Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

Mark Gerstein, Yale
Slides freely downloadable from Lectures.GersteinLab.org
& “tweetable” (via @markgerstein). See last slide for more info.
Moore’s Law: Exponential Scaling of Computer Technology

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

[Waldrop (‘15) Nature]
Kryder’s Law and S-curves underlying exponential growth

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

[Muir et al. (‘15) GenomeBiol.]
Sequencing Data Explosion: Faster than Moore’s Law for a Time (or a S-curve)

- DNA sequencing has gone through technological S-curves
  - In the early 2000’s, improvements in Sanger sequencing produced a scaling pattern similar to Moore’s law.
  - The advent of NGS was a shift to a new technology with dramatic decrease in cost).
Sequencing cost reductions have resulted in an explosion of data

- The type of sequence data deposited has changed as well.
  - Protected data represents an increasing fraction of all submitted sequences.
  - Data from techniques utilizing NGS machines has replaced that generated via microarray.

[Muir et al. (‘15) GenomeBiol.]
**Sequence Universe**

TCGA endpoint: ~2.5 Petabytes

~1.5 PB exome

~1 PB whole genome

SRA ~1 petabyte

TCGA 2.3 Petabytes in CGHub

222 TB

68 TB

40 TB

32 TB

29 TB

GTeX

ARRA Autism

NHLBI ESP

Star formation

100K Genomes England

Heidi Sofia, 7-16-15
Increasing diversity in sequence data sources

[Muir et al. ('15) GenomeBiol.]
The changing costs of a sequencing pipeline

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

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

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

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

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

Alignment algorithms scaling to keep pace with data generation

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

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

Alignment algorithms scaling to keep pace with data generation

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

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

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]
Personal Genomics:
Managing Rapid Data Scaling through Prioritizing High-impact Variants

• Introduction
  – The exponential scaling of data generation & processing
  – The landscape of variants in personal genomes suggests finding a few key ones

• Characterizing Rare Variants in Coding Regions
  – Identifying with STRESS cryptic allosteric sites
    • On surface & in interior bottlenecks
  – Using changes in localized Frustration to find sites sensitive to mutations
    • Difference betw. TSGs & oncogenes

• Evaluating the Impact of Non-coding Variants with Annotation
  – Annotating non-coding regions
  – Prioritizing rare variants with “sensitive sites” (human-conserved)
  – Prioritizing in terms of network connectivity (eg hubs)

• Putting it together in Workflows
  – Using LARVA to do burden testing on non-coding annotation
    • Need to correct for over-dispersion mutation counts
    • Parameterized according to replication timing
  – Using FunSeq to integrate evidence on variants
    • Systematically weighting all the features
    • suggesting non-coding drivers
    • Prioritizing rare germline variants
Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones

- Characterizing Rare Variants in Coding Regions
  - Identifying with **STRESS** cryptic allosteric sites
    - On surface & in interior bottlenecks
  - Using changes in localized **Frustration** to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes

- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)

- Putting it together in Workflows
  - Using **LARVA** to do burden testing on non-coding annotation
    - Need to correct for over-dispersion mutation counts
    - Parameterized according to replication timing
  - Using **FunSeq** to integrate evidence on variants
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
Human Genetic Variation

A Cancer Genome

A Typical Genome

Population of 2,504 peoples

<table>
<thead>
<tr>
<th>Origin of Variants</th>
<th>Class of Variants</th>
<th>Prevalence of Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germ-line</td>
<td>SNP</td>
<td></td>
</tr>
<tr>
<td>22K</td>
<td>3.5 – 4.3M</td>
<td></td>
</tr>
<tr>
<td>Non-coding</td>
<td>Indel</td>
<td>550 – 625K</td>
</tr>
<tr>
<td>4.1 – 5M</td>
<td>SV</td>
<td>2.1 – 2.5K (20Mb)</td>
</tr>
<tr>
<td>Somatic</td>
<td>Total</td>
<td>4.1 – 5M</td>
</tr>
<tr>
<td>~50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5K</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Passenger

Driver (~0.1%)

Common

Rare* (1-4%)

Rare (~75%)

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

References:
- The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Finding Key Variants

Germline

• **Common variants**
  • Can be associated with phenotype (ie disease) via a Genome-wide Association Study (GWAS), which tests whether the frequency of alleles differs between cases & controls.
  • 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.
  • They often have larger functional impact
  • Can be collapsed in the same element to gain statistical power (burden tests).
  • In some cases, causal variants can be identified through tracing inheritance of Mendelian subtypes of diseases in large families.

Finding Key Variants

Somatic

• Overall
  • Often these can be conceptualized as very rare variants
  • A challenge to identify somatic mutations contributing to cancer is to find driver mutations & distinguish them from passengers.

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

• Passengers
  • Conceptually, a passenger mutation has no direct or indirect effect on the selective growth advantage of the cell in which it occurred.

Vogelstein B. Science 2013. 339(6127):1546-1558
Association of Variants with Diseases

Healthy

Pooled Variants

Diseased

Burden Test

GWAS Positive

Common Variants

Rare or Somatic Variants

High Function Impact

Healthy

Pooled Variants

Diseased

Burden Test

GWAS Positive

Common Variants

Rare or Somatic Variants

High Function Impact
Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

• Introduction
  – The exponential scaling of data generation & processing
  – The landscape of variants in personal genomes suggests finding a few key ones

• Characterizing Rare Variants in Coding Regions
  – Identifying with **STRESS** cryptic allostERIC sites
    • On surface & in interior bottlenecks
  – Using changes in localized **Frustration** to find sites sensitive to mutations
    • Difference betw. TSGs & oncogenes

• Evaluating the Impact of Non-coding Variants with Annotation
  – Annotating non-coding regions
  – Prioritizing rare variants with “sensitive sites” (human-conserved)
  – Prioritizing in terms of network connectivity (eg hubs)

• Putting it together in Workflows
  – Using **LARVA** to do burden testing on non-coding annotation
    • Need to correct for over-dispersion mutation counts
    • Parameterized according to replication timing
  – Using **FunSeq** to integrate evidence on variants
    • Systematically weighting all the features
    • suggesting non-coding drivers
    • Prioritizing rare germline variants
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited.

Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated.

*Fibroblast growth factor receptor 2 (pdb: 1IIIL)*

- 1000G & ExAC SNVs (common | rare)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annotation overlap)
- HGMD site (w/annotation overlap)

[Sethi et al. COSB ('15)]
Models of Protein Conformational Change

Motion Vectors from Normal Modes (ANMs)

Characterizing uncharacterized variants
<= Finding Allosteric sites
<= Modeling motion
Predicting Allosterically-Important Residues at the Surface

1. MC simulations generate a large number of candidate sites
2. Score each candidate site by the degree to which it perturbs large-scale motions
3. Prioritize & threshold the list to identify the set of high confidence sites

\[
\text{binding leverage} = \sum_{m=1}^{10} \left( \sum_{i} \sum_{j} \Delta d_{ij(m)}^2 \right)
\]

Adapted from Clarke*, Sethi*, et al (in press)
Predicting Allosterically-Important Residues at the Surface

Adapted from Clarke*, Sethi*, et al (in press)
Predicting Allosterically-Important Residues within the Interior

Adapted from Clarke*, Sethi*, et al (in press)
Predicting Allosterically-Important Residues within the Interior

\[ \text{Cov}_{ij} = \langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle \]
\[ C_{ij} = \text{Cov}_{ij} / \sqrt{\langle \mathbf{r}_i^2 \rangle \langle \mathbf{r}_j^2 \rangle} \]
\[ D_{ij} = -\log(|C_{ij}|) \]

Adapted from Clarke*, Sethi*, et al. (in press)
Predicting Allosterically-Important Residues within the Interior

Adapted from Clarke*, Sethi*, et al (in press)
STRESS Server Architecture: Highlights
stress.molmovdb.org

- A light front-end server handles incoming requests, and powerful back-end servers perform calculations.
- Auto Scaling adjusts the number of back-end servers as needed.
- A typical structure takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.
- Input & output (i.e., predicted allosteric residues) are stored in S3 buckets.

Adapted from Clarke*, Sethi*, et al (in press)
Intra-species conservation of predicted allosteric residues

1000 Genomes

Surface

Interior

Adapted from Clarke*, Sethi*, et al (in press)
Intra-species conservation of predicted allosteric residues

ExAC

Surface

Interior

Adapted from Clarke*, Sethi*, et al (in press)
Unlike common SNVs, the statistical power with which we can evaluate rare SNVs in case-control studies is severely limited.

Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated.

Fibroblast growth factor receptor 2 (pdb: 1IIL)

[Sethi et al. COSB (’15)]
Protein structures may provide the needed alternative for evaluating rare SNVs, many of which may be disease-associated.

Rationalizing disease variants in the context of allosteric behavior with allosterity as an added annotation.

Fibroblast growth factor receptor 2 (pdb: 1IIL)

[Sethi et al. COSB ('15)]
Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

• Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones

• Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    • On surface & in interior bottlenecks
  - Using changes in localized Frustration to find sites sensitive to mutations
    • Difference between TSGs & oncogenes

• Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (e.g., hubs)

• Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    • Need to correct for over-dispersion mutation counts
    • Parameterized according to replication timing
  - Using FunSeq to integrate evidence on variants
    • Systematically weighting all the features
    • Suggesting non-coding drivers
    • Prioritizing rare germline variants
Schematic illustration of localized frustration

[Image: Diagram showing the comparison between ASN (NH$_3^+$) and ASP (carboxylate) with favorable and unfavorable interactions]

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

**Native Structure**

\[
\langle E \rangle - \frac{E_{\text{nat}}}{\sigma_E} = F_{\text{nat}} > 0
\]

**Mutated Structure**

\[
\langle E' \rangle - \frac{E_{\text{mut}}}{\sigma_E} = F_{\text{mut}} < 0
\]

\[
F_{\text{mut}} - F_{\text{nat}} = \Delta F < 0
\]

[Kumar et al. NAR (in press); bioRxiv 052027]
Comparing Frustration ($\Delta F$ values) across different SNV categories

1. **1000 Genomes**
   - Core: Red
   - Surface: Purple

2. **ExAC**
   - Core: Green
   - Surface: Olive

3. **HGMD**
   - Core: Blue
   - Surface: Light Blue

[Image: Box plots showing $\Delta F$ values for different SNV categories across three datasets: 1000 Genomes, ExAC, and HGMD. The plots compare Core and Surface regions for each dataset.]
$\Delta F$ distributions among rare and common SNVs

A

1000 Genomes

B

ExAC

[Kumar et al. NAR (in press); bioRxiv 052027]
Comparison between $\Delta F$ distributions: TSGs vs. oncogenes

[Kumar et al. NAR (in press); bioRxiv 052027]
Personal Genomics:
Managing Rapid Data Scaling through Prioritizing High-impact Variants

• Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones

• Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    • On surface & in interior bottlenecks
  - Using changes in localized Frustration to find sites sensitive to mutations
    • Difference betw. TSGs & oncogenes

• Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)

• Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    • Need to correct for over-dispersion mutation counts
    • Parameterized according to replication timing
  - Using FunSeq to integrate evidence on variants
    • Systematically weighting all the features
    • suggesting non-coding drivers
    • Prioritizing rare germline variants
Non-coding Annotations: Overview

Sequence features, incl. **Conservation**

- Large-scale sequence similarity comparison
- Identify large blocks of repeated and deleted sequence:
  - Within the human reference genome
  - Within the human population
  - Between closely related mammalian genomes
- Identify smaller-scale repeated blocks using statistical models

**Functional Genomics**

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

Signal processing of raw experimental data:
- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:
- Binding sites
- Transcriptionally active regions

Group active regions into larger annotation blocks

[Alexander et al., *Nat. Rev. Genet.* (’10)]
Summarizing the Signal: "Traditional" ChipSeq Peak Calling

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

Now an update: "PeakSeq 2" => MUSIC

[Rozowsky et al. ('09) Nat Biotech]
Finding "Conserved" Sites in the Human Population:
Negative selection in non-coding elements based on
Production ENCODE & 1000G Phase 1

Broad Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>0.68</td>
</tr>
<tr>
<td>Genomic Avg</td>
<td>0.64</td>
</tr>
<tr>
<td>Enhancer</td>
<td>0.61</td>
</tr>
<tr>
<td>ncRNA</td>
<td>0.60</td>
</tr>
<tr>
<td>DHS</td>
<td>0.59</td>
</tr>
<tr>
<td>TFSS</td>
<td>0.56</td>
</tr>
<tr>
<td>General</td>
<td>0.58</td>
</tr>
<tr>
<td>Chromatin</td>
<td>0.62</td>
</tr>
<tr>
<td>Pseudogene</td>
<td>0.57</td>
</tr>
</tbody>
</table>

• Broad categories of regulatory regions under negative selection

• Related to:
  
  Mu et al, *NAR*, 2011

[Khurana et al., *Science* ('13)]
Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

[Differential selective constraints among specific sub-categories]

[Khurana et al., Science ('13)]
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
Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

- **Introduction**
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones

- **Characterizing Rare Variants in Coding Regions**
  - Identifying with **STRESS**
    - Cryptic allosteric sites
      - On surface & in interior bottlenecks
  - Using changes in localized **Frustration** to find sites sensitive to mutations
    - Difference between TSGs & oncogenes

- **Evaluating the Impact of Non-coding Variants with Annotation**
  - Annotating non-coding regions
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (e.g., hubs)

- **Putting it together in Workflows**
  - Using **LARVA** to do burden testing on non-coding annotation
    - Need to correct for over-dispersion mutation counts
    - Parameterized according to replication timing
  - Using **FunSeq** to integrate evidence on variants
    - Systematically weighting all the features
    - Suggesting non-coding drivers
    - Prioritizing rare germline variants
Relating Non-coding Annotation to Protein-coding Genes via Networks

Regulatory elements

Assigning proximal sites (< 1Kb) to target genes

Assigning distal sites (10Kb-1Mb) to targets

Distal signals

Expression levels

Cell lines

GM12878
H1-hESC
HeLa-S3
Hep-G2
K562

Methylation H3K27ac

Gene 1  Gene 2  Gene 3

Scale

Strong

Weak

Other strategies to create linkage incl. eQTL and Hi-C. Much in recent Epigenomics Roadmap.

Connecting Distal Elements via **Activity Correlations**.
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
Regulatory Hubs are more Essential

[Khurana et al., PLOS Comp. Bio. ’13]
Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

- Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones

- Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes

- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)

- Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for over-dispersion mutation counts
    - Parameterized according to replication timing
  - Using FunSeq to integrate evidence on variants
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
Mutation recurrence

Cancer Type 1

Cancer Type 2

Cancer Type 3
Mutation recurrence

Cancer Type 1
- Early replicated regions
- Late replicated regions

Cancer Type 2
- Mutation recurrence

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
Late replicated regions

Cancer Type 1

Cancer Type 2

Cancer Type 3

Noncoding annotations

Early replicated regions

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

[Lochovsky et al. NAR ('15)]
Cancer Somatic Mutation Modeling

- 3 models to evaluate the significance of mutation burden
- 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
  - $R$: the replication timing bin of the element

<table>
<thead>
<tr>
<th>Model 1: Constant Background Mutation Rate (Model from Previous Work)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$: Binomial($n_i, p$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 2: Varying Mutation Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i</td>
</tr>
<tr>
<td>$p_i$: Beta($\mu, \sigma$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 3: Varying Mutation Rate with Replication Timing Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i</td>
</tr>
<tr>
<td>$p_i$: Beta($\mu</td>
</tr>
<tr>
<td>$\mu</td>
</tr>
</tbody>
</table>

[Lochovsky et al. NAR '15]
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)]
Adding DNA replication timing correction further improves the beta-binomial model.
LARVA Implementation

- http://larva.gersteinlab.org/
- Freely downloadable C++ program
  - Verified compilation and correct execution on Linux
- A Docker image is also available to download
  - Runs on any operating system supported by Docker
- Running time on transcription factor binding sites (a worst case input size) is ~80 min
  - Running time scales linearly with the number of annotations in the input
LARVA Results

TSS LARVA results

These have literature-verified cancer associations

noncoding annotation
p-values in sorted order

[Lochovsky et al. NAR ('15)]
Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

• Introduction
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones

• Characterizing Rare Variants in Coding Regions
  - Identifying with STRESS cryptic allosteric sites
    • On surface & in interior bottlenecks
  - Using changes in localized Frustration to find sites sensitive to mutations
    • Difference between TSGs & oncogenes

• Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (e.g., hubs)

• Putting it together in Workflows
  - Using LARVA to do burden testing on non-coding annotation
    • Need to correct for over-dispersion mutation counts
    • Parameterized according to replication timing
  - Using FunSeq to integrate evidence on variants
    • Systematically weighting all the features
    • Suggesting non-coding drivers
    • Prioritizing rare germline variants
Identification of non-coding candidate drivers amongst somatic variants: Scheme

[Khurana et al., Science ('13)]
Flowchart for 1 Prostate Cancer Genome

(from Berger et al. '11)
FunSeq2 - A flexible framework to prioritize regulatory mutations from cancer genome sequencing

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

Instructions
- Input File - BED or VCF formatted. Click ‘green’ button to add multiple files. With multiple files, the tool will do recurrent analysis. (Note: for BED format, user can put variants from multiple genomes in one file, see Sample input file.)
- Recurrence DB - User can choose particular cancer type from the database. The DB will continue to be updated with newly available WGS data.
- Gene List - Option to analyze variants associated with particular set of genes. Note: Please use Gene Symbols, one row per gene.
- Differential Gene Expression Analysis - Option to detect differentially expressed genes in RNA-Seq data. Two files needed: expression file & class label file. Please refer to Expression input files for instructions to prepare those files.

Site integrates user variants with large-scale context

Data Context

Variant Prioritization
Weighted scoring scheme
Highlighting variants

Variant Reports

FunSeq.gersteinlab.org

[Fu et al., GenomeBiology ('14)]
- Feature weight
  - Weighted with mutation patterns in natural polymorphisms
    (features frequently observed weight less)
  - entropy based method

[Fu et al., GenomeBiology ('14)]
- Feature weight
  - Weighted with mutation patterns in natural polymorphisms (features frequently observed weight less)
  - entropy based method

\[ p = \frac{3}{20} \]

[Fu et al., GenomeBiology ('14)]
- Feature weight
  - Weighted with mutation patterns in natural polymorphisms (features frequently observed weight less)
  - entropy based method

\[ p = \frac{3}{20} \]

Feature weight: \( w_d = 1 + p_d \log_2 p_d + (1 - p_d) \log_2 (1 - p_d) \)

\( p \uparrow \quad w_d \downarrow \quad p = \text{probability of the feature overlapping natural polymorphisms} \)

For a variant: \( \text{Score} = \sum w_d \quad \text{of observed features} \)

[Fu et al., GenomeBiology ('14)]
Germline pathogenic variants show higher core scores than controls

3 controls with natural polymorphisms (allele frequency >= 1%)
1. Matched region: 1kb around HGMD variants
2. Matched TSS: matched for distance to TSS
3. Unmatched: randomly selected

Ritchie et al., Nature Methods, 2014

[Fu et al., GenomeBiology ('14, in revision)]
Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

- **Introduction**
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones

- **Characterizing Rare Variants in Coding Regions**
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
  - Using changes in localized **Frustration** to find sites sensitive to mutations
    - Difference between TSGs & oncogenes

- **Evaluating the Impact of Non-coding Variants with Annotation**
  - Annotating non-coding regions
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (e.g., hubs)

- **Putting it together in Workflows**
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for over-dispersion mutation counts
    - Parameterized according to replication timing
  - Using FunSeq to integrate evidence on variants
    - Systematically weighting all the features
    - Suggesting non-coding drivers
    - Prioritizing rare germline variants
Personal Genomics: Managing Rapid Data Scaling through Prioritizing High-impact Variants

- **Introduction**
  - The exponential scaling of data generation & processing
  - The landscape of variants in personal genomes suggests finding a few key ones

- **Characterizing Rare Variants in Coding Regions**
  - Identifying with STRESS cryptic allosteric sites
    - On surface & in interior bottlenecks
  - Using changes in localized Frustration to find sites sensitive to mutations
    - Difference betw. TSGs & oncogenes

- **Evaluating the Impact of Non-coding Variants with Annotation**
  - Annotating non-coding regions
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)

- **Putting it together in Workflows**
  - Using LARVA to do burden testing on non-coding annotation
    - Need to correct for over-dispersion mutation counts
    - Parameterized according to replication timing
  - Using FunSeq to integrate evidence on variants
    - Systematically weighting all the features
    - suggesting non-coding drivers
    - Prioritizing rare germline variants
github.com/gersteinlab/

**Frustration**

S Kumar, D Clarke

**CostSeq2**

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

**FunSeq.gersteinlab.org**

- & -

**FunSeq2.gersteinlab.org**

Y Fu, E Khurana, Z Liu, S Lou, J Bedford, XJ Mu, KY Yip, V Colonna, XJ Mu, …

1000 Genomes Project Consortium, et al

LARVA.gersteinlab.org

L Lochovsky, J Zhang, Y Fu, E Khurana

STRESS.molmovdb.org

D Clarke, A Sethi, S Li, S Kumar, R W.F. Chang, J Chen

Acknowledgments

Hiring Postdocs. See gersteinlab.org/jobs
Info about content in this slide pack

• General PERMISSIONS
  – This Presentation is copyright Mark Gerstein, Yale University, 2016.
  – Please read permissions statement at www.gersteinlab.org/misc/permissions.html .
  – Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  – Paper references in the talk were mostly from Papers.GersteinLab.org.

• PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  – In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt