Identification of Expression Quantitative Trait Loci Among Breast Cancer-Associated SNPs

Fred Boehm
fjboehm@wisc.edu

Outline

- Motivation
- Methods
- Results
- Conclusions
- Next steps

Question

- Are genotypic differences in breast cancer-associated SNPs associated with differences in expression levels?
  - Might a breast cancer-associated SNP be an eQTL?
  - Perhaps a SNP's genotypic differences are associated with expression level differences for a set of (functionally) related genes

Approach

- Acquire genotypic data & expression data from breast tissue samples
- Develop a pipeline for analysis & characterization of SNP-expression probe associations
- Generate a prioritized list of breast cancer-associated SNPs for subsequent biological & toxicological studies
    - To understand gene-environment interactions

Tissue samples

- Reduction mammoplasty human mammary epithelial cell (HMEC) samples
- 62 female subjects without history of breast cancer
  - 30 from Berkeley, CA
  - 32 from Madison, WI
- Ages & other demographic info unavailable
  - Most are thought to be in their 20s in age
- Clinical data unavailable

SNP genotype data

- Illumina Omni chip data for all 62 subjects
    - ~1 million SNP loci with called genotypes
- Two subjects were genotyped more than once due to low genotype call rates

Affymetrix GeneChip expression data

- Human ST 1.0 chip
- ~25,000 normalized expression levels for all 62 subjects
- Raw data processed with Robust Multi-array Average (RMA) methods

Ghoussaini et al. (2012)

- 72 SNPs
  - Genotyped (ie, not imputed) & breast cancer-associated ($p < 0.0001$) in one or both of two UK breast cancer GWAS
- Meta-analysis of ~ 70,000 cases & 68,000 controls
  - 41 case-control studies & 9 breast cancer GWAS
- Identified 3 SNPs with very small p-values for association with breast cancer

Plan

- Characterize the 72 Ghoussaini SNPs using our newly developed analysis pipeline
  - Calculate a statistic to identify associations
  - Generate heatmaps for the associated traits
  - Perform gene set enrichment analysis for each SNP's associated traits
  - Generate neighborhood plots for our top SNPs

Methods

- Which statistical methods to use to identify associations?
  - Between genotypic variations at a given SNP and expression levels for a single probe
- Standard approach in the scientific literature is to use ANOVA-based methods
- Due to our small sample size (n=62) we don't want to use ANOVA
  - ANOVA assumes constant variance among genotype classes

Strategy #1: `pmax' statistic

- Consider a single expression probe and a single SNP (from among Ghoussaini's 72 SNPs)
- Regress

  expression ~ PC1 + PC2 + siteIndicator
- Save residuals for subsequent analysis

Strategy #1: `pmax' statistic

- Use above residuals to calculate t-test-based p-values for all 3 pairwise comparisons
  - Genotype = 0 v. Genotype = 1
  - Genotype = 0 v. Genotype = 2
  - Genotype = 1 v. Genotype = 2
- without equal variance assumption
- Assign pmax to be the maximum of the three p-values

Strategy #1: `pmax' statistic

- For SNPs with only two genotypes:
  - i.e., when there are no minor allele homozygotes in our sample
  - Assign pmax to be the p-value for comparison of the heterozygotes and major allele homozygotes
- Declare a SNP-gene association when pmax is below an arbitrary threshold
  - i.e., when all 3 p-values are small

Problem with pmax-based strategy

- Preferentially selects SNPs with only two genotypes in our sample
- We need to consider other statistics

Strategy #2: `phom' statistic

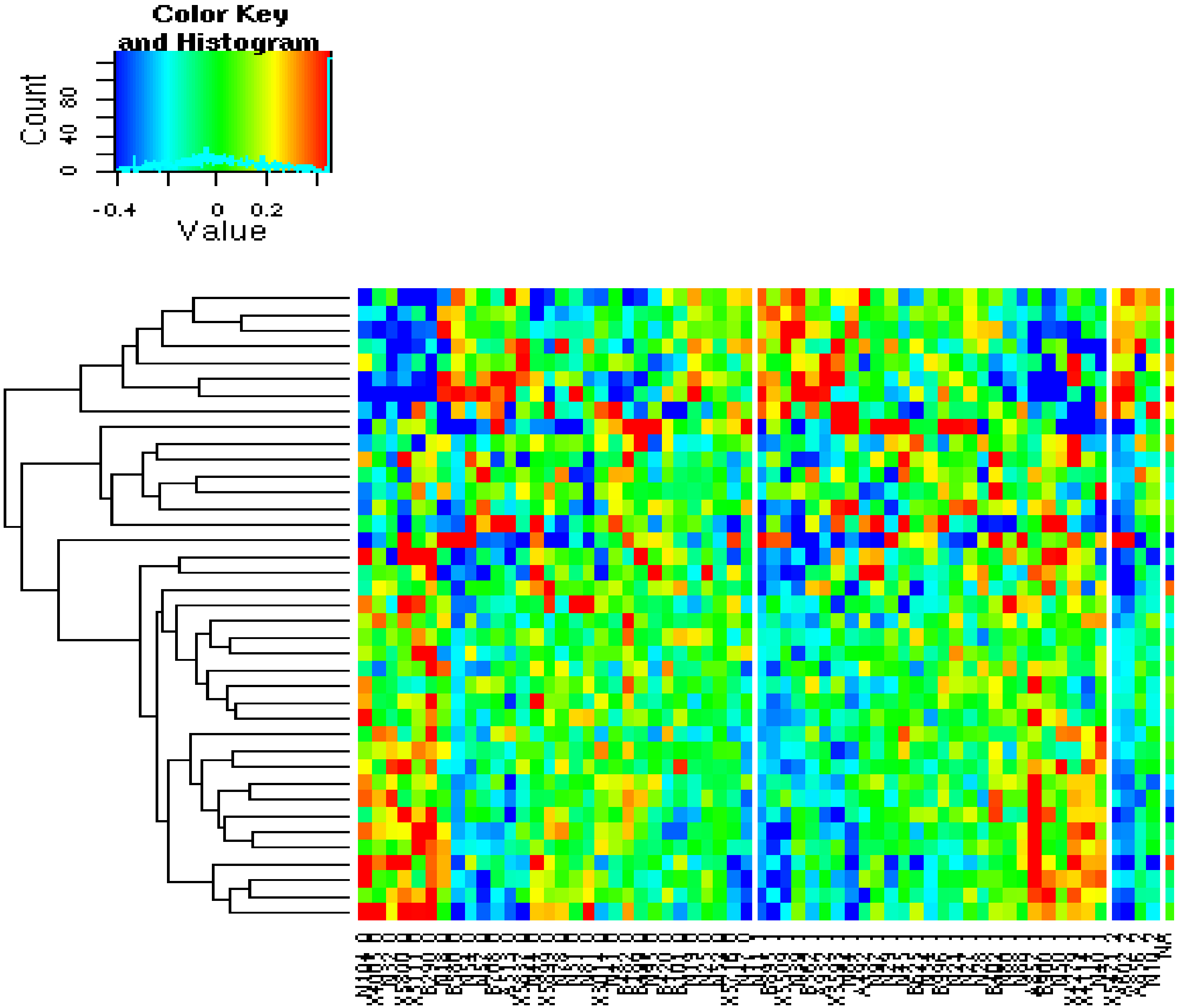
- Calculate the p-value for a t-test between the two homozygote classes
- Fewer genetic model assumptions
    - Since we don't require anything of the heterozygotes
- Rank SNPs by number of traits with phom < 10^-5
- Require that minor homozygote class have at least 4 subjects

Results

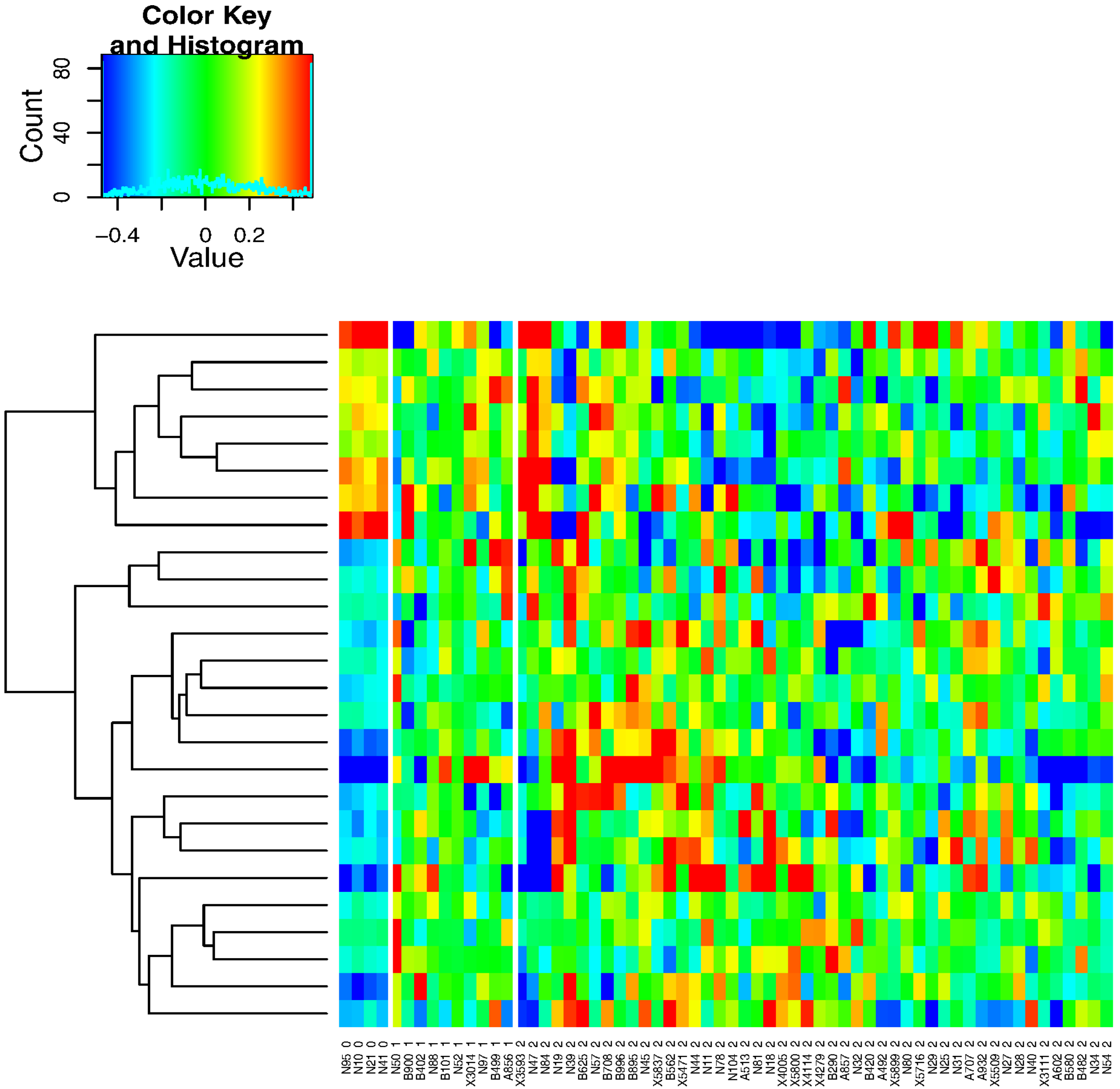
Lead SNP	Chrom	MAF	Proximal Genes	Proxy	R^2	n.AA	n.AB	n.BB	phome-5
rs11241138	5	0.41	NREP; 7SK	rs7718087	0.96	30	27	4	39
rs3101649	15	0.19	OCA2	rs3101649	1	4	10	48	26
rs2284424	12	0.22	GRIN2B	rs2284425	1	4	25	32	11

- NREP: neuronal regeneration related protein homolog
- 7SK: a small nuclear RNA
- OCA2: oculocutaneous albinism II
- GRIN2B: glutamate receptor, ionotropic, N-methyl D-aspartate 2B

Heatmap for rs7718087



Heatmap for rs3101649



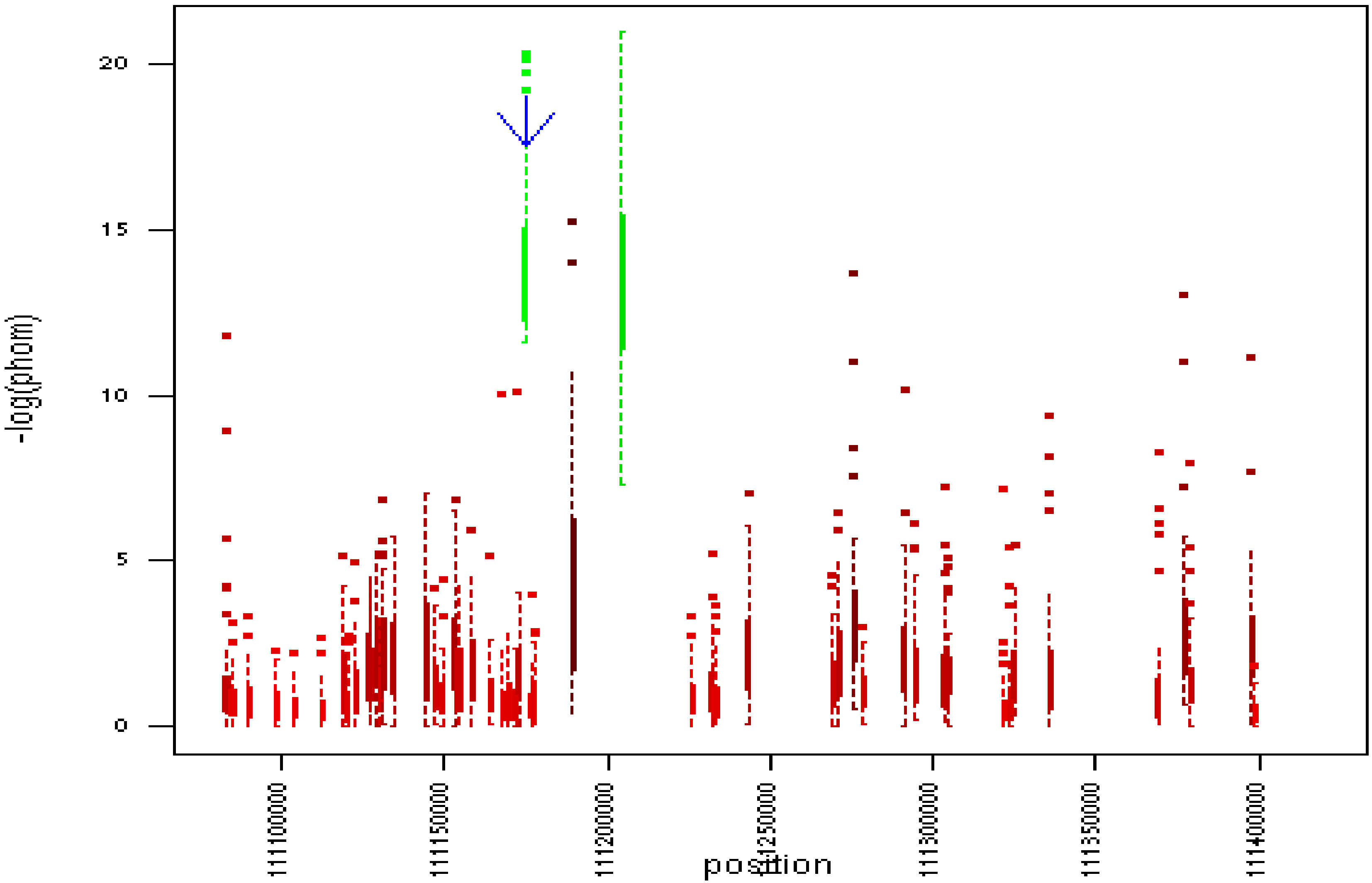
rs3101649 GSEA results

SetID	SetName	Count	NGenes	SetMean	ZScore	LeadGenes
GO:0032206	positive regulation of telomere maintenance	1	5	0.2	15.032	TNKS
GO:0001875	lipopolysaccharide receptor activity	1	5	0.2	15.032	LY96
GO:0051016	barbed-end actin filament capping	1	5	0.2	15.032	CAPG
GO:0004499	flavin-containing monooxygenase activity	1	5	0.2	15.032	FMO5
GO:0004791	thioredoxin-disulfide reductase activity	1	5	0.2	15.032	TXNRD3
GO:0031013	troponin I binding	1	5	0.2	15.032	RCAN3

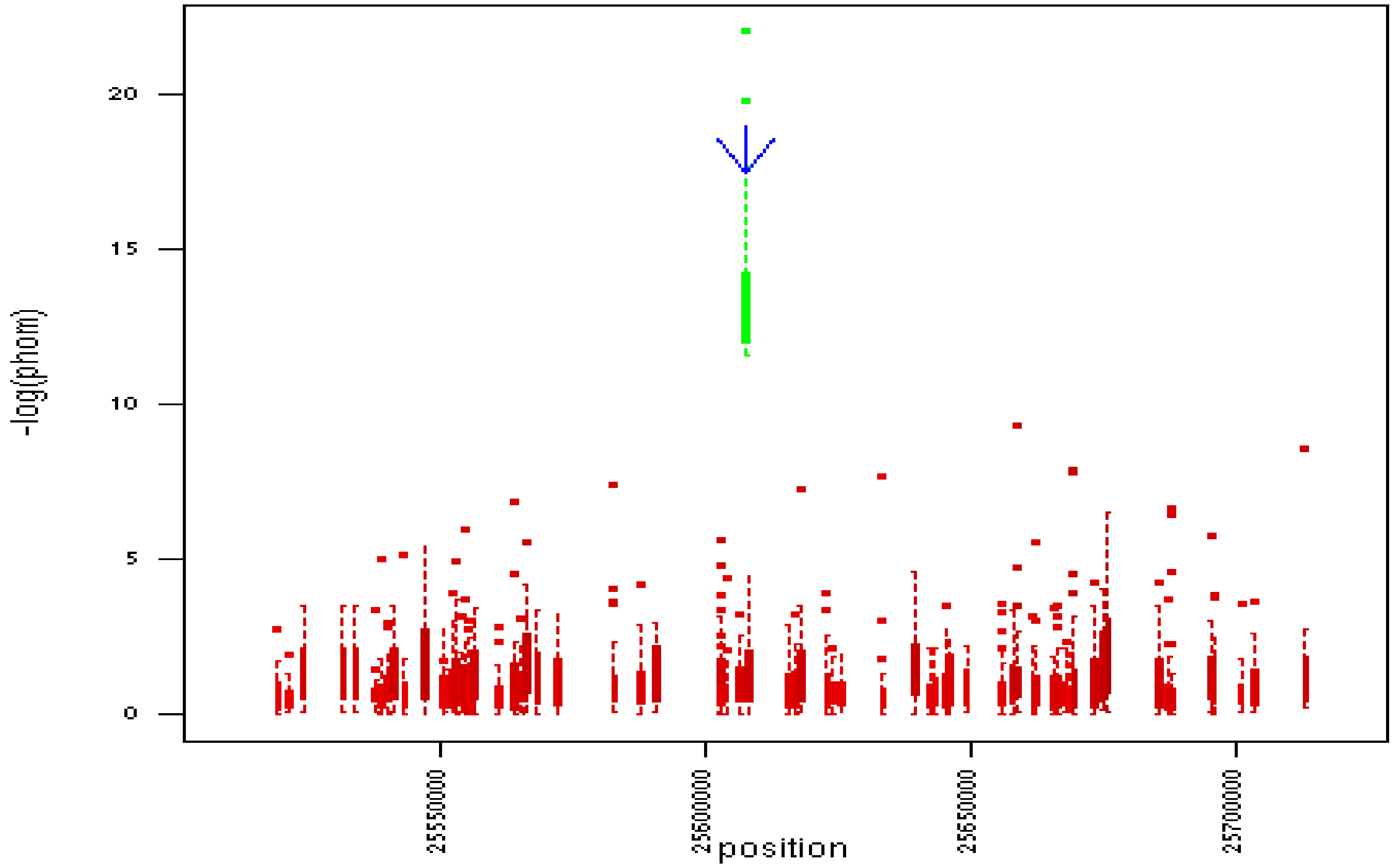
rs7718087 GSEA results

SetID	SetName	Count	NGenes	SetMean	ZScore	LeadGenes
GO:0016199	axon midline choice point recognition	1	5	0.2	10.412	ROBO3
GO:0031702	type 1 angiotensin receptor binding	1	5	0.2	10.412	BDKRB2
GO:0021877	forebrain neuron fate commitment	1	6	0.167	9.488	NKX2-1
GO:0022601	menstrual cycle phase	1	6	0.167	9.488	NKX2-1
GO:0042538	hyperosmotic salinity response	1	6	0.167	9.488	NKX2-1
GO:0006386	termination of RNA polymerase III transcription	2	12	0.167	13.421	POLR2H LZTS1
GO:0006385	transcription elongation from RNA polymerase III promoter	2	12	0.167	13.421	POLR2H LZTS1
GO:0016198	axon choice point recognition	1	6	0.167	9.488	ROBO3
GO:0044224	juxtaparanode region of axon	1	6	0.167	9.488	KCNAB2

Neighborhood plot for rs7718087 (39 traits)



Neighborhood plot for rs3101649 (26 traits)



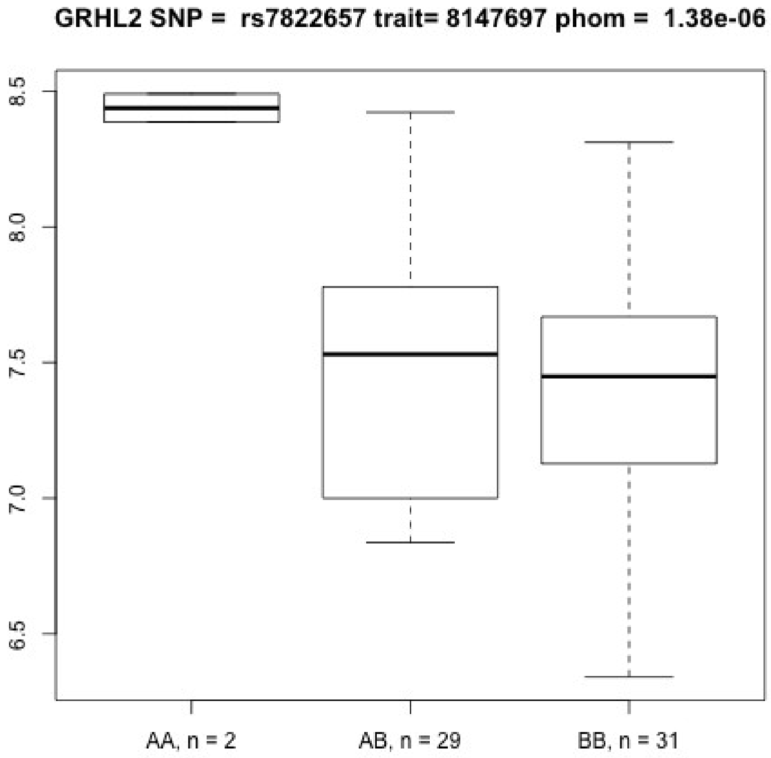
cis-eQTL

- For each of 494 proxy SNPs
  - Calculate phom for all genes within 300kb
- Rank all SNP-trait pairs by phom values

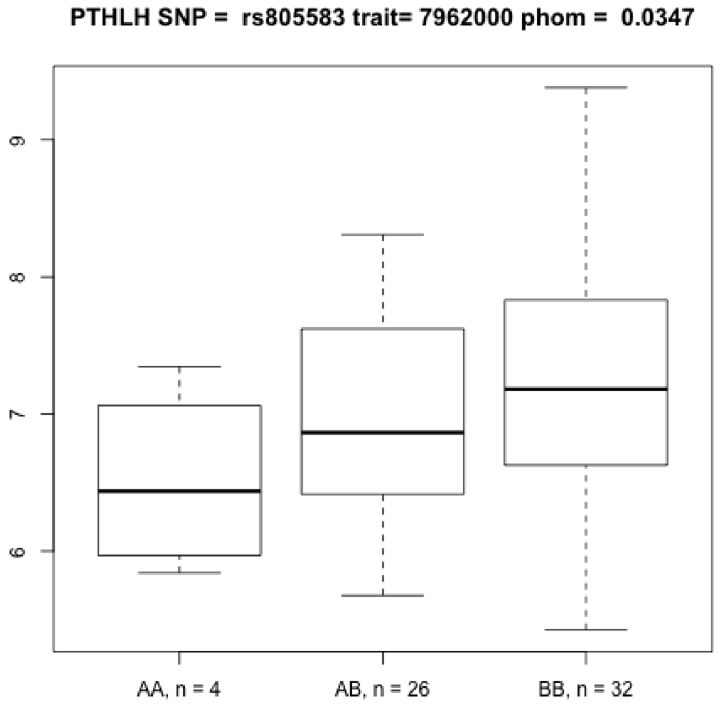
Cis-eQTL

GeneSymbol	Chromosome	Start	Stop	chromosome	PHOM	rs.id
OR3A3	chr17	3270612	3271577	chr17	3.11E-08	rs16953025
SPINT2	chr19	43447007	43474948	chr19	1.53E-07	rs2960337
GRHL2	chr8	102574162	102750995	chr8	1.38E-06	rs7822657
GRHL2	chr8	102574162	102750995	chr8	2.53E-06	rs1131863
GRHL2	chr8	102574162	102750995	chr8	2.53E-06	rs1131862
SNORA13	chr5	111525081	111525213	chr5	5.47E-05	rs980888

Boxplot of a top cis-eQTL



Box plot of a PTHLH cis-eQTL



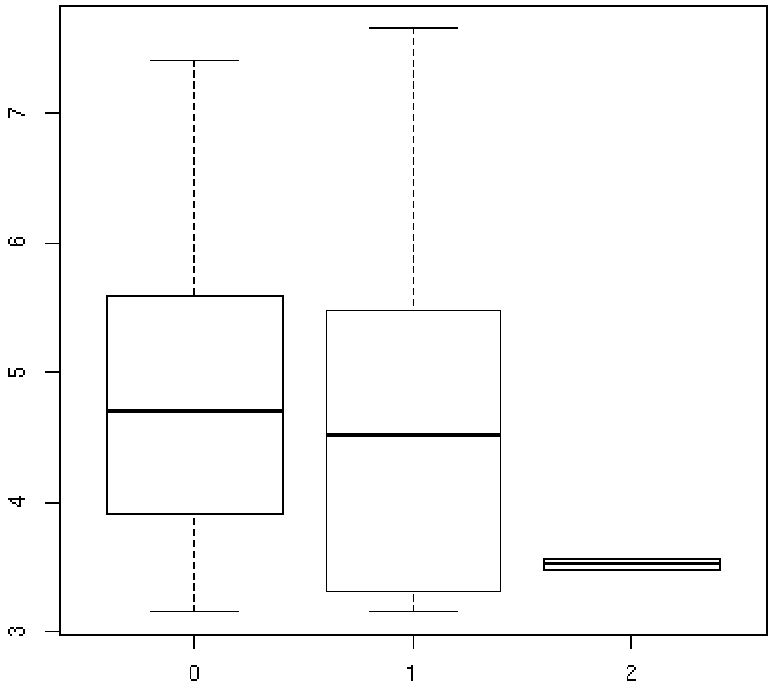
Mcs SNPs

- Two Mcs SNPs from Samuelson (2007, PNAS)
  - One is on our chip
  - A good proxy (R^2 = 1) for the other is on our chip
- 7 Genes involved in T cell biology

Mcs SNPs

GeneSymbol	Chromosome	Start	Stop	rs6476643	rs2182318
CD3E	chr11	117680662	117692100	0.83	0.596
CD3G	chr11	117720311	117729979	0.22	0.751
KLRD1	chr12	10269932	10360042	0.35	0.817
CD69	chr12	9796349	9804764	0.87	0.007
KLRK1	chr12	10416219	10451632	0.65	0.7
KLRK1	chr12	10453965	10454012	0.62	0.268
CD28	chr2	204279443	204310801	0.32	0.522

Top Mcs SNP-gene eQTL



Conclusions

- Several breast cancer-associated SNPs exhibit SNP-gene associations
- Low signal to noise ratio in our small sample
    - Similar studies often have n > 1000
    - GSEA results don't suggest extensive involvement of a single known biological pathway

Reporting what we've done

- Detailed analysis of our data on 62 subjects
  - 72 Ghoussaini SNPs and their proxies
  - Genome-wide expression data
  - Phom statistic
  - Analysis pipeline
- Limited by small minor allele homozygote class size
- Data resource to share

Future directions

- If sample size is the issue, how many subjects might we want to better prioritize our SNPs?
  - Address with simulations?
  - Other approaches?

References

- C.C. Laurie, et al. (2010). Genetic Epidemiology
- M. Ghoussaini, et al. (2012). Nature Genetics
- M. Newton (2009). allez software for R.
- N. Patterson, et al. (2006). PLOS Genetics
- S.M. Gogarten et al. GWASTools: Tools for Genome Wide Association Studies. R package version 1.2.1.