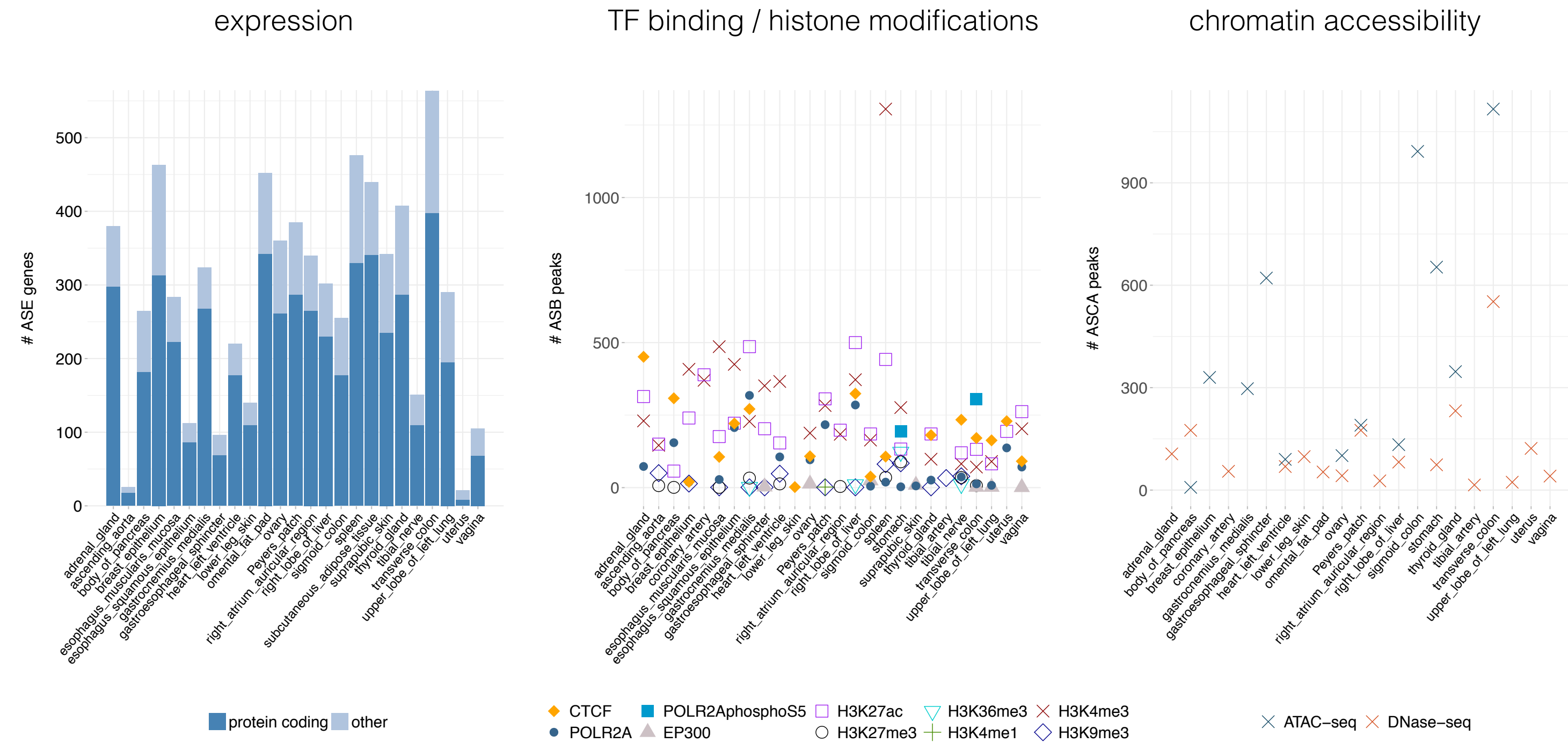


# Summary of allele-specific analyses and current plans

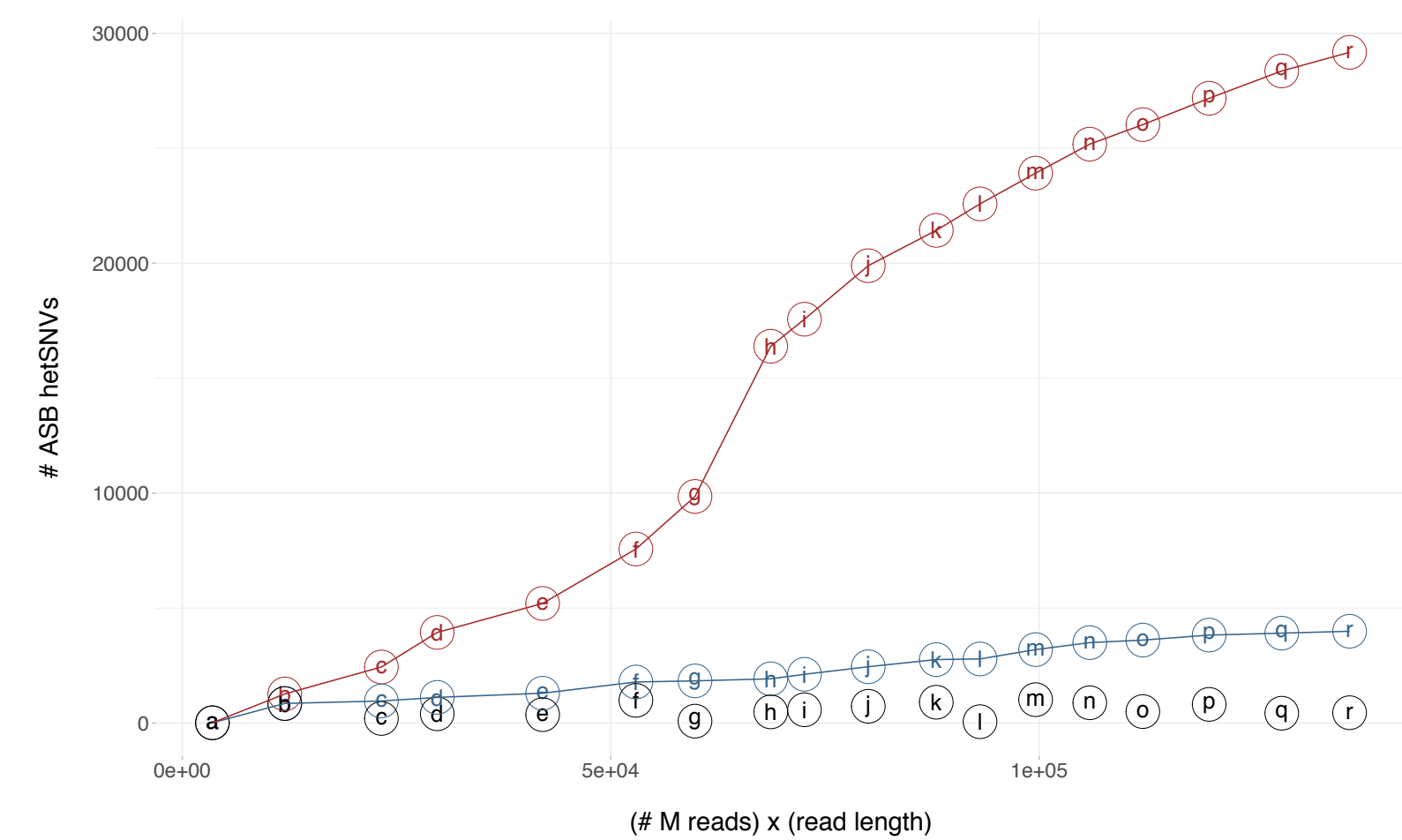
Timur Galeev, Joel Rozowsky, Mark Gerstein

2018-05-09

# Calling allele-specific hetSNVs, genes, and peaks (e.g. shown for ENC-003):



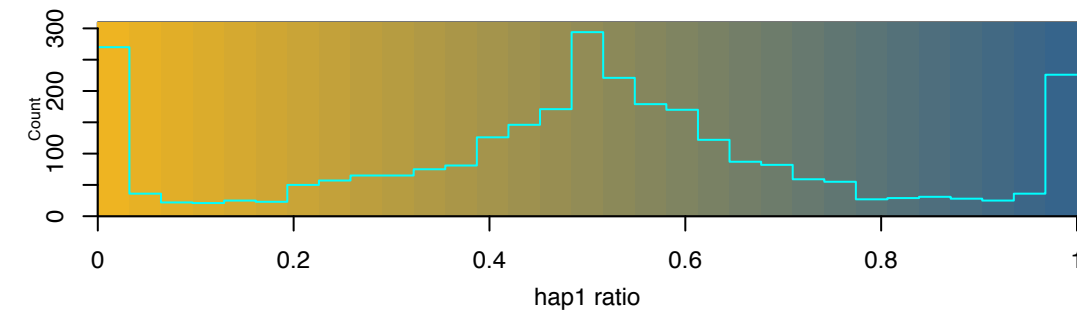
increasing detection power by pooling reads from multiple tissues



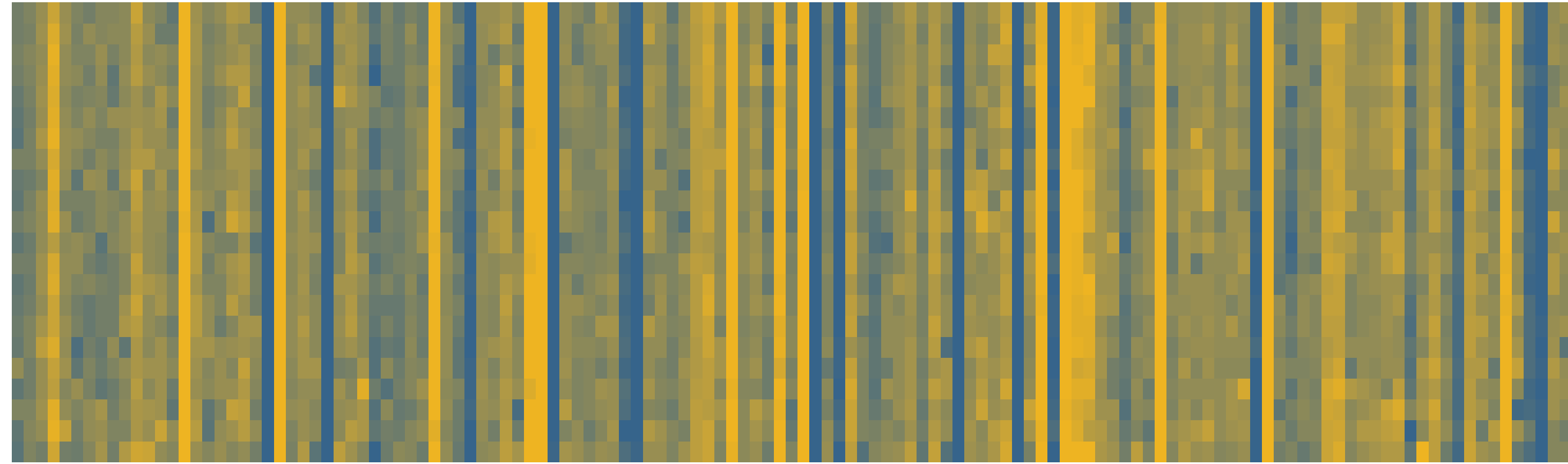
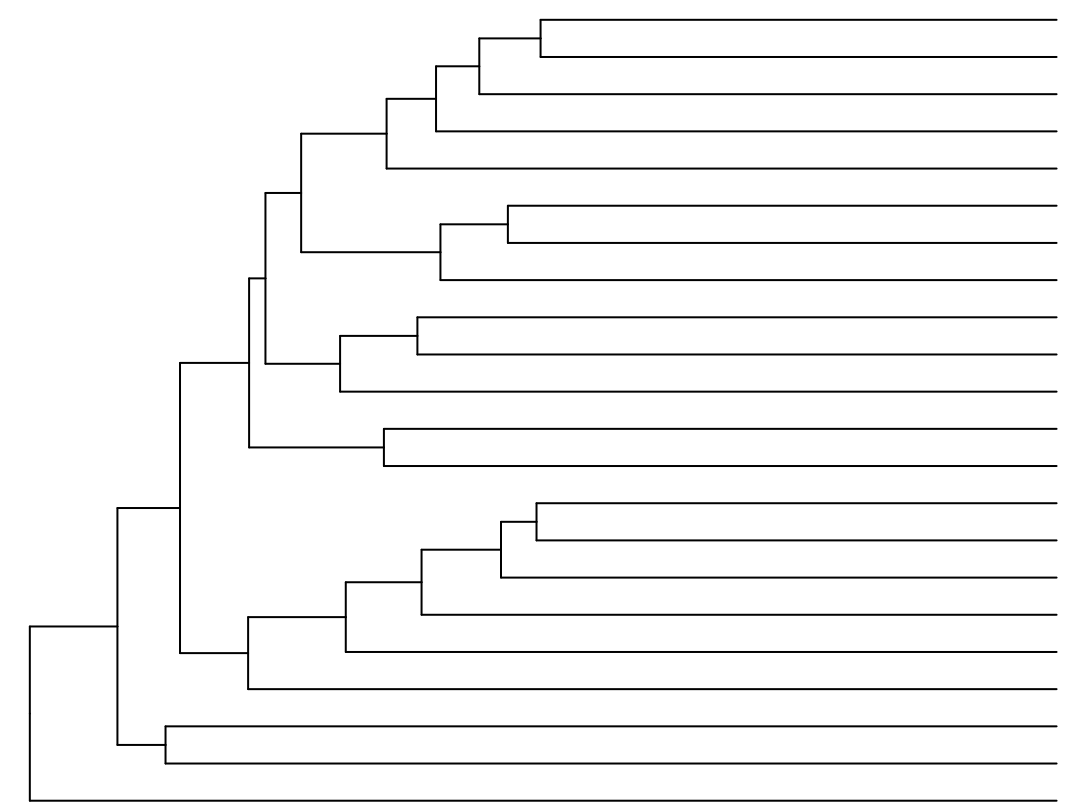
- Most analyses have been focused on ENC-002, ENC-003, will extend when high-quality genomes will be released for the two other individuals;
- add new datasets that became available since last processing;
- will also generate 'combined call-sets' using pooling samples from different tissues.

● called for single dataset	a ENCSR762PCG_CTCF-human_subcutaneous_adipose_tissue	j ENCSR692ILH_CTCF-human_spleen
● combining callsets	b ENCSR770IWO_CTCF-human_adrenal_gland	k ENCSR428BKN_CTCF-human_gastrocnemius_medialis
● called using combined read counts	c ENCSR304XUZ_CTCF-human_breast_epithelium	l ENCSR776AEL_CTCF-human_stomach
	d ENCSR074SFL_CTCF-human_esophagus_muscularis_mucosa	m ENCSR911GFJ_CTCF-human_right_lobe_of_liver
	e ENCSR925GDS_CTCF-human_sigmoid_colon	n ENCSR484DDO_CTCF-human_body_of_pancreas
	f ENCSR469POZ_CTCF-human_tibial_nerve	o ENCSR493APD_CTCF-human_ovary
	g ENCSR430TEE_CTCF-human_lower_leg_skin	p ENCSR970UZD_CTCF-human_upper_lobe_of_left_lung
	h ENCSR392SFJ_CTCF-human_uterus	q ENCSR744YJR_CTCF-human_thyroid_gland
	i ENCSR769WKR_CTCF-human_transverse_colon	r ENCSR606TNN_CTCF-human_vagina

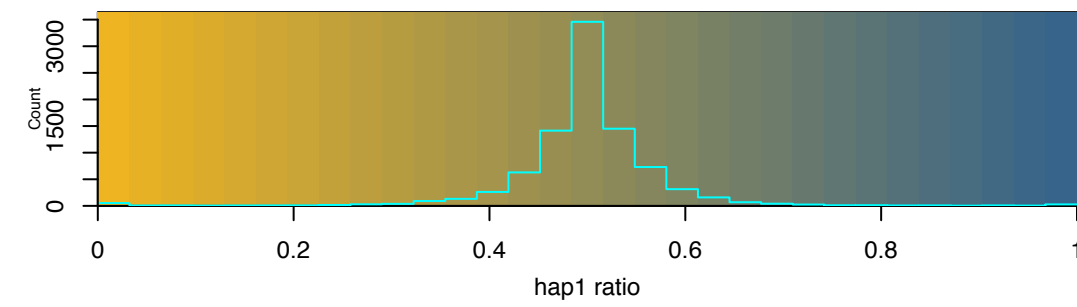
# ASB and ASE across tissues



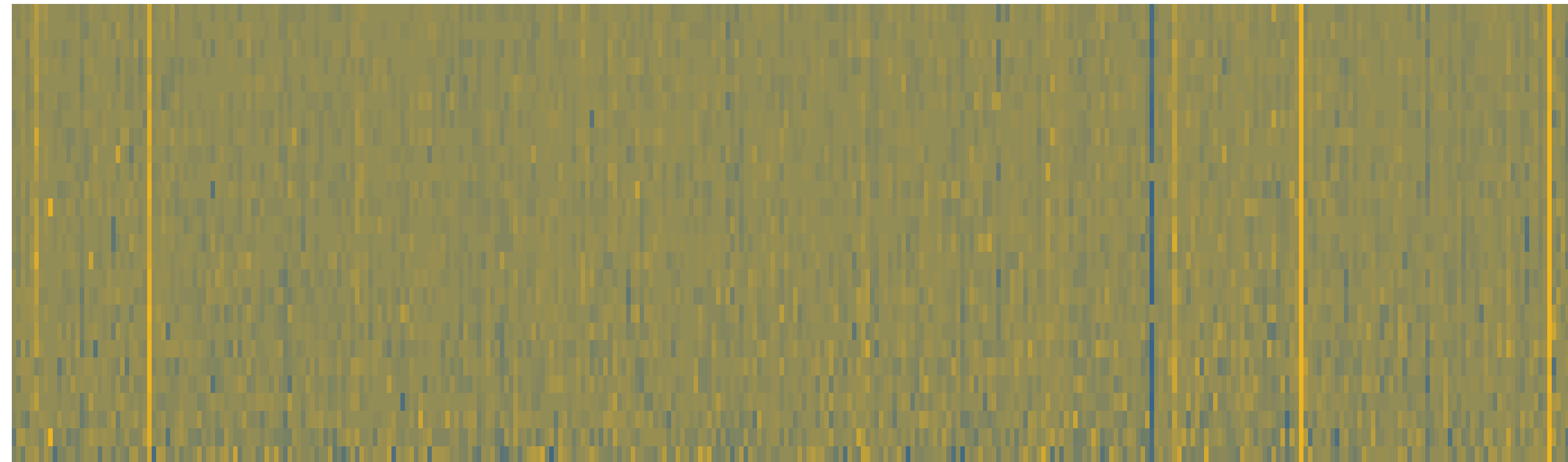
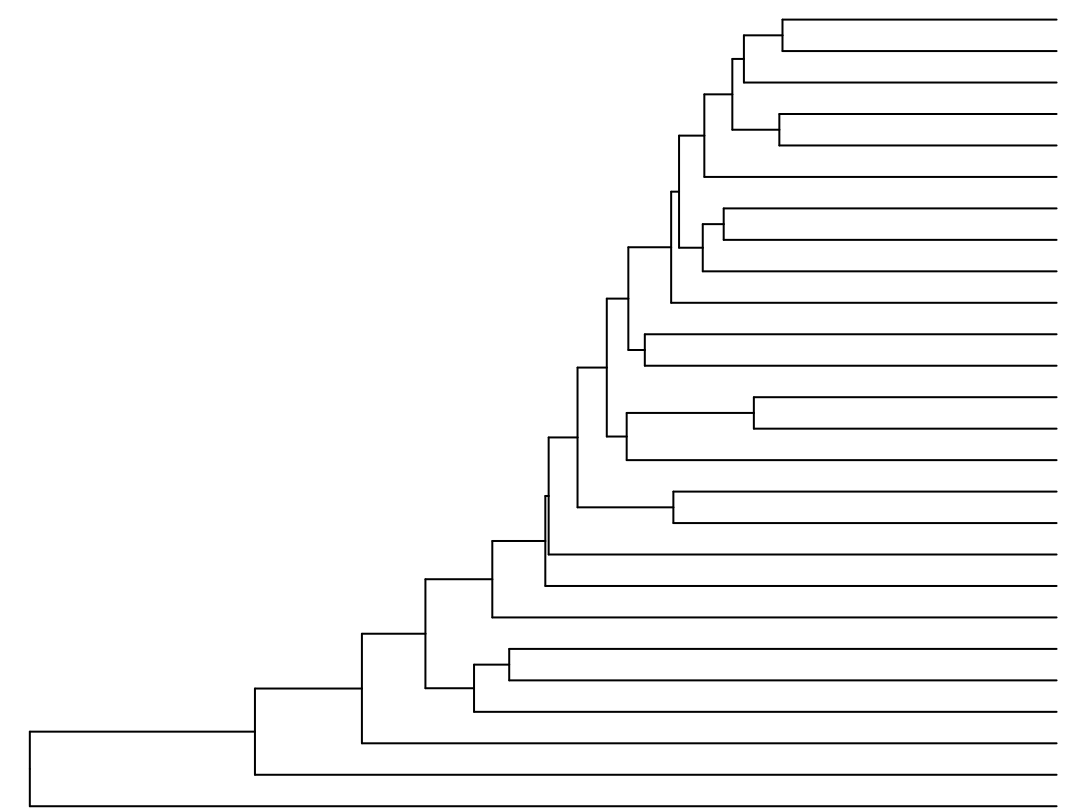
ASB (H3K27ac) in 'cell type-agnostic' candidate *cis*-Regulatory Elements



- ENC-003\_vagina
- ENC-003\_breast\_epithelium
- ENC-003\_coronary\_artery
- ENC-003\_gastroesophageal\_sphincter
- ENC-003\_adrenal\_gland
- ENC-003\_heart\_left\_ventricle
- ENC-003\_sigmoid\_colon
- ENC-003\_esophagus\_muscularis\_mucosa
- ENC-003\_tibial\_nerve
- ENC-003\_stomach
- ENC-003\_gastrocnemius\_medialis
- ENC-003\_transverse\_colon
- ENC-003\_right\_atrium\_auricular\_region
- ENC-003\_right\_lobe\_of\_liver
- ENC-003\_esophagus\_squamous\_epithelium
- ENC-003\_Peyers\_patch
- ENC-003\_spleen
- ENC-003\_upper\_lobe\_of\_left\_lung
- ENC-003\_thyroid\_gland
- ENC-003\_uterus
- ENC-003\_ascending\_aorta
- ENC-003\_body\_of\_pancreas



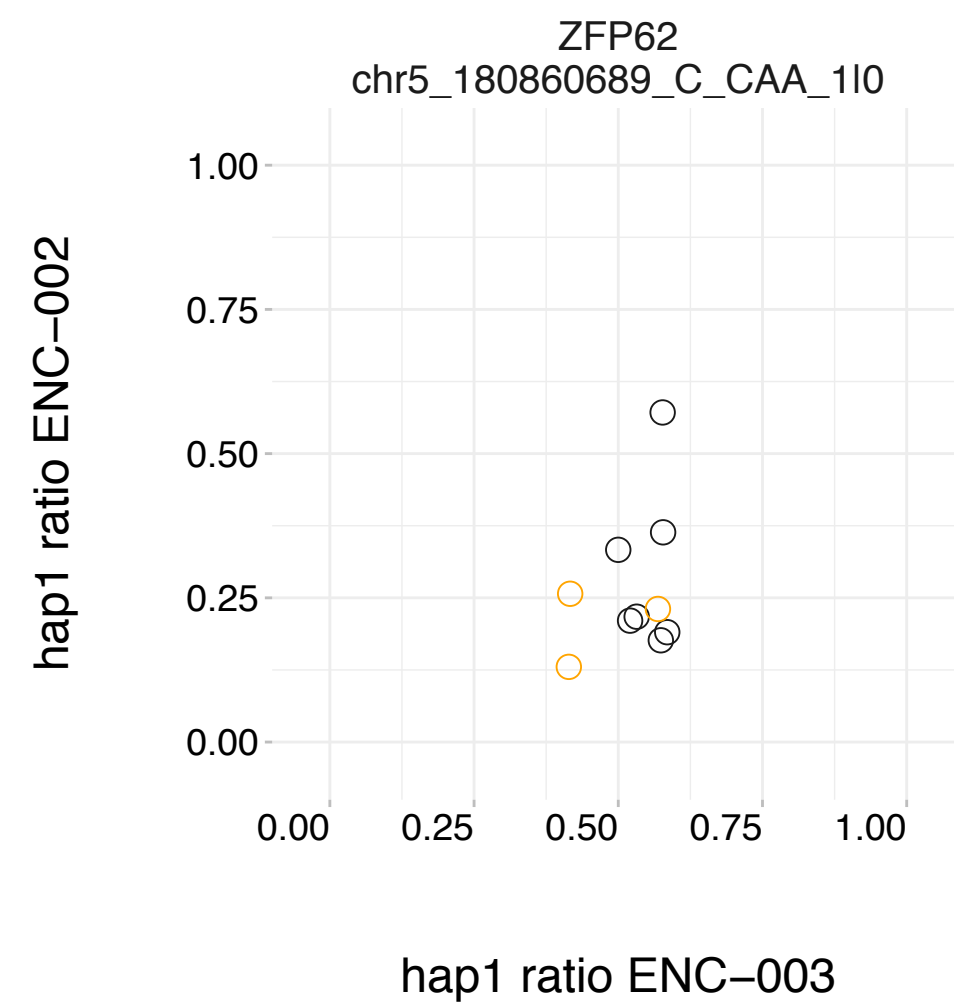
ASE, protein-coding genes



- ENC-003\_omental\_fat\_pad
- ENC-003\_breast\_epithelium
- ENC-003\_tibial\_nerve
- ENC-003\_thyroid\_gland
- ENC-003\_upper\_lobe\_of\_left\_lung
- ENC-003\_suprapubic\_skin
- ENC-003\_sigmoid\_colon
- ENC-003\_ovary
- ENC-003\_spleen
- ENC-003\_subcutaneous\_adipose\_tissue
- ENC-003\_esophagus\_muscularis\_mucosa
- ENC-003\_vagina
- ENC-003\_transverse\_colon
- ENC-003\_Peyers\_patch
- ENC-003\_body\_of\_pancreas
- ENC-003\_right\_atrium\_auricular\_region
- ENC-003\_heart\_left\_ventricle
- ENC-003\_adrenal\_gland
- ENC-003\_right\_lobe\_of\_liver
- ENC-003\_gastrocnemius\_medialis
- ENC-003\_esophagus\_squamous\_epithelium
- ENC-003\_gastroesophageal\_sphincter
- ENC-003\_lower\_leg\_skin
- ENC-003\_ascending\_aorta
- ENC-003\_uterus
- ENC-003\_stomach

# Relating allele-specific behavior with variation

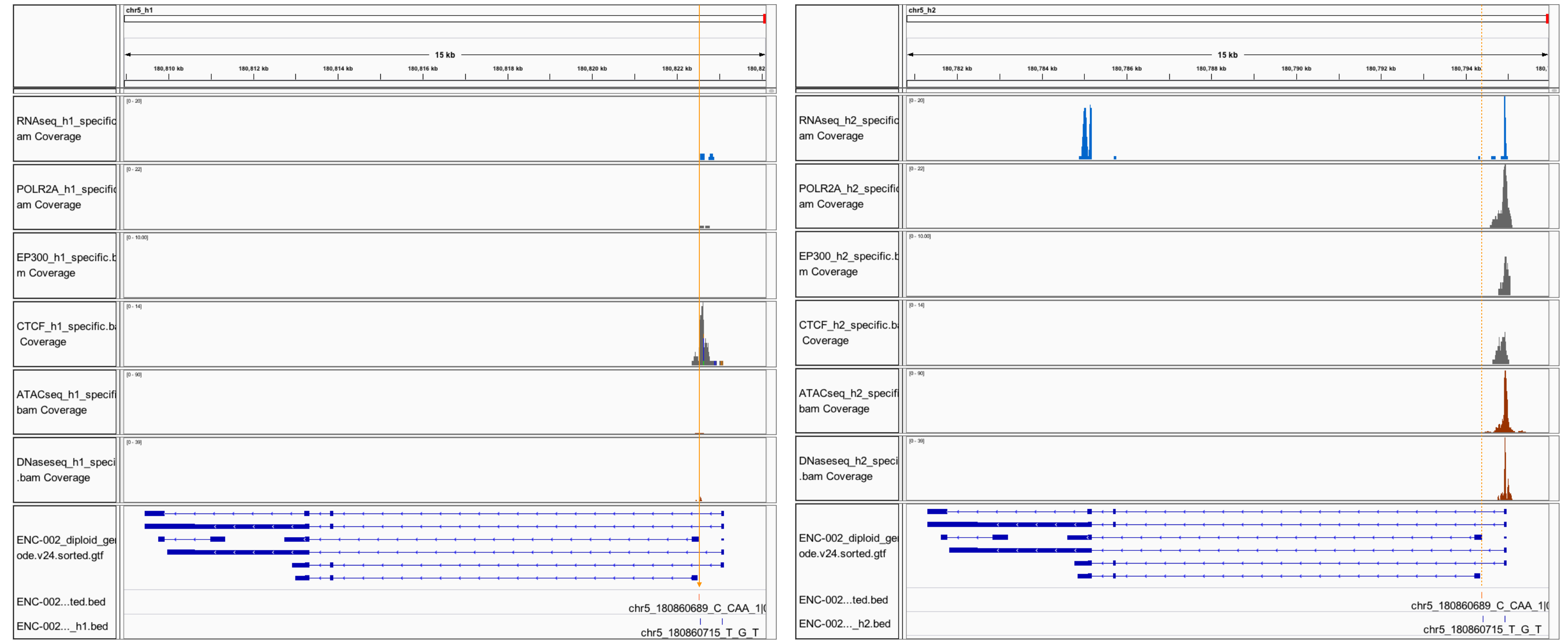
- Variants in regulatory elements



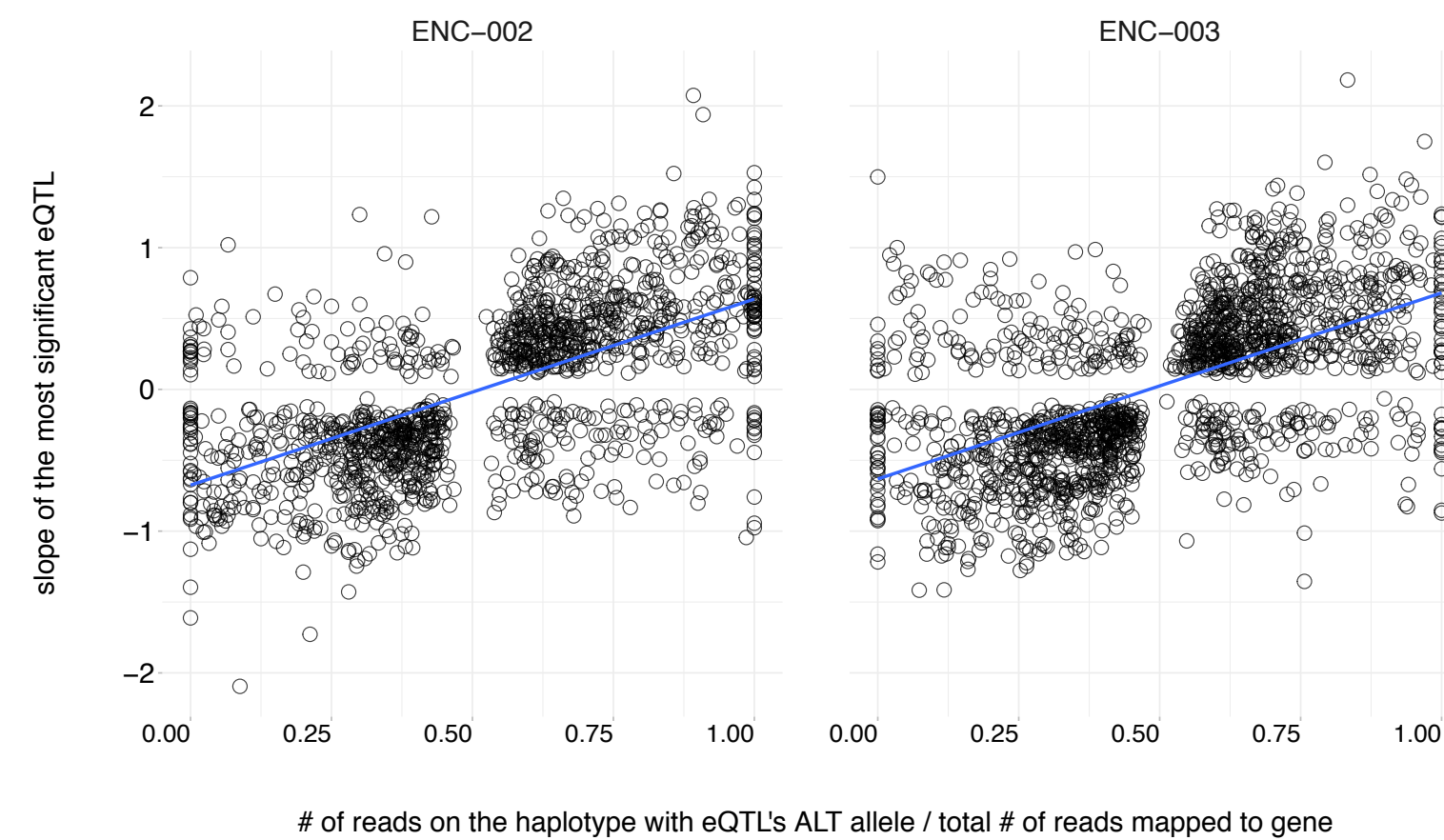
hap1

2bp INS in hap1

hap2



- Relationship with known eQTLs

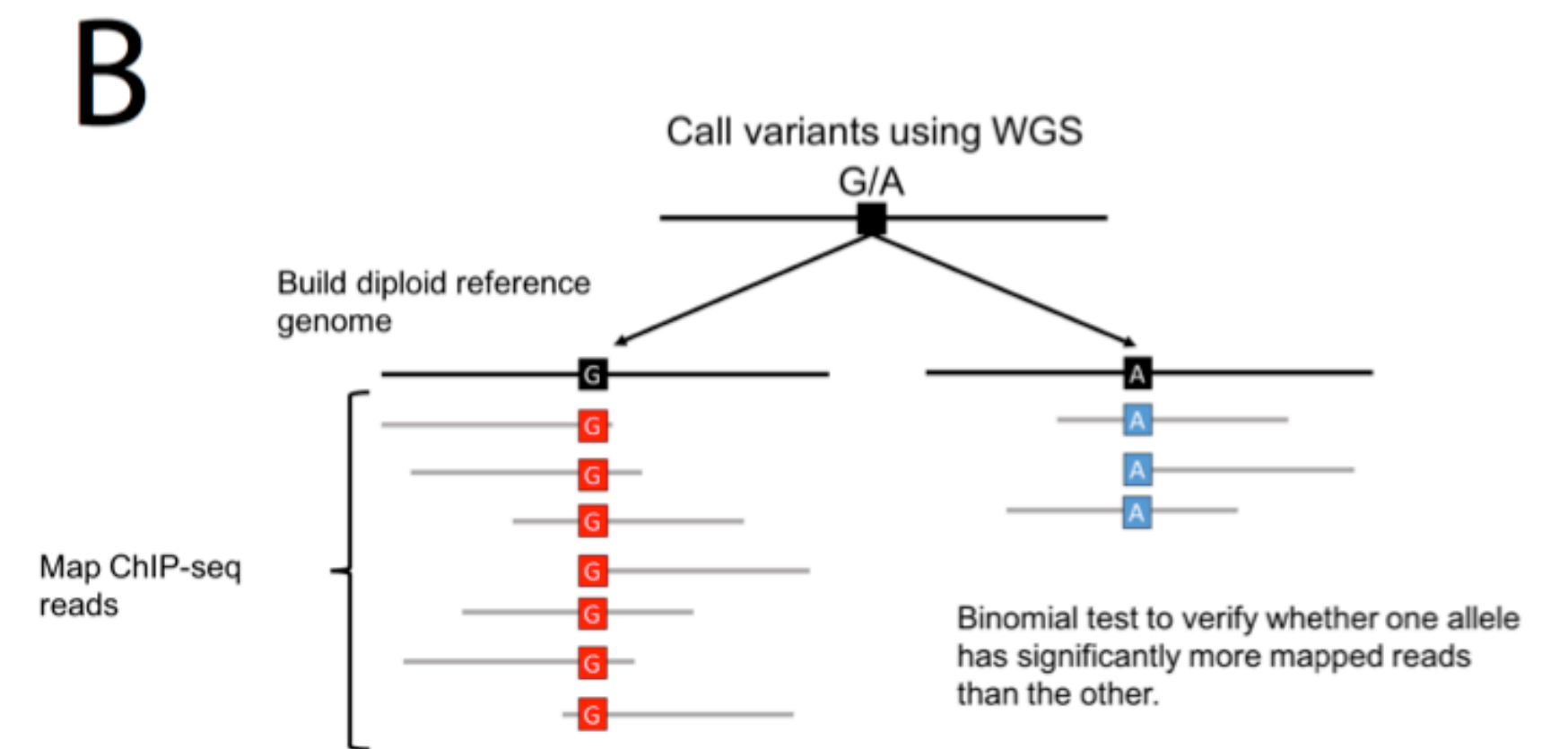
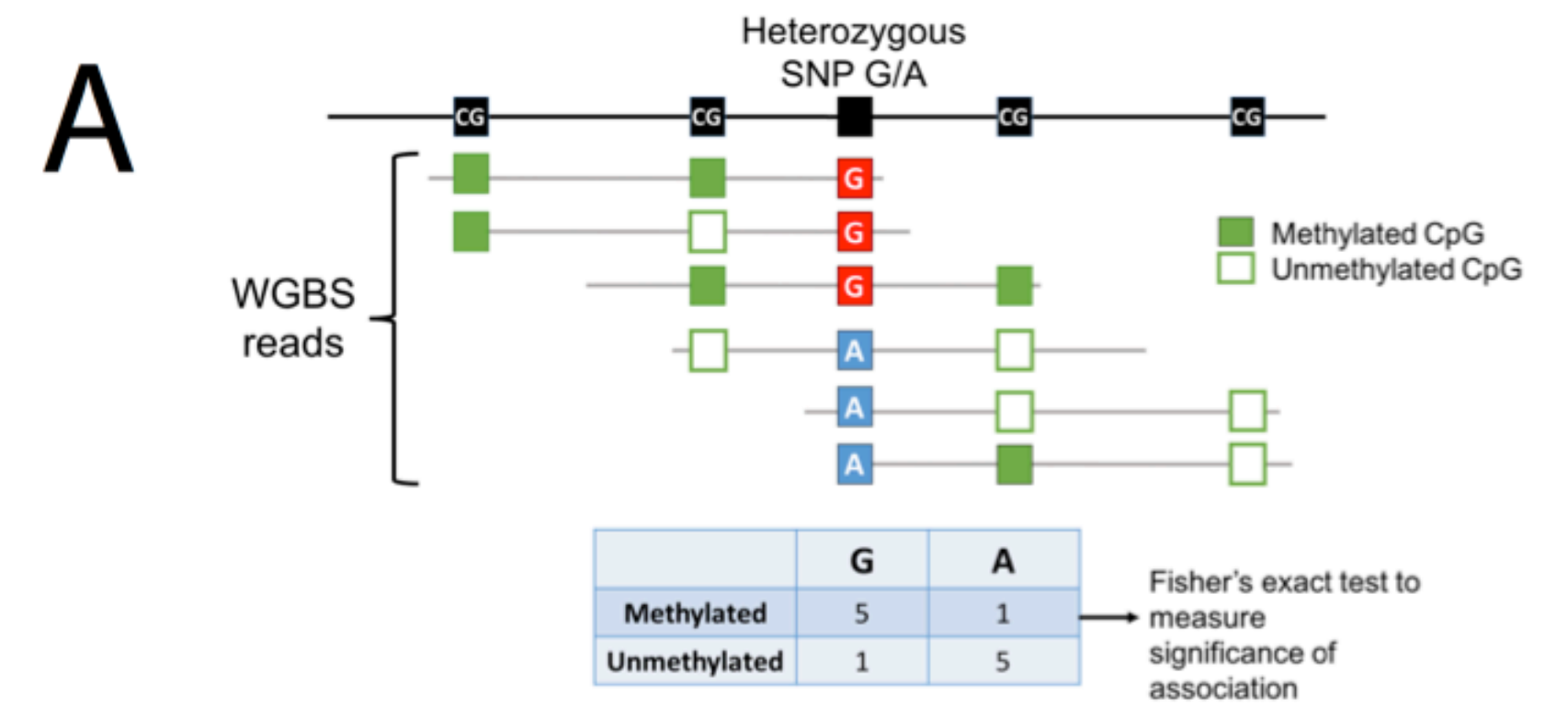


# Plans for calling allele-specific methylation (ASM) events

The screenshot shows the ENCODE Data Portal interface. At the top, there are navigation links for ENCODE, Data, Encyclopedia, Materials & Methods, and Help, along with a search bar. The main content area is titled "Experiment Matrix" and includes a search filter (currently "entex"). Below this, there are several data tables:

- Assay:** A list of assays with their counts: ChIP-seq (664), DNAm array (98), total RNA-seq (91), RAMPAGE (82), DNase-seq (69), small RNA-seq (50), ATAC-seq (46), microRNA counts (24), microRNA-seq (24), **WGBS (22)**, genotyping HTS (12), Hi-C (8), eCLIP (4), and genotyping array (4).
- Assay category:** DNA methylation (22).
- Date released:** January, 2018 (12), December, 2017 (9), April, 2018 (1).
- Available data:** fastq (22), bam (21), bed bedMethyl (21), bigBed bedMethyl (21), bigWig (21).

Below the tables are filter sections for Organism (Homo sapiens, 22), Biosample type (tissue, 22), Organ (gonad: 4, limb: 4, skin of body: 3, adrenal gland: 2, lung: 2, nerve: 2, ovary: 2, pancreas: 2, spleen: 2, stomach: 2, testis: 2, thyroid gland: 2, prostate gland: 1), Project (ENCODE, 22), Genome assembly (visualization) (GRCh38, 21), and Lab (Richard Myers, HAIB, 22). A central section shows "22 results" for the "WGBS" assay, listing various tissues and their counts: adrenal gland (2), body of pancreas (2), lower leg skin (2), ovary (2), spleen (2), stomach (2), testis (2), thyroid gland (2), tibial nerve (2), upper lobe of left lung (2), prostate gland (1), and suprapubic skin (1).



- and other assays, e.g. smallRNA-seq, RAMPAGE, Hi-C

Integration of multiple chromatin assays to  
understand allele-specific gene expression  
mechanism

GG, TG and Gerstein Lab

# Allele – specific regulation of MYBPC2 gene example from ENC-002 muscle tissue

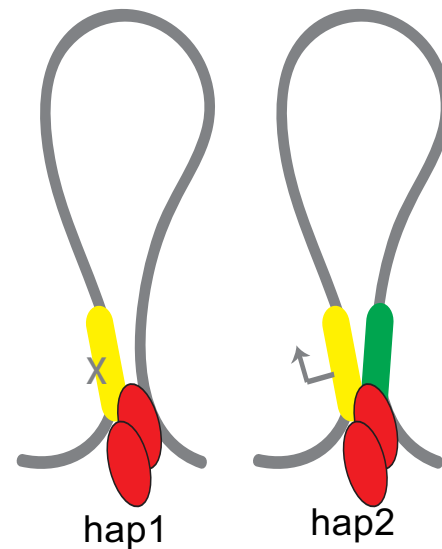


**ASE gene: MYBPC2**  
Chr19: 50,432,903 – 50,466,321

**ASB H3K27ac peaks:**  
~Chr19: 50,360,000 – 50,370,000

**Hi-C loop anchor 1:**  
Chr19: 50,365,000 – 50,370,000

**Hi-C loop anchor 2:**  
Chr19: 50,460,000 – 50,465,000  
(slightly more loops in hap2 but no significant allelic imbalance)





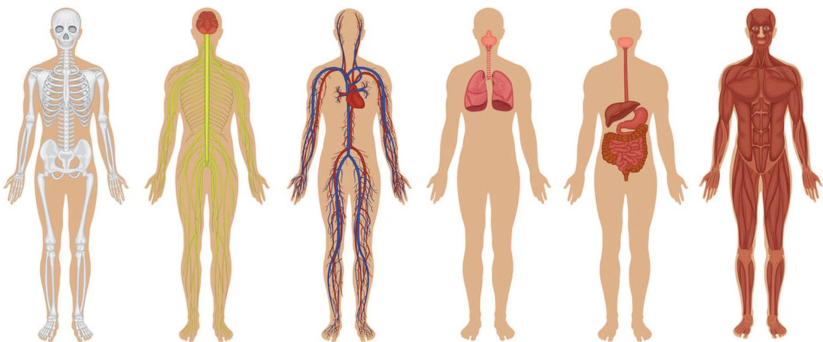
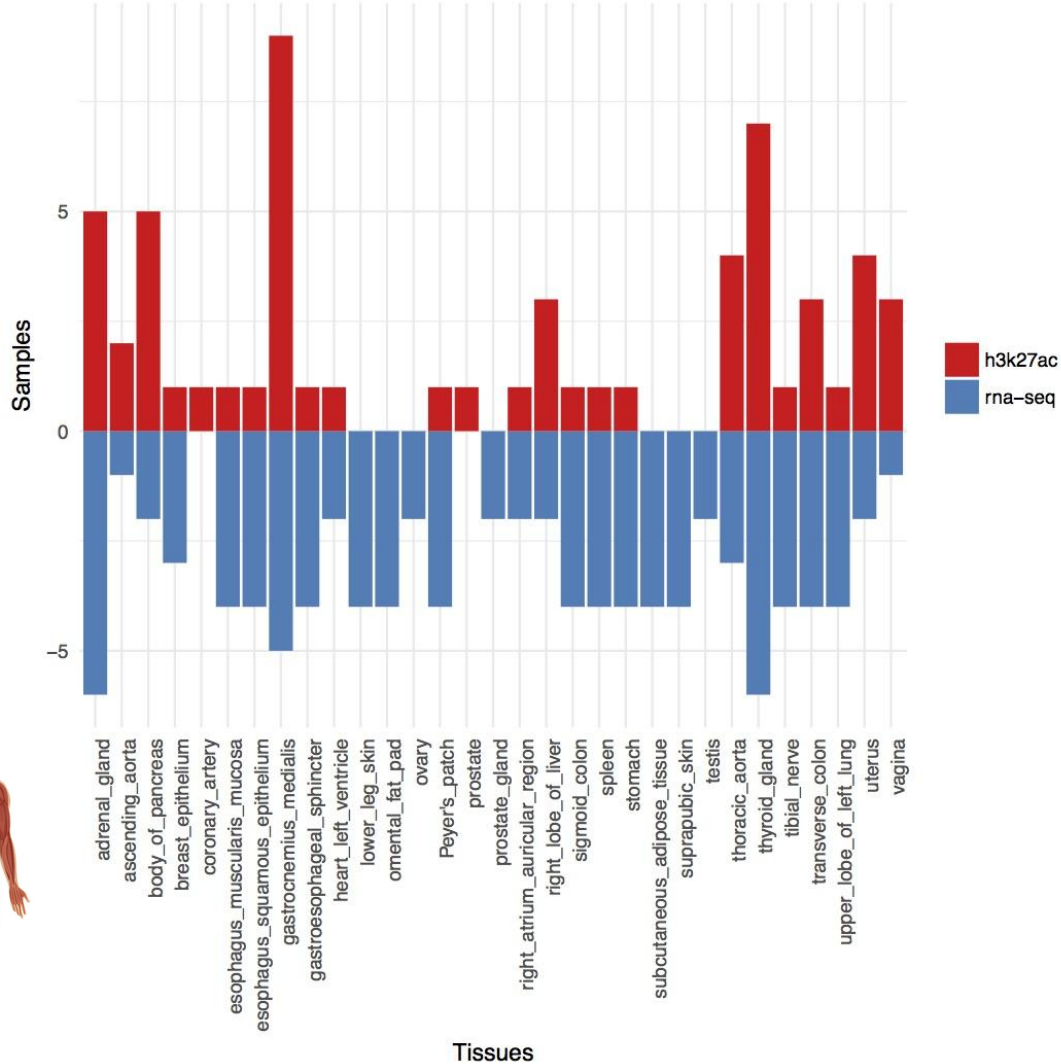


# Global Inter-tissue spectral analysis

Fabio Navarro

# Datasets

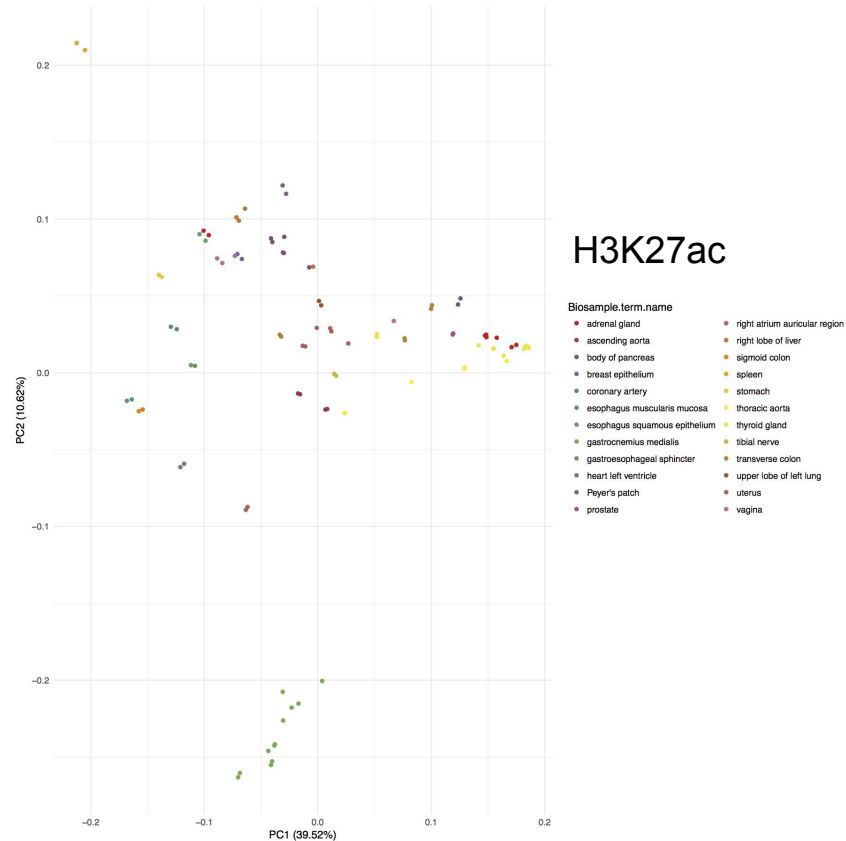
- Future analysis
  - RAMPAGE
  - ATAC-Seq/DNase
  - Other Hist mod/RNA-Pol2





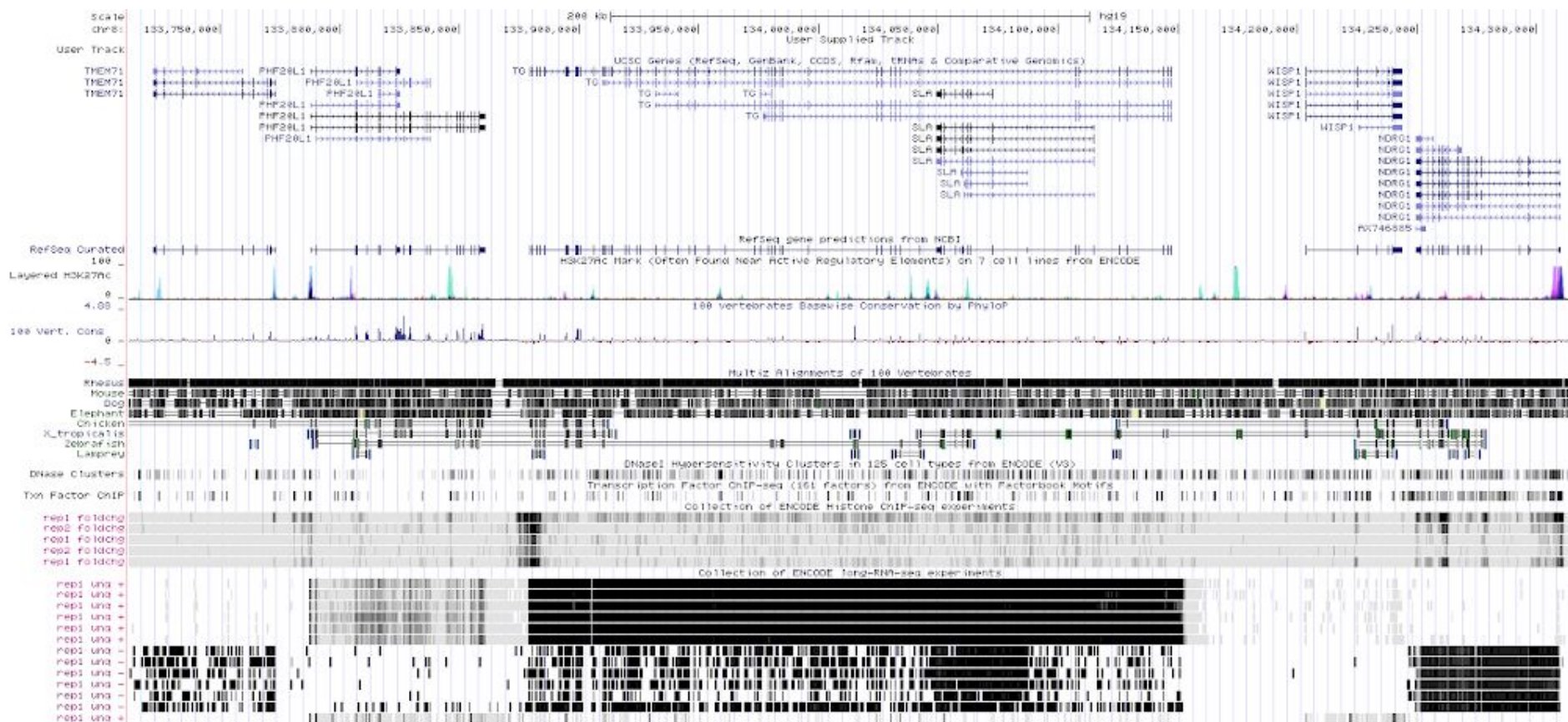


# Dimensionality Reduction (PCA)





# Inter individual/tissue variation



# eQTL enrichment analyses

Mengting Gu and Yucheng T. Yang  
Gerstein Lab



# Enrichment of eQTL/GWAS SNPs in functional regulatory regions in the genome

- Identify consistent potential regulatory regions in the genome
- Test for enrichment of eQTL/GWAS SNPs in active regions for each tissue

## ENC-003



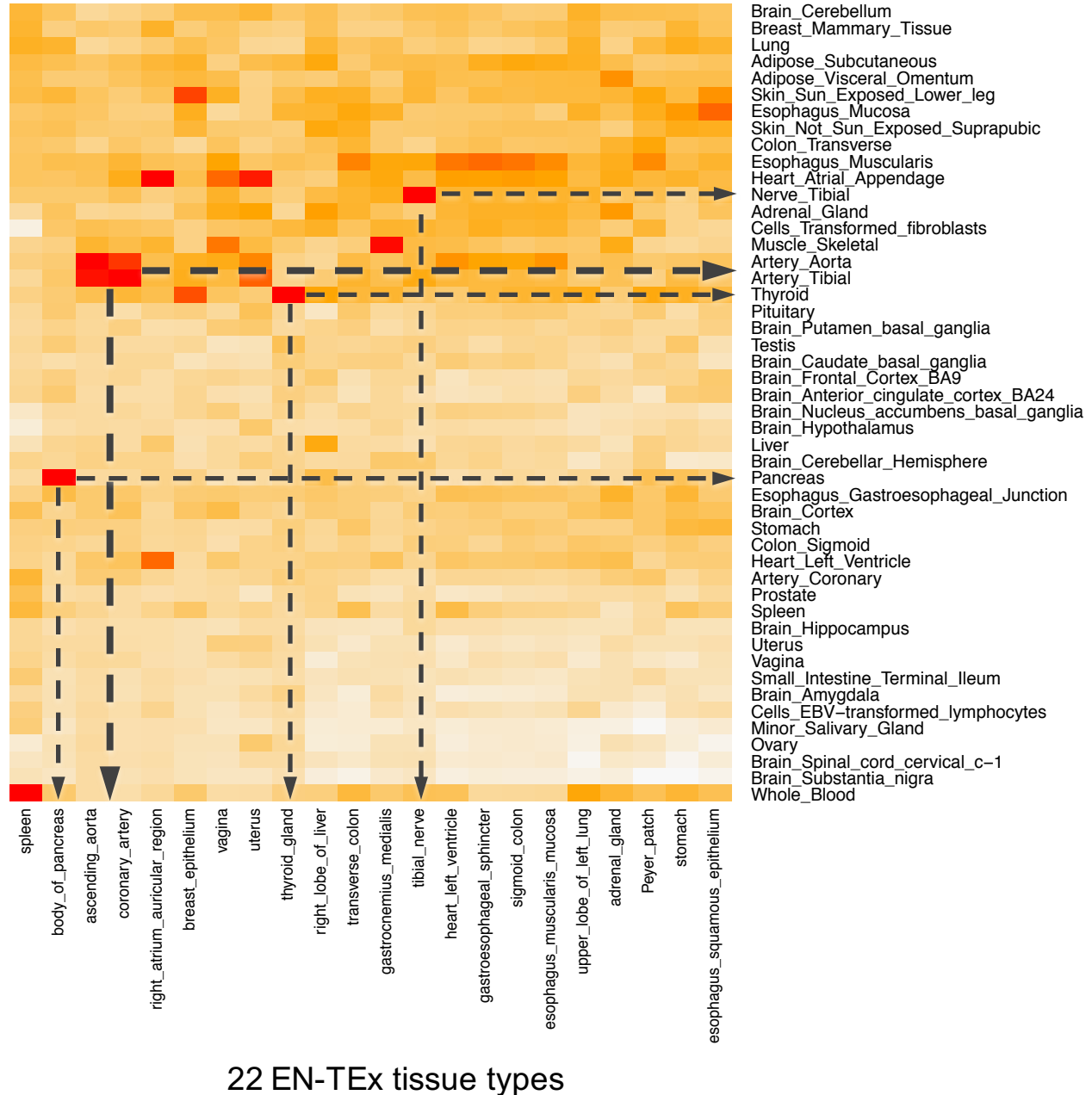
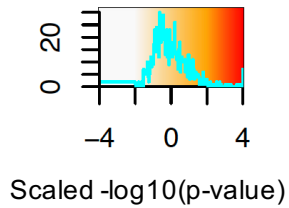
# Enrichment of eQTL/GWAS SNPs in functional regulatory regions in the genome

- Identify consistent potential regulatory regions in the genome
- Test for enrichment of eQTL/GWAS SNPs in active regions for each tissue
  
- We are developing a list of enhancers for EN-TEX tissues using DNA accessibility, histone modification and transcription factor binding data in each tissue
- Consistent candidate cis regulatory elements (ENCODE annotation)

# Enrichment of eQTL/GWAS SNPs in functional regulatory regions in the genome

- Identify consistent potential regulatory regions in the genome
- Test for enrichment of eQTL/GWAS SNPs in active regions for each tissue
  
- We are developing a list of enhancers for EN-TEX tissues using DNA accessibility, histone modification and transcription factor binding data in each tissue
- Consistent candidate cis regulatory elements (ENCODE annotation)
  
- Proof of concept analysis:
  - Enrichment test based on merged H3K27ac peaks
  - Use top ranked H3K27ac peaks in each tissue (excluding promoters) as putative enhancers

# Preliminary results of eQTL enrichment for H3K27ac peaks



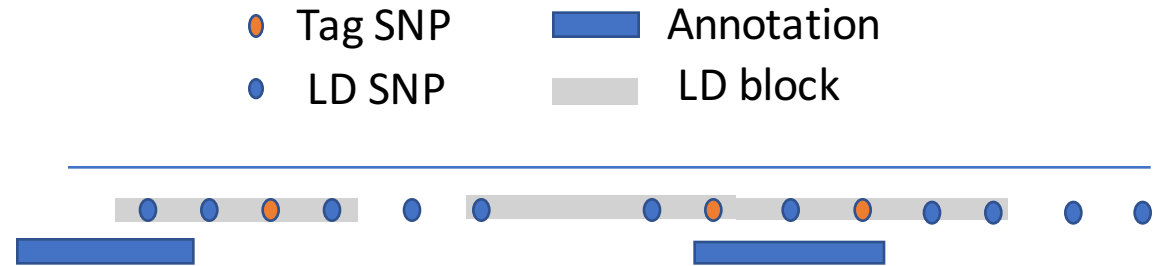
# Possible artifacts in GWAS/eQTL enrichment analysis

Jing Zhang  
Gerstein Lab

# Possible Artifacts in GWAS/eQTL enrichment analysis

## Artifacts from missing LD SNPs

- Especially a problem in GWAS
- LD pruning before the test
- Tag SNP may sit outside of annotation while causal SNP is its LD SNP
- True for eQTL analysis if 1kg SNPs are used as background



## Artifacts from duplicating LD SNPs

- Two LD SNPs are linked to two Tag SNPs



## Artifacts from size of LD blocks

- LD block changes in size up to several order

## Artifacts from size of SNP allele frequency blocks

- GWAS and eQTL SNPs are relatively common SNPs



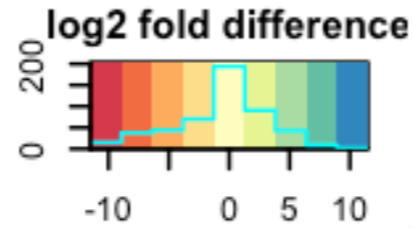
## Artifacts from distance to annotations

- Gene proximal and distal

# SV impact on allele-specific expression

Xiangmeng Kong, Mark Gerstein  
Yale University

# SV overlapped exons in ENC002



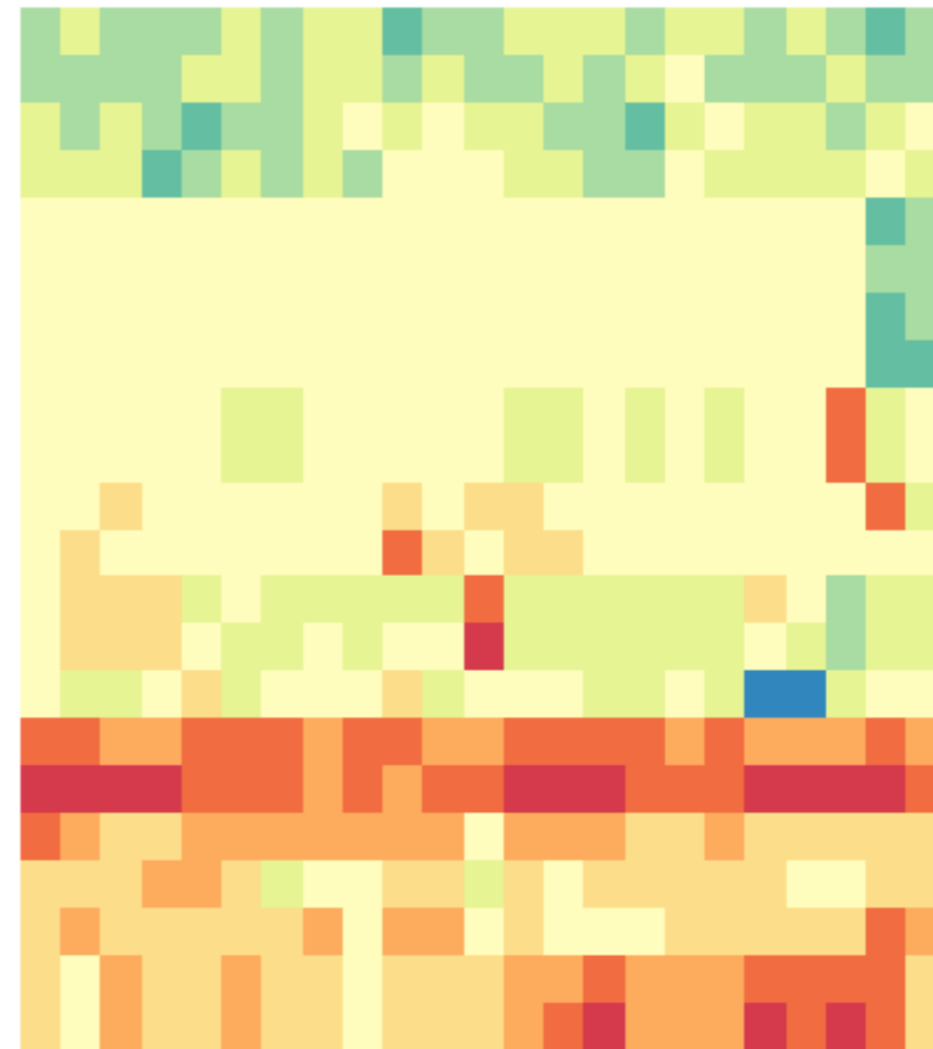
Total SVs: 21672

Overlapped exons: 926 (906 in Personal Genome)

Only reads uniquely mapped to one haplotype

$$\log_2 \left( \frac{h1 \text{ counts} + 1}{h2 \text{ counts} + 1} \right)$$

Significant fold change in  $\geq 1$  tissue: 265 exons



ENSE00001832945.1  
 ENSE00002309516.1  
 ENSE00001321274.5  
 ENSE00001412014.3  
 ENSE00003738589.1  
 ENSE00003723329.1  
 ENSE00003725530.1  
 ENSE00003723018.1  
 ENSE00003770192.1  
 ENSE00002069705.1  
 ENSE00003733255.1  
 ENSE00002019072.2  
 ENSE00003739352.1  
 ENSE00002252699.1  
 ENSE00001503744.5  
 ENSE00001757671.1  
 ENSE00001561414.3  
 ENSE00003759692.1  
 ENSE00001346022.4  
 ENSE00003504186.2  
 ENSE00002444536.1  
 ENSE00002436532.1

thyroid\_gland  
 thyroid\_gland  
 prostate\_gland  
 adrenal\_gland  
 adrenal\_gland  
 subcutaneous\_adipose\_tissue  
 upper\_lobe\_of\_left\_lung  
 spleen  
 testis  
 testis  
 gastrocnemius\_medialis  
 sigmoid\_colon  
 esophagus\_muscularis\_sphincter  
 thoracic\_mucosa  
 aorta  
 stomach  
 tibial\_nerve  
 lower\_leg\_skin  
 suprapubic\_skin  
 esophagus\_squamous\_epithelium  
 transverse\_colon  
 Peyers\_patch



# SV impact gene cases

p < 0.05 in all samples: 2

chr9 133019485 133020874 ENSE00001561414.3

chr13 110936758 110939784 ENSE00001757671.1

chr9 133019413 133021194 DEL GT:DR:DV1|0:35:17

chr13 110938283 110938283 INS GT:DR:DV1|0:29:35

