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

Mark Gerstein
Yale

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @MarkGerstein).
No Conflicts for this Talk
See last slide for more info.
Personal Genomics as a Gateway into Biology

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

Personal genomes will soon become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.
Keys to genome interpretation
Relating individuals' variants to DBs
**Scaling** DBs to the **population**
Identifying **key variants** - separating into rare, recurrent, common, &c
The **Scaling** of Genomic Data Science:

**Powered by exponential increases in data & computing**

*(Moore’s Law)*

[NHGRI website + Waldrop ('15) Nature]
Exponential Scaling Changes Fields Using Genomic Data

[Muir et al. ('15) GenomeBiol.]
Growth of ICGC datasets

ICGC Data Portal Cumulative Donor Count for Member Projects

Release 22 (August 2016):
- 70 projects
- 19,290 donors total
- 16,236 donors w/ molecular data
The changing costs of a sequencing pipeline

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

[Sboner et al. (’11), Muir et al. (’15) Genome Biology]
The changing costs of a sequencing pipeline

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

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]
The changing costs of a sequencing pipeline

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

Alignment algorithms scaling to keep pace with data generation

[Sboner et al. (’11), Muir et al. (’15) Genome Biology]
The changing costs of a sequencing pipeline

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

Alignment algorithms scaling to keep pace with data generation

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]
The changing costs of a sequencing pipeline

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

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

A Cancer Genome

A Typical Genome

Population of 2,504 peoples

Origin of Variants

<table>
<thead>
<tr>
<th></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>

Class of Variants

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

Prevalence of Variants

Common

Rare* (1-4%)

* 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
Finding Key Variants

Germline

- **Common variants**
  - Can be most readily associated with phenotype (ie disease) via GWAS
  - Usually their functional effect is weaker
  - Many are non-coding
  - Issue of LD in identifying the actual causal variant.

- **Rare variants**
  - Associations are usually underpowered due to low frequencies but often have larger functional impact
  - Can be collapsed in the same element to gain statistical power (burden tests).

Finding Key Variants

Somatic

• **Overall**
  - Often these can be thought of as *very rare variants*.

• **Drivers**
  - Driver mutation is a mutation that directly or indirectly confers a selective growth advantage to the cell in which it occurs.
  - A typical tumor contains 2-8 drivers; the remaining mutations are passengers.

• **Passengers**
  - Conceptually, a passenger mutation has no direct or indirect effect on the selective growth advantage of the cell in which it occurred.
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

• **Introduction**
  • An individual’s disease variants as the public's gateway into genomics & biology
  • **The exponential scaling** of data gen. & processing
  • Big-data mining to prioritize key variants as drivers

• **Functional impact #1: Coding**
  • **ALoFT**: Annotation of Loss-of-Function Transcripts.
  • **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs

• **Functional impact #2: Non-coding**
  • **uORFs**: Feature integration to find small subset of upstream mutations that potentially alter translation.
  • **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)

• **Recurrence:**
  **Statistics for driver identification**
  • **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  • Developed a variety of parametric & non-parametric methods taking this into account
  • **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  • **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs

**Recurrence #2:**
*(Low-power) application to pRCC*
  • WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  • Analysis of signatures & tumor evolution helps identify key mutations in different ways
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

• Introduction
  • An individual's disease variants as the public's gateway into genomics & biology
  • The exponential scaling of data gen. & processing
  • Big-data mining to prioritize key variants as drivers

• Functional impact #1: Coding
  • ALoFT: Annotation of Loss-of-Function Transcripts.
  • Frustration as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs

• Functional impact #2: Non-coding
  • uORFs: Feature integration to find small subset of upstream mutations that potentially alter translation.
  • FunSeq integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)

• Recurrence:
  Statistics for driver identification
  • BMR (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  • Developed a variety of parametric & non-parametric methods taking this into account
  • LARVA uses parametric beta-binomial model, explicitly modeling covariates
  • MOAT does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs

Recurrence #2:
(Low-power) application to pRCC
  • WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  • Analysis of signatures & tumor evolution helps identify key mutations in different ways
Variant Annotation Tool (VAT), developed for 1000G FIG

VCF Input
Output:
  • Annotated VCFs
  • Graphical representations of functional impact on transcripts
Access:
  • Webserver
  • AWS cloud instance
  • Source freely available

vat.gersteinlab.org

Complexities in LOF annotation

Transcript isoforms, distance to stop, functional domains, protein folding, etc.

Balasubramanian S. et al., *Genes Dev.*, ’11
Balasubramanian S.*, Fu Y.* et al., *NComms.*, ’17
**Annotation of Loss-of-Function Transcripts (ALoFT)**

Runs on top of VAT

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

Balasubramanian S.*, Fu Y.* et al., *NComms.*, ’17
LoF distribution varies as expected by mutation set (from healthy people v from disease)

Balasubramanian S.*, Fu Y.* et al., *NComms.*, '17
ALoFT identifies deleterious somatic LoF variants

Cancer genes:
- COSMIC consensus.
- *Enriched in deleterious LoFs.*

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

Balasubramanian S.*, Fu Y.* et al., *NComms.*, ’17
ALoFT refines cancer mutation characterization

Vogelstein et al. '13: if >20% of mutations in gene inactivating → tumor suppressor gene (TSG).
ALoFT further refines 20/20 rule predictions.

Balasubramanian S.*, Fu Y.* et al., NComms., '17
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- Introduction
  - An individual’s disease variants as the public’s gateway into genomics & biology
  - The exponential scaling of data gen. & processing
  - Big-data mining to prioritize key variants as drivers

- Functional impact #1: Coding
  - ALoFT: Annotation of Loss-of-Function Transcripts.
  - Frustration as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs

- Functional impact #2: Non-coding
  - uORFs: Feature integration to find small subset of upstream mutations that potentially alter translation.
  - FunSeq integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)

- Recurrence:
  - Statistics for driver identification
    - BMR (Background mutation rate) significantly varies & is correlated with replication timing & TADs
    - Developed a variety of parametric & non-parametric methods taking this into account
    - LARVA uses parametric beta-binomial model, explicitly modeling covariates
    - MOAT does a variety of non-parametric shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs

- Recurrence #2:
  (Low-power) application to pRCC
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways
What is localized frustration?

[Ferreiro et al., PNAS ('07)]
Workflow for evaluating localized frustration changes ($\Delta F$)

Energies calculated using the wild-type structure

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

WT structure

$\langle E \rangle - E_{\text{TYR}} \sigma_E^{-1} = F_{\text{TYR}} < 0$

$\langle E \rangle - E_{\text{TRP}} \sigma_E^{-1} = F_{\text{TRP}} > 0$

Model of mutated structure

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

$\langle E' \rangle - E'_{\text{TYR}} \sigma'_E^{-1} = F'_{\text{TYR}} < 0$

Energies calculated using the model of the mutated structure

[Refs: Kumar et al. NAR (2016)]
Complexity of the second order frustration calculation

First order frustration calculation (F)

Second order frustration calculation (ΔF)

MD-assisted free energy calculation (ΔG)
Comparing $\Delta F$ values across different SNV categories: disease v normal

Normal mutations (1000G) tend to unfavorably frustrate (less frustrated) surface more than core, but for disease mutations (HGMD) no trend & greater changes

[Kumar et al, NAR (2016)]
Comparison between $\Delta F$ distributions: TSGs v. oncogenes

SNVs in TSGs change frustration more in core than the surface, whereas those associated with oncogenes manifest the opposite pattern. This is consistent with differences in LOF v GOF mechanisms.

[Kumar et al, NAR (2016)]
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- **Introduction**
  - An individual's disease variants as the public's gateway into genomics & biology
  - The exponential scaling of data gen. & processing
  - Big-data mining to prioritize key variants as drivers

- **Functional impact #1: Coding**
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs

- **Functional impact #2: Non-coding**
  - **uORFs**: Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)

- **Recurrence:**
  - Statistics for driver identification
    - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
    - Developed a variety of parametric & non-parametric methods taking this into account
    - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
    - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs

- **Recurrence #2:**
  - (Low-power) application to **pRCC**
    - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
    - Analysis of signatures & tumor evolution helps identify key mutations in different ways
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.

[A] CDS [overlap] uORF

5' --- | --- | --- 3'

5' --- | --- | --- 3'

5' --- | --- | --- 3'

[B] uORF translation CDS

5' -- ATG ATG -- 3'

5' -- TTG ATG -- 3'

phenotype disease risk cancer

[McGillivray et al., NAR ('18)]
The population of functional uORFs may be significant

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

[McGillivray et al., NAR ('18)]
Prediction & validation of functional uORFs using 89 features

- All near-cognate start codons predicted.
- Cross-validation on independent ribosome profiling datasets and validation using in vivo protein levels and ribosome occupancy in humans (Battle et al. 2014).

[McGillivray et al., NAR ('18)]
A comprehensive catalog of functional uORFs

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

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

- **180K**: Large predicted positive set likely to affect translation

- Calibration on gold standards, suggests getting ~70% of known

[McGillivray et al., *NAR* ('18)]
Somatic alteration of uORFs disproportionately affects certain cancers and molecular pathways

- uORF gain and loss occurs in cancer (incl. in cancer associated genes, e.g., MYC, BCL2, etc.).
- Alteration of translation may contribute to cancer.
- These changes are concentrated in certain cancers and pathways.
- Mutations leading to uORFs differ in somatic vs. germline.

[McGillivray et al., NAR ('18)]
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- **Introduction**
  - An individual’s disease variants as the public’s gateway into genomics & biology
  - The *exponential scaling* of data gen. & processing
  - Big-data mining to prioritize key variants as drivers

- **Functional impact #1: Coding**
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs

- **Functional impact #2: Non-coding**
  - **uORFs**: Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)

- **Recurrence**
  **Statistics for driver identification**
  - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  - Developed a variety of parametric & non-parametric methods taking this into account
  - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  - **MOAT** does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs

- **Recurrence #2**
  **(Low-power) application to pRCC**
  - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  - Analysis of signatures & tumor evolution helps identify key mutations in different ways
Coding and non-coding elements may synergistically contribute to cancer

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

Broad categories of regulatory regions under negative selection Related to:

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)]
Defining Sensitive non-coding Regions

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

~0.4% genomic coverage (~ top 25)
~0.02% genomic coverage (top 5)

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

[Khurana et al., Science ('13)]
SNPs which break TF motifs are under stronger selection

[Khurana et al., Science (‘13)]
FunSeq.gersteinlab.org

- 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

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

[Fu et al., GenomeBiology ('14)]
3 controls with natural polymorphisms (allele frequency $\geq 1\%$)

1. Matched region: 1kb around HGMD variants
2. Matched TSS: matched for distance to TSS
3. Unmatched: randomly selected

Ritchie et al., Nature Methods, 2014

[Fu et al., GenomeBiology ('14, in revision)]
Flowchart for 1 Prostate Cancer Genome (from Berger et al. '11)
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

- **Introduction**
  - An individual’s disease variants as the public’s gateway into genomics & biology
  - The exponential scaling of data gen. & processing
  - Big-data mining to prioritize key variants as drivers

- **Functional impact #1: Coding**
  - ALoFT: Annotation of Loss-of-Function Transcripts.
  - Frustration as a localized metric of SNV impact.
    Differential profiles for oncogenes v. TSGs

- **Functional impact #2: Non-coding**
  - uORFs: Feature integration to find small subset of upstream mutations that potentially alter translation.
  - FunSeq integrates evidence, with a “surprisal” based weighting scheme.
    Prioritizing rare variants with “sensitive sites” (human conserved)

- **Recurrence:**
  - Statistics for driver identification
    - BMR (Background mutation rate) significantly varies & is correlated with replication timing & TADs
    - Developed a variety of parametric & non-parametric methods taking this into account
    - LARVA uses parametric beta-binomial model, explicitly modeling covariates
    - MOAT does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs

- **Recurrence #2:**
  - (Low-power) application to pRCC
    - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
    - Analysis of signatures & tumor evolution helps identify key mutations in different ways
Cancer Type 1

Cancer Type 2

Cancer Type 3

Mutation recurrence
Mutation recurrence

Cancer Type 1

Cancer Type 2

Cancer Type 3

Early replicated regions

Late replicated regions
Late replicated regions

Noncoding annotations

Cancer Type 1

Cancer Type 2

Cancer Type 3

Early replicated regions

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

Lochovsky et al. NAR ('15)
Variation in somatic mutations is closely associated with chromatin structure (TADs) & replication timing

Chromatin remodeling failure leads to more mutations in early-replicating regions

[Yan et al., PLOS Comp. Bio. ('17); S. Li et al., PLOS Genetics ('17)]
mrTADFinder:
Identifying TADs at multiple resolutions by maximizing modularity vs appropriate null

\[
Q = \frac{1}{2N} \sum_{i,j} (W_{ij} - \gamma E_{ij}) \delta_{\sigma_i,\sigma_j}
\]

\(\gamma\): resolution parameter

Choose a particular resolution \(\gamma\)
Optimize \(Q\) over all possible partitions

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

consensus boundaries based on the boundary scores

consensus TADs

output

[Yan et al., PLOS Comp. Bio. ('17)]
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

• Introduction
  • An individual’s disease variants as the public’s gateway into genomics & biology
  • The exponential scaling of data gen. & processing
  • Big-data mining to prioritize key variants as drivers
• Functional impact #1: Coding
  • ALoFT: Annotation of Loss-of-Function Transcripts.
  • Frustration as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs
• Functional impact #2: Non-coding
  • uORFs: Feature integration to find small subset of upstream mutations that potentially alter translation.
  • FunSeq integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)
• Recurrence:
  Statistics for driver identification
  • BMR (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  • Developed a variety of parametric & non-parametric methods taking this into account
  • LARVA uses parametric beta-binomial model, explicitly modeling covariates
  • MOAT does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs
• Recurrence #2:
  (Low-power) application to pRCC
  • WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  • Analysis of signatures & tumor evolution helps identify key mutations in different ways
Cancer Somatic Mutation Modeling

**PARAMETRIC MODELS**

*Model 1: Constant Background Mutation Rate (Model from Previous Work)*

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

*Model 2a: Varying Mutation Rate with Single Covariate Correction*

\[ x_i : \text{Binomial}(n_i, p_i) \]
\[ p_i : \text{Beta}(\mu | R_i, \sigma | R_i) \]
\[ \mu | R_i, \sigma | R_i : \text{constant within the same covariate rank} \]

*Model 2b: Varying Mutation Rate with Multiple Covariate Correction*

\[ x_i : \text{Binomial}(n_i, p_i) \]
\[ p_i : \text{Beta}(\mu | R_i, \sigma | R_i) \]
\[ \mu | R_i, \sigma | R_i : \text{constant within the same covariate rank} \]

**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

\[ \text{d}_\text{max} \]

\[ d_\text{min} \quad d_\text{min} \]

\[ \text{d}_\text{max} \]

[Lochovsky et al. Bioinformatics in press]
MOAT-v: Variant-based Permutation

Can preserve tri-nt context in shuffle

bin width $W$

$W \approx 2 \times d_{max}$

annotation

= original variants

= permuted variants

[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

```
Binning whole genome
```

```
Marking equivalence classes (bins with similar covariate vectors)
```

```
Overlaying variants (with tri-nucleotide indexing)
```

```
Shuffling variants
```

= original variants
= permuted variants

[Lochovsky et al. Bioinformatics in press]
LARVA Model Comparison

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

[Lochovsky et al. NAR ('15)]
LARVA Results

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

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>TMRSS13</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>
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

• **Introduction**
  - An individual's disease variants as the public's gateway into genomics & biology
  - *The exponential scaling* of data gen. & processing
  - Big-data mining to prioritize key variants as drivers

• **Functional impact #1: Coding**
  - **ALoFT**: Annotation of Loss-of-Function Transcripts.
  - **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs

• **Functional impact #2: Non-coding**
  - **uORFs**: Feature integration to find small subset of upstream mutations that potentially alter translation.
  - **FunSeq** integrates evidence, with a "surprisal" based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)

• **Recurrence:**
  - **Statistics for driver identification**
    - **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
    - Developed a variety of parametric & non-parametric methods taking this into account
    - **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
    - **MOAT** does a variety of non-parametric shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs

• **Recurrence #2:**
  - (Low-power) application to **pRCC**
    - WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
    - Analysis of signatures & tumor evolution helps identify key mutations in different ways
Power, as an issue in driver discovery

Better annotation or large number of samples could help.

[Kumar & Gerstein, Nature (’17)]
An (underpowered) case study: pRCC

- Kidney cancer lifetime risk of 1.6% & the papillary type (pRCC) counts for ~10% of all cases
- TCGA project sequenced 161 pRCC exomes & classified them into subtypes
  - Yet, cannot pin down the cause for a significant portion of cases....
- 35 WGS of TN pairs, perhaps useful? But not that definitive from a recurrence perspective

[Cancer Genome Atlas Research Network N Engl J Med. ('16)]
MET is long known pRCC driver
In MET, TCGA found somatic SNVs, duplications & an alt. splicing event as drivers (43/161).
In addition, from 35 WGS we found
- A noncoding hotspot associated with MET
- Lack of SVs & breakpoints disrupting MET
- Germline SNP (rs11762213) predicts survival in type 2 patients

[A. Gentile, L. Trusolino and PM. Comoglio, Cancer and Metastasis Reviews ('08); S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]
Beyond \textit{MET}: 2 non-coding hotspots in NEAT & ERRFI1, supported by expr. changes & survival analysis
Tumor Evolution: Highlight the Ordering of Key Mutations

Normal cell → MRCA cell → Distant metastasis

Time point X: diagnosis and treatment initiation
Time point Y: distant and local relapse

Driver mutations

Yates et al, NRG (2012)
Construct evolutionary trees in pRCC

• Infer mutation order and tree structure based on mutation abundance (PhyloWGS, Deshwar et al., 2015)
• Some of the key mutations occur in all the clones while others are just in some parts of the tree

DNMT3A: premature stop
NEAT1: noncoding
SMARCA4: missense

KDM6A: missense

MET: noncoding
ERRFI1: noncoding

[S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]
Mutation distance: 0.5

Germline

Populations (\%)

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

Populations (\%)

0.5

[S. Li, B. Shuch and M. Gerstein PLOS Genetics (’17)]
# Tree topology correlates with molecular subtypes

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

## Table of Histological type/Patient ID and Molecular Subtypes

| Histological type/Patient ID | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
|-----------------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| COCA                        |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Copy number gain            |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Somatic mutation            |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Splicing event              |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Germline mutation           |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| BAP1/PRM1/SETD2 mut.        |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| CDKN2A copy number loss     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| SDHB deletion               |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| MET                         |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Promoter mutation           |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 1-2 intronic mutation       |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| NEAT1 somatic mutation      |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| ERRFI1 promoter mutation    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Mutation Processes          |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Whole genome mutation rate  |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| DHS mutation percentage     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| SV number                   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Evolution tree topology     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

## Diagrams

- **No branch, less subclone**
- **Short branches**
- **No branch, more subclones**
- **Long branches**

**COCA**
- Affected
- Unaffected
- NA

**Mutation rate/percentage/SV number**
- High
- Medium
- Low

[Image of tree diagrams with color-coded branches and labels for COCA, CI, C2a, C2b, mutation rate/percentage/SV number, and evolutionary tree topology.]

[Image of color-coded grid with rows for histological type/Patient ID and columns for molecular subtypes, indicating affected or unaffected states with grey or white squares respectively, and NA for missing data.]
Mutational processes carry context-specific signatures

\[ M = S \times W + \varepsilon \]

\( S \): Mutation signature inferred

\( M \): Mutation spectrum observed

CpGs drive inter-patient variation in pRCC mutational spectra

- The loadings on PC1 are mostly [C>T]G
- Confirmed by higher C>T% in CpGs in the hypermethylated group (cluster1)

[S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]
Chromatin remodeling defect ("mut") leads to more mutations in open chromatin (raw number & fraction) in those pRCC cases w/ the mutation.

Key mutation affects mutational landscape which, in turn, affects overall burden in pRCC.

[S. Li, B. Shuch and M. Gerstein PLOS Genetics ('17)]
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

• Introduction
  • An individual’s disease variants as the public’s gateway into genomics & biology
  • The exponential scaling of data gen. & processing
  • Big-data mining to prioritize key variants as drivers

• Functional impact #1: Coding
  • **ALoFT**: Annotation of Loss-of-Function Transcripts.
  • **Frustration** as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs

• Functional impact #2: Non-coding
  • **uORFs**: Feature integration to find small subset of upstream mutations that potentially alter translation.
  • **FunSeq** integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)

• Recurrence:
  Statistics for driver identification
  • **BMR** (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  • Developed a variety of parametric & non-parametric methods taking this into account
  • **LARVA** uses parametric beta-binomial model, explicitly modeling covariates
  • **MOAT** does a variety of non-parametric shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs

Recurrence #2:
(Low-power) application to **pRCC**
  • WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  • Analysis of signatures & tumor evolution helps identify key mutations in different ways
Prioritizing Variants in Personal Genomes: Using functional impact & recurrence, with particular application to cancer

• Introduction
  • An individual’s disease variants as the public’s gateway into genomics & biology
  • The exponential scaling of data gen. & processing
  • Big-data mining to prioritize key variants as drivers

• Functional impact #1: Coding
  • ALoFT: Annotation of Loss-of-Function Transcripts.
  • Frustration as a localized metric of SNV impact. Differential profiles for oncogenes v. TSGs

• Functional impact #2: Non-coding
  • uORFs: Feature integration to find small subset of upstream mutations that potentially alter translation.
  • FunSeq integrates evidence, with a “surprisal” based weighting scheme. Prioritizing rare variants with “sensitive sites” (human conserved)

• Recurrence:
  Statistics for driver identification
  • BMR (Background mutation rate) significantly varies & is correlated with replication timing & TADs
  • Developed a variety of parametric & non-parametric methods taking this into account
  • LARVA uses parametric beta-binomial model, explicitly modeling covariates
  • MOAT does a variety of non-parm. shuffles (annotation, variants, &c). Useful when explicit covariates not available. Slower but speeded up w/ GPUs

Recurrence #2: (Low-power) application to pRCC
  • WGS finds additional facts on the canonical driver, MET. Other suggestive non-coding hotspots.
  • Analysis of signatures & tumor evolution helps identify key mutations in different ways
github.com/gersteinlab/Frustration
S Kumar, D Clarke

github.com/gersteinlab/MrTADfinder
KK Yan, S Lou

VAT.gersteinlab.org
L Habegger, S Balasubramanian, DZ Chen, E Khurana, A Sboner, A Harmanci, J Rozowsky, D Clarke, M Snyder

ALoFT.gersteinlab.org
S Balasubramanian, Y Fu, M Pwashe, P McGillivray, M Jin, J Liu, K Karczewski, D MacArthur

FunSeq.gersteinlab.org
Y Fu, E Khurana, Z Liu, S Lou, J Bedford, X Mu, K Yip

pRCC - S Li, B Shuch

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

LARVA.gersteinlab.org
L Lochofsky, J Zhang, Y Fu, E Khurana

github.gersteinlab.org/uORFs
P McGillivray, R Ault, M Pwashe, R Kitchen, S Balasubramanian
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/kwotppt