Human Genetics and Gene Mapping of Complex Traits

Advanced Genetics, Spring 2019
Human Genetics Series
Tuesday 4/09/19
Nancy L. Saccone
Dept of Genetics
nlims@genetics.wustl.edu / 314-747-3263
Genome-wide Association Studies (GWAS)

Successful by several metrics: have led to

- **Replicable** genetic variants underlying complex diseases
- Novel genes, pathways, biology
- Meaningful functional followup

Less successful by other metrics:

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

Additional good news:

- Polygenic risk scores explain more phenotypic variance

GWASes rely on linkage disequilibrium (LD) to "tag" variation, and thus must be interpreted in the context of LD: The signal SNP may not be the biologically causal SNP.
Outline and learning objectives

• Linkage disequilibrium (LD)
  • Measures of LD, how to compute
  • Sources of LD
  • Practical implications

• GWAS to "Post-GWAS"
  • Resources (e.g. available GWAS data, results)
  • GWAS meta-analyses, consortia
  • Diverse populations
  • Polygenic risk scores (PRS)
  • Translating to clinical care
  • Getting from statistical signal to biology
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: *CHRNA5-CHRNA3-CHRNB4* on chromosome 15q25

GWAS significance threshold: $P = 5 \times 10^{-8}$

Some (not all!) of these SNPs are highly correlated with a non-synonymous SNP rs16969968 in *CHRNA5*

Tobacco and Genetics Consortium, *Nat Genet* 2010
Example: \textit{CHRNA5-CHRNA3-CHRNB4} and rs16969968
Associated with nicotine dependence, smoking, lung cancer, COPD.

rs16969968
Saccone SF et al., 2007
Example: *CHRNA5-CHRNA3-CHRNB4* and rs16969968

Associated with nicotine dependence, smoking, lung cancer, COPD.
Example: *CHRNA5-CHRNA3-CHRNB4* and rs16969968
Associated with nicotine dependence, smoking, lung cancer, COPD.

A second independent "bin" of correlated SNPs represents a distinct statistical signal.
A typical GWAS signal consists of multiple SNPs due to LD

Often multiple distinct statistical signals exist

"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:

\[ D = h_{11} - p_{A₁}p_{B₁} \]
Common LD measures

Disequilibrium coefficient:
\[ D = h_{11} - p_{A_1} p_{B_1} \]

Normalized disequilibrium coefficient:
\[ D' = D / |D|_{\text{max}} \], where
\[
|D|_{\text{max}} = \begin{cases} 
\min(p_{A_1}p_{B_2}, p_{A_2}p_{B_1}) & \text{if } D > 0 \\
\min(p_{A_1}p_{B_1}, p_{A_2}p_{B_2}) & \text{if } D < 0
\end{cases}
\]

Range of D' is [-1,1]

Squared correlation coefficient:
\[ r^2 = D^2 / ( p_{A_1}p_{A_2}p_{B_1}p_{B_2} ) \]
Notes:

1. $D = h_{11} - p_{A1}p_{B1} = h_{22} - p_{A2}p_{B2}$

2. Choice of allele labeling may affect sign but not absolute value of $D$. 

$$
\begin{array}{c|cc}
A_1 & B_1 & B_2 \\
\hline
n_{11} & n_{12} \\
\hline
n_{21} & n_{22} \\
\hline
\end{array}
$$

2N
## Measuring LD

**Example:**

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

<table>
<thead>
<tr>
<th></th>
<th>(B_1)</th>
<th>(B_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A_1)</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>(A_2)</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

\[
D = h_{11} - p_{A1}p_{B1} = (0.5) - (0.5)(0.5) = 0.5 - 0.25 = 0.25
\]

\[
D_{max} = \min(p_{A1}p_{B2}, p_{A2}p_{B1}) = \min(0.25, 0.25) = 0.25
\]

\[
|D'| = 1, r^2 = 1
\]
Measuring LD

Example:

Only observe 2 haplotypes: $A_1B_1$ and $A_2B_2$

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>$A_2$</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A₁$</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>$A₂$</td>
<td>25</td>
<td>25</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>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>A₂</td>
<td>0</td>
<td>90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>A₂</td>
<td>10</td>
<td>80</td>
</tr>
</tbody>
</table>

\[ D' = \frac{(.1 - (.1)(.1))}{.09} = 1 \]

\[ \chi^2 = 100, \ p\text{-value} \sim 0.0 \]

\[ r^2 = 1 \]

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>A₂</td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>

\[ D' = \frac{0 - (.1)(.1)}{.01} = -1 \]

\[ \chi^2 = 1.23, \ p\text{-value} = 0.27 \]

\[ r^2 = 0.012 \]
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 marker allele frequencies.

$r^2$ equals 1 if and only if 1) the 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.
Describing empirical LD patterns

Haploview output: D’
Dick et al., 2007

Figure 2. Linkage disequilibrium across the single nucleotide polymorphisms (SNPs) genotyped in and around ACN9. D’ is illustrated by shading, with darker shades indicating higher D’. The \( r^2 \) is indicated by the number inside the shaded block.
LD across TCF7L2 in CEU HapMap.

Grant et al., Nat Genet 2006, Figure 1
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

ASW: African-American

CEU: European-American

\( r^2 \), Chromosome 15q25 region
The 1000 Genomes Project catalogs variants and LD structure

www.1000genomes.org
http://www.internationalgenome.org/

(Previously: The Haplotype Mapping (HapMap) Project)

Browsable via www.ensembl.org
Be aware of LD in **design** and in **interpretation**

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

The modern GWAS: "SNP chip" design
- Illumina
- Affymetrix
Where does LD come from?

• Potential sources of LD:
  1. Linkage between loci
  2. Random drift
  3. Founder effect
  4. Mutation
  5. Selection
  6. Population admixture / stratification
An important concern: population stratification

Spurious association between a locus and disease can occur if there are two (unknown) subpopulations.

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

**Population stratification in case-control association studies**

subpop1: allele A1

subpop2 alleles A1, A2

unshaded = affected
An example of spurious association due to admixture/stratification:

<table>
<thead>
<tr>
<th>population 1</th>
<th>population 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>81</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>90</td>
<td>50</td>
</tr>
<tr>
<td>90</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

chi-square = 0

χ² = 0

combined

| 34 | 26 | 60 |
| 106| 34 | 140|
| 140| 60 | 200|

χ² = 7.26

p-value = 0.007
How to protect against population stratification?

1. Match geographic/racial background of the case and control samples. (at the very least!)

2. Check to make sure there is no underlying stratification
   - Can use GWAS data
   - If study is not GWAS, would need a set of "ancestry-informative markers (AIMs)"

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

4. Software:
   - EIGENSTRAT software (principal components) (Price et al.)
   - STRUCTURE software (Pritchard et al.)
   - Devlin, Roeder, Bacanu "genomic control"
Next steps (post-GWAS)?

1. Narrow down to "true" (biologically causal) variants in the associated region
2. Determine how/why these variants alter disease risk
3. Translate to *clinical care, outcomes* (next lec)

How to do 1 and 2?

- Look across multiple diverse populations, leverage LD differences
- Bioinformatic prioritization
- Imputation, genotyping, sequencing to query the regions of association more thoroughly (1000 Genomes Project)
- Meta-analysis for larger, more powerful samples
- Functional follow-up
- Effects in model organisms
- Risk prediction using Polygenic Risk Scores
**CHRNA5-CHRNA3-CHRN4** region: the $r^2 \geq 0.8$ bin.

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

Conclusive evidence in European-ancestry populations

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

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

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

- **rs8034191**
  - Hung et al., 2008
  - Amos et al., 2008

Blue: nicotine dependence / smoking
Red: lung cancer
Green: cocaine dependence
**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 data through sharing/collaboration

For example, 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: genetic imputation
• Meta-analysis: statistically combines summary statistics across multiple datasets

• Thus meta-analysis can be applied to published data/results

• Collaborative meta-analyses goes further: not just retrospective literature review, but carrying out new, coordinated analyses across multiple datasets
  • Not limited to the "published analyses" for a given dataset
  • Can include unpublished datasets
Targeted smoking meta-analysis
(Saccone et al., PLoS Genetics 2010)

- 34 datasets, 17 sites
- rs16969968 on chr 15 genotyped in most but not all datasets
- Used proxy "tag" SNPs (rather than imputation):

<table>
<thead>
<tr>
<th>Locus</th>
<th>Target SNP</th>
<th>position (bp)</th>
<th>Proxy SNP</th>
<th>position (bp)</th>
<th>LD ($r^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs16969968</td>
<td>76669980</td>
<td>rs8034191</td>
<td>76593078</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs2036527</td>
<td>76638670</td>
<td>0.932</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs951266</td>
<td>76665596</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs1051730</td>
<td>76681394</td>
<td>0.900</td>
</tr>
</tbody>
</table>
Example: Targeted smoking meta-analysis (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>(1.13–1.66)</td>
<td>457/481</td>
</tr>
<tr>
<td>NYSFS</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>UVA–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 1.33 (1.26–1.39) 5.96 x 10^{-31}

Saccone et al., PLoS Genetics 2010
More and more data are available by request

• All NIH funded GWAS studies are required to make their data available through the Database of Genotypes and Phenotypes (dbGaP)

• Not just results – also individual-level genotype/phenotype data.

• (Open-access: for non-sensitive data)
• Controlled-access: Data access by request; granted after panel review - e.g. dbGAP DAC: Data Access Committee

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

Access dbGaP Data
- Advanced Search
- Controlled Access Data
- Public FTP Download
- Collections
- Summary Statistics

Resources
- dbGaP Data Browser
- Phenotype-Genotype Integrator
- dbGaP RSS Feed
- Software
- dbGaP Tutorial

Important Links
- How to Submit
- FAQ
- Code of Conduct
- Security Procedures
- Contact Us

Latest Studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Embargo Release</th>
<th>Details</th>
<th>Participants</th>
<th>Type Of Study</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>phs000895.v1.p1</td>
<td>Version 1 passed embargo</td>
<td>V D A S</td>
<td>17</td>
<td>Clinical Trial, Interventional</td>
<td>Links</td>
</tr>
<tr>
<td>A Phase II Trial of Sirolimus (Rapamune®) and Cyclosporine in Patients with Refractory Aplastic Anemia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phs001612.v1.p1</td>
<td>Version 1 passed embargo</td>
<td>V D A S</td>
<td>17</td>
<td>Methods Development, Longitudinal</td>
<td>Links</td>
</tr>
<tr>
<td>Platform for Assessing Immune Response Signatures in Healthy Human Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phs000830.v6.p1</td>
<td>Versions 1-6 passed embargo</td>
<td>V D A S</td>
<td></td>
<td>Cohort, Nested Case-Control, Family, Longitudinal</td>
<td>Links</td>
</tr>
<tr>
<td>CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Links
- [dbGaP RSS Feed](https://www.ncbi.nlm.nih.gov/gap)
- [Software](https://www.ncbi.nlm.nih.gov/gap)

How to Submit
- [FAQ](https://www.ncbi.nlm.nih.gov/gap)
- [Contact Us](https://www.ncbi.nlm.nih.gov/gap)
Ethical use of "publicly available" data

dbGaP Approved User Code of Conduct

The elements of the NIH Code of Conduct for Genomic Data Use include:

• Investigator(s) will use requested datasets solely in connection with the research project described in the approved Data Access Request for each dataset;

• Investigator(s) will make no attempt to identify or contact individual participants from whom these data were collected without appropriate approvals from the relevant IRBs;

• Investigator(s) will not distribute these data to any entity or individual beyond those specified in the approved Data Access Request;

• Investigator(s) will adhere to computer security practices that ensure that only authorized individuals can gain access to data files:

• Investigator(s) will not submit for publication or any other form of public dissemination analyses or other reports on work using or referencing NIH datasets prior to the embargo release date listed for the dataset (or dataset version) on dbGaP;

• Investigator(s) acknowledge the Intellectual Property Policies as specified in the Data Use Certification; and,

• Investigator(s) will report any inadvertent data release in accordance with the terms in the Data Use Certification, breach of data security, or other data management incidents contrary to the terms of data access.
Ethical use of "publicly available" data

• For dbGaP, investigators who request data sign an agreement not to *submit* analyses/manuscripts for publication before the embargo date.

• Requestors may still conduct analyses, get ready to submit a manuscript.

• This gives the PI who collected the data a proprietary period. Note that typically, data is made available on dbGaP at *the same time* it is available to the PI.
SCIENTIFIC PUBLISHING

Paper Retracted Following Genome Data Breach

Here’s a nightmare scenario: You go to the Web site of a leading journal, and there on your screen is a paper based on data you have painstakingly gathered but not yet had time to analyze.

That’s what happened to psychiatrist Laura Bierut, who discovered last week that other researchers had broken an embargo on use of data she and her colleagues had deposited in dbGaP, the National Institutes of Health’s (NIH’s) database of genotypes and phenotypes. Bierut, a professor at Washington University in St. Louis in Missouri, and colleagues had collected the data as part of genetic studies of alcoholism and other addictions collectively known as SAGE (Study of Addiction: Genetics and Environment).

dbGaP was established in 2006 to facilitate sharing of the oceans of genetic data generated by federal grantees. Other scientists can submit papers based on the material.

Scooped. Another team broke the database embargo and published a paper using Laura Bierut’s data.

after an embargo period of 9 to 12 months so those who generated the data can have first crack at analyzing them.

The SAGE embargo ends on 23 September. But on 31 August, a paper based on SAGE data appeared online in the Proceedings of the National Academy of Sciences (PNAS). In it, a team led by Heping Zhang, a biostatistician at Yale School of Public Health, reported a positive association between a gene called PKNOX2 and addictions in women of European origin. It was submitted to the journal last March, breaching the embargo by 6 months.

Bierut immediately shot off e-mails to Yale, Princeton University (home of co-author and National Academy of Sciences member Burton Singer, who contributed the paper), PNAS, various NIH officials, and colleagues. “[T]his was likely an unintentional act, [but] this incident remains very concerning,” she wrote, adding that it “sends
Ethical use of "publicly available" data

• Dr. Laura Bierut’s SAGE (Study of Addiction, Genes and Environment) study was at dbGaP, under embargo until September 23, 2009
• On August 31, 2009 a paper came out in PNAS by Yale / Princeton investigators – on the SAGE dataset.
• The paper had been submitted March 2009
• NIH cut off the investigators’ dbGaP access, launched investigation
• That paper was retracted; PI was banned from dbGaP for a period of time.
• Bierut et al. then published SAGE findings in PNAS.

• Lesson: Read and abide by rules of data use
Human genetic resources, cohort databases, opportunities

- dbGaP
- UK Biobank (UKB)
- Million Veteran Program (MVP)
- All of Us
- Consumer genomics companies, e.g. 23andMe
- Electronic Health Records (EHR) based research
UK Biobank

- Prospective cohort study of 500,000 individuals across the United Kingdom from 2006-2010
- Genetic, physical, and health data, some longitudinal
- Designed as a general resource for health research
- "Open access" database
  - Open summary data
  - Controlled-access individual-level data, available to qualified researchers who apply
UK Biobank is a national and international health resource with unparalleled research opportunities, open to all bona fide health researchers. UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. Scientists, please ensure you read the background materials before registering. To our participants, we say thank you for supporting this important resource to improve health. Without you, none of the research featured on this website would be possible.

Read more about Biobank UK
View all our publications

Scientists
- Data showcase
- Register and apply
- Genetic data
- Imaging data
- Biomarker panel details

Participants
- Update your contact details
- 2018 Newsletter
- UK Biobank in the press
- Digestive health questionnaire
- General practice linkage

Keeping in touch

Participant queries
Please call the UK Biobank Participant Resource Centre on 0800 0 276 276 (free from most phones) 8am-6pm Mon – Fri and Sat 8am-4pm or email us. You can also update your contact details via this website.

Researcher enquiries
Please study the information provided in this website and email us if you have any further enquiries.

Questions about the study
If you have any questions about our research as a participant or researcher, please view our Frequently Asked Questions or Contact Us.

Feedback
Data showcase:
http://biobank.ctsu.ox.ac.uk/crystal/

- Contains only anonymous summary information
- Searchable/browsable
- Info to help design your study and corresponding data request
Alternative: what if you are mainly interested in GWAS results, not the item-level data itself?

Example: UK Biobank summary GWAS results available from Ben Neale's lab at the Broad/MGH: http://www.nealelab.is/uk-biobank

- Over 2,000 phenotypes (August 2018)
http://www.nealelab.is/uk-biobank
GWAS RESULTS

GWAS round 2:
Results shared 1st August 2018.
Imputed genotypes from HRC plus UK10K & 1000 Genomes reference panels as released by UK Biobank in March 2018.

GWAS round 2 results can be found here

GWAS round 2 Github code repository

GWAS round 1:
Results shared 20th September 2017.
Imputed genotypes from HRC as released by UK Biobank in May 2017.

FAQ

Got questions about our GWAS of the UK Biobank?

What samples, markers and phenotypes we used? The format of our GWAS results files? Use permissions of the results? How to cite our results?

Read our FAQs.
For a description of the project and details of the analysis, please see [http://www.nealelab.is/uk-biobank](http://www.nealelab.is/uk-biobank).

To download GWAS results, see the links in the manifest tab below. At the top of each column in the manifest is a triangle. Click the triangle and search options become available for that column. Once you've found the code you are looking for, refer to the "wget command" column for the corresponding wget command to download the relevant results file.

The code used to generate the files described here is publicly available: [https://github.com/NealeLab/UK_Biobank_GWAS](https://github.com/NealeLab/UK_Biobank_GWAS).

Questions or concerns not addressed by this README, the project website, our FAQs ([http://www.nealelab.is/faq](http://www.nealelab.is/faq)) or the Github repository can be directed to nealelab.ukb@gmail.com.

---

### variants.tsv.bgz

This file contains annotations on each variant in the GWAS, calculated across the analysis subset of 361,194 samples.

**NOTE:** The order of variants in this file matches the order of variants in the results files described below. To join these annotations with a results file, either match on the "variant" field or simply paste the columns together (e.g. "paste variants.tsv K50.gwas.imputed_v3.both_sexes.tsv").

#### Contents:

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>variant</td>
<td>string</td>
<td>Variant identifier in the form &quot;chr:pos:ref:alt&quot;, where &quot;ref&quot; is aligned to the forward strand.</td>
</tr>
<tr>
<td>chr</td>
<td>string</td>
<td>Chromosome of the variant.</td>
</tr>
<tr>
<td>pos</td>
<td>Int</td>
<td>Position of the variant in GRCh37 coordinates.</td>
</tr>
<tr>
<td>ref</td>
<td>string</td>
<td>Reference allele on the forward strand.</td>
</tr>
<tr>
<td>alt</td>
<td>string</td>
<td>Alternate allele (not necessarily minor allele).</td>
</tr>
<tr>
<td>Phenotype Code</td>
<td>Phenotype Description</td>
<td>UK Biobank</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>N/A</td>
<td>README</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>List of European samples</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>List of samples used in GWAS - both sexes</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>List of samples used in GWAS - female</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>List of samples used in GWAS - male</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>Phenotype summary file - both sexes</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>Phenotype summary file - female</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>Phenotype summary file - male</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>List of variants used in GWAS, with annotations</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>List of phenotypes with updated sample counts in v2 phenotype summary file - female</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>List of phenotypes with updated sample counts in v2 phenotype summary file - male</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>Note on v2 phenotype updates</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>Updated v2 phenotype summary file - both sexes</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>Updated v2 phenotype summary file - female</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>List of phenotypes with updated v2 summary statistics - both sexes</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>List of phenotypes with updated v2 summary statistics - male</td>
<td>N/A</td>
</tr>
<tr>
<td>20</td>
<td>ls_female</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>21</td>
<td>age</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>Column</td>
<td>Description</td>
<td>URL</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>-----</td>
</tr>
<tr>
<td>6630</td>
<td>Number of cigarettes currently smoked daily (current cigarette smokers)</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6631</td>
<td>Number of cigarettes currently smoked daily (current cigarette smokers)</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6632</td>
<td>Number of cigarettes currently smoked daily (current cigarette smokers)</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6633</td>
<td>Time from waking to first cigarette</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6634</td>
<td>Time from waking to first cigarette</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6635</td>
<td>Time from waking to first cigarette</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6636</td>
<td>Difficulty not smoking for 1 day</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6637</td>
<td>Difficulty not smoking for 1 day</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6638</td>
<td>Difficulty not smoking for 1 day</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6639</td>
<td>Ever tried to stop smoking</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6640</td>
<td>Ever tried to stop smoking</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6641</td>
<td>Ever tried to stop smoking</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6642</td>
<td>Wants to stop smoking</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6643</td>
<td>Wants to stop smoking</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6644</td>
<td>Wants to stop smoking</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6645</td>
<td>Mother’s age at death</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6646</td>
<td>Mother’s age at death</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6647</td>
<td>Mother’s age at death</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6648</td>
<td>Mother’s age at death</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6649</td>
<td>Mother’s age at death</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
<tr>
<td>6650</td>
<td>Mother’s age at death</td>
<td><a href="http://www.nealelab.is/uk-biobank">http://www.nealelab.is/uk-biobank</a></td>
</tr>
</tbody>
</table>
Many sources for GWAS results

- UK Biobank
- Psychiatric Genomics Consortium (PGC)
- Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE)
- Published GWAS often link to sites for sharing results
Polygenic Risk Scores

• Motivation:
  – Want to estimate disease risk from genetic info
  – GWAS results are often freely available, more accessible than individual-level data
  – Can we use available summary data to estimate risk in a separate sample where we have individual data?
• GWAS results from large discovery samples like the PGC or UK Biobank are best
• The PRS can be generated from the discovery sample and applied to your data where you have individual level data.
Polygenic Risk Scores

• PRS for a given trait can be calculated for individuals: sum of genotypes at multiple loci, weighted by their effect sizes on that trait, usually obtained from published GWAS. (recall the regression formulas last week: betas -> effect sizes)

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

• Used for:
  – 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

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

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