Group meeting
Brain projects

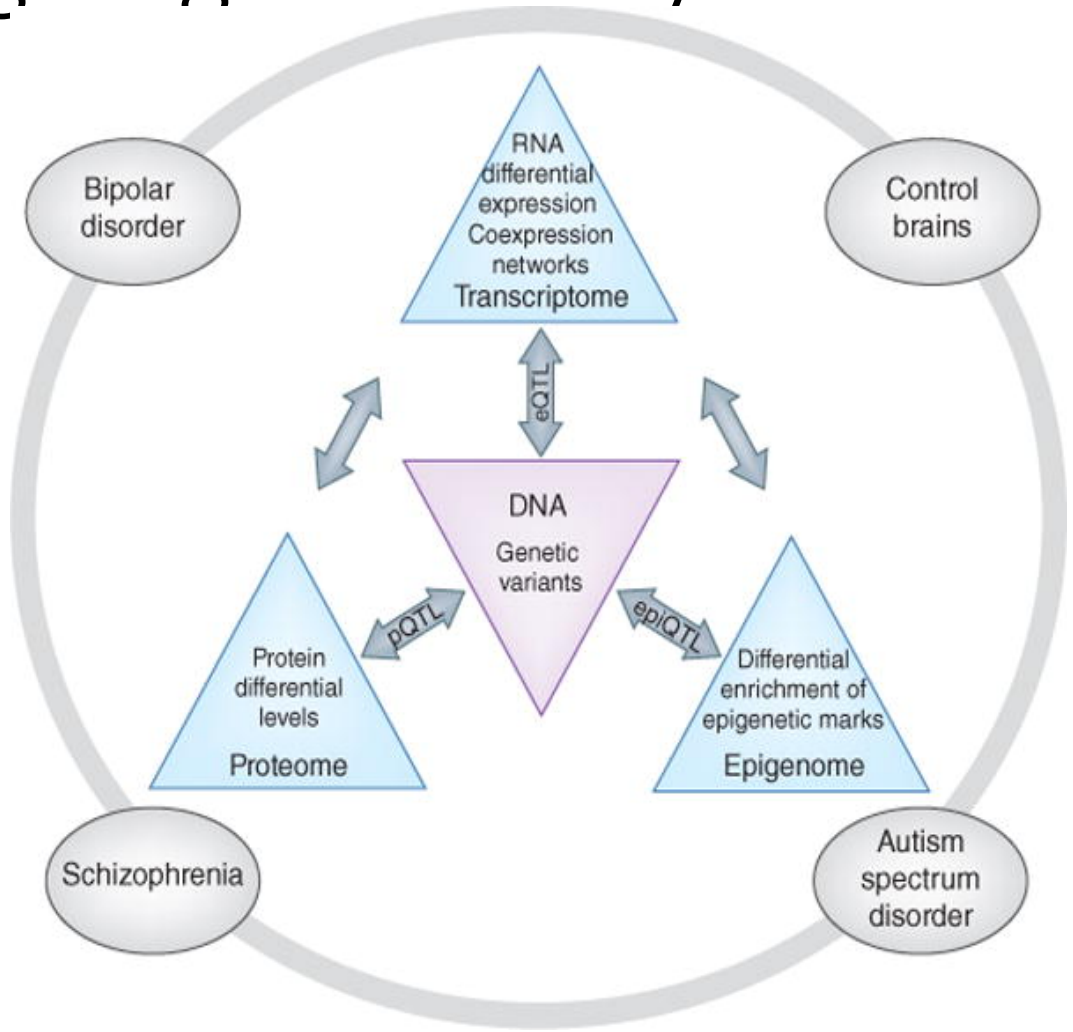
SL-05/26/2017

Work progress by brain group

Big Genomic Data to study Psychiatric Disease

Focus on psychiatric disease	PsychENCODE (>1000 disease and healthy control brains) CommonMind Consortium (600 disease and healthy control brains)
Disease SNPs	GWAS Catalog The NHGRI-EBI Catalog of published genome-wide association studies Psychiatric Genomics Consortium
Dozens of tissues	BRAINSPAN ATLAS OF THE DEVELOPING HUMAN BRAIN (13 developmental stages, 8-16 brain regions)
Various life stages	GTEx PROJECT Genotype-Tissue Expression (GTEx) (> 40 tissues including brain tissues)
Human epigenomic data	ROADMAP epigenomics PROJECT Different tissues (including healthy control brain) and cell lines

Ongoing work- PsychENCODE Project



Institutions

Yale, Mount Sinai, U Chicago, U MASS,UCLA ,USC, UIC, UCSF

Brain regions and cell types

DFC, CBC
NeuN+ and NeuN-

Sample size

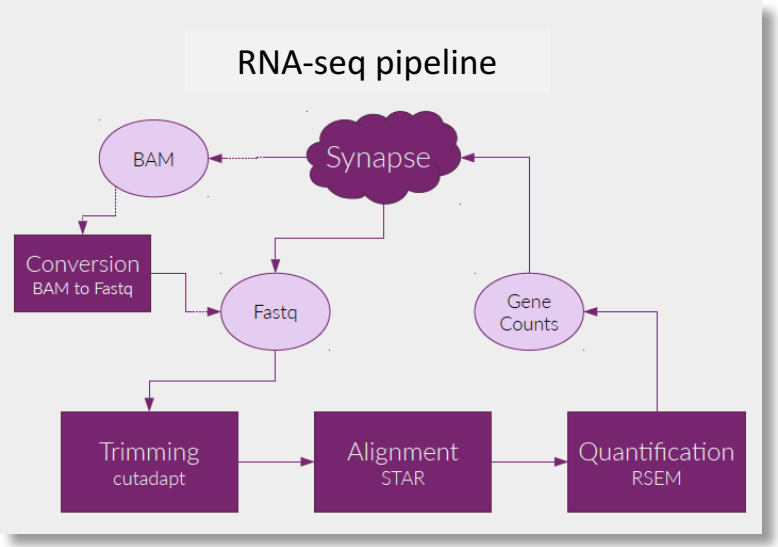
>1000 samples

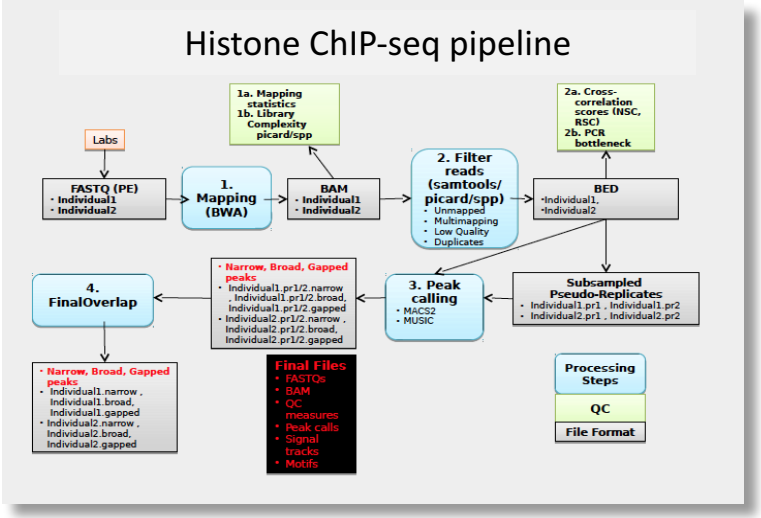
Datasets

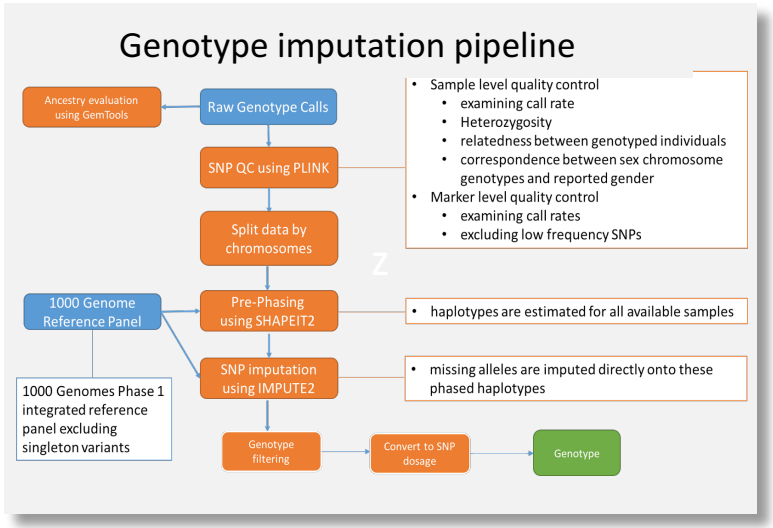
RNA-seq, ChIP-seq, genotype, ATAC-seq, WGBS

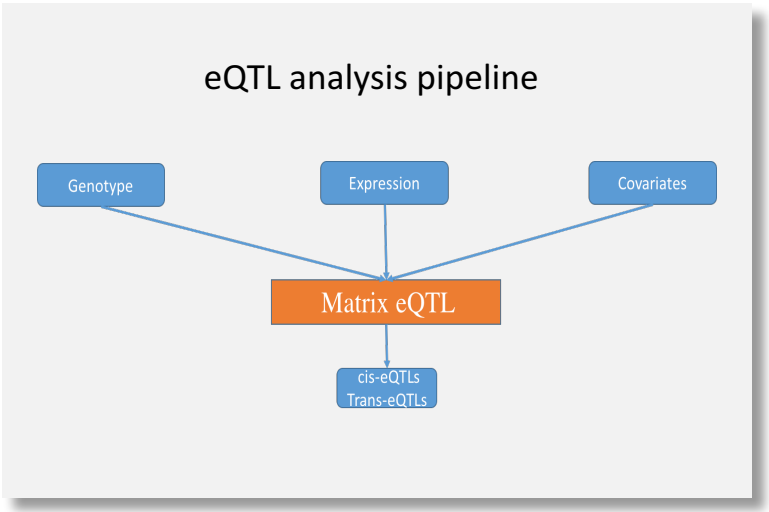
The PsychENCODE Consortium including **S. Liu**, *Nature Neuroscience* (2015)

PsychENCODE pipelines









Capstone projects

- **Capstone Project 1**: Cross – disorder gene expression analyses in autism, schizophrenia, and bipolar disorder
- **Capstone Project 2**: Adult and disease epigenetic
- **Capstone Project 3a**: Transcriptome and eQTL analyses across human brain development
- **Capstone project 3b**: Analysis of prompters/enhancer elements across early development: Construction of a Developmental EpiMap
- **Capstone Project 4**: Integrative analysis (with CommonMind, ENCODE, GTEx, and Roadmap)

Capstone 4 updates

Integrative analysis (with CommonMind, ENCODE, GTEx, and Roadmap)

Capstone 4 group

05/26/2017

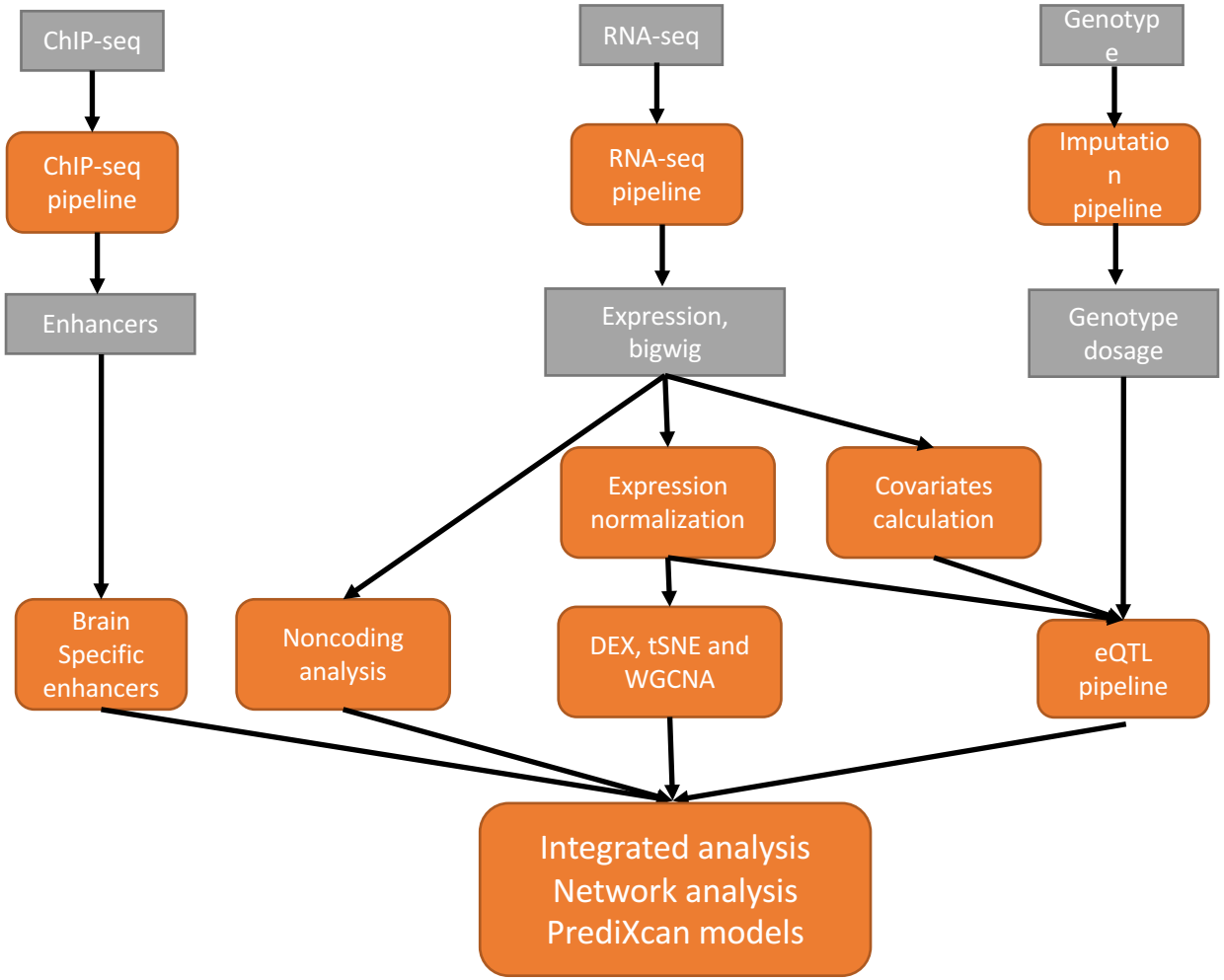
Capstone 4 - Analysis plan and updates

Analysis Tasks	Responsible Lab	Updates
1. Brain eQTLs and allelic expression using combination of PsychENCODE, GTEx, CommonMind and Roadmap data	Gerstein Lab, Sklar lab	Finished eQTL calculation for GTEx data
2. Identify brain specific enhancers using PsychENCODE and Roadmap data	Knowles Lab, Farnham Lab	Roadmap enhancers will start identify PEC enhancers
3. Analysis of eQTLs on brain specific enhancers and super enhancers	Gerstein Lab Knowles Lab, Farnham Lab	As soon as finish part of task 1 and 2
4. Develop PrediXcan models of the genes based on the new eQTL maps and applying these models to the GWAS risk alleles of different brain disorders	Knowles Lab, Farnham Lab	As soon as finish part of task 1,2 and 3
5. Protein-coding gene and ncRNA/TARs expression from RNA-seq	Gerstein Lab	
6.Specific gene expression and gene co-expression modules along with highly associated ncRNAs/TARs of different tissues and brain regions	Gerstein Lab	Identified brain specific genes using GTEx data
7. microRNA analysis for different tissues	Gerstein Lab	Fall 2017
8. Connection to Brain cancers (TCGA)	Gerstein Lab	Fall 2017
9. HiC data analysis-	Abyzov Lab	
10. Splicing/transcription eQTL	Jaffe Lab	

Capstone 4 - Datasets

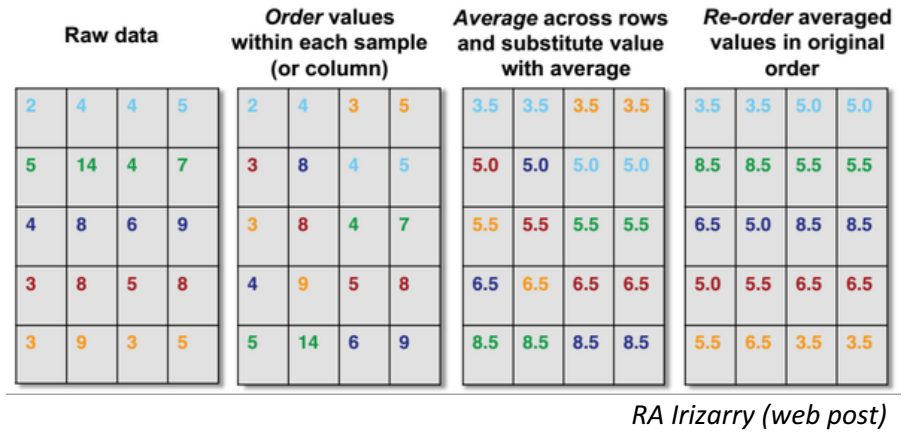
Study name	Individuals	Brain region	Disease status	Data type
UIC, Uchicago-BrainGVEX	428	Frontal cortex	SCZ, BD, ASD, control	RNA-seq (waiting for genotype data of half samples)
UCLA-ASD	73/96	PreFrontal cortex, CBC	ASD, control	RNA-seq , ChIP-seq, genotype data
Yale-ASD	68	DLPFC	ASD, control	RNA-seq, ChIP-seq (waiting for genotype data)
CommonMind v1.2	603	DLPFC	SCZ, BD, control	RNA-seq, genotype data (Illumina Infinium HumanOmniExpressExome 8 v 1.1b chip)
GTEx V6p (BA9)	92	DLPFC	control	RNA-seq, genotype data (Illumina OMNI 5M or 2.5M array)
Brainspan	43	16 regions	control	RNA-seq, genotype data
ENCODE/Roadmap		Different regions	control	RNA-seq , ChIP-seq
Total	1323(currently 754 with genotype) for DLPFC data			

Capstone 4 -Analysis flowchart



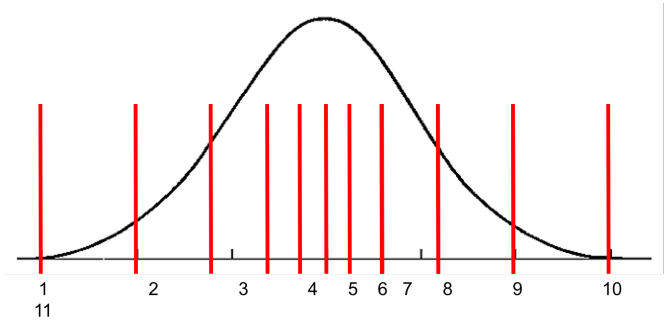
Conglomerated RPKM data files processed to generate RPKM counts for each tissue on relevant samples (GTEx only)

Quantile normalization



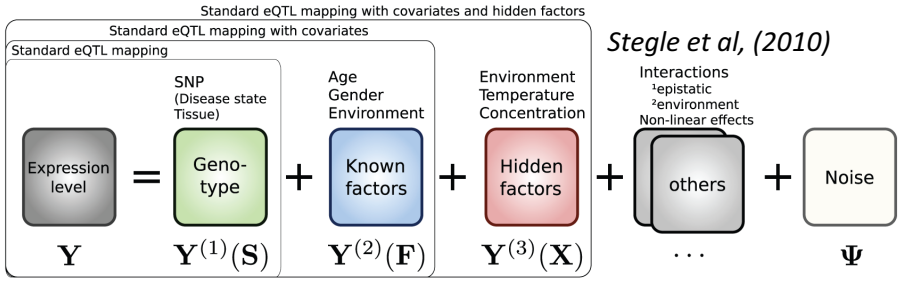
RA Irizarry (web post)

Inverse quantile normalization



PEER calculations on all datasets

Sums, differences, and original “residual” matrices



Stegle et al, (2010)

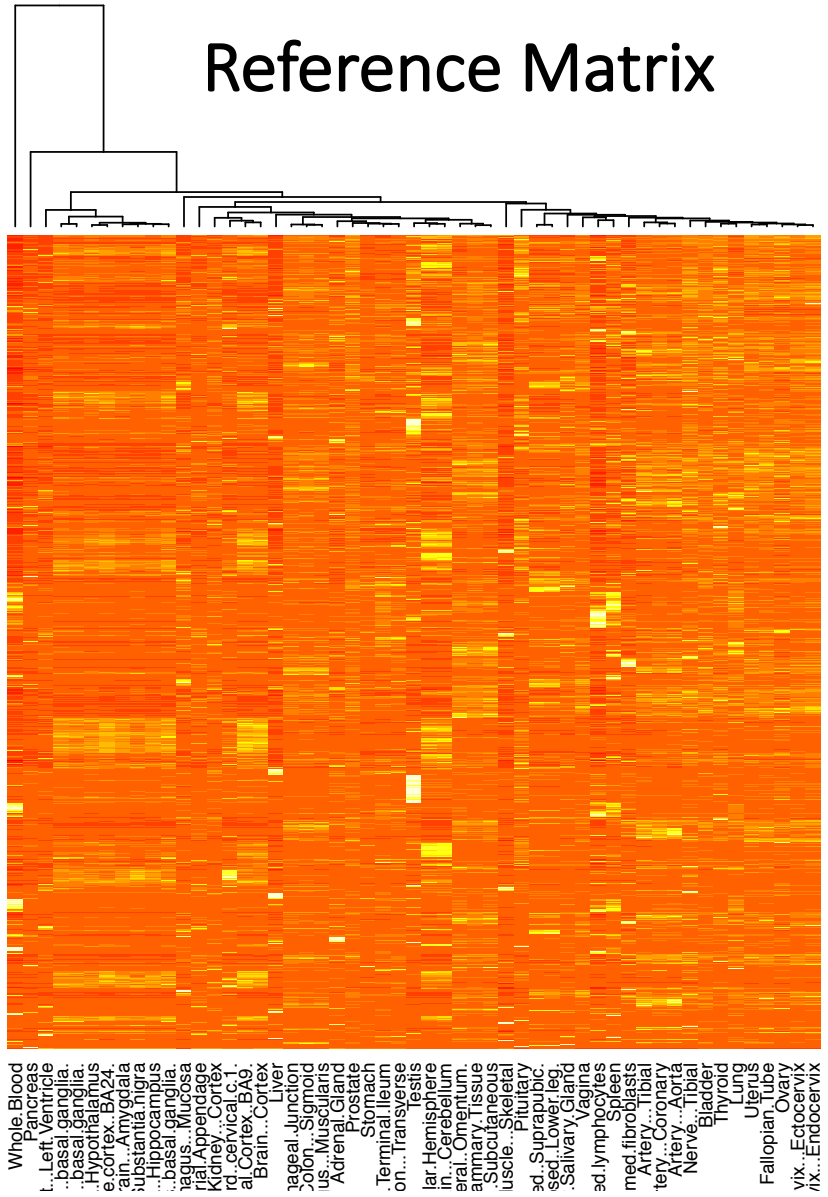
PEER-based processing on 50 GTEx tissues (including 15 brain tissues), plus:

- **BrainGVEX**
- **Brainspan**
- **CMC**
- **UCLA-ASD**
- **Yale-ASD**

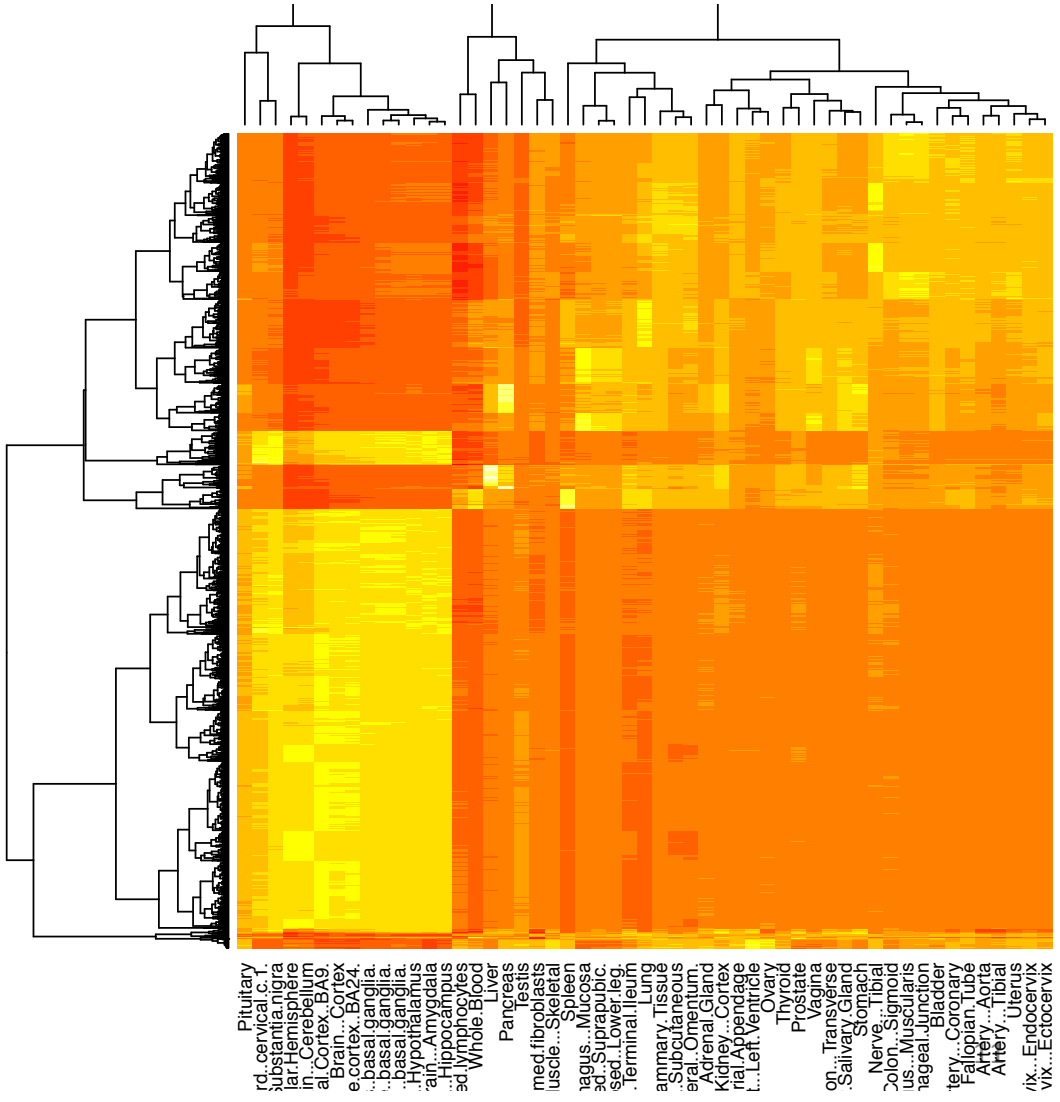
RCA strategy for different tissues

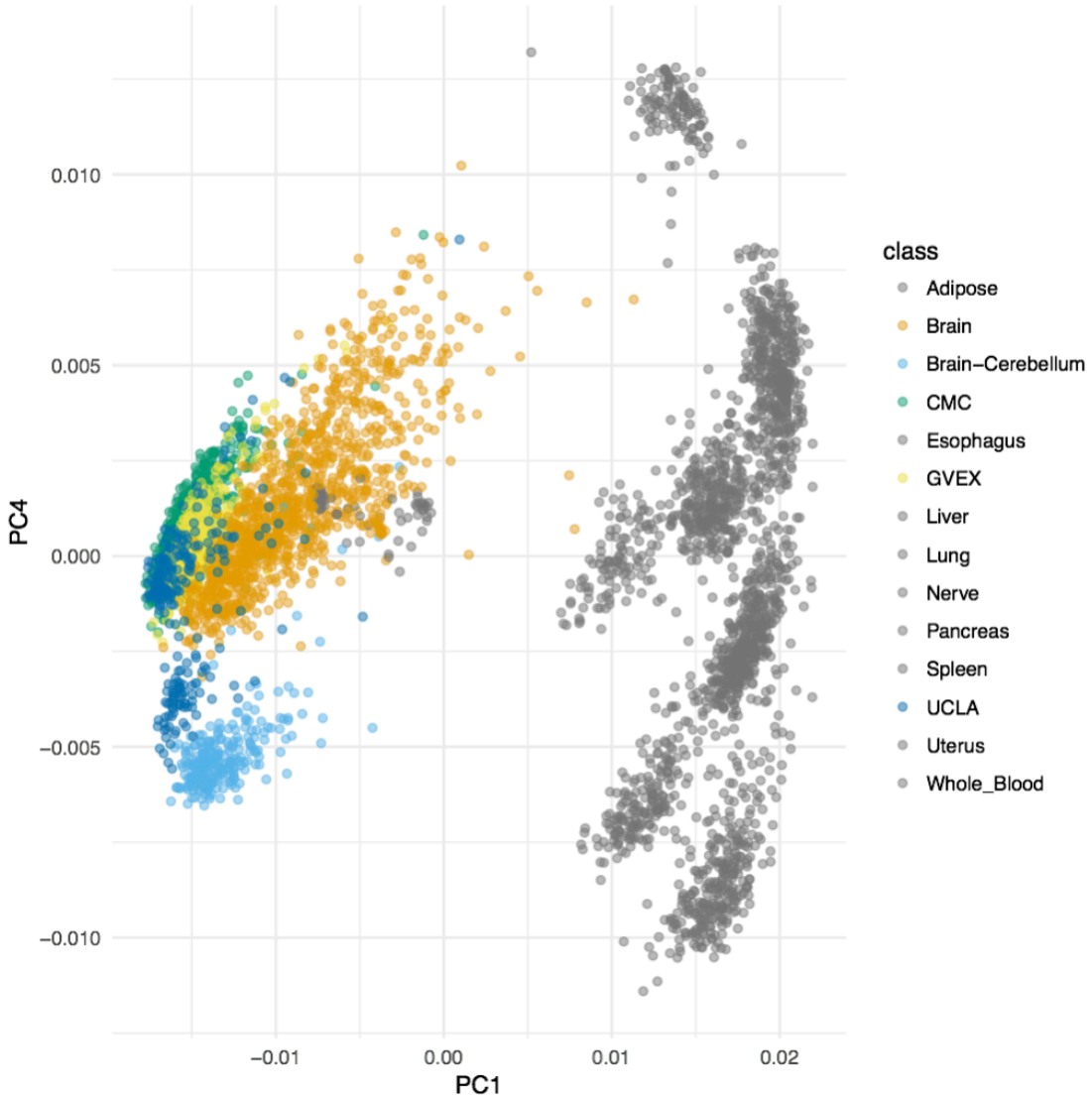
RCA = **R**eference **C**omponent **A**nalysis

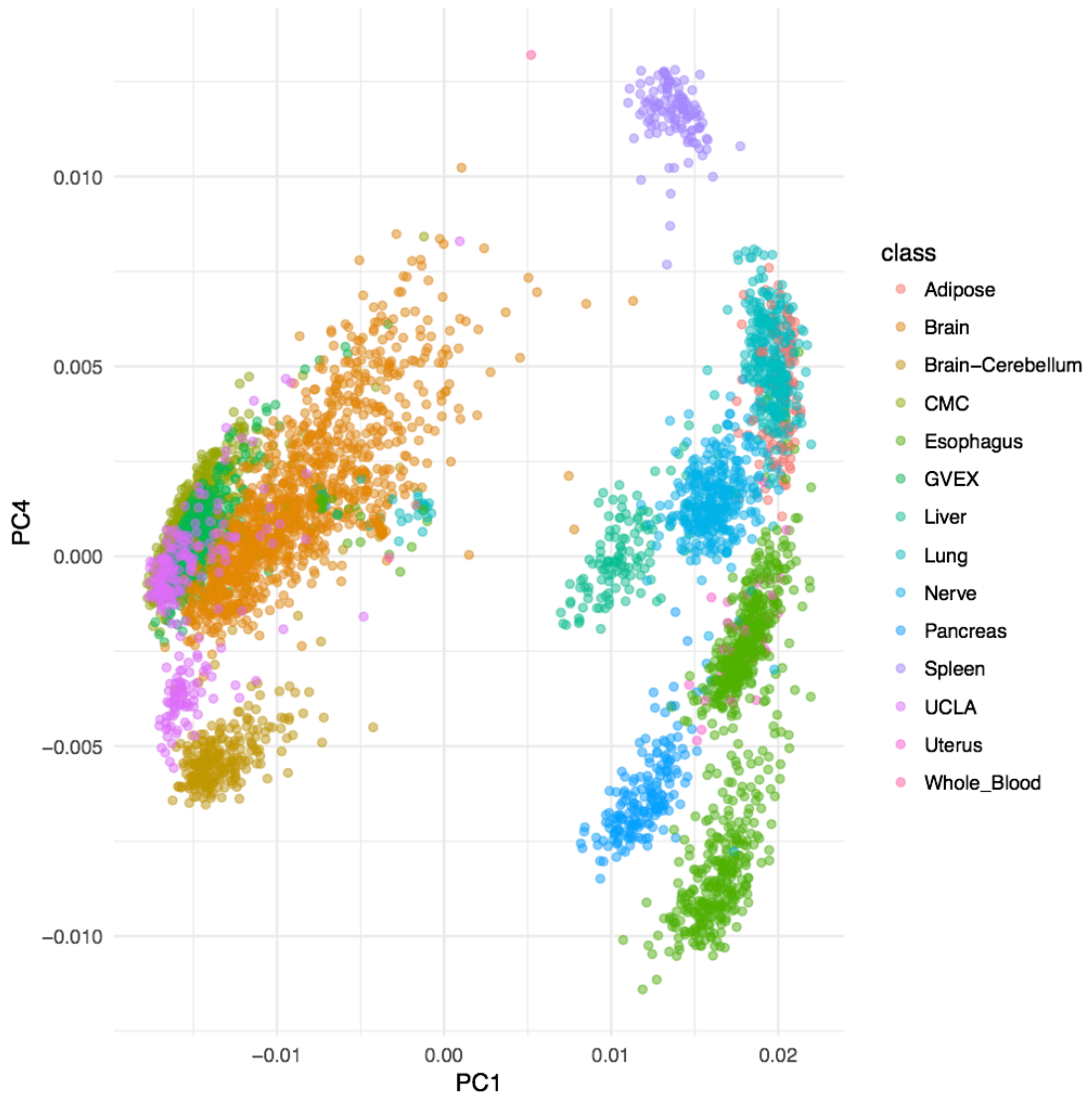
Reference Matrix

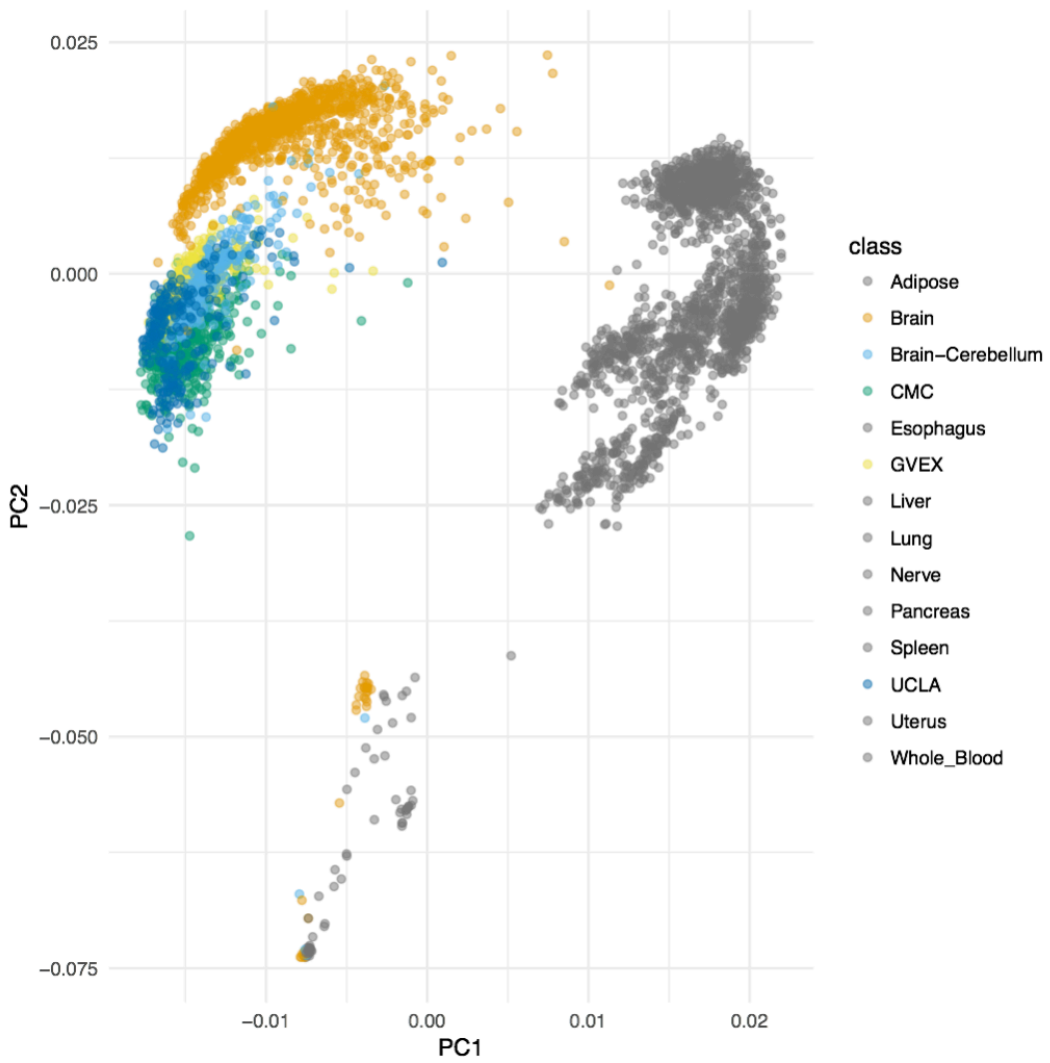


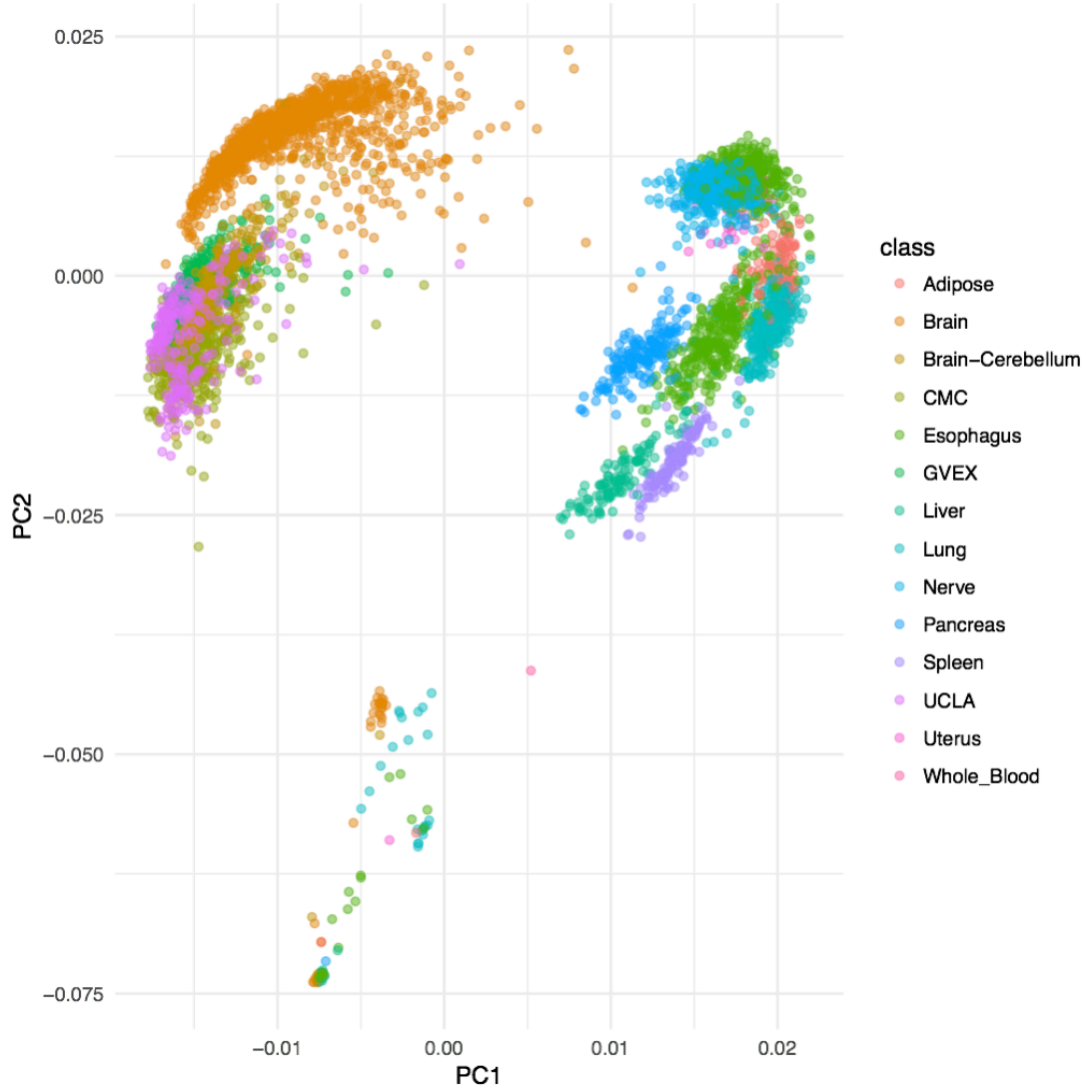
Correlation Matrix




0.010
0.005
0.000
−0.005
−0.010
PC4
−0.01
0.00
0.01
0.02
PC1
class
Adipose
Brain
Brain−Cerebellum
CMC
Esophagus
GVEX
Liver
Lung
Nerve
Pancreas
Spleen
UCLA
Uterus
Whole_Blood

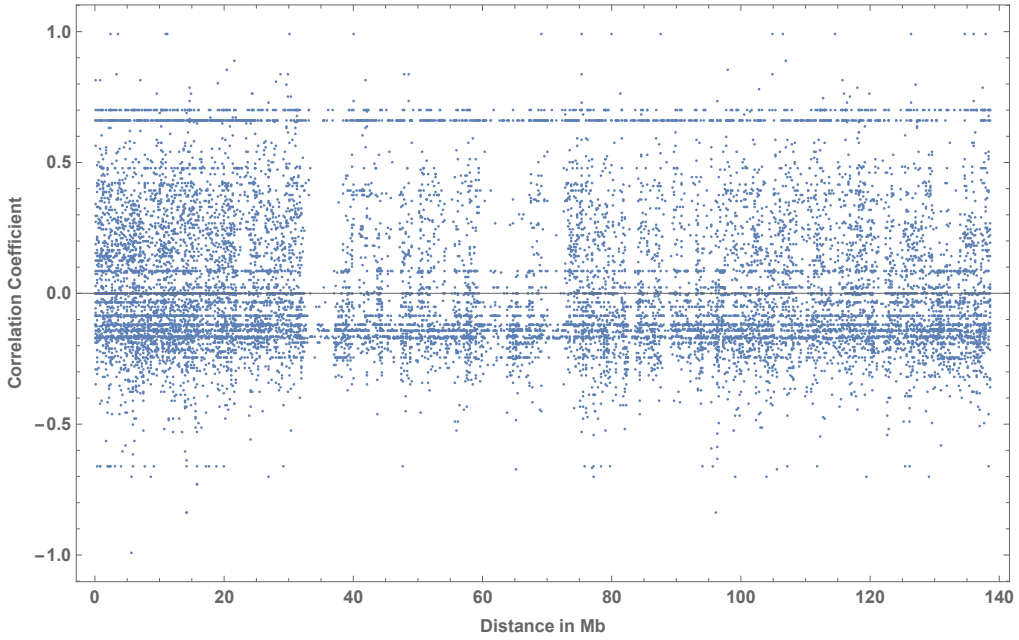

0.010
0.005
0.000
−0.005
−0.010
PC4
−0.01
0.00
0.01
0.02
PC1
class
Adipose
Brain
Brain−Cerebellum
CMC
Esophagus
GVEX
Liver
Lung
Nerve
Pancreas
Spleen
UCLA
Uterus
Whole_Blood

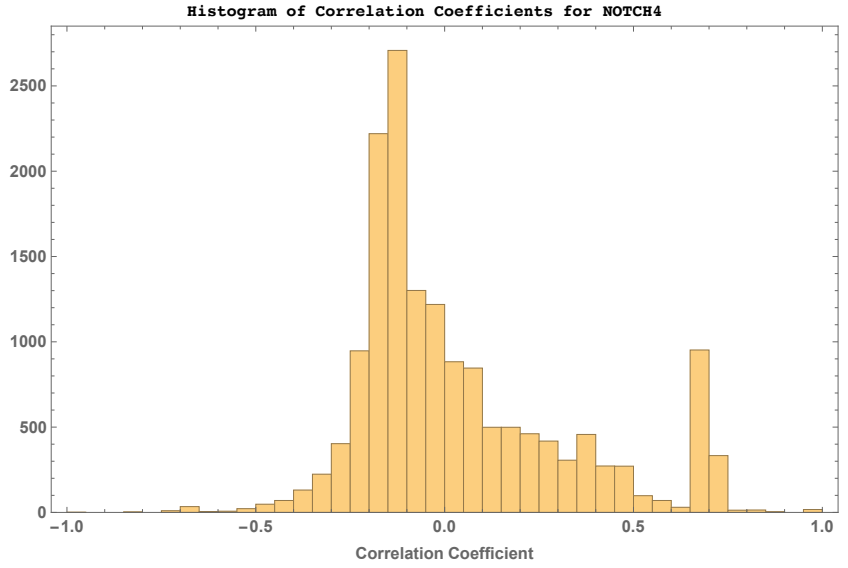

0.025
0.000
−0.025
−0.050
−0.075
PC2
−0.01
0.00
0.01
0.02
PC1
class
Adipose
Brain
Brain−Cerebellum
CMC
Esophagus
GVEX
Liver
Lung
Nerve
Pancreas
Spleen
UCLA
Uterus
Whole_Blood


0.025
0.000
−0.025
−0.050
−0.075
PC2
−0.01
0.00
0.01
0.02
PC1
class
Adipose
Brain
Brain−Cerebellum
CMC
Esophagus
GVEX
Liver
Lung
Nerve
Pancreas
Spleen
UCLA
Uterus
Whole_Blood

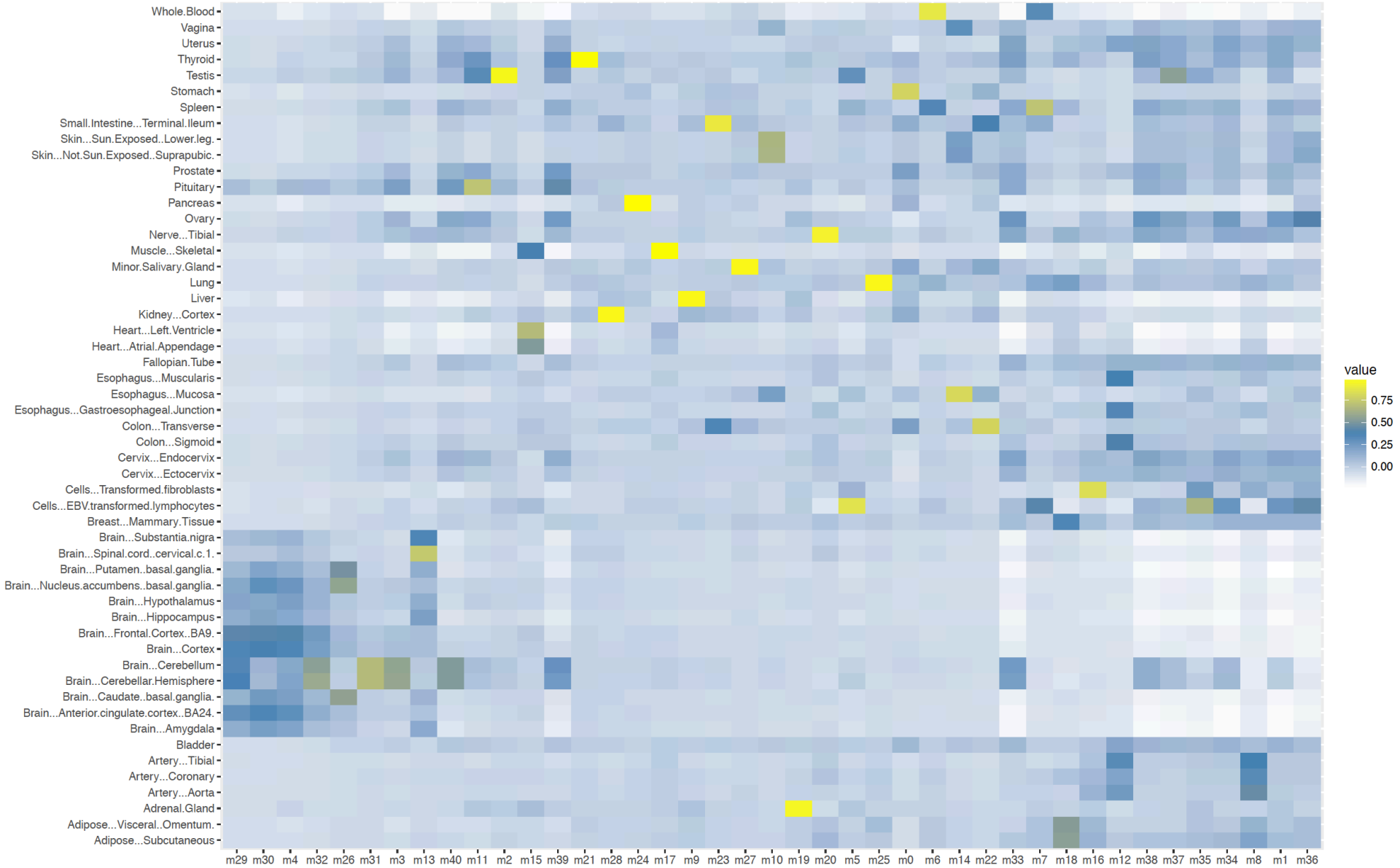
Enhancer and gene expression pattern correlation

Patterns for Notch4 in Chr. 6

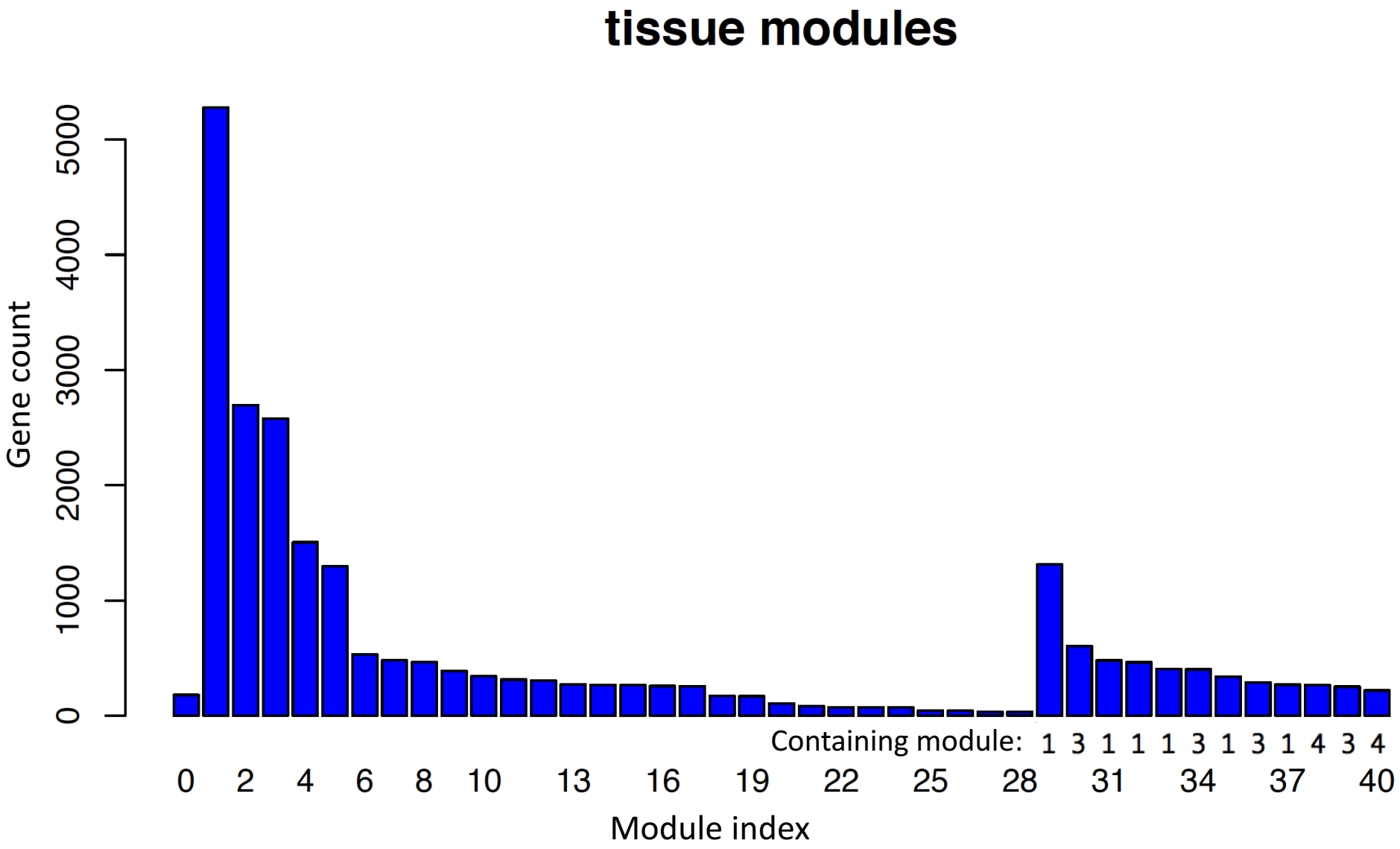




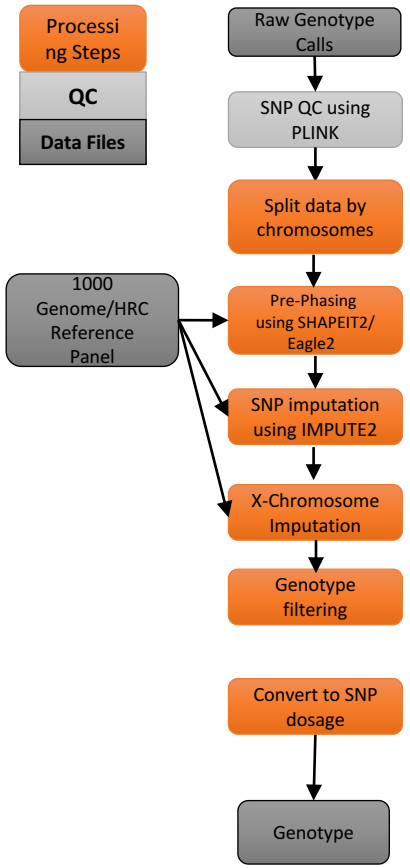
Module eigengenes ordered by brain region correlation

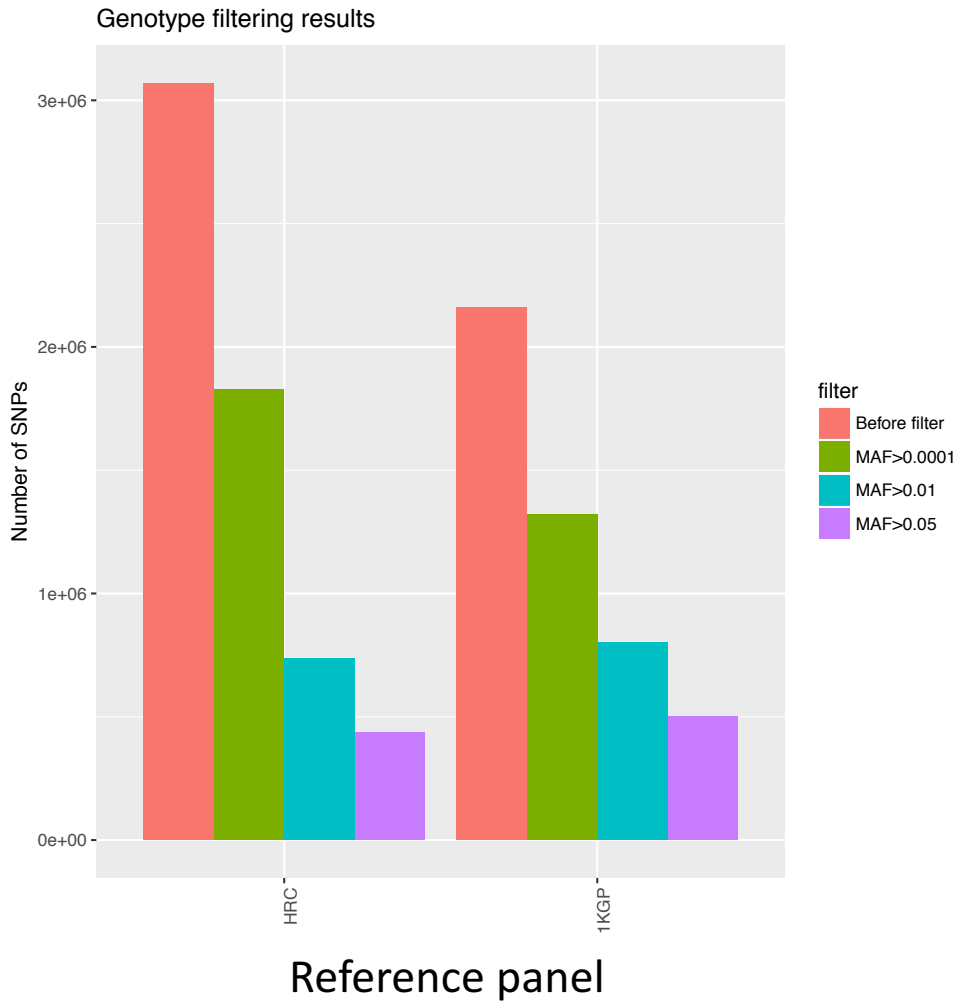


Module/Sub-module sizes in GTEx

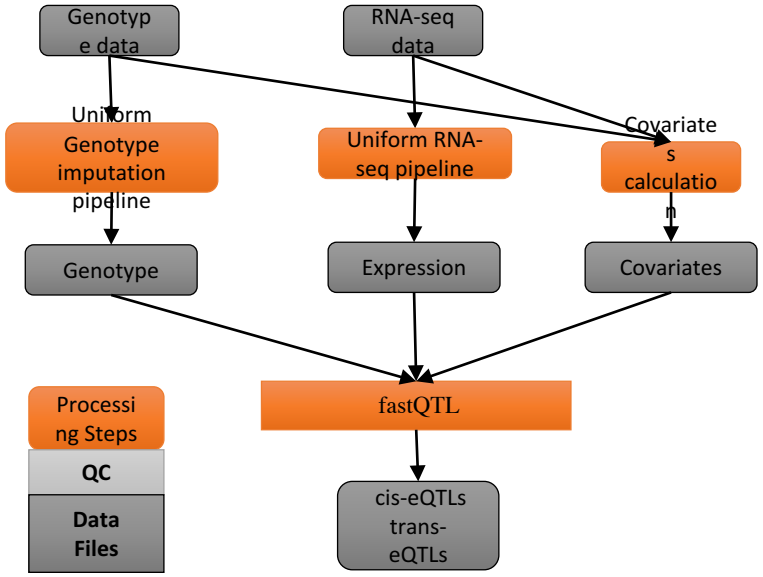


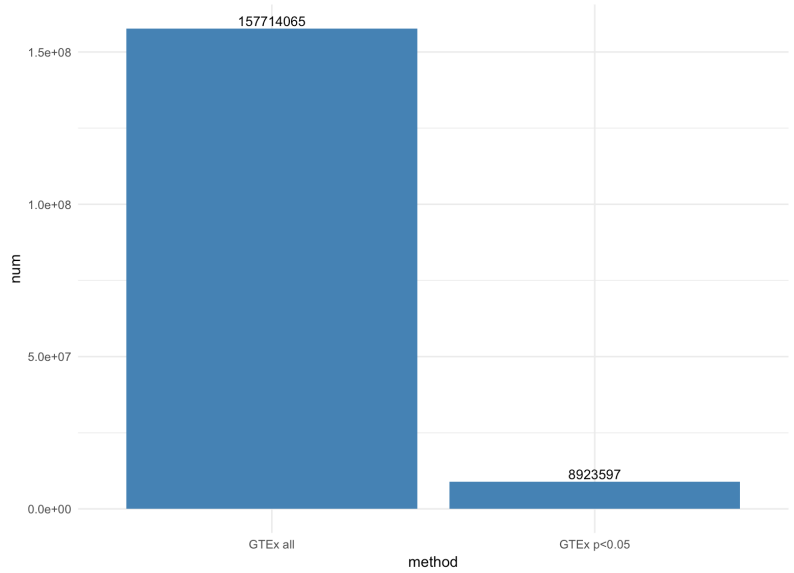
Genotype Imputation





eQTL analysis





Expression

Genes were selected based on expression thresholds of >0.1 RPKM in at least 10 individuals and ≥6 reads in at least 10 individuals.

Quantile normalized across samples.

For each gene, expression values were inverse quantile normalized to a standard normal distribution across samples.

Genotypes

Variants were imputed using HRC.

genotype filters applied:

- Call Rate Threshold 95%.
- R2 Threshold 0.6.
- MAF>= 1%

eQTL Analysis using FastQTL

Mapping window --1 megabase from TSS

To do list

- Merge GTEx, CommonMind and PsychENCODE data for eQTL calculation
- Identify significant eQTL
- Single-cell deconvolution

Paper structure

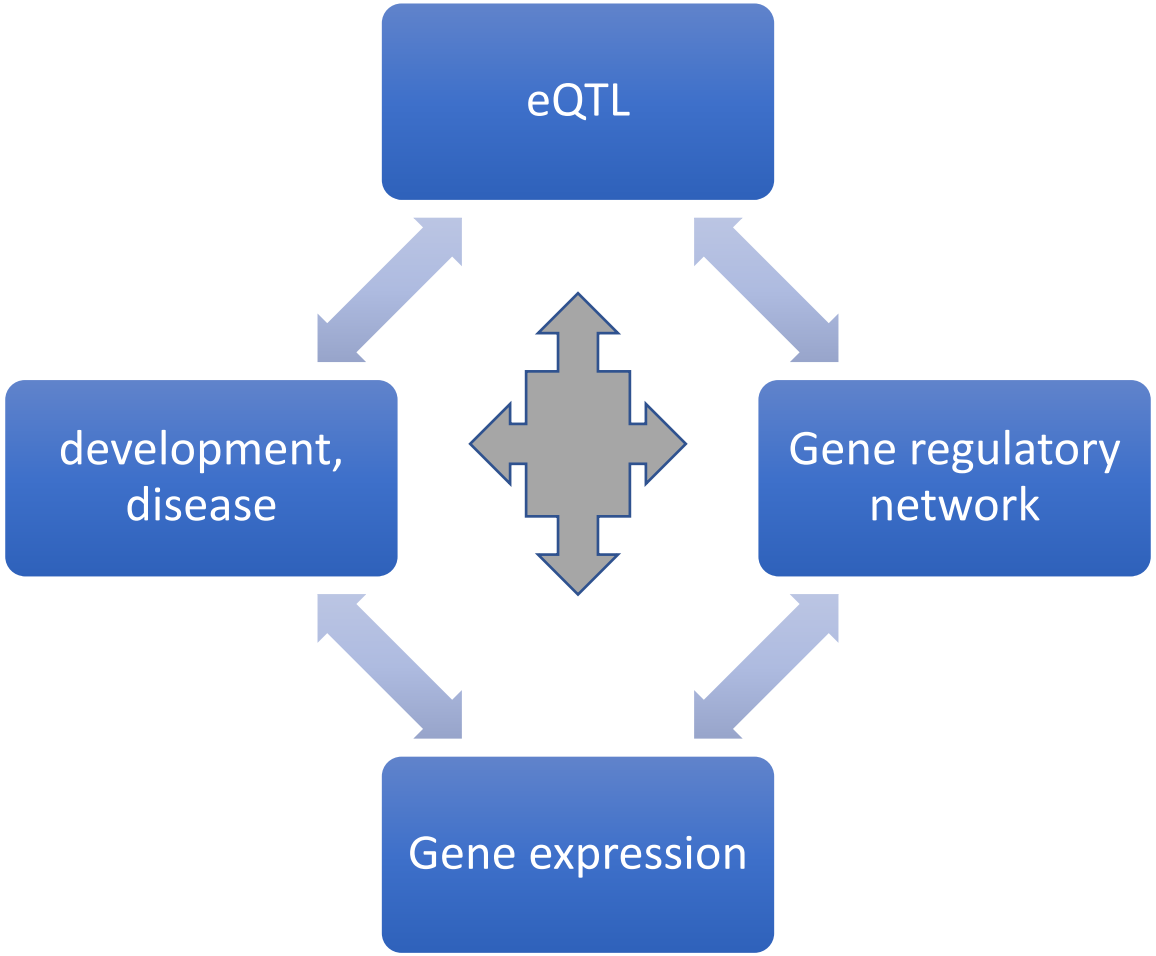
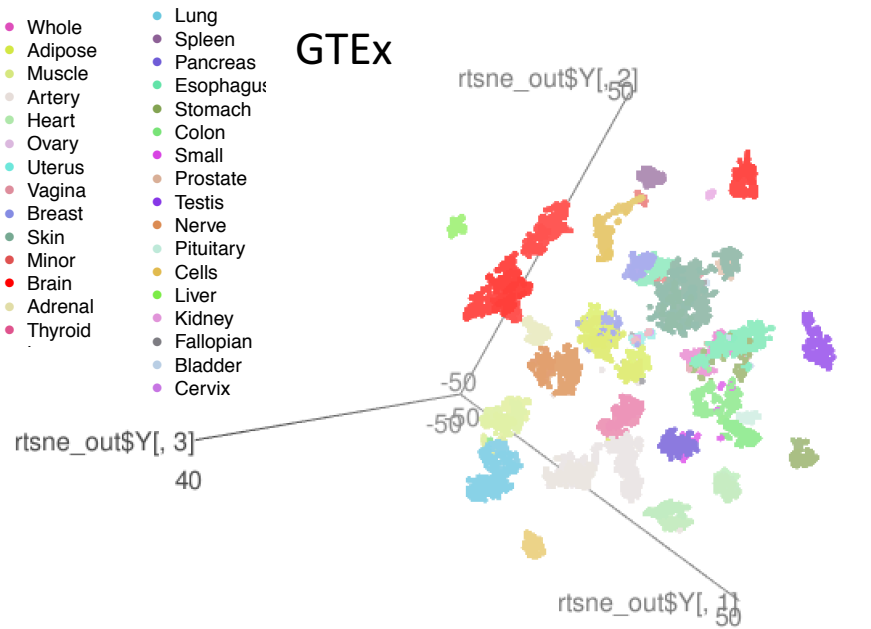
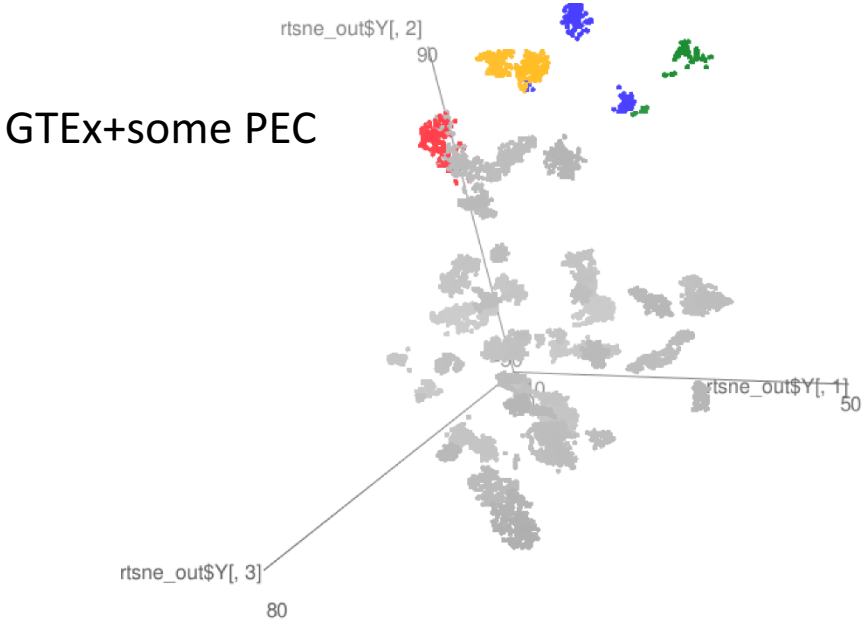


Figure 1 Brain vs. other tissues





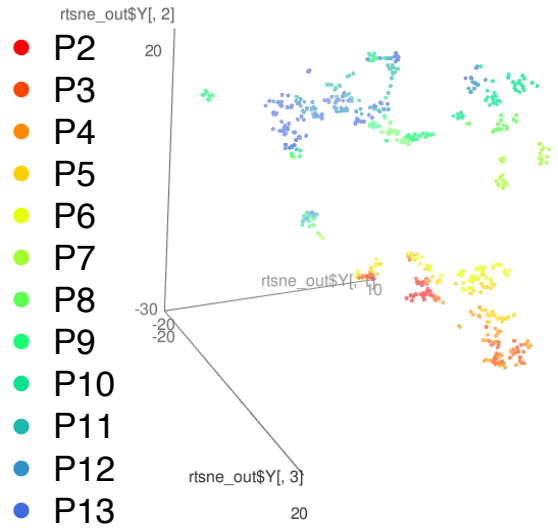
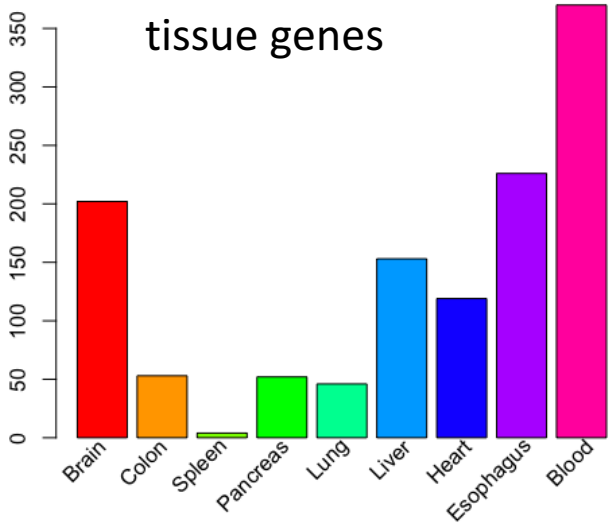


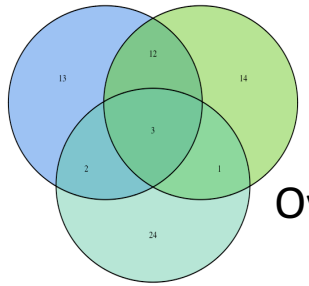
Figure 2 Brain and mental disease genes

A. Brain and other tissue genes



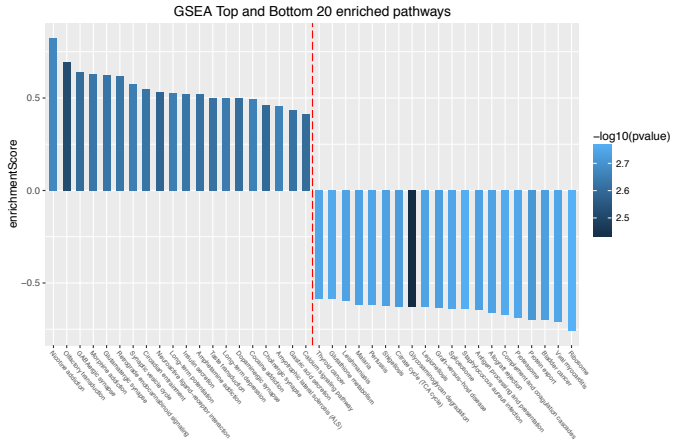
B. Mental disease genes

BP	AFF	SCZ	CMC_control	GTEx
47	8	275	291	222

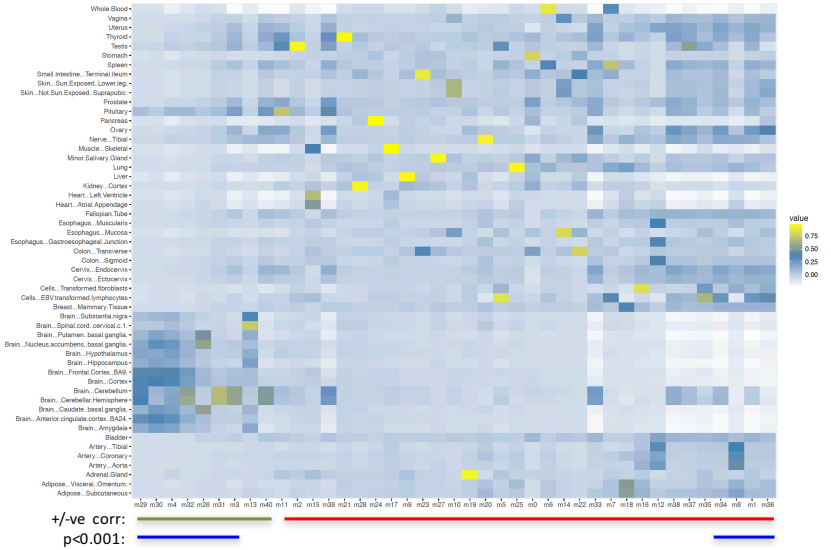


Overlap

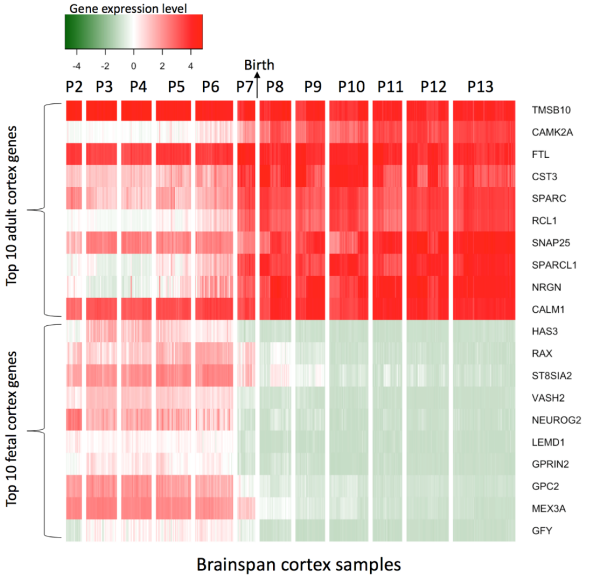
C. Gene functions



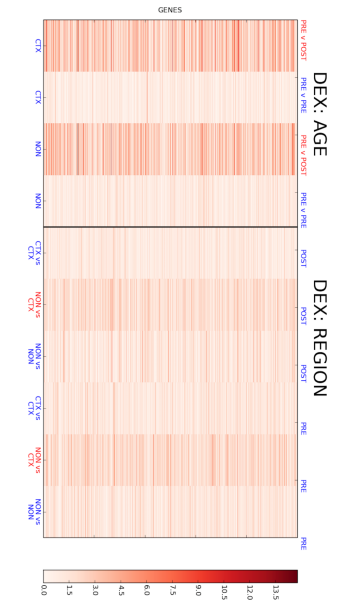
D. Brain gene expression patterns across tissues



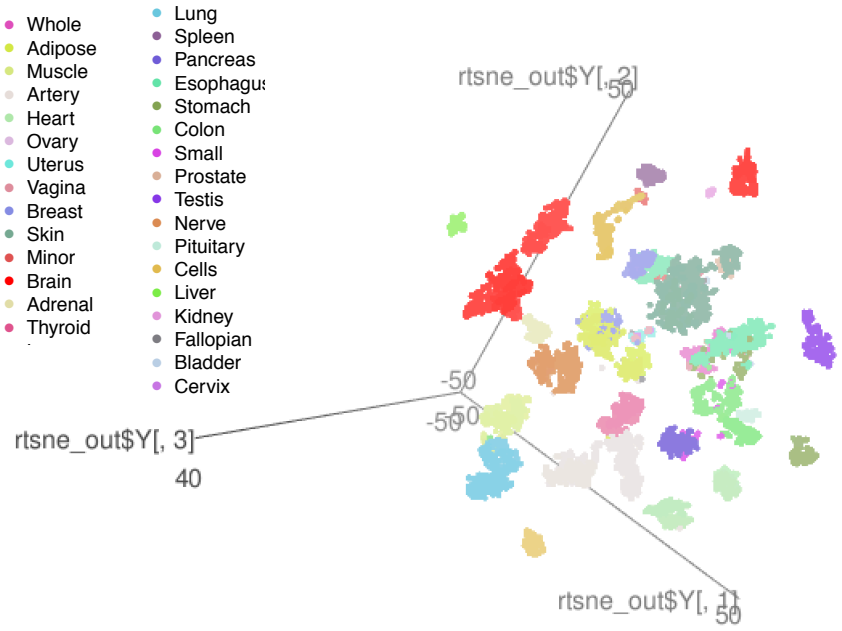
E. Brain gene expression pattern across developmental stages



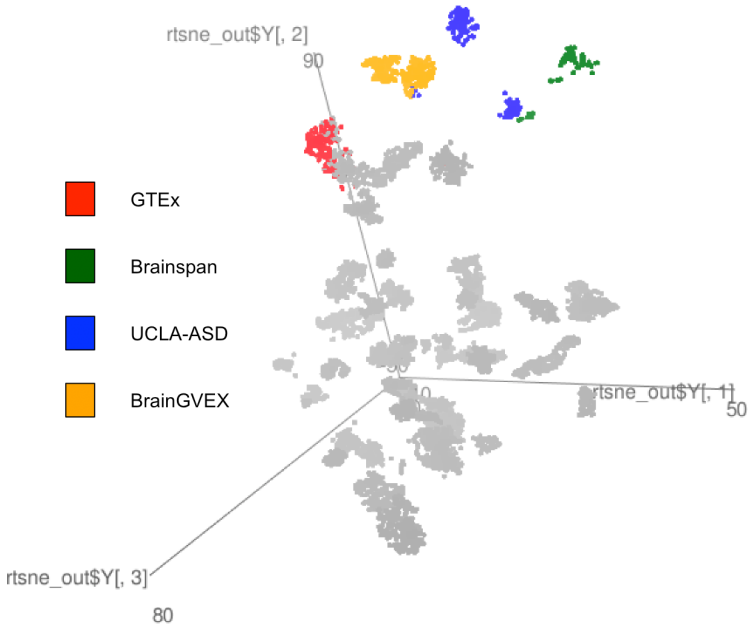
F. other features



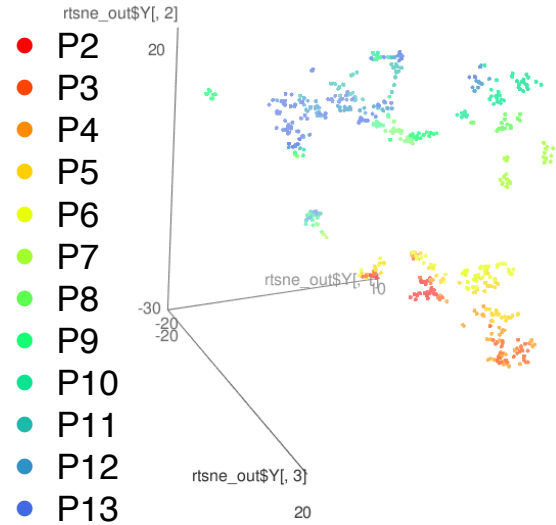
Brain sample clusters by tSNE, potentially driven by brain specific gene expression



GTEx



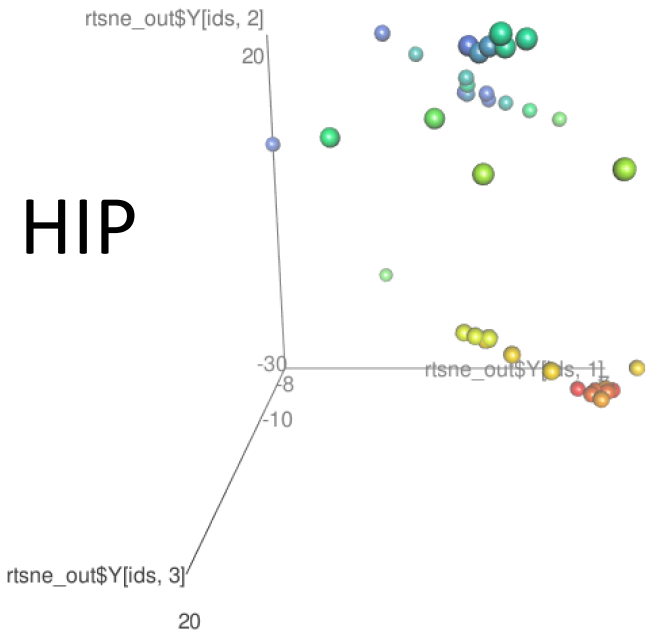
GTEx+some PEC



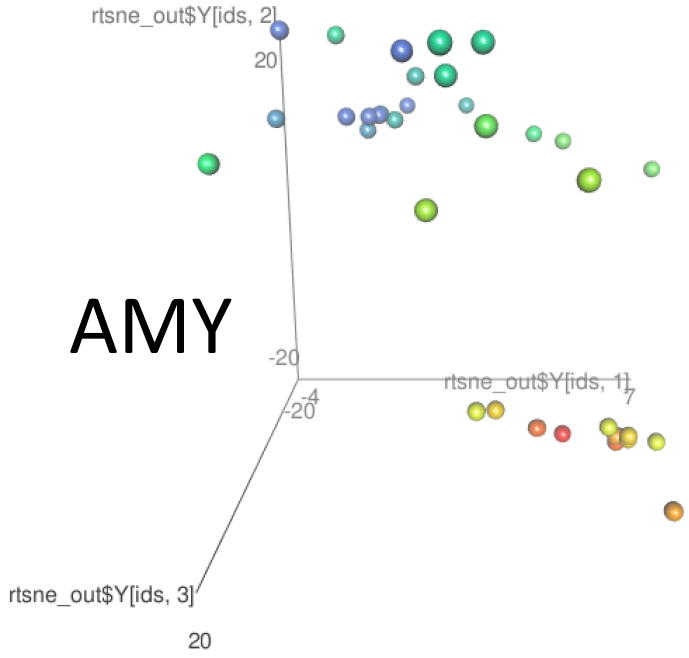
Brainspan

Brainspan samples by region by 3D-tSNE

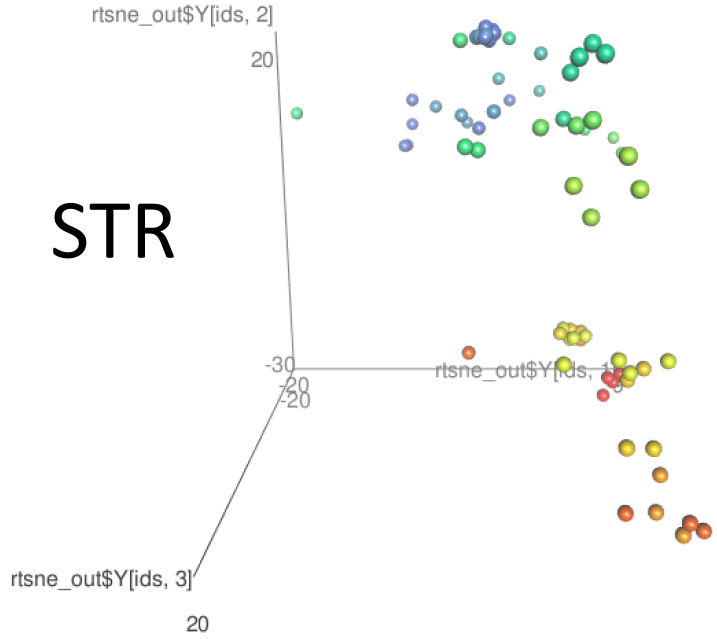
HIP



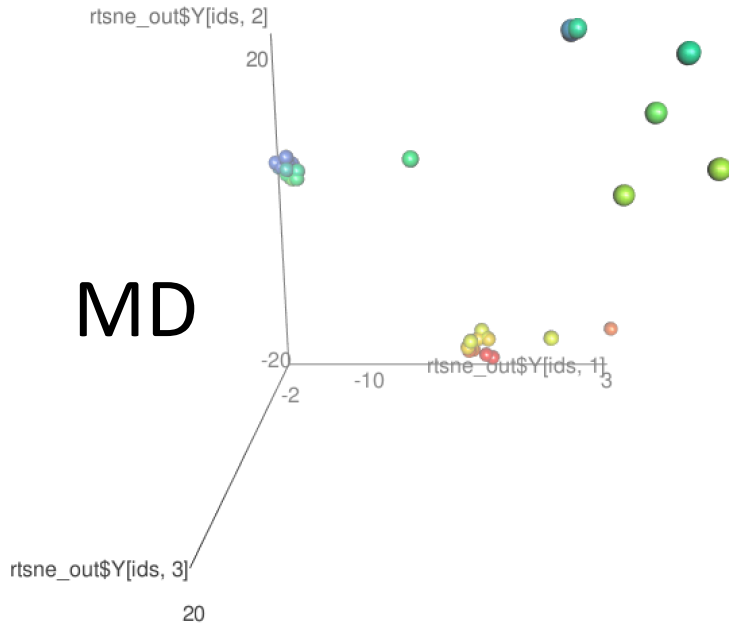
AMY



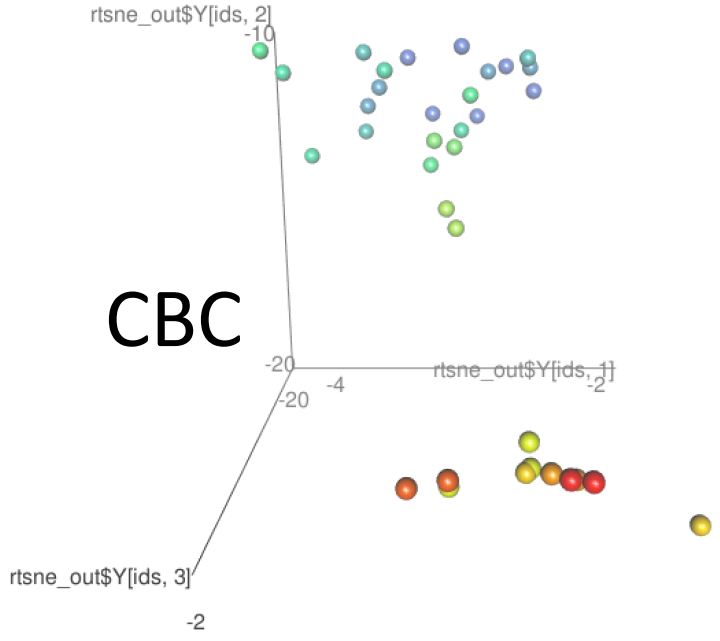
STR



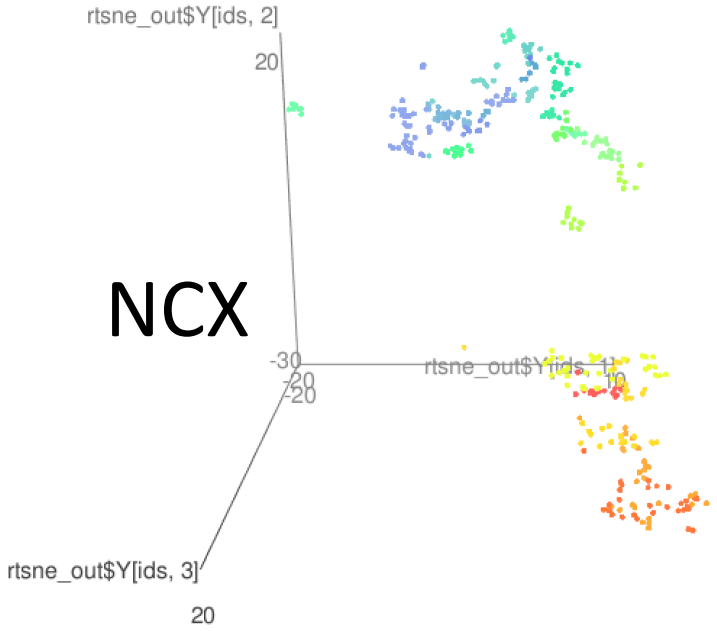
MD



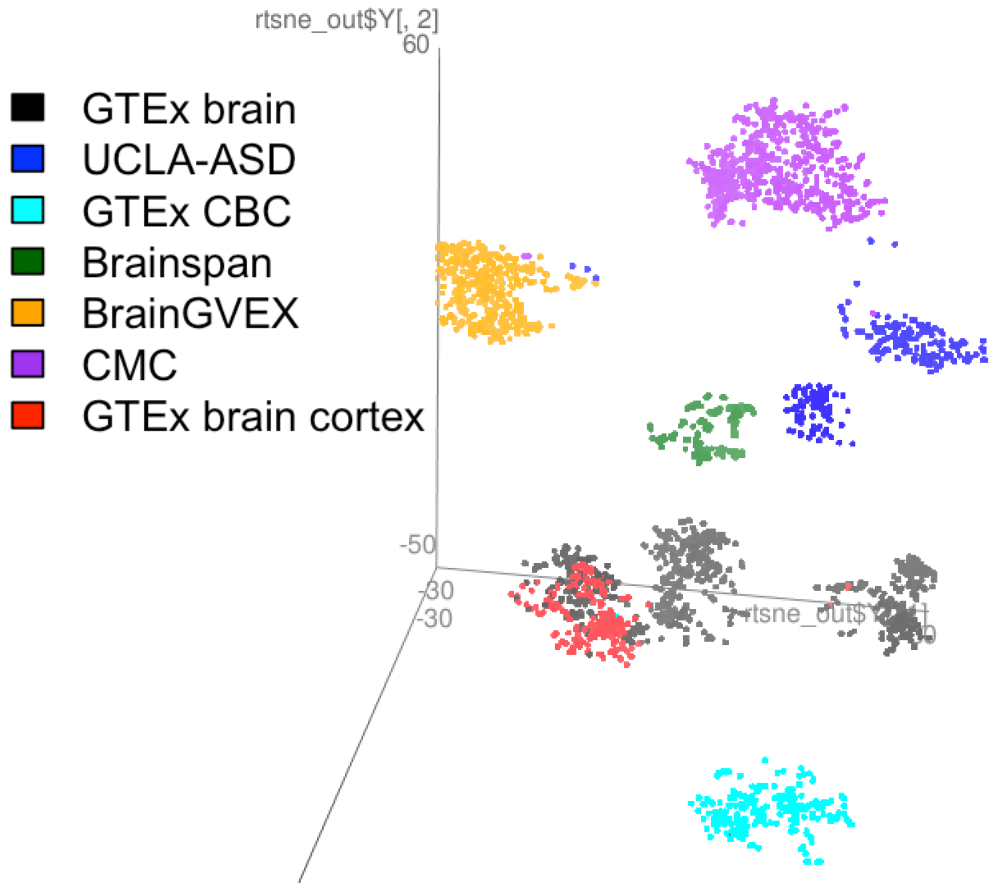
CBC

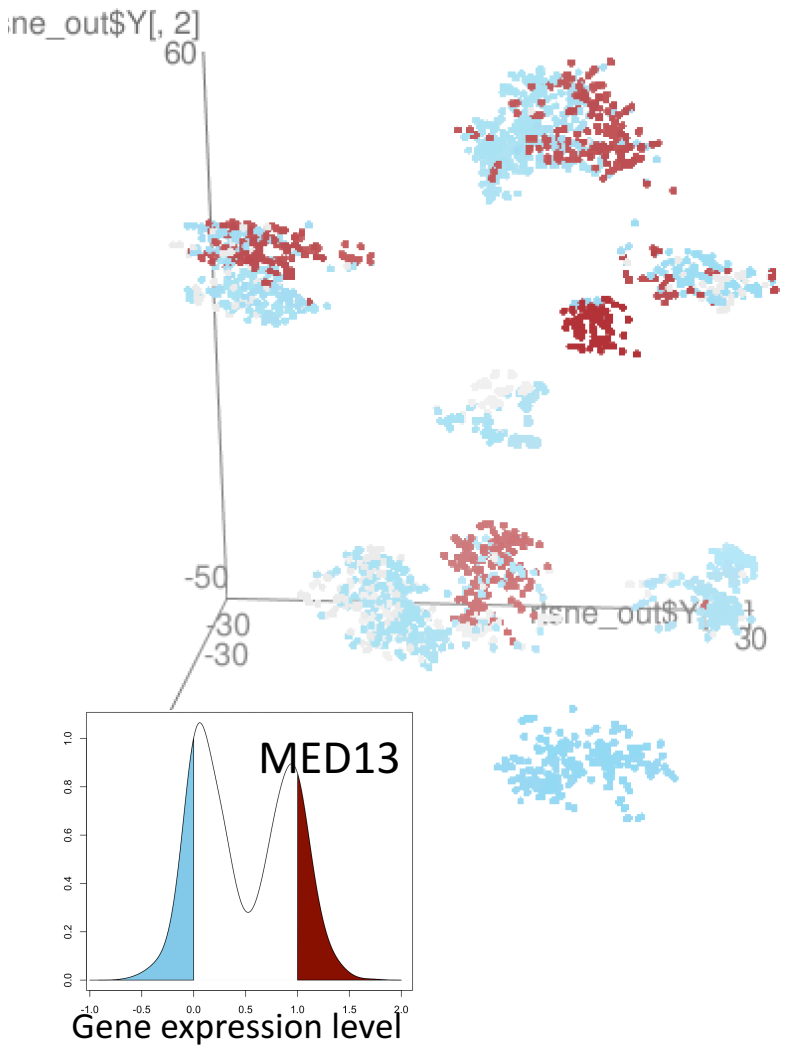


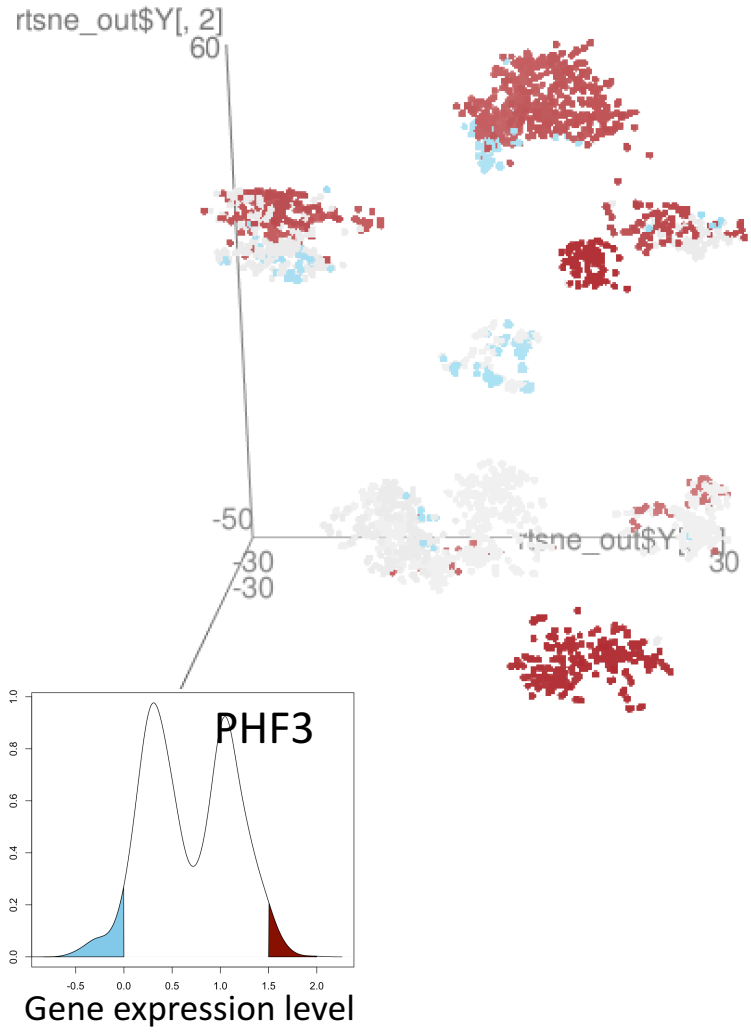
NCX



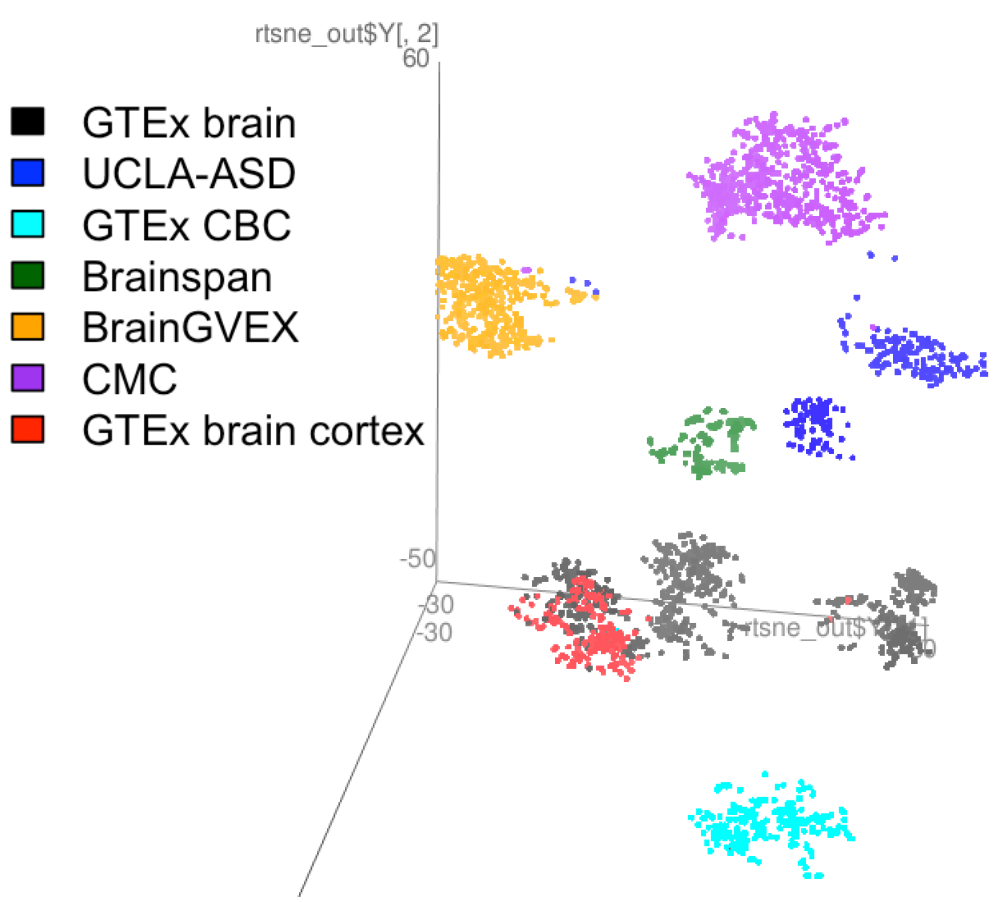
Example: Intellectual disability (Autism) gene expression distribution over tSNE

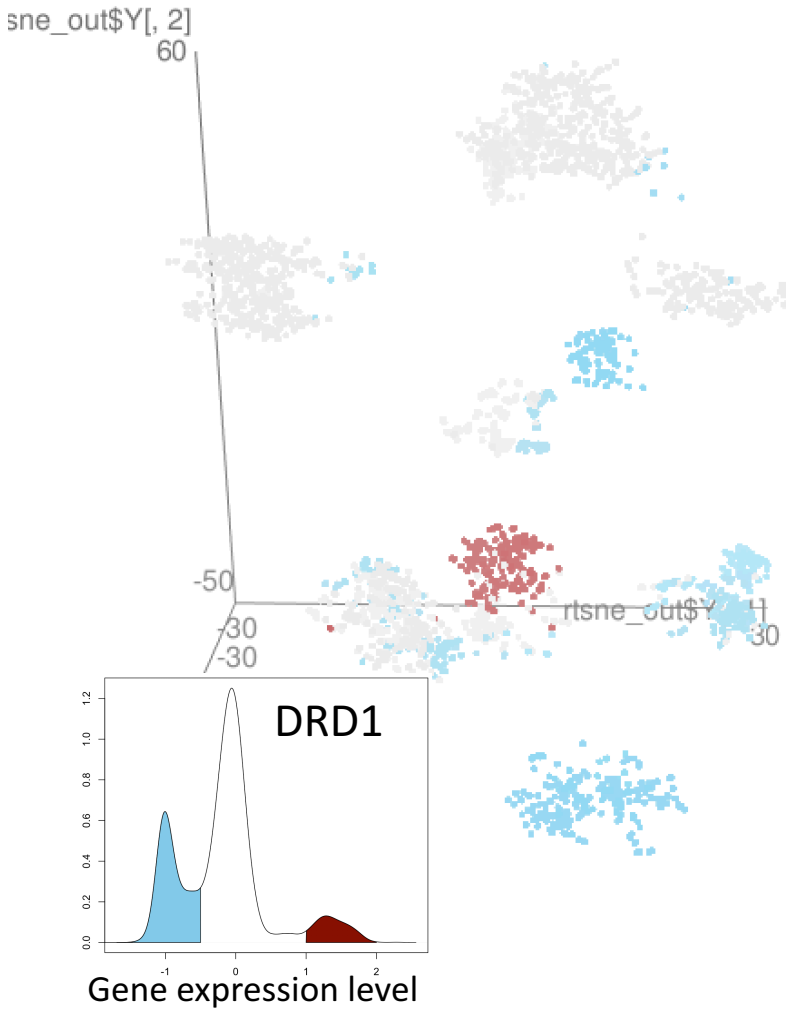


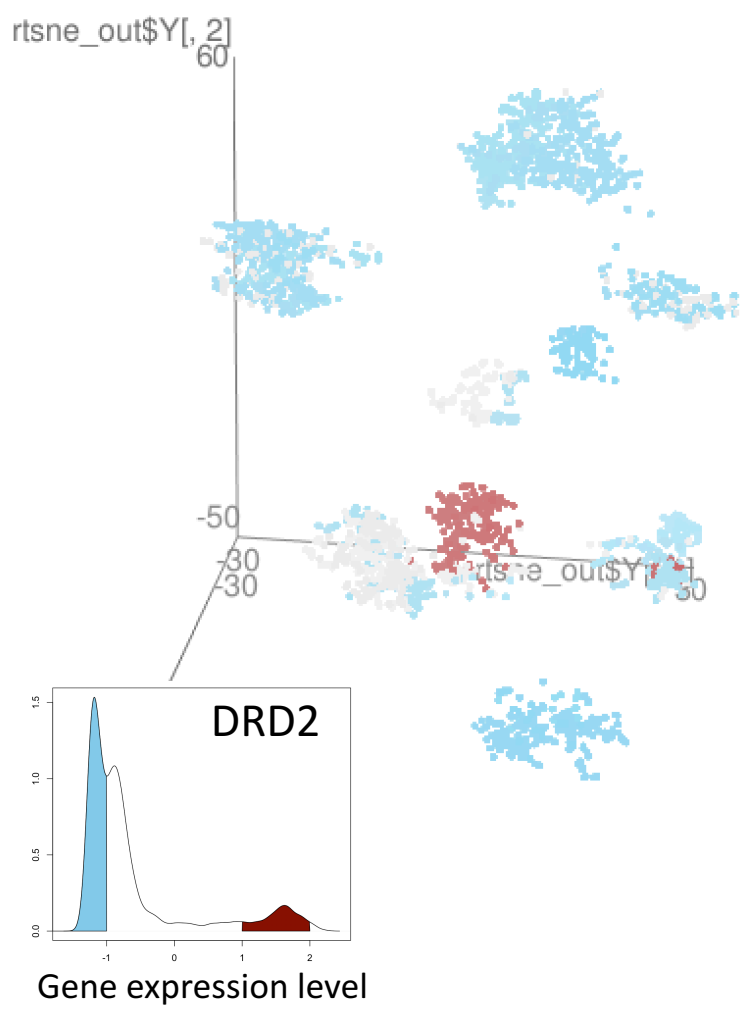




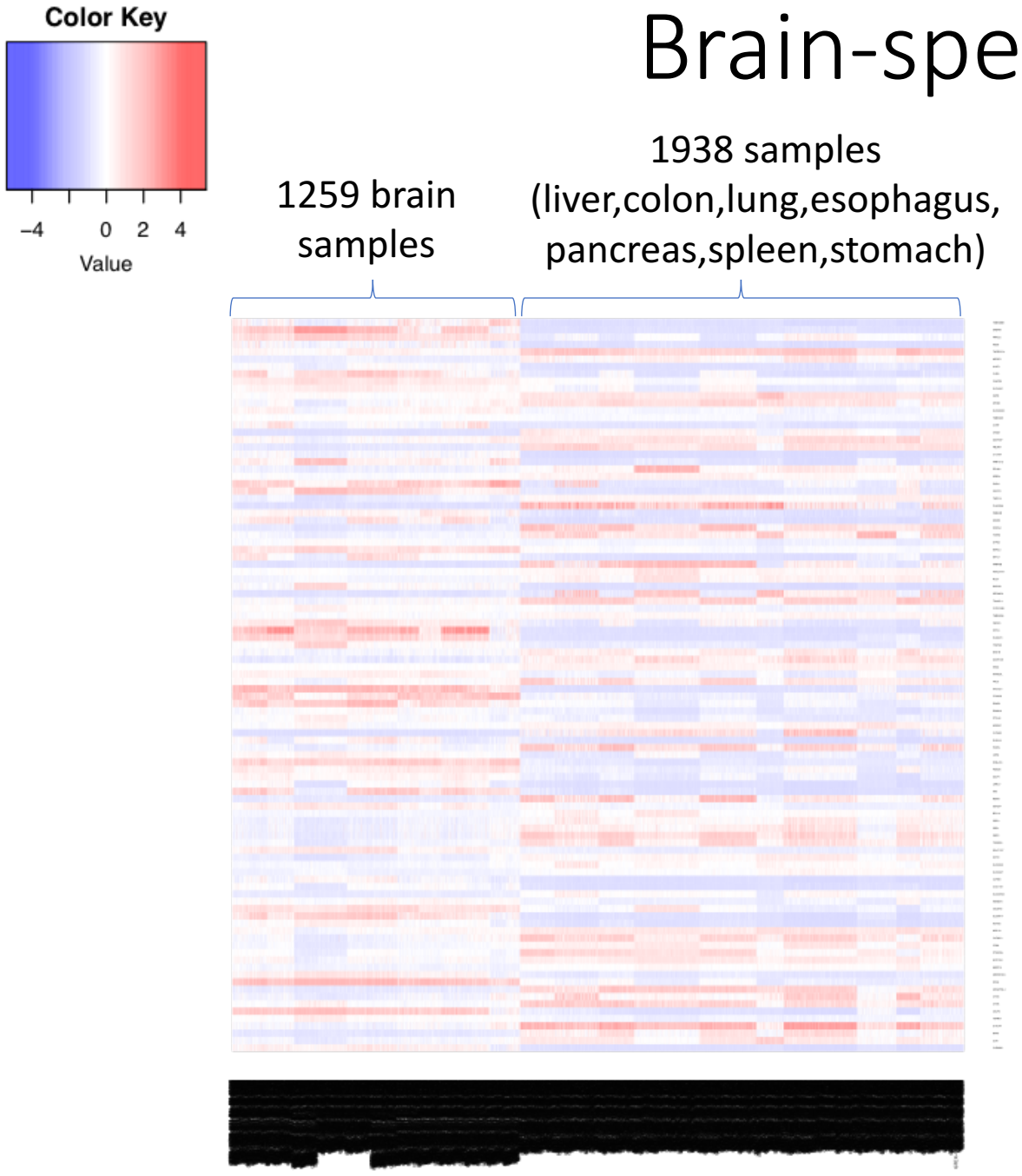
Example: Dopaminergic gene expression distribution over tSNE







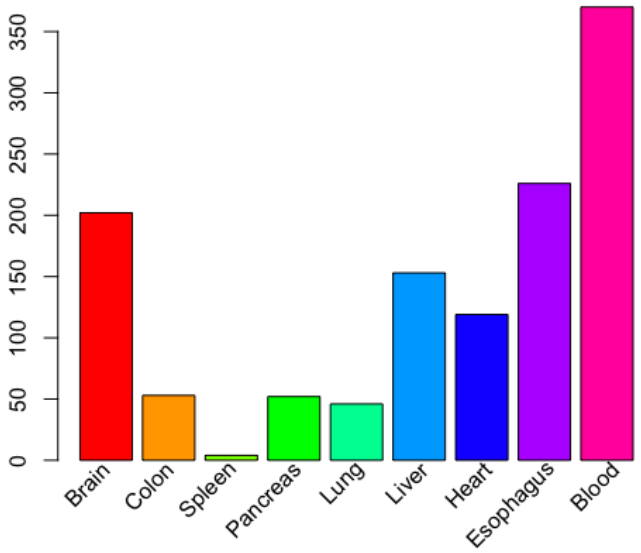
Brain-specific expressed genes



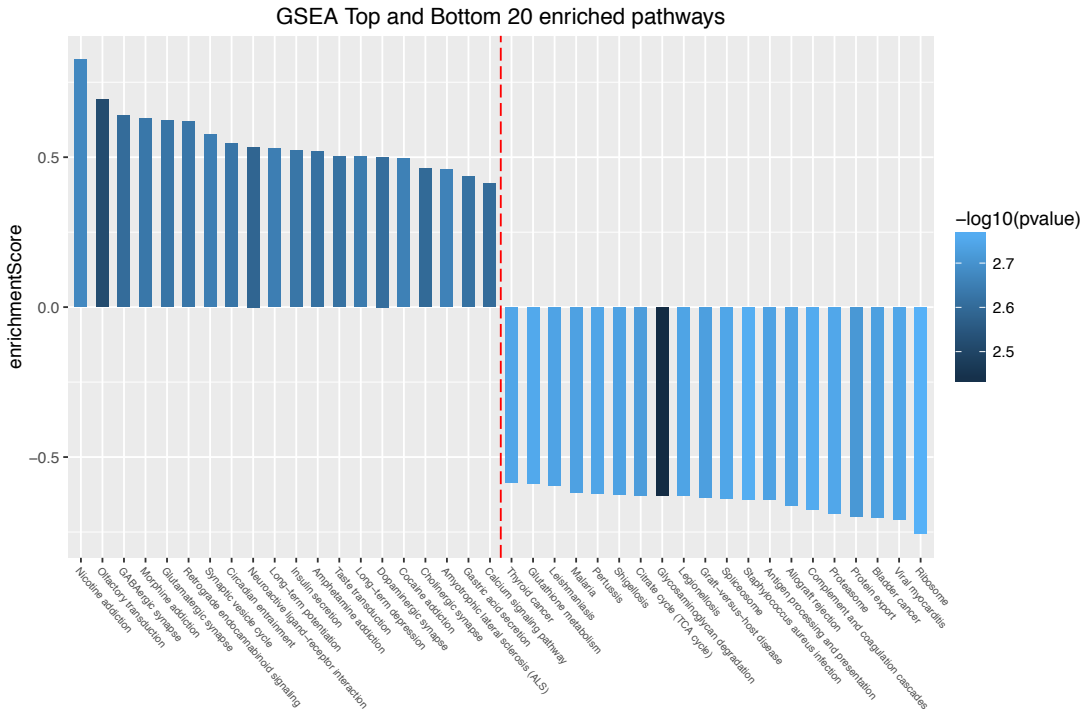
DEX brain genes (vs. other cons. tissues) in GTEx V6

Brain-specific expressed genes

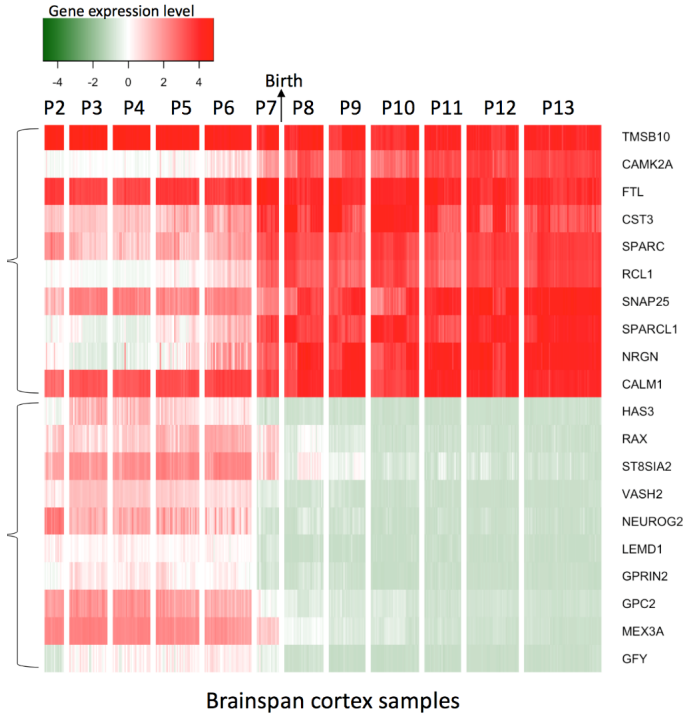
Brain and other tissue DEX genes



Enriched pathways and functions of Brain genes

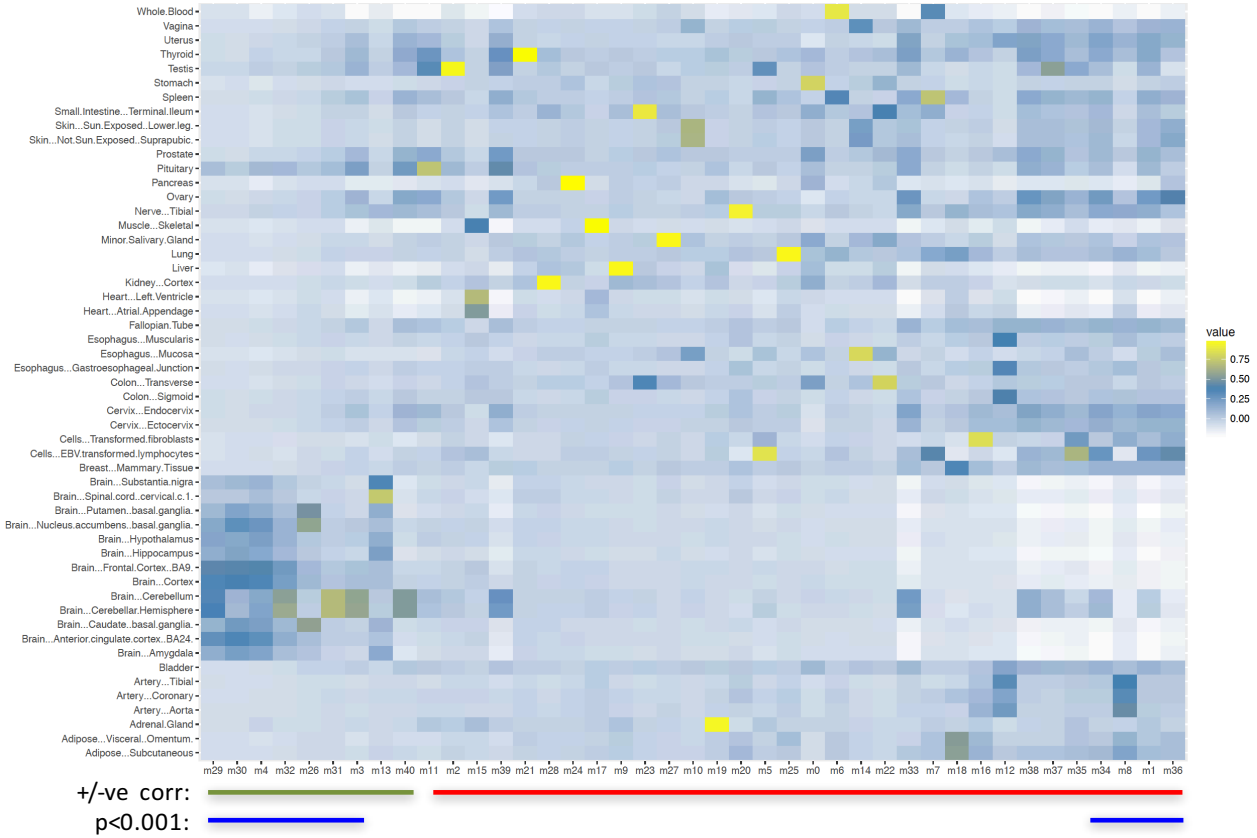


Top brain genes have specific developmental expression dynamics

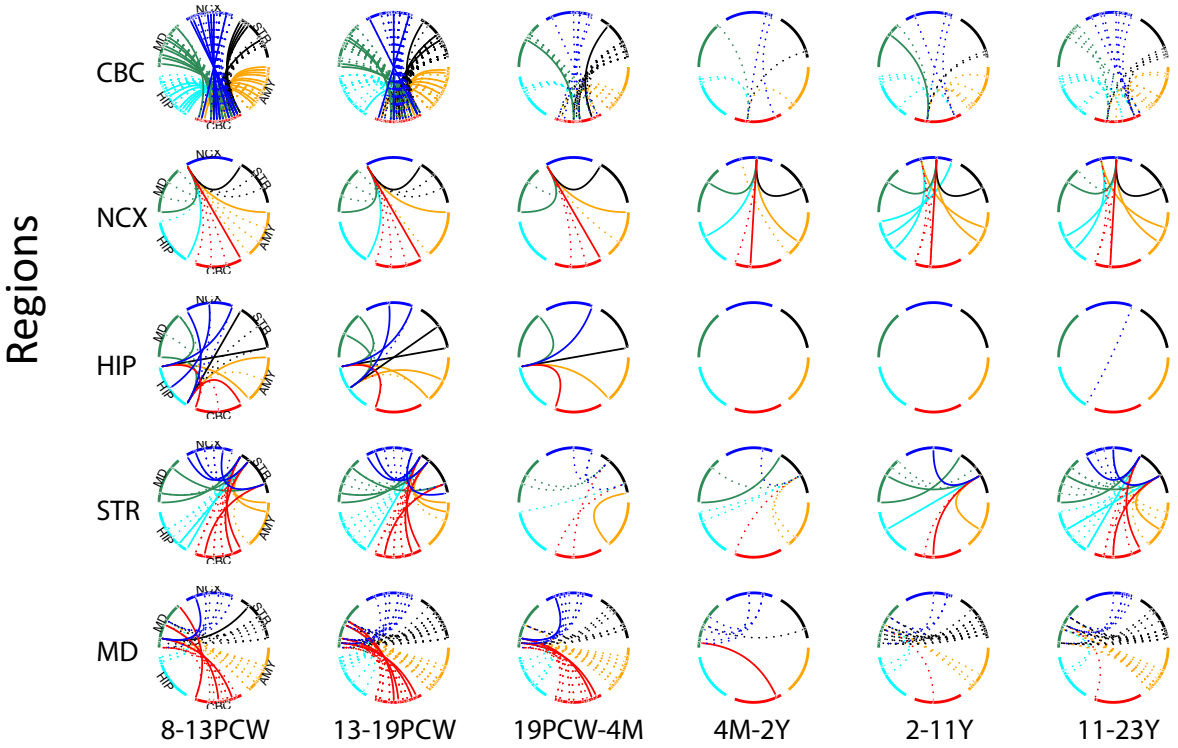


Gene co-expression modules by WGCNA

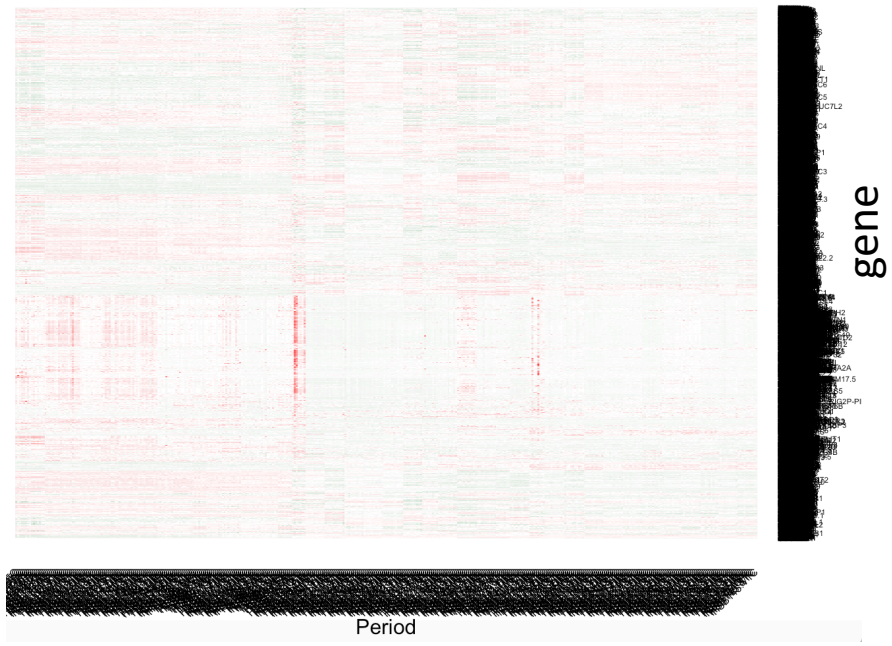
Co-expression across gtex tissues (JW)



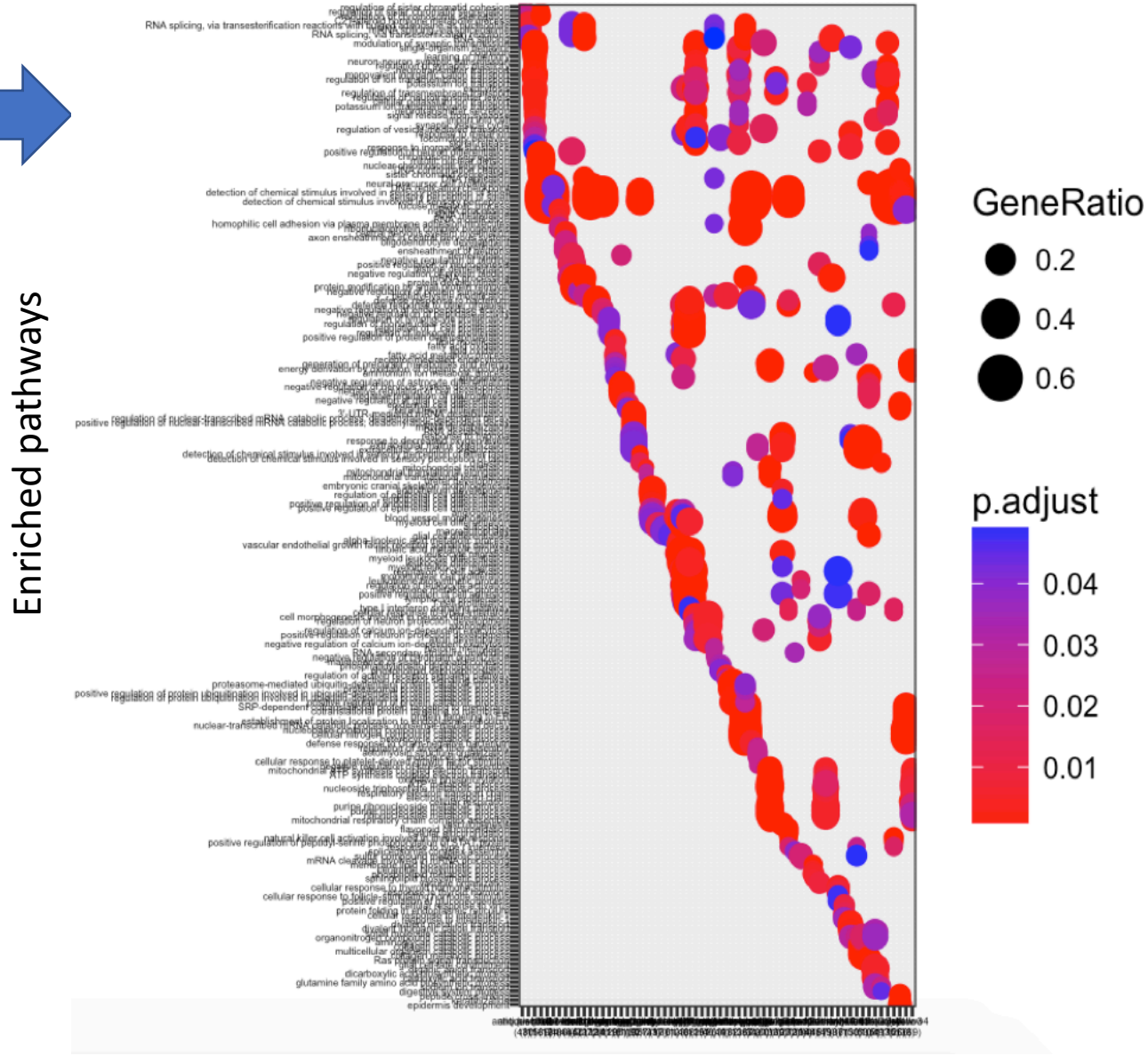
Co-expression across developmental stages



NCX gene expression matrix

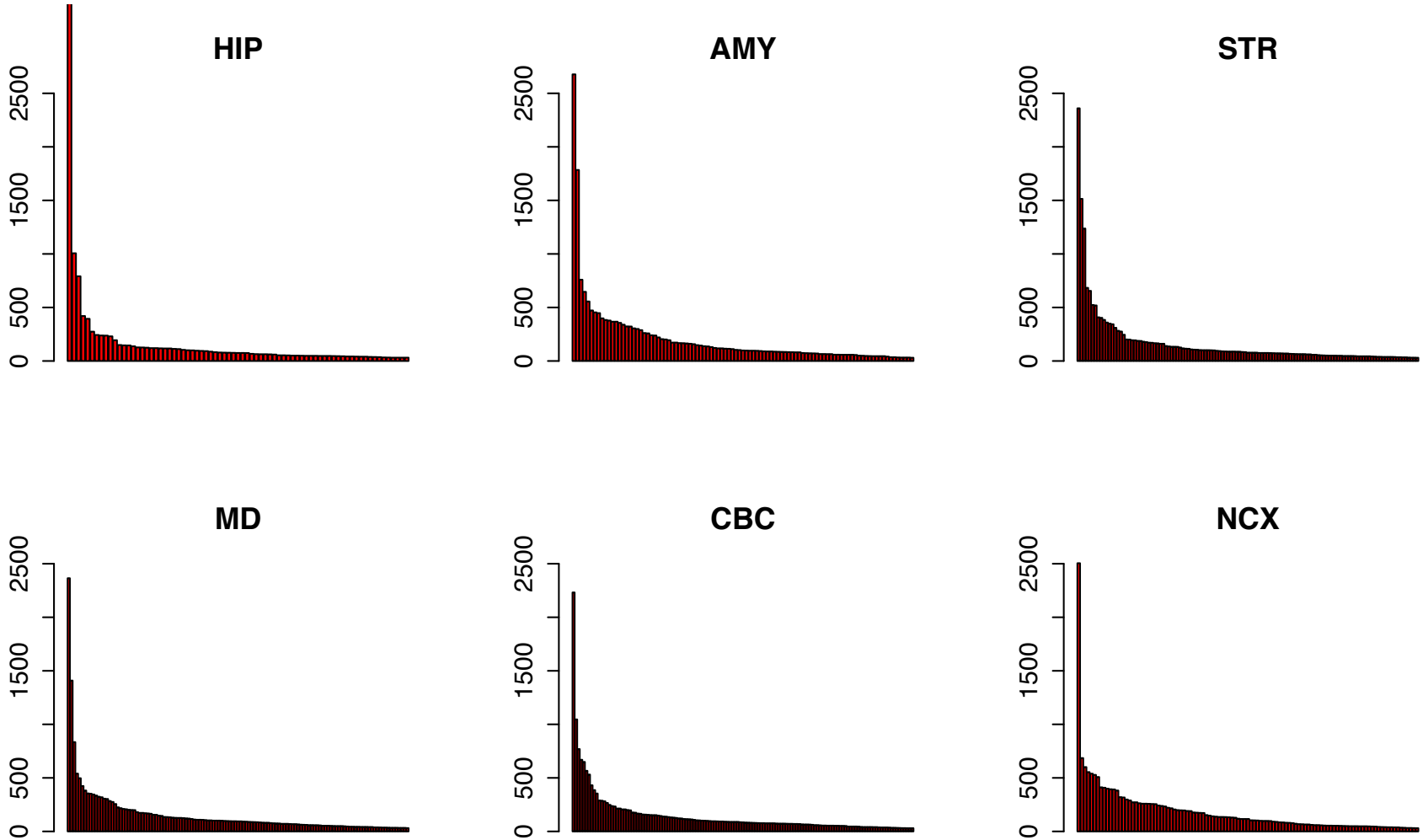


WGCNA



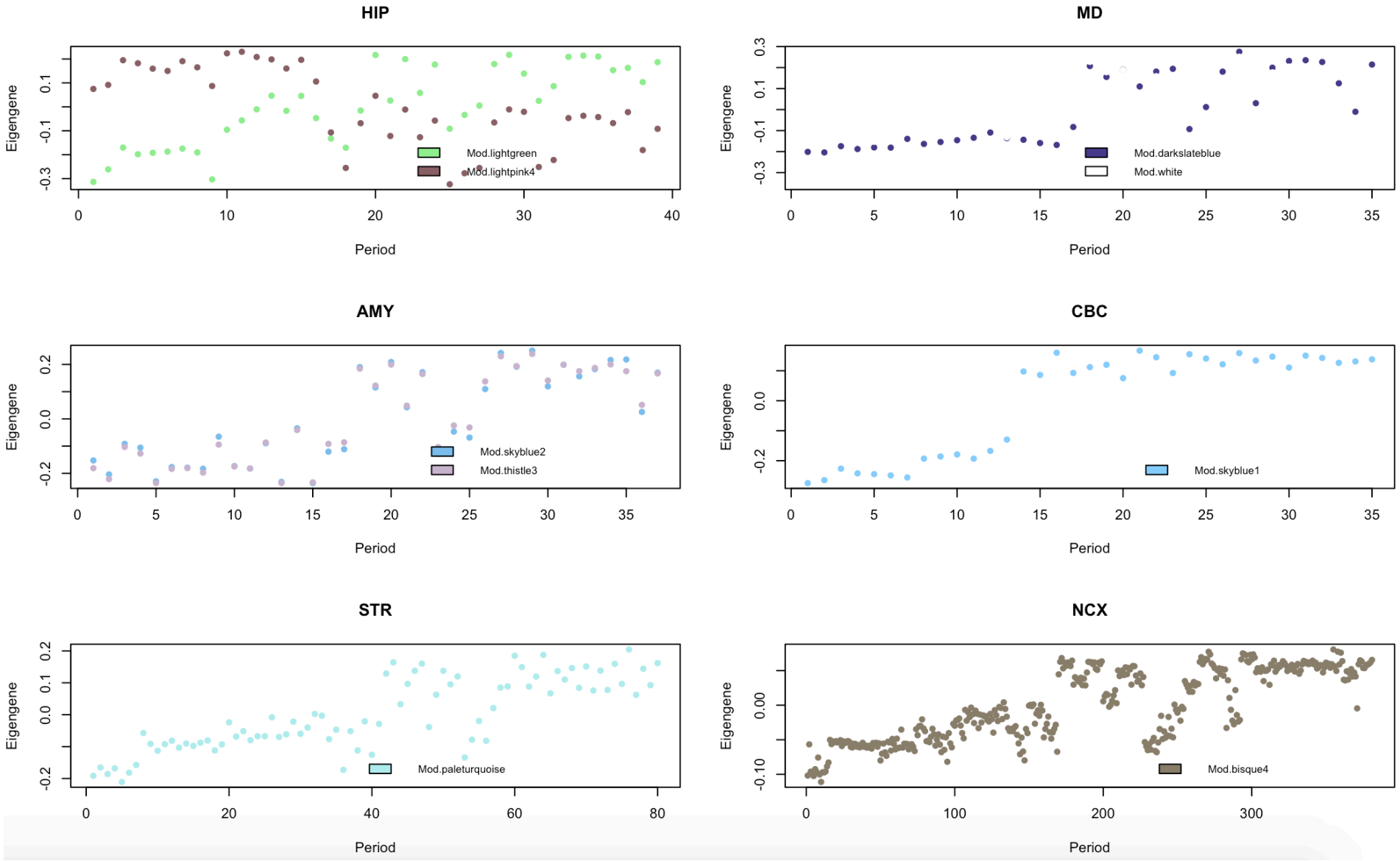
Brainspan region developmental gene co-expression network and modules by WGCNA

WGCNA gene co-expression modules across brainspan periods

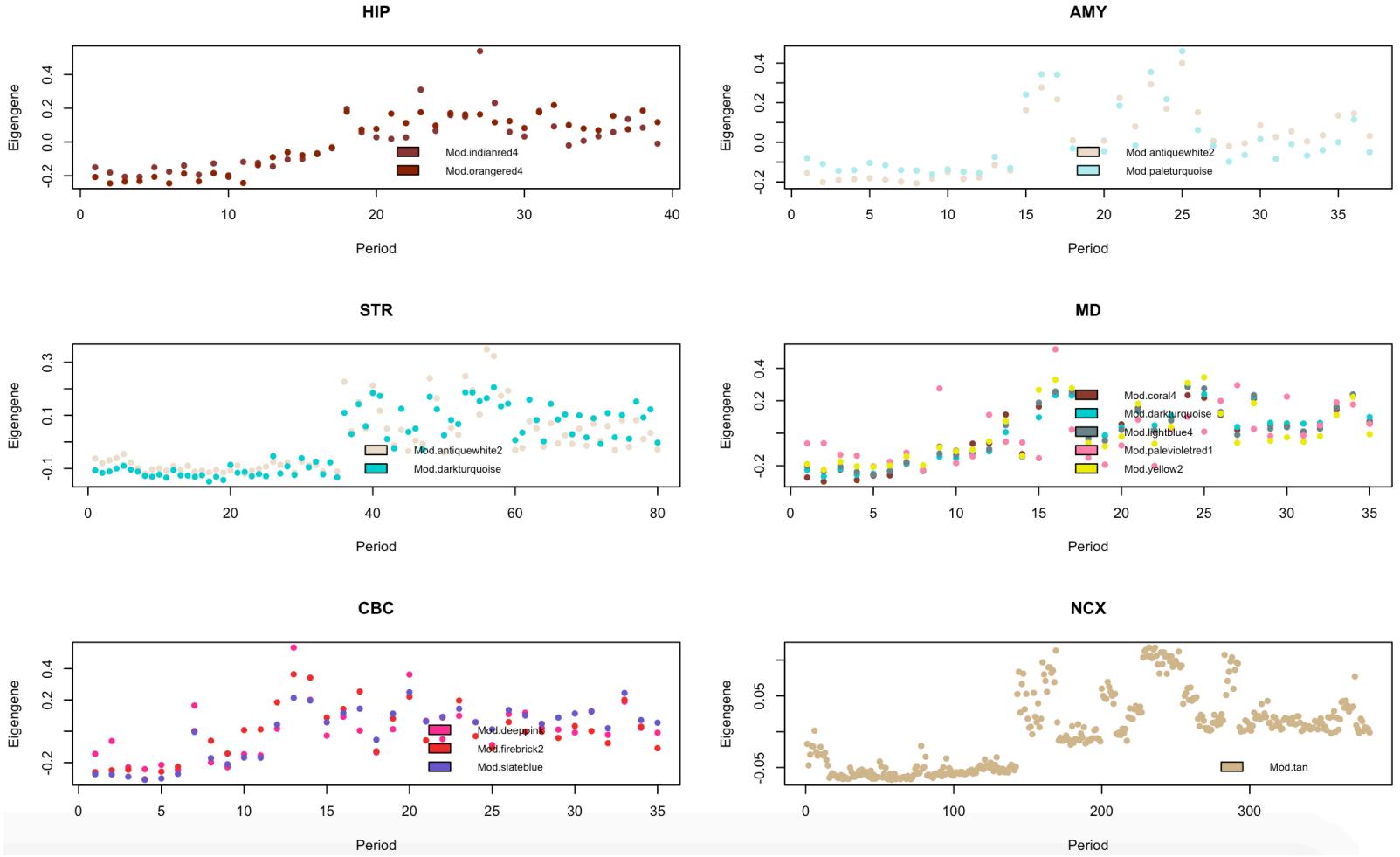


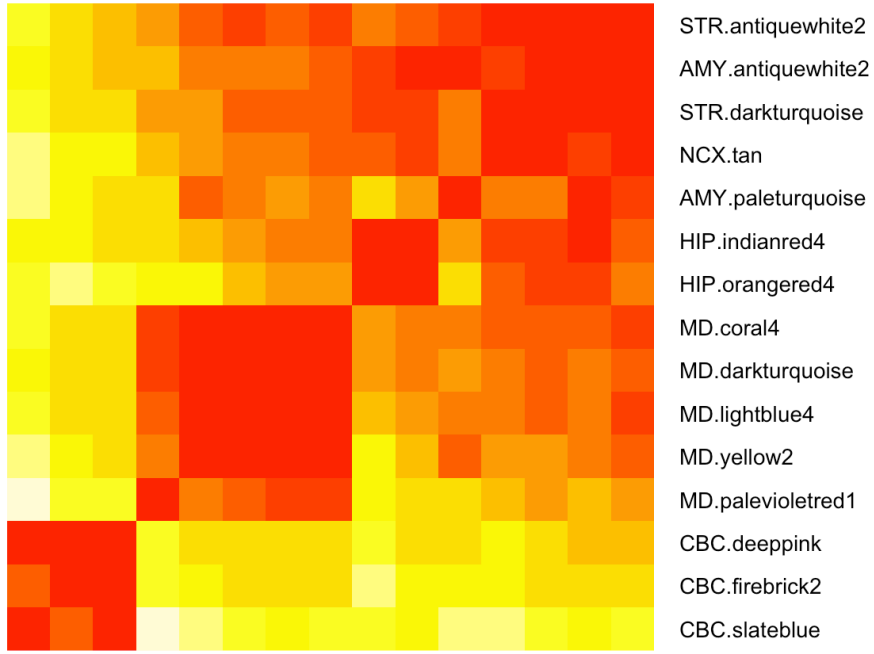
Pathway development: Morphine addiction

Eigengenes of "morphine addiction" modules



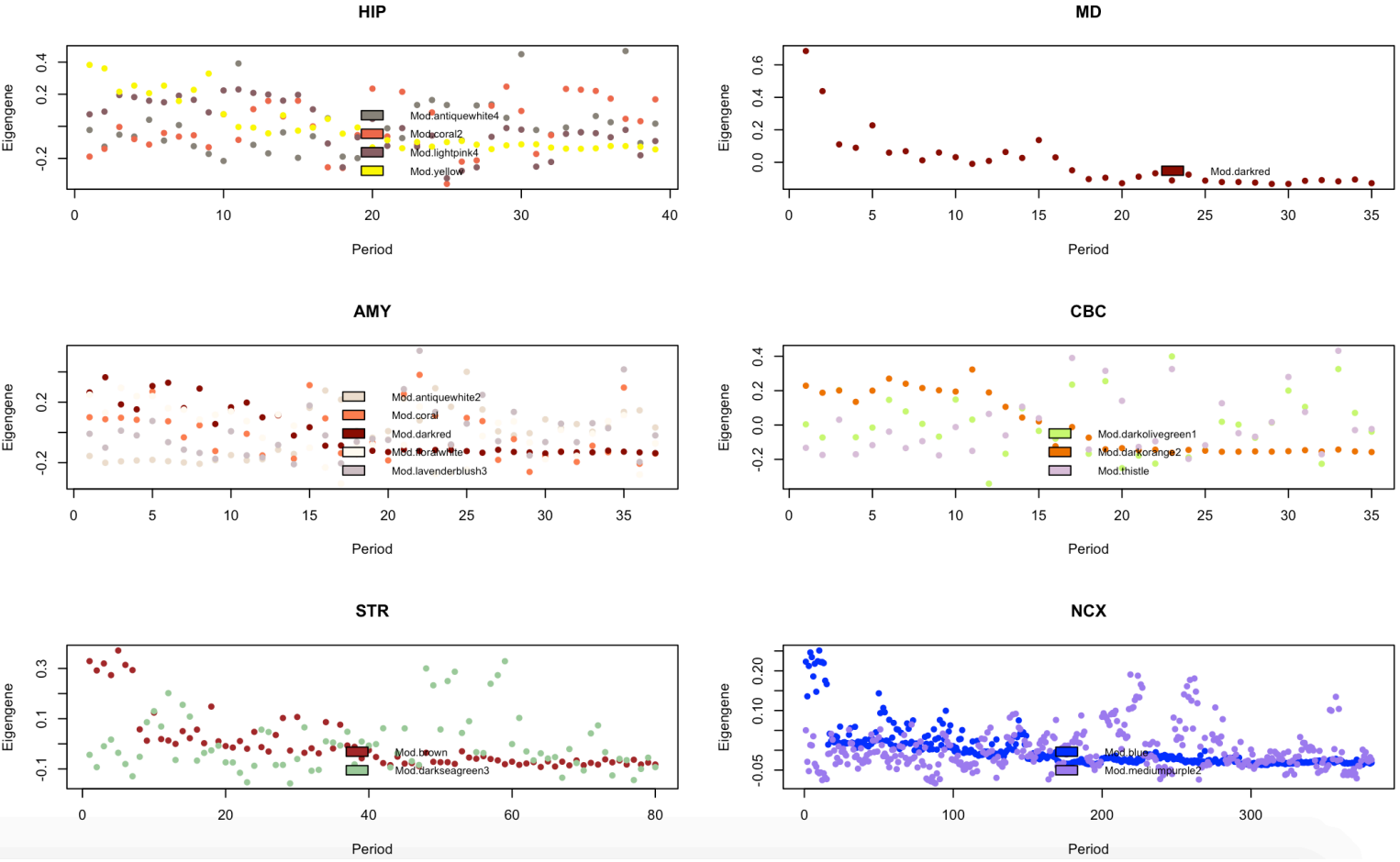
Another pathway: Interferon Signaling

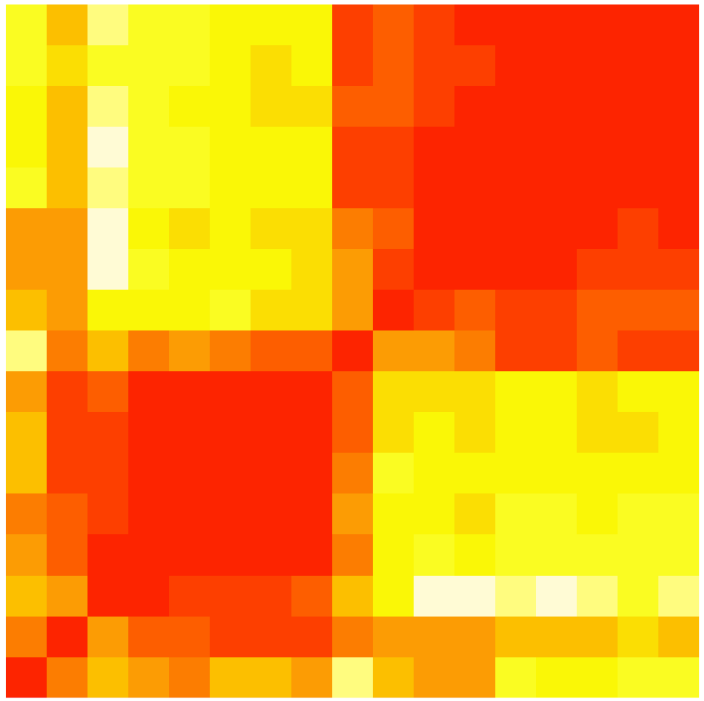




Gene membership correlations of interferon modules

Another pathway: Cell Cycle Checkpoint



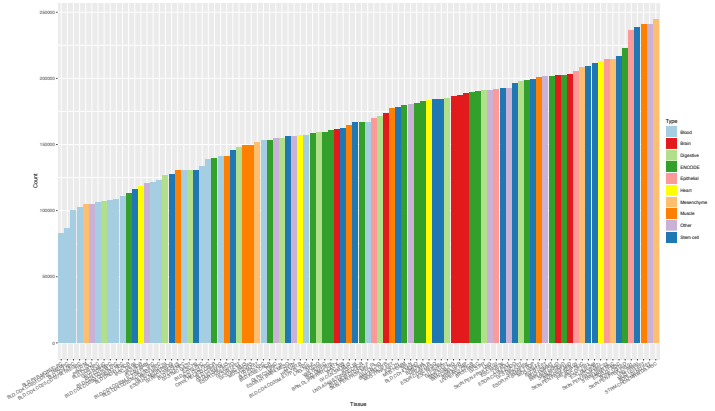


STR.brown
NCX.blue
MD.darkred
AMY.darkred
HIP.yellow
AMY.floralwhite
HIP.lightpink4
CBC.darkorange2
AMY.coral
NCX.mediumpurple2
STR.darkseagreen3
CBC.thistle
HIP.antiquewhite4
AMY.lavenderblush3
AMY.antiquewhite2
CBC.darkolivegreen1
HIP.coral2

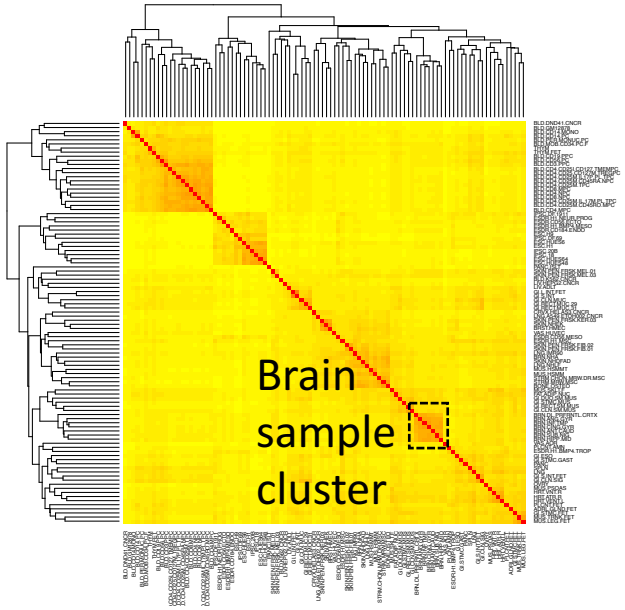
Gene membership correlations of cell cycle checkpoint modules

Brain gene regulation (MTG)

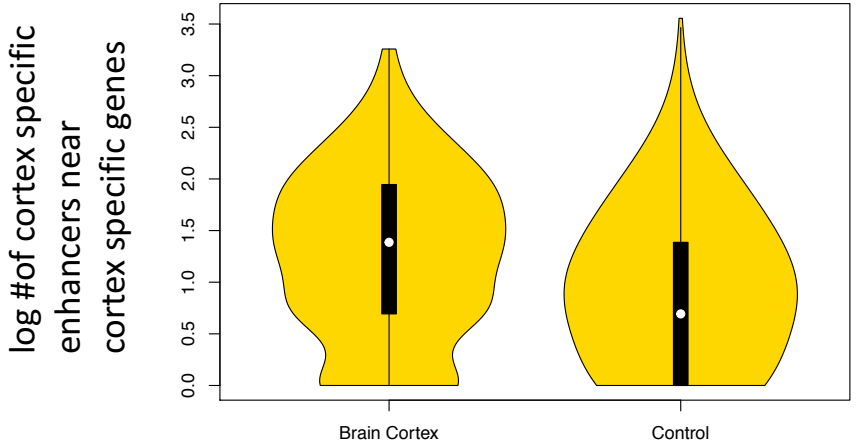
A. Brain enhancers



B. Brain samples clustered by enhancers

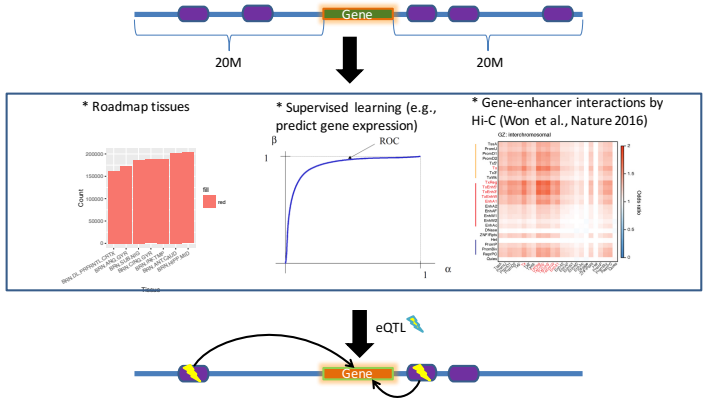


C. Brain enhancers vs. Brain genes



D. Gene regulatory networks

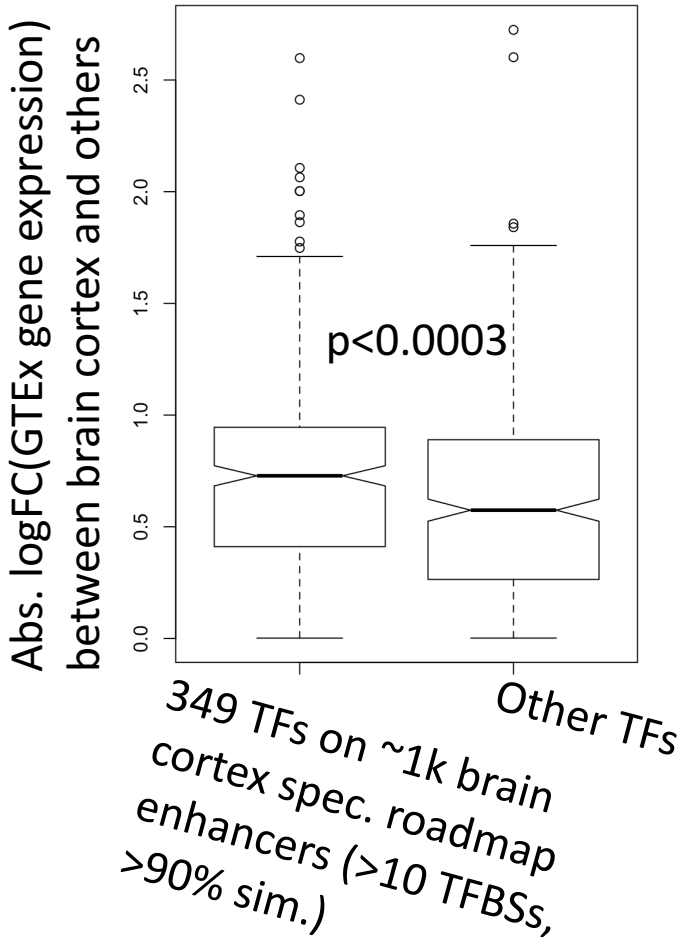
- Tissues-matched & Developmental-matched

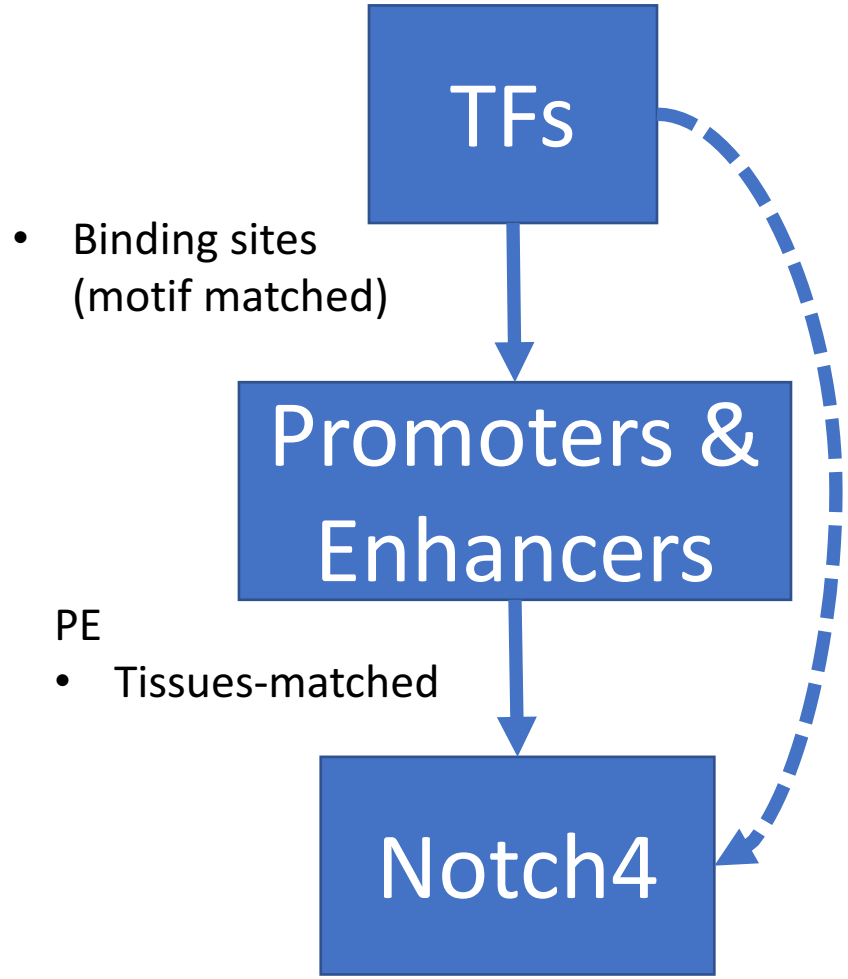


Building brain gene regulatory networks and circuits

- Binding sites (motif matched)

PE
- Tissues-matched
- Developmental-matched



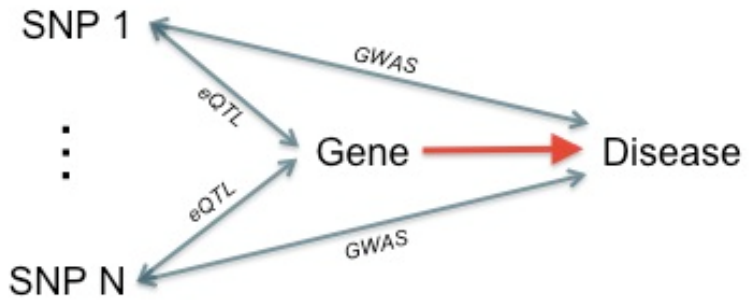


- Binding sites (motif matched)

PE

- Tissues-matched

Building brain gene regulatory networks and circuits

- Predict Target expression from TFs using regression (LASSO, Ridge, and Elastic Net)
- PEC BrainGVEX Ctrl and SCZ
- 

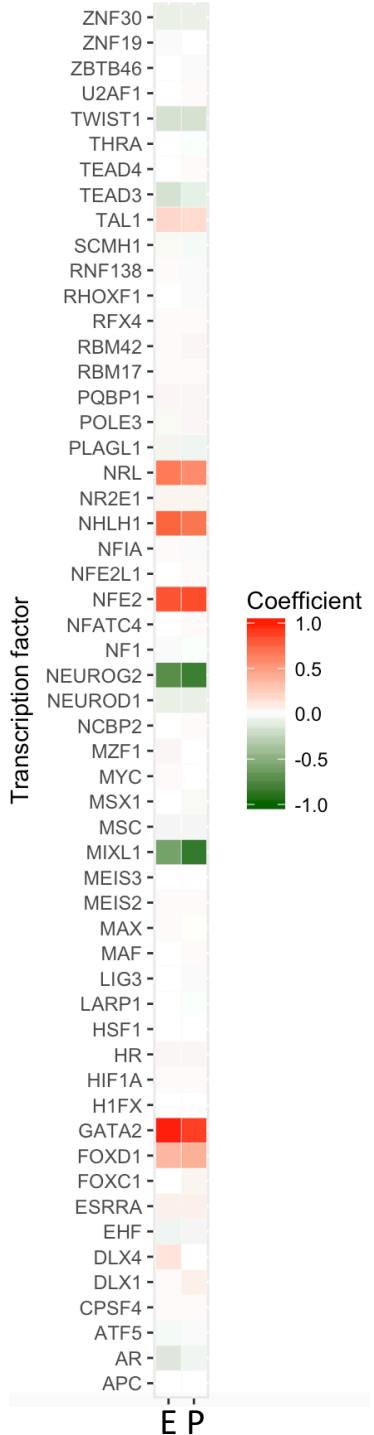
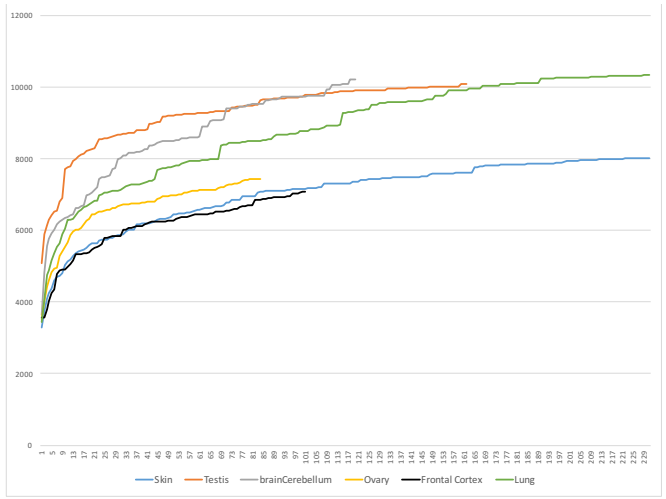
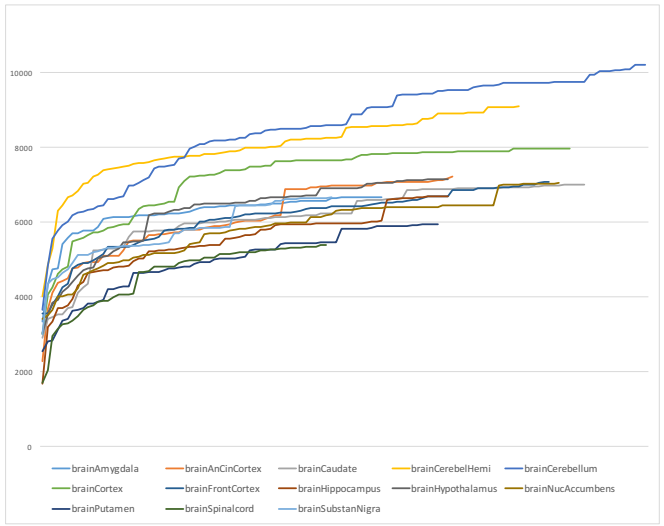


Figure 3 Novel transcriptional regions in Brain(FN)

A. Brain specific TARs

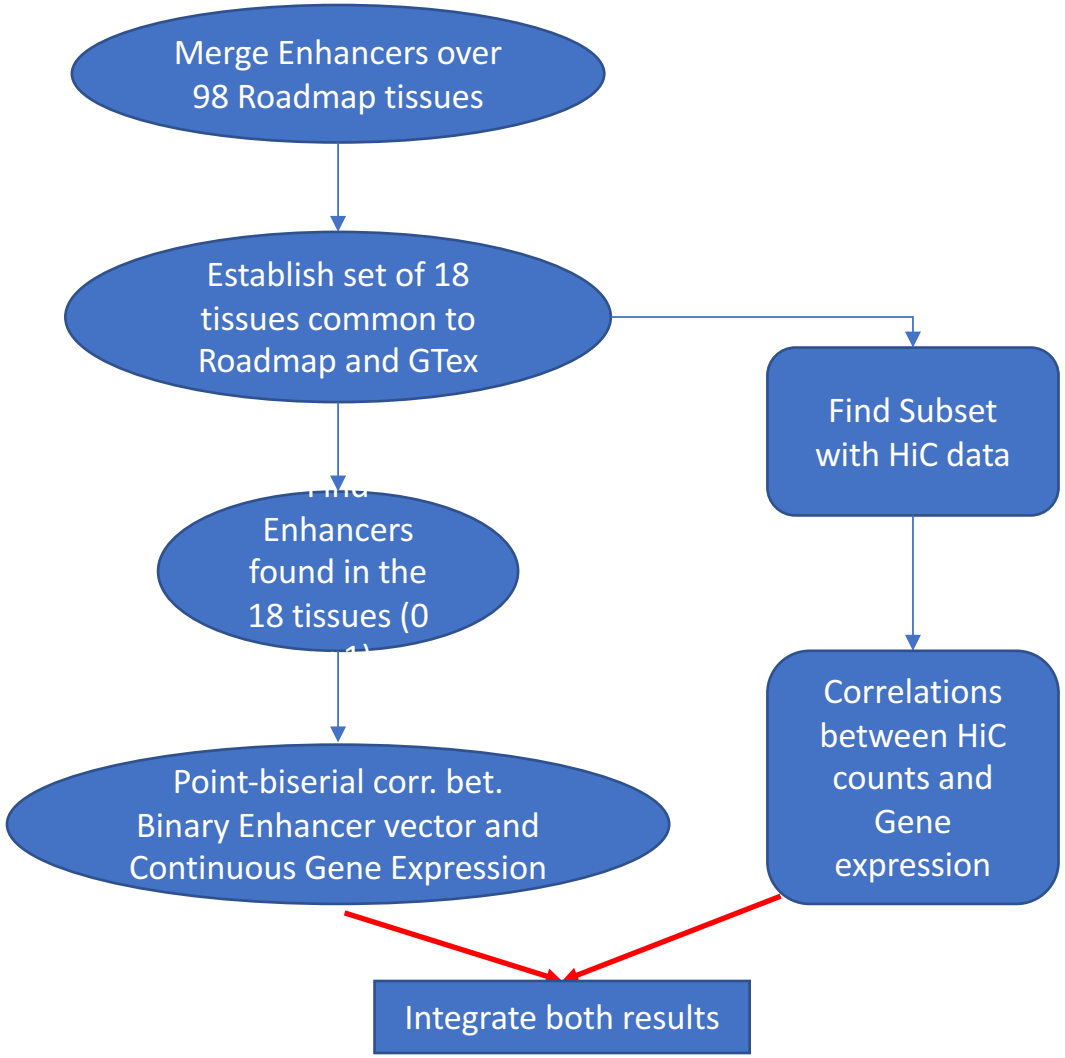


B. Brain region specific TARs



C. Associations with brain genes and functions

D. Summary table
# of miRNAs, lincRNAs, eRNAs, pseduogenes, …

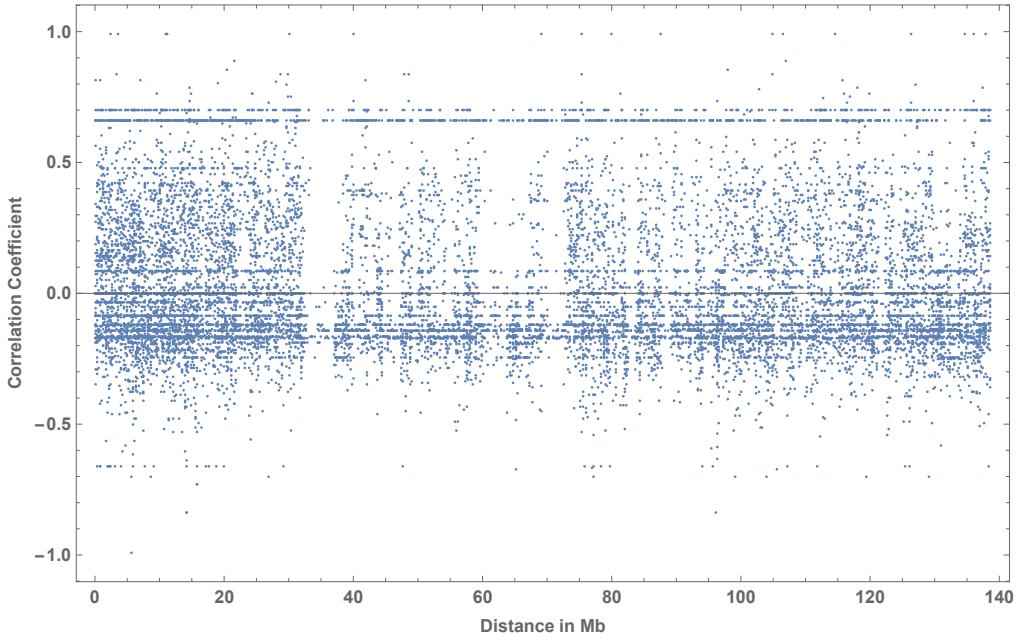


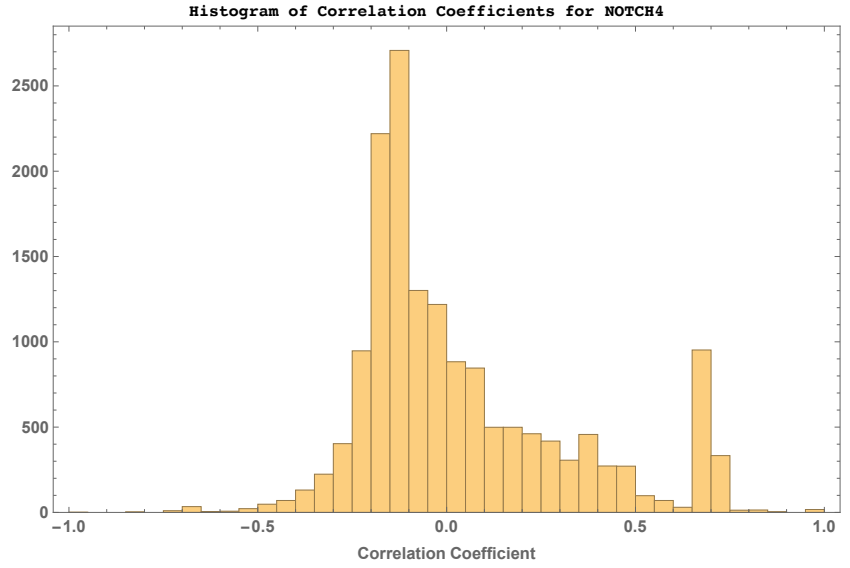
Analyses

- Rank Enhancers by correlation
- Build heatmaps of correlation and genomic location
- PCA of genes

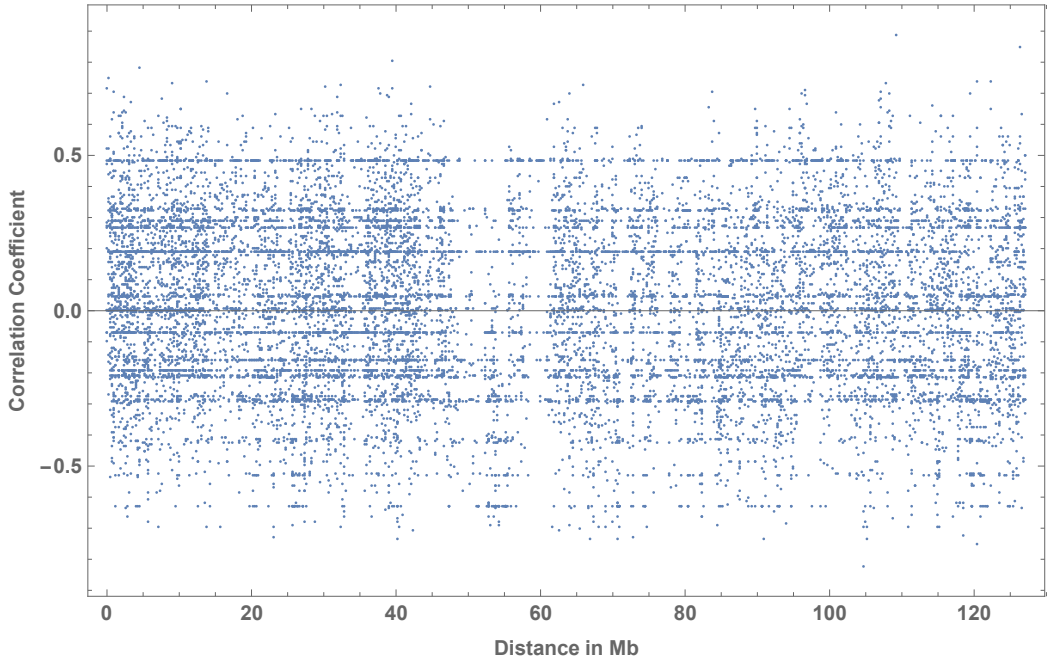
Enhancer and gene expression pattern correlation

Patterns for Notch4 in Chr. 6





Patterns for VegfA in Chr. 6



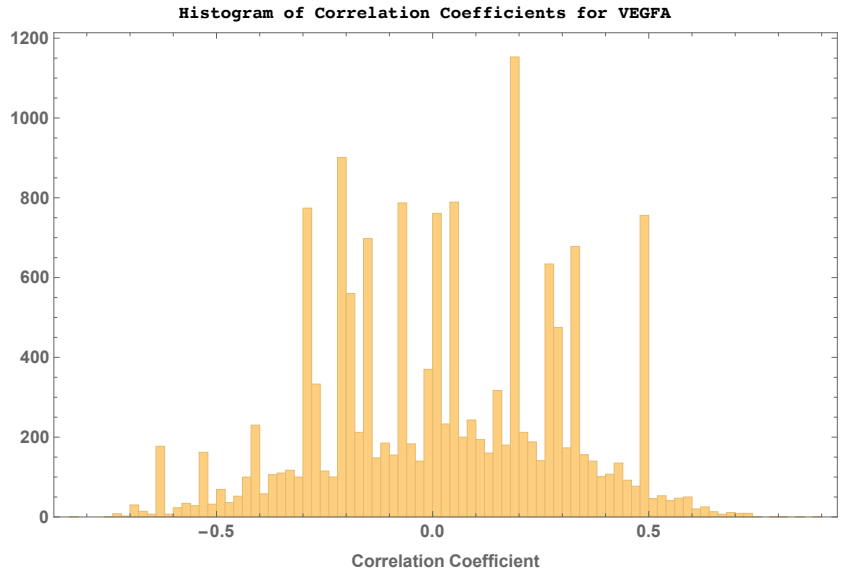
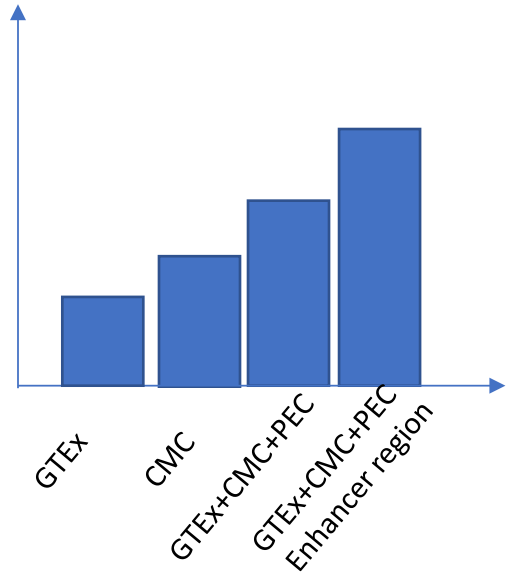
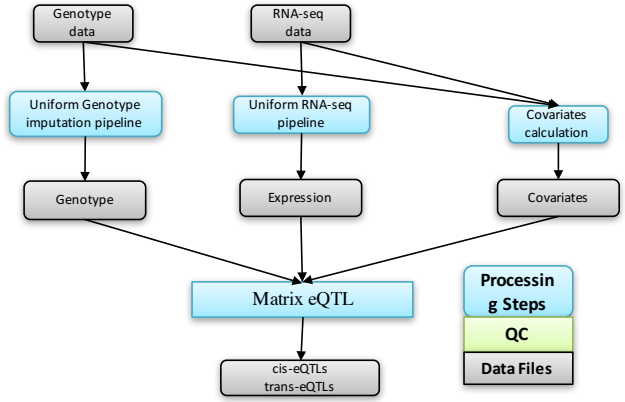


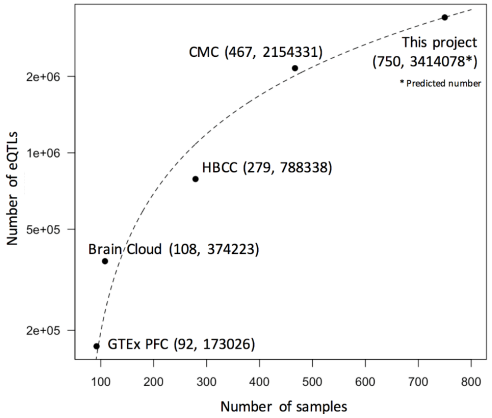
Figure 5 Brain eQTLs

A. Integration

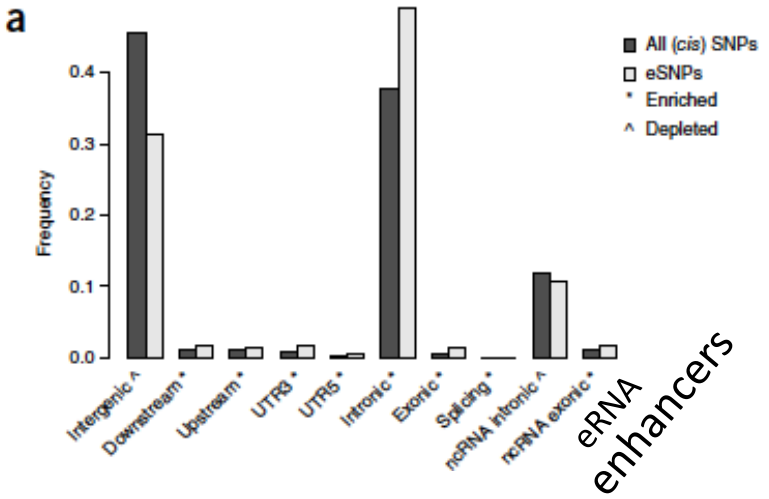


B. Pipeline for novel prediction

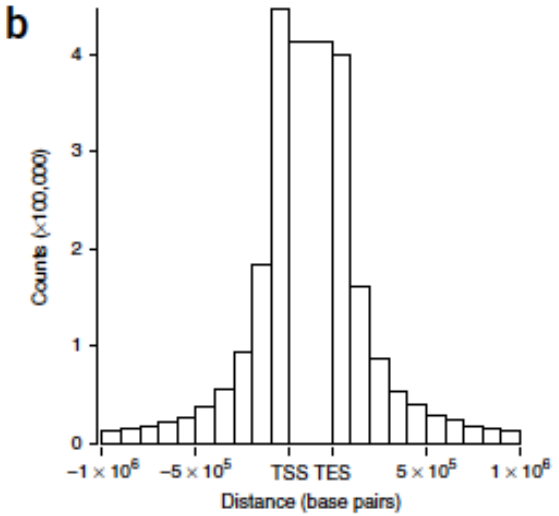




C. Enrichments of cis-eQTL compared to all eQTL in sequence-defined elements

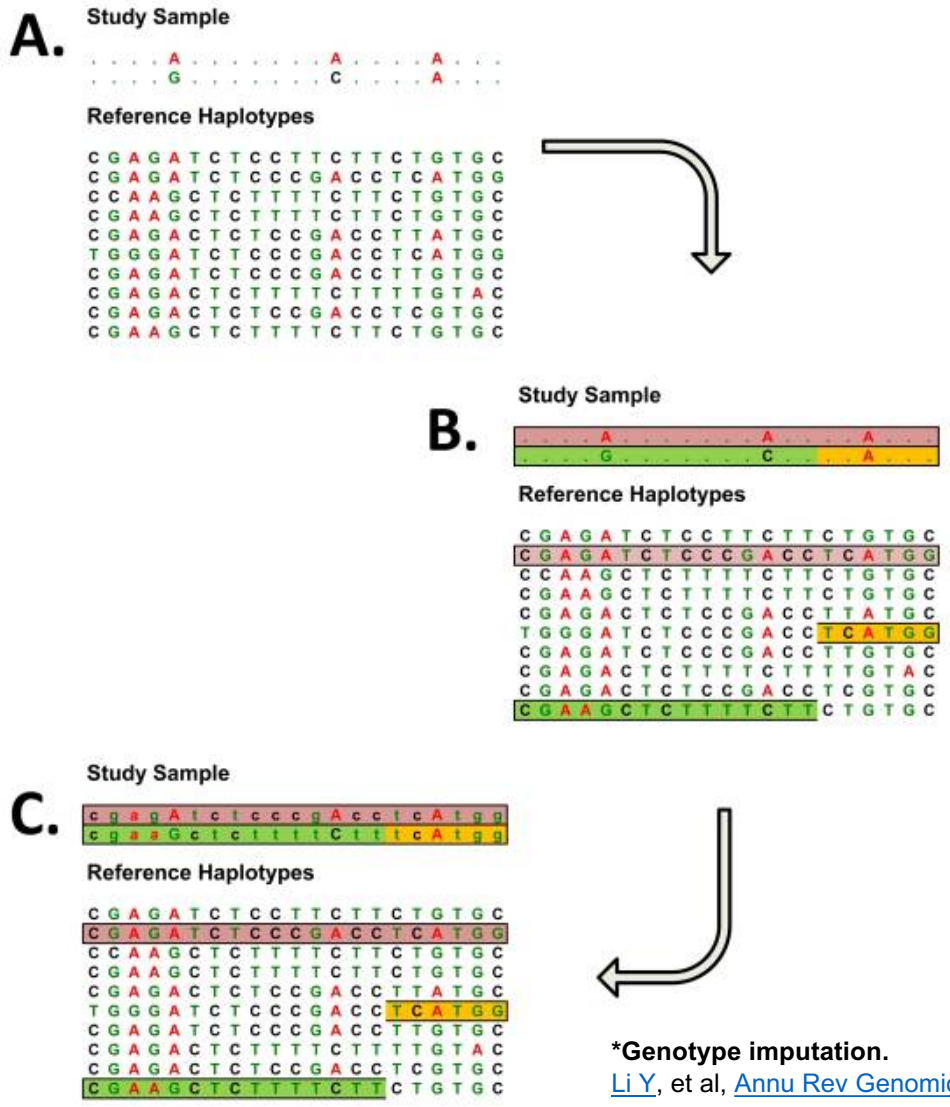


D. Distribution of Brain eQTLs?



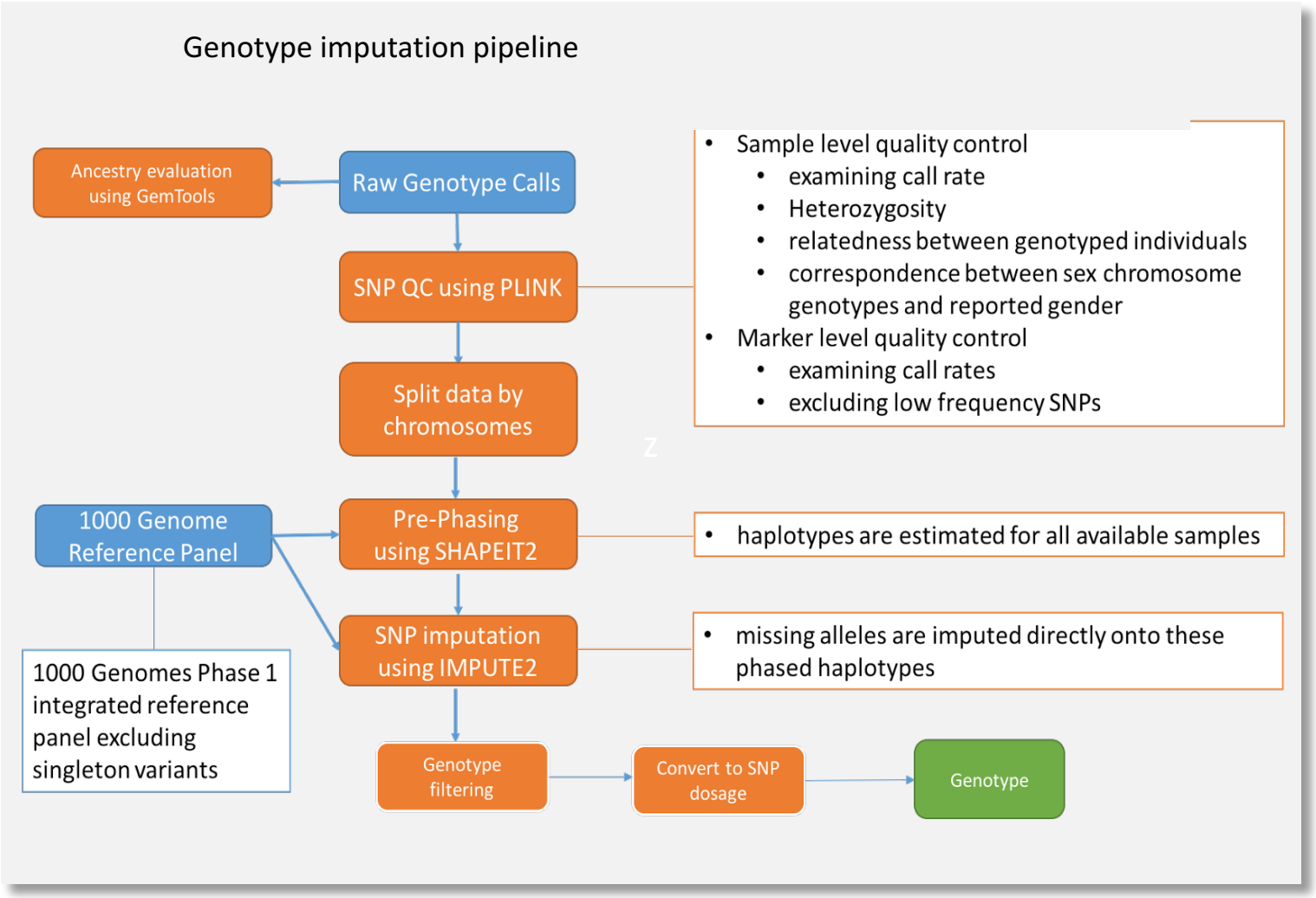
Genotype imputation

- Evaluate the evidence for association at genetic markers that are not directly genotyped
- Increases power of genomewide association scans
- Useful for combining data from studies that rely on different genotyping platforms



***Genotype imputation.**
Li Y, et al, Annu Rev Genomics Hum Genet. 2009

Genotype imputation pipeline



Compare HRC vs. 1KGP panel

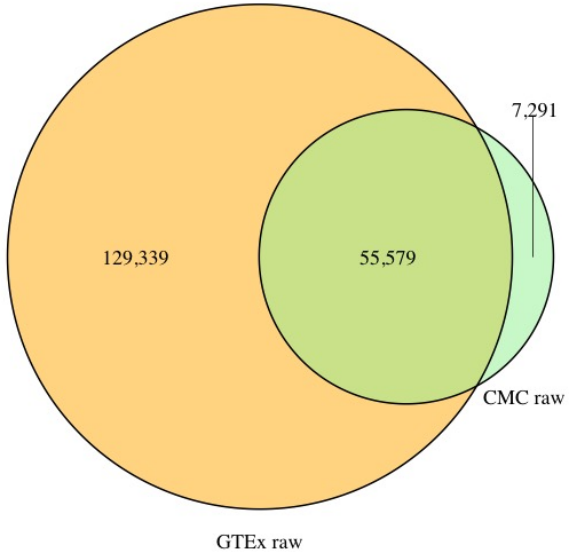
- 1000 Genomes Phase 1 panel (1KGP,1,092 samples genotyped at 28,975,367 sites spanning the autosomes)
- The Human Reference Consortium panel (HRC,32,470 samples genotyped at 39,635,008 sites spanning the autosomes and the X chromosome). McCarthy S, et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics, 48,* 1279-83.

		Imputation accuracy (mean r^2), MAF = 0.0001–0.5%			
1000G Phase 1	1,092	**0.45**	**0.45**	0.43	0.42
AMD	2,074	**0.54**	**0.54**	0.51	0.52
1000G Phase 3	2,504	**0.52**	**0.52**	0.49	**0.52**
SardiNIA	3,489	**0.55**	**0.55**	0.53	0.54
COMBINED	9,341	**0.76**	**0.76**	0.74	**0.76**
Mega	11,845	**0.76**	**0.76**	0.74	**0.76**
HRC v1.1	32,390	**0.77**	**0.77**	0.75	**0.77**

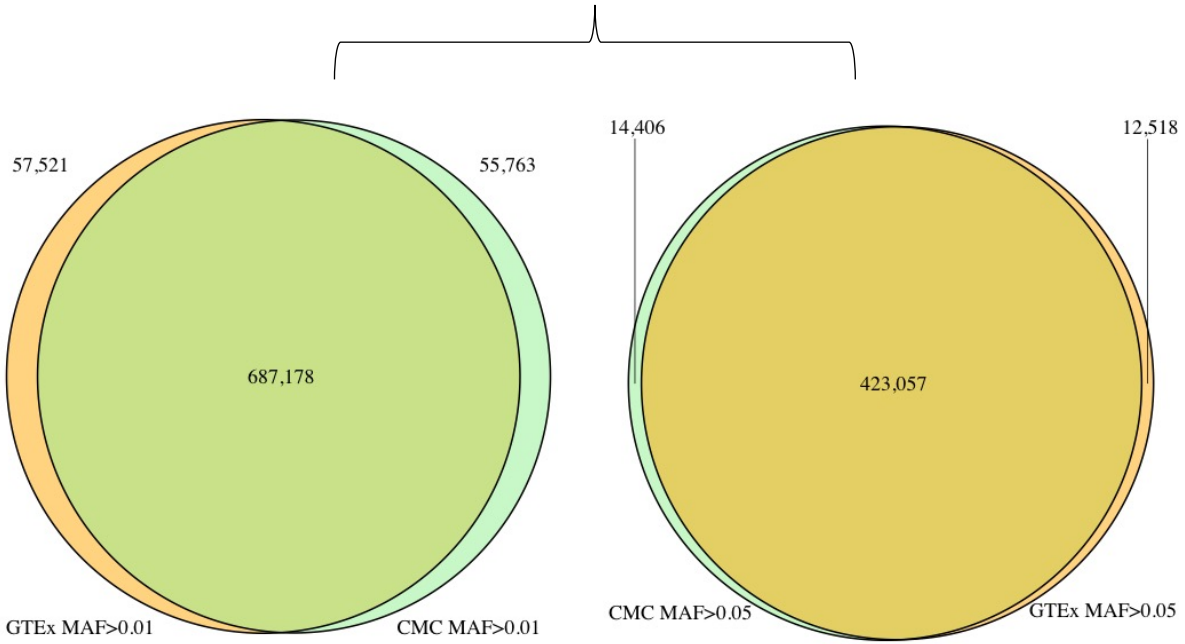
		Imputation accuracy (mean r^2), MAF = 0.5–5%			
1000G Phase 1	1,092	**0.77**	**0.77**	0.76	0.73
AMD	2,074	**0.82**	**0.82**	0.80	0.80
1000G Phase 3	2,504	**0.79**	**0.79**	0.78	**0.79**
SardiNIA	3,489	0.79	0.79	0.78	**0.80**
COMBINED	9,341	**0.89**	**0.89**	0.88	**0.89**
Mega	11,845	**0.89**	**0.89**	0.88	**0.89**
HRC v1.1	32,390	**0.90**	**0.90**	0.89	**0.90**

		Imputation accuracy (mean r^2), MAF = 5–50%			
1000G Phase 1	1,092	**0.96**	**0.96**	0.95	0.95
AMD	2,074	**0.96**	**0.96**	**0.96**	**0.96**
1000G Phase 3	2,504	**0.96**	**0.96**	**0.96**	**0.96**
SardiNIA	3,489	**0.96**	**0.96**	**0.96**	**0.96**
COMBINED	9,341	**0.97**	**0.97**	**0.97**	**0.97**
Mega	11,845	**0.97**	**0.97**	**0.97**	**0.97**
HRC v1.1	32,390	**0.98**	**0.98**	**0.98**	**0.98**

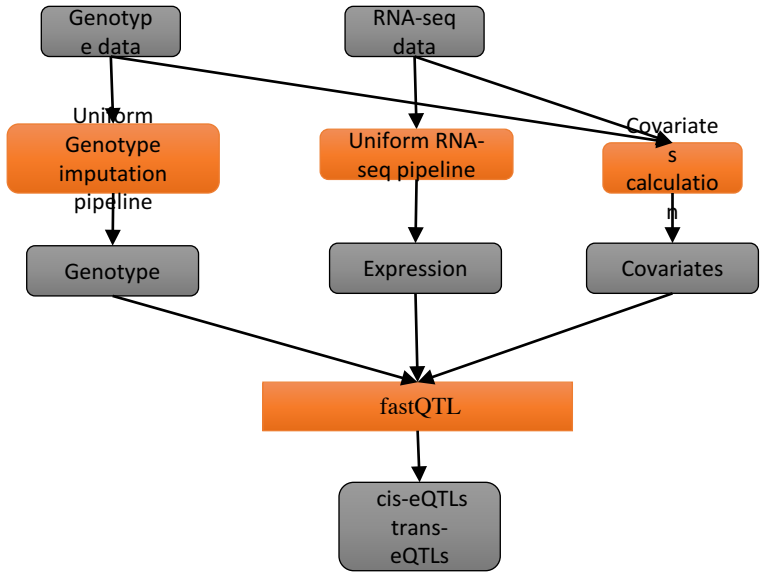
Before imputation



After imputation



eQTL analysis



Expression

- Genes were selected based on expression thresholds of >0.1 RPKM in at least 10 individuals and ≥6 reads in at least 10 individuals.
- Quantile normalized and inverse quantile normalized

Genotypes

- Variants were imputed using HRC panel.
- genotype filters applied:
  - Call Rate Threshold 95%.
  - R2 Threshold 0.6.
  - MAF>= 1%

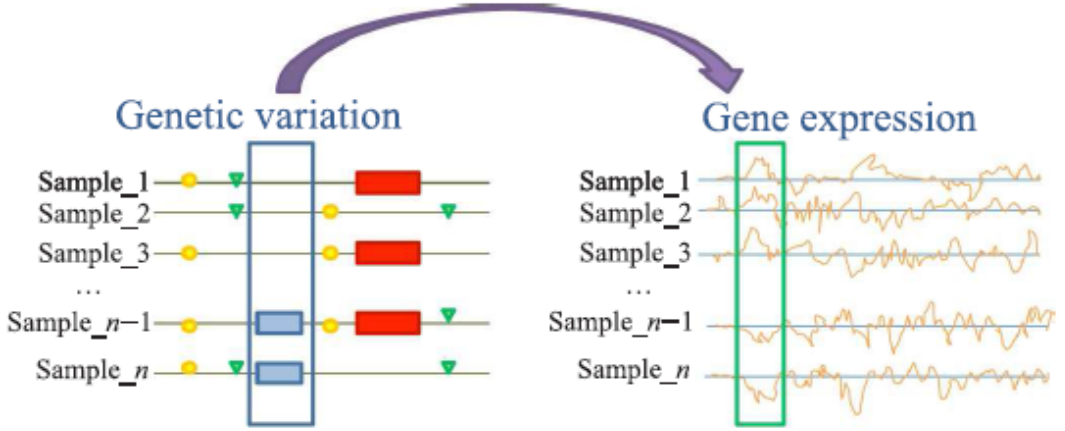
Covariates

- Top 3 genotyping principal components
- Probabilistic Estimation of Expression Residuals (PEER) factors
- Genotyping array platform
- Gender

eQTL Analysis using FastQTL

Mapping window --1 megabase from TSS

FastQTL



- eQTL
  - P: a single molecular phenotype (expression)
  - G :the set of genotype dosages
  - L: variant sites located within a cis-window of +/- W Mb of the genomic location of P
  - To discover the best candidate QTL for P, FastQTL measures Pearson product-moment correlation coefficients between P and all L variants in G, stores the most strongly correlated variant q
- Corrections:
  - Permutation-multiple genetic variants are tested per phenotype
  - FDR estimation-multiple phenotypes are tested genome-wide

Fastqtl inputs

Genotypes

```
##fileformat=VCFv4.1
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT UNR1 UNR2 UNR3 UNR4
chr7 123 SNP1 A G 100 PASS INFO GT:DS 0/0:0.001 0/0:0.000 0/1:0.999 1/1:1.999
chr7 456 SNP2 T C 100 PASS INFO GT:DS 0/0:0.001 0/0:0.000 0/1:1.100 0/0:0.100
chr7 789 SNP3 A T 100 PASS INFO GT:DS 1/1:2.000 0/1:1.001 0/0:0.010 0/1:0.890
```

Phenotypes

```
#Chr start end ID UNR1 UNR2 UNR3 UNR4 chr1 173863 173864 ENSG123 -0.50 0.82
-0.71 0.83
chr1 685395 685396 ENSG456 -1.13 1.18 -0.03 0.11
chr1 700304 700305 ENSG789 -1.18 1.32 -0.36 1.26
```

Covariates

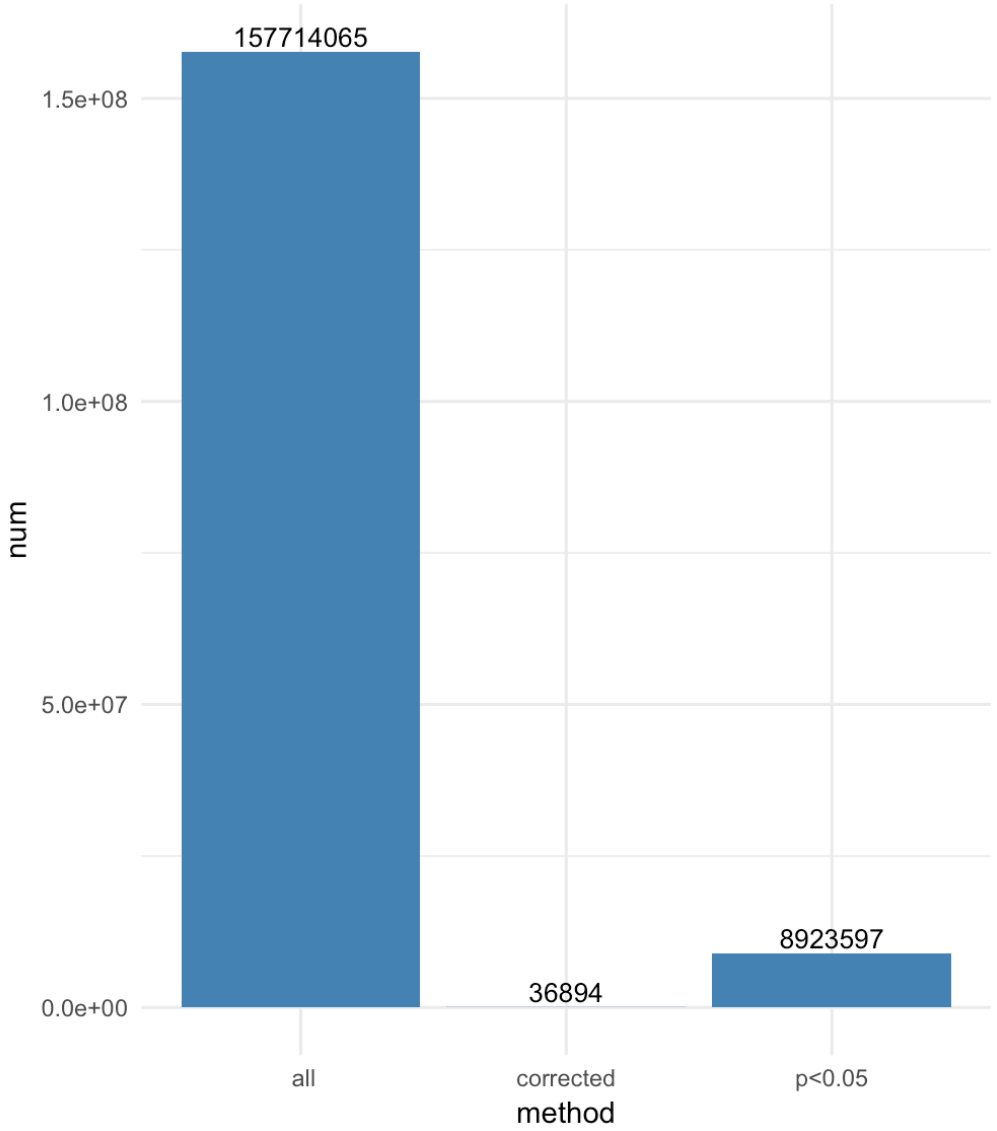
```
id UNR1 UNR2 UNR3 UNR4
PC1 -0.02 0.14 0.16 -0.02
PC2 0.01 0.11 0.10 0.01
PC3 0.03 0.05 0.08 0.07
BIN 1 0 0 1
```

Fastqtl outputs

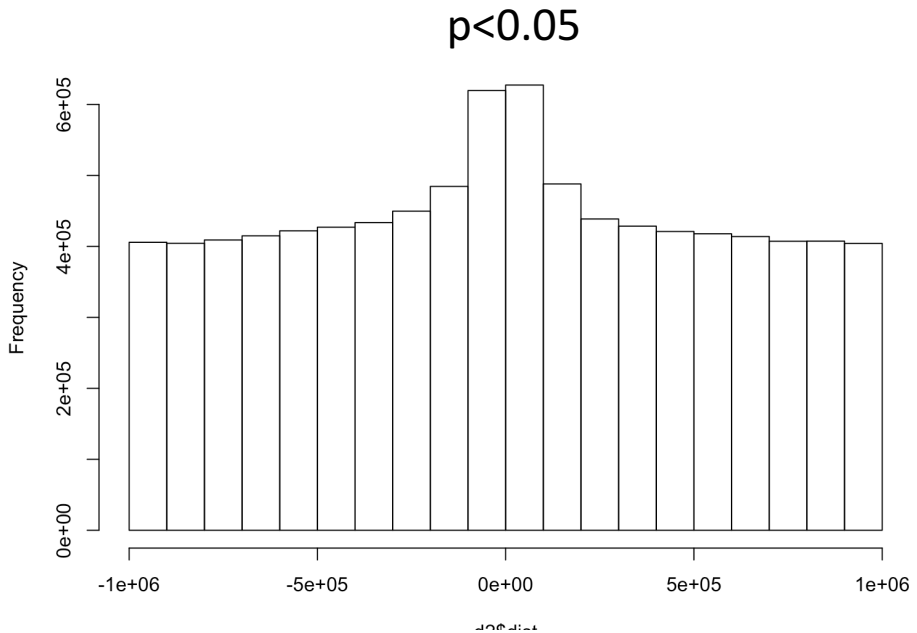
pid	nvar	shape1	shape2	dummy	sid	dist	npval	ppval	bpval
ENSG00000137960.5	6499	0.986176	521.431	85.9219	1:78991197	479611	0.00716333	0.991009	0.991204
ENSG00000122420.5	6748	1.06797	675.215	89.1562	1:79379411	609843	0.00154632	0.663337	0.642909
ENSG00000137959.11	6771	0.994923	488.644	87.0469	1:78845586	-240021	0.0013094	0.553447	0.540603
ENSG00000137965.6	6844	1.0706	367.636	79.0312	1:78830851	-284630	0.00495443	0.944056	0.950312
ENSG00000162618.8	7562	1.06351	518.869	84.9023	1:80128274	655871	0.00106882	0.562438	0.529677
ENSG00000117114.15	7430	1.03713	922.937	88.0312	1:81848542	76697	0.000744233	0.517483	0.527867
ENSG00000242598.1	7264	1.06631	782.403	84.6562	1:82024065	548	0.000559152	0.473526	0.457863
ENSG00000236268.1	6481	1.02308	555.127	84.9023	1:83085806	-366085	2.34E-05	0.021978	0.0219598
ENSG00000137941.12	6393	0.948159	699.688	88.0312	1:84269426	-195407	0.000596147	0.398601	0.40751
ENSG00000271576.1	6376	1.0122	801.111	89.2969	1:85443807	900193	0.00117491	0.637363	0.628244

1. ID of the tested molecular phenotype (in this particular case, the gene ID)
2. Number of variants tested in cis for this phenotype
3. MLE of the shape1 parameter of the Beta distribution
4. MLE of the shape2 parameter of the Beta distribution
5. Dummy [To be described later]
6. ID of the best variant found for this molecular phenotypes (i.e. with the smallest p-value)
7. Distance between the molecular phenotype - variant pair
8. The **nominal** p-value of association that quantifies how significant from 0, the regression coefficient is
9. The **slope** associated with the nominal p-value of association [only in version > v2-184]
10. A first **permutation** p-value directly obtained from the permutations with the direct method. This is basically a corrected version of the nominal p-value that accounts for the fact that multiple variants are tested per molecular phenotype.
11. A second **permutation** p-value obtained via beta approximation. We advice to use this one in any downstream analysis.

eQTL results



p<0.05



After Bonferroni correction

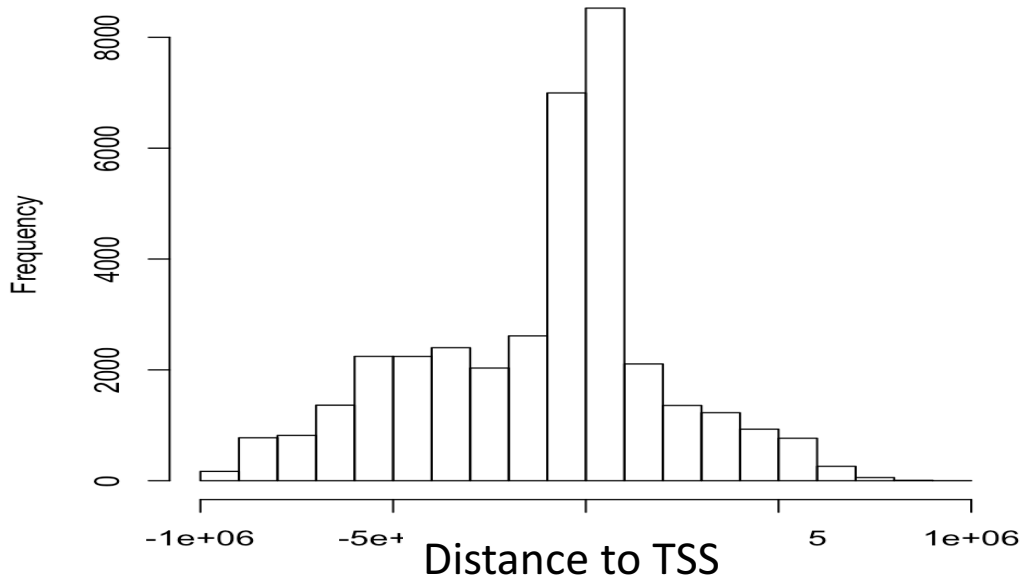
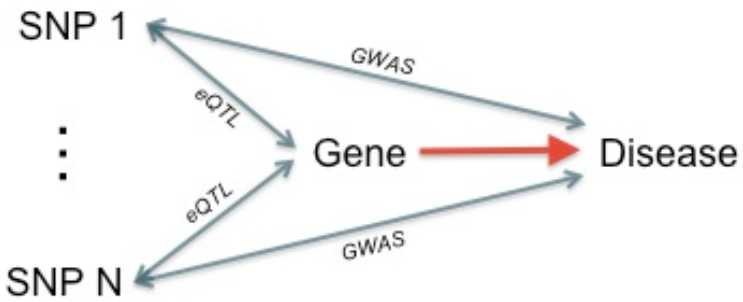


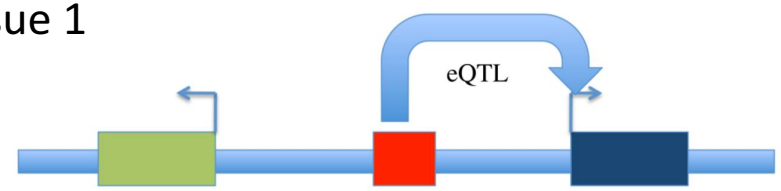
Figure 6. Integrative analysis

A. Overlap of GWAS SNPs and eQTL SNPs and Gene regulatory networks

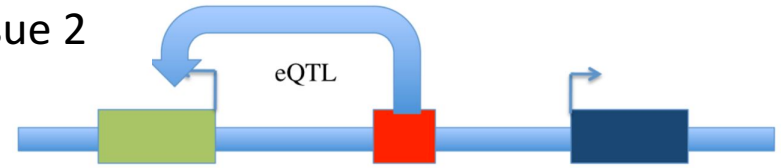


B. SNPs are different eQTLs in different tissues

Tissue 1



Tissue 2



Comparing cQTL & eQTLs.

C. Gene regulatory circuits

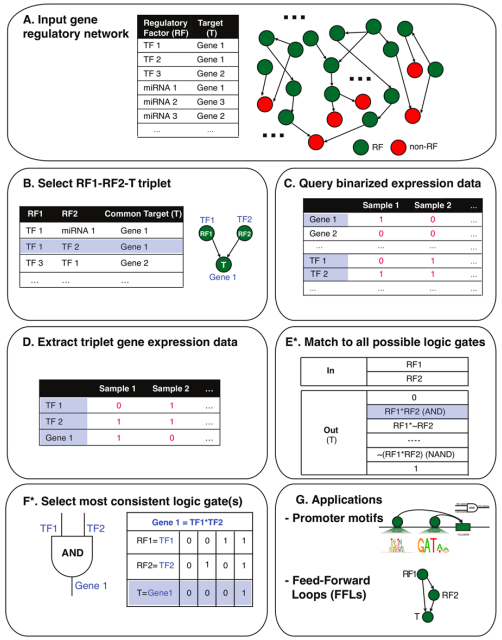


Table 1. Data resource: Brain genomics, transcriptomics and regulatomics

- RNA-seq, ChIP-seq, Hi-C data for # of brain regions, diseases
- Brain and disease genes
- eQTLs
- Expressed tissues: brain regions, other tissues
- Enhancers
  - Tissue or development
- Pathways and functions
- Gene regulation
  - eQTLs vs. TFBSs on enhancers

Acknowledgement

- Mark Gerstein
- Shuang Liu
- Daifeng Wang
- Fabio Navarro
- Declan Clarke
- Emani, Prashant
- Mengting Gu
- Aparna Nathan
- Jonathan Warrell
- Jonathan Park
- Timur Galeev
- James Knowles
- Suhn Kyong Rhie
- Peggy Farnham
- Pamela Sklar
- All PsychENCODE Consortium members

http://psychencode.org/

http://commonmind.org/

http://www.roadmapepigenomics.org/

http://brainspan.org/