Summer Institute in Statistical Genetics 2016

Integrative Genomics

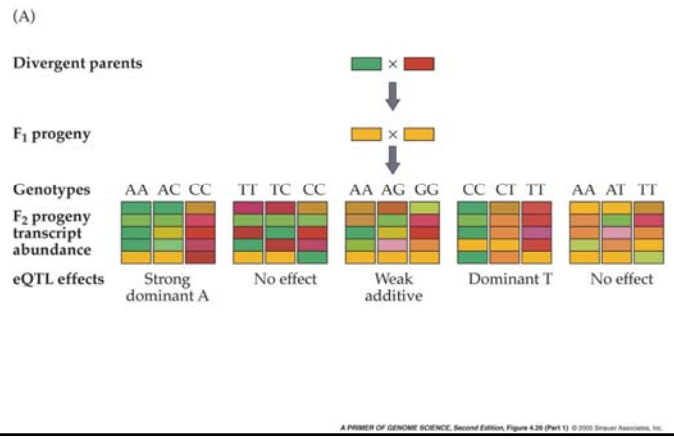
5a. Genetics of Gene Expression

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

Expression QTL analysis

- The architecture of transcription maps genotype onto phenotype
- Expression QTL (eQTL) are QTL that modulate transcript abundance in pedigrees or crosses
- The vast majority of GWAS variants (associated with disease or continuous traits) are now known to be regulatory, and hence to have eQTL effects.
- Estimates of heritability of transcription also suggest that it is remarkably high, in the range of 0.2 to 0.5, with transcription sometimes showing a higher genetic component than visible traits

Principle of eQTL analysis

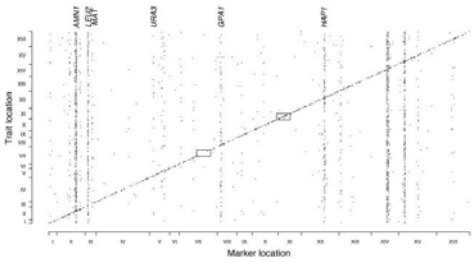


cis and *trans* eQTL

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

- 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

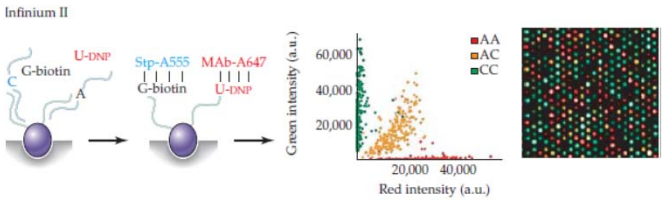


Limitations of eQTL analysis

- 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*

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

Gutenberg Heart Study example

Zeller et al (2011) *PLoS ONE* **5**: e10693

Significance level	Minimum R^2 [§]	Total number of associations	*cis*/*trans* ratio for associations	Total number of associated expressions (eQTLs)	*cis*/*trans* ratio for eQTLs	Total number of associated SNPs (eSNPs)	*cis*/*trans* ratio for eSNPs
$<10^{-6}$	0.016	93491	2.1	8575	0.5	67190	2.4
$<10^{-8}$	0.022	54749	7.3	3857	3.0	41425	11.2
$<10^{-10}$	0.028	42421	9.8	2998	6.0	33339	16.3
$<5.78\times10^{-12}$	**0.031**	**37403**	**10.7**	**2745**	**7.1**	**29912**	**17.1**
$<10^{-15}$	0.042	27330	12.7	2180	9.5	22591	17.8
$<10^{-20}$	0.057	19655	14.7	1725	12.8	16883	19.2
$<10^{-25}$	0.071	15015	16.4	1429	16.2	13045	21.5
$<10^{-35}$	0.099	9673	17.1	1031	21.6	8516	22.9
$<10^{-50}$	0.140	5873	14.0	712	28.8	5224	21.7
$<10^{-100}$	0.263	1790	10.5	290	28.1	1598	11.1
$<10^{-150}$	0.371	922	5.5	156	21.4	772	5.9
$<10^{-200}$	0.463	635	3.7	97	15.3	504	3.9
$<10^{-300}$	0.606	321	1.7	38	11.7	213	1.7

[§]*Minimum R^2 (proportion of gene expression variability explained by a SNP) observed for a given significance level. Numbers corresponding to study-wise significance are shown in bold. For investigating cis associations or performing any other hypothesis-based test, lower levels of significance may be considered.*

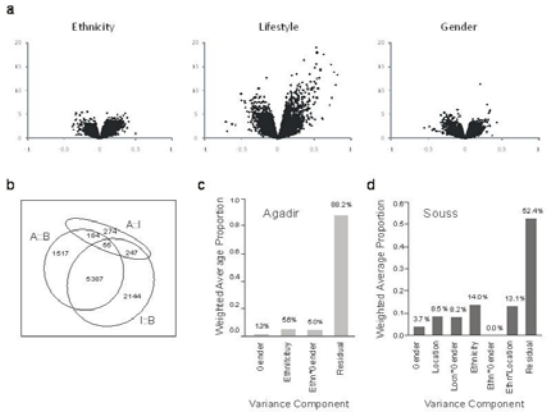
doi:10.1371/journal.pone.0010693.t002

Repeatability with GHS

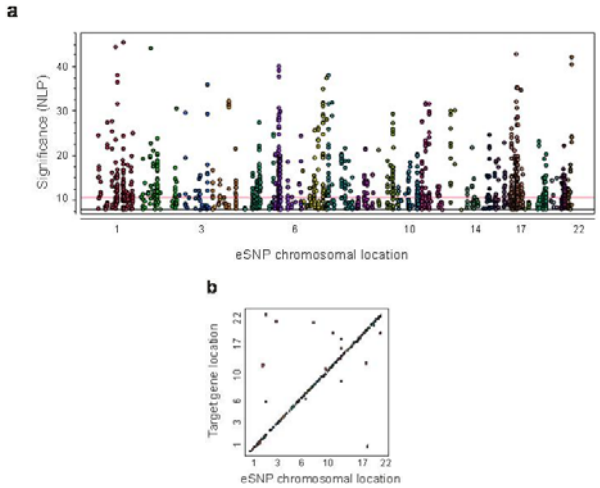
	Stranger et al.		Dixon et al.		Schadt et al.	
Level of signifi-cance	Number of eQTLs at level of significance	Percent signifi-cant in GHS*	Number of eQTLs at level of significance	Percent signif-icant in GHS*	Number of eQTLs at level of significance	Percent signifi-cant in GHS*
$>10^{-6}$	86	55.8	110	50.9	928	47.9
10^{-8}–10^{-10}	63	69.8	162	50.0	168	57.7
10^{-10}–10^{-15}	144	63.2	237	54.8	211	57.3
10^{-15}–10^{-20}	60	70.0	102	60.7	120	66.7
10^{-20}–10^{-25}	38	89.5	73	65.7	73	67.1
$\leq 10^{-25}$	48	70.8	89	67.4	103	73.8
All	439	66.7	773	56.5	1603	54.1

* Comparisons were based on sets of gene expressions overlapping between each study and GHS and were restricted to autosomal cis eQTLs. All cis eQTLs considered significant in each study were retrieved and replication was assessed in GHS ($P<3.9\times10^{-6}$ correcting for 12,808 gene expressions).
For Stranger et al [1], data were extracted from Table S2. We considered as significant the associations found in at least 3 HAPMAP populations. For Dixon et al [2], data were extracted from Table S1 and trans eQTLs were excluded. Matching of probes was done using a table provided by the authors on their web site. For Schadt et al [3], cis eQTLs considered significant (First.Pass.Indicator set to 1) were extracted from Table S3. For each eQTL, we selected in GHS the P-value of the best cis eSNP. The full data used to generate this table are provided in Files S2–S4.
doi:10.1371/journal.pone.0010693.t003

Variance components



eSNP plots



Linear modeling

Simple association:

Expression = μ + SNP + ε

Adjusted for fixed covariates:

Expression = μ + Location + SNP + SNP* Location + ε

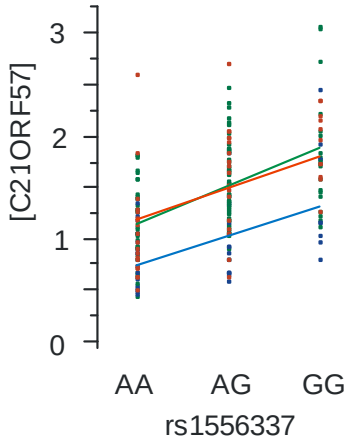
Adjusted for random and/or continuous covariates:

Expression = μ + Relatedness + Ethnicity + SNP + ε

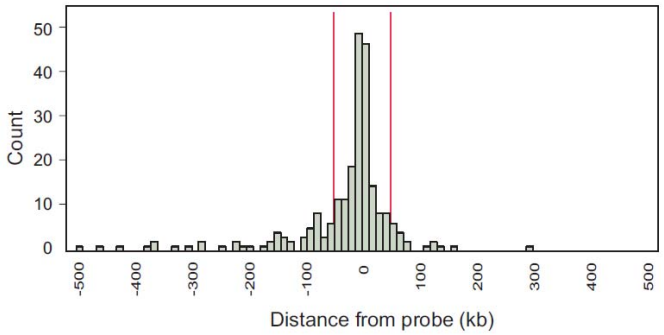
Alternate strategy to control for outliers if MAF < 5%:

Estimate Adjusted Expression Level, then perform SNP association on the rank order of the expression

Additive Genotype & Environment



Location of eSNPs



Effect of Normalization

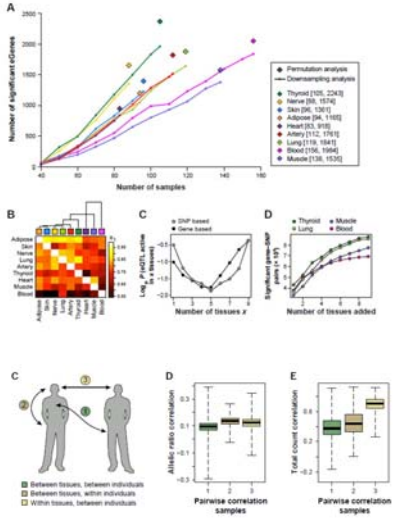
Table 3 **eSNP Analyses**

	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.

GTEx

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

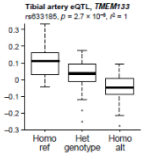


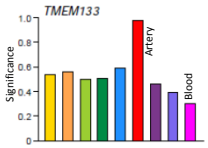
Science (2015) **9(5)**: e1003486

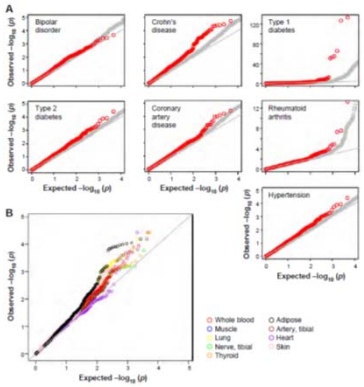
Tissue to trait from GTEx

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

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



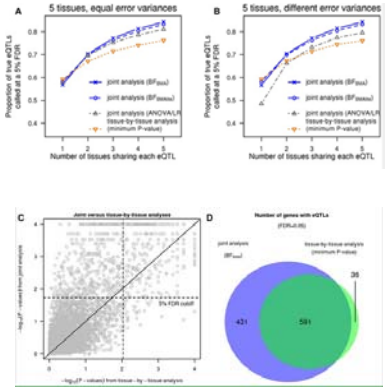




Cross-Tissue Heritability


A Statistical Framework for Joint eQTL Analysis in Multiple Tissues

Abstract



PLoS Genetics (2013) **9(5)**: e1003486

Multiple regression plus function

RESEARCH ARTICLE

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

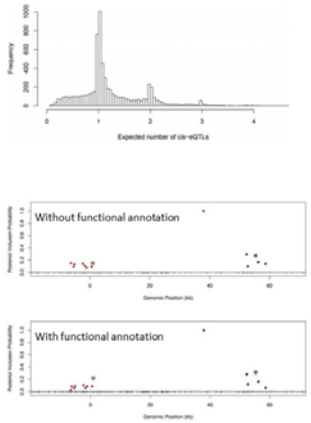
Xiaoquan Wen[1]*, Francesca Luca[2,3], Roger Pique-Regi[2,4]*

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}).



PLoS Genetics (2015) **11(4)**: e1005176

Some other software

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

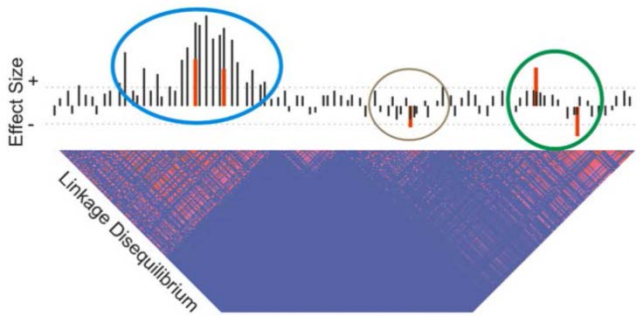
PLINK: The basic tool for GWAS
http://pngu.mgh.harvard.edu/~purcell/plink/tutorial.shtml

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

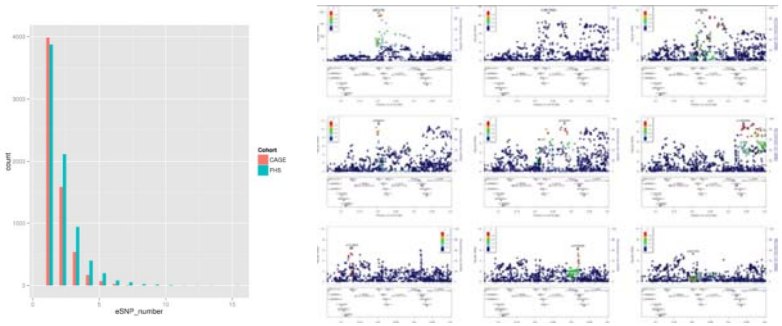
GEMMA: Genome-wide Efficient Mixed Model Association (GEMMA)
http://stephenslab.uchicago.edu/software.html#gemma

Etc etc

Secondary associations

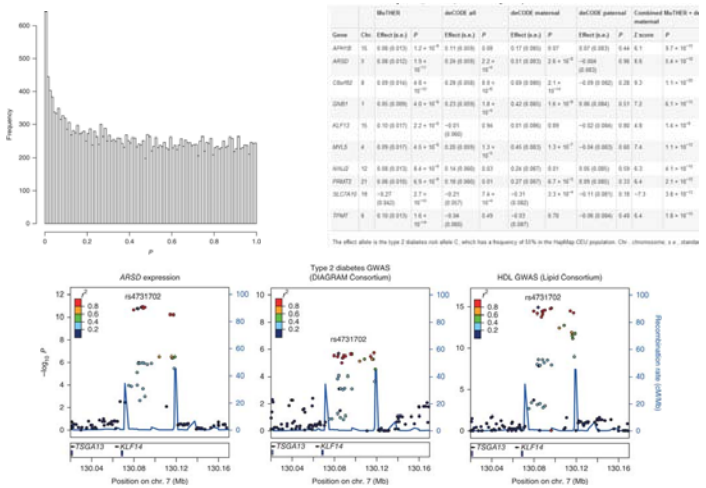


Multi-site regulation is common



Trans-Effect of KLF14

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



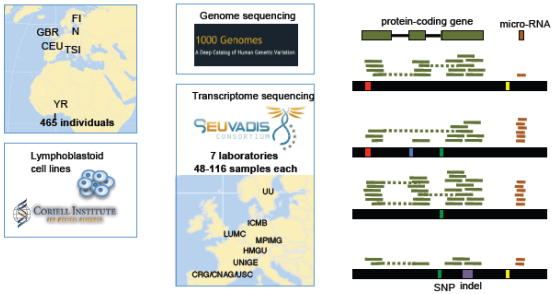
Challenges for eSNP analysis

- 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:

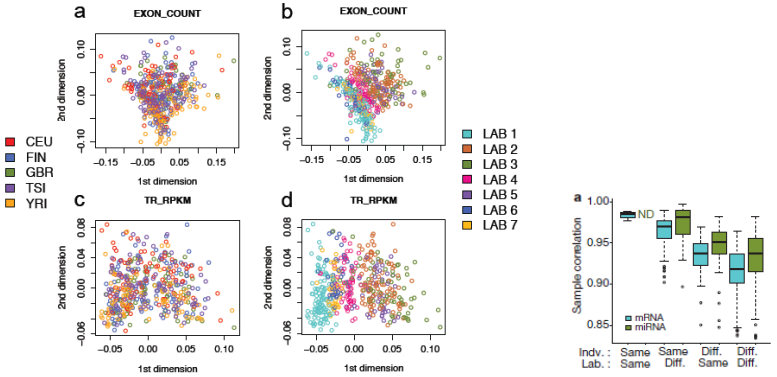
Lappalainen, Dermitzakis et al (2013) *Nature* **501**: 506-511

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



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

eSNP analysis by RNA-Seq

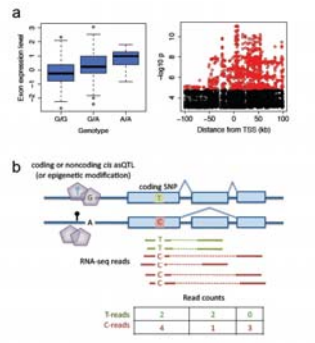
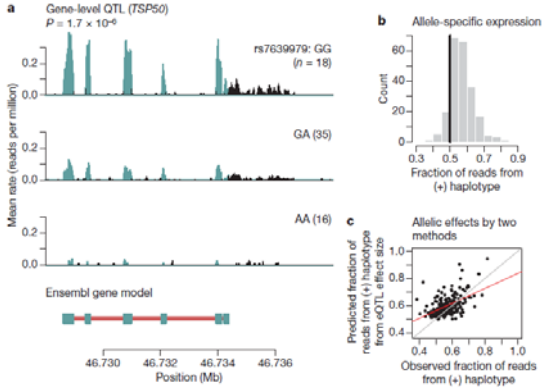


Table 1 | Numbers of transcriptome features with a QTL (FDR 5%)

	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

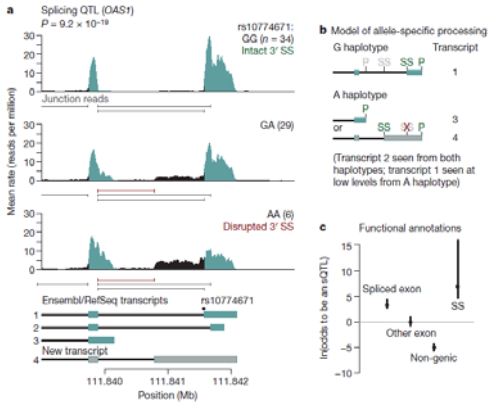
Lappalainen, Dermitzakis et al (2013) *Nature* **501**: 506-511

eQTL and Additivity



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

sQTL (Splicing QTL)



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

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