Lecture 23: PheWAS
Phenome-wide association studies

Slides credit:
Yue Li
Yongjin Park
<table>
<thead>
<tr>
<th>Project</th>
<th>Sets</th>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Lec</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>Thu</td>
<td>Sep 8</td>
<td>L1</td>
<td>Intro: Biology, Algorithms, Machine Learning, Course Overview</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Fri</td>
<td>Sep 9</td>
<td>R1</td>
<td>Recitation 1: Biology and Probability Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Thu</td>
<td>Sep 13</td>
<td>L2</td>
<td>Dynamic Programming, Global and local alignment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Tue</td>
<td>Sep 15</td>
<td>L3</td>
<td>Database search, Rapid string matching, BLAST, BLOSUM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Fri</td>
<td>Sep 15</td>
<td>R2</td>
<td>Recitation 2: Deriving Parameters of Alignment, Multiple Alignment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Thu</td>
<td>Sep 20</td>
<td>L4</td>
<td>Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Thu</td>
<td>Sep 22</td>
<td>L5</td>
<td>Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Fri</td>
<td>Sep 23</td>
<td></td>
<td>No classes - student holiday</td>
</tr>
<tr>
<td>Project Profile due</td>
<td>Tue 9/27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Thu</td>
<td>Sep 27</td>
<td>L6</td>
<td>Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Thu</td>
<td>Sep 29</td>
<td>L7</td>
<td>Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Fri</td>
<td>Sep 30</td>
<td>R3</td>
<td>Affinity Propagation Clustering and Random Forest Classification</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Tue</td>
<td>Oct 4</td>
<td>L8</td>
<td>Epigenomics: ChiP-Seq, Read mapping, Peak calling, IDR, Chromatin states</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Fri</td>
<td>Oct 6</td>
<td>L9</td>
<td>Three-dimensional chromatin interactions: 3C, 5C, HiC, Chi-A-Pet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Thu</td>
<td>Oct 7</td>
<td>R4</td>
<td>Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromMappe</td>
</tr>
<tr>
<td>Project Introduction about the projects, self introductions, mentor introduction, example projects, team mentor 32D-507</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Fri</td>
<td>Oct 7</td>
<td></td>
<td>No classes - Columbus Day Holiday</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Thu</td>
<td>Oct 11</td>
<td>L10</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Fri</td>
<td>Oct 14</td>
<td>R5</td>
<td>Recitation 6: Gapped Motif Discovery, DNAShape, PBMs, Selex</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Thu</td>
<td>Oct 13</td>
<td>L11</td>
<td>Neural Networks, Belief Networks, Deep learning</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Fri</td>
<td>Oct 21</td>
<td>L12</td>
<td>Networks II: Bayesian inference, Variational Bayes, approximate inference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Fri</td>
<td>Oct 21</td>
<td>R6</td>
<td>Networks II: Recommendation systems, SystM, Phewas</td>
</tr>
<tr>
<td>Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507 at 4-5pm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Thu</td>
<td>Oct 27</td>
<td>L13</td>
<td>Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Fri</td>
<td>Oct 28</td>
<td>L14</td>
<td>Disease Association Mapping, GWAS, organisinal phenotypes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Thu</td>
<td>Oct 27</td>
<td>L15</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Fri</td>
<td>Oct 28</td>
<td>L16</td>
<td>Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Fri</td>
<td>Oct 28</td>
<td></td>
<td>No Recitation, Veterans Day</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Thu</td>
<td>Nov 17</td>
<td>L17</td>
<td>Comparative genomics and evolutionary signatures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Fri</td>
<td>Nov 17</td>
<td>L18</td>
<td>Genome Scale Evolution, Genome Duplication</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Fri</td>
<td>Nov 24</td>
<td></td>
<td>No lecture, thanksgiving break - Thu Nov 25, 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>Thu</td>
<td>Nov 29</td>
<td>L19</td>
<td>Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17</td>
<td>Fri</td>
<td>Nov 29</td>
<td>L20</td>
<td>Phylogenetics: Gene/species trees, reconciliation, coalescent, ARGs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18</td>
<td>Thu</td>
<td>Dec 1</td>
<td>R9</td>
<td>Phylogenetic distance metrics, Coalescent Process</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19</td>
<td>Fri</td>
<td>Dec 2</td>
<td></td>
<td>No Recitation, thanksgiving break</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>Thu</td>
<td>Dec 8</td>
<td>L21</td>
<td>Single-cell genomics: technology, analysis, microfluidics, applications, insights</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21</td>
<td>Fri</td>
<td>Dec 8</td>
<td>L22</td>
<td>Mining human phenotypes, Phewas, UK Biobank, meta-phenotypes-imputation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22</td>
<td>Thu</td>
<td>Dec 9</td>
<td>R10</td>
<td>Recitation 10: Project Feedback, results, interpretation, directions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23</td>
<td>Fri</td>
<td>Dec 9</td>
<td>L23</td>
<td>Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
<td>Thu</td>
<td>Dec 13</td>
<td>L24</td>
<td>Genome Engineering with CRISPR/Cas9 and related technologies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>Fri</td>
<td>Dec 19</td>
<td>R11</td>
<td>Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills</td>
</tr>
<tr>
<td></td>
<td></td>
<td>26</td>
<td>Thu</td>
<td>Dec 13</td>
<td>L25</td>
<td>Final Presentations - Part I (11am), 32-80 reading room</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27</td>
<td>Fri</td>
<td>Dec 19</td>
<td>L25</td>
<td>Final Presentations - Part I (1pm), 32-141</td>
</tr>
<tr>
<td>Conference report due</td>
<td>Sun 12/11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Thu</td>
<td>Dec 1</td>
<td>L26</td>
<td>Intro: Biology, Algorithms, Machine Learning, Course Overview</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Fri</td>
<td>Dec 2</td>
<td>R10</td>
<td>Recitation 10: Project Feedback, results, interpretation, directions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Thu</td>
<td>Dec 8</td>
<td>L23</td>
<td>Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Fri</td>
<td>Dec 9</td>
<td>R11</td>
<td>Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Thu</td>
<td>Dec 13</td>
<td>L25</td>
<td>Final Presentations - Part I (11am), 32-80 reading room</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>Fri</td>
<td>Dec 19</td>
<td>L25</td>
<td>Final Presentations - Part I (1pm), 32-141</td>
</tr>
</tbody>
</table>
Module 6: Current research directions

• L21: Single-cell genomics
  • Measuring and analyzing biology at the single-cell level

• L22: PheWAS (Phenome-wide associations studies)
  • Multi-phenotype analyses, inferences, association, imputation

• L23: Cancer genomics
  • Mutational heterogeneity, tumor evolution, immune evasion

• L24: Genome engineering & high-throughput biology
  • From reading to writing, CRISPR-Cas9,
PheWAS: Multi-phenotype studies

1. Motivation of phenome-wide association studies
   - PheWAS-informed phenotyping, improved GWAS power, etc
   - Electronic health record (EHR) contain rich personalized information

2. Modeling multiple phenotypes in GWAS + epigenomics
   - Integration of multiple phenotypes in GWAS from Systems Genetics perspective (clustering approach)
   - Direct integration of multiple phenotypes by summary-based factored genetic model estimation

3. Epigenomics of PheWAS
   - Risk variants inference using epigenomic reference annotations
   - Using disease covariance to improve functional variants inference
   - Combining enrichment to improve causal pathway inference

4. Meta-phenotype inference and imputation
   - Models leveraging missing information and inferring missing mechanism
   - Modeling multimodal electronic health record data
   - Imputing missing EHR code and prioritizing patient disease risks
**Why Multi-phenotype analyses?**

- Associate unknown/underappreciated phenotypes with known
- Identify disease mechanisms by mediating phenotypes
- Predict disease with easy-to-measure ‘biomarker’ phenotypes
- Improve GWAS power by imputation of missing phenotypes
- Improve GWAS biological relevance by meta-phenotype association
- Understand pleiotropic effects affecting multiple phenotypes
- Guide experiments/diagnosis by predicting missing phenotypes
- Enable personalized medicine by treating combinations of symptoms
Electronic health records (EHRs)

EHR contains extremely rich information of a patient:
• Clinical notes (unstructured free-form text);
• Lab tests: Logical Obs. Identifiers Names & Codes (LOINC)
• Billing code: International Classification of Disease (ICD)
• Billing code: Diagnosis-related group (DRG)
• Pharmaceutical: Prescription data (RxNorm)
Rapid adoption of EHR technology in the US 2008-2015

As EHR data become more comprehensive, there is an urgent need for integrative machine learning approach to model them.
Ascertainment bias: Disease-dependent phenotypes

- Raw EHR data are indirect reflection of true patient state (due to recording process)
- State of patient varies
- Patient state influences:
  - Value of the measurements
  - Whether there is a measurement
  - Type of measurement
  - Timing of measurement

Hripcsak & Albers (2012)
PheWAS: Multi-phenotype studies

1. Motivation of phenome-wide association studies
   • PheWAS-informed phenotyping, improved GWAS power, etc
   • Electronic health record (EHR) contain rich personalized information

2. Modeling multiple phenotypes in GWAS + epigenomics
   • Integration of multiple phenotypes in GWAS from Systems Genetics perspective (clustering approach)
   • Direct integration of multiple phenotypes by summary-based factored genetic model estimation

3. Epigenomics of PheWAS
   • Risk variants inference using epigenomic reference annotations
   • Using disease covariance to improve functional variants inference
   • Combining enrichment to improve causal pathway inference

4. Meta-phenotype inference and imputation
   • Models leveraging missing information and inferring missing mechanism
   • Modeling multimodal electronic health record data
   • Imputing missing EHR code and prioritizing patient disease risks
2. Modeling PheWAS using genetic information: inferring pleiotropic variants and pathways implicated in multiple phenotypes

a. Shared risk variants
b. Summary statistics-based methods
c. Shared heritability. Using polygenic risk prediction.
d. Multiple phenotype sharing at the pathway level
Common polygenic risk variants for schizophrenia and bipolar disorder

- GWAS was performed on International Schizophrenia Consortium (ISC) with ~3.5k cases/controls
- The ISC-derived score was highly associated with disease in the two European schizophrenia samples (MGS-EA, O’Donovan)
- The ISC-derived score alleles were also associated with bipolar disorder in two independent samples.

The International Schizophrenia Consortium (Nat, 2009)
Genome-wide multi-phenotype association

Each diamond represents a unique phenotype association at each SNP.

Denny et al., (NatBiotech, 2013)
Reverse GWAS: PheWAS Manhanttan plots

X-axis: Phenotypes instead of SNPs
Y-axis: association
Each panel: a SNP
Each point: SNP-phenotype association
Each panel represents 1,358 phenotypes tested for association with a particular SNP.

Denny et al., (NatBiotech, 2013)
PheWAS scan using EHR ICD9 code

Denny et al., (Bioinf 2010)
PheWAS using NHANES data (National Health and Nutrition Examination Survey)

One “stone” (SNP) many “birds” (traits)

APOE/APOC1/C1P1/C2/C4, rs4420638, Coded Allele A

Length of the lines correspond to -logP

rs4420638 near APOC1 associated with LDL

Pendergrass et al., (PlosGen, 2013)
Leverage multiple traits in GWAS

• Modelling multiple traits with multivariate linear mixed model:

\[ Y = BX + G + E \]

\[ G \sim MN(0, V_g, D_k) \]

\[ E \sim MN(0, V_e, I_{n\times n}) \]

• \( Y \): D x N matrix of N individuals and D traits
• \( B \): D x M effect size matrix for M SNPs
• \( X \): M x N genotype matrix
• \( G \): D x N random effect matrix and follows matrix normal (MN)
• \( V_g \): D x D symmetric matrix of genetic variance component
• \( V_e \): D x D symmetric matrix of environmental variance component
• \( D_k \): N x N diagonal matrix filled with eigen values of the kinship matrix

Xiang & Stephens, (NatMeth, 2014)
Power gain in modeling multiple phenotypes

Xiang & Stephens, (NatMeth, 2014)
Genetic correlation w/ LD score regression

\[ E[z_{1j}z_{2j}l_j] = \frac{\sqrt{N_1N_2}\rho_g}{M}l_j + \frac{\rho N_s}{\sqrt{N_1N_2}} \]

\[ l_j = \sum_k r_{jk}^2 \text{ is the LD score of SNP } j \]

- \( N_i \) is the sample size for study \( i \)
- \( \rho_g \) is the genetic covariance, which can be efficiently estimated by fitting linear regression
- \( l_j = \sum_k r_{jk}^2 \) is the LD score of SNP \( j \)
- \( N_S \) is the number of individuals included in both studies
- \( \rho \) is the phenootypic correlation among the \( N_S \) overlapping samples

Sullivan et al., (NatGen, 2015)
LDSC-estimated genetic corr.
Detecting shared genetic influences by regional Bayes factor

\[
l(\Theta|D) = \sum_{i=1}^{M} \ln \left( \Pi_0 + \sum_{j=1}^{4} \pi_j RBF_{i}^{(j)} \right)
\]

Pickrell et al., (NatGen 2016)
Proportion of shared risk variants among 42 traits

Pickrell et al., (NatGen 2016)
Causal inference

**a)**
BMI vs. TG (BMI ascertainment)

**b)**
BMI vs. TG (TG ascertainment)

**c)**
LDL vs. CAD (LDL ascertainment)

**d)**
LDL vs. CAD (CAD ascertainment)

---

Pickrell et al., (NatGen 2016)
Causal inference

Pickrell et al., (NatGen 2016)
Probabilistically inferring pleiotropic variants

\[ \pi_{00} = \Pr (Z_{j00} = 1) : \quad (P_{j1}|Z_{j00} = 1) \sim U[0,1], \]
\[ (P_{j2}|Z_{j00} = 1) \sim U[0,1], \]
\[ \pi_{10} = \Pr (Z_{j10} = 1) : \quad (P_{j1}|Z_{j10} = 1) \sim Beta(\alpha_1,1), \]
\[ (P_{j2}|Z_{j10} = 1) \sim U[0,1], \]
\[ \pi_{01} = \Pr (Z_{j01} = 1) : \quad (P_{j1}|Z_{j01} = 1) \sim U[0,1], \]
\[ (P_{j2}|Z_{j01} = 1) \sim Beta(\alpha_2,1), \]
\[ \pi_{11} = \Pr (Z_{j11} = 1) : \quad (P_{j1}|Z_{j11} = 1) \sim Beta(\alpha_1,1), \]
\[ (P_{j2}|Z_{j11} = 1) \sim Beta(\alpha_2,1), \]

Need to enumerate all possibilities for any given D phenotypes

Chung et al., (PlosGen 2014)
2A. Integration of multiple phenotypes, genetics, and epigenomics
Integrative genetics analysis systematically ascertains mechanisms of complex GWAS traits

**Systems Genetics**

Joseph H. Nadeau and Almea M. Dudley

From studies with peas over 150 years ago, Gregor Mendel deduced the laws that govern the inheritance of traits in most organisms. The brilliance, but also the limitation, of Mendel’s work was its focus on single gene traits, such as stem color and plant height. However, phenotypic variation, including that which underlies health and disease in humans, often results from multiple interactions among numerous genetic and environmental factors. Systems genetics seeks to understand this complexity by integrating the questions and methods of systems biology with those of genetics to solve the fundamental problem of interpreting genotype and phenotype in complex traits and disease.

This global perspective is possible because of the technological infrastructure that derive from the Human Genome Project, which sequenced the genomic landscape of complex disorders

- **Characterization of multiple / polygenic landscape of complex disorders**
- **Goal:** Define what is “normal” range?
- **Genetic variation × Environmental “stress”**

**The State of Systems Genetics in 2017**

- Nilan S. Baliga,1,2,3,4,5,6,7,8,9 Johan L.M. Blijdorp,2,3,4,5,6,7,8,9 Jar D. Bok,2,3,4,5,6,7,8,9 Michael Bouttes,8,9,10 Nigel P.S. Crawford,7,8,9,10 Aimie M. Dudley,2,3,4,11 Charles R. Farber,2,3,4,11 Allan Jones,2,3,4,11 Allan I. Lowey,2,3,4,11 Alderson J. Lucie,2,3,4,11 M. Craig Mak,11,12 Joseph H. Nadeau,11,13 Marcus B. Noyes,11,13 Enrico Petretta,11,13 Nicholas T. Seyfried,11,13 Lars M. Steinmetz,11,13,14 Silbette G. Vanwetter,11,13,14

1. Characterization of systems-level genetic outcomes, not just one gene
2. Data integration across multi-modality
3. Networks of genes & genetic variations

Nadeau & Dudley, Science (2011)

Baliga et al., Cell (2017)
Ask systems genetics questions on the summary statistics of 47 UK Biobank traits

Mixed-model association for biobank-scale datasets

PheWAS by clustering

- Estimate genetic heritability based on our summary-based QTL modeling. Are they strongly genetic with respect to our reference genotype panel?
- Identify causal SNPs from the multivariate effects across 1,703 LD blocks of European population.
- Construct #LD × 47 Traits feature matrix based on #causal variants
- Perform clustering of LD blocks and traits
Estimation of multivariate genetic effects estimated from the univariate GWAS statistics

A (true) linear model for phenotypes:
\[ \mathbf{y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2 I). \]

How GWAS were carried out (SNP \( j \) by SNP):
\[ \mathbf{y} \sim \beta_j \mathbf{x}_j + \epsilon \]

Unknown / protected
- \( y_i \): pheno on individual \( i \)
- \( X_{ij} \): genotype, SNP \( j \), ind \( i \)

\( \beta_j \): univariate GWAS effect \( j \)
\( \tau_j \): standard error of effect \( j \)
\( \theta_j \): genetic effect of SNP \( j \) for all

Goal

• Why do we estimate \( \theta \)?
  1. Fine-mapping of causal variants (taking care of the LD); univariate stats are inflated by LD structure
  2. Heritability estimation: genetic variance = \( \theta^T R \theta \) where \( R \) is LD matrix
  3. Can enforce sparsity on \( \theta \)

Equivalent summary-based model:
\[ (\tau_j^{-1} \beta_j) \approx \mathcal{N}(\mathbf{R}\theta_j, (\sigma^2 R)_j). \]

Summary z-score Reference LD \( R \)
Traits are surprisingly heritable with respect to a linear model of common genetic variants

<table>
<thead>
<tr>
<th>Category</th>
<th>Trait</th>
<th>% heritability (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood</td>
<td>MEAN_PLATELET_VOL</td>
<td>56.8 (3.26)</td>
</tr>
<tr>
<td>Bone density</td>
<td>HEEL_TSCOREz</td>
<td>43.35 (4.34)</td>
</tr>
<tr>
<td>Body</td>
<td>HEIGHTz</td>
<td>54.01 (5.55)</td>
</tr>
<tr>
<td>BP</td>
<td>SYSTOLICadjMEDz</td>
<td>34.02 (4.1)</td>
</tr>
<tr>
<td>Disease</td>
<td>HYPOTHYROIDISM_SELF_REP</td>
<td>38.78 (2.88)</td>
</tr>
<tr>
<td>Impedance</td>
<td>BASAL_METABOLIC_RATEz</td>
<td>42.46 (6.57)</td>
</tr>
<tr>
<td>Lung</td>
<td>FEV1FVCzSMOKE</td>
<td>34.04 (3.74)</td>
</tr>
<tr>
<td>Mental</td>
<td>NEUROTICISM</td>
<td>25.22 (1.13)</td>
</tr>
<tr>
<td>other</td>
<td>MORNINGPERSON</td>
<td>23.84 (1.46)</td>
</tr>
<tr>
<td>Pigment</td>
<td>HAIR</td>
<td>53.77 (0.46)</td>
</tr>
<tr>
<td>Reproduction</td>
<td>MENARCHE_AGE</td>
<td>25.33 (3.95)</td>
</tr>
</tbody>
</table>

Subset of 47 UK Biobank Traits (Loh .. Price, 2018): heritability estimated by zQTL method.

- 20 – 57 % narrow-sense heritability
- Strongly agrees with Loh et al. (estimated by linear-mixed model in individual-level Data)
- Sparse effects of common SNPs can capture almost all variance.
A sparse set of strong common variants recapitulate almost all genetic variance

- No single SNP dominates amount of genetic variance explained
- Includes additive and constitutive genetic effects.
- Sparse effects of common SNPs can capture almost all heritability.
High heritability $\rightarrow$ “Omnigenic”? Not really.

- “Omnigenicity” implicates all SNPs contribute to the heritability of traits.
- Then, is GWAS just futile effort?
- If omnigenic, PheWAS might implicate everything is pleiotropic.
- We found clear LD-level enrichment of genetic variance if LD blocks are ranked by local genetic heritability (red curve) vs permuted selection LD blocks (gray area with 1k bootstrap).
- Locally polygenic, but not globally polygenic

Cumulative proportion of variance explained

Included LD blocks (# SNPs)
High heritability $\rightarrow$ “Omnigenic”? Not really.

High concentration of heritability for all 58 traits in Biobank Japan

High concentration of heritability for all the 47 traits in UK biobank
Clustering of causal LD blocks demonstrate the modularity of pleiotropic patterns

Clusters of LD blocks

Color = category; size = Pr{causal SNP}
2B. Factorization approaches for mining multiple phenotypes in population cohorts
A model-based approach to search for SNP-level pleiotropic patterns: factored GWAS

A vector of single trait z-score summary statistics

$$z_t \approx \mathcal{N}(R\theta_t, R)$$

Joint modeling of all the traits by factored regression

$$E[z_1, \ldots, z_T] \approx R \sum_k \lambda_k \Theta_k$$

$$\lambda_k = \text{Factor } k\text{-specific trait effect}$$

$$\Theta_k = \text{Factor } k\text{-specific genetic effect}$$

**Z**: univariate GWAS effect matrix (SNP x Trait) from UK Biobank

**R**: reference LD matrix (UK10K)

- $$\theta_{jk}$$: effect of SNP $$j$$ on factor $$k$$
- $$\lambda_{tk}$$: trait effect on trait $$k$$ of factor $$k$$
- Both $$\lambda$$ and $$\theta$$ are sparse with Bayesian prior
Factored GWAS approach reveals SNP-level pleiotropic locations and modularity

$$E[z_1, \ldots, z_T] \approx R \sum_k \lambda_k \theta_k$$

Can easily translate to genome-wide genetic correlation between traits
Factored GWAS reveals localized patterns of pleiotropy with explicit selection of causal SNPs

- Conventional PheWAS methods only highlight a single SNP-driven pleiotropy.
- SNP-by-SNP method is inflated by LD (not so exact)
- PheWAS tests for #SNPs x #traits (huge multiple hypothesis burden)
- Factored GWAS greatly reduces # parameters by factorization
Factored GWAS resolves 94 clusters of genomic regions based on pleiotropic patterns
Conditional probability sharing
pleiotropy cluster membership

Pr(row trait | significant in column trait)
Cluster#1: lung capacity, menarche, asthma, respiratory disease, cardiovascular disease, and BMI

- black dots: -log10 P GWAS of 6 pleiotropic traits
- yellow lines: average of the other unselected traits
- green bars: causal SNPs (fine-mapped) driving most of genetic variance in multiple traits
An example in cluster #5

Pigment hair, BMI, Height, T2D, Tanning, allergy, monocyte count, asthma, sunburn, respirator disease, lung capacity, BMI, bone density, lymphocyte count
# PheWAS: Multi-phenotype studies

1. **Motivation of phenome-wide association studies**
   - PheWAS-informed phenotyping, improved GWAS power, etc
   - Electronic health record (EHR) contain rich personalized information

2. **Modeling multiple phenotypes in GWAS + epigenomics**
   - Integration of multiple phenotypes in GWAS from Systems Genetics perspective (clustering approach)
   - Direct integration of multiple phenotypes by summary-based factored genetic model estimation

3. **Epigenomics of PheWAS**
   - Risk variants inference using epigenomic reference annotations
   - Using disease covariance to improve functional variants inference
   - Combining enrichment to improve causal pathway inference

4. **Meta-phenotype inference and imputation**
   - Models leveraging missing information and inferring missing mechanism
   - Modeling multimodal electronic health record data
   - Imputing missing EHR code and prioritizing patient disease risks
Epigenomes of 127 cell types

Roadmap Epigenome Consortium (Nat 2015)
Epigenomic annotations aid risk variants inference.
Multiple GWAS traits exhibit tissue-specific co-enrichments.
Inferring disease epigenomic-covariance

Li & Kellis (NAR 2016; bioRxiv)
Cell group epigenomic enrichments of 32 traits
Group-guided inferred cell-type-specific enrichments
Similar traits are correlated by epigenomic enrichments
Multi-trait inference improve functional enrichment

(Li & Kellis, NAR 2016)
Pathway-level co-enrichments

Psychiatric Genomics Consortium (NatNeu, 2014)
Combining enrichment improve causal pathway detection

**Table 2** Top results from integrative pathway analysis of three adult disorders

<table>
<thead>
<tr>
<th>Rank</th>
<th>BIP</th>
<th>MDD</th>
<th>SCZ</th>
<th>Combined $P$</th>
<th>q-value</th>
<th>Pathway ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0000</td>
<td>0.0592</td>
<td>0.0001</td>
<td>$5.75 \times 10^{-8}$</td>
<td>0.0003</td>
<td>GO:51568</td>
<td>Histone H3-K4 methylation</td>
</tr>
<tr>
<td>2</td>
<td>0.0004</td>
<td>0.0500</td>
<td>0.0006</td>
<td>$1.46 \times 10^{-5}$</td>
<td>0.0362</td>
<td>GO:16571</td>
<td>Histone methylation</td>
</tr>
<tr>
<td>3</td>
<td>0.0004</td>
<td>0.1462</td>
<td>0.0011</td>
<td>$4.73 \times 10^{-5}$</td>
<td>0.0414</td>
<td>GO:43414</td>
<td>Macromolecule methylation</td>
</tr>
<tr>
<td>4</td>
<td>0.0008</td>
<td>0.0630</td>
<td>0.0014</td>
<td>$5.10 \times 10^{-5}$</td>
<td>0.0414</td>
<td>GO:34968</td>
<td>Histone lysine methylation</td>
</tr>
<tr>
<td>5</td>
<td>0.4200</td>
<td>0.0001</td>
<td>0.0023</td>
<td>$5.58 \times 10^{-5}$</td>
<td>0.0414</td>
<td>GO:45216</td>
<td>Cell-cell junction organization</td>
</tr>
<tr>
<td>6</td>
<td>0.0001</td>
<td>0.0910</td>
<td>0.0064</td>
<td>$5.69 \times 10^{-5}$</td>
<td>0.0414</td>
<td>P000003</td>
<td>Alzheimer disease–amyloid secretase pathway</td>
</tr>
<tr>
<td>7</td>
<td>0.0007</td>
<td>0.0495</td>
<td>0.0024</td>
<td>$5.86 \times 10^{-5}$</td>
<td>0.0414</td>
<td>P04393</td>
<td>Ras pathway</td>
</tr>
<tr>
<td>8</td>
<td>0.3120</td>
<td>0.0000</td>
<td>0.1286</td>
<td>$7.12 \times 10^{-5}$</td>
<td>0.0422</td>
<td>GO:8601</td>
<td>Protein phosphatase type 2A regulator activity</td>
</tr>
<tr>
<td>9</td>
<td>0.8980</td>
<td>0.0001</td>
<td>0.0017</td>
<td>$7.83 \times 10^{-5}$</td>
<td>0.0422</td>
<td>GO:43297</td>
<td>Apical junction assembly</td>
</tr>
<tr>
<td>10</td>
<td>0.0013</td>
<td>0.0207</td>
<td>0.0055</td>
<td>$9.25 \times 10^{-5}$</td>
<td>0.0422</td>
<td>P00052</td>
<td>TGF-β signaling pathway</td>
</tr>
<tr>
<td>11</td>
<td>0.4890</td>
<td>0.0203</td>
<td>0.0000</td>
<td>$9.53 \times 10^{-5}$</td>
<td>0.0422</td>
<td>GO:14069</td>
<td>Postsynaptic density</td>
</tr>
<tr>
<td>12</td>
<td>0.0085</td>
<td>0.0009</td>
<td>0.0239</td>
<td>0.0001</td>
<td>0.0422</td>
<td>GO:32869</td>
<td>Cellular response to insulin stimulus</td>
</tr>
<tr>
<td>13</td>
<td>0.0188</td>
<td>0.0054</td>
<td>0.0022</td>
<td>0.0001</td>
<td>0.0450</td>
<td>P00010</td>
<td>B cell activation</td>
</tr>
<tr>
<td>14</td>
<td>0.0023</td>
<td>0.2988</td>
<td>0.0003</td>
<td>0.0001</td>
<td>0.0450</td>
<td>GO:8757</td>
<td>S-adenosylmethionine–dependent methyltransferase activity</td>
</tr>
<tr>
<td>15</td>
<td>0.0073</td>
<td>0.0080</td>
<td>0.0044</td>
<td>0.0001</td>
<td>0.0454</td>
<td>GO:23061</td>
<td>Signal release</td>
</tr>
<tr>
<td>16</td>
<td>0.4590</td>
<td>0.0000</td>
<td>0.0168</td>
<td>0.0002</td>
<td>0.0473</td>
<td>GO:34330</td>
<td>Cell junction organization</td>
</tr>
</tbody>
</table>

Psychiatric Genomics Consortium (NatNeu, 2014)
PheWAS: Multi-phenotype studies

1. Motivation of phenome-wide association studies
   • PheWAS-informed phenotyping, improved GWAS power, etc
   • Electronic health record (EHR) contain rich personalized information

2. Modeling multiple phenotypes in GWAS + epigenomics
   • Integration of multiple phenotypes in GWAS from Systems Genetics perspective (clustering approach)
   • Direct integration of multiple phenotypes by summary-based factored genetic model estimation

3. Epigenomics of PheWAS
   • Risk variants inference using epigenomic reference annotations
   • Using disease covariance to improve functional variants inference
   • Combining enrichment to improve causal pathway inference

4. Meta-phenotype inference and imputation
   • Models leveraging missing information and inferring missing mechanism
   • Modeling multimodal electronic health record data
   • Imputing missing EHR code and prioritizing patient disease risks
PheWAS without genetics information using EHR

- Genotype are often **not available** over large patient cohort
- Given the causal mediating phenotypes, **diseases of interest are conditionally independent of genotype**
- Deep phenotyping by electronic health records (EHR) technology as surrogates to mediating phenotypes is one promising direction

---

**Genetic**

**Mediating phenotypes**

**Disease phenotypes**
Medical recommendation system

- Trained Model
- EHR
- Recommended lab tests
- Recommended diagnoses
- Recommended Prescriptions
- Recommended Treatment Procedures
- Clinical recommendation
- Clinical Intelligent System®
Learning latent disease topics from diverse EHR data
Intuition behind phenotyping by matrix factorization

Patient similarity

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster j</th>
<th>Cluster K</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Phenotype similarity

<table>
<thead>
<tr>
<th>frequent urination</th>
<th>type 2 diabetes</th>
<th>high blood sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Cluster j</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cluster K</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Latent Dirichlet Allocation (LDA) (Blei et al, JMLR 2003)

Latent disease topics

\[ \theta_{jk} \]

Patient 1
Patient 2
Patient 3
Patient 4
Patient 5
Patient 6
Patient 7
Patient 8
Patient 9

\[ \Phi_{wk} \]

EHR features

Latent disease topics

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ z_{ijk} \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]

\[ \alpha \]

\[ \beta \]

\[ \theta_j \]

\[ \Phi_k \]

\[ x_{ij} \]

\[ \beta \]

\[ \theta_{jk} \]

\[ z_{ijk} \]

\[ E(x_{ij}^T) \]

1. marginalize latent topics dimension
2. marginalize EHR raw feature dimension
3. marginalize patient dimension

\[ x_{ij}^T \]
Grouping words by their topics

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants. An act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Challenge 1: EHR data are noisy and biased

- Unlike corpus documents in text mining, clinical notes are full of typos and arbitrary abbreviations
- Text mining commonly remove ‘stop words’, there no clearly defined ‘stop words’ in EHR
- Billing codes are often not meant to be disease-specific
Challenge 2: EHR data are extremely sparse

n(te

Patient fraction

icd

Patient fraction

x

x

x

x

Patient fraction

Patient fraction

Patient fraction

Patient fraction

Patient fraction

0.00 0.25 0.50 0.75 1.00

0.0 500 1500 1000

0 1000 2000 3000

0 2000 4000 6000

0 2500 5000 7500

0 200 400 600

0 500 1500 1000

0 1000 2000 3000
### Missing versus unrecorded EHR data

#### Lab tests

<table>
<thead>
<tr>
<th>Patient 1</th>
<th>Patient 2</th>
<th>Patient 3</th>
<th>Patient 4</th>
<th>Patient 5</th>
<th>Patient 6</th>
<th>Patient 7</th>
<th>Patient 8</th>
<th>Patient 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Missing</td>
<td>Missing</td>
<td>Missing</td>
<td>Missing</td>
<td>Missing</td>
<td>Missing</td>
<td>Missing</td>
<td>Missing</td>
<td>Missing</td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

#### Notes/Billing Codes/Drugs

<table>
<thead>
<tr>
<th>Recorded</th>
<th>Not recorded</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Recorded]</td>
<td>![Not recorded]</td>
</tr>
</tbody>
</table>

Note: The diagram shows the status of lab tests and notes/billing codes/drugs for different patients. The colors indicate the status: red for missing, blue for high, green for normal, and gray for low. Black squares represent recorded data, while white squares represent not recorded data.
Challenge 3: EHR data are not missing at random (NMAR)
NMAR is common in diverse domain: e.g., Distinct distribution for random and use-selected music ratings

(a) Yahoo! Random  
(b) Yahoo! User

Marlin & Zemel, ACM 2009
Missing mechanism concepts (Rubin & Little, 1987)

Data likelihood $L(\lambda) = p(y^{obs}, x|\lambda)$:

- **Missing completely at random (MCAR):**
  
  $$
  \mathcal{L}_{MCAR}(\lambda) = \int \int p(y^{obs}, z|\lambda) dz
  $$

  If data are MCAR, we can ignore the missing indicator $x$ in the data likelihood of $y$.

- **Missing at random (MAR):**
  
  $$
  \mathcal{L}_{MAR}(\lambda) = p(x|y^{obs}) \int_{ymis} \int_{z} p(y, z|\lambda) dz dy^{mis}
  $$

  If data are MAR, the missing indicator $x$ only depends on the observed results $y^o$.

- **Non-missing at random (NMAR):**
  
  $$
  \mathcal{L}_{NMAR}(\lambda) = \int_{ym} \int_{z} p(y, z|\lambda)p(x|y, z) dz dy^m
  $$

  If data are NMAR (i.e., $x$ always depend on $y$), we are stuck with the double integral.
Coding missing indicator as part of the data

Lin & Haug, JBI 2008
Using missing information improve medical predictions

Lin & Haug, JBI 2008
Jointly modeling lab test assignments and lab test results

Lab test decisions due to patients’ states
- Respiratory problem
  → C-REACTIVE PROTEIN (related)
- Kidney problem
  → Complete blood cell count (important for cancer)
  → UREA NITROGEN (related)

Observed and missing lab test results

<table>
<thead>
<tr>
<th></th>
<th>C-REAC PRO</th>
<th>CBC count</th>
<th>UERA NITRO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>HIGH</td>
<td>NORM</td>
<td>HIGH</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>NORM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Patient cluster

Missing mechanism

\[y_{ij} \rightarrow x_{ij} \rightarrow \Phi \]

Difficult to model because \( f(x|y,\Phi) \) is unknown

Test result weights \( \lambda \)

Patient meta-phenotype mixture membership \( \theta \)

Conditional hyper-independence latent variable parameters \( \phi \)
 Modeling missing mechanism in lab test is challenging

Lab test decisions due to patients’ states

- Respiratory problem → C-REACTIVE PROTEIN (related)
- Kidney problem → UREA NITROGEN (related)

Observed and missing lab test results

<table>
<thead>
<tr>
<th></th>
<th>C-REAC PRO</th>
<th>CBC count</th>
<th>UERA NITRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>... HIGH</td>
<td>NORM</td>
<td>?</td>
</tr>
<tr>
<td>Missing</td>
<td>... ?</td>
<td>NORM</td>
<td>HIGH</td>
</tr>
</tbody>
</table>

Patient cluster

- Test result weights $\lambda$
- Missing mechanism $y_{ij} \rightarrow x_{ij} \rightarrow \phi$
- Difficult to model because $f(x|y, \Phi)$ is unknown

Patient meta-phenotype mixture membership

- Latent variable $\theta$
- Hyper-parameters $\eta$
- Lab presence $x_{ij}$
- Lab results $y_{ij}$
- Lab frequency $\phi$
- Simplified by conditional independence

Modeling missing mechanism in lab test is challenging
Jointly modeling lab missing indicators and lab test results

\[
\ln q(z_{lj}^{QA} = k) \propto E_{q}[\ln(n_{jk}^{(-l)})]
\]

patient \(j\)'s score for topic \(k\)

\[
E_{q}[\ln p(y_{lj}|z_{lj} = k, \Theta_k)]
\]

score likelihood for question \(l\)

\[
E_{q}[\ln p(r_{lj}|z_{lj} = k, \Theta_k)]
\]

missing indicator likelihood

\[
E_{q(z_k)}[\ln p(r_{lj}|z_{lj} = k, \Theta)]_k =
\]

\[
E_{q}[\ln(\frac{p_{lk}^{-}(l,j)}{\nu l' p_{lk}^{-}(l,j) + q_{lk}^{-}(l,j)})^{r_{lj}}]
\]

\[
E_{q}[\ln(\frac{q_{lk}^{-}(l,j)}{\nu l' p_{lk}^{-}(l,j) + q_{lk}^{-}(l,j)})^{1-r_{lj}}]
\]

- \(p_{lk} = \sum j [z_{lj} = k][r \neq 1]\) is the number of patients assigned to topic \(k\) when lab \(l\) is observed

- \(q_{lk} = \sum j [z_{lj} = k][r = 0]\) is the number of patients assigned to topic \(k\) when lab \(l\) is not observed
Heterogeneous EHR data are related for the same patients.

**Related binary EHR phenotypes for the same patients**

<table>
<thead>
<tr>
<th>Chronic kidney disease</th>
<th>Pneumonia</th>
<th>Influenza</th>
<th>Asthma</th>
<th>Respiratory problem</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NA</td>
<td>NA</td>
<td>✔</td>
<td><img src="image1" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td><img src="image2" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td><img src="image3" alt="Image" /></td>
</tr>
</tbody>
</table>

**Lab tests of the same patients**

<table>
<thead>
<tr>
<th>CBC count</th>
<th>C-REAC PRO</th>
<th>UERA NITRO</th>
<th>Respiratory problem</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
</tr>
</tbody>
</table>

**Joint model of related data types**

Patient meta-phenotype mixture

- **phenotype frequency**
- **lab results frequency**
- **Binary clinical features**
- **lab presence**
- **lab results**
A. Multi-view learning of EHR data

B. Disease regression

MixEHR model

\[
\ln q(z_{lj}^{\text{Gene}} = k) \propto E_{q(z_{j}^{(-l)})} \left[ \ln(n_{jk}^{(-l)}) \right] E_{q(\Theta_k)} \left[ \ln p(y_{lj} | z_{lj} = k, \Theta_k) \right]
\]

\[
n_{jk}^{(-l)} = \sum_{I \neq l} [z_{ij}^{\text{Gene}} = k] + \sum_{l} [z_{lj}^{\text{ICD}} = k] + \sum_{l} [z_{lj}^{\text{Lab}} = k] + \ldots
\]
Meta-phenotype learning from multimodal patient-EHR

1. marginalize latent meta-phenotype dimension

2. marginalize EHR phenotype or lab test dimension

3. marginalize patient dimension

\[ E_q[\gamma_{ijk}^{(t)}] \]

\[ E_q[X_{ij}^{(t)}] \]

\[ E_q[Y_{lj}^{(t)}] \]

\[ E_q[n_{w.k}^{(t)}] \]

\[ E_q[m_{l.kv}] \]
MixEHR-NMAR model improves imputations

**likelihood(NMAR) > likelihood(MAR)**

NMAR model improves whereas MAR model overfits the data

NMAR model agrees with the true model

MixEHR outperforms the state-of-the-art model

- We simulated NMAR data from MixEHR trained on the real data;
- Simulated the missing indicators from the topic-specific frequency;
- We evaluated 4 models: (1) MixEHR nmar modeling NMAR using Questionnaire Only; (2) MixEHR nmar using All Data; (3) & (4) MixEHR mar assuming missing at random (MAR) using Questionnaire Only and All Data.
MixEHR learns meaningful disease topic embeddings

a. Top clinical features from select disease topics

<table>
<thead>
<tr>
<th>Disease topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cirrhosis</td>
</tr>
<tr>
<td>PulmEm</td>
</tr>
<tr>
<td>HIV</td>
</tr>
<tr>
<td>Diabetes</td>
</tr>
<tr>
<td>Leukemia</td>
</tr>
<tr>
<td>Dementia</td>
</tr>
<tr>
<td>Psycho</td>
</tr>
</tbody>
</table>

Select topics learned from MIMIC-III EHRdata
MixEHR learns meaningful disease topic embeddings

a. Topics correlated with age

<table>
<thead>
<tr>
<th>Disease topics correlated with age</th>
<th>Correlation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart failure</td>
<td>Inser temp pacemaker sys (3778)</td>
</tr>
<tr>
<td></td>
<td>Rt heart cardiac cath (3721)</td>
</tr>
<tr>
<td></td>
<td>Heart failure (APR, 1944)</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>Oth cardiac mon output (8968)</td>
</tr>
<tr>
<td>Dementia</td>
<td>Circulatory disorders with acute myocardial infarction &amp; major complic</td>
</tr>
<tr>
<td>Neoprene</td>
<td>Esophagoscopy nec (4223)</td>
</tr>
<tr>
<td>Neonate</td>
<td>Major respiratory infections &amp; inflammations (APR, 1374)</td>
</tr>
<tr>
<td>Preterm birth</td>
<td>Dementia w/o behav dist (29410)</td>
</tr>
</tbody>
</table>

b. Topics correlated with psychiatric traits

Schizophrenia-associated codes

<table>
<thead>
<tr>
<th>Correlation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
</tr>
<tr>
<td>notes</td>
</tr>
<tr>
<td>icd_cm</td>
</tr>
<tr>
<td>icd_cpt</td>
</tr>
<tr>
<td>presc</td>
</tr>
<tr>
<td>drg</td>
</tr>
</tbody>
</table>

PTSD-associated codes

<table>
<thead>
<tr>
<th>Correlation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
</tr>
<tr>
<td>notes</td>
</tr>
<tr>
<td>icd_cm</td>
</tr>
<tr>
<td>icd_cpt</td>
</tr>
<tr>
<td>presc</td>
</tr>
<tr>
<td>drg</td>
</tr>
</tbody>
</table>
MixEHR prioritizes patients by disease mixture

a. Top patients in select disease topics

b. Top EHR features

- Bone marrow biopsy (4131)
- Inject ca chemother nec (9925)
- Acute leukemia (APR.8004)
- Major o.r. procedures for lymphatic/hematopoietic/other neoplasia (APR.6004)
- Immunotherapy as anti (9925)
- Adv eff anti leukemic (9931)
- Other lymph node (9930)
- Acute leukemia w/o major o.r. procedure w moc (MS.834)
- Lymphoma & non-acute leukemia (APR.9716)
- Closed b.r. skin/subcut lns (9811)
- Pulm embol/infarct nec (41519)
- Other pleural ind (3409)
- Inject into thorax cav (3490)
- Vena cav angiocardiogram (8851)
- Pulmonary embolism (ICD.A78)
- Ectopic
- Pulmonary embolism w/o moc (MS.179)
- Anemia, anemia, anemia, anemia (45341)
- Pulmonary embolism (APR.1433)
- Enoxaparin sodium, enoxaparin sodium (92730, 75062803)
- Endosc destruc esoph lns (4233)
- Alcohol cirrhosis liver (5712)
- Anemia inject-spin canal (391)
- Hepatic coma & other major acute liver disorders (APR.2794)
- Disorders of liver except mali, etc, hep a w moc (MS.441)
- Cirrhosis of liver nos (5715)
- Closed liver biopsy (5011)
- Portal hypertension (723)
- Alcoholic liver disease (APR.2804)

M31

- Endosc destruc esoph lns (4233)
- Portal hypertension (723)
- Hepatic coma & other major acute liver disorders (APR.2794)
- Alcohol cirrhosis liver (5712)

M35

- Pulm embol/infarct nec (41519)
- Inject into thorax cav (3492)
- Pulmonary embolism (ICD.A78)
- Ectopic
- Pulmonary embolism w/o moc (MS.179)
- Anemia, anemia, anemia, anemia (45341)
- Pulmonary embolism (APR.1433)
- Enoxaparin sodium, enoxaparin sodium (92730, 75062803)
- Endosc destruc esoph lns (4233)
- Alcohol cirrhosis liver (5712)
- Anemia inject-spin canal (391)
- Hepatic coma & other major acute liver disorders (APR.2794)
- Disorders of liver except mali, etc, hep a w moc (MS.441)
- Cirrhosis of liver nos (5715)
- Closed liver biopsy (5011)
- Portal hypertension (723)
- Alcoholic liver disease (APR.2804)

M50

- Inject ca chemother nec (9925)
- Bone marrow biopsy (4131)
MixEHR accurately imputes missing EHR code

Top 1215 patients (sorted by M31 membership "liver cirrhosis")

EHR codes
- Observed EHR codes
- Imputed EHR codes

AUROC per EHR codes

Imputation accuracy
(5-fold cross-validation AUROC)
MixEHR accurately predicts future mortality events

a. mortality prediction

![Graph showing true positive rate vs. false positive rate for different models.]

- mixehr_K50 (85%)
- mixehr_K75 (79.8%)
- LDA_K50 (75.9%)
- LDA_K75 (74.6%)

b. mortality-related meta-phenotypes

![Table showing meta-phenotypes with associated probabilities and categories.]

<table>
<thead>
<tr>
<th>Meta-phenotypes</th>
<th>Prob.</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>

*Notes: Arrest, Anoxic brain damage, Morphine Sulfate_016467_74860d11, Acute necrosis of liver, Closed b/skin/subq tiss, Exploratory laparotomy, Septicemia w/mechanical ventilator, Respiratory system diagnosis with ventilator support, Ampicillin Sodium_008937_63323038810, Oxygen enrichment NEC.*
Summary

1. PheWAS elucidates phenotypic network by linking genetics to co-associated traits

2. Modeling multiple traits reveals genetics correlation between traits, helps elucidate co-morbidities, and improve causal signals detections at SNP and pathway levels

3. EHR data is extremely rich but modeling is challenging because the data are high dimensional, highly sparse and non-missing at random

4. Machine learning methods especially generative models hold great promise to learn the compressed latent dimensions of the high-dimension EHR data and impute missing EHR data to reveal underlying disease network
PheWAS: Multi-phenotype studies

1. **Motivation of phenome-wide association studies**
   - PheWAS-informed phenotyping, improved GWAS power, etc
   - Electronic health record (EHR) contain rich personalized information

2. **Modeling multiple phenotypes in GWAS + epigenomics**
   - Integration of multiple phenotypes in GWAS from Systems Genetics perspective (clustering approach)
   - Direct integration of multiple phenotypes by summary-based factored genetic model estimation

3. **Epigenomics of PheWAS**
   - Risk variants inference using epigenomic reference annotations
   - Using disease covariance to improve functional variants inference
   - Combining enrichment to improve causal pathway inference

4. **Meta-phenotype inference and imputation**
   - Models leveraging missing information and inferring missing mechanism
   - Modeling multimodal electronic health record data
   - Imputing missing EHR code and prioritizing patient disease risks