# GTEx in the UCSC Genome Browser

*Kate Rosenbloom*

*UCSC Genome Browser Group*

UNIVERSITY OF CALIFORNIA SANTA CRUZ | Genomics Institute

*February 2016*

# The GTEx project



https://commonfund.nih.gov/GTEx

NIH Common Fund sponsored initiative launched in September 2010 to create a biobank and database resources for studies on the relationship between genetic variation and gene expression in multiple human tissues.   *Aim:  50+ tissues in 1000 individuals.*
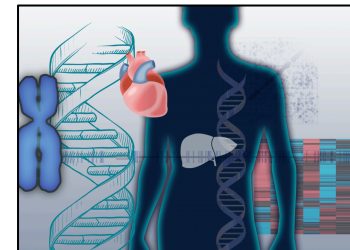
 NIH leads:  Jeff Streuwing and Simona Volpi (NHGRI)

## *The Genotype-Tissue Expression (GTEx) project*

# Goals

- To create a data resource to enable the systematic study of genetic variation and the regulation of gene expression in multiple reference human tissues

- To provide the scientific community with a biospecimen resource including tissues, nucleic acids and cell lines upon which to determine other molecular phenotypes

- To support and disseminate the results of a study of the ethical, legal and social issues related to donor recruitment and consent

- To support the development of novel statistical methods for the analysis of human eQTLs, alone and in the context of other molecular phenotypes

- To make data available to the research community as rapidly as possible

- To support the dissemination of knowledge, standards and protocols related to biospecimen collection and analysis methods developed during the project

# Status

**GTEx Reaches Midpoint Milestone!**
The GTExPortal ⤢ was updated in October 2015! This latest version of sequence data encompasses roughly half of the anticipated 960 postmortem donors. This release includes genotype data from approximately 450 donors and over 9600 RNA-seq samples

# Data: GTEx analysis releases

- Jun 2013: Freeze: Pilot dataset (9 tissues, 1641 samples, 178 donors)
  -> First round of publications

- Jun 2014: Gene and transcript expression levels for 53 tissues, 2921 samples, 214 donors (V4)

- Jan 2015: eQTL's for 9 tissues

- Oct 2015: Freeze: Midpoint (~7300 samples, 550 donors) (V6) ->
  -> Second round of publications

- Mid 2016: Freeze: Full dataset (15-20,000 samples, 900 donors)

V4 and V6 data are available on UCSC dev/preview servers. V6 will be public.

# GTEX Sample Metadata

- 60 variables reported

- From pathology reports
  - autolysis score
  - ischemic time
  - tissue quality

- 11 from sample processing and sequencing

- 45 metrics from analysis pipeline

Output of the RNA-seQC GenePattern module:
www.broadinstitute.org/cancer/cga/tools/rnaseqc/RNA-
SeQC_Help_v1.1.2.pdf

Data and visualization site at the Broad LDACC,     http://gtexportal.org
PI's:  Gad Getz, Kristin Ardlie

* Gene and transcript expression levels, eQTL's are hosted at GTEX portal
* RNA-seq reads and genotypes hosted at dbGAP

# UCSC 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
BM-CD105+ endothel
BM-CD34+
whole blood
PB-BDCA4+ dentritic
PB-CD14+ monocyt
PB-CD56+ NKCells
PB-CD4+ Tcells
PB-CD8+ Tcells
PB-CD19+ Bcells
uk lymphoblast(molt4)
721 B lymphoblast
lymphoma Burk Raji
uk promyelocyt(hl60)
ymphoma Burk Daudi
k chronic myel(k562)
colorect adenocarc
appendix

OXT:  Oxytocin precursor
Expression from GNF Atlas2

# UCSC GTEx  Grant Aims

The NIH Common Fund's Genotype-Tissue Expression (GTEx) program measures genotype and gene expression in multiple tissues for many human donors.

- UCSC will integrate GTEx data into the Genome Browser. The integration will include at least two new tracks in the Genome Browser, one that shows the tissue-by-tissue expression data associated with each gene, and another that shows significant allele-specific differences in expression. We will also add new columns for the GTEx data to the Gene Sorter, and new panels to the existing gene details display.

- UCSC will update the GTEx data in the Genome Browser at least once yearly, and will synchronize new releases to match major data releases from GTEx when possible.  UCSC will collaborate with the GTEx analysis team, integrating new analyses as well as new data, and advising on analysis methods when useful.

# 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

4. Integrate GTEx expression into browser gene details

5. Add GTEx expression columns to the Gene Sorter

# 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)

OXT

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

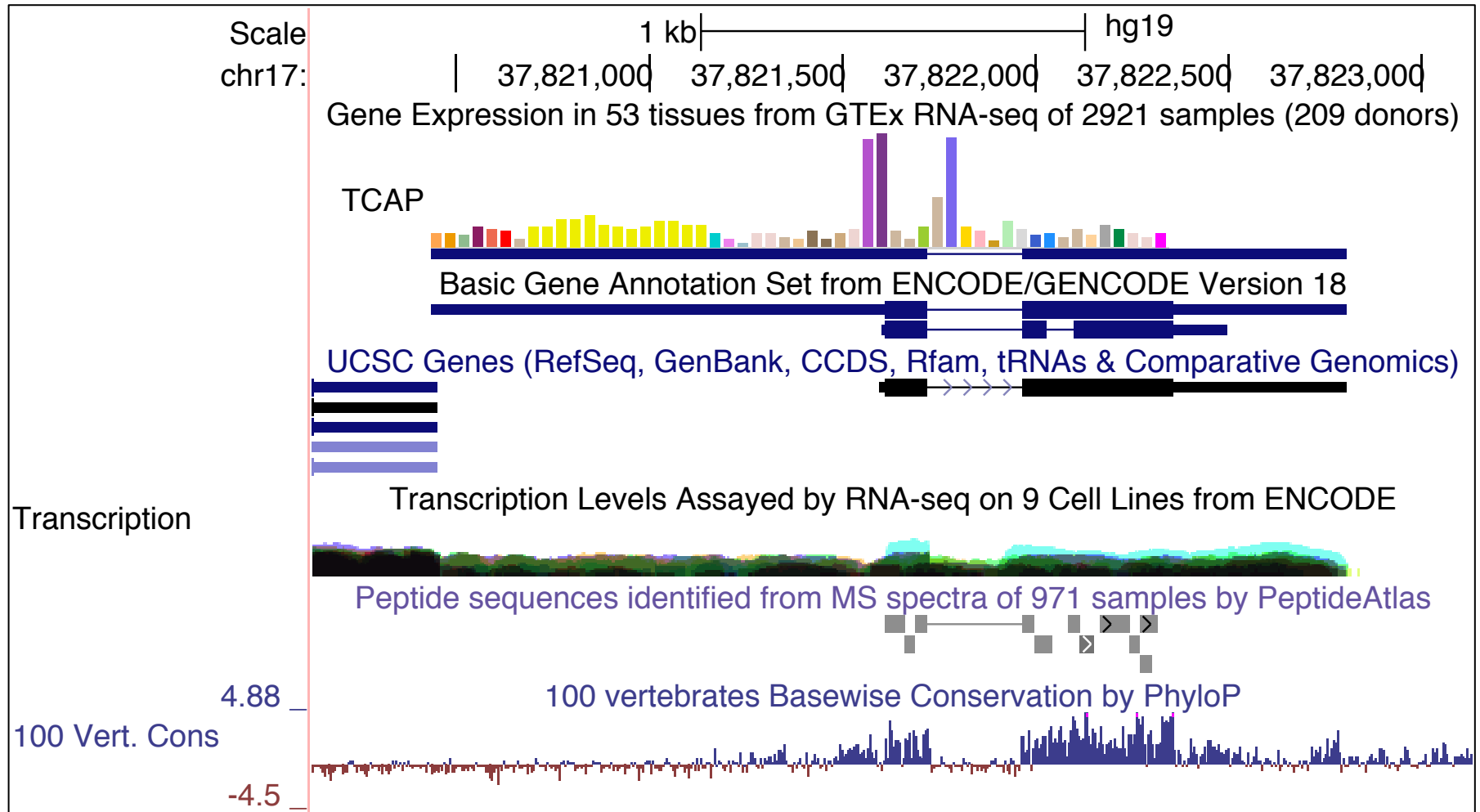4.  Create browser track for GTEx allele-specific expression

5.   Integrate GTEx expression into browser gene details

6.  Add GTEx expression columns to the Gene Sorter

-> *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 8555 samples from 570 donors (TCAP)

**Gene:** TCAP
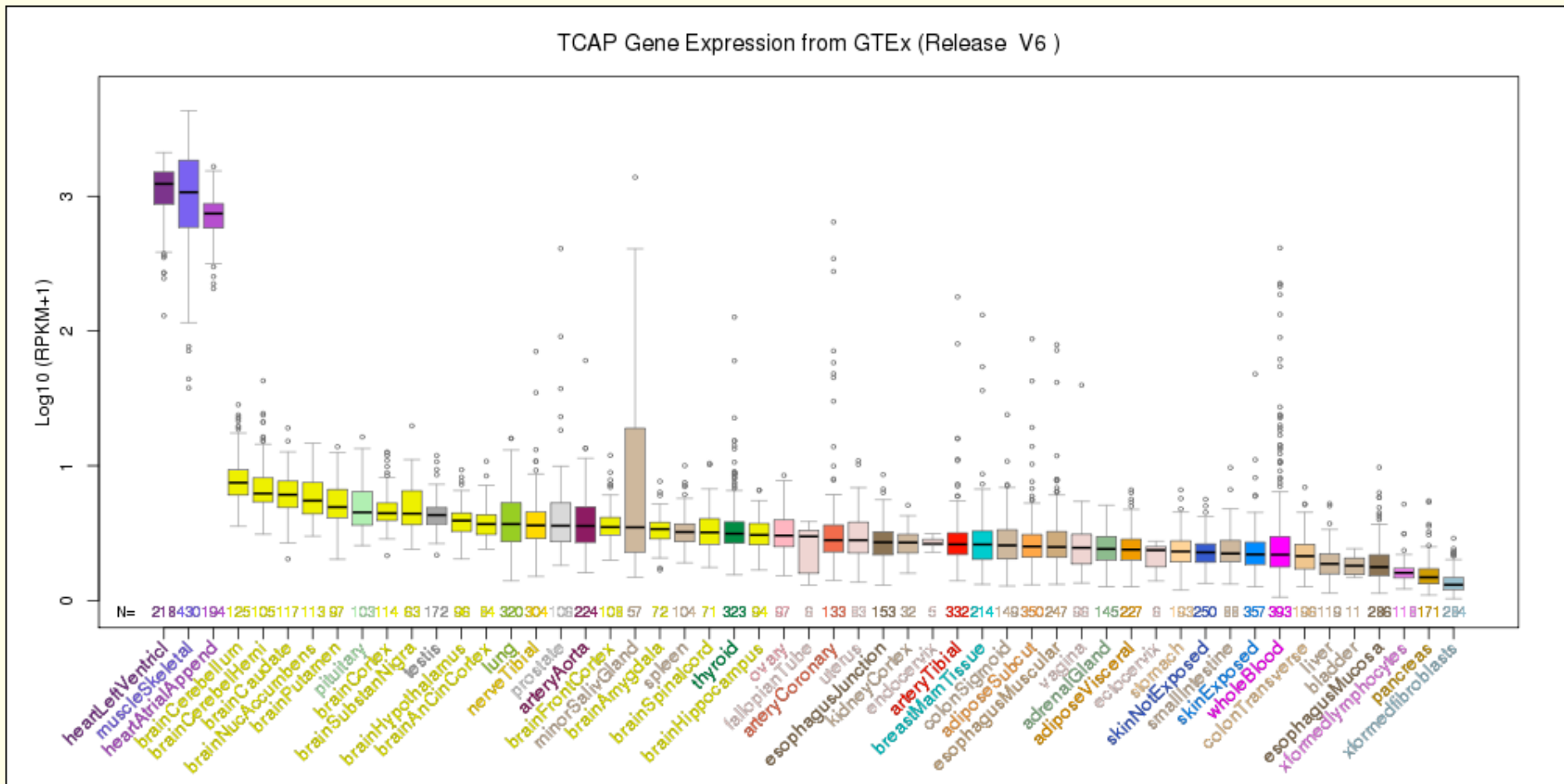**Description:** Homo sapiens titin-cap (TCAP), mRNA.
**Ensembl Gene ID:** ENSG00000173991.5
**Ensembl Transcript ID:** ENST00000309889.2
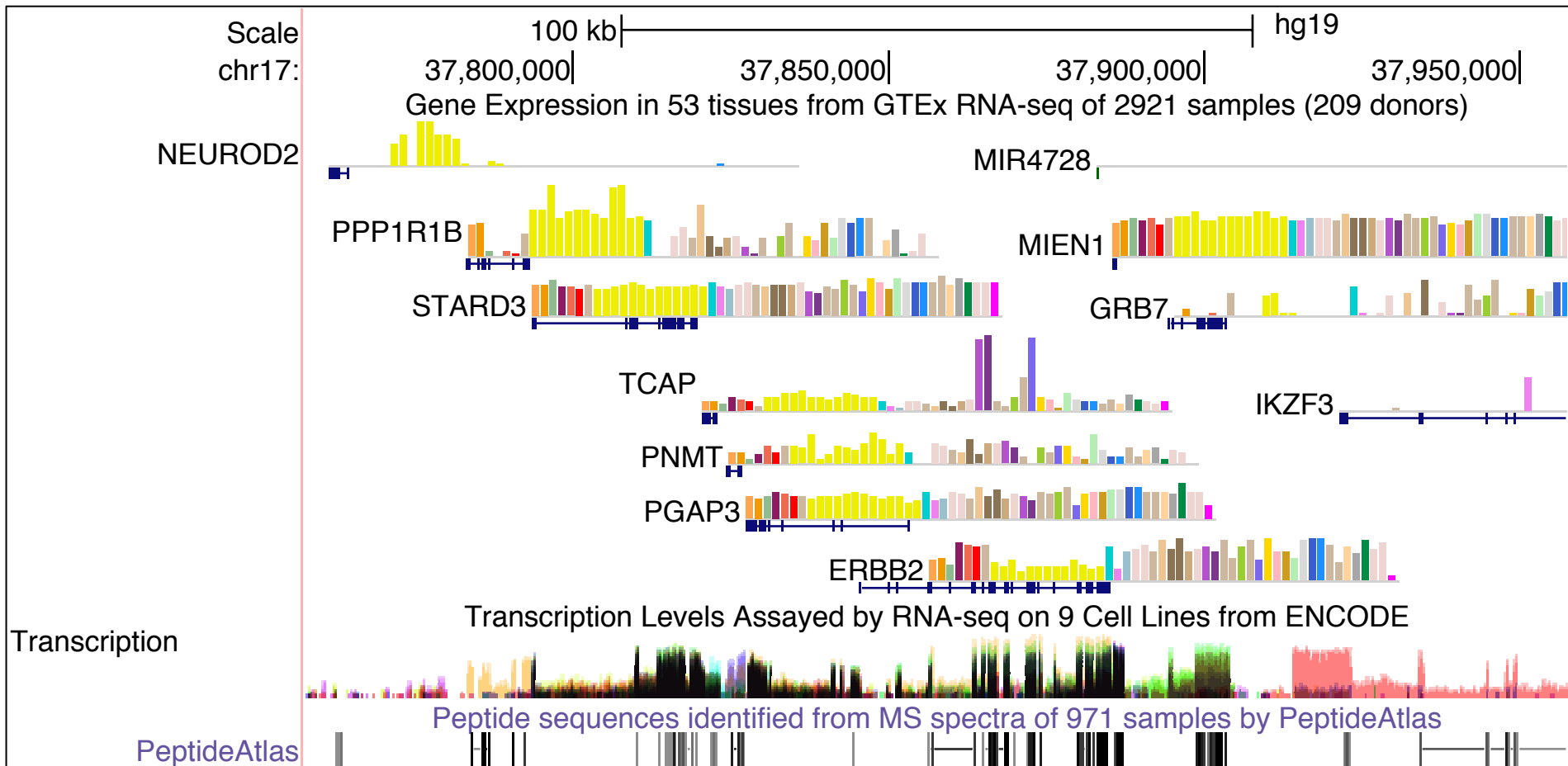**Ensembl Transcript Class:** coding
**Genomic Position:** chr17:37820440-37822808
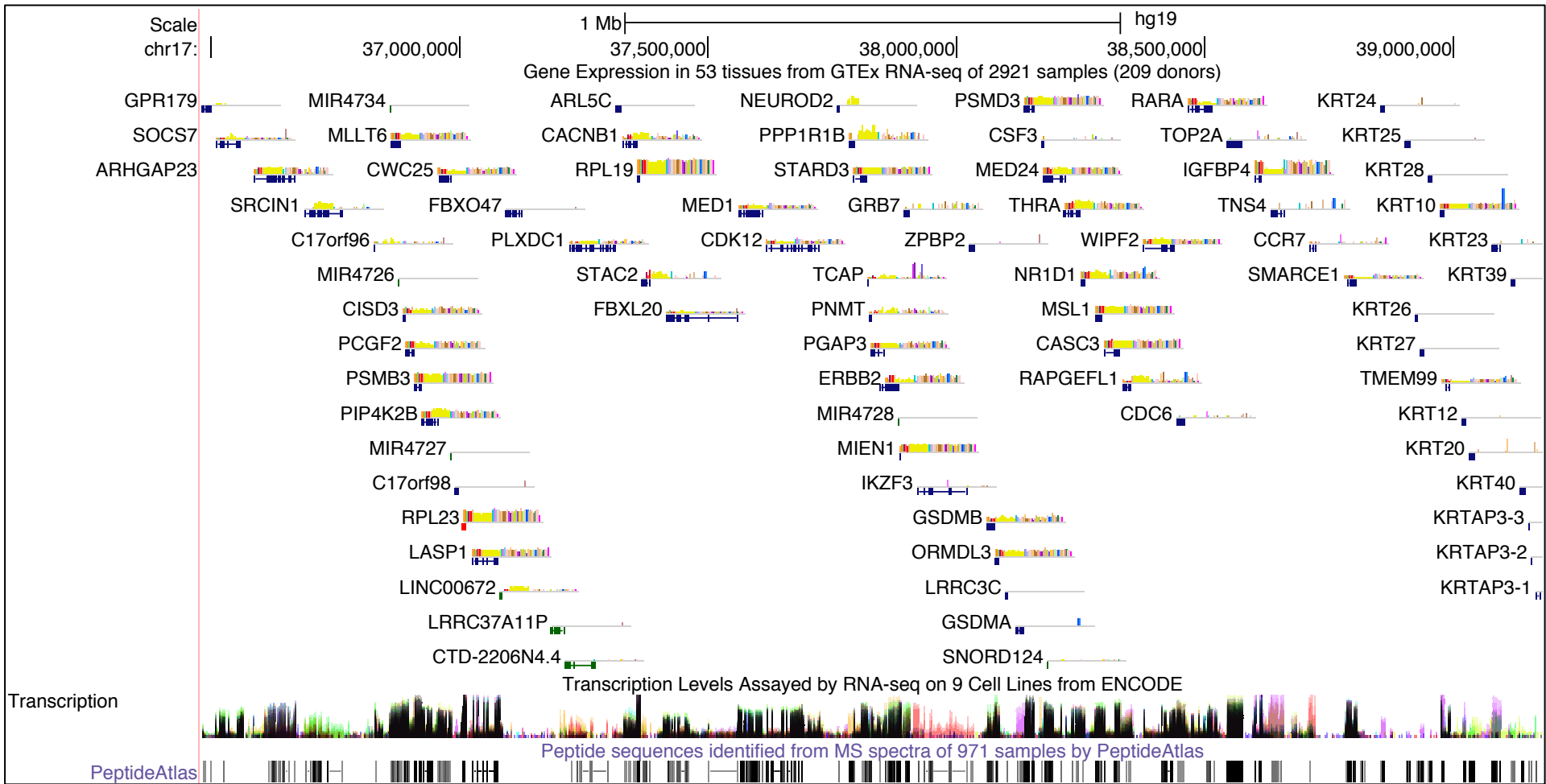View at GTEx portal



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

- Main browser displays a bargraph per gene, with bar height for each tissue based on the median expression level (RPKM)

- Mouseovers on gene shows description; on bars show tissue and score.

- Graph is configurable to draw bar heights based on raw score (with max limit selectable) or log transform

- Supports tissue selection via sortable table

- Provides comparison function to subset samples (currently, Male/Female). Comparisons are shown as a difference graph or mirror graphs.

- Details page shows boxplot of gene expression, including quartiles and outliers, on details page.

# Track Configuration



**GTEx Gene Expr Track Settings**

## Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples from 570 donors (▲All Expression tracks)

**Display mode:** [pack ▼] [Submit]

**Log10 transform:** ☑  **View limits maximum:** [20] RPKM (range 10-180000)

**Samples:** ○ All ● Compare by gender

**Comparison display:** ○ Difference graph ● Two graphs

Sample comparison

**Tissues:** [Set all] [Clear all]

Tissue selection

| Tissue↓1 | Samples↑4 | Organ↑2 | System↑3 |
|---|---|---|---|
| Adipose - Subcutaneous | 350 | Adipose Tissue | Other |
| Adipose - Visceral (Omentum) | 227 | Adipose Tissue | Other |
| Adrenal Gland | 145 | Adrenal Gland | Endocrine |
| Artery - Aorta | 224 | Blood Vessel | Cardiovascular |
| Artery - Coronary | 133 | Blood Vessel | Cardiovascular |
| Artery - Tibial | 332 | Blood Vessel | Cardiovascular |
| Bladder | 11 | Bladder | Other |
| Brain - Amygdala | 72 | Brain | Brain |
| Brain - Anterior cingulate cortex (BA24) | 84 | Brain | Brain |
| Brain - Caudate (basal ganglia) | 117 | Brain | Brain |
| Brain - Cerebellar Hemisphere | 105 | Brain | Brain |
| Brain - Cerebellum | 125 | Brain | Brain |
| Brain - Cortex | 114 | Brain | Brain |
| Brain - Frontal Cortex (BA9) | 108 | Brain | Brain |
| Brain - Hippocampus | 94 | Brain | Brain |
| Brain - Hypothalamus | 96 | Brain | Brain |
| Brain - Nucleus accumbens (basal ganglia) | 113 | Brain | Brain |
| Brain - Putamen (basal ganglia) | 97 | Brain | Brain |
| Brain - Spinal cord (cervical c-1) | 71 | Brain | Brain |
| Brain - Substantia nigra | 63 | Brain | Brain |
| Breast - Mammary Tissue | 214 | Breast | Reproductive |
| Cells - EBV-transformed lymphocytes | 118 | Blood | Other |

# Comparisons



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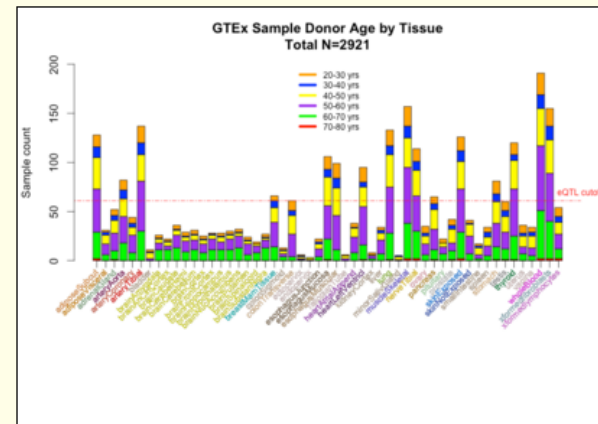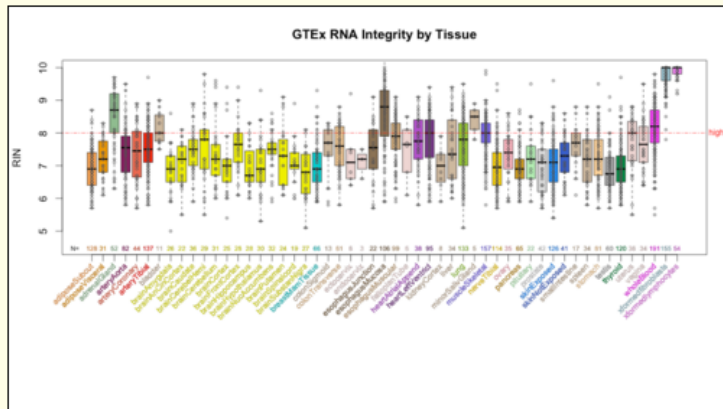
## Description

The [NIH Genotype-Tissue Expression (GTEX) project](#) was designed to establish a sample and data resource for studies on the relationship between genetic variation and gene expression in multiple human tissues. This track shows median gene expression levels in 51 tissues and 2 cell lines, based on RNA-seq data from the GTEx midpoint milestone data release (V6, October 2015). This release is based on data from 8555 tissue samples obtained from 570 adult post-mortem individuals.

## Methods

Gene-level expression files from GTEx analysis, along with the gene model (GENCODE V19) and sample metadata were downloaded from the [GTEX portal](#): [GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct](#). Median expression level (in RPKM) computed per gene/per tissue.

## Subject and Sample Characteristics

The scientific goal of the GTEx project required that the donors and their biospecimen present with no evidence of disease. The tissue types collected were chosen based on their clinical significance, logistical feasibility and their relevance to the scientific goal of the project and the research community. Postmortem samples were collected from non-diseased donors with ages ranging from 20 to 79. 34.4% of donors were female and 65.6% male.





Extensive summary plots of GTEx sample characteristics are available at the [GTEx portal](#).

# 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**

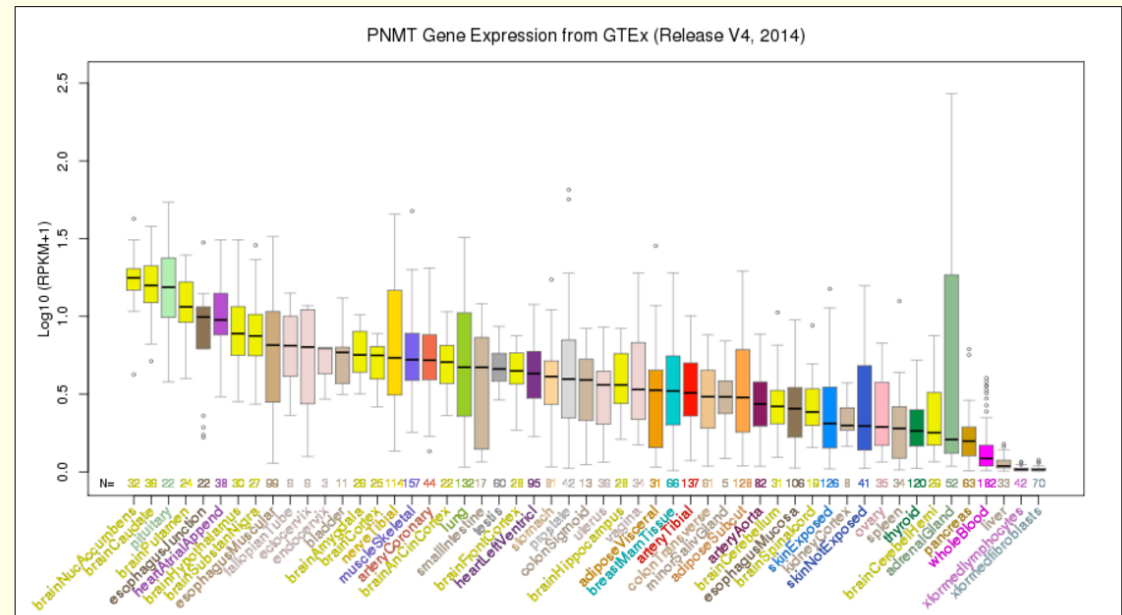| Genomic Sequence (chr17:37,827,375-37,841,002) | mRNA (may differ from genome) | | No protein |
|---|---|---|---|
| Gene Sorter | Genome Browser | Protein FASTA | VisiGene | Table Schema | BioGPS |
| CGAP | Ensembl | ExonPrimer | GeneNetwork | GTEx | H-INV |
| HGNC | Lynx | PubMed | Stanford SOURCE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

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



PNMT Gene Expression from GTEx (Release V4, 2014)

# 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 | 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0... | 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

# Community data:  GTEx data hub

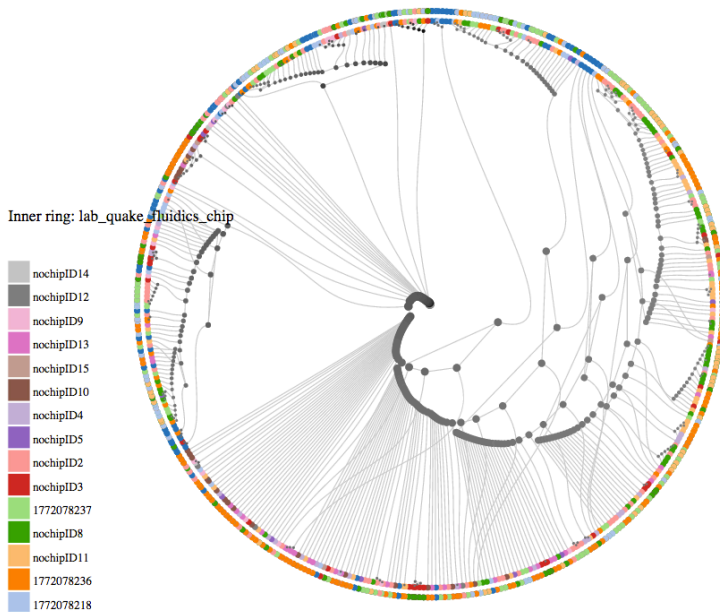- RNA-seq signal tracks for all samples
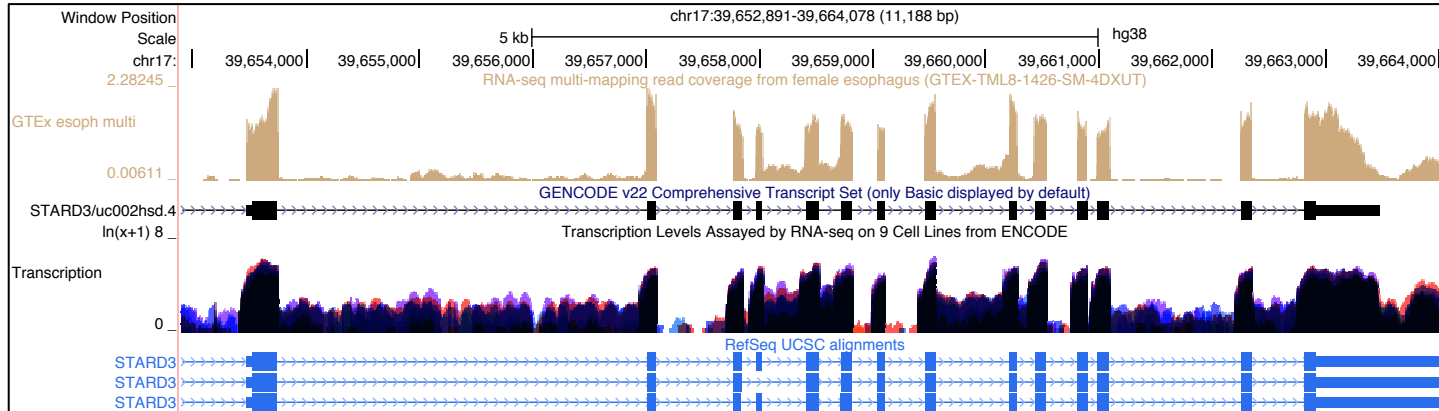


- Exon-level expression
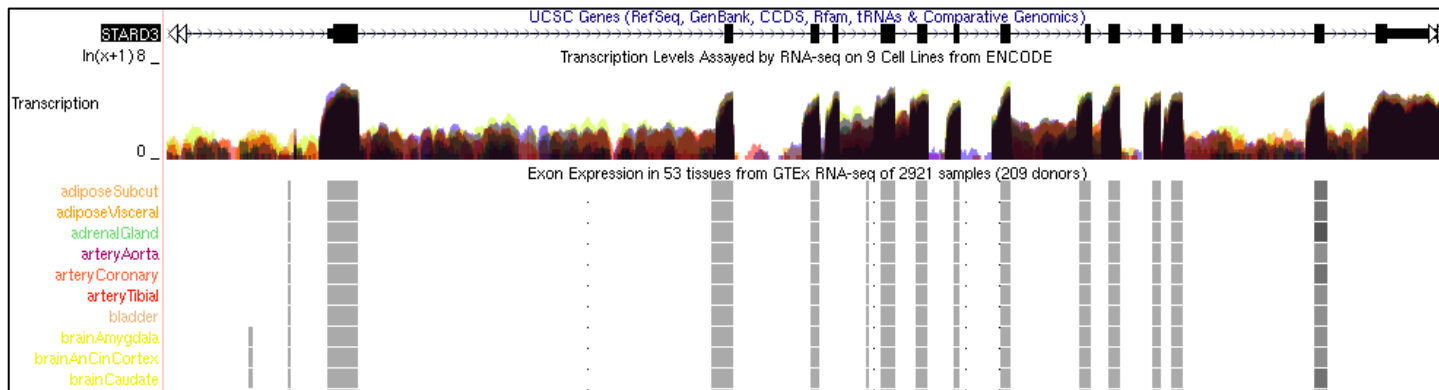


- Transcript-level expression
- Proteomics ?

# Plans

1. Release GTEx V6 gene expression track on hg19 and hg38 (in pushQ week of Feb 15)

2. Integrate GTEx V6 gene expression graph into hgGene page (replace microarray heatmap)

3. Build public track hub of GTEx RNA-seq signal from all V6 samples (~8000), organized by tissue.  (with Parisa Nejad, and in collaboration with Benedict Paten's group).

4. Add GTEx gene expression columns to Gene Sorter

5. Implement GTEx eQTL or allele-specific expression track (TBD)

6. Work with GTEx portal team (Jared Nedzel) and EBI Ensembl groups (Daniel Zerbino) to support each others GTEx visualization efforts.

# Further plans

- ## Tree of cells

- ## Anatomogram



Inner ring: lab_quake_fluidics_chip

- nochipID14
- nochipID12
- nochipID9
- nochipID13
- nochipID15
- nochipID10
- nochipID4
- nochipID5
- nochipID2
- nochipID3
- 1772078237
- nochipID8
- nochipID11
- 1772078236
- 1772078218

Outer ring: GEO_Sample_age

- postnatal 47 years
- postnatal 22 years
- postnatal 50 years
- postnatal 63 years
- postnatal 21 years
- prenatal 16-18 W
- postnatal 37 years
- postnatal 54 years

Credit: Chris Eisenhart

Courtesy of Robert Petryszak

EMBL-EBI

Expression Atlas

# Bonus feature (inspired by GTEx data): Apply button

# Acknowledgements