Human Genetics and Gene Mapping of Complex Traits

Advanced Genetics
Human Genetics Series
Thursday 4/04/24
Nancy L. Saccone
Dept of Genetics
nlims@genetics.wustl.edu
Announcements

• Remaining Tu/Th lectures: Th 4/4, Tu 4/9, Th 4/11, Tu 4/30.
• Small groups meet this week and next
• Mon 4/15 will be the last MTE-led proposal work session, in Connor
• Study sections to review proposals: In person, May 6, 7, 8 (M, Tu, W) from 1pm to ~3:30pm. Tues meeting may shift to 2-4:30pm. Please be available for all these times, and you will be given 1 section to attend.

• Announcements from last time:
  – Aims presentation in small groups next week. Suggestion: prepare 2 slides, 1 with background info, 1 with hypothesis/question plus Aims.
Outline and learning objectives

• Linkage disequilibrium (LD)
  • Practical implications, e.g. interpreting GWAS results
  • Measures of LD, how to compute
  • Sources of LD

• GWAS to "Post-GWAS"
  • Meta-analyses, consortia
  • Diverse populations
  • Model organism follow-up (getting from statistical signal to biology)
  • Polygenic risk scores (PRS)
  • Precision/personalized medicine
GWASes rely on linkage disequilibrium (LD) to "tag" variation, or to impute ungenotyped variants. Thus GWAS results must be interpreted in the context of LD: The signal SNP may not be a biologically causal variant.
Haplotype “tagging”

The International HapMap Project, Nature 2003
Interpretation of GWAS results must account for LD

- Suppose a SNP is significantly associated with a disease
- Other SNPs correlated (high $r^2$) with that SNP are additional, potentially “causative” variants

Example: GWAS of fetal hemoglobin detected BCL11A

Some (not all!) of these statistically significant SNPs are highly correlated with the lead SNP rs11886868 (intrinsic C $>$ T)

Uda et al.(2008), Genome-wide association study shows BCL11A... , PNAS 105:1620-1625
Interpretation of GWAS results must account for LD
Strongest signals are in intron 2 of BCL11A, point to group of SNPs

Uda et al. (2008), Genome-wide association study shows BCL11A…, PNAS 105:1620-1625
Displaying GWAS results: zooming in

Linkage Disequilibrium

Circle: genotyped SNP
X: imputed SNP

Non-daily vs daily smoking
Saccone et al., Nicotine and Tobacco Research 2016
• A typical GWAS signal consists of multiple SNPs due to LD

• Challenging to identify biologically involved or “causal” SNPs among correlated SNPs

• Multiple distinct statistical signals can exist under a peak

• "Conditional analysis" is used to identify distinct signals:
  • In the regression, include signal SNP as a covariate, observe if additional SNPs remain significant in the model
LD and Human Sequence Variation

ancestral chromosome

present day chromosomes:

alleles on the preserved "ancestral background" tend to be in linkage disequilibrium (LD)
Linkage Disequilibrium

- "Non-random" associations between alleles at different loci
- How to formally measure LD between alleles at 2 loci?
To measure LD between alleles at 2 biallelic loci

Locus A  Locus B
A₁, A₂  B₁, B₂

Given 2N haplotypes:
Haplotype freq for AᵢBⱼ is

\[ h_{ij} = \frac{n_{ij}}{2N} \]

Compare \( h_{ij} \) to the frequency expected under no association:

\[ p_{A₁} p_{B₁} = \left( \frac{n_{11} + n_{12}}{2N} \right) \left( \frac{n_{11} + n_{21}}{2N} \right) \]

Define the disequilibrium coefficient:

\[ D = h_{11} - p_{A₁} p_{B₁} \]

Choice of allele labeling may affect sign but not absolute value of D.
Common LD measures

Disequilibrium coefficient:
\[ D = h_{11} - p_{A1} p_{B1} \]

Normalized disequilibrium coefficient:
\[ D' = D / |D|_{\text{max}}, \text{ where} \]
\[
|D|_{\text{max}} = \begin{cases} 
\min(p_{A1}p_{B2}, p_{A2}p_{B1}) & \text{if } D > 0 \\
\min(p_{A1}p_{B1}, p_{A2}p_{B2}) & \text{if } D < 0 
\end{cases}
\]

Range of D' is \([-1,1]\)

Squared correlation coefficient:
\[ r^2 = D^2 / \left( p_{A1}p_{A2}p_{B1}p_{B2} \right) \]

For a given SNV pair, attainable range depends on their allele freqs

Chi-square statistic for a 2x2 table with 1 d.f.
\[ \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}} \]
Measuring LD

Example:

Only observe 2 haplotypes: \( A_1B_1 \) and \( A_2B_2 \)

\[
\begin{array}{c|c|c}
A_1 & B_1 & B_2 \\
50 & 0 & 50 \\
0 & 50 & 50 \\
50 & 50 & 100 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
A_1 & B_1 & B_2 \\
25 & 25 & 50 \\
25 & 25 & 50 \\
50 & 50 & 100 \\
\end{array}
\]

\[
D = h_{11} - p_{A_1}p_{B_1} = (0.5) - (0.5)(0.5) = 0.5 - 0.25 = 0.25
\]

\[
D_{\text{max}} = \min(p_{A_1}p_{B_2}, p_{A_2}p_{B_1}) = \min(0.25, 0.25) = 0.25
\]

\[
|D'| = 1
\]

\[
r^2 = 1
\]
Measuring LD

Example:

Only observe 2 haplotypes: A₁B₁ and A₂B₂

<table>
<thead>
<tr>
<th>A₁</th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A₂</th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

To measure significance: \( \chi^2 \) (1 df):

\[
\chi^2 = \frac{(50 - 25)^2}{25} + \frac{(0 - 25)^2}{25} + \frac{(0 - 25)^2}{25} + \frac{(50 - 25)^2}{25}
\]

Chi-sqr = 100, p-value very small
LD measures

Another useful example:

<table>
<thead>
<tr>
<th></th>
<th>$B_1$</th>
<th>$B_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>$A_2$</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>

$D' = (0.1 - (0.1)(0.1)) / (0.09) = 1$

$\chi^2 = 100$, p-value $\sim 0.0$

$r^2 = 1$

$D' = (0 - (0.1)(0.1)) / (0.01) = -1$

$\chi^2 = 1.23$, p-value $= 0.27$

$r^2 = 0.012$
LD measures

$|D'|$ is 1 when the alleles of the two markers are as correlated as they can be, given the allele frequencies of the co-occurring alleles.

The range of $r^2$ depends on the 2 marker allele frequencies.

$r^2$ equals 1 if and only if

1) the minor allele frequencies (MAFs) at the two loci match AND

2) the minor alleles always co-occur

$D'$ : useful for identifying regions of reduced recombination.

$r^2$ : useful for identifying markers that are good predictors of allelic status at other markers.
LD across TCF7L2 in CEU HapMap.

Grant et al., Nat Genet 2006, Figure 1
Where does LD come from?

Potential sources of LD:

1. Genetic linkage between loci
2. Random drift
3. Founder effect
4. Mutation
5. Selection
6. Population admixture / stratification
LD is not a simple monotonic function of distance
Dawson et al., Nature 2002

Panel (a): D' by distance between markers
Panel (b): r^2 by distance between markers
LD patterns can vary by population: differing population history, allele frequencies

Uda et al. (2008), Genome-wide association study shows BCL11A... PNAS 105:1620-1625
Where does LD come from?

Potential sources of LD:

1. Genetic linkage between loci
2. Random drift
3. Founder effect
4. Mutation
5. Selection
6. Population admixture / stratification
An example of LD association due to admixture/stratification:

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>A₂</td>
<td>81</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

Chi-square = 0

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>B₁</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>B₂</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

Chi-square = 0

**Combined**

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>34</td>
<td>26</td>
</tr>
<tr>
<td>A₂</td>
<td>106</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>60</td>
</tr>
</tbody>
</table>

Chi-square = 7.26

P-value = 0.007

“Simpsons paradox”
What if the rows denoted case/control status – would see spurious association

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case</strong></td>
<td>9</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>81</td>
<td>9</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

**population 1**

**population 2**

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case</strong></td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

chi-square = 0

Chi-square = 0

**Combined**

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case</strong></td>
<td>34</td>
<td>26</td>
<td>60</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>106</td>
<td>34</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>60</td>
<td>200</td>
</tr>
</tbody>
</table>

**Combined**

Chi-square = 7.26

p-value = 0.007
Spurious association between a locus and disease can occur if there are two (unknown) subpopulations.

Exaggerated example: if an allele occurs only in stratum 1, then any trait with very different prevalence in stratum 1 versus stratum 2 could appear to be associated with this SNP.

Subpop1 (i.e. stratum 1): \textbf{allele A1} \quad \text{Subpop2 (i.e. stratum 2): alleles A1, A2}

Unshaded = affected cases. Unaffected controls are more prevalent in stratum 1. Here allele A_1 appears to have much higher freq in controls compared to cases.
How to avoid spurious association from population stratification?

1. Match for consistent ancestry background in the case and control samples

2. Check to make sure there is no underlying stratification
   - Use the GWAS data for this analysis

3. If there is, account for this in the analysis

Software:

- EIGENSTRAT software (principal components) (Price et al.)
- STRUCTURE software (Pritchard et al.)
- Devlin, Roeder, Bacanu "genomic control"
Be aware of LD in **design** and in **interpretation**

GWAS SNP arrays rely on LD to “tag” common variation or impute untyped variation

A popular LD-based "tagging" SNP approach:
- "$r^2$ bin tags" (Carlson et al., 2004): greedy algorithm that identifies bins of SNPs such that at least one member of each bin has $r^2 > T$ (threshold, e.g. 0.8) with all bin members.
  - Bin members are not necessarily contiguous

Similar method: “haplotype tagging” – recall HapMap

Imputation of untyped SNPs is possible due to LD and haplotype structure
Genome-wide Association Studies (GWAS)

Successful by several metrics: have led to

- **Replicable** findings; novel genes, pathways, biology
- Meaningful functional follow-up and clinical benefit
  - Recall Jim Skeath’s lectures: fetal hemoglobin, \textit{BCL11A} and CRISPR-Cas9 for sickle cell & thalassemia

Less successful by other metrics:

- "Top" associated SNPs explain limited phenotypic variance
  (typical odds ratios ~ 1.3, variance explained ~ 1%)

Still, good news:

- Polygenic risk scores explain more phenotypic variance
“Post-GWAS” challenges for interpreting and following up results

1. Moving from associated SNVs to “causal variants” or “underlying biology”
   1. Identify the most promising variants for functional follow-up
      1. Narrow down amongst SNPs in strong LD using diverse cohorts with differing LD and population history
      2. Functional and bioinformatics annotation
   2. Functional assays
   3. Effects in model organisms, organoids

2. Large samples needed to detect or replicate moderate effects (OR < 1.3)
   1. Combine studies (meta-analysis of results, or combining samples)
**CHRNA5-CHRNA3-CHRNB4** region: the $r^2 \geq 0.8$ bin.

Associated with nicotine dependence, smoking, lung cancer, cocaine dependence.

**Conclusive evidence in European-ancestry populations**

<table>
<thead>
<tr>
<th>SNP</th>
<th>Study References</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs16969968</td>
<td>Saccone SF et al., 2007</td>
</tr>
<tr>
<td></td>
<td>Bierut et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Liu et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Saccone NL et al., 2010</td>
</tr>
<tr>
<td>rs8034191</td>
<td>Hung et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Amos et al., 2008</td>
</tr>
<tr>
<td>rs1317286</td>
<td>Berrettini et al., 2008</td>
</tr>
<tr>
<td>rs1051730</td>
<td>Thorgeirsson et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Hung et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Amos et al., 2008</td>
</tr>
</tbody>
</table>

Blue: nicotine dependence / smoking
Red: lung cancer
**CHRNA5-CHRNA3-CHRNB4** region: the $r^2 > 0.8$ bin. Associated with nicotine dependence, smoking, lung cancer

In HapMap YRI, there are only 2 non-trivial $r^2$ bins

The other SNPs are singleton bins!

Opportunity to narrow down the signal
Combining studies/datasets through sharing/collaboration

Via meta-analysis and/or combined ("mega") analysis

Benefits:

- Improved power
- Extends value of existing data (often costly to collect)

Challenges:

1. Harmonizing phenotypes
2. Harmonizing genotypes: imputation (prediction of genotypes “in silico” using LD information.)
Meta-analysis

• Meta-analysis: statistically combines summary statistics across multiple datasets
  • Increases power via the new larger sample size

• Meta-analysis can be applied to published data/results. Retrospective literature review.

• Collaborative meta-analyses goes further: new, coordinated analyses across multiple datasets
  • Can include unpublished datasets
Meta-analysis

- **Goal:** Given a variable (SNP) being tested, estimate its overall effect (e.g. beta coefficient or odds ratio) across all the studies

- Why not just take the mean of the effect sizes?
  - This does not account for the differing precision of each study

- **Solution:** Weight each study by some measure of its "precision" – related to sample size
  - **Inverse variance weighted meta-analysis**
Forest plot: Heavy / light smoking at rs16969968

<table>
<thead>
<tr>
<th>Study</th>
<th>OR</th>
<th>95% C.I.</th>
<th>Case/Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>COGEND</td>
<td>1.50</td>
<td>(1.28-1.75)</td>
<td>641/1010</td>
</tr>
<tr>
<td>ADD Health</td>
<td>2.20</td>
<td>(1.31-3.69)</td>
<td>41/293</td>
</tr>
<tr>
<td>BoMa-aff-bpd</td>
<td>1.38</td>
<td>(0.76-2.52)</td>
<td>109/35</td>
</tr>
<tr>
<td>BoMa-aff-mdd</td>
<td>1.52</td>
<td>(0.91-2.54)</td>
<td>92/54</td>
</tr>
<tr>
<td>BoMa-scz</td>
<td>1.31</td>
<td>(0.65-2.63)</td>
<td>90/28</td>
</tr>
<tr>
<td>CADD</td>
<td>0.90</td>
<td>(0.53-1.54)</td>
<td>51/141</td>
</tr>
<tr>
<td>CPS-II_CPD</td>
<td>1.37</td>
<td>(1.22-1.53)</td>
<td>1443/1377</td>
</tr>
<tr>
<td>CPS-II_LCA</td>
<td>1.34</td>
<td>(1.12-1.61)</td>
<td>461/624</td>
</tr>
<tr>
<td>ECLIPSE</td>
<td>1.81</td>
<td>(1.36-2.42)</td>
<td>778/137</td>
</tr>
<tr>
<td>GenMetS</td>
<td>1.49</td>
<td>(1.01-2.19)</td>
<td>77/319</td>
</tr>
<tr>
<td>HPFS_CHD</td>
<td>1.12</td>
<td>(0.85-1.48)</td>
<td>244/191</td>
</tr>
<tr>
<td>HPFS_KS</td>
<td>0.95</td>
<td>(0.57-1.59)</td>
<td>76/77</td>
</tr>
<tr>
<td>HPFS_T2D</td>
<td>1.42</td>
<td>(1.15-1.77)</td>
<td>481/309</td>
</tr>
<tr>
<td>LHS</td>
<td>1.48</td>
<td>(1.14-1.92)</td>
<td>1250/144</td>
</tr>
<tr>
<td>MD Anderson</td>
<td>1.25</td>
<td>(1.01-1.54)</td>
<td>1135/250</td>
</tr>
<tr>
<td>MUC12SCS</td>
<td>1.25</td>
<td>(0.86-1.81)</td>
<td>137/96</td>
</tr>
<tr>
<td>MUC12SCTL</td>
<td>1.25</td>
<td>(0.65-2.39)</td>
<td>33/118</td>
</tr>
<tr>
<td>MUCMDCS</td>
<td>1.19</td>
<td>(0.87-1.62)</td>
<td>200/153</td>
</tr>
<tr>
<td>MUCMDCTL</td>
<td>1.36</td>
<td>(1.02-1.81)</td>
<td>142/497</td>
</tr>
<tr>
<td>NAG–Aus/BigSib</td>
<td>1.13</td>
<td>(0.96-1.33)</td>
<td>734/592</td>
</tr>
<tr>
<td>NAG–Finland</td>
<td>1.94</td>
<td>(0.75-5.01)</td>
<td>45/29</td>
</tr>
<tr>
<td>NCI-EAGLE</td>
<td>1.50</td>
<td>(1.29-1.74)</td>
<td>841/699</td>
</tr>
<tr>
<td>NCI-PLCO</td>
<td>1.36</td>
<td>(1.13-1.62)</td>
<td>1254/379</td>
</tr>
<tr>
<td>NHS_BrCa</td>
<td>1.28</td>
<td>(1.02-1.61)</td>
<td>359/305</td>
</tr>
<tr>
<td>NHS_CHD</td>
<td>1.16</td>
<td>(0.88-1.53)</td>
<td>243/198</td>
</tr>
<tr>
<td>NHS_KS</td>
<td>1.46</td>
<td>(0.88-2.43)</td>
<td>63/72</td>
</tr>
<tr>
<td>NHS_T2D</td>
<td>1.37</td>
<td>(1.13-1.65)</td>
<td>457/481</td>
</tr>
<tr>
<td>NYFS</td>
<td>1.38</td>
<td>(0.87-2.18)</td>
<td>51/107</td>
</tr>
<tr>
<td>UK_Phase_II</td>
<td>1.26</td>
<td>(1.07-1.47)</td>
<td>962/563</td>
</tr>
<tr>
<td>Utah</td>
<td>0.90</td>
<td>(0.60-1.35)</td>
<td>234/62</td>
</tr>
<tr>
<td>UVAD-MSTF</td>
<td>3.09</td>
<td>(1.34-7.14)</td>
<td>143/22</td>
</tr>
<tr>
<td>VA-twin</td>
<td>1.14</td>
<td>(0.97-1.33)</td>
<td>1098/603</td>
</tr>
<tr>
<td>WSU</td>
<td>1.48</td>
<td>(1.12-1.95)</td>
<td>333/176</td>
</tr>
<tr>
<td>YALE-UConn</td>
<td>1.24</td>
<td>(0.91-1.70)</td>
<td>154/214</td>
</tr>
</tbody>
</table>

Total 14452/10355

Summary OR 1.33 (1.26-1.39) P-value 5.96 x 10^{-31}

Saccone et al., PLoS Genetics 2010
Smoking GWAS meta-analysis

$N = 38,181$ subjects

Tobacco and Genetics Consortium, Nat Genet 2010
Many sources for full GWAS results

- UK Biobank
- dbGaP
- Psychiatric Genomics Consortium (PGC)
- Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE)
- GWAS Catalog
- Published GWAS often link to sites for sharing results
Why are complete GWAS *results* useful?

- Enables meta-analyses and “lookups” for replication
- Hypothesis-generating – e.g. for functional follow-up
- Polygenic Risk Scores can be built with just the GWAS results from a “source” or “base” dataset
  - Typically applied to other “target” datasets that have individual level genotype and phenotype data
Human genetics results motivate model organism follow-up

- Patient sequencing and the Undiagnosed Diseases Network (UDN)

- GWAS: *CHRNA5* gene and smoking behavior
• Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR), 2021

• Undiagnosed Diseases Network (UDN)
WashU Model Organisms Screening Center (MOSC), Schedl and Solnica-Krezel

worm, zebrafish
Baylor, U of Oregon MOSC
zebrafish, fly

Screening for pathogenicity of variants identified in patients
Following up GWAS results: example from smoking behavior
N = 38,181 subjects

Tobacco and Genetics Consortium, Nat Genet 2010
**CHRNA5-CHRNA3-CHRN4 region:** the $r^2 \geq 0.8$ bin.

GWAS: associated with nicotine dependence, smoking, lung cancer

- **rs16969968**
  - Saccone SF, 2007
  - Bierut et al., 2008
  - Liu et al., 2008

- **rs1317286**
  - Berrettini et al., 2008

- **rs1051730**
  - Thorgeirsson et al., 2008
  - Hung et al., 2008
  - Amos et al., 2008

---

Blue: nicotine dependence / smoking
Red: lung cancer
Cross-population analyses point to **CHRNA5**
Associated with nicotine dependence, smoking, lung cancer

In HapMap YRI, there are only 2 non-trivial $r^2$ bins

The other SNPs are singleton bins!

Opportunity to narrow down the signal
Kenny lab: Increased nicotine intake in α5* knockout mice.

Kenny lab: Increased nicotine intake in α5* knockout mice.

Polygenic Risk Scores (PRS)

- Motivation:
  - Want to estimate disease risk from genetic info
  - GWAS results are often available, more accessible than individual-level data
- GWAS results from large discovery samples are best
- The PRS can be generated from the discovery sample results and applied to your dataset of individual level data.

- Uses:
  - Predicting disease risk (in individuals)
  - Determining shared genetic influences between 2 diseases or traits (e.g. test whether PRS for trait 1 is significantly associated with risk for trait 2)
  - As a covariate to detect/evaluate effect of other factors on disease
Polygenic Risk Scores

• How to calculate a PRS for a given trait, for an individual:
  – Sum of genotypes at multiple loci, weighted by their effect sizes on that trait, usually obtained from published GWAS.
  – Recall the regression formulas in prior lectures: beta coefficients are a measure of effect sizes

• Usually must set a p-value threshold to determine which variants/loci to include in the PRS (often very lenient)
GWAS Manhattan Plot

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
GWAS and Polygenic Risk Scores

Use GWAS results from “Source” cohort.

Choose threshold for which SNVs to include.

Can compute the PRS in a separate “Target” cohort by applying the same coefficients (weights) to the genotyped people in that sample:

\[ PRS \approx \alpha + \sum_{i=1}^{n} \beta_{1i}x_{1i} \]

\( n = \text{number of SNPs in the PRS} \)

So need only the results from a GWAS, e.g. the “betas” or odds ratios.
Figure 2c:
Compare to red dots – no predictive power
Polygenic Risk Scores, Martin et al 2019

- May not translate well across population groups
- Another reason we need diverse population representation in GWAS

Figure 3
Representation of ancestry categories by:
% of GWAS studies

European ancestry 52.27%

European ancestry 78.39%

Sirugo et al, Cell 177, March 21 2019
Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

• National Academies of Sciences, Engineering, Medicine Committee

• Co-chaired by Aravinda Chakravarti and Charmaine Royal

• March 2023 Report with recommended best practices:

• Webinar for the report release:
Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

• Guiding principles: Respect
  Beneficence
  Equity & Justice
  Validity & Reproducibility
  Transparency & Replicability

• Overarching themes of their recommendations, briefly:
  – Avoid typological thinking
  – Include environmental factors in study design (directly measured)
  – Engage communities and participants

• Some highlights, from several important recommendations:
  • Researchers should not use race as a proxy for human genetic variation.
  • Tailor use of population descriptors to the type and purpose of the study

• Think about how and why human cohorts and populations are described, in your own work and in papers you read.
Moving from GWAS to post-GWAS: Precision Medicine

– Reminder: Genomic medicine is a subset of Personalized/Precision medicine

– Can polygenic scores from GWAS identify clinically relevant elevations in disease risk?
  * Especially high-risk individuals that standard risk prediction might miss?

– Can SNPs or polygenic scores identify which individuals are most likely to benefit from a particular treatment?
Moving from GWAS to post-GWAS: Precision 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) – All of Us
    • Privacy protection
    • Modernizing regulatory landscape
    • Public-private partnerships
Target Timeline Towards Precision (Genomic) 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

Examples

– Warfarin dosing

– American Journal of Human Genetics now publishes a “Genomic Medicine Year in Review” feature

  • Manolio et al., AJHG December 2020

<table>
<thead>
<tr>
<th>Box 1. Criteria for Inclusion of Papers in Genomic Medicine Year in Review 2019 and 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Involve use of patients’ individual genomic variant information in clinical decision-making</td>
</tr>
<tr>
<td>• Demonstrate impact of direct clinical implementation</td>
</tr>
<tr>
<td>• Are likely to be generalizable beyond original setting</td>
</tr>
<tr>
<td>• Are likely to have implications for healthcare systems or practice guidelines</td>
</tr>
<tr>
<td>• Are of sufficient size to be robust to sampling error</td>
</tr>
<tr>
<td>• Are broadly representative of the field beyond NHGRI-sponsored or US-funded programs</td>
</tr>
</tbody>
</table>
Testing for interactions

- **Here:** GxT interaction, T=treatment

- **Others:** GxG (epistasis), GxE

- Does the 1st variable’s effect on the outcome differ for different values of the 2nd variable?

- Recall Heather Lawson’s quantitative genetics lectures

- Also the Gerke et al. discussion paper (this week), e.g. Figure 4
CHRNA5 Predicts Cessation & Response to Medication

Smokers with **CHRNA5**
- H1: LOW RISK
- H2
- H3: HIGH RISK

Abstinence

<table>
<thead>
<tr>
<th>Percentage</th>
<th>H1: LOW RISK</th>
<th>H2</th>
<th>H3: HIGH RISK</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>100%</td>
<td>90%</td>
<td>80%</td>
</tr>
<tr>
<td>10%</td>
<td>90%</td>
<td>80%</td>
<td>70%</td>
</tr>
<tr>
<td>20%</td>
<td>80%</td>
<td>70%</td>
<td>60%</td>
</tr>
<tr>
<td>30%</td>
<td>70%</td>
<td>60%</td>
<td>50%</td>
</tr>
<tr>
<td>40%</td>
<td>60%</td>
<td>50%</td>
<td>40%</td>
</tr>
<tr>
<td>50%</td>
<td>50%</td>
<td>40%</td>
<td>30%</td>
</tr>
<tr>
<td>60%</td>
<td>40%</td>
<td>30%</td>
<td>20%</td>
</tr>
<tr>
<td>70%</td>
<td>30%</td>
<td>20%</td>
<td>10%</td>
</tr>
<tr>
<td>80%</td>
<td>20%</td>
<td>10%</td>
<td>0%</td>
</tr>
</tbody>
</table>

This represents a GxE interaction!


Smoking cessation trial, N=1073, European Ancestry
*CHRNA5* Predicts Cessation & Response to Medication


**Smokers with CHRNA5 low risk haplotype**

**Smokers with CHRNA5 high risk haplotype**

This represents a GxE interaction!

Prospects for Genomic and Precision Medicine

Prospects for Genomic and Precision Medicine


**Box 5**

**Bold predictions for human genomics by 2030**

Some of the most impressive genomics achievements, when viewed in retrospect, could hardly have been imagined ten years earlier. Here are ten bold predictions for human genomics that might come true by 2030. Although most are unlikely to be fully attained, achieving one or more of these would require individuals to strive for something that currently seems out of reach. These predictions were crafted to be both inspirational and aspirational in nature, provoking discussions about what might be possible at the forefront of genomics in the coming decade.

1. Generating and analysing a complete human genome sequence will be routine for any research laboratory, becoming as straightforward as carrying out a DNA purification.
2. The biological function(s) of every human gene will be known; for non-coding elements in the human genome, such knowledge will be the rule rather than the exception.
3. The general features of the epigenetic landscape and transcriptional output will be routinely incorporated into predictive models of the effect of genotype on phenotype.
4. Research in human genomics will have moved beyond population descriptors based on historic social constructs such as race.
5. Studies that involve analyses of genome sequences and associated phenotypic information for millions of human participants will be regularly featured at school science fairs.
6. The regular use of genomic information will have transitioned from boutique to mainstream in all clinical settings, making genomic testing as routine as complete blood counts.
7. The clinical relevance of all encountered genomic variants will be readily predictable, rendering the diagnostic designation of 'variant of uncertain significance (VUS)' obsolete.
8. An individual's complete genome sequence along with informative annotations will, if desired, be securely and readily accessible on their smartphone.
9. Individuals from ancestrally diverse backgrounds will benefit equitably from advances in human genomics.
10. Breakthrough discoveries will lead to curative therapies involving genomic modifications for dozens of genetic diseases.