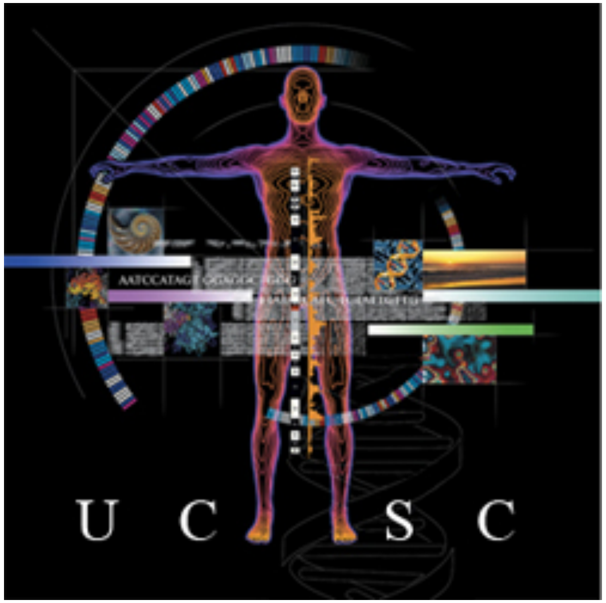
GTEx in the UCSC Genome Browser

Kate Rosenbloom

UCSC Genome Browser Group

UNIVERSITY OF CALIFORNIA
SANTA CRUZ
Genomics
Institute



December 2015

Motivation

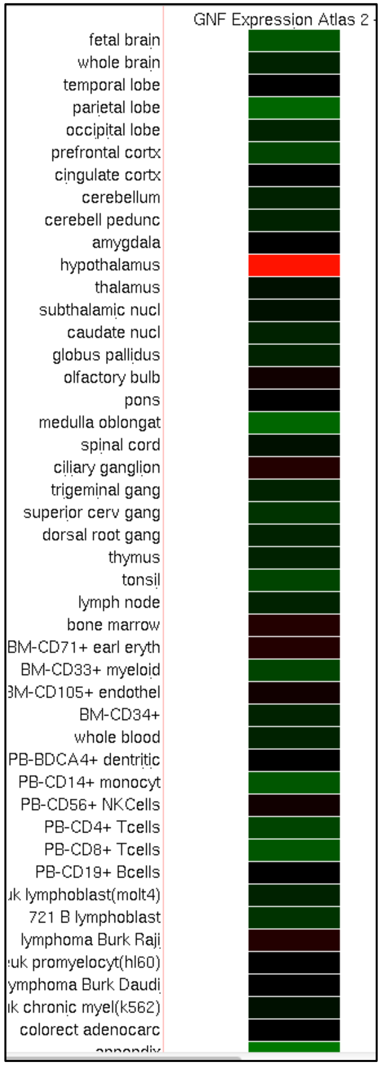
Incorporate newer gene expression data sets and explore new ways of displaying tissue-specific gene expression in the browser.

Current tissue expression tracks:

GNF Atlas 2 (2009)
ENCODE transcription signal (2008-2012)

➔

Browser supplement funding to integrate GTEx into the genome browser database and develop visualizations


GNF Expression Atlas 2

fetal brain
whole brain
temporal lobe
parietal lobe
occipital lobe
prefrontal cortx
cingulate cortx
cerebellum
cerebell pedunc
amygdala
hypothalamus
thalamus
subthalamic nucl
caudate nucl
globus pallidus
olfactory bulb
pons
medulla oblongat
spinal cord
ciliary ganglion
trigeminal gang
superior cerv gang
dorsal root gang
thymus
tonsil
lymph node
bone marrow
BM-CD71+ earl eryth
BM-CD33+ myeloid
3M-CD105+ endothel
BM-CD34+
whole blood
PB-BDCA4+ dentritic
PB-CD14+ monocyt
PB-CD56+ NKCells
PB-CD4+ Tcells
PB-CD8+ Tcells
PB-CD19+ Bcells
ık lymphoblast(molt4)
721 B lymphoblast
lymphoma Burk Raji
uk promyelocyt(hl60)
ymphoma Burk Daudi
ık chronic myel(k562)
colorect adenocarc

OXT: Oxytocin precursor
Expression from GNF Atlas2

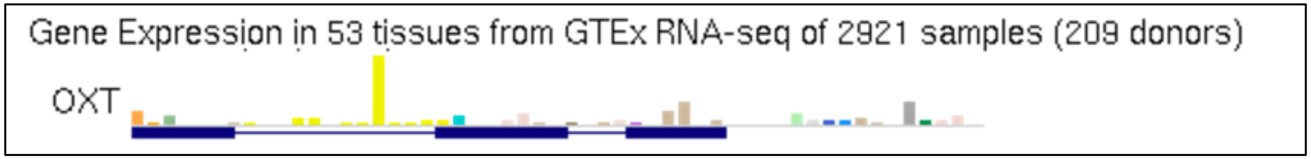
Aims

1. Integrate gene expression data and metadata into the browser database
2. Develop new track display with more compact layout for higher on-screen annotation density
3. Create browser tracks for GTEx gene-level expression and allele-specific expression

Progress

✓ 1. Integrate gene expression data and metadata into the browser database

✓ 2. Develop new bar-graph track display



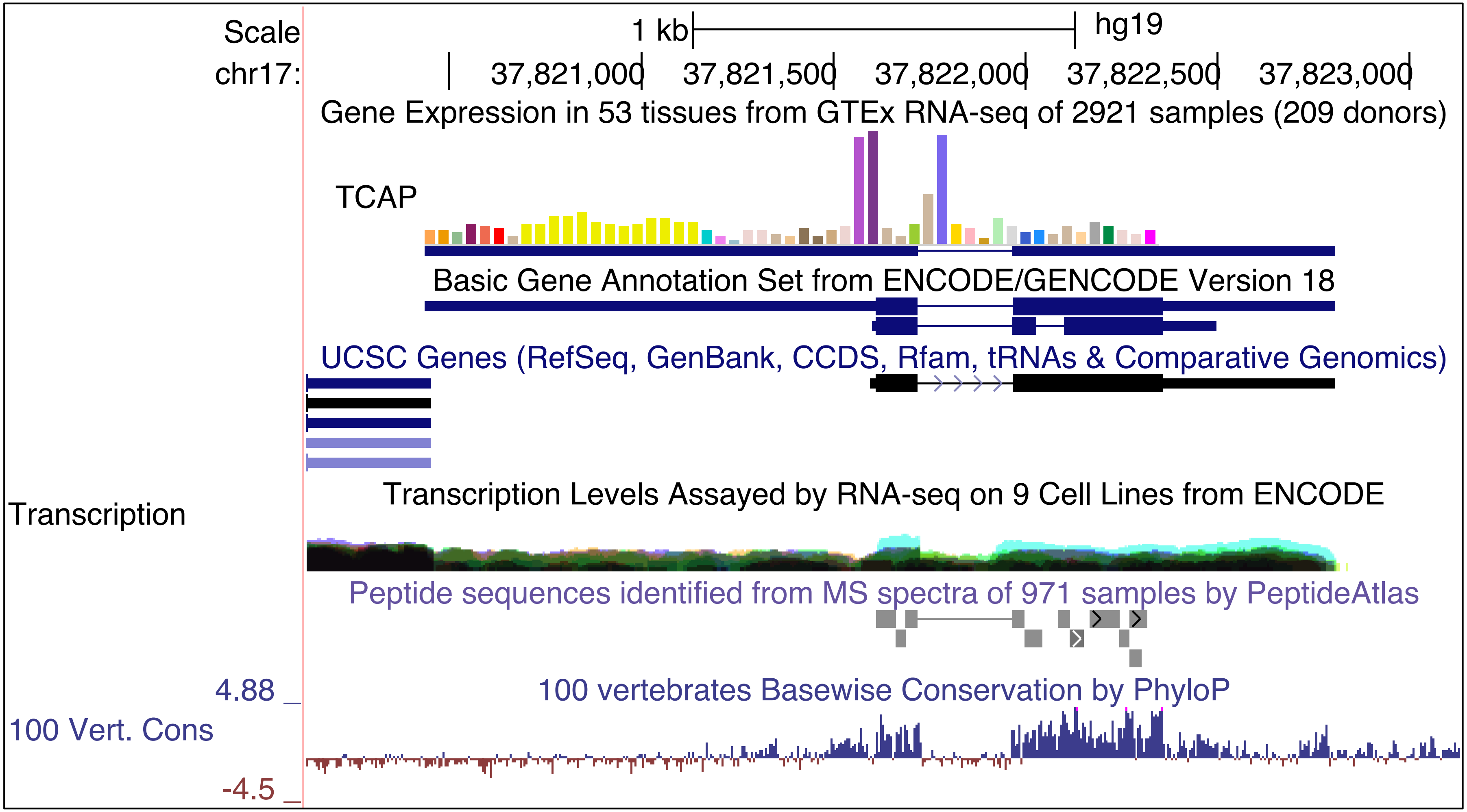
Gene Expression in 53 tissues from GTEx RNA-seq of 2921 samples (209 donors)

✓ 3. Create browser track for GTEx gene-level expression

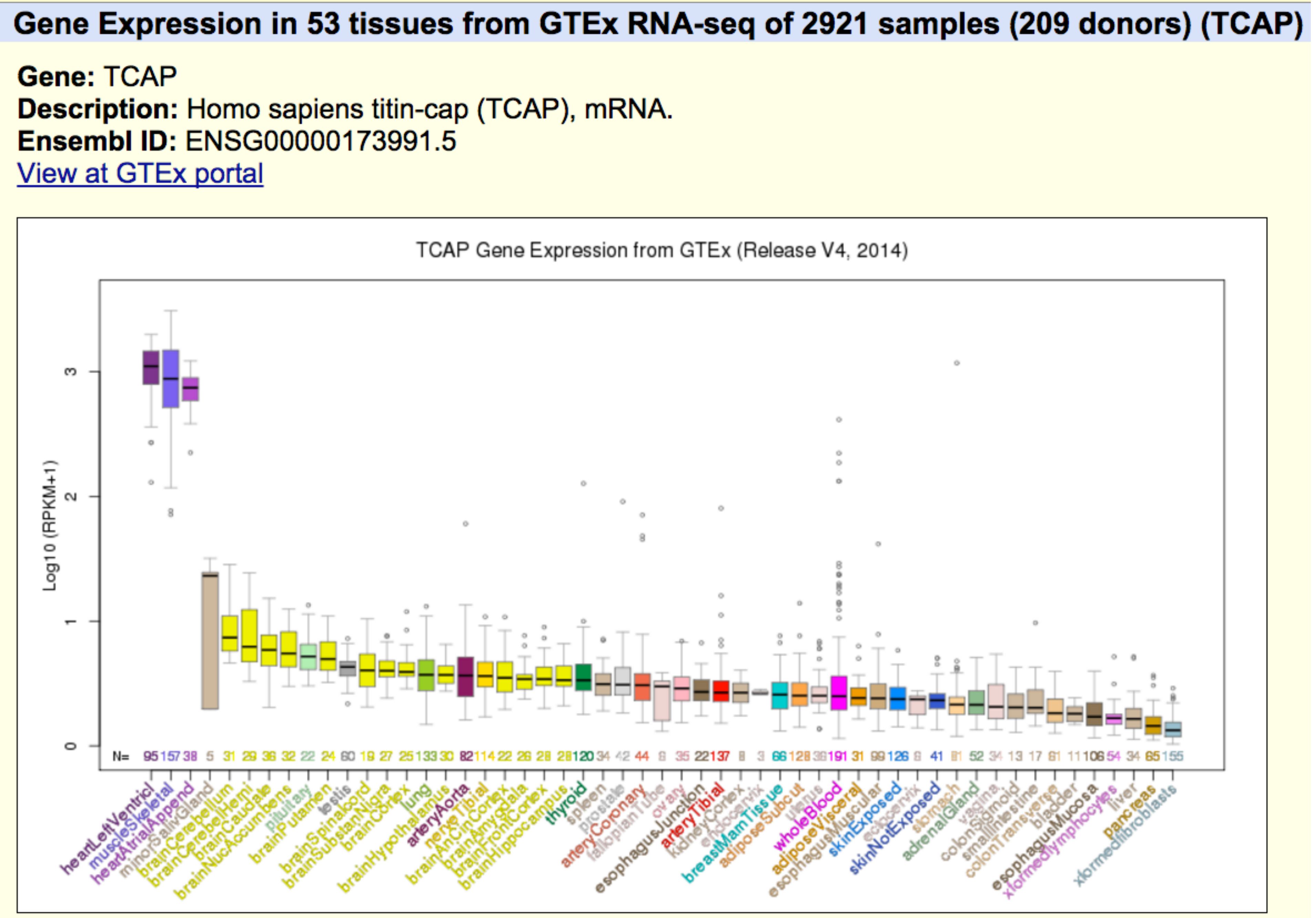
4. Create browser track for GTEx allele-specific expression

Try the GTEx gene expression track on the UCSC preview browser:
http://tinyurl.com/gtexUcsc

GTEx Gene Expression track



LCAP gene locus, showing highest expression in heart and skeletal muscle. Mouseover on bars shows tissue type and median score. Click-through shows boxplot with score ranges.

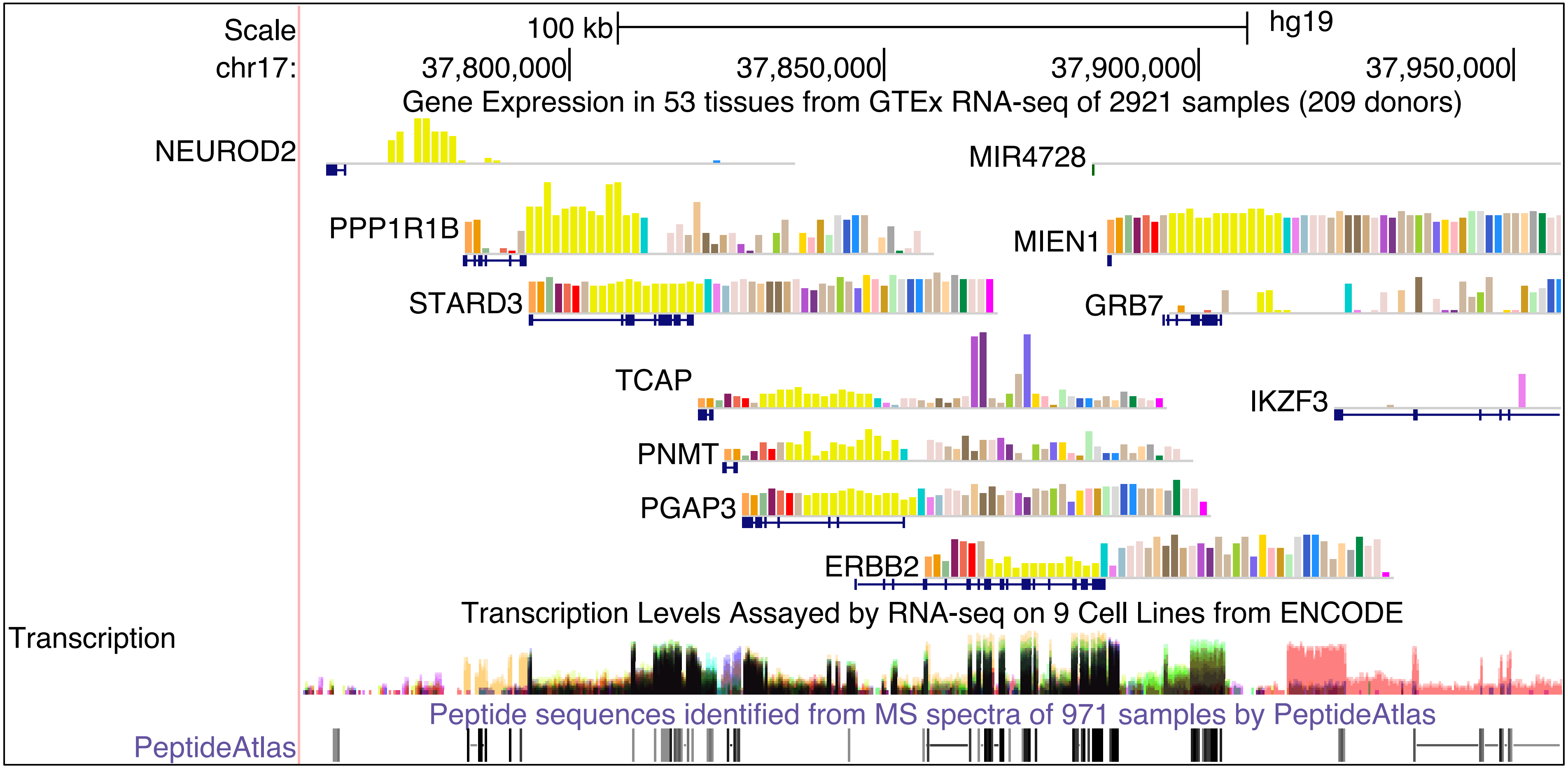


Gene Expression in 53 tissues from GTEx RNA-seq of 2921 samples (209 donors) (TCAP)

Gene: TCAP
Description: Homo sapiens titin-cap (TCAP), mRNA.
Ensembl ID: ENSG00000173991.5
[View at GTEx portal]

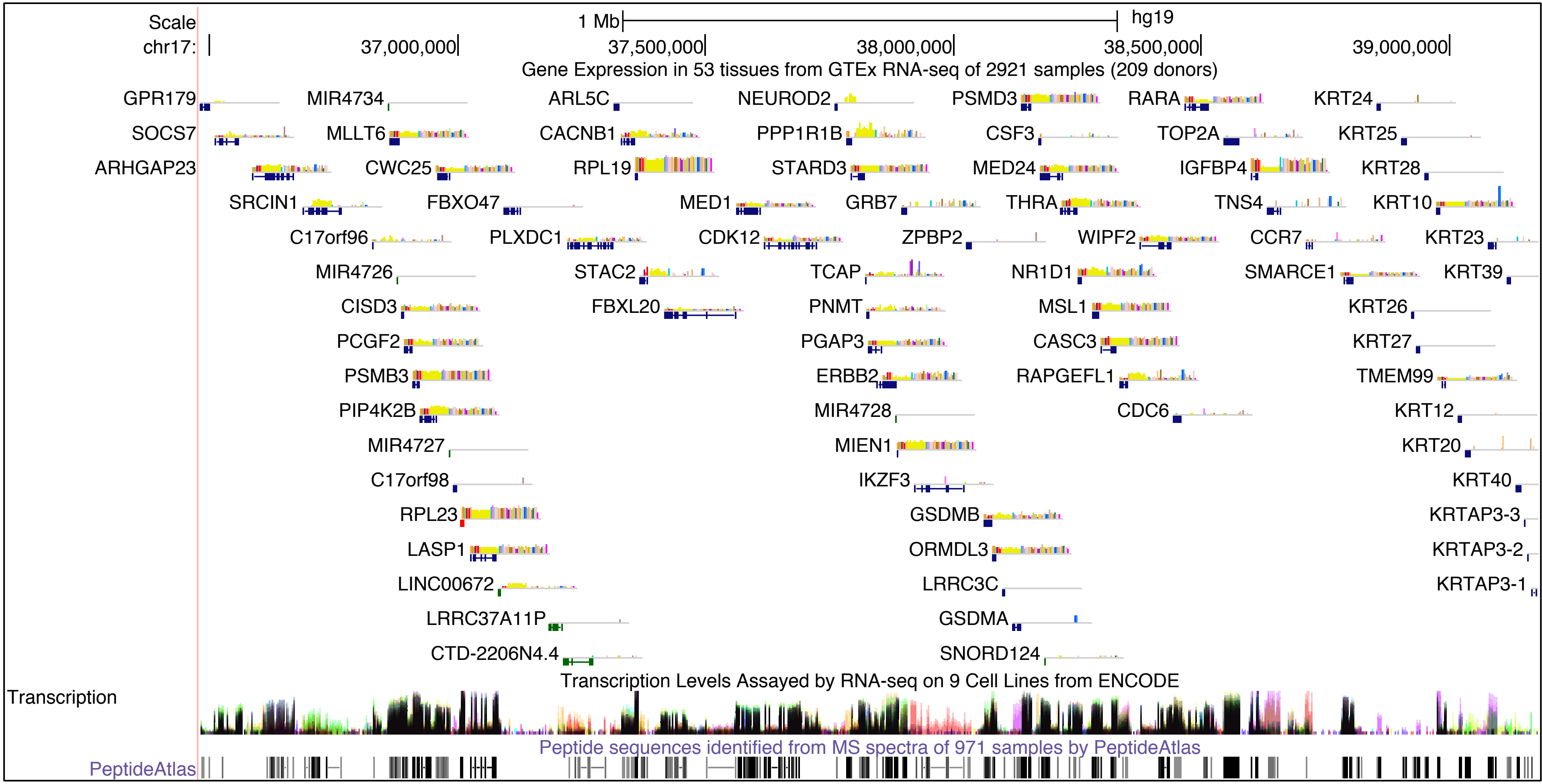
Gene expression details page

Multi-gene view



200 Kbp region of chromosome 17

Multi-megabase view

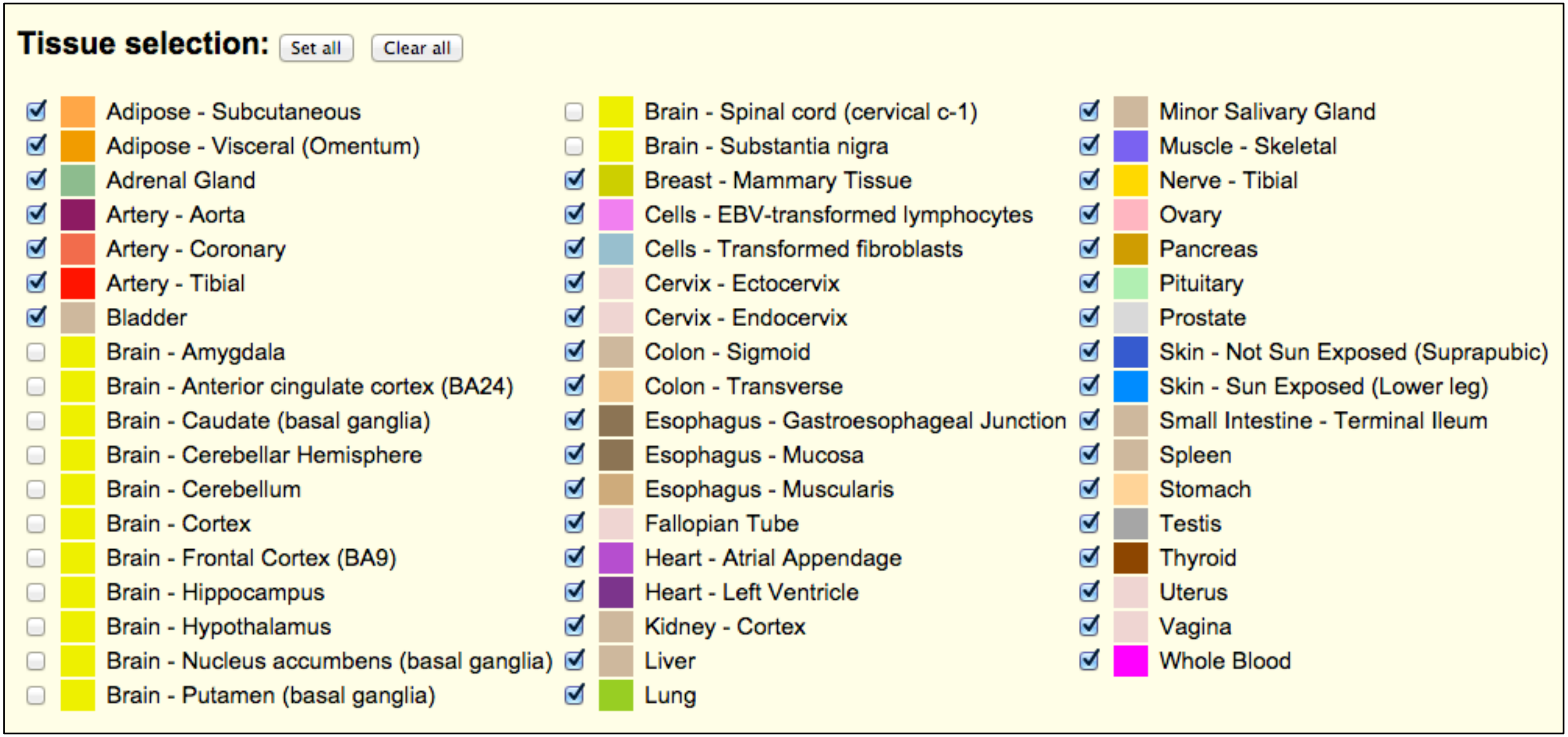


2.7 Mbp region of chromosome 17

Track features

- Displays bar graph height as median RPKM raw score (with configurable range) or log transform. Mouseovers on gene shows description; on bars show tissue and score.
- Shows sample quartiles and outliers on details page.
- Allows filtering by tissue.
- Provides comparison function to subset samples (currently, Male/Female).
- Displays comparisons as difference graph or mirror graphs.

Tissue selection

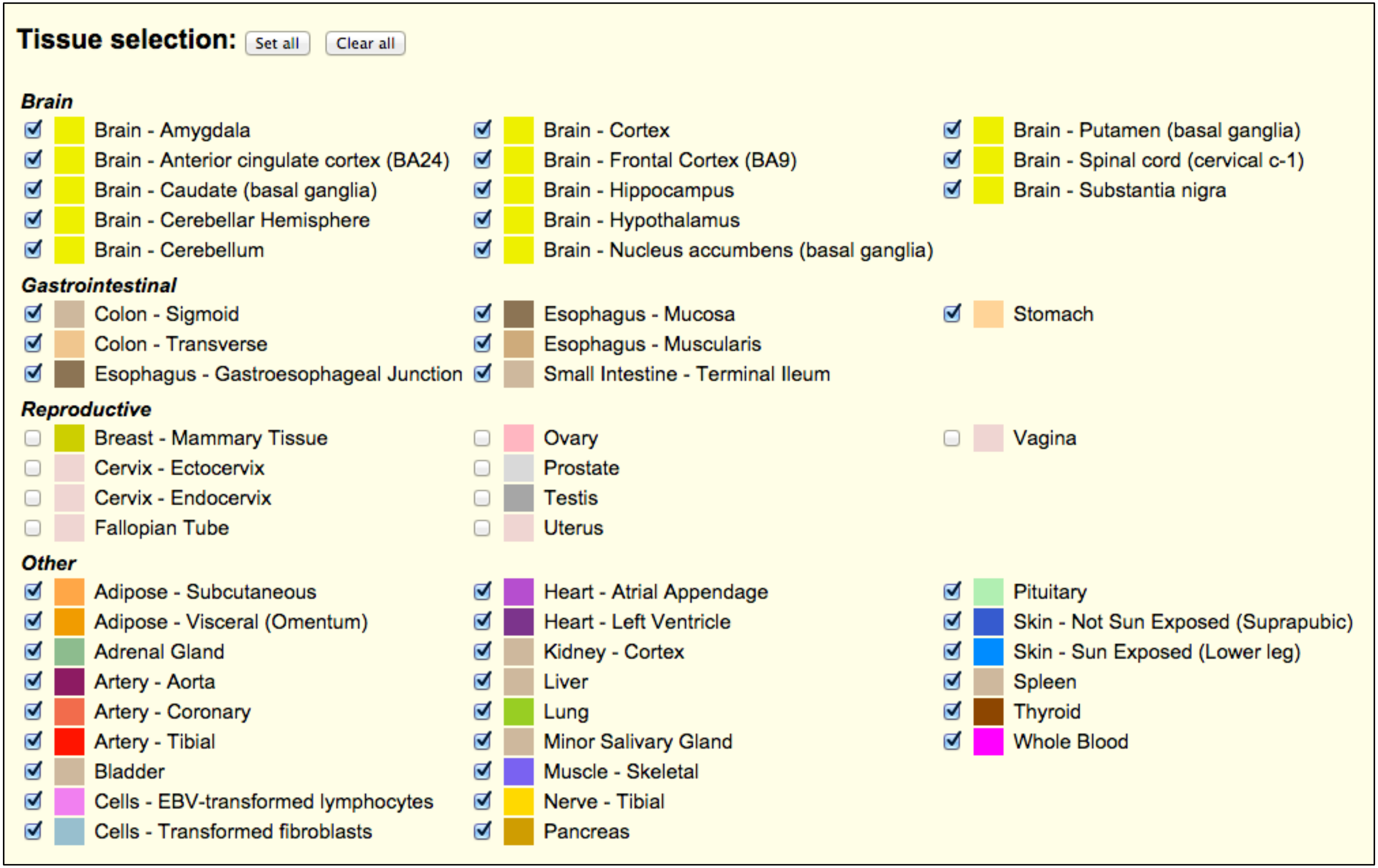


Tissue selection:

☑ Adipose - Subcutaneous
☑ Adipose - Visceral (Omentum)
☑ Adrenal Gland
☑ Artery - Aorta
☑ Artery - Coronary
☑ Artery - Tibial
☑ Bladder
☐ Brain - Amygdala
☐ Brain - Anterior cingulate cortex (BA24)
☐ Brain - Caudate (basal ganglia)
☐ Brain - Cerebellar Hemisphere
☐ Brain - Cerebellum
☐ Brain - Cortex
☐ Brain - Frontal Cortex (BA9)
☐ Brain - Hippocampus
☐ Brain - Hypothalamus
☐ Brain - Nucleus accumbens (basal ganglia)
☐ Brain - Putamen (basal ganglia)
☐ Brain - Spinal cord (cervical c-1)
☐ Brain - Substantia nigra
☑ Breast - Mammary Tissue
☑ Cells - EBV-transformed lymphocytes
☑ Cells - Transformed fibroblasts
☑ Cervix - Ectocervix
☑ Cervix - Endocervix
☑ Colon - Sigmoid
☑ Colon - Transverse
☑ Esophagus - Gastroesophageal Junction
☑ Esophagus - Mucosa
☑ Esophagus - Muscularis
☑ Fallopian Tube
☑ Heart - Atrial Appendage
☑ Heart - Left Ventricle
☑ Kidney - Cortex
☑ Liver
☑ Lung
☑ Minor Salivary Gland
☑ Muscle - Skeletal
☑ Nerve - Tibial
☑ Ovary
☑ Pancreas
☑ Pituitary
☑ Prostate
☑ Skin - Not Sun Exposed (Suprapubic)
☑ Skin - Sun Exposed (Lower leg)
☑ Small Intestine - Terminal Ileum
☑ Spleen
☑ Stomach
☑ Testis
☑ Thyroid
☑ Uterus
☑ Vagina
☑ Whole Blood

Track configuration supports user choice of tissues to display

Tissue selection, cont.



Tissue selection: Set all Clear all

Brain

- ☑ 🟨 Brain - Amygdala
- ☑ 🟨 Brain - Anterior cingulate cortex (BA24)
- ☑ 🟨 Brain - Caudate (basal ganglia)
- ☑ 🟨 Brain - Cerebellar Hemisphere
- ☑ 🟨 Brain - Cerebellum
- ☑ 🟨 Brain - Cortex
- ☑ 🟨 Brain - Frontal Cortex (BA9)
- ☑ 🟨 Brain - Hippocampus
- ☑ 🟨 Brain - Hypothalamus
- ☑ 🟨 Brain - Nucleus accumbens (basal ganglia)
- ☑ 🟨 Brain - Putamen (basal ganglia)
- ☑ 🟨 Brain - Spinal cord (cervical c-1)
- ☑ 🟨 Brain - Substantia nigra

Gastrointestinal

- ☑ 🟫 Colon - Sigmoid
- ☑ 🟧 Colon - Transverse
- ☑ 🟫 Esophagus - Gastroesophageal Junction
- ☑ 🟫 Esophagus - Mucosa
- ☑ 🟫 Esophagus - Muscularis
- ☑ 🟫 Small Intestine - Terminal Ileum
- ☑ 🟧 Stomach

Reproductive

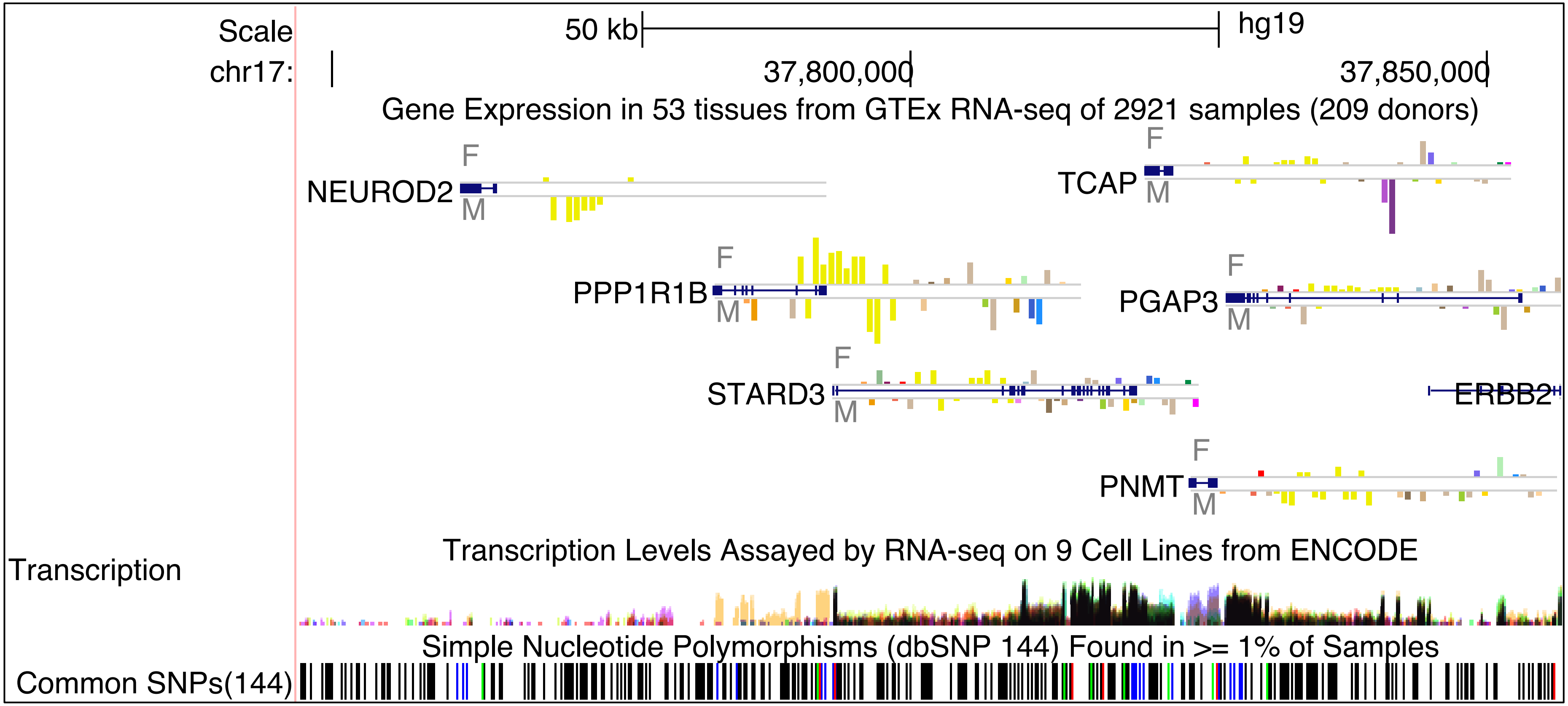
- ☐ 🟨 Breast - Mammary Tissue
- ☐ ⬜ Cervix - Ectocervix
- ☐ ⬜ Cervix - Endocervix
- ☐ ⬜ Fallopian Tube
- ☐ ⬜ Ovary
- ☐ ⬜ Prostate
- ☐ ⬜ Testis
- ☐ ⬜ Uterus
- ☐ ⬜ Vagina

Other

- ☑ 🟧 Adipose - Subcutaneous
- ☑ 🟧 Adipose - Visceral (Omentum)
- ☑ 🟩 Adrenal Gland
- ☑ 🟪 Artery - Aorta
- ☑ 🟥 Artery - Coronary
- ☑ 🟥 Artery - Tibial
- ☑ 🟫 Bladder
- ☑ 🟪 Cells - EBV-transformed lymphocytes
- ☑ 🟦 Cells - Transformed fibroblasts
- ☑ 🟪 Heart - Atrial Appendage
- ☑ 🟪 Heart - Left Ventricle
- ☑ 🟫 Kidney - Cortex
- ☑ 🟫 Liver
- ☑ 🟩 Lung
- ☑ 🟫 Minor Salivary Gland
- ☑ 🟪 Muscle - Skeletal
- ☑ 🟨 Nerve - Tibial
- ☑ 🟨 Pancreas
- ☑ 🟩 Pituitary
- ☑ 🟦 Skin - Not Sun Exposed (Suprapubic)
- ☑ 🟦 Skin - Sun Exposed (Lower leg)
- ☑ 🟫 Spleen
- ☑ 🟫 Thyroid
- ☑ 🟪 Whole Blood

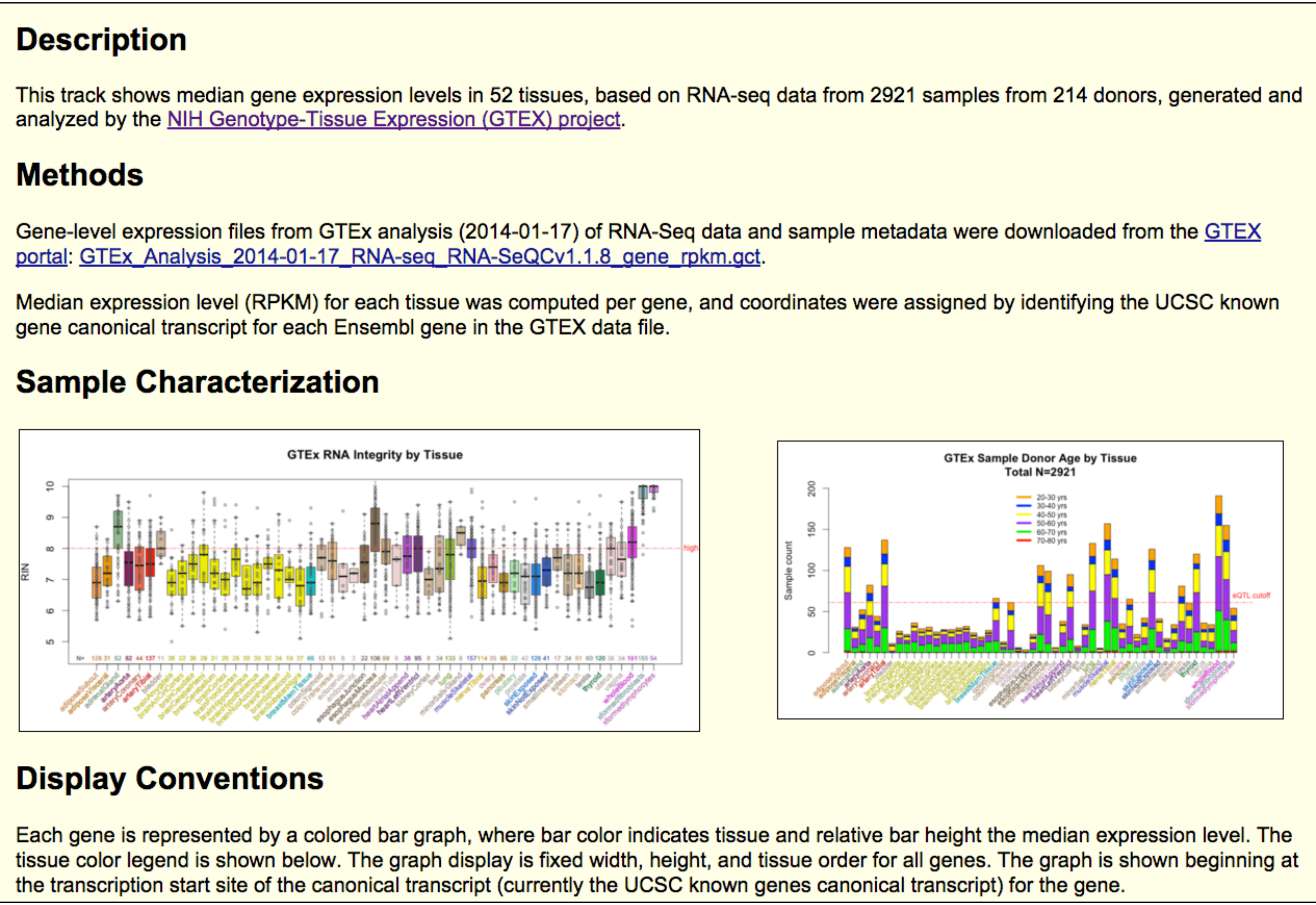
Alternative configuration panel, with tissues grouped by system

Comparisons



Gender expression differences graph. The tissue filter was applied here to exclude gender-specific tissues.

Track Description Page



Description

This track shows median gene expression levels in 52 tissues, based on RNA-seq data from 2921 samples from 214 donors, generated and analyzed by the NIH Genotype-Tissue Expression (GTEX) project.

Methods

Gene-level expression files from GTEx analysis (2014-01-17) of RNA-Seq data and sample metadata were downloaded from the GTEX portal: GTEx_Analysis_2014-01-17_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.

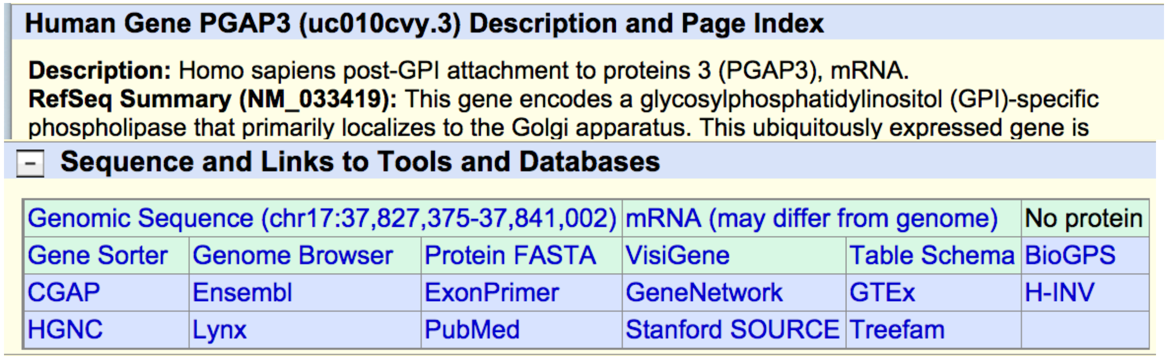
Median expression level (RPKM) for each tissue was computed per gene, and coordinates were assigned by identifying the UCSC known gene canonical transcript for each Ensembl gene in the GTEX data file.

Sample Characterization

Display Conventions

Each gene is represented by a colored bar graph, where bar color indicates tissue and relative bar height the median expression level. The tissue color legend is shown below. The graph display is fixed width, height, and tissue order for all genes. The graph is shown beginning at the transcription start site of the canonical transcript (currently the UCSC known genes canonical transcript) for the gene.

Linkouts to GTEx portal



Human Gene PGAP3 (uc010cvy.3) Description and Page Index

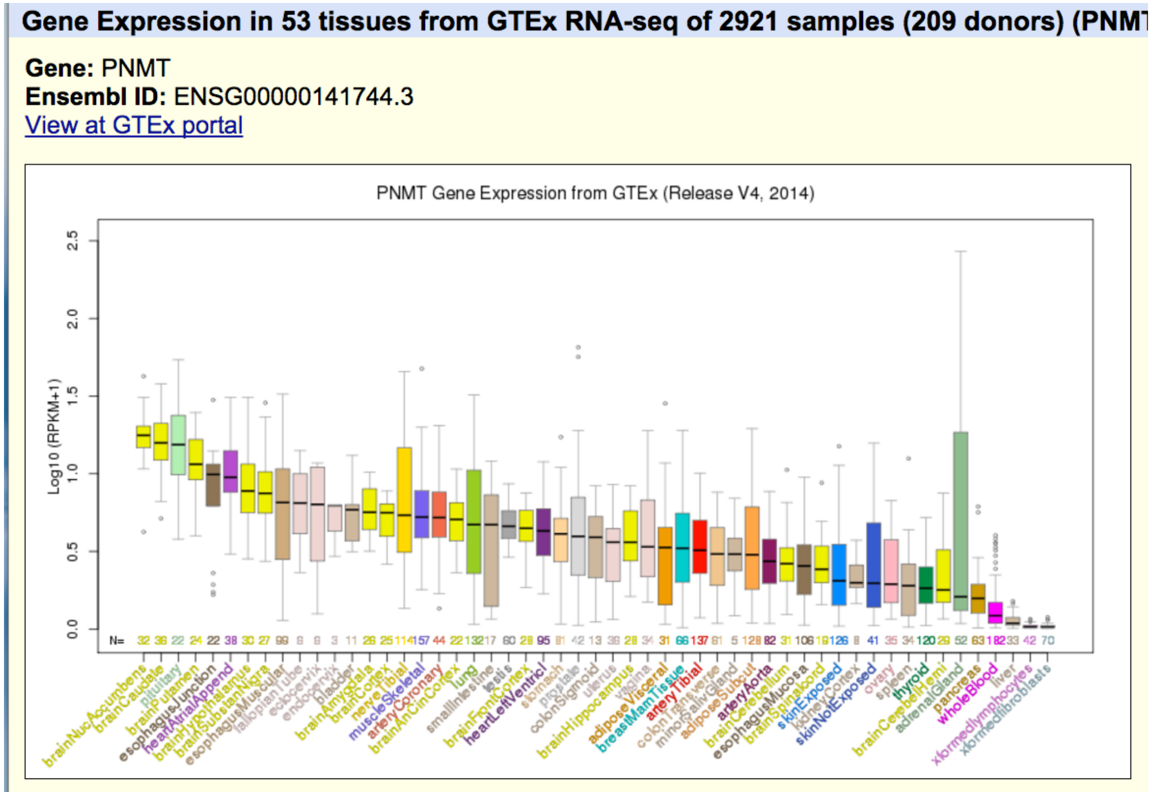
Description: Homo sapiens post-GPI attachment to proteins 3 (PGAP3), mRNA.
RefSeq Summary (NM_033419): This gene encodes a glycosylphosphatidylinositol (GPI)-specific phospholipase that primarily localizes to the Golgi apparatus. This ubiquitously expressed gene is

Sequence and Links to Tools and Databases

Genomic Sequence (chr17:37,827,375-37,841,002)			mRNA (may differ from genome)		No protein
Gene Sorter	Genome Browser	Protein FASTA	VisiGene	Table Schema	BioGPS
CGAP	Ensembl	ExonPrimer	GeneNetwork	GTEx	H-INV
HGNC	Lynx	PubMed	Stanford SOURCE	Treefam	

UCSC/GENCODE genes detail page

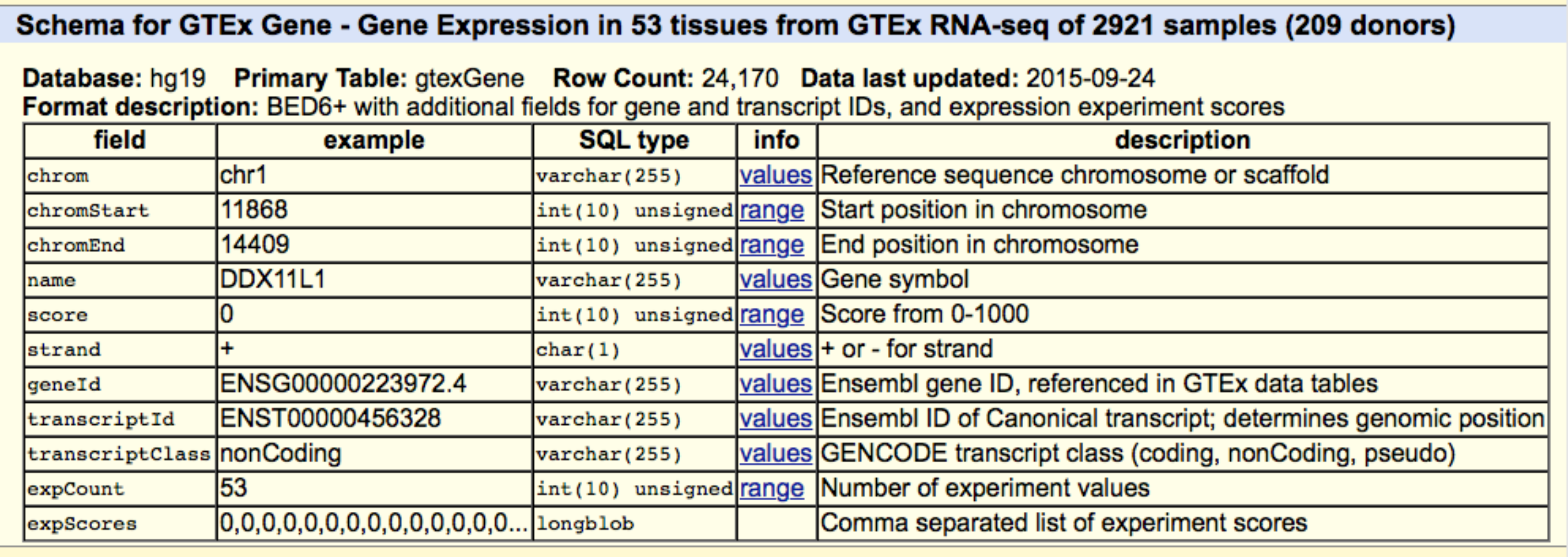
GTEx gene expression details page



Gene Expression in 53 tissues from GTEx RNA-seq of 2921 samples (209 donors) (PNM

Gene: PNMT
Ensembl ID: ENSG00000141744.3
View at GTEx portal

Under the covers:
GTEx Database tables



Schema for GTEx Gene - Gene Expression in 53 tissues from GTEx RNA-seq of 2921 samples (209 donors)

Database: hg19 **Primary Table:** gtexGene **Row Count:** 24,170 **Data last updated:** 2015-09-24
Format description: BED6+ with additional fields for gene and transcript IDs, and expression experiment scores

field	example	SQL type	info	description
`chrom`	chr1	`varchar(255)`	values	Reference sequence chromosome or scaffold
`chromStart`	11868	`int(10) unsigned`	range	Start position in chromosome
`chromEnd`	14409	`int(10) unsigned`	range	End position in chromosome
`name`	DDX11L1	`varchar(255)`	values	Gene symbol
`score`	0	`int(10) unsigned`	range	Score from 0-1000
`strand`	+	`char(1)`	values	+ or - for strand
`geneId`	ENSG00000223972.4	`varchar(255)`	values	Ensembl gene ID, referenced in GTEx data tables
`transcriptId`	ENST00000456328	`varchar(255)`	values	Ensembl ID of Canonical transcript; determines genomic position
`transcriptClass`	nonCoding	`varchar(255)`	values	GENCODE transcript class (coding, nonCoding, pseudo)
`expCount`	53	`int(10) unsigned`	range	Number of experiment values
`expScores`	0,0,0,0,0,0,0,0,0,0,0,0,0,0...	`longblob`		Comma separated list of experiment scores

Primary table for GTEx Gene Expression track. This is a summary table of tissue medians Per gene, anchored to the genomic location of the GTEx union transcript for the gene.

Related tables:

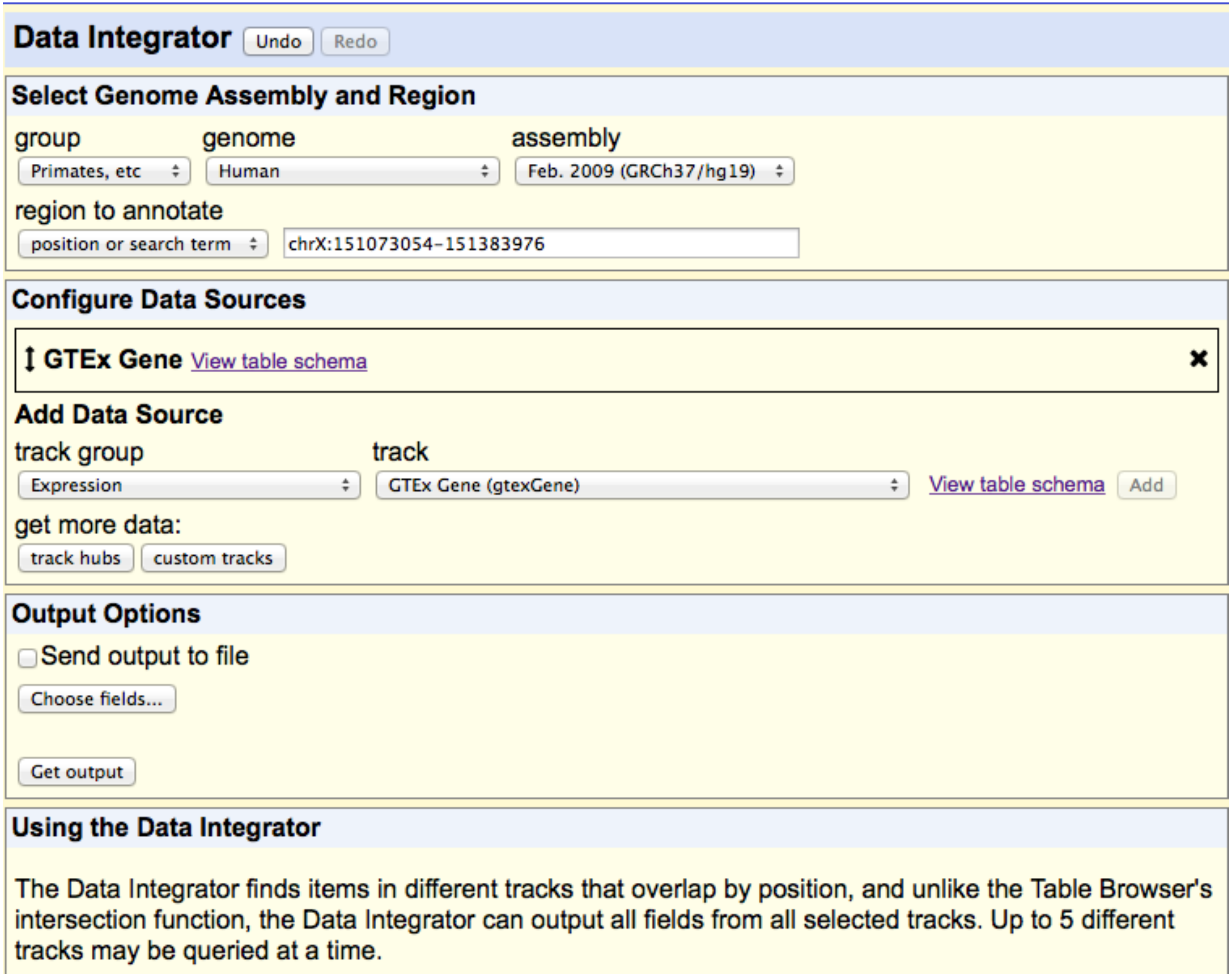
Data: gtexSampleData (RPKM scores for each gene by sample)
Metadata: gtexTissue, gtexSample, gtexDonor

Data mining & analysis: UCSC public MySQL server

```
$ mysql —user=genome —host=genome-mysql.cse.ucsc.edu —A \
   hgFixed -e ‘select sampleId, tissue, gender, age,\
   deathClass, ischemicTime, autolysisScore, rin,\
   collectionSites, batchId, isolationDate from \
   gtexSample, gtexDonor where \
   gtexSample.donor=gtexDonor.name' | \
      sed -e 's/ //' >   sampleDf.txt
$ R
>sampleDf <- read.table("sampleDf.txt", sep="\t",
   header=TRUE)
```

Example: Query GTEx metadata tables to create an R dataframe

Data mining & analysis: New tool for database queries/intersections



Data Integrator

Select Genome Assembly and Region

group: Primates, etc
genome: Human
assembly: Feb. 2009 (GRCh37/hg19)
region to annotate: position or search term: chrX:151073054-151383976

Configure Data Sources

GTEx Gene View table schema

Add Data Source

track group: Expression
track: GTEx Gene (gtexGene)

get more data:

Output Options

☐ Send output to file

Using the Data Integrator

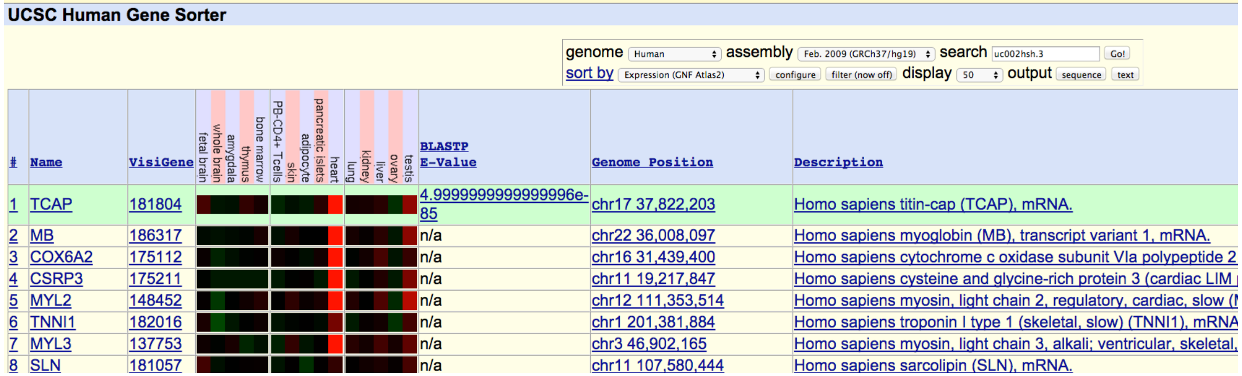
The Data Integrator finds items in different tracks that overlap by position, and unlike the Table Browser's intersection function, the Data Integrator can output all fields from all selected tracks. Up to 5 different tracks may be queried at a time.

Plans

- Release GTEx gene expression tracks on hg19 and hg38 (V4, then V6)
- Implement GTEx allele-specific expression track:
  - Display will be based on difference bar graph, as in gene expression track
  - Initial design will locate graph at eQTL SNP’s
  - Your input welcome!

Other plans

- Gene sorter



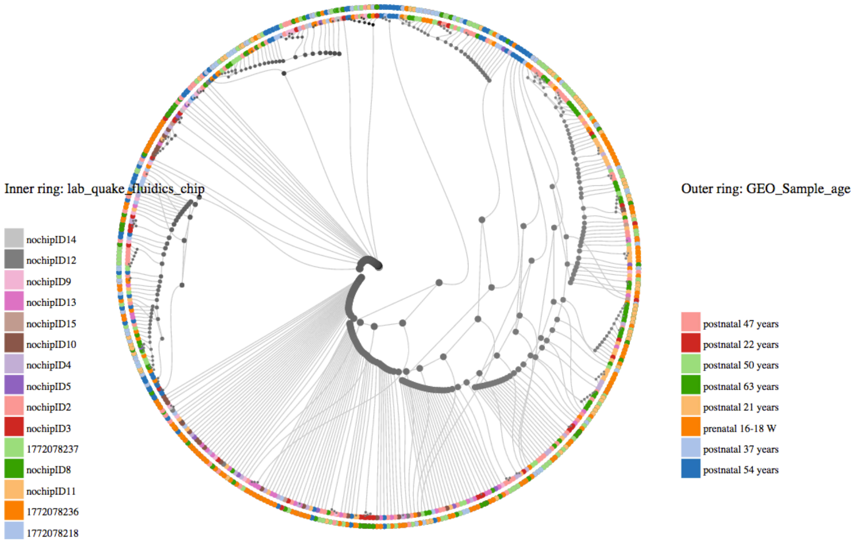
UCSC Human Gene Sorter

genome: Human; assembly: Feb. 2009 (GRCh37/hg19); search: uc002hsh.3

sort by: Expression (GNF Atlas2); filter (now off); display: 50

#	Name	VisiGene	fetal brain	whole brain	amygdala	thymus	bone marrow	PB-CD4+ Tcells	skin	adipocyte	pancreatic islets	heart	lung	kidney	liver	ovary	testis	BLASTP E-Value	Genome Position	Description
1	TCAP	181804																4.9999999999999996e-85	chr17 37,822,203	Homo sapiens titin-cap (TCAP), mRNA.
2	MB	186317																n/a	chr22 36,008,097	Homo sapiens myoglobin (MB), transcript variant 1, mRNA.
3	COX6A2	175112																n/a	chr16 31,439,400	Homo sapiens cytochrome c oxidase subunit VIa polypeptide 2
4	CSRP3	175211																n/a	chr11 19,217,847	Homo sapiens cysteine and glycine-rich protein 3 (cardiac LIM
5	MYL2	148452																n/a	chr12 111,353,514	Homo sapiens myosin, light chain 2, regulatory, cardiac, slow (
6	TNNI1	182016																n/a	chr1 201,381,884	Homo sapiens troponin I type 1 (skeletal, slow) (TNNI1), mRNA
7	MYL3	137753																n/a	chr3 46,902,165	Homo sapiens myosin, light chain 3, alkali; ventricular, skeletal,
8	SLN	181057																n/a	chr11 107,580,444	Homo sapiens sarcolipin (SLN), mRNA.

- Tree of cells visualization



Community data: GTEx data hub ?

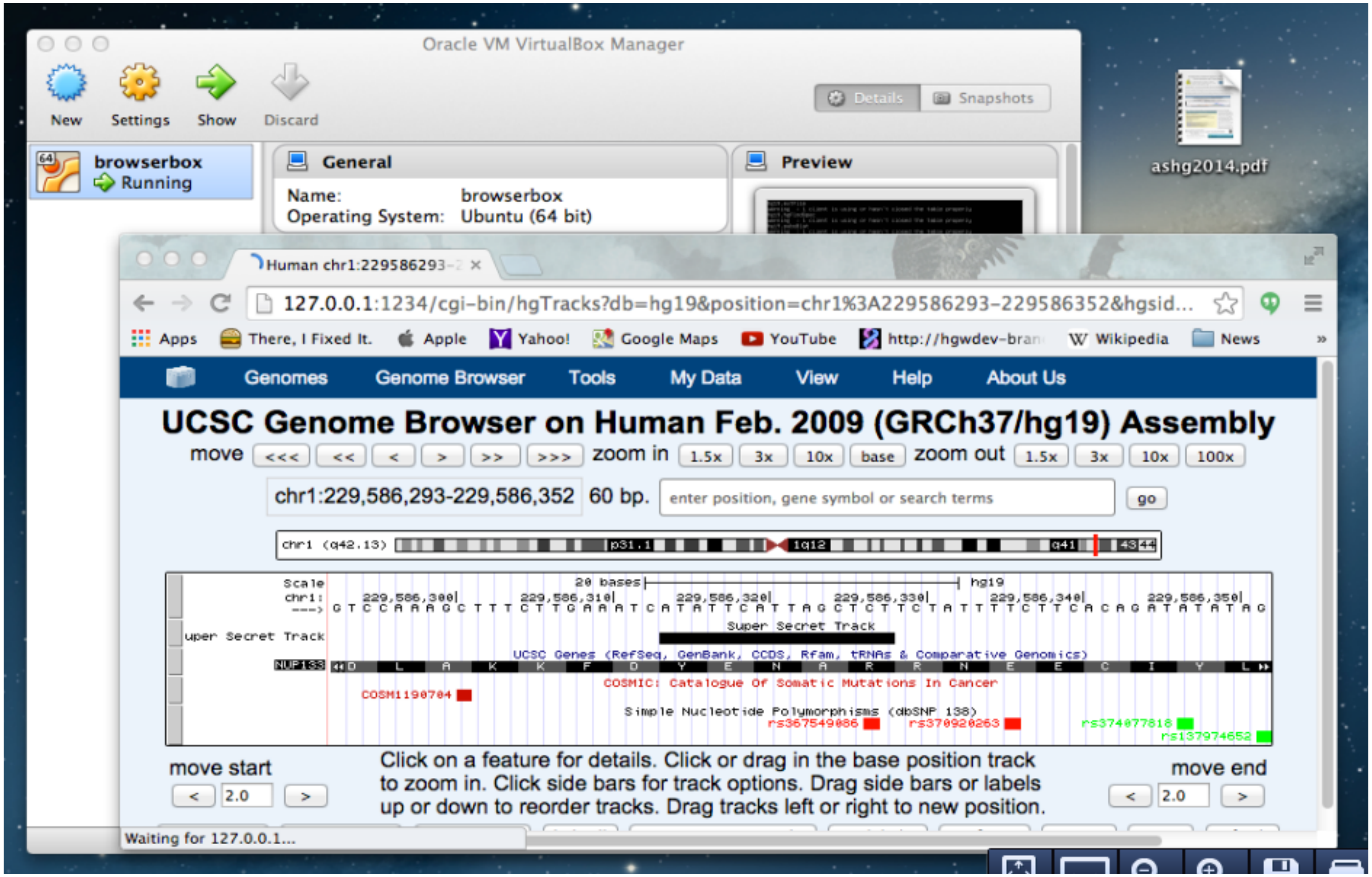
- RNA-seq signal tracks for all samples
- Exon-level expression
- Transcript-level expression
- Proteomics ?

-> Can display on UCSC or Ensembl (and later, NCBI sequence viewer)

-> Can view and intersect eQtl's with regulatory data from ENCODE, Epigenomics Roadmap, etc.

-> Can view and intersect with variant and medical annotations, comparative genomics

GTEx + private data:
Genome Browser in a Box (GBiB)



Name: browserbox
Operating System: Ubuntu (64 bit)

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr1:229,586,293-229,586,352 60 bp. enter position, gene symbol or search terms

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

GBiB running on a Macbook, with private data loaded as a custom track. The GBiB preinstalled image is run in a virtual machine so data can be kept local. UCSC tracks can be mirrored locally or retrieved from the UCSC public MySQL server.

Acknowledgements

Colleagues: UCSC Genome Browser group

PI: Jim Kent

Collaborators: GTEx Consortium

Funding: NHGRI (5 U41 HG002371 to UCSC Center for Genomic Science)

UCSC Browser help

Send us a message to our public mailing list:
genome@soe.ucsc.edu

View training info, videos, and user's guides:
http://genome.ucsc.edu/training/

Find us on:

Genome Browser

@GenomeBrowser

UCSC Genome Browser