Tissue-level classification of loci associated with type 2 diabetes

Jason Torres, PhD

McCarthy Lab Meeting

March 16, 2018

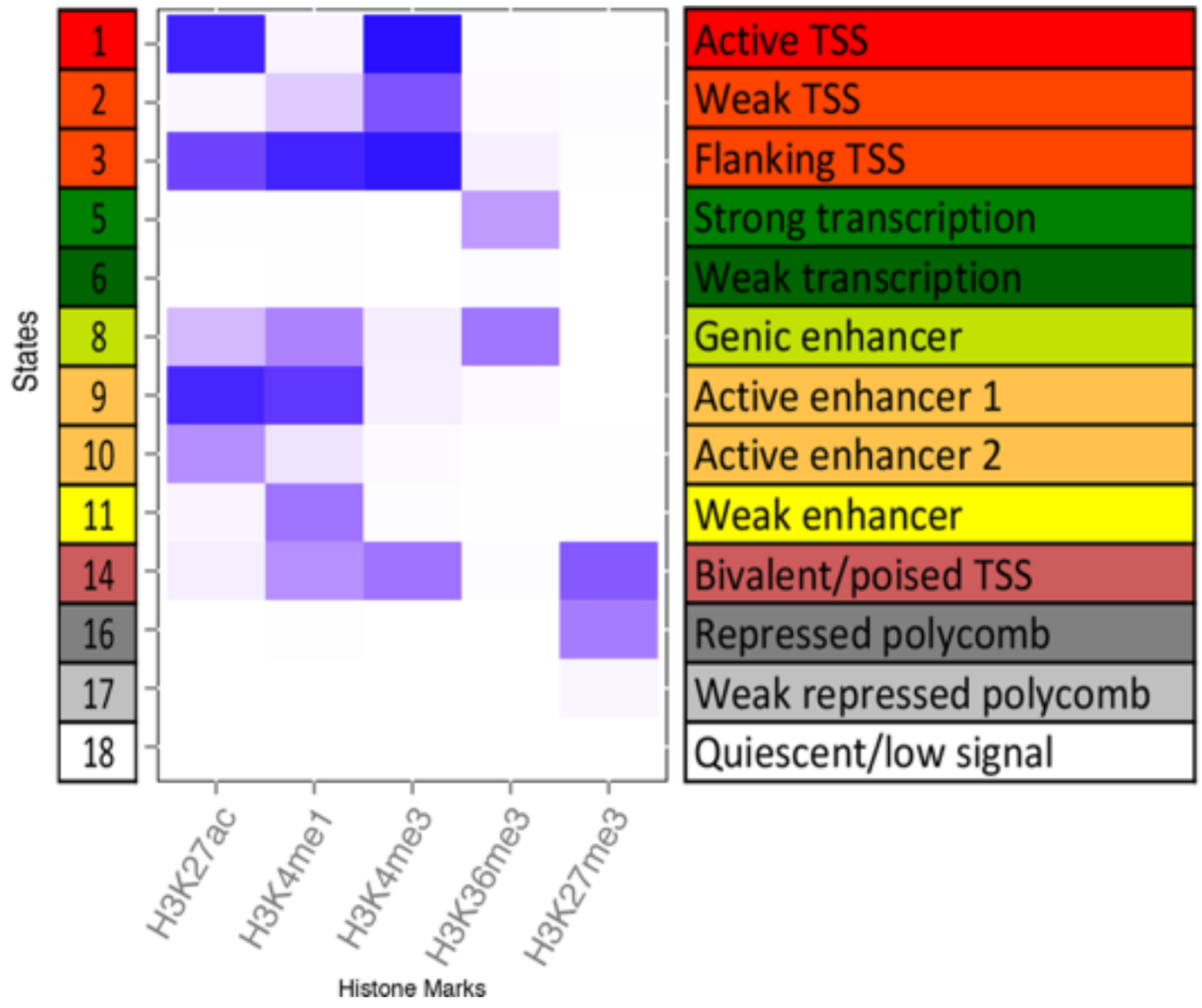
Goal: Develop an informative score for classifying T2D loci by tissue

- **Approach**: Use fgwas to fine-map "functional" credible sets with chromatin segmentation data from islet, adipose, muscle, and liver

- **Data**:
  - DIAMANTE-European GWAS (HRC-imputed)
  - Chromatin states (ChIP-seq) from Varshney et al. 2016.

- Pipeline for generating functional credible sets for all conditionally independent loci

ChromHMM Emission plot
Varshney *et* al. 2016

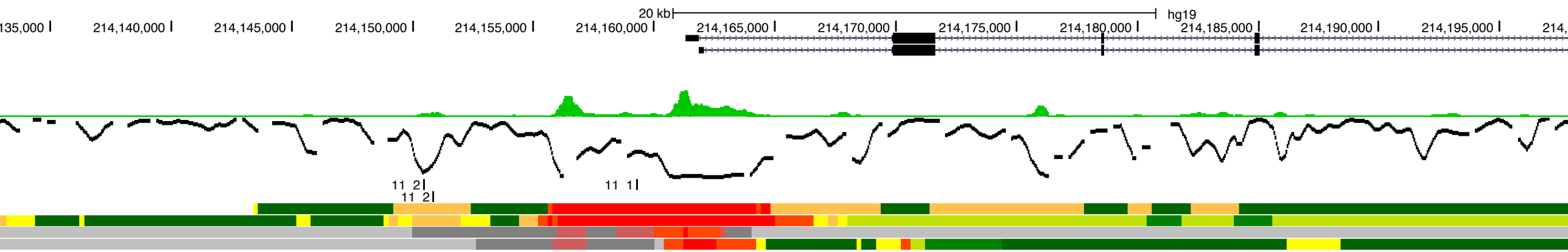


Scheme for partitioning GWAS data

- Estimates for genomic annotations that encompass a small number of SNPs fail to converge

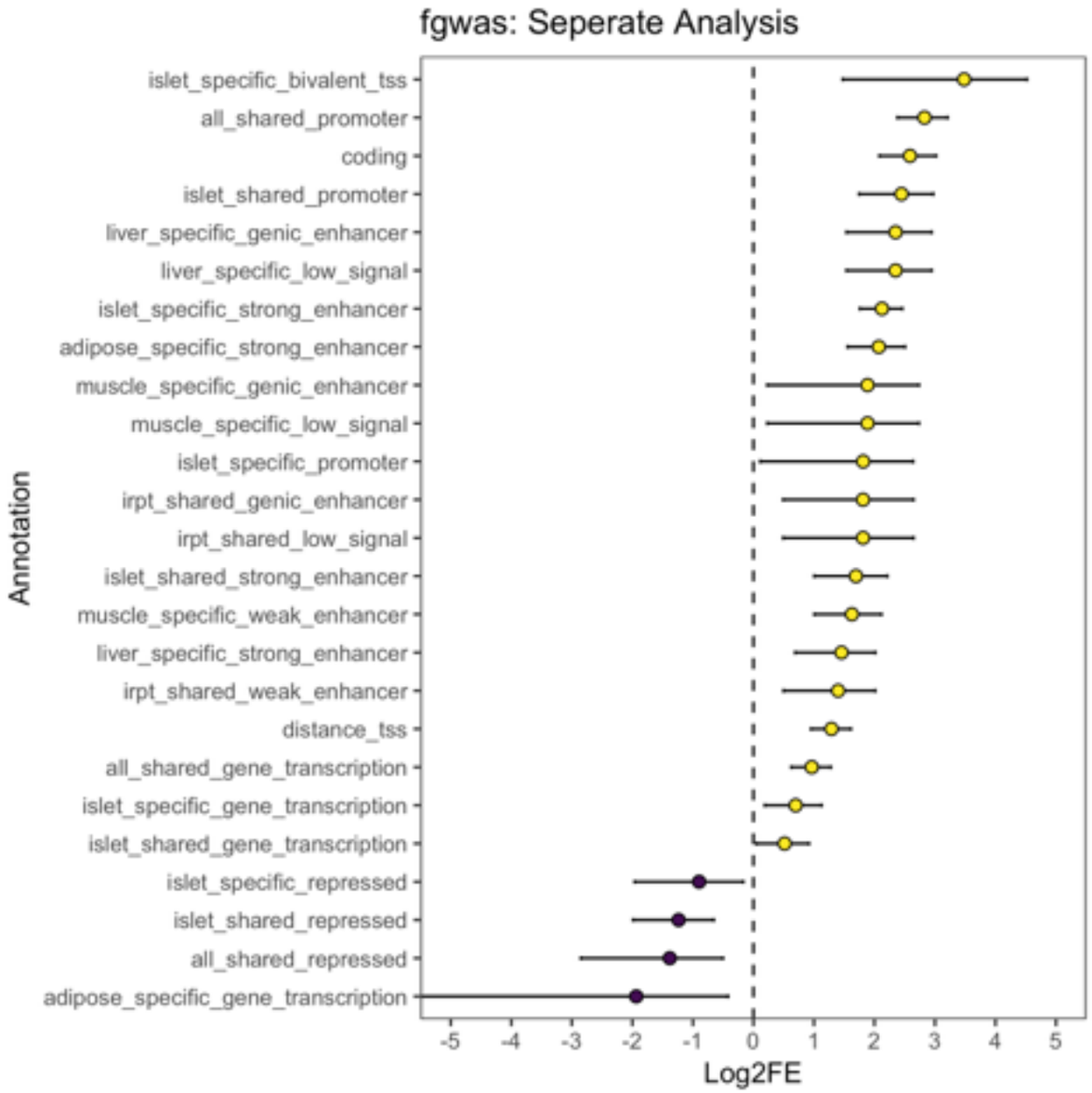
- Overlapping annotations may be difficult to interpret (i.e. sign flips) in joint model

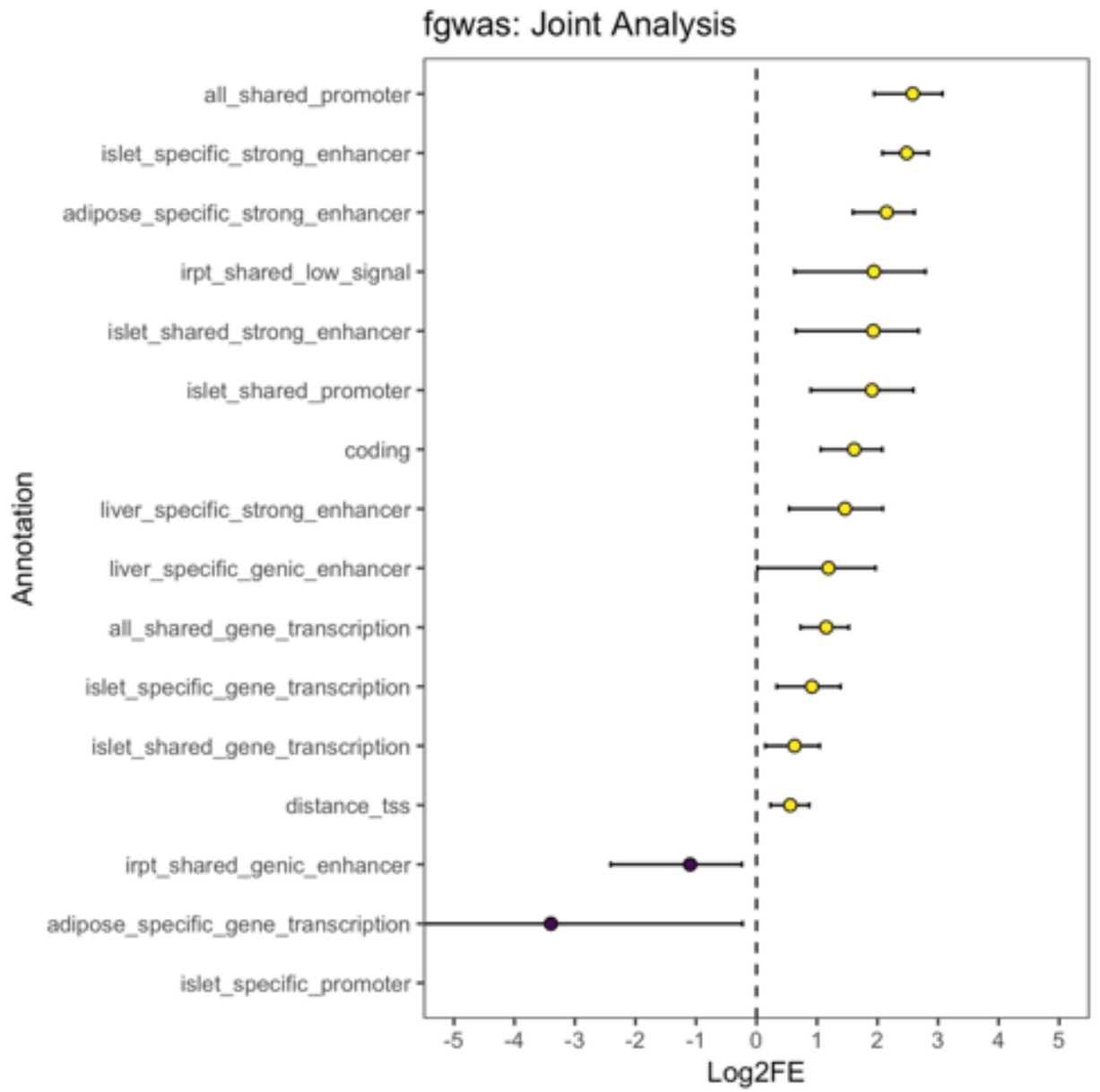
- **Strategy**: Partition the genome into subsets that are sensibly grouped



Tissue-specific, shared with islet, not shared with islet, coding, distance to TSS

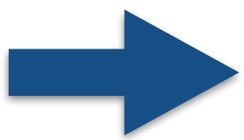
Tissue-specific and shared annotations show genome-wide enrichment





Scheme for obtaining tissue scores from “functional” fine mapping

Locus
credible set (*m* SNPs)

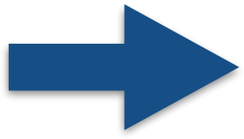


Tissue scores
(islet, adipose, muscle, liver, other)

SNP$_{m0}$
(fgwas PPA)

⋮

SNP$_{m}$
(fgwas PPA)



Divvy by genomic annotation
Coding
Strong enhancer
Weak enhancer
Promoter
Gene transcription
Genic enhancer

Expression specificity scores for divvying coding variant PPA

$$ESS_t = \frac{\text{median}(Expression_t)}{\sum_{x \in T} \text{median}(Expression_x)}$$

	GeneID	GeneName	islet.score	muscle.score	adipose.score	liver.score
1	ENSG00000181092.5	ADIPOQ	0.000000E+00	1.204194E-03	**9.987158E-01**	8.004492E-05
2	ENSG00000254647.2	INS	**9.999898E-01**	1.428814E-06	2.408706E-06	6.330036E-06
3	ENSG00000187758.3	ADH1A	8.960711E-04	4.900985E-04	1.228740E-02	**9.863264E-01**
4	ENSG00000125414.14	MYH2	1.559308E-05	**9.99565E-01**	3.473728E-04	7.200330E-05

General approach for divvying PPA value for each SNP in a credible set

$$s_{j,a} = \begin{bmatrix} s_{j,a,t_1} \\ s_{j,a,t_2} \\ s_{j,a,t_3} \\ s_{j,a,t_4} \end{bmatrix}$$

$$s_{j,a,t} = \frac{P_j}{\sum_{i \in T} \mathbb{1}(j,a,i)} \mathbb{1}(j,a,t)$$

$$\mathbb{1}(j,a,t) := \begin{cases} 1 & \text{if SNP}j \text{ overlaps annotation } a \text{ in tissue } t \\ 0 & \text{otherwise} \end{cases}$$

where $s_{j,a}$ is score vector for SNP j and annotation a

$$s_j = P_j \frac{q}{\sum_{i=1}^{4} q_i}$$

s_j is score vector for SNP j where $q = \sum_{a \in A} s_{j,a}$ and A is the set of all annotations

Example: *HNF4A*

Fine-mapped to one coding variant (PPA=1) that maps to islet strong enhancer and liver genic enhancer

	islet	muscle	adipose	liver
coding.vec	**0.188141**	9.012258E-05	0.0001536971	**0.8116152**
strongenh.vec	**1**	0.000000E+00	0.0000000000	0.0000000
weakenh.vec	0.000000	0.000000E+00	0.0000000000	0.0000000
genenh.vec	0.000000	0.000000E+00	0.0000000000	**1**
prom.vec	0.000000	0.000000E+00	0.0000000000	0.0000000
genetrans.vec	0.000000	0.000000E+00	0.0000000000	0.0000000

matrix of score values across tissues and annotations

sum for each tissue across annotations

	islet	muscle	adipose	liver
1	1.188141	9.012258E-05	0.0001536971	1.811615

scale and multiply by PPA

	islet	muscle	adipose	liver
1	0.396047	3.004086E-05	5.123236E-05	0.6038717

Approach for determining tissue score vector for each credible set (i.e. locus)

$$L = \begin{bmatrix} s_1 \\ s_2 \\ . \\ . \\ . \\ s_m \end{bmatrix}$$

L is a matrix where each row corresponds to the score vector for each of m SNPs in the credible set

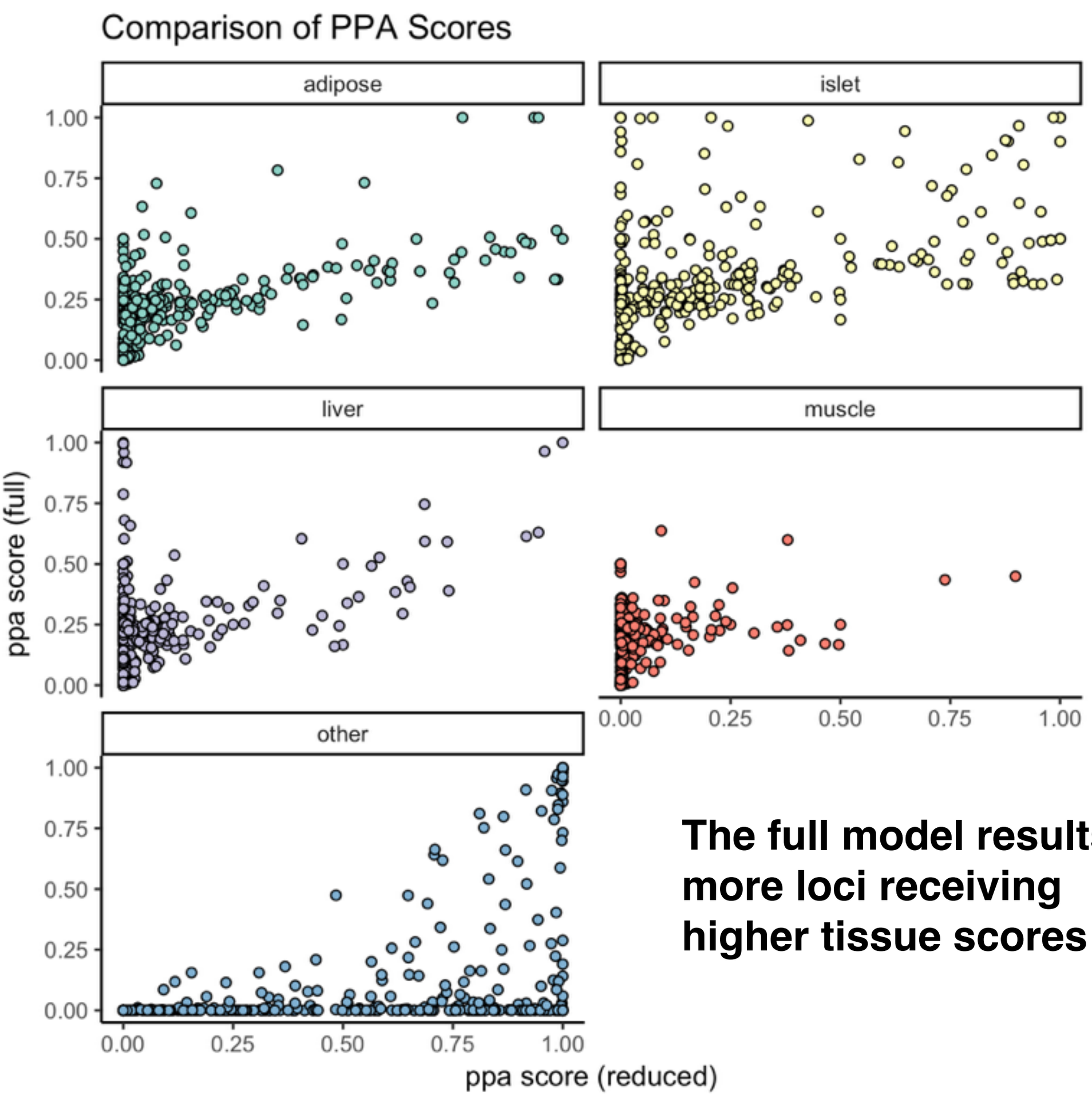
1. Take the sum of credible set PPA scores for each tissue
   - Yields a vector with four elements corresponding to each tissue

2. Determine the amount of credible set PPA not accounted for by variants mapping to tissue annotations
   - this is the “other” score for the locus

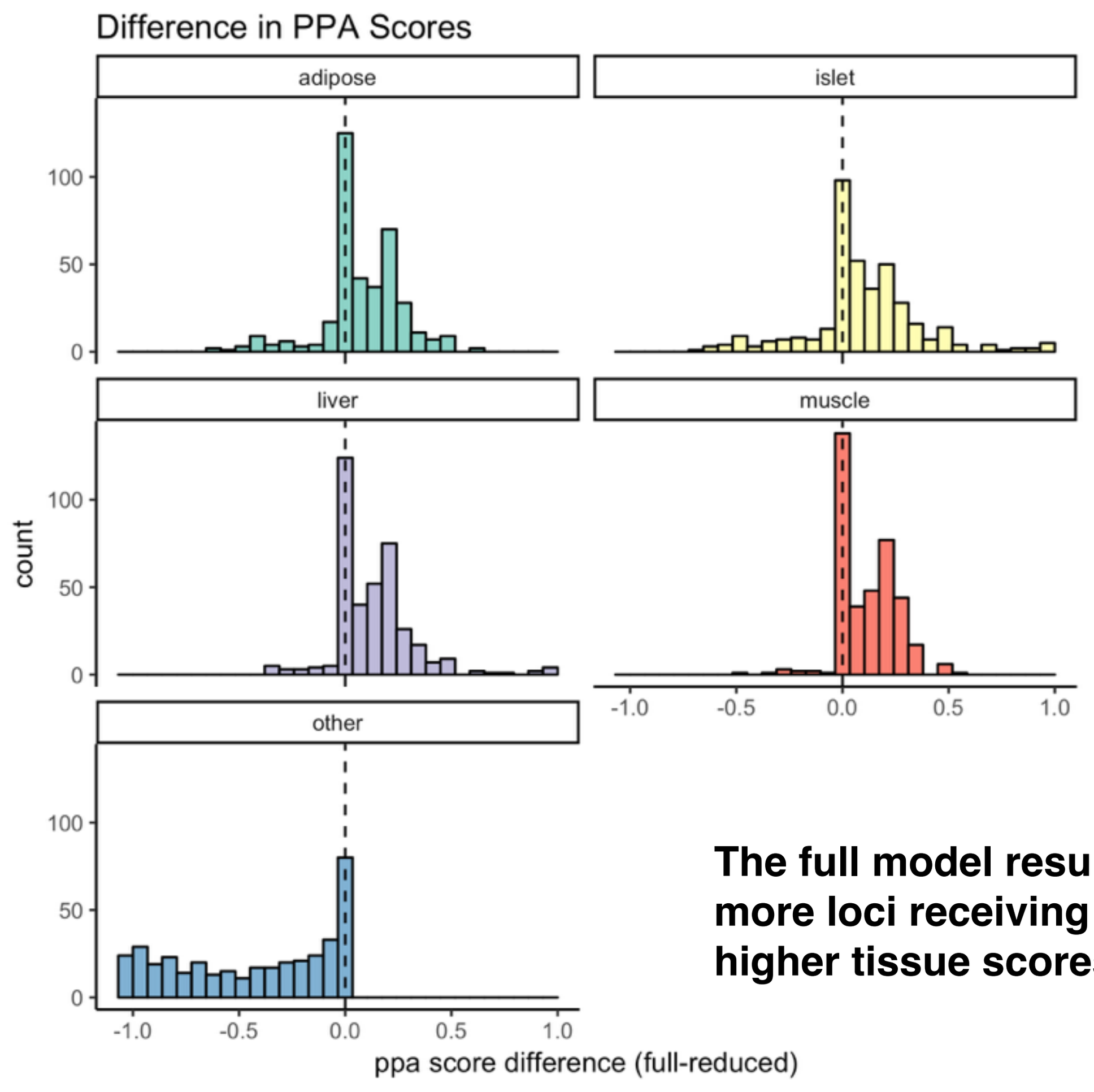
3. Scale the resulting score vector

What genomic annotations should be included in the scheme?

“**Reduced**” set	“**Full**” set
Coding	Coding
Strong enhancer	Strong enhancer
	Weak enhancer
	Promoter
	Gene transcription
	Genic enhancer



The full model results in more loci receiving higher tissue scores



The full model results in more loci receiving higher tissue scores

Dilution of signal in “full” model from assuming all annotations are equally important?

	refseq	Locus.ID	tissue	ppa.score.full	ppa.score.reduced	ppa.score.wfull	ppa.score.diff	ppa.score.diffW
1	MTNR1B	144_1	islet	**1**	**0**	1	1	0
2	MTNR1B	144_1	muscle	0	0	0	0	0
3	MTNR1B	144_1	adipose	0	0	0	0	0
4	MTNR1B	144_1	liver	0	0	0	0	0
5	MTNR1B	144_1	other	**0**	**1**	0	-1	0

	refseq	Locus.ID	tissue	ppa.score.full	ppa.score.reduced	ppa.score.wfull	ppa.score.diff	ppa.score.diffW
1	PPARG	30_1	islet	0.05492854	5.030831E-05	0.05493294	0.05487823	4.396757E-06
2	PPARG	30_1	muscle	0.05761088	5.669301E-05	0.05671312	0.05755418	-8.977524E-04
3	PPARG	30_1	adipose	**0.4450193**	**7.698313E-01**	0.52633933	-0.32481205	8.132003E-02
4	PPARG	30_1	liver	0.44244129	9.979674E-05	0.36201461	0.44234149	-8.042667E-02
5	PPARG	30_1	other	0.00000000	2.299619E-01	0.00000000	-0.22996186	0.000000E+00

	refseq	Locus.ID	tissue	ppa.score.full	ppa.score.reduced	ppa.score.wfull	ppa.score.diff	ppa.score.diffW
1	ADCY5	40	islet	**0.3131044**	**0.939313204**	0.3624188	-0.62620880	0.04931441
2	ADCY5	40	muscle	0.3016360	0.000000000	0.3265585	0.30163605	0.02492244
3	ADCY5	40	adipose	0.1973943	0.009529011	0.1621266	0.18786527	-0.03526772
4	ADCY5	40	liver	0.1878653	0.000000000	0.1488961	0.18786527	-0.03896913
5	ADCY5	40	other	0.0000000	0.051157784	0.0000000	-0.05115778	0.00000000

General approach for divvying PPA value for each SNP in a credible set

$$s_{j,a} = \begin{bmatrix} s_{j,a,t_1} \\ s_{j,a,t_2} \\ s_{j,a,t_3} \\ s_{j,a,t_4} \end{bmatrix}$$

$$s_{j,a,t} = \frac{P_j}{\sum_{i \in T} \mathbb{1}(j,a,i)} \mathbb{1}(j,a,t)$$

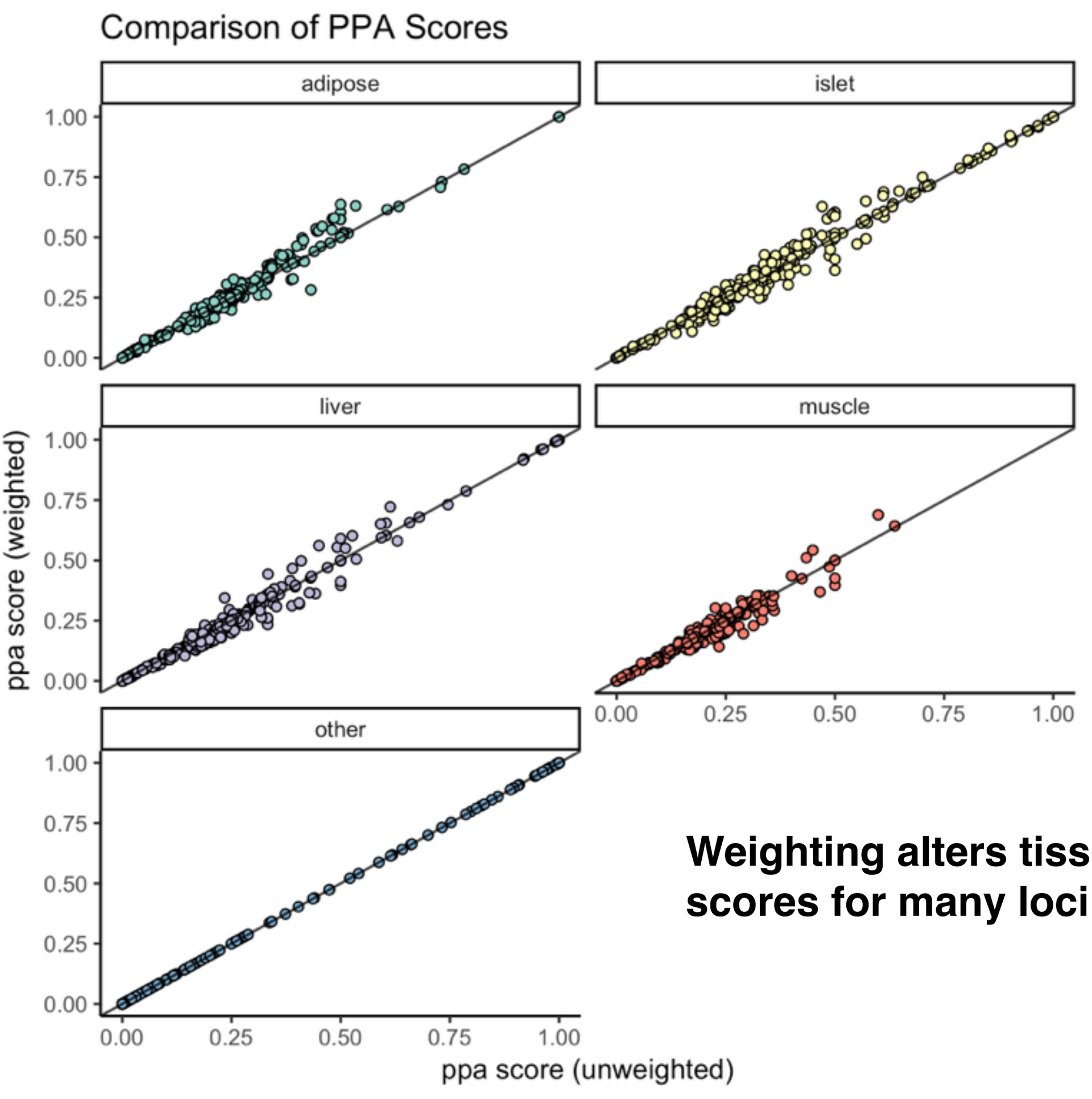
$$\mathbb{1}(j,a,t) := \begin{cases} 1 & \text{if SNP}j \text{ overlaps annotation } a \text{ in tissue } t \\ 0 & \text{otherwise} \end{cases}$$

where $s_{j,a}$ is score vector for SNP j and annotation a

$$s_j = P_j \frac{q}{\sum_{i=1}^{4} q_i}$$

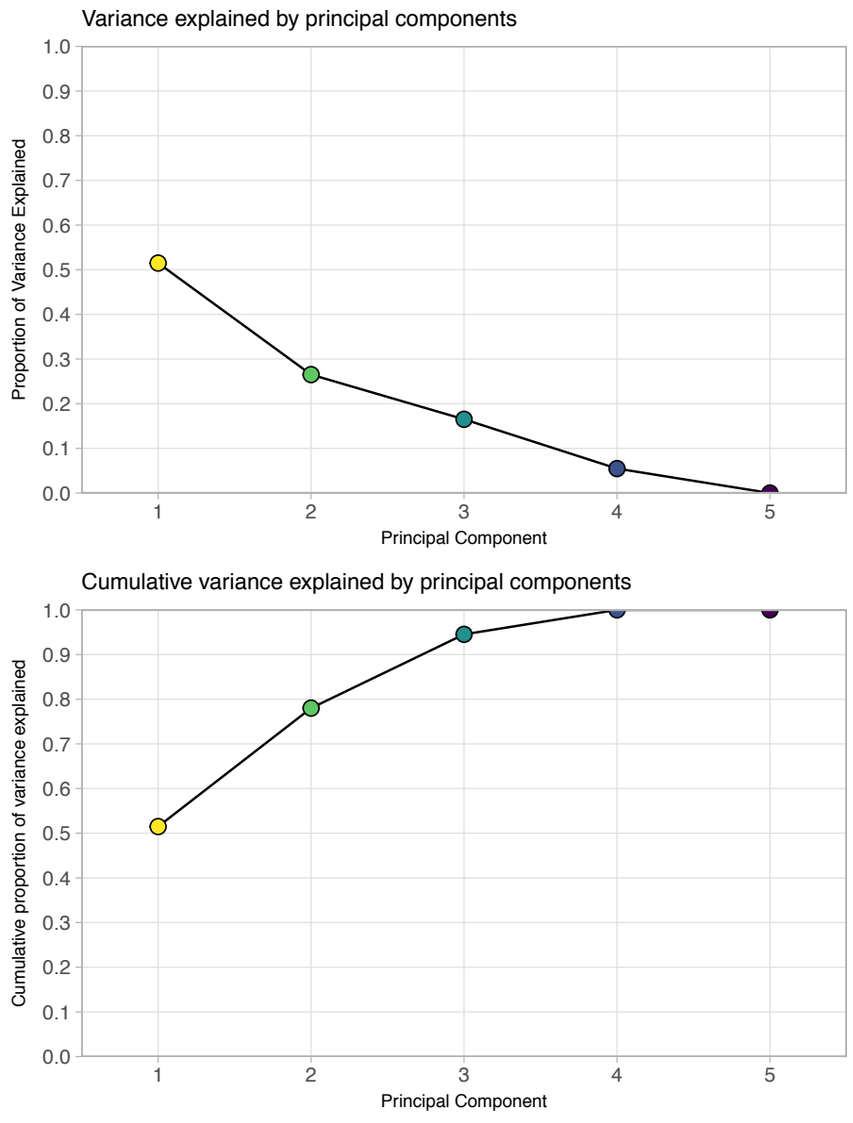
s_j is score vector for SNP j where A is the set of all annotations and w_a is a weight for annotation a

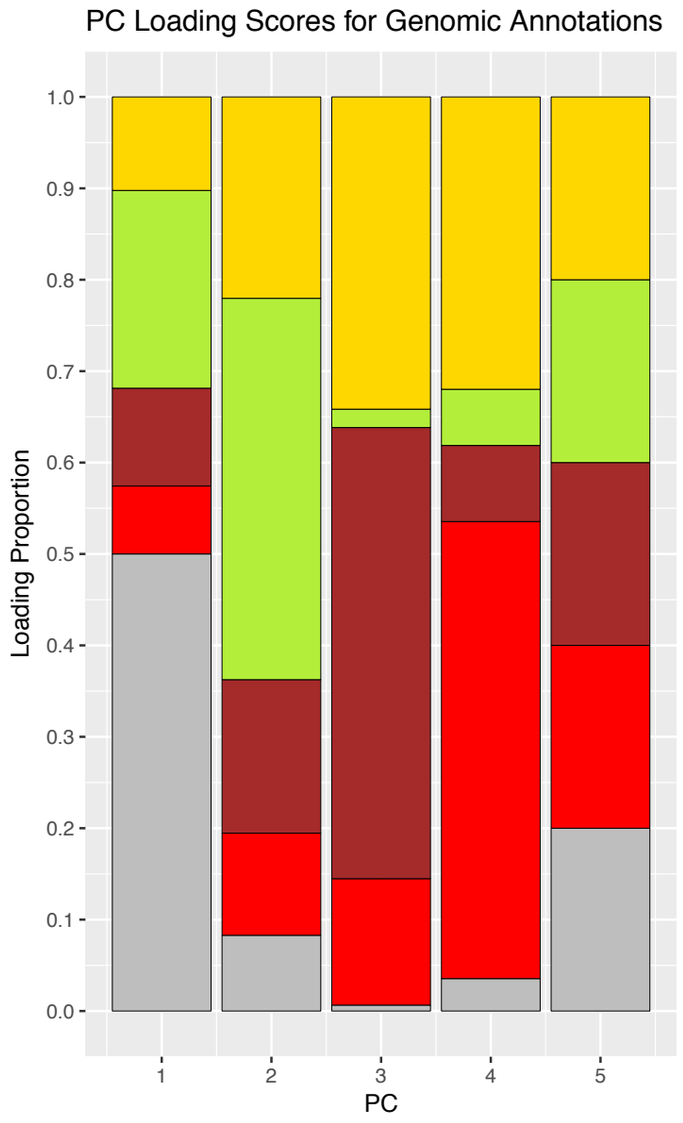
$$q = \sum_{a \in A} s_{j,a} w_a$$

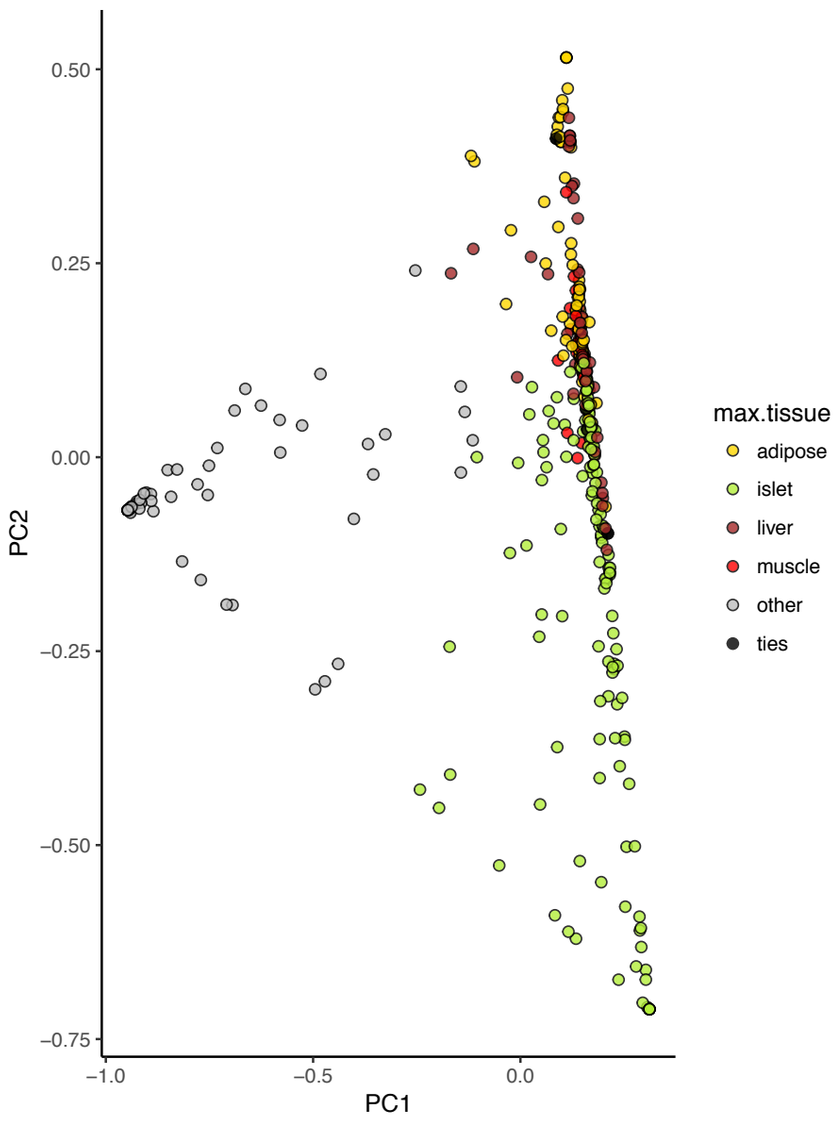


Weighting alters tissue scores for many loci

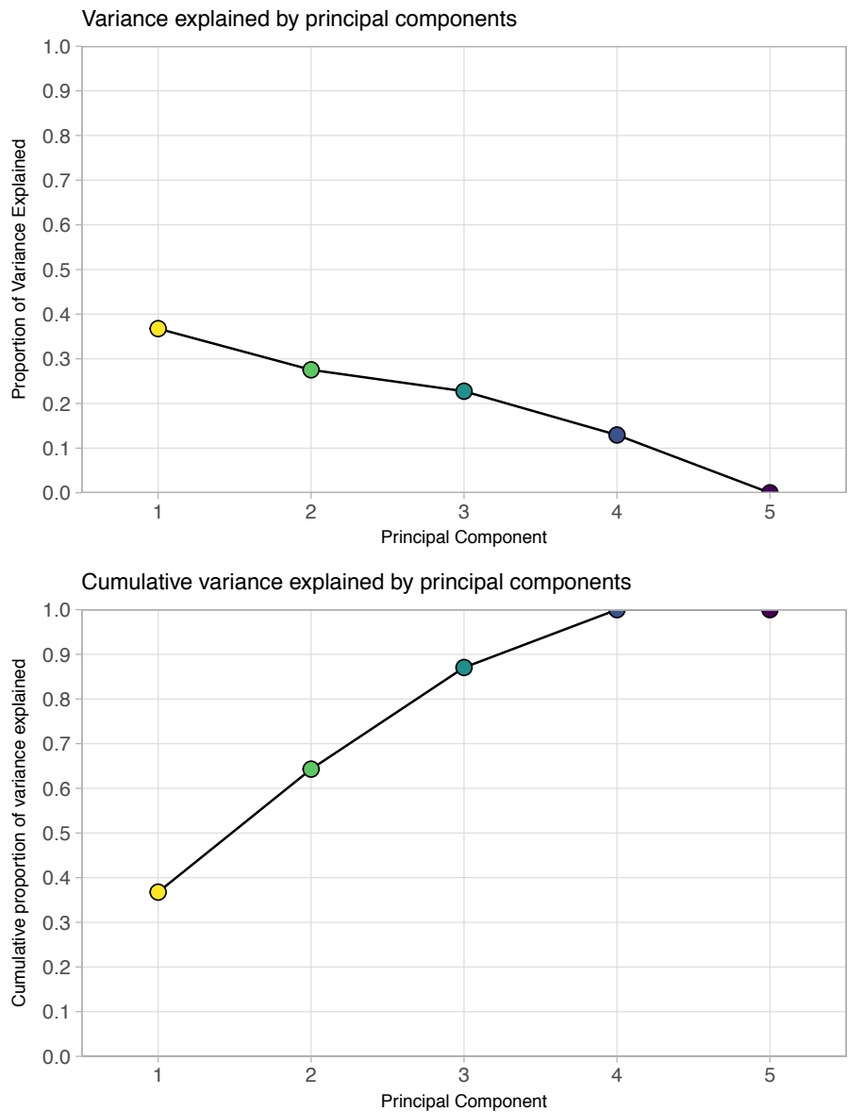
The first principal component separates “other” loci (n=93) from everything else

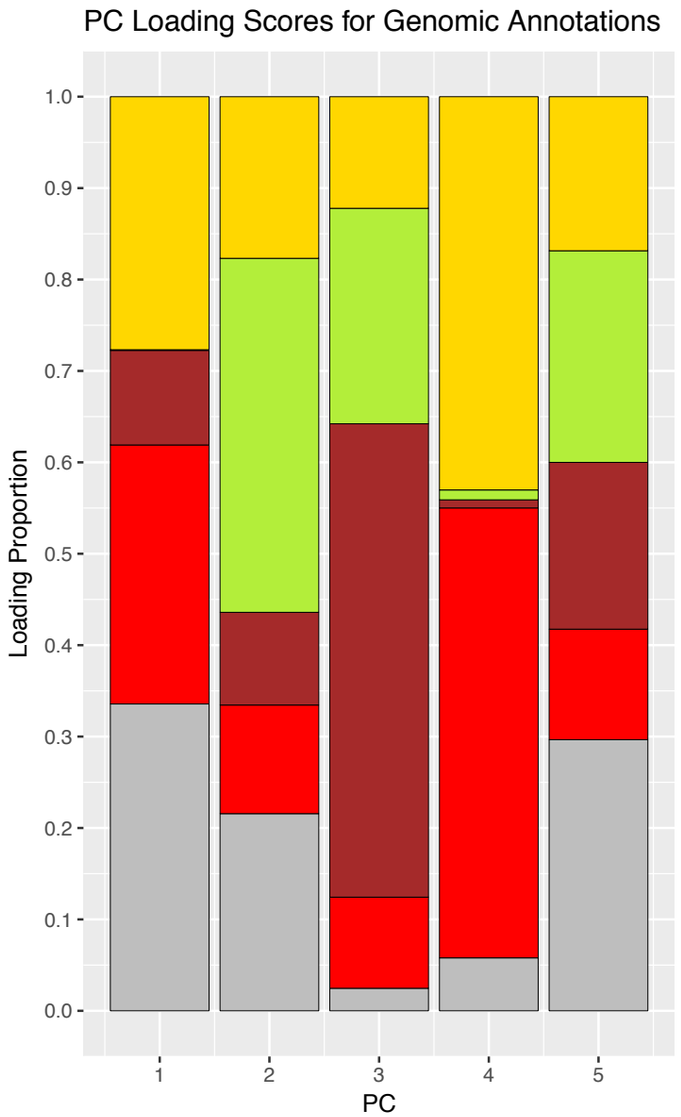


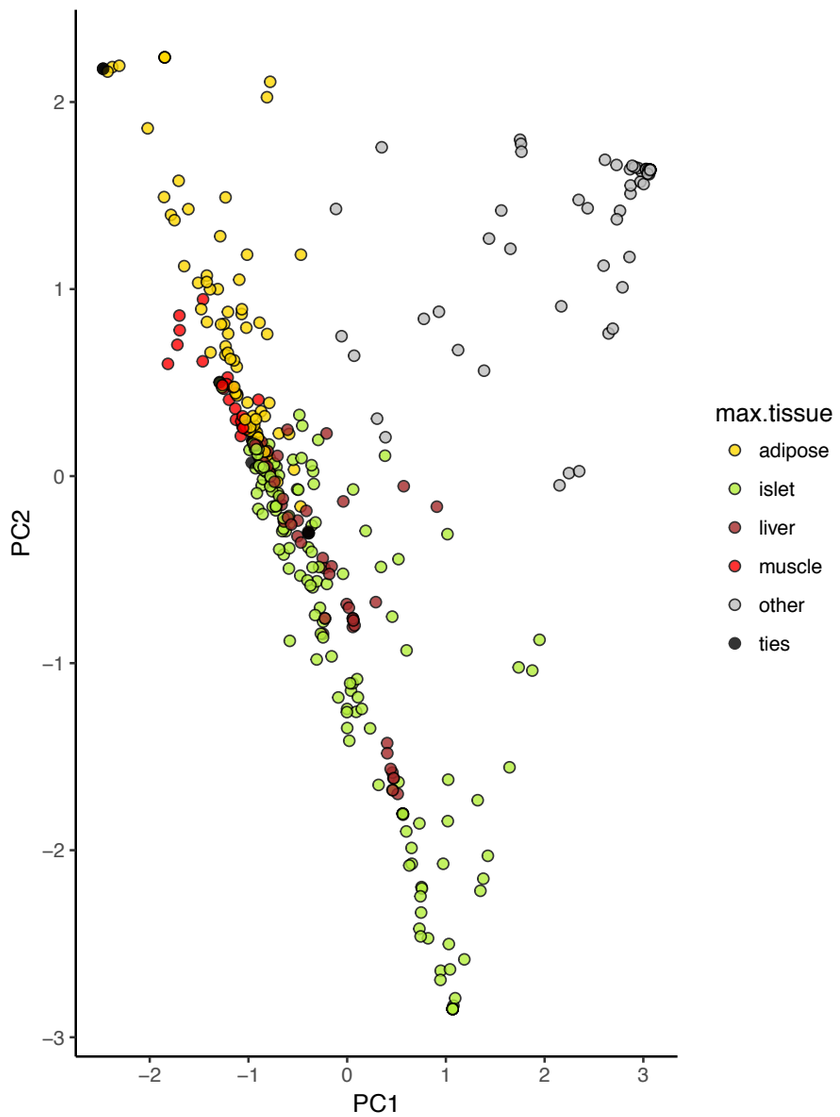




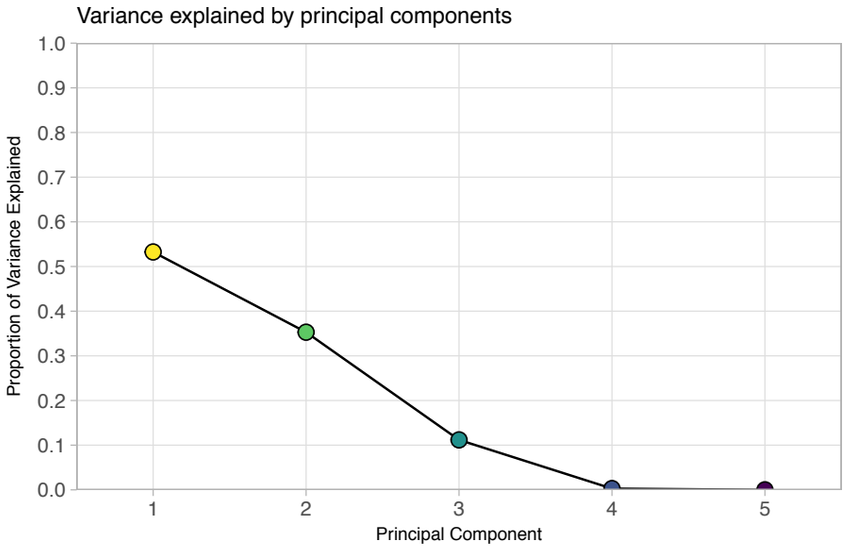
Similar profile when center and rescale input score matrix

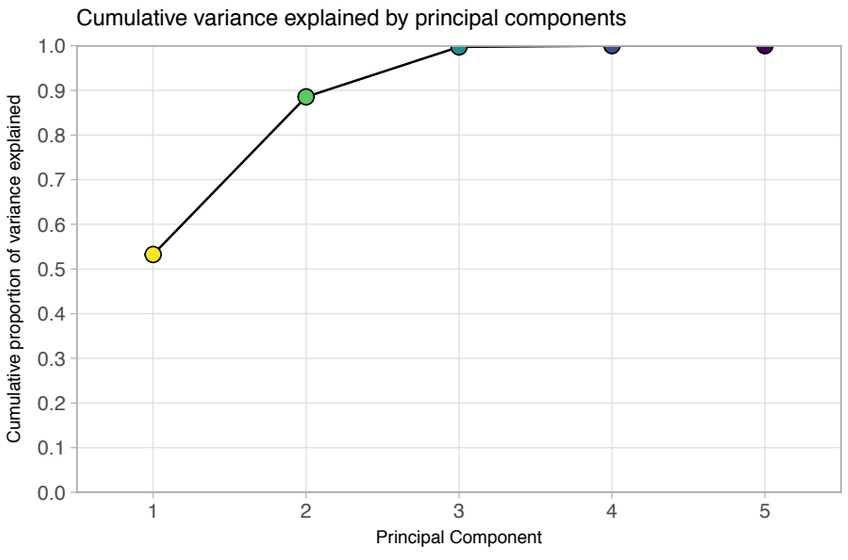


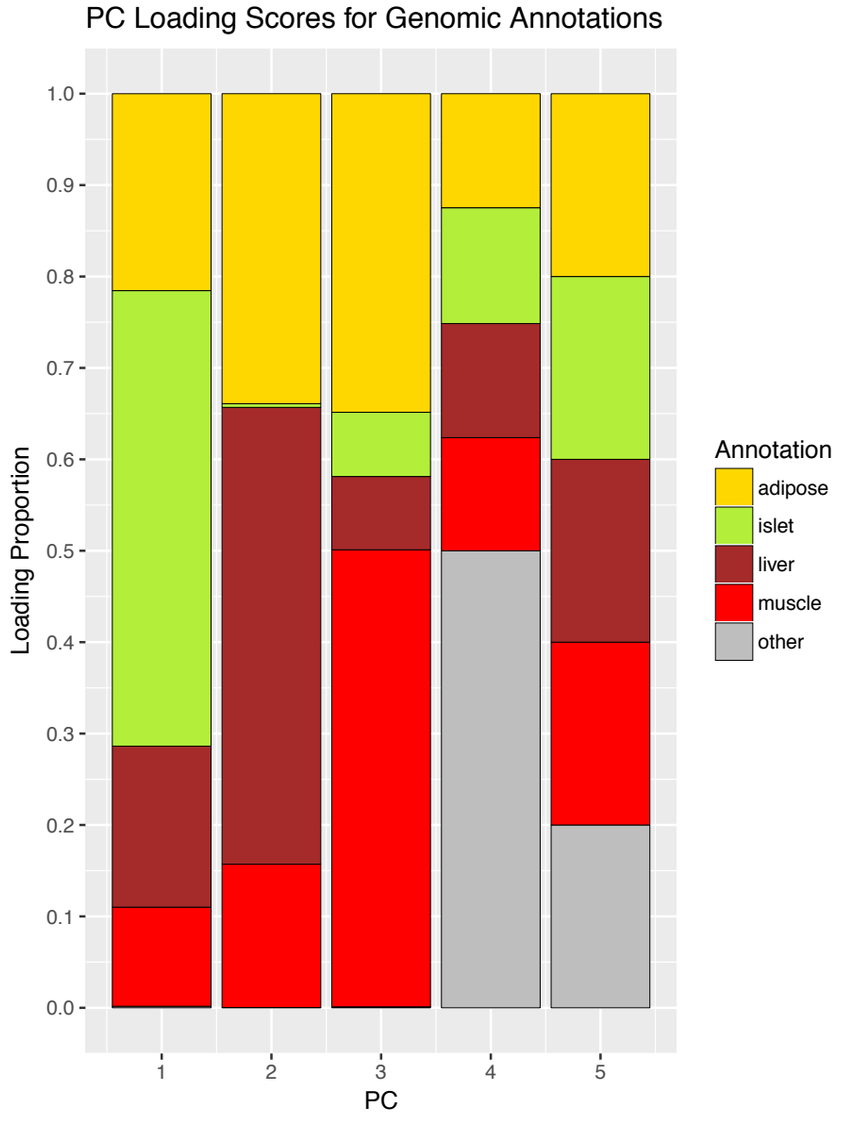


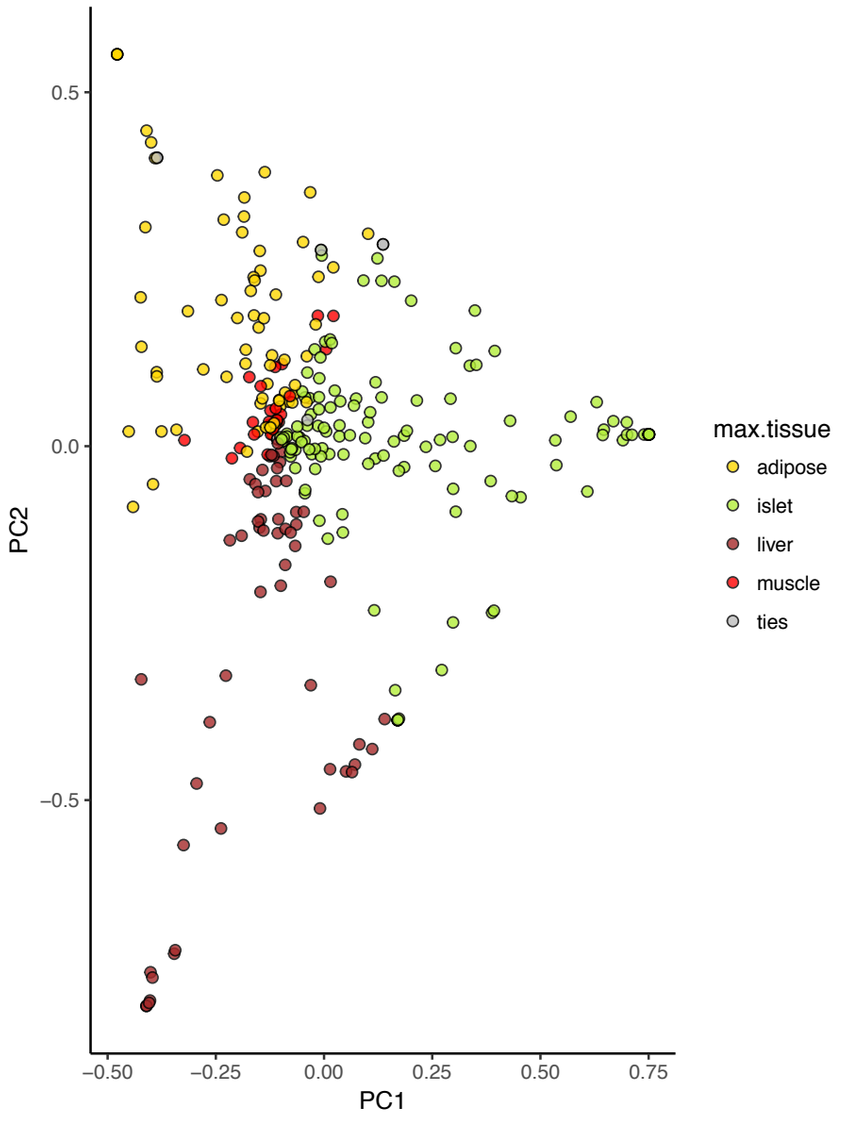


PC1 in pruned matrix (“other” score <= 0.10) separates islet loci from everything else. PC2 separates liver from adipose and muscle

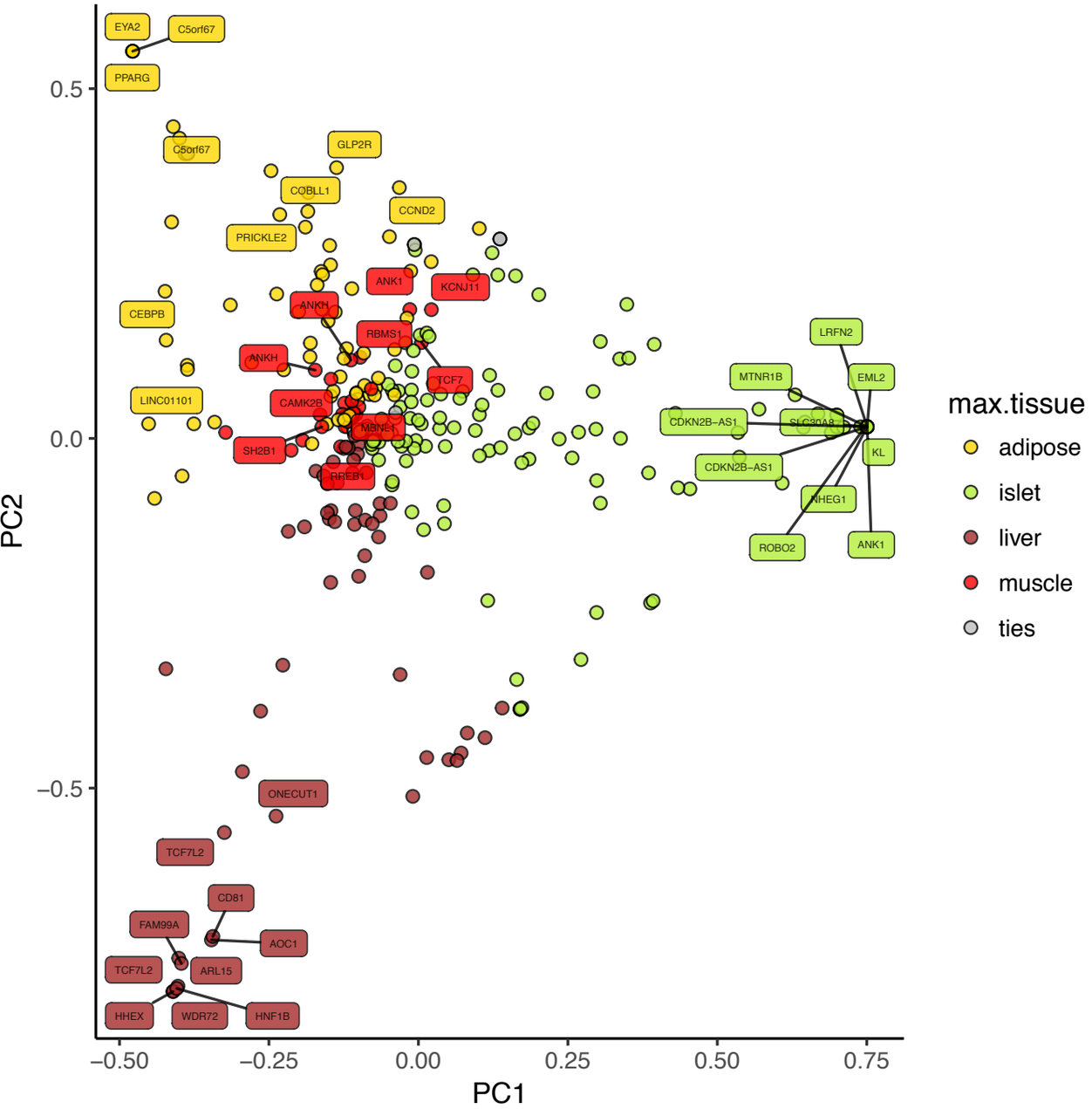


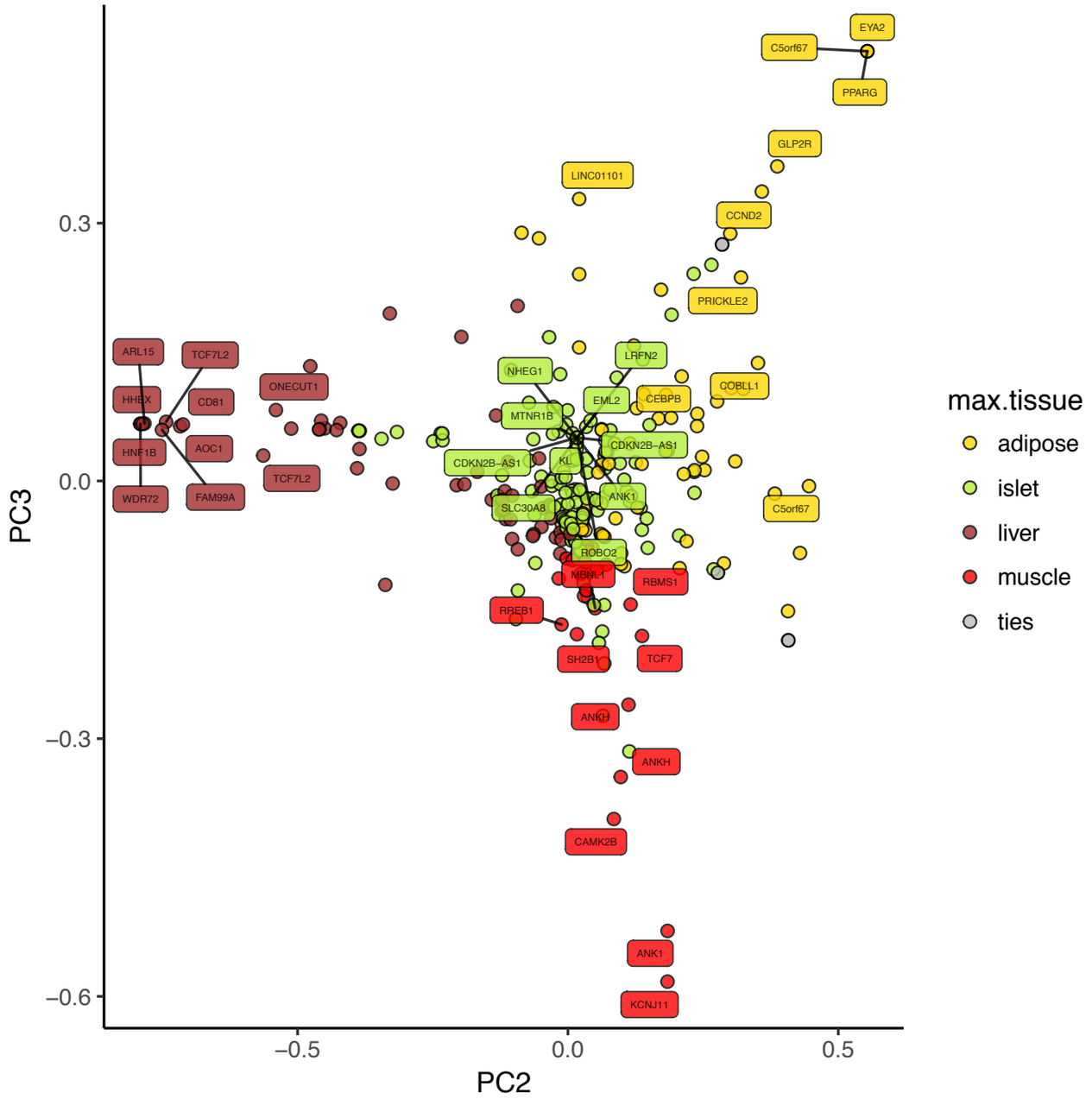




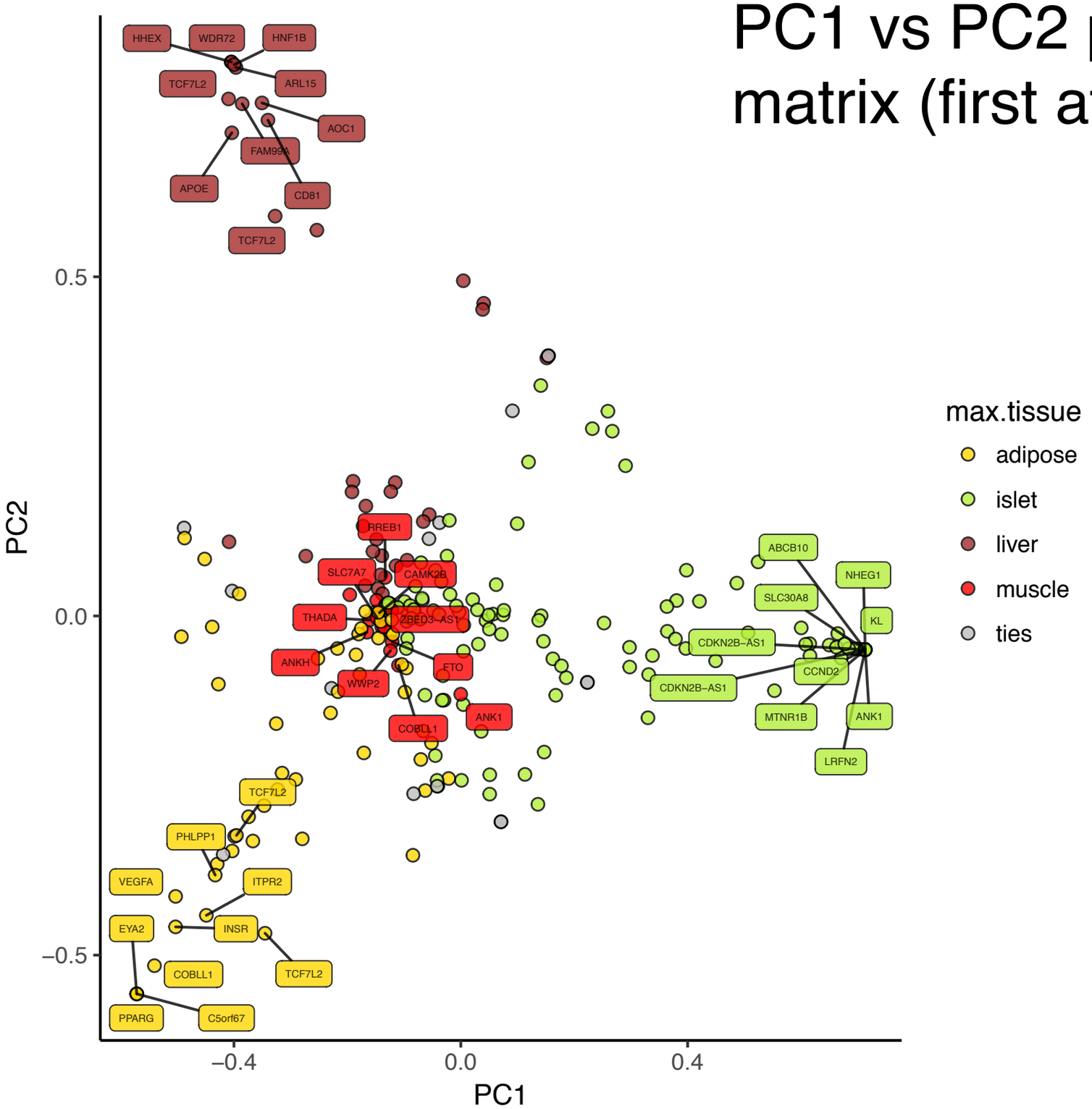


PC1 in pruned matrix (“other” score <= 0.10) separates islet loci from everything else. PC2 separates liver from adipose and muscle





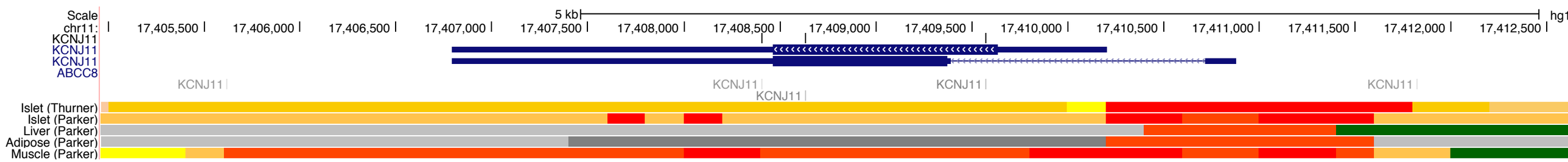
PC1 vs PC2 plot for unweighted “full” matrix (first attempt)



Tissues scores for *KCNJ11* credible set vary with model and weighting

	refseq	Locus.ID	tissue	ppa.score.full	ppa.score.reduced	ppa.score.wfull	ppa.score.diff	ppa.score.diffW
1	KCNJ11	135_1	islet	**0.392899363**	**0.612162221**	**0.303201465**	-0.219262859	-0.0896978980
2	KCNJ11	135_1	muscle	**0.59894185**	**0.380090176**	**0.688983349**	0.218851674	0.0900414998
3	KCNJ11	135_1	adipose	0.003256424	0.004884636	0.003797298	-0.001628212	0.0005408742
4	KCNJ11	135_1	liver	0.004902363	0.002862966	0.004017888	0.002039397	-0.0008844759
5	KCNJ11	135_1	other	0.000000000	0.000000000	0.000000000	0.000000000	0.0000000000

	Locus.ID	symbol	SEGNUMBER	SNPID	CHR	POS	logBF	Z	PPA
1	135_1	KCNJ11	54	chr11:17408630	chr11	17408630	54.6166	-10.8025	**0.47946011**
2	135_1	KCNJ11	54	chr11:17409572	chr11	17409572	54.4518	10.7872	**0.40660809**
3	135_1	KCNJ11	54	chr11:17408404	chr11	17408404	54.5897	-10.7996	0.09076952
4	135_1	KCNJ11	54	chr11:17418477	chr11	17418477	51.8248	-10.5403	0.01786720



annotation	weight
strong.enhancers	**1.5839700**
weak.enhancers	1.4815050
repressed.regions	-0.1424545
promoters	**2.5220150**
gene.transcription	1.0398415
bivalent.tss	1.5177900
genic.enhancer	1.8237250
low.signal	-1.0193400
coding	2.5870100

	Locus.ID	islet	muscle	adipose	liver	other
1	135_1	**0.3925058**	**0.6007978**	0.003443151	0.0032532	0

	GeneID	GeneName	islet.score	muscle.score	adipose.score	liver.score
1	ENSG00000187486.5	KCNJ11	0.1464343	**0.8365146**	0.01075026	0.00630090

	annotation	tissue	weight
1	strong.enhancers	adipose	1.7294600
2	**strong.enhancers**	**islet**	**2.4192650**
3	strong.enhancers	liver	1.0076935
4	strong.enhancers	muscle	0.5695666
5	weak.enhancers	adipose	0.8061710
6	weak.enhancers	islet	1.6125100
7	weak.enhancers	liver	1.4397500
8	weak.enhancers	muscle	1.5232600
9	repressed.regions	adipose	-0.0885795
10	repressed.regions	islet	-0.5018890
11	repressed.regions	liver	0.3082345
12	repressed.regions	muscle	0.0044020
13	promoters	adipose	2.2591100
14	promoters	islet	2.2379800
15	promoters	liver	1.8956730
16	**promoters**	**muscle**	**2.3708533**
17	gene.transcription	adipose	0.7433020
18	gene.transcription	islet	1.2535615
19	gene.transcription	liver	0.8084615
20	gene.transcription	muscle	0.9358600
21	bivalent.tss	adipose	1.0656500
22	bivalent.tss	islet	1.9313600
23	bivalent.tss	liver	1.8489800
24	bivalent.tss	muscle	1.1866000
25	genic.enhancer	adipose	1.6906900
26	genic.enhancer	islet	-23.9143000
27	genic.enhancer	liver	2.1158200
28	genic.enhancer	muscle	1.9567600
29	low.signal	adipose	-1.0189100
30	low.signal	islet	-1.0197700
31	low.signal	liver	-0.8916990
32	low.signal	muscle	-1.2294800
33	coding	adipose	2.5870100
34	coding	islet	2.5870100
35	coding	liver	2.5870100
36	coding	muscle	2.5870100

Proportion of PPA attributable to each ann
islet

218_2: EML2
161: KL
144_1: MTNR1B
117_4: CDKN2B–AS1
117_1: CDKN2B–AS1
110_1: SLC30A8
106_1: ANK1
88: NHEG1
81: LRFN2
39: ROBO2
68_2: PELO
58: SCD5
226_1: GNAS–AS1
144_2: DEUP1
146_4: CCND2–AS1
48_2: TRA2B
117_3: CDKN2B–AS1
233: RESP18
149: KLHL42
20_2: LINC01122
220: NKX2–2
173_1: RASGRP1
117_6: CDKN2B–AS1
153_1: RMST
236_2: RPS10–NUDT3
218_1: EML2
154: WSCD2
127_4: C10orf35
191_1: CTRB2
73: BHMT
45: EGFEM1P
155_1: KSR2
145_3: ETS1
8: SRGAP2D
132_1: MIR4686
92_2: STK31
188_1: SLX1B
106_2: ANK1
112: LINC00824
41: SLC12A8
173_2: RASGRP1
120: LOC101927450
66_2: OTULIN
44_3: P2RY1
14: TBCE
63: TMEM154
107_2: NDUFAF6
48_1: IGF2BP2
122: LOC101927502
133_4: KCNQ1

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Proportion

Locus

Proportion of PPA attributable to each ann
muscle

106_3: ANK1
135_1: KCNJ11
210_1: BCL2
37: ADAMTS9–AS2
31_1: UBE2E2
16: FAM49A
90: LOC102724152
70_3: C5orf67
95_1: CAMK2B
66_3: ANKH
170: NRXN3
66_1: ANKH
28_2: COBLL1
151: TSPAN8
237_1: LINC01625
148: ITPR2
95_2: CAMK2B
77_2: TCF7
167: SLC7A7
133_1: KCNQ1
126: CDC123
133_7: KCNQ1
78_3: RREB1
237_2: LOC100507477
133_8: KCNQ1
43: LINC01214
5: PTGFRN
72: ZBED3–AS1
27: RBMS1
187: SH2B1
189_1: FTO
165: LINC01080
234: PIK3R1
184: PRC1
99_2: CPA1
180: MAP2K5
31_2: UBE2E2
159: RNF6
40: ADCY5
96_2: LOC100630923
36_1: ATXN7
19_2: THADA
166_2: IRS2
113_1: BOP1
44_1: MBNL1
89: SLC22A3
189_3: FTO
107_3: NDUFAF6
196: RAI1
190: WWP2

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Proportion

Proportion of PPA attributable to each annc
adipose

223_1: EYA2
70_2: C5orf67
30_2: PPARG
87_2: RSPO3
235: LOC101927697
195: GLP2R
24_3: GLI2
36_2: PRICKLE2
70_3: C5orf67
12_1: LOC102723886
90: LOC102724152
239: WDR11
210_1: BCL2
146_6: CCND2
146_1: CCND2
146_5: CCND2–AS1
37: ADAMTS9–AS2
31_1: UBE2E2
24_1: LINC01101
237_1: LINC01625
224: CEBPB
28_1: COBLL1
148: ITPR2
98: CTTNBP2
212: INSR
133_3: KCNQ1
122: LOC101927502
70_1: C5orf67
30_1: PPARG
82_2: VEGFA
133_4: KCNQ1
137_2: EHF
162: DLEU1
210_2: PHLPP1
130_1: TCF7L2
6: NOTCH2
130_6: TCF7L2
31_2: UBE2E2
230: PNPLA3
147: CDKN1B
173_2: RASGRP1
69_3: ARL15
117_2: CDKN2B–AS1
94: CRHR2
9: SRGAP2D
130_5: TCF7L2
241_2: NCOR2
241_1: CCDC92
59_1: FAM13A
134: PDE3B

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Proportion

Locus

Proportion of PPA attributable to each ann
liver

176: WDR72
129_1: HHEX
198_2: HNF1B
69_1: ARL15
132_2: FAM99A
130_3: TCF7L2
100: AOC1
132_4: CD81
130_4: TCF7L2
175: ONECUT1
91_3: AGMO
146_2: PARP11
18: GCKR
217_1: APOE
222_1: HNF4A
129_3: EXOC6
156_4: HNF1A
230: PNPLA3
242_1: SHBG
46: SLC2A2
156_3: HNF1A–AS1
222_3: HNF4A
198_1: HNF1B
52: FGFR3
50: LOC105374266
24_1: LINC01101
11_1: PROX1–AS1
11_2: PROX1–AS1
198_3: HNF1B
145_2: LOC105369564
12_1: LOC102723886
1: MACF1
30_1: PPARG
156_5: SPPL3
239: WDR11
141: LTBP3
11_3: PROX1
22_1: CEP68
206_1: TCF4
197: NF1
114_2: GLIS3
14: TBCE
41: SLC12A8
125_3: LINC01451
42_1: PLXND1
174_1: RPAP1
132_1: MIR4686
112: LINC00824
202_1: ERN1
19_3: ZFP36L2

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Proportion

Proportion of PPA attributable to each ann
other

208: GRP
204: LAMA1
146_3: CCND2–AS1
127_2: NEUROG3
121: LOC101927450
118: LINGO2
111: CASC11
105: PURG
91_1: DGKB
86_1: CENPW
75: SLCO6A1
62: LINC00613
21: MIR4432HG
15_1: LOC727944
164_2: LINC00374
163_2: LOC101926897
150_1: RPSAP52
109: TRHR
97_1: RELN
124: ABO
127_1: C10orf35
99_1: KLF14
205: MIR924HG
35: CACNA2D3
15_2: TMEM18
164_1: LINC00374
169: CLEC14A
116_1: MLLT3
209_1: MC4R
25: PABPC1P2
92_1: STK31
83_1: TFAP2B
91_2: DGKB
29: LOC646736
136_2: WT1–AS
163_1: PCDH17
225: LINC01524
20_1: LINC01122
60: LOC101929210
166_1: LINC00399
142_2: MYEOV
209_2: MC4R
44_2: P2RY1
168: AKAP6
104: LPL
56: GNPDA2
83_2: TFAP2B
211_1: KDM4B
57_2: SPATA18
150_2: HMGA2

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Proportion

Annotation
islet
muscle
adipose
liver
other

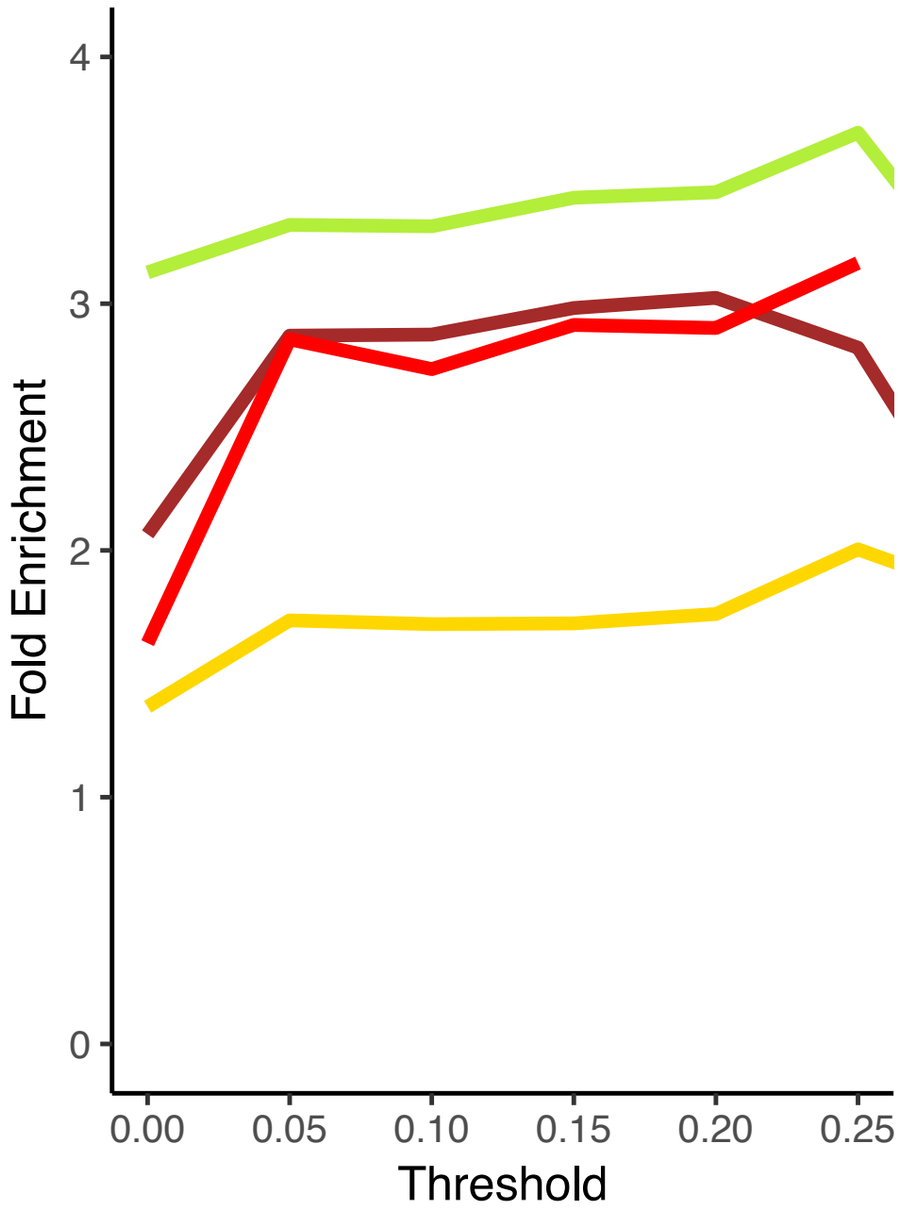
Locus

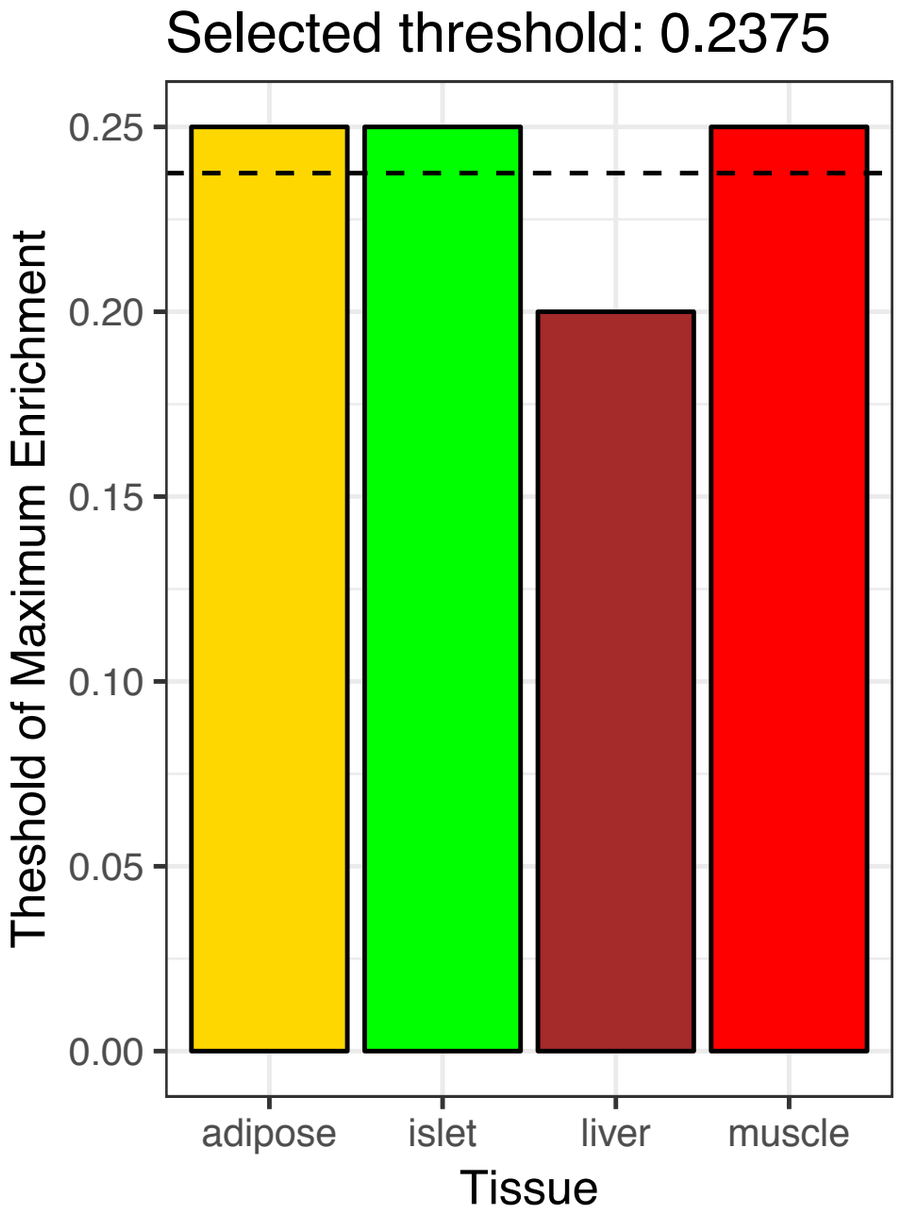
Scale
chr7:
MC4R
KLF14
Islet (Thurner)
Islet (Parker)
Liver (Parker)
Adipose (Parker)
Muscle (Parker)

130,430,000
130,435,000
130,440,000
130,445,000
20 kb
130,450,000
130,455,000
130,460,000
130,465,000
hg19
130,470,000

Establishing a rule-based classifier

- Tissue classifier applied the score values for each locus
- For a given locus, assign to tissue t if score value for that tissue is the greatest in the set **and** greater than threshold r
- Threshold r is determined through an eQTL enrichment analysis
  - for each evaluated threshold on interval [0,1], classify loci into tissue groups
  - for each tissue group, take the combined set of credible SNPs across all assigned loci and perform and eQTL enrichment analysis (MAF-adjusted)
  - Assess eQTL enrichment for the corresponding tissue
  - Identify threshold that maximizes enrichment for each tissue
    - subject to (a) enrichment being significant and (b) enrichment values being available for all tissues at threshold
  - Take mean of maximizing thresholds

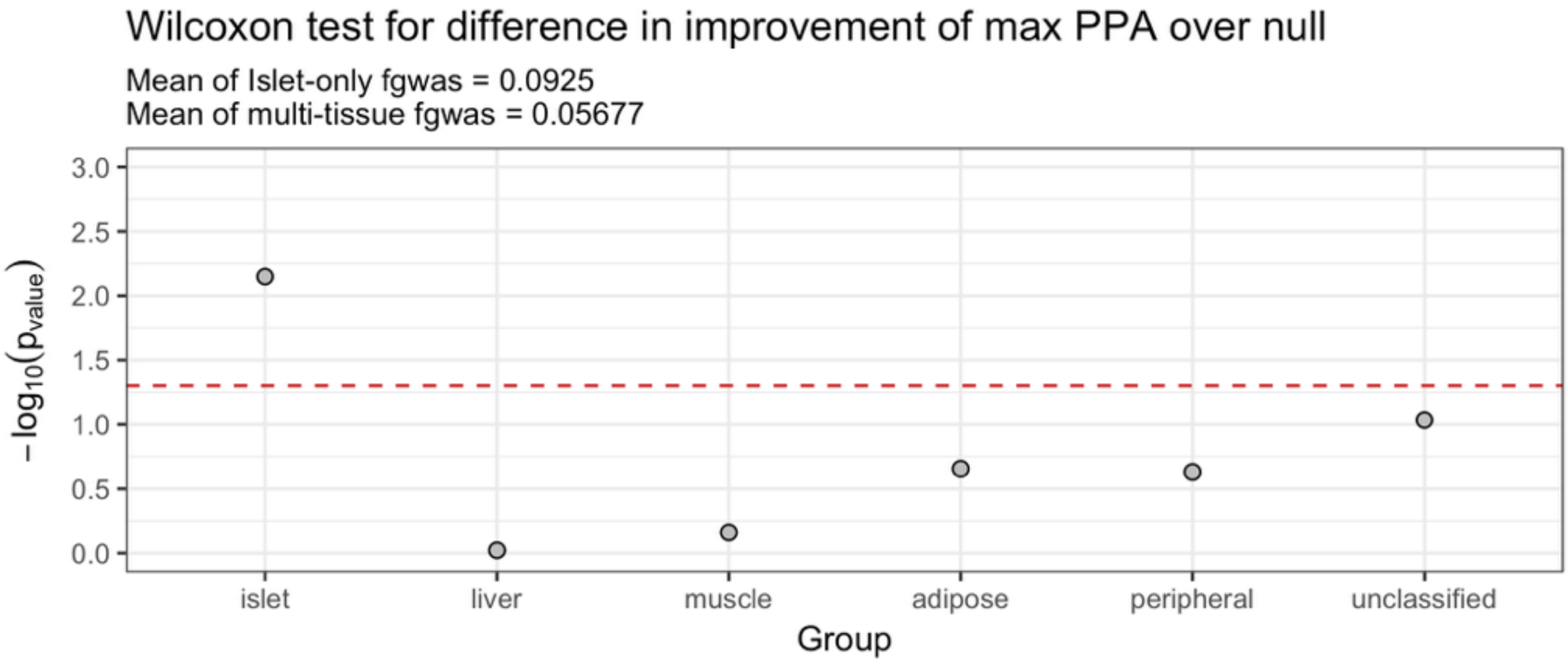




	islet	liver	adipose	muscle
1	136	39	59	21

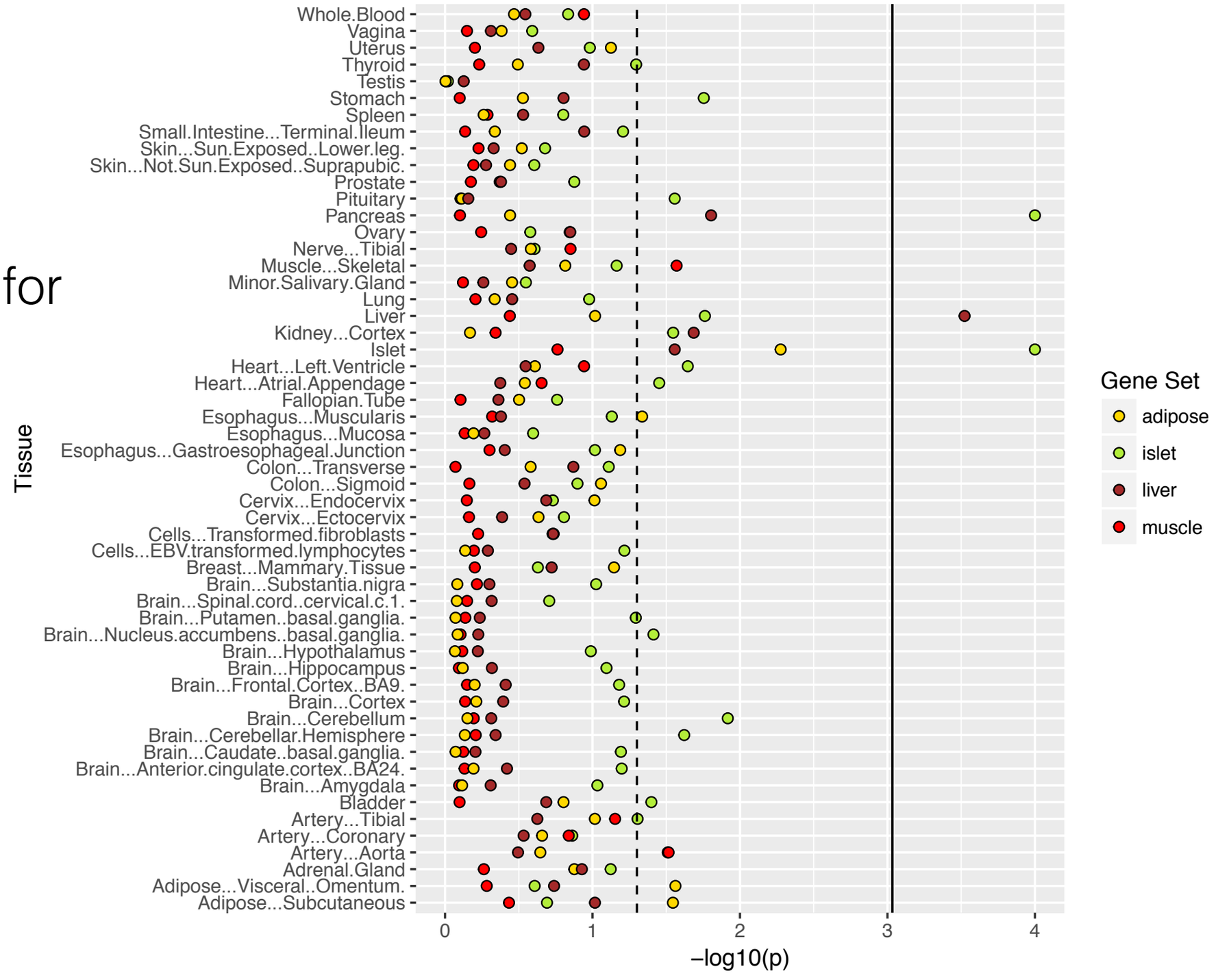
Validating the tissue classifier

Validation 1: Test if fgwas fine-mapping of loci classified as **islet** improves with additional islet epigenetic data



Validation 2: Gene co-expression analysis of annotated genes for each classified loci across tissues

Rank-based permutation test for coexpression



Conclusions and current work

- “functional” credible set-based tissue scores separate islet and insulin-responsive peripheral tissue loci
- Secondary separation of liver from adipose and muscle
- Top loci for each group are biologically plausible although some require special consideration (e.g. *KNCJ11*)
- Some loci unclassified due to low/quiescent signals in chromatin segmentation (e.g. *MC4R*, *KLF14*)
- Rule-based classifier results in tissue groups that are supported by validation analyses
  - only loci classified as islet are more finely-mapped if additional islet epigenetic data
  - tissue-classified loci genes are most highly co-expressed in the corresponding tissue

Special thanks

- Anubha
- Moustafa
- Juan
- Agata
- Matthias
- Mark
- and you too ;)

islet shared enhancer variant

	refseq	Locus.ID	tissue	ppa.score.full	ppa.score.reduced	ppa.score.diff
1	CDC123	126	islet	1.666667e-01	0.5	-3.333333e-01
2	CDC123	126	muscle	3.333333e-01	0.0	3.333333e-01
3	CDC123	126	adipose	3.333333e-01	0.0	3.333333e-01
4	CDC123	126	liver	1.666667e-01	0.5	-3.333333e-01
5	CDC123	126	other	1.110225e-16	0.0	1.110225e-16

Coding and islet-specific strong enhancer

	refseq	Locus.ID	tissue	ppa.score.full	ppa.score.reduced	ppa.score.diff
1	PPARG	30_1	islet	0.05492854	5.030831e-05	0.05487823
2	PPARG	30_1	muscle	0.05761088	5.669301e-05	0.05755418
3	PPARG	30_1	adipose	0.43986191	7.698313e-01	-0.32996944
4	PPARG	30_1	liver	0.44759867	9.979674e-05	0.44749888
5	PPARG	30_1	other	0.00000000	2.299619e-01	-0.22996186

(filter(full.df,other<0.50) %>% dim(.))[1] # 314 loci with other scores < 0.50
(filter(cse.df,other<0.50) %>% dim(.))[1] # 148 loci with other scores < 0.50

(filter(full.df,other<0.10) %>% dim(.))[1] # 237 loci with other scores < 0.10
(filter(cse.df,other<0.10) %>% dim(.))[1] # 67 loci with other scores < 0.10

There is a slight correlation between the number of snps in credible set and other score
0.11, pval=0.03 (reduced);

Conclusion: The number of SNPs in the credible set doesn't influence the difference in ppa scores between full and reduced methods, altough slight correlation with number of snps and value of the "other" score