Genomic Approaches to the Study of Complex Genetic Diseases

Karen Mohlke, PhD
Department of Genetics
University of North Carolina
April 20, 2016

No Relevant Financial Relationships with Commercial Interests
Complex diseases & traits

Gene mapping in populations

Genome-wide association studies

- Test a large portion of the common single nucleotide genetic variation in the genome for association with a disease or variation in a quantitative trait

- Find disease/quantitative trait-related variants without a prior hypothesis of gene function

Genetic architecture

- Rare alleles causing Mendelian disease
- Few examples of high-effect common variants influencing common disease
- Low-frequency variants with intermediate effect
- Common variants implicated in common disease by GWA
- Rare variants of small effect very hard to identify by genetic means
- Very rare, Rare, Low frequency, Common allele frequencies

Genome-wide association studies identify loci

Outline

- Genome-wide association study design
  - Samples/study participants
  - Genotyping
  - Tests of association
  - Imputation and meta-analysis
- Interpretation of results
  - Effect size and significance
  - Example locus characteristics
- Sequencing/rare variant studies
Study designs

Population-based cohort

- Enroll subjects regardless of health or disease

Prospective cohort

- Enroll subjects; measure X, Y, Z over time, wait for disease onset

Case-control

- What happened prior to disease onset?
- Identify/enroll cases and controls

Matching of cases and controls

Cases and controls should be comparable in all respects except disease status (e.g. age, sex, demographics)
Selection of cases

- Potential criteria to enrich genetic effect size
  - More severely affected individuals
  - Require other family member to have disease
  - Younger age-of-disease onset

Cases

Selection of controls

- Potential criterion to enrich genetic effect size
  - Low risk of disease rather than population-based samples

Controls
Comparable ancestry

Cases

Controls

Ancestry differences

Cases

Controls

May have inadequate ancestry information prior to genotyping
Confounding and population stratification

Confounding

True risk factor

Exposure of interest

Confounded association

Disease

Correlation, not causal

Causal

Population stratification

Ethnicity

Genotype of interest

Correlation, not causal

Correlation, not causal

Genotype of interest

Correlation, not causal

Causal

Disease

Population stratification

- Systematic differences in allele frequencies between subpopulations that may be due to different ancestry
- Oversampled individuals from one subpopulation for cases in a case-control genetic association study can produce spurious associations

Cancer Epidemiol Biomarkers Prev 11: 513
Account for or avoid population stratification

- Match cases with controls
- Restrict to one subgroup
- Adjust for genetic background
  E.g. Use principle components (PCs) to infer ancestry from genotype data and adjust for PCs in association analysis
- Family-based study design – genotype relatives and analyze transmission of alleles from heterozygous parents to offspring
  Transmission disequilibrium test (TDT), family-based association test (FBAT)

Genome-wide genotyping panels

- 10,000 - 5 million variants
- Affymetrix, Illumina
  - Random SNPs
  - Selected haplotype tag variants
  - Copy number probes
  - More lower frequency variants
  - Exome variants
  - Some arrays allow variants to be added
Selecting ‘haplotype tag’ SNPs

a SNPs

\[\begin{align*}
\text{SNP} & \rightarrow \text{SNP} & \rightarrow \\
\text{SNP} & \rightarrow \\
\end{align*}\]

AACACGCCA\ldots TTCGGGGTC\ldots AGTCGACCG\ldots \\
AACACGCCA\ldots TTCAGGTC\ldots AGTCACCCG\ldots \\
AACATGCCA\ldots TTCGGGGTC\ldots AGTCACCCG\ldots \\
AACACGCCA\ldots TTCGGGGTC\ldots AGTCGACCG\ldots \\

b Haplotypes

- Haplotype 1: CTCAGTACGGTTAGGCA
- Haplotype 2: TTAGTCGCAAACGTAATA
- Haplotype 3: CCCATCTTGATACCTGGTG
- Haplotype 4: TGCTAACCGGTTCAGACA

\[\begin{align*}
\text{Tag SNPs} & \rightarrow \text{Tag SNPs} \\
\text{A} & \rightarrow \text{T} & \rightarrow \\
\text{G} & \rightarrow \\
\end{align*}\]


Illumina Infinium Assay

Whole genome amplification

\[\text{Fragment DNA} \rightarrow \text{gDNA} \rightarrow \text{BeadArray of capture probes}\]

Allele-specific primer extension with labeled nucleotides

\[\begin{align*}
\text{Bead type} & \rightarrow \text{Two haptons/colors} \\
\text{Captured human gDNA} & \\
\end{align*}\]

Dual-color fluorescent staining
Detect fluorescent color and intensity

Illumina.com and adapted from Gunderson (2005) NatGen 37:549
**Affymetrix Axiom Array**

- **Target prep**: Amplify
  - 25-125 bp

- **Hybridization**: Capture 30-mers

- **Ligation**: Differentiate
  - Ligase closes the gap between capture and label probe if complete complementarity; wash off others

- **Signal amplification**: Stain and image
  - Stain to detect label
  - GeneTitan platform

---

**Global genomic coverage**

Global coverage (%) by SNP chips

<table>
<thead>
<tr>
<th>SNP chip</th>
<th>CEU</th>
<th>CHB+JPT</th>
<th>YRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP Array 5.0</td>
<td>64</td>
<td>66</td>
<td>41</td>
</tr>
<tr>
<td>SNP Array 6.0</td>
<td>83</td>
<td>84</td>
<td>62</td>
</tr>
<tr>
<td>HumanHap300</td>
<td>77</td>
<td>66</td>
<td>29</td>
</tr>
<tr>
<td>HumanHap550</td>
<td>87</td>
<td>83</td>
<td>50</td>
</tr>
<tr>
<td>HumanHap650Y</td>
<td>87</td>
<td>84</td>
<td>60</td>
</tr>
<tr>
<td>Human1M</td>
<td>93</td>
<td>92</td>
<td>68</td>
</tr>
</tbody>
</table>

Percent of SNPs present on the chip or tagged at $r^2 > 0.8$ by at least one SNP in the chip within 250 kb

*Li (2008) EJHG 16:625*
Quality control: Identify and remove bad samples

- Poor quality samples
  - Sample success rate < 95%  
  - Excess heterozygous genotypes
- Sample switches
  - Wrong sex
- Unexpected related individuals
  - Pair-wise comparisons of genotype similarity
  - Duplicates
- Ancestry different from the rest of sample

Quality control: Identify and remove bad SNPs

- Genotyping success rate < 95%
- Different genotypes in duplicate samples
- Expected proportions of genotypes are not consistent with observed allele frequencies
- Non-Mendelian inheritance in trios
- Differential missingness in cases and controls
Quality control: Identify and remove bad SNPs

Ideal genotyping plot  Clusters mis-called  Clusters overlap

Statistical analysis: linear regression
Two main parameters: p-value and effect size

\[ y = \beta_0 + \beta_1 x \]

Trait = \( \beta_0 + \beta_1 \text{SNP}_1 \)

Toe size = \( \beta_0 + \beta_1 \text{rs123456} \)
Statistical analysis: linear regression

Two main parameters: p-value and effect size

\[ y = \beta_0 + \beta_1 x \]

\[ \text{Trait} = \beta_0 + \beta_1 \text{SNP}_1 \]

\[ \text{Toe size} = \beta_0 + \beta_1 \text{rs123456} \]

\[ \text{Toe size} = \beta_0 + \beta_1 \text{rs123456} + \beta_2 \text{sex} + \beta_3 \text{age} + \beta_4 \text{age}^2 + \beta_5 \text{BMI} \]

- **Assumptions**
  - Trait is normally distributed for each genotype, with a common variance
  - Subjects independent (e.g. unrelated)

---

**Odds ratio**

- Surrogate measure of effect of allele on risk of developing disease

<table>
<thead>
<tr>
<th>Allele</th>
<th>A</th>
<th>C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>860</td>
<td>1140</td>
<td>2000</td>
</tr>
<tr>
<td>Control</td>
<td>1000</td>
<td>1000</td>
<td>2000</td>
</tr>
<tr>
<td>Total</td>
<td>1860</td>
<td>2140</td>
<td>4000</td>
</tr>
</tbody>
</table>

Odds of C allele given case status = \( \frac{\text{Case C}}{\text{Case A}} \)

Odds of C allele given control status = \( \frac{\text{Control C}}{\text{Control A}} \)

Odds Ratio = \( \frac{\text{Case C} / \text{Case A}}{\text{Control C} / \text{Control A}} = \frac{1140 / 860}{1000 / 1000} = 1.33 \)
Association study odds ratio plot

References
Hedegård, Nat Genet, 2007
Grettí, Nat Genet, 2006
Fleming, Nat Genet, 2007
Suzuki, Nat Genet, 2005
Suzuki, Diabetes, 2004
Suzuki, Diabetes, 2004
Suzuki, Diabetes, 2005
Suzuki, Diabetes, 2004

Relationship between GWAS sample size and power

Adjust for population structure: genomic control

- With population structure, the distribution of Cochran-Armitage trend tests, genome-wide, is inflated by a constant multiplicative factor $\lambda$.

- That factor can be estimated from the association results $\lambda = \text{median}(X_i^2)/0.456$.

- Inflation factor $\lambda > 1$ indicates population structure, unknown relatives or other errors.

- The tests of association can be adjusted by this factor. $X_i^2_{\text{adjusted}} = X_i^2/\lambda$

![Quantile-Quantile (Q-Q) plot](Image)


---

'Manhattan plot’ for HDL-cholesterol

- Novel HDL loci
- GWS HDL loci
- GWAS, New for HDL

Global Lipids Genetics Consortium
188,577 individuals from 60 studies, GWAS + metabochip variants

GLGC (2013) Nat Gen 45:1274
Multiple testing

- Genotype and test > 300K – 5M SNPs
- Correct for the multiple tests

\[ P-value = 5 \times 10^{-8} \]

- Need large effect or large sample size

Imputation of ungenotyped variants

Imputation: Observed genotypes

Observed Genotypes
. . . A . . . . . . . . A . . .
. . . G . . . . . . . . C . . .

Reference Haplotypes
C G A G A T C T C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T C T T C T G T G C
C G A A G C T C T T T T T T T T G T G C
C G A G A C T C T C C G A C T T A T G C
T G G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T T T T G A C
C G A A G C T C T T T T T T T T G A C
C G A G A C T C T C C G A C C T C T G T G C
C G A A G C T C T T T T T T T G A C

Study Sample

HapMap or 1000 Genomes or ...

Identify match among reference

Observed Genotypes
. . . A . . . . . . . . A . . .
. . . G . . . . . . . . C . . .

Reference Haplotypes
C G A G A T C T C T T C T T C T G T G C
CGAGATICCTCCGACCTCATGG
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T C T T C T G T G C
C G A A G C T C T T T T T T T T G T G C
C G A G A C T C T C C G A C T T A T G C
T G G G G A T C T C C C G A C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T T T T T G A C
C G A A G C T C T T T T T T T T G A C
C G A A G C T C T C C G A C C T C T G T G C
C G A G A C T C T C C C G A C C T C T T T T T T T T G A C
C G A A G C T C T C C C G A C C T C T G T G C
C G A A G C T C T C C C G A C C T C T T T T T T T T T T T T T T G A C
CGAGATICCTCTCTCTGTG
Phase chromosomes, impute missing genotypes

**Observed Genotypes**

- c g a g a
- c g a g

**Reference Haplotypes**

- C G A G A T C T C T T T T T T C T T G T G C
- C G A G A T C T C T C C G A C C T C A T G C
- C C A A G C T C T T T T T C T T G T G C
Combining GWAS by meta-analysis

- Combine studies giving more weight to studies with greater precision
- Increase power vs individual studies
- Can investigate consistency of effects across studies
- Potential sources of heterogeneity:
  - Phenotype definitions are different
  - Different genotyping and analysis strategies
  - Environmental effects may differ

Common meta-analysis methods

- P-value meta-analysis (should take direction of association under account)
- Effect size meta-analysis applied on normalized effects (e.g. natural logarithm of odds ratio for binary outcomes, mean difference or standardized mean difference for continuous traits)
- Fixed effects (between-study variance is assumed to be zero)
- Random effects (between-study variance estimated and incorporated)
- Bayesian meta-analysis (incorporates uncertainty in prior beliefs about parameters such as between-study variance, effect size, genetic model)

Zeggini (2009) Pharmacogenomics 10:191
Outline

- Genome-wide association study design
  - Samples/study participants
  - Genotyping
  - Tests of association
  - Imputation and meta-analysis

- Interpretation of results
  - Effect size and significance
  - Example locus characteristics

- Sequencing/rare variant studies

‘Manhattan plot’ for HDL-cholesterol

Global Lipids Genetics Consortium
188,577 individuals from 60 studies, GWAS + metabochip variants

GLGC (2013) Nat Gen 45:1274
Table 1: New loci primarily associated with HDL cholesterol levels obtained from joint GWAS and MetaboChip meta-analysis

<table>
<thead>
<tr>
<th>Locus</th>
<th>Marker name</th>
<th>Chr.</th>
<th>Position (Mb)</th>
<th>Associated trait</th>
<th>MAF</th>
<th>Minor/Major allele</th>
<th>Effect of allele</th>
<th>Joint p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>rs26745152</td>
<td>1</td>
<td>27.14</td>
<td>HDL, LDL, TG</td>
<td>0.09</td>
<td>TC</td>
<td>-0.051, 0.050, 0.037</td>
<td>1.87, 1.73, 1.78</td>
</tr>
<tr>
<td>rs1246743</td>
<td>DGFT/MVX</td>
<td>3</td>
<td>56.70</td>
<td>HDL</td>
<td>0.34</td>
<td>G/T</td>
<td>0.020</td>
<td>1.87, 2 x 10^-10</td>
</tr>
<tr>
<td>rs6509944</td>
<td>APOB/PLA1</td>
<td>12</td>
<td>1.60</td>
<td>HDL</td>
<td>0.49</td>
<td>GA</td>
<td>-0.021</td>
<td>1.87, 3 x 10^-10</td>
</tr>
<tr>
<td>rs10476736</td>
<td>CPS1</td>
<td>2</td>
<td>23.54</td>
<td>HDL, LDL, TG</td>
<td>0.33</td>
<td>TC</td>
<td>-0.007</td>
<td>1.87, 9 x 10^-10</td>
</tr>
<tr>
<td>rs2066736</td>
<td>APOE</td>
<td>5</td>
<td>1.60</td>
<td>HDL, TG</td>
<td>0.49</td>
<td>G/T</td>
<td>0.025</td>
<td>1.87, 3 x 10^-10</td>
</tr>
<tr>
<td>rs2006477</td>
<td>APOC1</td>
<td>20</td>
<td>1.60</td>
<td>HDL</td>
<td>0.20</td>
<td>AG</td>
<td>-0.005</td>
<td>1.87, 4 x 10^-10</td>
</tr>
<tr>
<td>rs1335058</td>
<td>APOA1</td>
<td>3</td>
<td>23.54</td>
<td>HDL, LDL, TG</td>
<td>0.50</td>
<td>TC</td>
<td>0.025</td>
<td>1.87, 7 x 10^-10</td>
</tr>
<tr>
<td>rs1335065</td>
<td>LPL</td>
<td>3</td>
<td>56.70</td>
<td>HDL, LDL, TG</td>
<td>0.21</td>
<td>AG</td>
<td>0.029</td>
<td>1.87, 9 x 10^-10</td>
</tr>
<tr>
<td>rs8505261</td>
<td>GCK</td>
<td>1</td>
<td>1.60</td>
<td>HDL</td>
<td>0.39</td>
<td>TC</td>
<td>0.020</td>
<td>1.87, 1 x 10^-10</td>
</tr>
<tr>
<td>rs30013888</td>
<td>CYP27B1</td>
<td>1</td>
<td>1.60</td>
<td>HDL</td>
<td>0.18</td>
<td>GA</td>
<td>-0.027</td>
<td>1.87, 5 x 10^-10</td>
</tr>
<tr>
<td>rs802072</td>
<td>FABP4</td>
<td>4</td>
<td>23.54</td>
<td>HDL</td>
<td>0.46</td>
<td>AG</td>
<td>-0.005</td>
<td>1.87, 4 x 10^-10</td>
</tr>
<tr>
<td>rs9602836</td>
<td>APOA2</td>
<td>1</td>
<td>1.60</td>
<td>HDL</td>
<td>0.44</td>
<td>AG</td>
<td>-0.019</td>
<td>1.87, 5 x 10^-10</td>
</tr>
<tr>
<td>rs9656620</td>
<td>APOE</td>
<td>12</td>
<td>1.60</td>
<td>HDL, LDL, TG</td>
<td>0.49</td>
<td>G/T</td>
<td>0.029, -0.020</td>
<td>1.87, 1.68</td>
</tr>
<tr>
<td>rs202485</td>
<td>VRK1</td>
<td>2</td>
<td>23.54</td>
<td>HDL, LDL, TG</td>
<td>0.45</td>
<td>G/A</td>
<td>0.004</td>
<td>1.87, 6 x 10^-10</td>
</tr>
<tr>
<td>rs1427995</td>
<td>SNX13</td>
<td>7</td>
<td>17.92</td>
<td>HDL, LDL, TG</td>
<td>0.38</td>
<td>TG</td>
<td>-0.006</td>
<td>1.87, 9 x 10^-12</td>
</tr>
<tr>
<td>rs917014</td>
<td>AKT7</td>
<td>7</td>
<td>17.92</td>
<td>HDL, LDL, TG</td>
<td>0.32</td>
<td>G/T</td>
<td>0.022</td>
<td>1.87, 1 x 10^-10</td>
</tr>
<tr>
<td>rs17172637</td>
<td>TMEM176A</td>
<td>7</td>
<td>150.53</td>
<td>HDL</td>
<td>0.12</td>
<td>CT</td>
<td>-0.006</td>
<td>1.84, 2 x 10^-10</td>
</tr>
<tr>
<td>rs1949808A</td>
<td>APOE</td>
<td>3</td>
<td>56.70</td>
<td>HDL, LDL, TG</td>
<td>0.26</td>
<td>GA</td>
<td>0.025, 0.026, 0.025</td>
<td>1.87, 1.87</td>
</tr>
<tr>
<td>rs1246602</td>
<td>APOA1</td>
<td>1</td>
<td>1.60</td>
<td>HDL</td>
<td>0.15</td>
<td>CT</td>
<td>0.004</td>
<td>1.76, 2 x 10^-10</td>
</tr>
<tr>
<td>rs1280636</td>
<td>AKT5</td>
<td>7</td>
<td>17.92</td>
<td>HDL, LDL, TG</td>
<td>0.23</td>
<td>AG</td>
<td>0.024</td>
<td>1.87, 3 x 10^-10</td>
</tr>
<tr>
<td>rs4999774</td>
<td>MAF</td>
<td>11</td>
<td>17.92</td>
<td>HDL, LDL, TG</td>
<td>0.19</td>
<td>AC</td>
<td>-0.006</td>
<td>1.87, 1 x 10^-10</td>
</tr>
<tr>
<td>rs9830659</td>
<td>SLC2A1</td>
<td>14</td>
<td>10.38</td>
<td>HDL, LDL, TG</td>
<td>0.40</td>
<td>G/A</td>
<td>0.020</td>
<td>1.84, 1 x 10^-10</td>
</tr>
<tr>
<td>rs1257982</td>
<td>FTO</td>
<td>16</td>
<td>53.81</td>
<td>HDL, LDL, TG</td>
<td>0.43</td>
<td>AG</td>
<td>0.020, 0.033</td>
<td>1.86, 1.25</td>
</tr>
<tr>
<td>rs11695224</td>
<td>FTO</td>
<td>17</td>
<td>53.81</td>
<td>HDL, LDL, TG</td>
<td>0.26</td>
<td>AG</td>
<td>-0.009</td>
<td>1.85, 2 x 10^-10</td>
</tr>
</tbody>
</table>

Chr. = chromosome; A1 = minor allele; A2 = major allele; TG = triglycerides; TC = total cholesterol. Effect sizes are given with respect to the minor allele (A1) in uL. For loci associated with two or more traits, the trait with the strongest effect is listed first.

*The secondary trait was most strongly associated with a different SNP (rs2054029) within 1 Mb of rs129316 (rs1293013 within 1 Mb of rs121960, r² = 0.386).

GLGC (2013) Nat Gen 45:1274

Single good candidate gene

Signal outside of genes

Many candidate genes

Interpret GWA locus names with caution; many are merely the nearest gene to the signal

Interpret plausible candidate genes

<table>
<thead>
<tr>
<th>Locus</th>
<th>Nearest Gene</th>
<th>Nearest Gene (kb)</th>
<th>No. of Genes within 100kb</th>
<th>Literature Candidate</th>
<th>Gene with Nonsynonymous SNP (p&lt;0.5)</th>
<th>eQTL Gene (p&lt;0.01)</th>
<th>Pathway Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIGV-NR0B2</td>
<td>PIGV</td>
<td>13.5</td>
<td>7</td>
<td>PIGV, NR0B2</td>
<td>NR0B1*, C1orf172*, NR0B2</td>
<td>NR0B2</td>
<td></td>
</tr>
<tr>
<td>HDGF-FMVK*</td>
<td>RRM4D1</td>
<td>0</td>
<td>10</td>
<td>HDGF, CRABP2</td>
<td>HDGF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANGPTL1*</td>
<td>C1orf220</td>
<td>0</td>
<td>3</td>
<td>C1orf220</td>
<td>C1orf220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPS1</td>
<td>CPS1</td>
<td>0</td>
<td>2</td>
<td>CPS1</td>
<td>CPS1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATG7</td>
<td>ATG7</td>
<td>0</td>
<td>2</td>
<td>ATG7</td>
<td>ATG7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SETD2</td>
<td>SETD2</td>
<td>0</td>
<td>4</td>
<td>SETD2</td>
<td>SETD2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBMS5</td>
<td>RBMS5</td>
<td>0</td>
<td>4</td>
<td>RBMS5</td>
<td>RBMS5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAB1</td>
<td>STAB1</td>
<td>0</td>
<td>10</td>
<td>STAB1, NISCH</td>
<td>STAB1, NISCH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSK3B</td>
<td>GSK3B</td>
<td>0</td>
<td>3</td>
<td>GSK3B, NR12</td>
<td>GSK3B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4orf52*</td>
<td>C4orf52*</td>
<td>131.5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAM15A</td>
<td>FAM15A</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADH5</td>
<td>ADH5</td>
<td>4.9</td>
<td>4</td>
<td></td>
<td>ADH5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSPO3</td>
<td>RSPO3</td>
<td>4</td>
<td>1</td>
<td></td>
<td>RSPO3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAE6B</td>
<td>DAE6B</td>
<td>5</td>
<td>5</td>
<td></td>
<td>DAE6B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNX13</td>
<td>SNX13</td>
<td>1</td>
<td>1</td>
<td></td>
<td>SNX13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IKZF1</td>
<td>IKZF1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>IKZF1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMEM176A</td>
<td>ABP1</td>
<td>20.1</td>
<td>5</td>
<td></td>
<td>TMEM176A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAREC8-ALX5</td>
<td>MAREC8</td>
<td>3</td>
<td>3</td>
<td></td>
<td>MAREC8, ALX5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR4A4B</td>
<td>OR4A4B</td>
<td>3.2</td>
<td>2</td>
<td></td>
<td>OR4A4B, OR5D12, OR5D13, OR5S11*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GLGC (2013) Nat Gen 45:1274

Nearby independent signals

- p1+2 = 2e-15
- p1+2 = 3e-20

CEU: D’ = .07, r² < .01, p-values remain unchanged with other SNP as covariate

Cristen Willer
**Conditional analysis**

\[ y = \beta_0 + \beta_1 x \]

\[ \text{Trait} = \beta_0 + \beta_1 \text{SNP}_1 + \beta_2 \text{SNP}_2 \]

\[ [\text{HDL}] = \beta_0 + \beta_1 \text{rs261332} + \beta_2 \text{rs4775041} \]

\[ [\text{HDL}] = \beta_0 + \beta_1 \text{rs261332} + \beta_2 \text{rs4775041} + \beta_3 \text{sex} + \beta_4 \text{age} + \beta_5 \text{age}^2 \]

**Tests independence of SNP effects**

If \( \beta_1 \) changes when \( \beta_2 \) is included in the model, then SNP \( \text{SNP}_1 \) is sometimes inherited with SNP \( \text{SNP}_2 \)

If neither \( \beta \) changes in reciprocal tests, then the two SNPs independently affect the trait

---

**Fine-mapping across populations**

- **Europeans**
  - HDL-C locus near PPP1R3B
  - Position on chr8 (Mb)

- **African Americans**
  - HDL-C locus near PPP1R3B
  - Position on chr8 (Mb)
Outline

• Genome-wide association study design
  – Samples/study participants
  – Genotyping
  – Tests of association
  – Imputation and meta-analysis

• Interpretation of results
  – Effect size and significance
  – Example locus characteristics

• Sequencing/rare variant studies

Figure 1. An overview of steps taken in the search for low-frequency and rare variants affecting complex traits.

Some sequencing study designs for complex traits

- Sequence selected individuals
  - extreme trait values (>95% vs <5% level)
  - cases and controls

- Increase the number of individuals
  - by decreasing sequencing coverage ($)
  - by collecting rare variants onto a less expensive genotyping array

- Sequence population isolates, where rare variants may have drifted to higher frequencies and LD may be longer

---

**Table 4. Functional Characterization of MC4R Nonsynonymous Variants in the Obese Cohort**

<table>
<thead>
<tr>
<th>Variant</th>
<th>Sequence</th>
<th>Known or Novel</th>
<th>n</th>
<th>alpha-MSH Activation (IC50)</th>
<th>Basal Activity</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>S30F</td>
<td>tgaac/t/ccttg</td>
<td>Known^100</td>
<td>1</td>
<td>Not tested alone[^112]</td>
<td>Not tested alone[^112]</td>
<td>...</td>
</tr>
<tr>
<td>G32E</td>
<td>ccttg/a/aagaag</td>
<td>Novel</td>
<td>1</td>
<td>.3 nM</td>
<td>70%</td>
<td>Minor</td>
</tr>
<tr>
<td>E61K</td>
<td>tgtg/g/aagaaat</td>
<td>Novel</td>
<td>1</td>
<td>Low</td>
<td>&lt;10%</td>
<td>Severe</td>
</tr>
<tr>
<td>S127L</td>
<td>tptct/c/tgagta</td>
<td>Known[^112]</td>
<td>1</td>
<td>29 nM</td>
<td>80%</td>
<td>Intermediate</td>
</tr>
<tr>
<td>L211Del</td>
<td>tctc[c/t]/tactg</td>
<td>Known[^112]</td>
<td>2</td>
<td>Truncated receptor</td>
<td>Truncated receptor</td>
<td>Severe</td>
</tr>
<tr>
<td>P299H(^1)</td>
<td>cgactc/a/atcga</td>
<td>Known[^112]</td>
<td>2</td>
<td>Negative</td>
<td>&lt;=10%</td>
<td>Severe</td>
</tr>
<tr>
<td>A303T</td>
<td>ttt[t/a]/caactc</td>
<td>Novel</td>
<td>1</td>
<td>Low</td>
<td>&lt;10%</td>
<td>Severe</td>
</tr>
<tr>
<td>C326R</td>
<td>gcttt/c/tgagac</td>
<td>Novel</td>
<td>1</td>
<td>.4 nM</td>
<td>150%</td>
<td>Minor</td>
</tr>
<tr>
<td>Wild type</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>.3 nM</td>
<td>100%</td>
<td>...</td>
</tr>
</tbody>
</table>

^1 Individuals who had the L211Del also had the P299H variant.

Variant discovery at GWAS locus

- Sequence ‘positional candidate’ genes in cases & controls or individuals with extreme trait values
- Identify variants in cases (one extreme) that are absent from controls (other extreme)
- Hypothesize that occasional ‘smoking gun’ variants with strong effect will be identified
- Use evidence that variants affect gene function and lead to the same disease/trait to implicate that gene at the association signal
- Does not require finding the variant(s) responsible for association signal that may have a weaker effect
Rare variants confirmed to be associated with T1D in more samples

Establishes the role of IFIH1 in T1D and demonstrates that resequencing studies can pinpoint disease-causing genes in regions initially identified by GWASs.

Identify an increased ‘burden’ of variants in a single gene or locus

- Many individually important variants will be too rare to detect the association with the trait; however, there will often be more than one important variant in a gene

- Gene-based tests combine information from multiple variants into a single test statistic to be used as predictor in genetic association tests

Rare variant burden (gene-based) tests

- Collapse information from multiple variants into single test (e.g. count risk alleles across a set of variants)

- Some tests allow the direction of effect of each variant to be different (gain of function versus lost of function)

- Choice of variants to include in tests has a large impact on the test. Including too many neutral variants reduces statistical power, but so can not including the right ones

  - Filter missense variants on minor allele frequency and predictive function

  - Restrict tests to obvious functional variants (nonsense, frameshift indels, splice errors)

Gene-based rare variant association methods

<table>
<thead>
<tr>
<th>Method name</th>
<th>Citation</th>
<th>Software</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined Multivariate and Collapsing (CMC)</td>
<td>Liu &amp; Leal, PLoS Comp. Bio, 2008</td>
<td>EPACTS</td>
<td>All rare variants collapsed into a single variant; individual dosage for the collapsed variant is regressed against phenotype.</td>
</tr>
<tr>
<td>Variable threshold (VT)</td>
<td>Price et al, AJHG, 2010</td>
<td>PLINK-Seq</td>
<td>Sum of rare allele count in cases vs. controls; allele frequency threshold for inclusion is varied to maximize test statistic.</td>
</tr>
<tr>
<td>Weighted Sum Statistic (F-RHOT)</td>
<td>Madsen &amp; Browning, PLoS Gen. 2008</td>
<td>PLINK-Seq</td>
<td>Permutation-based test comparing inverse-frequency-weighted rare variant counts per individual in cases vs. controls.</td>
</tr>
<tr>
<td>Kernel-Based Adaptive Cluster (KBAC)</td>
<td>Liu &amp; Leal, PLoS Gen. 2010</td>
<td>PLINK-Seq</td>
<td>Variant weights are determined adaptively, and are based on observed effect sizes; individuals scored by weighted sum of allele counts.</td>
</tr>
<tr>
<td>Summary case-control count methods</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BURDEN method</td>
<td>Purcell (PLINK-Seq)</td>
<td>PLINK-Seq</td>
<td>Permutation-based test comparing raw allele counts in cases vs. controls.</td>
</tr>
<tr>
<td>UNO test</td>
<td>Purcell (PLINK-Seq)</td>
<td>PLINK-Seq</td>
<td>Simple count of total case-unique rare alleles; permutations to assess significance.</td>
</tr>
<tr>
<td>Bi-directional variance-component gene-based tests</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-ALPHA</td>
<td>Neale et al, PLoS Gen. 2011</td>
<td>PLINK-Seq</td>
<td>Detects deviation of observed case control variant counts from expected binomial distribution.</td>
</tr>
<tr>
<td>Sequence Kernel Association Test (SKAT)</td>
<td>Wu et al, AJHG 2011</td>
<td>EPACTS</td>
<td>Generalized form of C-ALPHA with variants weighted by allele frequency.</td>
</tr>
<tr>
<td>Linear combination of unidirectional and variance-component tests</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKAT-O (Optimal SKAT)</td>
<td>Lee et al, AJHG, 2012</td>
<td>EPACTS</td>
<td>Adaptive linear combination of unidirectional burden test and variance-component (SKAT) tests.</td>
</tr>
<tr>
<td>Mixed Effects Score Test (MST)</td>
<td>Sun et al, Genetic Epidemiology 2013</td>
<td>Public R package</td>
<td>Hierarchical regression model combining two independent test statistics which quantify variant effect sizes and ‘heterogeneity’.</td>
</tr>
</tbody>
</table>

An example of a gene-based test

Loss-of-function mutations in SLC30A8 protect against type 2 diabetes

- Initially sequenced 352 young lean T2D cases, 406 elderly obese euglycemic controls
- Then tested variants in 6,388 cases and 7,496 controls
- Found a nonsense variant in 7 cases and 21 controls, odds ratio (OR) = 0.38, $P = 0.05$
- Added this variant to the exome array and tested more individuals (N= 48,115, $P = 0.0067$).
- Difficult to increase sample size because variant mostly restricted to western Finland
- Expanded to look at more variants in the gene in other populations...

UK10K sequencing study


UK10K association results

Figure 3 | Summary of association results across the UK10K-cohorts study.

Clinical translation

- Identification of susceptibility variants
  - Novel biological insights
  - Clinical advances
    - Therapeutic targets
    - Biomarkers
    - Prevention
  - Improved measures of individual aetiological processes
    - Personalized medicine
      - Diagnostics
      - Prognostics
      - Therapeutic optimization


Future of complex trait analyses

- More and more loci identified
- Larger meta-analyses
- Deeper follow-up of signals
- More diverse populations
- Gene-based results from rare variants
- Gene-gene and -environment interactions
- Molecular and biological mechanisms