

Disclosure for:

ASHG Interactive Workshop: Overview and Interpretation of GTEx Resources: eQTLs and Gene Expression

No Relevant Conflicts to Disclose:

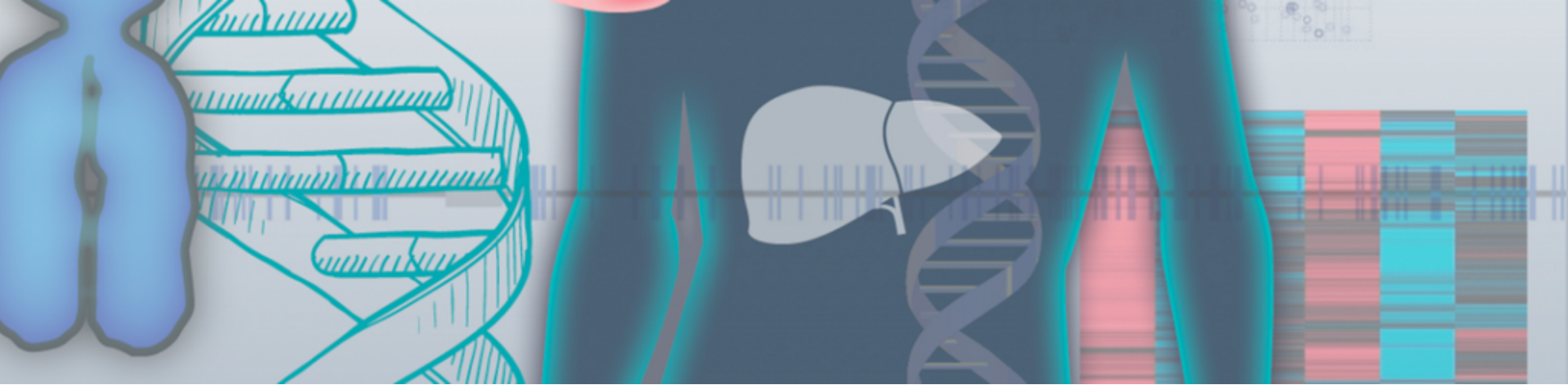
Kristin G. Ardlie

François Aguet

Ayellet V. Segrè

Jared L. Nedzel

Stephen Montgomery



Overview and Interpretation of GTEx Resources: eQTLs and Gene Expression

ASHG 2017 Annual Meeting
10/18/2017

GTEx Workshop Agenda

- Overview of study and data
- Portal demonstration
- Jupyter notebook
- GWAS-eQTL challenges

Association of common DNA variants with diseases and traits

Controls



ACGGGCAATCACGT
ACGGGCAAACACGT
ACGGGCAATCACGT
ACGGGCAAACACGT
ACGGACAATCAAGT
ACGGACAAACAAGT

Cases



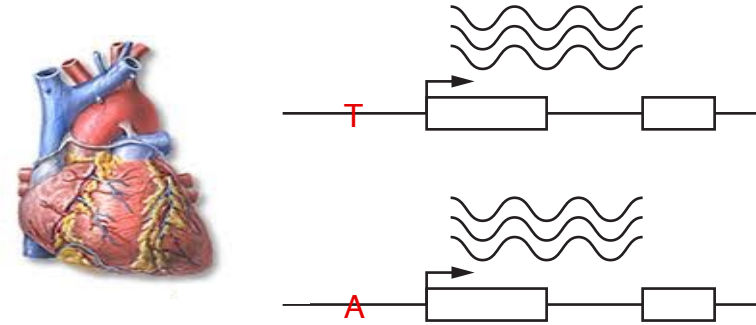
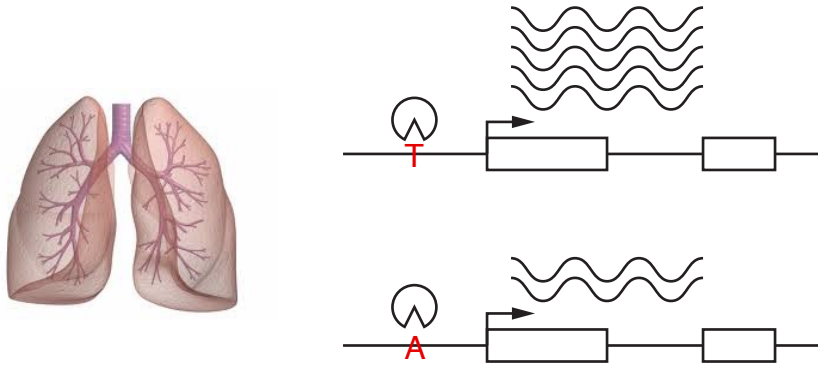
ACGGGCAATCACGT
ACGGACAAACAAGT
ACGGACAAACAAGT
ACGGACAATCAAGT
ACGGACAATCAAGT
ACGGACAAACAAGT

Genome-wide association studies (GWAS) led to discovery of
>10,000 common DNA variants associated with >600 diseases/traits.

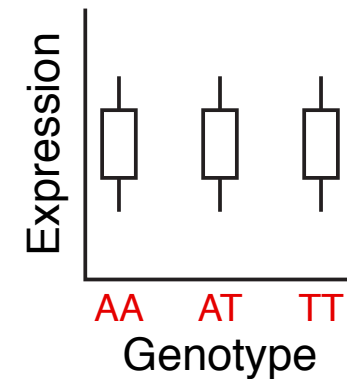
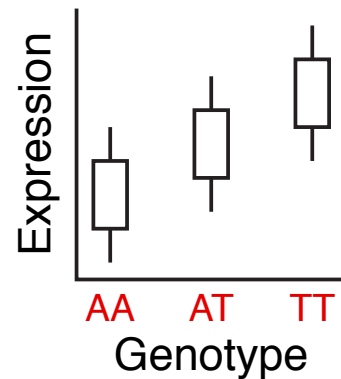
~95% GWAS SNPs located in **non-coding regions**

eQTLs: expression quantitative trait loci

Hypothesis: the functional effect of most (non-coding) GWAS variants is modification of gene expression

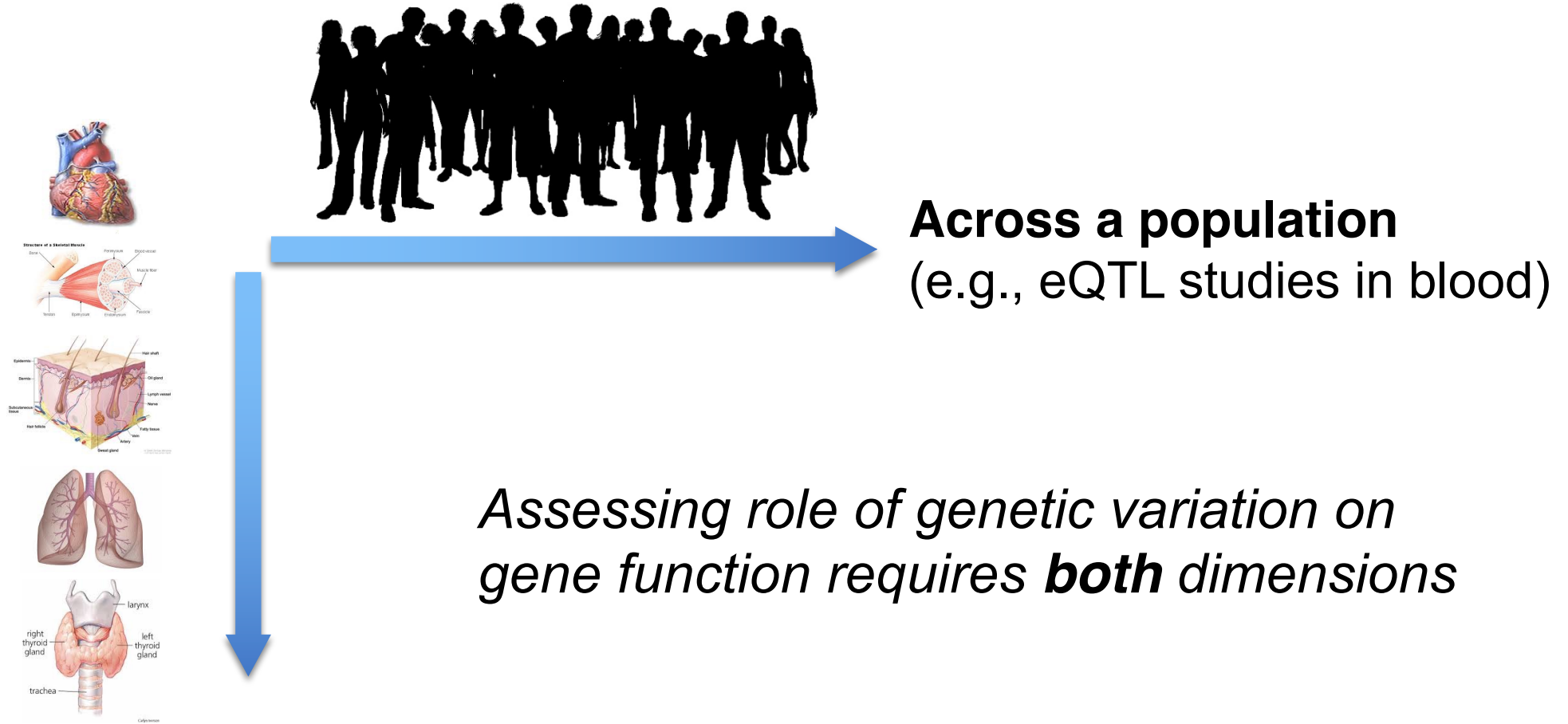


Measured in a population:



Regulatory variation is measured as expression quantitative trait loci (eQTLs)

Regulation of gene expression: multi-tissue and multi-individual



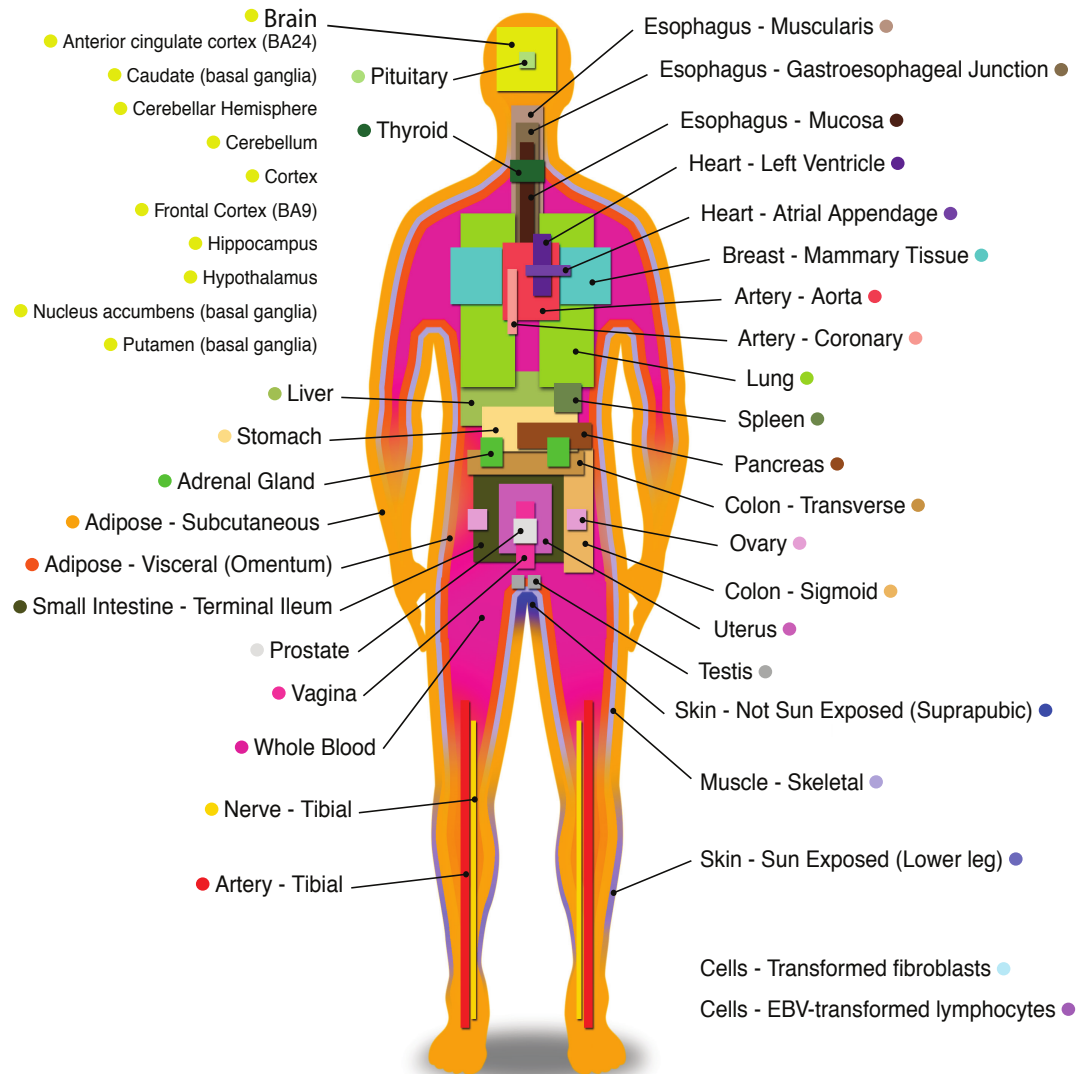
Across tissues or cell types

Functional genomic maps

(e.g., ENCODE, Roadmap Epigenomics)

The Genotype Tissue-Expression project

Atlas of gene expression and eQTLs in non-diseased human tissues from up to 960 recently deceased donors



- **53 tissue sites**

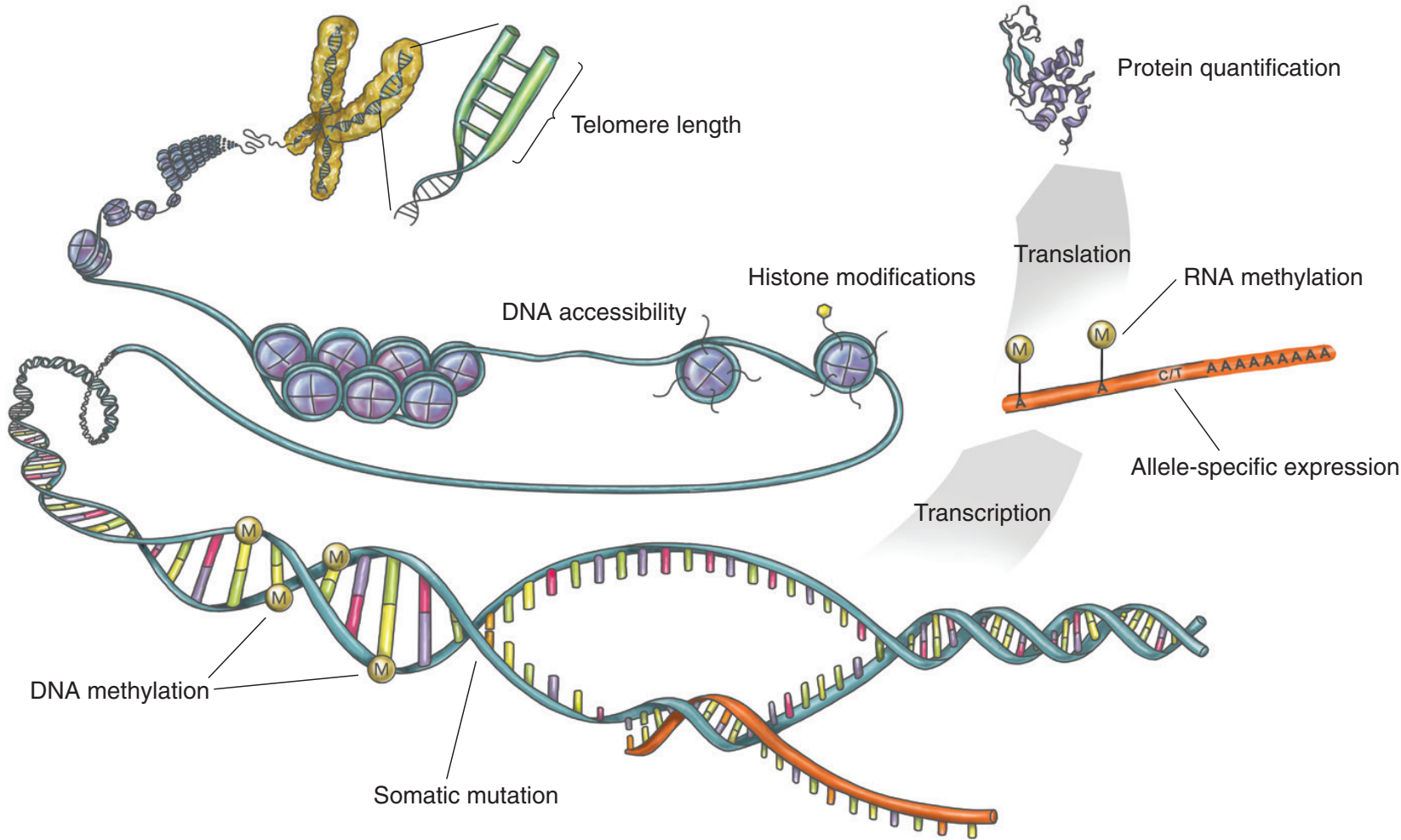
- 11 distinct brain regions
- 2 cell lines

This workshop

- **Core molecular assays:**

- WGS/WES (primarily whole blood)
- RNA-seq
- Small RNA-seq (future)

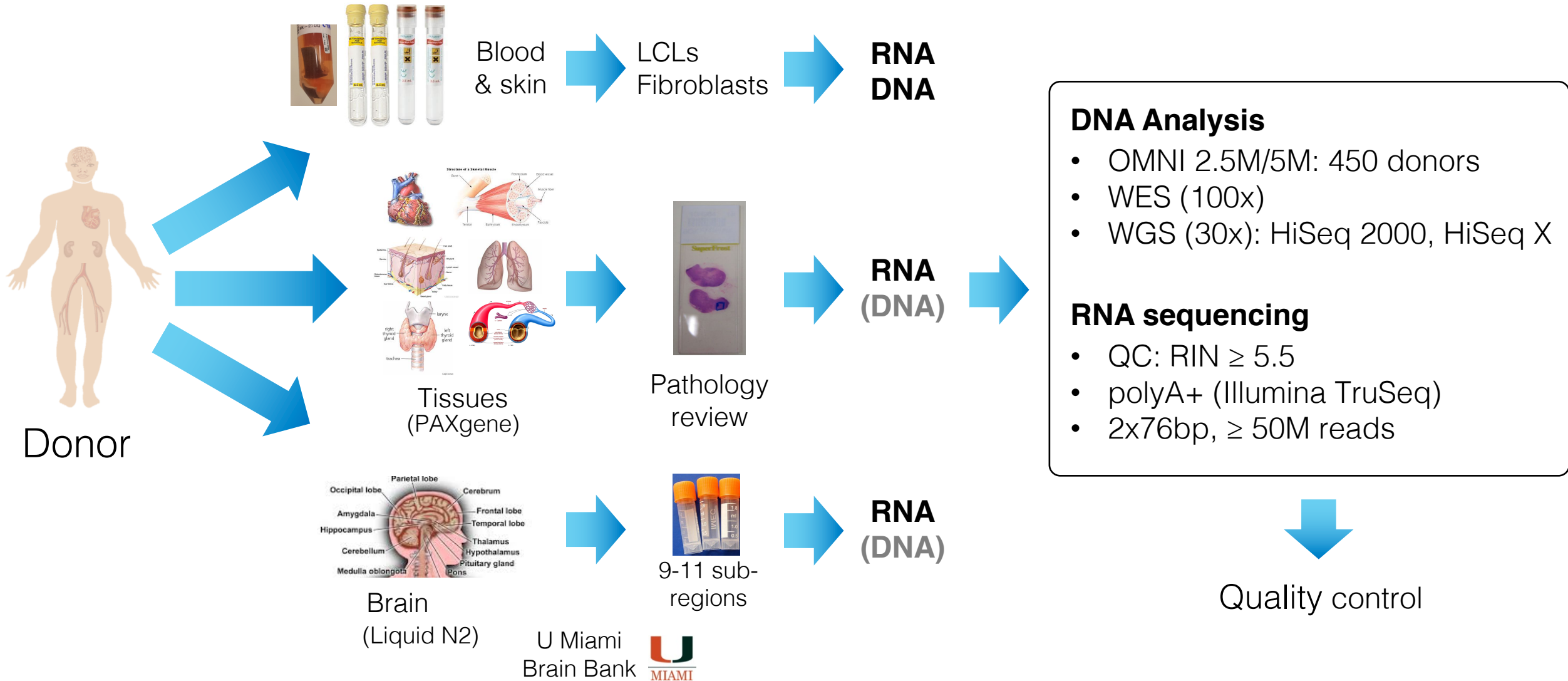
eGTEx: the Enhancing GTEx project



eGTEx data types

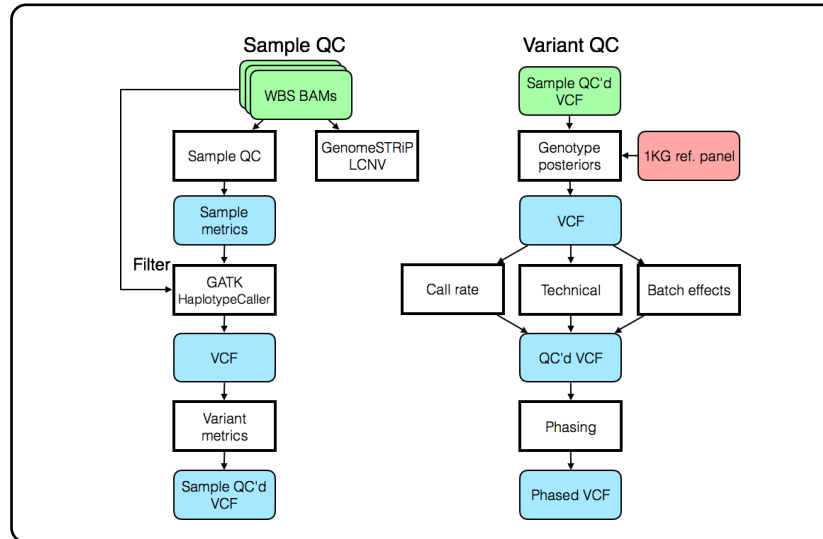
- Protein quantifications (x2)
- Methylation (WGBS)
- Histone modifications (ChIP-seq)
- Dnase-seq
- mmPCR-seq (deep ASE)
- Somatic DNA-seq (deep exome seq)
- Analysis of telomere structure

Sample and data processing overview

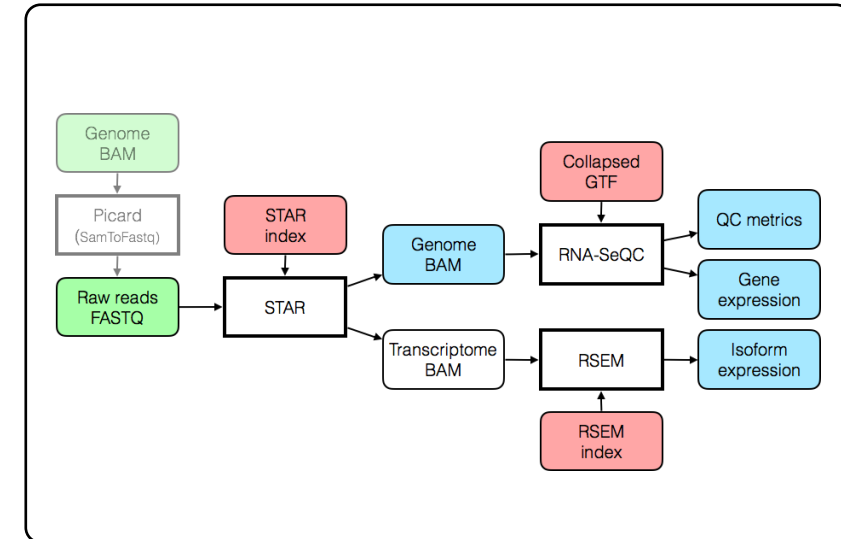


Data processing and quality control pipelines

Genotype QC: samples & variants



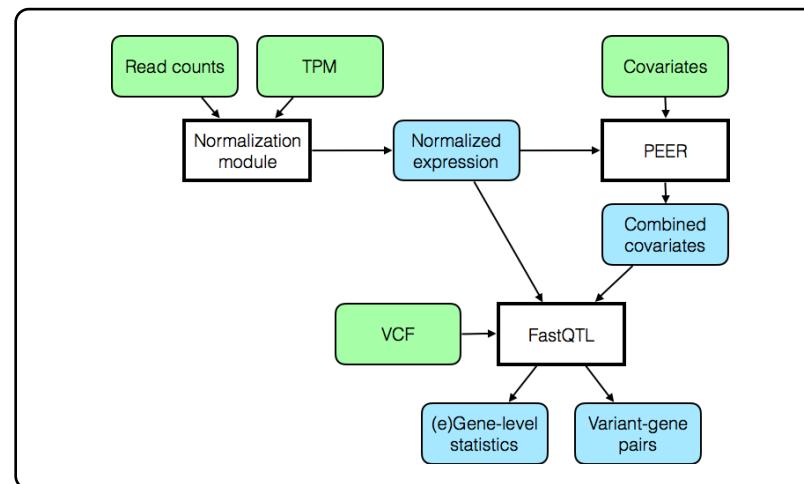
RNA-seq alignment, quantification & QC



VCF



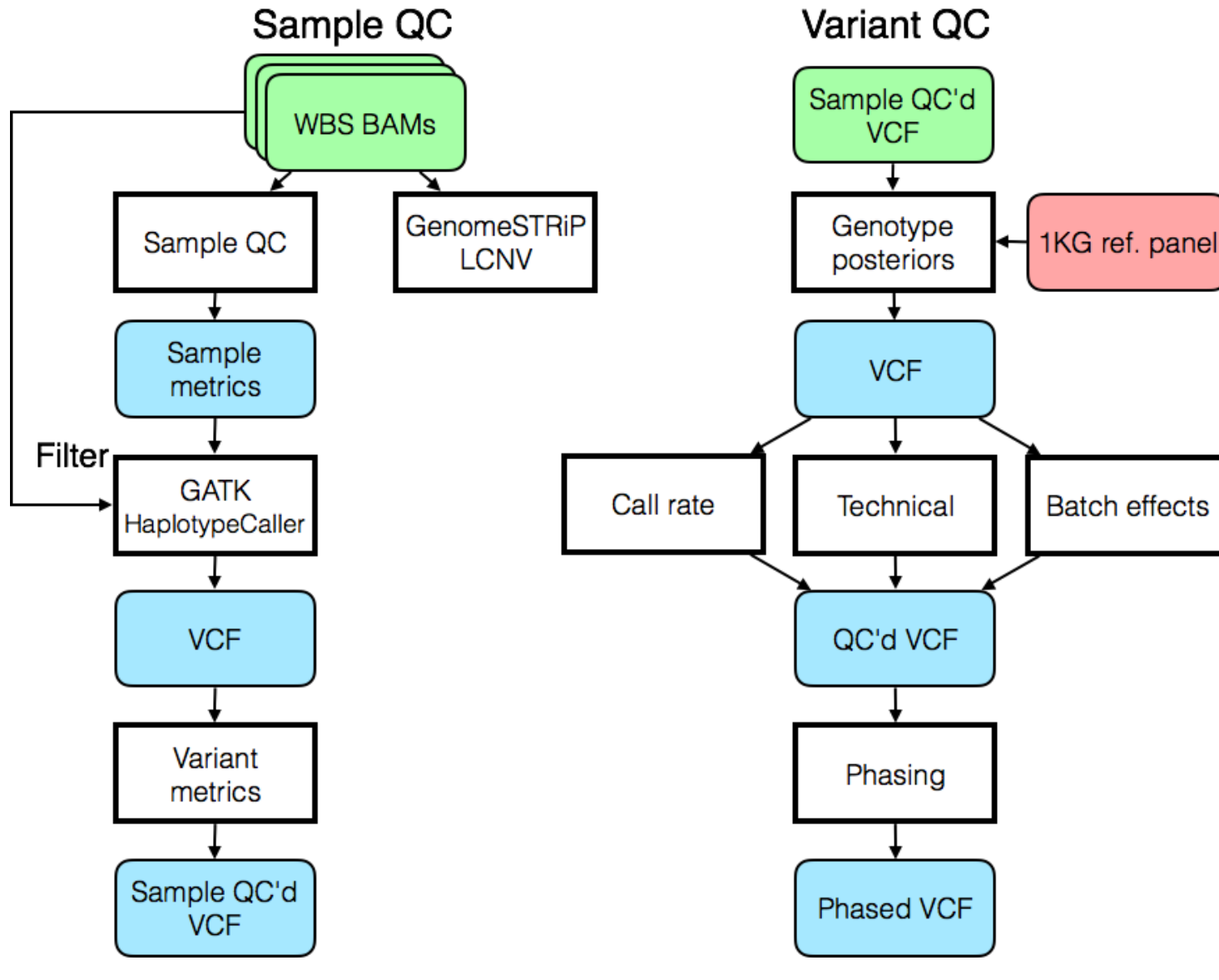
eQTL mapping



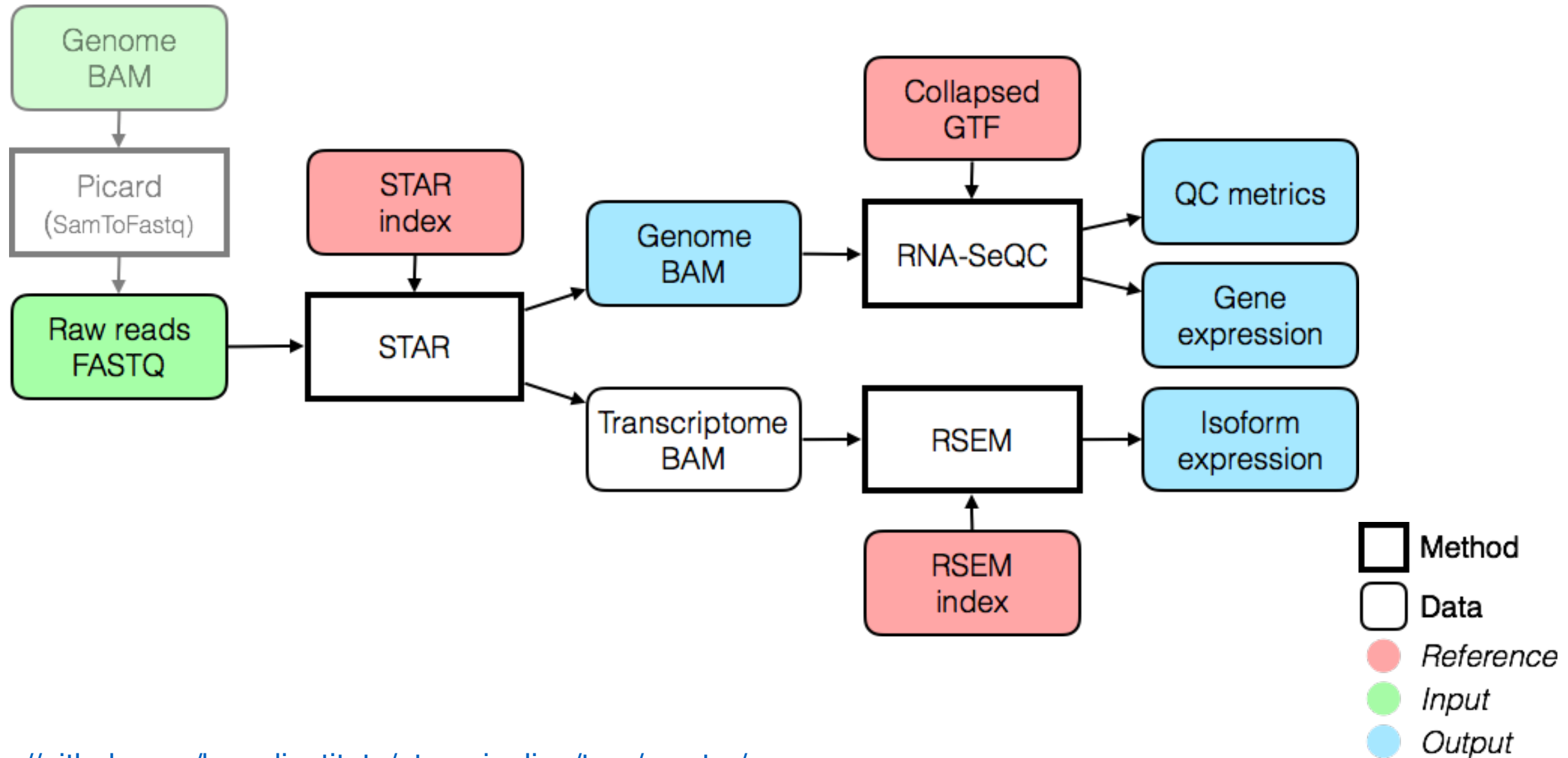
Expression tables,
Covariates



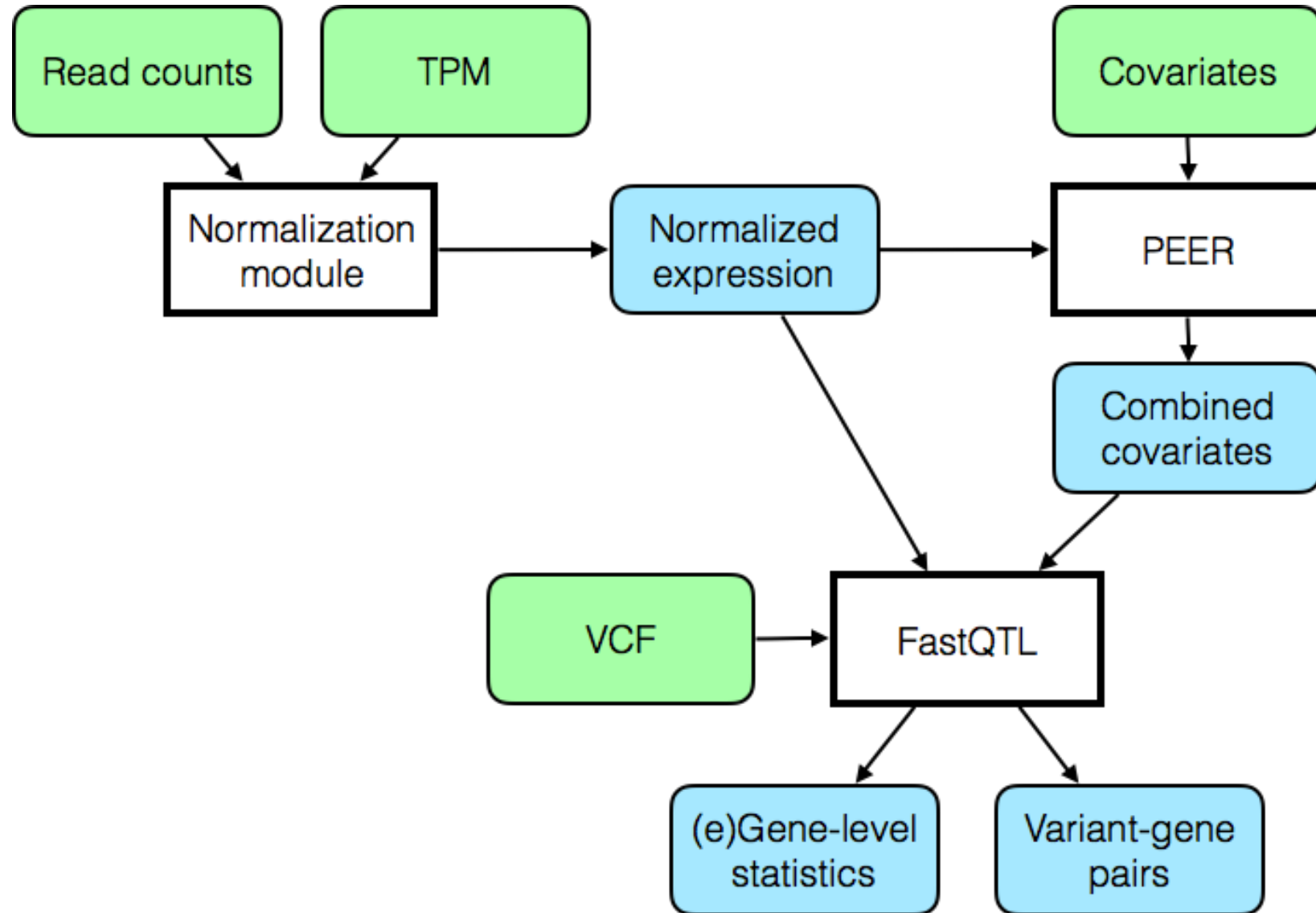
Genotype QC pipeline



RNA-seq pipeline: alignment, quantification, QC



eQTL mapping pipeline



RNA-seq and eQTL pipeline details

Current public release

Release	V6p	V7	V8	V9
Genome build	GRCh37	GRCh37	GRCh38	GRCh38
GENCODE annotation	v19	v19	v26	v26
Aligner	TopHat 1.4.1	STAR 2.4.2a	STAR 2.5.3a	STAR 2.5.3a
Gene expression	RNA-SeQC 1.1.8	RNA-SeQC 1.1.9	RNA-SeQC 1.1.9	RNA-SeQC 1.1.9
Transcript expression	FluxCapacitor 1.6	RSEM 1.2.22	RSEM 1.3.0	RSEM 1.3.0
Quality control metrics	RNA-SeQC 1.1.8	RNA-SeQC 1.1.9	RNA-SeQC 1.1.9	RNA-SeQC 1.1.9
QTL mapper	FastQTL			

- Pipeline components selected and updated based on internal and published benchmarks (e.g., Teng et al., Genome Biology, 2016).

Overview of GTEx resources: open-access data

- Expression
 - Gene-level expression (TPM, counts)
 - Transcript-level expression (TPM, counts, isoform proportions)
 - Exon read counts
- QTLs
 - Single-tissue eQTLs (*cis*- and *trans*-)
 - Multi-tissue eQTLs
 - Future: splicing QTLs
- Histology images
- De-identified public access sample and subject metadata

All open-access data is available at gtexportal.org

Overview of GTEx resources: protected data

- Sequence data:
 - RNA-seq (2x76 bp, unstranded, >50M reads/sample)
 - WGS (30x coverage) and WES (100x coverage)
 - Illumina Omni2.5/5 microarray genotypes (subset of 450 donors)
- Allele-specific expression (ASE)
- Full sample and subject metadata
- Future: eGTEx sequence data
 - ChIP-seq
 - WGBS-seq

All protected-access data is available at dbGaP, under accession phs000424

GTEx data releases

Release	V6/V6p	V7	V8	V9
RNA-seq	8,555	11,688	17,382	~20,000
WGS	148	635	838	~960
WES	520	603		~960
OMNI	450	450	450	450
RNA-seq w/ GT	7333	10361	15253	~20,000
eQTL tissues	44	48	49	49

Current public release

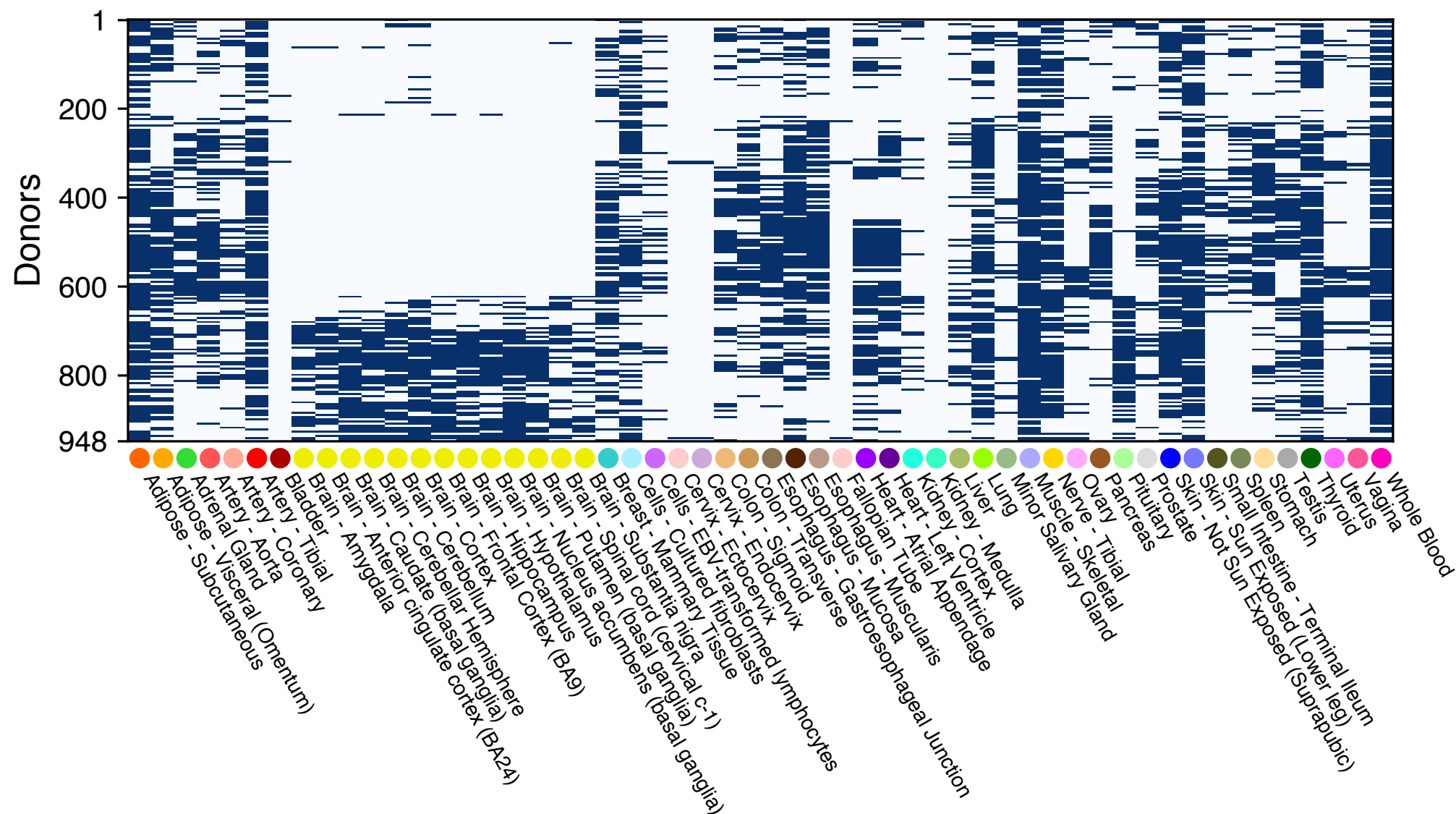
Analysis freezes

Midpoint publications: V6p

- Full list available at <https://gtexportal.org/home/publicationPage>
- Data remains available on GTEx Portal

No publication embargo on V7

GTEx data production: samples per donor



Expression data on GTEx Portal

Gene-level expression

- Based on collapsed GENCODE annotation
- Quantified with RNA-SeQC
- TPM
- Read counts
- No covariate correction

Transcript-level expression

- Based on full GENCODE annotation
- Quantified with RSEM
- TPM
- Expected read counts
- No covariate correction

eQTL inputs

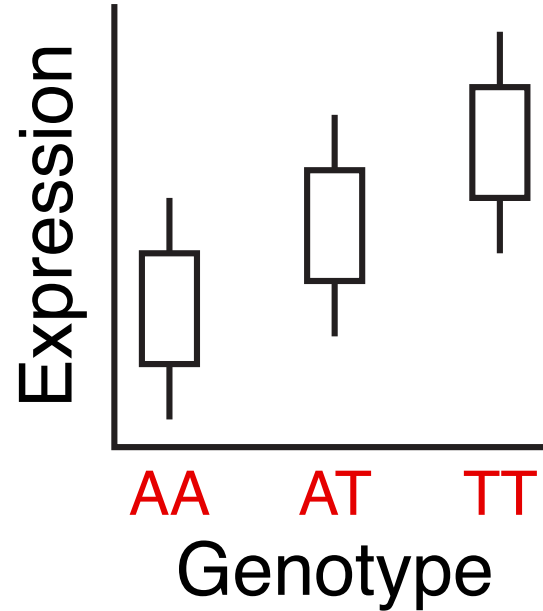
- Based on gene-level quantifications
- Additional normalization: TMM of read counts; inverse normal transform
- Covariates (hidden + known) in separate file

Annotation used for gene-level expression quantification

- RNA-seq protocol:
 - polyA+
 - Unstranded
- Ambiguity in quantifying exon domains shared between sense and anti-sense transcripts
- Collapsing procedure:
 - Masks overlapping intervals
 - Mask 'readthrough' and 'retained intron' transcripts



Definition of *cis*-eQTLs in GTEx

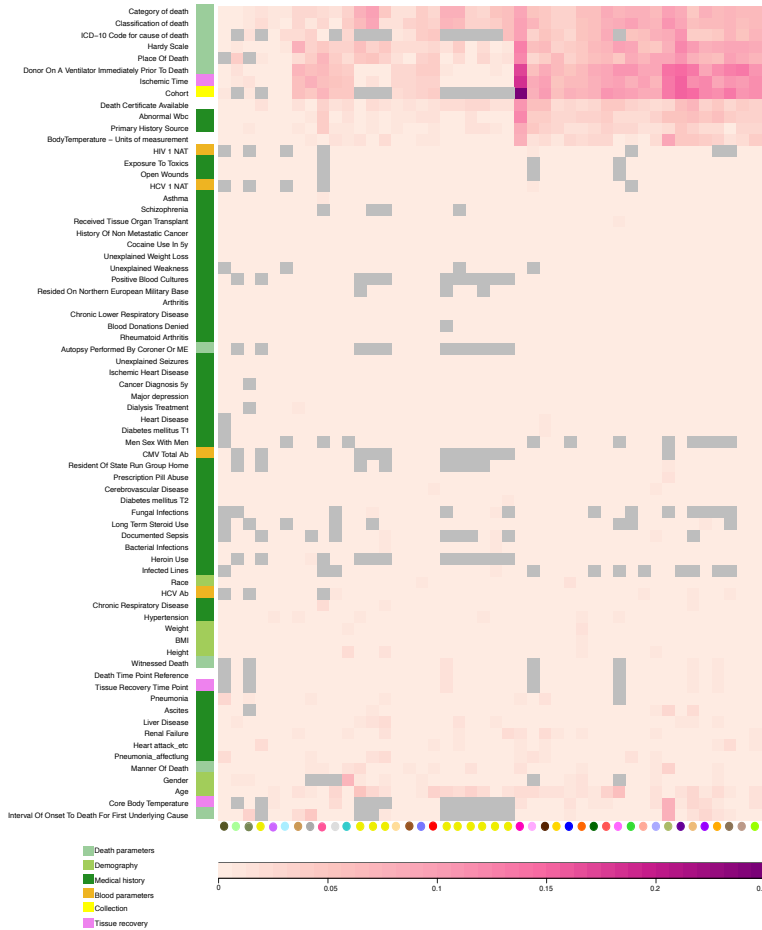
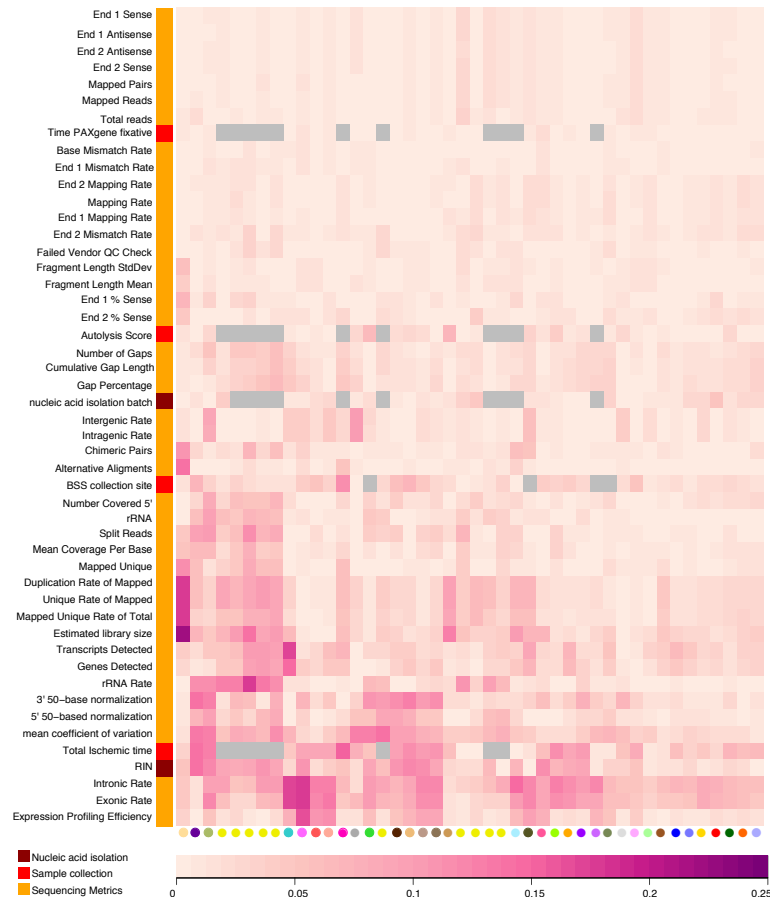


- ***cis*-eQTL**: genome-wide significant association between ≥ 1 eVariant and eGene, with associations tested within $\pm 1\text{Mb}$ *cis*-window around TSS. Does not imply evidence of allelic effects at each locus.
- **eGene**: gene with at least one significant eQTL (at 5% FDR).
- **eVariant**: variant with a significant association to ≥ 1 eGene.
- **Effect allele**: ALT allele (not necessarily the minor allele).

Data normalization for eQTL analyses

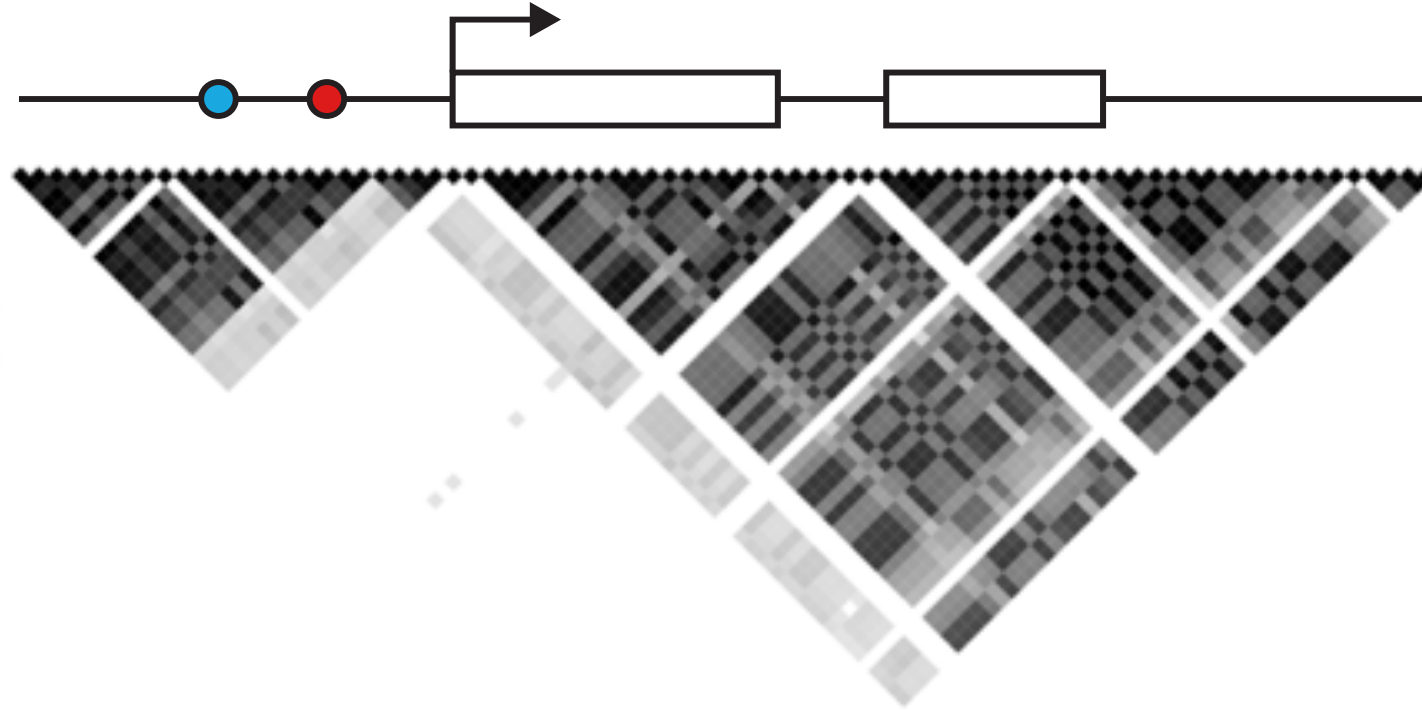
- Expression thresholds:
 - ≥ 6 counts in $\geq 20\%$ of samples AND
 - ≥ 0.1 TPM in $\geq 20\%$ of samples
- Normalization:
 - Between sample normalization: TMM (from edgeR)
 - Corrects for library size differences and expression outlier effects
 - Within-gene normalization: inverse normal transform
 - Attenuates outliers

Covariate correction in eQTL analyses



- Genotype: top 3 PCs, sex, sequencing platform (HiSeq 2000, HiSeq X)
- Expression: significant technical confounders may be unknown; estimation of hidden confounders is key (e.g., through PEER factors)

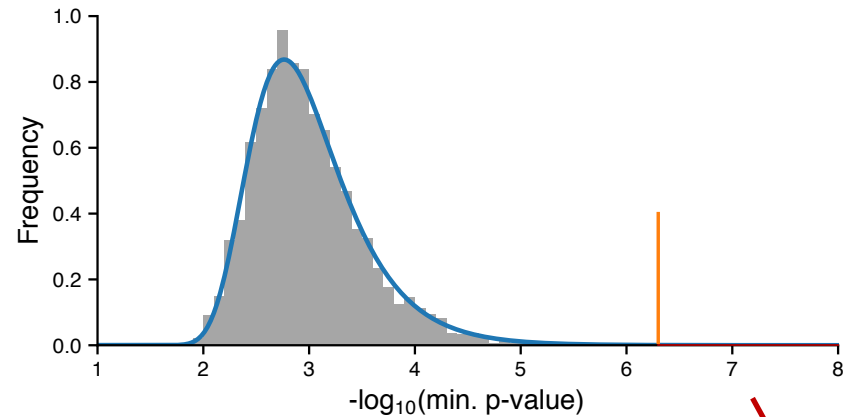
eQTL mapping and eGene discovery



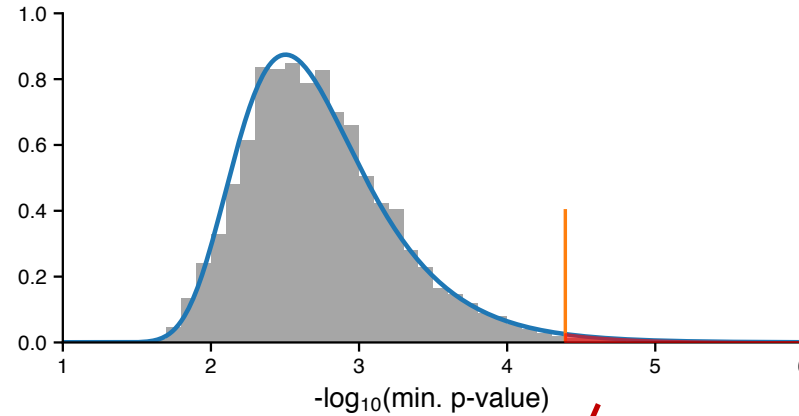
- Variants in *cis*-window ($\pm 1\text{Mb}$ from TSS) may be correlated due to linkage disequilibrium (LD)
- LD must be incorporated in multiple hypothesis testing correction when establishing genome-wide significance
 - Empirical p-values from permutation of genotypes

Multiple hypothesis correction for eGene detection

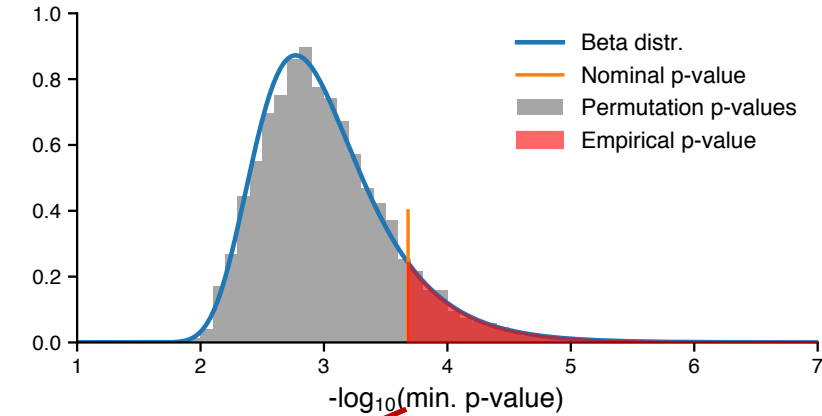
Gene A



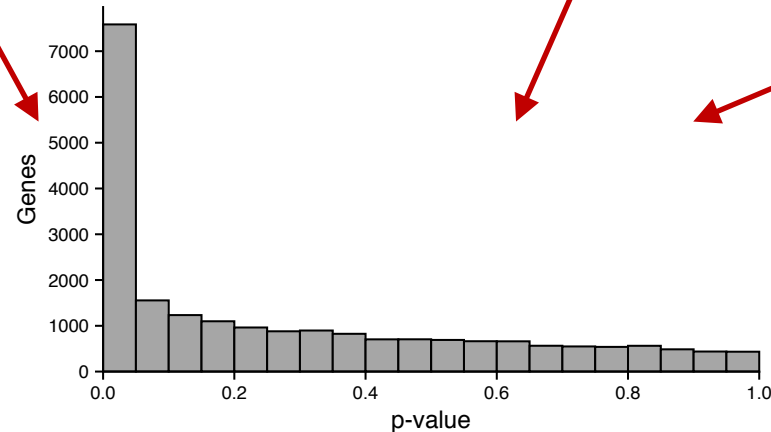
Gene B



Gene C

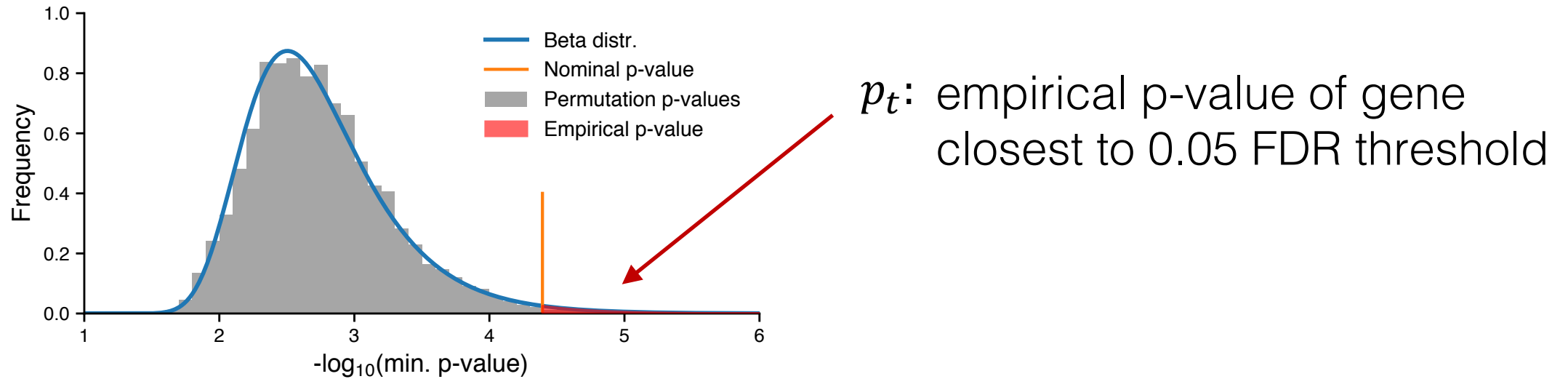


Empirical p-value
distribution



q-values (Storey)
eGenes at ≤ 0.05 FDR

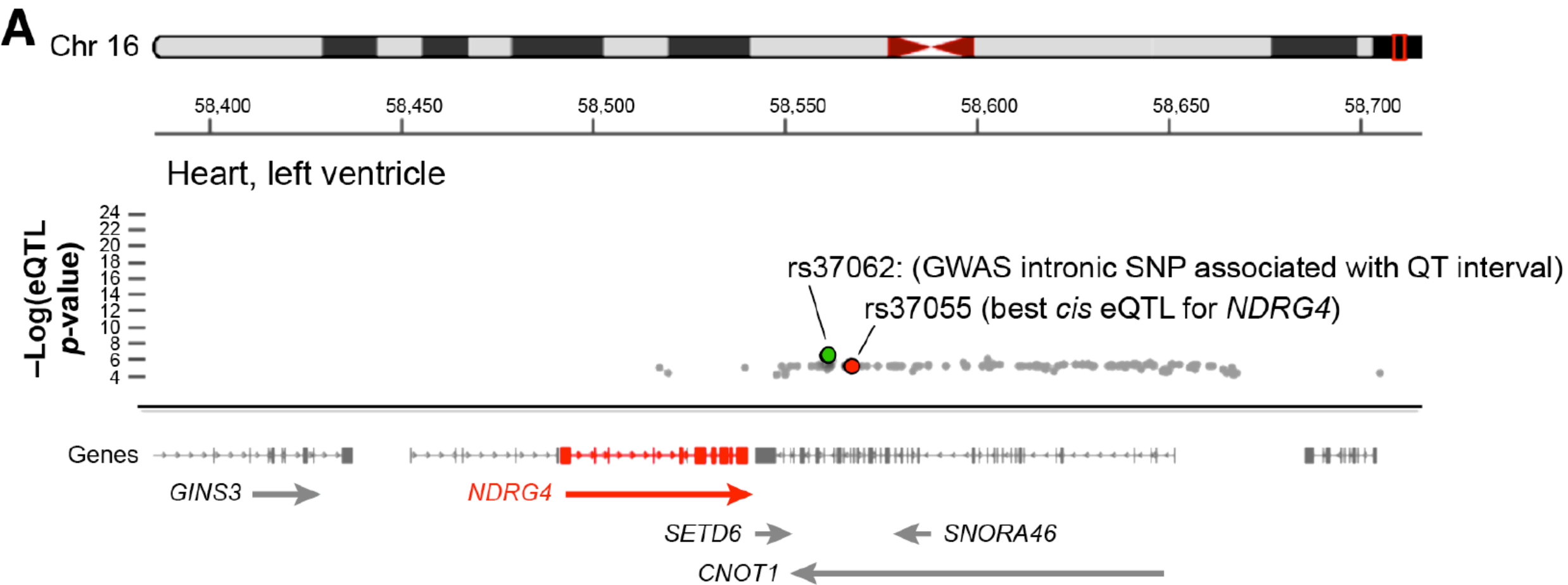
Threshold for significant variant-gene pairs



Nominal p-value threshold for each gene g :

$F_g^{-1}(p_t)$ where F_g^{-1} is the inverse cumulative Beta distribution of the gene.

Example for portal demonstration



NDRG4, *SETD6*, *CNOT1* near QT interval-associated variant, rs37062

jupyter notebook: overview of expression and eQTL data

- The interactive parts of the workshop will be conducted using a jupyter notebook, GTEx_ASHG17_workshop.ipynb
- On the GTEx Portal, go to <https://gtexportal.org/workshop.html>
- Click on “Start the notebook” to begin. This will launch a cloud-based instance of the notebook, with access to all data examples.
Please note that the notebook is read-only.
- The notebook is also available for download at <https://github.com/broadinstitute/gtex-ashg2017-workshop>

Organization of GTEx data: common identifiers

Sample ID:

GTEx-1117F-0226-SM-5GZZ7

Donor ID

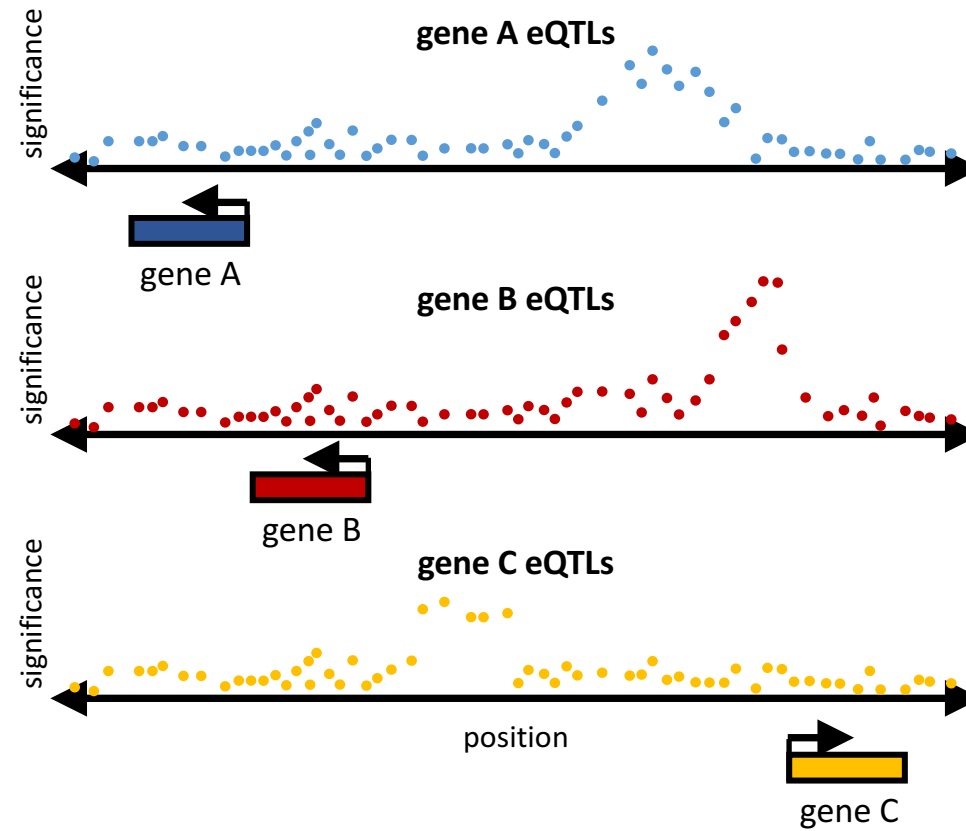
Aliquot ID

Donor-specific
tissue collection ID

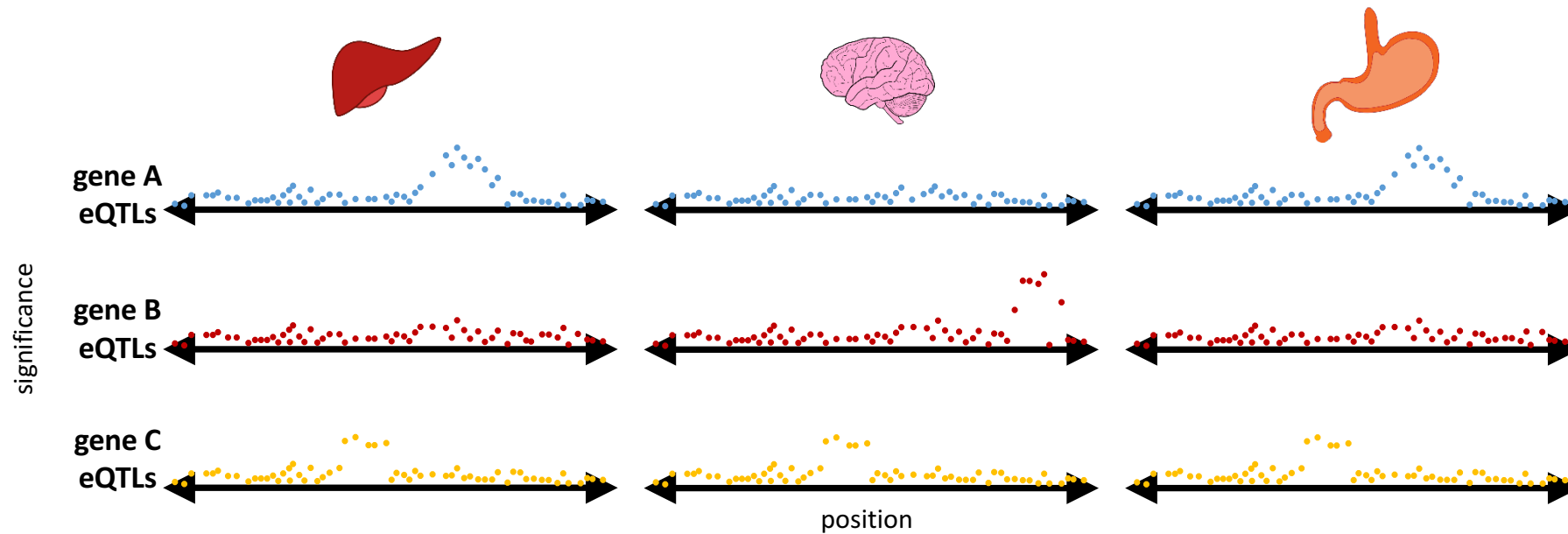
- All sample attributes are indexed by Sample ID
- All donor attributes are indexed by Donor ID
- The donor-specific tissue collection ID is not a proxy for tissue type

Implications of GTEx for interpreting GWAS signals

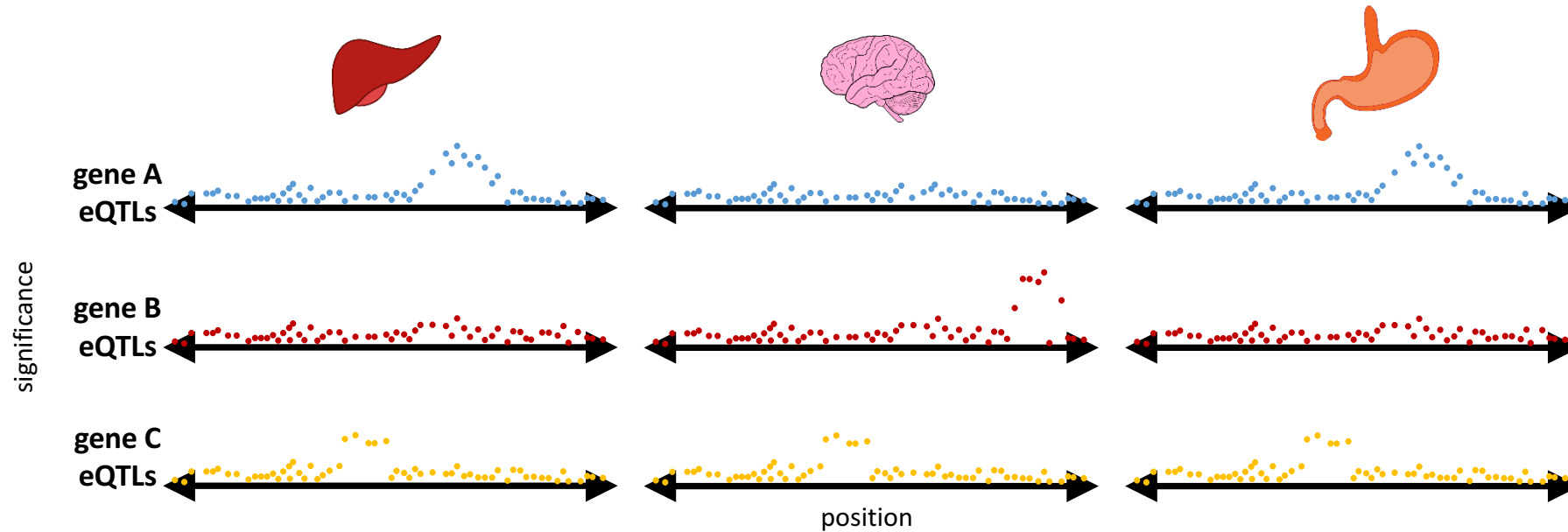
Many genes in the same region have eQTLs



...with different effects across tissues

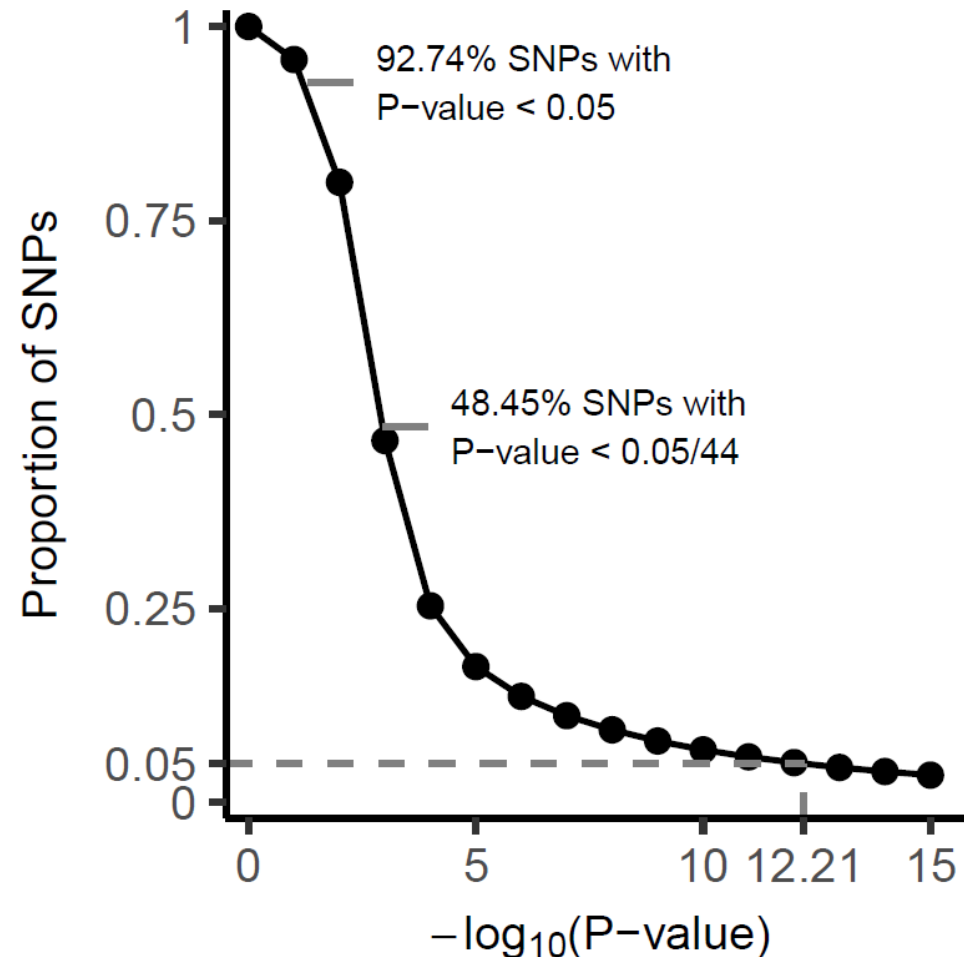


...with different effects across tissues



Which one(s) explain the disease risk?

Lots of eQTL data means that seemingly significant associations are the norm

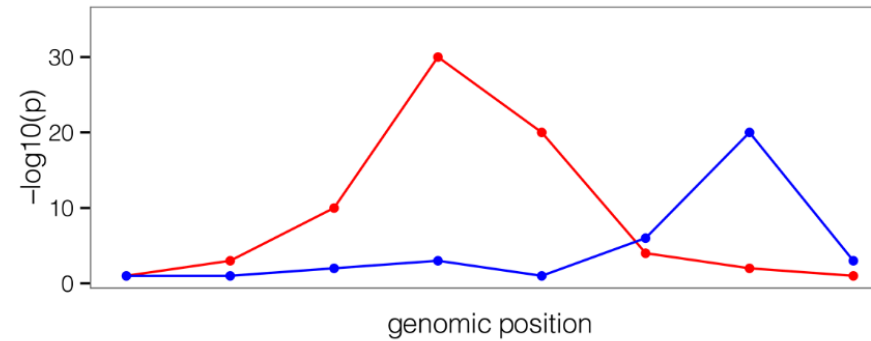
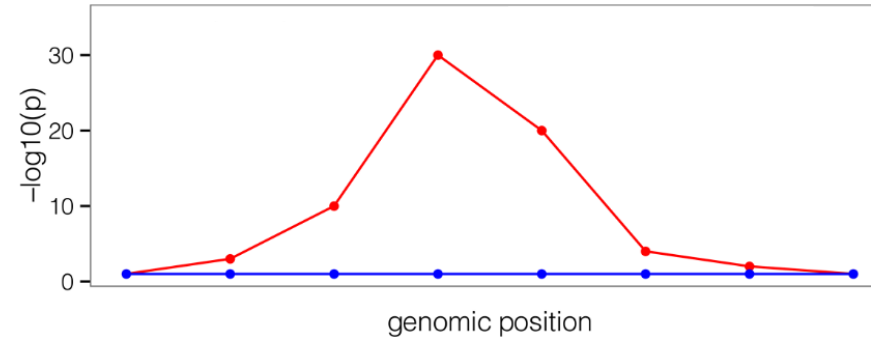


eQTL/GWAS interpretation needs to be examined more cautiously

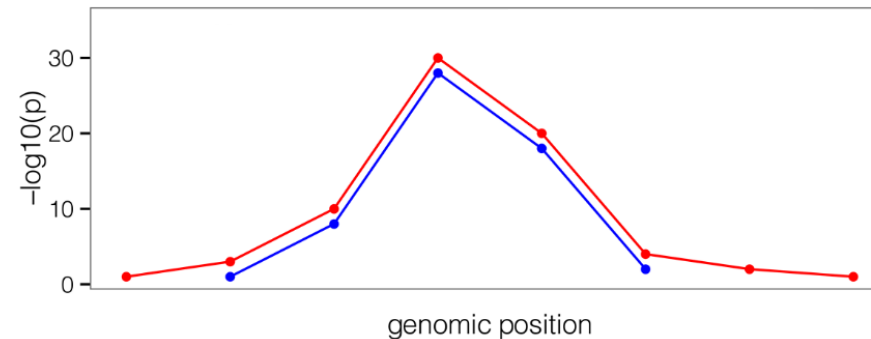
The possibility of a central, tissue-specific effect of the *ESR1* variant rs67229052 is supported by its demonstration as an eQTL for *ESR1* in only one of ~50 Genotype-Tissue Expression (GTEx) tissues (brain_caudate_basal_ganglia; using the proxy SNP, rs4305732, with $r^2 = 0.98$); the allele associated with higher *ESR1* expression ($P = 0.0004$)

~1/3 of all variants could meet this criterion

Co-localization approaches combine eQTL and GWAS signals

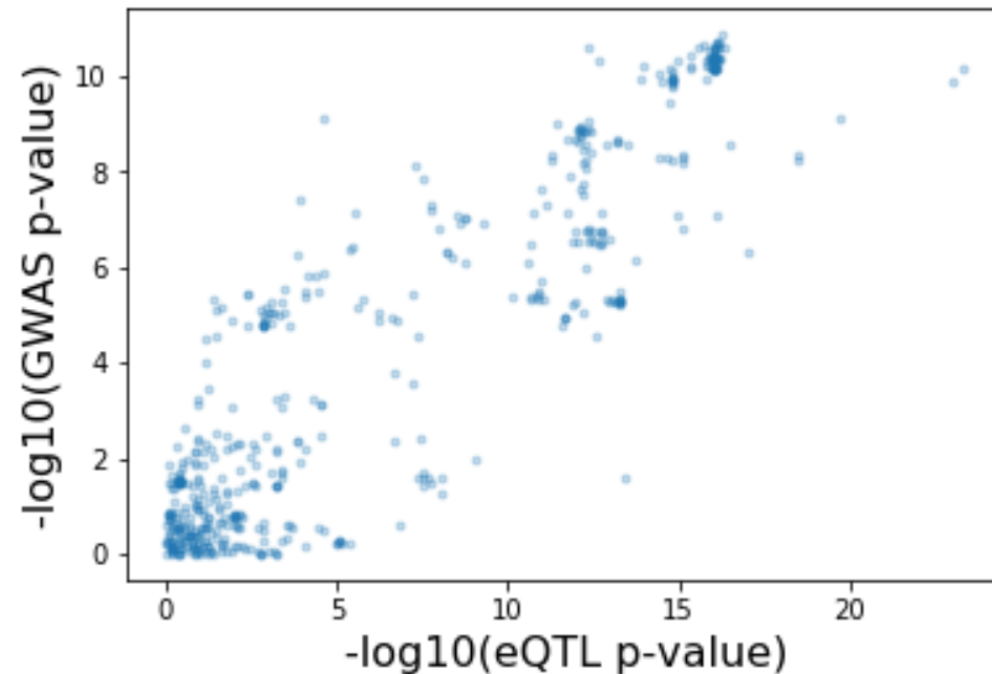


GWAS signal
eQTL signal

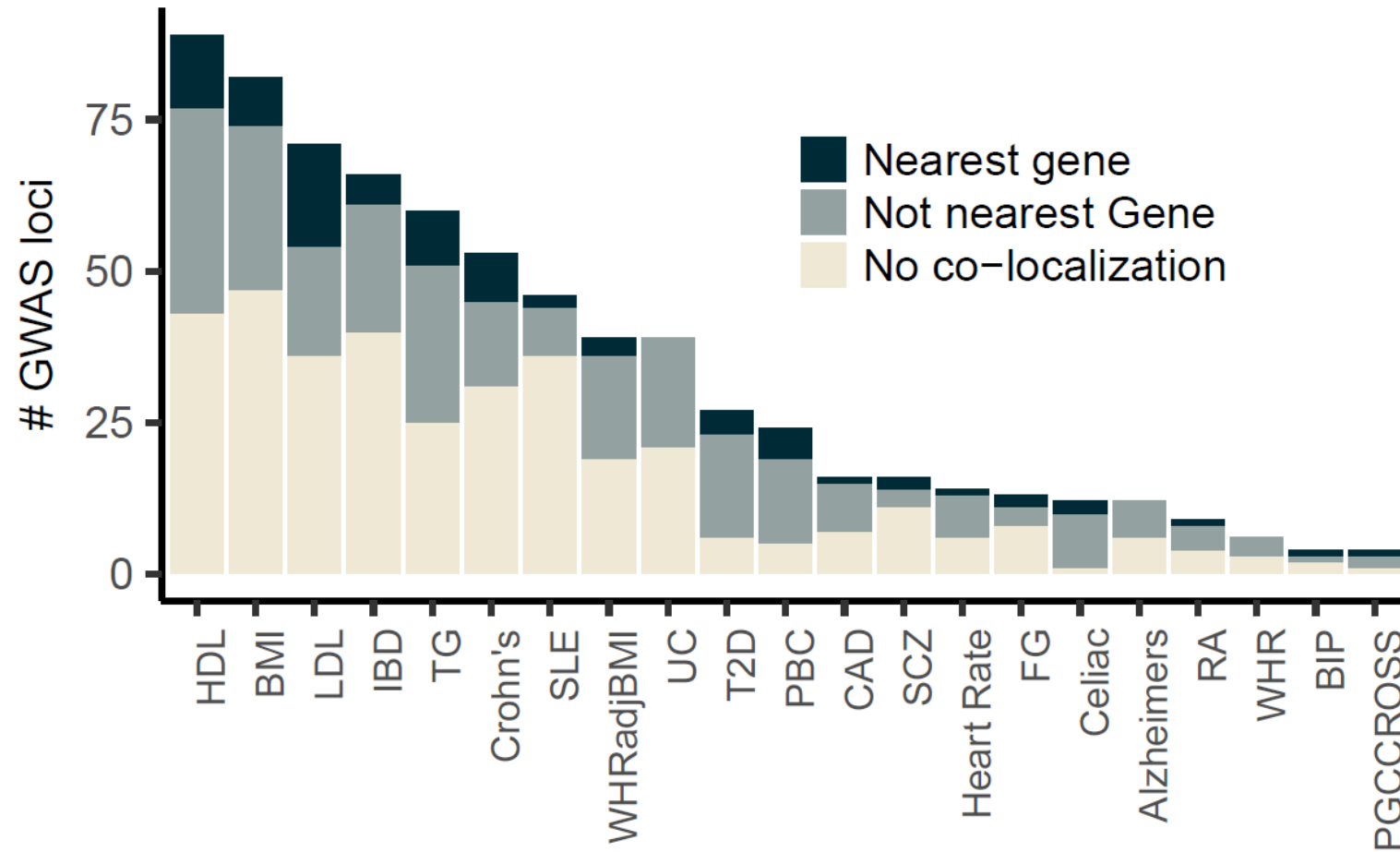


iPython notebook task

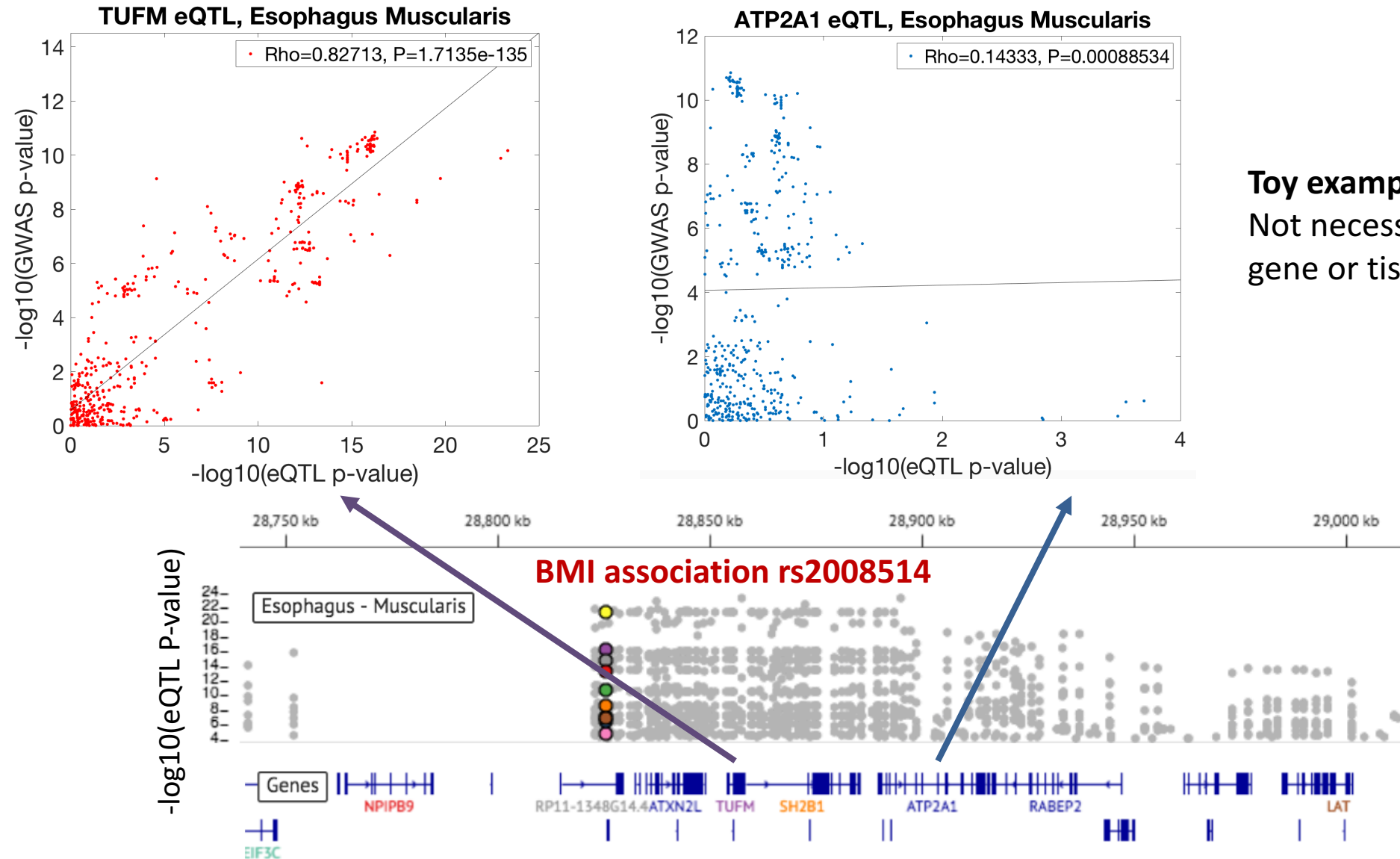
- Correlation of GWAS and eQTL summary statistics over an associated hit for BMI.



Co-localization of eQTLs and GWAS in GTEx

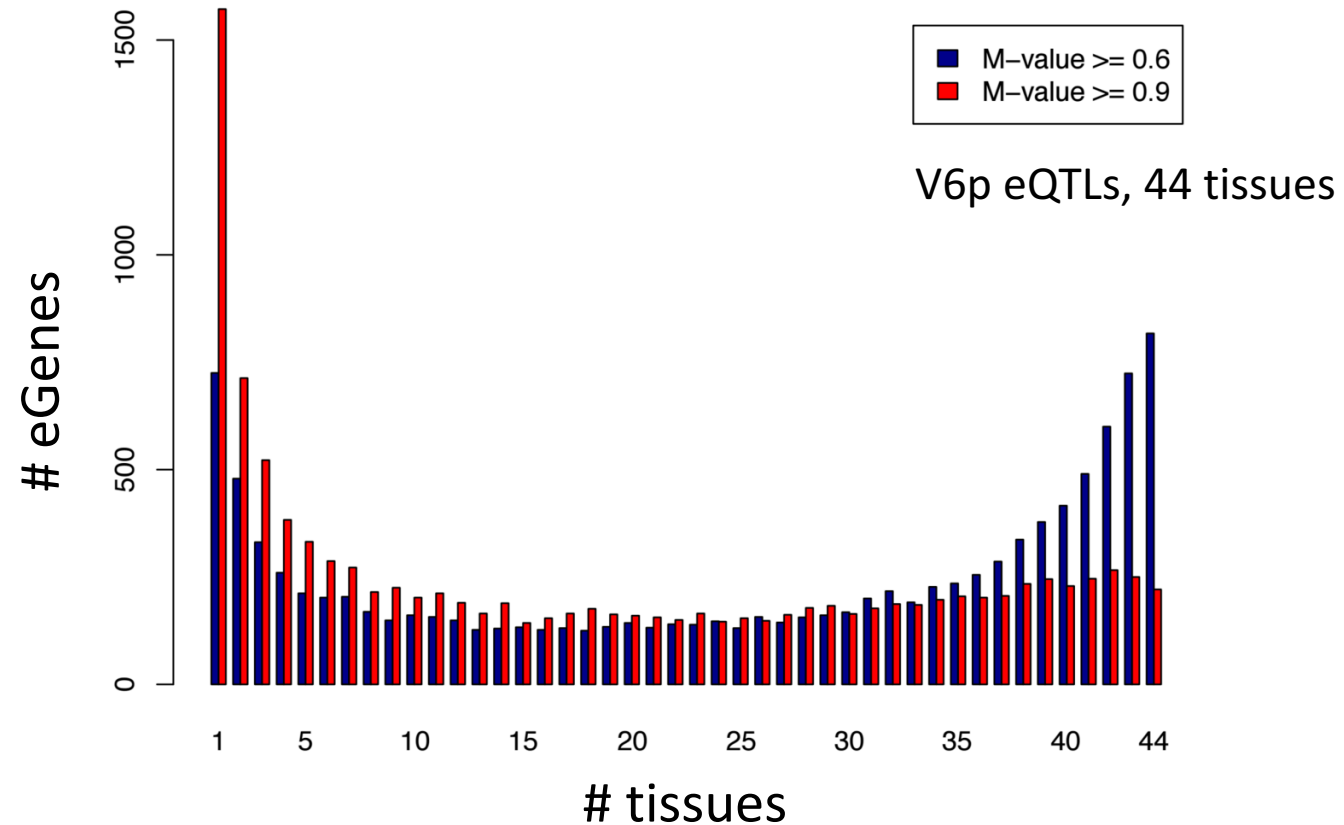


- Correlation of GWAS and eQTL summary statistics for two separate genes over an associated hit for BMI



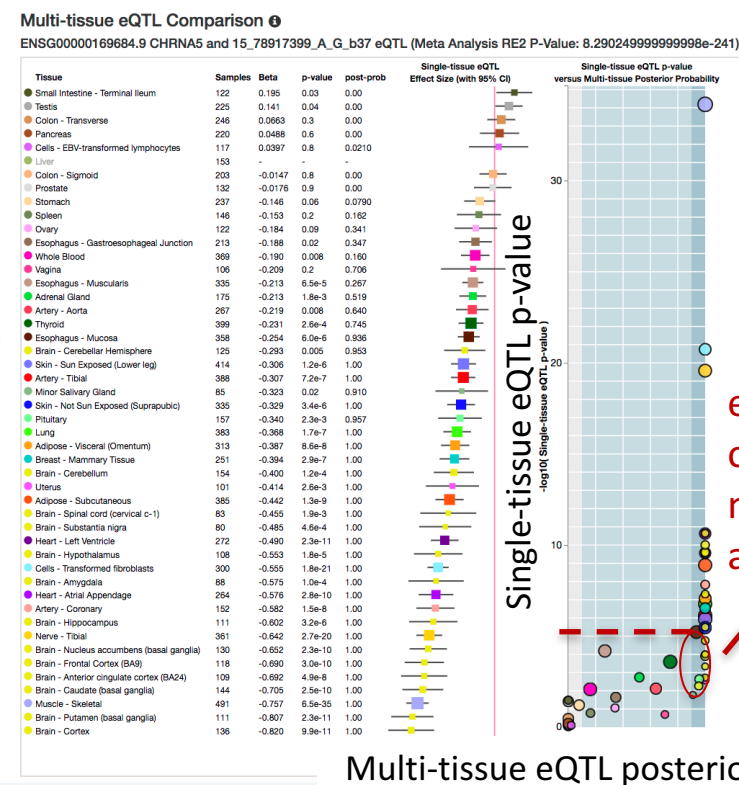
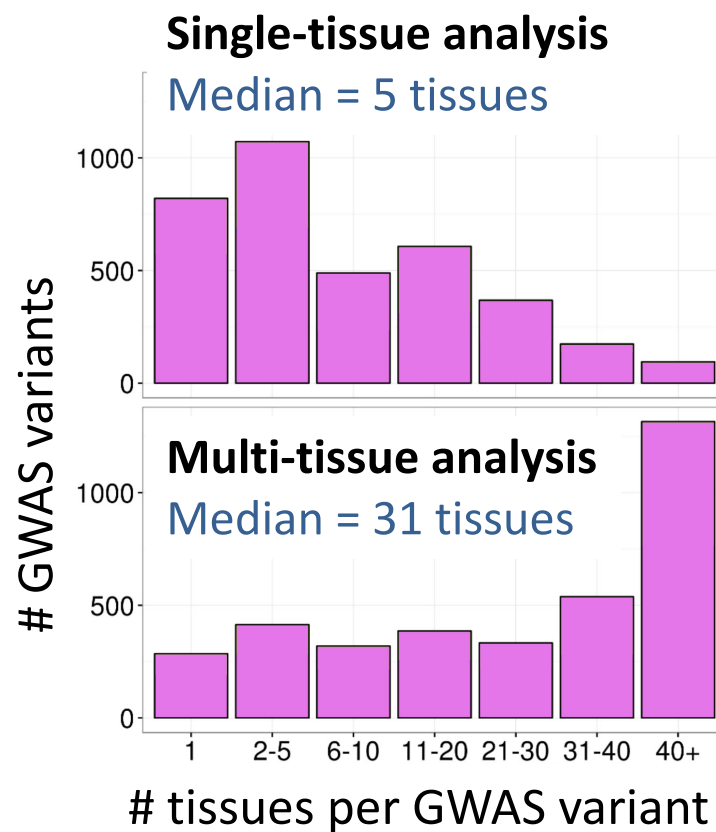
Toy example
Not necessarily the causal gene or tissue

Bimodal distribution of tissue-specificity of *cis*-eQTLs

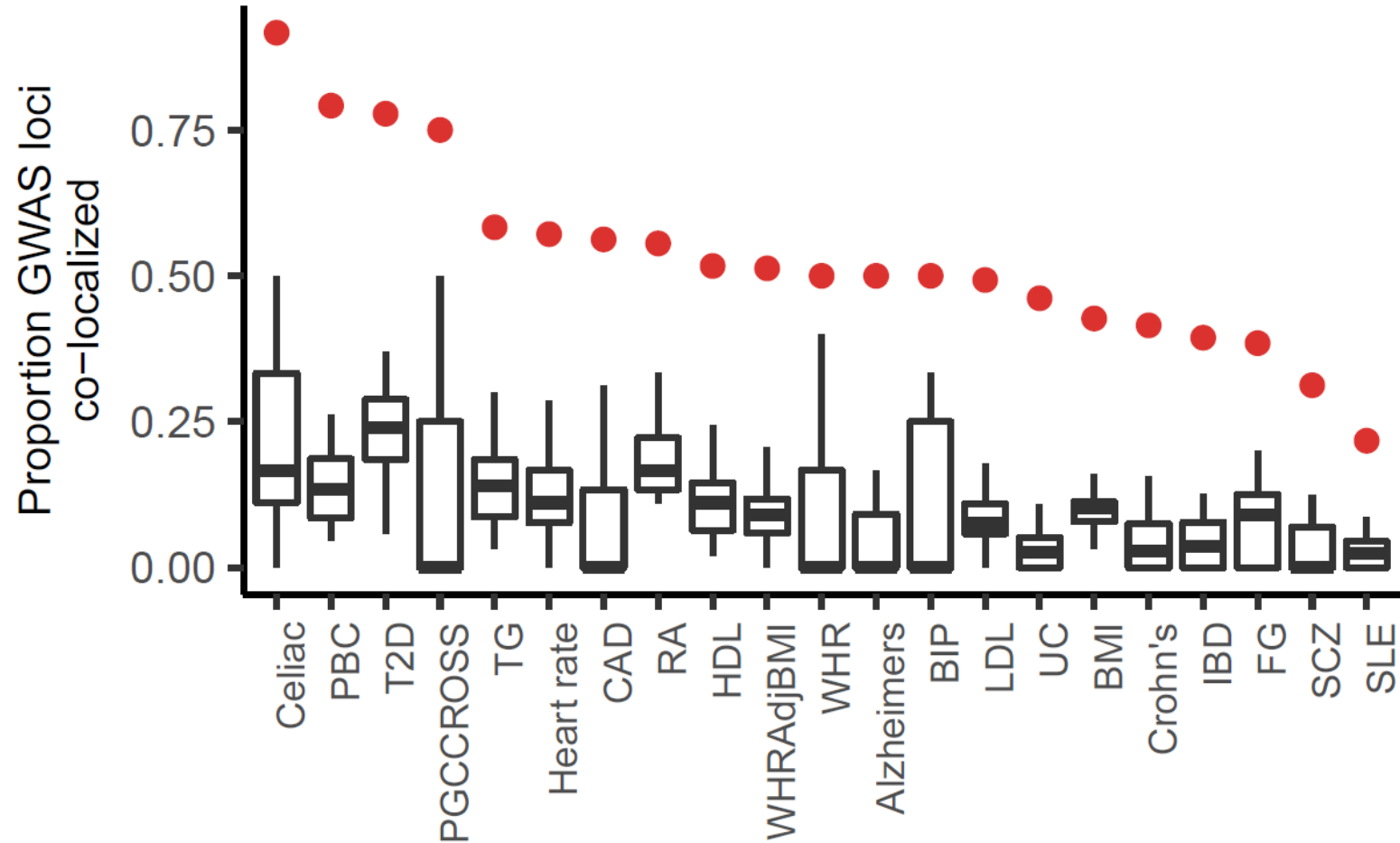


Multi-tissue eQTL meta-analysis: [Metasoft](#) (Han, B and Eskin, E, AJHG 2011)

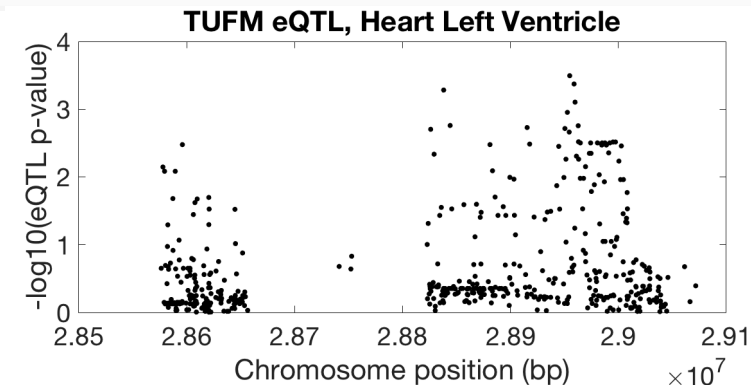
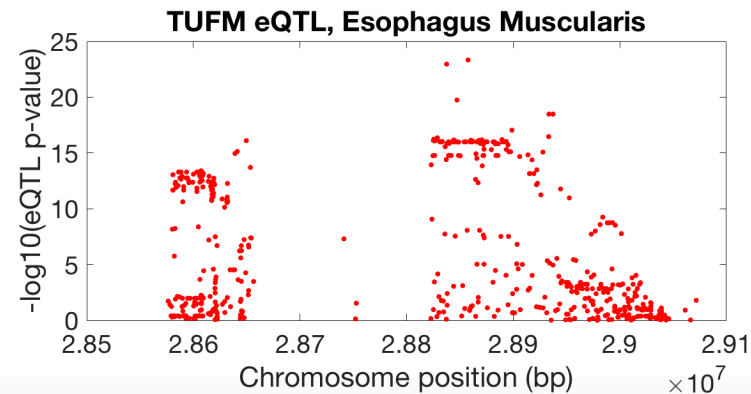
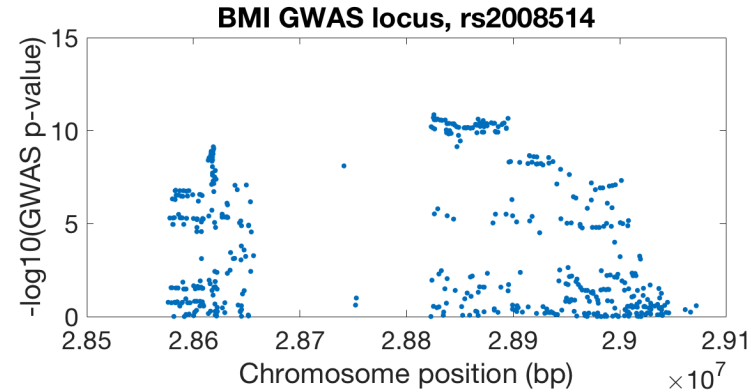
Number of tissues per eQTL in LD with GWAS variants increases with increased power (multi-tissue analysis)



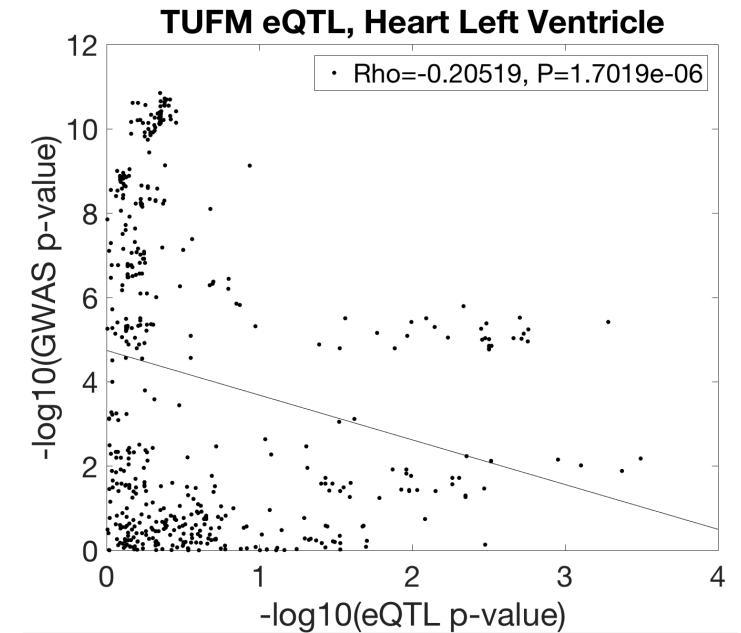
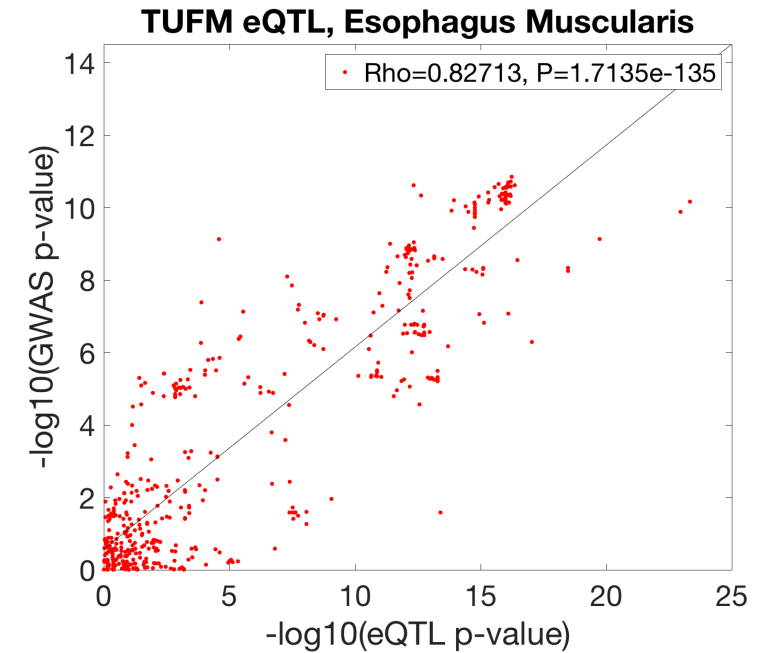
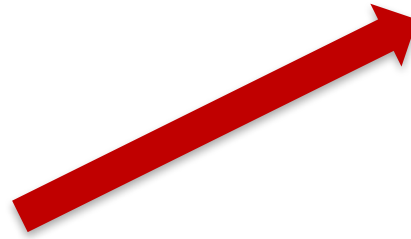
Co-localization of eQTLs and GWAS in GTEx



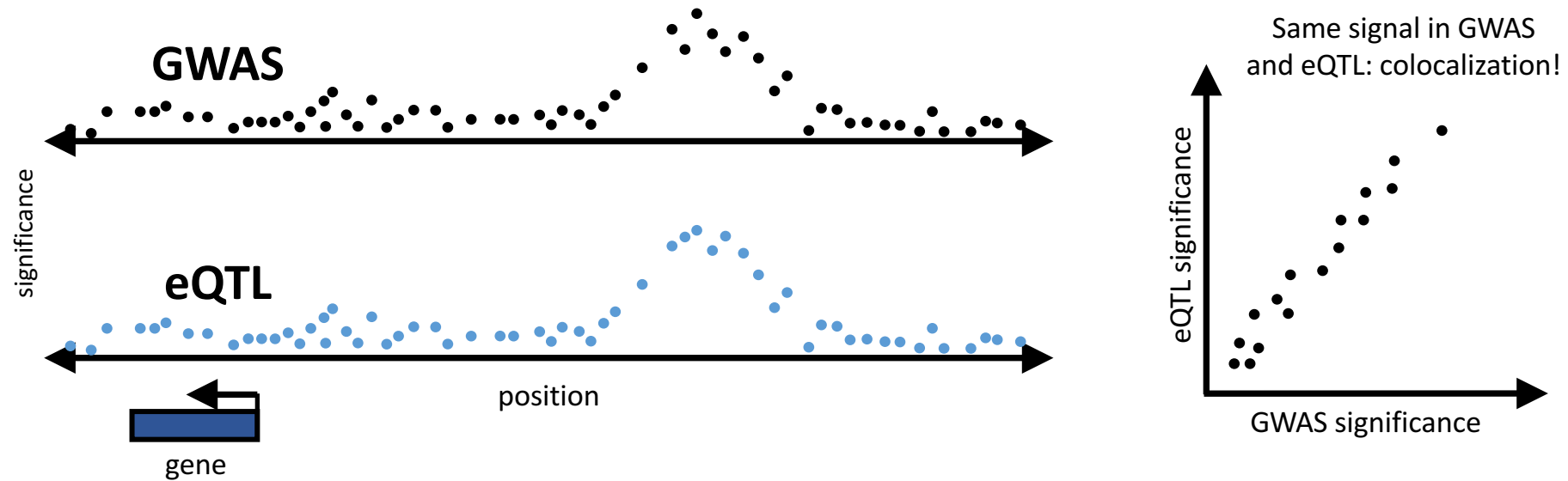
- Correlation of GWAS and eQTL summary statistics for two separate tissues over an associated hit for BMI



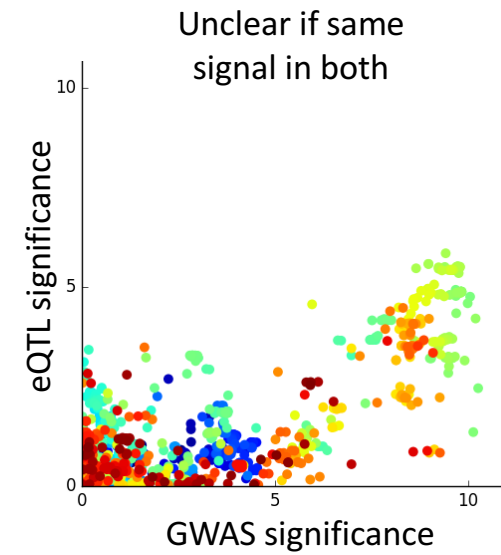
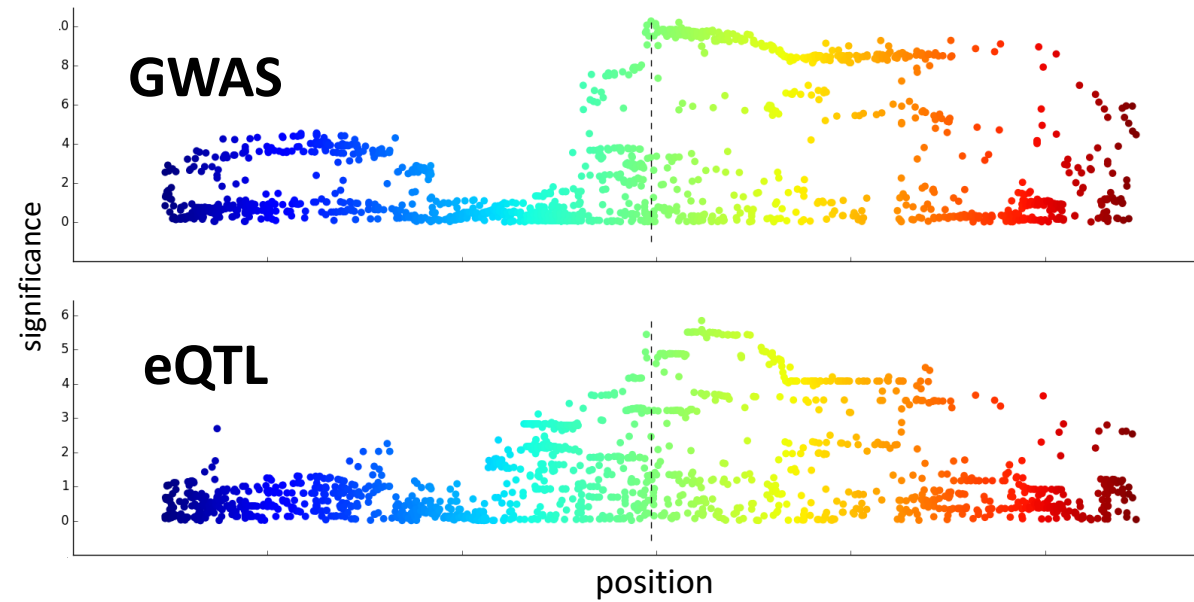
Toy example
Not necessarily the causal
gene or tissue



Detecting GWAS/eQTL overlap is easy in principle



Difficult in practice

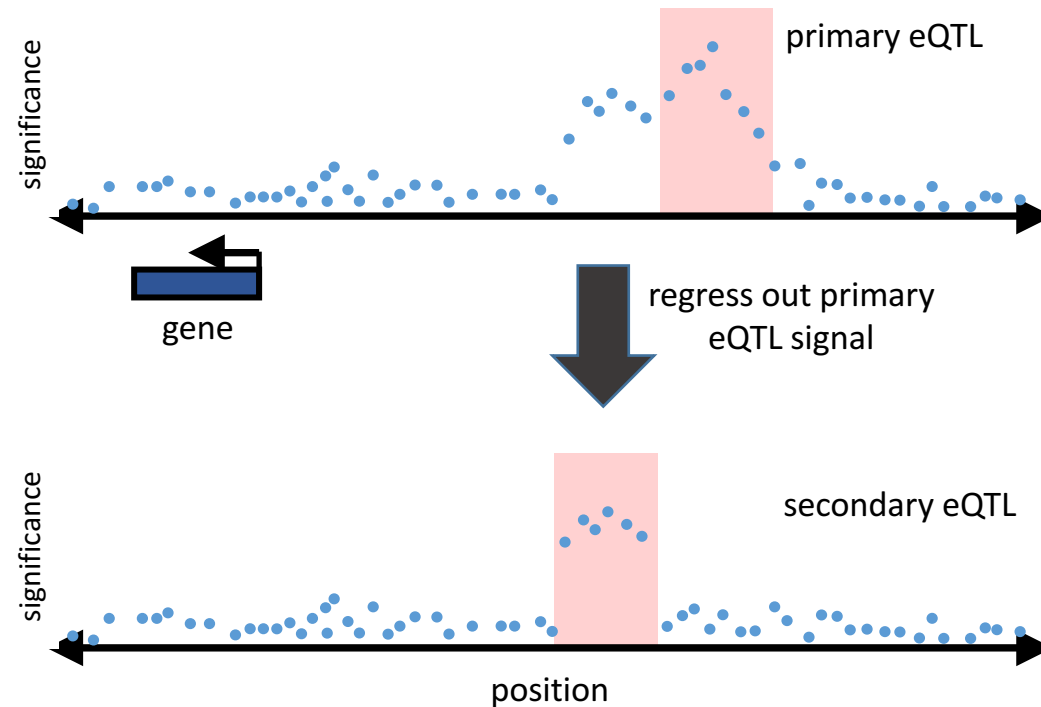


Methods to detect colocalization

Method	method archetype	identifies causal variants?	multiple causal variants?
COLOC	Bayesian	No	No
Sherlock	Bayesian	No	Yes
eCAVIAR	complete likelihood (exhaustive search)	Yes	Yes, but intractable
FINEMAP	complete likelihood (stochastic search)	Yes	Yes
SMR	Mendelian randomization	No	No
TWAS	TWAS	No	Yes
MetaXcan	TWAS	No	Yes

Further caveats: Some genes have multiple independent eQTLs

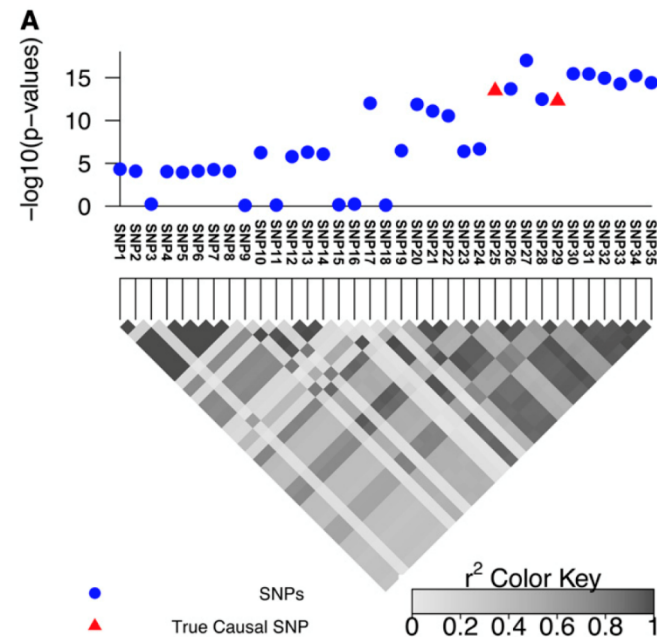
- Could explain complexity of GWAS/eQTL signals
- Conditional eQTLs not yet tested for colocalization with GWAS



Zeng et al. (2016) *BioRxiv*.

Another challenge: Identifying causal variants in eQTL regions

Fine-mapping methods propose credible sets of causal variants for an eQTL



- CAVIAR (Hormozdiari *et al.* Genetics 2014) results will be on GTEx Portal soon!

eQTL limited in capturing rare variant effects

Gene expression outliers can point to rare variants with large effects

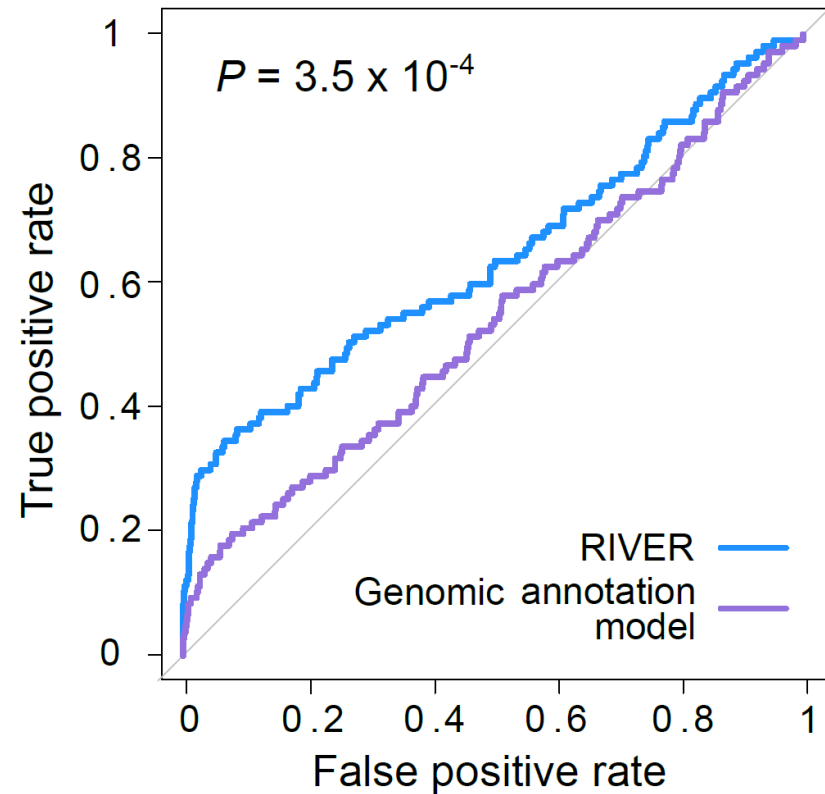
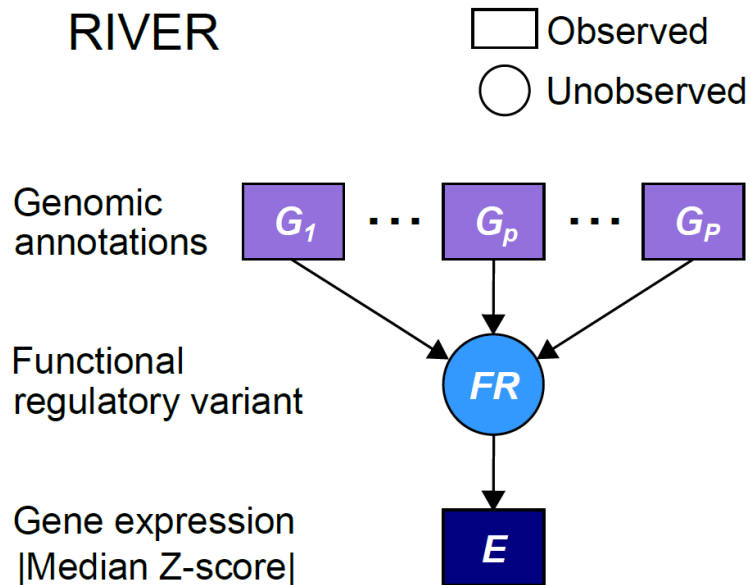


Underexpression
outlier

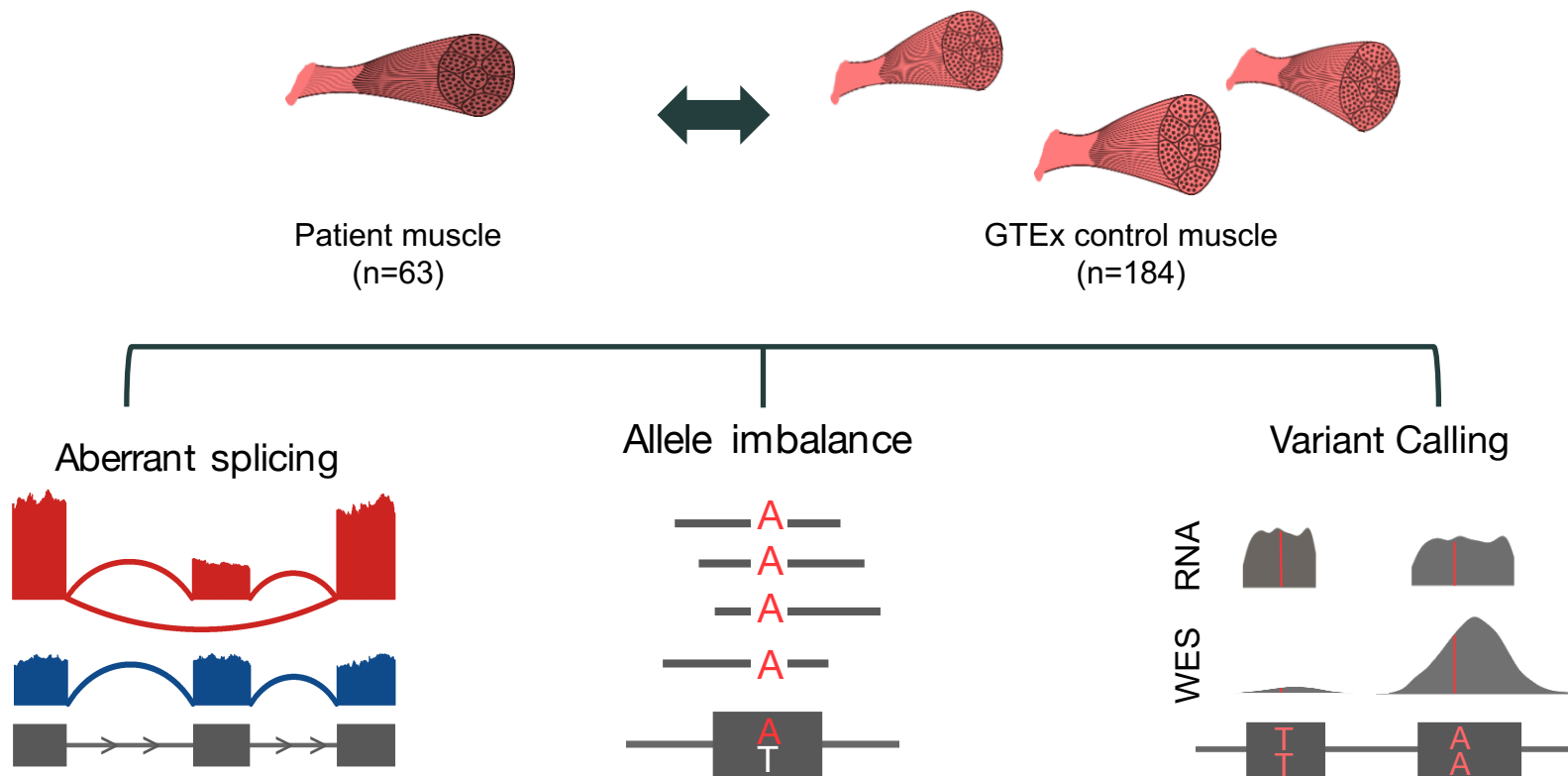


Overexpression
outlier

Interpreting personal variants using genetic and functional genomics data



Using GTEx to help solve rare disease cases.



See Talk by Beryl Cummings Friday 10:45 AM

Interpreting genetic variants in disease

Genetic variation influence gene expression of ~90% of all known protein-coding genes

Abundance of eQTL data requires care when conducting GWAS follow-up

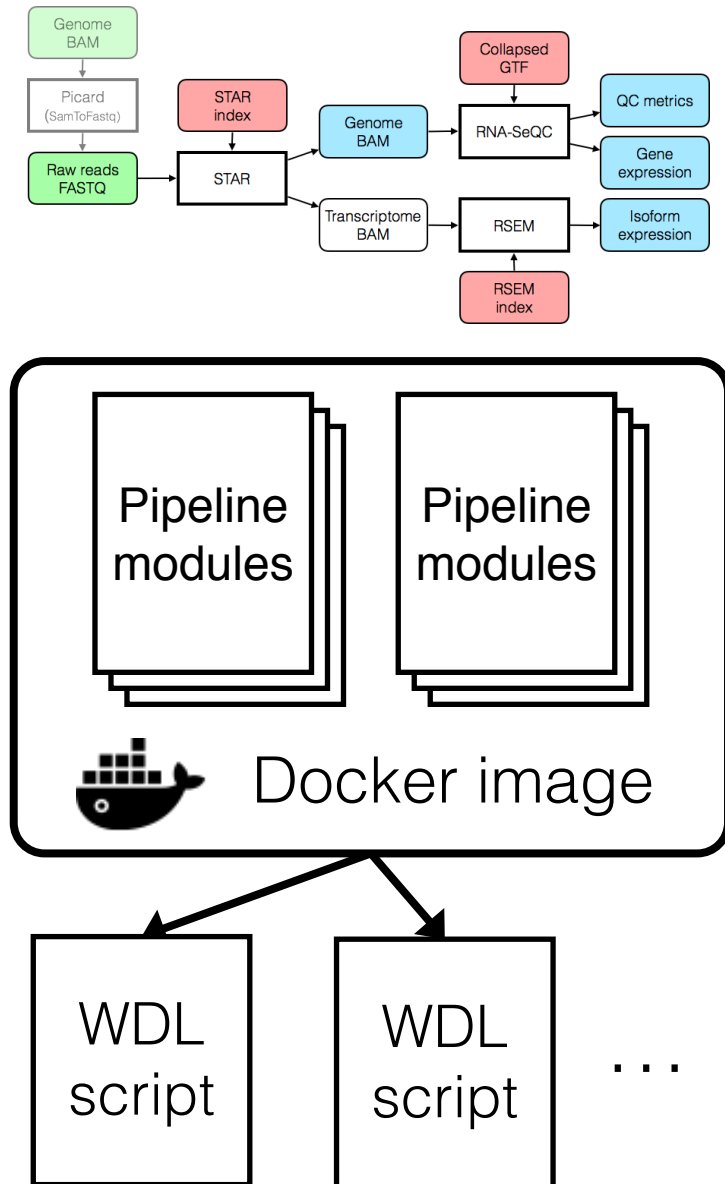
- Multiple testing can lead to false discoveries
 - Co-localization methods required
- 40% of all variants do not co-localize with their nearest gene

Gene expression outliers can identify large-effect rare variants

- Can be used to interpret individual risk factors and identify rare disease genes and variants

GTEx pipelines

- Source code is available at <https://github.com/broadinstitute/gtex-pipeline>
 - Includes wrapper scripts, Dockerfiles
- Pipelines are available on FireCloud (<http://firecloud.org>)
 - Namespace: broadinstitute_gtex



Biobank

- The biobank from the GTEx project is hosted at the Broad Institute.
- Samples can be searched and requested at <https://gtexportal.org/home/samplesPage>.
- Sample requests for research complementing the primary project are welcome.

Acknowledgements

GTEx Consortium

Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group

François Aguet¹, Kristin G. Ardlie¹, Beryl B. Cummings^{1,2}, Ellen T. Gelfand¹, Gad Getz^{1,3}, Kane Hadley¹, Robert E. Handsaker^{1,4}, Katherine H. Huang¹, Seva Kashin^{1,4}, Konrad J. Karczewski^{1,2}, Monkol Lek^{1,2}, Xiao Li¹, Daniel G. MacArthur^{1,2}, Jared L. Nedzel¹, Duyen T. Nguyen¹, Michael S. Noble¹, Ayellet V. Segrè¹, Casandra A. Trowbridge¹, Taru Tukiainen^{1,2}

Statistical Methods groups—Analysis Working Group Nathan S. Abell^{5,6}, Brunilda Balliu⁶, Ruth Barshir⁷, Omer Basha⁷, Alexis Battle⁸, Gireesh K. Bogu^{9,10}, Andrew Brown^{11,12,13}, Christopher D. Brown¹⁴, Stephane E. Castel^{15,16}, Lin S. Chen¹⁷, Colby Chiang¹⁸, Donald F. Conrad^{19,20}, Nancy J. Cox²¹, Farhan N. Damani⁸, Joe R. Davis^{5,6}, Olivier Delaneau^{11,12,13}, Emmanouil T. Dermitzakis^{11,12,13}, Barbara E. Engelhardt²², Eleazar Eskin^{23,24}, Pedro G. Ferreira^{25,26}, Laure Frésard^{5,6}, Eric R. Gamazon^{21,27,28}, Diego Garrido-Martín^{9,10}, Ariel D.H. Gewirtz²⁹, Genna Gliner³⁰, Michael J. Gloudemans^{5,6,31}, Roderic Guigo^{9,10,32}, Ira M. Hall^{18,19,33}, Buhm Han³⁴, Yuan He³⁵, Farhad Hormozdizadeh²³, Cedric Howald^{11,12,13}, Hae Kyung Im³⁶, Brian Jo²⁹, Eun Yong Kang²³, Yungil Kim⁸, Sarah Kim-Hellmuth^{15,16}, Tuuli Lappalainen^{15,16}, Gen Li³⁷, Xin Li⁶, Boxiang Liu^{5,6,38}, Serghei Mangul²³, Mark I. McCarthy^{39,40,41}, Ian C. McDowell⁴², Pejman Mohammadi^{15,16}, Jean Monlong^{9,10,43}, Stephen B. Montgomery^{5,6}, Manuel Muñoz-Aguirre^{9,10,44}, Anne W. Ndungu³⁹, Dan L. Nicolae^{36,45,46}, Andrew B. Nobel^{47,48}, Meritxell Oliva^{36,49}, Halit Ongen^{11,12,13}, John J. Palowitch⁴⁷, Nikolaos Panousis^{11,12,13}, Panagiotis Papasaiakas^{9,10}, YoSon Park¹⁴, Princy Parsana⁸, Anthony J. Payne³⁹, Christine B. Peterson⁵⁰, Jie Quan⁵¹, Ferran Reverter^{9,10,52}, Chiara Sabatti^{53,54}, Ashis Saha⁸, Michael Sammeth⁵⁵, Alexandra J. Scott¹⁸, Andrey A. Shabalin⁵⁶, Reza Sodaei^{9,10}, Matthew Stephens^{45,46}, Barbara E. Stranger^{36,49,57}, Benjamin J. Strober³⁵, Jae Hoon Sul⁵⁸, Emily K. Tsang^{6,31}, Sarah Uebachs⁴⁶, Martijn van de Bunt^{39,40}, Gao Wang⁴⁶, Xiaohan Wen⁵⁹, Fred A. Wright⁶⁰, Hualin S. Xi⁵¹, Esti Yeger-Lotem^{7,61}, Zachary Zappala^{5,6}, Judith B. Zaugg⁶², Yi-Hui Zhou⁶⁰

Enhancing GTEx (eGTEx) groups Joshua M. Akey^{29,63}, Daniel Bates⁶⁴, Joanne Chan⁵, Lin S. Chen¹⁷, Melina Claussnitzer^{1,65,66}, Kathryn Demanelis¹⁷, Morgan Diegel⁶⁴, Jennifer A. Doherty⁶⁷, Andrew P. Feinberg^{35,68,69,70}, Marian S. Fernando^{36,49}, Jessica Halow⁶⁴, Kasper D. Hansen^{68,71,72}, Eric Haugen⁶⁴, Peter F. Hickey⁷², Lei Hou^{1,73}, Farzana Jasmine¹⁷, Ruiqi Jian⁵, Lihua Jiang⁵, Audra Johnson⁶⁴, Rajinder Kaul⁶⁴, Manolis Kellis^{1,73}, Muhammad G. Kibriya¹⁷, Kristen Lee⁶⁴, Jin Billy Li⁵, Qin Li⁵, Xiao Li⁵, Jessica Lin^{5,74}, Shin Lin^{5,75}, Sandra Linder^{5,6}, Caroline Linke^{36,49}, Yaping Liu^{1,73}, Matthew T. Maurano⁷⁶, Benoit Molinier¹, Stephen B. Montgomery^{5,6}, Jemma Nelson⁶⁴, Fidencio J. Neri⁶⁴, Meritxell Oliva^{36,49}, Yongjin Park^{1,73}, Brandon L. Pierce¹⁷, Nicola J. Rinaldi^{1,73}, Lindsay F. Rizzardi⁶⁸, Richard Sandstrom⁶⁴, Andrew Skol^{36,49,57}, Kevin S. Smith^{5,6}, Michael P. Snyder⁵, John Stamatoyannopoulos^{64,74,77}, Barbara E. Stranger^{36,49,57}, Hua Tang⁵, Emily K. Tsang^{6,31}, Li Wang¹, Meng Wang⁵, Nicholas Van Wittenberghe¹, Fan Wu^{36,49}, Rui Zhang⁵



NIH Common Fund Concepcion R. Nierras⁷⁸

NIH/NCI Philip A. Branton⁷⁹, Latarsha J. Carithers^{79,80}, Ping Guan⁷⁹, Helen M. Moore⁷⁹, Abhi Rao⁷⁹, Jimmie B. Vaughn⁷⁹

NIH/NHGRI Sarah E. Gould⁸¹, Nicole C. Lockart⁸¹, Casey Martin⁸¹, Jeffery P. Struwing⁸¹, Simona Volpi⁸¹

NIH/NIMH Anjene M. Addington⁸², Susan E. Koester⁸²

NIH/NIDA A. Roger Little⁸³

Biospecimen Collection Source Site—NDRI Lori E. Brigham⁸⁴, Richard Hasz⁸⁵, Marcus Hunter⁸⁶, Christopher Johns⁸⁷, Mark Johnson⁸⁸, Gene Kopen⁸⁹, William F. Leinweber⁸⁹, John T. Lonsdale⁸⁹, Alisa McDonald⁸⁹, Bernadette Mestichelli⁸⁹, Kevin Myer⁸⁶, Brian Roe⁸⁶, Michael Salvatore⁸⁹, Saboor Shad⁸⁹, Jeffrey A. Thomas⁸⁹, Gary Walters⁸⁸, Michael Washington⁸⁸, Joseph Wheeler⁸⁷

Biospecimen Collection Source Site—RPCI Jason Bridge⁹⁰, Barbara A. Foster⁹¹, Bryan M. Gillard⁹¹, Ellen Karasik⁹¹, Rachna Kumar⁹¹, Mark Miklos⁹⁰, Michael T. Moser⁹¹

Biospecimen Core Resource—VARI Scott D. Jewell⁹², Robert G. Montroy⁹², Daniel C. Rohrer⁹², Dana R. Valley⁹²

Brain Bank Repository—University of Miami Brain Endowment Bank

David A. Davis⁹³, Deborah C. Mash⁹³

Leidos Biomedical—Project Management Anita H. Undale⁹⁴, Anna M. Smith⁹⁵, David E. Tabor⁹⁵, Nancy V. Roche⁹⁵, Jeffrey A. McLean⁹⁵, Negin Vatanian⁹⁵, Karna L. Robinson⁹⁵, Leslie Sobin⁹⁵, Mary E. Barcus⁹⁶, Kimberly M. Valentino⁹⁵, Liqun Qi⁹⁵, Steven Hunter⁹⁵, Pushpa Hariharan⁹⁵, Shilpi Singh⁹⁵, Ki Sung Um⁹⁵, Takunda Matose⁹⁵, Maria M. Tomaszewski⁹⁵

ELSI Study Laura K. Barker⁹⁷, Maghboeba Mosavel⁹⁸, Laura A. Siminoff⁹⁷, Heather M. Traino⁹⁷

Genome Browser Data Integration & Visualization—EBI Paul Flicek⁹⁹, Thomas Juettemann⁹⁹, Magali Ruffier⁹⁹, Dan Sheppard⁹⁹, Kieron Taylor⁹⁹, Stephen J. Trevanion⁹⁹, Daniel R. Zerbino⁹⁹

Genome Browser Data Integration & Visualization—UCSC Genomics Institute, University of California Santa Cruz Brian Craft¹⁰⁰, Mary Goldman¹⁰⁰, Maximilian Haeussler¹⁰⁰, W. James Kent¹⁰⁰, Christopher M. Lee¹⁰⁰, Benedict Paten¹⁰⁰, Kate R. Rosenbloom¹⁰⁰, John Vivian¹⁰⁰, Jingchun Zhu¹⁰⁰



***Donors and
their families***