Topics in Precision Oncology: Addressing the role of the non-coding genome in cancer

Mark Gerstein
Yale

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @MarkGerstein).
No Conflicts for this Talk. See last slide for more info.
What if matching a cancer cure to our genetic code was just as easy

- Sub-topic of precision medicine
- Analysis of the exact somatic mutations in an individual, suggesting individualized treatment

"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

https://obamawhitehouse.archives.gov/blog/2016/02/25/precision-medicine-health-care-tailored-you
Coding regions are only ~1-2% of the genome yet contain almost all the drivers.

Open Q: what is the role of the non-coding genome in cancer?
PCAWG: most comprehensive resource for cancer whole genome analysis

- Union of TCGA-ICGC efforts
  - Jointly analyzing ~2800 whole genome tumor/normal pairs
    - > 580 researchers
    - ~30M total somatic SNVs

Adapted from Campbell et. al., bioRxiv ('17).
Comprehensive non-coding Annotation

Applicable to cancer genomics
Topics in Precision Oncology:
Addressing the role of the non-coding genome in cancer

• **Background**
  • Drivers v passenger
  • Coding v noncoding
  • Pcawg & encode 3

• **Additive-Effects model to measure the Impact of non-coding v coding mutations**
  • Repurposing a formalism from germline genetics for missing heritability to cancer
  • Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  • Notable effect, particularly for non-coding passengers, in addition to known coding drivers.
  • Recasting as a predictive model to est. number of weak drivers

• **Network Rewiring in Cancer**
  • Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  • TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
Topics in Precision Oncology: Addressing the role of the non-coding genome in cancer

- **Background**
  - Drivers v passenger
  - Coding v noncoding
  - PcaWG & encode 3

- **Additive-Effects model to measure the Impact of non-coding v coding mutations**
  - Repurposing a formalism from germline genetics for missing heritability to cancer
  - Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  - Notable effect, particularly for non-coding passengers, in addition to known coding drivers.
  - Recasting as a predictive model to est. number of weak drivers

- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
Relating Germline Missing Heritability to Cancer Studies

Organismal trait: Height

Subclonal trait in cancer: Growth rate

Population level definitions:
- Parent-offspring heritability;
- Twin-based heritability...

SNP-based polygenic & additive model:

\[ h^2 = \sigma_u \]

\[ y = X\beta + Z\mu + \varepsilon \]

- Trait
- Covariates & fixed effects
- Genetic predictors & random effects
- Environmental noise

Tumor sample taken
Missing heritability for height & other traits

• Height is a highly polygenic trait:

<table>
<thead>
<tr>
<th>SNP category</th>
<th># SNPs</th>
<th>Heritability estimate ($h^2$)</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWAS SNPs$^1$</td>
<td>50</td>
<td>~0.05</td>
<td>2008</td>
</tr>
<tr>
<td>Common SNPs$^2$</td>
<td>~295K</td>
<td>0.54 (SE 0.1)</td>
<td>2010</td>
</tr>
<tr>
<td>Common-rare SNPs$^3$</td>
<td>47.1M</td>
<td>0.79 (SE 0.09)</td>
<td>2019</td>
</tr>
<tr>
<td>Population estimate (twins)$^4$</td>
<td>-</td>
<td>0.8</td>
<td>(2012)</td>
</tr>
</tbody>
</table>

SE = standard error

• Many other traits have substantial missing GWAS-based heritability$^5$:

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of loci</th>
<th>Proportion of heritability explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age-related macular degeneration$^7$</td>
<td>5</td>
<td>50%</td>
</tr>
<tr>
<td>Crohn’s disease$^{21}$</td>
<td>32</td>
<td>20%</td>
</tr>
<tr>
<td>Systemic lupus erythematosus$^3$</td>
<td>6</td>
<td>15%</td>
</tr>
<tr>
<td>Type 2 diabetes$^{24}$</td>
<td>18</td>
<td>6%</td>
</tr>
<tr>
<td>HDL cholesterol$^{25}$</td>
<td>7</td>
<td>5.2%</td>
</tr>
<tr>
<td>Height$^{15}$</td>
<td>40</td>
<td>5%</td>
</tr>
<tr>
<td>Early onset myocardial infarction$^{76}$</td>
<td>9</td>
<td>2.8%</td>
</tr>
<tr>
<td>Fasting glucose$^{27}$</td>
<td>4</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

* Residual is after adjustment for age, gender, diabetes.

References:

Additive effects model to quantify cumulative effect of nominal passengers in PCAWG

- Model for the effect of an individual SNP on a phenotype
  \[ y_j = \mu + z_{ij}u_i + e_j \]

- Extension to model the combined effects of multiple SNPs
  \[ y_j = \mu + g_j + e_j \quad \text{and} \quad g_j = \sum_{i=1}^{m} z_{ij}u_i \]
  \[ g_j \sim N(0, \sigma_g^2 = m\sigma_u^2) \quad u \sim N(0, I\sigma_u^2) \]
Using additive effects to compare different categories of variants

Model: \[ y_j = \mu + z_{j}^{\text{drv}}u_1 + \sum_{k \in \{2,3,4\}} z_{ijk}u_{ik} + e_j \]

Parameters: \( (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_E^2) \)

Variant categories:
- \( k = 1 \): coding drivers
- \( k = 2 \): coding other
- \( k = 3 \): promoters
- \( k = 4 \): other non-coding
Overall additive variance increase for multiple cancer cohorts in PCAWG with the inclusion of passengers

Increase in the variance from ~50% using drivers alone to ~59% with putative passengers included, averaged across all cohorts.
Element level additive variance for multiple cancer cohorts in PCAWG, comparing coding & non-coding

In addition to coding mutations, promoter & other non-coding mutations contributed significant amounts of extra variance (~2% & 7%).

[Kumar, Warrell et al. (20, in press) Cell + bioRxiv]
Recasting the additive effects model in a predictive context: Best Linear Unbiased Predictor (BLUP) analysis

Additive variance

Cumulative

SNVs added

All SNPs

95% CI

SNVs, ordered by descending BLUP (\( \hat{u} \)):

BLUP predictor:

\[
\hat{u} = \arg\max_u (P(u|y, \sigma_u^2)) \\
= \arg\max_u (P(y|u)P(u|\sigma_u^2))
\]

Lower bound on # weak drivers (8.4 pan-cancer average; enriched for PCAWG genes w/ FDR<0.25)
Topics in Precision Oncology:
Addressing the role of the non-coding genome in cancer

• **Background**
  • Drivers v passenger
  • Coding v noncoding
  • Pcawg & encode 3

• **Additive-Effects model to measure the Impact of non-coding v coding mutations**
  • Repurposing a formalism from germline genetics for missing heritability to cancer
  • Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  • Notable effect, particularly for non-coding passengers, in addition to known coding drivers.
  • Recasting as a predictive model to est. number of weak drivers

• **Network Rewiring in Cancer**
  • Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  • TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
Regulatory Network Construction

[Zhang et al. ('20), Nat. Comm. + bioRxiv]
Rewired edges in comparison of GM12878 to K562
109 node TF-TF network (approx. CML)
Simplifying Network Rewiring

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

[Image of network rewiring diagrams]
TopicNet: Measuring transcriptional regulatory network change using LDA

[Lou et al. bioxriv + Bioinformatics ('20)]
[Zhang et al. ('20), biorxiv + Nat. Comm. (in press)]

Lectures.gersteinlab.org
Topics in Precision Oncology:
Addressing the role of the non-coding genome in cancer

• **Background**
  • Drivers v passenger
  • Coding v noncoding
  • PcaWG & encode 3

• **Additive-Effects model to measure the Impact of non-coding v coding mutations**
  • Repurposing a formalism from germline genetics for missing heritability to cancer
  • Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  • Notable effect, particularly for non-coding passengers, in addition to known coding drivers.
  • Recasting as a predictive model to est. number of weak drivers

• **Network Rewiring in Cancer**
  • Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  • TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
Topics in Precision Oncology: 
Addressing the role of the non-coding genome in cancer

- **Background**
  - Drivers v passenger
  - Coding v noncoding
  - Pcawg & encode 3

- **Additive-Effects model to measure the Impact of non-coding v coding mutations**
  - Repurposing a formalism from germline genetics for missing heritability to cancer
  - Using it to assess the overall Impact of passengers v drivers, non-coding vs coding, distal vs proximal non-coding
  - Notable effect, particularly for non-coding passengers, in addition to known coding drivers.
  - Recasting as a predictive model to est. number of weak drivers

- **Network Rewiring in Cancer**
  - Large-scale ENCODE chip-seq data highlights TFs changing targets greatly in oncogenesis. (Focus on CML)
  - TopicNet LDA approach (from text-mining) finds regulators that greatly change their gene communities
Acknowledgments

PCAWG.gersteinlab.org
S Kumar, J Warrell, S Li, P McGillivray, W Meyerson, L Salichos, A Harmanci, A Martinez-Fundichely, C Chan, M Nielsen, L Lochovsky, Y Zhang, X Li, S Lou, J Skou Pedersen, C H, G Getz, E Khurana

ENCODEC.gersteinlab.org
J Zhang, D Lee, V Dhiman, P Jiang, J Xu, P McGillivray…. S Liu, K White

github.com/gersteinlab/TopicNet
S Lou, T Li, X Kong, J Zhang, J Liu, D Li
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, 2019.
- 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