From GWAS to PRS

• Recall: GWAS uses regression-based tests
Outcome: case/control status (P = prob of being a case)
Predictors: SNP genotype ($x_1$), covariates (age, sex, population principal components)

$$\logit(P) = \ln \left( \frac{P}{1-P} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

$\beta_1$ represents SNP effect size
Coronary Artery Disease (N ~ 185,000): Nikpay et al., Nat Genet 2015
Visualizing thresholds for PRS derivation

PRSes can be derived at various significance thresholds, with pruning to remove correlated SNPs in LD.

Coronary Artery Disease (N ~ 185,000): Nikpay et al., Nat Genet 2015
Background: GWAS and Polygenic Risk Scores

- Once we have results from a GWAS, and a threshold for which SNPs to include, we can compute the PRS in a separate sample by applying the same coefficients (weights) to genotypes in that sample.

\[
\text{PRS} \sim \sum_{i=1}^{n} \beta_{1i} x_{1i}
\]

\[n = \text{number of SNPs in the PRS}\]

- So need only the summary results from the GWAS, e.g. betas (odds ratios)
Polygenic Risk Scores

Classic example: Schizophrenia (SZ) and Bipolar Disorder (BP)

Purcell et al. (2009), Nature 460; 748-752
Cautions regarding GWAS

- Quality control is essential
- Replication is key
- Expect signals to be supported by multiple correlated SNPs
- Hidden population stratification can lead to spurious signals
- When combining or meta-analyzing samples: pay attention to strand
Imputation

- Method(s) for "in silico genotyping" or predicting/estimating genotype status at un-typed variants

- Obtain probabilities of having the 3 genotypes:
  \[ p_{11}, p_{12}, p_{22} \]

- Benefits:
  - Improved detection of a signal at an imputed SNP
  - Ease of combining data/results across studies with differing SNP coverage.
Imputation overview

a. Data for samples. 
? = untyped SNPs

b. Testing typed SNPs

c. Phase each sample; model haplos as mosaic of those in reference

d. Reference haplotypes

e. Impute alleles for the samples (orange)

f. Testing typed and imputed SNPs

(adapted from Marchini & Howie, Nat Rev Genet 2010)
Imputation

Reference panel sources:
- HapMap
- 1000 Genomes
- Haplotype Reference Consortium (HRC) – primarily European-ancestry
- Others

“Cosmopolitan” reference panel usually recommended: combine populations.
- Represents the widest range of haplotypes
Feasibility of identifying risk variants, by allele frequency and effect size (odds ratio)

Manolio et al., 2009
What is different about Human Genetics (recall from Cristina Strong's lectures)

• Can study complex behaviors and cognition, neurgenetics
• Extensive sequence variation leads to common/complex disease
  1. Common disease, common variant hypothesis
  2. Large # of small-effect variants
  3. Large # of large-effect rare variants
  4. Combo of genotypic, environmental, epigenetic interactions
• Imprinting – uniquely mammalian
• Trinucleotide repeat diseases – "anticipation"

Greg Gibson, Nature Rev Genet 2012

Addition:

• Potential for the most direct clinical impact
Testing for interactions

• GxG (epistasis), GxE

• Does the effect of 1\textsuperscript{st} variable on the outcome differ for different values of the 2\textsuperscript{nd} variable?

• Recall Heather Lawson's lectures
GxE interaction example – not in humans

Cooper and Zubek, 1958

2 rat strains: "maze dull" and "maze bright" - selectively bred based on maze performance.

Then, two environments:
  - normal lab rearing
  - "enriched" environment (e.g. playthings, open space)
Animal models

Cooper and Zubek, 1958: mean number of learning errors

Rowe, 2003

as presented in

Rowe, 2003
Other possible outcomes

Another way to get a GxE interaction: cross-over
Possible outcomes

Here's what we might see if there were main effects only (*no GxE interaction*)
Possible outcomes

another possibility: if *no* GxE interaction AND no main effect of environment
Testing for interactions

• Traditional approach is to use a “product term” in the regression model
• Example (case-control logistic regression analysis)
  Let \( P = \) probability of being a case
  Let \( x_1 = \) SNP genotype (additive coding)
  \( x_2 = \) sex (0=male, 1=female)
  \[
  \ln \left( \frac{P}{1 - P} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2
  \]

  To test for interaction between \( x_1 \) and \( x_2 \), i.e. does the SNP effect differ in men vs women:
  \[
  \ln \left( \frac{P}{1 - P} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2)
  \]

  Is there significant evidence that \( \beta_3 \) is different from 0?
Example: $x_1 = \text{SNP genotype, } x_2 = \text{sex}$

- Full model with interaction:

$$\left( \frac{P}{1 - P} \right) = e^{[\alpha + \beta_2 x_2] + \beta_1 x_1 + \beta_3 x_1 x_2}$$

- Odds Ratio for SNP when $x_2 = 0$ (male)

$$OR_{\text{male}} = \frac{P_1/(1 - P_1)}{P_0/(1 - P_0)} = \frac{e^{[\ldots] + \beta_1 x_1 + \beta_3 x_1 x_2} \bigg|_{x_1=0,x_2=0}}{e^{[\ldots] + \beta_1 + 0} \bigg|_{x_1=0,x_2=0}} = \frac{e^{[\ldots] + \beta_1 + 0}}{e^{[\ldots] + \beta_1 + 0}} = e^{\beta_1}$$

- Odds Ratio for SNP when $x_2 = 1$ (female)

$$OR_{\text{female}} = \frac{P_1/(1 - P_1)}{P_0/(1 - P_0)} = \frac{e^{[\ldots] + \beta_1 x_1 + \beta_3 x_1 x_2} \bigg|_{x_1=1,x_2=1}}{e^{[\ldots] + \beta_1 + \beta_3} \bigg|_{x_1=0,x_2=1}} = \frac{e^{[\ldots] + \beta_1 + \beta_3}}{e^{[\ldots]}} = e^{\beta_1+\beta_3}$$

Non-zero $\beta_3$ allows the two ORs to differ
Executive Summary

“Personalized medicine” refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not.
Moving from GWAS to post-GWAS: Genetics, clinical impact and personalized medicine

- 2011: Green and Guyer (NHGRI), Nature 2011: “Base pairs to bedside” – not just “bench to bedside”

- 2015: President Obama announced $215 million Precision Medicine Initiative
  - Objectives:
    - Cancer treatment
    - Voluntary national research cohort (compare with UK research facilitated by nationalized healthcare)
    - Privacy protection
    - Modernizing regulatory landscape
    - Public-private partnerships
Target Timeline Towards Precision Medicine

- Understanding the structure of genomes
- Understanding the biology of genomes
- Understanding the biology of disease
- Advancing the science of medicine
- Improving the effectiveness of healthcare

1990-2003
Human Genome Project

2004-2010

2011-2020

Beyond 2020

Major goal in human genetics: personalized medicine

– Can polygenic scores from GWAS identify clinically relevant elevations in disease risk?

– Can polygenic scores identify high-risk individuals that standard risk prediction might miss?

– Can SNPs or polygenic scores identify which individuals are most likely to benefit from a given treatment?
Khera et al., 2019

- 5 Common diseases: Coronary artery disease (CAD)
  Atrial fibrillation (Afib)
  Type 2 diabetes (T2D)
  Inflammatory bowel disease (IBD)
  Breast Cancer (BC)

- UK Biobank data

- Used genome-wide polygenic scores (GPS) to identify high-GPS subgroups at > 3-fold increased risk of disease.

- These subgroups were a meaningful proportion of the population (e.g. 8% for CAD)

- They propose it may be time to include GPS in clinical care

What questions might you ask here?
Questions to ask:

• What is the baseline prevalence of the diseases studied?

• What population samples were used for each step
  – GPS derivation
  – Validation
  – identification of thresholds for "high risk GPS" individuals?

• How did they define the GPS (many methods exist that can be implemented multiple ways)?

• How many variants needed for a "good" GPS?
Results: Coronary Artery Disease

Scaled to mean 0 and SD 1
Compare PRS to rare, monogenic mutation:

- High PRS group
  - over 3-fold increased risk of coronary artery disease
  - 8% of the population

- Compare to rare, monogenic mutation in familial hypercholesterolemia
  - increases cholesterol
  - 3-fold increase in heart attack risk
  - 0.5% of the population

Khera et al., 2019
Coronary Artery Disease

![Box plot showing polygenic score percentile for Control and Case groups in CAD](image)
Coronary Artery Disease

The key figure:
If observed the red dots – no predictive power

The key type of figure:
Khera et al, summary for Coronary Artery Disease (CAD)

• 8% of population had "high" GPS with > 3-fold increased risk compared to "low" GPS

• Compared to familial hypercholesterolemia mutations:
  – At comparable or greater risk
  – But 20-fold fewer individuals have such mutations

• This 8% subset could **not** be readily identified using conventional risk factors
  – Only 20% had high cholesterol
  – 44% had family history
  – (Q: what % had "at least 1" conventional risk factors?)
Pharmacologic Treatment

Genomics can guide personalized medicine

Smoking cessation pharmacotherapy as an example
## Slow decline in U.S. smoking rates (CDC data)

<table>
<thead>
<tr>
<th>Year</th>
<th>% current smokers, of US adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004-2006</td>
<td>21%</td>
</tr>
<tr>
<td>2007</td>
<td>20%</td>
</tr>
<tr>
<td>2008-2009</td>
<td>21%</td>
</tr>
<tr>
<td>2010</td>
<td>19.3%</td>
</tr>
<tr>
<td>2011</td>
<td>19.0%</td>
</tr>
<tr>
<td>2012</td>
<td>18.1%</td>
</tr>
<tr>
<td>2013</td>
<td>17.8%</td>
</tr>
<tr>
<td>2014</td>
<td>16.8%</td>
</tr>
<tr>
<td>2015</td>
<td>15.1%</td>
</tr>
<tr>
<td>2016</td>
<td>15.5%</td>
</tr>
<tr>
<td>2020</td>
<td>Healthy people target: 12%</td>
</tr>
</tbody>
</table>
CHRNA5 predicts cessation success and response to medication

Pharmacogenetics and Genomics:
February 2013 - Volume 23 - Issue 2 - p 94–103

Nicotinic acetylcholine receptor variation and response to smoking cessation therapies

Bergen, Andrew W.; Javitz, Harold S.; Krasnow, Ruth; Nishita, Denise; Michel, Martha; Conti, David V.; Liu, Jinghua; Lee, Won; Edlund, Christopher K.; Hall, Sharon; Kwok, Pui-Yan; Benowitz, Neal L.; Baker, Timothy B.; Tyndale, Rachel F.; Lerman, Caryn; Swan, Gary E.
Study Design

University of Wisconsin TTURC
- N=1073, European Ancestry
- All subjects received intensive behavioral counseling
- Pharmacotherapy arms (NRT, bupropion, combo) and 1 placebo arm
- Cessation Abstinence at 60 days
  Time to relapse over 60 days

CHRNA5 haplotype
- rs16969968
  Functional amino acid change in CHRNA5
- rs680244
  CHRNA5 mRNA levels in brain and lung
- Haplotype of 2 variants
  - H1 (GC, 20.8%) (low risk for dependence)
  - H2 (GT, 43.7%)
  - H3 (AC, 35.5%), high risk for dependence

"CHRNA5 Predicts Cessation & Response to Medication"

Smokers with

**CHRNA5**

low risk

Smokers with

**CHRNA5**

high risk

Abstinence

This represents a GxE interaction!

Number Needed to Treat (NNT) Varies by Genetic Status

NNT > 1000

NNT = 4

NNT: Number of patients to treat for 1 to benefit

Abstinence

-placebo + counseling

- medication + counseling

H1: LOW RISK  
H2

H3: HIGH RISK

In smokers with high genetic risk, NNT for cessation medication compares very well to current evidence-based clinical treatments.

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Treatment</th>
<th>Number needed to treat (NNT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duodenal Ulcer</td>
<td>H Pylori Antibiotics</td>
<td>2</td>
</tr>
<tr>
<td>COPD</td>
<td>Anticholinergic</td>
<td>16</td>
</tr>
<tr>
<td>Stroke</td>
<td>Alteplase</td>
<td>4.5-21.4</td>
</tr>
<tr>
<td>Hypercholesterolemia</td>
<td>Statins</td>
<td>63-256</td>
</tr>
<tr>
<td>Nicotine Dependence</td>
<td>NRT</td>
<td>4 in CHRNA5 high risk gp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000 in CHRNA5 low risk gp</td>
</tr>
</tbody>
</table>

UpToDate®, 2013
Summary

• Human gene mapping for complex diseases/traits:
  – Linkage in families; LOD scores
  – Association; linear and logistic regression; effect size; GWAS
• Linkage disequilibrium (LD)
  – Computing measures of LD
  – Importance of LD in study design and interpretation
  – Genetic imputation ("in silico genotyping")
• Consortium science / data sharing / meta-analysis
• Polygenic risk scores
• GxG and GxE interactions
  – Visualizing interactions vs main effects
  – Testing with a product term
• Genomics and “personalized medicine”
  – Polygenic risk scores
  – Genotype-by-treatment interactions