Human Genetics and Gene Mapping of Complex Traits

Advanced Genetics
Human Genetics Series
Thursday 3/29/23
Nancy L. Saccone, Ph.D.
Dept of Genetics
nlims@genetics.wustl.edu
Outline and learning objectives

• Human disease mapping approaches:
  • Part I: Linkage analysis
    • Disease models
    • Recombination, LOD scores
  • Part II: Genetic association analyses
    • Use of regression-based tests
    • Effect sizes: beta and odds ratio, “variance explained”

• Genome-Wide Association Studies (GWAS)

• Multiple test correction: what and why

• Precision Medicine

Mapping disease genes

– Linkage

• quantify co-segregation of trait and genotype in families

- LOD score traditionally used to measure statistical evidence for linkage
- Based on counting recombinants

• Association

• Common design: case-control sample of unrelateds, analyzed for allele frequency differences
### Comparing Linkage and Association

<table>
<thead>
<tr>
<th>Linkage mapping:</th>
<th>Association mapping:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requires family data</td>
<td>Unrelated cases/controls OR Case/parents OR family design</td>
</tr>
<tr>
<td>Disease travels with marker allele within families (close genetic distance between disease locus and marker)</td>
<td>Disease is associated with marker allele that may be either causative or in linkage disequilibrium with causal variant</td>
</tr>
<tr>
<td>Relationship between same allele and trait need not exist across the full sample (e.g. across different families)</td>
<td>Works only if association exists at the population level</td>
</tr>
<tr>
<td>robust to allelic heterogeneity: if different mutations occur within the same gene/locus, the method works</td>
<td>not robust to allelic heterogeneity</td>
</tr>
<tr>
<td>signals for complex traits tend to be broad (~20 Mb)</td>
<td>association signals generally not as broad</td>
</tr>
</tbody>
</table>
**Classic models of disease**

Penetrance table: probabilities of disease by genotype

Classical autosomal dominant inheritance (no phenocopies, fully penetrant).

Penetrance table:

<table>
<thead>
<tr>
<th></th>
<th>f_{++}</th>
<th>f_{+D}</th>
<th>f_{DD}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Often the dominant allele is rare, so that probability of homozygous DD individuals occurring is negligible.

Classical autosomal recessive inheritance (no phenocopies, fully penetrant).

Penetrance table:

<table>
<thead>
<tr>
<th></th>
<th>f_{++}</th>
<th>f_{+d}</th>
<th>f_{dd}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Genetic models of disease

Other examples of penetrance tables (locus-specific):

\[
\begin{array}{ccc}
 f_{++} & f_{+d} & f_{dd} \\
 0 & 1 & 1 \\
 0 & 0 & 1 \\
 0 & 0 & 0.9 \\
 0.1 & 1 & 1 \\
 0.1 & 0.8 & 0.8 \\
\end{array}
\]

**Incomplete/reduced penetrance:** when the risk genotype's effect on phenotype is not always expressed/observed. (e.g. due to environmental interaction, modifier genes)

**Phenocopy:** individual who develops the disease/phenotype in the absence of "the" risk genotype (e.g. through environmental effects, heterogeneity of genetic effects)
Part I: Human linkage studies

Need to track co-segregation of trait and markers (number of recombination events among observed meioses)

General “linkage screen" approach:

• Recruit families

• Genotype individuals at marker loci along the genome

• If a marker locus is "near" the trait-influencing locus, the parental alleles from the same **grandparent** at these two loci "tend to be inherited together" (recombination between the two loci is rare)

• Historically used very polymorphic markers (e.g. microsatellites) rather than bi-allelic SNPs
Linkage analysis concepts

Example of meiosis resulting in some recombinant gametes

Parent

\[ \begin{align*}
A_1 & \quad A_2 \\
B_1 & \quad B_2
\end{align*} \]

meiosis

\[ \begin{align*}
A_1 & \quad A_1 \quad A_2 \\
B_1 & \quad B_1 \quad B_2 \\
A_2 & \quad A_2
\end{align*} \]

From one grandparent

Non-sister chromatids may cross over between the loci

sister chromatids

Recombination with respect to 2 loci: has occurred when in the individual's haplotype (ie. in a gamete inherited from the parent), the alleles at the 2 loci come from the 2 different grandparents.
Linkage analysis concepts

Example of meiosis resulting in some recombinant gametes

Parent

\[ A_1 | A_2 \]
\[ B_1 | B_2 \]

meiosis

\[ A_1 | A_1 A_2 | A_2 \]
\[ B_1 | B_1 B_2 | B_2 \]

Non-sister chromatids may cross over between the loci

\[ A_1 \]
\[ B_1 \]
\[ A_1 A_2 \]
\[ B_2 B_1 \]
\[ A_2 \]
\[ B_2 \]

From one grandparent

sister chromatids

General expectation:

If loci are close together, less likely to see a recombination

So want to count recombinants and non-recombinants

gametes
Testing for linkage

Let $\theta = \text{the probability of recombination between 2 given loci}$

(range, 0 to $\frac{1}{2}$)

Null hypothesis: $\theta = \frac{1}{2}$

Alternative hypothesis: $\theta = \hat{\theta}$, the max likelihood estimate of $\theta$

$$\text{maximum LOD score} = \log_{10}\left(\frac{L(\hat{\theta})}{L(1/2)}\right)$$

- $L = \text{likelihood function for the observed data}$

- Traditionally, a LOD $\geq 3$ (or 3.6) is considered "significant"
Linkage analysis in pedigrees

In general, for a case in which all recombinants and non-recombinants can be explicitly counted:

• Suppose we observe $k$ recombinants in a total of $m$ meioses. We estimate the “maximum likelihood estimate” for $\theta$ to be

$$\hat{\theta} = \left( \frac{\# \text{ recombs}}{\# \text{ meioses}} \right) = \frac{k}{m}$$

(Unless $k/m > 1/2$, in which case $\hat{\theta} = 1/2$)

• The likelihood function comes from the binomial probability distribution

$$L(\theta) = \binom{m}{k} \theta^k (1 - \theta)^{m-k}$$

where

$$\binom{m}{k} = \frac{m!}{k!(m-k)!}$$

Then the maximum LOD score = $\log_{10} \left( \frac{L(\hat{\theta})}{L(1/2)} \right)$

So how do we count recombinants and nonrecombinants?
First, how to count recombs/nonrecombs between two genotyped marker loci, and compute the LOD score

Example: phase known (vertical lines), fully heterozygous parents

4 informative meioses.
Can explicitly count: 1 R, 3 NR.
First, how to count recombs/nonrecombs between two genotyped marker loci, and compute the LOD score

Example: phase known (vertical lines), fully heterozygous parents

\[
\hat{\theta} = \frac{\text{# recombs}}{\text{# meioses}} = \frac{1}{4}
\]

4 informative meioses.
Can explicitly count: 1 R, 3 NR.

Binomial probability distribution gives us a likelihood function for the observed data:

\[
L(\theta) = \binom{4}{1} \theta^1 (1 - \theta)^3
\]
First, how to count recombs/nonrecombs between two genotyped marker loci, and compute the LOD score

1 Recombinant, 3 Nonrecombinants

\[
L(\theta) = \binom{4}{1} \theta (1 - \theta)^3 = 4 \theta (1 - \theta)^3
\]

\[\hat{\theta} = \frac{\# \text{ recombs}}{\# \text{ meioses}} = \frac{1}{4}\]

\[
L(1/4) = 4 \times (1/4) \times (3/4)^3 =
\]

\[
L(1/2) = 4 \times (1/2) \times (1/2)^3 =
\]

\[
\text{LOD} = \log_{10} \left( \frac{L(1/4)}{L(1/2)} \right) = \log_{10} (27/16)
\]

\[= 0.227 \ll \text{traditional threshold of 3 or 3.6} \]
Linkage analysis in pedigrees

In cases where cannot definitively count all recombs/non-recombs, the likelihood function becomes more complex.

E.g. if phase is unknown, need to account for multiple possible phases, and build the likelihood as a weighted sum of multiple likelihoods.
Notes on linkage analysis

• To be informative for linkage, meioses must come from fully heterozygous parents (het and with different alleles)

• The likelihood function is more complex in cases where cannot definitively count all recombs/non-recombs
  • E.g. if phase is unknown, then account for multiple possible phases, and build the likelihood as a weighted sum of multiple likelihoods

• LOD curves are additive (because of properties of the log function and likelihoods)
  • Recall our example family, max LOD = 0.227 at $\hat{\theta} = \frac{1}{4}$. If we have 15 such families with max LOD = 0.227 at $\hat{\theta} = \frac{1}{4}$, then together, LOD is $0.227 + 0.227 + \ldots + 0.227$ or $15*0.227 = 3.405$ (!)
Linkage analysis in pedigrees

Next, how do we compute the LOD score indicating linkage between an **unknown disease locus** and a marker locus?

Key: must assume a disease model. **This allows us to assign possible genotypes at the disease locus.** Disease model specifies:

- disease gene frequency $p_D$
- penetrances for the 3 genotypic states
Linkage analysis in pedigrees

Example: Suppose the disease allele D is fully penetrant, dominant, rare. Suppose phenocopies do not occur.

Penetrance table:

<table>
<thead>
<tr>
<th>Transmission</th>
<th>d d</th>
<th>D d</th>
<th>DD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penetrance</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

![Pedigree Diagram]
**Linkage analysis in pedigrees**

Example: Suppose the disease allele D is fully penetrant, dominant, rare. Suppose phenocopies do not occur.

Penetrance table: \( \begin{array}{ccc}
    d & d & D \ d & DD \\
    0.0 & 1.0 & 1.0
\end{array} \) This allows us to fill in genotypes!

Mother “not informative”

Phase of father unknown.
Linkage analysis in pedigrees

Example: Suppose the disease allele D is fully penetrant, dominant, rare. Suppose phenocopies do not occur.

Penetrance table:  

<table>
<thead>
<tr>
<th></th>
<th>d d</th>
<th>D d</th>
<th>DD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

This allows us to fill in genotypes!

Mother “not informative”

Phase of father unknown.

For simplicity, suppose we know phase of the father is 1 D / 2 d

Then:

0 recombs, 2 non-recombs

\[ L(\theta) = \binom{2}{0} \theta^0 (1 - \theta)^2, \hat{\theta} = 0 \]
Linkage analysis for complex diseases

- Often the mode of inheritance (disease model) is unclear

- Would like to perform analysis without assuming a specific genetic model

- "Model free" approach (covered in Human Genetic Analysis course in the Fall)
Part II: Genetic Association Testing

Simplest statistical test: Compare allele frequency in cases vs controls. Suppose we have $N_{\text{case}}$ cases, $N_{\text{ctrl}}$ controls.

<table>
<thead>
<tr>
<th></th>
<th># of minor alleles</th>
<th># of major alleles</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>$N_{11}$</td>
<td>$N_{12}$</td>
<td>$2*N_{\text{case}}$</td>
</tr>
<tr>
<td>Controls</td>
<td>$N_{21}$</td>
<td>$N_{22}$</td>
<td>$2*N_{\text{ctrl}}$</td>
</tr>
</tbody>
</table>

To test for association, can use chi-squared test, 1 df: $\sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$

(or use Fisher’s exact test)
Part II: Genetic Association Testing

In practice, statistical analysis models used:

Quantitative continuous trait:
- linear regression

Dichotomous trait – e.g. case/control:
- logistic regression
  - more flexible than chi-square / Fisher’s exact test
  - can include covariates
  - provides estimate of odds ratio
Linear regression  
(recall Heather Lawson’s lectures & Gerke et al)

Let $y =$ quantitative trait value. Model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \text{error}$$

$x_1 =$ SNP genotype (e.g. # copies of designated allele: 0,1,2)  
$x_2, \ldots, x_n$ are covariate values (e.g. age, sex, environment)

And let

$$\hat{y} = \alpha + \boxed{\beta_1 x_1} + \beta_2 x_2 + \ldots + \beta_n x_n$$

= predicted quantitative trait value

Null hypothesis $H_0: \beta_1 = 0$.

The SNP “effect size” is represented by $\beta_1$, the coefficient of $x_1$.

The key test: Is there significant evidence that $\beta_1$ is non-zero?
The least squares solution finds $\alpha$ and $\beta$ that minimize the sum of the squared residuals.

Least squares linear regression: general example

$$y = \alpha + \beta x$$

Fitted line, Slope = $\beta$

Residual deviations

Residual = The difference between the observed value and model-predicted value
The least squares solution finds $\alpha$ and $\beta$ that minimize the sum of the squared residuals.

$y = \alpha + \beta x$
SNP Marker Additive Coding:

<table>
<thead>
<tr>
<th>Genotype</th>
<th>$x_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
<td>0</td>
</tr>
<tr>
<td>1/2</td>
<td>1</td>
</tr>
<tr>
<td>2/2</td>
<td>2</td>
</tr>
</tbody>
</table>

Codes number of “2” alleles
Least squares linear regression

\[ y = \alpha + \beta x \]

In this example, as # of coded alleles increases, the trait value increases
\[ y = \alpha + \beta \, x \] 

\( \beta \) is a measure of the EFFECT SIZE for the SNP

Alternative: \( r^2 \) = squared correlation coefficient, indicates proportion of phenotypic variance in \( y \) that’s explained by \( x \)
Logistic regression for dichotomous traits (GWAS)

Let \( y = 1 \) if case, \( 0 \) if control (2 values)

Let \( P = \) probability that \( y = 1 \) (case)

**Let** \( x_1 = \) genotype (additive coding)

Let \( x_2, \ldots, x_n \) denote other covariate variables

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

Why?
Logit function

• Usual regression expects a dependent variable that can take on any value, \((-\infty, \infty)\)
• A probability \(P\) is in \([0,1]\), so not a good dependent variable
• Odds = \(p/(1-p)\) is in \([0,\infty)\)
• Logit = \(\ln(\text{odds})\) is in \((-\infty, \infty)\)
Think of the shapes of the graphs

- \( y = \frac{x}{1-x} \) (x in place of P)

As x varies from 0 to 1, y varies from 0 to \( \infty \)

- \( y = \ln(x) \) varies from \(- \infty \) to \( \infty \)
Logistic regression

Let $y = 1$ if case, $0$ if control (2 possible values)

Let $P =$ probability that $y = 1$ (case)

Note that can exponentiate both sides to get 

$$\text{odds} = \frac{P}{1-P} = e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n} = e^\Omega$$

What about the “effect size”? It’s the “odds ratio”, and it is still related to $\beta_1$!
Odds ratio

• The number $e (=2.718...)$ is the base of natural logarithms

• $e^0 = 1$

• $e^{\beta_1}$ is the odds ratio; if $\beta_1=0$ then odds ratio is 1 (no effect of the SNP)
To get odds ratio per copy of the allele ("effect size")

• Full model: \( \left( \frac{P}{1 - P} \right) = e^{[\alpha + \beta_1 x_1 + \ldots + \beta_n x_n]} \)

• Odds when \( x_1 = 1 \) (1 copy of the allele)
  \( P_1 / (1 - P_1) = e^{[\alpha + \beta_1 x_1 + \ldots + \beta_n x_n]} \bigg|_{x_1=1} = e^{[\alpha_0 + \beta_2 x_2 + \ldots + \beta_n x_n]} + \beta_1 \)

• Odds when \( x_1 = 0 \) (0 copies of the allele)
  \( P_0 / (1 - P_0) = e^{[\alpha + \beta_1 x_1 + \ldots + \beta_n x_n]} \bigg|_{x_1=0} = e^{[\alpha_0 + \beta_2 x_2 + \ldots + \beta_n x_n]} \)

• Odds Ratio:
  \( \left( \frac{P_1 / (1 - P_1)}{P_0 / (1 - P_0)} \right) = \frac{e^{[\alpha + \beta_2 x_2 + \ldots + \beta_n x_n]} + \beta_1}{e^{[\alpha + \beta_2 x_2 + \ldots + \beta_n x_n]}} = e^{\beta_1} \)
Logistic regression summary

Let y = 1 if case, 0 if control (2 values)
Let P = probability that y = 1 (case), ranges from 0 to 1
Then logit(P) ranges from -∞ to ∞

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

Odds ratio

\[ e^{\beta_1} \]

Similar to case of linear regression, can compute an analog to “variance explained,” usually also called \( r^2 \)
Displaying GWAS results: “Manhattan” plot

- x-axis: chromosomal position
- y-axis: $-\log_{10}(p\text{-value})$

So $p = 1 \times 10^{-8}$ is plotted at $y=8$, $p = 5 \times 10^{-8}$ at $y = 7.3$

N = 2,564 (778 “cases,” 1786 “controls”)

Non-daily vs daily smoking
Saccone et al., Nicotine and Tobacco Research 2016
Displaying GWAS results: zooming in

Linkage Disequilibrium

Circle: genotyped SNP
X: imputed SNP

rs77876433

Non-daily vs daily smoking
Saccone et al., Nicotine and Tobacco Research 2016
Displaying GWAS results: Q-Q plot

Quantile-Quantile plot

Idea: Rank tested SNPs by association evidence; compare number of observed associations versus expected associations under the null at a given significance level

Saccone et al., Nicotine and Tobacco Research 2016
Displaying GWAS results
– “Q-Q plot”: Quantile-quantile plot

Helps detect systematic bias in data:
Most datapoints should be close to the y=x line
Exception: signals (lowest, most significant p-values)
Multiple test correction

Why do we correct for multiple tests?

Recall for GWAS, usual p-value threshold requires $p \leq 5 \times 10^{-8}$

Corresponding to $-\log(p) \sim 7.3$

In contrast: threshold of $p \leq 0.05$ corresponds to $-\log(p) \sim 1.3$
Multiple test correction

P-value: probability of observing the data or more "extreme" data when the null hypothesis holds = \( \text{Pr}(\text{as or more extreme} \mid \text{null}) \)

For a single test, traditional required threshold is p-value \( \leq 0.05 \)

**Bonferroni correction**

- If \( n = \) number of independent tests, then under Bonferroni correction, p-value threshold for "significance" is \( 0.05/n \), or more generally, require \( P_{\text{single test}} \leq P_{\text{experimentwide}} / n \)

In the GWAS setting – how to determine the number of "independent tests"?

- Account for linkage disequilibrium (LD), or correlation between alleles (more on this later)
- Typical threshold corrects for 1 million independent tests, requiring \( p \leq 5 \times 10^{-8} \)
Why is multiple test correction important?

- Example: suppose you didn't correct for multiple tests?

10 independent tests, and suppose the null hypothesis holds. What's the probability $P$ that when the null hypothesis holds, we'll still see at least one $p \leq 0.05$ just by chance?

$$P = 1 - (\text{probability that all 10 tests have } p > 0.05 \mid \text{null})$$

$$= 1 - (\text{probability that one test has } p > 0.05 \mid \text{null})^{10}$$

$$= 1 - (1 - 0.05)^{10}$$

$$= 1 - (0.95)^{10} = 1 - 0.5987 = 0.401$$

Greater than 40% chance!
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

1990-2003 Human Genome Project

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

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**

Abstinence

<table>
<thead>
<tr>
<th>Smokers with CHRNA5 low risk haplotype</th>
<th>Smokers with CHRNA5 high risk haplotype</th>
</tr>
</thead>
</table>

This represents a GxE interaction!

**CHRNA5 Predicts Cessation & Response to Medication**

**Smokers with CHRNA5 low risk haplotype**

**Smokers with CHRNA5 high risk haplotype**

**Abstinence**

This represents a GxE interaction!


**Smoking cessation trial, N=1073, European Ancestry**

- **H1: LOW RISK**
- **H2**
- **H3: HIGH RISK**

- Placebo + Counseling
- Medication + Counseling
Prospects for Genomic and Precision Medicine

Prospects for Genomic and Precision Medicine

• NHGRI strategic vision – Green et al. (2020), Nature 586: 683-692

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 genomics 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 “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.
Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research

• National Academies of Sciences, Engineering, Medicine Committee

Dr Ann Morning, NYU Dept of Sociology, gave the WashU Genetics Seminar Seminar March 16, 2023

• 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