GWAS and prior knowledge to uncover gene-gene interactions

Marylyn D. Ritchie, PhD
Director, Center for Systems Genomics
The Pennsylvania State University
Biochemistry and Molecular Biology
July 18, 2013
As of 7/9/2013, the catalog includes 1,654 publications and 10,976 SNPs.
Distribution of Effects

Median = 1.28

Mostly tiny effects

Odds Ratio (upper inclusive bound)
Distribution of Effects
Missing Heritability

- Under our nose
- Out of sight
- In the architecture
- Underground networks
- Lost in diagnosis
- The great beyond

Biology is complex
Statistical vs. biological epistasis

Moore and Williams, BioEssays 27:637–646, 2005
If interactions with minimal main effects are the norm rather than the exception, can we analyze all possible combinations of loci with traditional approaches to detect purely interaction effects? NO
How many combinations are there?

- ~500,000 SNPs to span the genome (HapMap)

2 x 10^{26} combinations
* 1 combination per second
* 86400 seconds per day

\[
\frac{2 \times 10^{26}}{1 \times 86400} = 2.979536 \times 10^{21} \text{ days to complete}
\]

(8.163113 \times 10^{18} \text{ years})
How many combinations are there?

- ~500,000 SNPs to span the genome (HapMap)

**5 Million SNPs in current technology**

<table>
<thead>
<tr>
<th># SNPs</th>
<th># models</th>
<th>time**</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 SNP</td>
<td>5.00x10^6</td>
<td>5 sec</td>
</tr>
<tr>
<td>2 SNPs</td>
<td>1.25x10^{13}</td>
<td>144 days</td>
</tr>
<tr>
<td>3 SNPs</td>
<td>2.08x10^{19}</td>
<td>2.4x10^8 days</td>
</tr>
<tr>
<td>4 SNPs</td>
<td>2.60x10^{25}</td>
<td>3.01x10^{14} days</td>
</tr>
<tr>
<td>5 SNPs</td>
<td>2.60x10^{31}</td>
<td>3.01x10^{20} days</td>
</tr>
</tbody>
</table>

**assuming 1 CPU that performs 1 million tests per second**
**5 Million SNPs in current technology**

<table>
<thead>
<tr>
<th># SNPs</th>
<th># models</th>
<th>time**</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 SNP</td>
<td>$5.00 \times 10^6$</td>
<td>$5 \text{ sec}$</td>
</tr>
<tr>
<td>2 SNPs</td>
<td>$1.25 \times 10^{13}$</td>
<td>$144 \text{ days}$</td>
</tr>
<tr>
<td>3 SNPs</td>
<td>$2.08 \times 10^{19}$</td>
<td>$2.4 \times 10^8 \text{ days}$</td>
</tr>
<tr>
<td>4 SNPs</td>
<td>$2.60 \times 10^{25}$</td>
<td>$3.01 \times 10^{14} \text{ days}$</td>
</tr>
<tr>
<td>5 SNPs</td>
<td>$2.60 \times 10^{31}$</td>
<td>$3.01 \times 10^{20} \text{ days}$</td>
</tr>
</tbody>
</table>

**assuming 1 CPU that performs 1 million tests per second**
Epistasis Analysis in GWAS data

- Exhaustive evaluation
- Evaluate interactions in top hits from single-SNP analysis
- Use prior biological knowledge to evaluate specific combinations – “Candidate Epistasis”

The Biofilter

- Use publicly available databases to establish relationships between gene-products
- Suggestions of biological epistasis between genes
- Integrating information from the genome, transcriptome, and proteome into analysis

LOKI: Library of Knowledge Integration

- Genes and SNPs:
  - dbSNP
  - NCBI Entrez Gene
- Gene and Protein Interactions:
  - BioGRID
  - MINT
  - PharmGKB
- Pathways:
  - Gene Ontology
  - KEGG
  - Reactome
  - NetPath
- Protein Families:
  - Pfam
- Regulatory Regions:
  - ORegAnno
- ECRs:
  - UCSC
- Annotations:
  - NHGRI GWAS Catalog

The Biofilter

• Method described: Bush et al. 2009 *Pacific Symposium on Biocomputing*

• Applications
  - Multiple Sclerosis
    • Bush et al. 2009 ASHG talk, 2011 *Genes & Immunity*
  - HDL
    • Turner et al. 2010 ASHG Talk, 2011 *PLoS ONE*
  - HIV Pharmacogenomics
    • Grady et al. 2010 ASHG poster, 2011 *Pacific Symposium on Biocomputing*
  - Lipid traits
    • Holzinger et al. in preparation
  - BMI
    • Verma et al., in preparation
  - Cataracts
    • Hall et al., in preparation
Using Biofilter: GWAS Annotation

Are there biological relationships between significant results?

Single Locus Statistical Results
- SNP 1, Rs101841, \( p = 0.000163 \)
- SNP 2, Rs182645, \( p = 0.000268 \)
- SNP 3, Rs23876, \( p = 0.00324 \)
- SNP 4, Rs378645, \( p = 0.004354 \)
- SNP 5, Rs37564, \( p = 0.02341 \)
- SNP 6, Rs8751, \( p = 0.03412 \)
- SNP 7, Rs86745, \( p = 0.03685 \)
- SNP 8, Rs41254, \( p = 0.04675 \)

Biofilter Analysis

Annotated Statistical Results

- **Results in the Same Gene**
  - SNP 3, Rs23876, \( p = 0.00324 \)
  - SNP 6, Rs8751, \( p = 0.03412 \)

- **Results in the Same Pathway**
  - SNP 2, Rs182645, \( p = 0.000268 \)
  - SNP 3, Rs23876, \( p = 0.00324 \)
  - SNP 6, Rs8751, \( p = 0.03412 \)
  - SNP 7, Rs86745, \( p = 0.03685 \)

- **Results with Biological Interaction**
  - SNP 3, Rs23876, \( p = 0.00324 \)
  - SNP 6, Rs8751, \( p = 0.03412 \)
  - SNP 7, Rs86745, \( p = 0.03685 \)
Using Biofilter: Prioritizing Analysis

Is there epistasis in genes whose products interact either directly or through a metabolic intermediate?
Using Biofilter: Prioritizing Analysis

Is there epistasis between genes of two related pathways?

Pathway 1

- SNP 1
- SNP 2
- SNP 3
- SNP 4
- SNP 5
- SNP 6

Pathway 2

- SNP 7
- SNP 8
- SNP 9
- SNP 10

Disease Phenotype

Models

[1, 7]
[1, 8]
[1, 9]
[1, 10]
[2, 7]
[2, 8]
Etc.
## Candidate Approaches

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Smaller set of genes to explore</td>
<td>• Limited by current state of knowledge</td>
</tr>
<tr>
<td>• Fewer statistical tests</td>
<td>• Limitations of learning completely novel biology</td>
</tr>
<tr>
<td>• Results will have solid interpretations</td>
<td></td>
</tr>
</tbody>
</table>
A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility

930 trio families
Genotyped on Affymetrix 500K array
Post QC ~300,000 SNPs

Figure 1
Knowledge-Driven Multi-Locus Analysis Reveals Gene-Gene Interactions Influencing HDL Cholesterol Level in Two Independent EMR-Linked Biobanks

Stephen D. Turner¹, Richard L. Berg², James G. Linneman², Peggy L. Peissig², Dana C. Crawford¹, Joshua C. Denny³, Dan M. Roden⁴,⁵, Catherine A. McCarty⁶, Marylyn D. Ritchie¹, Russell A. Wilke⁴*

1 Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, 2 Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America, 3 Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, 4 Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, 5 Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, 6 Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America

- eMERGE Genome-wide association study (Illumina 660)
- Phenotype: median HDL for anyone having 2+ HDL measurements in their EMR
- Marshfield PMRP  n=3903
- Vanderbilt BioVU  n=1858
Knowledge-Driven Multi-Locus Analysis Reveals Gene-Gene Interactions Influencing HDL Cholesterol Level in Two Independent EMR-Linked Biobanks

Stephen D. Turner¹, Richard L. Berg², James G. Linneman², Peggy L. Peissig², Dana C. Crawford¹, Joshua C. Denny³, Dan M. Roden⁴,⁵, Catherine A. McCarty⁶, Marylyn D. Ritchie¹, Russell A. Wilke⁴*

¹ Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, ² Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America, ³ Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, ⁴ Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, ⁵ Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, ⁶ Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America

[Diagram of lipid metabolism]
Future Directions

1) SNPs from GWAS catalog for a particular disease-trait association

2) Map SNPs -> gene -> pathway using Biofilter

3) SNPs from KEGG, Reactome, or Netpath linked to SNPs from GWAS Catalog in LOKI

4) Exhaustive SNP-SNP models
   - SNP1 – SNP2
   - SNP1 – SNP3
   - SNP1 – SNP4
   - SNP1 – SNP5
   - ...

Pathway 1

SNP 1
SNP 2
SNP 3
SNP 4
SNP 5
SNP 6
Summary

• Biofilter is a bioinformatics application to annotate, filter, and construct gene-gene models for evaluation

• We have successfully used Biofilter in a number of genome-wide interaction analyses to identify replicating/confirmatory gene-gene models

• The GWAS catalog is an important and useful public database incorporated into LOKI – the knowledge base from which Biofilter draws its information
Future Directions

• Integrate more public databases into LOKI
  – Regulatory regions
  – Non-coding regions

• Develop additional filtering and model construction strategies based on specific hypotheses

• Develop a user-interface for ease of use
Acknowledgements

Ritchie Lab
Gretta Armstrong, project manager
Carrie Buchanan Moore, MD/PhD student*
Scott Dudek, software developer
Alex Frase, software developer*
Molly Hall, PhD student
Neerja Katiyar, PhD student*
Dokyoong Kim PhD, Postdoctoral fellow
Ruowang Li, PhD student
Sarah Pendergrass PhD, Research Associate*
Anurag Verma, Bioinformatics Programmer
Shefali Verma, Bioinformatics Analyst
John Wallace, software developer*
Dan Wolfe, bioinformatics research assistant*

HDL project - eMERGE

MS project - IMSGC

* - working on Biofilter
Just because we have not found it yet, doesn’t mean it’s not there.....