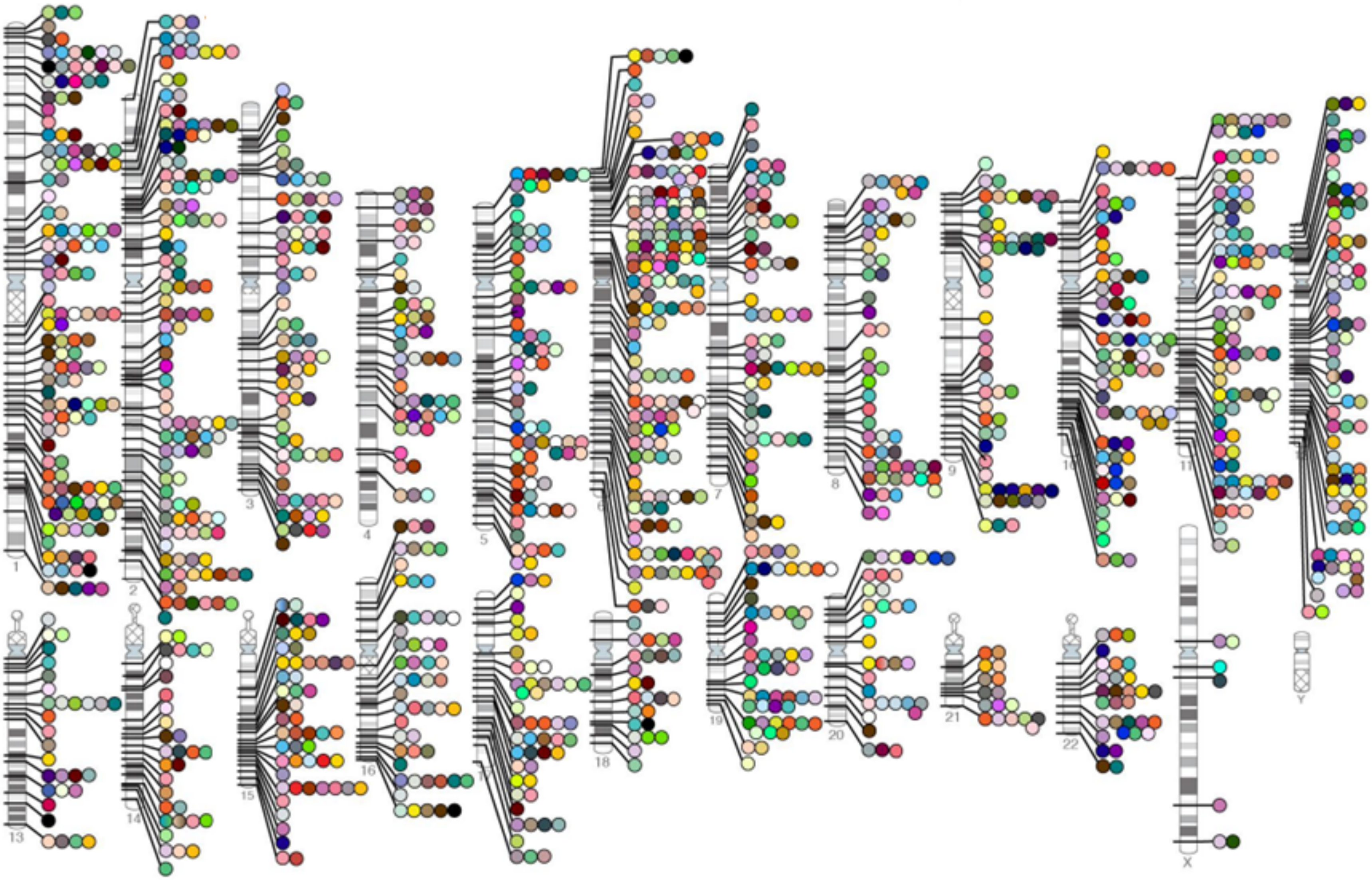
Genetics of gene expression

Stephen Montgomery
smontgom@stanford.edu
montgomerylab.stanford.edu
Stanford University School of Medicine

Chromosome map of disease-associated regions



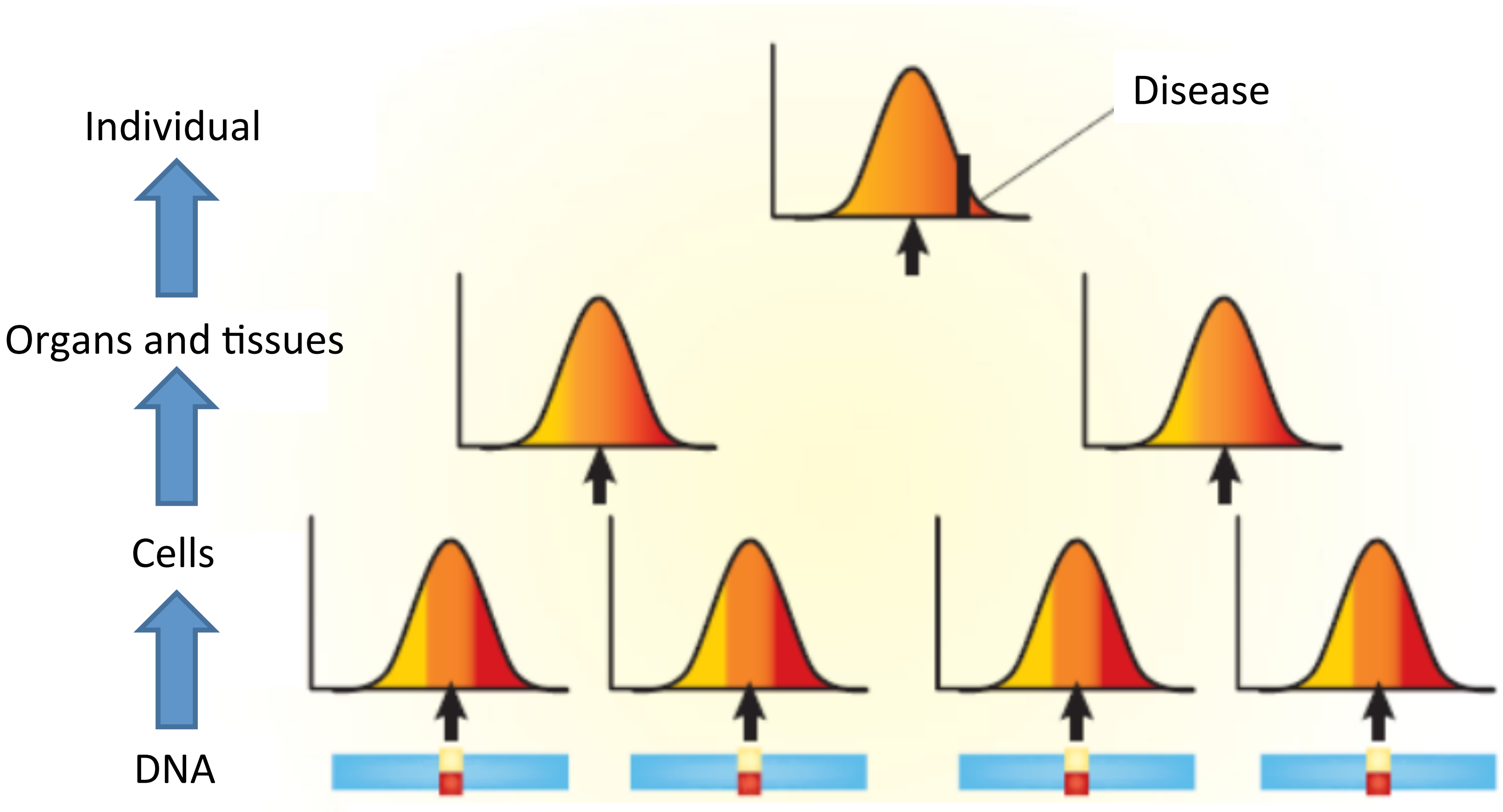
“GWAS have so far identified only a small fraction of the heritability of common diseases, so the ability to make meaningful predictions is still quite limited”

Francis Collins, Director of the NIH, *Nature*, April 2010

Trait	Heritability	Individuals studied	Heritability explained
Coronary artery disease	40%	86995	10%
Type 2 Diabetes	40%	47117	10%
BMI	50%	249796	3%
Blood pressure	50%	34433	1%
Circulating lipids	50%	100000	25%
Height	80%	183727	12.5%

Where is the missing heritability?

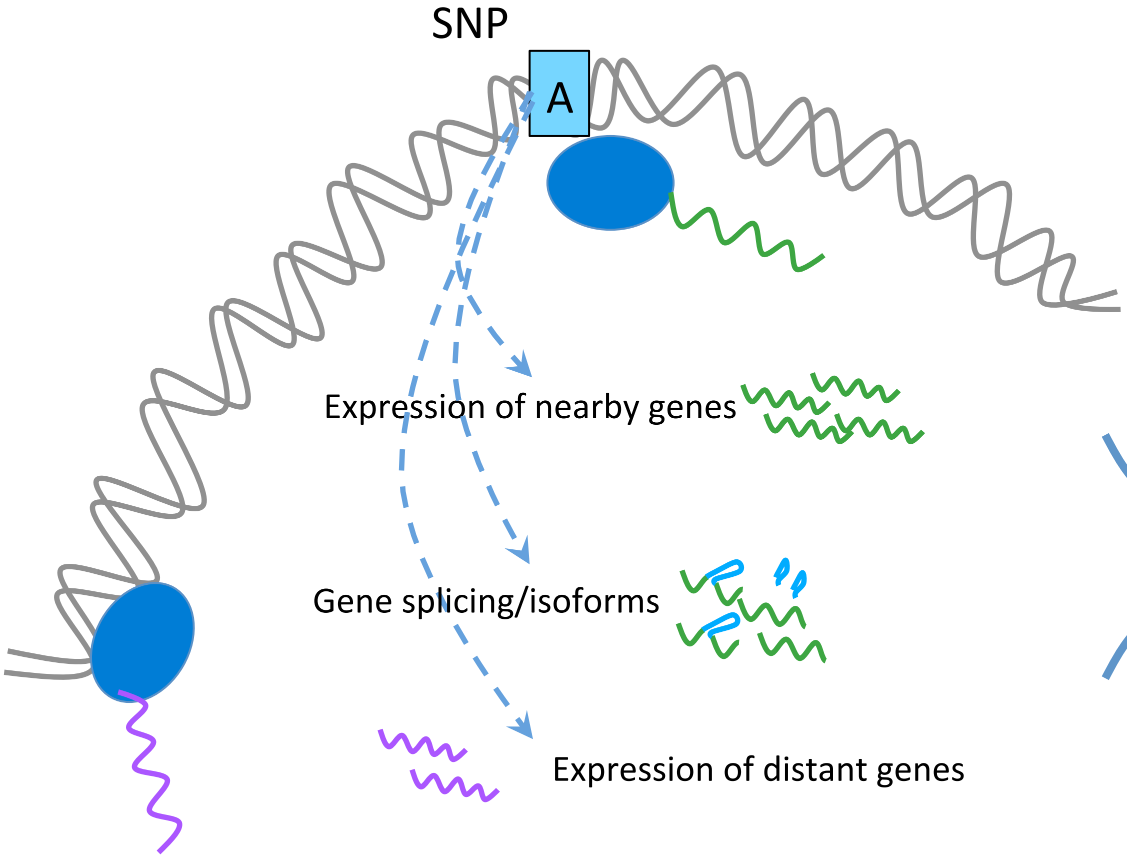
Disease starts at a cellular level



Understanding the influence of genetics on cells will improve our ability to predict disease risk

Genetic studies of gene expression

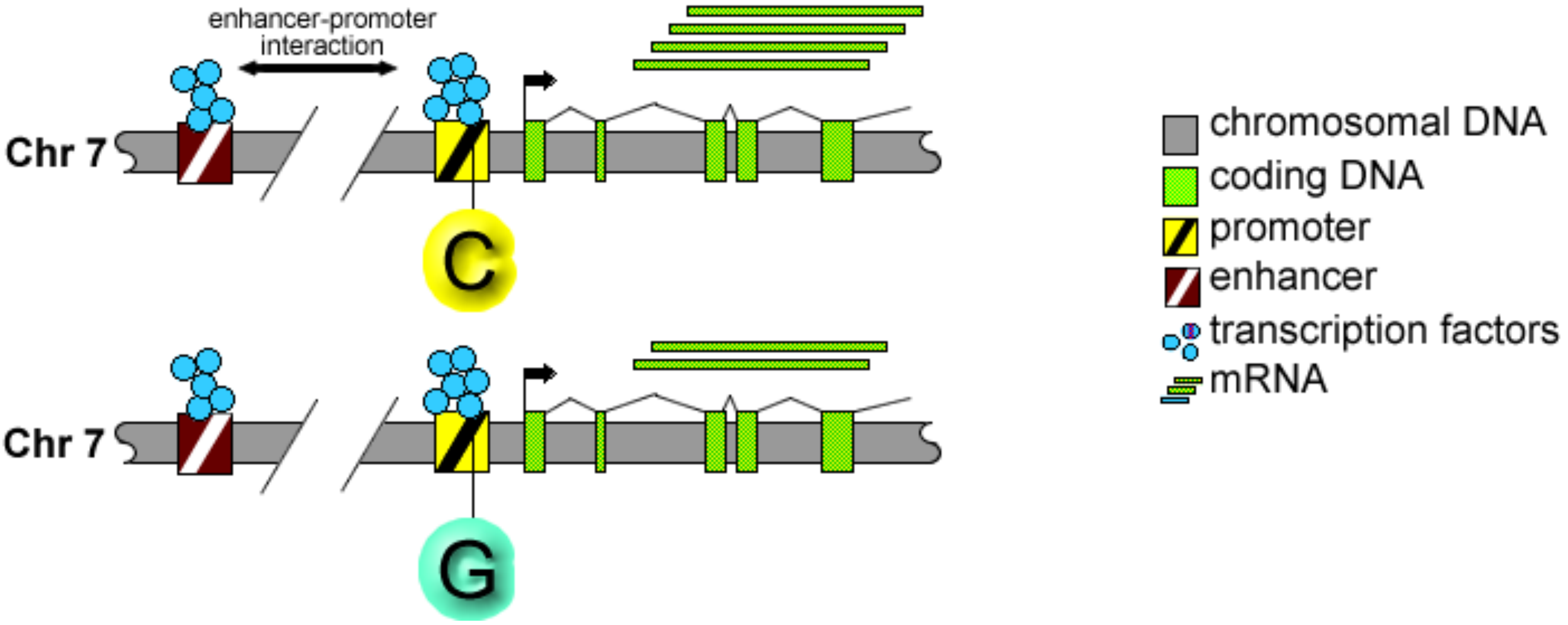
Explore impact of genetic variation on transcriptome diversity



Cellular processes

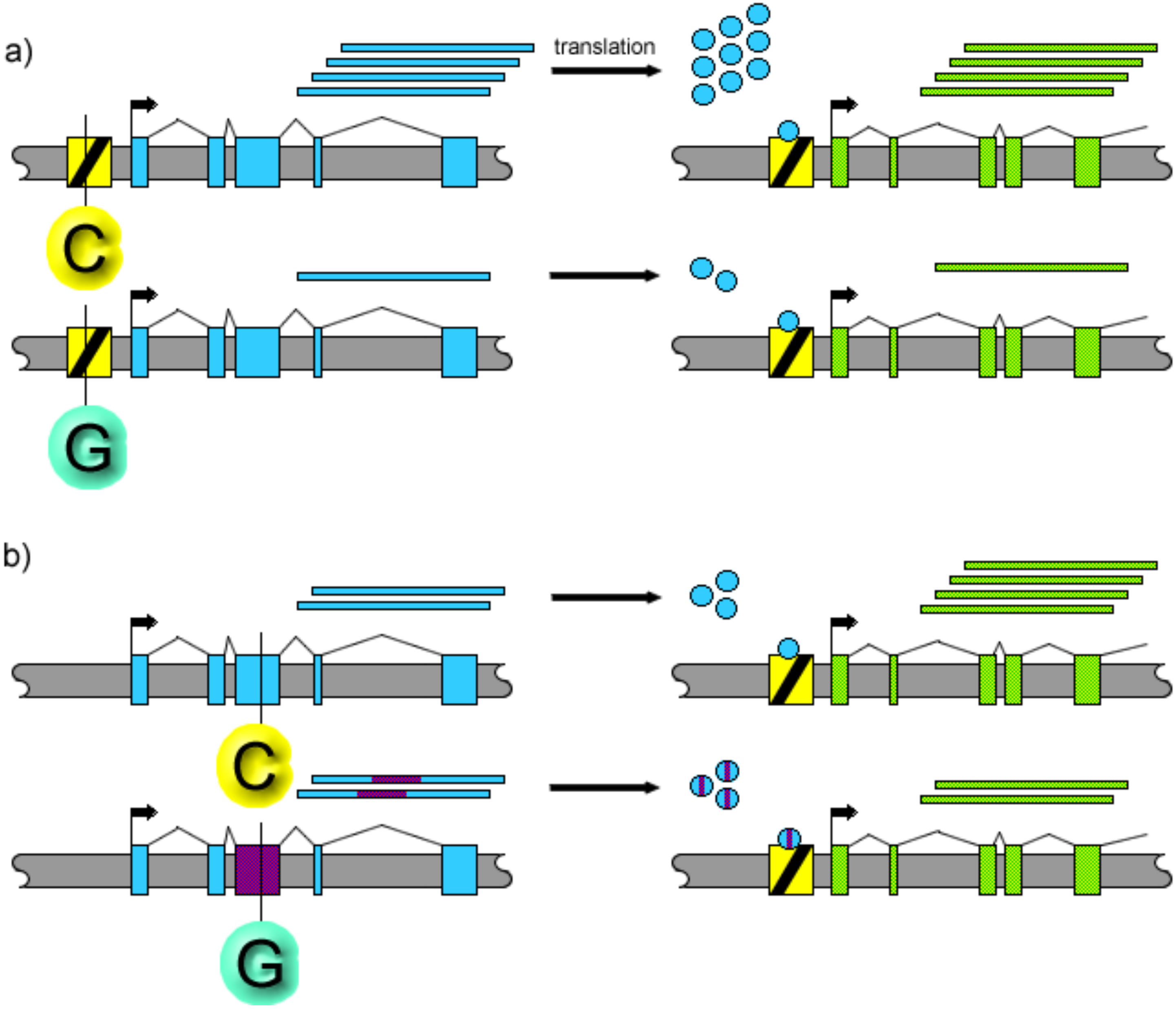
Disease risk

cis- effect

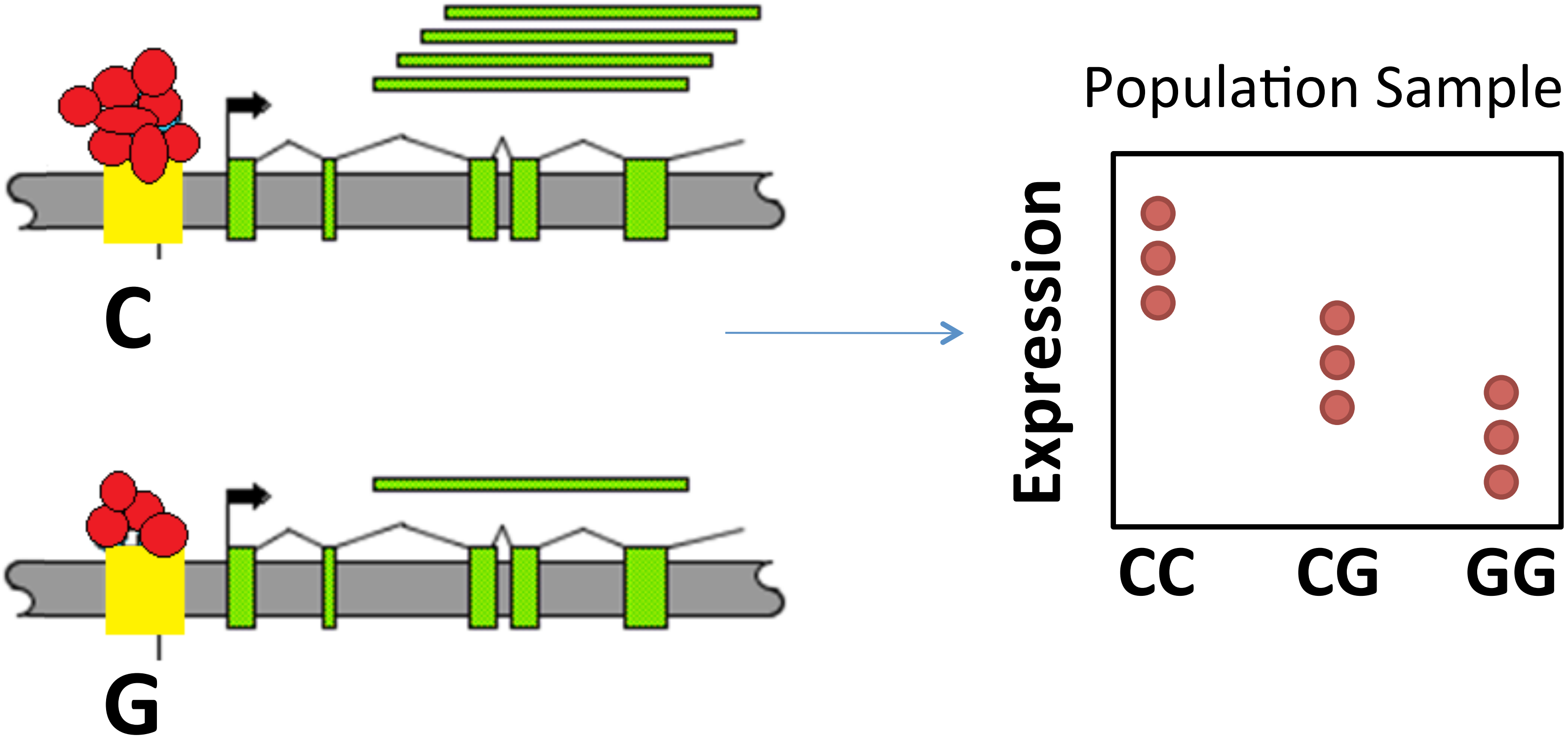


Canonical model

trans- effect

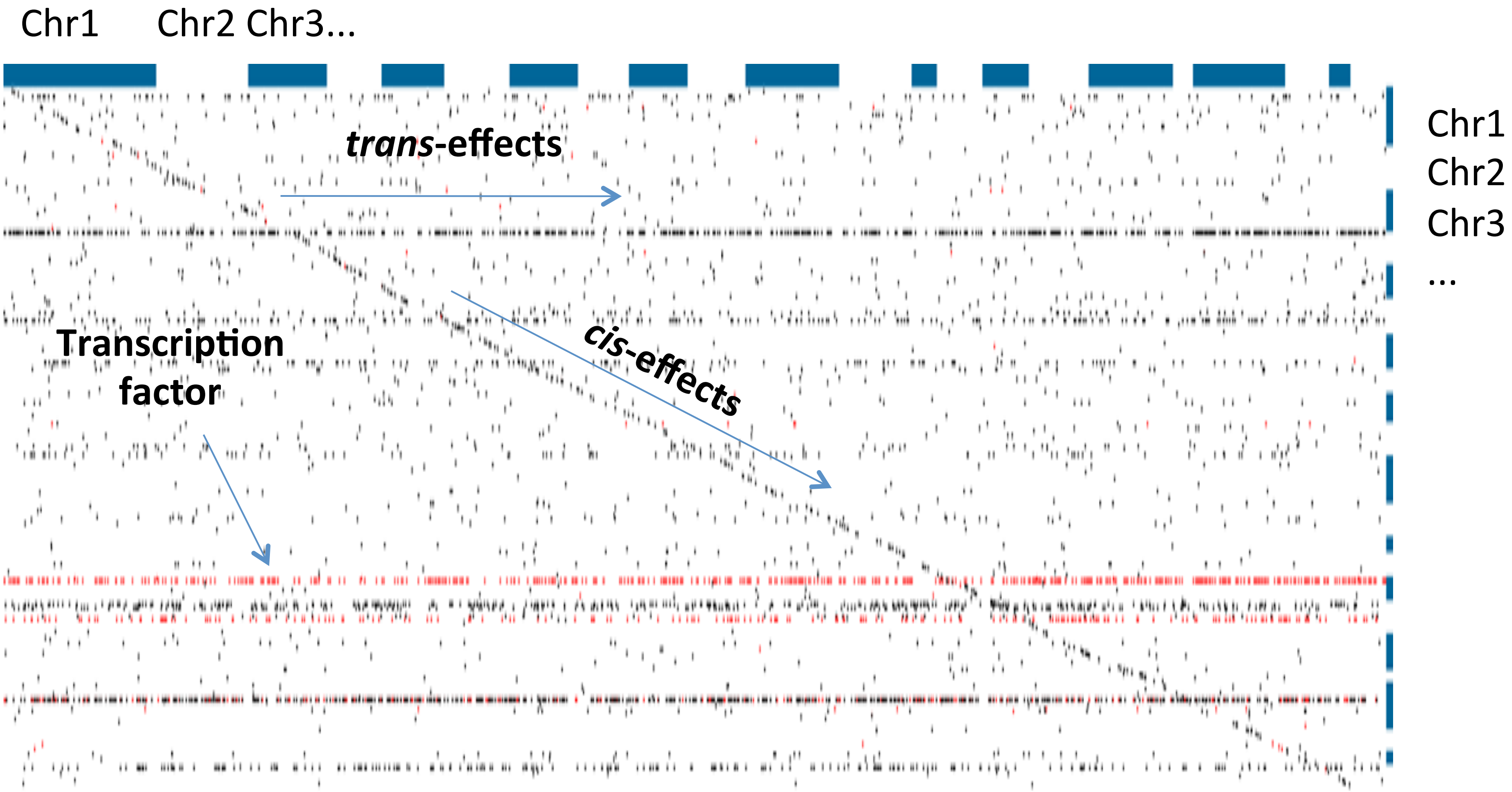


Genetic association can pinpoint regulatory haplotypes



We can identify genetic variants impacting gene expression (eQTLs)

The landscape of regulatory variation



Location of genetic variants by the gene's whose expression they impact

Advantages to studying the genetics of gene expression

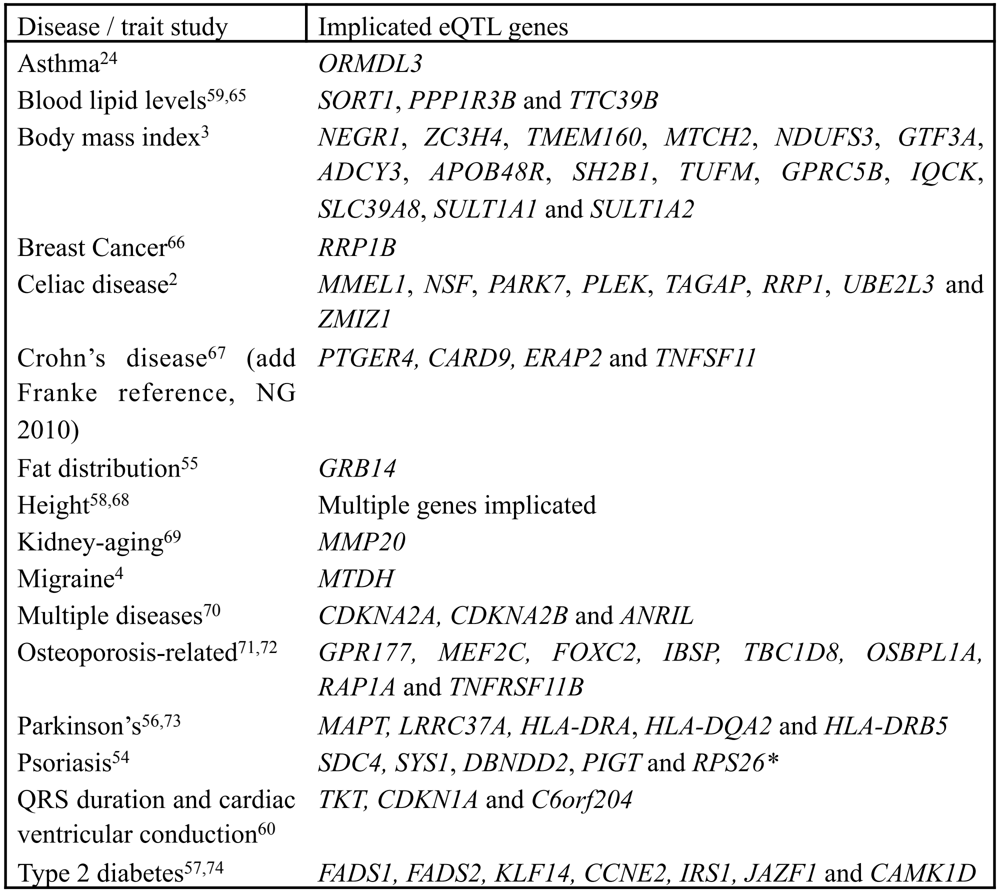
Can rapidly evaluate 1000s of quantitative traits

Can identify genetic regulatory networks

Can easily transform or perturb the system.

Variants are directly connected to cellular mechanism.

Genetic differences in gene expression can identify candidate genes for GWAS variants



Disease / trait study	Implicated eQTL genes
Asthma[24]	*ORMDL3*
Blood lipid levels[59,65]	*SORT1*, *PPP1R3B* and *TTC39B*
Body mass index[3]	*NEGR1*, *ZC3H4*, *TMEM160*, *MTCH2*, *NDUFS3*, *GTF3A*, *ADCY3*, *APOB48R*, *SH2B1*, *TUFM*, *GPRC5B*, *IQCK*, *SLC39A8*, *SULT1A1* and *SULT1A2*
Breast Cancer[66]	*RRP1B*
Celiac disease[2]	*MMEL1*, *NSF*, *PARK7*, *PLEK*, *TAGAP*, *RRP1*, *UBE2L3* and *ZMIZ1*
Crohn's disease[67] (add Franke reference, NG 2010)	*PTGER4, CARD9, ERAP2* and *TNFSF11*
Fat distribution[55]	*GRB14*
Height[58,68]	Multiple genes implicated
Kidney-aging[69]	*MMP20*
Migraine[4]	*MTDH*
Multiple diseases[70]	*CDKNA2A, CDKNA2B* and *ANRIL*
Osteoporosis-related[71,72]	*GPR177, MEF2C, FOXC2, IBSP, TBC1D8, OSBPL1A, RAP1A* and *TNFRSF11B*
Parkinson's[56,73]	*MAPT, LRRC37A, HLA-DRA*, *HLA-DQA2* and *HLA-DRB5*
Psoriasis[54]	*SDC4, SYS1*, *DBNDD2*, *PIGT* and *RPS26**
QRS duration and cardiac ventricular conduction[60]	*TKT, CDKN1A* and *C6orf204*
Type 2 diabetes[57,74]	*FADS1, FADS2, KLF14, CCNE2, IRS1, JAZF1* and *CAMK1D*

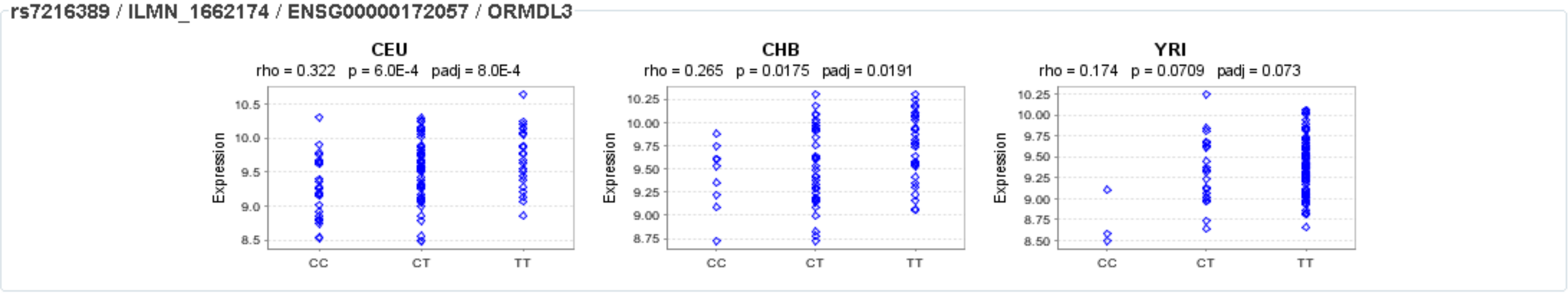
eQTL correlation helps pinpoint implicated genes and mode of effect

Montgomery, Nat Rev Genetics, 2011

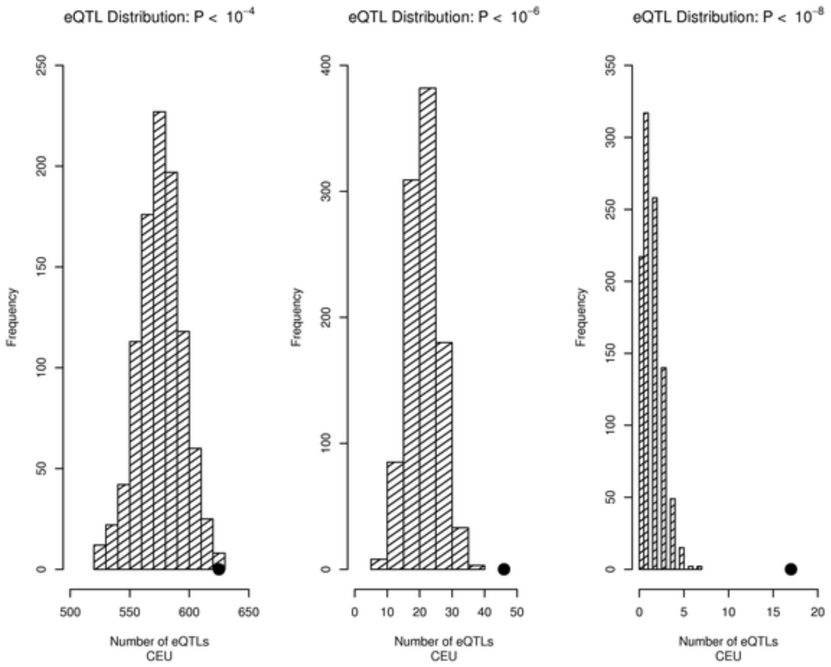
Class activity: What are my asthma variants doing?

In the subset of individuals for whom expression data are available, the T nucleotide allele at *rs7216389* (the marker most strongly associated with disease in the combined GWA analysis) has a frequency of 62% amongst asthmatics compared to 52% in non-asthmatics (*P* = 0.005 in this sample).

Moffatt, Nature, 2007



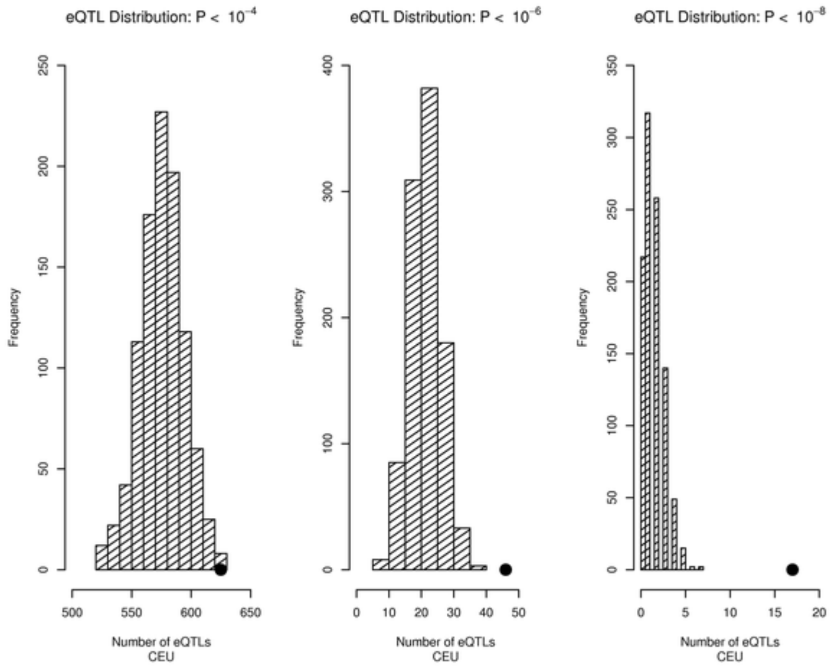
eQTL are more likely to be trait SNPs

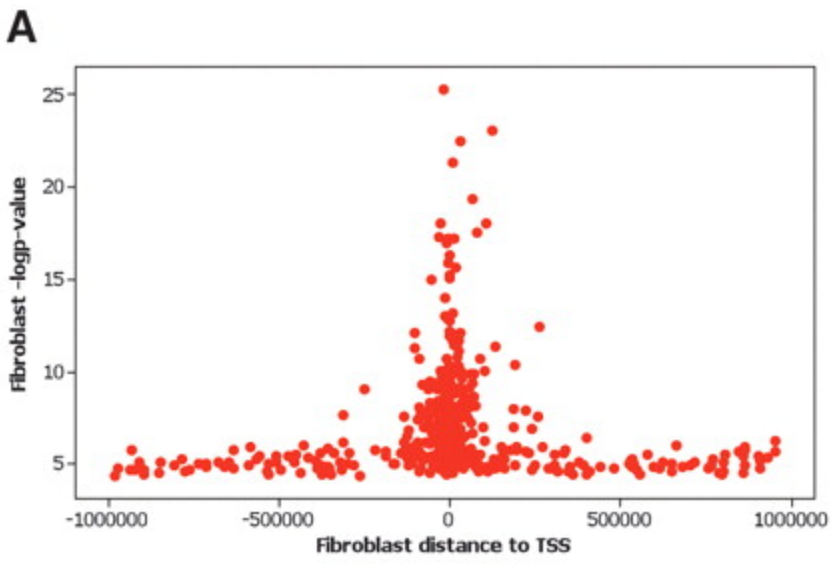


The null was a set of SNPs frequency matched to the GWAS SNPs.
Any problem with this?

Nicolae et al., PLoS Genetics, 2010

GWAS SNPs more likely to be near genes





The null was a set of SNPs frequency matched to the GWAS SNPs.
Any problem with this?

How are eQTL detected and reported?

Reported as the number of genes with significant heritability, linkage or association compared to an FDR

Example 1:
“Of the total set of genes, 2,340 were found to be expressed, of which 31% had significant heritability when a false-discovery rate of 0.05 was used.”

- Monks, AJHG, 75(6): 1094–1105. 2004

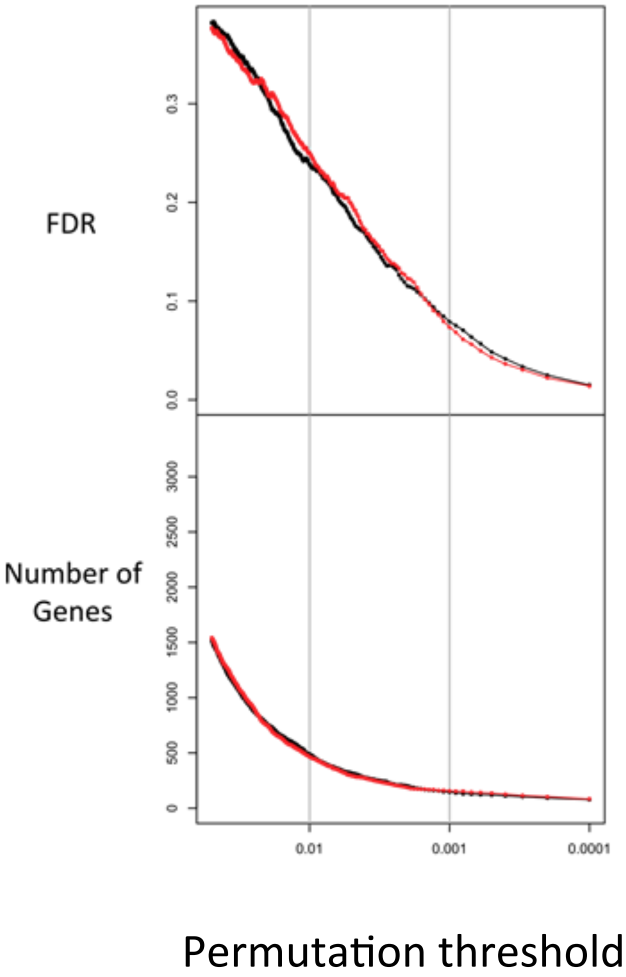
Example 2:
“Applying this genome-wide threshold to 3,554 scans we would expect only 3.5 genome scans to show any linkage evidence with a *P*-value this extreme by chance. Instead we found 142 expression phenotypes with evidence for linkage beyond the *P*-value threshold, and in some cases far beyond, so we conclude that false-positive linkage findings are at most a small fraction of the significant results.”

- Morley, Nature, 430(7001): 743–747. 2004

Example 3:
“We detected 293, 274, 326 and 363 cis associations for CEU, CHB, JPT and YRI, respectively, corresponding to 783 distinct genes and an FDR of 4–5%.”

- Stranger, Nat Genetics, 39, 1217–1224. 2007

eQTL definition depends on false discovery reported

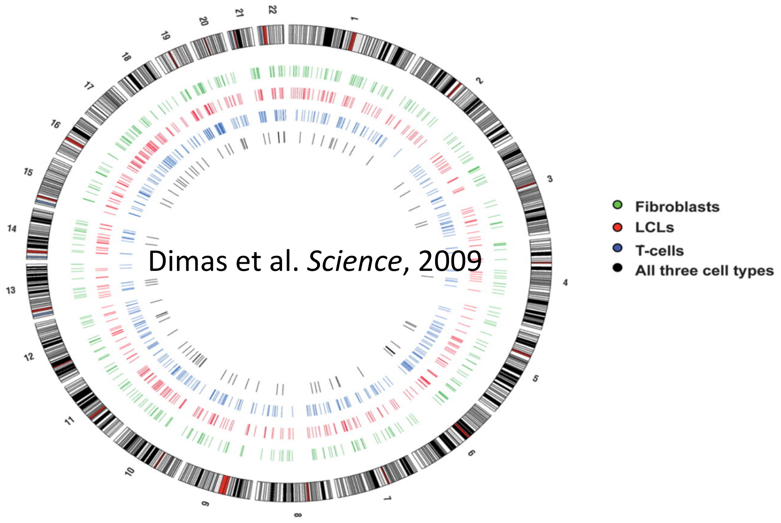


IMPORTANT: Understand the relationship Between false positive rate and eQTL reported!

Discovery of eQTL depends on:

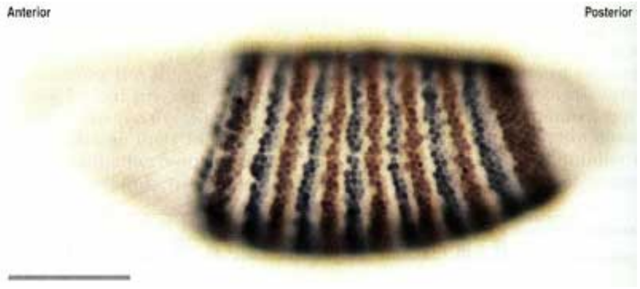
(A)Biological factors
(B) Technological factors

Biological factors influencing eQTL discovery



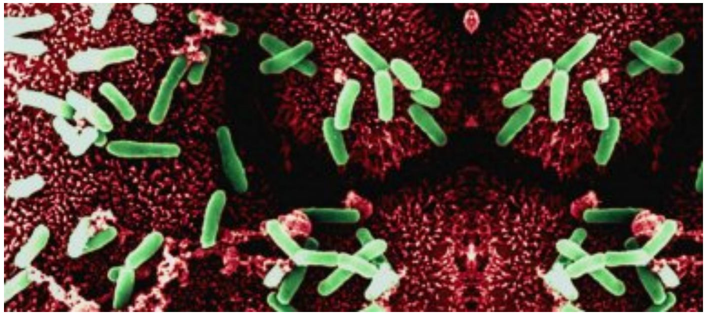
Dimas et al. *Science*, 2009

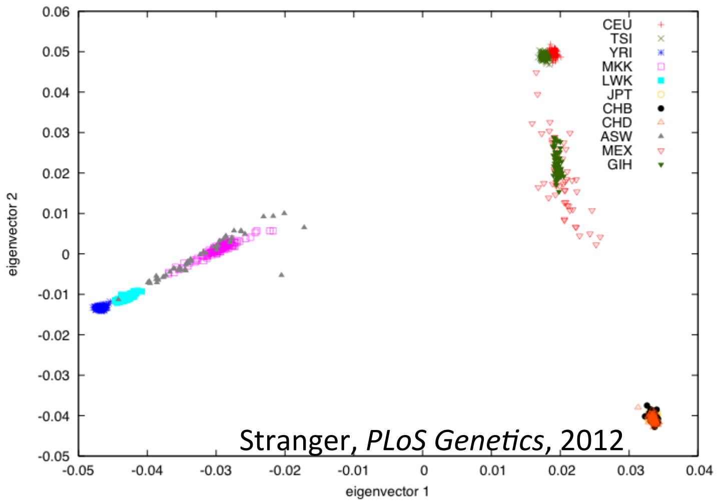
Trait biology



Ancestry

Environment





Stranger, *PLoS Genetics*, 2012

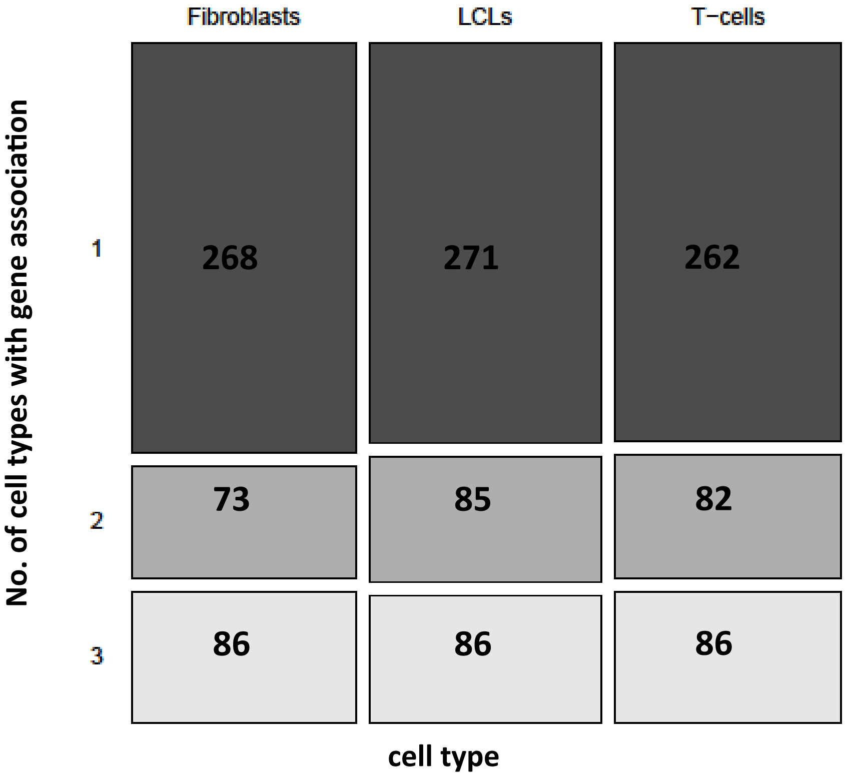
Biological factor: Cell or tissue type

Determining how ubiquitous eQTL signals (and potential disease mechanism) are in different tissues.

i.e. if I find an eQTL in fat will it be informative of mechanism underlying disease risk for a disease based in muscle.

Probably not

Cell type-specific and cell type-shared gene associations (0.001 permutation threshold)



69-80% of cis associations are cell type-specific

Dimas *et al Science* 2009

50% specific (adipose and blood)

Emilsson *et al Nature* 2008

>50% specific (cortical tissue and peripheral blood)

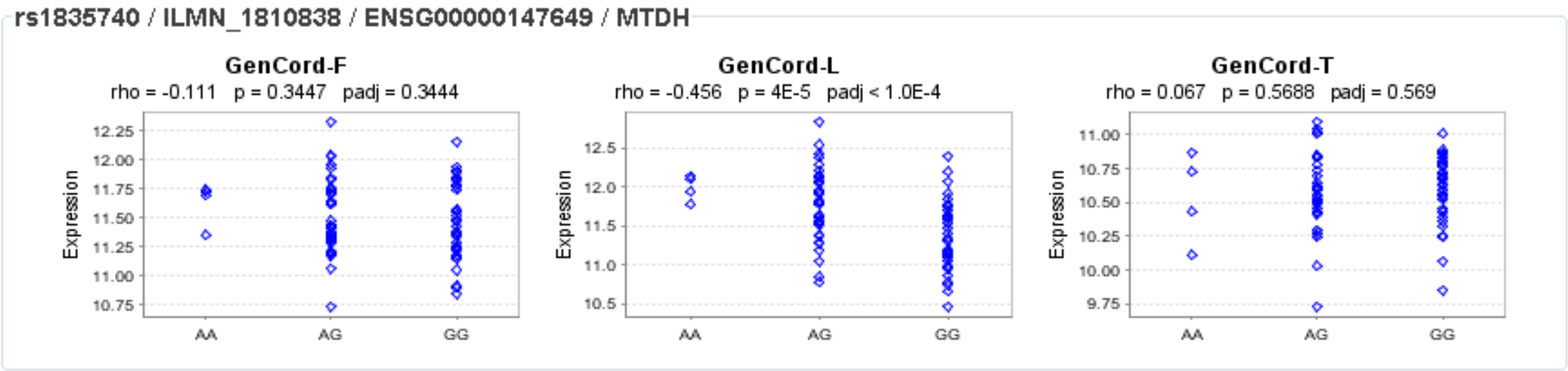
Heinzen *et al PloS Biology* 2008

However, all estimates depend on eQTL discovery FDR and method for assessing sharing

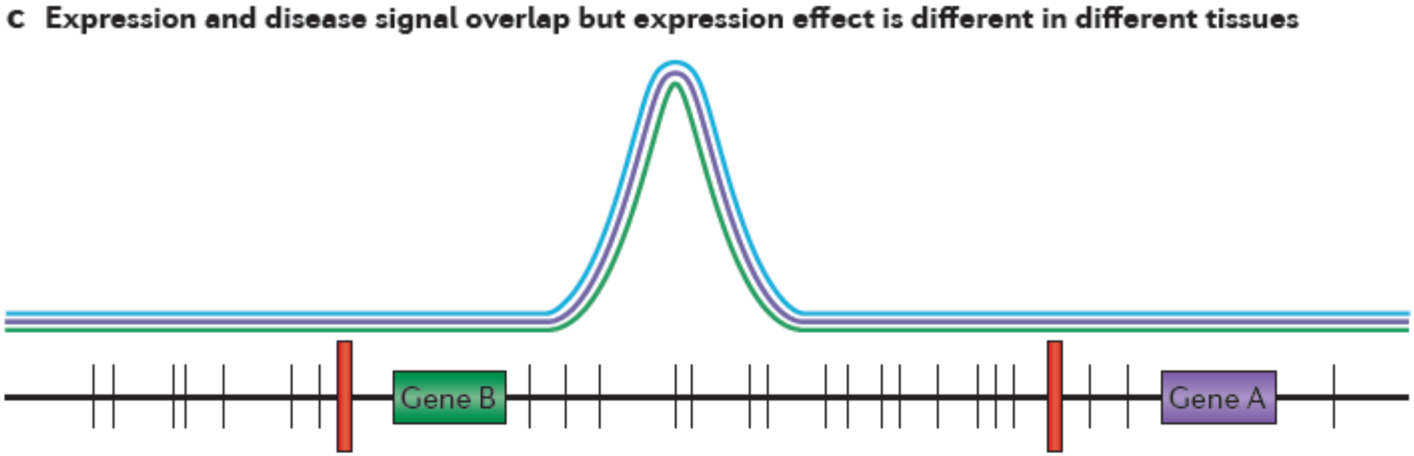
Class activity: What are my migraine variants doing in different tissues?

We identified the minor allele of rs1835740 on chromosome 8q22.1 to be associated with migraine ($P = 5.38 \times 10^{-9}$, odds ratio = 1.23, 95% CI 1.150–1.324) in a genome-wide association study of 2,731 migraine cases ascertained from three European headache clinics and 10,747 population-matched controls. In an expression quantitative trait study in lymphoblastoid cell lines, transcript levels of the *MTDH* were found to have a significant correlation to rs1835740 ($P = 3.96 \times 10^{-5}$, permuted threshold for genome-wide significance 7.7×10^{-5}).

Anttila, Nature Genetics, 2011



Predictive value of eQTL dependent on proximity to pathological tissue


c Expression and disease signal overlap but expression effect is different in different tissues
Gene B
Gene A

We have limited understanding of the Type I and II error rate

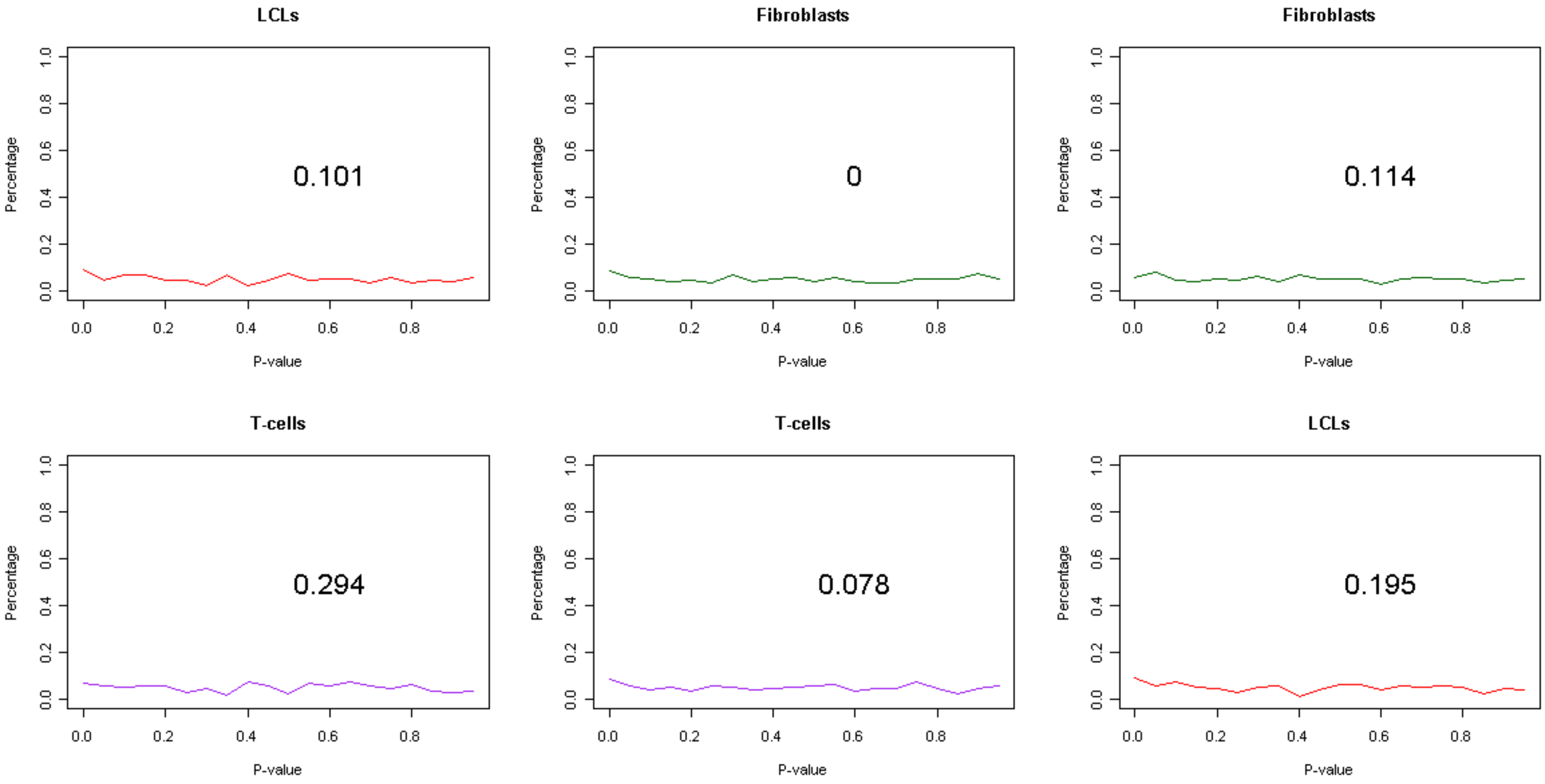
However, a lack of sharing may allow us to discover the pathological tissue

Example of tissue-specific GWAS-eQTL sharing

F: 306 eQTL genes

L: 377 eQTL genes

T: 299 eQTL genes

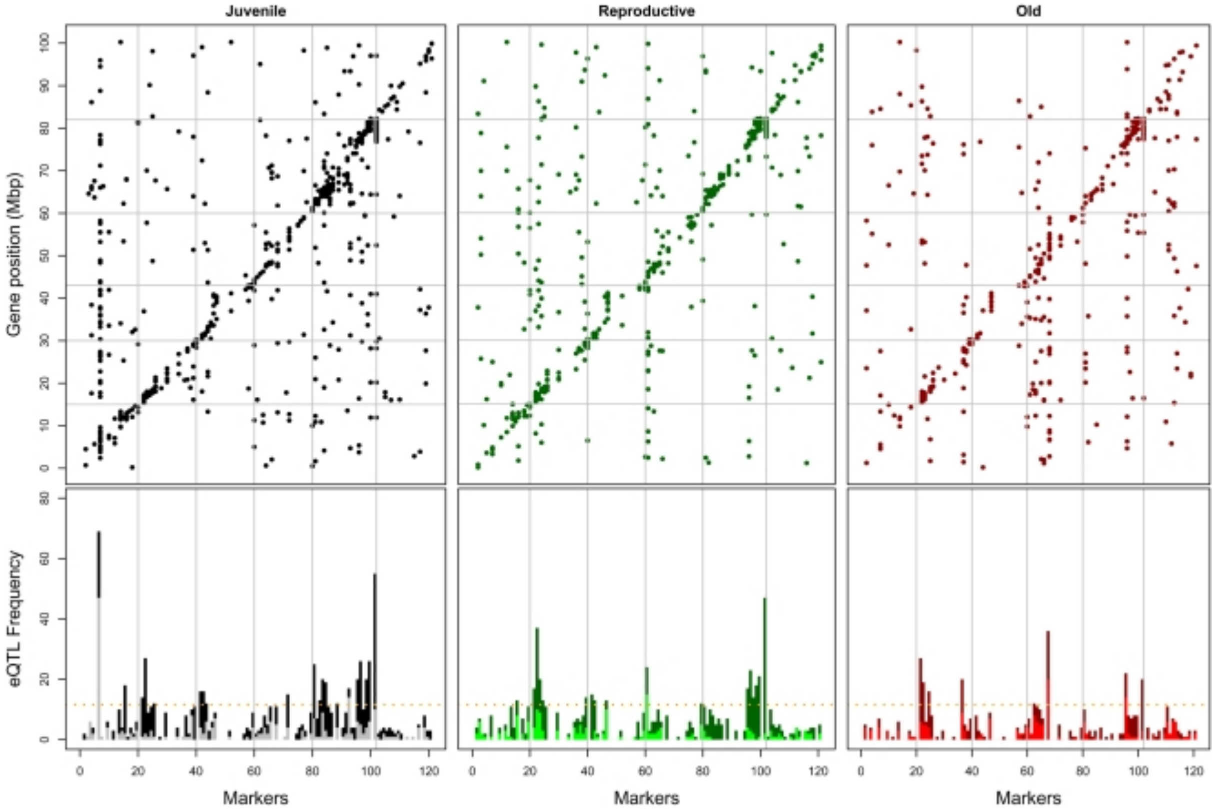


Guiterrez-Arcelus, submitted

Biological factor: Development and aging

Determining how eQTL behave over time (development and aging).

Less eQTL in older individuals



Recombinant inbred *C.elegans*

More interruption by somatic or environmental effects?

Viñuela A, Genome Research 20(7):929-37. 2010

Biological factor: Studied population

Determining how ubiquitous eQTL signals (and potential disease mechanism) are in different populations.

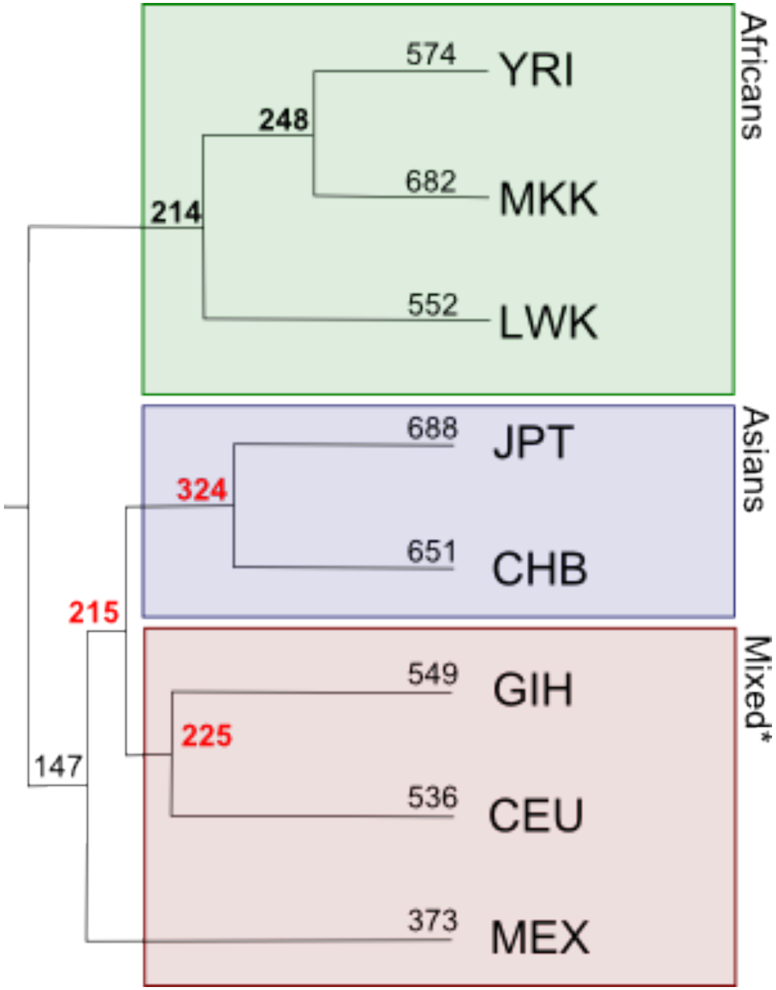
i.e. if I find an eQTL in Europeans will it be informative of mechanism underlying disease risk for a disease found in Chinese.

Not all eQTL shared across populations

"We have reported that many genes showing cis associations at the 0.001 permutation threshold are shared (about 37%) in at least two populations … In 95–97% of the shared associations, the direction of the allelic effect was the same across populations, and the discordant 3–5% was of the same order as the FDR."

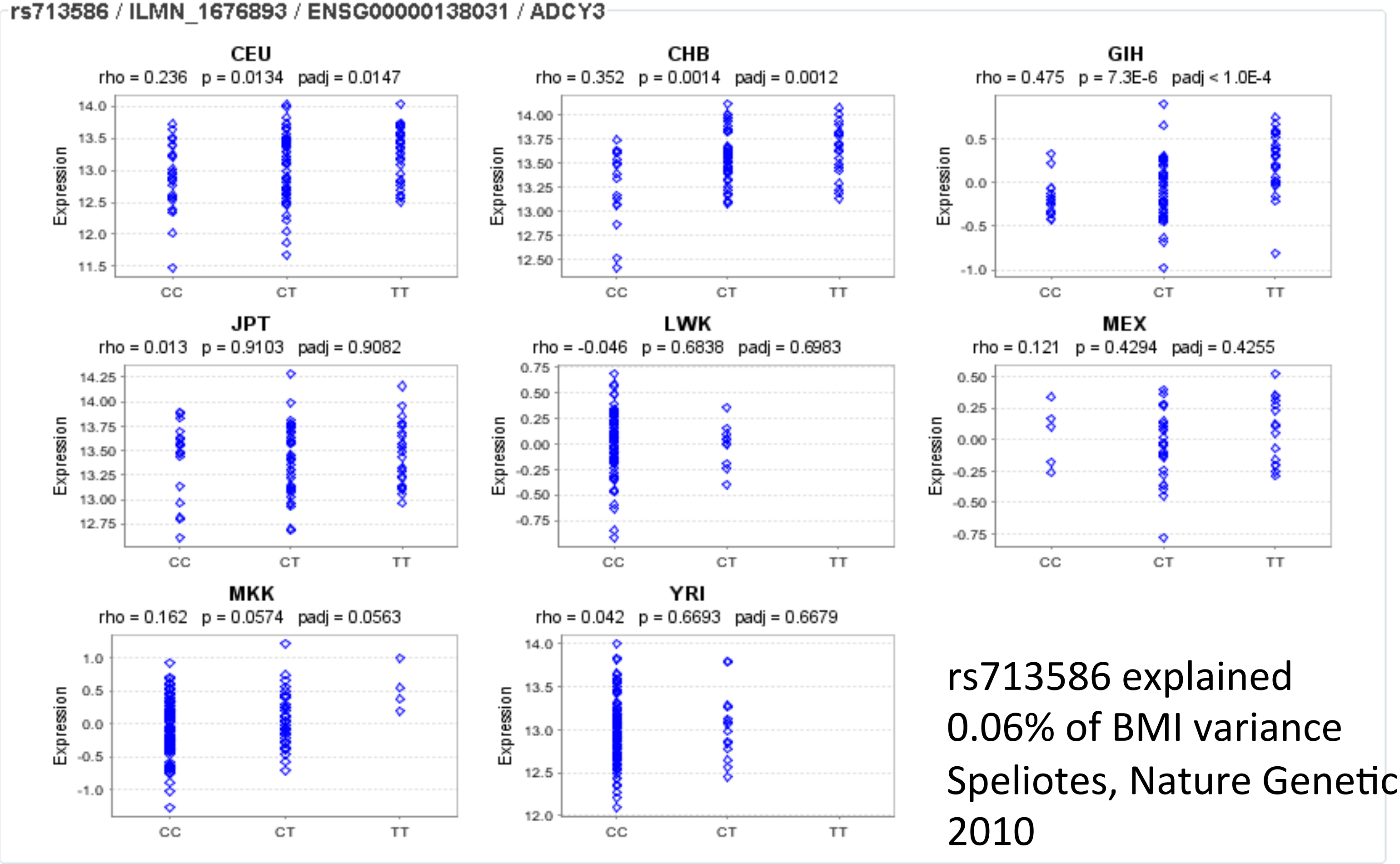
Stranger et al, Nat Genetics, 2007

If we know the etiology of a disease can we predict its population frequency from cellular models of that disease?



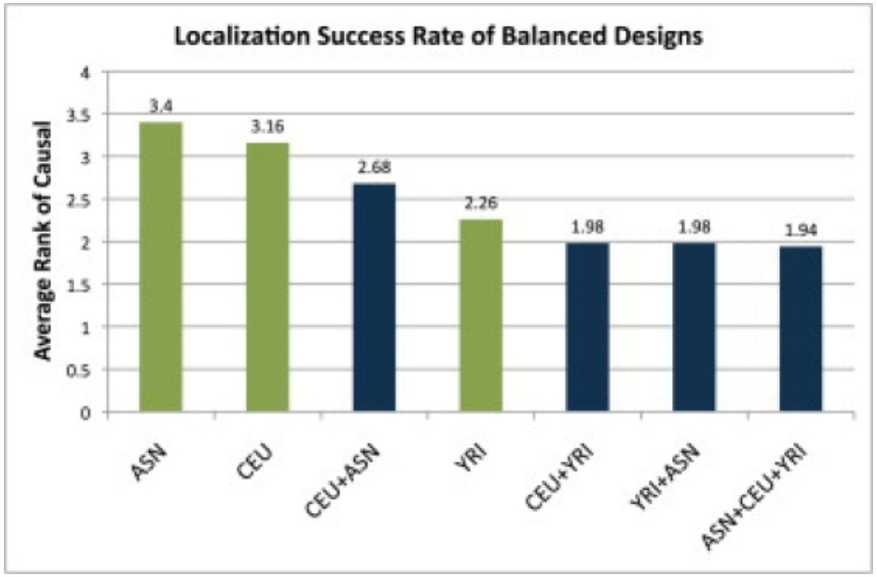
Stranger et al, PLoS Genetics, 2012

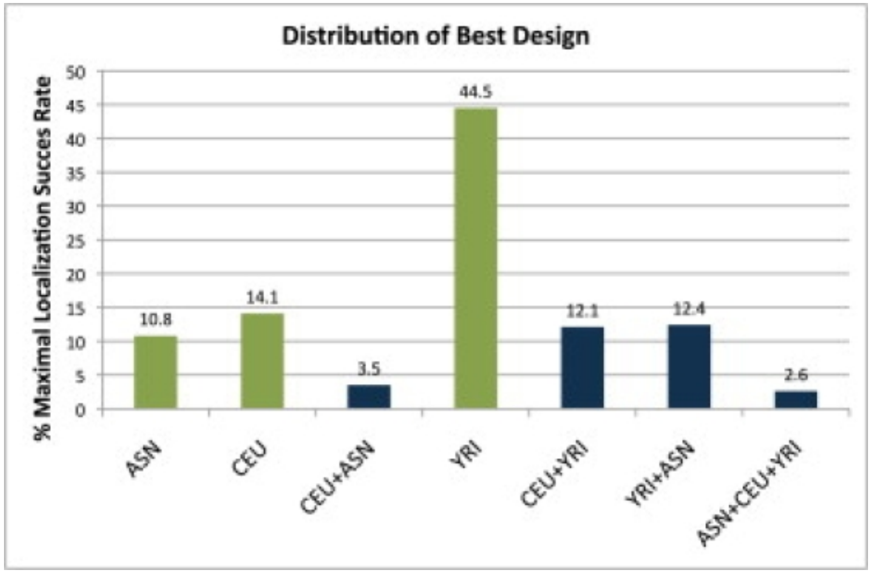
Class activity: What are my BMI variants doing in different populations?



rs713586 explained 0.06% of BMI variance
Speliotes, Nature Genetics, 2010

Multiple population study designs: Recombination mapping can get at causal variants

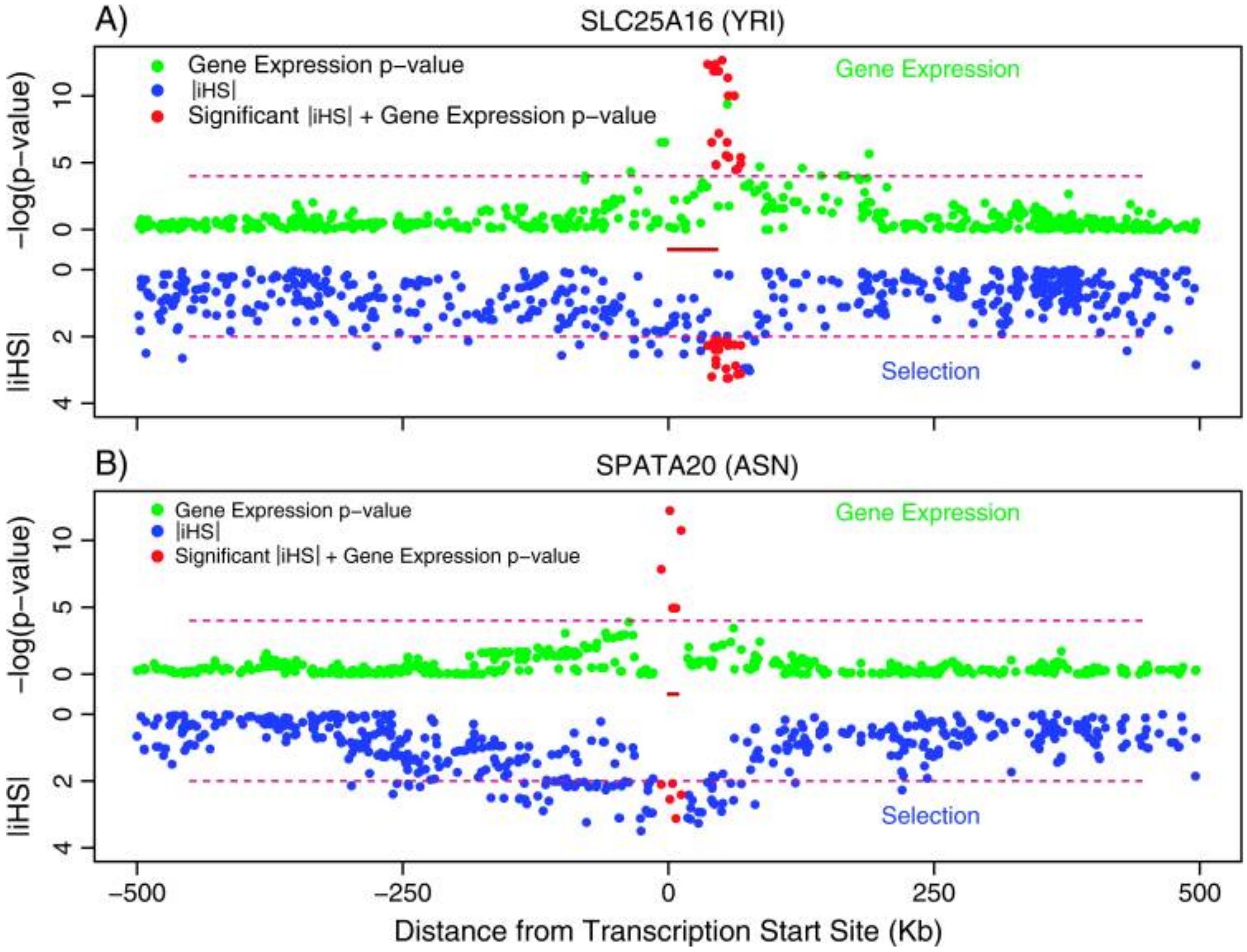




Zaitlen, AJHG, ; 86(1): 23–33. 2010

Multiple populations do well at mapping causal variants; however their design results in a reduction of power

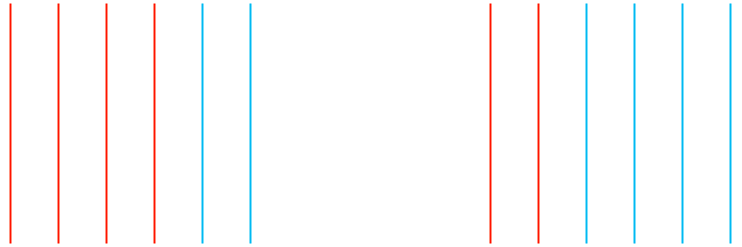
eQTLs under selection



Kudaravalli, MBE, 2009

Admixed populations

- Challenges: Loss of power if local ancestry not known or inflation in significance if frequency differences are large and effect is trans-acting.



Eur: mean 3.0

Afr: mean 4.0

If mean expression invariant to genotype then
allele frequency differences will create false association

Solution: Add local ancestry as a covariate

Biological factor: Environment studies

Determining how eQTLs behave under stimulus

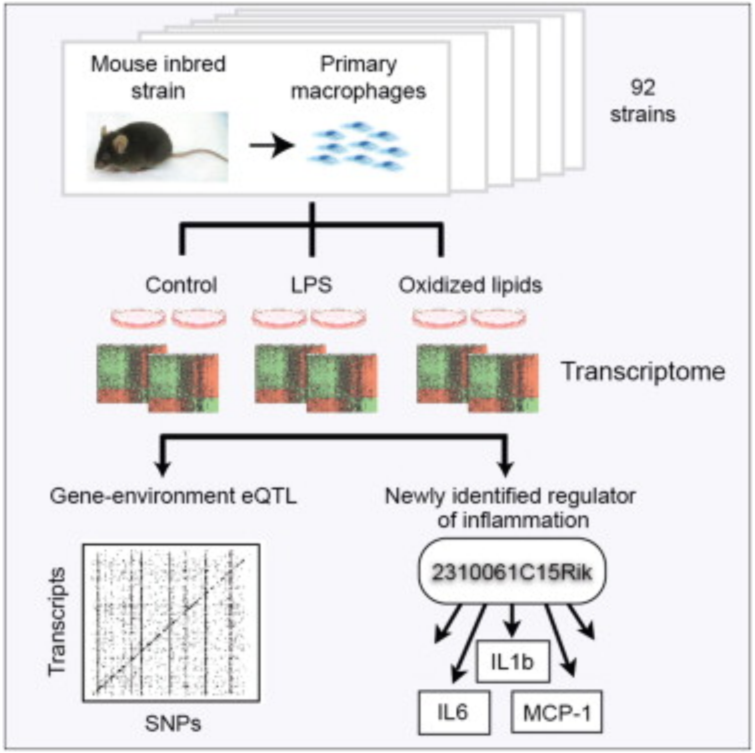
i.e. if I find an eQTL in resting state will it be informative of mechanism underlying an responsive state.

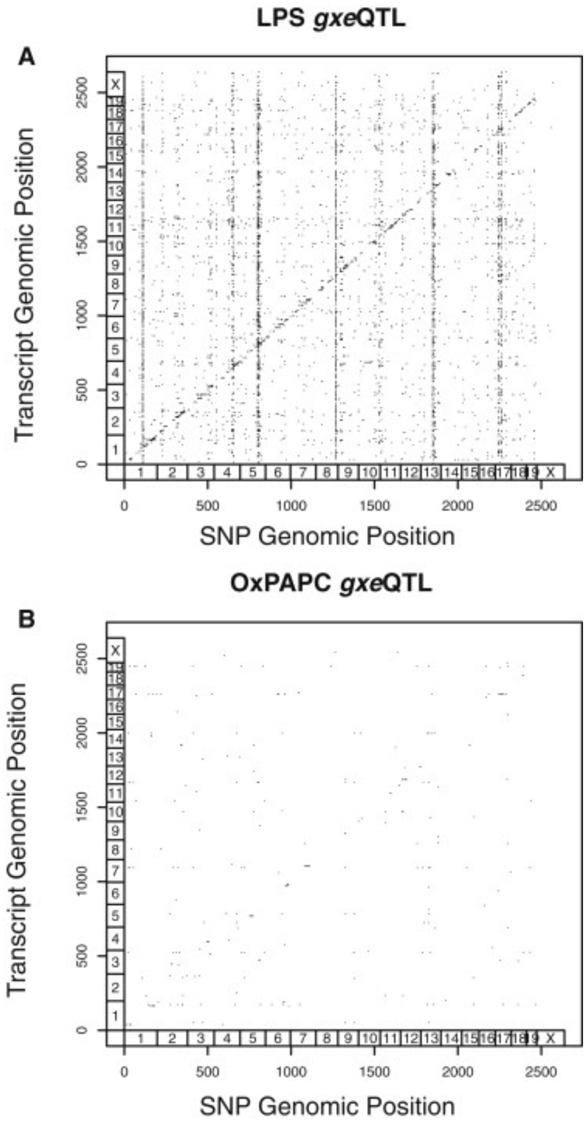
Answer: GxE discoveries have been study dependent

“We carried out large-scale induction experiments using primary human bone cells derived from unrelated donors of Swedish origin treated with 18 different stimuli (7 treatments and 2 controls, each assessed at 2 time points). … We found that 93% of cis-eQTLs at 1% FDR were observed in at least one additional treatment, and in fact, on average, only 1.4% of the cis-eQTLs were considered as treatment-specific at high confidence. “

- Grundberg PloS Genetics 7(1). 2011

LPS-stimulation eQTL





Orozco et al, Cell, 2012

Discovery of eQTL depends on technological factors

Gene expression technology

PCR-based, array-based, **sequencing-based**

Genotyping technology

array-based, sequencing-based

Sample size

More individuals and/or families yields more power to detect association with particular effect sizes. (Lowers FDR). Early studies used 18-30 families or 45-60 unrelated individuals.

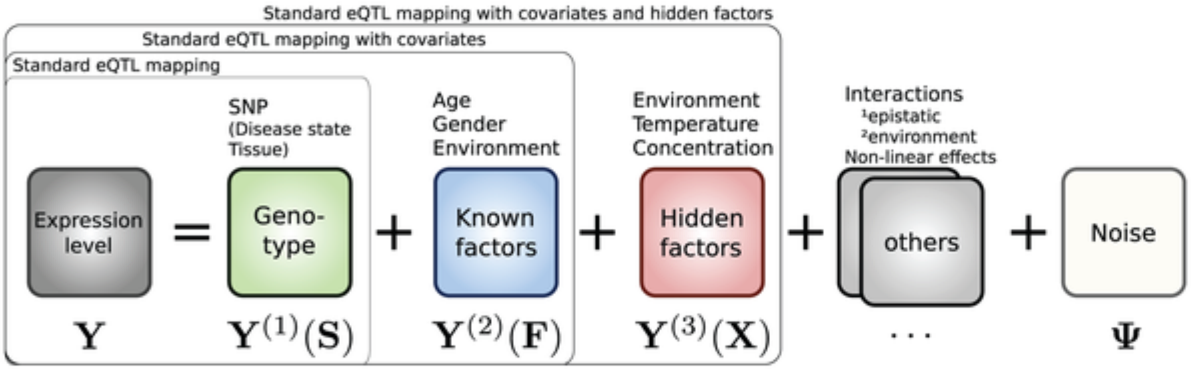
THE biases we don’t know about: Hidden factors can cause false associations

- Hidden technical and biological variables. i.e. population, sex, date of processing
- However, correcting these factors can remove true signals (i.e. master regulators)

Methods to correct hidden factors

- Factor analysis on *40* global factors has tripled eQTL discovery.

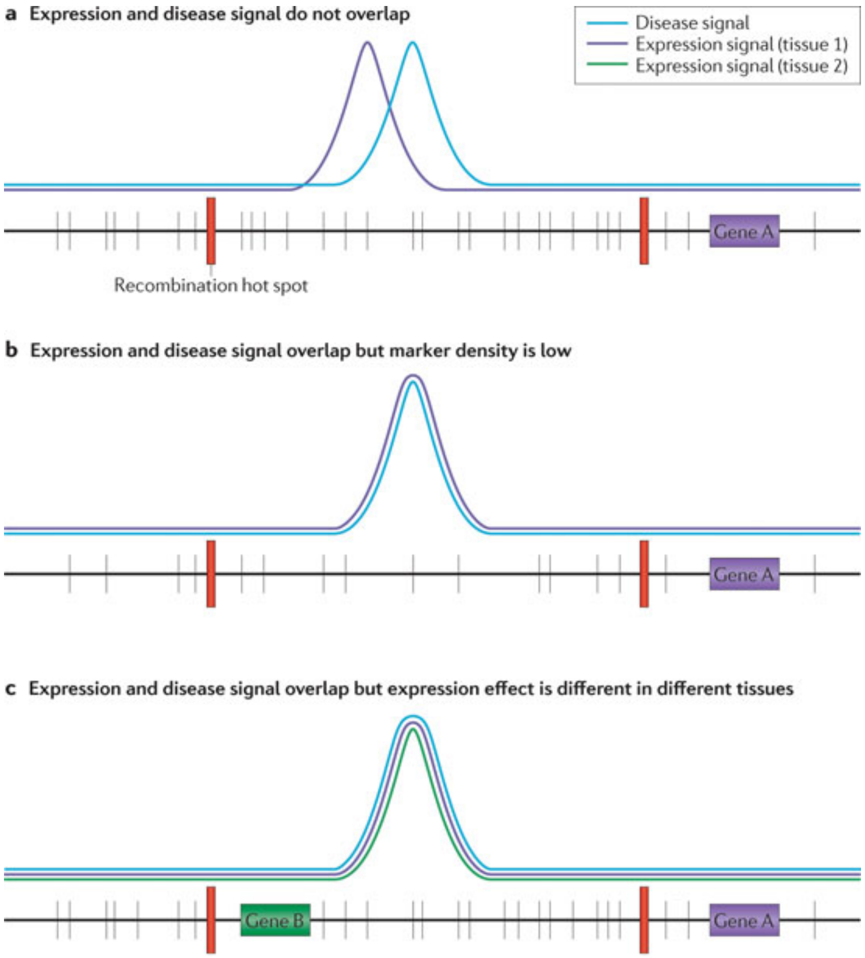
  - Stegle, PLoS Computational Biology, 2010



- Surrogate variable analysis, has increased by 20% eQTL discovery

  - Leek, PLoS Genetics, 2007

Why are biological and technological contexts important for understanding eQTL role in disease?



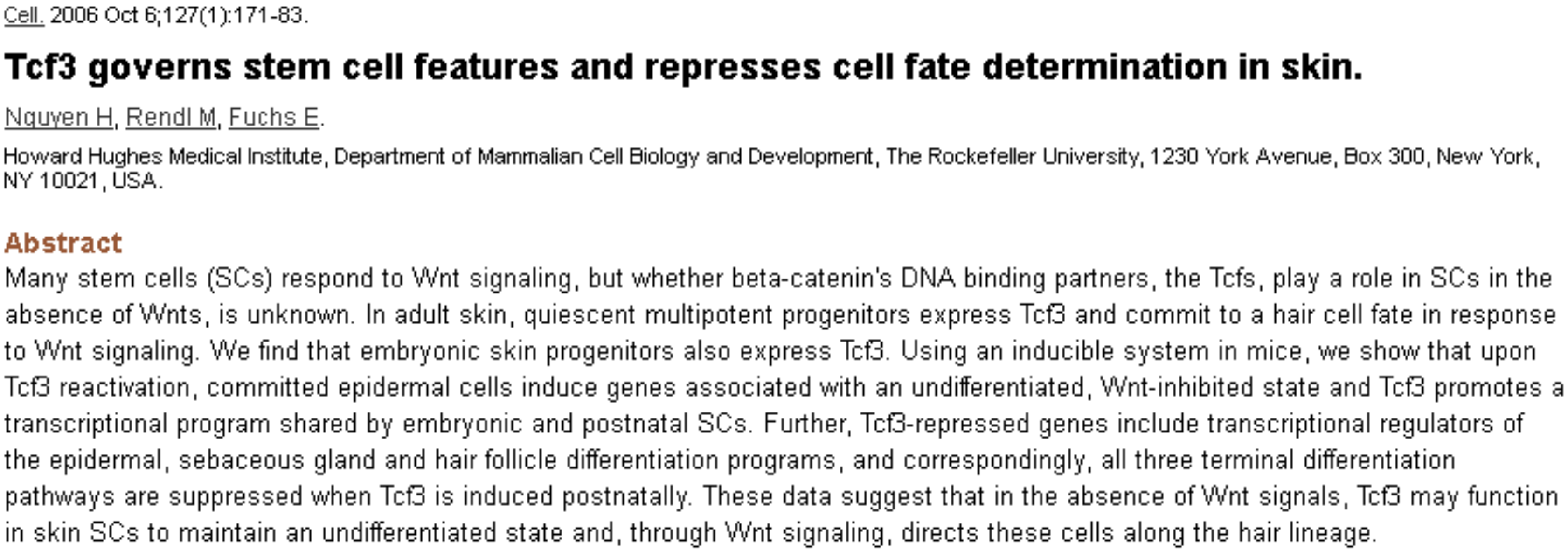
a Expression and disease signal do not overlap
Disease signal
Expression signal (tissue 1)
Expression signal (tissue 2)
Gene A
Recombination hot spot
b Expression and disease signal overlap but marker density is low
Gene A
c Expression and disease signal overlap but expression effect is different in different tissues
Gene B
Gene A

Nature Reviews | Genetics

eQTL data can open up new biology through reverse genetic approaches

- Without traits and disease we can find variants influencing expression level.
- We can speculate and investigate what these effects might do.

Class activity: What are my TCF3 variants doing



Cell. 2006 Oct 6;127(1):171-83.

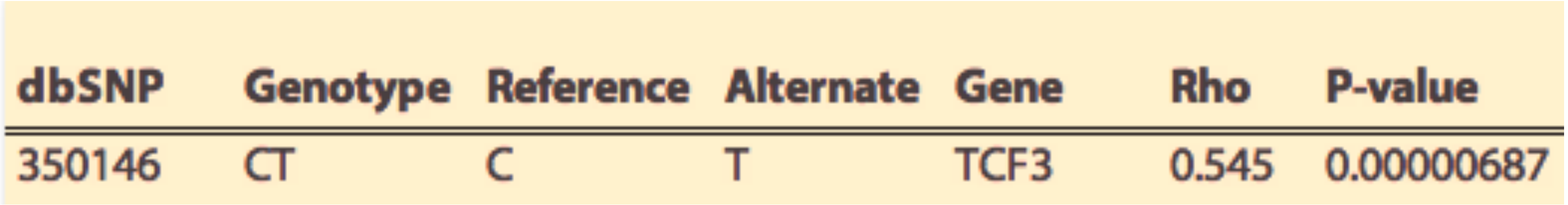
Tcf3 governs stem cell features and represses cell fate determination in skin.

Nguyen H, Rendl M, Fuchs E.

Howard Hughes Medical Institute, Department of Mammalian Cell Biology and Development, The Rockefeller University, 1230 York Avenue, Box 300, New York, NY 10021, USA.

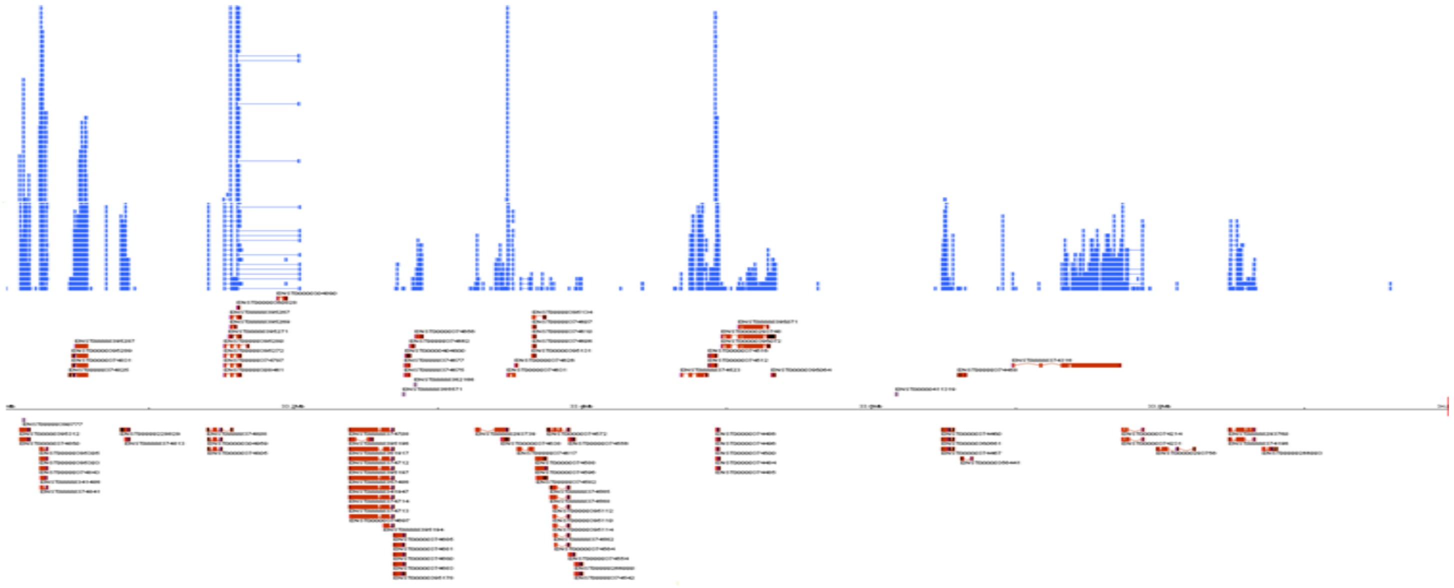
Abstract

Many stem cells (SCs) respond to Wnt signaling, but whether beta-catenin's DNA binding partners, the Tcfs, play a role in SCs in the absence of Wnts, is unknown. In adult skin, quiescent multipotent progenitors express Tcf3 and commit to a hair cell fate in response to Wnt signaling. We find that embryonic skin progenitors also express Tcf3. Using an inducible system in mice, we show that upon Tcf3 reactivation, committed epidermal cells induce genes associated with an undifferentiated, Wnt-inhibited state and Tcf3 promotes a transcriptional program shared by embryonic and postnatal SCs. Further, Tcf3-repressed genes include transcriptional regulators of the epidermal, sebaceous gland and hair follicle differentiation programs, and correspondingly, all three terminal differentiation pathways are suppressed when Tcf3 is induced postnatally. These data suggest that in the absence of Wnt signals, Tcf3 may function in skin SCs to maintain an undifferentiated state and, through Wnt signaling, directs these cells along the hair lineage.



dbSNP	Genotype	Reference	Alternate	Gene	Rho	P-value
350146	CT	C	T	TCF3	0.545	0.00000687

Next generation sequencing has increased our ability to survey the transcriptome.

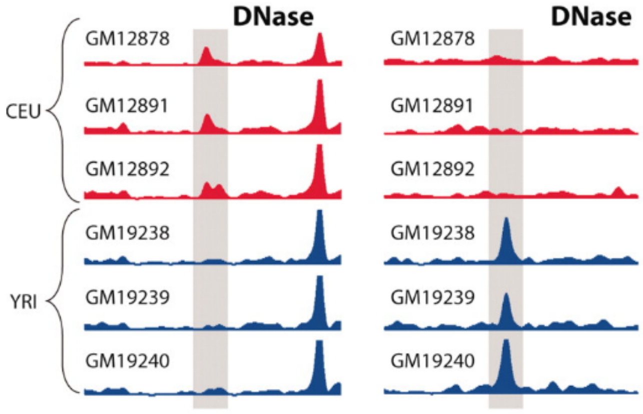


RNA-Seq

Montgomery, Nature 2010
Pickrell, Nature 2010

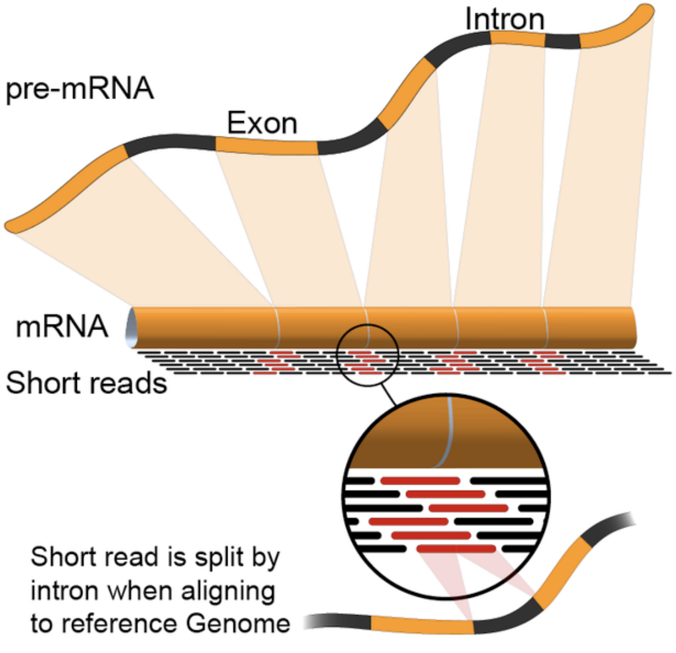
ChIP-Seq

McDaniell, Science 2010

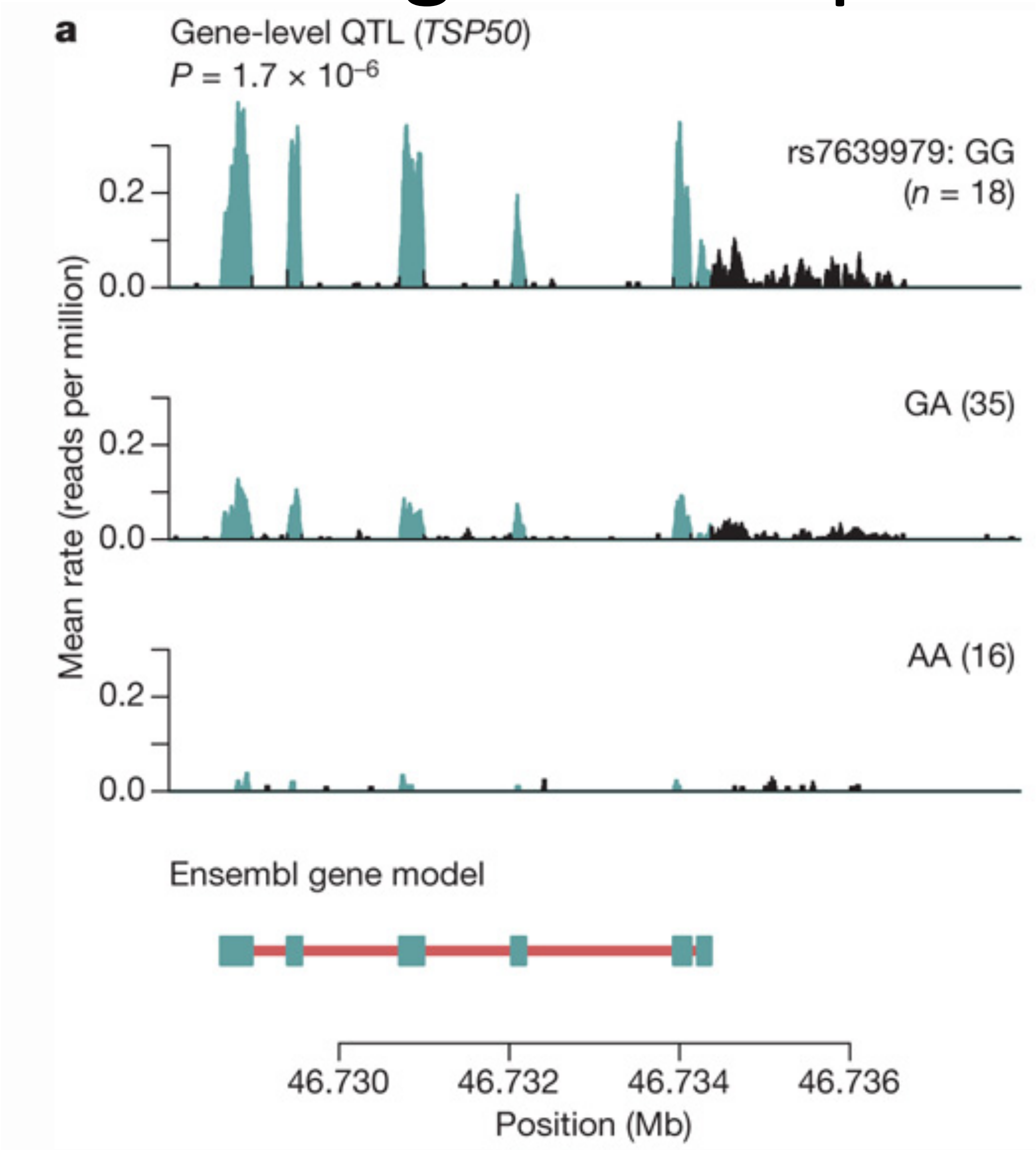


What is RNA-seq

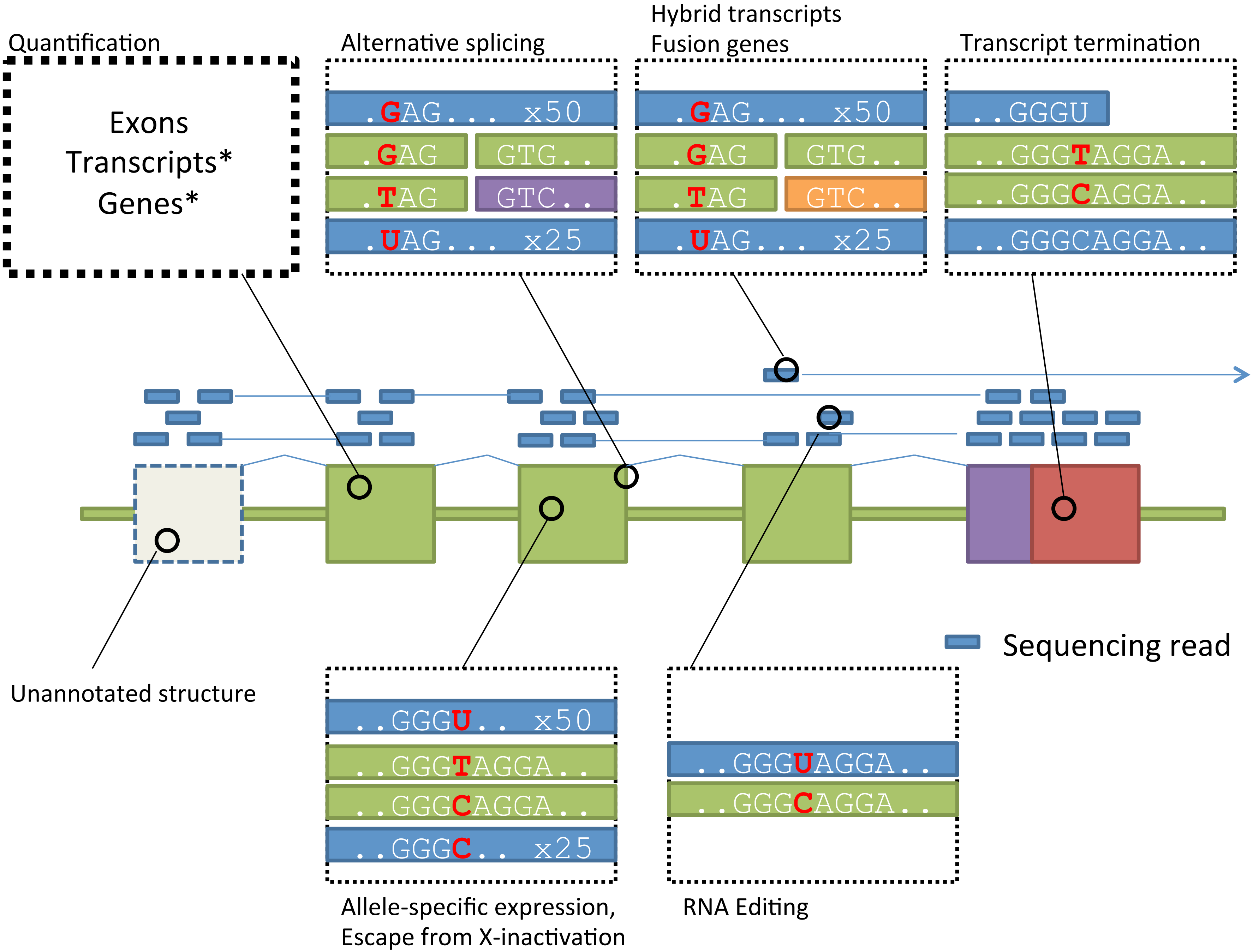
High-throughput sequencing of cDNA to
understand/quantify a sample's gene expression profile
Output: millions of short, single or paired-end sequences (reads)



Genetics of gene expression using RNA-Seq



Increased resolution of transcriptome through RNA- sequencing



Quantification

Exons
Transcripts*
Genes*

Alternative splicing

Hybrid transcripts
Fusion genes

Transcript termination

Sequencing read

Unannotated structure

Allele-specific expression,
Escape from X-inactivation

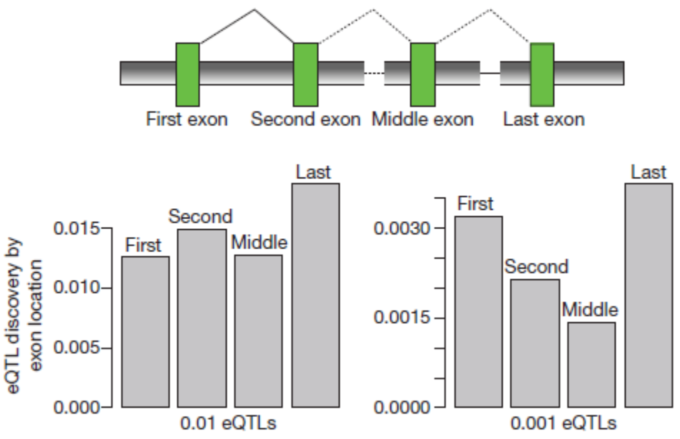
RNA Editing

RNA-seq provides resolution of more QTLs

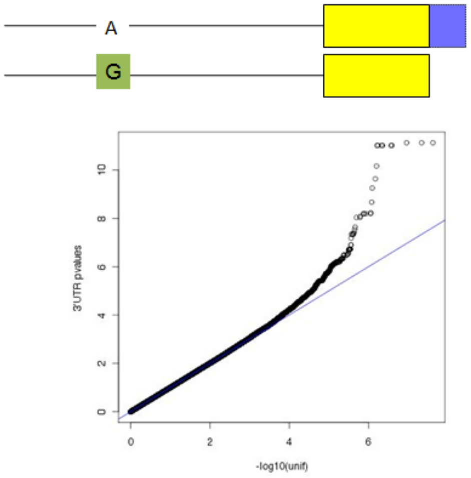
RNA-sequencing in 60 Europeans (HapMap genotypes; LCLs)

Found 2x more expression Quantitative Trait Loci (eQTLs) and...

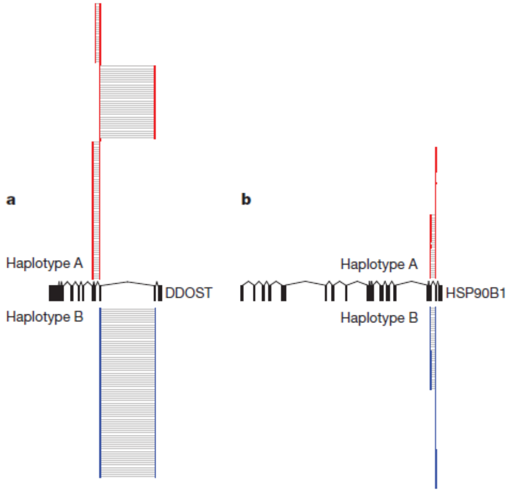
Exon-eQTLs



UTR Length-QTLs

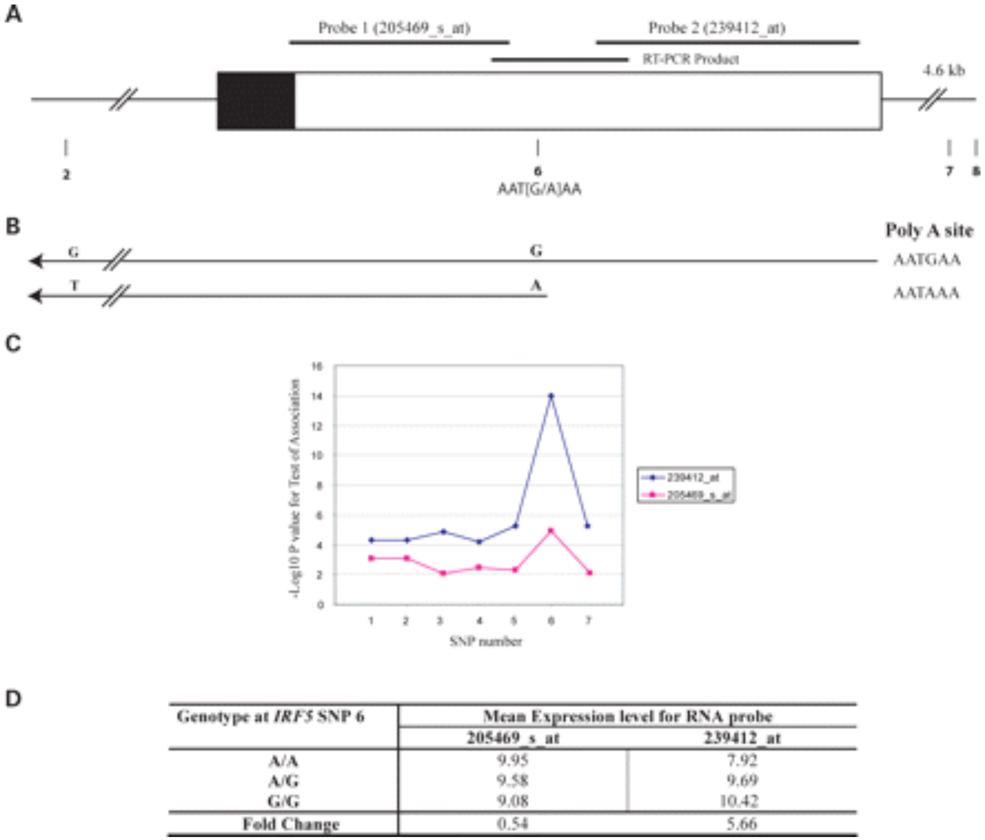


Splicing eQTLs



Rare eQTLs with allele specific expression-based approaches

Class activity: rs10954213 creates a functional polyadenylation site



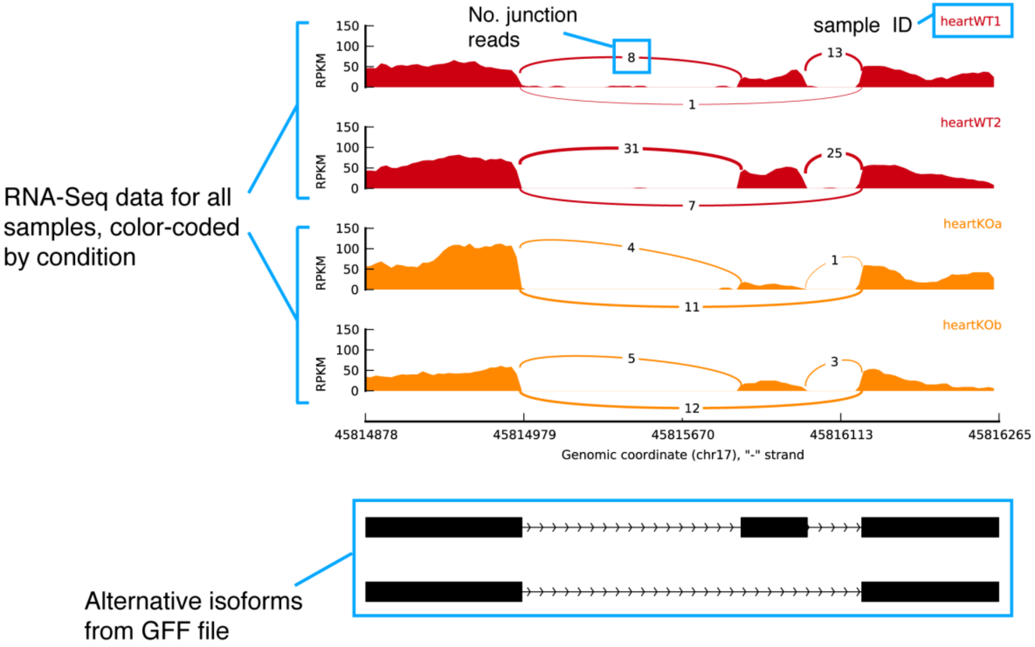
Genotype at *IRF5* SNP 6	Mean Expression level for RNA probe	
	205469_s_at	239412_at
A/A	9.95	7.92
A/G	9.58	9.69
G/G	9.08	10.42
Fold Change	0.54	5.66

The A allele of rs10954213 creates a functional polyadenylation site and the A genotype correlates with increased expression of a transcript variant containing a shorter 3′-UTR. Expression levels of transcript variants with the shorter or longer 3′-UTRs are inversely correlated. Our data support a new mechanism by which an *IRF5* polymorphism controls the expression of alternate transcript variants which may have different effects on interferon signaling

Cunninghame Graham et al., HMG, 2007

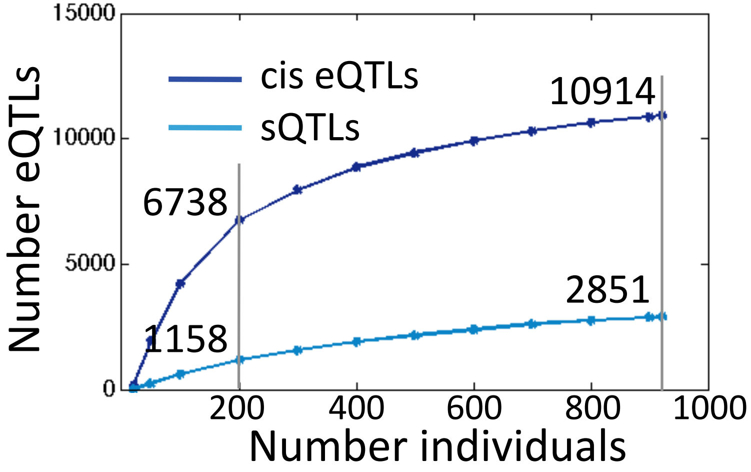
Splicing eQTL

Can investigate relative transcript ratios or reads across junctions.



Katz et al, Nature Methods, 2010

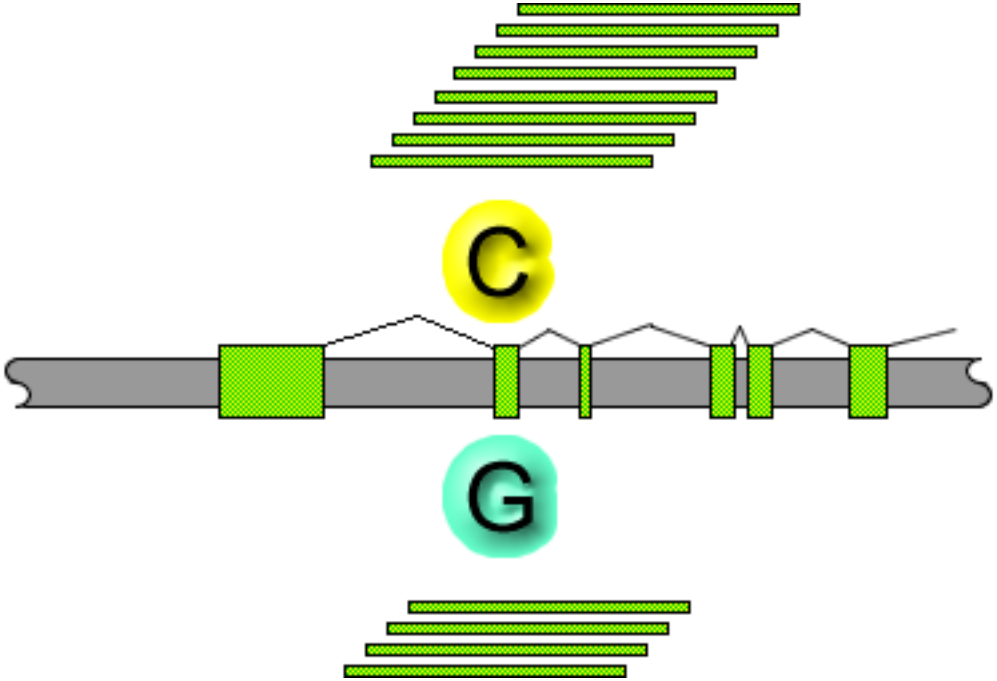
- Splicing also affected for many genes



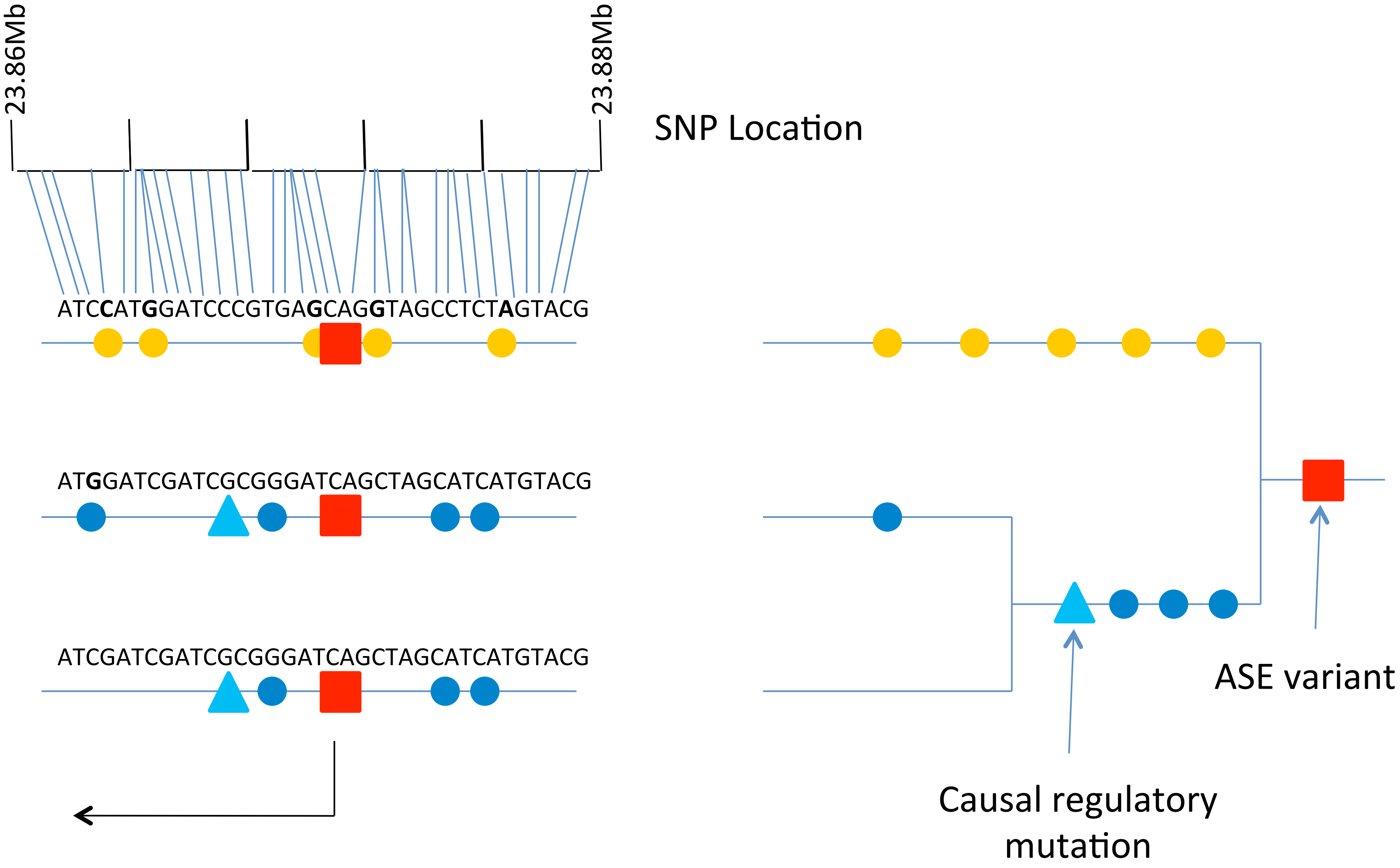
Battle et al, submitted

Advantages of ASE

- Test within an individual allelic imbalance, given one has sufficient reads.

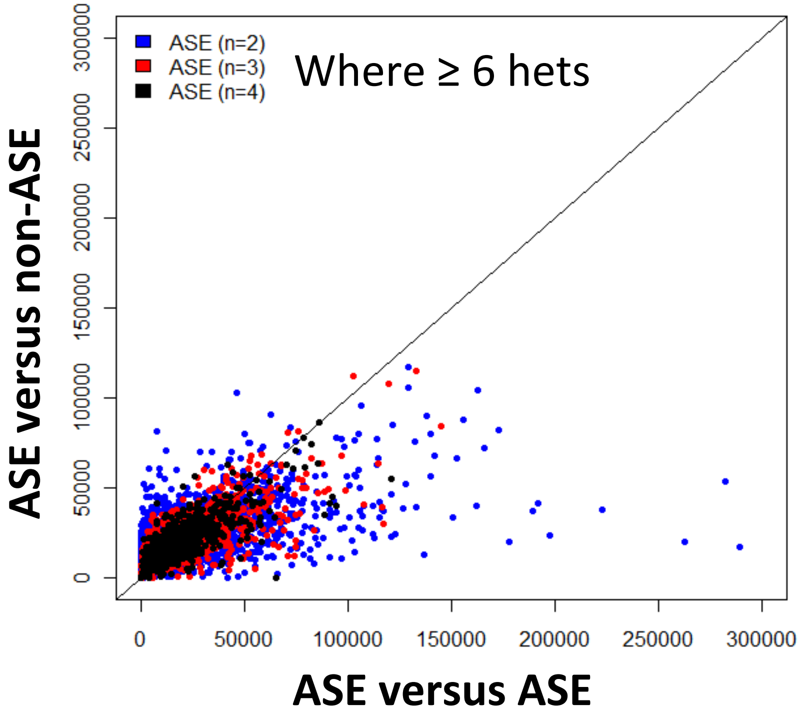


Looking for rare regulatory haplotypes


23.86Mb
23.88Mb
SNP Location
ATC**C**AT**G**GATCCCGTGA**G**CAG**G**TAGCCTCT**A**GTACG
AT**G**GATCGATCGCGGGATCAGCTAGCATCATGTACG
ATCGATCGATCGCGGGATCAGCTAGCATCATGTACG
ASE variant
Causal regulatory mutation

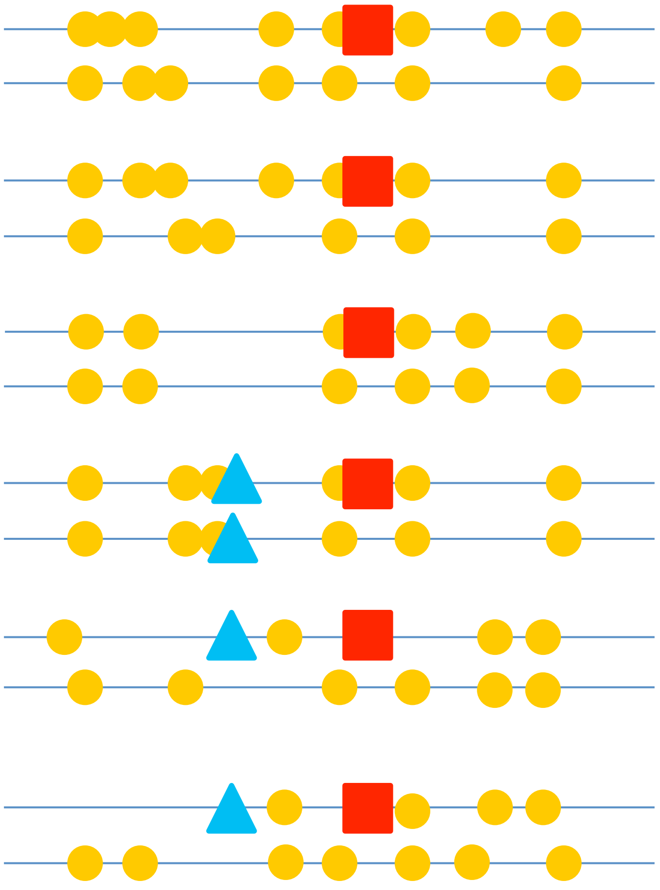
Can measure extent of haplotype homozygosity
since shared ASE assumed to have common genealogy

Evidence of recent and rare eQTLs



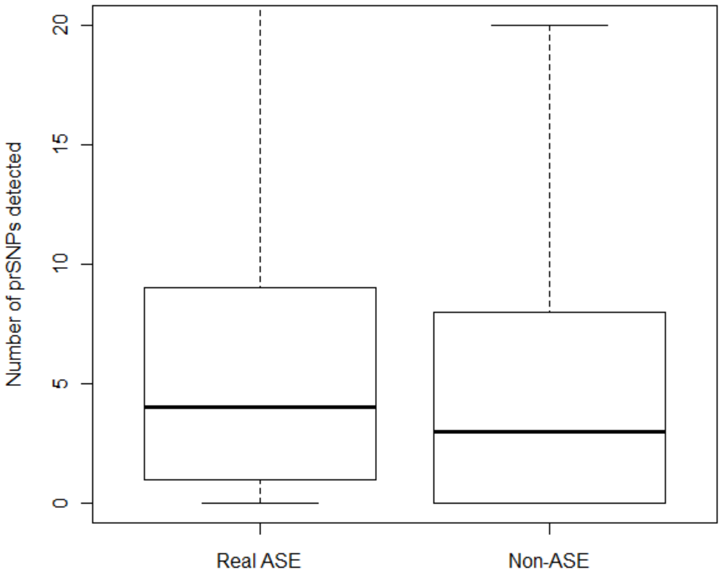
When ASE individuals compared, we observed longer tracts of haplotype homozygosity

Can we find the recent and rare causal regulatory variants?



POOL OF INDIVIDUALS	Putative regulatory SNP
NO ASE	- / -
NO ASE	- / -
NO ASE	- / -
NO ASE	▲ / ▲
ASE	▲ / -
ASE	▲ / -

More putative regulatory SNPs found for real ASE versus non-ASE

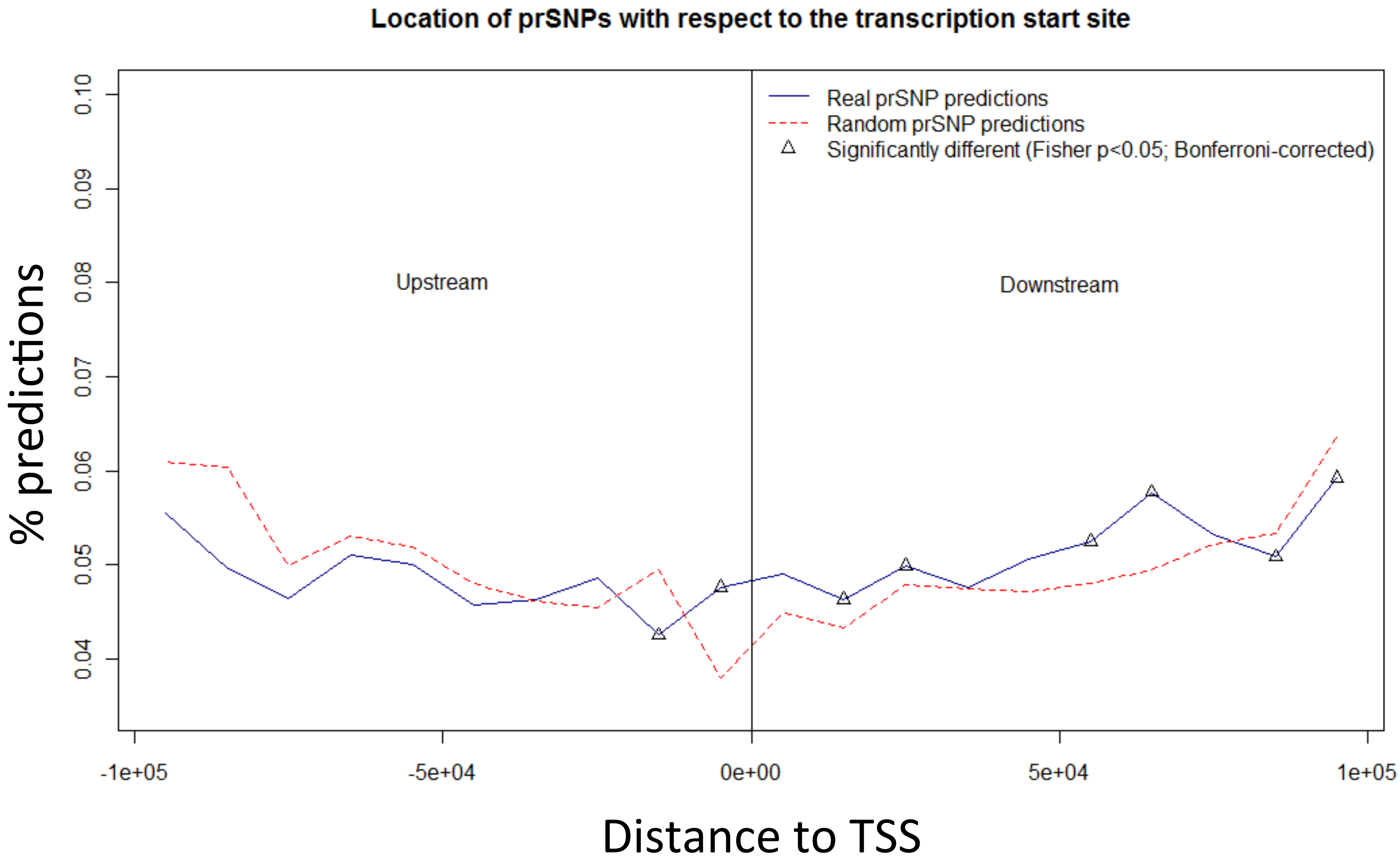


Mann Whitney
p<2e-16

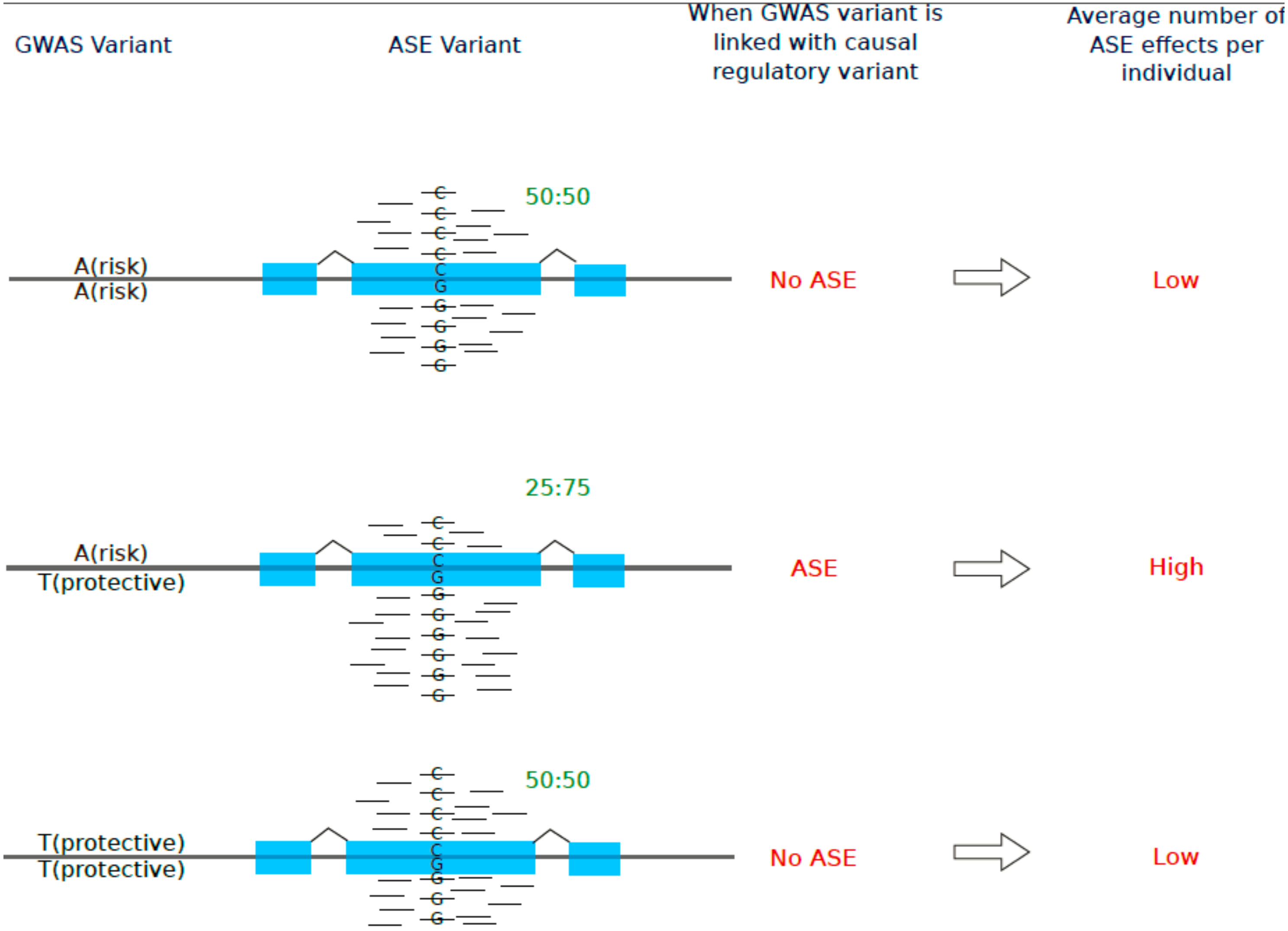
We see 1 more prSNP on average in real ASE versus non-ASE

Montgomery, PLoS Genetics, 2011

Putative regulatory SNPs are enriched around TSS



Using ASE to detect rare and common variation underlying GWA-variants



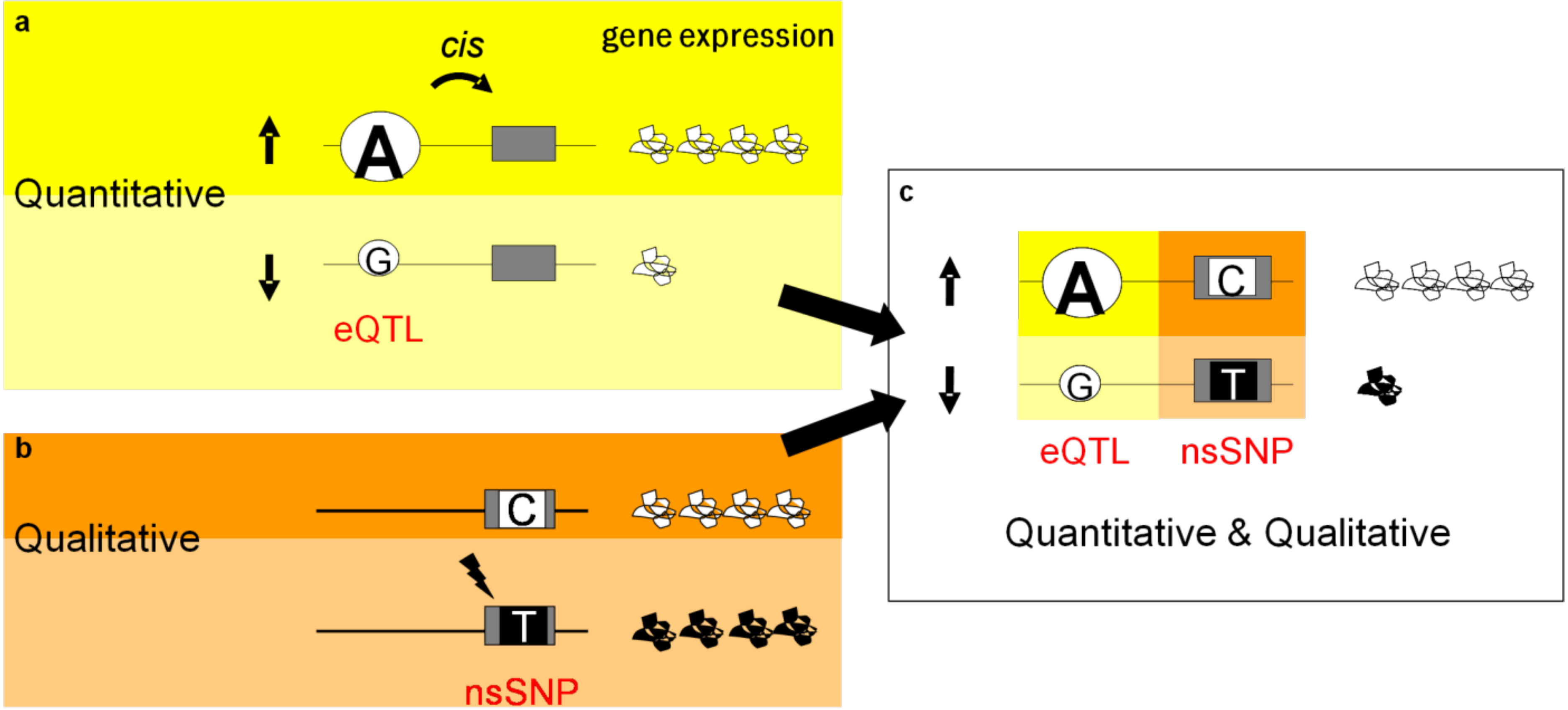
GWAS Variant	ASE Variant	When GWAS variant is linked with causal regulatory variant		Average number of ASE effects per individual
A(risk) A(risk)	50:50	No ASE	⇨	Low

A(risk)
T(protective)
	25:75	ASE	⇨	High

T(protective) T(protective)	50:50	No ASE	⇨	Low

Conde et al, AJHG, Jan 2013

Abundant epistasis between regulatory and protein coding variation

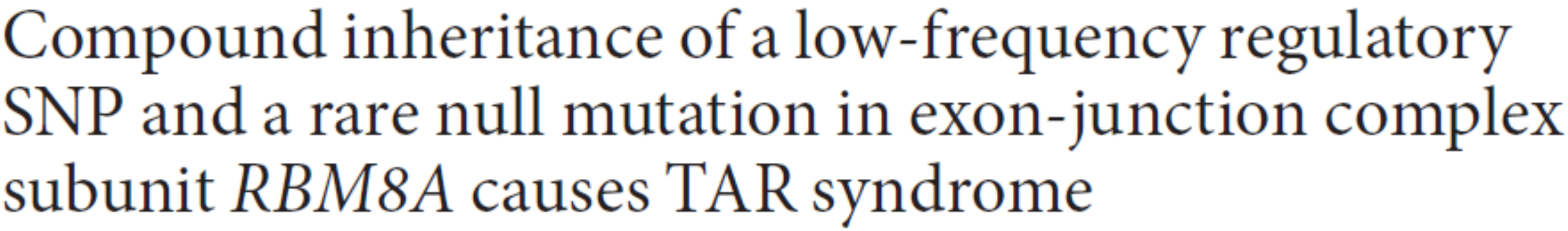


18.2% (1502 of 8233) Dimas, 2008

46.2% nonsynonymous sites where ASE can be detected are significant in 1 indiv.

Montgomery et al., PLoS Genetics, 2011
Lappalainen et al., AJHG, 2011

Compound inheritance of regulatory and coding polymorphism causes disease



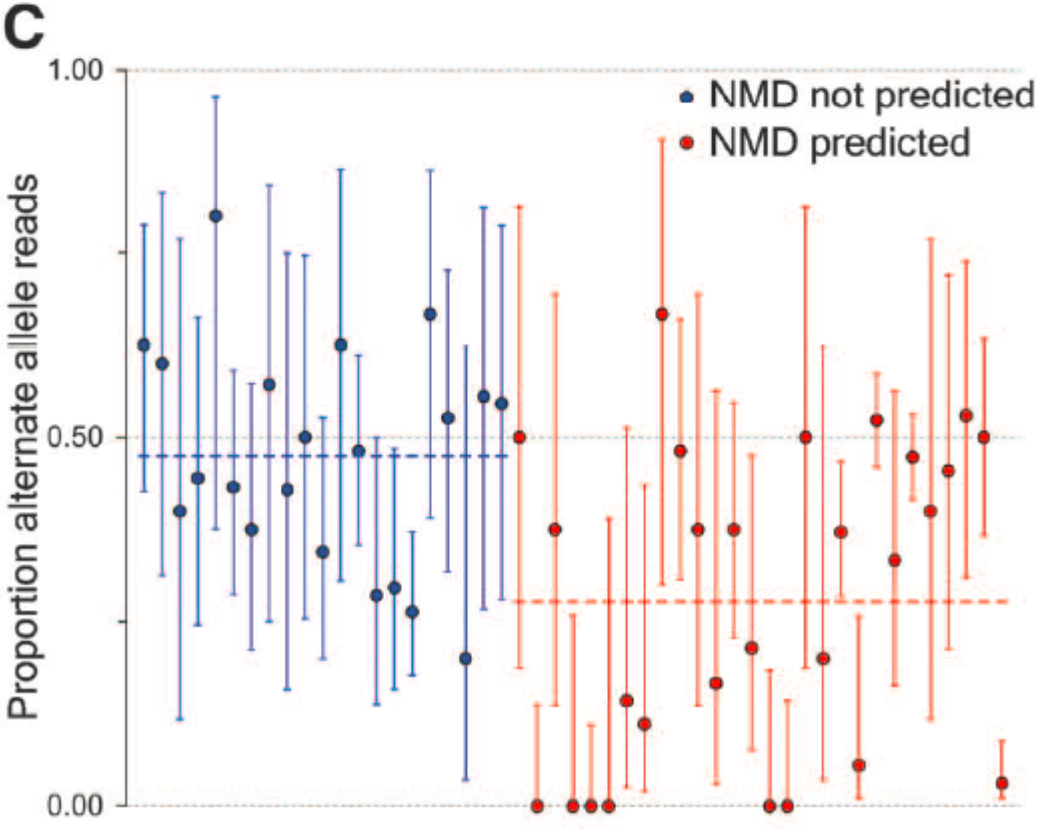
Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit *RBM8A* causes TAR syndrome

The exon-junction complex (EJC) performs essential RNA processing tasks1–5. Here, we describe the first human disorder, thrombocytopenia with absent radii (TAR)6, caused by deficiency in one of the four EJC subunits.

The thrombocytopenia with absent radii (TAR) syndrome is characterized by a reduction in the number of platelets (the cells that make blood clot)

Albers, Nature Genetics, 2012

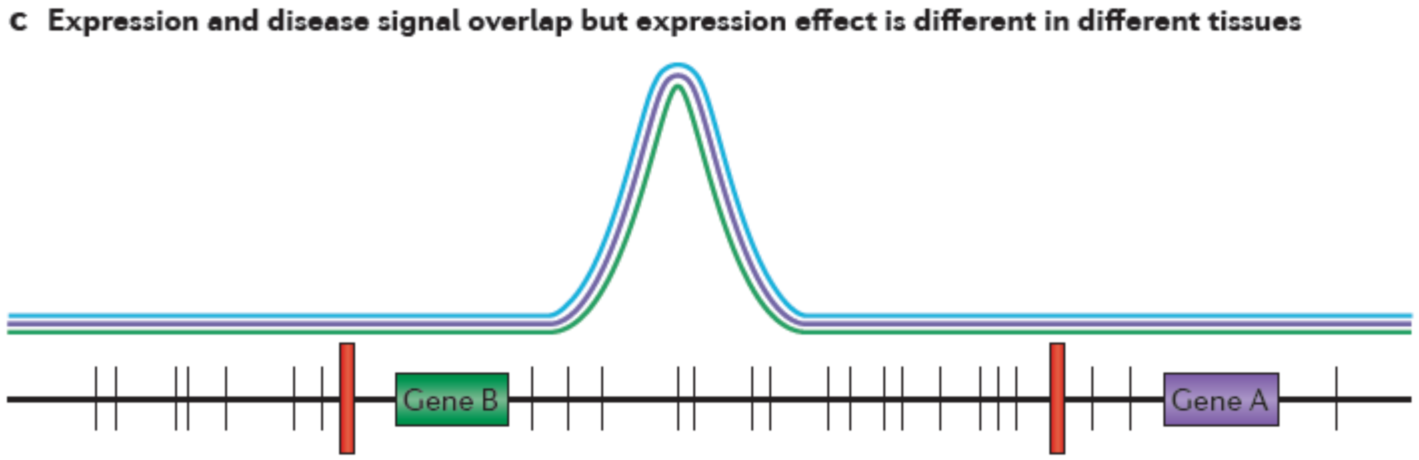
Interpreting completed genomes with gene expression



MacArthur et al., Science, 2012

Understanding disease mechanism

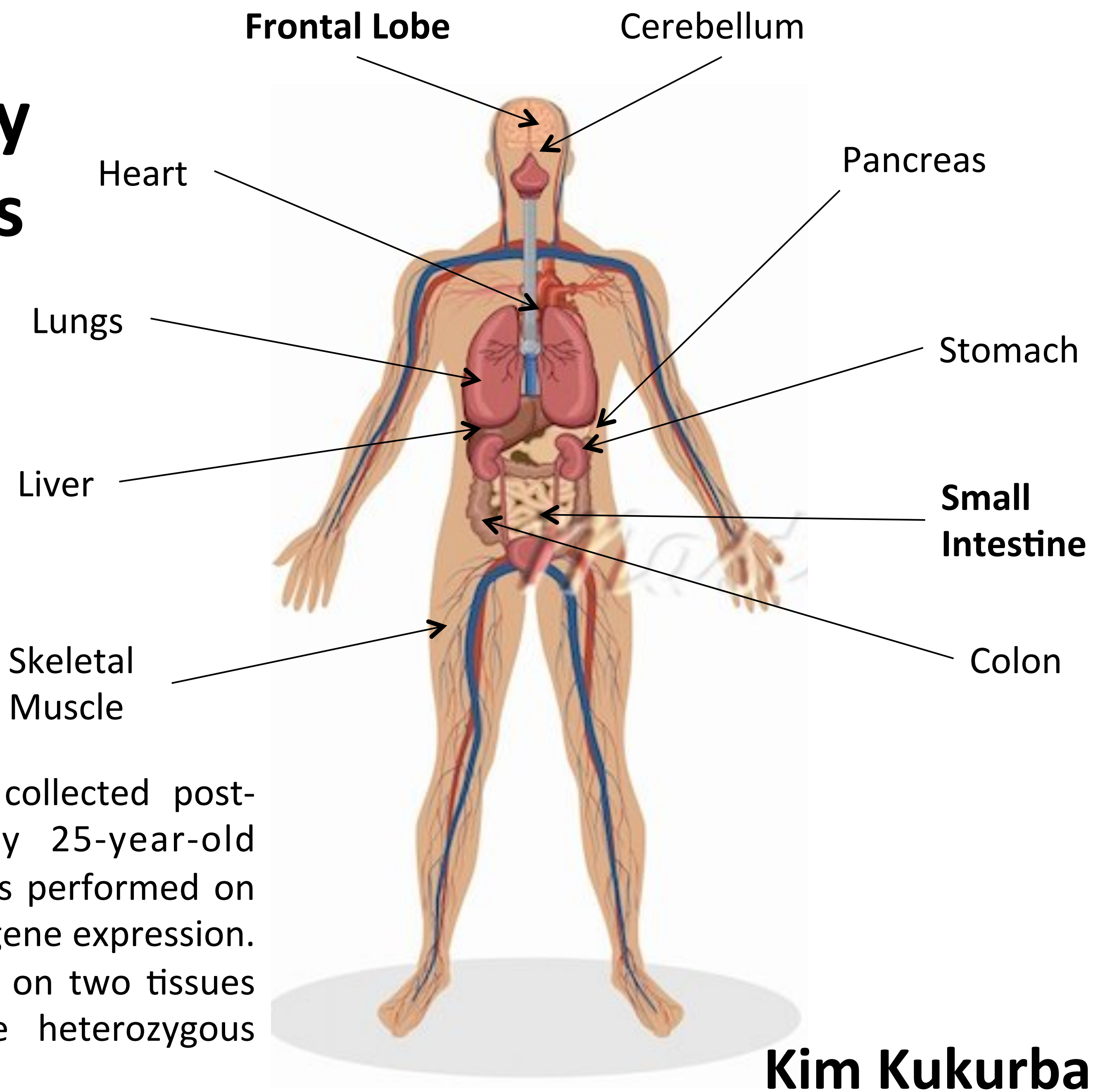
Predictive value of gene expression dependent on proximity to pathological tissue


c Expression and disease signal overlap but expression effect is different in different tissues
Gene B
Gene A

We have limited understanding of the Type I and II error rate

However, a lack of sharing may allow us to discover the pathological tissue

Allelic heterogeneity of deleterious variation



Frontal Lobe

Cerebellum

Heart

Pancreas

Lungs

Stomach

Liver

Small Intestine

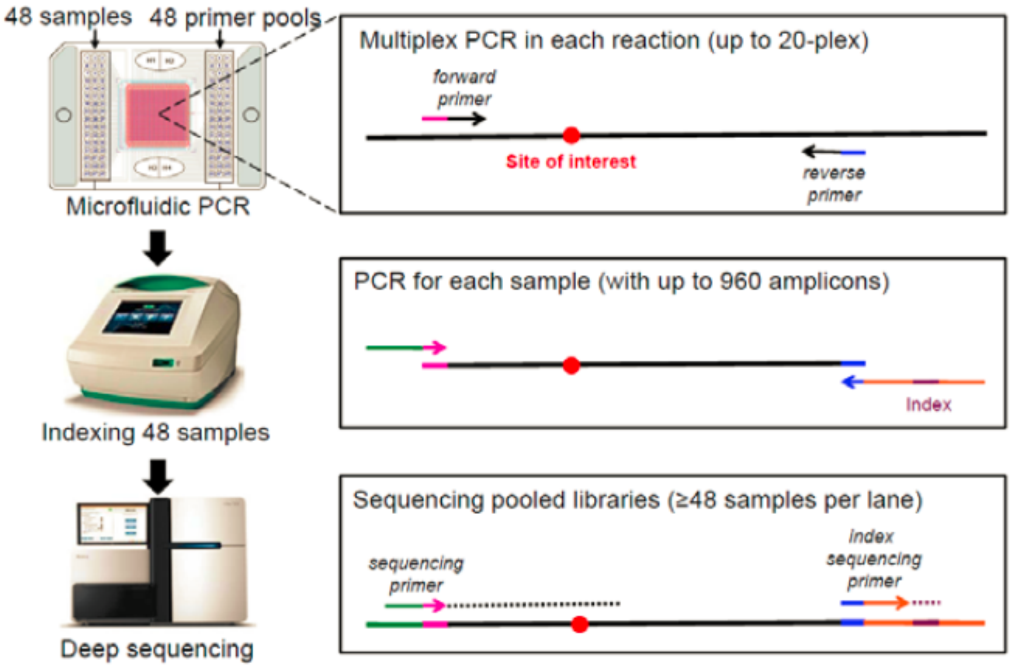
Skeletal Muscle

Colon

Ten human tissues were collected post-mortem from a healthy 25-year-old Chinese male. RNA-Seq was performed on the ten tissues to quantify gene expression. Exome-Seq was performed on two tissues (bolded) to ascertain the heterozygous sites in the genome.

Kim Kukurba

Targeted ASE using the microfluidics-based multiplex PCR and deep sequencing (mmPCR-Seq)

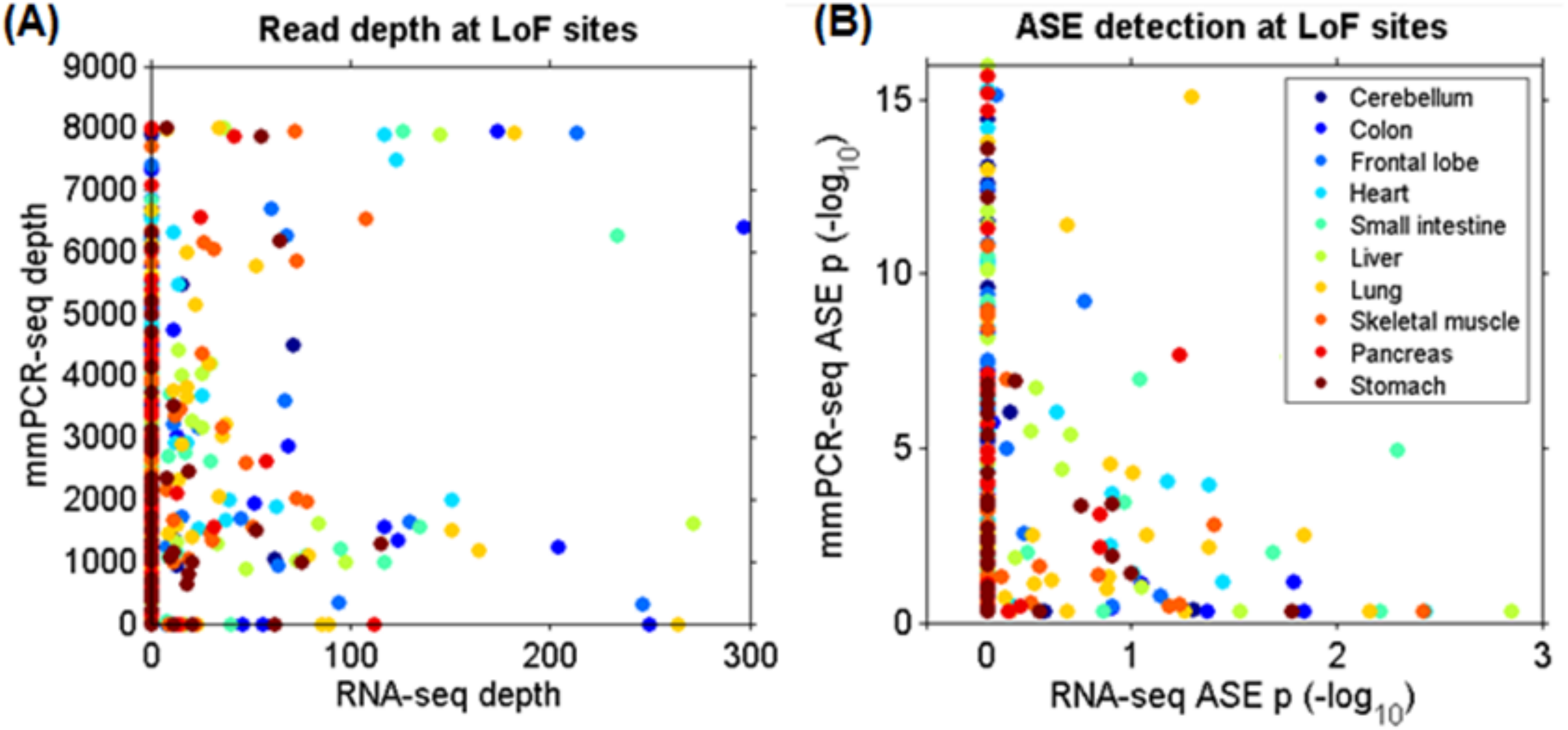


Zhang et al., submitted

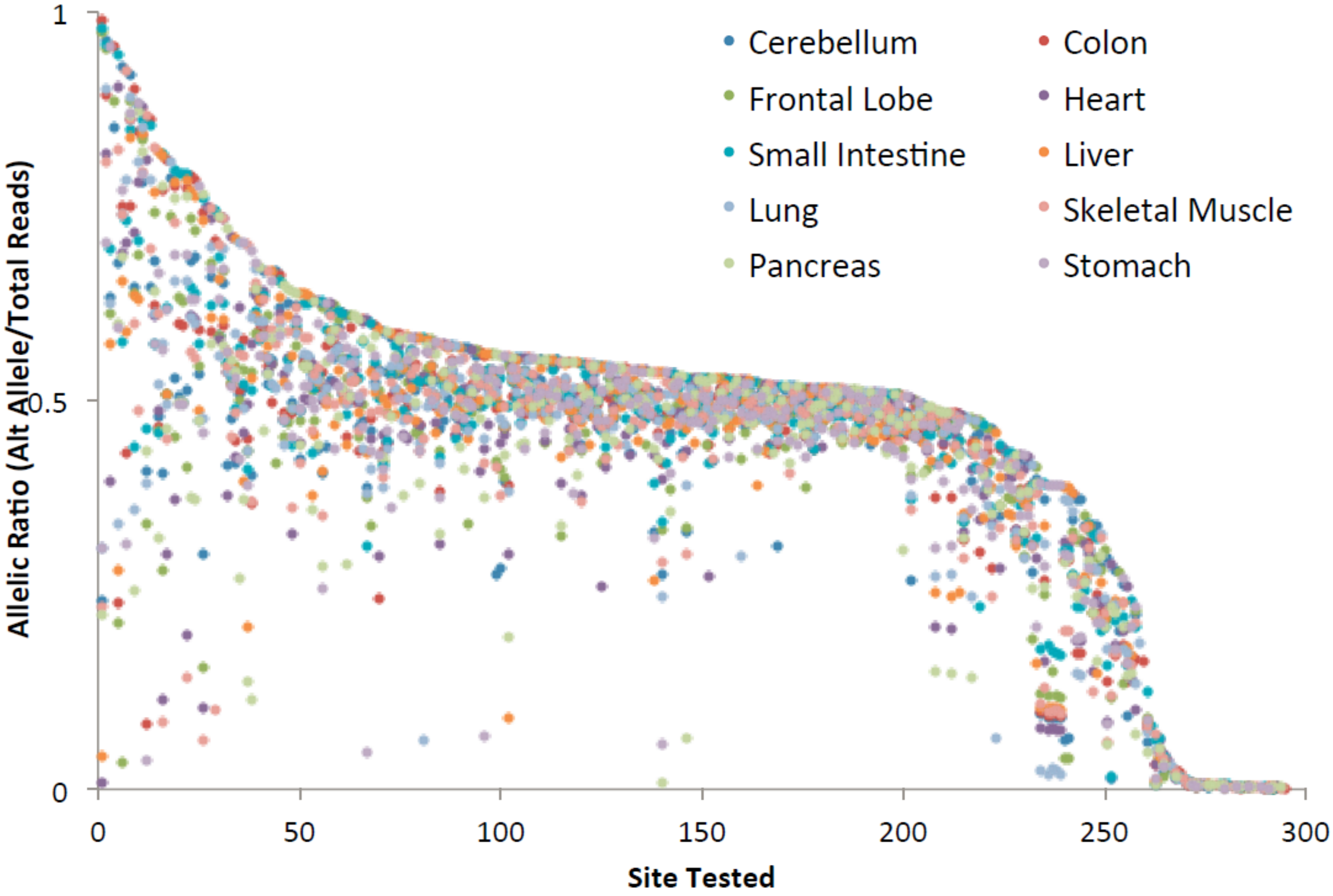
Application of mmPCR-Seq to deleterious and LoF alleles

- Selected all rare and predicted deleterious and damaaging nsSNPs (74 sites)
- Selected all complete stop-gain sites (50 sites)
- Control sites (160)

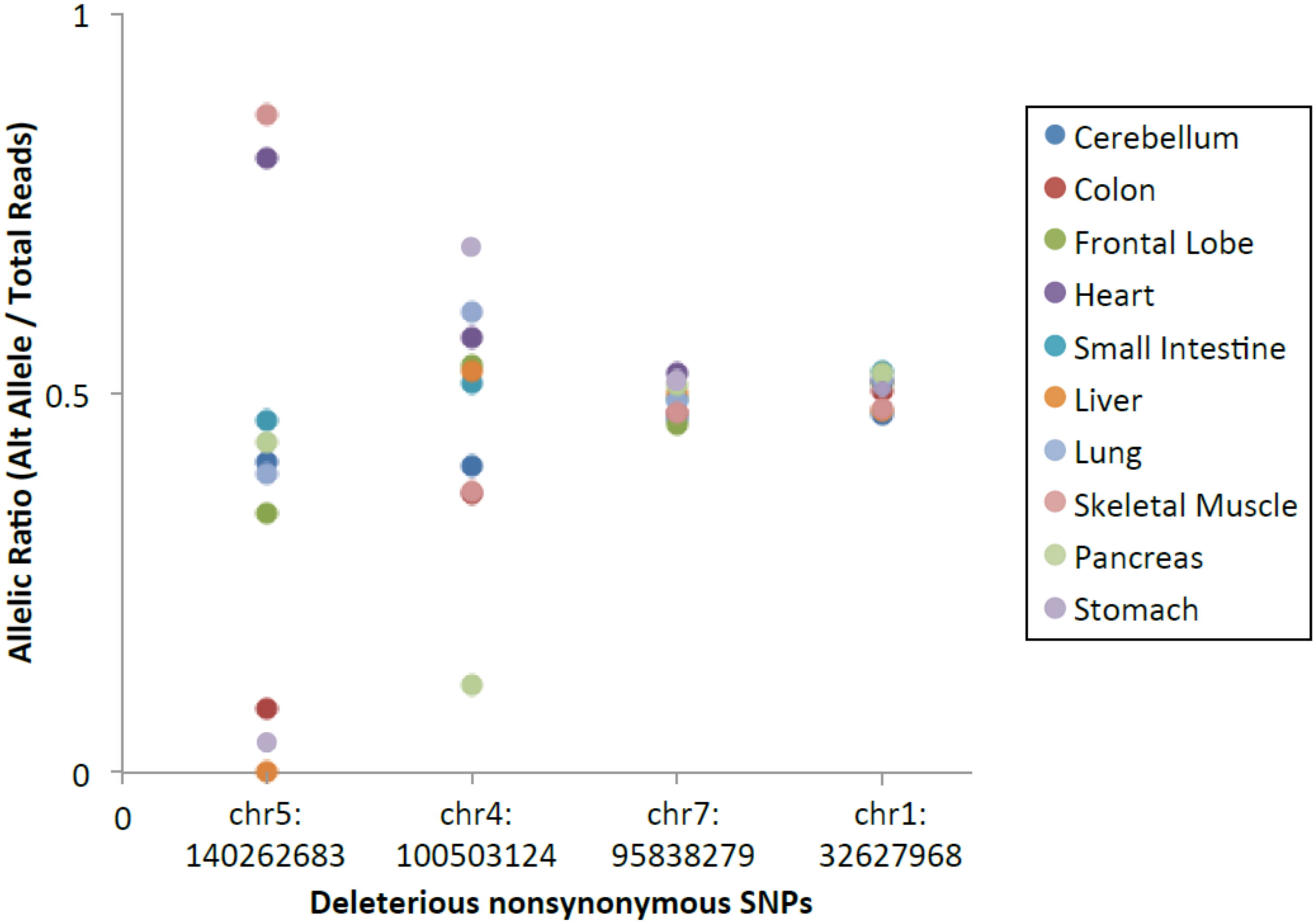
mmPCR-Seq versus RNA-Seq



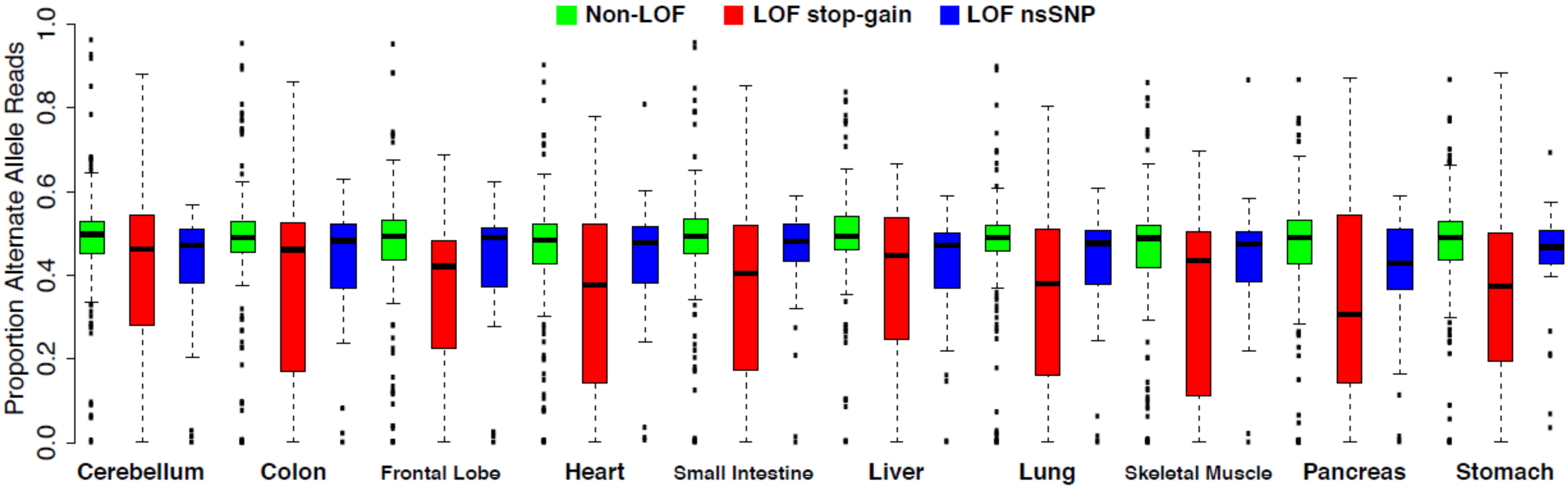
Allelic ratios across tissues with mmPCR-Seq



Variable expression of deleterious alleles in different tissues



Stop gain alleles exhibit lower expression across tissues



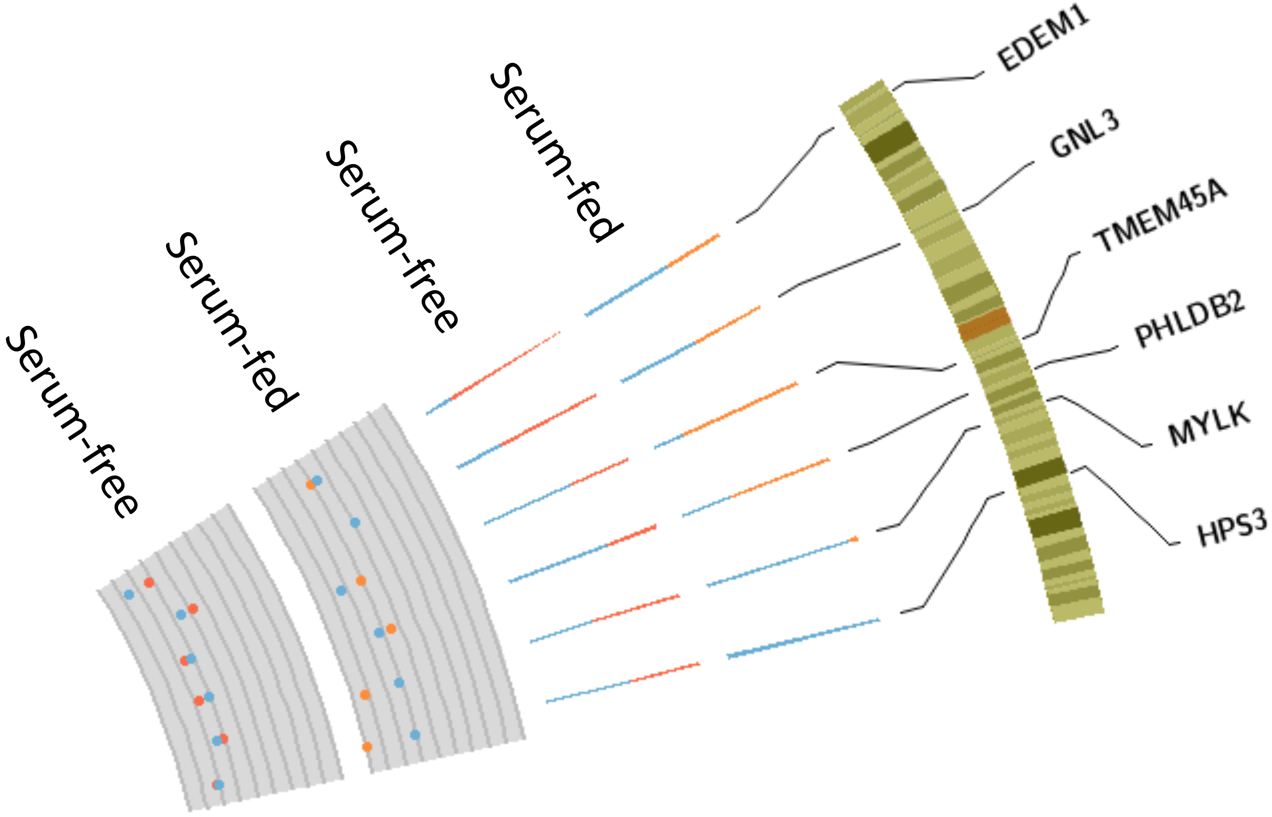
Identification of ASE in genes with rare, deleterious nsSNPs

No ASE	NLRP3 SLC8A3 DOCK8 CSPG4 UTP20 DTX1 FAM129B CEP128 ANKRD27 GPR75 MTMR9 TMEM168 ALPK1 SLC25A13 EAPP KPNA6 C12orf10 RUSC2 XAB2 EIF2AK4 MKL2
Shared ASE	FMO3 ACOT11 SRGAP2 NOTCH2NL CACNA1A UNC45A PLIN1 ABCA7 CLIP4 AQP7 LMTK3 USP49 TBX21 FAM120A
Variable ASE	*PCDHA13* CIDEA MTTP CLEC11A TRPC3

Kim Kukurba

Using RNA-Sequencing to survey differential allelic expression in cardiovascular disease

Compared serum-starved and serum-fed coronary artery smooth muscle cells



Olga Sazonova and Thomas Quertermous

How will gene expression influence decisions in the clinic?

Build cellular models of disease

Survey diagnostic responses to treatments

Identify diverse disease mechanisms; move us beyond protein coding mutations alone

Identify pathological tissues

Allow us to identify effects (or transferability) in different populations

Classify undiagnosed conditions

Cost-effective

“**The field will transition from doing primarily association work to figuring out what implicated variants do biologically**.”

David Goldstein, Director of the Center for Human Genome Variation, Duke University, *Nature*, Feb 2012

Projects in the Montgomery lab

EQTL and reverse genetics approaches in Sardinia – Mauro Pala, Zach Zappala, Xin Li

Rare non-coding variants in a large family – Xin Li, Konrad Karczewski

mmPCR-Seq methods development and application – Rui Zhang, Xin Li, Billy Li

Genetics of gene expression in exosomes – Kevin Smith, Xin Li

Pinpointing causal regulatory variants – Marianne DeGorter

DNaseI causal variant mapping and population genetics – Zach Zappala, William Greenleaf

Indels in 179 genomes (just out in Genome Research)– Gerton Lunter, Oxford

Idiopathic Pulmonary Fibrosis RNA-Seq – Tracy Nance, Glenn Rosen

Population and demographic modeling of allelic effects – Joe Davis, Carlos Bustamante

Long read RNA-Seq and AST – Hoon Cho, Alexis Battle

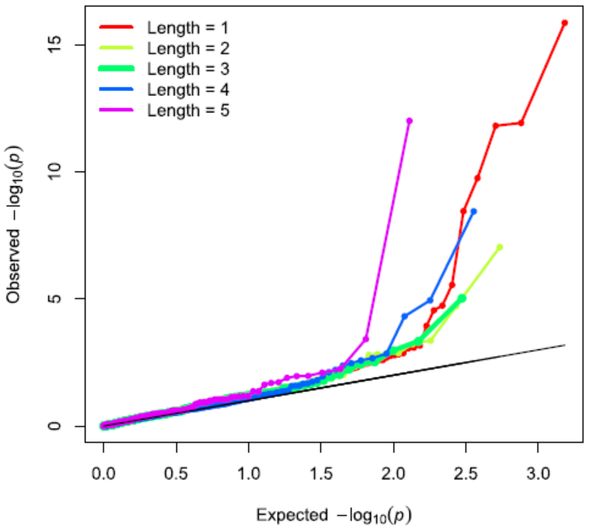
Trans-eQTLs and family NF-kappaB ChipSeq and Disease ChIP-Seq (PNAS, in press) – Konrad Karczewski, Mike Snyder

PML Cancer transcriptomes – Graham Dellaire / Dalhousie

Clinically-actionable fusion gene identification (AJSP) – Tripp Sweeney, Rob West

Gencord RNA-Seq mQTL and eQTL (eLife, in press)– Maria Gutierrez-Arcelus, Manolis Dermitzakis

Metagenomic diagnostic for C. diff.– Niaz Banaei, Merck



montgomerylab.stanford.edu

Further recommended reading:

1) Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis (2010, Nature)

2) 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response (2011, Nature)