



Summer Institute  
in Statistical Genetics 2018

## Integrative Genomics

### 5b. Genetics of Gene Expression: eSNPs



[ggibson.gt@gmail.com](mailto:ggibson.gt@gmail.com)

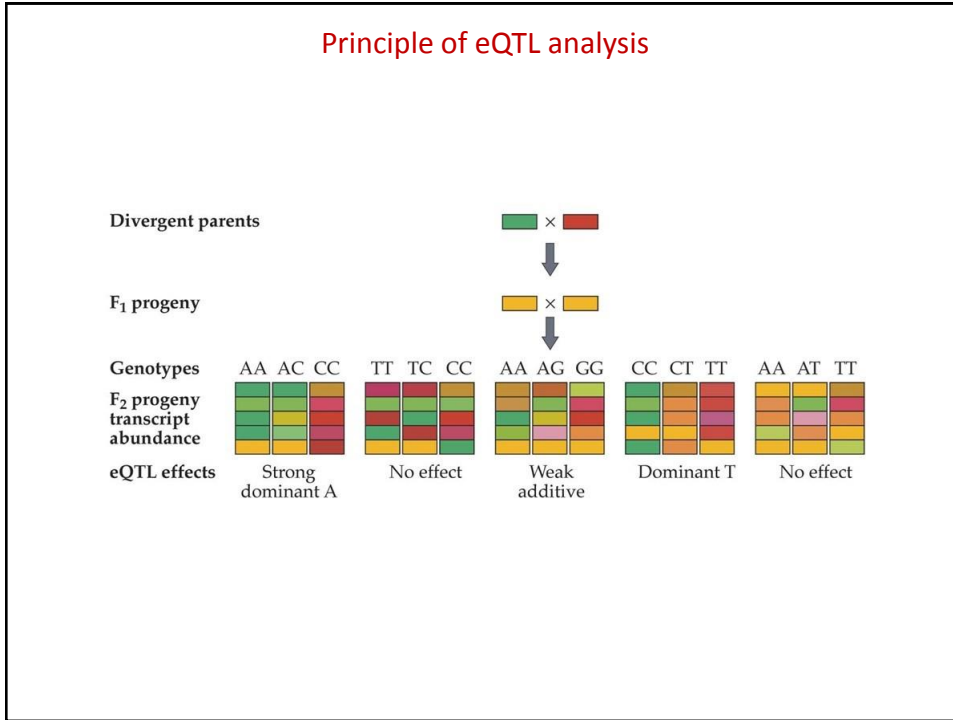
<http://www.gibsongroup.biology.gatech.edu>

Georgia  
Tech



### 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



### cis and trans eQTL

- Liver samples from 111 F<sub>2</sub> 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

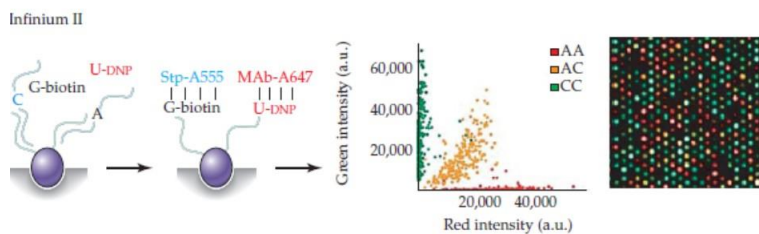
Schadt, Friend et al (2003) *Nature* **422**: 297-302

### Limitations of eQTL analysis

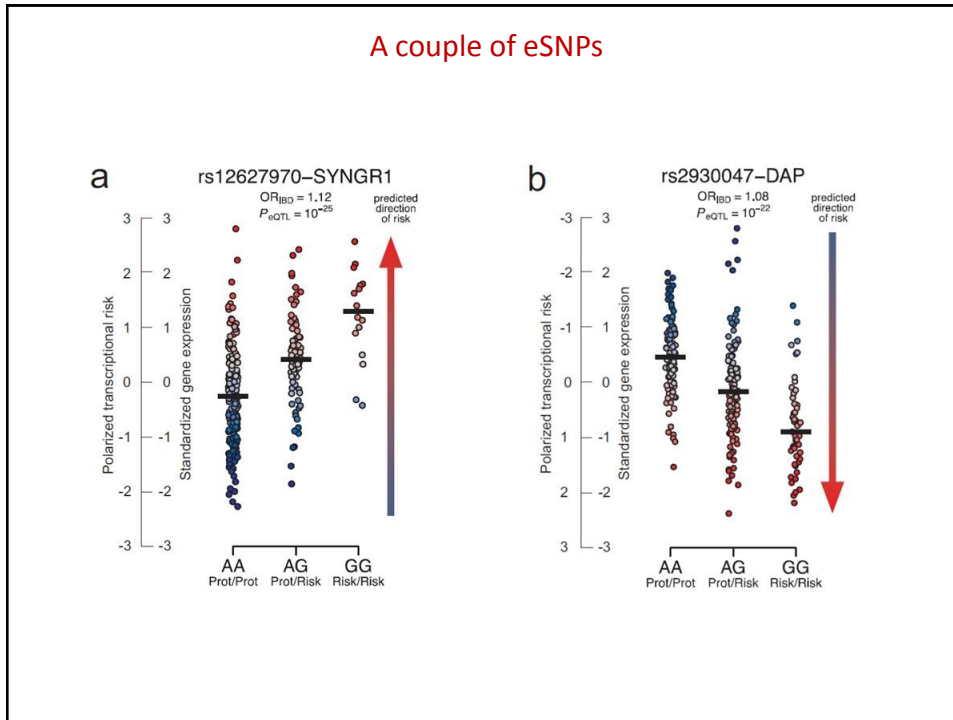
- Any QTL experiment is only a comparison of two lines, so does not say anything about the frequency of QTL effects in a population
- If the number of F2 or BC progeny is less than 100, QTL analysis is prone to false positives, particularly for *trans*-hotspots
- Consequently, significance must be evaluated 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  $\sim$  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



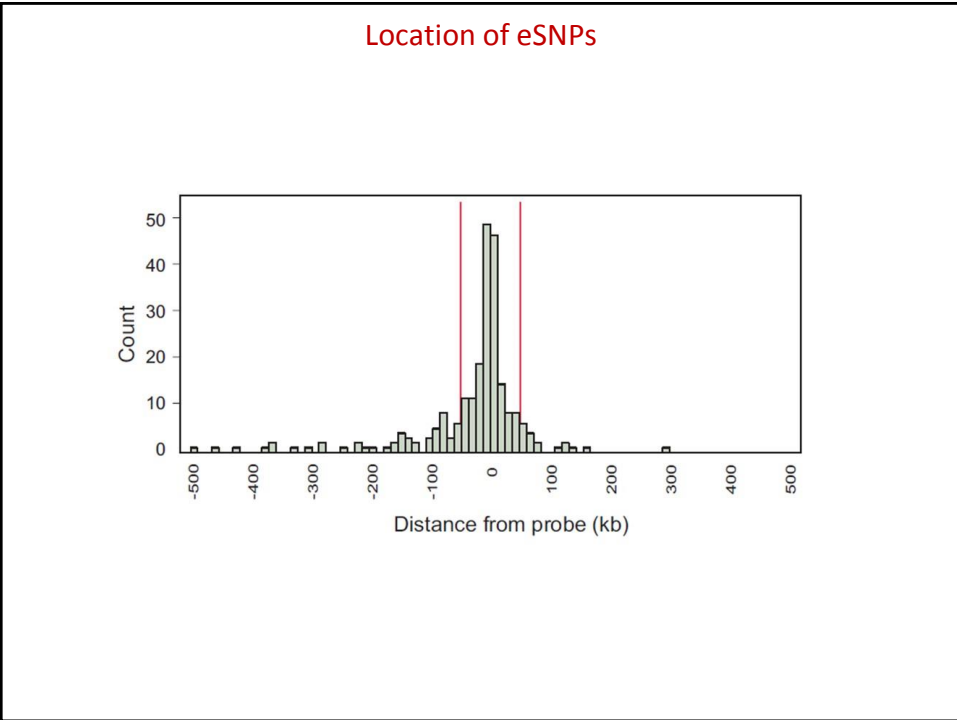
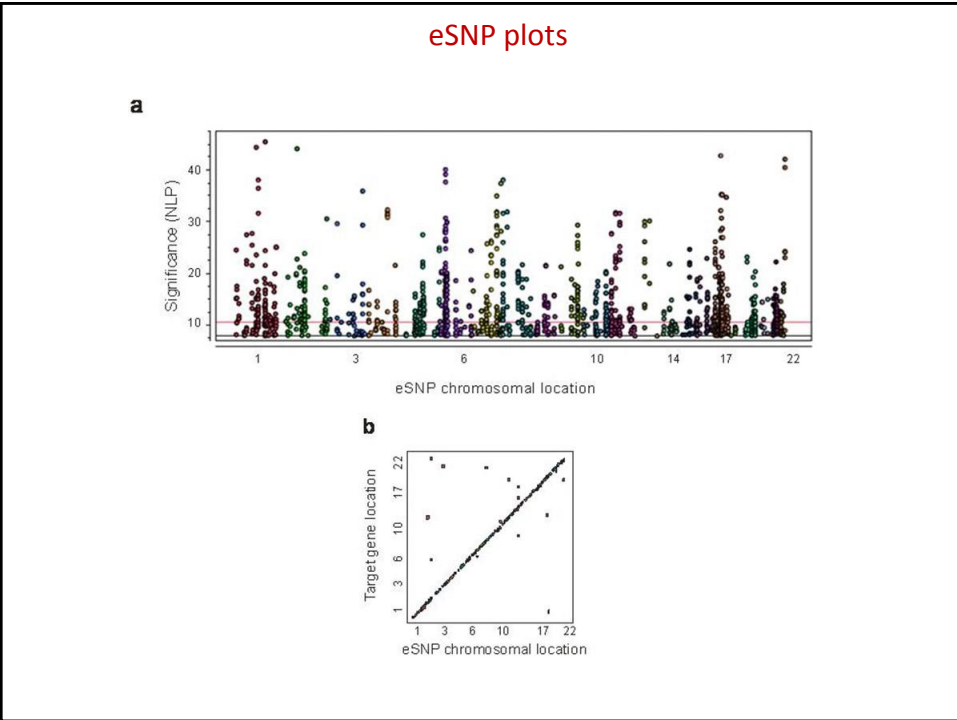
- Whole transcriptome profiling of the same individuals
- GWAS (Genome-wide association study) for transcription -> precise localization of regulatory SNPs in *cis* and *trans*



### Significance thresholds

- Bonferroni for *cis*-linkages:  
 $0.05 / (20,000 \text{ genes} \times 250 \text{ SNPs}) = 1 \times 10^{-8}$
- Permutation for *cis*-linkages:  
 Random sets of  $n$  SNPs from distribution of 2Mb windows
- Bonferroni for *trans*-linkages:  
 $0.05 / (20,000 \text{ genes} \times 500,000 \text{ SNPs}) = 5 \times 10^{-12}$
- 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



### Effect of Normalization

Table 3 eSNP Analyses

Normalization	Pearson Correlation				Spearman Rank Correlation	
	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.

### GTEx (Genotype-Tissue-Expression Project)

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

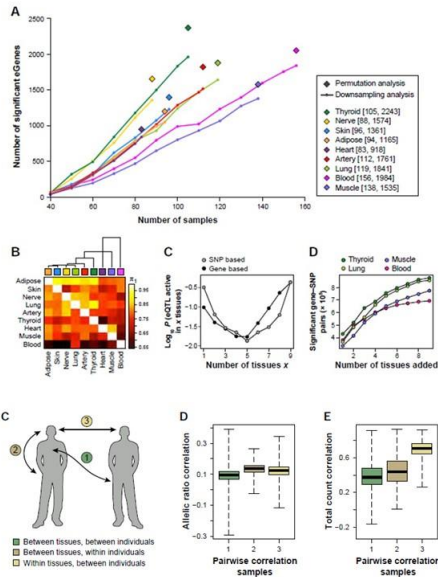
The GTEx Consortium<sup>1,2</sup>

1. Author Affiliations

2. Corresponding author: Kristin G. Ardlie (kardlie@ccpadsinstitute.org) or Emmanouil T. Dermizakis (emmanouil.dermizakis@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 1644 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.

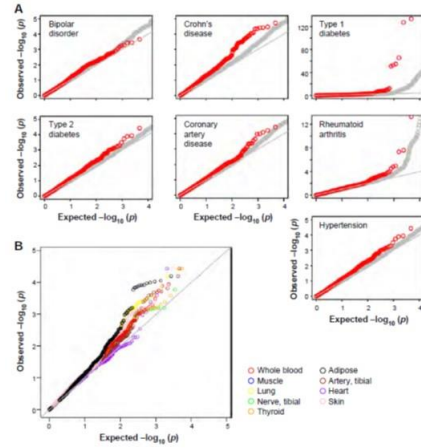
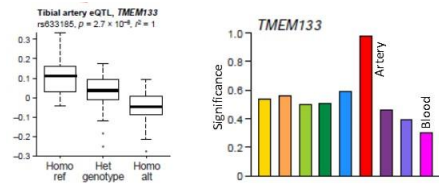


Science (2015) 9(5): e1003486

### 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

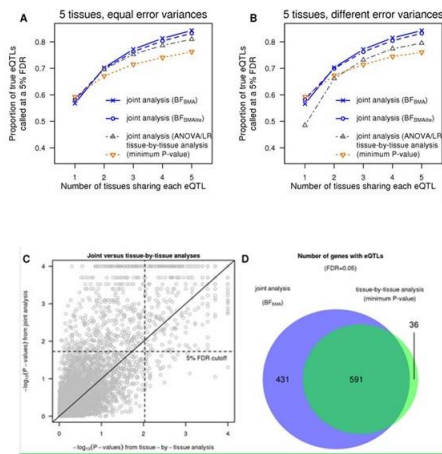
#### A Statistical Framework for Joint eQTL Analysis in Multiple Tissues

Timothée Flutre, Xiaoqun Wen, Jonathan Pritchard, Matthew Stephens

Article	Authors	Metrics	Comments	Related Content

**Abstract**

Mapping expression Quantitative Trait Loci (eQTLs) represents a powerful and widely adopted approach to identifying putative regulatory variants and linking them to specific genes. Up to now eQTL studies have been conducted in a relatively narrow range of tissues or cell types. However, understanding the biology of organismal phenotypes will involve understanding regulation in multiple tissues, and ongoing studies are collecting eQTL data in dozens of cell types. Here we present a statistical framework for powerfully detecting eQTLs in multiple tissues or cell types (or more generally, multiple subgroups). The framework explicitly models the potential for each eQTL to be active in some tissues and inactive in others. By modeling the sharing of active eQTLs among tissues, this framework increases power to detect eQTLs that are present in more than one tissue compared with "tissue-by-tissue" analyses that examine each tissue separately. Conversely, by modeling the inactivity of eQTLs in some tissues, the framework allows the proportion of eQTLs shared across different tissues to be formally estimated as parameters of a model, addressing the difficulties of accounting for incomplete power when comparing overlaps of eQTLs identified by tissue-by-tissue analysis. Applying our framework to re-analyze data from transformed B cells, T cells, and fibroblasts, we find that it substantially increases power compared with tissue-by-tissue analysis, identifying 63% more genes with eQTLs (at FDR = 0.05). Further, the results suggest that, in contrast to previous analyses of the same data, the majority of eQTLs detectable in these data are shared among all three tissues.



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

## Multiple regression plus function

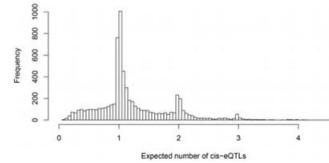
RESEARCH ARTICLE

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

Xiaoquan Wen<sup>1\*</sup>, Francesca Luca<sup>2,3</sup>, Roger Pique-Regi<sup>2,4\*</sup>

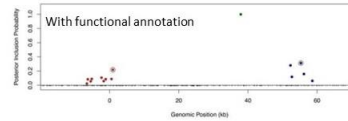
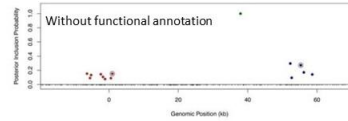
<sup>1</sup> Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA, <sup>2</sup> Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI, USA, <sup>3</sup> Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, USA, <sup>4</sup> Department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

\* [xwen@umich.edu](mailto:xwen@umich.edu) (XW); [rpique@wayne.edu](mailto:rpique@wayne.edu) (RPR)



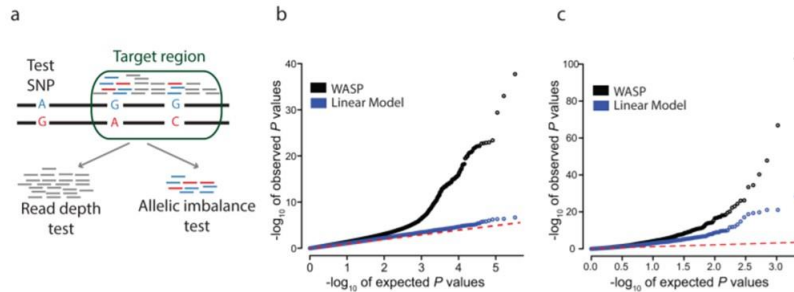
#### Abstract

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 ( $p\text{-value} = 4.93 \times 10^{-25}$ ).



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

## 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/eqt-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/](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 Approximation of Posteriors (Fast Bayesian)  
<https://github.com/xqwen/dap>

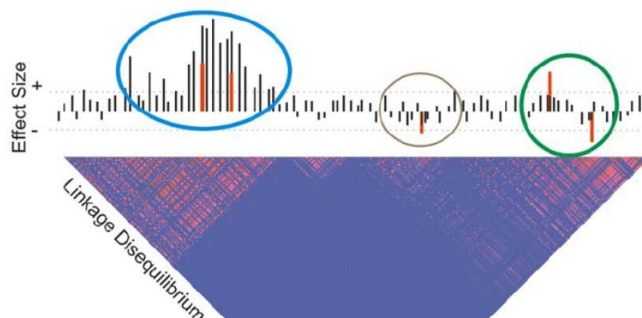
CAVIAR: Bayesian Fine Mapping  
<http://genetics.cs.ucla.edu/caviar/>

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

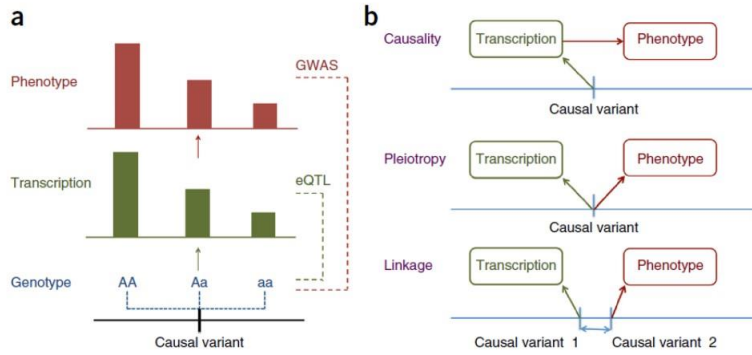
### Why Colocalized Signals do not alone imply Causation

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

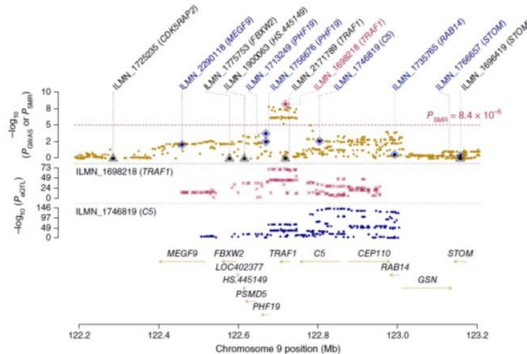
### An Example: TRAF1 not C5 explains the RA association

The GWAS credible interval in the vicinity of TRAF1 contains the peak  $P_{SMR}$  with one of the TRAF1 probes, not with the C5 probe.

The eQTL peak in the region is actually stronger for C5 than TRAF1, but it is due to SNPs in variable LD, not to pleiotropy.

HEIDI confirms this since there is more heterogeneity for the blue than purple SNPs.

So, the TRAF1 region SNPs are GWAS significant and strong C5 eQTL, but probably mediate their effect through TRAF1.



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

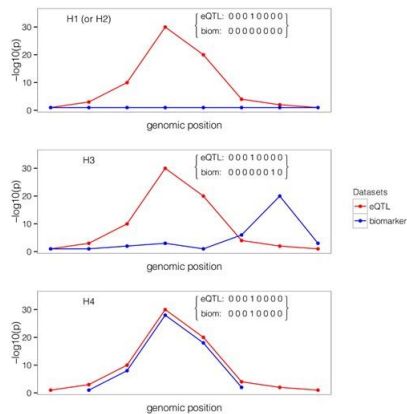
## Coloc: A Bayesian test for colocalization of pairs of association signals

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

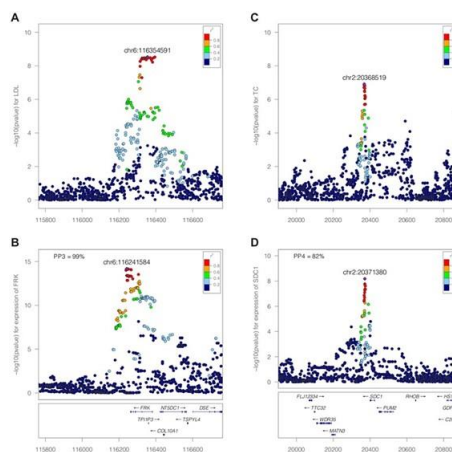
On the left, the profile of association at the FRK locus with LDL (top) is very different from that with *FRK* expression.

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.

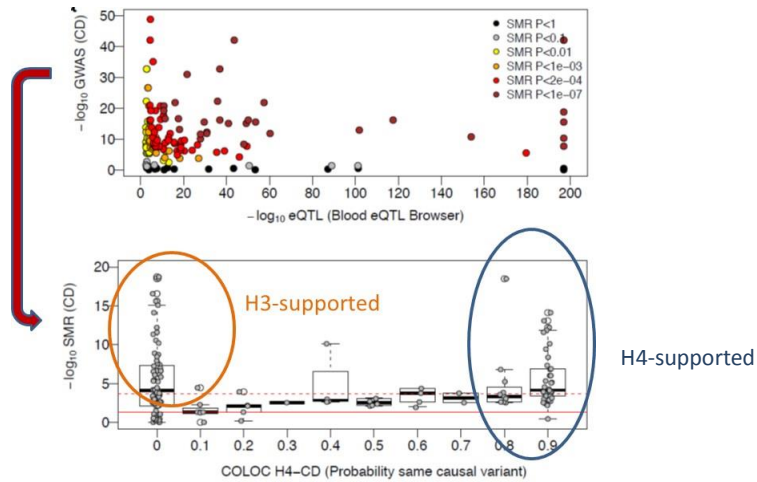
H4 is the supported hypothesis.

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

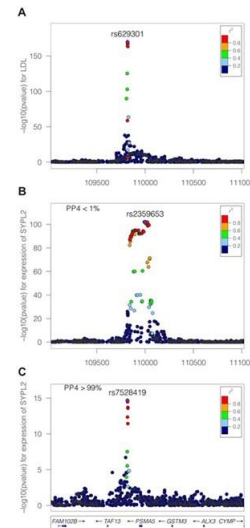
Heavily dependent on statistical power of the contributing analyses, which is generally relatively low

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

## 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.

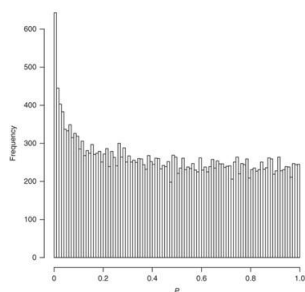
Cross-population joint analysis of eQTLs: Fine mapping and functional annotation.  
 Wen X, Luca F, Pique-Regi R. *PLoS Genet.* **11**(4): e1005176.

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.

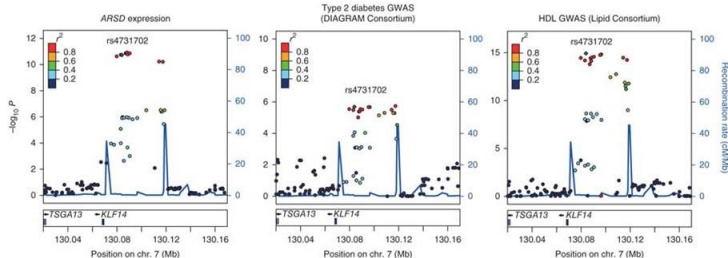
This paper extends the framework for incorporating ENCODE data while allowing for multiple causal variants at each locus. Implemented in DAP software: <http://github.com/xqwen/dap/>

## Trans-Effect of KLF14



Gene	Chr.	MUTHER		deCODE all		deCODE maternal		deCODE paternal		Combined MUTHER + de maternal	
		Effect (s.e.)	P	Effect (s.e.)	P	Effect (s.e.)	P	Effect (s.e.)	P	Z score	P
APHB	15	0.08 (0.013)	1.2 × 10 <sup>-8</sup>	0.11 (0.059)	0.08	0.17 (0.085)	0.07	0.07 (0.063)	0.44	6.1	9.7 × 10 <sup>-10</sup>
ARSD	X	0.08 (0.012)	1.9 × 10 <sup>-8</sup>	0.24 (0.059)	2.2 × 10 <sup>-7</sup>	0.51 (0.083)	2.6 × 10 <sup>-7</sup>	-0.04 (0.052)	0.96	8.6	5.4 × 10 <sup>-10</sup>
C6orf92	8	0.09 (0.014)	4.8 × 10 <sup>-9</sup>	0.28 (0.059)	8.9 × 10 <sup>-7</sup>	0.69 (0.080)	2.1 × 10 <sup>-7</sup>	-0.09 (0.052)	0.28	9.3	1.1 × 10 <sup>-10</sup>
GWB1	1	0.06 (0.009)	4.0 × 10 <sup>-8</sup>	0.23 (0.059)	1.8 × 10 <sup>-7</sup>	0.42 (0.085)	1.6 × 10 <sup>-7</sup>	0.06 (0.084)	0.51	7.2	6.1 × 10 <sup>-10</sup>
KLF13	15	0.10 (0.017)	2.2 × 10 <sup>-8</sup>	-0.01 (0.060)	0.94	0.01 (0.086)	0.89	-0.02 (0.084)	0.80	4.8	1.4 × 10 <sup>-9</sup>
MYL5	4	0.09 (0.017)	4.5 × 10 <sup>-9</sup>	0.20 (0.059)	1.3 × 10 <sup>-7</sup>	0.45 (0.083)	1.3 × 10 <sup>-7</sup>	-0.04 (0.083)	0.60	7.4	1.1 × 10 <sup>-10</sup>
NAL2	12	0.08 (0.013)	8.4 × 10 <sup>-9</sup>	0.14 (0.060)	0.03	0.24 (0.087)	0.01	0.05 (0.085)	0.59	6.3	4.1 × 10 <sup>-10</sup>
PRMT2	21	0.06 (0.010)	6.9 × 10 <sup>-9</sup>	0.18 (0.060)	0.01	0.27 (0.087)	6.7 × 10 <sup>-7</sup>	0.09 (0.085)	0.33	6.4	2.1 × 10 <sup>-10</sup>
SLC7A10	19	-0.27 (0.042)	2.7 × 10 <sup>-10</sup>	-0.21 (0.057)	7.4 × 10 <sup>-9</sup>	-0.31 (0.082)	3.3 × 10 <sup>-7</sup>	-0.11 (0.081)	0.18	-7.3	3.8 × 10 <sup>-10</sup>
TPMT	6	0.10 (0.013)	1.6 × 10 <sup>-8</sup>	-0.04 (0.060)	0.49	-0.03 (0.087)	0.78	-0.06 (0.084)	0.49	6.4	1.6 × 10 <sup>-10</sup>

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., standard error.



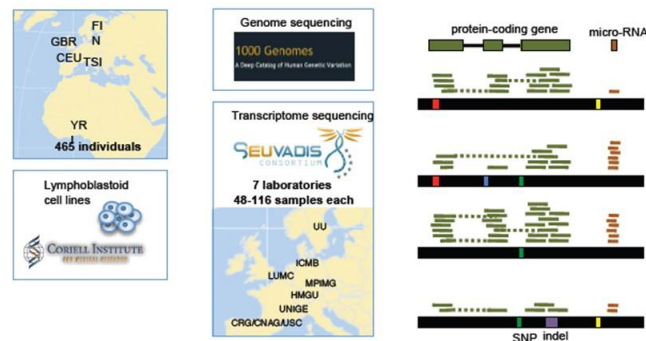
Small et al (2011) *Nature Genetics* **43**: 561-564

### 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  $10^{-8}$

### 1000G eSNP study

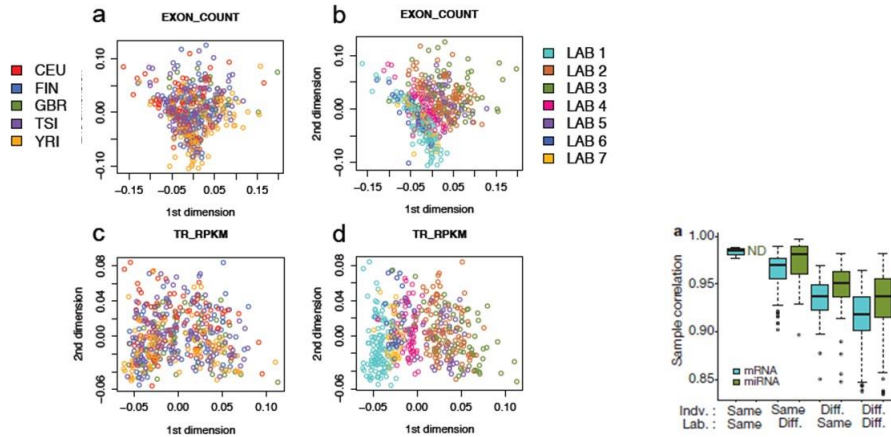
Performed RNA-Seq and miRNA-Seq on LCL for ~90 people each from five 1000G populations: Utah (CEPH), Finland, Britain, Tuscany and Nigeria (Yorubans)



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

### Technical effects in the study

Sequencing in 7 laboratories showed inter-lab variance is less than among individual, yet there clearly are lab effects, particularly at transcript level

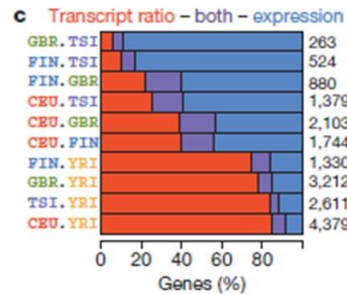
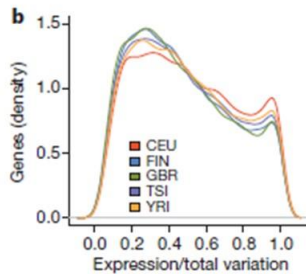


Lappalainen, Dermitzakis et al (2013) *Nature* 501: 506-511

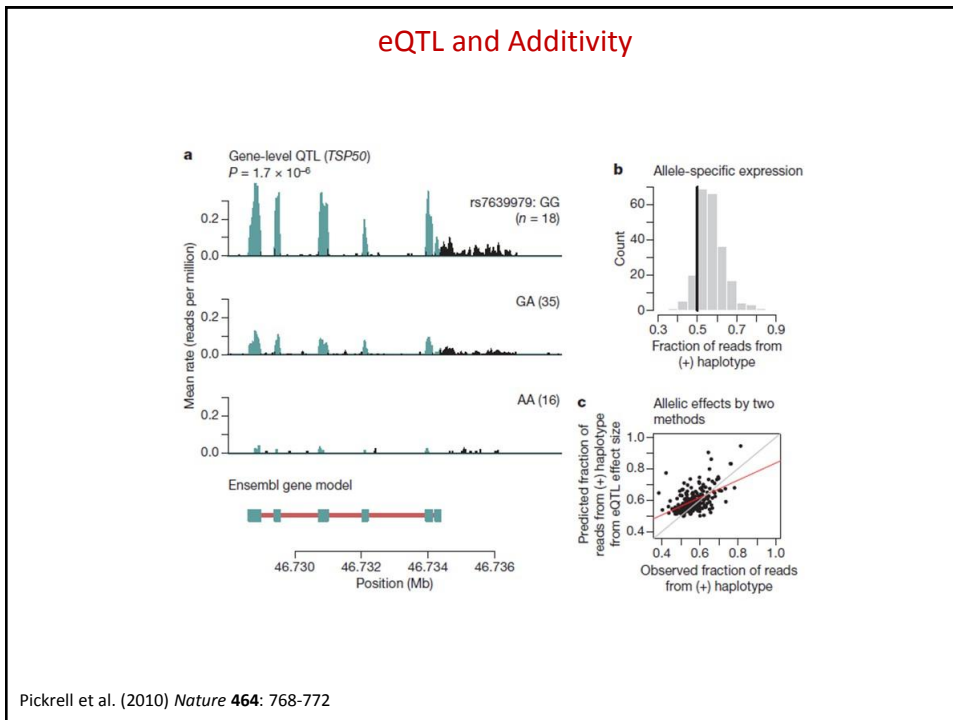
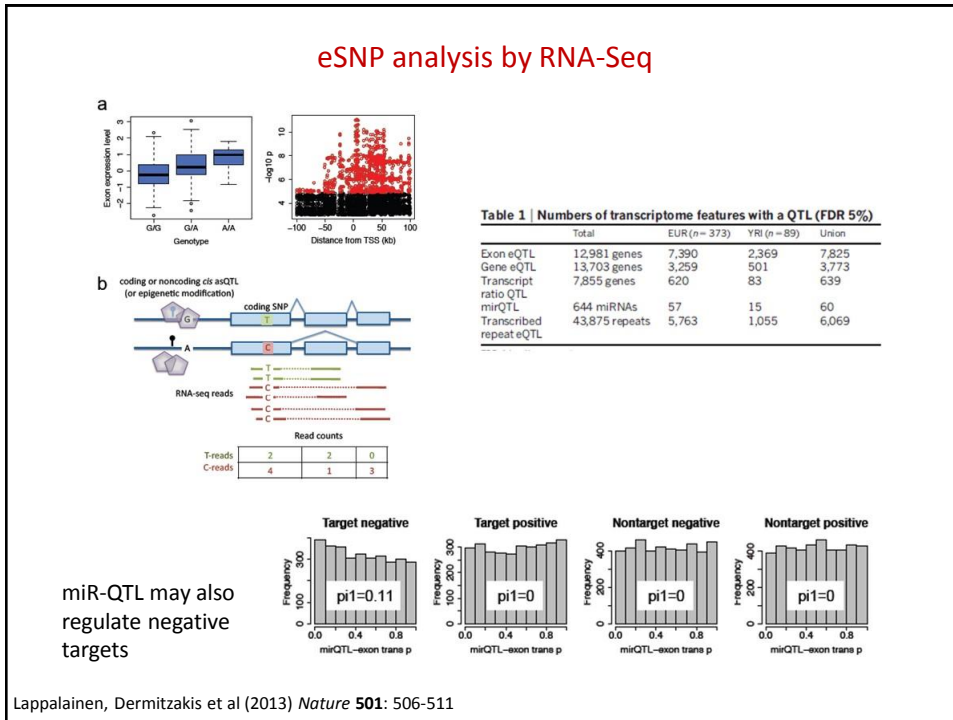
### Expression and Isoform components

Proportion of expression variation among individuals within populations ranges from 20% to 95%

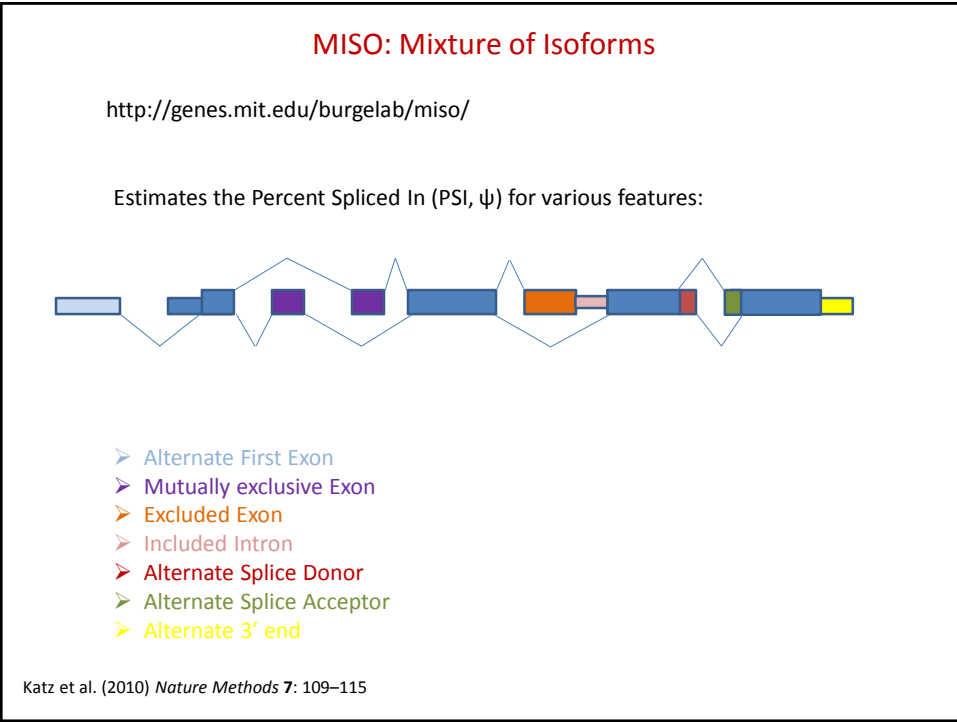
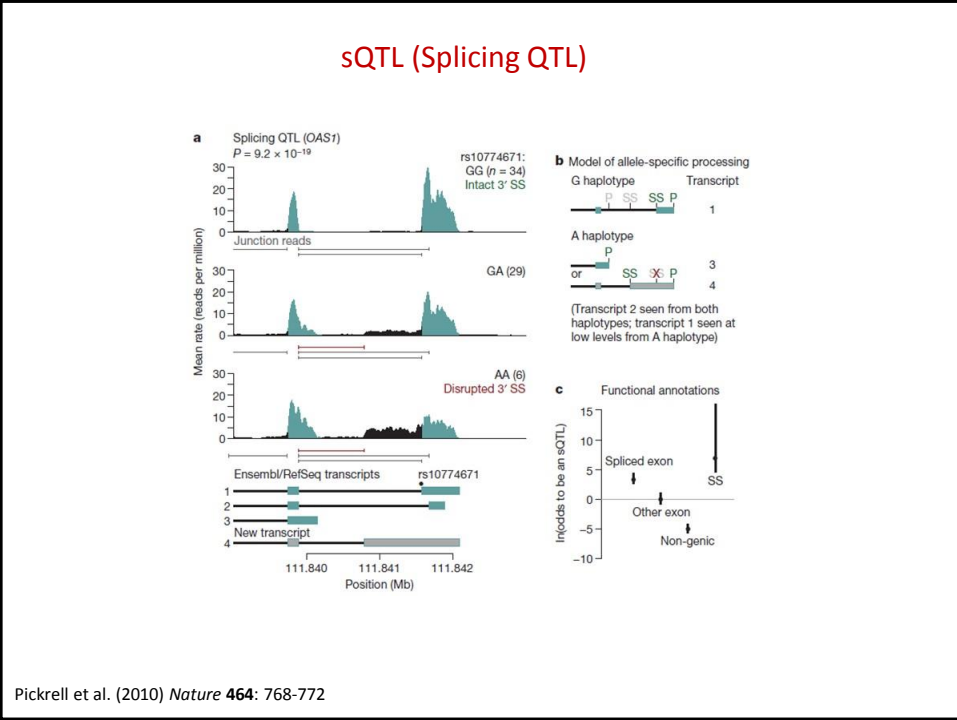
Transcript ratio (isoform abundance) is greater than overall expression variation, and varies among populations, especially wrt Yorubans.



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







## Meta-analysis

<http://genenetwork.nl/bloodeqtlbrowser/>

### Blood eQTL browser

This web page accompanies the manuscript titled: Systematic identification of *trans*-eQTLs as putative drivers of known disease associations. By Westra et al. which has been published in Nature Genetics. If you want to use any of the cis- or *trans*-eQTL results displayed on this page in your publication, please cite this paper as indicated below. For further questions, contact the corresponding author: [voide@uio.no](mailto:voide@uio.no)

### Download eQTL Results

You can download the cis- and *trans*-eQTLs selected at a false discovery rate of 0.00.  
Cis-eQTLs (FDR 0.0)  
*Trans*-eQTLs (FDR 0.0)

### How to cite

If you use the eQTLs present on this website in your paper or research, please cite our work. Download citation directly from Nature Genetics

### Query eQTL Results

Or, you can query the cis- and *trans*-eQTLs below (examples: rs7607916 or VEGFA)

Gene or SNP name:

NATURE GENETICS | LETTER



日本語要約

## Systematic identification of *trans* eQTLs as putative drivers of known disease associations

Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, Alexandra Zhernakova, Daria V Zhernakova, Jan H Veldink, Leonard H Van den Berg, Juha Karjalainen, Sebo Withoff, André G Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter A C 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Meitner *et al.*

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