RNA-seq analysis of EN-TEx data

Anna Vlasova
Computational Biology of RNA processing Lab, CRG
Thomas Gingeras, Michael Schatz, Roderic Guigó

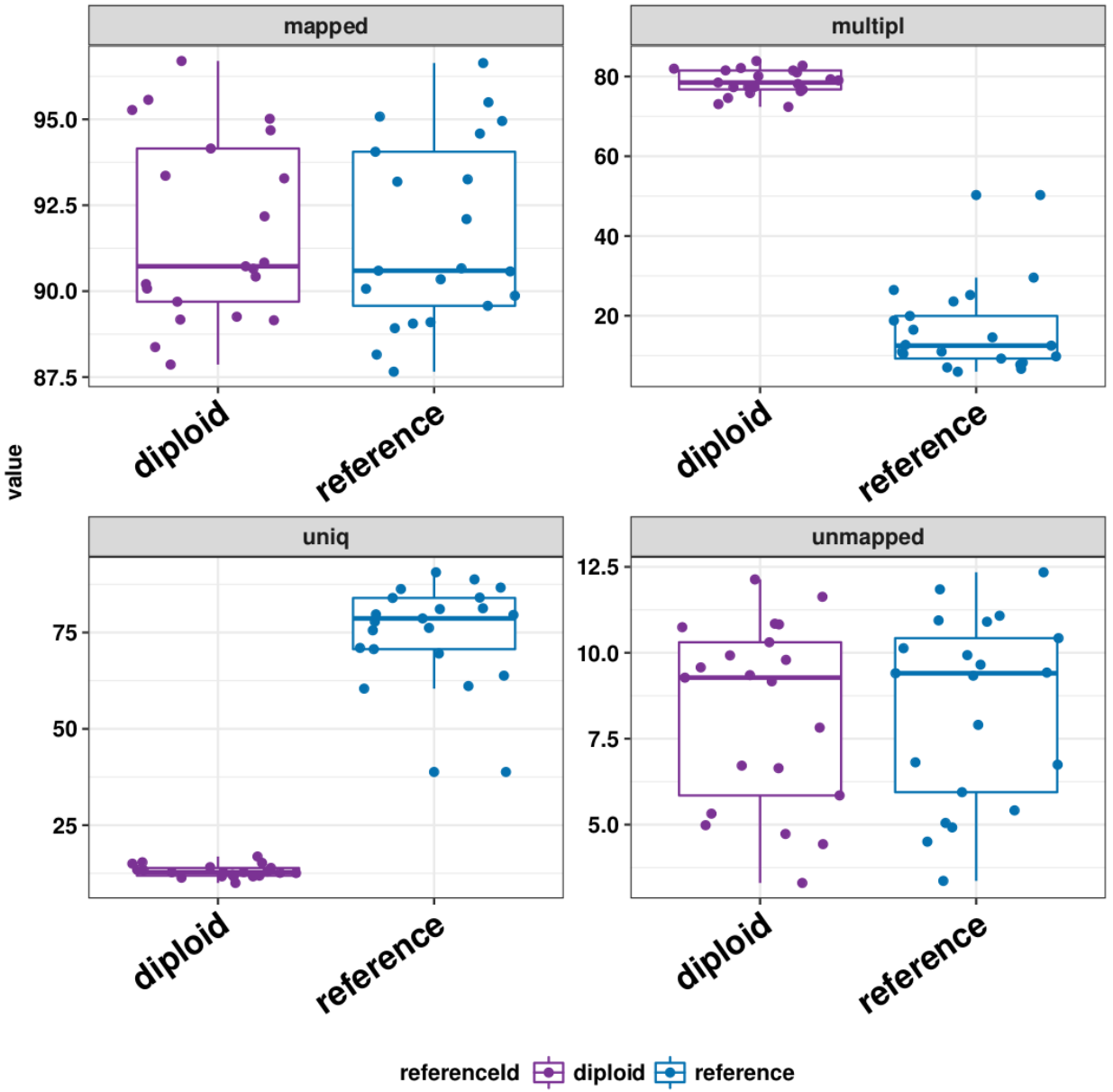
20/11/17
EN-TEx working group call

Outline

- RNAseq mappings to the diploid and reference genomes
- Effect of deletions(DEL) to the gene expression
- Novel transcriptional elements in insertions (INS)

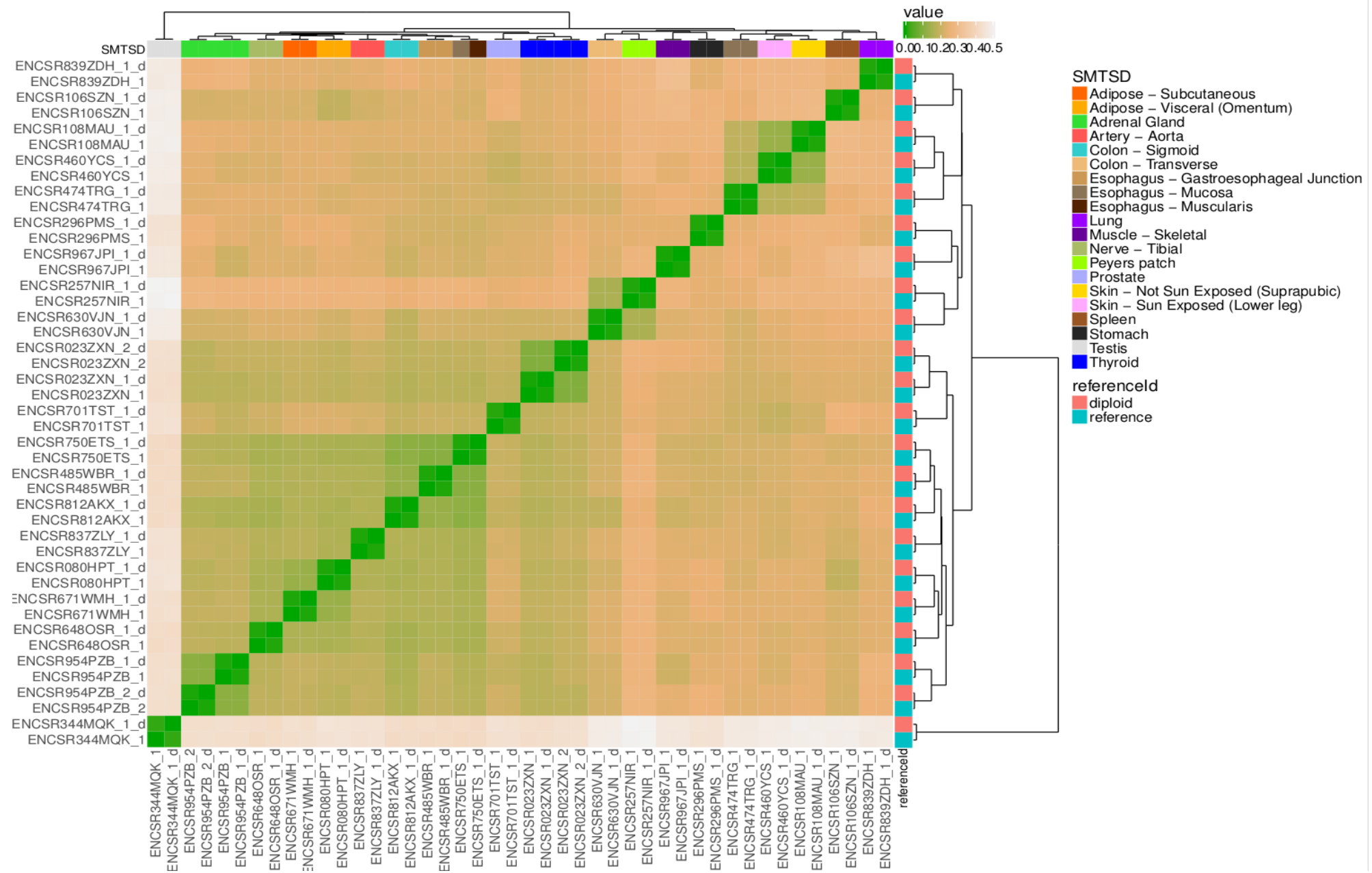
RNAseq mappings to the diploid and reference genomes

Mapping statistics



In the diploid genome 10,000 – 80,000 reads more mapped, compare to the ref.

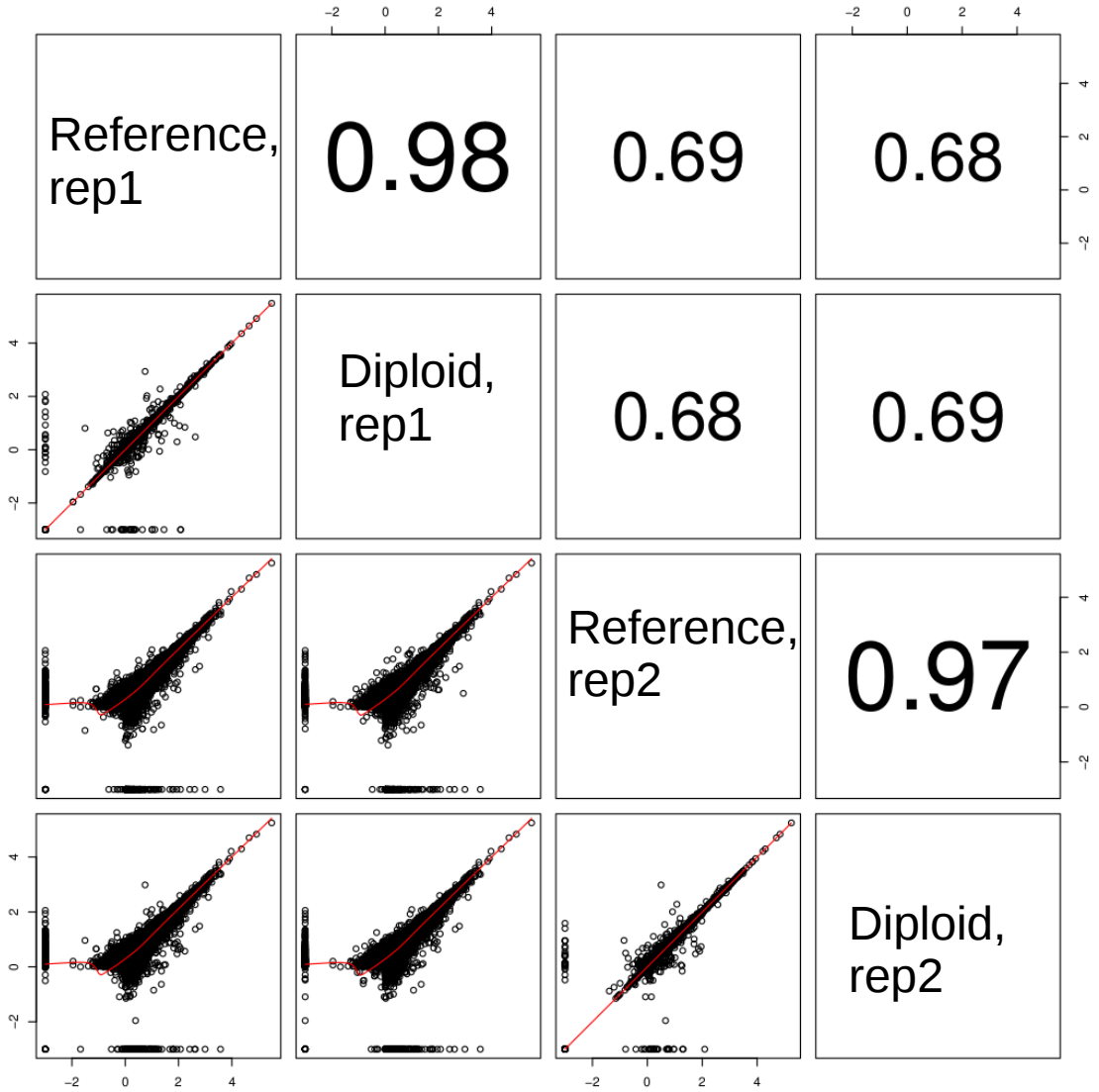
Clustering of the samples; reference vs diploid



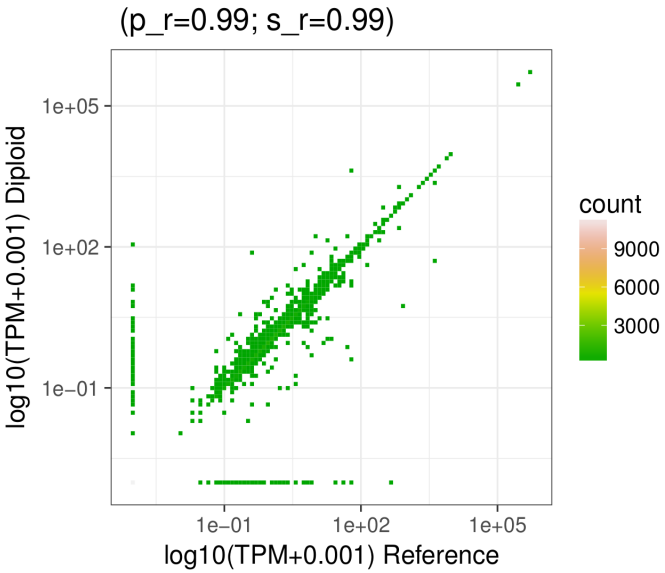
distance = (1-pearson correlation), ENC002

Correlation between samples; reference vs diploid

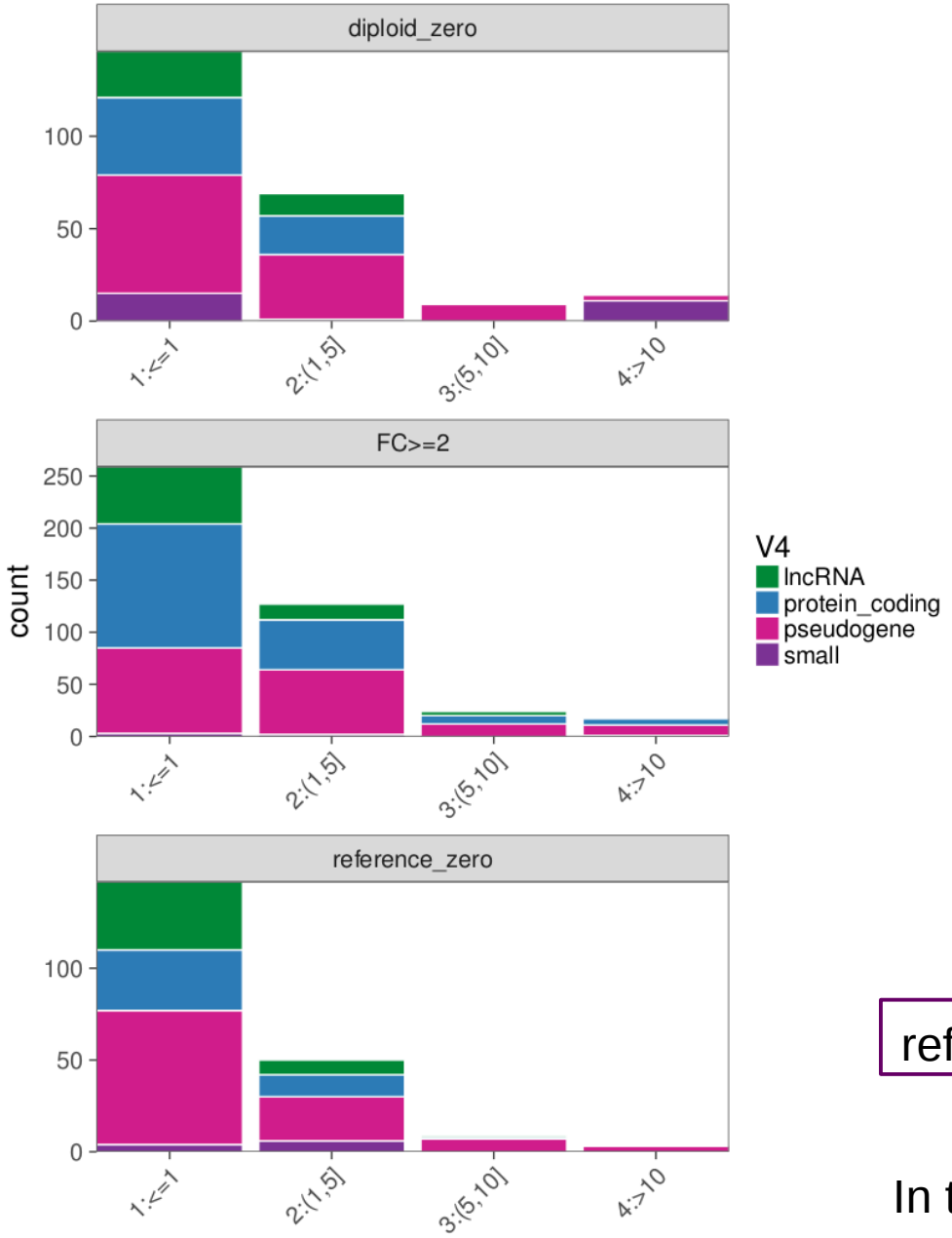
Thyroid-gland_ENCSR023ZXN, sample with replicates
17,295 genes TPM ≥ 1 in at least one sample



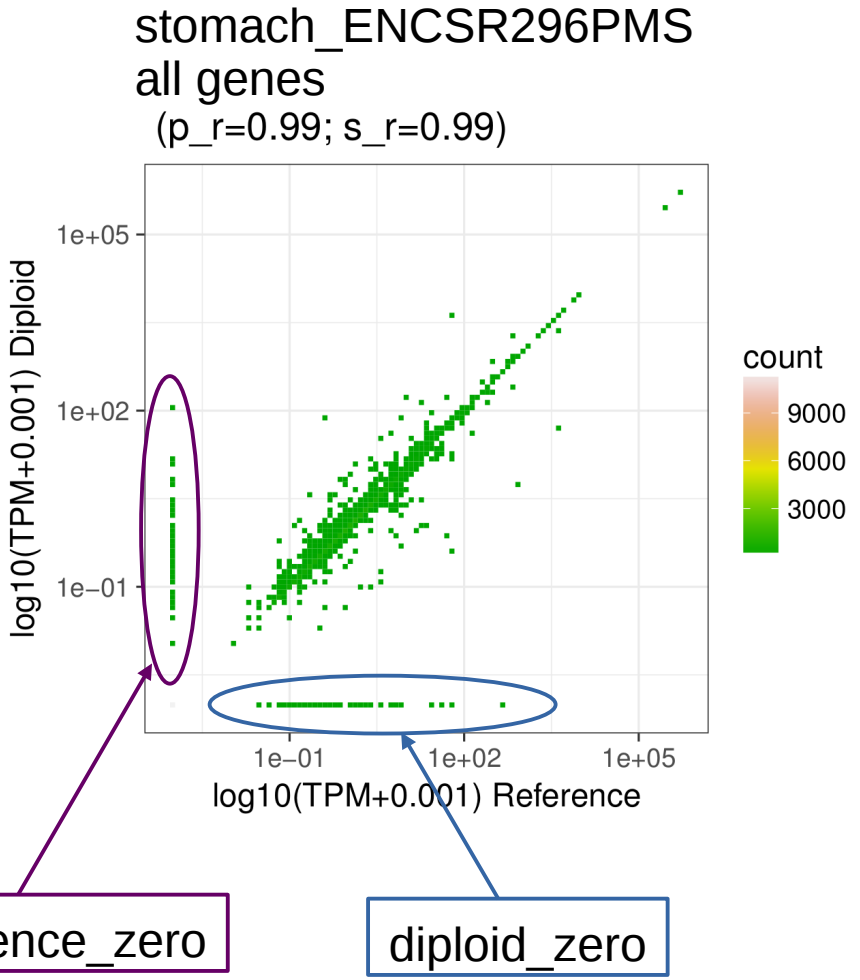
For all genes, 61,467 pearson correlation >0.99 across all samples



Gene expression quantification



Number of tissues



In total, 728 genes quantified differently

In total there are **252** protein coding genes changed their expression in one or few tissues

GO enrychment terms

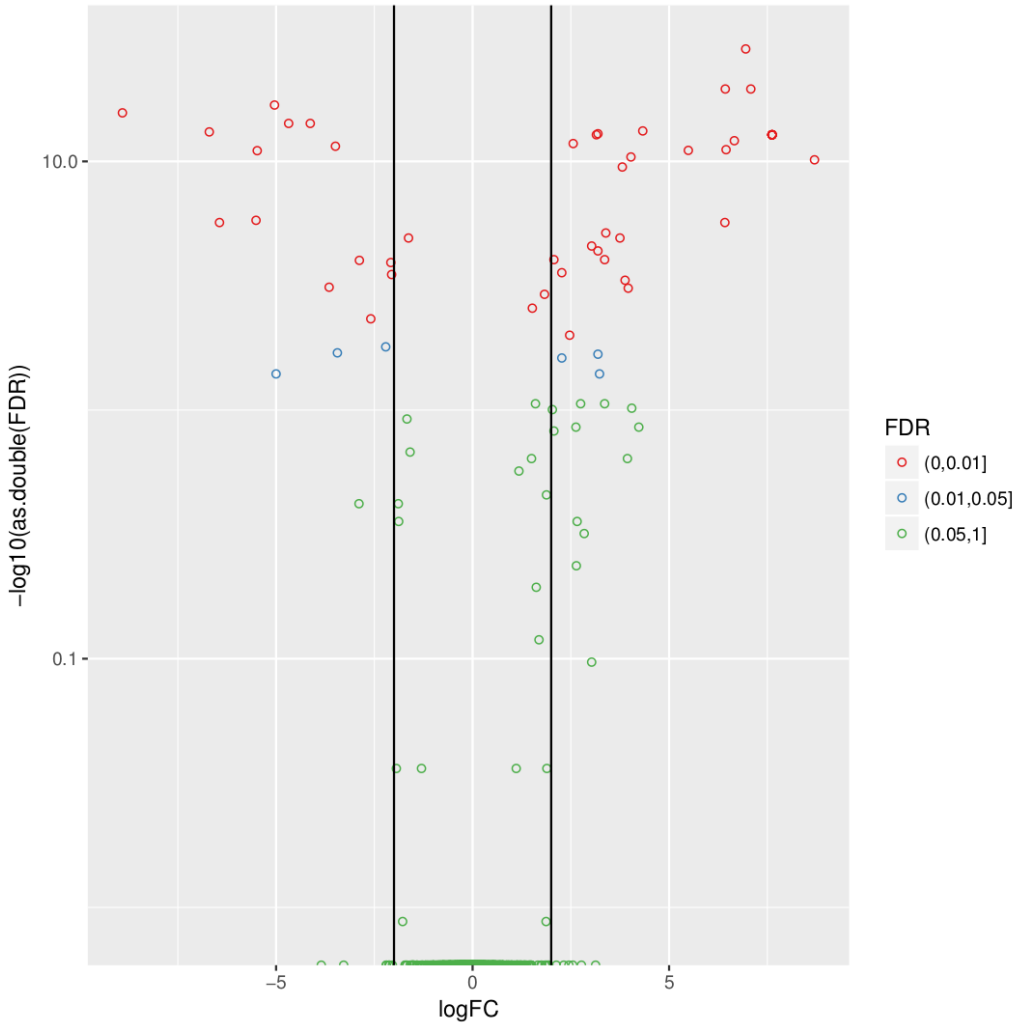
GO:0060333	interferon-gamma-mediated signaling pathway
GO:0031295	T cell costimulation
GO:0034341	response to interferon-gamma
GO:0019884	antigen processing and presentation of exogenous antigen
GO:0048002	antigen processing and presentation of peptide antigen
GO:0042605	peptide antigen binding
GO:0032395	MHC class II receptor activity

KEGG pathway enrychment for selected genes

Gene to KEGG test for over-representation

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05330	0.000	35.055	0	7	35	Allograft rejection
05332	0.000	32.707	0	7	37	Graft-versus-host disease
04940	0.000	28.839	0	7	41	Type I diabetes mellitus
04612	0.000	18.741	1	8	69	Antigen processing and presentation
05320	0.000	22.767	0	7	50	Autoimmune thyroid disease
05322	0.000	10.434	1	9	134	Systemic lupus erythematosus
05416	0.000	15.998	1	7	68	Viral myocarditis
04672	0.000	20.489	0	6	46	Intestinal immune network for IgA production
05310	0.000	29.090	0	5	28	Asthma
05150	0.000	17.416	0	6	53	Staphylococcus aureus infection
04514	0.000	9.193	1	8	131	Cell adhesion molecules (CAMs)
05140	0.000	12.752	1	6	70	Leishmaniasis
04145	0.000	7.937	1	8	150	Phagosome

Genes with different expression quantification in reference and diploid genomes across multiple tissues



logFC>=2, FDR<=0.05

Up-regulated in reference: 55 genes

- 30 pseudogene
- 12 small
- 9 protein_coding
- 4 lncRNA

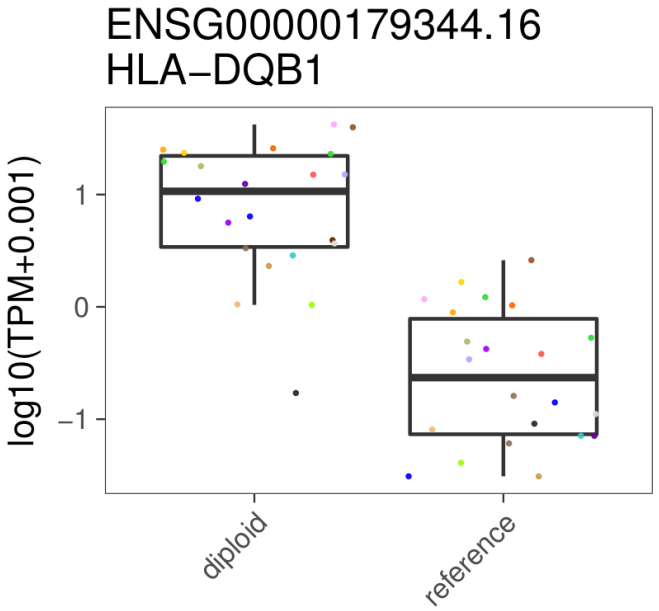
NBPF26
UGT2B15
HLA-DRB5
AC073333.1
RP11-514P8.6
FOXD4L3
NANOG
CTD-3126B10.5
RIMBP3

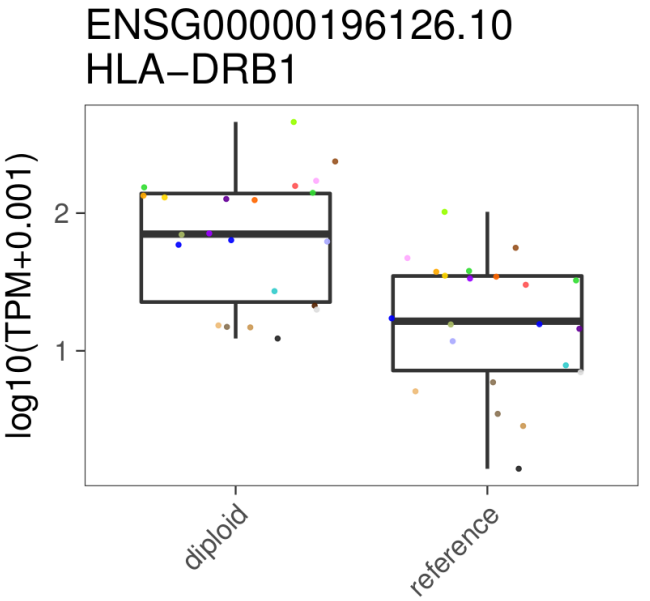
Up-regulated in diploid: 23 genes

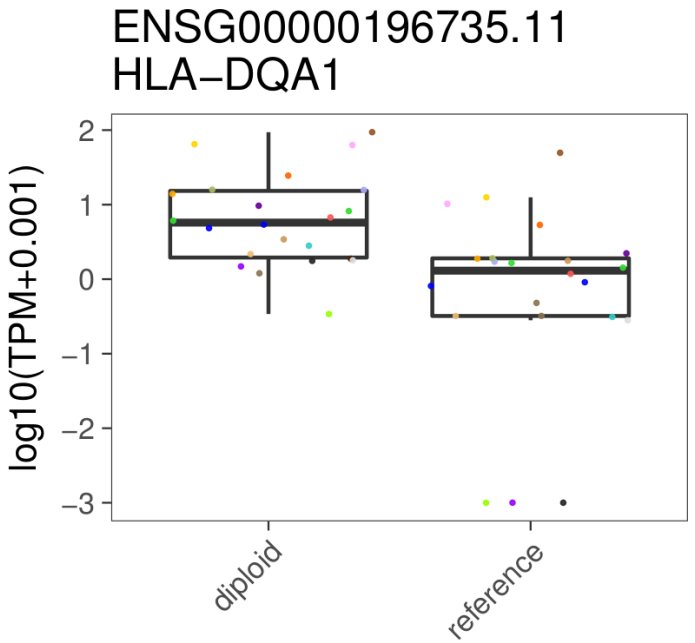
- 13 pseudogene
- 4 lncRNA
- 4 protein_coding
- 2 small

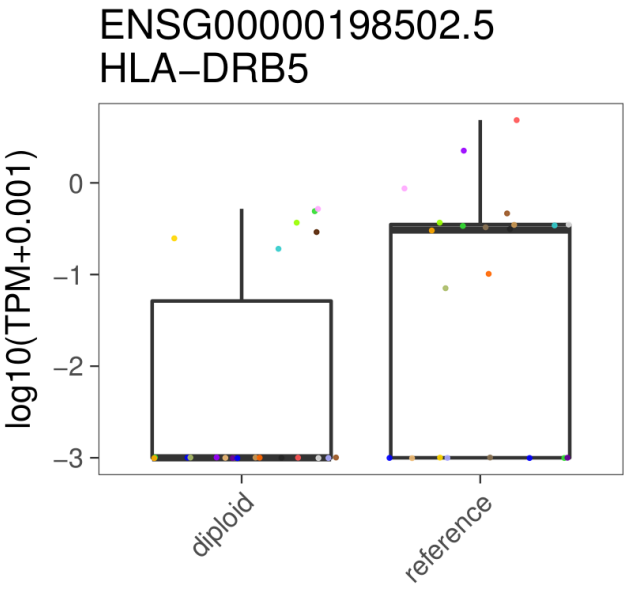
HLA-DRB1
HLA-DQA1
HLA-DQB1
IGHV4-31

Examples


ENSG00000179344.16
HLA−DQB1
log10(TPM+0.001)
1
0
−1
diploid
reference


ENSG00000196126.10
HLA−DRB1
log10(TPM+0.001)
2
1
diploid
reference

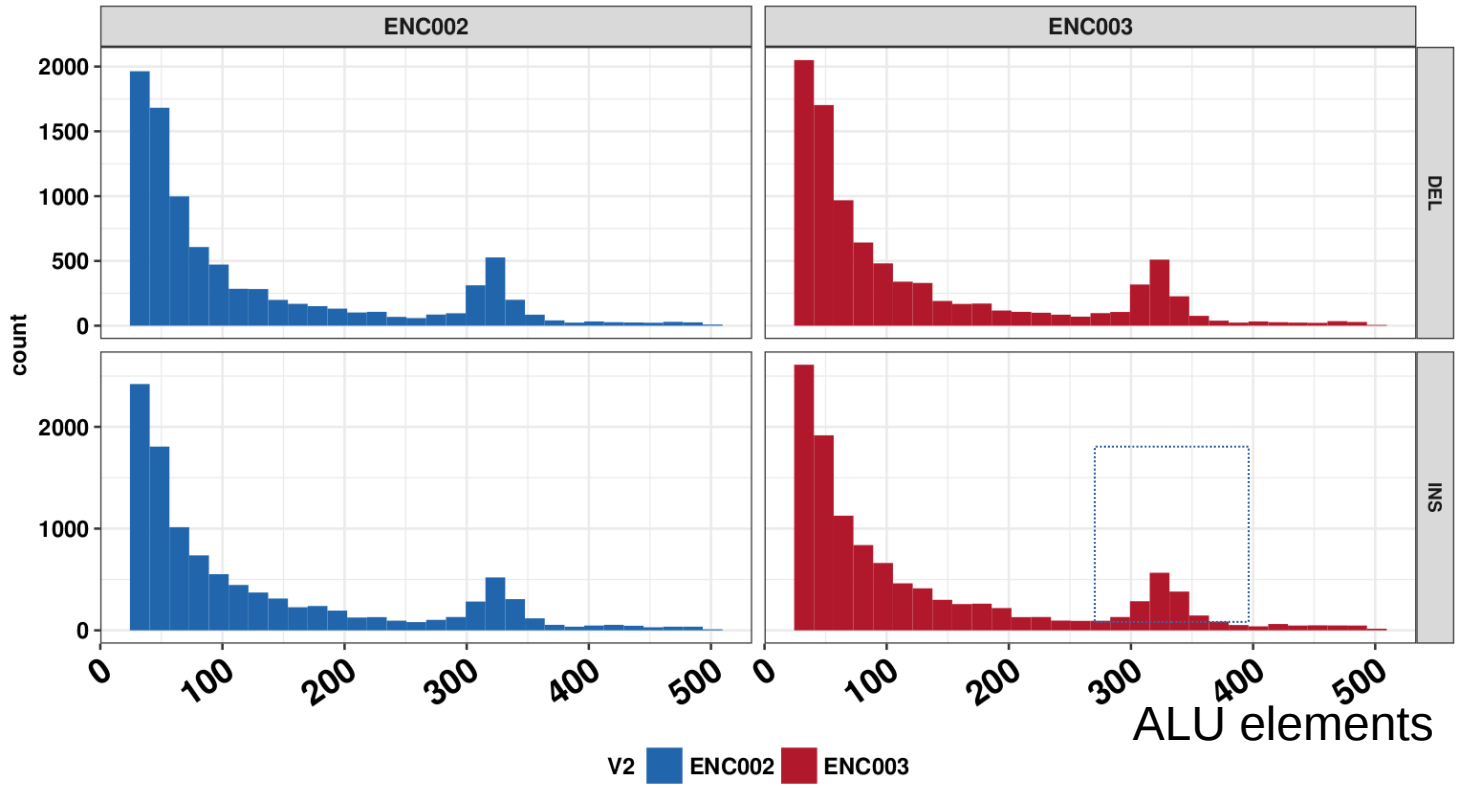

ENSG00000196735.11
HLA−DQA1
log10(TPM+0.001)
2
1
0
−1
−2
−3
diploid
reference


ENSG00000198502.5
HLA−DRB5
log10(TPM+0.001)
0
−1
−2
−3
diploid
reference

Effect of deletions(DEL) to the gene expression

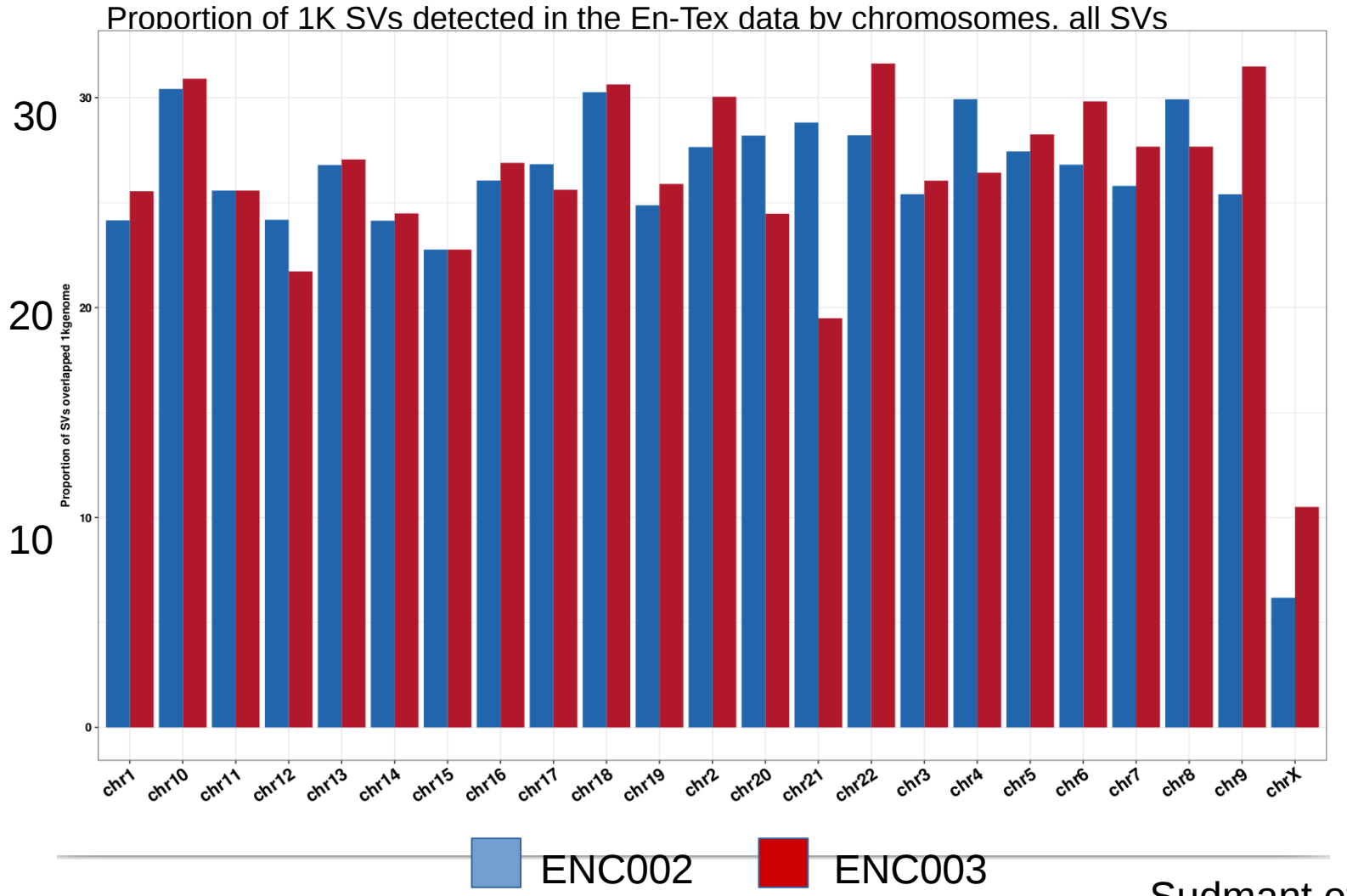
Structural variants in ENC002 and ENC003

	ENC002				ENC003			
	#SVs	Size			#SVs	Size		
		min	max	median		min	max	median
DEL	10,018	31	19,118	82	10,399	31	19,916	85
INS	11,556	31	7,372	84	12,866	31	7,709	88
INV	98	51	19,629	1041	111	75	19,628	622
Total	21,672				23,376			



SVs in EN-TEx and 1000 genome(1K) project

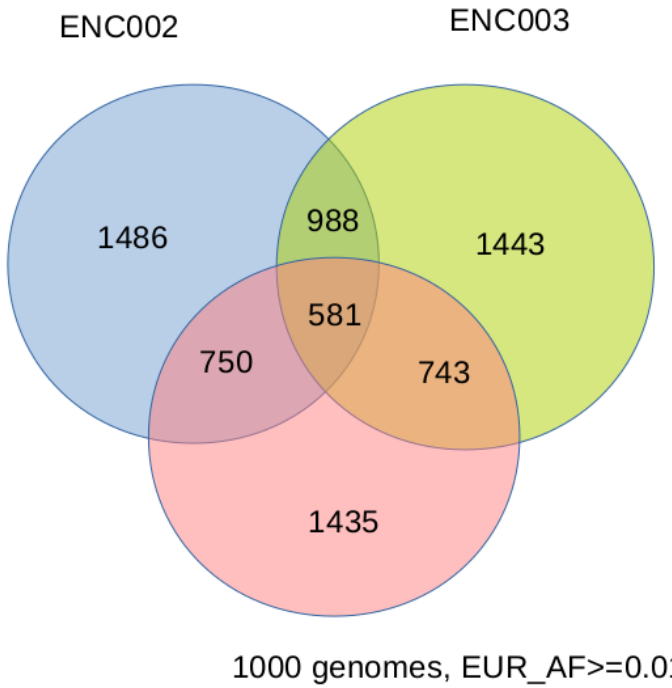
- 1K SVs in hg19 coordinates → lifted over into GRCh38 coordinates
- EUR allele frequency >=0.01, in total there are 10,014 SV in 1K data set
- Used all SV categories from 1K: DEL, ALU, DUP, INS, INV, …
- Partial overlap statistic - 1bp was sufficient to detect an overlap
- ENC002 =27.6% , ENC003 =27.7%

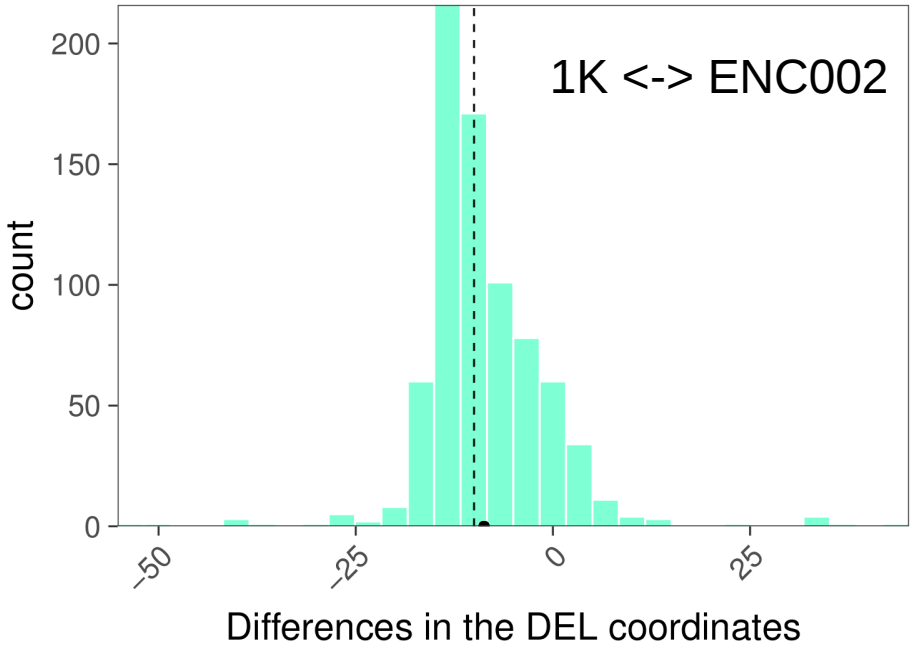


Sudmant et al.,Nature 2015

SVs in EN-TEx and 1000 genome(1K) project

Deletions with size 200-500nt, ALU elements





Structural variants that <u>have no overlap</u> to the SVs from 1K genome:

Number of SVs per individual

ENC002 = 14,213
ENC003 = 15,353

Number of SVs that are **unique** in each individual

	ENC002	ENC003
DEL	2,467	2,670
INS	5,812	6,731
INV	14	25

Genomic elements affects by SVs

Calculated partial (1bp) and complete overlap – element is fully imbedded in SV interval

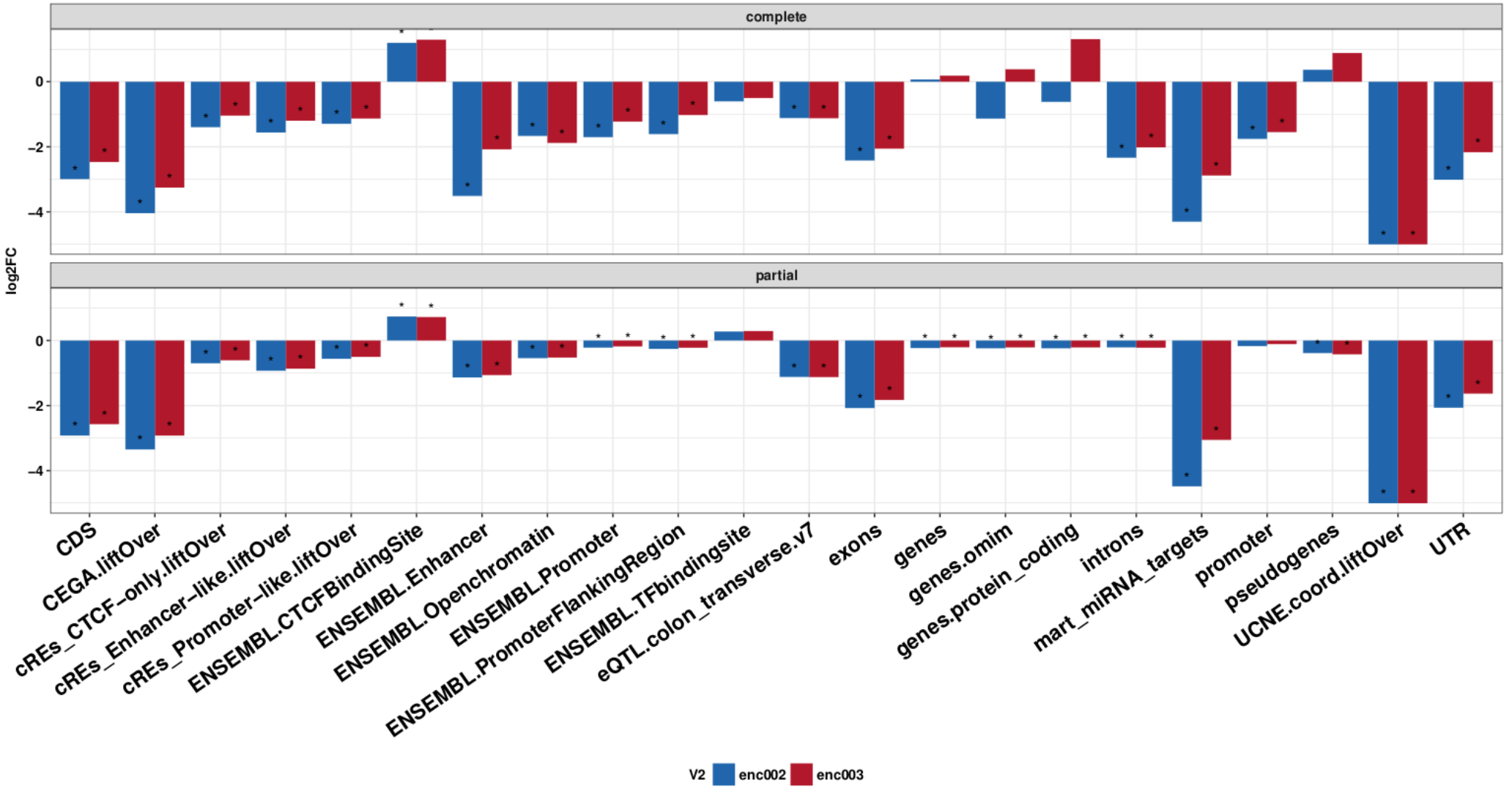
Regions used to calculate an overlap:

- Gencode24 genes, exons, introns, CDs, UTRs
- Protein coding genes
- Pseudo-genes
- Genes present in OMIM database
- Ultra-conserved non-coding elements (UCNEs) from http://ccg.vital-it.ch/UCNEbase/ *
- CEGA, Conserved Elements from Genomic Alignments, from http://cega.ezlab.org/ *
- Promoter regions (+/-500bp from the annotated TSS, gencode24)
- ENSEMBL regulatory elements (GRCh38, via bioMart)
  - TF binding sites; CTCF Binding Site; Enhancer; Open chromatin; Promoter; Promoter Flanking Region; miRNA target sites
- ENCODE cREs elements *
  - CTCF-only; Enhancer-like; Promoter-like

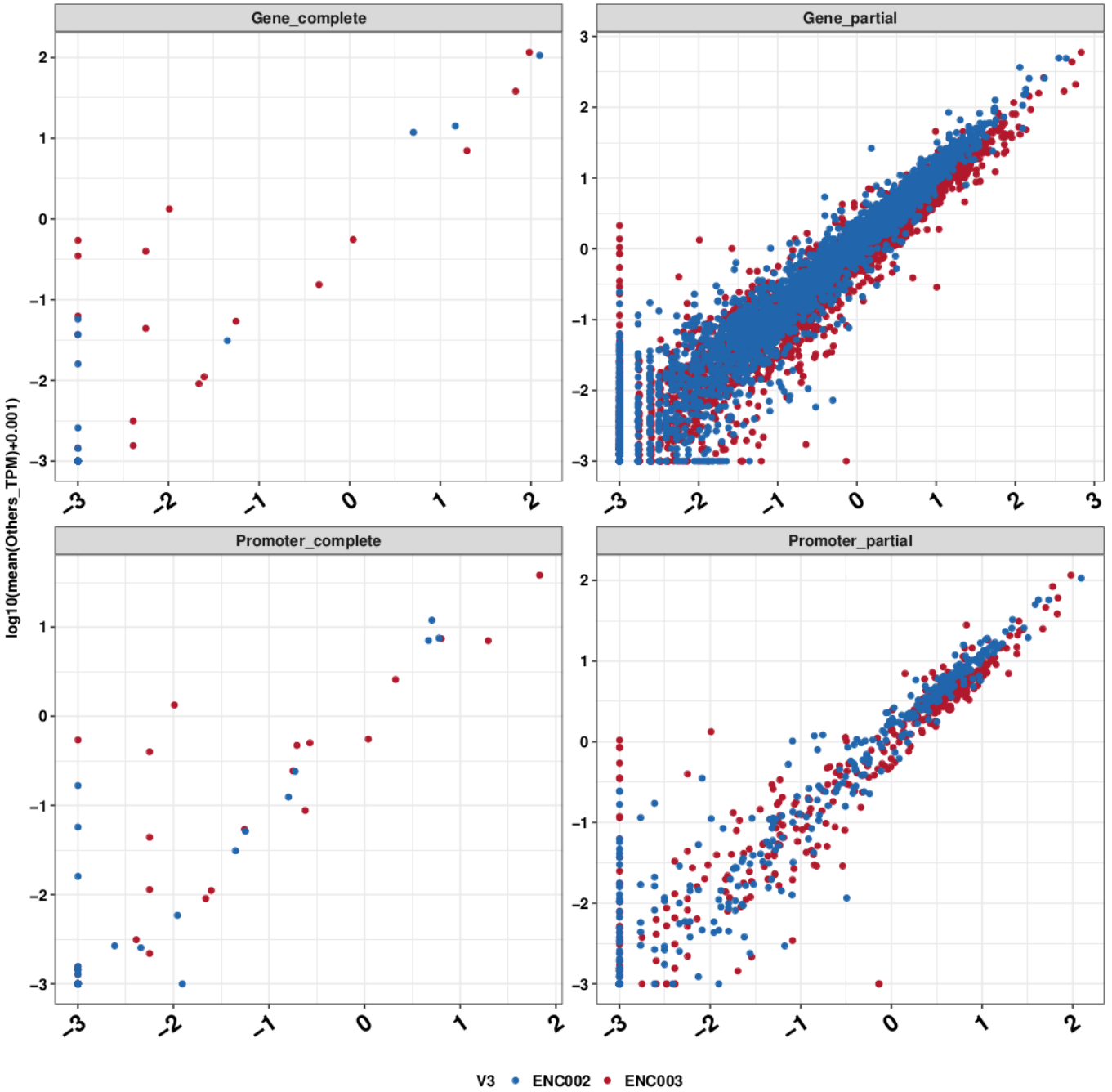
* These elements were originally in hg19 coordinates -> lifted over to GRCh38 coordinates

Genomic elements affects by SVs

Enrichment of functional elements intersecting SVs



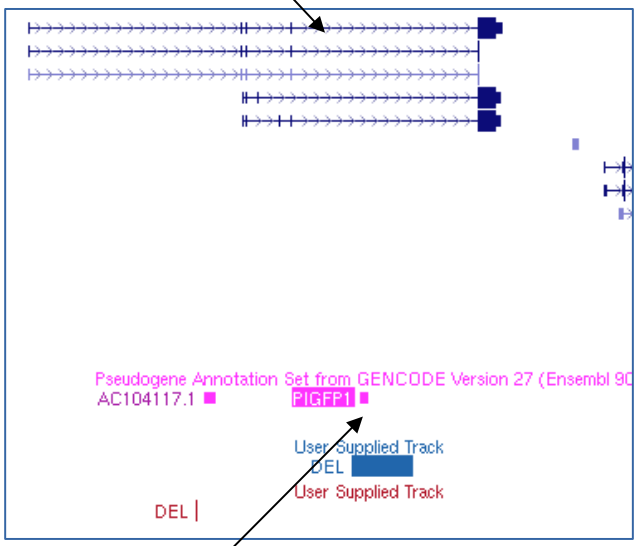
Changes in expression for genes in deletions



log10(Selected individual_TPM+0.001)

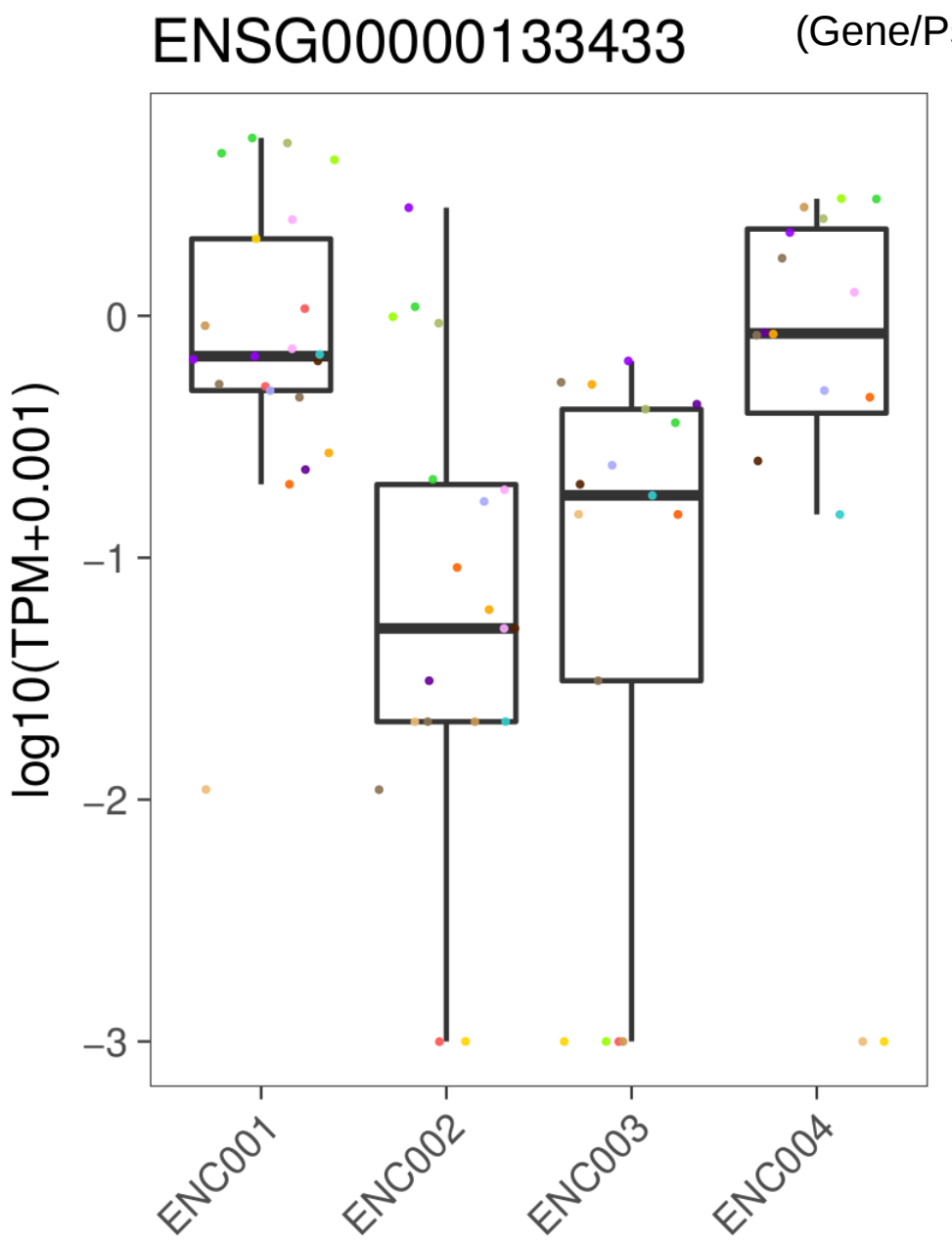
		gene	promoter
ENC002	partial	3,896	533
	complete	44	39
ENC003	partial	4,100	548
	complete	43	36

Partial overlap



Complete overlap

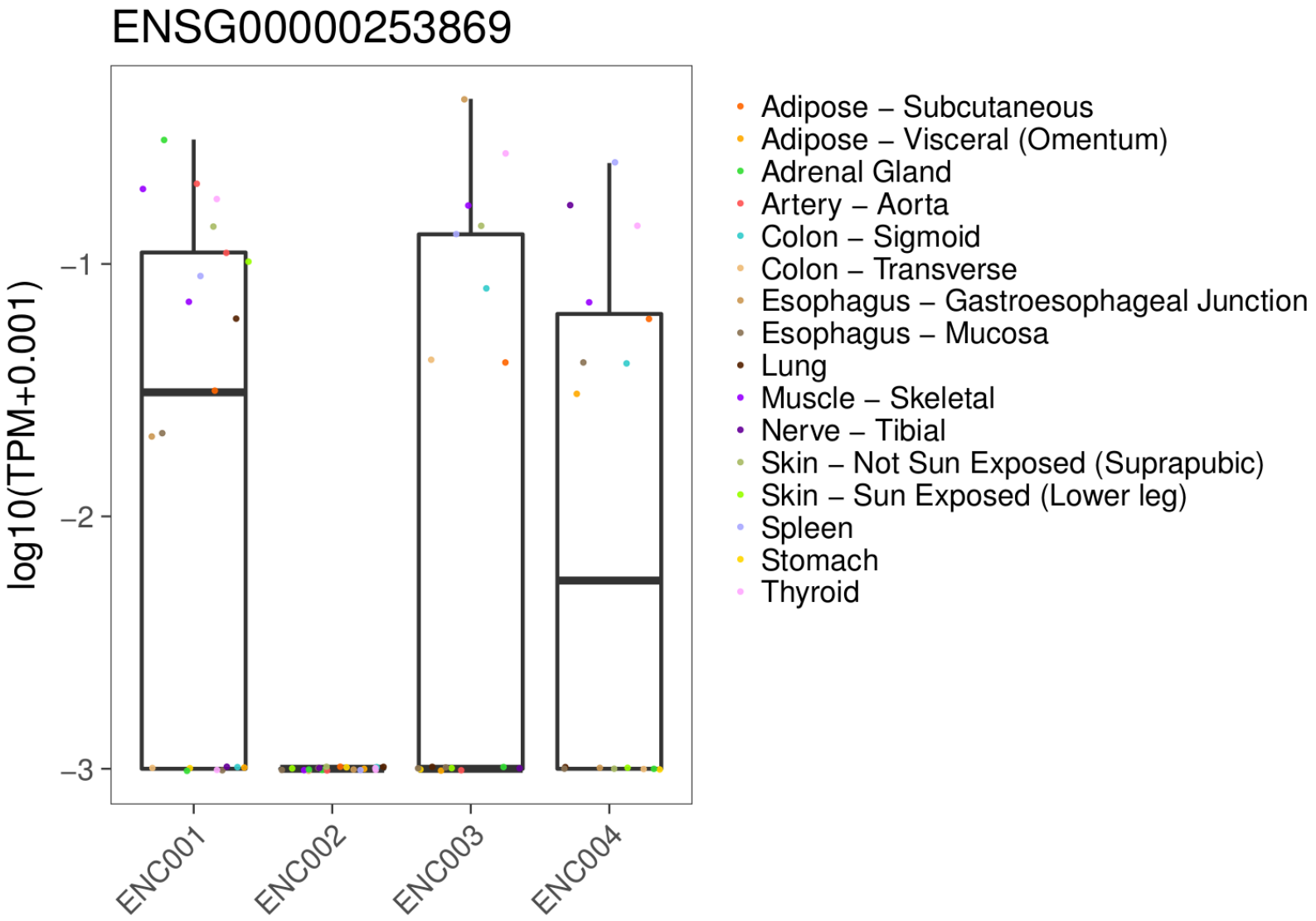
Glutathione S-Transferase Theta 2B
(Gene/Pseudogene)GSTT2B



- Adipose – Subcutaneous
- Adipose – Visceral (Omentum)
- Adrenal Gland
- Artery – Aorta
- Colon – Sigmoid
- Colon – Transverse
- Esophagus – Gastroesophageal Junction
- Esophagus – Mucosa
- Lung
- Muscle – Skeletal
- Nerve – Tibial
- Skin – Not Sun Exposed (Suprapubic)
- Skin – Sun Exposed (Lower leg)
- Spleen
- Stomach
- Thyroid

Complete overlap in both individuals, ENC002 and ENC003

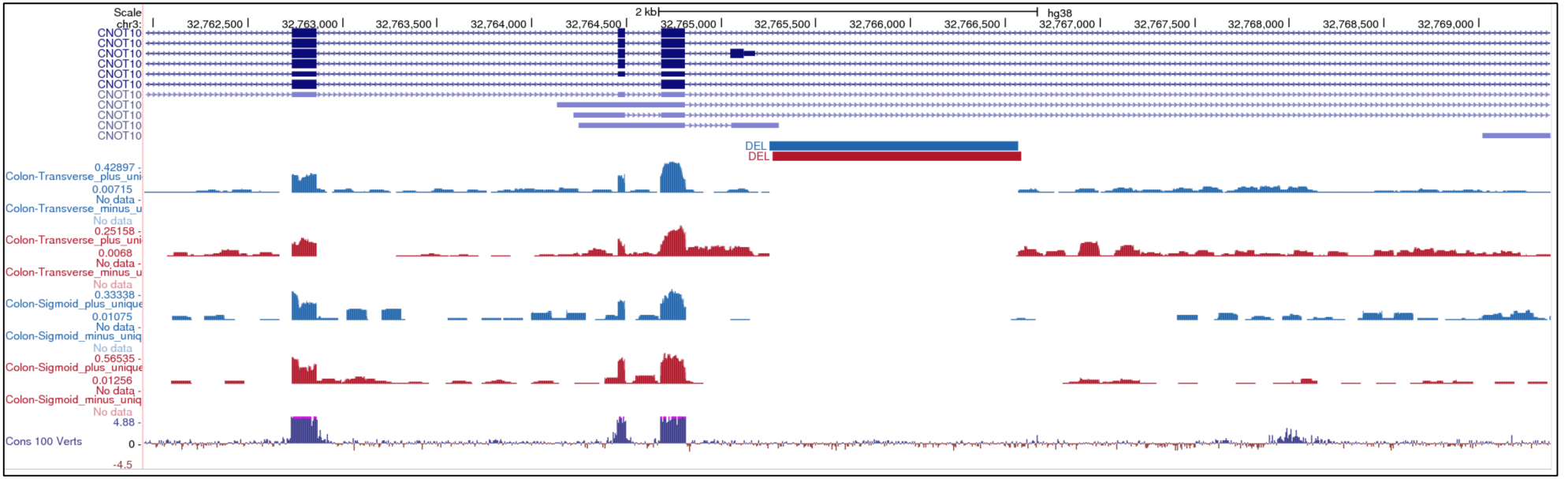
This deletions is eQTL in GTEx dataset

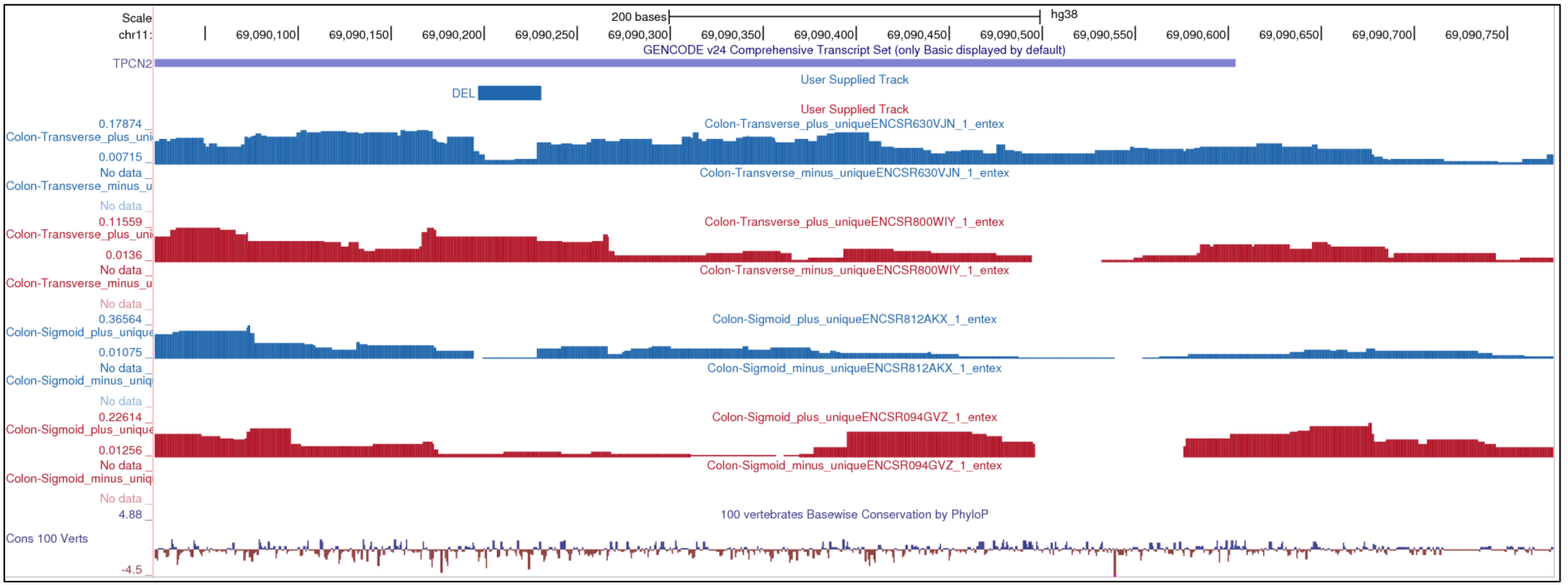


Homozygous deletion in ENC002

Pseudogene, PIGFP1, Phosphatidylinositol Glycan Anchor Biosynthesis Class F

In the reference-based mappings we observe novel splice junction, while it is a deletion in the genome.

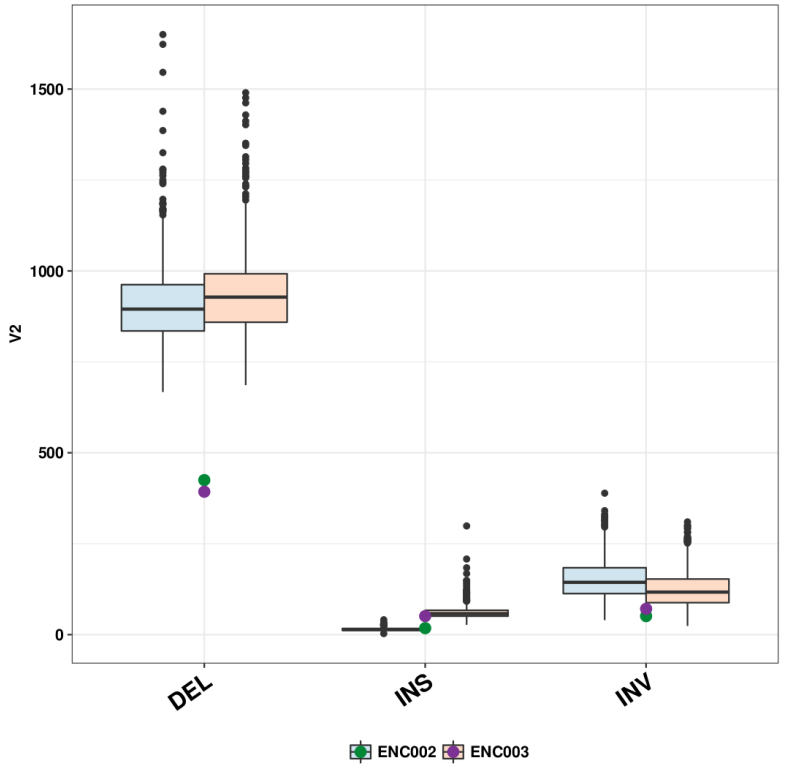


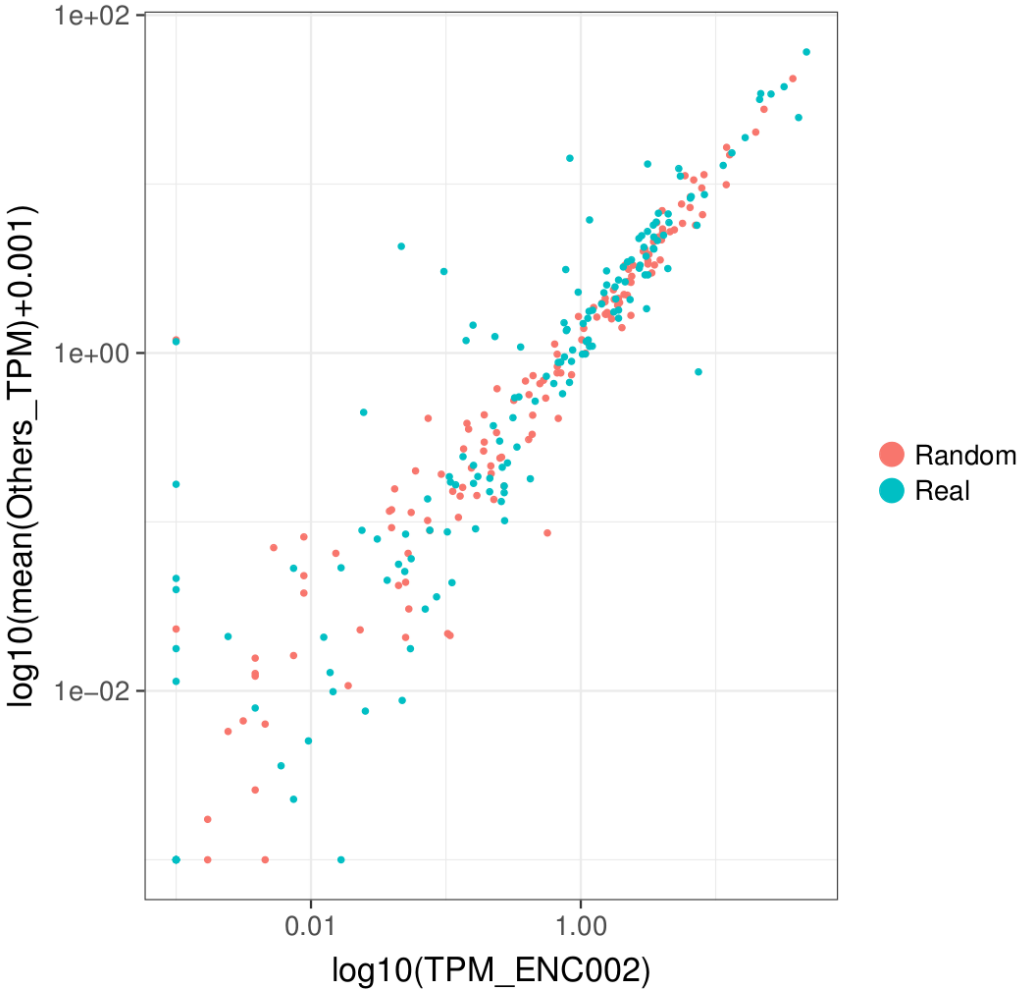


Structural variations overlapping eQTLs

Colon-transverse eQTLs, GTEx v7 lifted over to GRCh38: 829,332 (SNPs=524,031, genes=8,090)

Number of eQTL SNPs overlapping SV

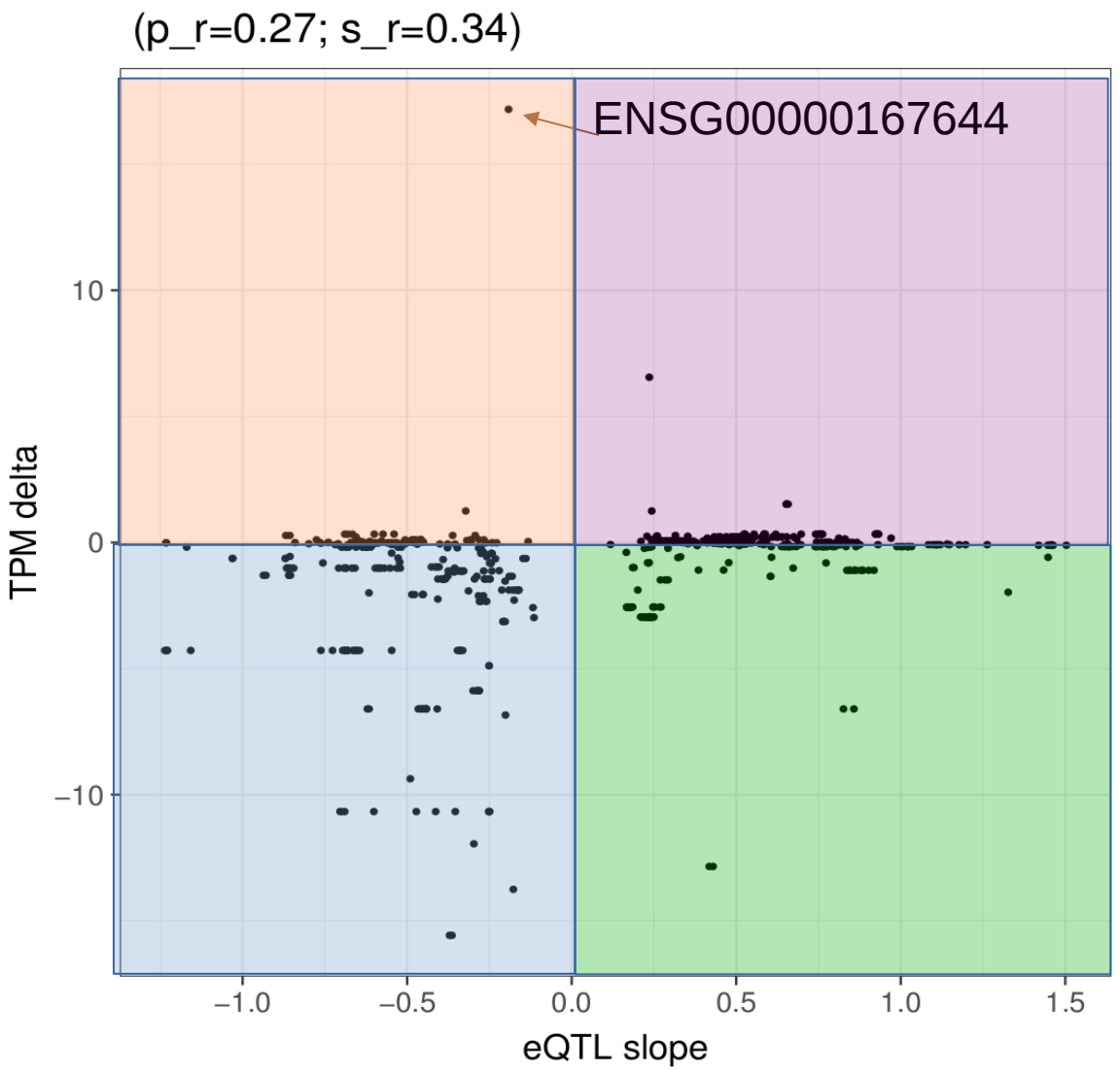




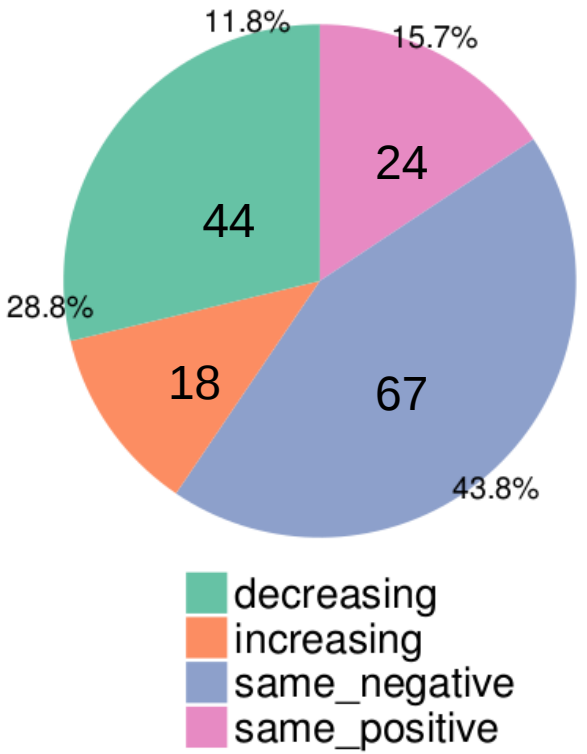
Boxplots – random overlap
Solid dots – real overlap

ENC002 DEL are overlapping
423 SNPs that regulate 160 genes

Changes in target gene expression caused by deletions eliminating eQTLs



Changes in the gene expression

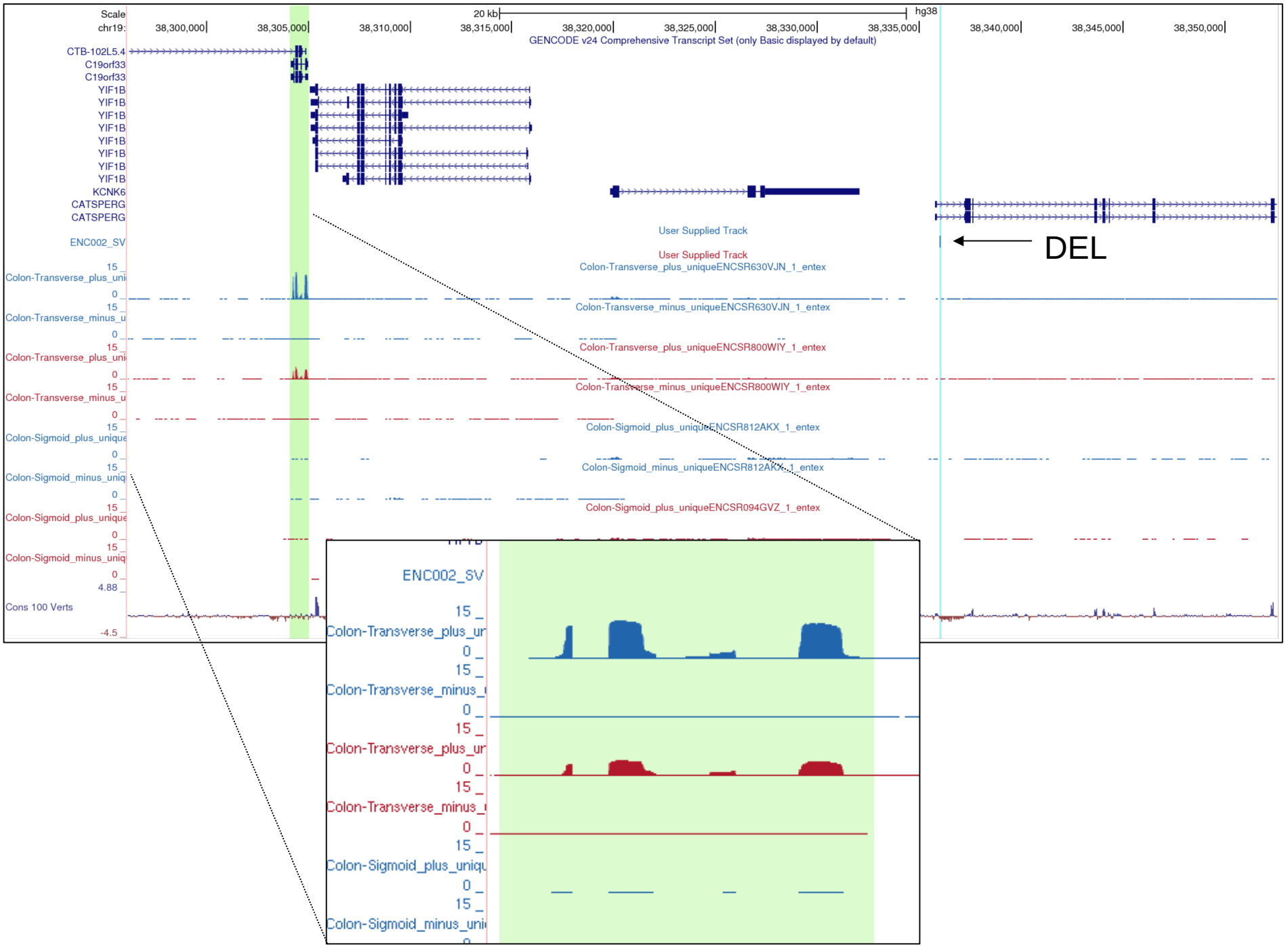


eQTL slope – slope coefficient in a linear model predicting significance of eQTL

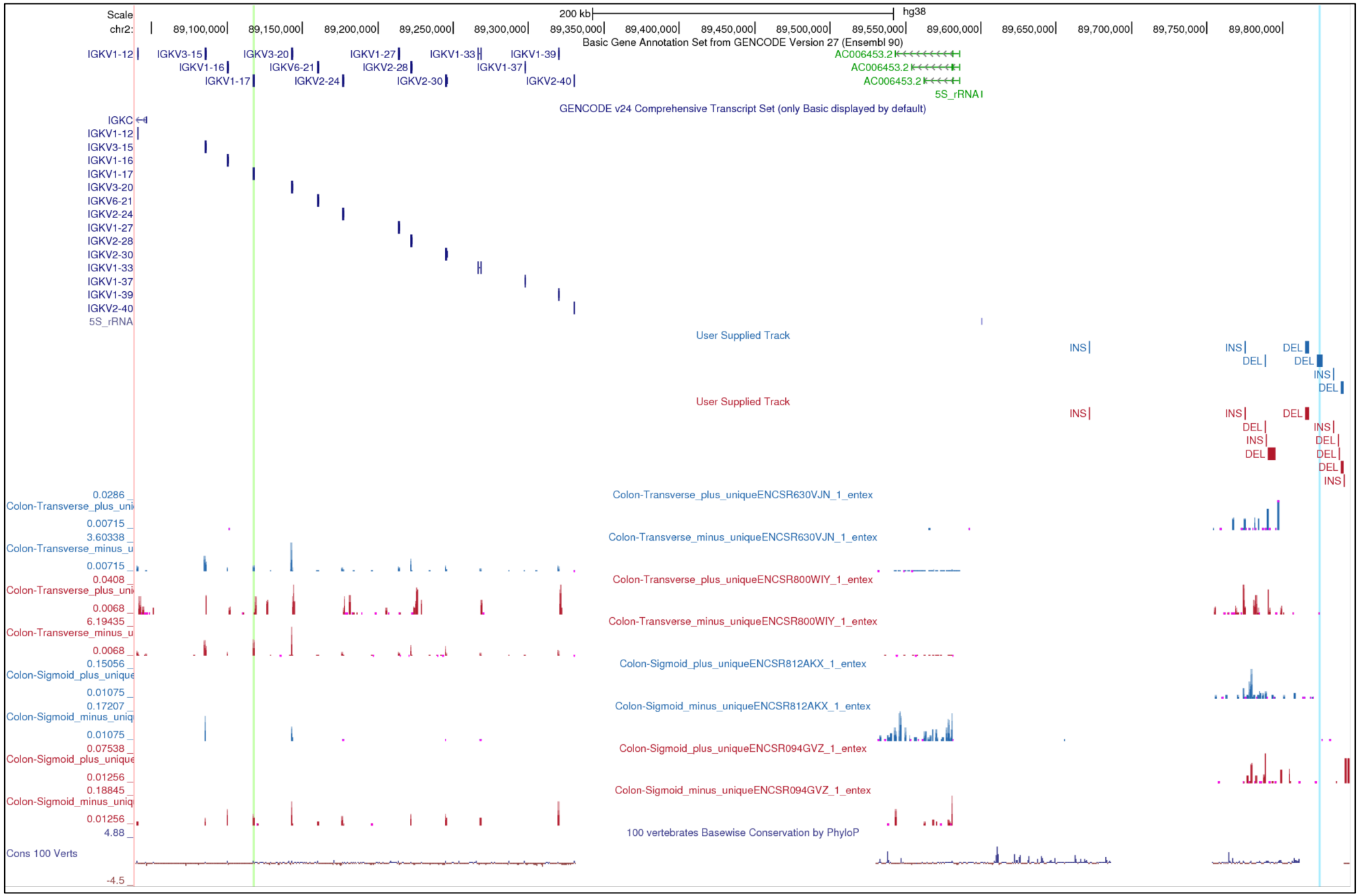
Note: all this genes have other eQTL SNPs, up to 3,000 per gene

TPM delta= TPM(ENC002) – mean(TPM rest)

ENSG00000167644, C19orf33, Hepatocyte Growth Factor Activator Inhibitor Type 2-Related Small Protein

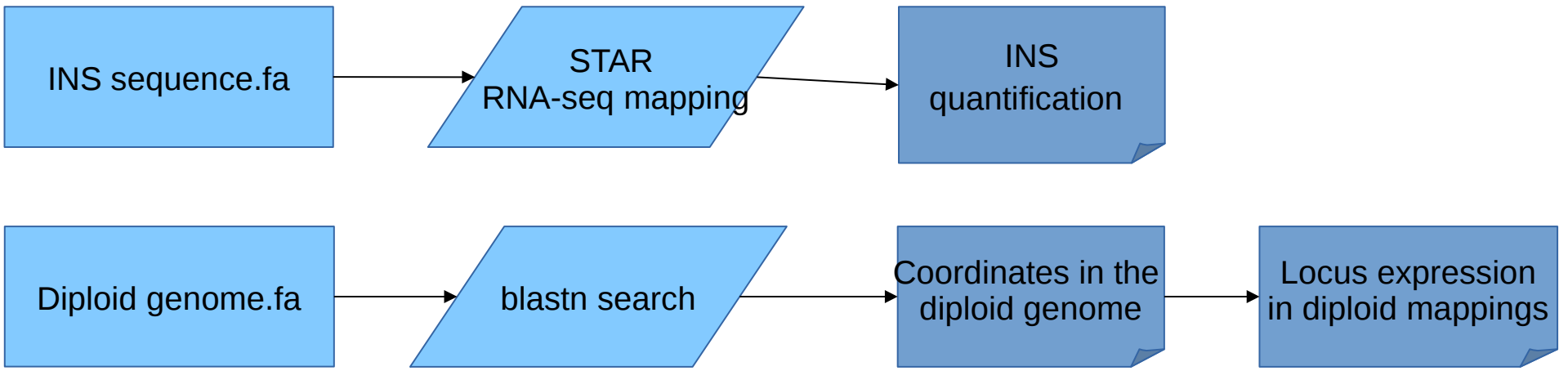


ENSG00000241755, IGKV1-9



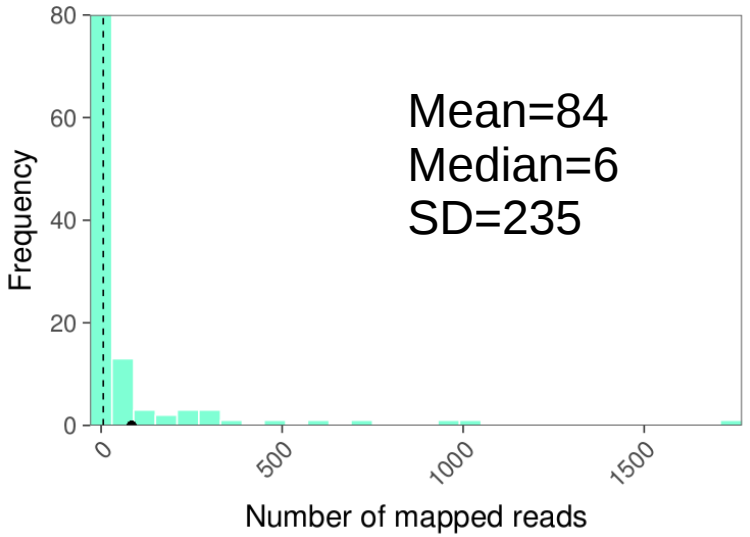
Novel transcriptional elements in insertions (INS)

Putative functional effect of insertions



ENC002 - 332 INS with mapped reads, majority of the alignments are with mismatches

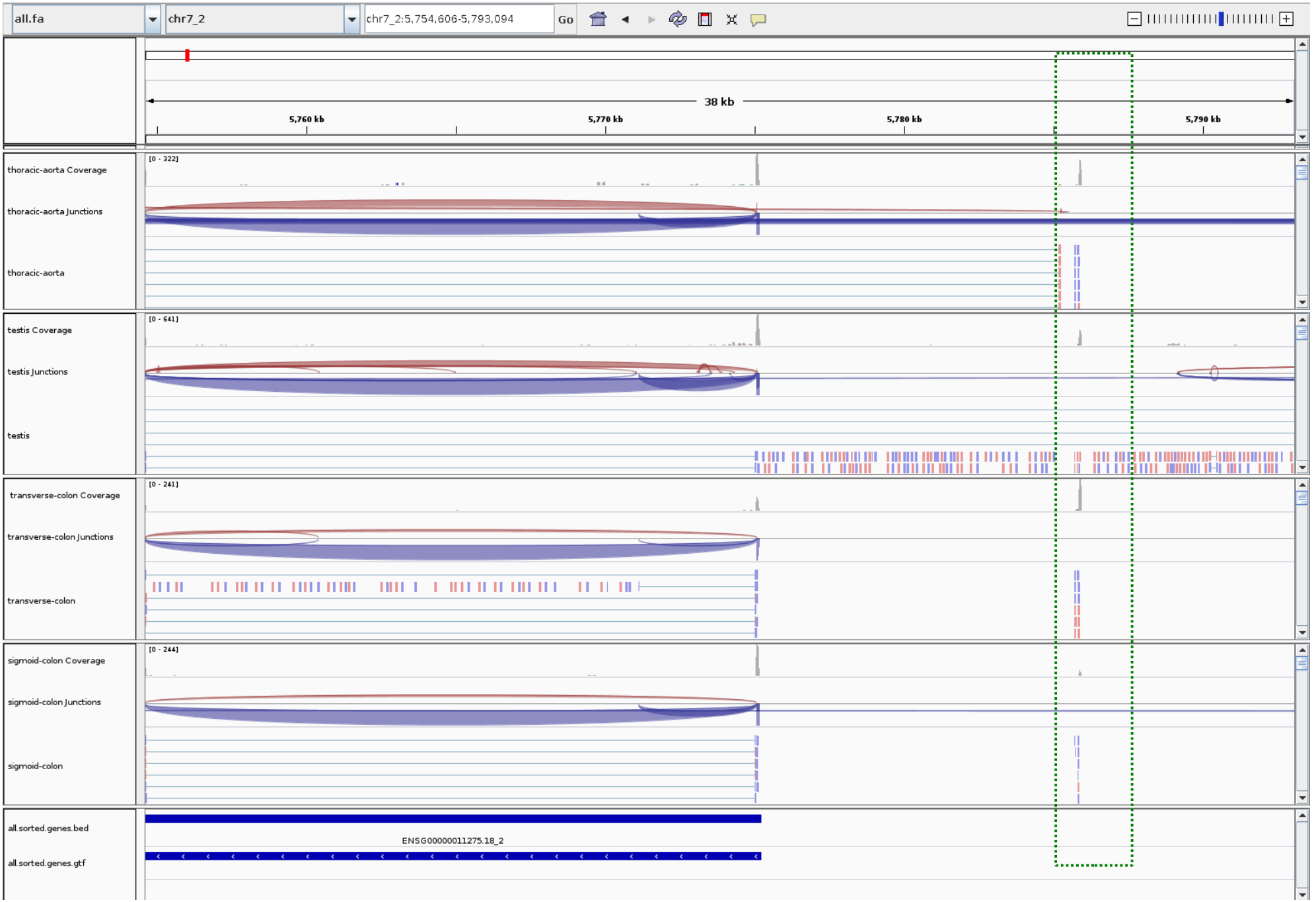
Number of reads per INS



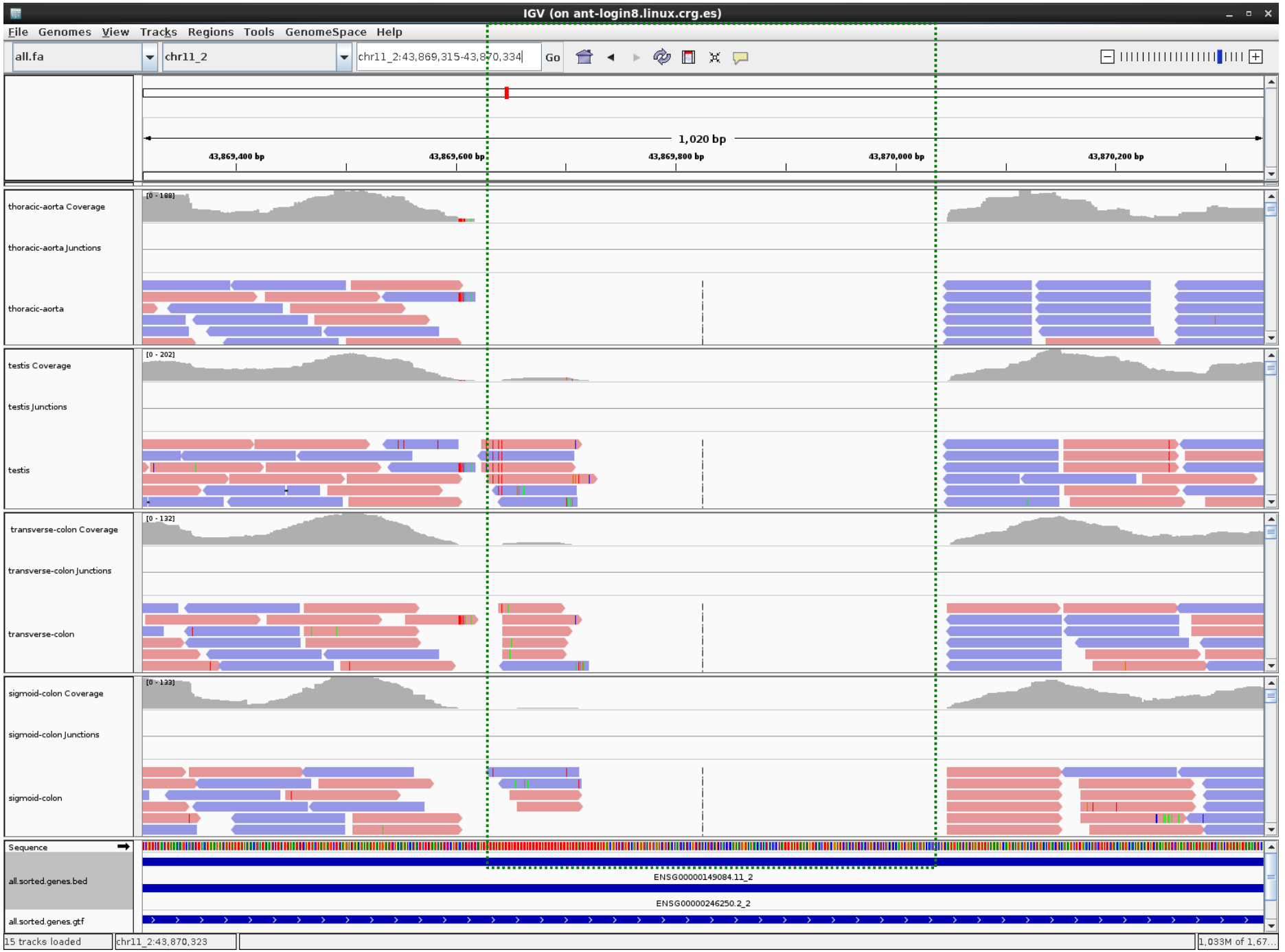
There are 60 insertions that have ≥50 reads mapped

Some of this INS are transposone elements

Diploid genome, insertion chr7_2:5785322-5786144



Diploid genome, insertion chr11_2:43869603-43870046, overlapping exon



Putative alu element insertion

Work in progress ...

- Construct personal genome and personal annotation for all individuals
- Expression changes due to SVs overlapping functional elements, i.e. enhancers, eQTLs
- SNPs and short indel analysis
- Novel transcription elements in insertions
- Chimeric transcripts in reference and personal genomes
- Allele specific expression, with Gerstein group
- Integrate other functional assays to perform tissue specific analysis, i.e. smallRNAs, RAMPAGE, ChiP-seq

Acknowledgments

Roderic Guigo Lab
Beatrice Borsari
Julien Lagarde
Alessandra Breschi

Michael Schatz Lab

Thomas Gingeras Lab
Alex Dobin
Dinar Yunusov

Mark Gerstein Lab

+ all ENCODE Partners