A Primer on Genetic Data Available in the Wisconsin Longitudinal Study

Craig S. Atwood; Tina Gonzales; Carol Roan; Huey-Chi V. Chang; Robert M. Hauser; Pamela Herd

Laboratory of Endocrinology, Aging and Disease
Geriatrics Research, Education and Clinical Center, William S. Middleton VA Hospital;
Departments of Medicine and Sociology
University of Wisconsin-Madison
Genotyping the Wisconsin Longitudinal Study (WLS)

- WLS is rich in sociobehavioral/environmental (E) variables
  - social background, youthful aspirations, schooling, military service, family formation, labor market experiences, health, and social participation

- Determine all genetic (G) alterations between individuals
  - ~5,000-10,000 WLS participants

- Unique cohort containing detailed E and G data available for association studies
Genetic Alterations 101

Single nucleotide polymorphism (SNP)

DNA molecule 1 differs from DNA molecule 2 at a single base-pair location (a C/T polymorphism).

3.1 billion base pairs in the human genome

17 million unique SNPs (1 in every 180 base pairs), but for most 1 in every 1000 base pairs

Current DNA arrays – a few 100,000 SNPs
Genetic Alterations 101

Single nucleotide polymorphism (SNP)

Types of Genetic Alterations
- SNPs
- Deletions
- Insertions
- Repeat expansion (e.g. CGA repeats, Huntington’s disease)

Coding vs. Non-coding Alterations
- Exonic genetic alterations
  - Missense mutation
  - amino acid change in protein
  - Nonsense mutation
  - prematurely ends the building of a protein
- Intronic genetic alterations
  - alter splice sites, stop and start codons
  - alter miRNA

DNA molecule 1 differs from DNA molecule 2 at a single base-pair location (a C/T polymorphism).
Genetic Variables

• Traditional DNA sequencing
  • low cost ($100-$1,000)
  • few 10-100s genetic alterations (84 SNPs measured in pilot WLS data)
  • Measures all mutations (SNPs, deletions, insertions, repeats)

• DNA arrays
  • moderate cost ($1,000-$10,000)
  • measure ~500,000s SNPs (of the ~3-4 million SNPs per genome)
  • Can’t measure deletions, insertions, repeats

• Whole genome sequencing (WGS)
  • higher cost ($1,000-$5,000 per genome currently)
    • 2000: + $1 billion to sequence 1st genome
    • 2014/15: 5,000 WLS participants x $1,000 = $5 million
  • captures ALL genetic modifications
  • better than measuring SNPs on an array
Genetic Data

- Saliva collected - DNA extracted
- Genotyping performed (Kbiosciences, LLC)
- Genotypic data for 84 SNPs for 7,101 WLS participants (4,569 graduates and 2,532 siblings)
- Genes with known associations to:

  Alcohol addiction  Fragile X syndrome  Muscle Mass
  Alzheimer's disease  Heart disease  Nicotine addiction
  Autism  Impulsivity  Obesity
  Bipolar disorder  Insulin resistance  Parkinson's disease
  Cancer (breast)  IQ  Premature Ovarian Failure
  Cerebrovascular disease  Liver disease  Reproductive
  Cognition  Longevity  Social Behavior
  Cognitive Aging  Memory  Schizophrenia
  Depression  Menstrual Patterns  Social Behavior
  Diabetes mellitus type II  Mental Retardation  Social Behavior
  Fertility  Motor Coordination
Experimental Design

- Saliva collected - DNA extracted
- Genotyping performed (Kbiosciences, LLC)
- Genotypic data for 84 SNPs for 7,101 WLS participants (4,569 graduates and 2,532 siblings)
  - Genes with known associations to:

<table>
<thead>
<tr>
<th>Alcoholic addiction</th>
<th>Fragile X syndrome</th>
<th>Muscle Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer's disease</td>
<td>Heart disease</td>
<td>Nicotine addiction</td>
</tr>
<tr>
<td>Autism</td>
<td>Impulsivity</td>
<td>Obesity</td>
</tr>
<tr>
<td>Bipolar</td>
<td>Insulin resistance</td>
<td>Parkinson's disease</td>
</tr>
<tr>
<td>Cancer (breast)</td>
<td>IQ</td>
<td>Premature Ovarian Failure</td>
</tr>
<tr>
<td>Cerebrovascular disease</td>
<td>Liver disease</td>
<td>Reproductive</td>
</tr>
<tr>
<td>Cognition</td>
<td>Longevity</td>
<td>Social Behavior</td>
</tr>
<tr>
<td>Cognitive Aging</td>
<td>Memory</td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>Menstrual Patterns</td>
<td></td>
</tr>
<tr>
<td>Diabetes mellitus type II</td>
<td>Mental Retardation</td>
<td></td>
</tr>
<tr>
<td>Fertility</td>
<td>Motor Coordination</td>
<td></td>
</tr>
</tbody>
</table>
Age-related Diseases are the Leading Causes of Death

Table 1. Leading causes of death in the US

<table>
<thead>
<tr>
<th>Cause of Death</th>
<th>All Ages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Deaths</td>
<td>2,426,264</td>
</tr>
<tr>
<td>Diseases of the heart</td>
<td>631,636</td>
</tr>
<tr>
<td>Malignant Neoplasms (Cancer)</td>
<td>559,888</td>
</tr>
<tr>
<td>Cerebrovascular diseases</td>
<td>137,119</td>
</tr>
<tr>
<td>Chronic Lower Respiratory Disease</td>
<td>124,583</td>
</tr>
<tr>
<td>Accidents (unintentional injuries)</td>
<td>121,599</td>
</tr>
<tr>
<td>Diabetes Mellitus</td>
<td>72,449</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>72,432</td>
</tr>
<tr>
<td>Influenza and Pneumonia</td>
<td>56,326</td>
</tr>
</tbody>
</table>

### Table 2. Sample sizes by disease and gender.

<table>
<thead>
<tr>
<th>Disease</th>
<th>$n_{\text{diseased}}$ (male/female)</th>
<th>$n_{\text{non-diseased}}$ (male/female)</th>
<th>$n_{\text{total}}$ (male/female)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>777 (357/420)</td>
<td>6160 (2988/3172)</td>
<td>6937 (3345/3592)</td>
</tr>
<tr>
<td>11.2% Dementia*</td>
<td>383 (203/180)</td>
<td>6352 (3039/3313)</td>
<td>6735 (3242/3493)</td>
</tr>
<tr>
<td>5.7% Diabetes*</td>
<td>776 (455/321)</td>
<td>6161 (455/321)</td>
<td>6937 (3346/3591)</td>
</tr>
<tr>
<td>11.2% Heart disease*</td>
<td>1006 (658/348)</td>
<td>5729 (2584/3145)</td>
<td>6735 (3242/3493)</td>
</tr>
<tr>
<td>14.9% Stroke</td>
<td>169 (85/84)</td>
<td>5431 (2608/2823)</td>
<td>5600 (2693/2907)</td>
</tr>
<tr>
<td>3.0% Depression</td>
<td>711 (230/481)</td>
<td>4081 (2103/1978)</td>
<td>4792 (2333/2459)</td>
</tr>
<tr>
<td>14.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Genotyping the Wisconsin Longitudinal Study (WLS) LEAD ‘14

Morbidity and mortality will continue to increase as the WLS ages
Genetic Alterations in the Etiology of Diseases: Single Main Factor Effects Versus Gene-Gene Interactions

- Genome Wide Association Studies (GWAS) have not identified a single main factor that explains late-onset AD (Harold et al., 2009, *Nat. Genet.* 2009).

- Humans are complex biological systems - interest to focusing on the detection of effects that, owing to their interaction with other genetic or environmental factors, might not be identified by using standard single-locus tests

- Continued development of analytical tools to identify genetic associations with increased predictability
Statistical Analyses

- Multifactor dimensionality reduction (MDR) (Velez et. al., 2007 Genet. Epidemiol. 2007)
  - Screening tool

- Recursive partitioning
  - Screening tool

- Logistic regression
  - Verification of multi-gene or gene-environment interactions

- Machine Learning Methodologies
  - multiple classification algorithms, using the freely available data-mining software Weka, version 3.6.6
Recursive Partitioning: creates a decision tree that strives to correctly classify members of the population based on several dichotomous dependent variables (Breiman L, Friedman J, Stone CJ, and Olshen RA. Classification and Regression Trees. 1984. Chapman and Hall, London).
Objective

• To mine WLS genotypic data for single-gene effects, and multi-gene and gene-environment interactions, for the following age-related diseases:
  • depression
  • cancer
  • dementia
  • diabetes
  • heart disease
  • stroke

(All referenced WLS variables ascertained in 2004-2005)
Objective

• To mine WLS genotypic data for single-gene effects, and multi-gene and gene-environment interactions, for the following age-related diseases:
  • depression
  • cancer
  • dementia
  • diabetes
  • heart disease
  • stroke

(All referenced WLS variables ascertained in 2004-2005)
Criteria for CIDI-SF Depression in WLS

Have you ever had a time in your life lasting two weeks or more when nearly every day you felt sad, blue, depressed, or when you lost interest in most things like work, hobbies, or things you usually liked to do for fun?

+ PLUS +

During which you experienced 3 or more of following symptoms?

- weight loss
- trouble sleeping
- feeling tired
- feeling bad upon waking
- losing interest
- trouble concentrating
- thoughts about death

Recursive Partitioning

CIDI-SF Depression in Men (G x G)

**DRD2**—dopamine receptor D2
SNP associations: reduced dopamine binding sites in brain, smoking, alcoholism, neuropsychiatric disorders

GNRH1—gonadogonadotropin-releasing hormone 1
Novel SNP: GNRH1 gene important for reproduction (GNRH1 aka: luteinizing-releasing hormone)

Identify increased and decreased prevalence, i.e. disease/condition causing and protective gene-gene interactions

Roetker et al., 2012; *BMJ Open*
Recursive Partitioning

CIDI-SF Depression in Men (G x E)

*NOT MARRIED: Current marital status was separated, divorced, widowed, or never married.

Identify increased and decreased prevalence, i.e. disease/condition causing and protective E x E interactions
Logistic Regression

CIDI-SF Depression Risk in Men

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Risk ratio</th>
<th>P-value (95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOT MARRIED</td>
<td>2.497</td>
<td>&lt; 5x10^{-6} ** ***</td>
</tr>
<tr>
<td>DEATH OF CHILD</td>
<td>2.347</td>
<td>&lt; 2x10^{-4} ** ***</td>
</tr>
<tr>
<td>HAS ANY PAIN</td>
<td>1.716</td>
<td>&lt; 4x10^{-4} ** ***</td>
</tr>
<tr>
<td>DRD2.2 T/T</td>
<td>3.335</td>
<td>&lt; 2x10^{-4} ** ***</td>
</tr>
<tr>
<td>GNRH1 T/T</td>
<td>1.515</td>
<td>&lt; 7x10^{-3} ** ***</td>
</tr>
</tbody>
</table>

*** Significant after controlling for multiple testing

Despite DRD2.2 and GNRH1 not appearing in G x E recursive partitioning tree, still found to be significant in logistic regression model
Recursive Partitioning

CIDI-SF Depression in Women (G x G)

DRD2—dopamine receptor D2
SNP association: migraines (potentially)

APOC3—apolipoprotein C3
SNP association: nonalcoholic fatty liver disease

ACVR2B—activin A receptor, type IIB
Gene associated with growth and differentiation

FTO—fat mass and obesity-associated
SNP association: obesity

1.3 times increase in prevalence

19.6%
n=2459
DRD2.3=T/T

17.4%
n=1227
APOC3=C/C,T/C

21.5%
n=469
APOC3=T/T

15.0%
n=748
ACVR2B=T/C

21.9%
n=1201
FTO=T/C

17.5%
n=576
IL6=C/G

25.8%
n=597
IL6=C/C,T/T

21.9%
n=288
IL6=C/C,G/G

15.1%
n=225
ACVR2B=C/C,T/T

27.7%
n=238
IL6=C/C

13.2%
n=281
ACVR2B=C/C

17.5%
n=288
ACVR2B=C/C,T/T

15.0%
n=748
ACVR2B=C/C,T/T

Roetker et al., 2012; BMJ Open
Recursive Partitioning

CIDI-SF Depression in Women (G x E)

DISC1—disrupted in schizophrenia 1
SNP associations: cognitive decline in women, bipolar disorder, verbal fluency

1.5 times increase in prevalence

**CHANCE LOSE JOB**: Do you think there is that you will lose your job completely in the next two years?

*NOT MARRIED*: Current marital status was separated, divorced, widowed, or never married.

**NOT MARRIED**: 29.6%
n=27

MARRIED ONCE

14.2%
n=1507

MARRIED MORE THAN ONCE

25.8%
n=299

NO DECEASED CHILD

12.7%
n=1309

DISC1=A/A,T/T

10.2%
n=766

NO CHANCE LOSE JOB**

5.6%
n=54

CHANCE LOSE JOB**

29.6%
n=27

DISC1=T/A

16.4%
n=532

HAS DECEASED CHILD

26.8%
n=138

DISC1=A/A,T/T

**NOT MARRIED**: 29.1%
n=652

LEAD ‘14

CIDI-SF Depression in Women (G x E)
## Logistic Regression

### CIDI-SF Depression Risk in Women

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Risk ratio</th>
<th>P-value (95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOT MARRIED</td>
<td>2.330</td>
<td>&lt; 2x10^{-12} ***</td>
</tr>
<tr>
<td>MULTIPLE MARRIAGES</td>
<td>2.023</td>
<td>&lt; 5x10^{-8} ***</td>
</tr>
<tr>
<td>DEATH OF CHILD</td>
<td>1.539</td>
<td>&lt; 9x10^{-3} ***</td>
</tr>
<tr>
<td>DISC1 T/A</td>
<td>1.299</td>
<td>&lt; 2x10^{-2} *</td>
</tr>
<tr>
<td>DRD2.3 C/C or T/C</td>
<td>1.310</td>
<td>&lt; 2x10^{-2} *</td>
</tr>
<tr>
<td>FTO C/C or T/T</td>
<td>1.335</td>
<td>&lt; 1x10^{-2} ***</td>
</tr>
<tr>
<td>APOC3 T/T</td>
<td>1.351</td>
<td>&lt; 8x10^{-3} ***</td>
</tr>
</tbody>
</table>

* Marginally significant
*** Significant after controlling for multiple testing
CIDI-SF Depression Summary

• In general, environmental factors were more predictive of depression
  – marital status most important factor for both men and women

• Polymorphisms in neurotransmitter and neuroendocrine genes were most frequently associated with CIDI-SF
  – DRD2.2 and GNRH1 play significant role in men
  – DRD2.3, FTO, DISC1 and APOC3 play significant role in women
  – twice the depression rate in women
  – continuous dopamine exposure associated with depression

• Have yet to identify the most predictive genetic alterations
Conclusions

• Recursive Partitioning, MDR, LR and Machine Learning Analyses
  • Utilizing all these techniques to identify meaningful G x G and G x E interactions.

• This study also demonstrates the utility of recursive partitioning analyses as an efficient and powerful exploratory analysis technique for uncovering genetic and molecular pathway interactions associated with disease etiology.

• The exhaustive search properties and non-parametric nature of these analytic techniques suggest applicability to extremely high-dimensional genetic data such as those generated for GWAS

• These results demonstrate the importance of assessing G x G and G x E interactions in identifying genetic and environmental predictors of increased and decreased predictive value for diseases and conditions
Future WLS Genotyping Studies

• WLS is an ideal candidate for future GWAS studies given its large sample size, rich covariate composition and longitudinal nature

• Graduates – 75 years of age

• Cusp of large increases in age-related diseases and mortality

• The next 10 years will provide a windfall of genetic association data for this unique cohort
The WLS Genotyping Team
(Past and Present)

Department of Sociology
Pamela Herd, Ph.D.
Carol Roan, Ph.D.
Robert M. Hauser, Ph.D.
Taissa Hauser, M.A.
Vicky Chang, Ph.D.
James A. Jonker, M.A.
Chee Lee, B.A.
Susan Bright, B.A.
Nicholas Roetker, B.A.
Tina Gonzales, M.S.

Department of Medicine
Craig S. Atwood, Ph.D.
Jacob Basson, B.S.

Funding: National Institute on Aging, (AG-9775, AG-21079, and AG-033285).