Identification of deleterious variation
GGD Pathway Lecture

Don Conrad
September 8, 2014
The playing field

• Types of variants
  – Protein coding
    • missense
    • nonsense
  – Non-coding
    • Splice site disruptors
    • Promoter disrupters
    • Regulatory element disruptors
  – Structural variants
    • (may affect multiple features)

• Types of Data
  – Annotations (evolutionary constraint, biochemical consequences, etc)
  – Population frequencies
  – In vitro or in vivo experimental assays
    • ENCODE, CRE-seq, etc.
  – Phenotypes
    • Mouse knockouts, cellular assays, etc.
  – Database of known disease mutations
Tools for evaluating performance

- Databases of disease variants and benign variants
- ROC curves
- Simulations
Direct identification of the causal gene for a monogenic disorder by exome sequencing.

<table>
<thead>
<tr>
<th>Number of genes in which each affected has at least one...</th>
<th>Any 3 of 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)</td>
<td>3,768</td>
</tr>
<tr>
<td>NS/SS/I not in dbSNP</td>
<td>119</td>
</tr>
<tr>
<td>NS/SS/I not in eight HapMap exomes</td>
<td>160</td>
</tr>
<tr>
<td>NS/SS/I neither in dbSNP nor eight HapMap exomes</td>
<td>22</td>
</tr>
<tr>
<td>*...And predicted to be damaging</td>
<td>3</td>
</tr>
</tbody>
</table>


*By PolyPhen*
An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge

The Facts:

- 30 groups competed
- WGS from 3 families (muscular dystrophy, muscular dystrophy + hearing loss, cardiac defects)
- “only 2 groups identified the consensus candidate variants in all cases, demonstrating a need for fine-tuning of the generally accepted methods”
Lessons from CLARITY

• Amino-acid substitutions: “The most common tools to tackle the problem of determining the effect of amino acid substitutions on protein function for missense mutations were SIFT and Polyphen”

• Splice-disrupting mutations: 15% of all annotated disease mutations (e.g. in HGMD) are splice-site mutations

• Other non-coding variants: ½ of the teams assessed non-coding variants: 46% of these looked at TFBS “motif breakers” and 23% considered changes in known promoter/enhancer elements. None annotated any non-coding variants as pathogenic
SIFT: Sorting Intolerant From Tolerant (2002)
• $a_i = \text{probability of amino acid } i \text{ at a position}$
• $a_{\text{max}} = \text{max } \{a\}$
• $a'_i = a_i / a_{\text{max}} = \text{“SIFT Score”}$
• If $a'_i < \text{threshold}$, $T \Rightarrow \text{“deleterious”}$
PolyPhen-2, Adzhubei, et al. 2010
Nature Methods
"naïve" Bayes Classifier

\[ X = \text{classification, e.g. "disease" or "benign"} \]
\[ F = \text{feature annotation} \]

\[ \Pr(X \mid F_1, \ldots, F_n) = \frac{1}{C} \Pr(X) \prod_{i=1}^{n} \Pr(F_i \mid X) \]

<p>| Pr(X|F₁,...,Fₙ) | PolyPhen-2 Label          | HumDiv FDR |
|-----------------|---------------------------|------------|
| &gt;0.85           | “probably damaging”       | 0.1        |
| 0.15-0.85       | “possibly damaging”       | 0.18       |
| &lt; 0.15          | “benign”                  |            |</p>
<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Values with ranges in HumDiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>nt1</td>
<td>wild type allele nucleotide</td>
<td>A, C, G, T</td>
</tr>
<tr>
<td>nt2</td>
<td>mutation allele nucleotide</td>
<td>A, C, G, T</td>
</tr>
<tr>
<td>site</td>
<td>SITE annotation from UniProt/Swiss-Prot</td>
<td>Yes, No</td>
</tr>
<tr>
<td>region</td>
<td>REGION annotation from UniProt/Swiss-Prot</td>
<td>NO, PROPER, SIGNAL, TRANSMEM</td>
</tr>
<tr>
<td>phat</td>
<td>PHAT matrix element in the TRANSMEM region</td>
<td>[-8.0, 4.0], mean = -0.04</td>
</tr>
<tr>
<td>score1</td>
<td>PSIC score for the wild type allele</td>
<td>[-1.1], mean = 1.07</td>
</tr>
<tr>
<td>score2</td>
<td>PSIC score for the mutant allele</td>
<td>[-1.39, 2.64], mean = 0.166</td>
</tr>
<tr>
<td>score_delta</td>
<td>difference of PSIC scores (Score1-Score2)</td>
<td>[-3.23, 4.57], mean = 0.905</td>
</tr>
<tr>
<td>num_observ</td>
<td>number of residues observed at the position of the multiple alignment</td>
<td>[1, 432], mean 69.3</td>
</tr>
<tr>
<td>delta_volume</td>
<td>change in residue side chain volume</td>
<td>[-167, 167], mean = -1.93</td>
</tr>
<tr>
<td>transv</td>
<td>mutation origin by transversion or transition</td>
<td>Yes, No</td>
</tr>
<tr>
<td>Cpg</td>
<td>mutation origin in the CpG hypermutable context</td>
<td>Yes, No</td>
</tr>
<tr>
<td>pfam_hit</td>
<td>position of the mutation within/outside a protein domain as defined by Pfam</td>
<td>Yes, No</td>
</tr>
<tr>
<td>id_p_max</td>
<td>congruency of the mutant allele to the multiple alignment</td>
<td>[0, 95.5], mean = 24</td>
</tr>
<tr>
<td>id_q_min</td>
<td>sequence identity with the closest homologue deviating from wild type allele</td>
<td>[1.56, 95.5], mean 68.76</td>
</tr>
<tr>
<td>cpgVar1Var2</td>
<td>presence of the CpG context combined with wild type and mutant amino acid types</td>
<td>NO, AA1_AA2</td>
</tr>
<tr>
<td>cpg_transition</td>
<td>whether variant happened as transition in CpG context</td>
<td>No, Transition, Transversion</td>
</tr>
<tr>
<td>charge_change</td>
<td>change in electrostatic charge</td>
<td>0.12</td>
</tr>
<tr>
<td>hydroph_change</td>
<td>change in hydrophobicity</td>
<td>[0, 0.8], mean 0.08</td>
</tr>
<tr>
<td>ali_iide</td>
<td>sequence identity with the closest homolog with known 3D structure</td>
<td>[0, 1], mean 0.33</td>
</tr>
<tr>
<td>ali_len</td>
<td>alignment length with the closest homolog with known 3D structure</td>
<td>[0, 1213], mean 130.0</td>
</tr>
<tr>
<td>acc_normed</td>
<td>normalized accessible surface area of amino acid residue</td>
<td>[0, 1.55], mean 0.35</td>
</tr>
<tr>
<td>sec_str</td>
<td>secondary structure</td>
<td>HELIX, SHEET, OTHER</td>
</tr>
<tr>
<td>map_region</td>
<td>region of the Ramachandran map</td>
<td>ALPHA, BETA, OTHER</td>
</tr>
<tr>
<td>delta_prop</td>
<td>change in accessible surface area propensity</td>
<td>[-2.89, 2.89], mean -0.07</td>
</tr>
<tr>
<td>b_fact</td>
<td>crystallographic beta-factor</td>
<td>[-1.85, 5.17], mean 0.0</td>
</tr>
<tr>
<td>hct_cont_ave_num</td>
<td>average number of contact with heteroatoms</td>
<td>Yes, No</td>
</tr>
<tr>
<td>hct_cont_min_dist</td>
<td>minimal distance to a heteroatom</td>
<td>Yes, No</td>
</tr>
<tr>
<td>inter_cont_ave_num</td>
<td>average number of interchain contacts in a protein complex</td>
<td>Yes, No</td>
</tr>
<tr>
<td>inter_cont_min_dist</td>
<td>average minimal interchain distance</td>
<td>Yes, No</td>
</tr>
<tr>
<td>delta_volume_new</td>
<td>change in residue volume for buried residues</td>
<td>[-119, 138], mean -0.5</td>
</tr>
<tr>
<td>delta_prop_new</td>
<td>change in accessible surface area propensity for buried residues</td>
<td>[-1.83, 2.89], mean 0.0026</td>
</tr>
</tbody>
</table>

**Supplementary Table 1.** Complete list of all 32 initial features considered, with 11 final features selected for use in PolyPhen-2 classifier highlighted in blue font.
PolyPhen-2 Training Data

**HumDiv dataset**
- 3,155 Mendelian disease mutations from UniProt
- 6,312 inter-species substitutions
- ftp://genetics.bwh.harvard.edu/datasets/HumDiv.tar.gz

**HumVar dataset**
- all 13,032 human disease-causing mutations from UniProt
- 8,946 human nsSNPs without annotated involvement in disease
Historical methods for annotating amino acid substitutions are not concordant.

Percent pairwise agreement between existing methods

<table>
<thead>
<tr>
<th>Method</th>
<th>PhyloP</th>
<th>SIFT</th>
<th>Polyphen2</th>
<th>LRT</th>
<th>MutationTaster</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhyloP</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.189</td>
<td>-</td>
<td>68.82</td>
<td>62.14</td>
<td>61.69</td>
</tr>
<tr>
<td>Polyphen2</td>
<td>0.306</td>
<td>0.517</td>
<td></td>
<td>66.82</td>
<td>64.91</td>
</tr>
<tr>
<td>LRT</td>
<td>0.475</td>
<td>0.267</td>
<td>0.457</td>
<td></td>
<td>77.19</td>
</tr>
<tr>
<td>MutationTaster</td>
<td>0.389</td>
<td>0.313</td>
<td>0.444</td>
<td>0.618</td>
<td>-</td>
</tr>
</tbody>
</table>
Performance of existing methods in classifying known disease and benign SNVs

Data composition:
HGMD (n=36140)
1000 Genomes (n=45115)
Combining existing methods improves classification performance.

Data composition:
- HGMD (n=36140)
- 1000 Genomes (n=45115)
Genes: Beyond Amino Acid Substitutions
Integrating RNA data - splicing

- 15% of Mendelian disease mutations are splice site disruptions
- “Fuzzy” sequence motifs control splicing
- Mutations may activate cryptic splice sites

Integrating RNA data - ASE

- ASE: Allele-specific expression
Integrating RNA-seq- NMD

- NMD= Nonsense-mediated decay
- “50 bp rule” – only nonsense mutations 50-55bp upstream of the 3’-most exon-exon junction trigger NMD
Integrating RNA-seq – interpreting personal genomes

- Assess the extent of biallelic expression at recessive disease genes harboring heterozygous genotypes
- Assess the extent of NMD in LoF-containing transcripts
- Detect “weird” isoforms
- Feasibility
  - sampling issues
  - Chronic versus developmental disease
Non-coding variation
GERP: Genome Evolutionary Rate Profiling

Measures ratio : observed/expected nucleotide substitutions
The LD “Problem”
Both rs10782001 and rs12924903, which are in strong LD with each other, regulate r–lo. The strongest differential expression between psoriatic and normal skin (2.66-fold increase for the observed association signal, combined acting P = 0.02). FBXL10 and FBXL11 contain rich repeats in addition to the F-box protein family, only three (FBXL10, FBXL11, and FBXL19) were associated region could be responsible for the observed association signal, for PsC in a subphenotype analysis, it also yielded a strong association. Imputation based on the 1000 Genomes Project led us to select this SNP for typing and revealed that the association signals could be driven by SNPs in the GWAS discovery and a 1.43 fold increase for PsA than with PsC at a statistically significant level (OR = 1.12, acting P = 0.02). Evidence for psoriasis association in four genomic regions including the three new loci attaining genome-wide significance and the confirmed region. The upper portion of each plot depicts association combined --lo gP, and the lower portion depicts RefSeq genes and LD plots from the phase 2 HapMap CEU sample. For each region, Linkage disequilibrium for the CEPH (CEU) from phased genotypes.

Hierarchical Model:
Background

\[
\begin{align*}
\text{Hierarchical Model:} & \\
\text{Background} & \\
\end{align*}
\]
Hierarchical Model: Background

Level I

\[ \lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_L \]

Level II

\[ BF'_{jK} = BF_{jk} \pi_{jk} \]

\[ BF_{jK} = BF_{jk} \pi_{jk} \]

\[ BF_{jK} = BF_{jk} \pi_{jk} \]
Bayesian Regression

SNP priors based on annotations

Annotation I
Annotation II
Annotation III

Posterior combines regression signal with prior

Posterior probability of being an eQTN

Genomic Position

Slide Credit: Daniel Gaffney
Previous SNP-based results

Features predictive of LCL eQTLs (>50 tested)

- Proximity to gene
- Histone marks ($n = 5$ types)
- DNaseI hypersensitivity sites
- Core promoter motifs ($n = 2$)
- Transcription factor binding sites ($n = 4$)
Annotation effect sizes in the combined model

Chromatin Marks:
>5kb Upstream
H3k27ac
H3k4me2
DNase1
H3k4me1
H3k9ac
H3k4me3

Chromatin Marks:
Within 5kb of TSS
DNase1
H3k9ac
H3k4me3
H3k27ac

Chromatin Marks:
>5kb Downstream
H3k36me3
H3k27ac
DNase1
H3k4me1
H3k9ac
H3k4me2
H3k4me3

Inferred binding sites
ETS Family
Interferon Response
NFkB

Transcription Factor
ChIP−seq
regions
Jun−D
NFkB
SRF
TAFII
MAX
GABP
NRSF
C−fos

Core promoter
motifs
Overrep. 6−mers
regPotential
Known Elements

\log_{10}(eQTL Enrichment)
### Annotations

- **DNase1**
- **H3K27ac**
- **H3K4me1**
- **H3K4me3**
- **H3K9ac**
- **H3K36me3**
- **GABP**
- **Jun-D**
- **NF-kB**
- **MAX**
- **TAF II**
- **SRF**
- **regPotential**

### Variation in NF-κB binding between individuals

#### NfKB ChIP-seq read depth

**CC: High exp**

**TT: Low exp**

**TC**

Distance from SNP

Gaffney, et al 2012 Genome Biology
BG: distance model only
Experimental: model using ENCODE-like annotations
Bins correspond to sets of 25 SNPs
Rank: rank using just the prior probability
Percentage: % of all putative eQTNs
CADD: combined annotation-dependent depletion

Method: SVM (support vector machine classifier)

Training data: - full genome, not just coding variants
  - "Simulated" variants – 44M SNVs, 2.1M insertions, 3.1M deletions
  - "Observed" variants – 14M SNVs, 627K insertions, 1.1M deletions

Annotations: 63 annotations – evolutionary constraint, ENCODE, transcription metrics, protein metrics (polyPhen, SIFT, etc)
Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics


4 OCTOBER 2013 VOL 342 SCIENCE www.sciencemag.org

“FunSeq”
The n=1 problem: How do we put a p-value on a unique case?

- 7,000-11,000 non-synonymous SNVs (nsSNVs) per individual
- Most nsSNVs are benign, but some contribute to disease
- Candidate nsSNVs in many rare diseases and unique cases are difficult to identify

Doctors typically see one individual, not a population, with the same idiopathic Mendelian disorder
The coloring problem
The classification problem

Disease population

Healthy population
Gene-level model

Incorporate information that captures the genetics of the variant and physiology of the gene (probability of haploinsufficiency or recessivity)

Combining existing methods and gene physiology further improves performance

Data composition:
HGMD (n=36140)
1000 Genomes (n=45115)
Simulating the n=1 problem

Select known mendelian disease variant (HGMD)

Evaluate and repeat with a new disease variant

Compare disease variant score to variants in healthy exome (1000 Genomes)
Population-level model

Coding variants from 26,000 exomes were scored using pDa and used to establish null distributions of observed scores in healthy individuals by gene.

Allele frequencies for observed variants were established by sequencing 26,000 exomes, variants not observed in the exome data were annotated with gene based mutation rates (μ).

\[ P(V | G, S) = P(S | G, V) \times P(G, V) \times P(S | G) \times P(G) \]

S= score
G=Gene
V=genotype

All possible coding variants were simulated and used to establish the distribution of possible scores by gene.
Evaluation of simulated $n=1$ problem
Population data greatly reduces false negative signals
Some concluding thoughts

• Best methods are likely to be gene-specific (or “module”-specific)
• Best methods will someday be disease-specific and tissue-specific (and sex-specific, age-specific, and...?)
• We want to annotate diplotypes, not variants
• Best methods will someday integrate many data types from patient