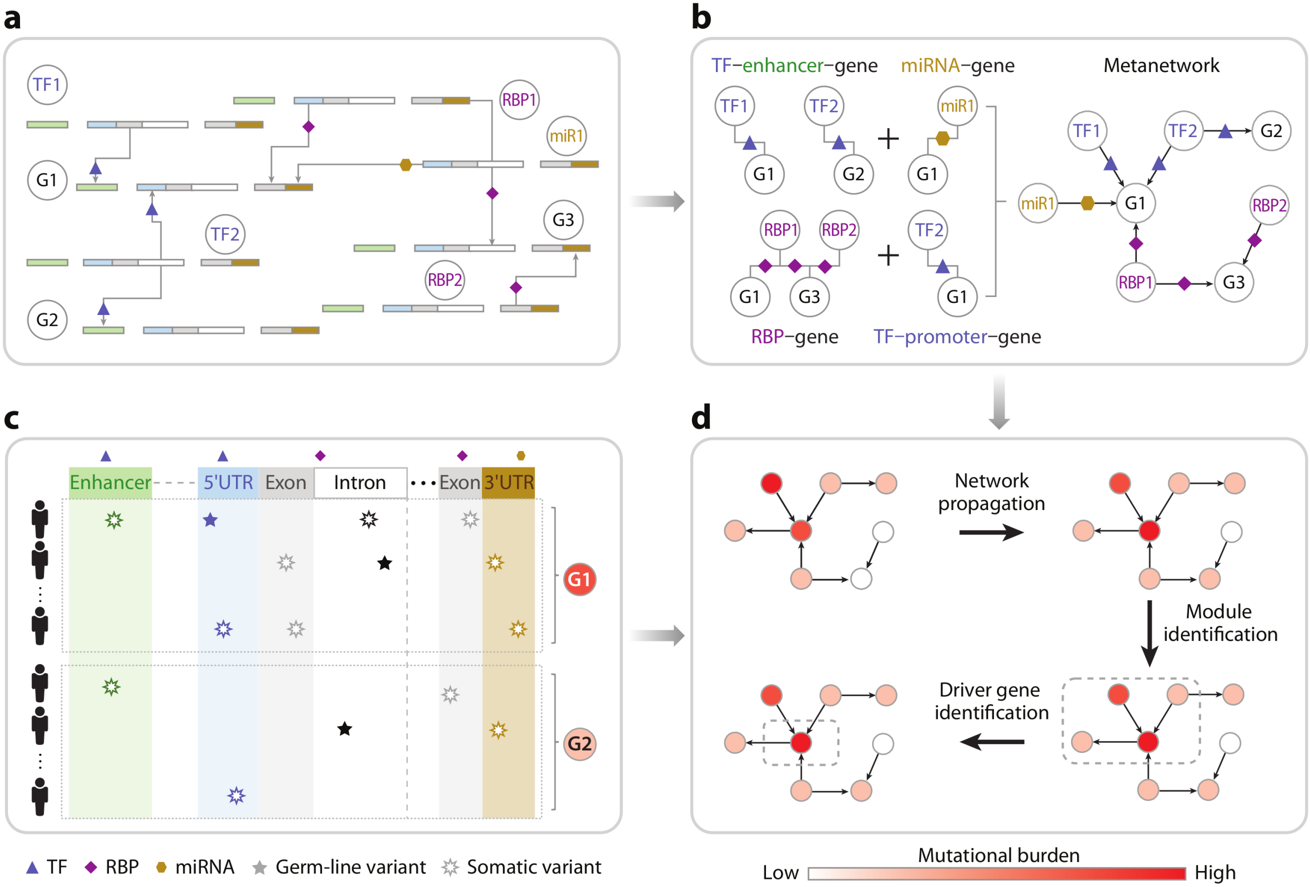
| Disease | Scale |
|---|---|
| rare | a few with high impact |
| common | many with weak effect |
| cancer | a few drivers |

Example of power issue in disease studies



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

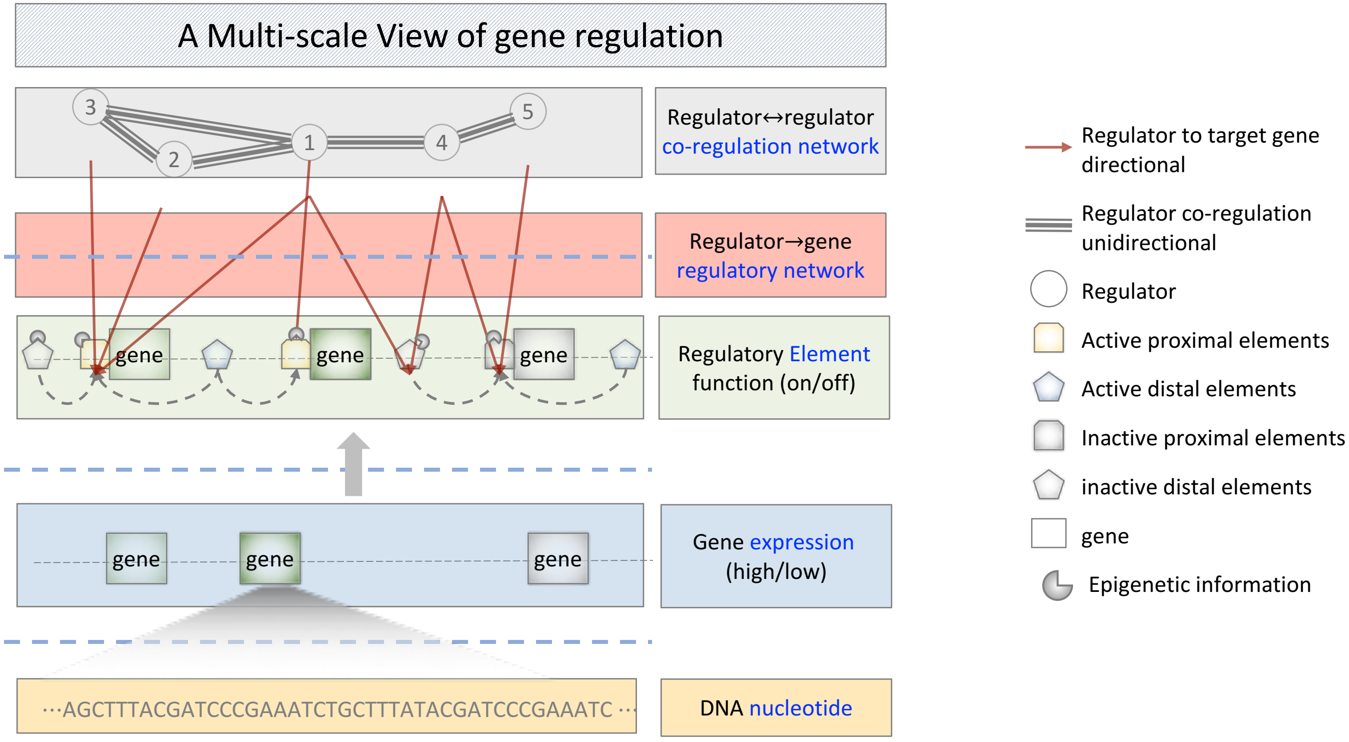
# Coding and non-coding elements may synergistically contribute to cancer



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

Major Challenges:

- Many levels of dysregulations related to disease status



A Multi-scale View of gene regulation

Regulator↔regulator co-regulation network

Regulator→gene regulatory network

Regulatory Element function (on/off)

Gene expression (high/low)

DNA nucleotide

Regulator to target gene directional

Regulator co-regulation unidirectional

Regulator

Active proximal elements

Active distal elements

Inactive proximal elements

inactive distal elements

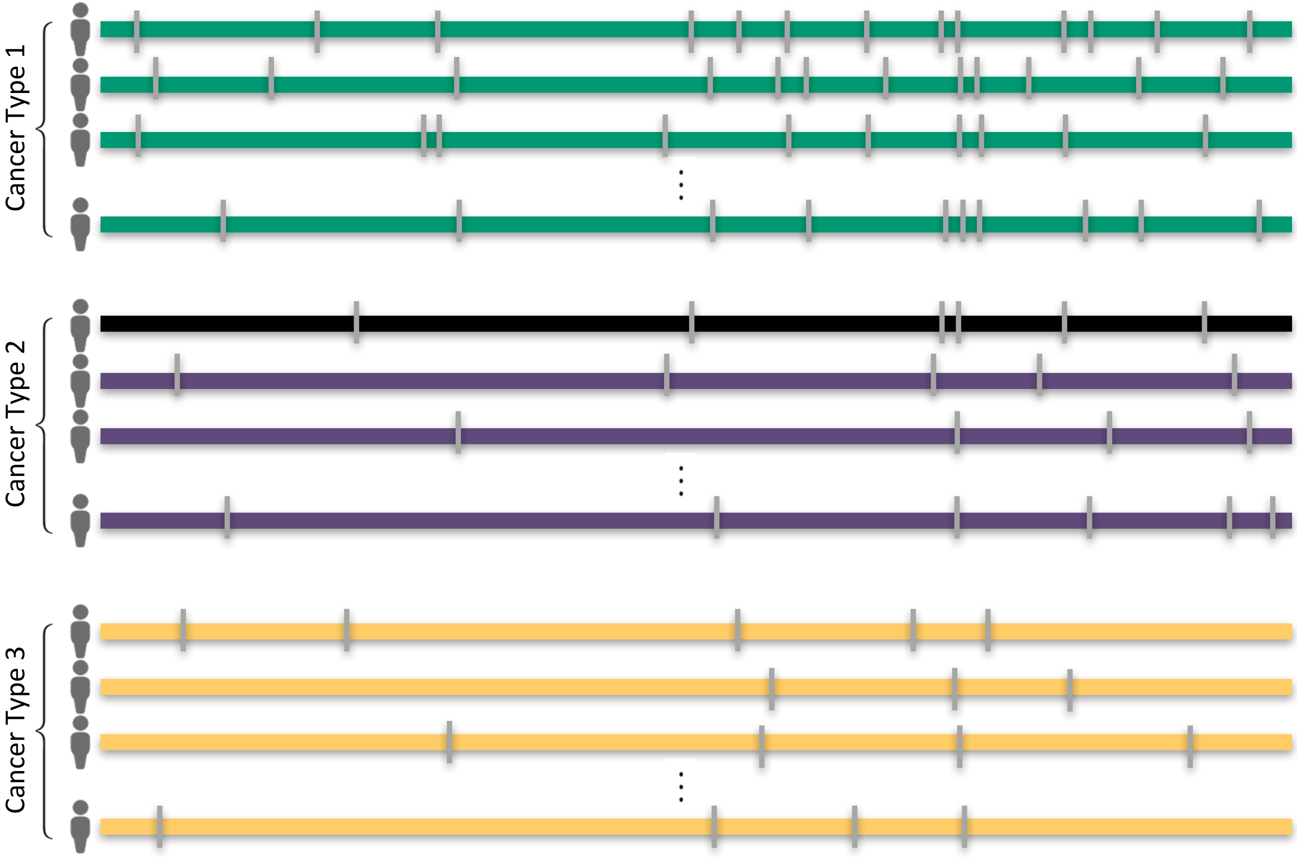
gene

Epigenetic information

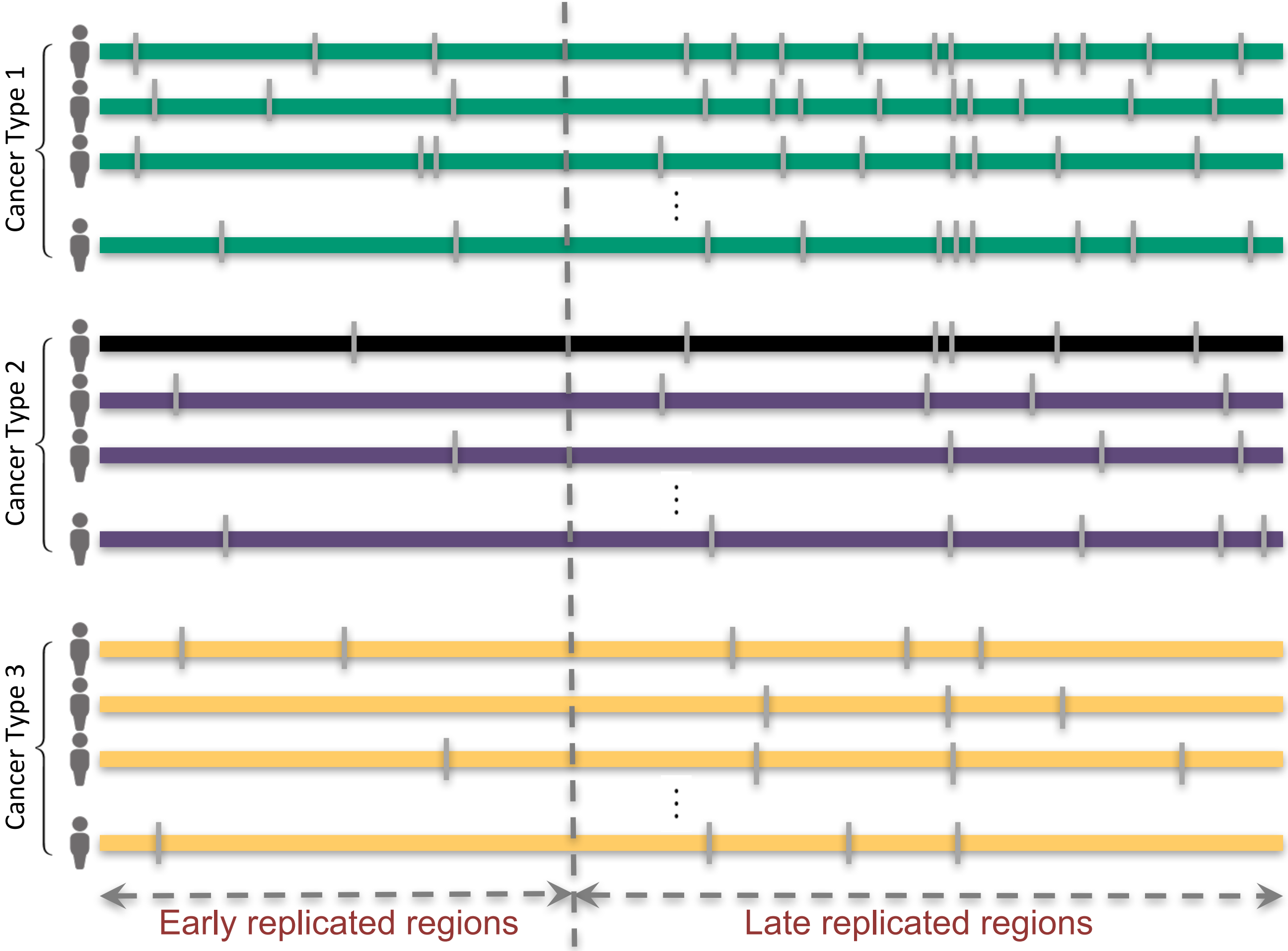
# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

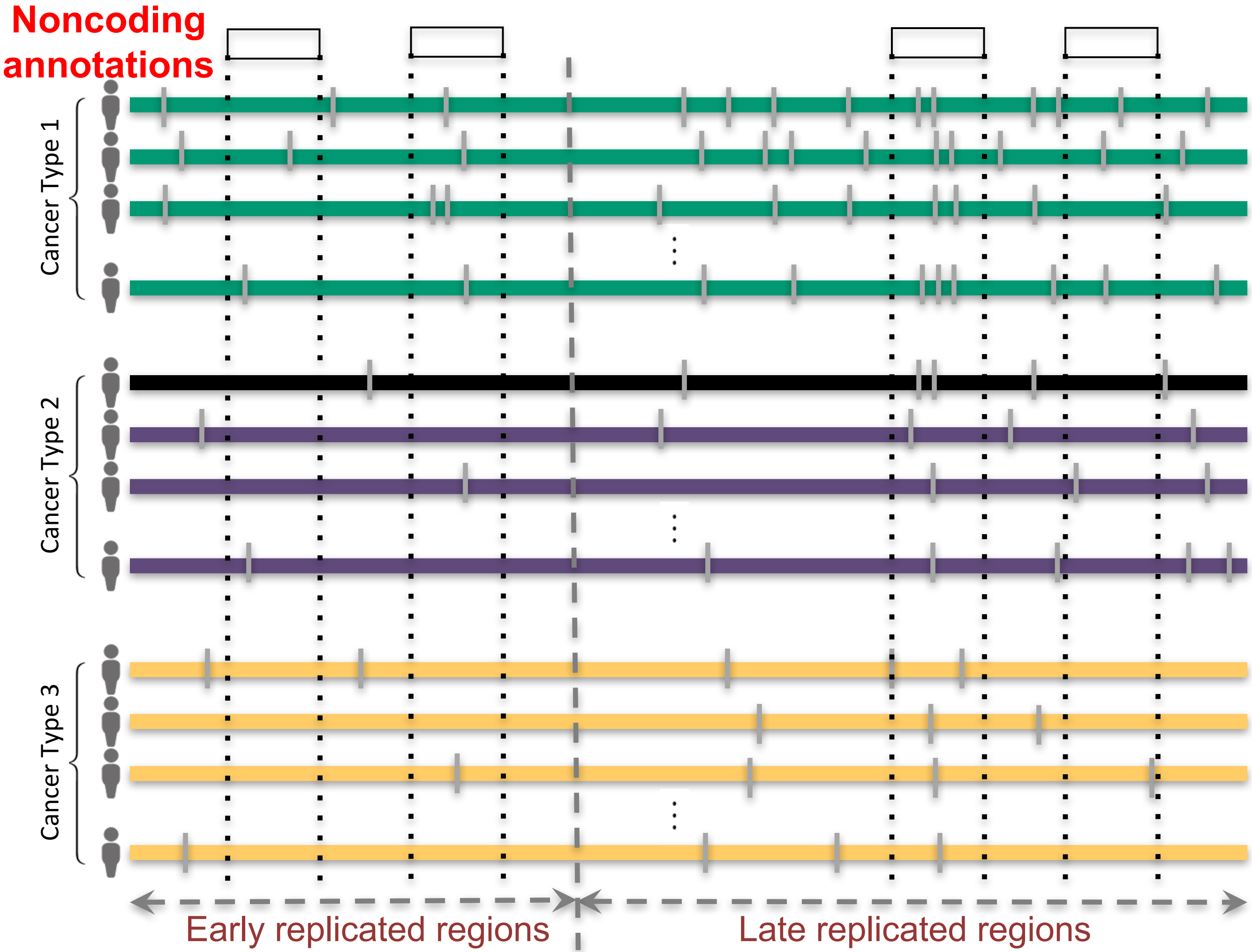
- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

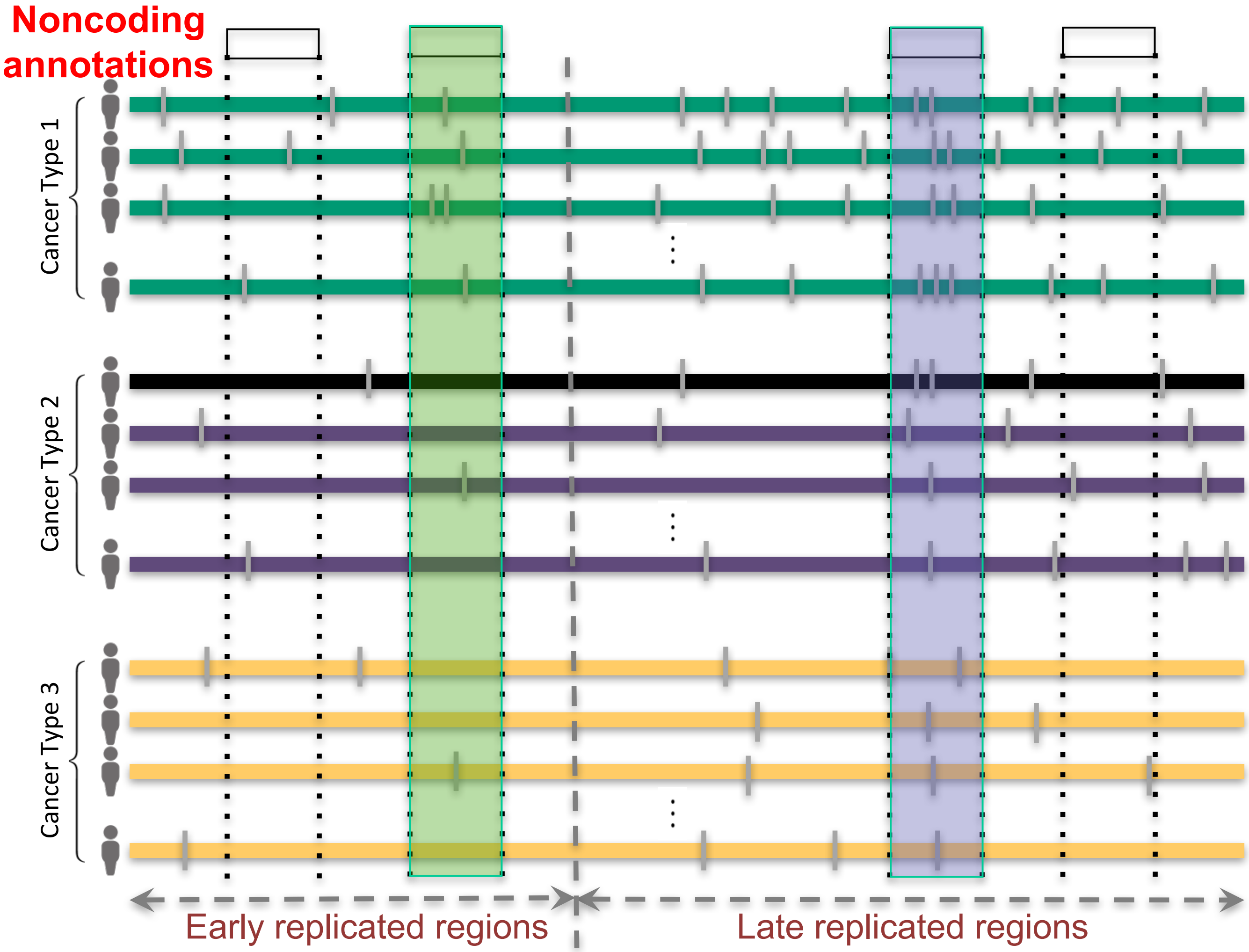
# Mutation recurrence



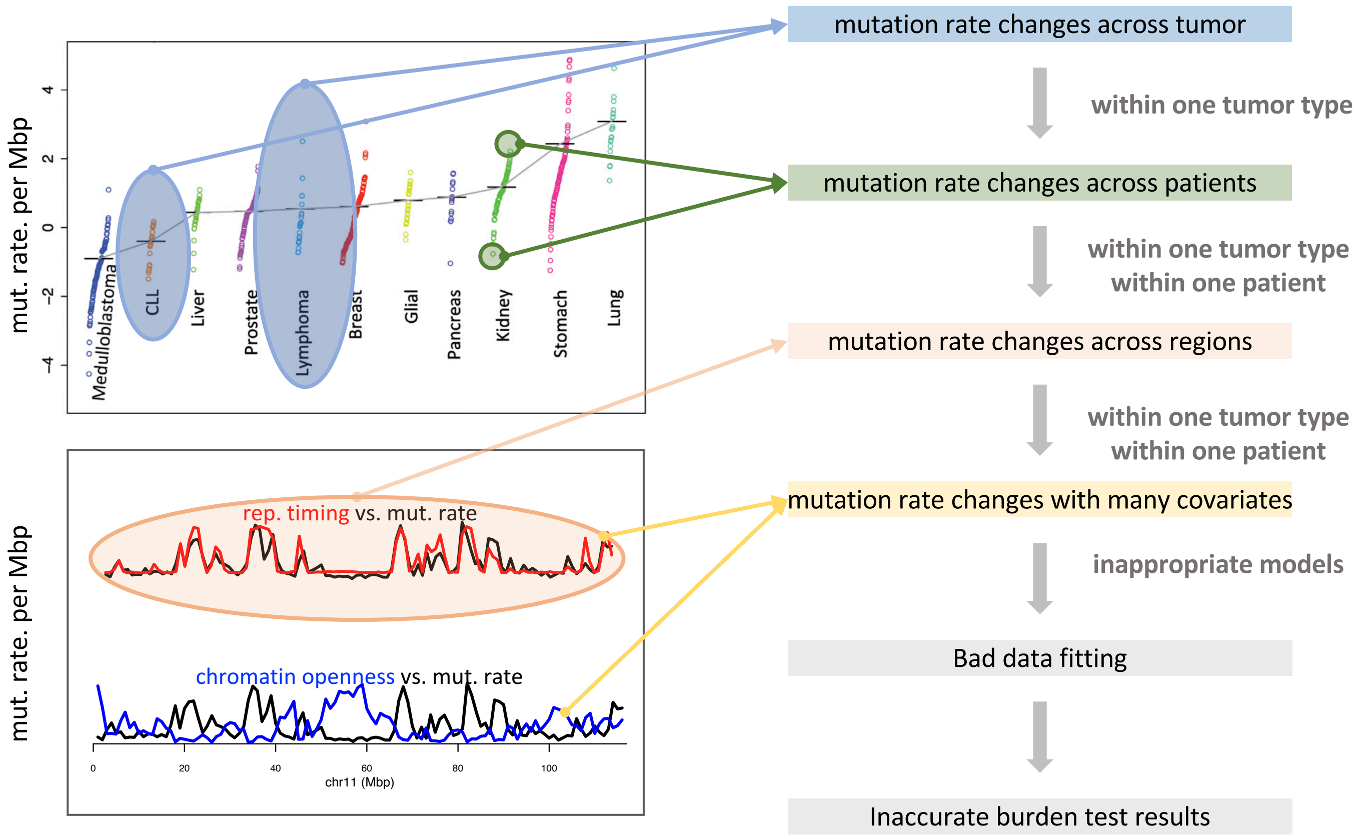
# Mutation recurrence




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


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

# violation of the constant mutation rate assumption



mutation rate changes across tumor

within one tumor type

mutation rate changes across patients

within one tumor type
within one patient

mutation rate changes across regions

within one tumor type
within one patient

mutation rate changes with many covariates

inappropriate models

Bad data fitting

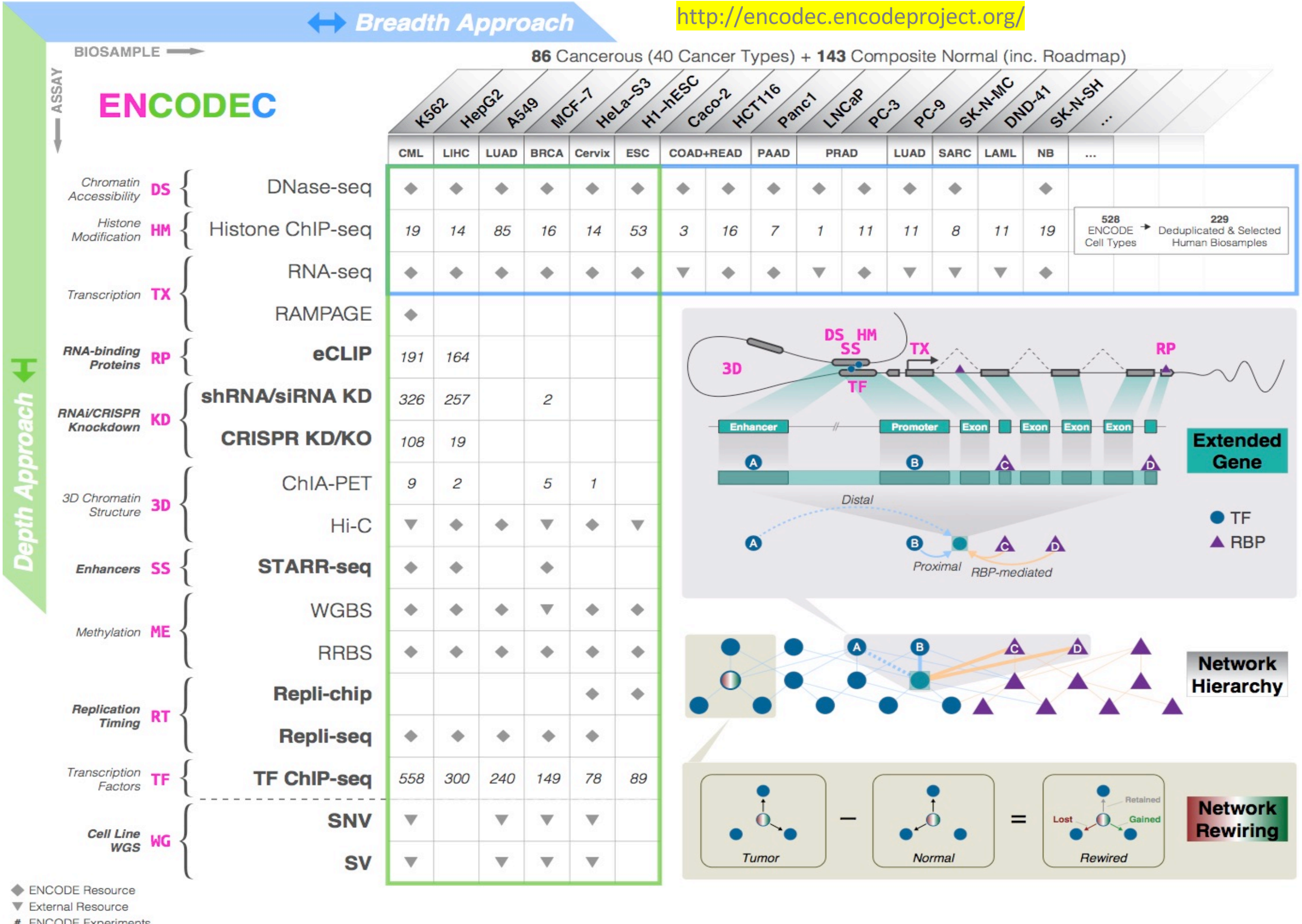
Inaccurate burden test results

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

# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

http://encodec.encodeproject.org/



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

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

528 ENCODE Cell Types → 229 Deduplicated & Selected Human Biosamples

◆ ENCODE Resource
▼ External Resource
# ENCODE Experiments

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

http://encodec.encodeproject.org/

↔ Breadth Approach

Depth Approach

BIOSAMPLE →

ASSAY ↓

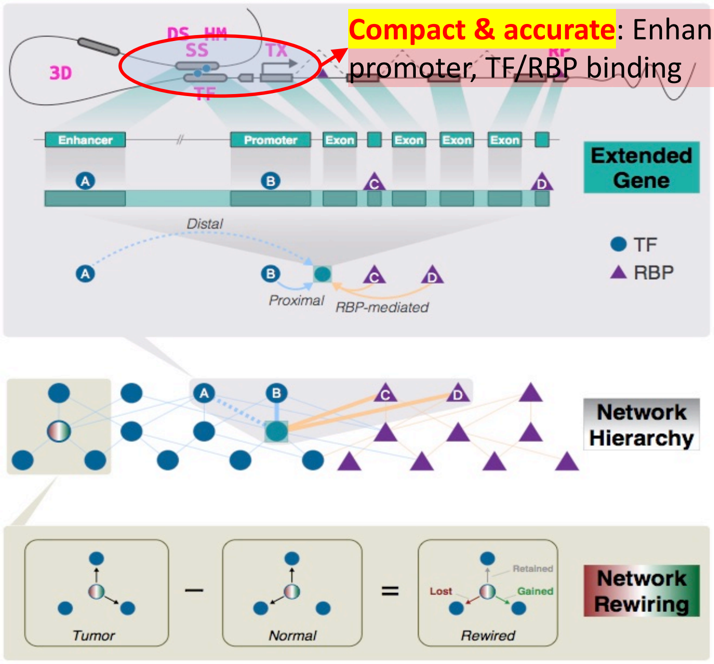
# ENCODEC

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

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

528 ENCODE Cell Types → 229 Deduplicated & Selected Human Biosamples

◆ ENCODE Resource
▼ External Resource
# ENCODE Experiments



**Compact & accurate**: Enhancer, promoter, TF/RBP binding

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

http://encodec.encodeproject.org/

# ENCODEC

↔ Breadth Approach

Depth Approach

BIOSAMPLE →

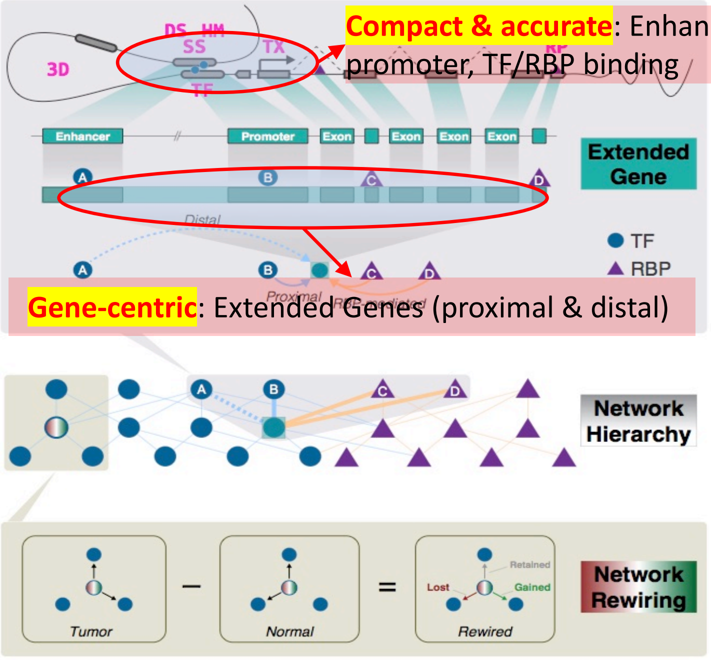
ASSAY ↓

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

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

528 ENCODE Cell Types → 229 Deduplicated & Selected Human Biosamples

◆ ENCODE Resource
▼ External Resource
# ENCODE Experiments

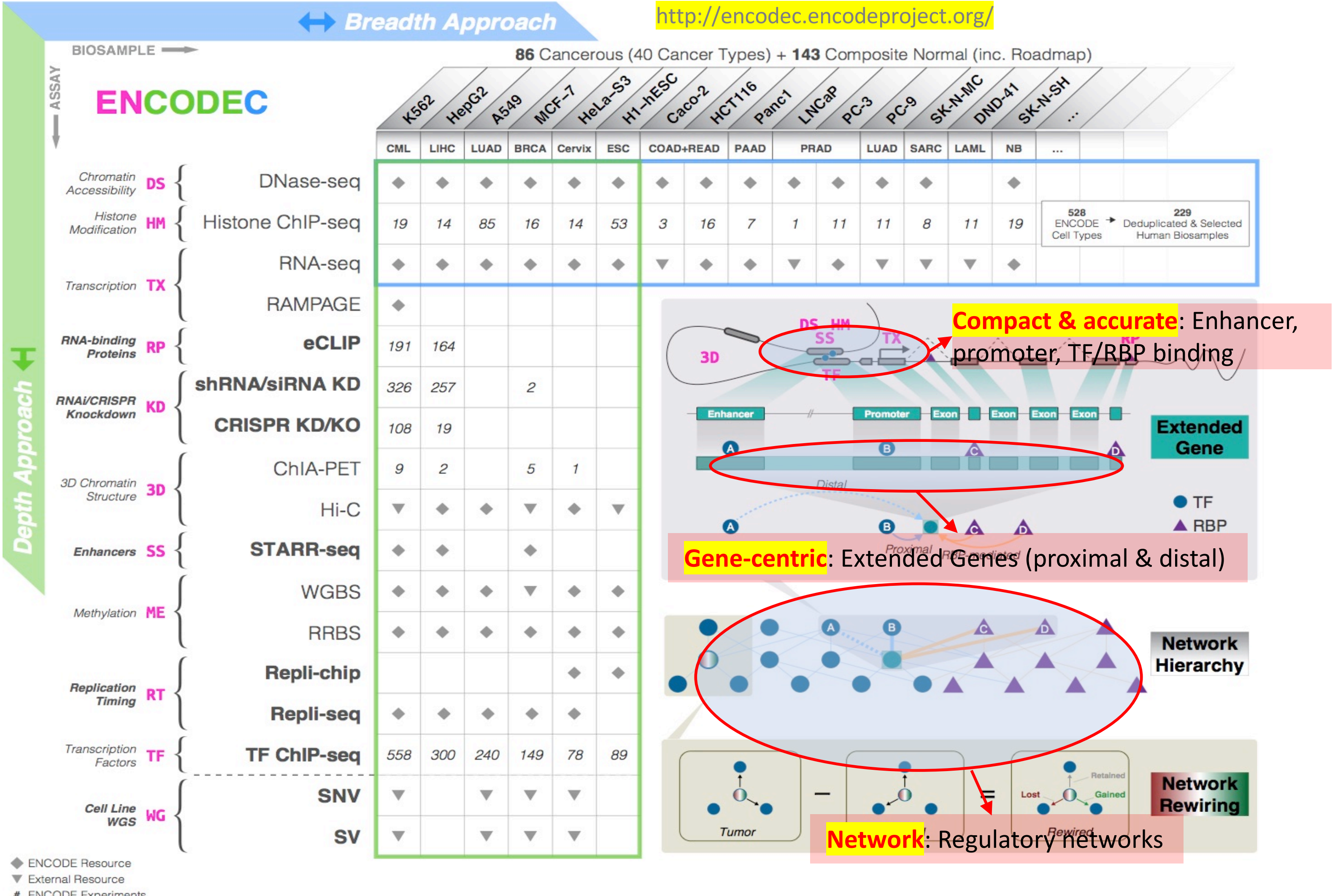


**Compact & accurate**: Enhancer, promoter, TF/RBP binding

**Gene-centric**: Extended Genes (proximal & distal)

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

http://encodec.encodeproject.org/



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

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

528 ENCODE Cell Types → 229 Deduplicated & Selected Human Biosamples

◆ ENCODE Resource
▼ External Resource
# ENCODE Experiments

**Compact & accurate**: Enhancer, promoter, TF/RBP binding

**Gene-centric**: Extended Genes (proximal & distal)

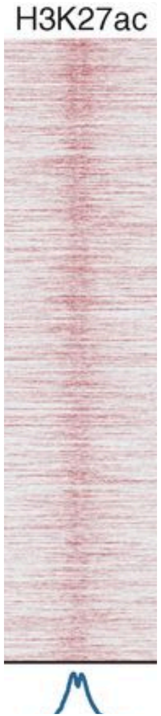
**Network**: Regulatory networks

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

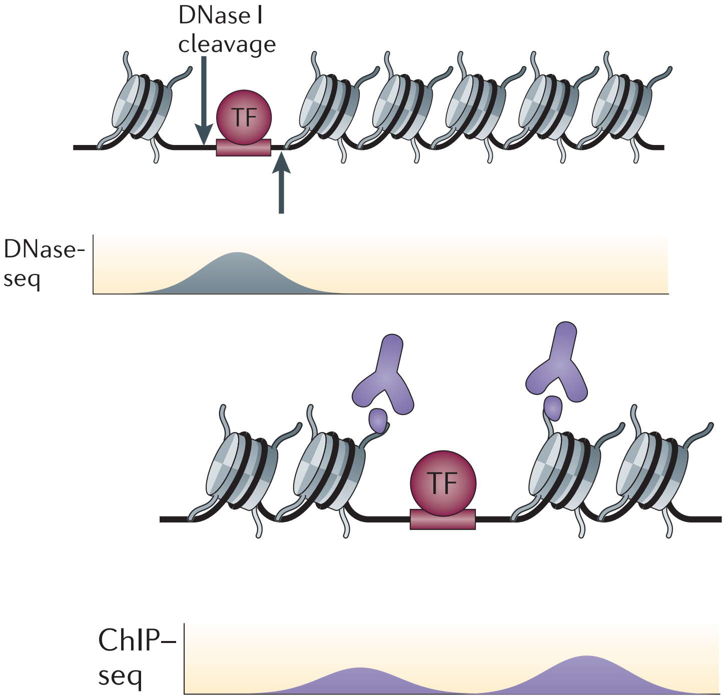
# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

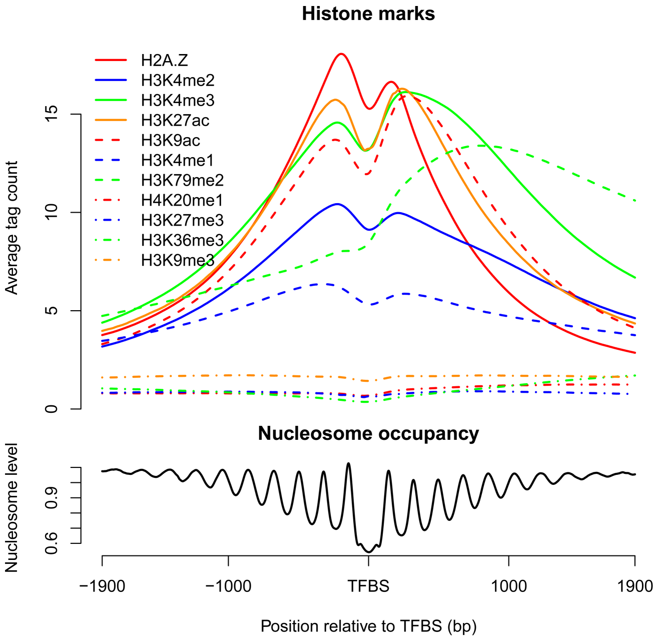
# Unique shape associated histone signals flanking active enhancers identified through STARR-seq



Arnold, et al., Science

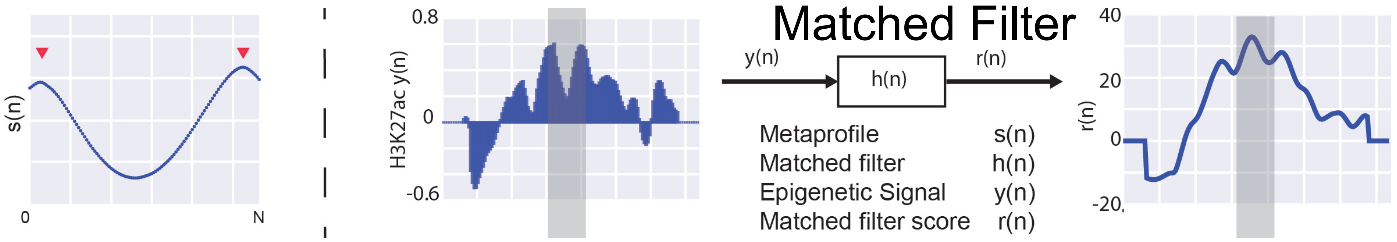


Shlyueva, et al., Nat Rev Genet



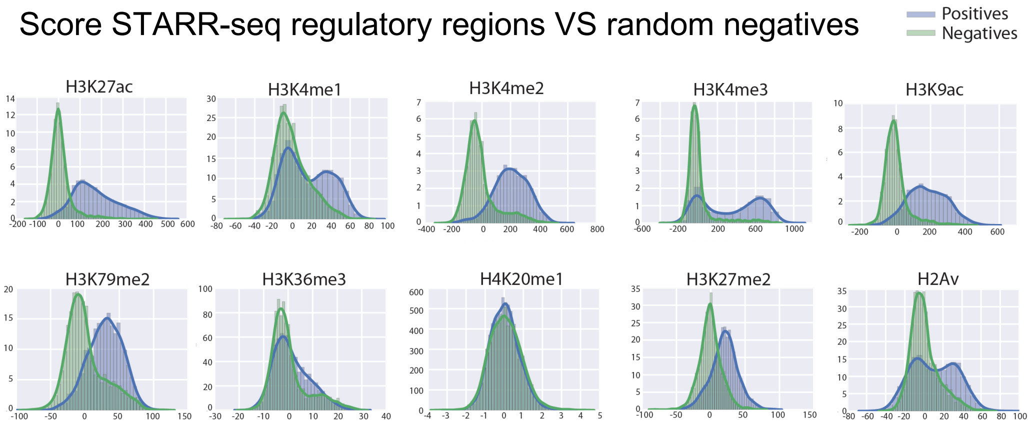
Nie, et al., PloS one

# Matched Filter recognize shape patterns

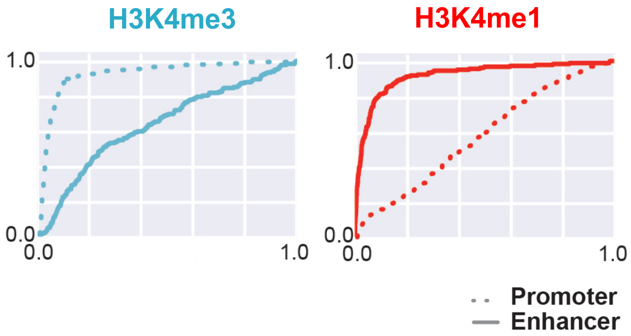


| | |
|---|---|
| Metaprofile | s(n) |
| Matched filter | h(n) |
| Epigenetic Signal | y(n) |
| Matched filter score | r(n) |

Score STARR-seq regulatory regions VS random negatives



Evaluate using ROC curve

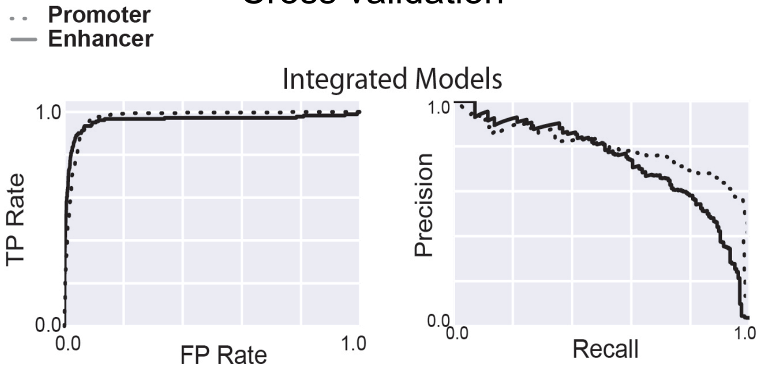


[ biorxiv.org/content/early/2018/08/05/385237 ]

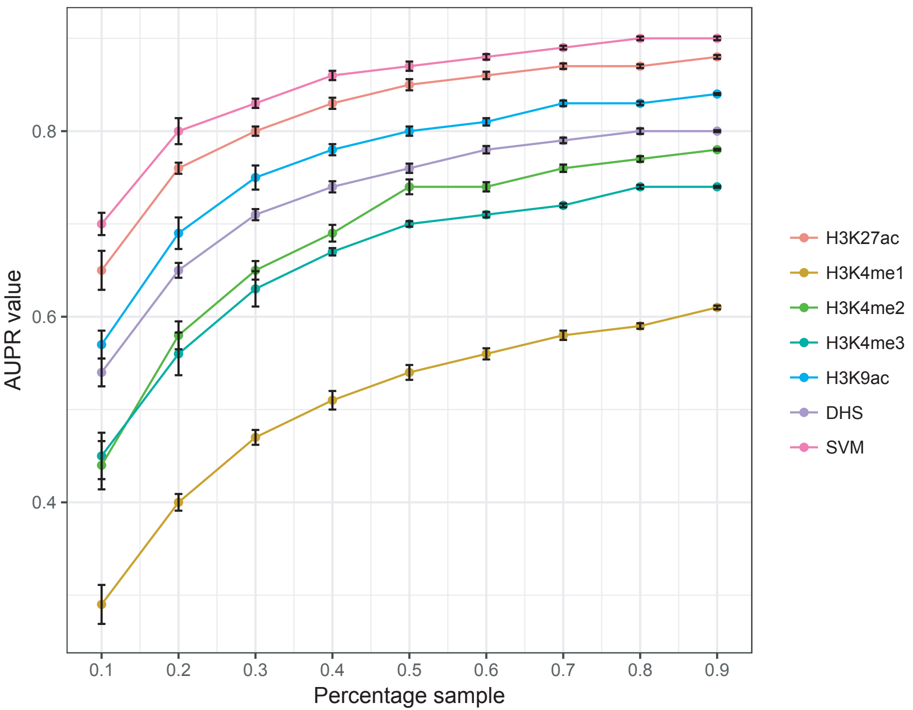
# Integrate matched filter scores of multiple features

| Model | AUROC | AUPR |
|---|---|---|
| Random Forest | 0.96 (0.95) | 0.91 (0.79) |
| Ridge Regression | 0.95 (0.94) | 0.90 (0.77) |
| Linear SVM | 0.96 (0.95) | 0.91 (0.78) |
| Naive Bayes | 0.95 (0.93) | 0.89 (0.72) |

Cross validation

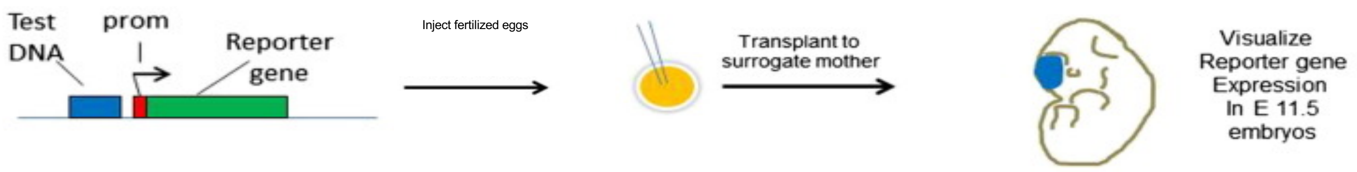


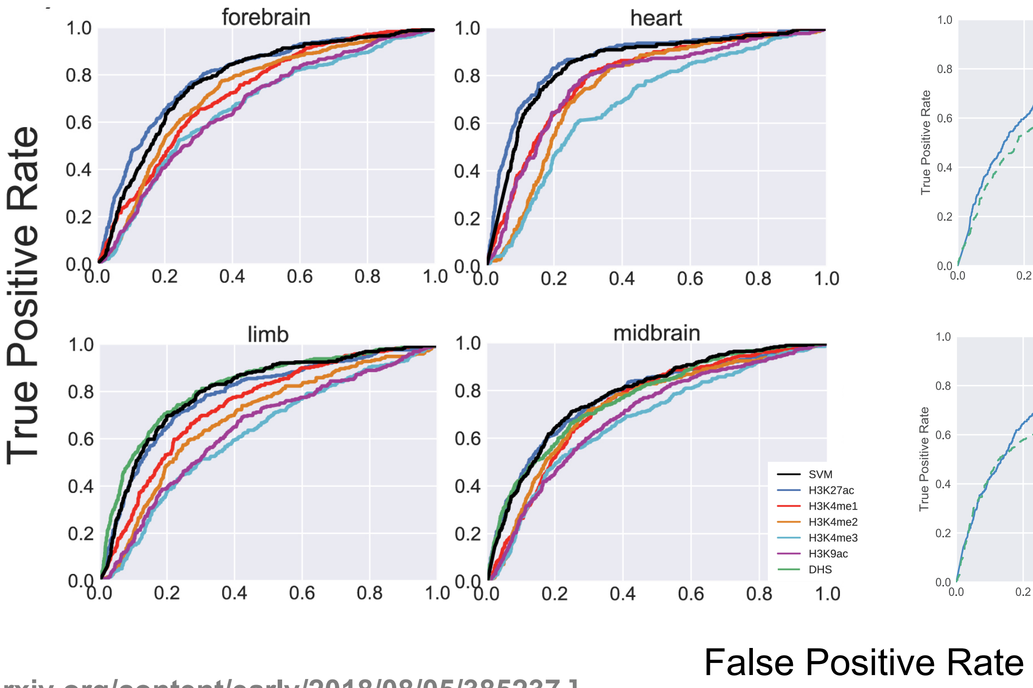
Large scale STARR-seq experiment data helps to improve the performance of integrated model

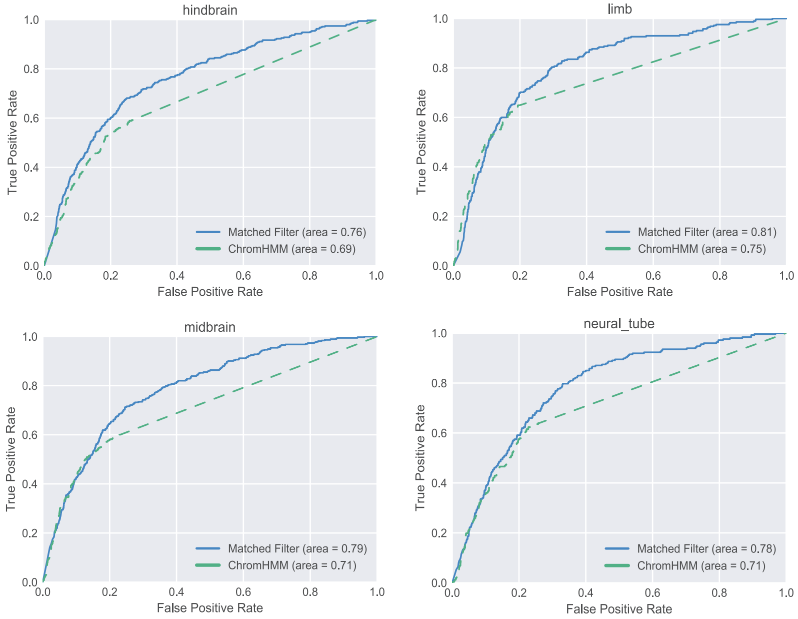


[ biorxiv.org/content/early/2018/08/05/385237 ]

# Validation with transgenic mouse enhancer assay

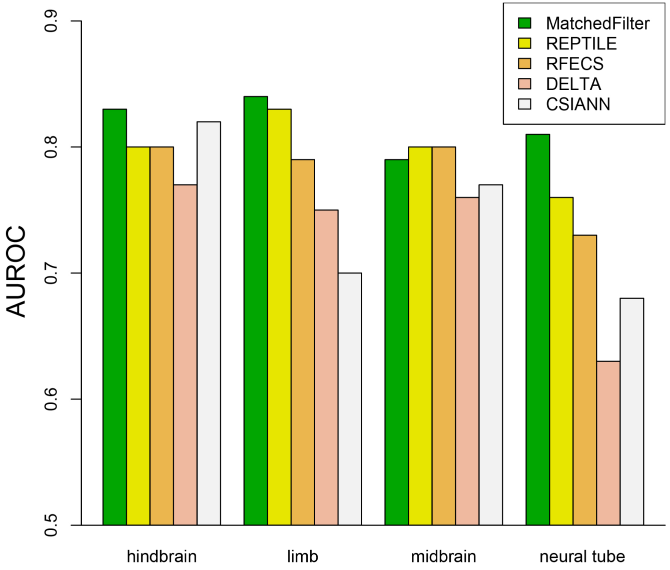

Test DNA
prom
Reporter gene
Inject fertilized eggs
Transplant to surrogate mother
Visualize Reporter gene Expression In E 11.5 embryos



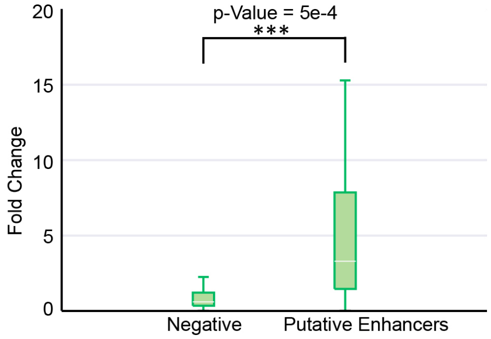


[ biorxiv.org/content/early/2018/08/05/385237 ]

# Matched-Filter can be applied across different organisms

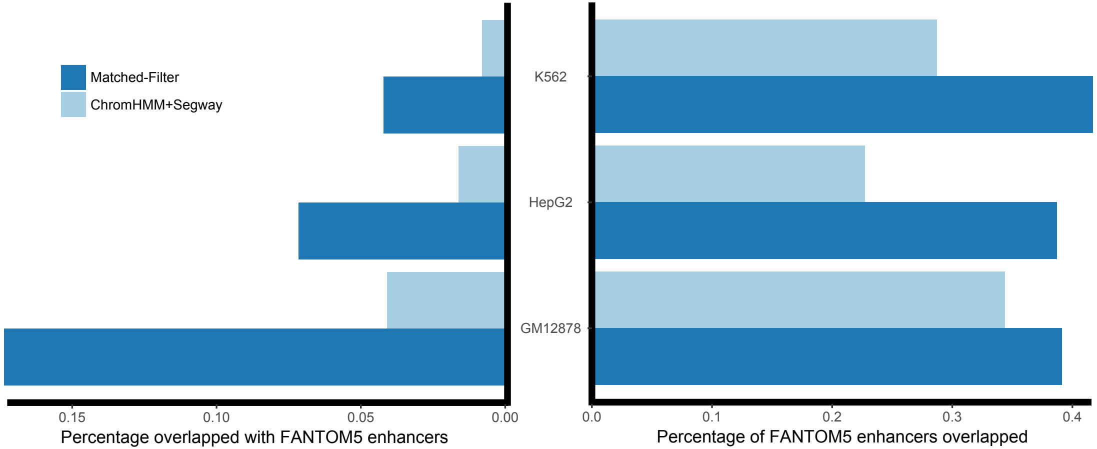


Compare Matched-Filter performance with other state-of-the-art methods



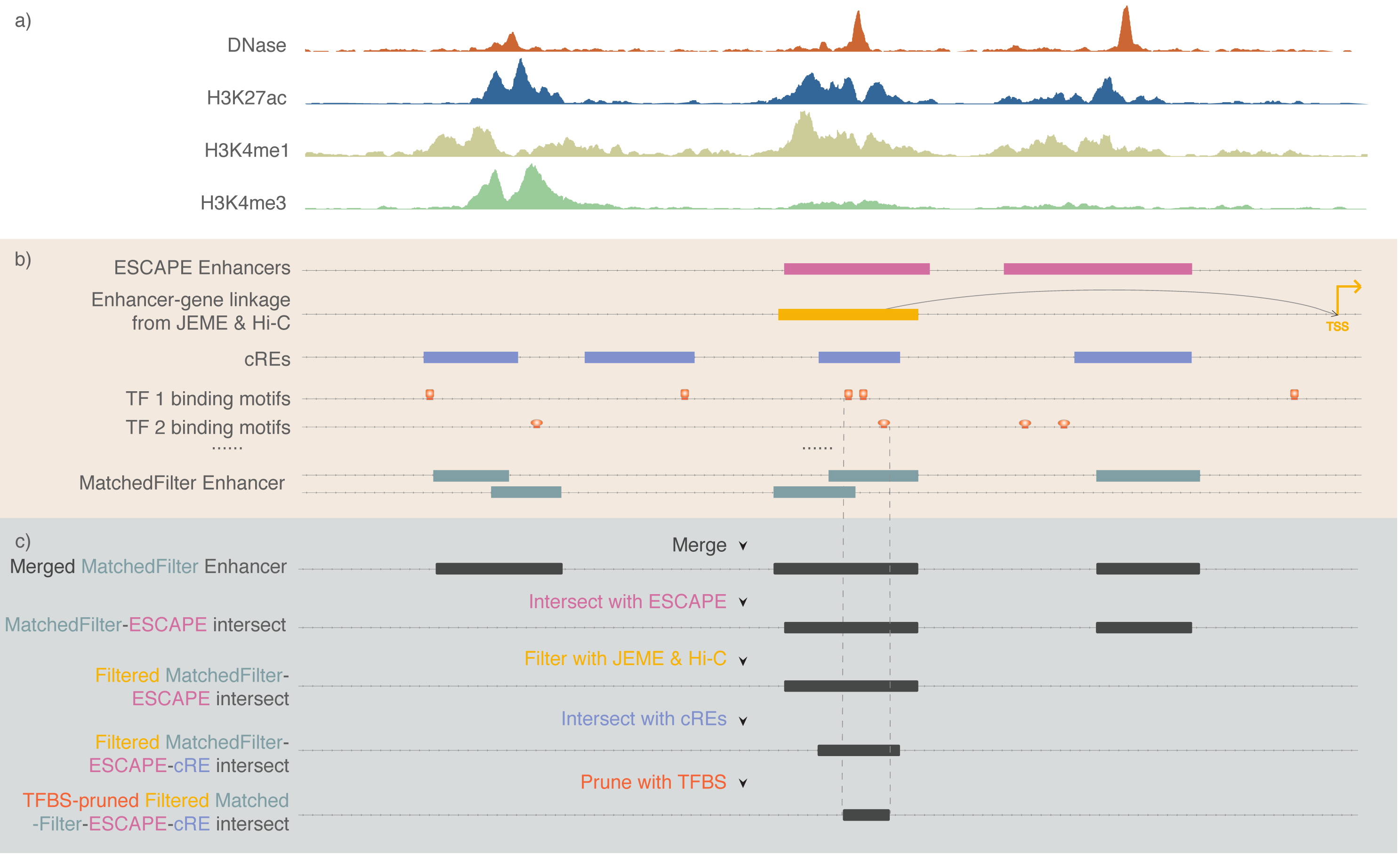
Validation using transduction-based reporter assay (H1-hESC, HOS, A549 and TZMBL)

Compare overlap with FANTOM5 enhancers



[ biorxiv.org/content/early/2018/08/05/385237 ]

# Constructing a high-confidence set of cell-specific enhancers


a)
DNase
H3K27ac
H3K4me1
H3K4me3
b)
ESCAPE Enhancers
Enhancer-gene linkage from JEME & Hi-C
TSS
cREs
TF 1 binding motifs
TF 2 binding motifs
......
......
MatchedFilter Enhancer
c)
Merge
Merged MatchedFilter Enhancer
Intersect with ESCAPE
MatchedFilter-ESCAPE intersect
Filter with JEME & Hi-C
Filtered MatchedFilter-ESCAPE intersect
Intersect with cREs
Filtered MatchedFilter-ESCAPE-cRE intersect
Prune with TFBS
TFBS-pruned Filtered Matched-Filter-ESCAPE-cRE intersect

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

# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

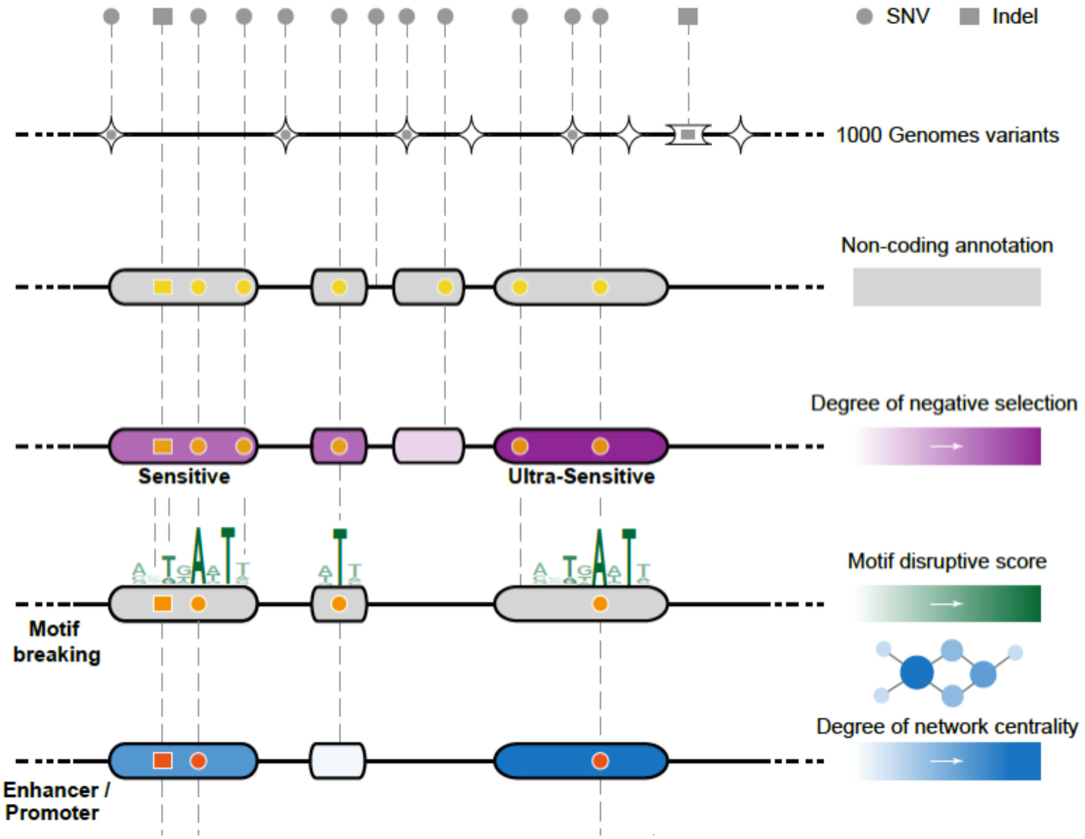
# Funseq: a flexible framework to determine functional impact & use this to prioritize variants

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

**Conservation (GERP, allele freq.)**

**Mutational impact (motif breaking, Lof)**

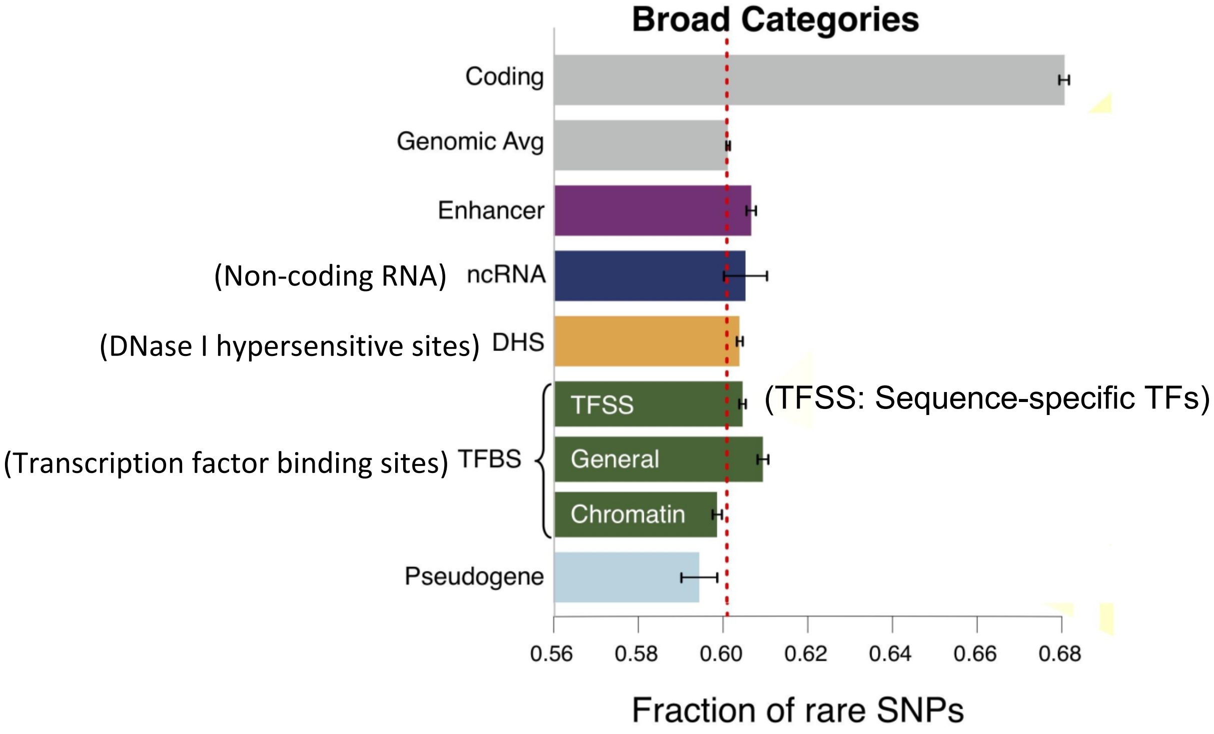
**Network (centrality position)**



[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

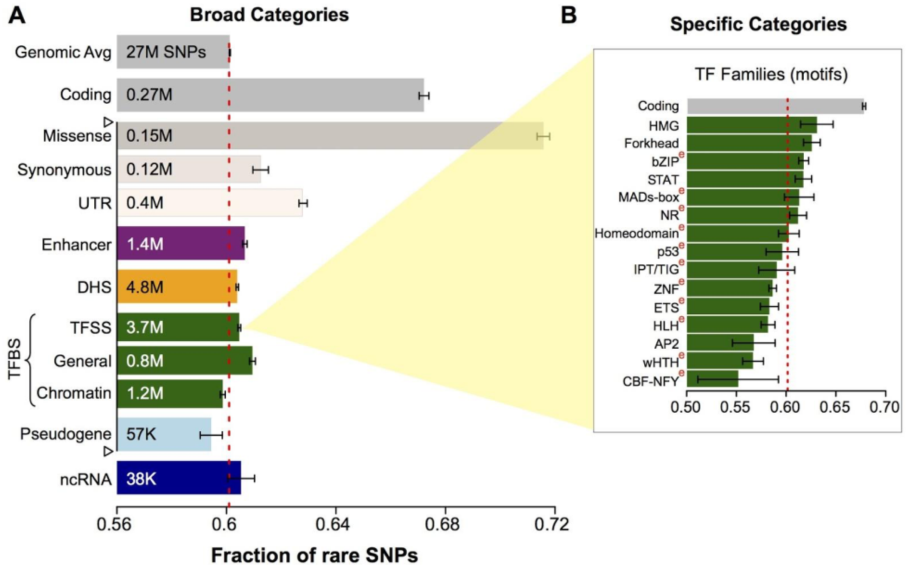


**Depletion of Common Variants in the Human Population**

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

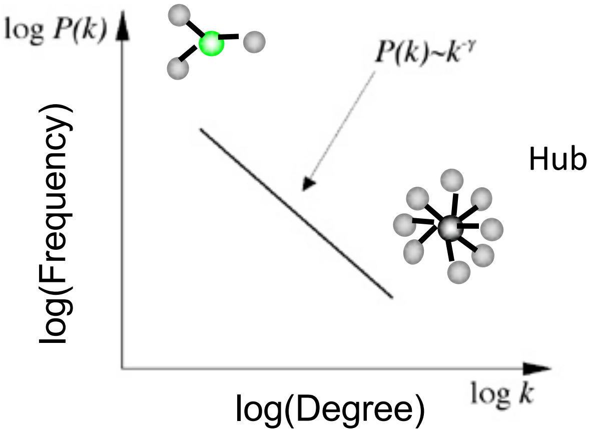


Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

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

# Hubs Under Constraint: A Finding from the Network Biology Community

Power-law distribution



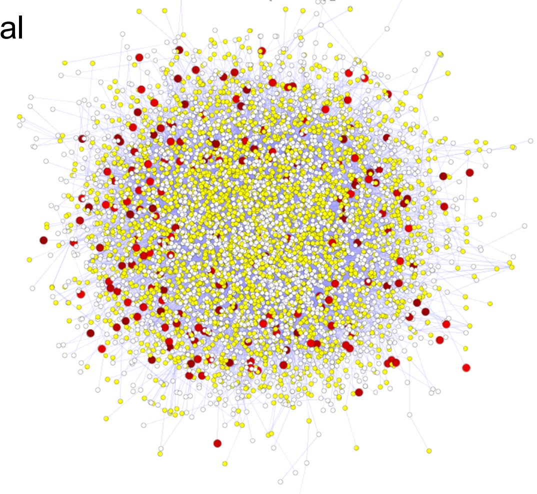
● High likelihood of positive selection

● Lower likelihood of positive selection

● Not under positive selection

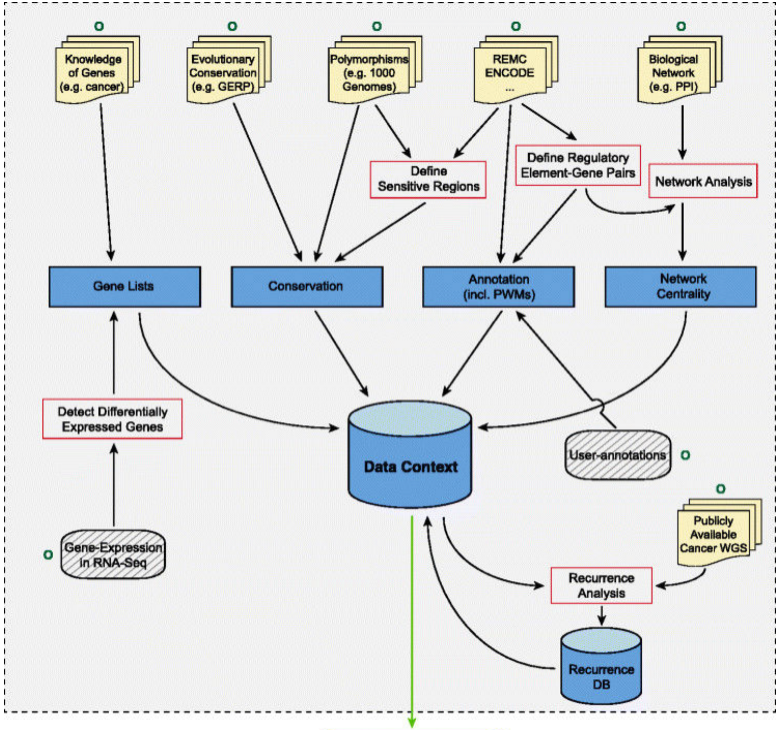
○ No data about positive selection

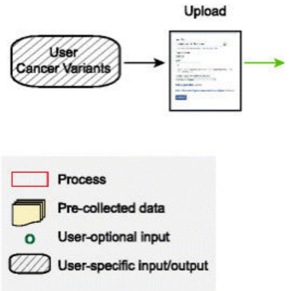
[Nielsen et al. *PLoS Biol.* (2005), HPRD, Kim et al. PNAS (2007)]

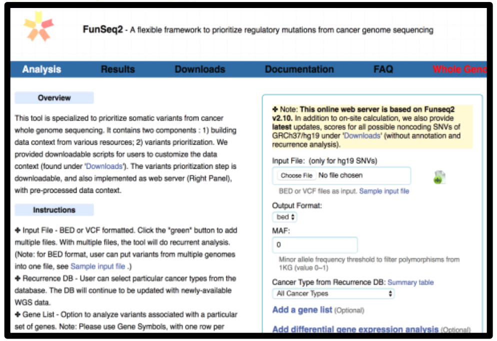


- More Connectivity, More Constraint: Genes & proteins that have a more central position in the network tend to evolve more slowly and are more likely to be essential.
- This phenomenon is observed in **many organisms & different kinds of networks**
    - **yeast PPI** - Fraser et al ('02) Science, ('03) BMC Evo. Bio.
    - **Ecoli PPI** - Butland et al ('04) Nature
    - **Worm/fly PPI** - Hahn et al ('05) MBE
    - **miRNA net** - Cheng et al ('09) BMC Genomics

# FunSeq.gersteinlab.org


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)

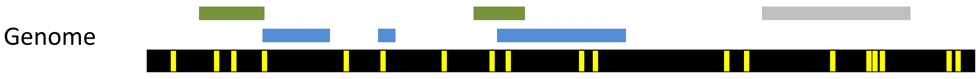
Cancer Type from Recurrence DB: Summary table

All Cancer Types

Add a gene list (Optional)

Add differential gene expression analysis (Optional)

HOT region 🟩
Sensitive region 🟦
Polymorphisms 🟨



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

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

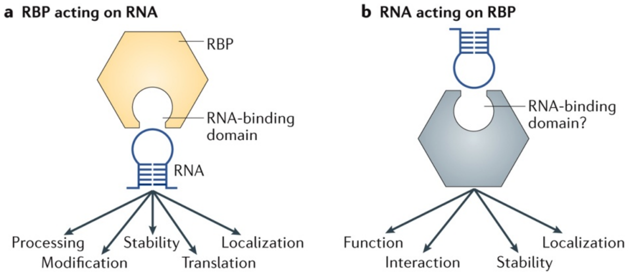
- Practical web server
- Submission of variants & pre-computed large data context from uniformly processing large-scale datasets

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

# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a "surprisal" based weighting scheme.
  - Prioritizing variants within "sensitive sites" (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

# RNA Binding Proteins (RBPs)

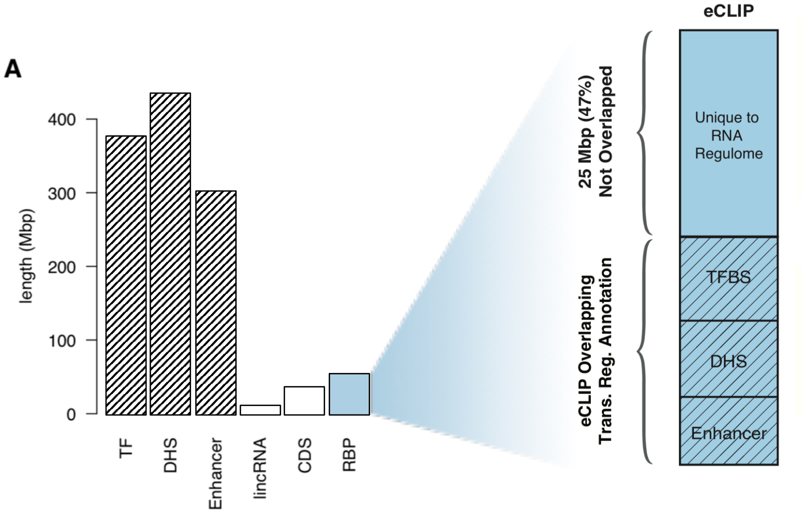


Nature Reviews | Molecular Cell Biology

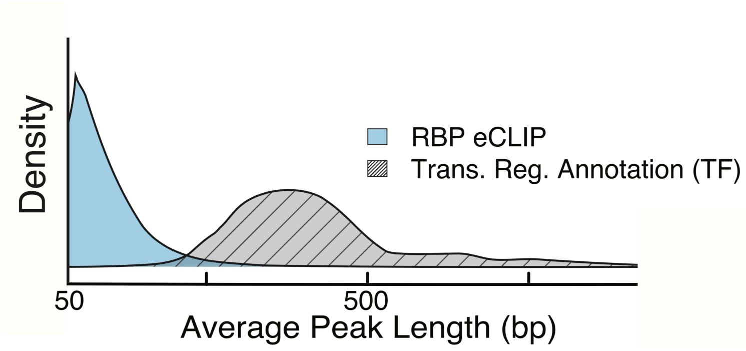
Nat Rev Mol Cell Biol. 2018 May;19(5):327-341. doi: 10.1038/nrm.2017.130. Epub 2018 Jan 17.

- **Before ENCODE3: >150 expt**.
  in many different cell types

- **ENCODE3 did ~350 focused eCLIP expt.**
  for >110 RBPs on HepG2 & K562
  (Van Nostrand...Yeo. Nat. Meth. '16;
  Van Nostrand...Graveley, Yeo
  (submitted in relation to ENCODE3))

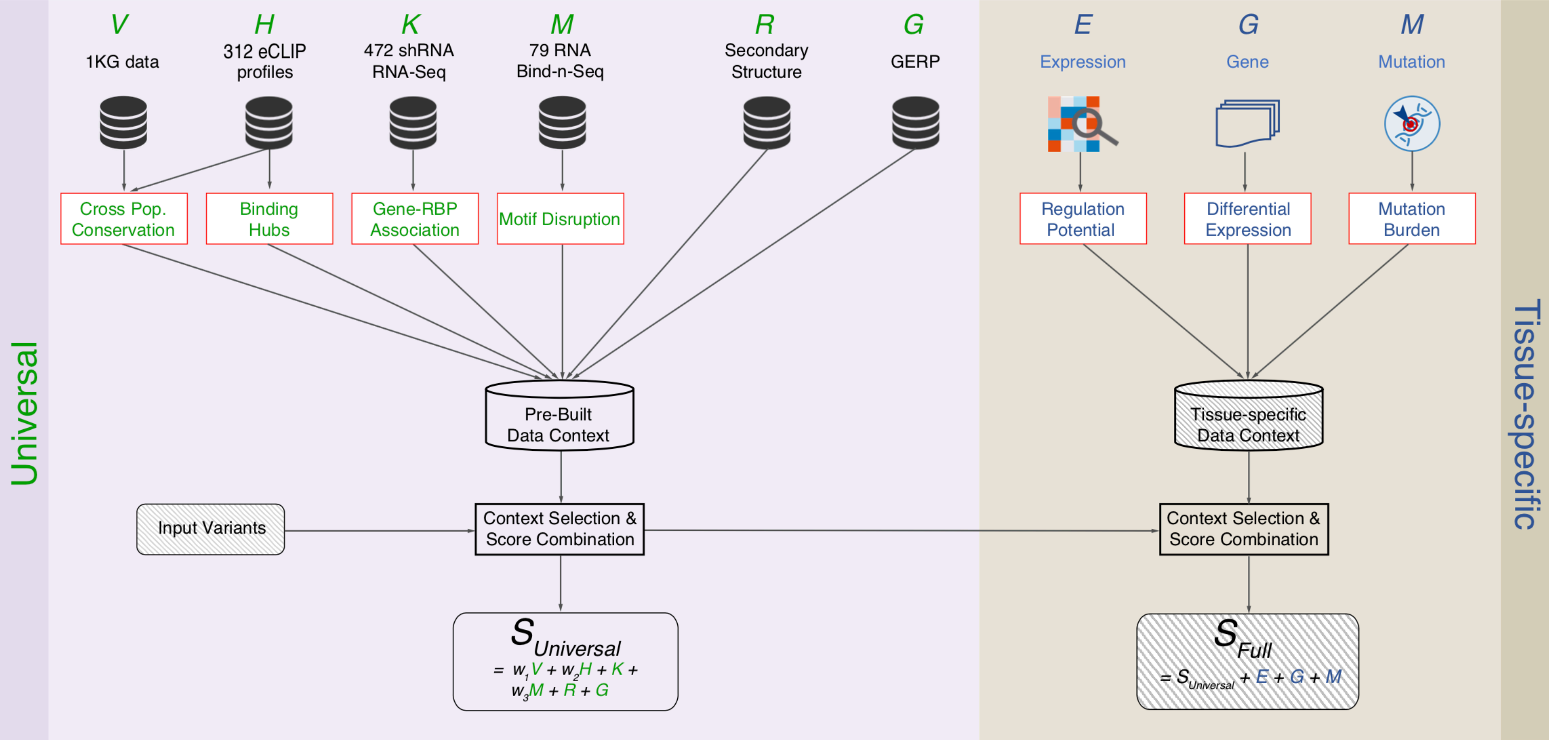


## ENCODE 3 - eCLIP peaks



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

# Schematic of RADAR Scoring



## Data and Procedure

| | |
|---|---|
| ▭ | Input/output |
| ▨ | Tissue-specific (Optional) |
| 🟥 | Procedure |
| ● | Pre-collected data |

## Context Selection & Score Combination

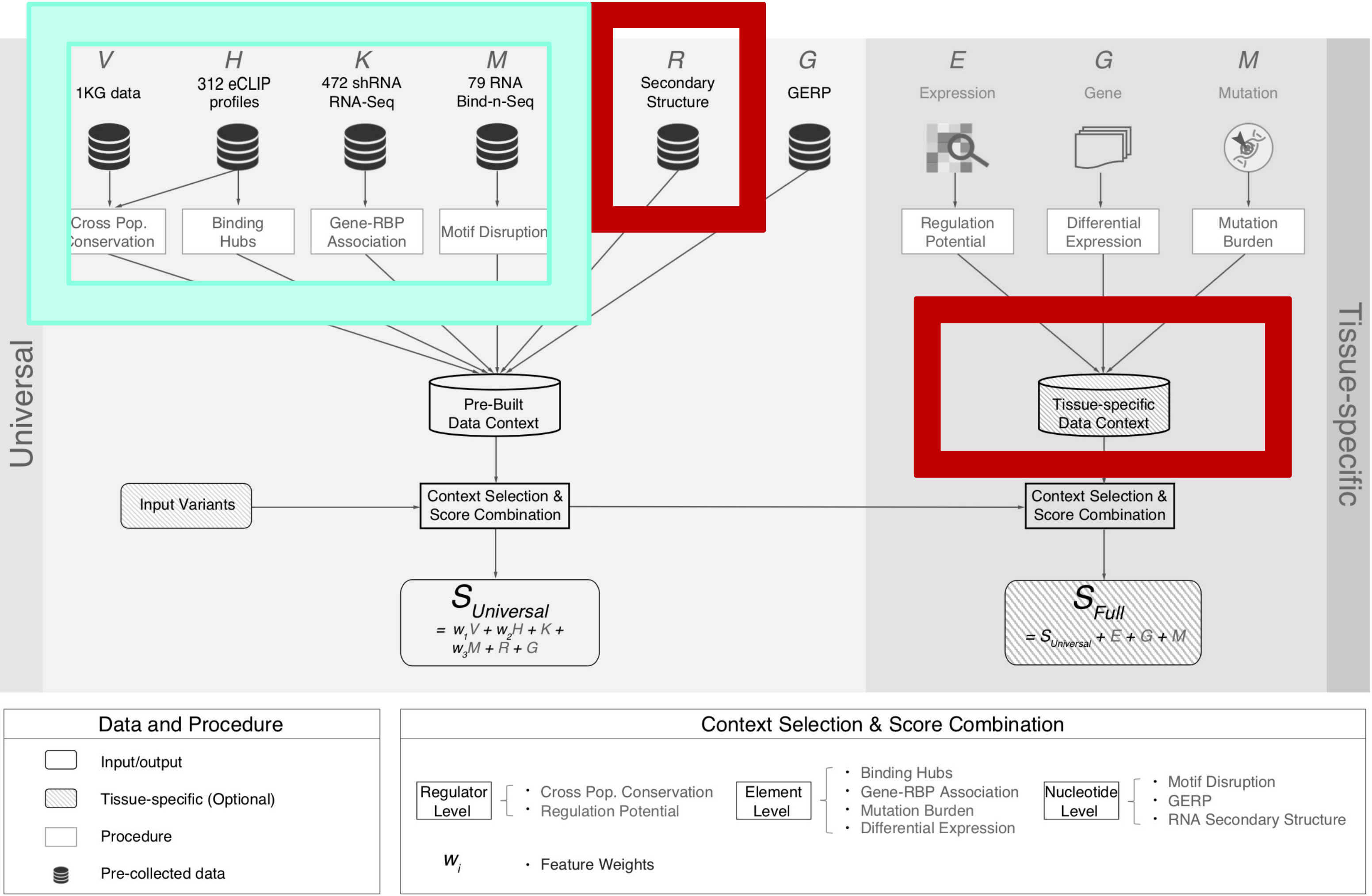
Regulator Level
- Cross Pop. Conservation
- Regulation Potential

Element Level
- Binding Hubs
- Gene-RBP Association
- Mutation Burden
- Differential Expression

Nucleotide Level
- Motif Disruption
- GERP
- RNA Secondary Structure

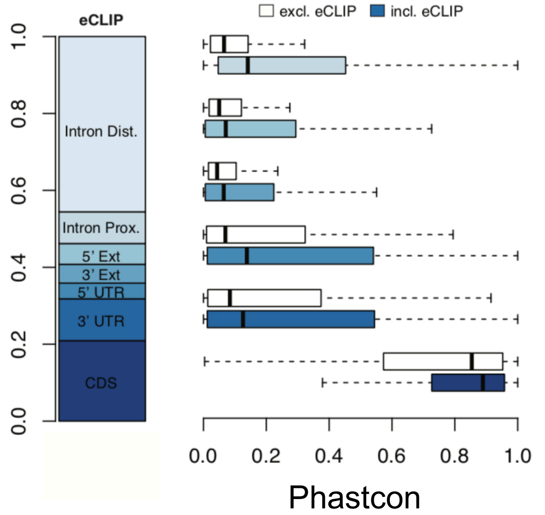
$w_i$
- Feature Weights

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

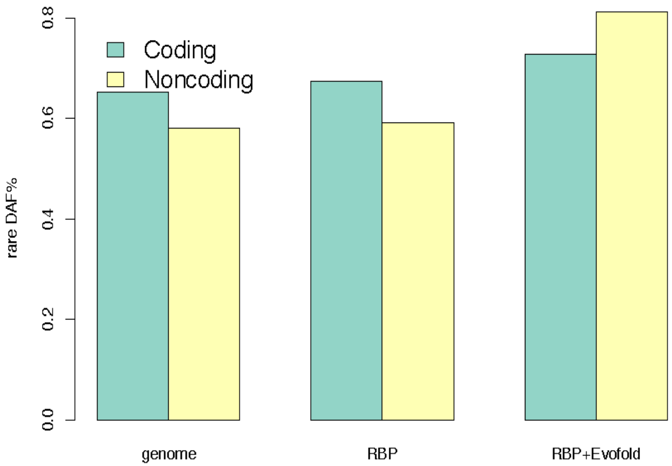


[Zhang*, Liu* et al., *Genome Biology* (in review ‘18)]

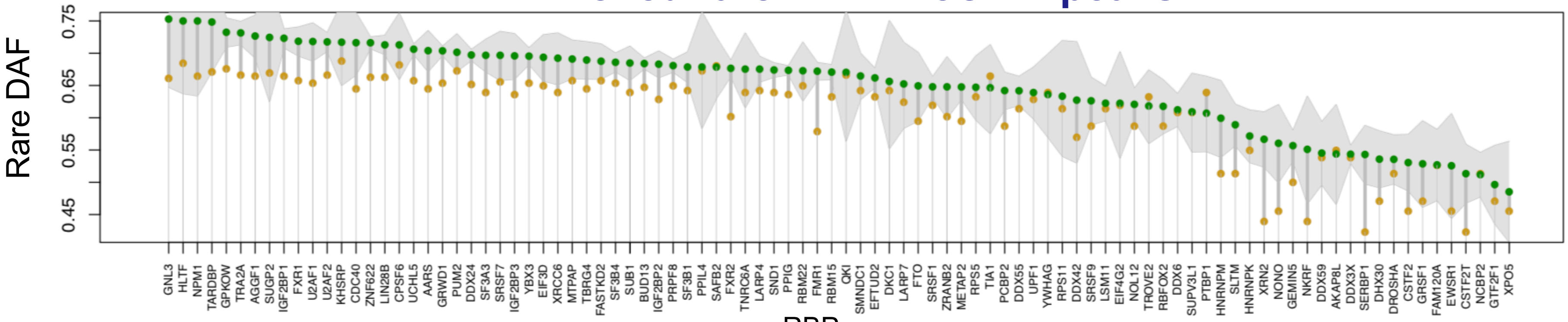
## High Phastcon in RBP-overlapped annotations



## RNA Structure Cons. from Evofold



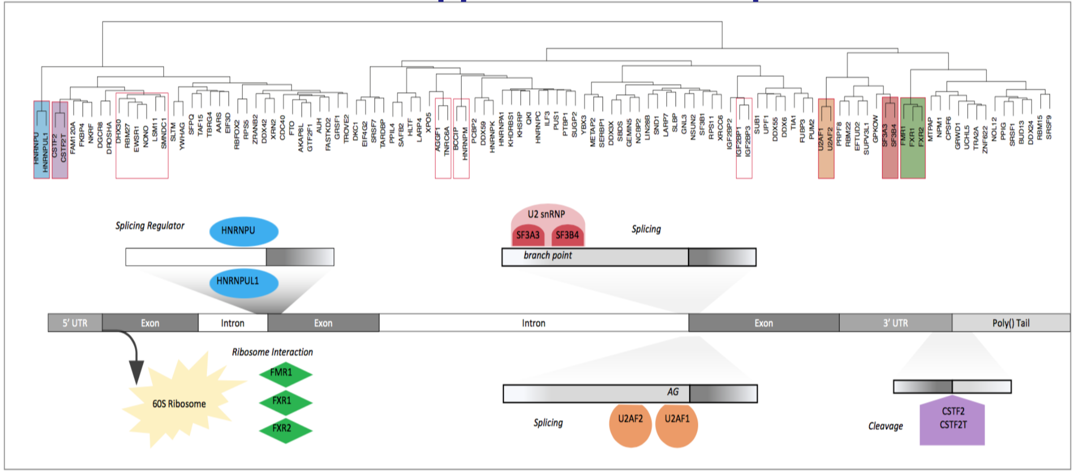
## Enriched rare DAF in eCLIP peaks



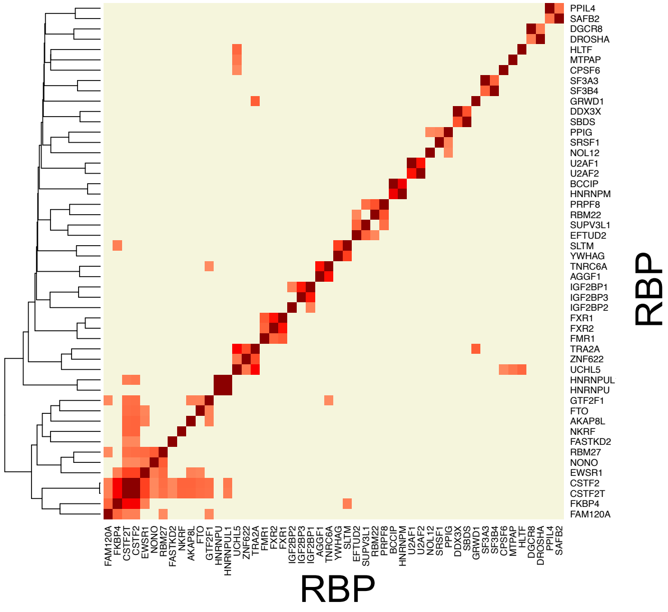
[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

# Co-binding of RBPs form biologically relevant complexes

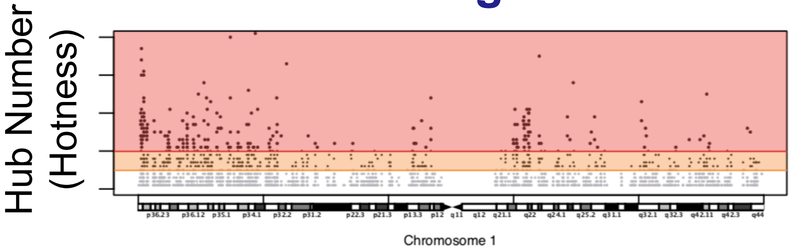
## Literature supported RBP complexes

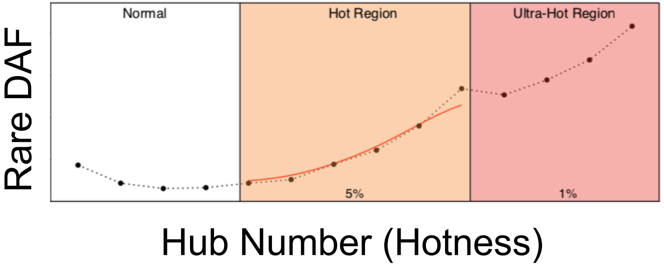


## Unique co-binding patterns of RBPs



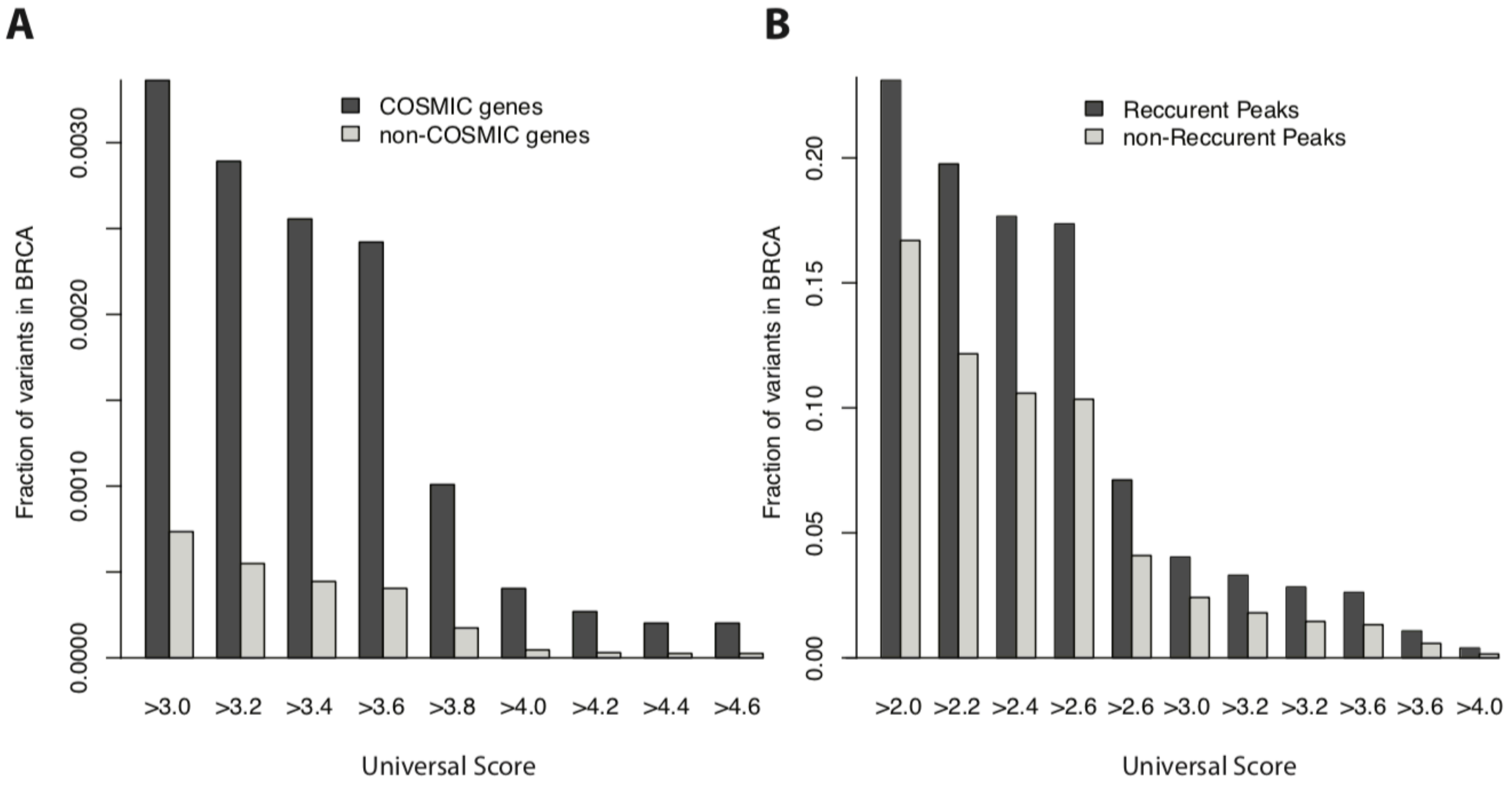
## Binding hubs are enriched for rare variants





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

# RADAR Scores enriched in COSMIC genes and recurrently mutated regions

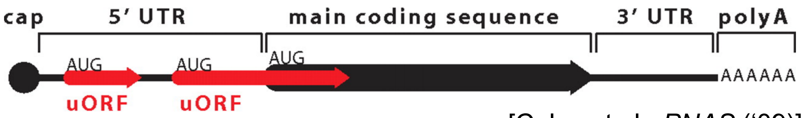


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]

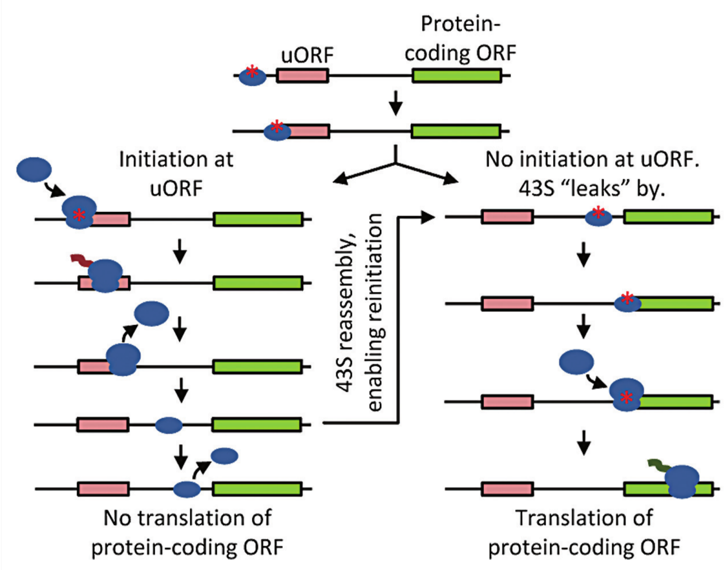
# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a "surprisal" based weighting scheme.
  - Prioritizing variants within "sensitive sites" (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

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

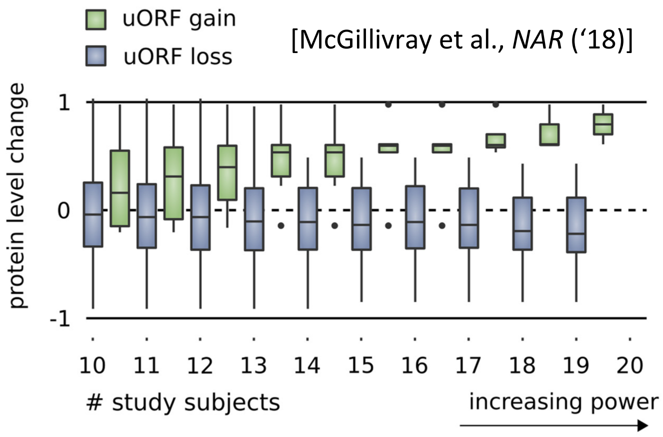


[Calvo et al., *PNAS* (‘09)]

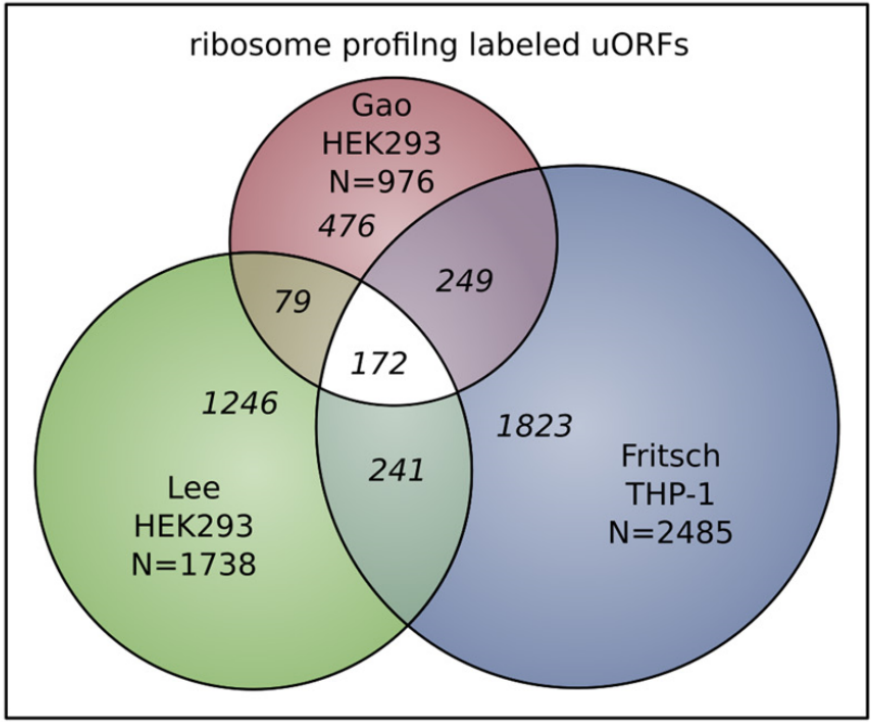

uORF
Protein-coding ORF
Initiation at uORF
No initiation at uORF. 43S “leaks” by.
43S reassembly, enabling reinitiation
No translation of protein-coding ORF
Translation of protein-coding ORF

[Ferreira et al., *Bioengineered* (‘14)]

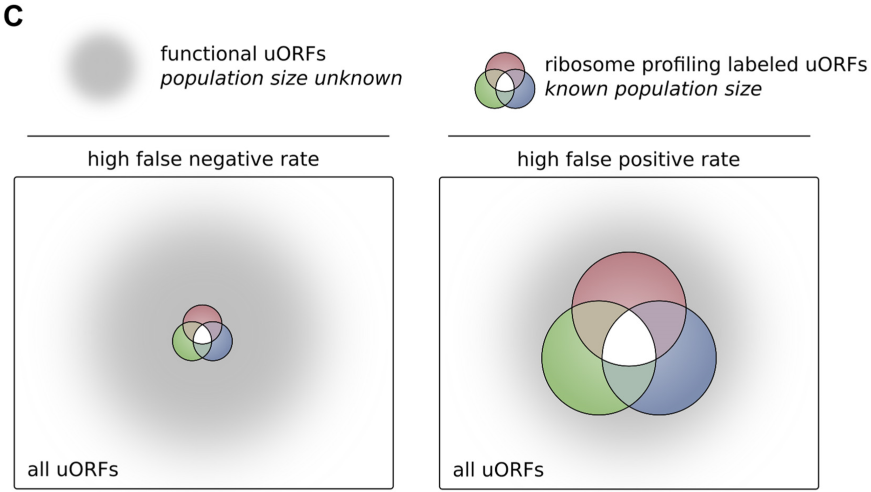
- uORFs regulate the translation of downstream coding regions.
- This regulation may be altered by somatic mutation in cancer.
- In Battle et al. 2014 data uORF gain & loss assoc. protein level change.



[McGillivray et al., *NAR* (‘18)]



# The population of functional uORFs may be significant



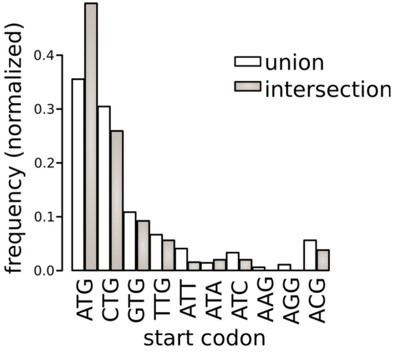
## From a “Universe” of 1.3 M pot. uORFs

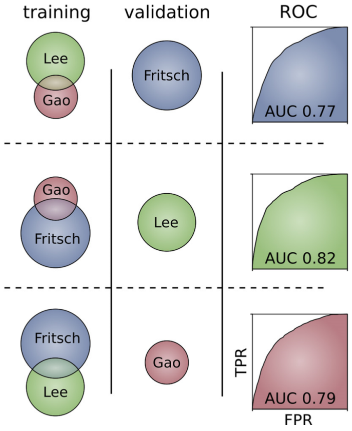
- Ribosome profiling experiments have low overlap in identified uORFs.
- This suggests high false-negative rate, and more functional uORFs than currently known.

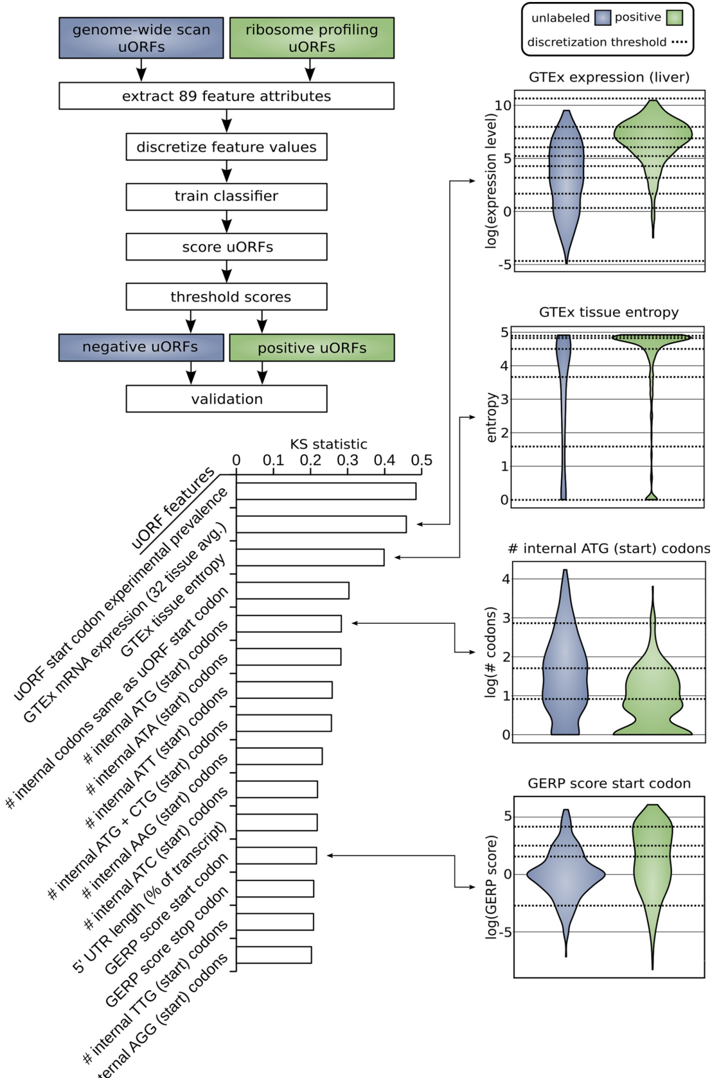
[McGillivray et al., *NAR* (‘18)]

# 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).







Expr. Level

Tissue Dist.

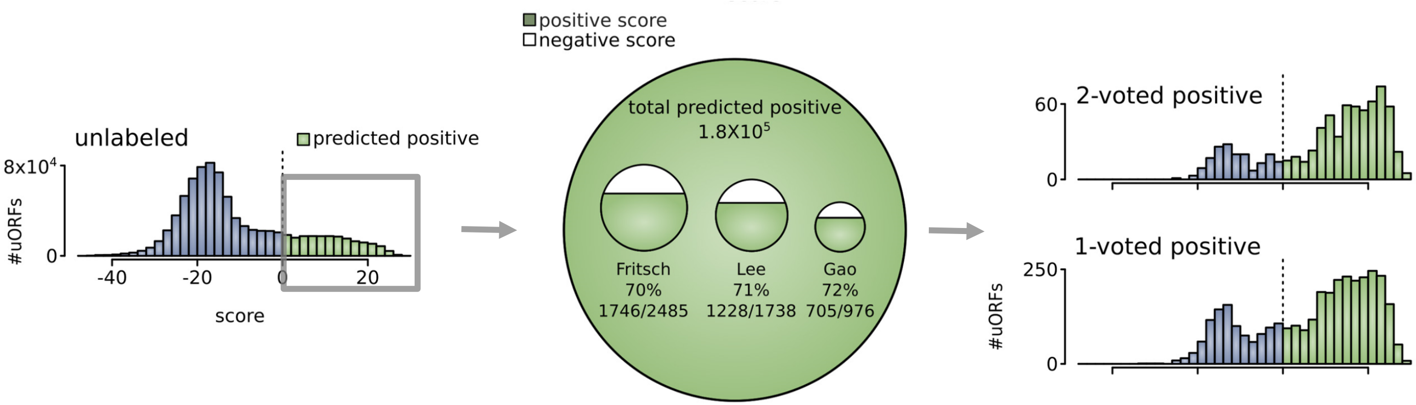
Int. ATG Start

Conser-vation

[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)]

# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a "surprisal" based weighting scheme.
  - Prioritizing variants within "sensitive sites" (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

# 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**

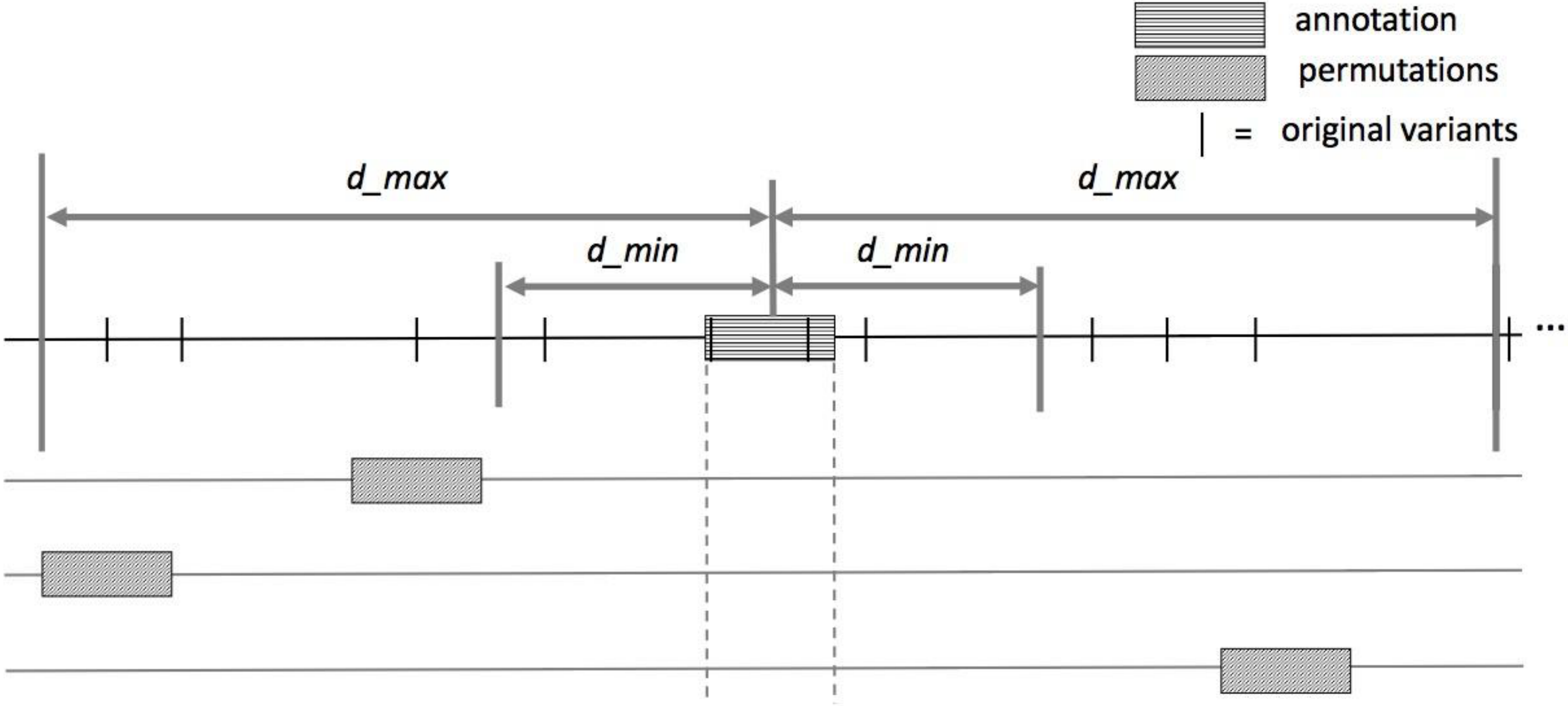
Shuffle annotations within local region to assess background mutation rate.

**Model 3b: Random Permutation of Input Variants**

Shuffle variants within local region to assess background mutation rate.

[Lochovsky et al. *Bioinformatics* in press]

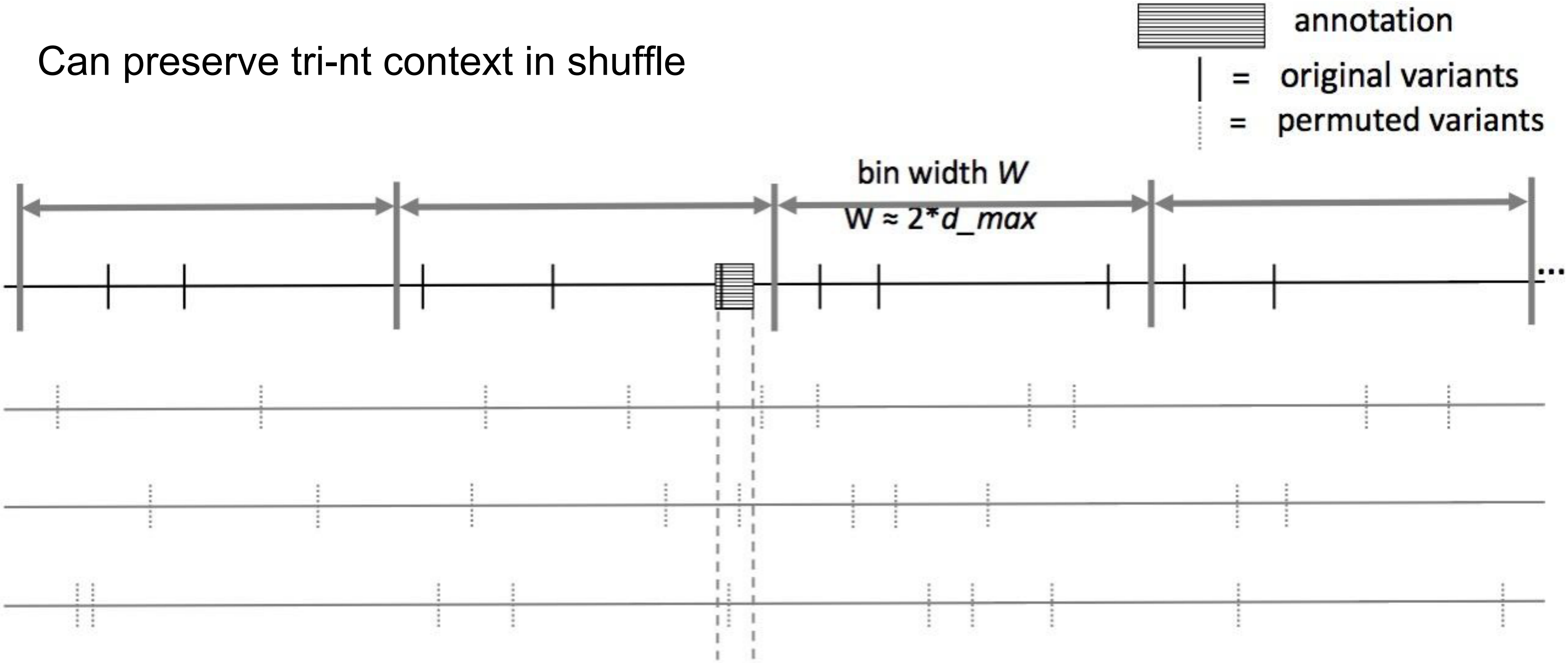
# MOAT-a: Annotation-based permutation



[Lochovsky et al. *Bioinformatics* in press]

# MOAT-v: Variant-based Permutation

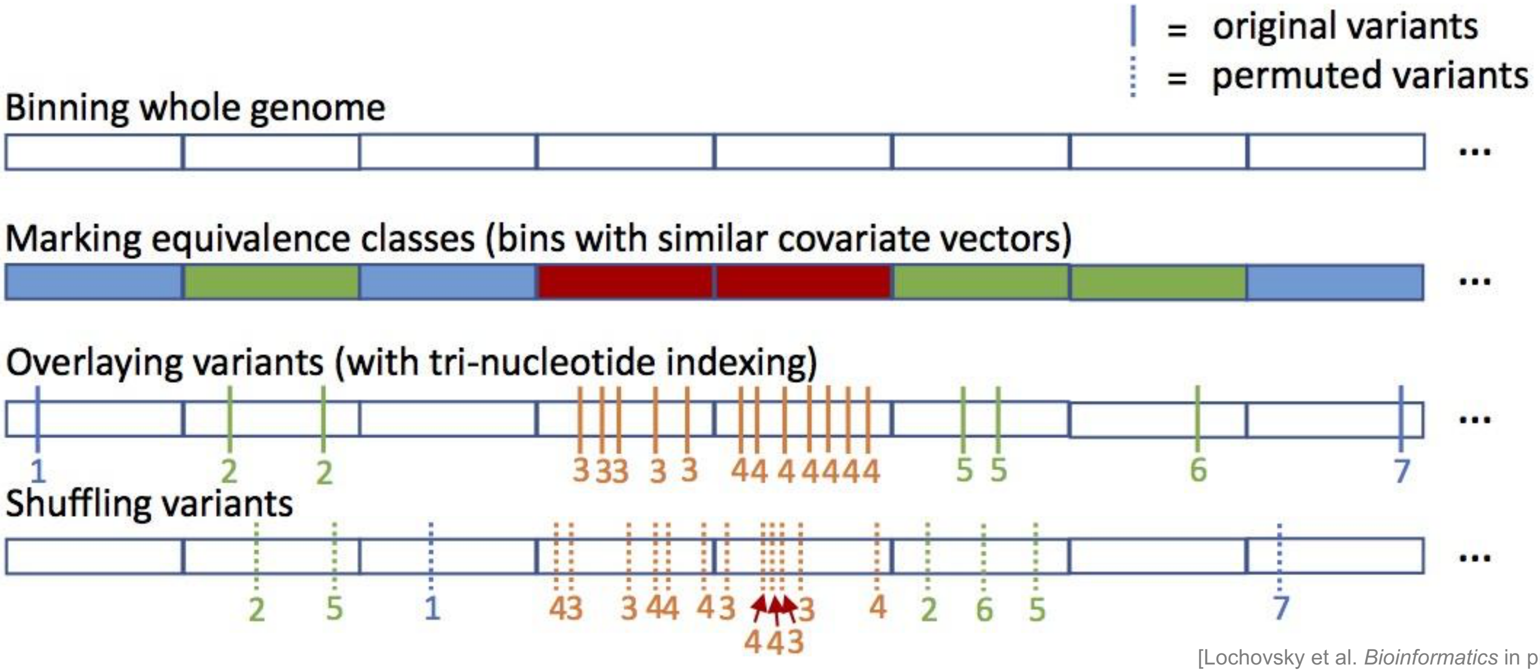
Can preserve tri-nt context in shuffle



[Lochovsky et al. *Bioinformatics* in press]

# MOAT-s: a variant on MOAT-v

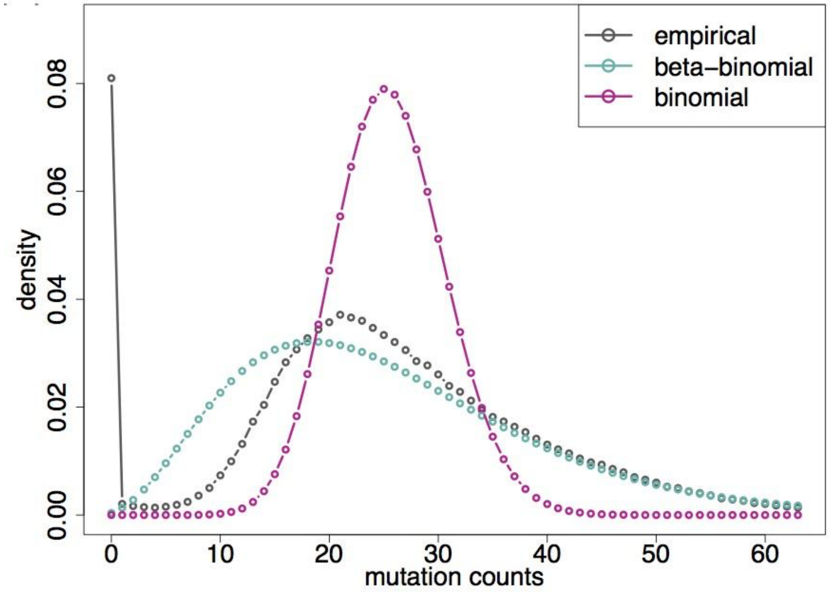
- A somatic variant simulator
  - Given a set of input variants, shuffle to new locations, taking genome structure into account



[Lochovsky et al. *Bioinformatics* in press]

# LARVA Model Comparison

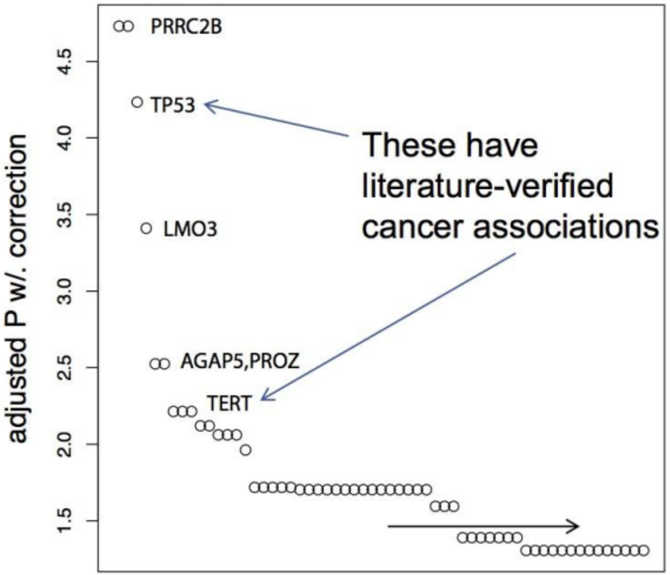
- Comparison of mutation count frequency implied by the binomial model (model 1) and the beta-binomial model (model 2) relative to the empirical distribution
- The beta-binomial distribution is significantly better, especially for accurately modeling the over-dispersion of the empirical distribution



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

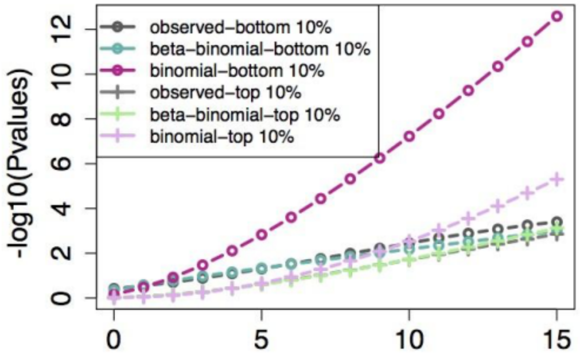
# LARVA Results

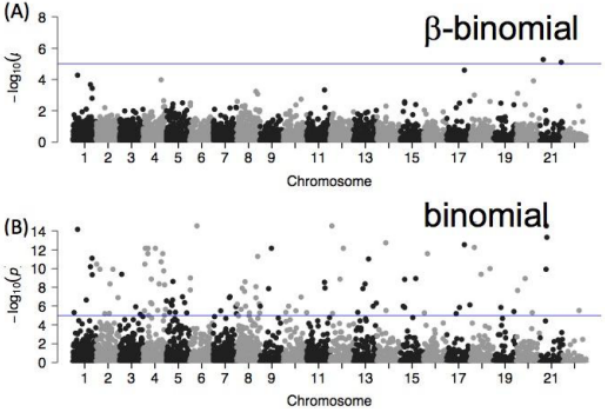
TSS LARVA results



These have literature-verified cancer associations

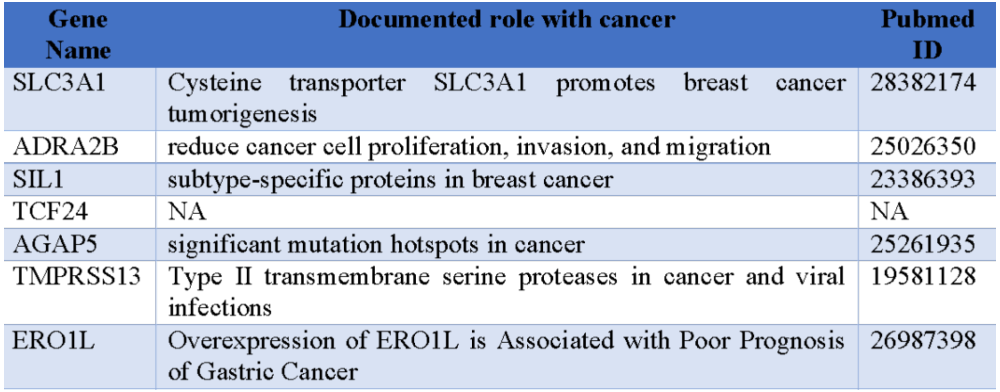
noncoding annotation p-values in sorted order





[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 |

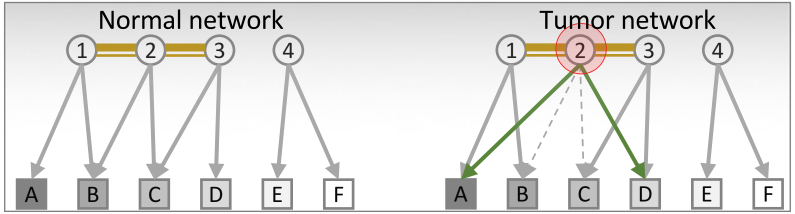
[Lochovsky et al. *Bioinformatics* in press]

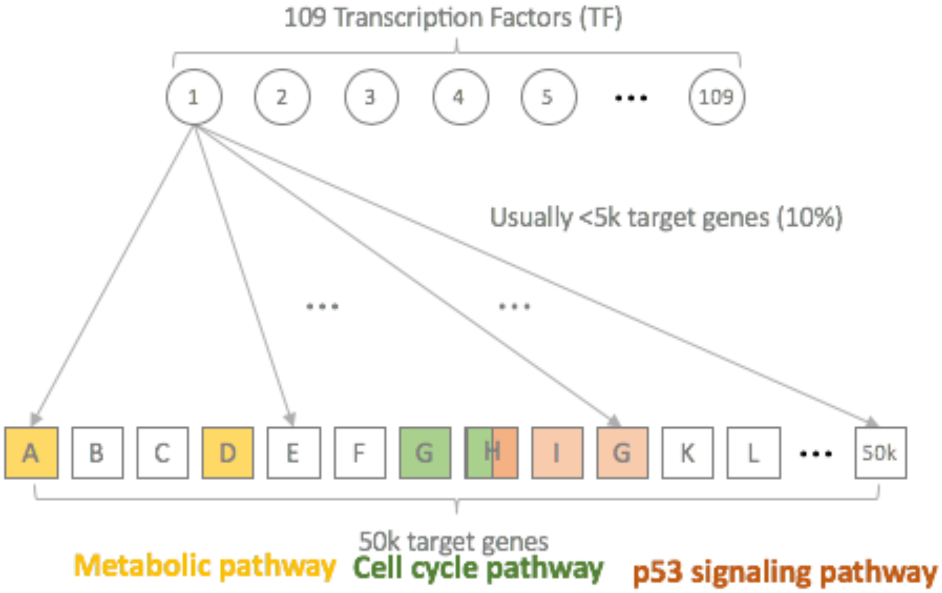
# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a "surprisal" based weighting scheme.
  - Prioritizing variants within "sensitive sites" (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

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

| | |
|---|---|
| Fact | TF→ *gene* regulation is important |
| Hypothesis | Disease-associated TFs have target gain or loss events |
| Method | Latent Dirichlet Allocation |





## Biology Intuition

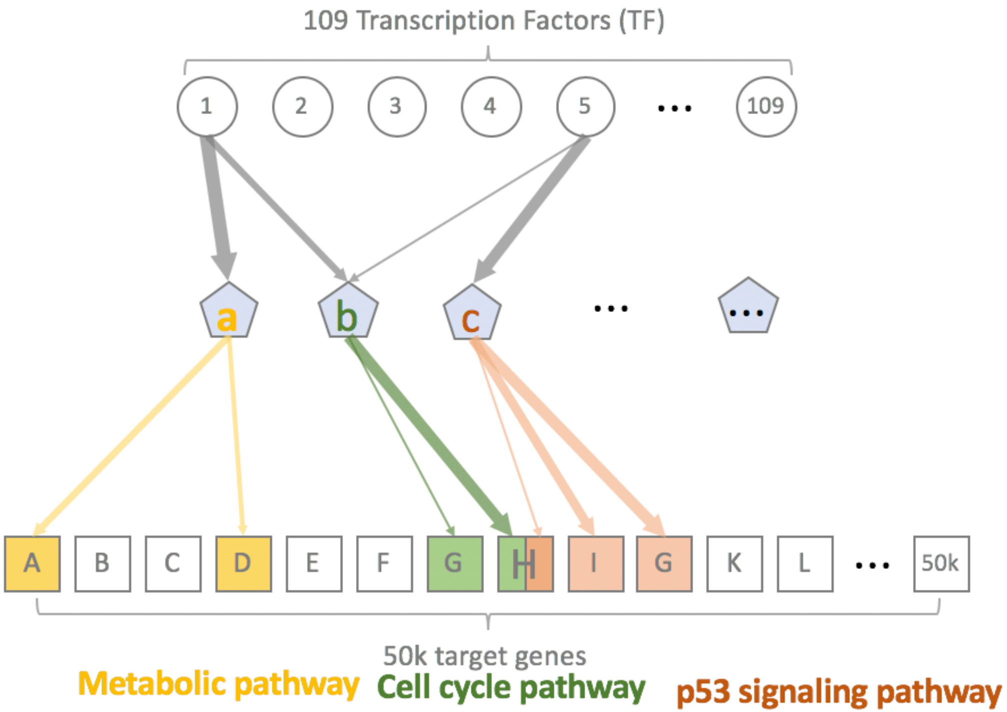
**Sparse & noisy** network: ~50k target genes in total, <10% active in one cell type

**Interpretability**: natural units are molecular pathways (unobserved)

**Soft clustering**: may have significant overlapping between pathways

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

# De-noising process by dimension reduction



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

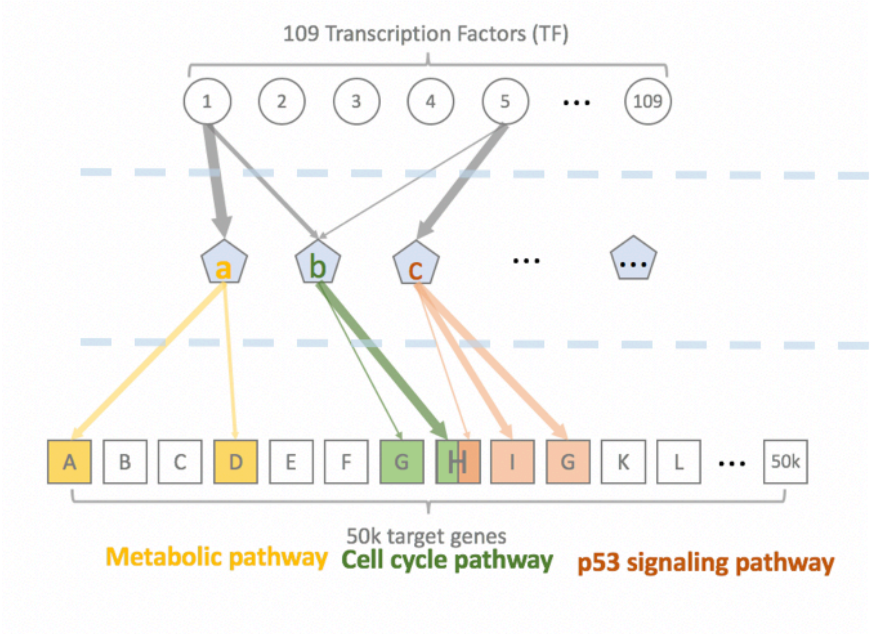
Hidden Layer
(50 biological pathways?)

Challenge: how to define appropriate pathways?

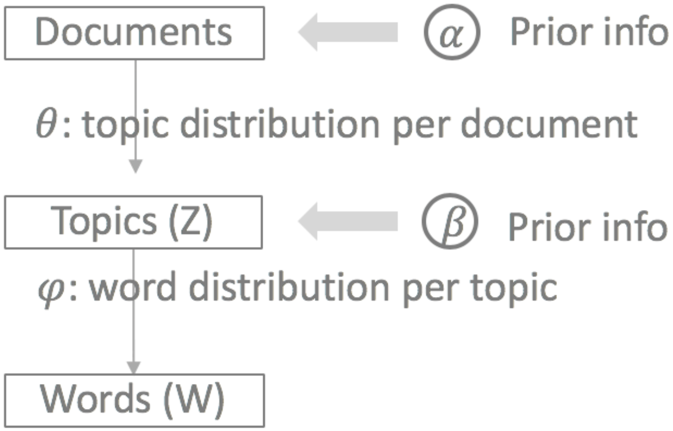
[Zhang et al. ('19), biorxiv.org]

# RegLDA: automatic gene topic identification based on Latent Dirichlet Allocation

$TF \rightarrow gene$ network



Latent Dirichlet Allocation



Documents

$\alpha$ Prior info

$\theta$: topic distribution per document

Topics (Z)

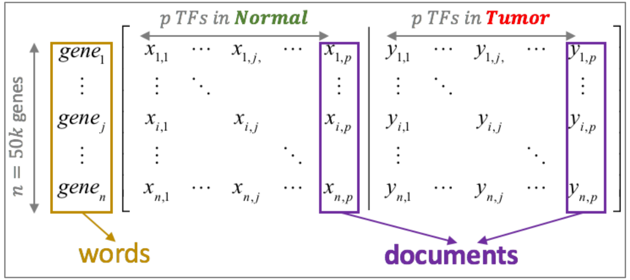
$\beta$ Prior info

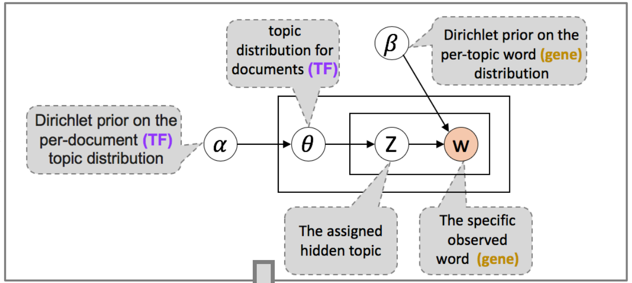
$\varphi$: word distribution per topic

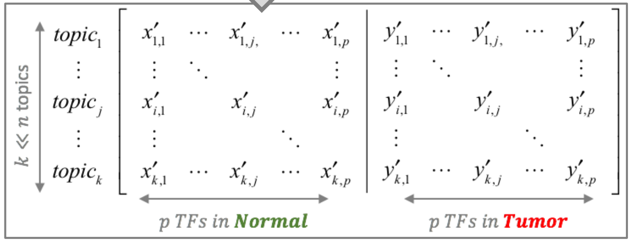
Words (W)

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

# latent Dirichlet allocation


p TFs in Normal
p TFs in Tumor
n = 50k genes
gene$_1$
gene$_j$
gene$_n$
$x_{1,1}$ ⋯ $x_{1,j,}$ ⋯ $x_{1,p}$
$x_{i,1}$ $x_{i,j}$ $x_{i,p}$
$x_{n,1}$ ⋯ $x_{n,j}$ ⋯ $x_{n,p}$
$y_{1,1}$ ⋯ $y_{1,j,}$ ⋯ $y_{1,p}$
$y_{i,1}$ $y_{i,j}$ $y_{i,p}$
$y_{n,1}$ ⋯ $y_{n,j}$ ⋯ $y_{n,p}$
words
documents


topic distribution for documents (TF)
β
Dirichlet prior on the per-topic word (gene) distribution
Dirichlet prior on the per-document (TF) topic distribution
α
θ
Z
w
The assigned hidden topic
The specific observed word (gene)

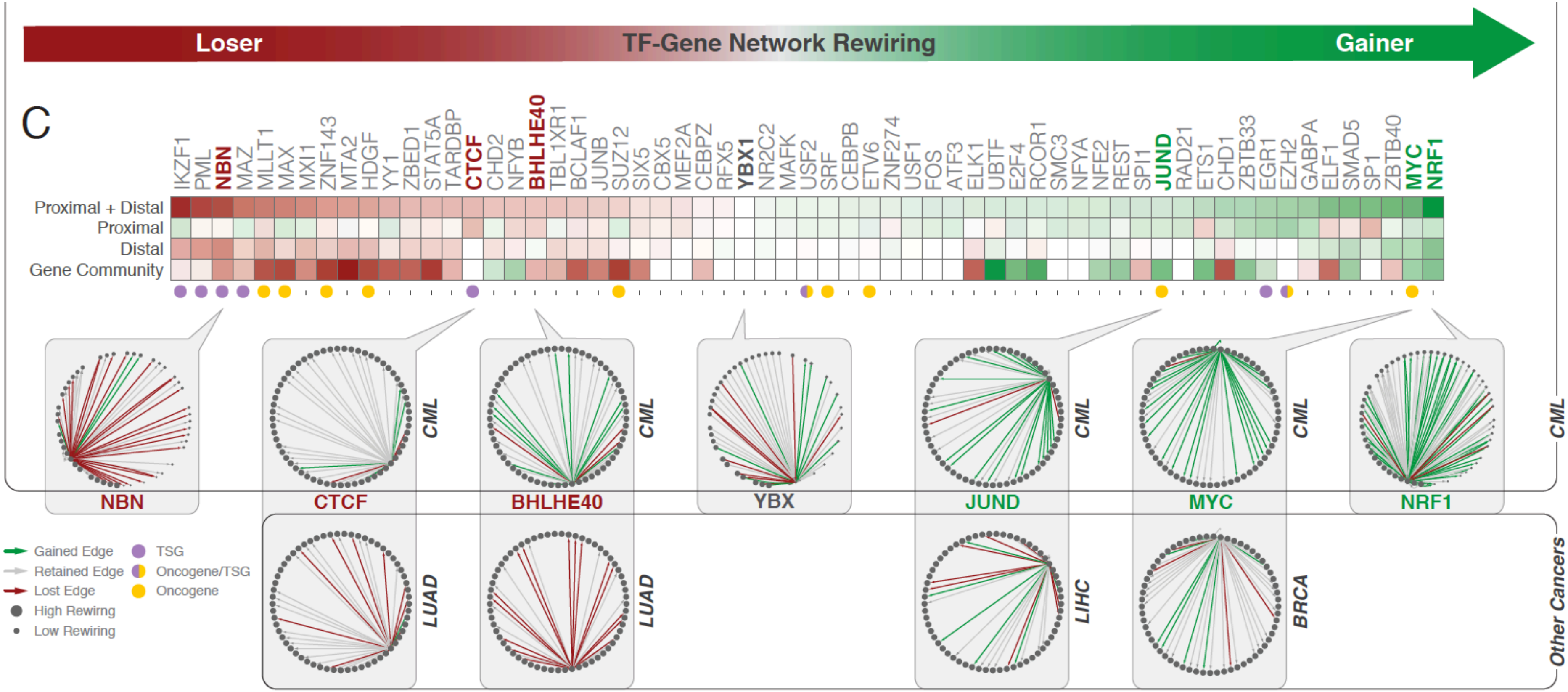

k ≪ n topics
topic$_1$
topic$_j$
topic$_k$
$x'_{1,1}$ ⋯ $x'_{1,j,}$ ⋯ $x'_{1,p}$
$x'_{i,1}$ $x'_{i,j}$ $x'_{i,p}$
$x'_{k,1}$ ⋯ $x'_{k,j}$ ⋯ $x'_{k,p}$
$y'_{1,1}$ ⋯ $y'_{1,j,}$ ⋯ $y'_{1,p}$
$y'_{i,1}$ $y'_{i,j}$ $y'_{i,p}$
$y'_{k,1}$ ⋯ $y'_{k,j}$ ⋯ $y'_{k,p}$
p TFs in Normal
p TFs in Tumor

Gain/Loss Summary Statistic on Topics

$$\theta^{tumor} = (0.9,0.05,0.05)$$
$$\theta^{normal} = (0.05,0.05,0.9)$$

$$\theta^{tumor} = (0.9,0.05,0.05)$$
$$\theta^{normal} = (0.85,0.05,0.1)$$

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

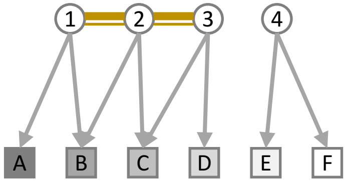


[Zhang et al. ('19), biorxiv.org]

# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a “surprisal” based weighting scheme.
  - Prioritizing variants within “sensitive sites” (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

Normal Network



═ Co-regulation

→ TF/RBP to gene

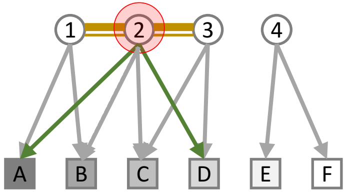
○ TF/RBP

□ gene

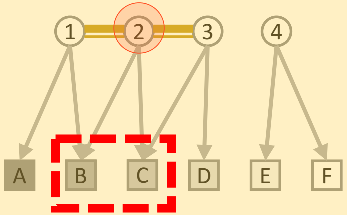
High expression ▨ low expression

Disease Network ：
dotted line = lost edge

Principles

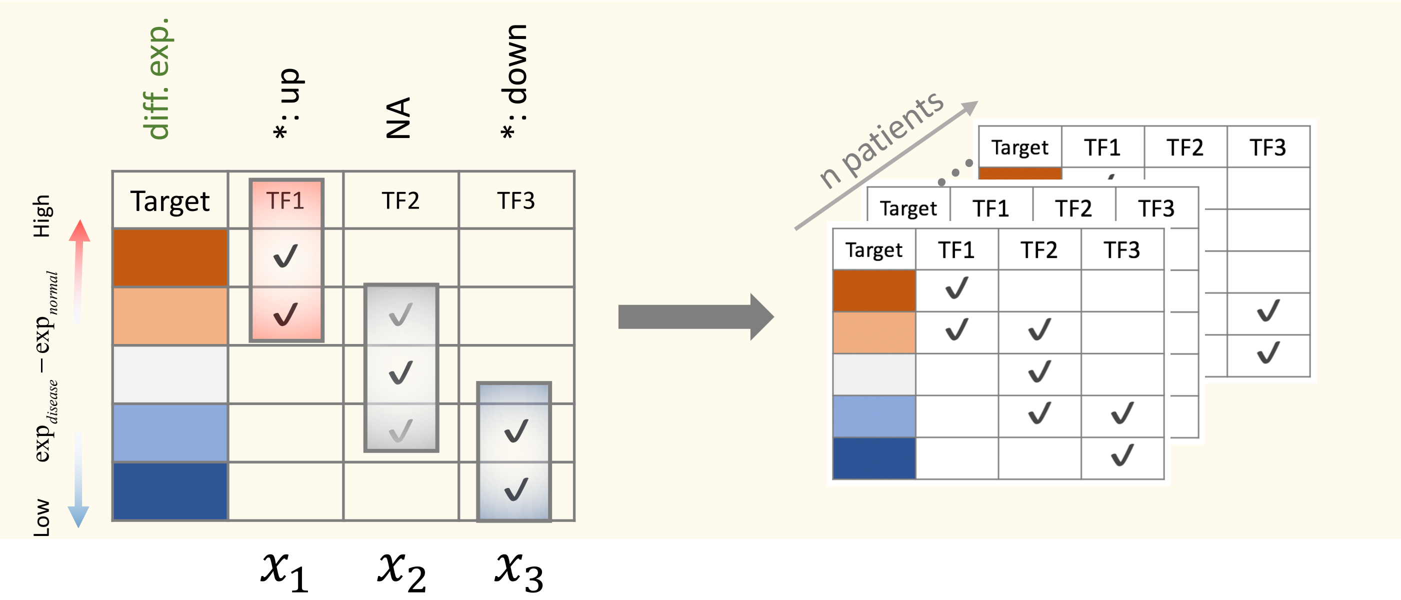


Direct target
gain/loss



Target gene
expression changes

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



| Target | TF1 | TF2 | TF3 |
|---|---|---|---|
| 🟧 | ✓ | | |
| 🟧 | ✓ | ✓ | |
| ⬜ | | ✓ | |
| 🟦 | | ✓ | ✓ |
| 🟦 | | | ✓ |

| Target | TF1 | TF2 | TF3 |
|---|---|---|---|
| 🟧 | ✓ | | |
| 🟧 | ✓ | ✓ | |
| ⬜ | | ✓ | |
| 🟦 | | ✓ | ✓ |
| 🟦 | | | ✓ |

$\beta_1 = 2.5$ $\beta_3 = -3.1$

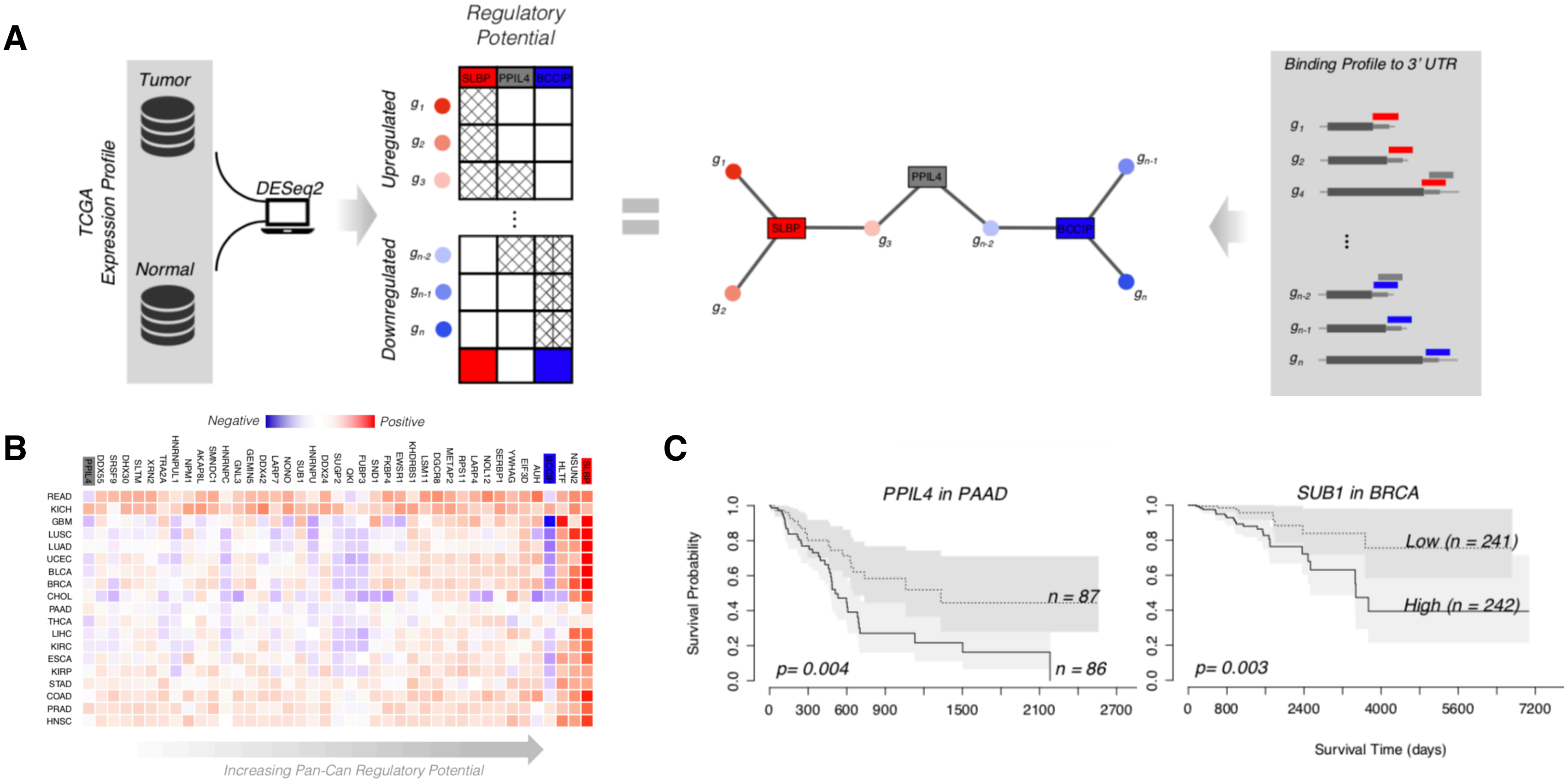
2198 ChIP-seq

459 eCLIP

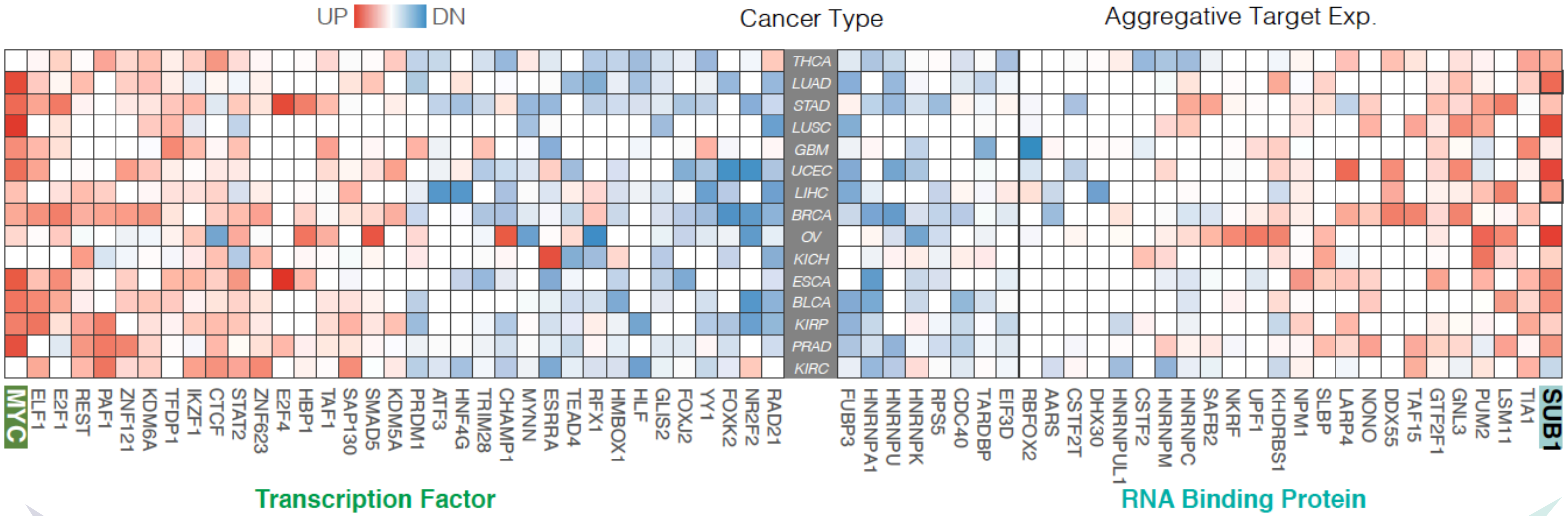
$$\underbrace{y=\left(\exp_{disease}-\exp_{normal}\right)}_{\text{differential expression}} \sim \beta_0+\underbrace{\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}_{\text{Network for Regulator 1 to k}}$$

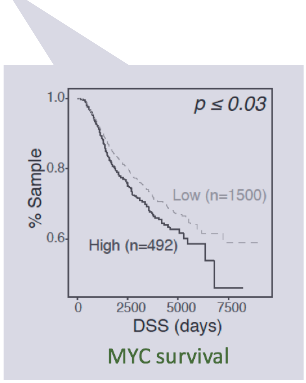
[Zhang et al. ('19), biorxiv.org]

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

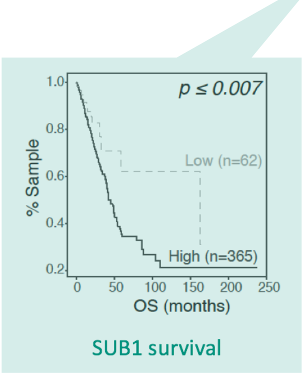


[Zhang*, Liu* et al., *Genome Biology* (in review '18)]



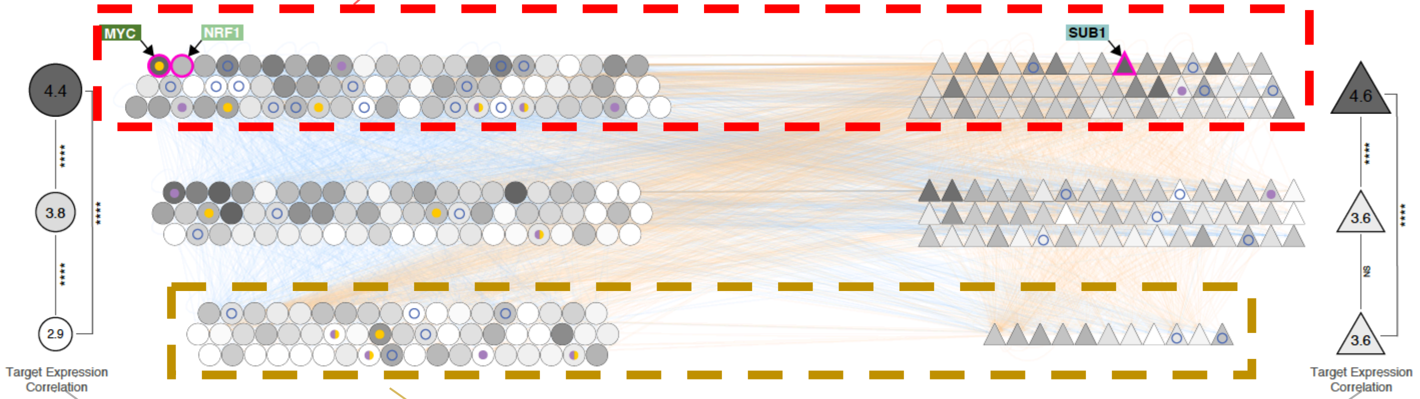


**Aggregated** t-statistic in regression over TCGA samples



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

Top Layer: Master regulators, regulating others more than being regulated

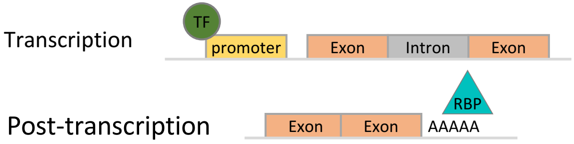


**Bottom Layer: follower regulators, being regulated more than regulating others**

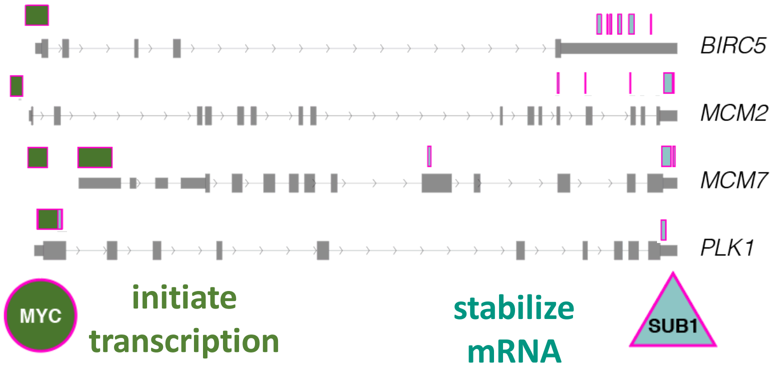
How much power each regulator has in driving tumor-normal differential expressions

TF-RBP crosstalk

TF-RBP regulate the same gene at different levels



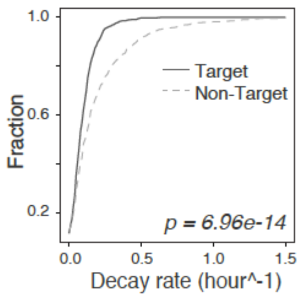
[Zhang et al. ('19), biorxiv.org]

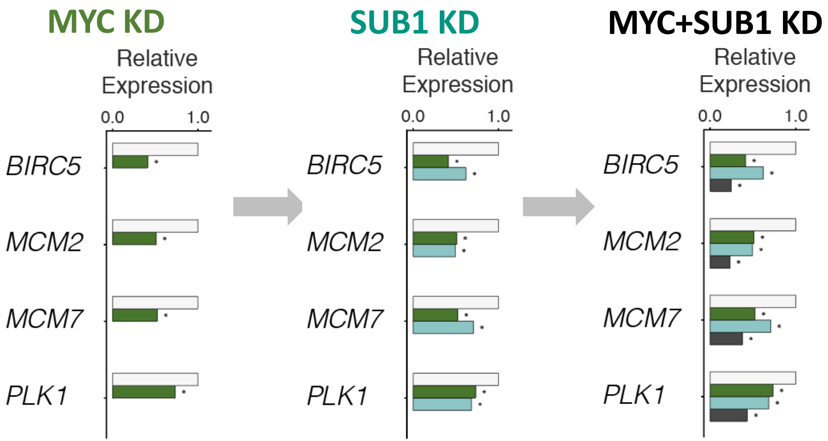


initiate transcription

stabilize mRNA

Slower mRNA decay rate in SUB1 targets




MYC KD

SUB1 KD

MYC+SUB1 KD

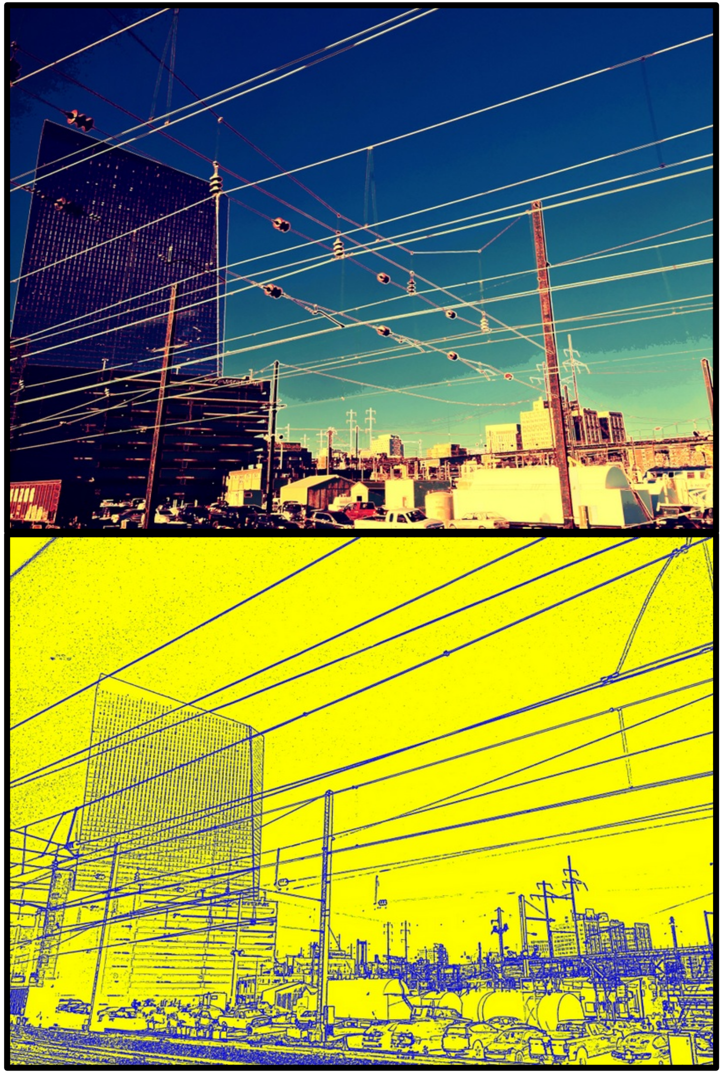
# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a "surprisal" based weighting scheme.
  - Prioritizing variants within "sensitive sites" (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

# Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

- **Background**
  - PMI & Variant Prioritization
  - Types of annotations: peaks, segmentations, regulators
  - Genomic covariates
  - ENCODEC: ENCODE cancer annotation resource
- **Matched Filter Annotation**
  - Integrating cross-assay signal-track patterns associated with enhancers
  - Trained on high throughput STARR-seq experiments
  - Validation in many different contexts
- **FunSeq Prioritization**
  - Integrates evidence, with a "surprisal" based weighting scheme.
  - Prioritizing variants within "sensitive sites" (human conserved)
- **RADAR Prioritization**
  - Adapts FunSeq approach to RBPs
  - Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP
  - Incorporates new features related to RNA sec. struc & tissue specific effects
- **uORF Prioritization**
  - Feature integration to find small subset of upstream mutations that potentially alter translation
- **LARVA & MOAT**
  - Uses parametric beta-binomial model, explicitly modeling genomic covariates
  - Non-parametric shuffles. Useful when explicit covariates not available.
- **Network Rewiring**
  - Network rewiring highlights regulators that change their targets greatly.
  - LDA approach specifically finds those that greatly change their gene communities
- **Regulatory Drivers of Differential Expression**
  - Highlighting regulators in terms of their power to drive differential expression.
  - Relationship of this to network hierarchy & RBP-TF cross talk
  - Example of MYC & SUB1

**Ack**nowledgments! Also, Hiring: See **Jobs**.gersteinlab.org



**ENCODEC**.gersteinlab.org

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

github.com/gersteinlab/**MatchedFilter**

A **Sethi**, M **Gu**, E Gumusgoz, L Chan, KK Yan, J Rozowsky, A Harrington, B Mannion, E Lee, Y Fukuda-Yuzawa, A Visel, D Dickel, K Yip, R Sutton, LA Pennacchio

**FunSeq**.gersteinlab.org

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

**RADAR.**gersteinlab.org

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

github.gersteinlab.org/**uORFs**

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

**{LARVA,MOAT}.**gersteinlab.org

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

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