eQTLs and sQTLs….

Robert Gentleman
Sahar Mozaffari
Rob Tunney
other members of Computational Biology at 23andMe

Outline

- why is this interesting…
- what is an eQTL or sQTL
- how do we find them
- what are the technical issues
- what are the inferential issues

A classic DE experiment

- take some number of samples with a disease and some number without out (nD, nCtrl)
- obtain RNA-seq (or similar, could be protein levels or methylation or ….)
- then perform an Differential Expression analysis
- find differentially expressed genes and then try to understand which are causal and which are consequences
- in practice most (sometimes almost all) DE genes are not causal and hence will not be good targets for therapeutic intervention
- also not likely biomarkers in the sense that they are not necessarily predictive of long term consequences

DE experiment

- sample sizes tend to be small
- it is not clear how representative of all people with the disease the sample is
- many of the DE genes are consequences
  - eg a TF is expressed in a tissue it should be silent
  - eg in cancer large genomic rearrangements yield many correlated changes, some small number may be causal – the remainder are passengers

A different approach

- in many cases someone has performed a genome wide association study for the disease
- the idea behind a GWAS is to identify genetic variants that associate with the disease
- the GWAS variant is more likely to be causal since genome comes first and phenotype temporally second
- sample sizes tend to be large (eg UK Biobank has about 500K people and provides useful GWAS for many diseases)
- the problem with a GWAS is that the variant we identify may not be the causal variant

GWAS

- so, what we want to do is to try to find some functional variant that is in high linkage disequilibrium with the tag variant
- simple examples: the tag variant is a coding variant in some gene
  - CFTR for cystic fibrosis
  - BRCA1/BRCA2 for breast cancer
- some disease causing variants mediate their affect through changes in protein abundance
  - mRNA is a easy to measure surrogate for protein abundance

The two approaches are complementary not competitive

- both differential expression experiments and GWAS can help identify likely causal genes
- each has their strengths and weaknesses and the better we are able to combine them the easier it will be to identify novel regulatory genes that may be useful as drug targets or as biomarkers for therapy …
- there is no good reason you could not extend either to use features/approaches from the other

The Human Genome

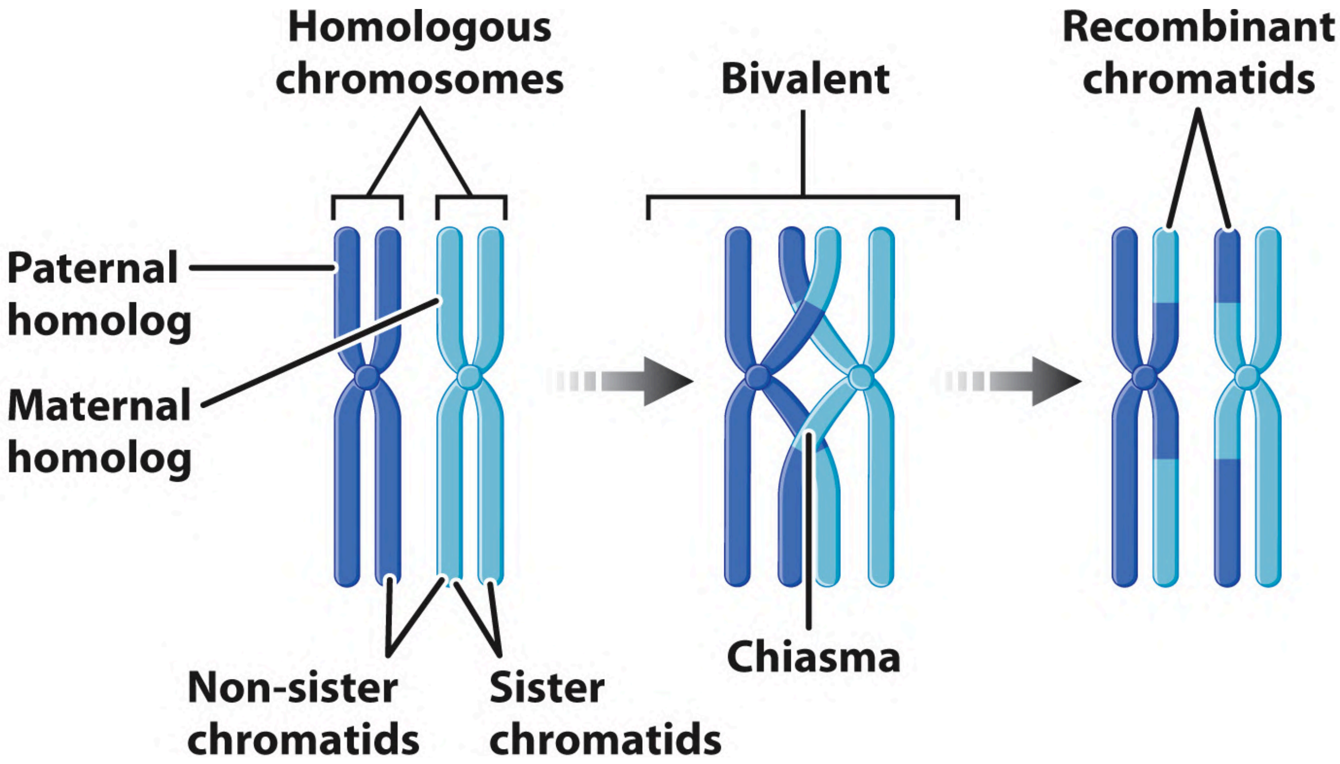
- approximately 3.5 billion nucleotides in a single copy of the human genome
- we can sequence the genome and attempt to measure it at every location
  - the cost of Whole Genome Sequencing (WGS) is about $1500/per person (depends on depth and scale)
- we can genotype individuals and then impute
  - the cost of using a genotyping array is less than $50/per person (~700K variants); total cost about $100/pp

Genetics Primer

- the human genome is encoded on 22 autosomes (each of us has a pair) and the sex chromosomes (X and Y) – making 23 *pairs*
- it is about 3.5B nucleotides long (ACTG)
- variation in the sequence of the genome is associated with human disease
  - but finding the actual cause can be challenging
  - this is referred to as the fine mapping problem

Crossing Over

- primarily occurs during meiosis (creation of gametes)
- two or three events per chromosome per meiosis



http://www.macmillanhighered.com/BrainHoney/Resource/6716/digital_first_content/trunk/test/morris2e/asset/img_ch11/morris2e_ch11_fig_11_09.html

A few complications

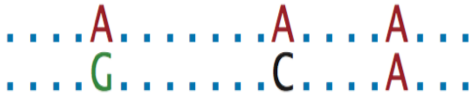
- Linkage disequilibrium (a strong association between nearby variants) causes confounding and makes it hard to identify the likely causal variant
- we don’t really have a perfect reference
  - there is lots of variation that is not yet accounted for in the reference sequence
  - the reference should be population specific
  - we know little about larger (structural) variants
- **phasing**: we have 2 copies of each chromosome, phasing is used to assign a variant to a specific one of the pair

Genotyping at scale

- to genotype at scale a good strategy is to use arrays for most people and **impute**
- basically with imputation we take advantage of the linkage disequilibrium
  - individuals that are identical at a subset of genetic variants will likely be identical in between those variants

Imputation

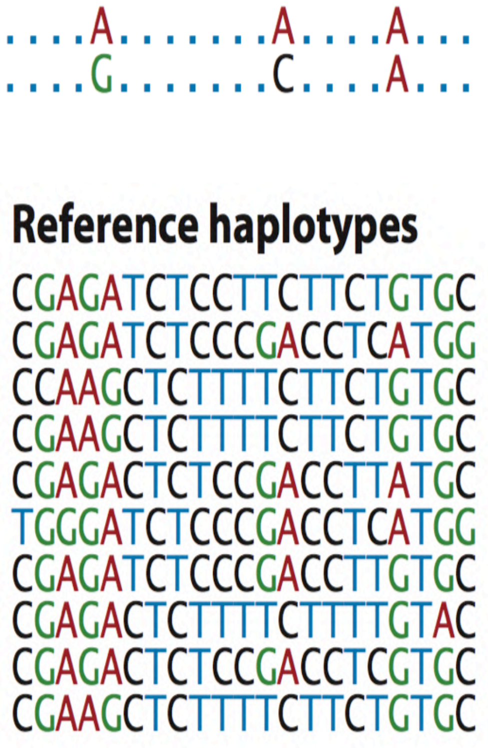
- Our SNP arrays have only ~700k markers on them (sparse compared to the size of the human genome)



Figures from Li *et al.*, Annu Rev Genomics Hum Genet 2009.

Imputation

- Our SNP arrays have only ~700k markers on them (sparse compared to the size of the human genome)



```
....A.......A....A...
....G.......C....A...
```

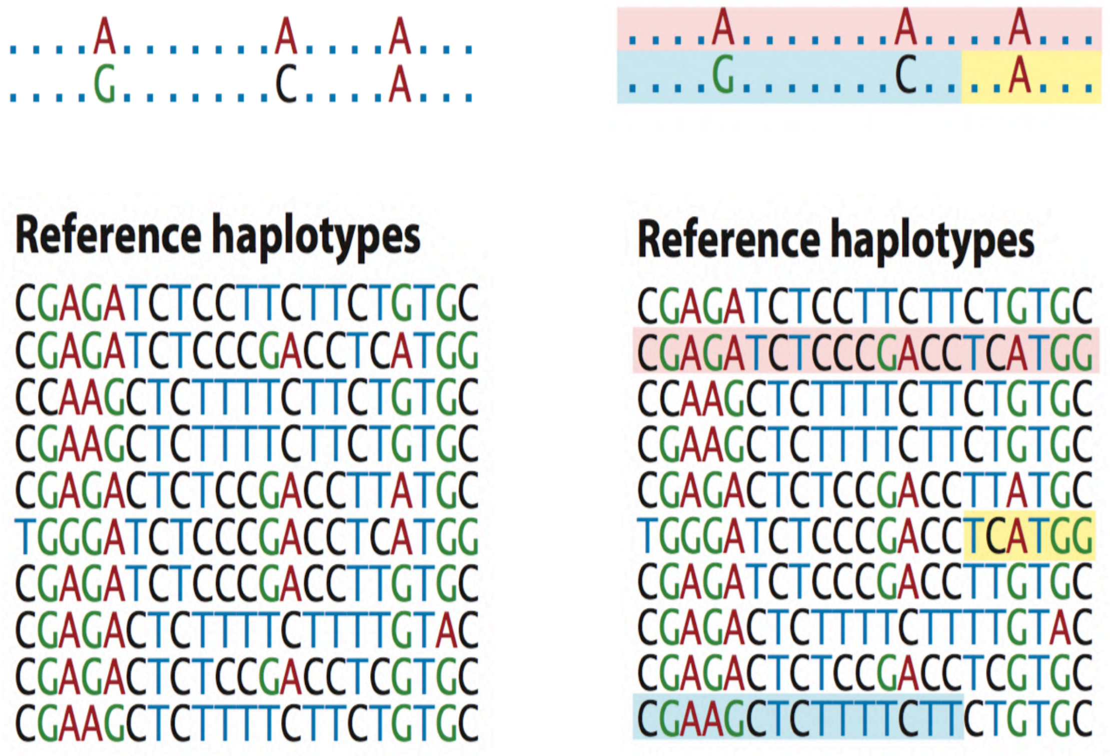
Reference haplotypes

```
CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAAGCTCTTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTTCTTCTGTGC
```

Figures from Li *et al.*, Annu Rev Genomics Hum Genet 2009.

Imputation

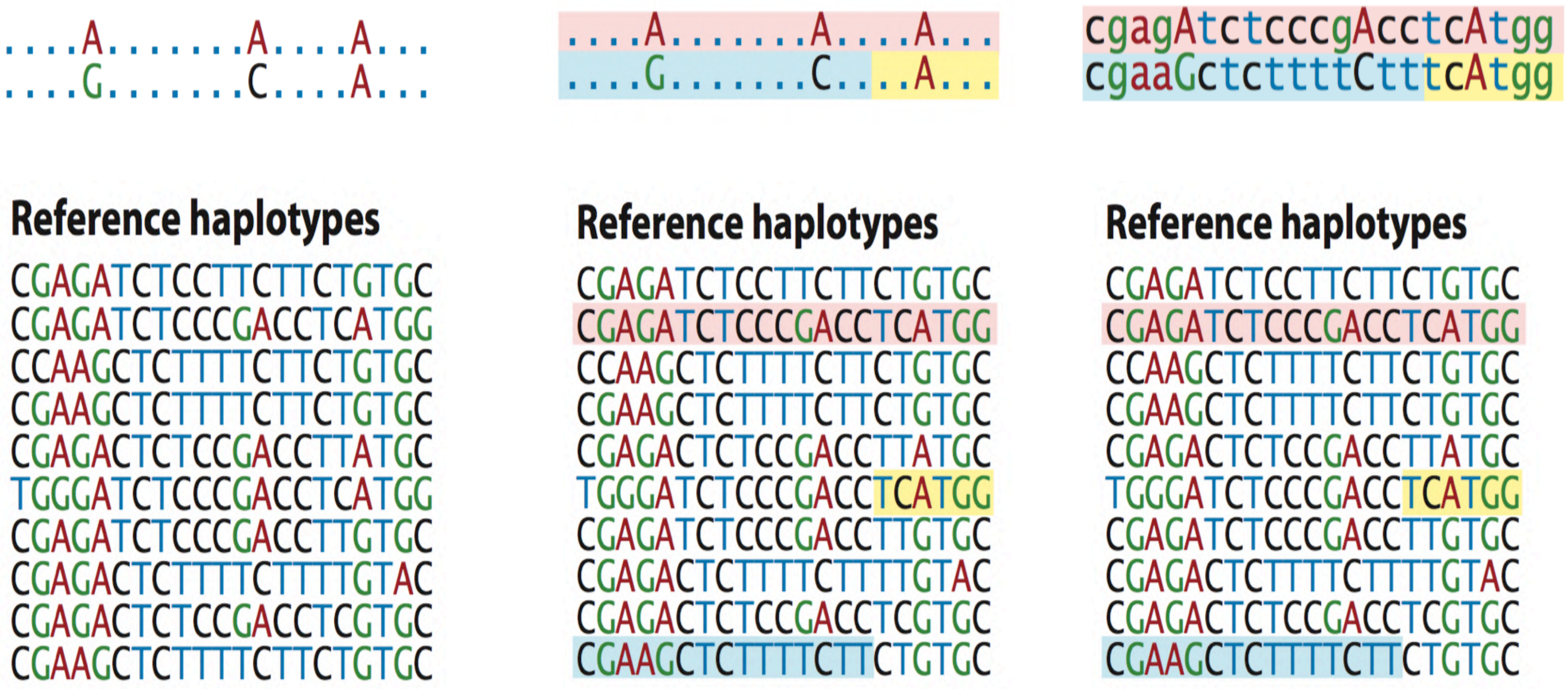
- Our SNP arrays have only ~700k markers on them (sparse compared to the size of the human genome)



Figures from Li *et al.*, Annu Rev Genomics Hum Genet 2009.

Imputation

- Our SNP arrays have only ~700k markers on them (sparse compared to the size of the human genome)



....A.......A....A...
....G.......C....A...

Reference haplotypes

CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAAGCTCTTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTTCTTCTGTGC

....A.......A....A...
....G.......C....A...

Reference haplotypes

CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAAGCTCTTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTTCTTCTGTGC

cgagAtctcccgAcctcAtgg
cgaaGctcttttCtttcAtgg

Reference haplotypes

CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAAGCTCTTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTTCTTCTGTGC

Figures from Li *et al.*, Annu Rev Genomics Hum Genet 2009.

Imputation

- the quality of the imputation depends on the size (number of indviduals) in the reference panel
- and on how well the reference panel matches the population that was genotyped
- we currently impute up to 25M variants accurately
- we are developing reference panels for different ethnicities

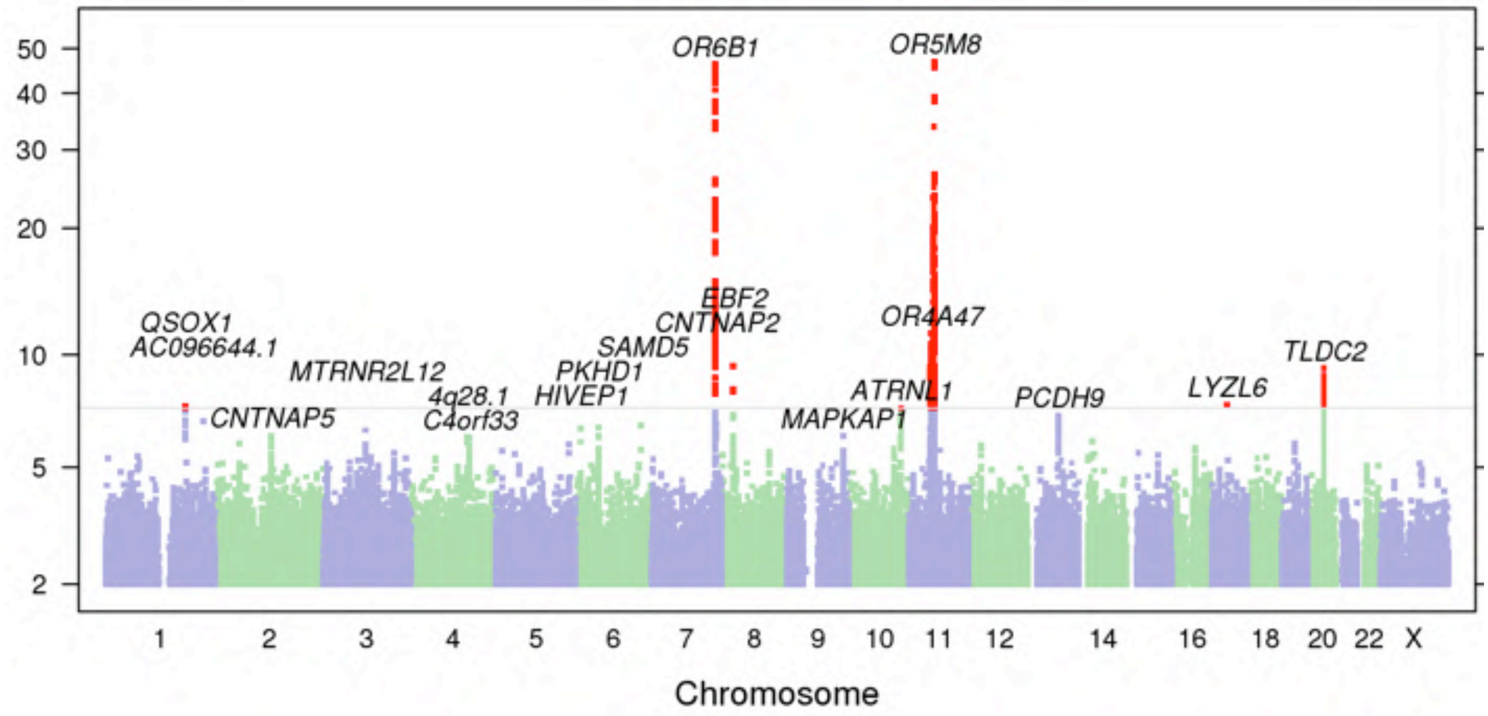
GWAS

- Genome-wide association study
- basically a logistic regression at every locus to associate a **phenotype** (presence/absence) with genetic variation
- we then examine those for which the p-value is less than 5 e-8 (or there-abouts) – which are called **hits**
- the hit indicates an association between variation at the locus and risk of the phenotype/disease

Our genotype affects our characteristics

- what food we like
- how tall we are
- how fat or thin
- what diseases we are susceptible to
- behaviors – risk taking, depression etc.

From the Fun

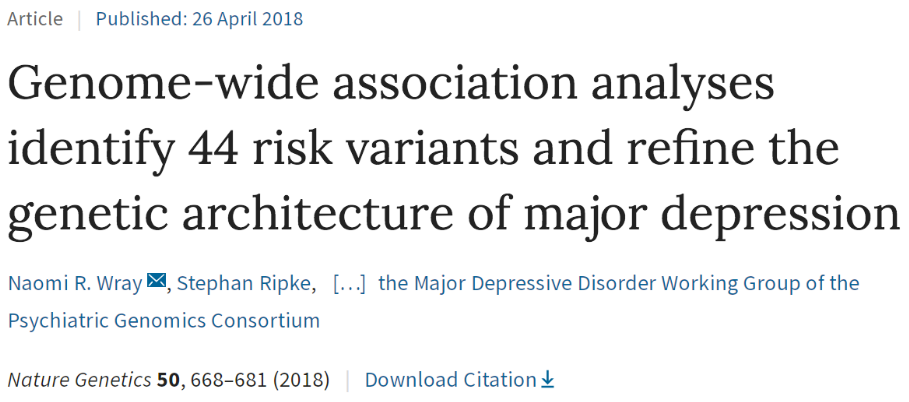


variants that associate with a preference for Strawberry ice cream over vanilla

they are in olfactory receptors

23andMe Blog

To the Serious

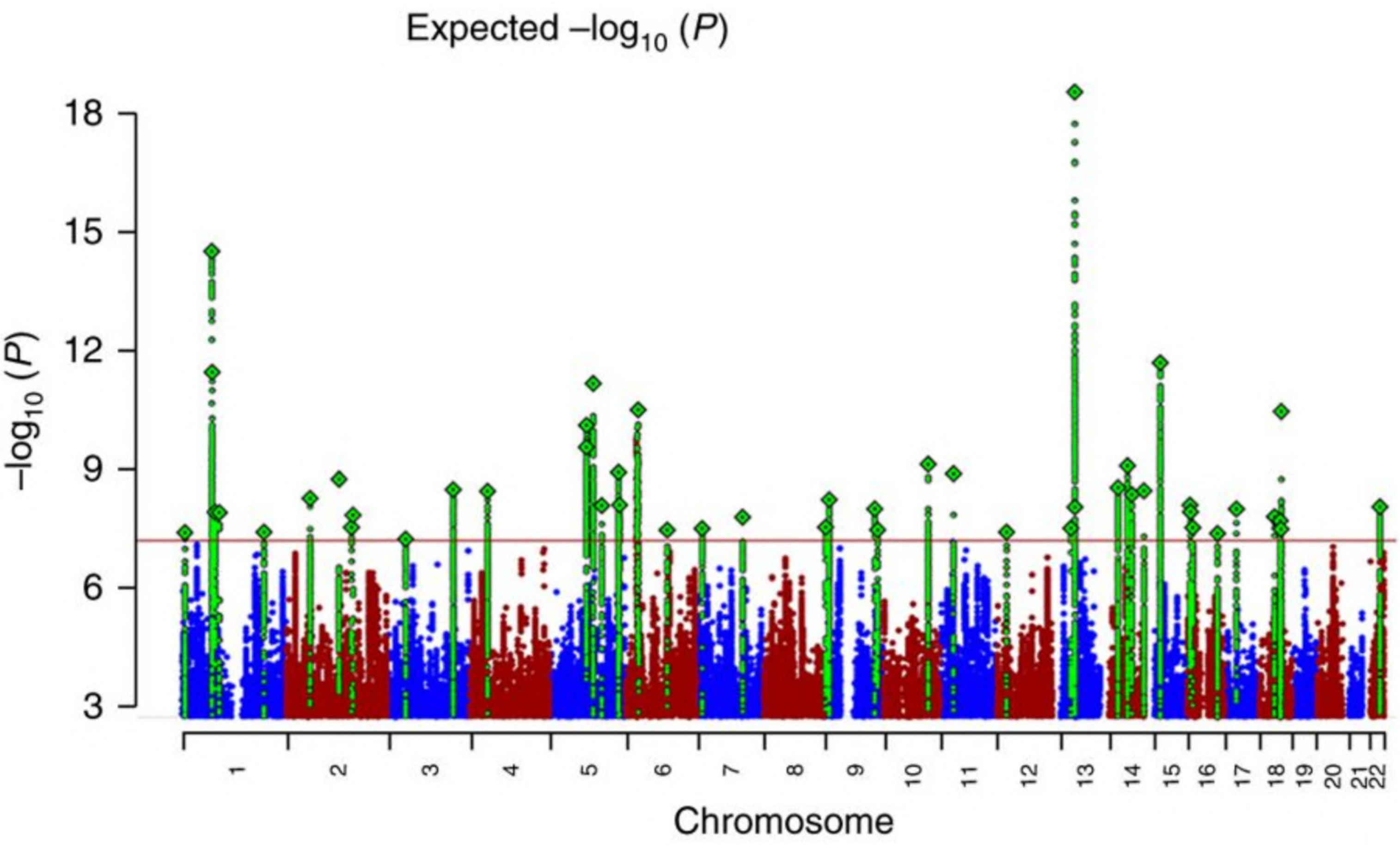


Article | Published: 26 April 2018

Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression

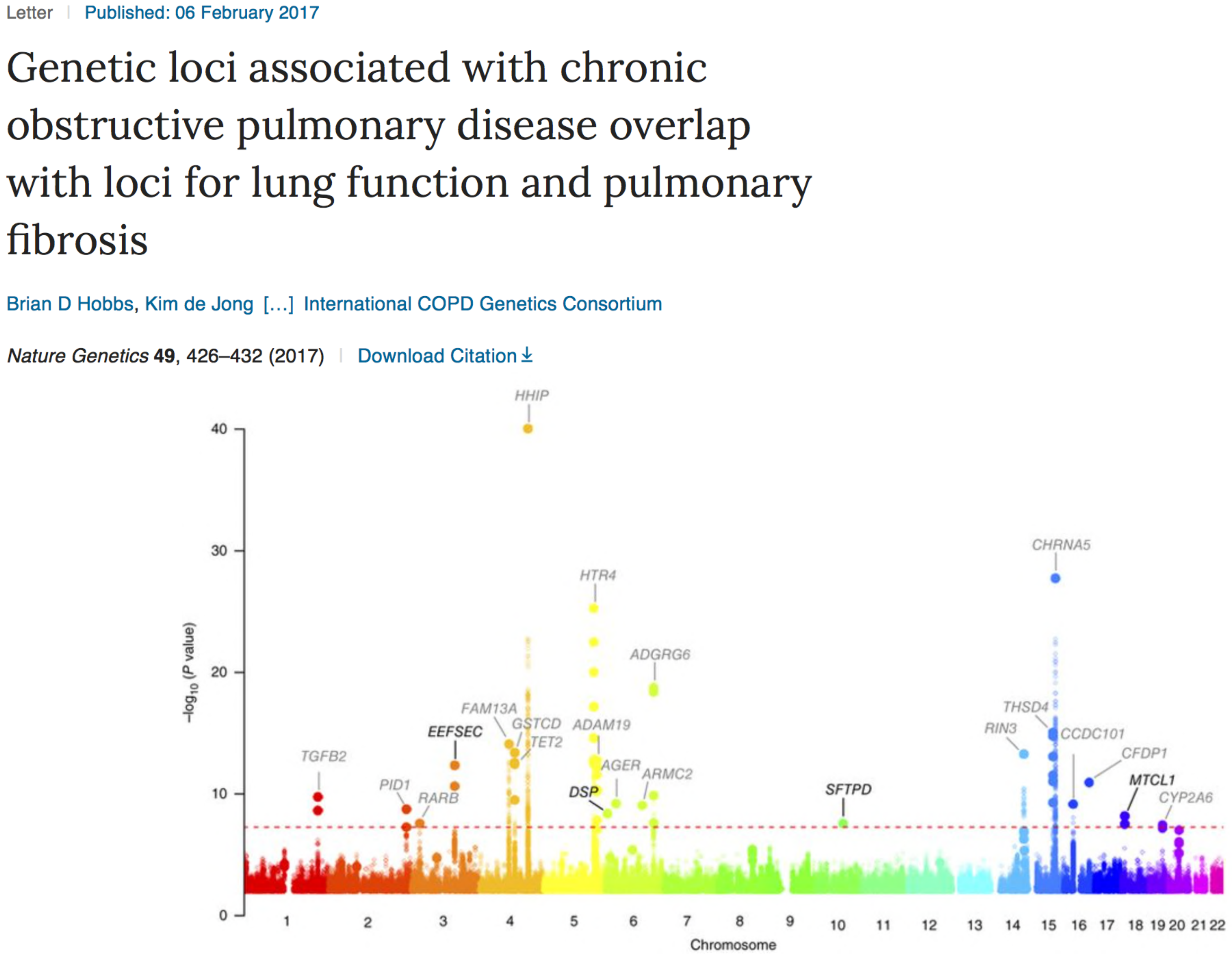
Naomi R. Wray ✉, Stephan Ripke, […] the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium

Nature Genetics **50**, 668–681 (2018) | Download Citation ⤓



Some complexities

- there can be very complex genetic interactions that lead to disease
- there can be gene by environment interactions that affect risk
  - eg risk of smoking and risk of lung diseases (cancer, COPD etc)
    - Spitz MR, Amos CI, Dong Q, Lin J, Wu X. **The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer**. J Natl Cancer Inst2008;100:1552-6. doi:10.1093/jnci/djn363 pmid:18957677



Letter | Published: 06 February 2017

Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis

Brian D Hobbs, Kim de Jong […] International COPD Genetics Consortium

Nature Genetics **49**, 426–432 (2017) | Download Citation

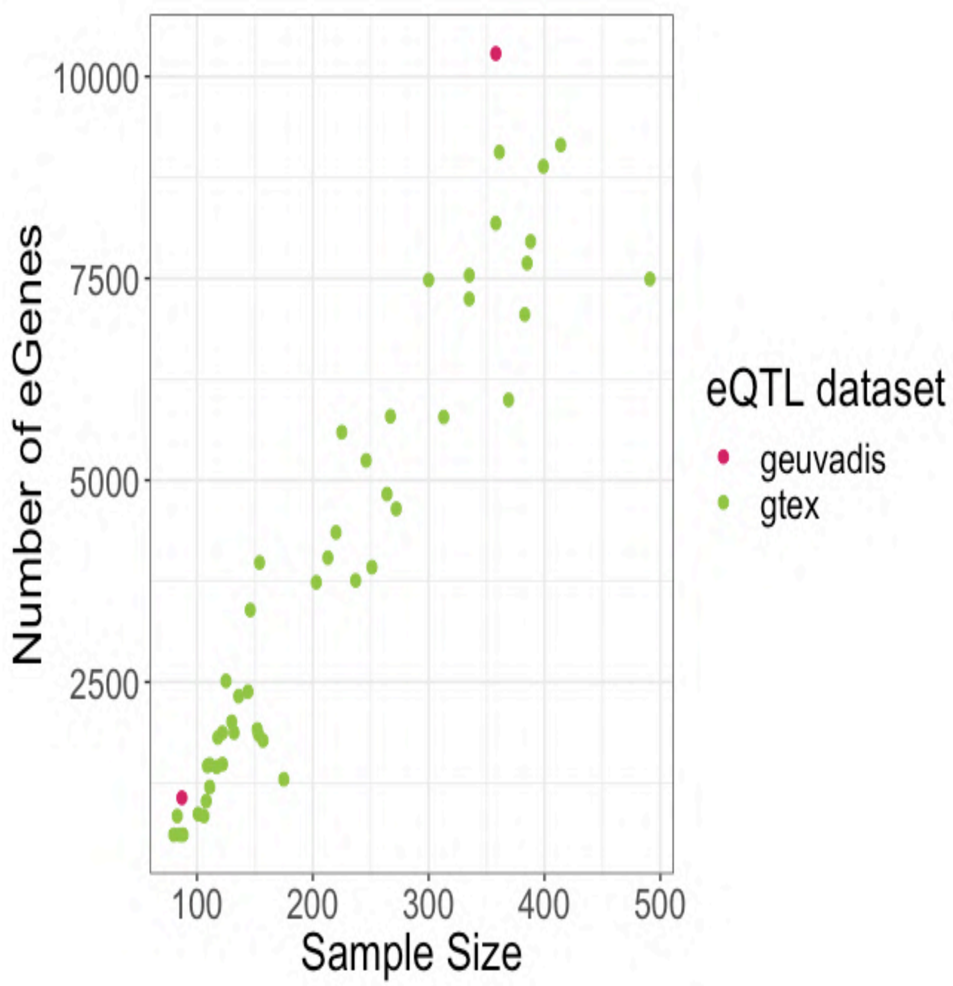
eQTL

- an eqtl is an ***e**xpression* ***Q**uantitative* ***T**rait* ***L**ocus*
- we essentially perform a GWAS using the expression value of the gene as a trait
  - but there are 20K human genes, so this would lead to an amazing amount of computation and multiple testing correction
  - so most people do some form of cis-eQTL analysis using only SNPs within some kMB of the TSS or TSE. Often k = 0.5MB

eQTL

- expression Quantitative Trait Locus
  - a location in the genome where there is polymorphic expression (the nucleotide at that position varies in the population)
  - a gene, whose expression appears to be associated with that variation at some level

It is all about power….



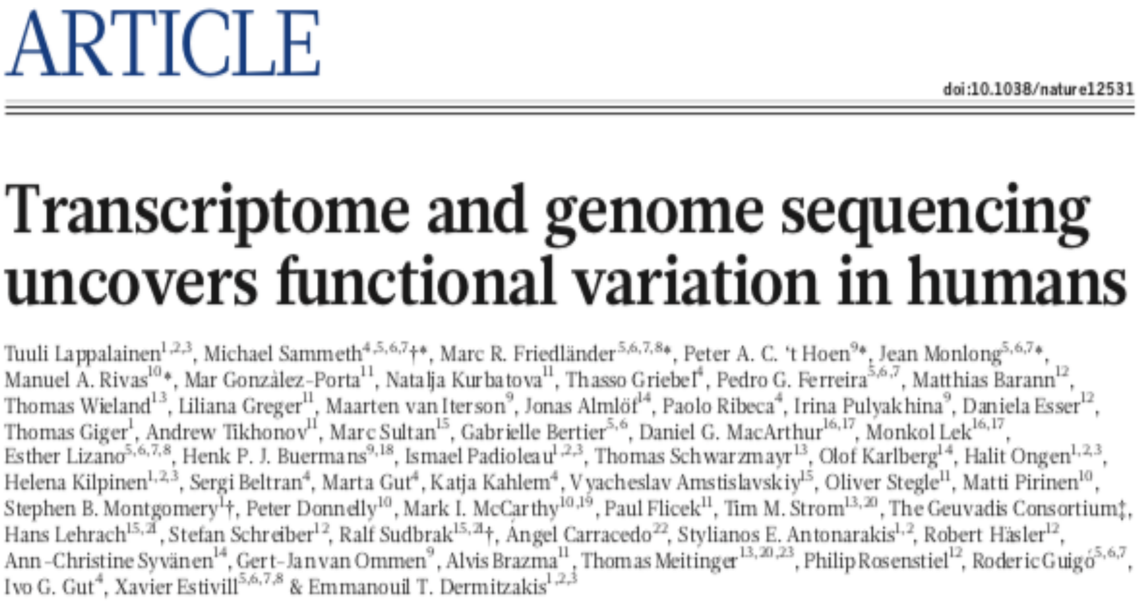
eQTLs in practice

- the actual modeling is complex as there is often a need to correct for:
  - unknown expression batch effects (PEER)
  - unknown population structure (genetic PCs)
  - other known, or possible confounders (often age and sex)
- to date there is little concern with performing conditional analysis and usually just the top hit in some locus is obtained
  - this tends to favor common alleles..due to power

Data requirement

- we need some number of individuals who have been both genotyped and had RNA-seq (or similar) carried out on them
- from the RNA-seq we can compute expression levels (in some units) and assess local structure (eg are some exons skipped)

GEUVADIS



ARTICLE

doi:10.1038/nature12531

Transcriptome and genome sequencing uncovers functional variation in humans

Tuuli Lappalainen[1,2,3], Michael Sammeth[4,5,6,7]†*, Marc R. Friedländer[5,6,7,8]*, Peter A. C. 't Hoen[9]*, Jean Monlong[5,6,7]*, Manuel A. Rivas[10]*, Mar Gonzàlez-Porta[11], Natalja Kurbatova[11], Thasso Griebel[4], Pedro G. Ferreira[5,6,7], Matthias Barann[12], Thomas Wieland[13], Liliana Greger[11], Maarten van Iterson[9], Jonas Almlöf[14], Paolo Ribeca[4], Irina Pulyakhina[9], Daniela Esser[12], Thomas Giger[1], Andrew Tikhonov[11], Marc Sultan[15], Gabrielle Bertier[5,6], Daniel G. MacArthur[16,17], Monkol Lek[16,17], Esther Lizano[5,6,7,8], Henk P. J. Buermans[9,18], Ismael Padioleau[1,2,3], Thomas Schwarzmayr[13], Olof Karlberg[14], Halit Ongen[1,2,3], Helena Kilpinen[1,2,3], Sergi Beltran[4], Marta Gut[4], Katja Kahlem[4], Vyacheslav Amstislavskiy[15], Oliver Stegle[11], Matti Pirinen[10], Stephen B. Montgomery[1]†, Peter Donnelly[10], Mark I. McCarthy[10,19], Paul Flicek[11], Tim M. Strom[13,20], The Geuvadis Consortium‡, Hans Lehrach[15,21], Stefan Schreiber[12], Ralf Sudbrak[15,21]†, Àngel Carracedo[22], Stylianos E. Antonarakis[1,2], Robert Häsler[12], Ann-Christine Syvänen[14], Gert-Jan van Ommen[9], Alvis Brazma[11], Thomas Meitinger[13,20,23], Philip Rosenstiel[12], Roderic Guigó[5,6,7], Ivo G. Gut[4], Xavier Estivill[5,6,7,8] & Emmanouil T. Dermitzakis[1,2,3]

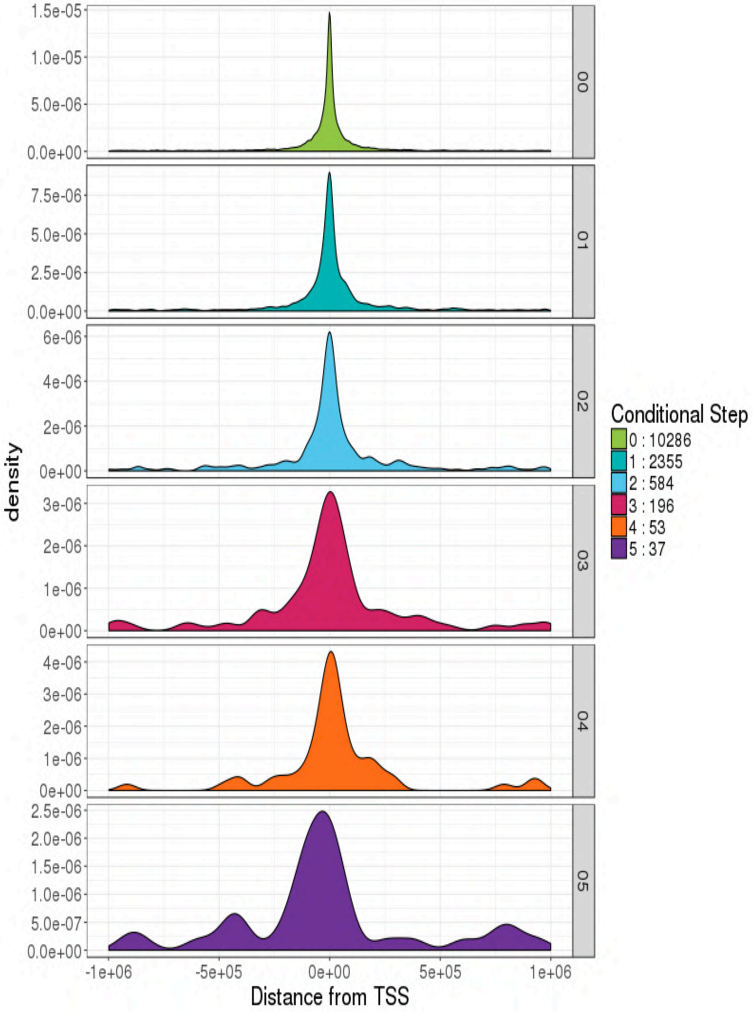
Lappalainen et al. 2013 Nature http://dx.doi.org/10.1038/nature12531

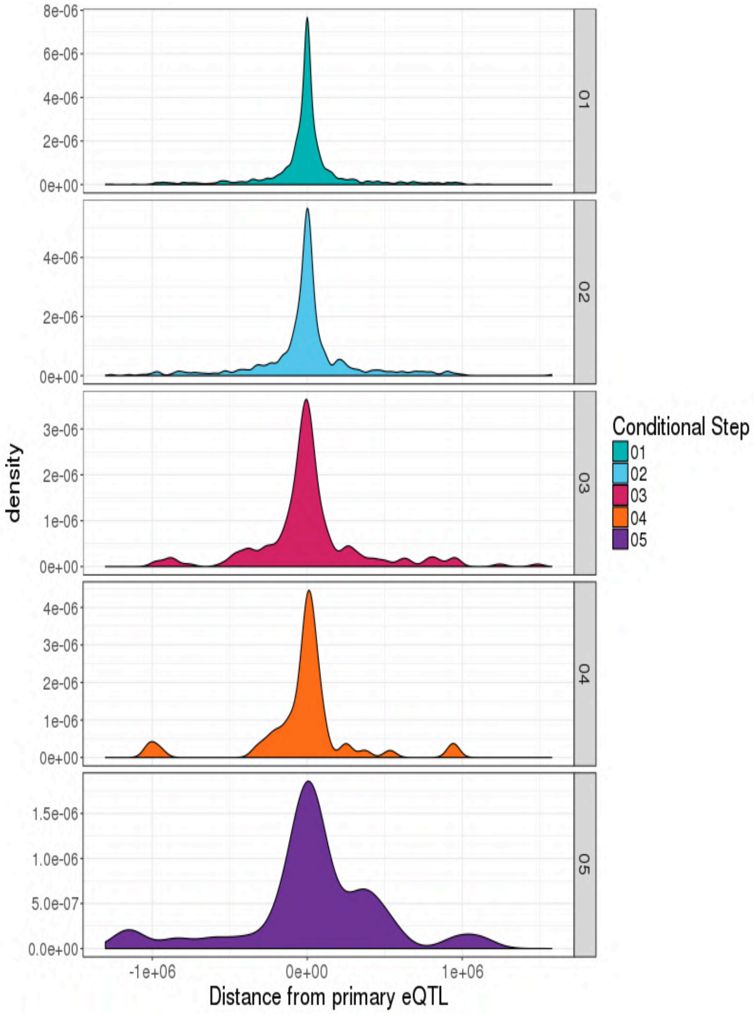
GEUVADIS

- 462 individuals with expression
- 445 pass 1000 Genomes Phase 3 QC
  - 358 EUR
  - 87 YRI
- Across 7 labs

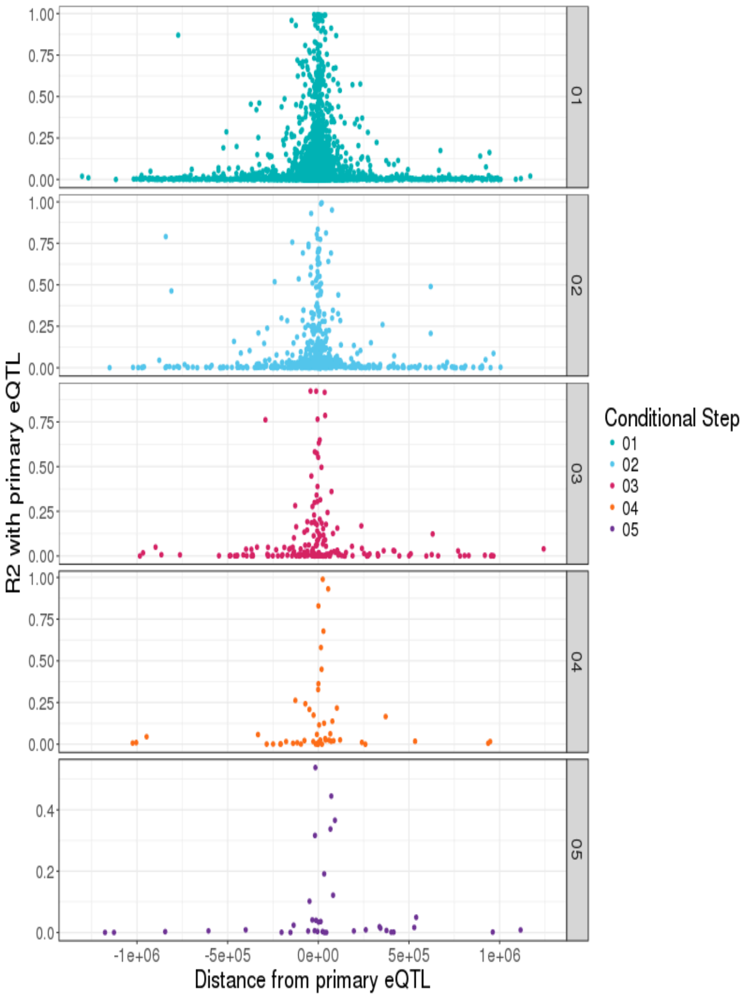
Population	Sample size used
CEU	89
GBR	92
FIN	86
TSI	91
YRI	87

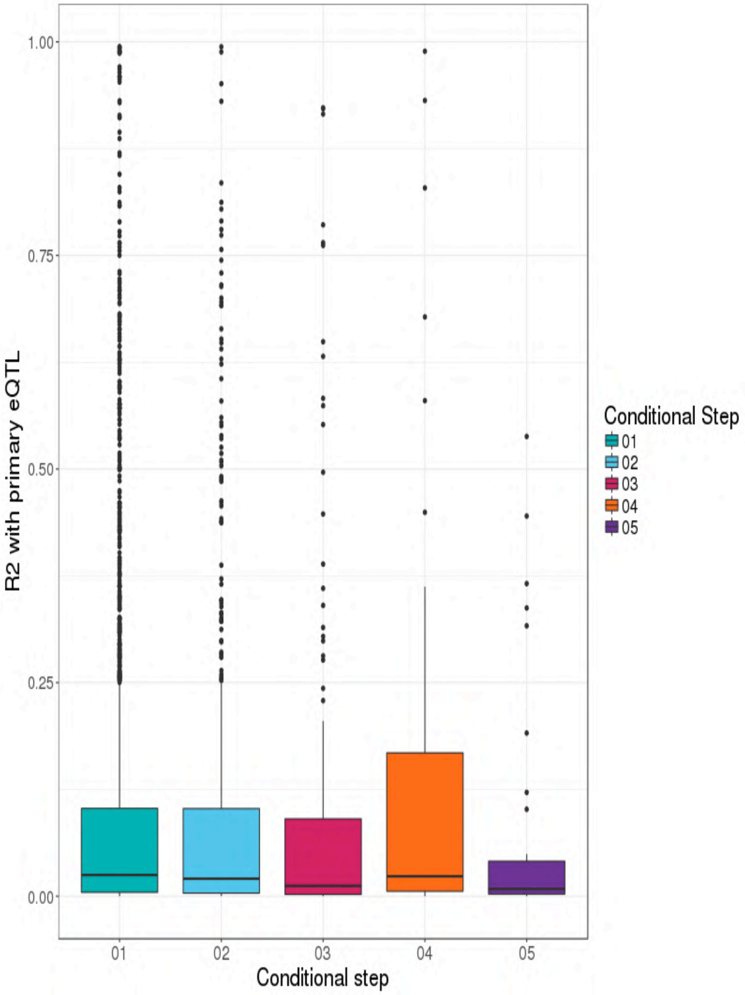
Distance of conditional eQTL from TSS & primary eQTL



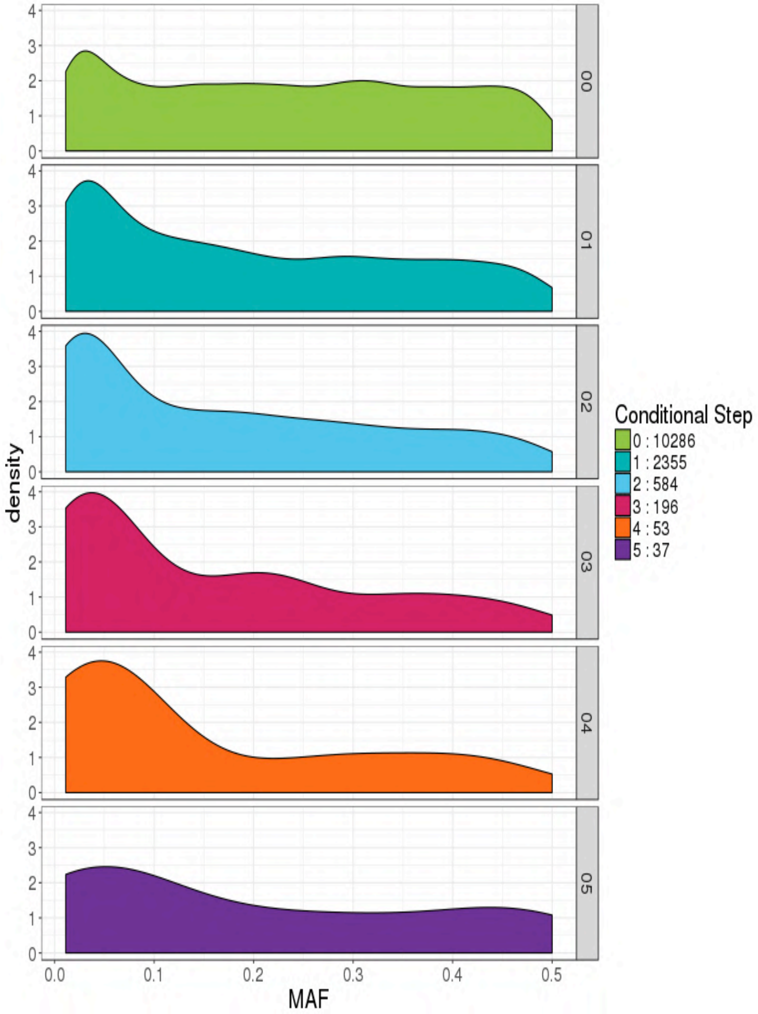


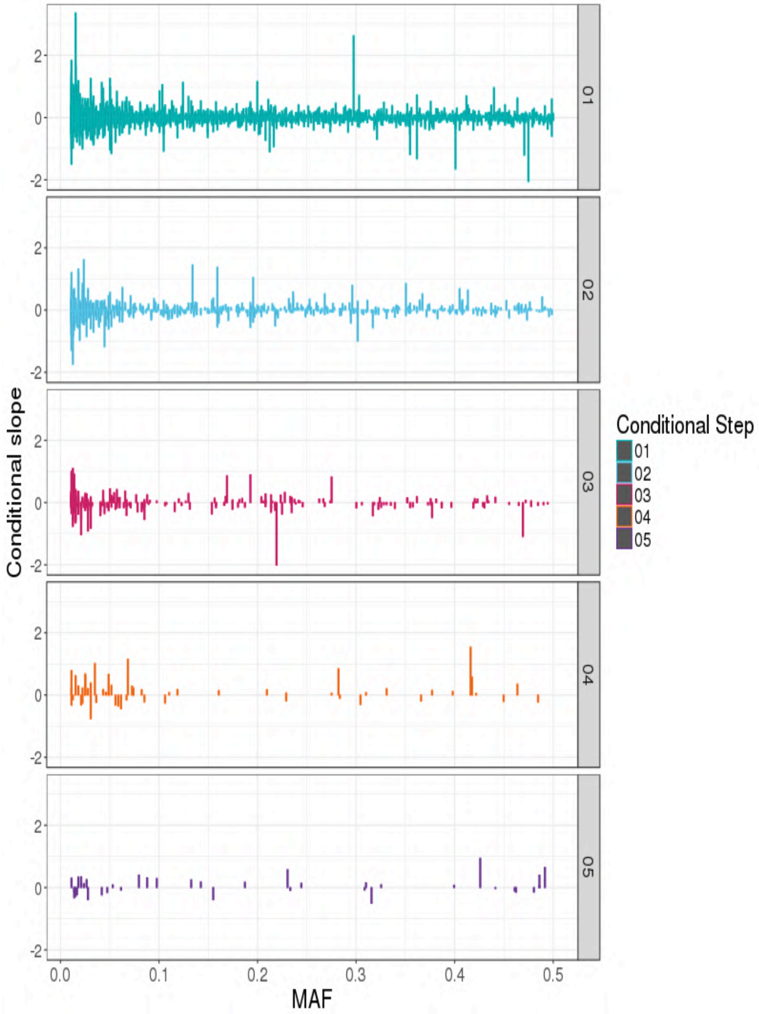
LD R^2 of conditional eQTL with primary eQTL



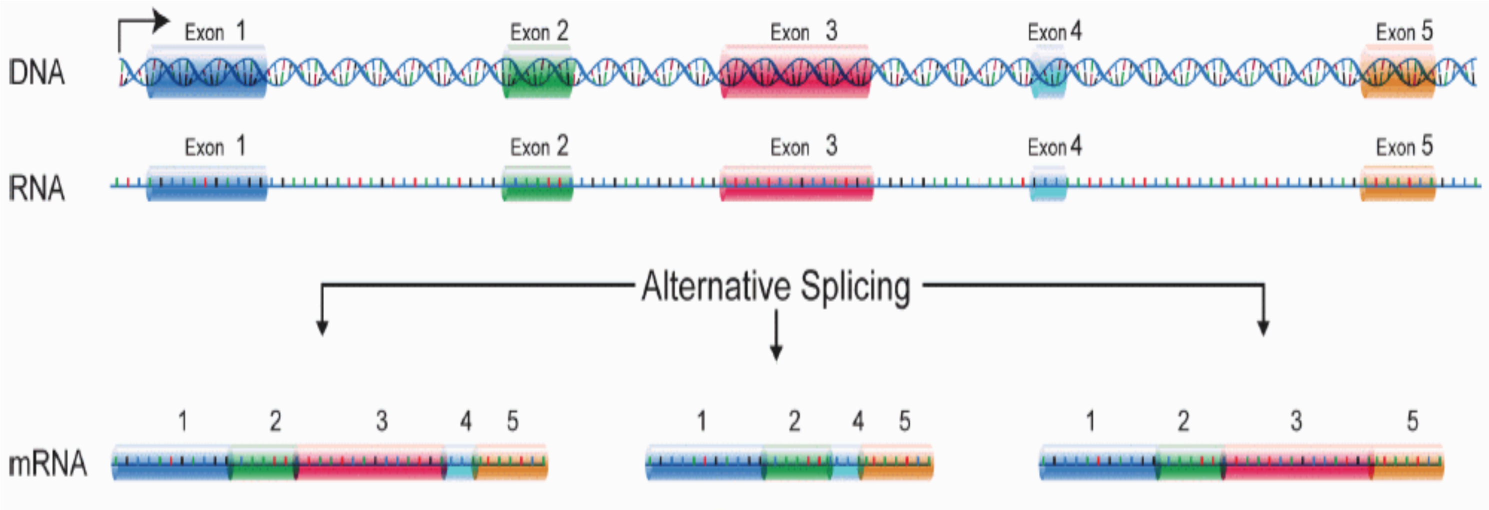


Conditional eQTL MAF & estimated effect size





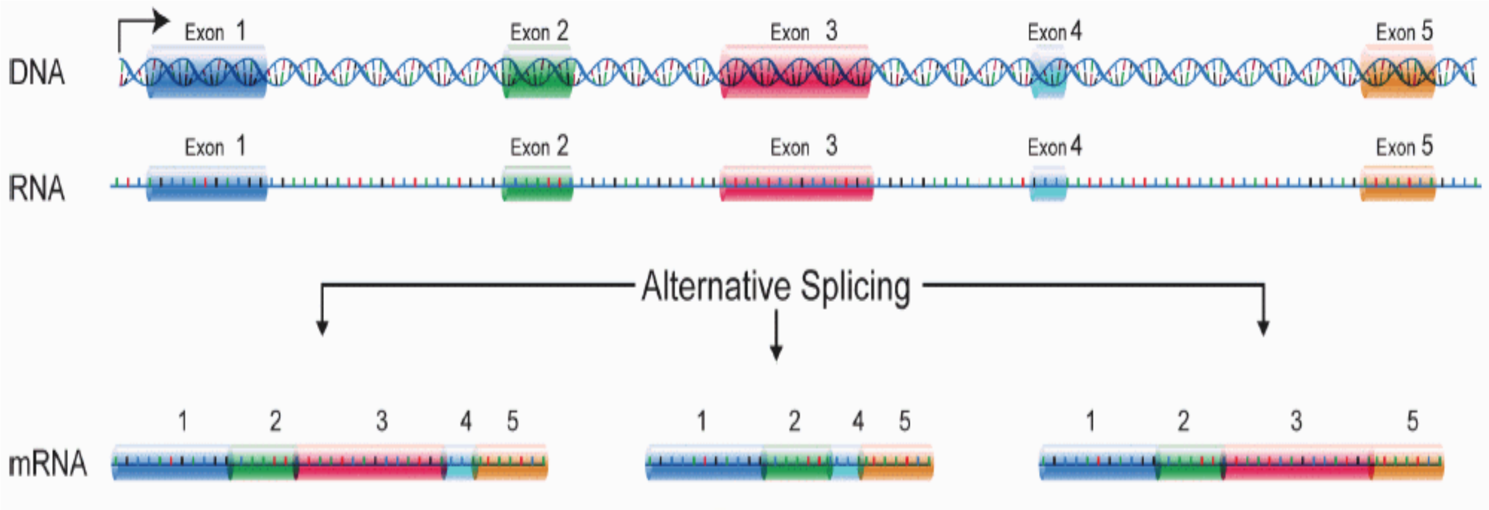
sQTL analysis can compute effects on isoform abundances...



X% Y% Z%

NCBI: http://www.genome.gov/Images/EdKit/bio2j_large.gif

...or on the splicing frequency of individual exons



X%

Y%

Z%

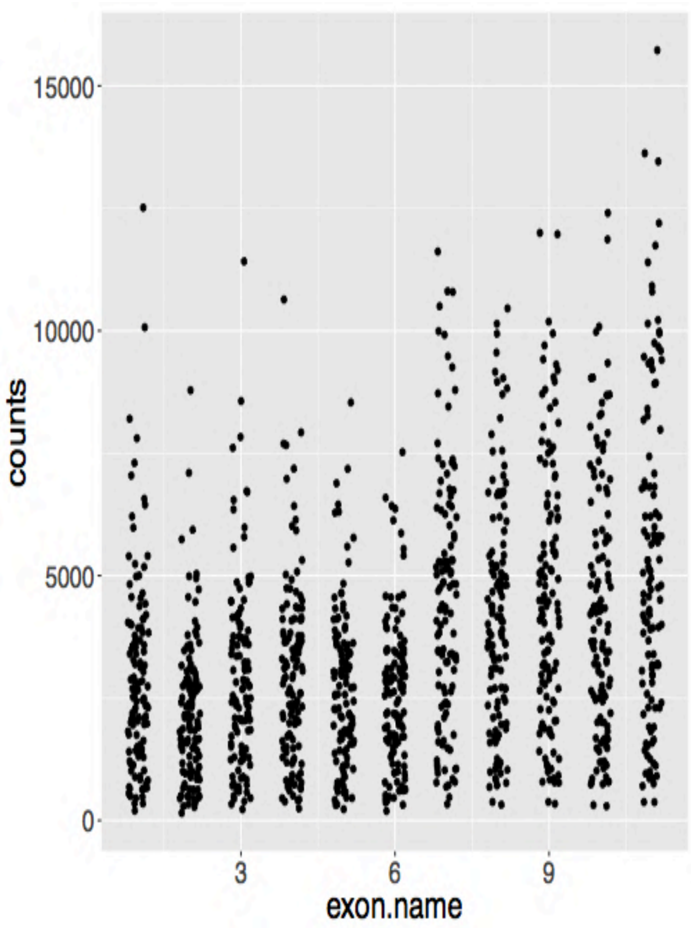
Included:
X+Z%
Excluded:
Y%

NCBI: http://www.genome.gov/Images/EdKit/bio2j_large.gif

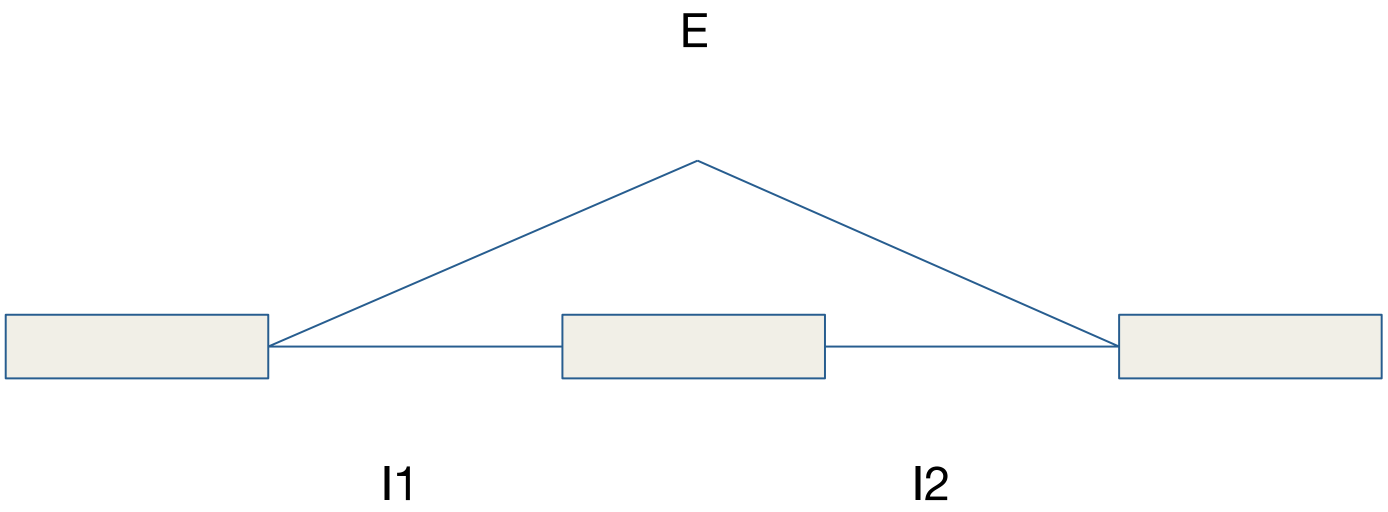
Exon Level sQTL analysis

1. Less sensitive to 3′ recovery bias
2. Less computationally complex
   a. 2 vs. 2^n outcomes
3. Interpretable for TX
   a. How often is a domain spliced in?

SLC221A Read Counts vs. Exon Position

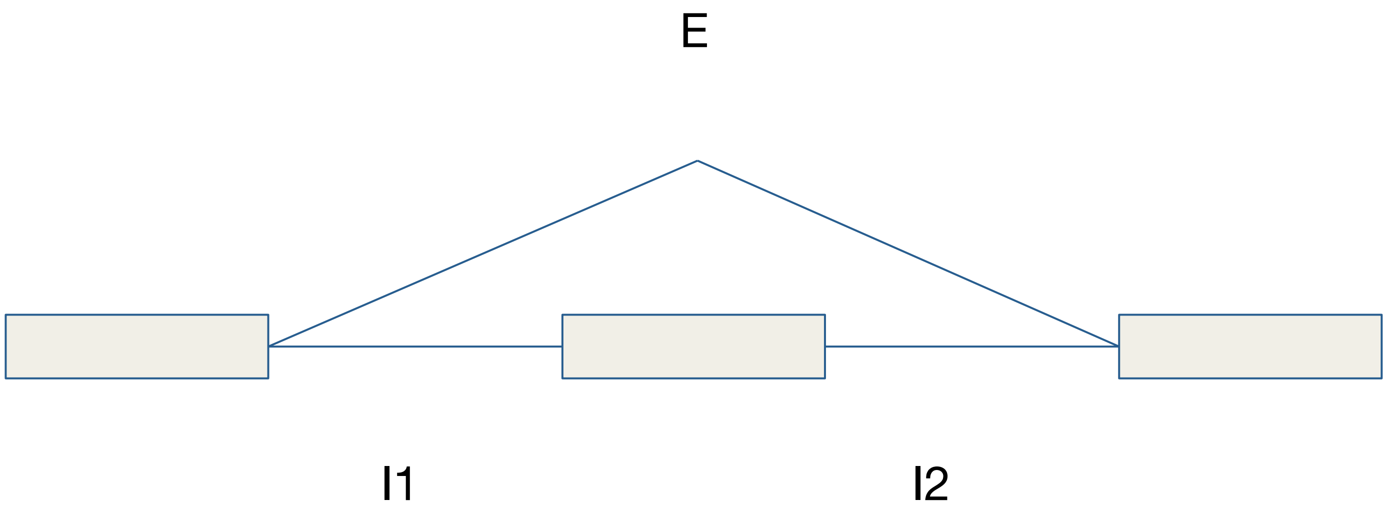


Cassette exons are simple alternative splicing events



Use junction reads to assess splicing ratios

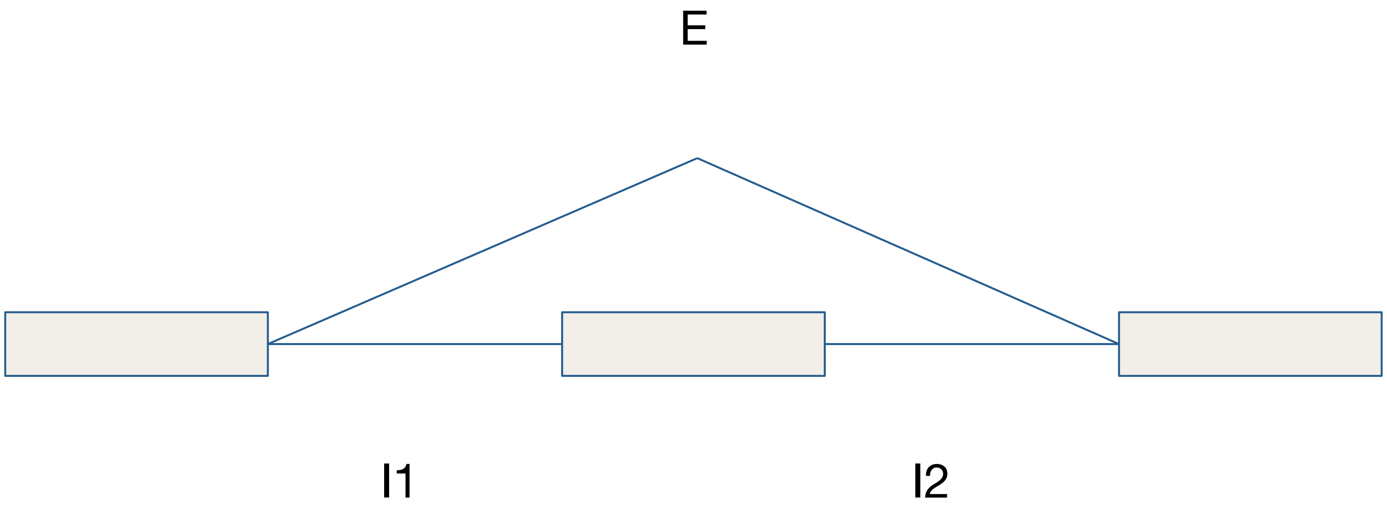
Exclusion counts



Inclusion counts = 0.5 * (I1 + I2)

Use junction reads to assess splicing ratios

Exclusion counts



Inclusion counts = 0.5 * (I1 + I2)

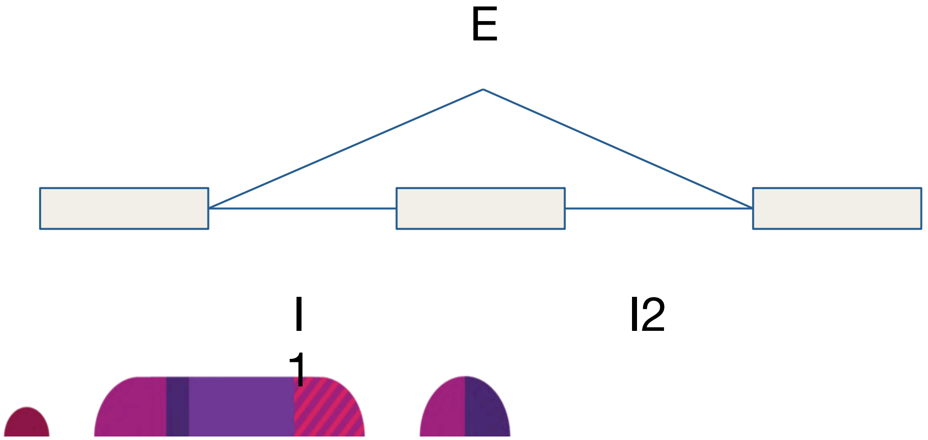
Percent spliced in (ψ) = Incl. / (Incl. + Excl.)

Splicing Data and Modeling

1. Each event has 2 counts: inclusion, exclusion
2. How to associate genotype with these counts?
   a. Common: compute PSI, pass to FastQTL
   b. GLM on one count, using total counts as offset term

3. Inclusion/exclusion controls for gene expression

$$\log(E[x_i]) = \log(N_i) + \beta_c^T c + \beta_g g$$



sQTL Analysis - Overview

1. Data
    a. 114 GTEx Liver samples
2. Splicing quantification
    a. Spliced RNA-Seq alignment with STAR
    b. Annotated cassette exons from VastDB
    c. Compute inclusion/exclusion from junction reads
3. sQTL association
    a. Test each exon for association with cis-SNPs
    b. SNPs in window 20 kb 5′ and 3′ of exon

sQTL Association Testing - Model

1. Negative binomial regression (glm.nb in R)
2. For each cassette exon, for sample i:
    a. x_i = inclusion counts
    b. N_i = inclusion counts + exclusion counts
3. Covariates: age, sex, WGS platform, surgical/postmortem, 5 genetic PCs

$$\log(E[x_i]) = \log(N_i) + \beta_c^T c + \beta_g g$$

$$\mathrm{Var}(x_i) = E[x_i] + \frac{1}{\theta} E[x_i]^2$$

sQTL Association Testing - Model

1. Data requirements to test exons
   a. ≥10 junction reads in ≥40 samples
   b. ≥2% minor allele in sample
   c. ≥10% samples with alternative splicing

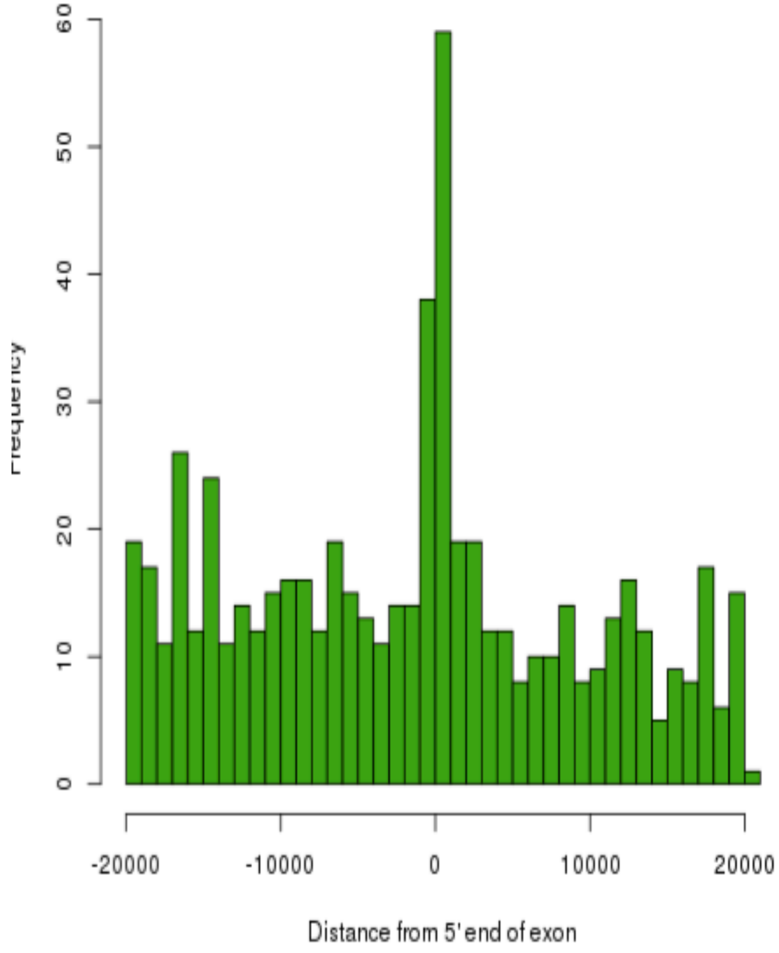
2. Test exons for overdispersion w.r.t poisson regression model
   a. Overdispersed -> NB regression
   b. Not overdispersed -> Poisson regression

$$\mathrm{Var}(x_i) = E[x_i] + \frac{1}{\theta} E[x_i]^2$$

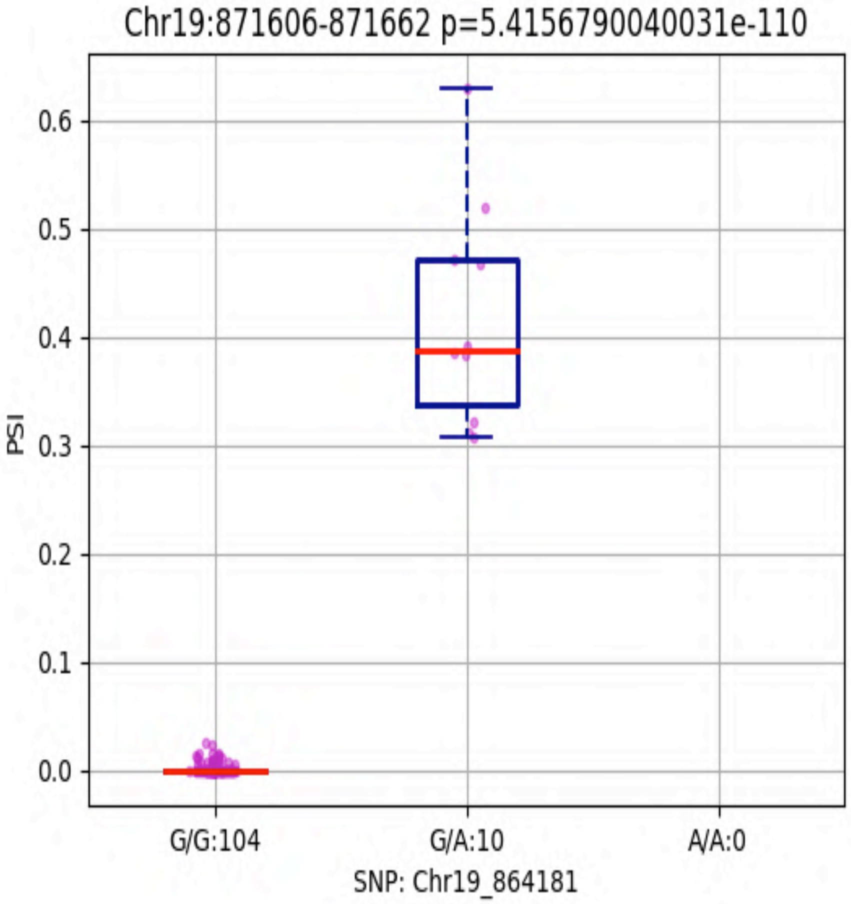
sQTL Results

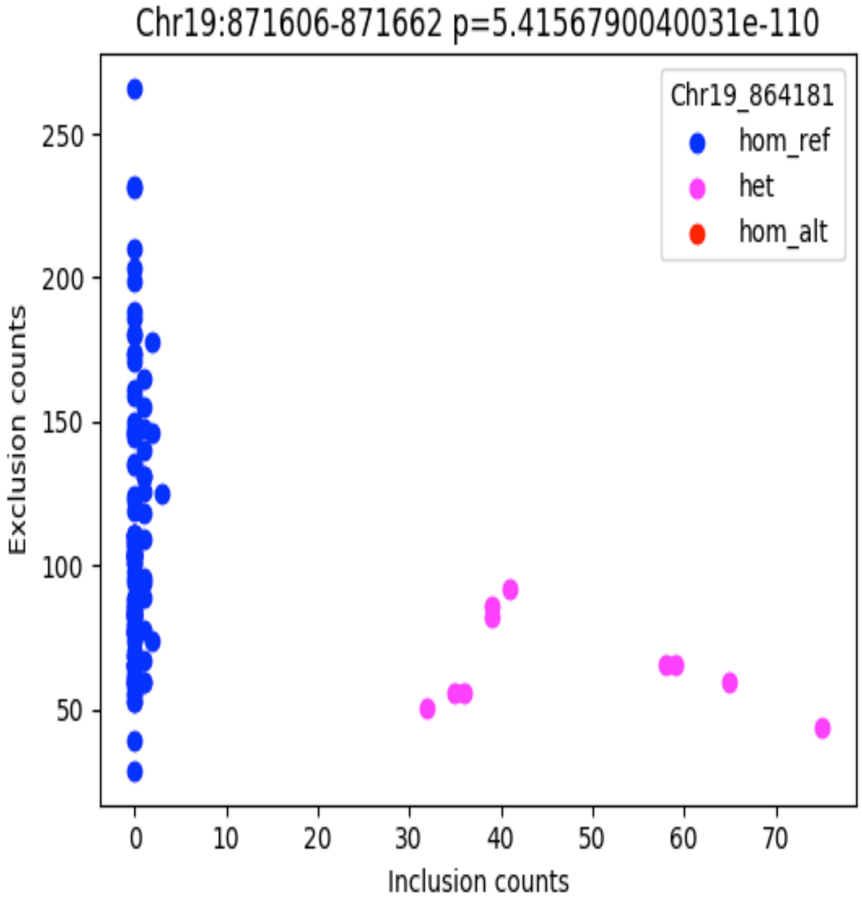
GTEx Liver sQTL results

1. 17355 candidate exons
2. 8723 exons pass data filters
3. 611 exons with ≥1 sQTL hit

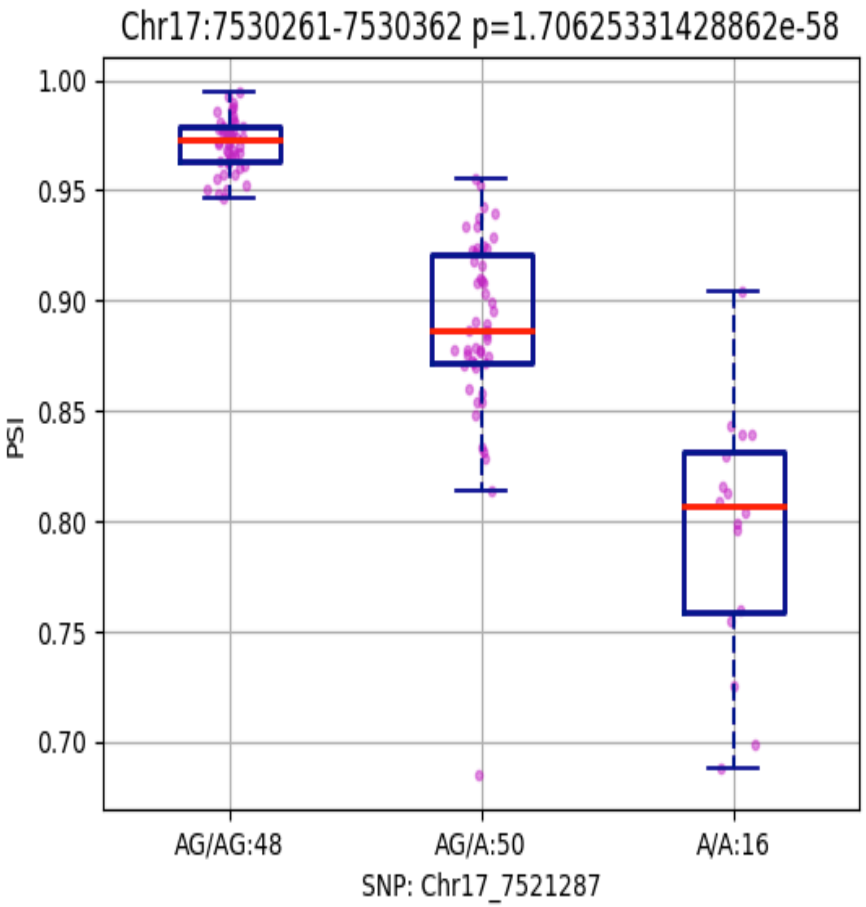


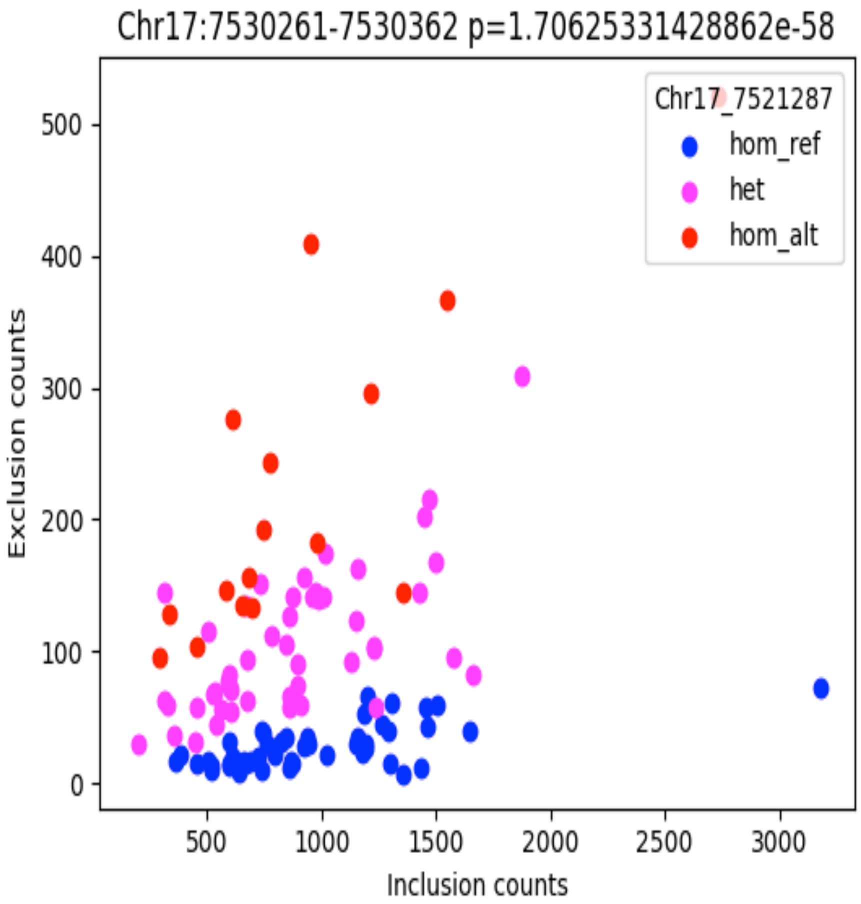
Top Hits


Chr19:871606-871662 p=5.4156790040031e-110

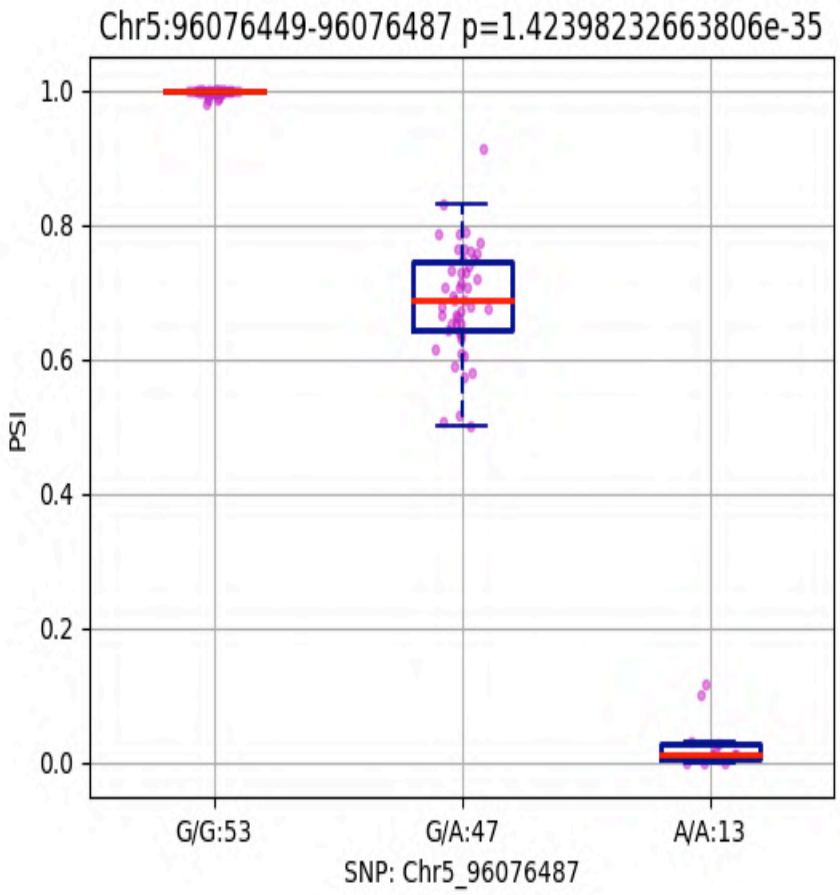

Chr19:871606-871662 p=5.4156790040031e-110

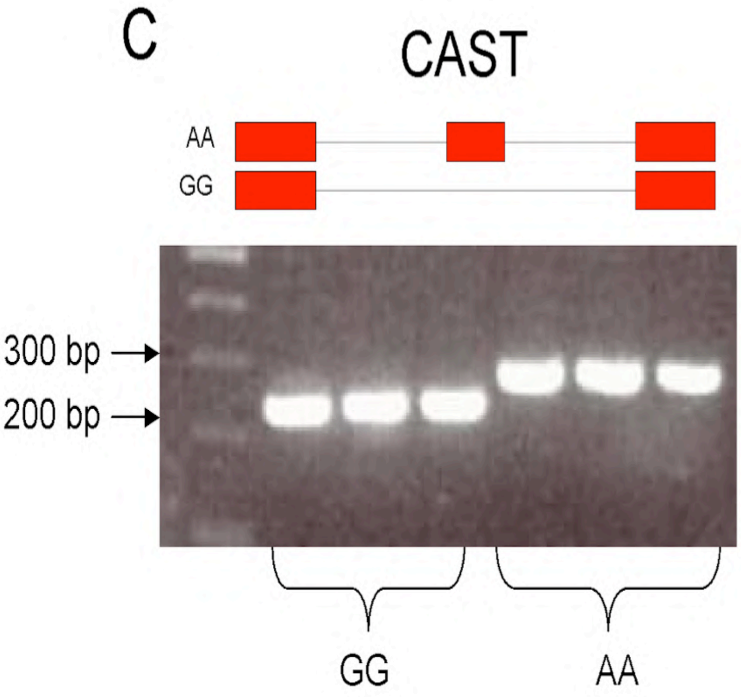
Top Hits


Chr17:7530261-7530362 p=1.70625331428862e-58
PSI
1.00
0.95
0.90
0.85
0.80
0.75
0.70
AG/AG:48
AG/A:50
A/A:16
SNP: Chr17_7521287


Chr17:7530261-7530362 p=1.70625331428862e-58
Exclusion counts
500
400
300
200
100
0
500
1000
1500
2000
2500
3000
Inclusion counts
Chr17_7521287
● hom_ref
● het
● hom_alt

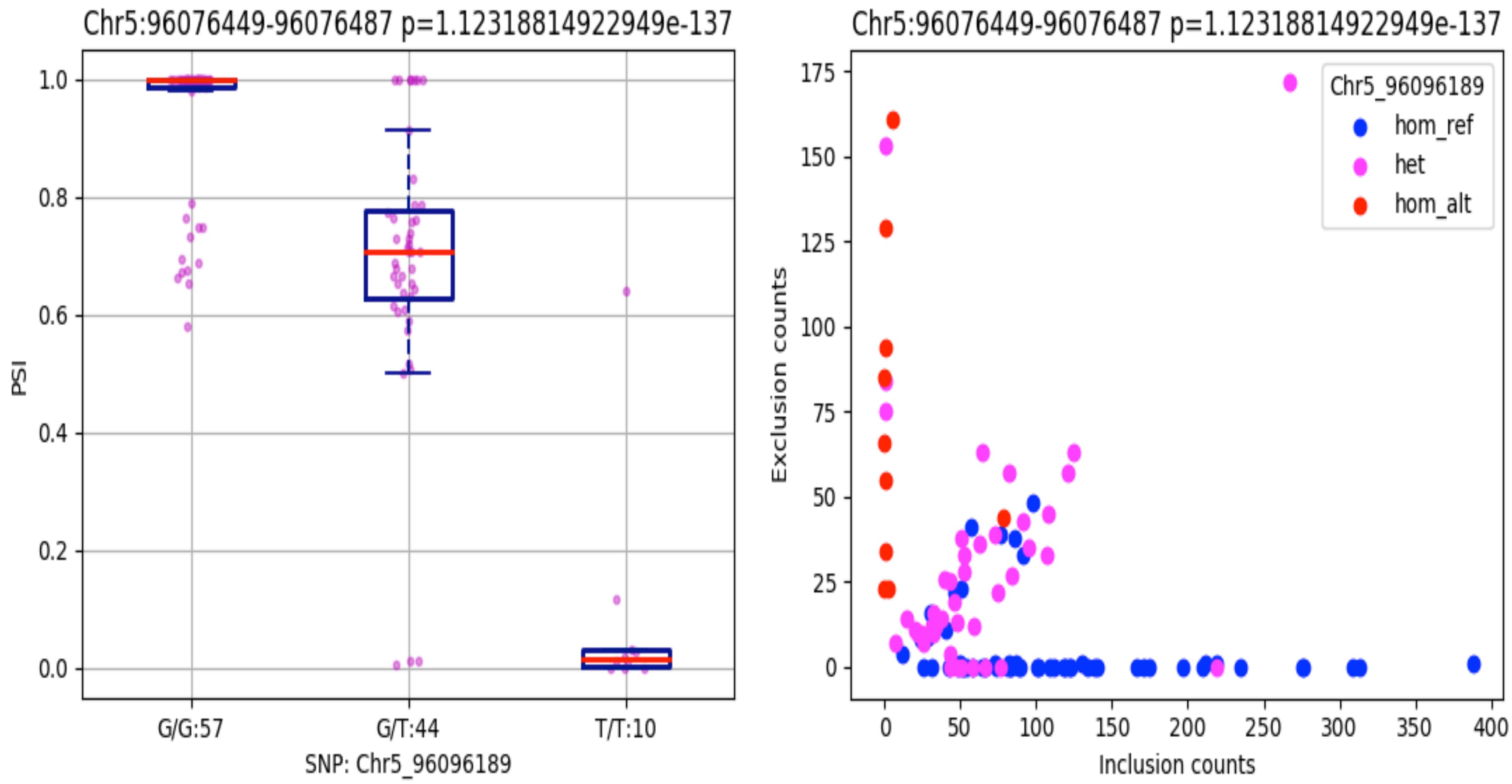
Splice Site sQTLs - CAST



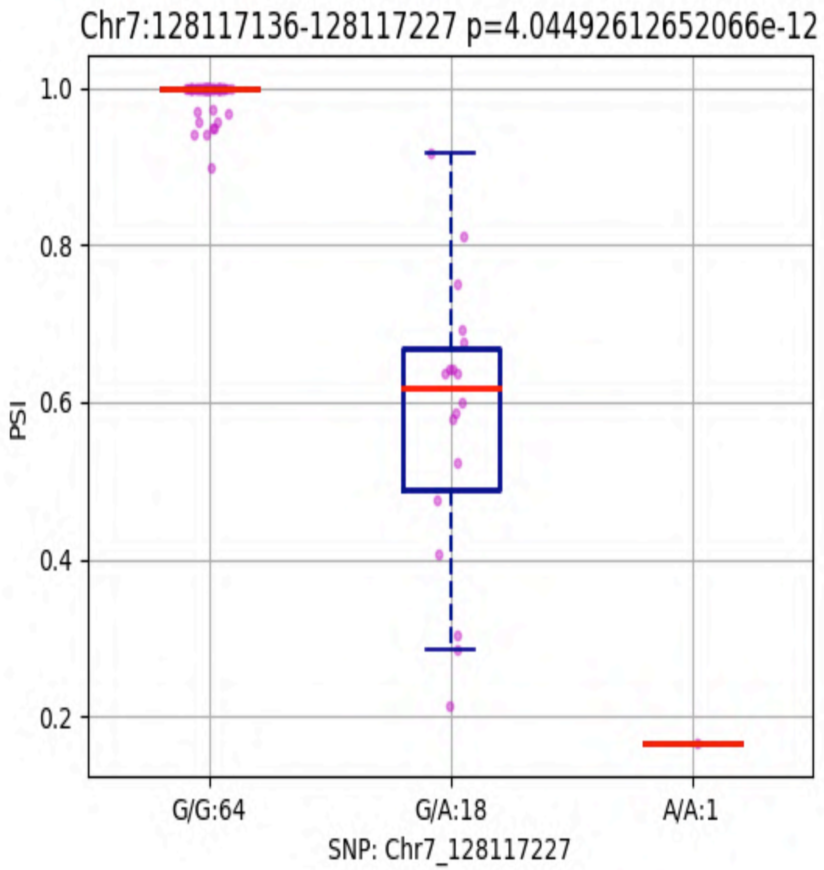


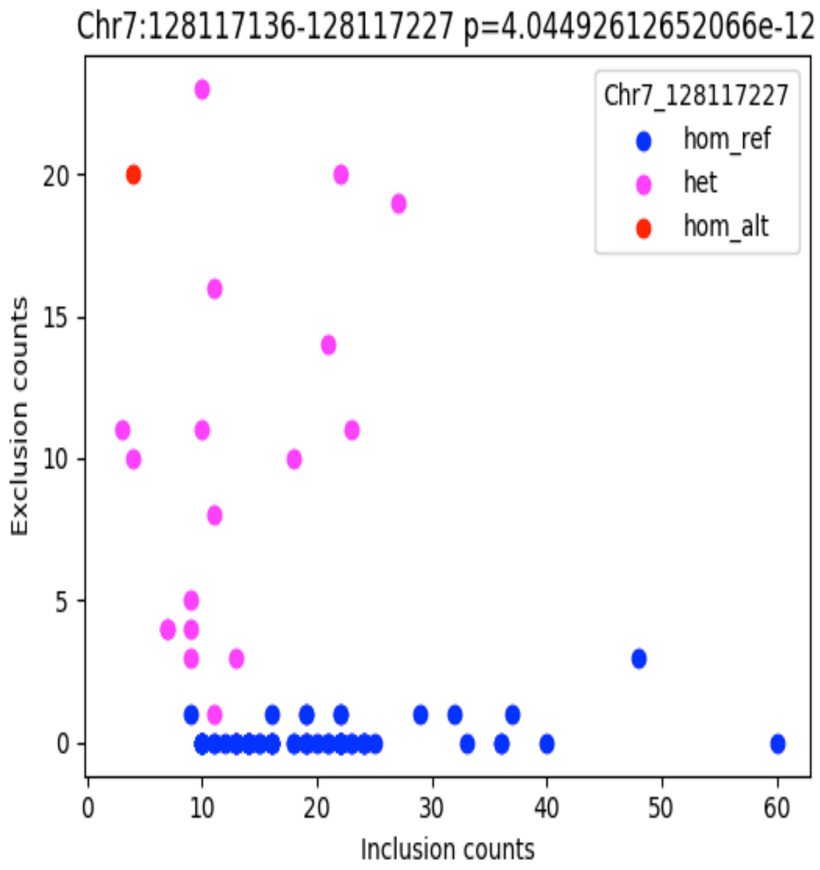
Coulombe-Huntington et al. 2009, PLOS Genetics

Top Hit, chr5:96076449-96076487

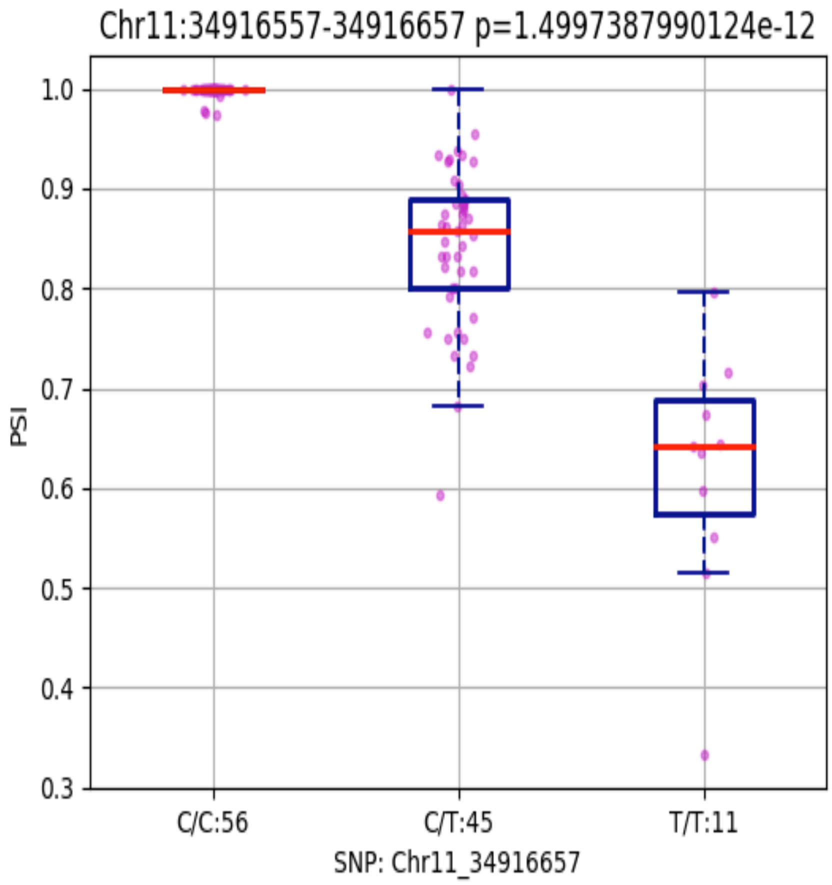


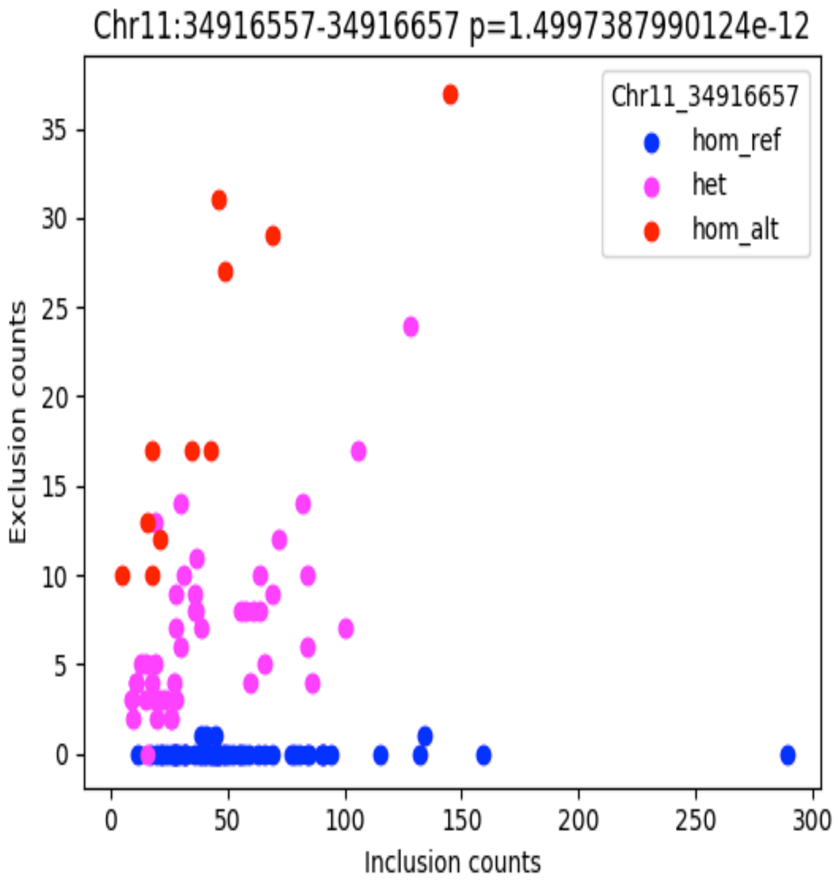
Splice Site sQTLs - METTL2B





Splice Site sQTLs - APIP


Chr11:34916557-34916657 p=1.4997387990124e-12
PSI
C/C:56
C/T:45
T/T:11
SNP: Chr11_34916657


Chr11:34916557-34916657 p=1.4997387990124e-12
Exclusion counts
Inclusion counts
Chr11_34916657
● hom_ref
● het
● hom_alt