The Promise of Genome-wide Association Studies (GWAS)

Stephen J. Chanock, M.D.
Division of Cancer Epidemiology and Genetics
Center for Cancer Research

December 1, 2006
Value of GWAS

• Identification of promising low-penetrance, high-frequency susceptibility loci

• Evaluation of gene-gene interactions and genetic interactions with environmental exposures

• Tool for identifying novel mechanisms in cancer

• Foundation for strategies for prevention and intervention
GWAS & NCI Priorities

• Capitalize on revolution in genetics
  ➢ Annotation of common genetic variation
  ➢ Technology platforms
• Intramural capabilities at Core Genotyping Facility
• NCI investment in cohorts
• Informatics and access: NCICB (caBIG portal)
• Coordinate with NIH-wide activities
Mission of Cancer Genetic Markers of Susceptibility

- Conduct GWAS in 2 cancers
  - Prostate (1 in 8 men)
  - Breast (1 in 9 women)
- Rapid sequential replication studies
- Aggressive timeline
- Initial scans in nested case-control studies
  - Prostate, Lung, Colon, Ovary (PLCO) Project
  - Nurses’ Health Study

http://cgems.cancer.gov
Strategy for Prostate & Breast Cancer

Initial Study
Nested case/control study in prospective cohorts
1,150 cases/1,150 controls

Follow-up Study #1
3,500 cases/3,500 controls

Follow-up Study #2
3,500 cases/3,500 controls

Fine Mapping

540,000 Tag SNPs
~28,000 SNPs
at least 1,500 SNPs
30±20 loci
N demonstrated susceptibility loci
Power of the First Two Phases of CGEMS

Point-wise significance $10^{-7}$; "genome wide" significance 0.05

<table>
<thead>
<tr>
<th>Mode</th>
<th>GRR</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recessive</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Dominant</td>
<td>1.5</td>
<td>1.0</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Additive</td>
<td>1.4</td>
<td>1.0</td>
<td>1.4</td>
<td>1.8</td>
</tr>
<tr>
<td>Multiplicative</td>
<td>1.3</td>
<td>1.0</td>
<td>1.3</td>
<td>1.69</td>
</tr>
</tbody>
</table>

Continuous line: power for direct detection ($r^2 = 1$)
Dashed line: power for $r^2 = 0.8$

What is available for GWAS in 2006?

Coverage analysis based on HapMap II Data

<table>
<thead>
<tr>
<th></th>
<th>Coverage Analysis</th>
<th>CEU</th>
<th>YRI</th>
<th>JPT/CHB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina</td>
<td>HumanHap300</td>
<td>80%</td>
<td>35%</td>
<td>40%</td>
</tr>
<tr>
<td>Illumina</td>
<td>HumanHap500</td>
<td>91%</td>
<td>58%</td>
<td>88%</td>
</tr>
<tr>
<td>Affymetrix*</td>
<td>500k Mapping</td>
<td>63%*</td>
<td>41%</td>
<td>63%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*77% (with 50k MegA)

http://tagzilla.nc.nih.gov
CGEMS Scans

Prostate Cancer

Two Scans
Illumina
317k (available)

Breast Cancer

One Scan
Illumina
550k (March 2007)

240k (February 2007)
Discordance Rates in Genotype Analysis

PLCO
49 duplicate pairs
Mean discordance rate \(2 \times 10^{-4}\)

CEPH-CGEMS
74 duplicate pairs
Mean discordance rate \(2 \times 10^{-4}\)

CEPH-HapMap
28 individuals (with 24 duplicates)
Mean Discordance rate \(1.4 \times 10^{-3}\)
QQ Plot for p-values of ~300k SNPs in Prostate Scan 1A

http://cgems.cancer.gov
Log-Log Quantile Plot for p-values for the Four Statistical Tests Used

307,256 SNPs

http://cgems.cancer.gov
PLCO Recruitment Sites
Opportunity to Look at Geographic Differences
Admixture Coefficient in PLCO Prostate Study Samples

Method:
Run merged PLCO data + HapMap data on STRUCTURE with 6,000 SNPs having no pair-wise r² and high FST values. The population of origin of the HapMap samples is specified.

Result:
Reliable identification of 3 outliers. They are all 3 control DNAs and have to be removed from subsequent analysis.
Log$_{10}$ p-value of the 4 d.f. $\chi^2$ test Plotted Against the Position of the 8q24 SNP (rs#1447295)* in Build 35

*Amundadottir Nat Genet 2006
*Freedman PNAS 2006
Prostate Scan
8q24 Region

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1447295</td>
<td>A</td>
<td>0.1</td>
<td>1.08</td>
<td>1.45</td>
<td>1.24</td>
<td>1.46</td>
</tr>
<tr>
<td>rs4242382</td>
<td>A</td>
<td>0.1</td>
<td>1.13</td>
<td>1.39</td>
<td>1.27</td>
<td>1.39</td>
</tr>
<tr>
<td>rs7017300</td>
<td>C</td>
<td>0.13</td>
<td>1.14</td>
<td>1.63</td>
<td>1.17</td>
<td>1.37</td>
</tr>
<tr>
<td>rs7837688</td>
<td>T</td>
<td>0.1</td>
<td>1.14</td>
<td>1.36</td>
<td>1.26</td>
<td>1.54</td>
</tr>
</tbody>
</table>

Key Findings:
1. Comparable risk as original reports in Nat Genet and PNAS
2. Comparable risk for BPC3 (~6,500 cases/controls)
3. Discovery of 1 and perhaps 2 additional loci
SNP1
rs1447295
SNP2
SNP3
SNP4

8q24 Region & PLCO Data
HapMap CEU D’
Value-added Analysis in CGEMS

Opportunity to Investigate

- Determinants of risk factors
  - BMI, smoking, hormone levels
- Multi-SNP analysis
- Gene: Gene interactions
  - Explore pathways
CGEMS: caBIG Posting
Pre-computed Analysis

Pre-computed Analysis
No Restrictions

Raw Genotype
Case/control
Age (in 5 yrs)
Family Hx (+/-)
Registration

This is the home page of the Cancer Genetic Markers of Susceptibility (CGEMS) data access. The following links provide information on the project and background. The CGEMS study design uses cases and controls drawn from well designed epidemiological studies of prostate and breast cancer. DNA from these subjects is being used to generate genotypes to perform a Genome-Wide Association Study (GWAS) on over 500,000 genetic variants to determine their role in cancer susceptibility.

CGEMS Prostate Scan Phase 1
A GWAS has been conducted in a large, national study in the U.S.A., the Prostate, Lung, Colorectal, and Ovary study (PLCO). The analysis includes 1,177 subjects who developed prostate cancer during the observational period and 1,105 individuals who did not develop prostate cancer during the same time period. The prostate scan is being conducted in two parts, Phase 1A and Phase 1B.

The data generated from these scans can be accessed through this portal. The first posting includes data from Phase 1A of the prostate cancer scan and includes:

- Association test results for over 300,000 SNPs
- Frequency and descriptive statistics on these SNPs
- Individual phenotypic and genotypic data for the study participants and control samples. Note that these data can only be made available to eligible investigators after a registration process (link).

The results of Phase 1B will be available in February 2007.

For more information on:
- About CGEMS Study
- How to use the CGEMS data portal
- Register to access raw data

Click the question mark icon for context sensitive help throughout the application.

CGEMS updates:
- This release, Version 1.0, was deployed on Oct 10, 2006.
- The current dataset in use was deployed on Oct 10, 2006

http://cgems.cancer.gov
**Association Tests**

8q24

Scan 1A

~300,000 SNPs

---

**CGEMS SNP Association Finding Report**

**Study:** CGEMS Prostate Cancer WGAS Phase 1A

<table>
<thead>
<tr>
<th>dbSNP ID</th>
<th>Chromosome</th>
<th>Physical Position (bp)</th>
<th>Associated Genes</th>
<th>Analysis Name</th>
<th>p-value</th>
<th>Whole Genome Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1234567</td>
<td>B</td>
<td>128523110</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.02361</td>
<td>7583</td>
<td></td>
</tr>
<tr>
<td>rs8765432</td>
<td>B</td>
<td>128527862</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.61895</td>
<td>187681</td>
<td></td>
</tr>
<tr>
<td>rs9087654</td>
<td>B</td>
<td>128529872</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.569441</td>
<td>172475</td>
<td></td>
</tr>
<tr>
<td>rs3213456</td>
<td>B</td>
<td>128529858</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.353344</td>
<td>106901</td>
<td></td>
</tr>
<tr>
<td>rs7865432</td>
<td>B</td>
<td>128530060</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.753514</td>
<td>228046</td>
<td></td>
</tr>
<tr>
<td>rs7865432</td>
<td>B</td>
<td>128533996</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.101896</td>
<td>30853</td>
<td></td>
</tr>
<tr>
<td>rs1234567</td>
<td>B</td>
<td>128541502</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.026153</td>
<td>7829</td>
<td></td>
</tr>
<tr>
<td>rs9087654</td>
<td>B</td>
<td>128544367</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.109914</td>
<td>33305</td>
<td></td>
</tr>
<tr>
<td>rs3213456</td>
<td>B</td>
<td>128551824</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.009611</td>
<td>21001</td>
<td></td>
</tr>
<tr>
<td>rs7865432</td>
<td>B</td>
<td>128554220</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>4.18E-4</td>
<td>149</td>
<td></td>
</tr>
<tr>
<td>rs9087654</td>
<td>B</td>
<td>128555770</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.572839</td>
<td>173461</td>
<td></td>
</tr>
<tr>
<td>rs3213456</td>
<td>B</td>
<td>128565958</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.281571</td>
<td>85131</td>
<td></td>
</tr>
<tr>
<td>rs1234567</td>
<td>B</td>
<td>128576206</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.282396</td>
<td>85399</td>
<td></td>
</tr>
<tr>
<td>rs9087654</td>
<td>B</td>
<td>128576784</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.285649</td>
<td>85401</td>
<td></td>
</tr>
<tr>
<td>rs3213456</td>
<td>B</td>
<td>128586755</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>9.6E-5</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>rs7865432</td>
<td>B</td>
<td>128594450</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>1.58E-4</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>rs1234567</td>
<td>B</td>
<td>128608542</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>3.8E-5</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>rs9087654</td>
<td>B</td>
<td>128614565</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.106728</td>
<td>32421</td>
<td></td>
</tr>
<tr>
<td>rs3213456</td>
<td>B</td>
<td>128619305</td>
<td>Incidence density sampling, Adjusted score test</td>
<td>0.854811</td>
<td>258529</td>
<td></td>
</tr>
</tbody>
</table>

---

http://cgems.cancer.gov
GWAS in Pancreatic Cancer: PanScan Objectives

- Identify loci associated with pancreatic cancer
  - 1,200 cases and 1,200 controls drawn from 12 cohort studies
- Define susceptibility loci for common genetic variants (MAF > 5%)
- Follow-up studies in cohort and case control studies
- Public access for data
  - Pre-computed association testing – Open
  - Raw genotype data with limited phenotype data-registered access
What is down the road?

2-4 Year Forecast
- Cheaper and denser SNP technologies
  - Better coverage of genome

4-8 Year Forecast
- Whole Genome Sequencing
  - Replace SNPs
  - Magnification of Challenge of Confidentiality
  - Challenge to Epidemiologic Rigor
Follow-up of GWAS: Steps to Clinical Implementation

• Fine mapping of notable regions
• Functional determination of causal variants
• Design issue for analysis in clinical studies
  ➢ Population-based studies
  ➢ Sequence of clinical studies
• Validation criteria