Measuring the Phenotype:
What disease endpoint or trait are you studying?

Erin M. Ramos, PhD MPH
Office of Population Genomics
NHGRI

July 18, 2008
Course and Lecture Objectives

- **Course Objective #2**: To understand the various methods, their advantages, and their disadvantages for the definition of phenotypes (disease endpoints, quantitative traits, etc) for use in association studies.

- **Learning Objectives**:
  - Convey the importance of selecting appropriate phenotypes for your genomic research study
  - Describe the properties of a good measure and the consequences of measurement error on study results
  - Consider the advantages of using standard measures for your phenotype of interest
1) Phenotype definition
   1. Discrete verse quantitative traits
   2. Complex disease and natural history of disease
   3. Selecting your phenotype

2) Measurement error
   1. Properties of a good measure
   2. Consequences of measurement error

3) Advantages of using standard phenotypes
   1. Why is it important to use standard measures?
   2. Example of successful cross-study analyses
   3. Introduction to PhenX
1. Phenotype (φαινότυπος)

- Means “the form of what appears”
- Root φαίνειν (phanein) also found in φαινόμενον (phenomenon)
- Also linked to φως, φωτός (light, of the light)
- In order for something to appear, we need light to see it
- A phenotype is the observable expression of an individual’s genotype

‘In writing the history of a disease... [T]he clear and natural phenomenon of the disease should be noted ... accurately, and in all their minuteness; in imitation of the exquisite industry of those painters who represent in their portraits the smallest moles and faintest spots.’

Discrete Trait

- **Discrete/Dichotomous**
  - Two values
  - e.g. Type II Diabetes (No/Yes)
  - Typically of direct clinical relevance (e.g. cancer, hypertension, arthritis)

Distribution of Measured Values for Type II Diabetes

![Bar chart showing distribution of Type II Diabetes cases and controls.](image)


Data from The Finland-United States Investigation of NIDDM Genetics (Fusion) Study
Quantitative Trait

- **Quantitative/Continuous**
  - Range of possible values (e.g. Systolic blood pressure, BMI)
  - Can be reduced to a discrete/dichotomous trait by using a predefined threshold value

---


Data from the GAIN: Search for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes Study
A. **Monogenic Disease.** A variant in a single gene is the primary determinant of a monogenic disease or trait, responsible for most of the disease risk or trait variation (dark blue sector), with possible minor contributions of modifier genes (yellow sectors) or environment (light blue sector).  

B. **Complex disease.** Many variants of small effect (yellow sectors) contribute to disease risk or trait variation, along with many environmental factors (blue sector).

(Manolio TA, et al, JCI, 118: 1590, 2008)
Complex Disease

- Characterized by high levels of genetic complexity; multiple genes may act independently or interact with other genes to influence the phenotype

- Multiple manifestations with varying degrees of genetic influence
  - e.g. Myocardial Infarction, Coronary Artery Atherosclerosis, and Sudden Cardiac Death are forms of Coronary Disease

- Multiple causes, which may be attributed to separate or overlapping genetic influences
  - e.g. Atherosclerosis caused by lipid accumulation, inflammation, endothelial disruption, and thrombosis.
Complex Disease

- Difficult to distinguish individuals with “sub-clinical” disease from “non-diseased” individuals if early stage diagnosis is inadequate

- Characterized by variable age of onset of clinical symptoms

- Environmental factors may modify the genotype-phenotype relationship; thus, disease expression range from nearly undetectable to severely debilitating

(Ellsworth DL and Manolio TA, Annals of Epi, 1999)
Natural history of disease

Non-diseased → Sub-Clinical Phase → Clinical Phase → Outcome:
- Cure
- Control
- Disability
- Death

Some Sources of Data:
- Interviews
- Clinical records
- Hospital records

Time

(Adapted from Gordis, 3rd Ed, 2004)
Some Limitations of Hospital Data

- Hospital admissions are selective in relation to:
  - Demographics
  - Severity of disease
  - Associated conditions
  - Admission patterns

- Hospital records are not typically designed for research. They may be:
  - Incomplete, illegible, or missing
  - Variable in diagnostic quality

- Populations at risk are not generally defined
Some Limitations of Clinical Data

- Can be a rich source of patient specific data (clinical exam, diagnostic tests, and procedures), but...
- Chart extraction can be difficult
- Patients might receive care from additional sources
- Uneven organization, incompleteness, legibility, etc...
- Clinical diagnostic criteria can vary and change over time
Some Limitations of Interview Data

- The respondent:
  - Has the disease, but does not have symptoms and does not report the disease
  - Has the disease, sought medication attention, but reports a different disease
  - May provide disease information accurately, but it is recorded inaccurately

- The interviewer may know the hypothesis being tested, thus probing more intensively in one group of respondents than another

- Incomplete or missing data
Selecting your phenotype

Goal: Reduce heterogeneity in your phenotype to increase your chance of finding genes!!

- What disease/trait interests you?
- Evidence for genetic influence on your disease/trait of interest
- Homogenous cases (highly specific disease criteria)
- Intermediate phenotypes (closer proximity to genes)
Evidence for genetic influence

Familial Clustering:
- Risk of disease in relative of case > risk in relative of non-case or general population
- Discrete Trait: Familial relative risk, Risch’s $\lambda_s$
- Continuous Trait: parent-offspring correlation & sib-sib correlation

- Twin studies
  - Comparing Concordance between Monozygotic Twins and Dizygotic Twins
Association of rs10033464 & Atrial Fibrillation (AF)

Variants conferring risk of atrial fibrillation on chromosome 4q25

Daniel F. Gudbjartsson¹, David O. Arnar², Anna Helgadottir¹, Solveig Gretarsdottir¹, Hilma Holm²,
Asgeir Sigurdsson¹, Adalbjorg Jonasdottir¹, Adam Baker¹, Gudmar Thorleifsson¹, Kristleifur Kristjansson¹,
Arnar Palsson¹, Thorarinn Blondal¹, Patrick Sulem¹, Valgerdur M. Backman¹, Gudmundur A. Hardarson¹,
Ebba Palsdottir¹, Agnar Helgason¹, Runa Sigurjonsdottir², Jon T. Sverrisson³, Konstantinos Kostulas⁴,
Maggie C. Y. Ng⁵, Larry Baum⁵, Wing Yee So⁵, Ka Sing Wong⁵, Juliana C. N. Chan⁵, Karen L. Furie⁶,
Steven M. Greenberg⁶, Michelle Sale⁶, Peter Kelly⁶, Calum A. MacRae⁷, Eric E. Smith⁶, Jonathan Rosand⁶,
Jan Hillert⁴, Ronald C. W. Ma⁶, Patrick T. Ellinor⁷, Gudmundur Thorgeirsson⁷, Jeffrey R. Gulcher¹, Augustine Kong¹,
Unnur Thorsteinsdottir¹ & Kari Stefansson¹

- **Discovery Study (Iceland Cases):**
  - All cases of AF at two large hospitals

- **Replication Study (U.S. Cases):**
  - Younger patients with lone AF
  - AF with co-existing hypertension
  - Stroke patients with AF

(Gudbjartsson DF et al, Nature, June 2007)
# Association of rs10033464* & Atrial Fibrillation (AF)

<table>
<thead>
<tr>
<th>Case / Control</th>
<th>Mean Age (yr)</th>
<th>OR</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iceland</td>
<td>2801 / 17,714</td>
<td>1.40</td>
<td>9.4 x 10^{-9}</td>
</tr>
<tr>
<td>United States</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lone AF</td>
<td>251 / 804</td>
<td>46.1</td>
<td>1.68</td>
</tr>
<tr>
<td>AF &amp; Hyp</td>
<td>67 / 804</td>
<td>54.5</td>
<td>1.66</td>
</tr>
<tr>
<td>Other AF</td>
<td>318 / 804</td>
<td>75.2</td>
<td>0.97</td>
</tr>
</tbody>
</table>

OR = Odds Ratio; *PITX2 gene, known to be involved in early heart development

(Gudbjartsson DF et al, Nature, June 2007)
Late-Onset Alzheimer’s Disease (LOAD)

- Discovery set: clinically & neuropathologically confirmed LOAD cases and neuropathologically normal controls

- Rationale: to exclude misdiagnosed cases & cognitively normal controls who have LOAD neuropathology

## Association of GAB2 alleles & LOAD

<table>
<thead>
<tr>
<th>Stage</th>
<th>Cases (N)</th>
<th>Controls (N)</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>446</td>
<td>290</td>
<td>312,316</td>
</tr>
<tr>
<td></td>
<td>Neuropathology Discovery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>197</td>
<td>114</td>
<td>“</td>
</tr>
<tr>
<td></td>
<td>Neuropathology Replication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>218</td>
<td>146</td>
<td>“</td>
</tr>
<tr>
<td></td>
<td>Clinical Replication*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>861</strong></td>
<td><strong>550</strong></td>
<td><strong>7</strong></td>
</tr>
</tbody>
</table>

* To confirm genetic association independent of brain donor selection bias

# Association of GAB2 alleles & LOAD

<table>
<thead>
<tr>
<th>SNP</th>
<th>P-Value</th>
<th>Freq. in Controls</th>
<th>OR</th>
<th>[95%CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1385600</td>
<td>2.81 E-09</td>
<td>0.71</td>
<td>3.65</td>
<td>[2.34,5.71]</td>
</tr>
<tr>
<td>rs1007837</td>
<td>3.97 E-07</td>
<td>0.73</td>
<td>3.01</td>
<td>[1.94,4.68]</td>
</tr>
<tr>
<td>rs4945261</td>
<td>3.08 E-08</td>
<td>0.72</td>
<td>3.44</td>
<td>[2.18,5.43]</td>
</tr>
<tr>
<td>rs10793294</td>
<td>1.59 E-07</td>
<td>0.66</td>
<td>2.83</td>
<td>[1.90,4.21]</td>
</tr>
<tr>
<td>rs4291702</td>
<td>5.88 E-07</td>
<td>0.70</td>
<td>2.96</td>
<td>[1.91,4.59]</td>
</tr>
<tr>
<td>rs7115850</td>
<td>2.80 E-10</td>
<td>0.67</td>
<td>3.92</td>
<td>[2.51,6.11]</td>
</tr>
<tr>
<td><strong>rs2373115</strong></td>
<td><strong>9.66 E-11</strong></td>
<td><strong>0.70</strong></td>
<td><strong>4.06</strong></td>
<td><strong>[2.81,14.69]</strong></td>
</tr>
</tbody>
</table>

* Sample Size = 861 Cases & 550 Controls

**Rs2373115 interacts w/ APOE to modify risk**

<table>
<thead>
<tr>
<th>APOE*e4 Group</th>
<th>APOE*e4 OR [95% CI]</th>
<th>rs2373115 OR [95%CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOE*e4 -</td>
<td>1.12 [0.82,1.53]</td>
<td></td>
</tr>
<tr>
<td>APOE*e4 +</td>
<td>2.88 [1.90,4.36]</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>6.07 [4.63-7.95]</td>
<td>1.34 [1.06,1.70]</td>
</tr>
</tbody>
</table>

OR = Odds Ratio; ORs compare GG to GT/TT

Suggests GAB2 modifies LOAD risk in APOE e4 carriers

Intermediate Phenotypes

- Phenotype that is heritable, measurable, and has a closer relationship to the biological process involved in culmination of disease
- Represents a more elementary phenomenon
- The # of genes affecting intermediate phenotype variation is smaller than the number of genes affecting the full disease/trait phenotypic variation
- The genes affecting intermediate phenotypes have larger effect size
- **For an intermediate phenotype to be useful, it should be heritable & associated with disease/trait of interest!**
Intermediate Phenotypes

P = Gene Product; LL = Low Level Phenotype; INT = Intermediate Phenotype; HL = High Level Phenotype

Alzheimer’s Disease

Disease (Clinical Signs)

Cognitive function (MMSE score)

Hippocampal atrophy (from MRI)

A-Beta protein level

Environmental Contribution

P = Gene Product; LL = Low Level Phenotype; INT = Intermediate Phenotype; HL = High Level Phenotype

(Schork, NJ, Am J Respir Crit Care Med, 1997)
Lecture Outline

1) Phenotype definition
   1. Discrete verse quantitative traits
   2. Complex disease and natural history of disease
   3. Selecting your phenotype

2) Measurement error
   1. Properties of a good measure
   2. Consequences of measurement error

3) Advantages of using standard phenotypes
   1. Why is it important to use standard measures?
   2. Example of successful cross-study analyses
   3. Introduction to PhenX
2. Measurement Error

- Measure refers broadly to any way of capturing data on a certain characteristic of study subjects.
- Method could be self-administered questionnaire, personal interview, physical exam, lab test, medical records extraction, etc.
- Regardless of characteristic or data collection method, there is a TRUE value of the characteristic for each study subject.
- Any discrepancy between the TRUE value and the MEASURED value is Measurement Error.
Properties of a good measure

- **Reliability**
  - describes consistency, reproducibility of a measurement
  - A good measurement should yield the same value if applied repeatedly under similar conditions

- **Validity**
  - describes accuracy of a measurement
  - A good measurement should yield the correct value/reflect the truth

- Reliability is a prerequisite for validity

- Reliability is necessary, but not a sufficient condition for validity
Properties of a good measure

- The goal is to hit the Bullseye with each dart:
  - Results are neither reliable or valid
  - Results are reliable, but not valid
  - Results are both reliable and valid

Modified from D.S. Bhola, PhD
Quantifying Reliability

- Discrete/Categorical Traits
  - To what degree do the measurements agree beyond what we would expect by chance alone?
  - Kappa ($\kappa$) ranges from 0 to 1
  - Guidelines for Interpretation of Kappa (Source: Landis & Koch, 1977)

- Kappa | Interpretation
--- | ---
> .80 | Almost perfect
.61-.80 | Substantial
.41-.60 | Moderate
.21-.40 | Fair
.00-.20 | Slight
Kappa (κ)

Data layout for Calculating Kappa

<table>
<thead>
<tr>
<th>Measure #1</th>
<th>Measure # 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>-</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

\[ \kappa = \frac{P_o - P_e}{1 - P_e} \]

Where:

- \( P_o \) = observed concordance (% agreement observed)
  \( \frac{(a+d)}{N} \)

- \( P_e \) = concordance expected by chance (% agreement expected by chance alone)
  \[ \frac{(a+b)(a+c)}{N} + \frac{(b+d)(c+d)}{N} \] / N
Kappa (κ): Example

Wright and colleagues (2000) studied genital-tract human papillomavirus (HPV) testing as possible screening test for cervical cancer. The examined agreement between test results on swabs obtained by clinicians with swabs obtained by screeners themselves. For 1415 women, both kinds of specimens were obtained:

<table>
<thead>
<tr>
<th>Clinician collected</th>
<th>Self-collected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>170</td>
<td>132</td>
</tr>
<tr>
<td>-</td>
<td>128</td>
<td>985</td>
</tr>
<tr>
<td>Total</td>
<td>298</td>
<td>1117</td>
</tr>
</tbody>
</table>

\[ P_0 = \text{observed concordance} \]
\[ (170+985) / 1415 = 0.816 \]

\[ P_e = \text{concordance expected by chance} \]
\[ \frac{(302)(298) + (1113)(1117)}{1415} = 0.666 \]

\[ \kappa = P_0 - P_e / 1 - P_e \]
\[ \kappa = 0.816 - 0.666 / 1-0.666 \]
\[ = 0.45, \text{moderate agreement} \]

(from Koepsell & Weiss, page 221)
Quantifying Validity

- True status of characteristic of interest must be known ("gold standard")
- Compare measure of your characteristic of interest to the gold standard

Data layout for assessing validity of binary test

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Condition present</th>
<th>a = # of true positives</th>
<th>b = # of false positives</th>
<th>c = # of false negatives</th>
<th>d = # of true negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>a</td>
<td>b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>c</td>
<td>d</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity = when condition truly present, how often does the test detect it?
= \( \frac{a}{a+c} \)

Specificity = when condition is truly absent, how often does test give a neg. result?
= \( \frac{d}{b+d} \)

(from Koepsell & Weiss, page 223)
Consequences of Measurement Error

- Impact of measurement error on results depends on the way the error has arisen
- Measurement error of a discrete/binary outcome is termed misclassification
- Nondifferential (nonselective) misclassification of outcome
  - Present whether errors in assessing subject’s status are similar regardless whether that subject has been exposed or not
  - Generally leads to an attenuation of the estimated size of a true association between exposure and disease
  - i.e. bias towards null
- Improving the resolution of measurement tools will allow more accurate characterization of the relationship between exposure (genotype) and disease!
Lecture Outline

1) Phenotype definition
   1. Discrete verse quantitative traits
   2. Complex disease and natural history of disease
   3. Selecting your phenotype

2) Measurement error
   1. Properties of a good measure
   2. Consequences of measurement error

3) Advantages of using standard phenotypes
   1. Why is it important to use standard measures?
   2. Example of successful cross-study analyses
   3. Introduction to PhenX
3. Genome-Wide Association (GWA) Studies

- GWA studies measure > 100,000 single nucleotide polymorphisms (SNPs) across the genome & relate them to common diseases and traits
- Since 2005, over 160 GWA studies have identified robust SNP associations (P < 10^{-7}) for nearly 60 diseases and traits

- Type 2 Diabetes
- 386,731 markers

(http://www.broad.mit.edu/diabetes/scandinavs/type2.html)
Unique Aspects of GWA Studies

- Permit examination of genetic variation at an unprecedented level of resolution
- Allow “agnostic” genome-wide evaluation
- Once genome measured, can be related to any trait
- Most robust associations in GWAS reports have not been with genes previously suspected of being related to the disease
- Some significant associations are in regions that are not currently known to harbor genes

“The chief strength of the new approach also contains its chief problem: with more than 500,000 comparisons per study, the potential for false positive results is unprecedented.”

“Thus, the sine qua non for belief in any specific result from a GWAS is not the strength of the P value in the initial study, but the consistency and strength of the association across one or more large-scale replication studies.”

Courtesy, Teri Manolio, NHGRI (Hunter DJ and Kraft P, NEJM, 2007)
Cross-Study Analysis is Essential

- More bang for the buck!
  - GWA and related studies are expensive
  - Combining studies increases ability to detect loci with moderate effect size
  - Once genome is characterized it can be related to traits beyond those focused on in the initial study (with appropriate consent)

- Potential for cross-study analysis limited by lack of standardized measures being included in GWAS
  - despite many risk factors common to multiple diseases (e.g. obesity, smoking, etc)
## Association of rs1042725 (HMGA2) & height

<table>
<thead>
<tr>
<th>Study</th>
<th>Mean Age</th>
<th>N</th>
<th>Mean height (cm) by genotype</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>TT</td>
<td>CT</td>
</tr>
<tr>
<td>GWA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WTCCC (T2D)</td>
<td>57.9</td>
<td>792</td>
<td>160.4</td>
<td>161.5</td>
</tr>
<tr>
<td>DGI (T2D)</td>
<td>65.2</td>
<td>638</td>
<td>160.0</td>
<td>161.3</td>
</tr>
<tr>
<td>DGI (Controls)</td>
<td>58.5</td>
<td>546</td>
<td>162.1</td>
<td>162.8</td>
</tr>
<tr>
<td>Combined</td>
<td>&gt;4K</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Weedon et al, Nature Genet 2007; 39:1245-50)
## Association of rs1042725 (HMGA2) & height

<table>
<thead>
<tr>
<th>Study (women only)</th>
<th>Mean Age</th>
<th>N</th>
<th>Mean height (cm) by genotype</th>
<th>Mean height (cm) by genotype</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>TT</td>
<td>CT</td>
<td>CC</td>
</tr>
<tr>
<td>GWA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WTCCC (T2D)</td>
<td>57.9</td>
<td>792</td>
<td>160.4</td>
<td>161.5</td>
<td>162.2</td>
</tr>
<tr>
<td>DGI (T2D)</td>
<td>65.2</td>
<td>638</td>
<td>160.0</td>
<td>161.3</td>
<td>162.1</td>
</tr>
<tr>
<td>DGI (Controls)</td>
<td>58.5</td>
<td>546</td>
<td>162.1</td>
<td>162.8</td>
<td>163.7</td>
</tr>
<tr>
<td>Combined</td>
<td>&gt;4K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Replication              |          |     |                              |                              |         |
| UKGCC T2D                | 64.0     | 820 | 159.0                        | 159.3                        | 159.9   | 0.037   |
| EFSOCH parents           | 32.9     | 936 | 164.6                        | 165.0                        | 165.4   | 0.004   |
| Combined                 | >19K     |     |                              |                              |         | 3x10^-11|

(Weedon et al, Nature Genet 2007; 39:1245-50)
## Association of rs1042725 (HMGA2) & height

<table>
<thead>
<tr>
<th>Study (women only)</th>
<th>Mean Age</th>
<th>N</th>
<th>Mean height (cm) by genotype</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>TT</td>
<td>CT</td>
</tr>
<tr>
<td><strong>GWA</strong></td>
<td></td>
<td></td>
<td>160.4</td>
<td>161.5</td>
</tr>
<tr>
<td>WTCCC (T2D)</td>
<td>57.9</td>
<td>792</td>
<td>160.4</td>
<td>161.5</td>
</tr>
<tr>
<td>DGI (T2D)</td>
<td>65.2</td>
<td>638</td>
<td>160.0</td>
<td>161.3</td>
</tr>
<tr>
<td>DGI (Controls)</td>
<td>58.5</td>
<td>546</td>
<td>162.1</td>
<td>162.8</td>
</tr>
<tr>
<td>Combined</td>
<td>&gt;4K</td>
<td></td>
<td>161.5</td>
<td>161.3</td>
</tr>
</tbody>
</table>

| Replication                |          |     | 162.2          | 162.8    | 163.7    |           |
|                            |          |     | 0.0006         | 0.003    | 0.003    |           |
| UKGCC T2D                  | 64.0     | 820 | 159.0          | 159.3    | 159.9    | 0.037     |
| EFSOCH parents             | 32.9     | 936 | 164.6          | 165.0    | 165.4    | 0.004     |
| Combined                   | >19K     |     | 162.2          | 162.8    | 163.7    | 3x10^-11  |
| **All studies**            | >23K     |     | effect size/C allele ~0.4cm | 4x10^-16 |

(Weedon et al, Nature Genet 2007; 39:1245-50)
**FTO Variant (rs9939609), T2 Diabetes, & Obesity**

<table>
<thead>
<tr>
<th>Study</th>
<th>Cases</th>
<th>Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTCCC (TD2 Ph 1)</td>
<td>1,924</td>
<td>2,938</td>
<td>1.27</td>
<td>[1.16-1.37]</td>
<td>5 x 10(^{-8})</td>
</tr>
<tr>
<td>WTCCC (TD2 Ph 2)</td>
<td>3,757</td>
<td>5,346</td>
<td>1.15</td>
<td>[1.09-1.23]</td>
<td>9 x 10(^{-6})</td>
</tr>
<tr>
<td>- Adjusted for BMI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.44</td>
</tr>
</tbody>
</table>

Association of rs9939609 with T2D risk mediated through BMI

<table>
<thead>
<tr>
<th>Study</th>
<th>% ♂</th>
<th>N</th>
<th>TT</th>
<th>AT</th>
<th>AA</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTCCC TD2 cases</td>
<td>58</td>
<td>1,913</td>
<td>30.2</td>
<td>30.5</td>
<td>32.0</td>
<td>8 x 10(^{-6})</td>
</tr>
<tr>
<td>UKGCC TD2 cases</td>
<td>57</td>
<td>2,961</td>
<td>30.6</td>
<td>31.0</td>
<td>32.0</td>
<td>3 x 10(^{-5})</td>
</tr>
<tr>
<td>EFSOCH controls</td>
<td>51</td>
<td>1,746</td>
<td>24.5</td>
<td>25.2</td>
<td>25.4</td>
<td>0.0002</td>
</tr>
<tr>
<td>EPIC-Norfolk (pop-based)</td>
<td>47</td>
<td>2,425</td>
<td>25.9</td>
<td>26.2</td>
<td>26.6</td>
<td>0.001</td>
</tr>
<tr>
<td>All studies (Bonferonni correction)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.2x10(^{-29})</td>
</tr>
</tbody>
</table>

Frayling et al, Science 2007; 316:889-893
PhenX

consensus measures for Phenotypes and exposures

Building consensus for standard measures of phenotypes and exposures

- PhenX is a three-year project led by RTI International and funded by the National Human Genome Research Institute (NHGRI) to contribute to the integration of genetics and epidemiologic research
- PhenX will prioritize up to 20 research domains related to complex diseases and environmental exposures
- Consensus building will lead to a recommended minimal set of standard measures for use in Genome-wide Association Studies (GWAS) and other large-scale genomic research efforts
- Standard measures will maximize benefits of future research by enabling cross-study comparisons and analysis
- Selection and specification of the measures will be driven by the scientific community via the PhenX Steering Committee, Working Groups, and Surveys
- The PhenX Toolkit will make the standard measures available to the scientific community

STEERING COMMITTEE
A Steering Committee of distinguished experts from the scientific community will guide the selection of the measures and promote their use. Domains may include diseases and conditions, lifestyle factors, and environmental and occupational exposures.

WORKING GROUPS
Working Groups will be constituted for specific domains for the purpose of identifying a small set of measures and corresponding methods for measurement. The measures will be vetted with the scientific community through periodic surveys accessed through this web site.

SURVEYS
Surveys will be periodically available on this web site for the scientific community to review and comment on selected measures. The Demographics Survey is now available at www.phenx.org/surveys.

PHENX TOOLKIT
The PhenX Toolkit will make the results of the project readily accessible via the Internet and enable researchers to implement the standard measures.
PhenX Domains

- Aging
- Alcohol, Tobacco, and Other Substances*
- Anthropometrics*
- Cancer
- Cardiovascular
- Central Nervous System
- Demographics*
- Child development
- Diet
- Diabetes

- Exposures & Responses
- Gastrointestinal
- Immunity
- Lung Function
- Ocular
- Oral Health
- Physical Activity
- Psychosocial
- Renal Function
- Reproduction
- Skin/Bone/Muscle
Demographic Measures

- Age
- Ancestry
- Race/Ethnicity
- Sex/Gender
- Current Marital Status
- Current Employment Status
- Education
- Income/Wealth
- Health Care
- Years in the U.S.
Summary Points

- Selecting appropriate phenotypes for your genomic research study is important
- Use reliable and valid measures to capture the information about your disease/trait and relevant covariates
- To increase potential for cross-study analysis, think about using commonly used measures with standard assessment protocols
Take home message...

PHENOTYPE,

PHENOTYPE,

PHENOTYPE!!!
Association of rs563694 & fasting glucose

Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels

Wei-Min Chen,1,2 Michael R. Erdos,3 Anne U. Jackson,4 Richa Saxena,5 Serena Sanna,4,6 Kristi D. Silver,7 Nicholas J. Timpson,8 Torben Hansen,9 Marco Orrù,6 Maria Grazia Piras,5 Lori L. Bonnycastle,3 Cristen J. Willer,4 Valeriya Lyssenko,10 Haiqing Shen,7 Johanna Kuusisto,11 Shah Ebrahimi,12 Natascia Sestu,13 William L. Duren,4 Maria Cristina Spada,6 Heather M. Stringham,4 Laura J. Scott,4 Nazario Olla,6 Amy J. Swift,3 Samer Najjar,13 Braxton D. Mitchell,7 Debbie A. Lawlor,8 George Davey Smith,8 Yoav Ben-Shlomo,14 Gitte Andersen,9 Knut Borch-Johnsen,9,15,16 Torben Jørgensen,15 Jouko Saramies,17 Timo T. Valle,18 Thomas A. Buchanan,19,20 Alan R. Shuldiner,7 Edward Lakatta,13 Richard N. Bergman,20 Manuela Uda,9 Jaakko Tuomilehto,18,21 Oluf Pedersen,9,18 Antonio Cao,6 Leif Groop,10 Karen L. Mohlke,22 Markku Laakso,11 David Schlessinger,13 Francis S. Collins,3 David Altschuler,5 Gonçalo R. Abecasis,4 Michael Boehnke,4 Angelo Scuteri,23,24 and Richard M. Watanabe20,25

- Rationale: Understanding genetic variants that regulate fasting glucose concentrations may further our understanding of the pathogenesis of diabetes

(Chen, W et al., JCI, July, 2008)
## Association of rs563694 & fasting glucose*

<table>
<thead>
<tr>
<th>Study</th>
<th>n</th>
<th>Mean fasting glucose (mM)</th>
<th>Effect Size (mM)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CC</td>
<td>AC</td>
<td>AA</td>
</tr>
<tr>
<td>FUSION stage I</td>
<td>1,233</td>
<td>5.26</td>
<td>5.31</td>
<td>5.33</td>
</tr>
<tr>
<td>SardiNIA</td>
<td>3,855</td>
<td>4.88</td>
<td>4.95</td>
<td>5.00</td>
</tr>
<tr>
<td>GWA meta analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up FUSION</td>
<td>655</td>
<td>5.28</td>
<td>5.44</td>
<td>5.46</td>
</tr>
<tr>
<td>Amish</td>
<td>1,655</td>
<td>4.90</td>
<td>4.89</td>
<td>5.03</td>
</tr>
<tr>
<td>METSIM</td>
<td>4,386</td>
<td>5.55</td>
<td>5.64</td>
<td>5.71</td>
</tr>
<tr>
<td>Follow-up meta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall meta analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* In non-diabetic individuals

(Chen, W *et al.*, JCI, July, 2008)
Association of rs563694 & fasting glucose*

<table>
<thead>
<tr>
<th>Study</th>
<th>n</th>
<th>Mean fasting glucose (mM)</th>
<th>Effect Size (mM)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CC</td>
<td>AC</td>
<td>AA</td>
</tr>
<tr>
<td>GWA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUSION stage I</td>
<td>1,233</td>
<td>5.26</td>
<td>5.31</td>
<td>5.33</td>
</tr>
<tr>
<td>SardiNIA</td>
<td>3,855</td>
<td>4.88</td>
<td>4.95</td>
<td>5.00</td>
</tr>
<tr>
<td>GWA meta analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUSION stage II</td>
<td>655</td>
<td>5.28</td>
<td>5.44</td>
<td>5.46</td>
</tr>
<tr>
<td>Amish</td>
<td>1,655</td>
<td>4.90</td>
<td>4.89</td>
<td>5.03</td>
</tr>
<tr>
<td>METSIM</td>
<td>4,386</td>
<td>5.55</td>
<td>5.64</td>
<td>5.71</td>
</tr>
<tr>
<td>Follow-up meta analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall meta analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Concluded that G6PC2, a glucose-6-phosphatase (expressed in pancreatic cells), may underlie variation in fasting glucose