Penetrance Estimates for Incidental Genomic Findings

James Diao, Arjun Manrai, Isaac Kohane

Division of Health Sciences and Technology, Harvard-MIT
Department of Biomedical Informatics, Harvard Medical School

INTRODUCTION

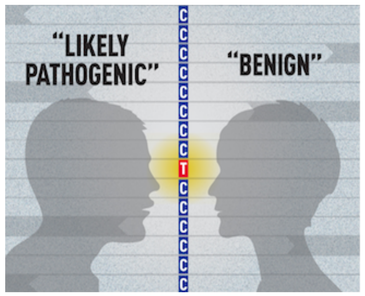
(Genetic Testing and Relevant Datasets)

Genetic testing: a difference from the reference genome (variant) may indicate disease.

Incidental finding: variant in gene unrelated to diagnostic indication that prompted sequencing.

-Due to multiple testing and low priors, these typically have high rates of false positives, so we normally don't report them.

ACMG (American College of Medical Genetics & Genomics): recommends an exception for 56 genes thought to be more indicative of disease.



1000 Genomes Project: contains whole-genome sequence data for 2,504 healthy adults from diverse ethnic populations.

ClinVar: central repository of interpretations for genetic variants (benign vs. pathogenic).

OBJECTIVES

1. Develop an ETL workflow for extraction, transformation, and loading of genomic and interpretation data from relevant sources.

2. Evaluate variant distribution across a healthy, diverse cohort (1000 Genomes).

3. Estimate plausible penetrance ranges for the ACMG recommendations.

PENETRANCE MODEL

$$Penetrance = P(D|V) = \frac{P(D) * P(V|D)}{P(V)} = \frac{(prevalence)(allelic\ heterogeneity)}{(allele\ frequency)}$$

where D = disease, V = any variant

Penetrance: Probability of developing disease, given a positive genetic test result.
Prevalence: Proportion of general population with disease.
Allelic Heterogeneity: Proportion of diseased population with a pathogenic variant.
Allele Frequency: Proportion of general population with a pathogenic variant.

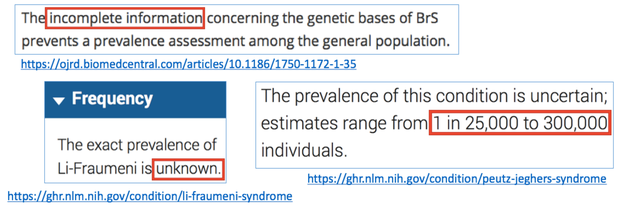
METHODS & WORKFLOW

Literature Search

1. Group disease subtypes into 30 categories.

2. Query Google Scholar for "[disease name] prevalence." Prioritize studies with PubMed IDs, more citations, and larger sample sizes.

3. Record prevalence values + URL, year, etc.



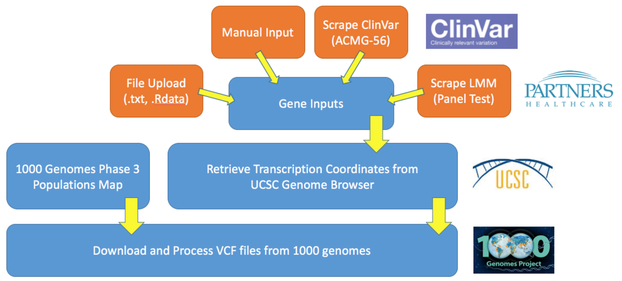
ETL for Datasets

Pipeline + UI using R/Shiny/Markdown

1. Extract: query UCSC Genome Browser for gene regions and retrieve corresponding VCF files from 1000 Genomes.

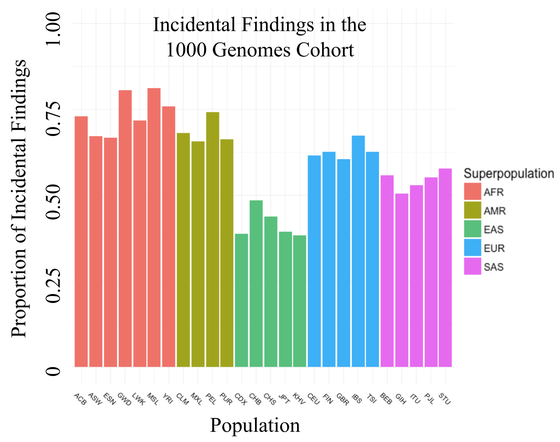
2. Transform: separate variants with multiple alternates; convert genotypes to allele counts.

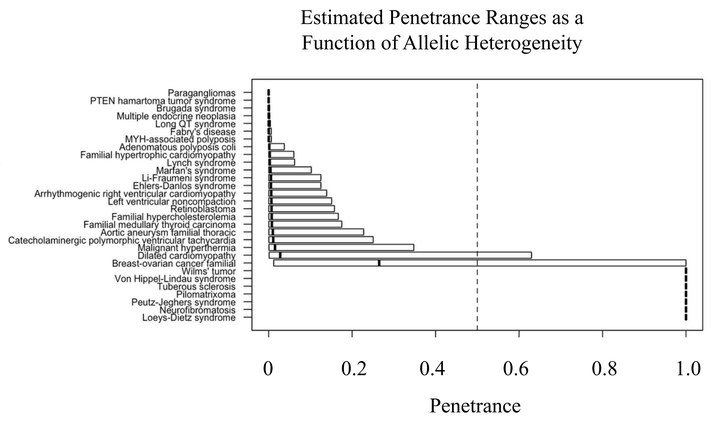
3. Load: collect labels from the Phase 3 Populations Map. Stage final data objects.



https://github.com/jamesdiao/2016-paper-ACMG-penetrance

KEY FIGURES





CONCLUSIONS

1. High counts: 40-80% of individuals have an incidental finding under ACMG guidelines, far higher than empirical disease prevalences.

2. Clustered distribution: by ethnicity – AFR (African) have the most findings, EAS (East Asian) have the fewest.

3. High sensitivity: findings dominated by a few high-frequency variants.

4. Very low penetrance estimates:
Out of the 30 diseases (22 with data):
(a) 20 have max theoretical penetrance **< 50%**
(b) 12 have max theoretical penetrance **< 5%**

5. High uncertainty around parameters: translates into very large errors bars.

*-This is a preliminary "letter-of-the-law" evaluation and does **not** yet demonstrate real-world effects on patients.*

NEXT STEPS

1. Identify questionable variants:
(a) high-frequency (common findings)
(b) highly enriched in 1 ethnic population.

2. Validation with empirical penetrance values and other sequencing datasets (e.g. gnomAD).

3. Model biases in parameter estimates (prevalence, pathogenicity, etc.)

4. Confer with clinical collaborators to determine alternate protocols at Laboratory of Molecular Medicine and Partners HealthCare.

ACKNOWLEDGEMENTS

Raj: for his mentorship throughout this project.

Zak: for looking over & shaping my presentations.

HST Summer Institute Administration - (Susanne, Barbara, Dominique, Jean, and Sonal): for making everything possible.

DBMI: for all the AC and coffee.