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

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>174,650</td>
<td>268,600</td>
</tr>
<tr>
<td>Lung &amp; bronchus</td>
<td>116,440</td>
<td>111,710</td>
</tr>
<tr>
<td>Colon &amp; rectum</td>
<td>78,500</td>
<td>67,100</td>
</tr>
<tr>
<td>Urinary bladder</td>
<td>61,700</td>
<td>61,880</td>
</tr>
<tr>
<td>Melanoma of the skin</td>
<td>57,220</td>
<td>39,260</td>
</tr>
<tr>
<td>Kidney &amp; renal pelvis</td>
<td>44,120</td>
<td>37,810</td>
</tr>
<tr>
<td>Non-Hodgkin lymphoma</td>
<td>41,090</td>
<td>33,110</td>
</tr>
<tr>
<td>Oral cavity &amp; pharynx</td>
<td>38,140</td>
<td>29,700</td>
</tr>
<tr>
<td>Leukemia</td>
<td>35,920</td>
<td>26,830</td>
</tr>
<tr>
<td>Pancreas</td>
<td>29,940</td>
<td>25,860</td>
</tr>
<tr>
<td><strong>All Sites</strong></td>
<td><strong>870,970</strong></td>
<td><strong>891,480</strong></td>
</tr>
</tbody>
</table>

**Estimated numbers:**
- 1,762,450 new cases per year
- ~4,800 new cases per day

Segiel et al, Cancer statistics, 2019
Much Interest in Precision Oncology

- Analysis of the exact somatic mutations in an 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

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

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

<table>
<thead>
<tr>
<th>Origin of Variants</th>
<th>Coding</th>
<th>Non-coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germ-line</td>
<td>22K</td>
<td>4.1 – 5M</td>
</tr>
<tr>
<td>Somatic</td>
<td>~50</td>
<td>5K</td>
</tr>
</tbody>
</table>

A Typical Genome

<table>
<thead>
<tr>
<th>Class of Variants</th>
<th>SNP</th>
<th>Indel</th>
<th>SV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.5 – 4.3M</td>
<td>550 – 625K</td>
<td>2.1 – 2.5K (20Mb)</td>
</tr>
<tr>
<td>Total</td>
<td>4.1 – 5M</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Population of 2,504 peoples

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

**Origin of Variants**

- **Coding**: 22K (Germ-line) / ~50 (Somatic)
- **Non-coding**: 4.1 – 5M (Germ-line) / 5K (Somatic)

**Class of Variants**

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

**Prevalence of Variants**

- **Common**: ~75%
- **Rare* (1-4%)**: ~25%

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


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

**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

Image adapted from NHGRI
Non-coding Annotations: Overview

Features are often present on multiple "scale" (e.g., elements and connected networks).

Sequence features, incl. **Conservation**

- 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**

- 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

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

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
  - ....


J. Zhou, O.G. Troyanskaya, Nat. Methods, 2015
Major takeaway from annotation experience for disease studies: *less is more*

<table>
<thead>
<tr>
<th>Disease</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>rare</td>
<td>a few with high impact</td>
</tr>
<tr>
<td>common</td>
<td>many with weak effect</td>
</tr>
<tr>
<td>cancer</td>
<td>a few drivers</td>
</tr>
</tbody>
</table>

Example of power issue in disease studies

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

- DNA nucleotide
- Gene expression (high/low)
- Active proximal elements
- Active distal elements
- Inactive proximal elements
- Inactive distal elements
- Regulator
- Regulator co-regulation network
- Regulator→gene regulatory network
- Regulator→regulator co-regulation network
- Regulator to target gene directional
- Regulator co-regulation unidirectional

Gene regulatory network

Active proximal elements

Inactive distal elements

Epigenetic information

Major Challenges:

- Many levels of dysregulations related to disease status

...AGCTTTACGATCCGAAATCTGGTATACGATCCGAAATC...
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
  - ENCODE: 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

Cancer Type 1

Cancer Type 2

Cancer Type 3
Mutation recurrence

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
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
## ENCODEC

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

### Assay Approach
- **Breadth Approach**: 86 Cancerous (40 Cancer Types) + 143 Composite Normal (inc. Roadmap)
- **Depth Approach**: 528 ENCODE Cell Types, 229 ENCODED & Selected Human Biosamples

### Assays

<table>
<thead>
<tr>
<th>Assay</th>
<th>K562</th>
<th>HeLaG3</th>
<th>A549</th>
<th>MCF-7</th>
<th>HeLa-S3</th>
<th>HT-1080</th>
<th>Caco-2</th>
<th>HCT116</th>
<th>Panc1</th>
<th>LNCaP</th>
<th>PC-3</th>
<th>PC-9</th>
<th>SK-H-406</th>
<th>DLD-1</th>
<th>SW-620</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Histone ChIP-seq</strong></td>
<td>19</td>
<td>14</td>
<td>85</td>
<td>16</td>
<td>14</td>
<td>53</td>
<td>3</td>
<td>16</td>
<td>7</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td>8</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td><strong>RNA-seq</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RAMPAGE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>eCLIP</strong></td>
<td>191</td>
<td>164</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>shRNA/siRNA KD</strong></td>
<td>326</td>
<td>257</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CRISPR KD/KO</strong></td>
<td>108</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ChIA-PET</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hi-C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>STARR-seq</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>WGBS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RRBS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Repli-chip</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Repli-seq</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TF ChIP-seq</strong></td>
<td>558</td>
<td>300</td>
<td>240</td>
<td>149</td>
<td>78</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Cytoscape Network Diagram
- **Extended Gene**: TF, RBP
- **Network Hierarchy**: Transcription Factors, Cell Line WGS
- **Network Rewiring**: Tumor, Normal
ENCODEC

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

Gene-centric: Extended Genes (proximal & distal)

Network Hierarchy

Network Rewiring

[Zhang et al. ('19), biorxiv.org]
Compact & accurate: Enhancer, promoter, TF/RBP binding

Gene-centric: Extended Genes (proximal & distal)

Network: Regulatory networks

Network Hierarchy

Encodec.enodeproject.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
Matched Filter recognize shape patterns

Matched Filter

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

<table>
<thead>
<tr>
<th>Model</th>
<th>AUROC</th>
<th>AUPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.96 (0.95)</td>
<td>0.91 (0.79)</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>0.95 (0.94)</td>
<td>0.90 (0.77)</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>0.96 (0.95)</td>
<td>0.91 (0.78)</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.95 (0.93)</td>
<td>0.89 (0.72)</td>
</tr>
</tbody>
</table>

Cross validation

Large scale STARR-seq experiment data helps to improve the performance of integrated model
Validation with transgenic mouse enhancer assay

Inject fertilized eggs

Transplant to surrogate mother

Visualize Reporter gene Expression In E 11.5 embryos

[ biornxiv.org/content/early/2018/08/05/385237 ]
Matched-Filter can be applied across different organisms

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

Compare overlap with FANTOM5 enhancers

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

[ biorxiv.org/content/early/2018/08/05/385237 ]
Constructing a high-confidence set of cell-specific enhancers

[Zhang et al. ('19), biorxiv.org]
Disease Genomics: Thoughts on Genome Annotation, Prioritizing Variants, Highlighting Dysregulation, & the Application of all of these to Cancer

<table>
<thead>
<tr>
<th>Background</th>
<th>RADAR Prioritization</th>
<th>Network Rewiring</th>
<th>Regulatory Drivers of Differential Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMI &amp; Variant Prioritization</td>
<td>Adapts FunSeq approach to RBPs</td>
<td>Network rewiring highlights regulators that change their targets greatly.</td>
<td>Highlighting regulators in terms of their power to drive differential expression.</td>
</tr>
<tr>
<td>Types of annotations: peaks, segmentations, regulators</td>
<td>Prioritizes variants based on post-transcriptional regulome using ENCODE eCLIP</td>
<td>LDA approach specifically finds those that greatly change their gene communities</td>
<td>Relationship of this to network hierarchy &amp; RBP-TF cross talk</td>
</tr>
<tr>
<td>Genomic covariates</td>
<td>Incorporates new features related to RNA sec. struc &amp; tissue specific effects</td>
<td></td>
<td>Example of MYC &amp; SUB1</td>
</tr>
<tr>
<td>ENCODEC: ENCODE cancer annotation resource</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Matched Filter Annotation</th>
<th>uORF Prioritization</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrating cross-assay signal-track patterns associated with enhancers</td>
<td>Feature integration to find small subset of upstream mutations that potentially alter translation</td>
<td></td>
</tr>
<tr>
<td>Trained on high throughput STARR-seq experiments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation in many different contexts</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FunSeq Prioritization</th>
<th>LARVA &amp; MOAT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrates evidence, with a “surprisal” based weighting scheme.</td>
<td>Uses parametric beta-binomial model, explicitly modeling genomic covariates</td>
<td></td>
</tr>
<tr>
<td>Prioritizing variants within “sensitive sites” (human conserved)</td>
<td>Non-parametric shuffles. Useful when explicit covariates not available.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network Rewiring</th>
<th>Regulatory Drivers of Differential Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>LARVA &amp; MOAT</td>
<td>Highlighting regulators in terms of their power to drive differential expression.</td>
</tr>
<tr>
<td></td>
<td>Relationship of this to network hierarchy &amp; RBP-TF cross talk</td>
</tr>
<tr>
<td></td>
<td>Example of MYC &amp; SUB1</td>
</tr>
</tbody>
</table>
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)
Finding "Conserved" Sites in the Human Population:

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

Broad categories of regulatory regions under negative selection
Related to:

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)]
• 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
  - 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

\[ w_d = 1 + p_d \log_2 p_d + (1 - p_d) \log_2 (1 - p_d) \]

- 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)

- **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))

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

\[ S_{\text{Universal}} = w_V V + w_H H + K + w_R R + G \]

\[ S_{\text{Full}} = S_{\text{Universal}} + E + G + M \]

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

Enriched rare DAF in eCLIP peaks

RNA Structure Cons. from Evofold

[Zhang*, Liu* et al., Genome Biology (in review '18)]
Co-binding of RBPs form biologically relevant complexes

Literature supported RBP complexes

Binding hubs are enriched for rare variants

[Rare DAF]

Unusual co-binding patterns of RBPs

[Hub Number (Hotness)]

[Hub Number (Hotness)]

[Zhang*, Liu* et al., Genome Biology (in review ‘18)]
RADAR Scores enriched in COSMIC genes and recurrently mutated regions

A

B

[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

- 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.

[Ferreira et al., Bioengineered ('14)]

[Calvo et al., PNAS ('09)]

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

[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)
\[ x_i \sim \text{Binomial}(n_i, p) \]

Model 2a: Varying Mutation Rate with Single Covariate Correction
\[ x_i \sim \text{Binomial}(n_i, p_i) \]
\[ p_i \sim \text{Beta}(\mu | R_i, \sigma | R_i) \]
\[ \mu | R_i, \sigma | R_i : \text{constant within the same covariate rank} \]

Model 2b: Varying Mutation Rate with Multiple Covariate Correction
\[ x_i \sim \text{Binomial}(n_i, p_i) \]
\[ p_i \sim \text{Beta}(\mu | R_i, \sigma | R_i) \]
\[ \mu | R_i, \sigma | R_i : \text{constant within the same covariate rank} \]

NON-PARAMETRIC MODELS

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.

• 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)

[Lochovsky et al. NAR ('15)]

[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

[Lochovsky et al. NAR ('15)]
MOAT: recapitulates LARVA with GPU-driven runtime scalability

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

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

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

<table>
<thead>
<tr>
<th>Number of permutations</th>
<th>Fold speedup of CUDA version</th>
</tr>
</thead>
<tbody>
<tr>
<td>1k</td>
<td>14x</td>
</tr>
<tr>
<td>10k</td>
<td>100x</td>
</tr>
<tr>
<td>100k</td>
<td>256x</td>
</tr>
</tbody>
</table>

[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
  - ENCODE: 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 re-wiring analyses**: key cancer-associated regulator identification through network comparisons

<table>
<thead>
<tr>
<th>Fact</th>
<th>TF → gene regulation is important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis</td>
<td>Disease-associated TFs have target gain or loss events</td>
</tr>
<tr>
<td>Method</td>
<td>Latent Dirichlet Allocation</td>
</tr>
</tbody>
</table>

**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
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

[Zhang et al. ('19), biorxiv.org]
Gain/Loss Summary Statistic on Topics

\[ \theta^{\text{tumor}} = (0.9, 0.05, 0.05) \]
\[ \theta^{\text{normal}} = (0.05, 0.05, 0.9) \]

\[ \theta^{\text{tumor}} = (0.9, 0.05, 0.05) \]
\[ \theta^{\text{normal}} = (0.85, 0.05, 0.1) \]

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

TF-Gene Network Rewiring

Loser  Gainer

Proximal + Distal  Proximal  Distal  Gene Community

NBN  CTCF  BHLHE40  YBX  JUND  MYC  NRF1

Gained Edge  TSG  Retained Edge  Oncogene/TSG  Lost Edge  Oncogene

High Rewiring  Low Rewiring

[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

Disease Network:
dotted line = lost edge

Co-regulation
TF/RBP to gene
TF/RBP
Gene
High expression
low expression

Direct target gain/loss

Target gene expression changes

Principles

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

Lectures.gersteinlab.org
\[
\begin{align*}
\beta_1 &= 2.5 \\
\beta_3 &= -3.1
\end{align*}
\]

2198 ChIP-seq
459 eCLIP

\[
y = \left( \frac{\text{exp}_{\text{disease}} - \text{exp}_{\text{normal}}}{\text{differential expression}} \right) \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k
\]

Network for Regulator 1 to k

Zhang et al. ('19), biorxiv.org
Lectures.gersteinlab.org
Regulatory Potential of RBPs derived from regression between gene network and expression levels

A

B

C

[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
initiate transcription  stabilize mRNA

Slower mRNA decay rate in SUB1 targets

MYC KD  SUB1 KD  MYC+SUB1 KD

<table>
<thead>
<tr>
<th></th>
<th>Relative Expression</th>
<th></th>
<th>Relative Expression</th>
<th></th>
<th>Relative Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>BIRC5</td>
<td>*</td>
<td>*</td>
<td>BIRC5</td>
<td>*</td>
<td>BIRC5</td>
</tr>
<tr>
<td>MCM2</td>
<td>*</td>
<td>*</td>
<td>MCM2</td>
<td>*</td>
<td>MCM2</td>
</tr>
<tr>
<td>MCM7</td>
<td>*</td>
<td>*</td>
<td>MCM7</td>
<td>*</td>
<td>MCM7</td>
</tr>
<tr>
<td>PLK1</td>
<td>*</td>
<td>*</td>
<td>PLK1</td>
<td>*</td>
<td>PLK1</td>
</tr>
</tbody>
</table>
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
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