Using Protein-Protein Interactions to Identify Functional Connections Between Disease Loci

MPG Workshop
May 5th, 2010
Liz Rossin
Premise

• GWAS identifies regions throughout the genome that harbor risk for disease
• The genes in these regions may participate in a common pathway or process

➤ Can we use protein-protein interaction information to uncover these processes?
➤ If so, can we leverage from discovered pathways hypotheses about candidate genes?
What is a Protein-Protein Interaction?

- Sourced from large/small scale experiments
  - Y2H, complex purification, tandem affinity purification
- What data do we use?
  - The InWeb database (Lage et al. 2007)
Methods: Building Networks

Associated Regions

Direct Network

InWeb Network

Indirect Network
Statistical Analysis: How Surprising are the Networks?

- Networks uncovered may be spurious
  - Selection of 30 random SNPs may yield up to 15 connections

- **Approach**
  - Define empirical null distribution for connectivity
  - Generate 1000 random networks
  - Compare disease network to null distribution
Application of Method

• Mendelian disease
  – Genes harboring causal mutations for the same Mendelian disease often physically interact
  – Fanconi Anemia

• Complex disease
  – We hypothesize that genetic variants affect a common mechanism underlying complex disease
  
  ➢ Does the connectivity seen in Mendelian disease hold true for complex disease?
Benchmark: Fanconi Anemia

• Autosomal recessive disorder characterized by physical abnormalities, bone marrow failure, and increased risk of malignancy
• Cause by mutation in genes encoding the DNA repair machinery
• At least 13 genes have been discovered and many are known to participate in protein-protein interactions

➤ Can our method detect an enrichment for connectivity among these genes?
Benchmark: Fanconi Anemia

23 Direct Connections
Enrichment: $p < 0.001$
What is Different When Approaching Complex Disease?

• In Mendelian disease, the method works extremely well
  – Few pathways, most genes known

• In complex disease:
  – Many pathways, few genes known
  – Causal genes are uncertain
  – Current genetic associations may be capturing only subsets of the underlying mechanism
Complex Disease: Rheumatoid Arthritis

• Chronic inflammatory disease of the joints
• 29 Replicated Loci (Raychaudhuri et al. 2010)
• Speculation about underlying pathways makes it a promising disease to study in this context
  – Literature mining has already revealed strong connectivity (Soumya Raychaudhuri)

➤ Can we uncover physical connections between proteins encoded in associated regions?
Rheumatoid Arthritis
Rheumatoid Arthritis

Connectivity Score

Probability

1000 Randomized Networks
Hypothesis Generation for Candidate Genes

rs3761847

Genes Within Locus

Connectivity Score - log(p)

0
0.5
1
1.5
2
2.5
3

TRAF1  PSMD5  GSN  RAB14  FBXW2  C5
Conclusions

• The use of protein-protein interaction data facilitates mechanistic hypotheses for complex disease
• Defining the process helps generate hypotheses about candidate genes
• Common variants may influence phenotype at a process level, rather than gene level
• Extensions:
  – Tissue specificity: which cell types is the disease network specific to?
  – New associations: Networks can highlight previously unidentified risk genes
Available Tools

• Paper is near submission and is available upon request

• Perl and R Scripts available upon request
  (rossin@broadinstitute.org)
  – Network Building
  – Randomization Analysis
  – Prioritization of Candidate Genes
  – Tissue Specificity (Benita et al. Blood 2010)
  – Identifying New Risk Candidates

• InWeb Protein-Protein Interaction Data
  (klage.kasper@mgh.harvard.edu)
Acknowledgements

• Mark Daly
• Chris Cotsapas
• Soumya Raychaudhuri
• David Altshuler
• Ayellet Segre

• Kasper Lage
• Diana Tartar

• Robert Plenge
• Eli Stahl

• Ramnik Xavier
• Yair Benita