Implicating Sequence Variants in Human Disease
NHGRI Workshop
Sept 12-13, 2012

Integrated Approach Working Group

Don Conrad

Mark Daly

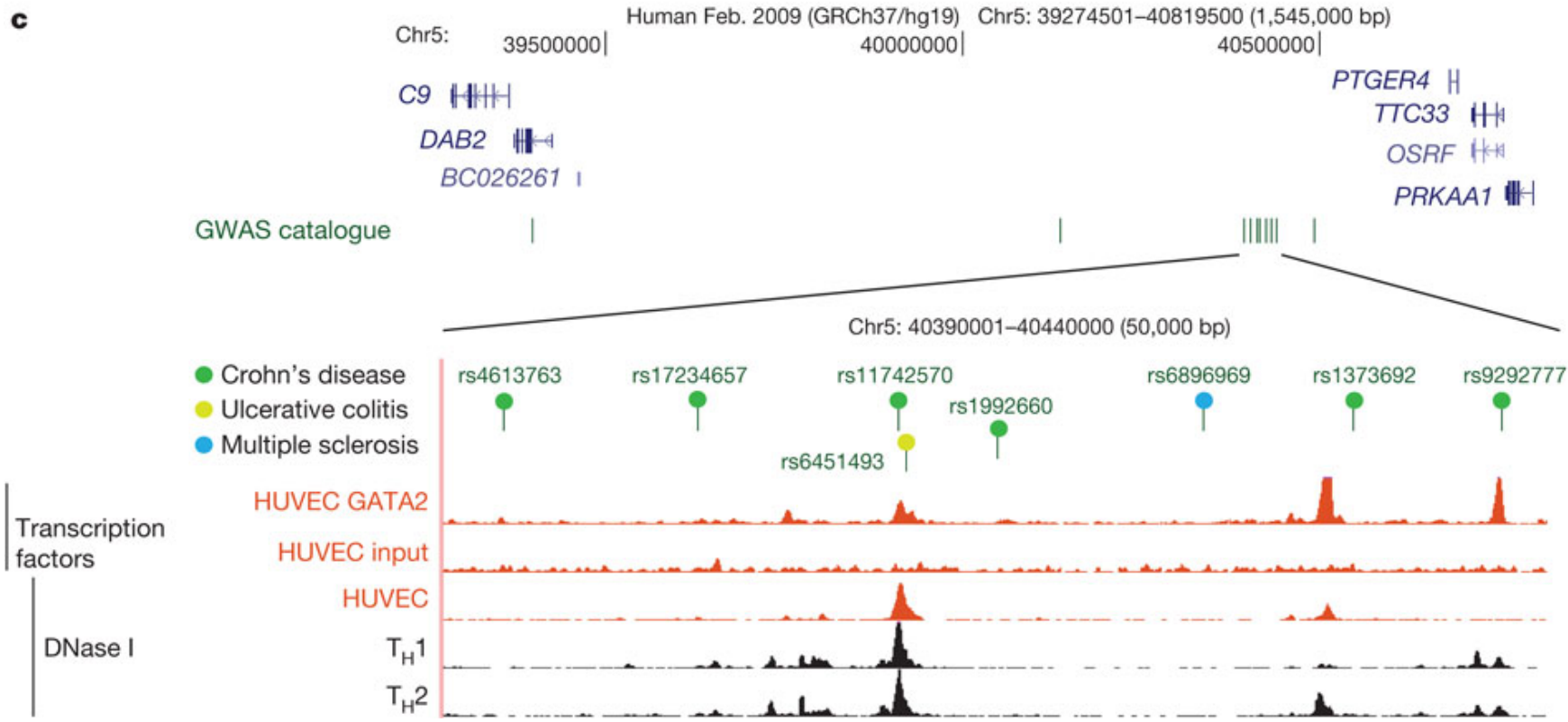
Mark Gerstein

Chris Gunter

Integration of data types

- Existing lists of causal genes, known associations
- Unbiased genome screening data sets (e.g. PPI, expression, RNAi) and genome annotation data (e.g. TF binding sites, epigenetic marks)
- Semi-structured Literature
- Model organism data

The challenge

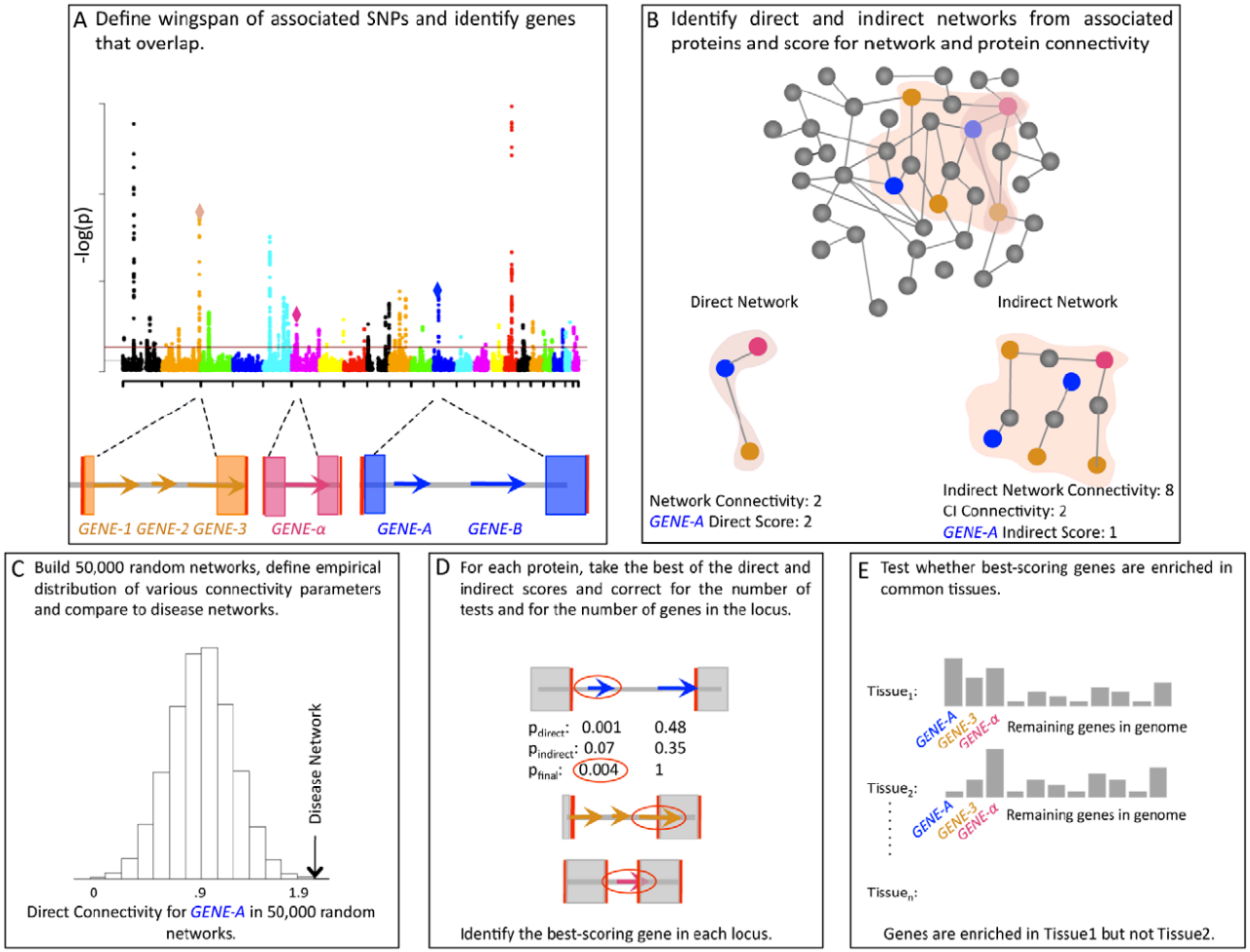


Use Case I

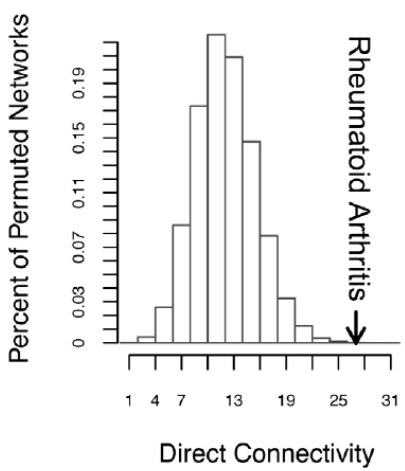
- Using networks to implicate a gene in a complex and highly multigenic scenario
- Vidal, Barabasi, Gerstein, Hurles, many others

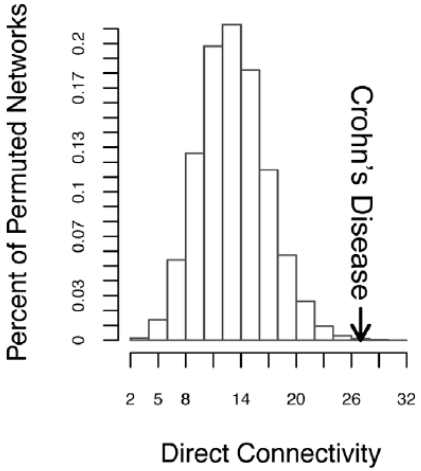
Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology

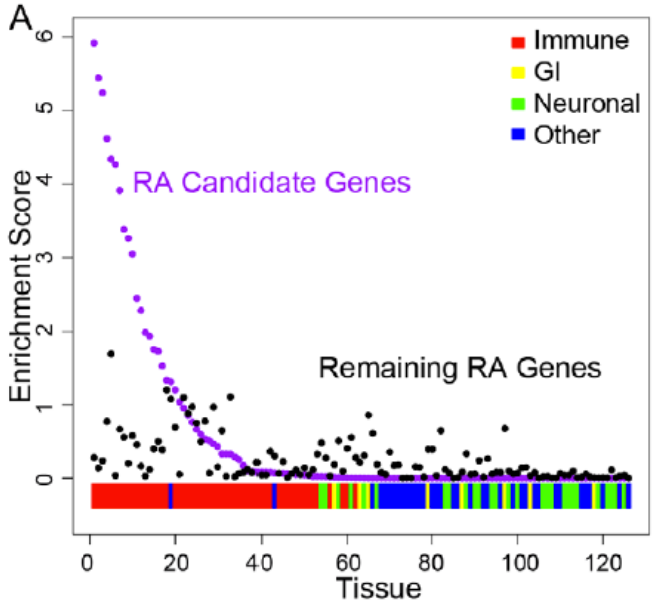
Elizabeth J. Rossin[1,2,3,4,5], Kasper Lage[2,3,6,7], Soumya Raychaudhuri[1,2,8], Ramnik J. Xavier[1,2,3], Diana Tatar[6], Yair Benita[1], International Inflammatory Bowel Disease Genetics Constortium[¶], Chris Cotsapas[1,2☯], Mark J. Daly[1,2,3,4,5☯*]

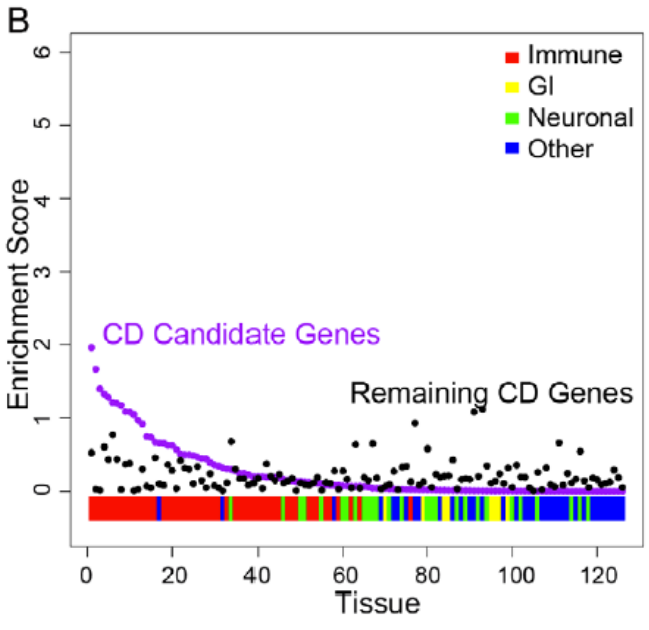


DAPPLE results









DAPPLE predicts new associations

- 293 genes in CD network and expressed in relevant tissues
- 10/293 predicted CD genes later confirmed by GWAS meta analysis ($p<0.001$)

Use Case II

- Establishing causality for non-coding DNA variants

Non-coding annotation

- Rapidly growing large body of NC annotation
- Disruption of these annotations are interpretable from a sequence perspective
- Annotations allow integration of information on cell types and tissues (epigenome roadmap, ENCODE)

Gaffney et al. Genome Biology 2012, 13:R7
http://genomebiology.com/2012/13/1/R7

Genome Biology

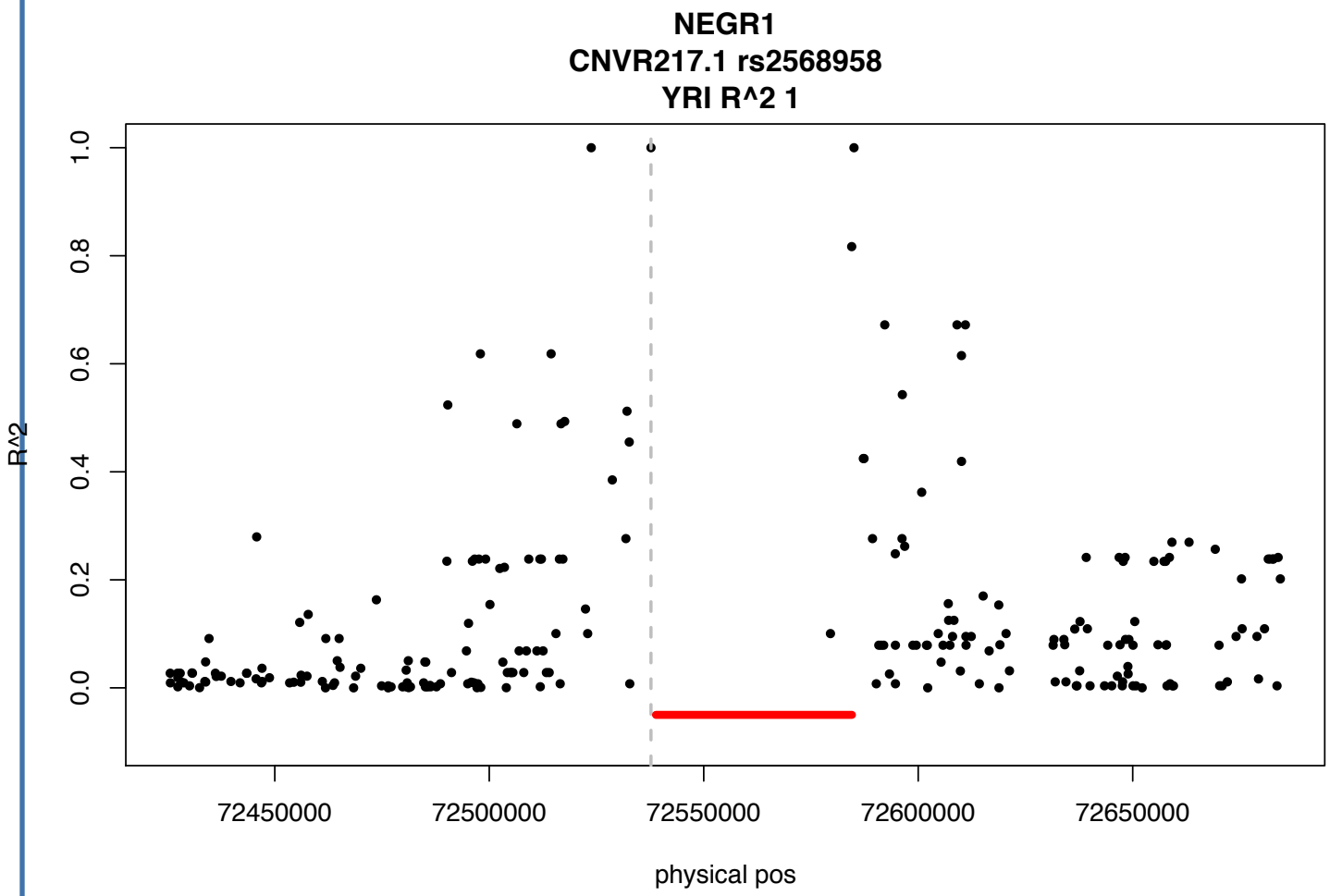
RESEARCH Open Access

Dissecting the regulatory architecture of gene expression QTLs

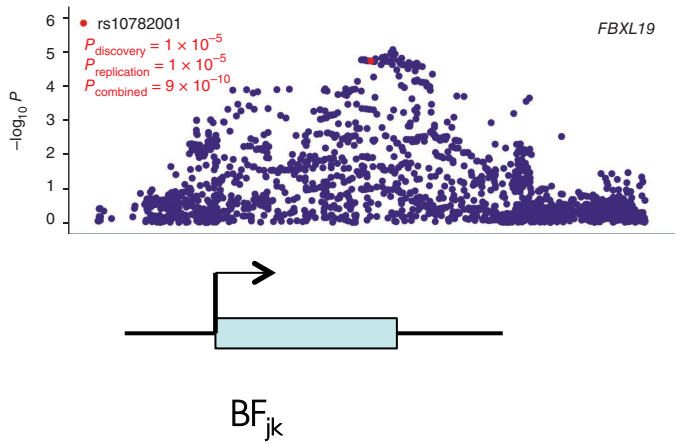
Daniel J Gaffney[1,2,5*], Jean-Baptiste Veyrieras[1], Jacob F Degner[1], Roger Pique-Regi[1], Athma A Pai[1], Gregory E Crawford[3], Matthew Stephens[1,4], Yoav Gilad[1] and Jonathan K Pritchard[1,2]

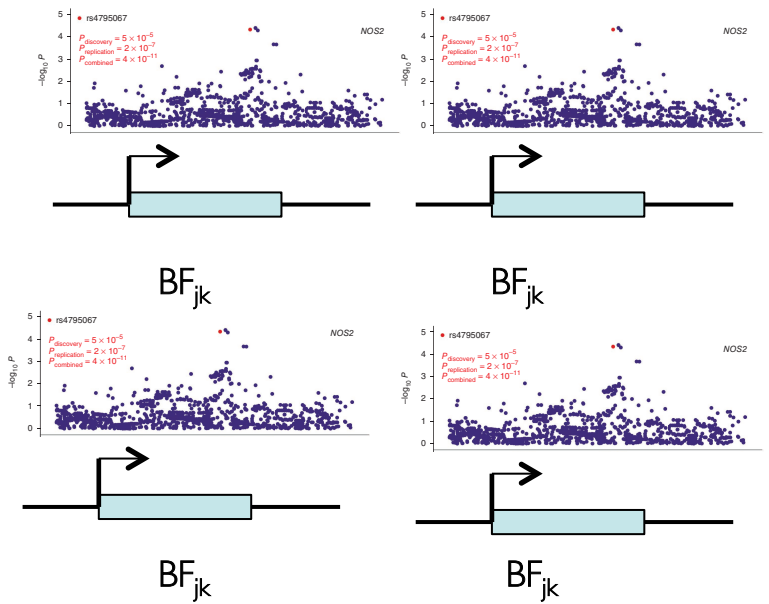
- Cis-eQTL study: whole-genome LCL expression and 13M SNPS from 210 samples (1000 genomes project)
- Key example of how non-coding annotation can be used to form priors on "causality".

The LD "Problem"



Hierarchical Model: Background



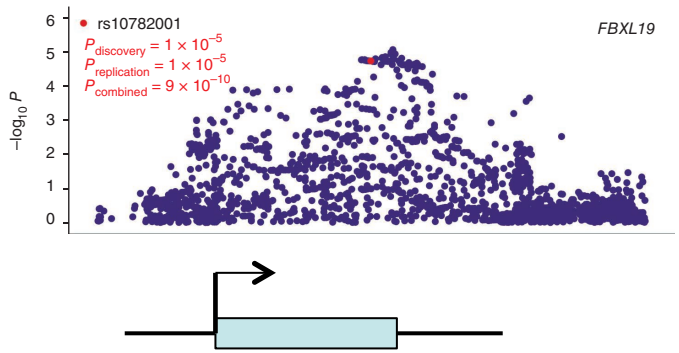


Hierarchical Model: Background

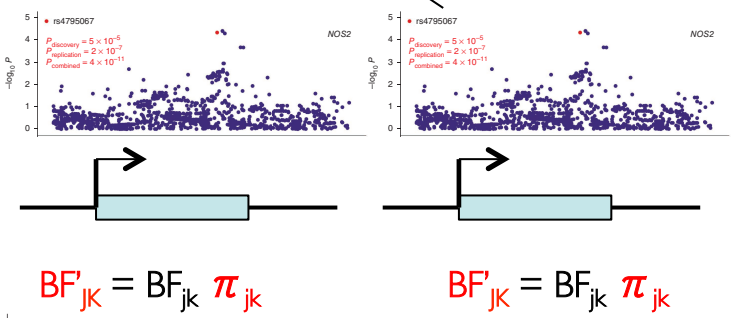
Level I

λ_1
λ_2
λ_3
.
.
.
λ_L

Level II

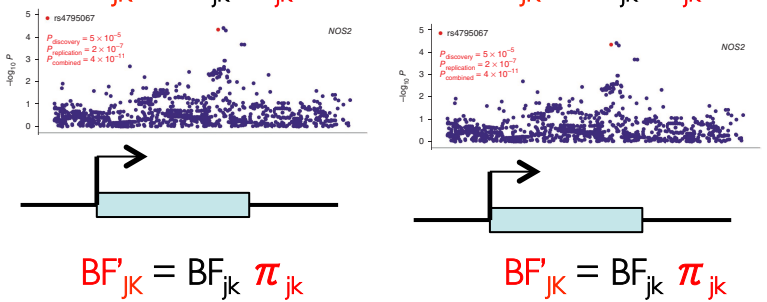


$BF'_{JK} = BF_{jk}\ \pi_{jk}$



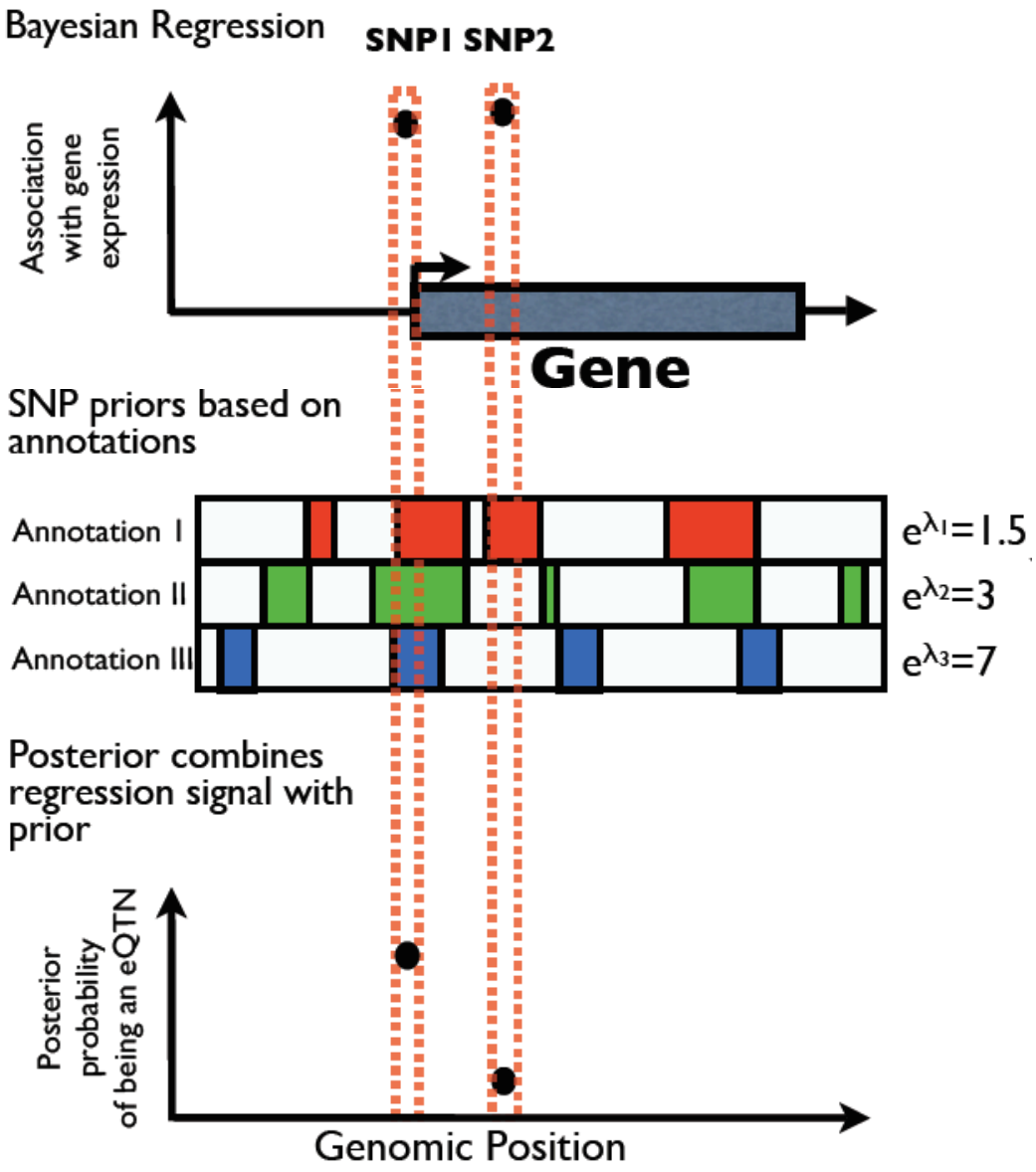
$BF'_{JK} = BF_{jk}\ \pi_{jk}$

$BF'_{JK} = BF_{jk}\ \pi_{jk}$



$BF'_{JK} = BF_{jk}\ \pi_{jk}$

$BF'_{JK} = BF_{jk}\ \pi_{jk}$



Slide Credit: Daniel Gaffney

SNP-based results

Features predictive of LCL eQTLs (>50 tested)

- Proximity to gene
- Histone marks (n = 5 types)
- DNase I hypersensitivity sites
- Core promoter motifs (n = 2)
- Transcription factor binding sites (n = 4)

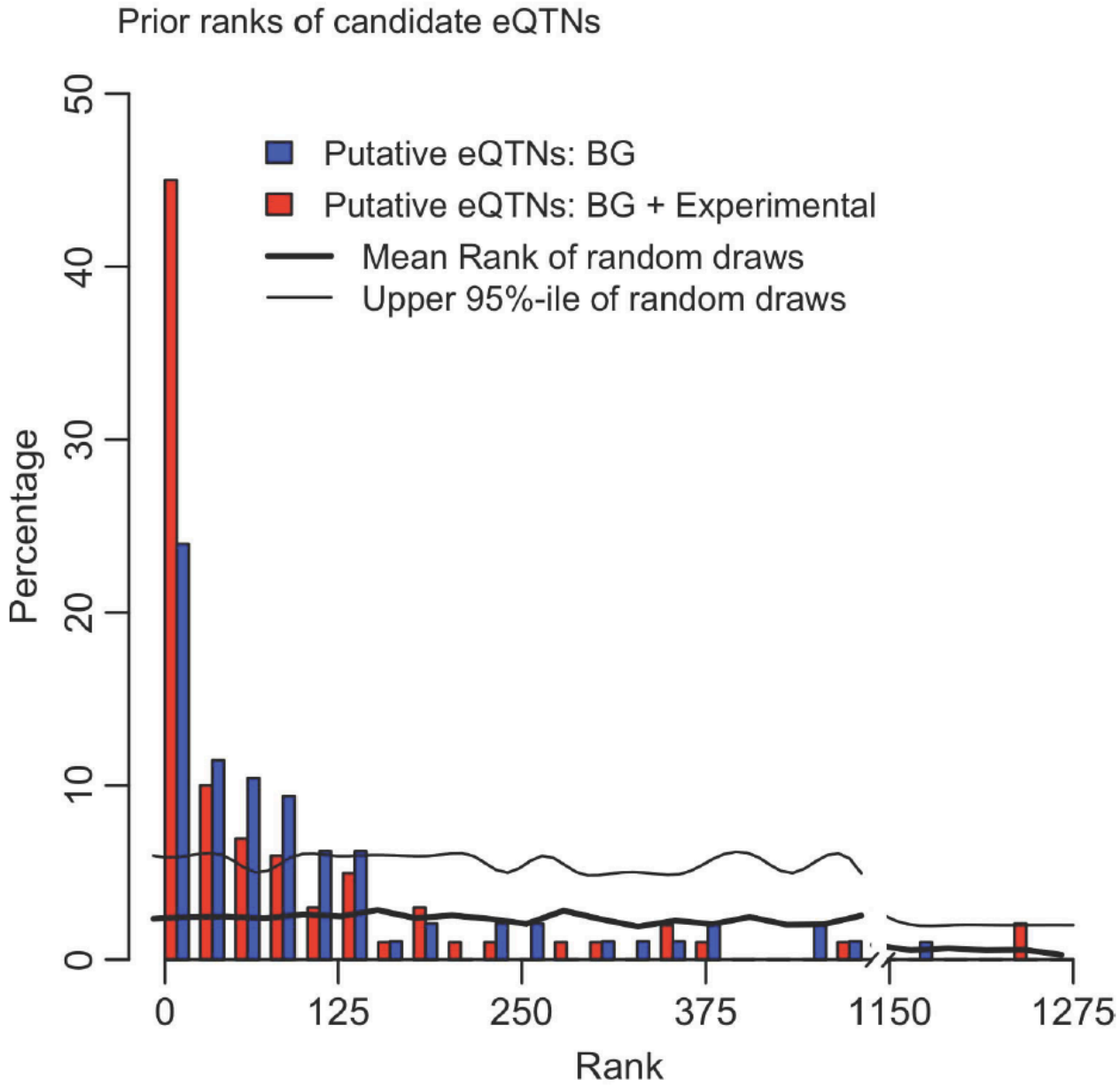


Figure 7 Prior rankings of SNPs for 100 genes where a single SNP is a clear best candidate for being the 'true' eQTN using the prior probability from the hierarchical model. The histogram shows the percentage of genes for which the putative causal site is ranked by the prior among the top 1 to 15 SNPs, 15 to 30 SNPs, and so on. Typically, the candidate region for each gene contains approximately 1,200 SNPs. The two prior models correspond to the distance model only (blue) and the distance model plus experimental annotations (red). The 100 genes analyzed here were excluded from all other analyses. BG, background.

Use Case III

- Identifying causal variant for a presumed Mendelian disease, high locus heterogeneity

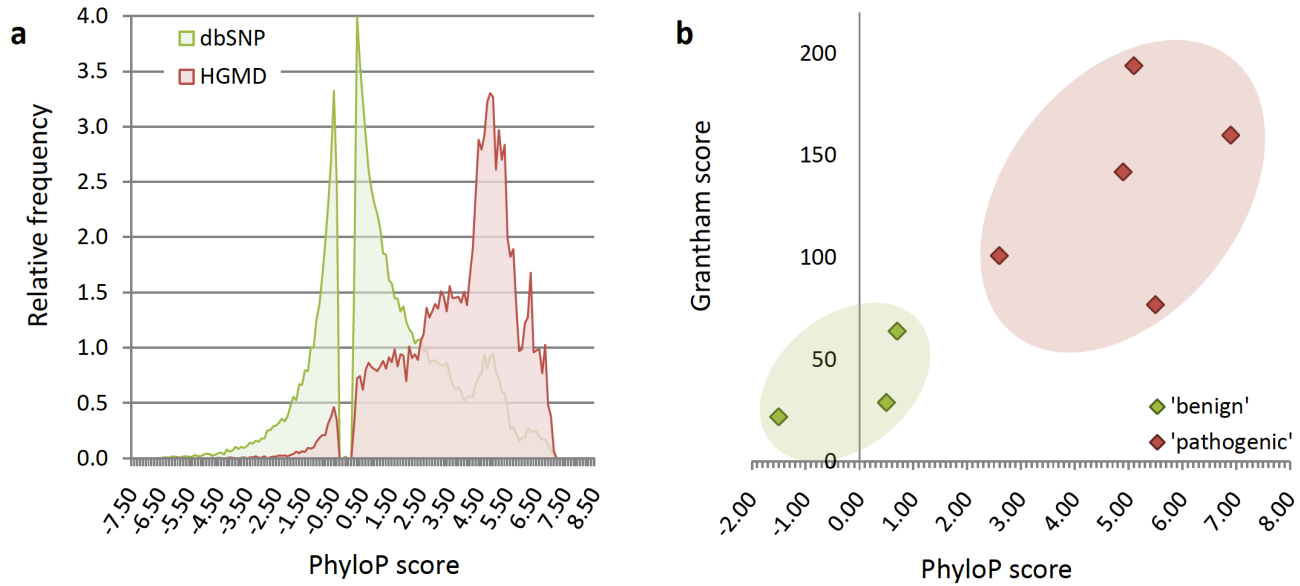
A *de novo* paradigm for mental retardation

Lisenka E L M Vissers[1,2], Joep de Ligt[1,2], Christian Gilissen[1], Irene Janssen[1], Marloes Steehouwer[1], Petra de Vries[1], Bart van Lier[1], Peer Arts[1], Nienke Wieskamp[1], Marisol del Rosario[1], Bregje W M van Bon[1], Alexander Hoischen[1], Bert B A de Vries[1], Han G Brunner[1,3] & Joris A Veltman[1,3]

NATURE GENETICS ADVANCE ONLINE PUBLICATION

- Identified 9 *de novo* mutations in 10 samples with idiopathic MR
- "6 of these … are likely to be pathogenic based on gene function, evolutionary conservation, and mutation impact".

Supplementary Figure 5: Distribution of PhyloP and Grantham scores for dbSNP, HGMD and the *de novo* mutations identified in this study



Take-home points

- Rigorous model-based approaches are essential for integrating secondary data
- Networks are key data structure for integration
- Wealth of non-coding annotation available for integration
- “N=1” may be tractable with data integration with natural variation

Discussion Questions

- What are the data types to be used?
- What are the key untapped data resources today?
- How do we best collect data with the *a priori* intention of integration?
  - Especially with respect to databases and literature
- In what research areas will the integrated approach be most important for progress?
- How does one establish the statistical validity of integrated approaches? Are there situations in which experimental validation of an approach is optional?
- What are the considerations for validity unique to each use case 1, 2 and 3?
- What are the ways in which the integrated approach can be abused?
- What are the optimal ways to leverage ENCODE data for establishing causality?
- How do we store and share the results of integrated analyses?