Variation in the human genome & disease

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

April 7, 2010

Linkage

From Goncalo Abecasis
Linkage versus association

From Goncalo Abecasis
Ingo Ruczinski  Variation in the human genome & disease

Single nucleotide polymorphisms

From Goncalo Abecasis
Ingo Ruczinski  Variation in the human genome & disease

urgi.versailles.inra.fr
Haplotypes


Ingo Ruczinski  Variation in the human genome & disease

Haplotypes

Haplotype blocks

Kim and Dionne (2007)

Case-control design


Population stratification


Principal components


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 GeneChip 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 \( P < 5 \times 10^{-8} \), in bipolar disorder, 1 in coronary artery disease, 9 in Crohn’s disease, 7 in rheumatoid arthritis, 3 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 association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 38 loci with single-variant \( P \) values between \( 10^{-7} \) and \( 5 \times 10^{-7} \)) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes observed at these loci. 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 analysis of multiple disease phenotypes. It has generated a genome-wide genotype database for future studies of common diseases in the British population, and shows 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.
Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout

Owen M. Woodward,1,3 Anna Kötting,1,3 Josef Cernohorsky,1,3 Eric Boenssens,1,3 William B. Guggino,2 and Michael Kötting1,3

1Department of Physiology, Johns Hopkins Medical Institutions, Baltimore, MD 21205; 2Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; 3Human Genetics Center and Division of Epidemiology, University of Texas, Houston, TX 77030; and 4Department of Biological Chemistry, Johns Hopkins Medical Institutions, Baltimore, MD 21205

Edited by Maurina B. Burg, National Institutes of Health, Bethesda, MD, and approved May 4, 2000 (received for review February 4, 2000)

Genome-wide association studies (GWAS) have successfully identified common single nucleotide polymorphisms (SNPs) associated with a wide variety of complex diseases, but do not address gene function or establish causality of disease-associated SNPs. We recently used GWAS to identify SNPs in a genomic region on chromosome 4 that associate with serum urate levels and gout, a consequence of elevated urate levels. Here we use functional assays that human ATP-binding cassette, subfamily G, 2 (ABCG2), encoded by the ABCG2 gene contained in this region, is a hitherto unknown urate efflux transporter. We further show that native ABCG2 is located in the brush border membrane of kidney proximal tubule cells, where it mediates renal urate secretion. Introduction of the mutations Q141K encoded by the common SNP rs2331142 by site-directed mutagenesis resulted in 53% reduced urate transport rates compared to wild-type ABCG2 (P = 0.001). Data from a population-based study of 14,783 individuals support rs2331142 as the causal variant in the region and show highly significant associations with urate levels [odds: P = 1.9 - 30, minor allele frequency (MAF) 0.11; blacks P = 10^{-3}, MAF 0.03] and gout (adjusted odds ratio 1.68 per risk allele, both races). Our data indicate that at least 15% of all gout cases in whites are attributable to this causal variant. With approximately 3 million US individuals suffering from often insufficiently treated gout, ABCG2 represents an attractive drug target. Our study completes the chain of evidence from association to causation and supports the common disease-common variant hypothesis in the etiology of gout.
A HapMap harvest of insights into the genetics of common disease

Teri A. Manolio, Lisa D. Brooks, and Francis S. Collins
National Human Genome Research Institute, Bethesda, Maryland, USA.

The International HapMap Project was designed to create a genome-wide database of patterns of human genetic variation, with the expectation that these patterns would be useful for genetic association studies of common diseases. This expectation has been amply fulfilled with just the initial output of genome-wide association studies, identifying nearly 100 loci for nearly 40 common diseases and traits. These associations provided new insights into pathophysiology, suggesting previously unsuspected etiologic pathways for common diseases that will be of use in identifying new therapeutic targets and developing targeted interventions based on genetically defined risk. In addition, HapMap-based discoveries have shed new light on the impact of evolutionary pressures on the human genome, suggesting multiple loci important for adapting to disease-causing pathogens and new environments. In this review we examine the origin, development, and current status of the HapMap; its prospects for continued evolution; and its current and potential future impact on biomedical science.

Results

Catalogued findings for the following phenotypes:

- Eye diseases
- Diabetes
- Cancer
- Gastrointestinal disorders
- Cardiovascular conditions and lipid metabolism
- Neuropsychiatric conditions
- Autoimmune and infectious diseases
- Various traits


http://www.genome.gov/GWAstudies/
**Results**

Phenotypes under investigation in collaborative GWA studies

<table>
<thead>
<tr>
<th>GAIN®</th>
<th>GEIP</th>
<th>STAMpeed®</th>
<th>CGEMS®</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention deficit hyperactivity disorder</td>
<td>Type 2 diabetes</td>
<td>Early-onset myocardial infarction</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>Major depressive disorder</td>
<td>Maternal metabolism and birth weight</td>
<td>Asthma</td>
<td>Breast cancer</td>
</tr>
<tr>
<td>Bipolar I disorder</td>
<td>Preterm birth</td>
<td>Platelet phenotypes</td>
<td>Pancreatic cancer</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>Oral clefts</td>
<td>CHD and other heart, lung and blood disorders</td>
<td>Lung cancer</td>
</tr>
<tr>
<td>Type 1 diabetic nephropathy</td>
<td>Dental caries</td>
<td>Childhood respiratory outcomes</td>
<td>Bladder cancer</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>Coronary disease</td>
<td>Hematopoietic cell transplant outcome</td>
<td>Renal cancer</td>
</tr>
<tr>
<td></td>
<td>Lung cancer</td>
<td>Atherosclerosis in hypertensives</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Addicition</td>
<td>Asthma and lung function</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cardiovascular risk factors</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Atherosclerosis pathway genes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CV events</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Early coronary artery disease</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phenotypic variability in sickle cell anemia</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Centenarians</td>
<td></td>
</tr>
</tbody>
</table>


**Missing heritability**

**Finding the missing heritability of complex diseases**


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 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.

Missing heritability

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 degeneration</td>
<td>5</td>
<td>50%</td>
<td>Sibling recurrence risk</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>32</td>
<td>20%</td>
<td>Genetic risk (liability)</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>6</td>
<td>15%</td>
<td>Sibling recurrence risk</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>18</td>
<td>6%</td>
<td>Sibling recurrence risk</td>
</tr>
<tr>
<td>HDL cholesterol</td>
<td>7</td>
<td>5.2%</td>
<td>Residual phenotypic variance</td>
</tr>
<tr>
<td>Height</td>
<td>40</td>
<td>5%</td>
<td>Phenotypic variance</td>
</tr>
<tr>
<td>Early onset myocardial infarction</td>
<td>9</td>
<td>2.8%</td>
<td>Phenotypic variance</td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>4</td>
<td>1.5%</td>
<td>Phenotypic variance</td>
</tr>
</tbody>
</table>

*Residual is after adjustment for age, gender, diabetes.

1000 Genomes

A Deep Catalog of Human Genetic Variation

1000 Genomes Project

The 1000 Genomes Project is an international research consortium formed to create the most detailed and medically useful picture to date of human genetic variation. The project involves sequencing the genomes of approximately 1200 people from around the world and receives major support from the Wellcome Trust Sanger Institute in Hinxton, England, the Beijing Genomics Institute Shenzhen in China and the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH).

Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

On 4 September 2008, the co-chairs of the analysis group and overall project co-chairs drafted a letter to the NIH Council about 1000 Genomes Project. This letter, available here, reviews the goals, describes the current status, and provides an update on the critical tasks the Analysis Group must accomplish in order to deliver a valuable community resource and achieve the Project’s goals. Additional project details are available in the September 2007 meeting report.

http://www.1000genomes.org

Ingo Ruczinski

Variation in the human genome & disease

Trisomy
Karyotypes

General Cytogenetics Information  http://members.aol.com/chrominfo/

FISH

Courtesy of the Pevsner Laboratory

Ingo Ruczinski  Variation in the human genome & disease

Ingo Ruczinski  Variation in the human genome & disease
**De novo deletion**

![Image of de novo deletion analysis]

**Copy number variants**

Selected disease associations with rare CNVs and common CNPs

<table>
<thead>
<tr>
<th>Disease</th>
<th>Locus</th>
<th>Type of CNV</th>
<th>Size (kb)</th>
<th>Population frequency</th>
<th>Case frequency</th>
<th>Effect size (OR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rare CNVs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism [19]</td>
<td>16p11.2</td>
<td>De novo deletion</td>
<td>600</td>
<td>$1 \times 10^{-4}$</td>
<td>1%</td>
<td>100</td>
</tr>
<tr>
<td>Autism [19]</td>
<td>16p11.2</td>
<td>Rare duplication</td>
<td>600</td>
<td>$3 \times 10^{-4}$</td>
<td>0.50%</td>
<td>16</td>
</tr>
<tr>
<td>Schizophrenia [60, 78]</td>
<td>1q21.1</td>
<td>Rare deletion</td>
<td>1,400</td>
<td>$2 \times 10^{-4}$</td>
<td>0.30%</td>
<td>15</td>
</tr>
<tr>
<td>Schizophrenia [60, 78]</td>
<td>1q21.1</td>
<td>Rare deletion</td>
<td>1,400</td>
<td>$2 \times 10^{-4}$</td>
<td>0.47%</td>
<td>Not observed in 4,737 controls</td>
</tr>
<tr>
<td>Schizophrenia [60, 78]</td>
<td>15q13.3</td>
<td>Rare deletion</td>
<td>1,600</td>
<td>$2 \times 10^{-4}$</td>
<td>0.20%</td>
<td>12</td>
</tr>
<tr>
<td>Schizophrenia [60, 78]</td>
<td>15q13.3</td>
<td>Rare deletion</td>
<td>1,600</td>
<td>$2 \times 10^{-4}$</td>
<td>1.0%</td>
<td>Not observed in 3,696 controls</td>
</tr>
<tr>
<td>Schizophrenia [60, 78]</td>
<td>22q11.2</td>
<td>Rare deletion</td>
<td>3,000</td>
<td>$2.5 \times 10^{-4}$</td>
<td>1%</td>
<td>40</td>
</tr>
<tr>
<td>Common CNPs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crohn’s disease [43]</td>
<td>IRGM</td>
<td>Deletion polymorphism</td>
<td>20</td>
<td>7%</td>
<td>11%</td>
<td>1.5</td>
</tr>
<tr>
<td>Body mass index [61]</td>
<td>NEGR1</td>
<td>Deletion polymorphism</td>
<td>45</td>
<td>66%</td>
<td>Quantitative trait</td>
<td>&lt;1 kg</td>
</tr>
<tr>
<td>Psoriasis [64]</td>
<td>LCE3C</td>
<td>Deletion polymorphism</td>
<td>30</td>
<td>56%</td>
<td>66%</td>
<td>1.3</td>
</tr>
</tbody>
</table>

More information

The confidence in genotype calls can differ substantially between SNPs!
Plate effects

Bipolar GWAS (EA controls) from dbGap

A versus B plots
A versus B plots

SNP_A--6415968

SNP_A--2026121
A versus B plots

SNP_A--4298545

log(B)

log(A)

A versus B plots

SNP_A--4298545

log(B)

log(A)
A versus B plots

\[
\log_2(A) + \log_2(B)
\]

plates ordered by median date

Ingo Ruczinski  Variation in the human genome & disease
A versus B plots

SNP_A–1833154

SNP_A–4232920
A versus B plots

SNP_A–8700561

SNP_A–1895536

Ingo Ruczinski  Variation in the human genome & disease
A versus B plots

SNP_A–8525194

log₂(A)

log₂(B)

plate 1
plate 2

Trisomy 21

Chr 21

copy number

samples

Samples from Aravinda Chakravarti and Betty Doan
Trisomy 21

Samples from Aravinda Chakravarti and Betty Doan

A versus B plots

Ingo Ruczinski  Variation in the human genome & disease
A versus B plots

SNP_A–8341330

SNP_A–8339372

Ingo Ruczinski  Variation in the human genome & disease
A versus B plots

SNP_A–8340560

SNP_A–1969323
A versus B plots

SNP_A–1969323

Ingo Ruczinski  Variation in the human genome & disease

Trisomy 21

Chr 21

Samples from Aravinda Chakravarti and Betty Doan
Samples from Aravinda Chakravarti and Betty Doan

Ingo Ruczinski  Variation in the human genome & disease

Samples from Aravinda Chakravarti and Betty Doan

Ingo Ruczinski  Variation in the human genome & disease
Prediction regions for copy number


Vanilla and ICE HMMs for genotype and copy number

Open source software


Ingo Ruczinski
Variation in the human genome & disease

Acknowledgments

Collaborators: Rob Scharpf
Benilton Carvalho, Rafael Irizarry, Giovanni Parmigiani, Holger Schwender.
Terri Beaty & group, Kathleen Barnes & group.

Computing support: Marvin Newhouse, Jiong Yang.

Funding: NIH R01 DK061662, GM083084, HL090577, and a CTSA grant to the Johns Hopkins Medical Institutions.
http://biostat.jhsph.edu/~iruczins/