Assessing variants in the human genome

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

November 19, 2010

Very large data sets

Statistical Methods for Very Large Datasets Conference 2011

Wednesday, June 01, 2011 6:00 AM -
Friday, June 03, 2011 5:30 PM (Eastern Time)

InterContinental Harbor Court Hotel
800-824-0076
550 Light Street
Baltimore, Maryland 21202
Map and Directions

Welcome!
The Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health invites you to a 3-day conference on Very Large Data Sets. The conference is scheduled from June 1-3, 2011 and will be hosted in beautiful, downtown Baltimore, Maryland, USA at the InterContinental Harbor Court Hotel.

The conference has a one-track session for invited presentations and a high profile session for contributed poster presentations. A panel discussion will attempt to define what large data sets are, anticipate new challenges, and identify possible solutions.

Conference Overview
There is an acute and increasing need to adapt standard statistical methods and to develop new approaches for the analysis of very large data sets. A data set is very large if it raises difficult or insurmountable computational problems for standard data analysis using available computing systems. The continuous increase in size and complexity of data sets is due in part to increased computational and storage capabilities, new measurement technologies and study designs, and an increasing number of study "units".
Very large data sets

Ingo Ruczinski
Assessing variants in the human genome

Important Dates:
June 1, 2010: Call for poster presentation abstracts
February 1, 2011: Final date for early bird registration fee
March 1, 2011: Final date for submission of poster abstracts
April 1, 2011: Notification for poster abstract acceptance
May 31, 2011: Final date for reduced conference rate at hotel
June 1-3, 2011: Conference dates

Confirmed Speakers include:
Goncalo Abecasis, University of Michigan
DuBois Bowman, Emory University
Brian Caflisch, Johns Hopkins University
Raymond Carroll, Texas A&M University
Ciprian Crainiceanu, Johns Hopkins University
Francesca Dominici, Harvard University
William DuMouchel, Phase Forward Lincoln Safety Group
Sandrine Dutto, University of California at Berkeley
Jay Emerson, Yale University
Stephen Eubank, Virginia Tech
Montass Fuentes, North Carolina State University
Robert Gentleman, Fred Hutchinson Cancer Research Center
Rafael Irizarry, Johns Hopkins University
Hongkai Ji, Johns Hopkins University
Nicole Lazar, University of Illinois at Chicago
Jeffrey Morris, MD Anderson Cancer Center
Hans-Georg Muller, University of California at Davis
Doug Nyhuis, National Center for Atmospheric Research
Todd Ogden, Columbia University
Roger Peng, Johns Hopkins University
James Ramsay, McGill University
Ingo Ruczinski, Johns Hopkins University
Steven Salzberg, University of Maryland
Terry Speed, University of California at Berkeley
John Storey, Princeton University
Alex Szalay, Johns Hopkins University
Jonathan Taylor, Stanford University
Chris Volinsky, AT&T Labs-Research

http://biostat.jhsph.edu/~iruczins/
ingo@jhu.edu
Acknowledgments

Collaborators: Kathleen Barnes, Terri Beaty, Benilton Carvalho, Bob Cole, Rafael Irizarry, Tom Louis, Rasika Mathias, Matt Ritchie, Rob Scharpf, Holger Schwender, Keith West.

Computing support: Marvin Newhouse, Jiong Yang.

Funding: NIH R01 DK061662, GM083084, HL090577, and a CTSA grant to the Johns Hopkins Medical Institutions.

Single nucleotide polymorphisms

[Diagram showing two DNA sequences with a SNP highlighted]
Ingo Ruczinski  Assessing variants in the human genome

WTCCC

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study using the Affymetrix GenoChip 500K Mapping Array Set undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases, and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at \( P < 5 \times 10^{-7} \) in bipolar disorder, 1 in coronary artery disease, 9 in Crohn’s disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 5 in type 2 diabetes. On the basis of prior findings and replication studies thus far completed, almost all of these signals reflect genuine susceptibility effects. We observed associations at many previously identified loci, and found compelling evidence that some but not all risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-variant \( P \) values between \( 10^{-5} \) and \( 5 \times 10^{-7} \)) likely to yield additional susceptibility loci. The importance of appropriately large sample sizes confirmed by the modest effect sizes observed at most loci identified. This study thus represents a thorough validation of the GWA approach. It has also demonstrated that careful use of a shared control group represents a safe and effective approach to GWA analyses of multiple diseases: phenotypes has generated a genome-wide genotype database for future studies of common diseases in the British population, and it is clear that, provided individuals with non-European ancestry are excluded, the extent of population stratification in the British population is generally modest. Our findings offer new avenues for exploring the pathophysiology of these important disorders. We anticipate that our data, results and software, which will be widely available to other investigators, will provide a powerful resource for human genetics research.


Ingo Ruczinski

Assessing variants in the human genome

Results

http://www.genome.gov/GWAstudies/
Efficient Evaluation of Ranking Procedures when the Number of Units is Large, With Application to SNP Identification

Thomas A. Louis * and Ingo Ruczinski
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore MD.

Prepared in honor of Hans van Houwelingen.

Summary
Simulation-based assessment is a popular and frequently necessary approach to evaluation of statistical procedures. Sometimes overlooked is the ability to take advantage of underlying mathematical relations and we focus on this aspect. We show how to take advantage of large-sample theory when conducting a simulation using the analysis of genomic data as a motivating example. The approach uses convergence results to provide an approximation to smaller-sample results, results that are available only by simulation. We consider evaluating and comparing a variety of ranking-based methods for identifying the most highly associated SNPs in a genome-wide association study, derive integral equation representations of the posterior distribution of percentiles produced by three ranking methods, and provide examples comparing performance. These results are of interest in their own right and set the framework for a more extensive set of comparisons.

Key words: Efficient simulation, ranking procedures, SNP identification.

Ingo Ruczinski  Assessing variants in the human genome

Case-parent trios

<table>
<thead>
<tr>
<th>Recruitment Site</th>
<th>CL</th>
<th>CLP</th>
<th>CP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utah</td>
<td>68</td>
<td>96</td>
<td>52</td>
<td>216</td>
</tr>
<tr>
<td>Norway</td>
<td>106</td>
<td>174</td>
<td>107</td>
<td>387</td>
</tr>
<tr>
<td>Korea</td>
<td>19</td>
<td>40</td>
<td>5</td>
<td>64</td>
</tr>
<tr>
<td>Maryland</td>
<td>19</td>
<td>71</td>
<td>25</td>
<td>115</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>26</td>
<td>70</td>
<td>11</td>
<td>107</td>
</tr>
<tr>
<td>Singapore</td>
<td>15</td>
<td>45</td>
<td>53</td>
<td>113</td>
</tr>
<tr>
<td>Taiwan</td>
<td>42</td>
<td>176</td>
<td>74</td>
<td>292</td>
</tr>
<tr>
<td>Iowa</td>
<td>16</td>
<td>29</td>
<td>24</td>
<td>69</td>
</tr>
<tr>
<td>Denmark</td>
<td>6</td>
<td>15</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>Philippines</td>
<td>0</td>
<td>94</td>
<td>0</td>
<td>94</td>
</tr>
<tr>
<td>WuHan</td>
<td>39</td>
<td>136</td>
<td>42</td>
<td>217</td>
</tr>
<tr>
<td>Shandong Province</td>
<td>54</td>
<td>129</td>
<td>30</td>
<td>213</td>
</tr>
<tr>
<td>Western China</td>
<td>43</td>
<td>63</td>
<td>38</td>
<td>144</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>453</td>
<td>1138</td>
<td>466</td>
<td>2057</td>
</tr>
</tbody>
</table>
Case-parent trios

F : 12  M : 12  C : 11
F : 12  M : 12  C : 12
F : 12  M : 12  C : 22

F : 11  M : 12  C : 11
F : 11  M : 12  C : 12
F : 11  M : 12  C : 12

F : 12  M : 22  C : 12
F : 12  M : 22  C : 12
F : 12  M : 22  C : 22

Ingo Ruczinski  Assessing variants in the human genome

Genotypic TDT

Assume that at a certain locus the father has alleles $11$ and the mother has alleles $12$. The four Mendelian children thus have alleles $11, 12, 11, \text{ and } 12$.

Assume the affected proband has genotype $11$.

The three Pseudo controls then have the genotypes $11, 12, \text{ and } 12$.

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affected proband</td>
<td>1</td>
</tr>
<tr>
<td>Pseudo control #1</td>
<td>0</td>
</tr>
<tr>
<td>Pseudo control #2</td>
<td>0</td>
</tr>
<tr>
<td>Pseudo control #3</td>
<td>0</td>
</tr>
</tbody>
</table>

We can use conditional logistic regression to analyze the data.
The transmission disequilibrium test measures the over-transmission of an allele from parents to affected offspring. For a set of n parents with alleles 1 and 2 at a genetic locus, each parent can be summarized by the transmitted and the non-transmitted allele:

<table>
<thead>
<tr>
<th></th>
<th>Non-TA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2</td>
<td>∑</td>
</tr>
<tr>
<td>TA</td>
<td>a b</td>
<td>a+b</td>
</tr>
<tr>
<td></td>
<td>c d</td>
<td>c+d</td>
</tr>
<tr>
<td>∑</td>
<td>a+c</td>
<td>b+d</td>
</tr>
<tr>
<td></td>
<td>2n</td>
<td></td>
</tr>
</tbody>
</table>

Only the heterozygous parents contribute information!

Under the null of no association, \( \frac{(b - c)^2}{b + c} \sim \chi^2_1 \)

→ Even better, use `binom.test()` in R.

---

**GWAs results**

a. All CL/P

b. Asian CL/P

c. European CL/P

[BEA ... RUC ... SCO | NAT-GEN 2010]
### Fast genotypic TDT

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Computing Time [in sec]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>11</td>
<td>45</td>
</tr>
<tr>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>13</td>
<td>35</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>22</td>
<td>30</td>
</tr>
</tbody>
</table>

---

### Candidate genes

**Genetic Determinants of Facial Clefting: Analysis of 357 Candidate Genes Using Two National Cleft Studies from Scandinavia**

Astanand Jugessur, Min Shi, Håkon Kristian Gjesling, Rolf Terje Lia, Allen James Wilcox, Clarice Ring Weinberg, Kaare Christensen, Abe Lowman Boyles, Sandra Daacke-Hirsch, True Nguyen Trung, Camilla Bille, Andrew Carl Lidral, Jeffrey Clark Murray

1. Center for Development, Musculoskeletal Disorders, Murdoch Children’s Research Institute, Royal Children’s Hospital, Parkville, Australia; 2. Department of Epidemiology, National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, Durham, North Carolina, United States of America; 3. Department of Epidemiology, National Institute of Public Health, Oslo, Norway; 4. Department of Epidemiology and Medical Genetics, University of Southern Denmark, Odense, Denmark; 5. College of Nursing, University of Iowa, Iowa City, Iowa, United States of America; 6. Department of Pediatrics, Epidemiology and Biological Sciences, University of Iowa, Iowa City, Iowa, United States of America

**Abstract**

**Background:** Facial clefts are common birth defects with a strong genetic component. To identify fetal genetic risk factors for clefting, 1538 SNPs in 357 candidate genes were genotyped in two population-based samples from Scandinavia (Norway 562 case-parent and 592 control-parent triads; Denmark 235 case-parent triads).

**Methodology/Principal Findings:** We used two complementary statistical methods, TRIMM and HAPLIN, to look for associations across these two national samples. TRIMM tests for association in each gene by using multi-SNP genotypes from case-parent triads directly without the need to infer haplotypes. HAPLIN on the other hand estimates the full haplotype distribution over a set of SNPs and estimates relative risks associated with each haplotype. For isolated cleft lip with or without cleft palate (I-CLP), TRIMM and HAPLIN both identified significant associations with I-CLP and ADHFS in both populations, but only HAPLIN found an association with FGFR2. For isolated cleft palate (O-CLP), TRIMM found associations with CHD2, ABCC6, and PDZC4 in both populations, but only the association with PDZC4 was identified by HAPLIN. In addition, HAPLIN identified an association with FTS that was not detected by TRIMM.

**Conclusion/Significance:** Strong associations with seven genes were replicated in the Scandinavian samples and our approach effectively replicated the strongest previously known association with I-CLP. Based on two national cleft cohorts of similar ancestry, two robust statistical methods and a large panel of SNPs in the most promising cleft condition genes to date, this study identified a previously unknown association with clefting for ADHFS and provides additional candidates and analytic approaches to advance the field.

**Keywords:** Facial clefting, Genetics, Candidate genes, Population-based samples, Scandinavia.
Parent-of-origin effects

<table>
<thead>
<tr>
<th>Parental</th>
<th>Maternal</th>
<th>PO-LRT&lt;sup&gt;6&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>SNP name</td>
<td>T</td>
</tr>
<tr>
<td>1</td>
<td>rs7771980</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>rs2677104</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>rs8019055</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>rs8189554</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>rs910586</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>rs2819953</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>rs795724</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>rs1345799</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>rs2819861</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>rs2791013</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>rs2796093</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>rs2796098</td>
<td>15</td>
</tr>
<tr>
<td>13</td>
<td>rs4714854</td>
<td>15</td>
</tr>
<tr>
<td>14</td>
<td>rs4724834</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>rs2364442</td>
<td>17</td>
</tr>
<tr>
<td>16</td>
<td>rs1934328</td>
<td>41</td>
</tr>
<tr>
<td>17</td>
<td>rs7773575</td>
<td>33</td>
</tr>
<tr>
<td>18</td>
<td>rs7771899</td>
<td>36</td>
</tr>
<tr>
<td>19</td>
<td>rs10455422</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>rs6634383</td>
<td>13</td>
</tr>
<tr>
<td>21</td>
<td>rs13207392</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>rs7748231</td>
<td>13</td>
</tr>
<tr>
<td>23</td>
<td>rs10948237</td>
<td>13</td>
</tr>
<tr>
<td>24</td>
<td>rs1928533</td>
<td>12</td>
</tr>
</tbody>
</table>

[ Sul · · · Ruc · · · Bea | Gen·Epi 2008 ]

Ingo Ruczinski  Assessing variants in the human genome

Copy number estimates are noisy
Plate effects

Confounding of plate and disease
Genotype estimates are more robust.

Birdseed

CRLMM

De-novo deletions
De-novo deletions
De-novo deletions

Ingo Ruczinski
Assessing variants in the human genome
At locus $i$, for subject $j$ in plate $p$, we have for allele $k \in \{A, B\}$

\[
I_{kip} = \nu_{kip} \delta_{kip} + \phi_{kip} c_{kip} \epsilon_{kip} \implies \hat{c}_{kip} = \max \left\{ \frac{1}{\phi_{kip}} (I_{kip} - \nu_{kip}), 0 \right\}
\]
Trisomy 21

Samples from Aravinda Chakravarti and Betty Doan

Ingo Ruczinski  Assessing variants in the human genome
Ingo Ruczinski  
Assessing variants in the human genome

A versus B plots

SNP_A–8339372

Ingo Ruczinski  
Assessing variants in the human genome

A versus B plots

SNP_A–8339372
A versus B plots

SNP_A–8340560

Ingo Ruczinski
Assessing variants in the human genome
A versus B plots

SNP_A–1969323

Ingo Ruczinski  Assessing variants in the human genome
Trisomy 21

Samples from Aravinda Chakravarti and Betty Doan

Ingo Ruczinski  Assessing variants in the human genome
Trisomy 21

Assessing variants in the human genome

Prediction regions for copy number

[ SCH · RUC · CAR · DOA · CHA · IRI | BIOSTAT 2010 ]
Vanilla and ICE HMMs for genotype and copy number

[ SCH · PAR · PEV · RUC | AOAS 2008 ]

Ingo Ruczinski  Assessing variants in the human genome

Open source software

[ SCH · RUC | BIOINF 2007 ]  ●  [ SCH · RUC | M-MOL-BIO 2010 ]  ●  [ SCH · RUC · · · IRI | BIOSTAT 2010 ]

Ingo Ruczinski  Assessing variants in the human genome
Using the R Package crrlm for Genotyping and Copy Number Estimation

Robert B Scharpf  
Johns Hopkins University

Rafael A Irizarry  
Johns Hopkins University

Matthew E Ritchie  
Walter+Eliza Hall Institute of Medical Research

Benilton Carvalho  
University of Cambridge

Ingo Ruczinski  
Johns Hopkins University

Abstract

Genotyping platforms such as Affymetrix can be used to assess genotypes-phenotype as well as copy number-phenotype associations at millions of markers. While genotyping algorithms are largely concordant when assessed on HapMap samples, tools to assess copy number changes are more variable and often discordant. One explanation for the discordance is that copy number estimates are susceptible to systematic differences between groups of samples that were processed at different times or by different labs. Analysis algorithms that do not adjust for batch effects are prone to spurious measures of association. The R package crrlm implements a multilevel model that adjusts for batch effects and provides allele-specific estimates of copy number. This paper illustrates a workflow for the estimation of allele-specific copy number, develops marker- and study-level summaries of batch effects, and demonstrates how the marker-level estimates can be integrated with complimentary Bioconductor software for inferring regions of copy number gains or losses. All analyses are performed in the statistical environment R. A compendium for reproducing the analysis is available from the author’s website (http://www.bioinfo.jhmi.edu/research/crrlmCompendium/index.html).

Keywords: copy number, batch effects, robust, multilevel model, high-throughput, oligonucleotide array.

Ingo Ruczinski  
Assessing variants in the human genome

Compendium

The crrlmCompendium package contains the raw, data, and R functions used to make the figures in this paper. Users should be able to reproduce the figures upon successful installation of the compendium. The compendium requires R >= 2.12. To install the compendium and its dependencies you will need an internet connection.

```
install.packages("crrlmCompendium", dependencies=TRUE)
```

To install the crrlmCompendium, download the zip file of the latest build.

```
R package - build
crrlmCompendium.rda
```

The package can be installed from the command line by R CMD INSTALL crrlmCompendium_1.4.4.tar.gz, or from an R session in the same directory by:

```
install.packages("crrlmCompendium_1.4.4.tar.gz", repos="")
```

Windows users would first need to install the appropriate R/packaga(Release) environment.

R code extracted from the manuscript.Rnw vignette for reproducing the figures is available from the Code link adjacent to the figures below. To reproduce the figures, simply copy the code into R.

2.2 Reproducing the Manuscript

The complete analysis of the HapMap phase III data is contained in the manuscript.Rnw source file. This document is located in the Lincs/variab/ directory of the crrlmCompendium package. Three additional steps are required for the complete analysis. First, one must download and install the HapMap Phase 3.1.2 data for the Affymetrix 6.0 platform. Secondly, one must change the following lines in the manuscript.Rnw vignette as appropriate:

```
path/variable <- "path/to/HapMap_phase_3.1.2"
```

Finally, one must install additional package dependencies that were not required for installing the crrlmCompendium. In particular, the packages Rcpp, genefilter, ellipse, and dada. Note that the genotyping and copy number estimation steps in the manuscript.Rnw also require long computations. We suggest submitting the code using R CMD BATCH. Provided that LATEX is installed, a pdf version of the manuscript can be generated by issuing the following commands from R:

```
lualatex manuscript.Rnw
lualatex manuscript.Rnw
```

3 Figures and Code

<table>
<thead>
<tr>
<th>Figure</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="code1.R" alt="Code" /></td>
</tr>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="code2.R" alt="Code" /></td>
</tr>
</tbody>
</table>
Finding the missing heritability of complex diseases

Teri A. Manolio¹, Francis S. Collins¹, Nancy J. Cox², David B. Goldstein⁶, Lucia A. Hindorff⁵, David J. Hunter⁶, Mark I. McCarthy³, Erin M. Ramos², Lon R. Cardon⁶, Aravinda Chakravarti², Judy H. Cho⁷, Alan E. Guttmacher¹, Augustine Kong¹, Leonid Kruglyak¹, Elaine Mardis¹, Charles N. Rotimi¹, Montgomery Slakkin⁷, David Valle⁷, Alice S. Whittemore⁶, Michael Boehnke⁵, Andrew G. Clark⁴, Evan E. Eichler⁵, Greg Gibson⁷, Jonathan L. Haines⁷, Trudy F. C. Mackay⁵, Steven A. McCarroll⁷ and Peter M. Visscher⁷

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained. Here we examine potential sources of missing heritability and propose research strategies, including and extending beyond current genome-wide association approaches, to illuminate the genetics of complex diseases and enhance its potential to enable effective disease prevention or treatment.

Estimates of heritability and number of loci for several complex traits

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of loci</th>
<th>Proportion of heritability explained</th>
<th>Heritability measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age-related macular degeneration²</td>
<td>5</td>
<td>50%</td>
<td>Sibling recurrence risk</td>
</tr>
<tr>
<td>Crohn's disease²¹</td>
<td>32</td>
<td>20%</td>
<td>Genetic risk (liability)</td>
</tr>
<tr>
<td>Systemic lupus erythematosus³</td>
<td>6</td>
<td>15%</td>
<td>Sibling recurrence risk</td>
</tr>
<tr>
<td>Type 2 diabetes²⁴</td>
<td>18</td>
<td>6%</td>
<td>Sibling recurrence risk</td>
</tr>
<tr>
<td>HDL cholesterol⁷⁶</td>
<td>7</td>
<td>5.2%</td>
<td>Residual phenotypic variance</td>
</tr>
<tr>
<td>Height⁴⁵</td>
<td>40</td>
<td>5%</td>
<td>Phenotypic variance</td>
</tr>
<tr>
<td>Early onset myocardial infarction⁷⁰</td>
<td>9</td>
<td>2.8%</td>
<td>Phenotypic variance</td>
</tr>
<tr>
<td>Fasting glucose⁷⁷</td>
<td>4</td>
<td>1.5%</td>
<td>Phenotypic variance</td>
</tr>
</tbody>
</table>

¹Residual is after adjustment for age, gender, diabetes.
Genetic Heterogeneity in Human Disease

Jon McClellan* and Mary-Claire King**

*Department of Psychiatry
**Departments of Medicine and Genome Sciences
University of Washington, Seattle, WA 98195-7720, USA
*Correspondence: drjack@uw.edu (J.M.), mcking@uw.edu (M.-C.K.)
DOI 10.1016/j.cell.2010.03.002

Strong evidence suggests that rare mutations of severe effect are responsible for a substantial portion of complex human disease. Evolutionary forces generate vast genetic heterogeneity in human illness by introducing many new variants in each generation. Current sequencing technologies offer the possibility of finding rare disease-causing mutations and the genes that harbor them.

[McClellan and King | Cell 141: 2010]
Genetic heterogeneity

From the New Yorker

"O.K., let's slowly lower in the grant money."

Todd Bearson
Arlington, Mass.
### NHLBI exome sequencing project

**Program Officer:** Deborah Applebaum-Bowden

<table>
<thead>
<tr>
<th>Heart GO</th>
<th>Lung GO</th>
<th>Women’s Health Initiative Sequencing Program</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PI:</strong> Stephen Rich</td>
<td><strong>PIs:</strong> Michael Bamshad</td>
<td><strong>PI:</strong> Rebecca Jackson</td>
</tr>
<tr>
<td>University of Virginia</td>
<td>U. of Washington</td>
<td>Ohio State University</td>
</tr>
<tr>
<td></td>
<td>Kathleen Barnes, Johns Hopkins University</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Pls:</strong> Debbie Nickerson, Mark Rieder, Jay Shendure, Phil Green</td>
<td><strong>Pls:</strong> Stacey Gabriel &amp; David Altshuler</td>
</tr>
</tbody>
</table>