Summer Institute in Statistical Genetics 2019

Integrative Genomics

10. Genetics of Gene Expression: eSNPs

ggibson.gt@gmail.com
http://www.gibsongroup.biology.gatech.edu

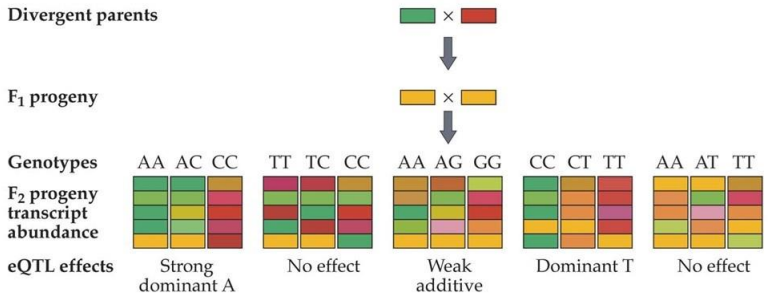
Georgia Tech

Center for Integrative Genomics

Expression QTL analysis

- The architecture of transcription maps genotype onto phenotype
- Expression QTL (eQTL) are QTL that modulate transcript abundance in pedigrees or crosses
- It is estimated that at least 10% of transcripts differ in abundance between any two strains of most organisms; as much as 50% across a species
- Estimates of heritability of transcription also suggest that it is remarkably high, with transcription often showing a higher genetic component than visible traits

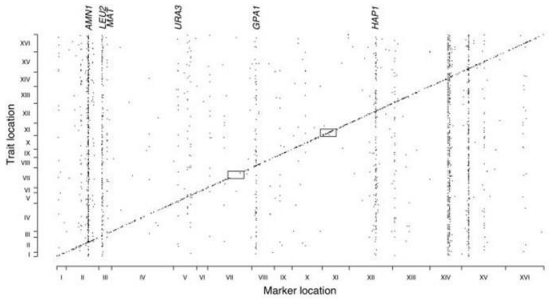
Principle of eQTL analysis



cis and *trans* eQTL

- Liver samples from 111 F2 mice from an obesity cross
- 15% of 23,500 genes with at least one eQTL explaining ~ 25% of the variance
- Tendency for strong eQTL to be in *cis* to the actual gene
- eQTL clustered in 7 hotspots (each 0.2% of the genome but >1% of the eQTLs)

Similarly for yeast:
Ronald and Akey,
PLoS ONE (2007) e678



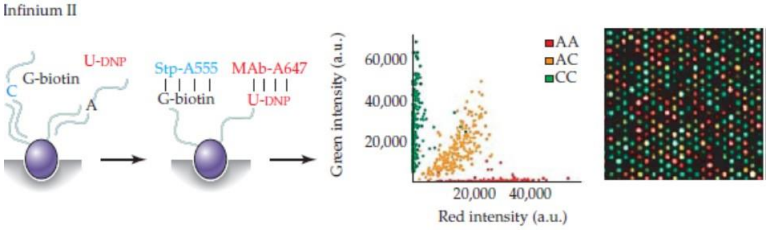
Schadt, Friend et al (2003) *Nature* **422**: 297-302

Limitations of eQTL analysis

- Any QTL experiment is only a comparison of two lines, so does not say anything about the frequency of QTL effects in a population
- If the number of F2 or BC progeny is less than 100, QTL analysis is prone to false positives, particularly for *trans*-hotspots
- Consequently, significance must be evaluated by permutation *being sure to permute the full genotype matrix against the full transcript abundance profile to preserve correlation structure*
- Resolution of QTL analysis is generally low (5 cM ~ 100-1,000genes), although enrichment for cis => most will be in the gene itself
- With pedigree analyses, ensure that one family is not driving the entire experiment

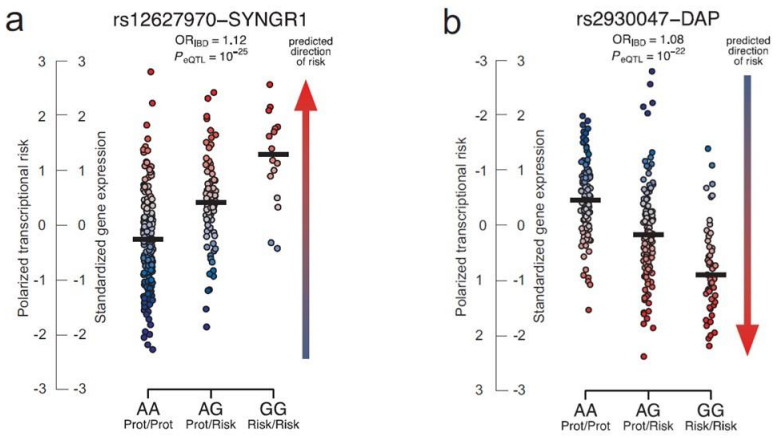
Principle of eSNP analysis

- Whole genome genotyping of >100 unrelated individuals



- Whole transcriptome profiling of the same individuals
- GWAS (Genome-wide association study) for transcription -> precise localization of regulatory SNPs in *cis* and *trans*

A couple of eSNPs

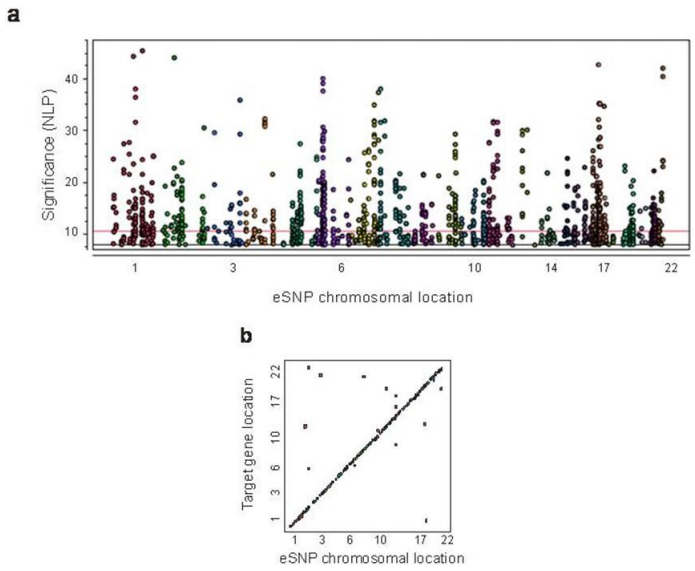


Significance thresholds

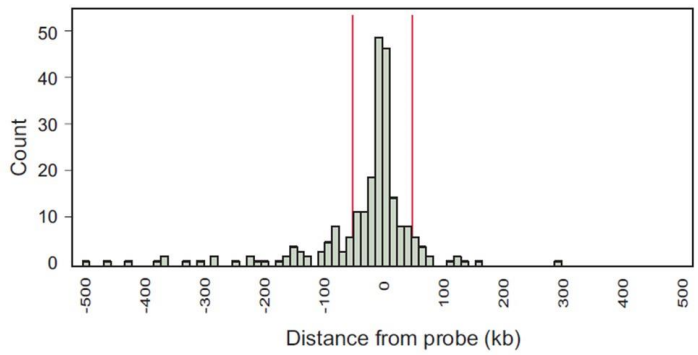
- Bonferroni for *cis*-linkages:
  0.05 / (20,000 genes x 250 SNPs) = 1 x 10^{-8}
- Permutation for *cis*-linkages:
  Random sets of n SNPs from distribtion of 2Mb windows
- Bonferroni for *trans*-linkages:
  0.05 / (20,000 genes x 500,000 SNPs) = 5 x 10^{-12}
- Permutation for *trans*-linkages:
  Randomize complete genotype and transcript matrices

OR adopt FDR criteria, although power not generally an issue
AND consider step-wise regression to adjust for LD

eSNP plots



Location of eSNPs



Effect of Normalization

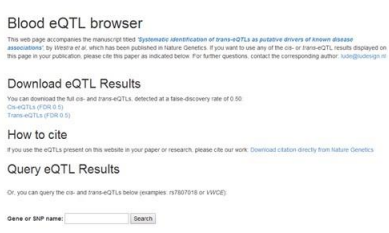
Table 3 eSNP Analyses

	Pearson Correlation				Spearman Rank Correlation	
Normalization	Total (NLP 8)	Cis (NLP 5)	Cis (NLP 8)	Probes (NLP 8)	Cis (NLP 8)	Probes (NLP 8)
RAW	552	1183	411	39	324	36
MEA	1082	2009	743	77	703	71
dr3	627	1362	455	44	407	46
DRM	959	2150	761	87	747	77
IQR	935	1708	603	71	565	73
LMN	484	1281	439	44	394	44
QNM	1211	2288	842	88	791	81
SNM	969	2084	825	86	821	81
PCA	602	1563	585	73	505	74

The Table reports the total number of associations detected between 34,548 Chromosome 6 SNPs and 732 Chromosome 6 Probes, respectively including total (trans and cis) associations at NLP 8; just cis associations at NLP 5 or NLP 8 (defining cis as eSNPs within 250 kb of the probe); the number of independent probes with eSNPs at NLP 8 (all using Pearson correlation with the transcript abundance); and then the cis associations and number of independent probes at NLP 8 using Spearman rank correlation.

Meta-analysis

http://genenetwork.nl/bloodeqtlbrowser/



eQTL meta-analysis on 5,311 individuals replicated in 2,775 more

Found trans-eQTL for 233 SNPs at 103 loci many of which are also disease QTL

Also generates local cis-eSNPs for almost half the genome

Westra et al. (2013) *Nature Genetics* **45**: 1238–1243

GTEx (Genotype-Tissue-Expression Project)

The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans

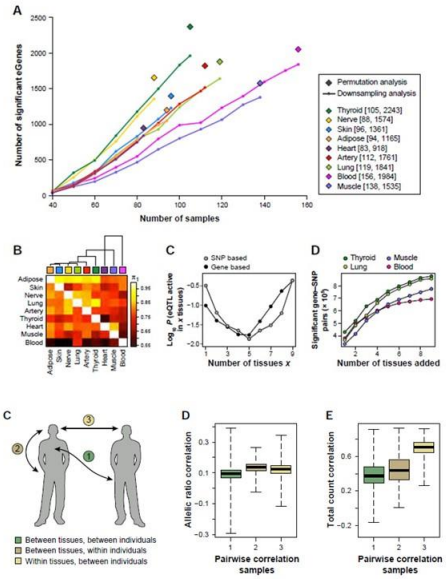
The GTEx Consortium[+]

[+] Author Affiliations

[✉]Corresponding author: Kristin G. Ardlie (kardlie@broadinstitute.org) or Emmanouil T. Dermitzakis (emmanouil.dermitzakis@unige.ch)

ABSTRACT | EDITOR'S SUMMARY

Understanding the functional consequences of genetic variation, and how it affects complex human disease and quantitative traits, remains a critical challenge for biomedicine. We present an analysis of RNA sequencing data from 1641 samples across 43 tissues from 175 individuals, generated as part of the pilot phase of the Genotype-Tissue Expression (GTEx) project. We describe the landscape of gene expression across tissues, catalog thousands of tissue-specific and shared regulatory expression quantitative trait loci (eQTL) variants, describe complex network relationships, and identify signals from genome-wide association studies explained by eQTLs. These findings provide a systematic understanding of the cellular and biological consequences of human genetic variation and of the heterogeneity of such effects among a diverse set of human tissues.

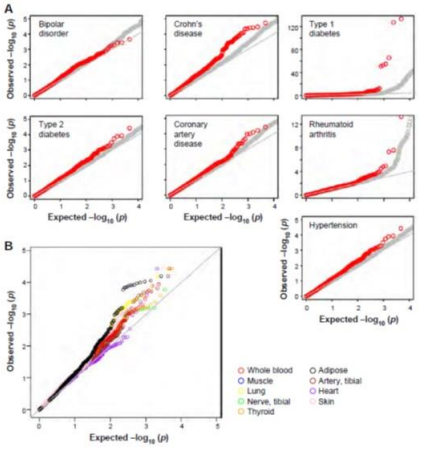


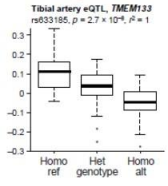
Science (2015) **9(5)**: e1003486

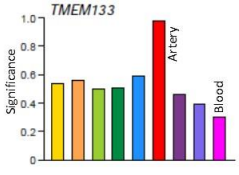
Tissue to trait from GTEx

Whole blood eQTL are enriched for trait associations for CD, T1D, RA

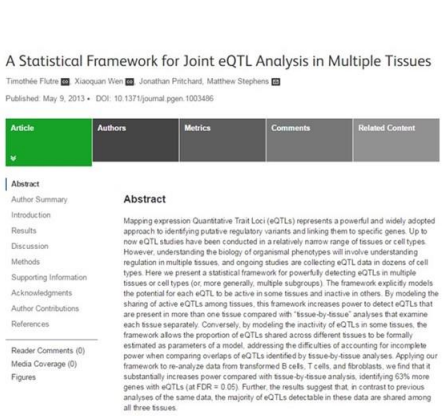
Adipose, Lung, Blood and Artery eQTL are enriched for Hypertension GWAS associations

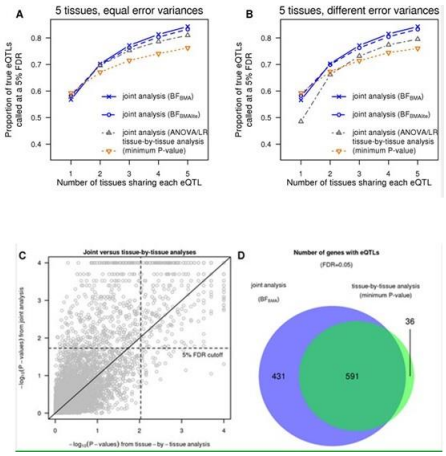






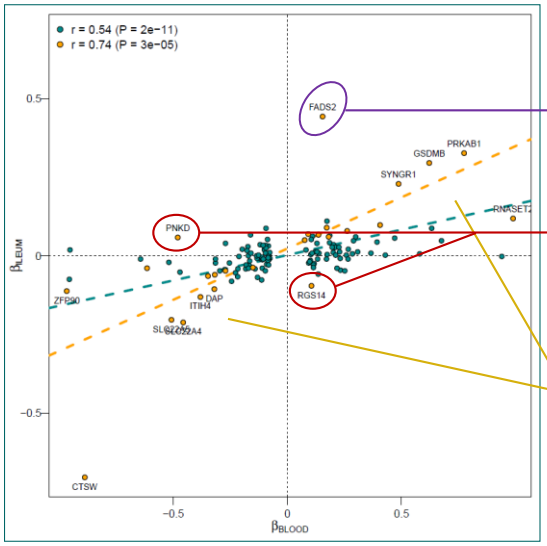
Cross-Tissue Heritability





Flutre *et al* (2013) *PLoS Genetics* **9**: e1003486

Blood and Rectal eQTL

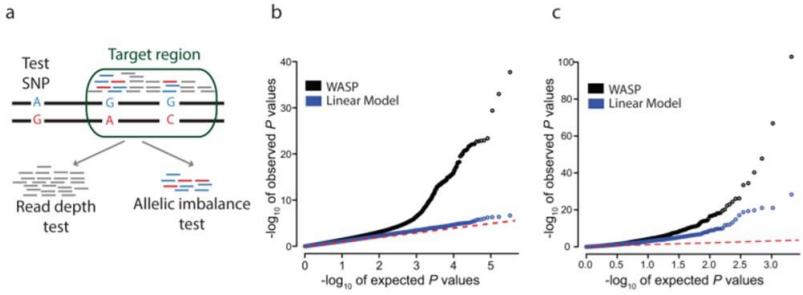


FADS2 eQTL effect is disproportionately large in ileum

PNKD and *RGS14* eQTL have opposite effects in blood and ileum

Overall there is good concordance in sign of effect

WASP



https://github.com/bmvdgeijn/WASP

Formulates a CHT: Combined Haplotype Test
"The CHT jointly models two components: the allelic imbalance at phased heterozygous SNPs and the total read depth in the target region"

Van de Geijn, *et al* (2015) *Nature Methods* **12**: 1061-1063

Some other software

http://omictools.com/eqtl-mapping-c1260-p1.html

PLINK: The basic tool for GWAS
http://pngu.mgh.harvard.edu/~purcell/plink/tutorial.shtml

Matrix eQTL: Ultra-fast eQTL analysis
http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/

GEMMA: Genome-wide Efficient Mixed Model Association (GEMMA)
http://stephenslab.uchicago.edu/software.html#gemma

FMeQTL: Bayesian Joint mapping
https://github.com/xqwen/fmeqtl

DAP: Deterministic Approcimation of Posteriors (Fast Bayesian)
https://github.com/xqwen/dap

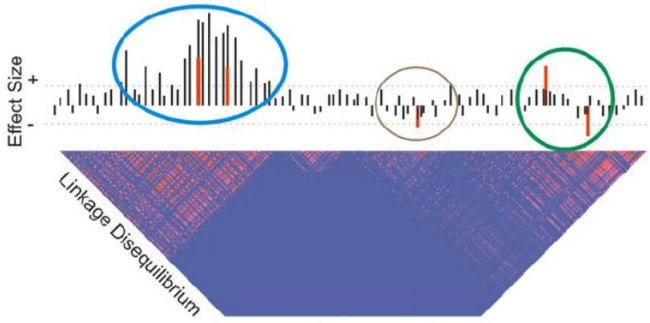
CAVIAR: Bayesian Fine Mapping
http://genetics.cs.ucla.edu/caviar/

Ventham et al (2016) *Nature Communications* **7**: 13507

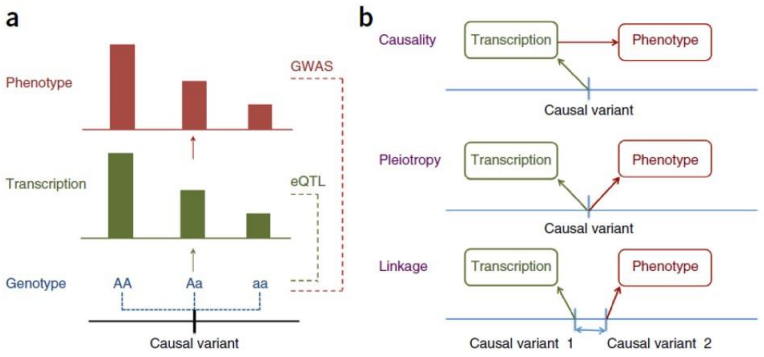
Why Colocalized Signals do not alone imply Causation

Sampling variance means that we can only map "credible intervals"
Many genes harbor multiple eSNPs, and possibly multiple trait associated SNPs
LD means that multiple sites can interfere with one another in estimation of peak locations

The nearest gene is only sometimes the one affected by a SNP!



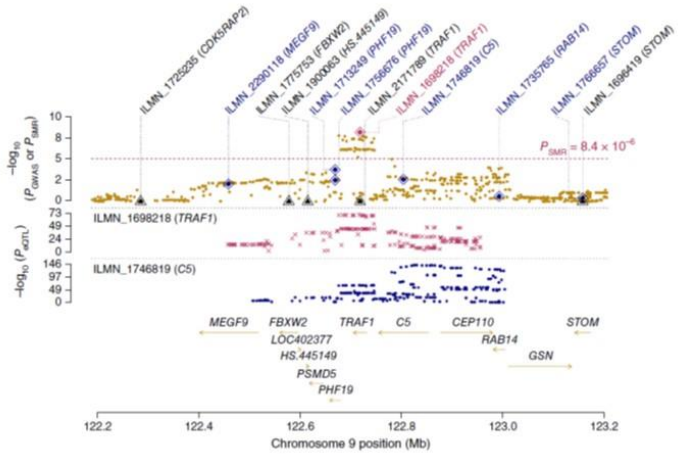
SMR: Causality and Correlation for eQTL and GWAS



Statistical power of MR is proportional to: variance of SNP on transcript, variance of transcript on phenotype, N

Zhu et al (2016) *Nature Genetics* **48**: 481-487

An Example: TRAF1 not C5 likely explains the RA association



Zhu et al (2016) *Nature Genetics* **48**: 481-487

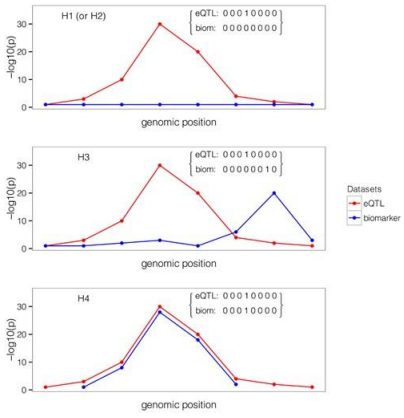
Coloc: A Bayesian test for colocalization of pairs of association signals

H1 is the hypothesis that there is only an eQTL signal at a locus

H2 is the hypothesis that there is only a GWAS signal at a locus.

H3 is the hypothesis that there are two independent eQTL and GWAS signals in linkage.

H4 is the strong hypothesis that the same SNP (not just the locus) is responsible for both the GWAS and eQTL.



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383

Examples of H3 and H4

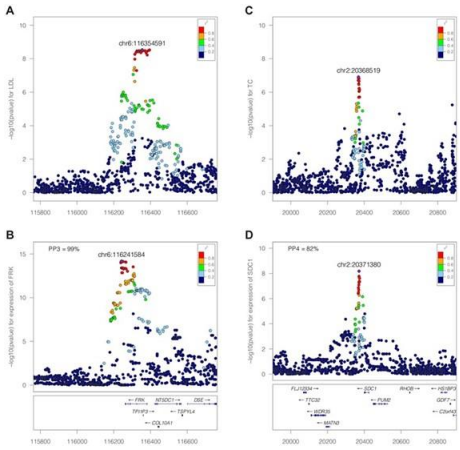
On the left, the profile of association at the FRK locus with LDL (top) is very different from that with *FRK* expression.

H3 is the supported hypothesis.

On the right, even though there are two different peak SNPs, they are in the same strong LD region and the profiles are almost the same for Total Cholesterol and *Soc1* expression.

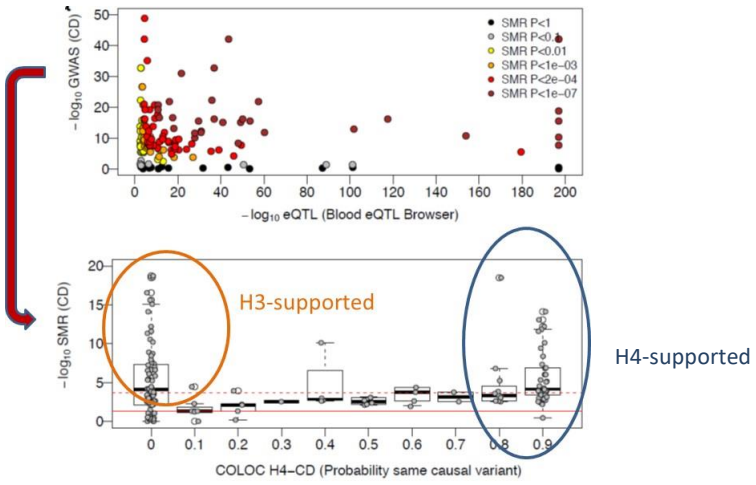
H4 is the supported hypothesis.

Bayesian analysis evaluate each H relative to the other four and generates a confidence level for the most likely one.



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383

SMR and coloc are complementary in our Crohns Disease study



Limitations of colocalization analyses

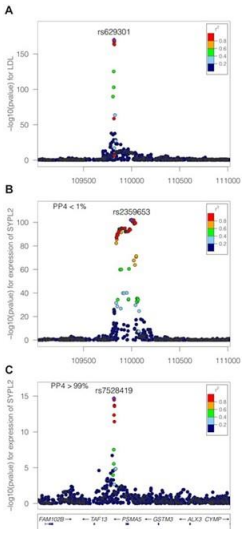
Heavily dependent on statistical power of the contributing analyses, which is generally relatively low

Depends on high quality imputation if the SNPs are not directly typed

Assumes that the GWAS and eQTL are evaluated on the same population (there is no stratification)

A negative result may arise if the incorrect tissue is being studied for the gene expression

Assumes there is a single causal variant at each locus for each effect (which is very unlikely) although this example shows that conditional analysis has the potential to resolve joint effects



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383

Multiple regression plus function

RESEARCH ARTICLE

Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation

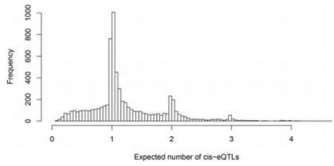
Xiaoquan Wen[1]*, Francesca Luca[2,3], Roger Pique-Regi[2,4]*

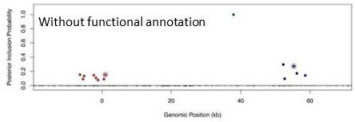
1 Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA, **2** Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI, USA, **3** Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, USA, **4** Department of Clinical and Translational Sciences, Wayne State University, Detroit, MI, USA

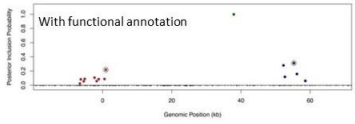
* xwen@umich.edu (XW); rpique@wayne.edu (RPR)

Abstract

Mapping expression quantitative trait loci (eQTLs) has been shown as a powerful tool to uncover the genetic underpinnings of many complex traits at the molecular level. In this paper, we present an integrative analysis approach that leverages eQTL data collected from multiple population groups. In particular, our approach effectively identifies multiple independent cis-eQTL signals that are consistent across populations, accounting for population heterogeneity in allele frequencies and linkage disequilibrium patterns. Furthermore, by integrating genomic annotations, our analysis framework enables high-resolution functional analysis of eQTLs. We applied our statistical approach to analyze the GEUVADIS data consisting of samples from five population groups. From this analysis, we concluded that i) jointly analysis across population groups greatly improves the power of eQTL discovery and the resolution of fine mapping of causal eQTL ii) many genes harbor multiple independent eQTLs in their *cis* regions iii) genetic variants that disrupt transcription factor binding are significantly enriched in eQTLs (p-value = 4.93×10^{-22}).







Wen *et al* (2015) *PLoS Genetics* **11**: e1005176

Joint Mapping

A variety of open source methods are appearing that utilize Bayesian methods to perform joint mapping of eQTL

A statistical framework for joint eQTL analysis in multiple tissues.
Flutre T, Wen X, Pritchard J, Stephens M. *PLoS Genet*. 2013 **9**(5): e1003486.

This paper shows that combining signals across tissues increases power while also allowing assessment of whether the effect sizes are different in different cell types. Implemented in eQTLBMA software.

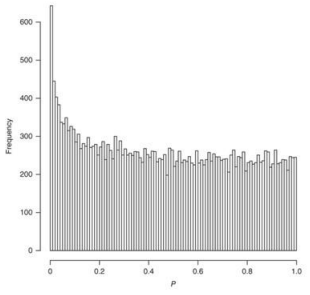
Cross-population joint analysis of eQTLs: Fine mapping and functional annotation.
Wen X, Luca F, Pique-Regi R. *PLoS Genet.* **11**(4): e1005176.

This paper shows that combining signals across populations increases power while also allowing assessment of how incorporating ENCODE data improves resolution. Implemented in FM QTL software.

Efficient integrative multi-SNP association analysis via Deterministic Approximation of Posteriors
Wen X, Lee Y, Luca F, Pique-Regi R. *AM J Hum. Genet.* **98**(6): 1114-1129.

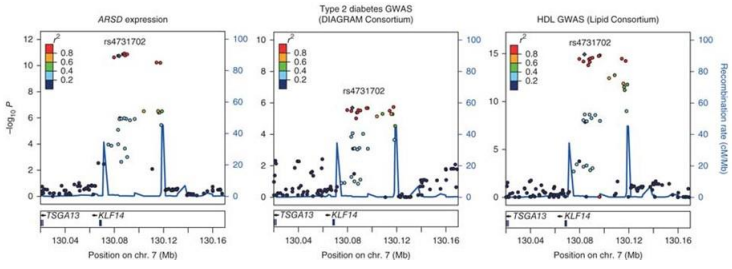
This paper extends the framework for incorporating ENCODE data while allowing for multiple causal variants at each locus. Implemented in DAP software: http://github.com/xqwen/dap/

Trans-Effect of KLF14



		MuTHER		deCODE all		deCODE maternal		deCODE paternal		Combined MuTHER + deCODE maternal	
Gene	Chr.	Effect (s.e.)	P	Effect (s.e.)	P	Effect (s.e.)	P	Effect (s.e.)	P	Z score	P
APH1B	15	0.08 (0.013)	1.2×10^{-8}	0.11 (0.059)	0.08	0.17 (0.085)	0.07	0.07 (0.083)	0.44	6.1	9.7×10^{-10}
ARSD	X	0.08 (0.012)	1.9×10^{-11}	0.24 (0.059)	2.2×10^{-4}	0.51 (0.083)	2.6×10^{-9}	−0.004 (0.083)	0.96	8.6	5.4×10^{-18}
C8orf82	8	0.09 (0.014)	4.8×10^{-10}	0.28 (0.058)	8.9×10^{-6}	0.69 (0.080)	2.1×10^{-14}	−0.09 (0.082)	0.28	9.3	1.1×10^{-20}
GNB1	1	0.05 (0.009)	4.0×10^{-8}	0.23 (0.059)	1.8×10^{-4}	0.42 (0.085)	1.6×10^{-6}	0.06 (0.084)	0.51	7.2	6.1×10^{-13}
KLF13	15	0.10 (0.017)	2.2×10^{-9}	−0.01 (0.060)	0.94	0.01 (0.086)	0.89	−0.02 (0.084)	0.80	4.8	1.4×10^{-6}
MYL5	4	0.09 (0.017)	4.5×10^{-8}	0.20 (0.059)	1.3×10^{-3}	0.45 (0.083)	1.3×10^{-7}	−0.04 (0.083)	0.60	7.4	1.1×10^{-13}
NINJ2	12	0.08 (0.013)	8.4×10^{-9}	0.14 (0.060)	0.03	0.24 (0.087)	0.01	0.05 (0.085)	0.59	6.3	4.1×10^{-10}
PRMT2	21	0.06 (0.010)	6.9×10^{-9}	0.18 (0.060)	0.01	0.27 (0.087)	6.7×10^{-3}	0.09 (0.085)	0.33	6.4	2.1×10^{-10}
SLC7A10	19	−0.27 (0.042)	2.7×10^{-10}	−0.21 (0.057)	7.4×10^{-4}	−0.31 (0.082)	3.3×10^{-4}	−0.11 (0.081)	0.18	−7.3	3.8×10^{-13}
TPMT	6	0.10 (0.013)	1.6×10^{-14}	−0.04 (0.060)	0.49	−0.03 (0.087)	0.78	−0.06 (0.084)	0.49	6.4	1.8×10^{-10}

The effect allele is the type 2 diabetes risk allele C, which has a frequency of 55% in the HapMap CEU population. Chr., chromosome; s.e., standar



Small et al (2011) *Nature Genetics* **43**: 561-564

Challenges for eSNP analysis

- Great for finding transcripts regulated by one or two major effect SNPs that explain 20-60% of variance – but these are a minority
- Multiple comparison issues limit the power to detect weaker effects and to map several sites per transcript (unless N>10,000 ?)
- Outliers can produce very small p-values when MAF<5% and are quite common; PARTICULARLY with respect to interaction effects because one or two individuals will by chance be in a sub-group
- Only a few human tissues are accessible, and cost/ethics preclude recurrent sampling in many cases: hard to get longitudinal data
- Overlap between tissues estimated as only 10-20%, not much less than power to replicate 'marginal' associations at 10^{-8}

eSNP analysis by RNA-Seq

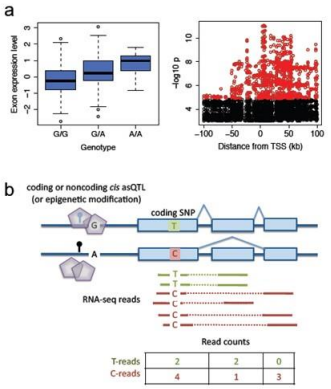
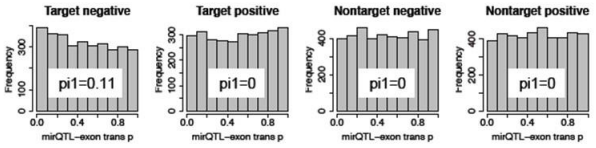


Table 1 | Numbers of transcriptome features with a QTL (FDR 5%)

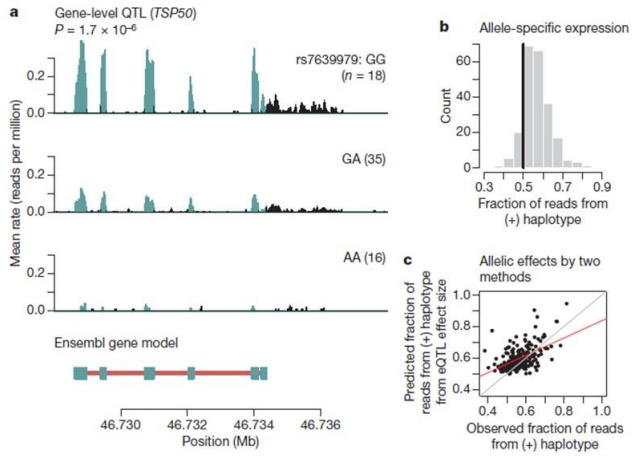
	Total	EUR (n = 373)	YRI (n = 89)	Union
Exon eQTL	12,981 genes	7,390	2,369	7,825
Gene eQTL	13,703 genes	3,259	501	3,773
Transcript ratio QTL	7,855 genes	620	83	639
mirQTL	644 miRNAs	57	15	60
Transcribed repeat eQTL	43,875 repeats	5,763	1,055	6,069

miR-QTL may also regulate negative targets



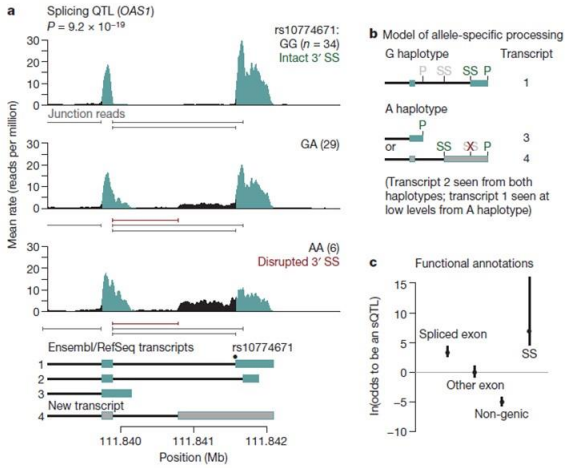
Lappalainen, Dermitzakis et al (2013) *Nature* **501**: 506-511

eQTL and Additivity



Pickrell et al. (2010) *Nature* **464**: 768-772

sQTL (Splicing QTL)



Pickrell et al. (2010) *Nature* **464**: 768-772

MISO: Mixture of Isoforms

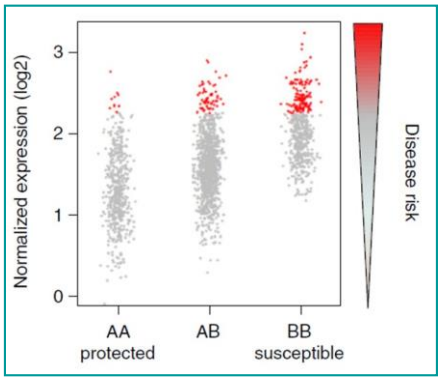
http://genes.mit.edu/burgelab/miso/

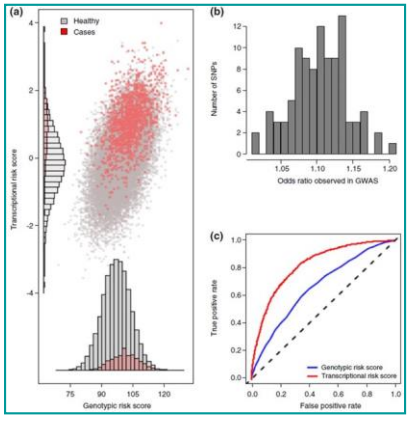
Estimates the Percent Spliced In (PSI, ψ) for various features:

- Alternate First Exon
- Mutually exclusive Exon
- Excluded Exon
- Included Intron
- Alternate Splice Donor
- Alternate Splice Acceptor
- Alternate 3′ end

Katz et al. (2010) *Nature Methods* **7**: 109–115

Transcriptional Risk Scores - theory

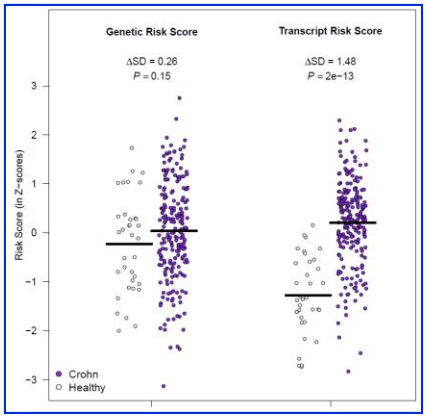




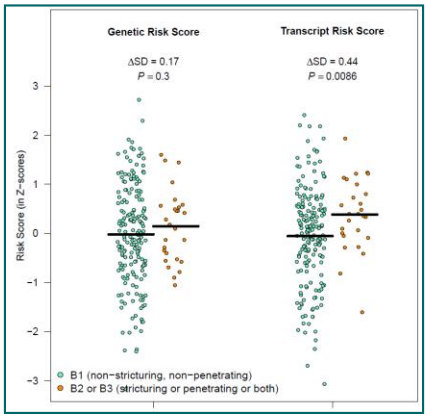
Gibson et al (2015) *Genome Medicine* **7**: 60

Transcriptional Risk Scores for Crohn's Disease

Healthy-Disease



Disease Progression



Coherent and Incoherent associations