Personal Genomics: Managing Exponential 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

1000 Genomes
A Deep Catalog of Human Genetic Variation

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

[Figure: Graph showing the number of bases sequenced over time, categorized by publication and species, with specific years (2009-2015) and species sequenced by year.]

[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]
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]
Personal Genomics: Managing Exponential 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

• Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
  - Prioritizing using AlleleDB in terms of allelic elements
    • Having observed difference in molecular activity in many contexts

• 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

• Postscript: Analysis of the Evolution of a Consortium
  - Differences in “modularity” for members & users
  - Key role for brokers
Personal Genomics: Managing Exponential 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

• Evaluating the Impact of Non-coding Variants with Annotation
  – Annotating non-coding regions on different scales with MUSIC
  – Prioritizing rare variants with “sensitive sites” (human-conserved)
  – Prioritizing in terms of network connectivity (e.g., hubs)
  – Prioritizing using AlleleDB in terms of allelic elements
    • Having observed difference in molecular activity in many contexts

• 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

• Postscript: Analysis of the Evolution of a Consortium
  – Differences in “modularity” for members & users
  – Key role for brokers
Personal Genomics as a Gateway into Biology

Personal genomes soon will 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 soon will 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.
Human Genetic Variation

A Cancer Genome

A Typical Genome

Population of 2,504 peoples

<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>

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

<table>
<thead>
<tr>
<th>Prevalence of Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common</td>
</tr>
<tr>
<td>Rare* (1-4%)</td>
</tr>
<tr>
<td>Rare (~75%)</td>
</tr>
</tbody>
</table>

* 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 associated with phenotype (i.e., 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 Exponential 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

• Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “*sensitive sites*” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
  - Prioritizing using AlleleDB in terms of allelic elements
    • Having observed difference in molecular activity in many contexts

• 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

• Postscript: Analysis of the Evolution of a Consortium
  - Differences in “modularity” for members & users
  - Key role for brokers
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)

- Score against the control

Now an update: "PeakSeq 2" => MUSIC [Rozowsky et al. (09) Nat Biotech]
MulOscale Analysis, Minima/Maxima based Coarse Segmentation

Maxima
Minima

Window Length
1 kb
4 kb
16 kb
64 kb

Harmanci et al., Genome Biology 2014, MUSIC.gersteinlab.org
Multiscale Decomposition

[Harmanci et al, Genome Biol. ('14)]
Multiscale Decomposition

[Harmanci et al, Genome Biol. ('14)]
Personal Genomics: Managing Exponential 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

• Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
  - Prioritizing using AlleleDB in terms of allelic elements
    • Having observed difference in molecular activity in many contexts

• 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

• Postscript: Analysis of the Evolution of a Consortium
  - Differences in “modularity” for members & users
  - Key role for brokers
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
Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

[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)

Defining Sensitive non-coding Regions

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)]
Personal Genomics: Managing Exponential 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

• Evaluating the Impact of Non-coding Variants with Annotation
  – Annotating non-coding regions on different scales with MUSIC
  – Prioritizing rare variants with “sensitive sites” (human-conserved)
  – Prioritizing in terms of network connectivity (eg hubs)
  – Prioritizing using AlleleDB in terms of allelic elements
    • Having observed difference in molecular activity in many contexts

• 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

• Postscript: Analysis of the Evolution of a Consortium
  – Differences in “modularity” for members & users
  – Key role for brokers
Relating Non-coding Annotation to Protein-coding Genes via Networks

Regulatory elements

Assigning proximal sites (< 1 Kb) to target genes

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

Distal signals

Expression levels

Methylation H3K27ac ...
GM12878
H1-hESC
HeLa-S3
Hep-G2
K562 ...

Gene 1 Gene 2 Gene 3 ...

Scale

Strong

Weak

Connecting Distal Elements via Activity Correlations.

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

Connecting Distal Elements via Activity Correlations.

Other strategies to create linkage incl. eQTL and Hi-C. Much in recent Epigenomics Roadmap.
Hubs Under Constraint: A Finding from the Network Biology Community

- 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

LoF-tolerant genes
Essential genes

Size of nodes scaled by total degree

[Khurana et al., PLOS Comp. Bio. '13]
Personal Genomics: Managing Exponential 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

- **Evaluating the Impact of Non-coding Variants with Annotation**
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts

- **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

- **Postscript: Analysis of the Evolution of a Consortium**
  - Differences in “modularity” for members & users
  - Key role for brokers
Allele-specific binding and expression

Genomic variants affecting allele-specific behavior
  e.g. allele-specific binding (ASB)

e.g. allele-specific expression (ASE)
Inferring Allele Specific Binding/Expression using Sequence Reads

RNA/ChIP-Seq Reads

- ACTTTGATAGCGTCAA TG
- CTTTGATAGCGTCAA TG C
- CTTTGATAGCGTCAA CGC
- TTGACAGCGTCAA TGCAC
- TGATAGCGTCAA TGCACG
- ATAGCGTCAA TGCACGTC
- TAGCGTCAA TGCACGTCG
- CGTCAA CGCACGTCGGGA
- GTCAA TGCACGTCGAGAG
- CAA TGCACGTCGGGAGTT
- AATGCACGTCGGGAGTTG
- TGCACGTTGGGAGTTGGC

Haplotypes with a Heterozygous Polymorphism

10 x T
2 x C
AlleleDB: Building 382 personal genomes to detect allele-specific variants on a large-scale

1. Build personal genomes

2. Align ChIP-seq & RNA-seq reads

3. Detect allele-specific variants via a series of filters and tests

Many Technical Issues:
- Reference bias,
- Ambiguous mapping bias,
- Over-dispersed (non binomial null)

allelebd.gersteinlab.org

[Chen et al. (‘16) Nat. Comm.]
AlleleDB: Annotating rare & common allele-specific variants over a population

- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

[Chen et al. ('16) Nat. Comm.]
AlleleDB: Annotating rare & common allele-specific variants over a population

- Interfaces with UCSC genome browser
- Showing ZNF331 gene structure

[Chen et al. ('16) Nat. Comm.]
Collecting ASE/ASB variants into allele-specific genomic regions

Does a particular genomic element have a higher tendency to be allele-specific? Fisher’s exact test, for the enrichment of allele-specific variants in the element (with respect to non-allele-specific variants that could potentially be called as allelic)

Human reference genome

[Chen et al. (‘16) Nat. Comm.]
Groups of elements that are enriched or depleted in allelic activity

[Chen et al. (‘16) Nat. Comm.]
Personal Genomics:
Managing Exponential 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

• **Evaluating the Impact of Non-coding Variants with Annotation**
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “**sensitive sites**” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
  - Prioritizing using AlleleDB in terms of allelic elements
    • Having observed difference in molecular activity in many contexts

• **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

• **Postscript: Analysis of the Evolution of a Consortium**
  - Differences in “modularity” for members & users
  - Key role for brokers
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
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>$x_i \sim Binomial(n_i, p)$</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 \sim 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 \sim 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.

Bottom 10% of rep. timing bins requires large correction

Top 10% of rep. timing bins requires little correction
LARVA Implementation

- [http://larva.gersteinlab.org/](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

-chromosome
-log10(p)

These have literature-verified cancer associations

noncoding annotation
p-values in sorted order

-log10(p)

[Lochovsky et al. NAR (’15)]
Personal Genomics: Managing Exponential 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

• Evaluating the Impact of Non-coding Variants with Annotation
  – Annotating non-coding regions on different scales with MUSIC
  – Prioritizing rare variants with “sensitive sites” (human-conserved)
  – Prioritizing in terms of network connectivity (eg hubs)
  – Prioritizing using AlleleDB in terms of allelic elements
    • Having observed difference in molecular activity in many contexts

• 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

• Postscript: Analysis of the Evolution of a Consortium
  – Differences in “modularity” for members & users
  – Key role for brokers
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)

L. 1000 Genomes Screen

II. Functional annotation
   Coding
   Regulatory

III. a. Sensitive
   b. Disruptive

IV. Network connectivity

V. Recurrence

Candidate drivers

1829 somatic SNVs

Found in 1000 Genomes?

N Y

123

1706

Annotated?

N Y

300

500

In ultra-sensitive region?

N Y

3

8

Target gene known?

N Y

2

1

Target gene is a hub?

N Y

Y

Recurrent?

N Y

0

1

Gene under strong selection?

N Y

2

5

Lot?

N Y

1

2

Gene is a hub?

N Y

3

6

Target gene known?

N Y

1

2

Target gene is a hub?

N Y

3

5

Recurrent?

N Y

0

1

123

Unlikely to be driver
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 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

User Variants

Highlighting variants

Variant Reports
- 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

\[ p = \frac{3}{20} \]
- Feature weight
  - Weighted with mutation patterns in natural polymorphisms (features frequently observed weight less)
  - entropy based method

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

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

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

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

[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 Exponential 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

- Evaluating the Impact of Non-coding Variants with Annotation
  - Annotating non-coding regions on different scales with MUSIC
  - Prioritizing rare variants with “sensitive sites” (human-conserved)
  - Prioritizing in terms of network connectivity (eg hubs)
  - Prioritizing using AlleleDB in terms of allelic elements
    - Having observed difference in molecular activity in many contexts

- 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

- Postscript: Analysis of the Evolution of a Consortium
  - Differences in “modularity” for members & users
  - Key role for brokers
Personal Genomics: Managing Exponential 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

• Evaluating the Impact of Non-coding Variants with Annotation
  – Annotating non-coding regions on different scales with MUSIC
  – Prioritizing rare variants with “sensitive sites” (human-conserved)
  – Prioritizing in terms of network connectivity (eg hubs)
  – Prioritizing using AlleleDB in terms of allelic elements
    • Having observed difference in molecular activity in many contexts

• 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

• Postscript: Analysis of the Evolution of a Consortium
  – Differences in “modularity” for members & users
  – Key role for brokers
Increase in Consortium Science

[Graph showing the increase in number of papers in Pubmed from 2004 to 2015 for all papers (red squares) and consortium-related papers (blue circles). The x-axis represents the years from 2004 to 2015, and the y-axis represents the number of papers in Pubmed, ranging from 0 to 12000000. The graph indicates a significant increase over the years, with a sharper rise for consortium-related papers.]
With help of NHGRI, identified:
1,786 ENCODE members & 8,263 non-members
from 558 consortium papers supported by ENCODE funding &
702 community papers that used ENCODE data but were not supported by
ENCODE funding

[Wang et al., TIG ('16)]
ENCODE co-authorship network

[Wang et al., TIG ('16)]
Network statistics highlight change in modularity with consortium rollouts (L) & importance of broker role (R)

Coalesced into single module due to ENCODE consortium papers in 2007

Some separation but retention of a unified modular structure

Number of clusters

[Wang et al., TIG (‘16)]
Similar Findings in terms of modularity & broker scientists in the modENCODE consortium as for ENCODE

modENCODE

[Wang et al., TIG ('16)]
Personal Genomics:
Managing Exponential 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

• Evaluating the Impact of Non-coding Variants with Annotation
  – Annotating non-coding regions on different scales with MUSIC
  – Prioritizing rare variants with “sensitive sites” (human-conserved)
  – Prioritizing in terms of network connectivity (eg hubs)
  – Prioritizing using AlleleDB in terms of allelic elements
    • Having observed difference in molecular activity in many contexts

• 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

• Postscript: Analysis of the Evolution of a Consortium
  – Differences in “modularity” for members & users
  – Key role for brokers
AlleleDB.gersteinlab.org

J Chen, J Rozowsky, TR Galeev, A Harmanci, R Kitchen, J Bedford, A Abyzov, Y Kong, L Regan

“Cost Seq 2”

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

LARVA.gersteinlab.org

L Lochovsky, J Zhang, Y Fu, E Khurana

archive.gersteinlab.org/proj/netsnp

E Khurana, Y Fu, J Chen

“Encodeauthors”

D Wang, K Yan, E Pan, J Rozowsky

FunSeq.gersteinlab.org

FunSeq2.gersteinlab.org


MUSIC.gersteinlab.org

A Harmanci, J Rozowsky

Acknowledgments

Hiring Postdocs. See gersteinlab.org/jobs
Extra

STOP
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