Gene Expression Matrix

ENCODE AWG, Nov 13 2015

Sarah Djebali

Anna Vlasova

Gene Expression Matrix

- STAR-RSEM pipeline, gencode v.19
- TPM and FPKM values for each replicate
- Tabular-separated (tsv) and Json formats
- long RNA-seq (>200nt)

- Total number of experiments =163
- Total number of bioreplicates = 320
- Total number of genes = 58,540
- Labs: Thomas Gingeras, Barbara Wold, Brenton Graveley

http://genome.crg.es/~sdjebali/STAR-RSEM/geneid_genename_with_tpmallrep_fpkmallrep.tsv.gz
http://genome.crg.es/~sdjebali/STAR-RSEM/geneid_genename_with_tpmallrep_fpkmallrep.json.g

Json format

```
{ "gene_name": "SEC62", "ensembl_id": "ENSG00000008952.12",
  "expression_values": [
   { "dataset": "ENCSR000AAA",
    "rep1_tpm": 10.70, "rep2_tpm": 4.78, "rep1_fpkm": 26.43, "rep2_fpkm": 21.64 },
       { "dataset": "ENCSR000AAB",
  "rep1_tpm": 10.51, "rep2_tpm": 1.40, "rep1_fpkm": 24.57, "rep2_fpkm": 9.69 },
{ "dataset": "ENCSR000AAC",
 "rep1_tpm": 3.40, "rep2_tpm": 10.81, "rep1_fpkm": 17.47, "rep2_fpkm": 38.94 },
{ "dataset": "ENCSR000AAD",
 "rep1_tpm": 1.04, "rep2_tpm": 3.11, "rep1_fpkm": 9.34, "rep2_fpkm": 15.16 },
 { "dataset": "ENCSR000AAE",
 "rep1_tpm": 7.32, "rep2_tpm": 2.44, "rep1_fpkm": 26.35, "rep2_fpkm": 13.35 }....
```

Gene Expression Matrix

Lab	**# Experiments/ Bioreplicates**	**Fraction (experiments)**			**Preparation (experiments)**		
		Whole cell	**Nucleus**	**Cytosol**	**Total**	**polyA+**	**nonPolyA+**
Barbara Wold	15 / 30	15	0	0	13	2	0
Brenton Graveley	9 / 18	9	0	0	9	0	0
Thomas Gingeras	139 / 272	131	8	8	107	21	11
All	**163 / 320**	**155**	**8**	**8**	**129**	**23**	**11**

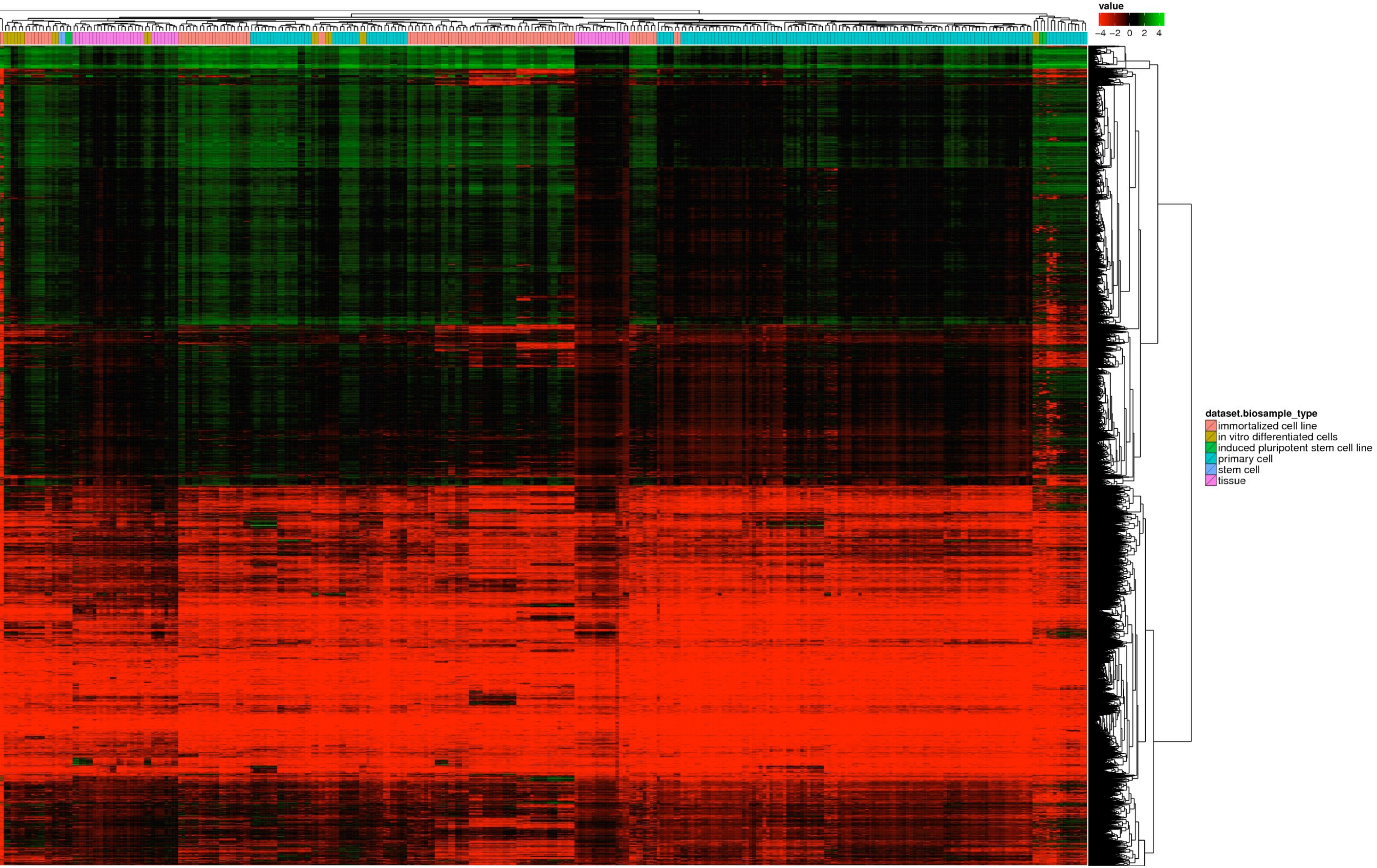
Grouping by biosample_type

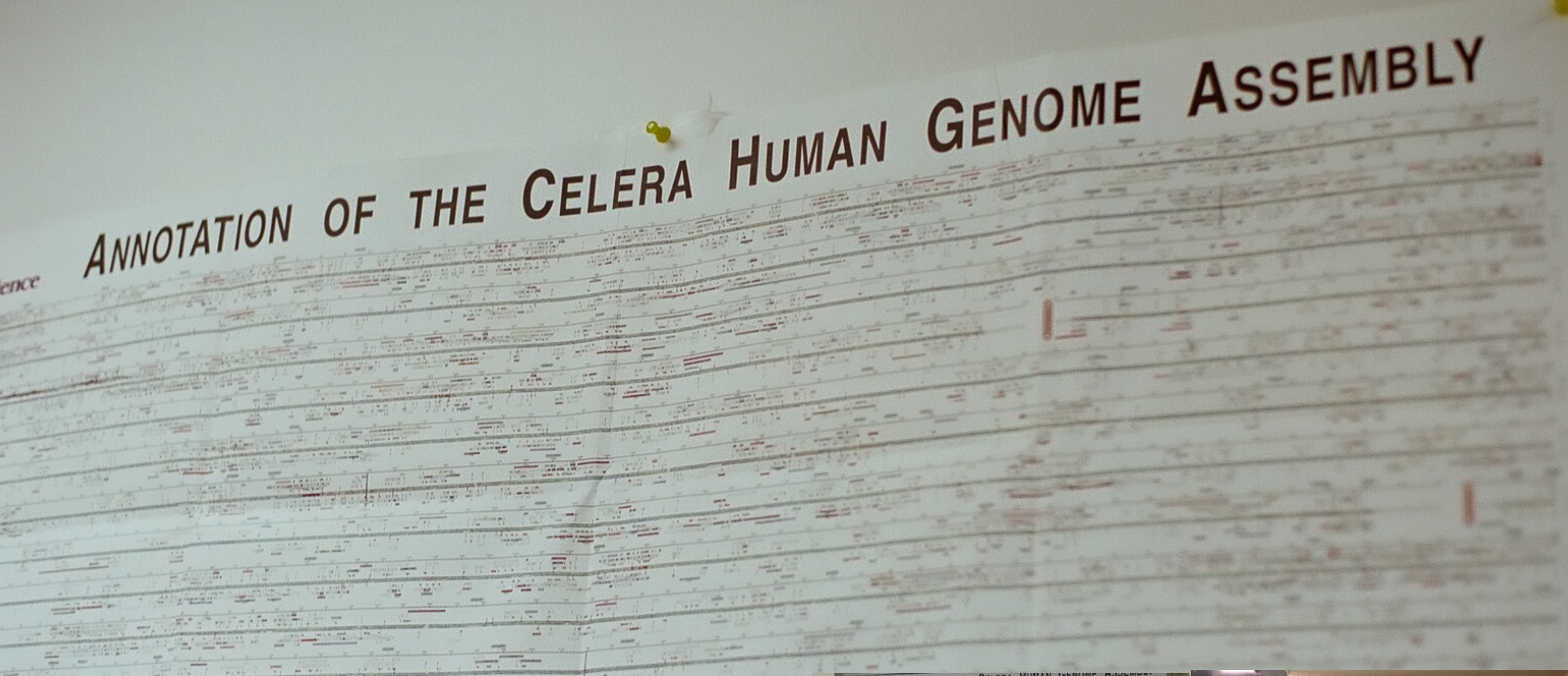
primary cells	81
immortalized cell lines	47
tissues	23
in vitro differentiated cells	9
induced pluripotent stem cell lines	2
stem cells	1

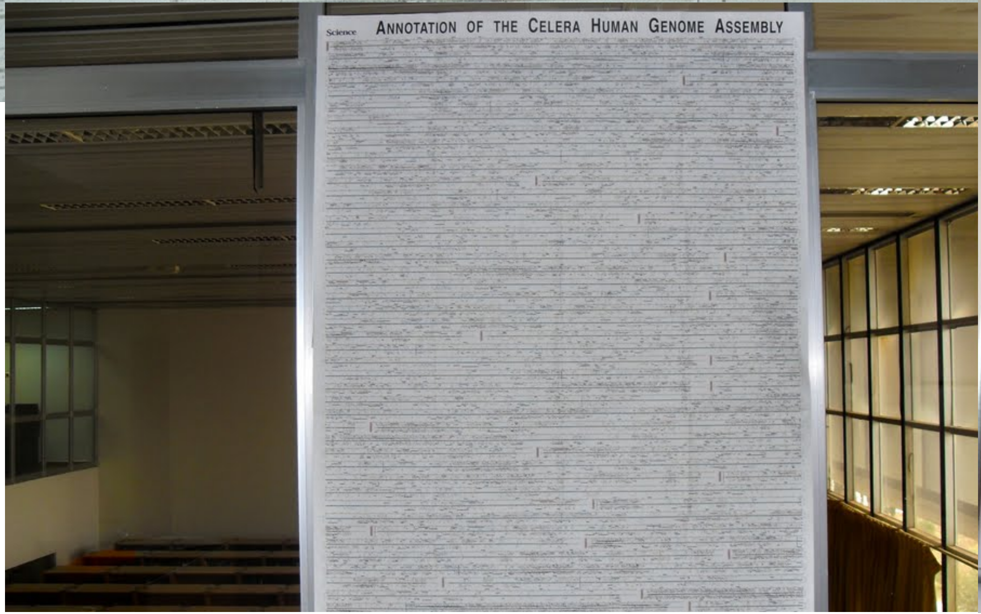
Experiments can also be summarized

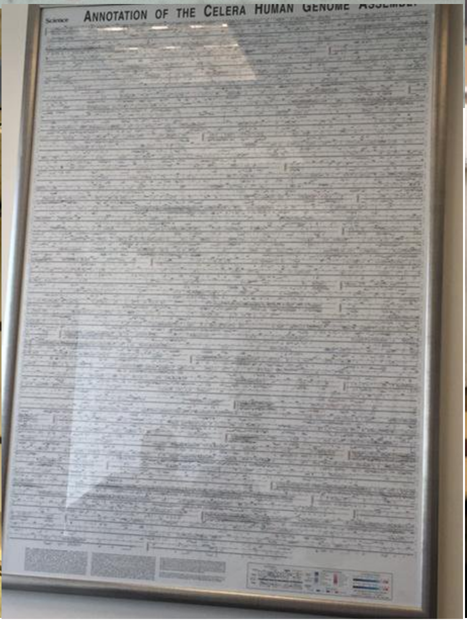
- organ_slims
- system_slims
- develomental_slims

Gene Expression Matrix



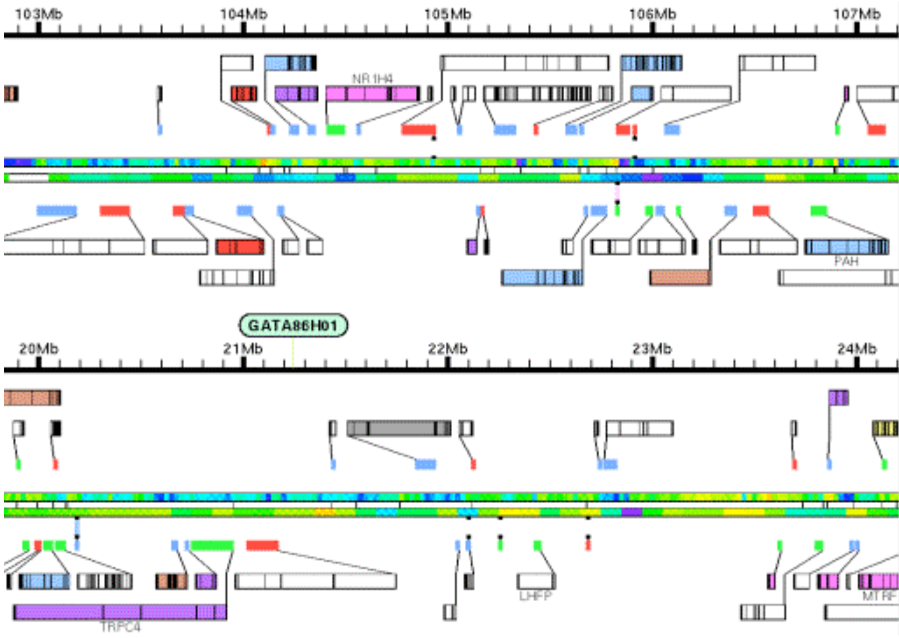

Science
ANNOTATION OF THE CELERA HUMAN GENOME ASSEMBLY


Science
ANNOTATION OF THE CELERA HUMAN GENOME ASSEMBLY


Science
ANNOTATION OF THE CELERA HUMAN GENOME ASSEMBLY

ANNOTATION OF THE CELERA HUMAN GENOME ASSEMBLY
Home
Home

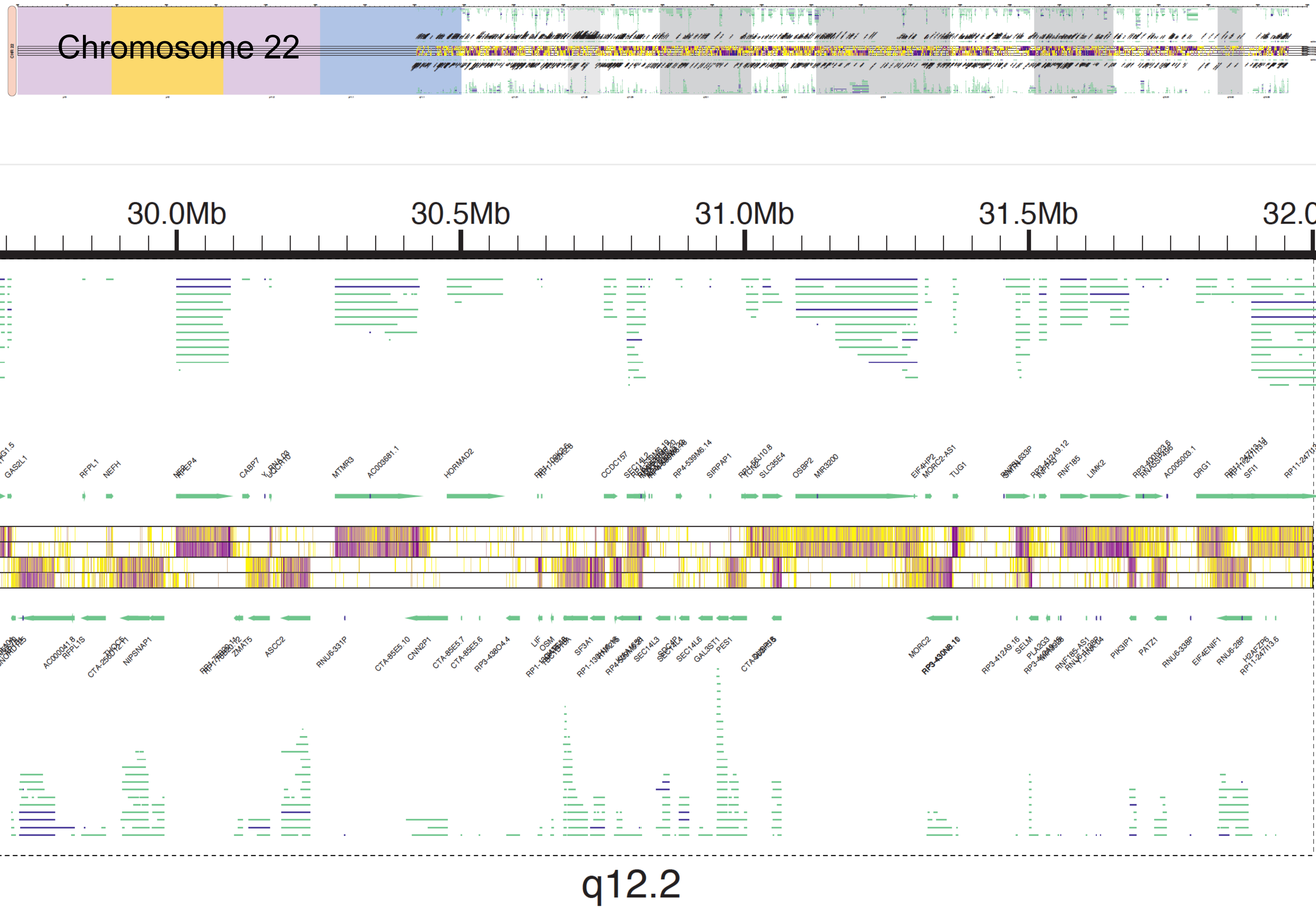
Annotation of the Celera Human Genome Assembly



https://www.sciencemag.org/content/291/5507/1304/suppl/DC2

ANNOTATION OF THE CELERA HUMAN GENOME
KOBLOG

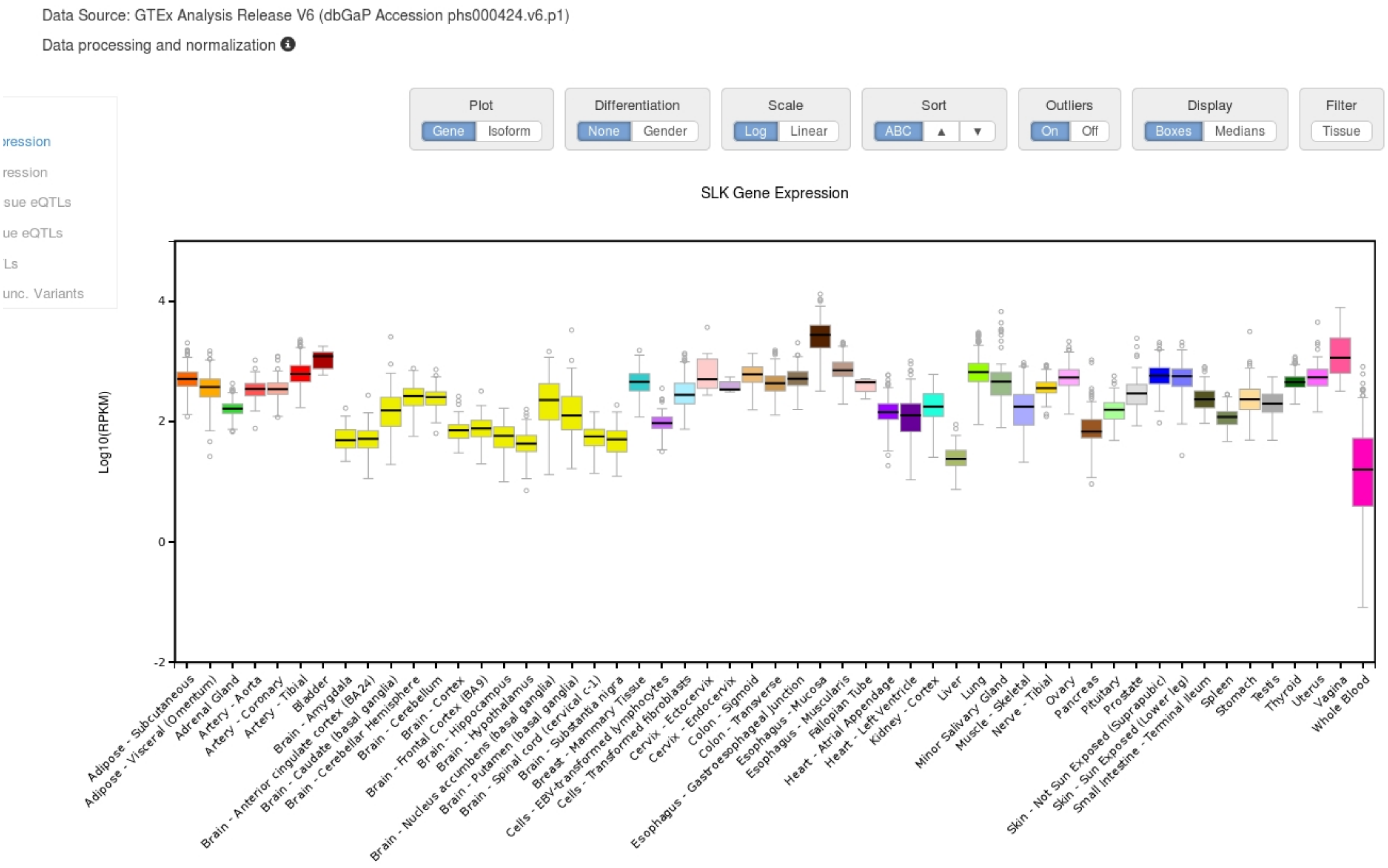
KOBLOG


Chromosome 22
CHR 22
30.0Mb
30.5Mb
31.0Mb
31.5Mb
32.0
GAS2L1
RFPL1
NEFH
NF2
CABP7
UQCR10
MTMR3
AC003681.1
HORMAD2
CCDC157
SEC14L2
RP4-539M6.14
SIRPAP1
RP1-56J10.8
TCN2
SLC35E4
OSBP2
MIR3200
EIF4HP2
MORC2-AS1
TUG1
SMTN
RP3-412A9.12
INPP5J
RNF185
LIMK2
RP3-400N23.6
RNA5SP496
AC005003.1
DRG1
SFI1
AC000041.8
RFPL1S
THOC5
CTA-256D12.11
NIPSNAP1
ZMAT5
ASCC2
RNU6-331P
CTA-85E5.10
CNN2P1
CTA-85E5.7
CTA-85E5.6
RP3-438O4.4
LIF
OSM
SF3A1
SEC14L3
SDC4P
SEC14L4
SEC14L6
GAL3ST1
PES1
MORC2
RP3-430N8.10
RP3-412A9.16
SELM
PLA2G3
RNF185-AS1
PIK3IP1
PATZ1
RNU6-338P
EIF4ENIF1
RNU6-28P
H2AFZP6
RP11-247I13.6
q12.2

Gene based expression plots

GTEx

http://www.gtexportal.org/home/gene/SLK



Data Source: GTEx Analysis Release V6 (dbGaP Accession phs000424.v6.p1)

Data processing and normalization

Plot [✓ Gene] Isoform

Differentiation [✓ None] Gender

Scale [✓ Log] Linear

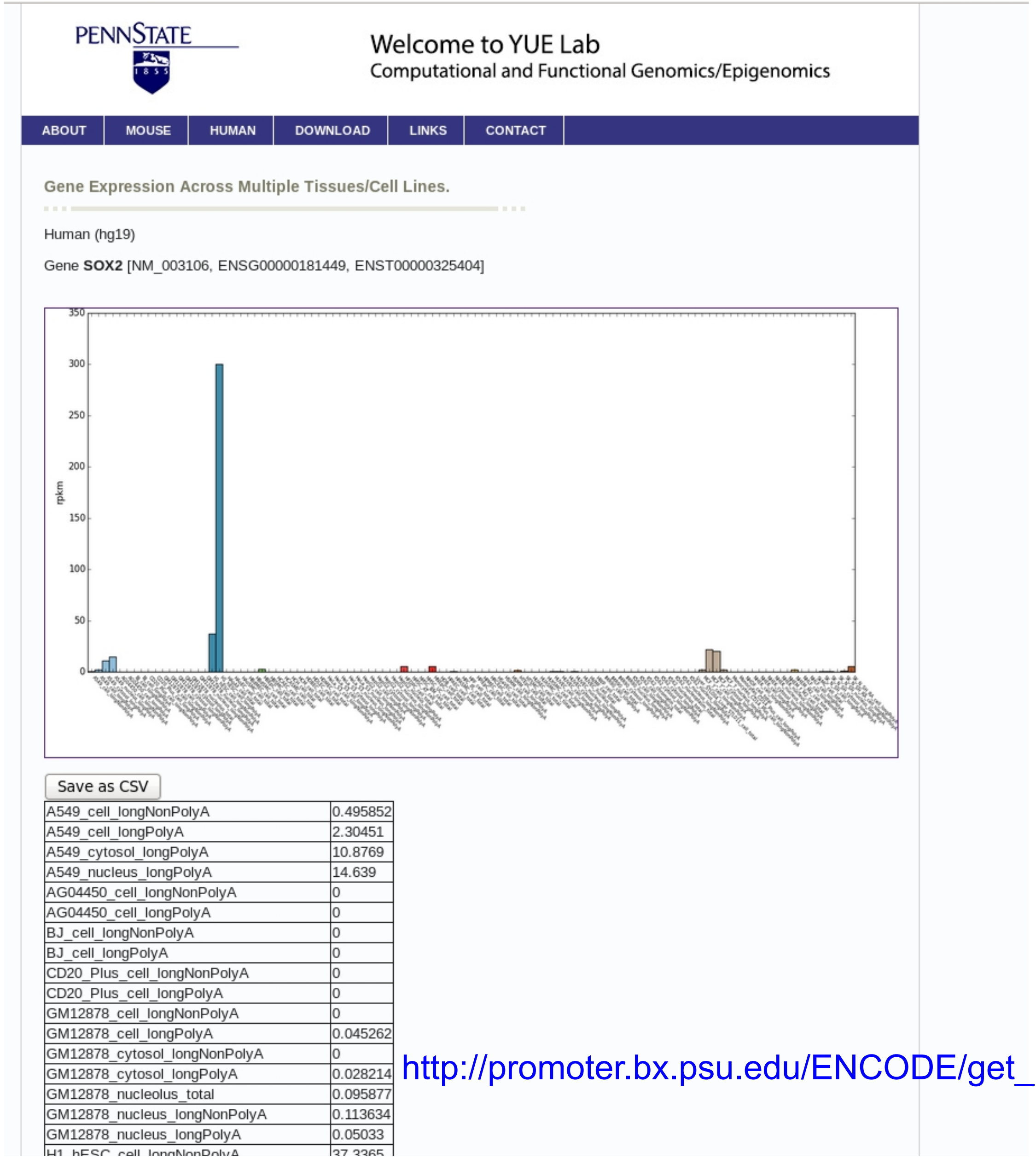
Sort [✓ ABC] ▲ ▼

Outliers [✓ On] Off

Display [✓ Boxes] Medians

Filter Tissue

ENCODE browser



Welcome to YUE Lab
Computational and Functional Genomics/Epigenomics

Gene Expression Across Multiple Tissues/Cell Lines.

Human (hg19)

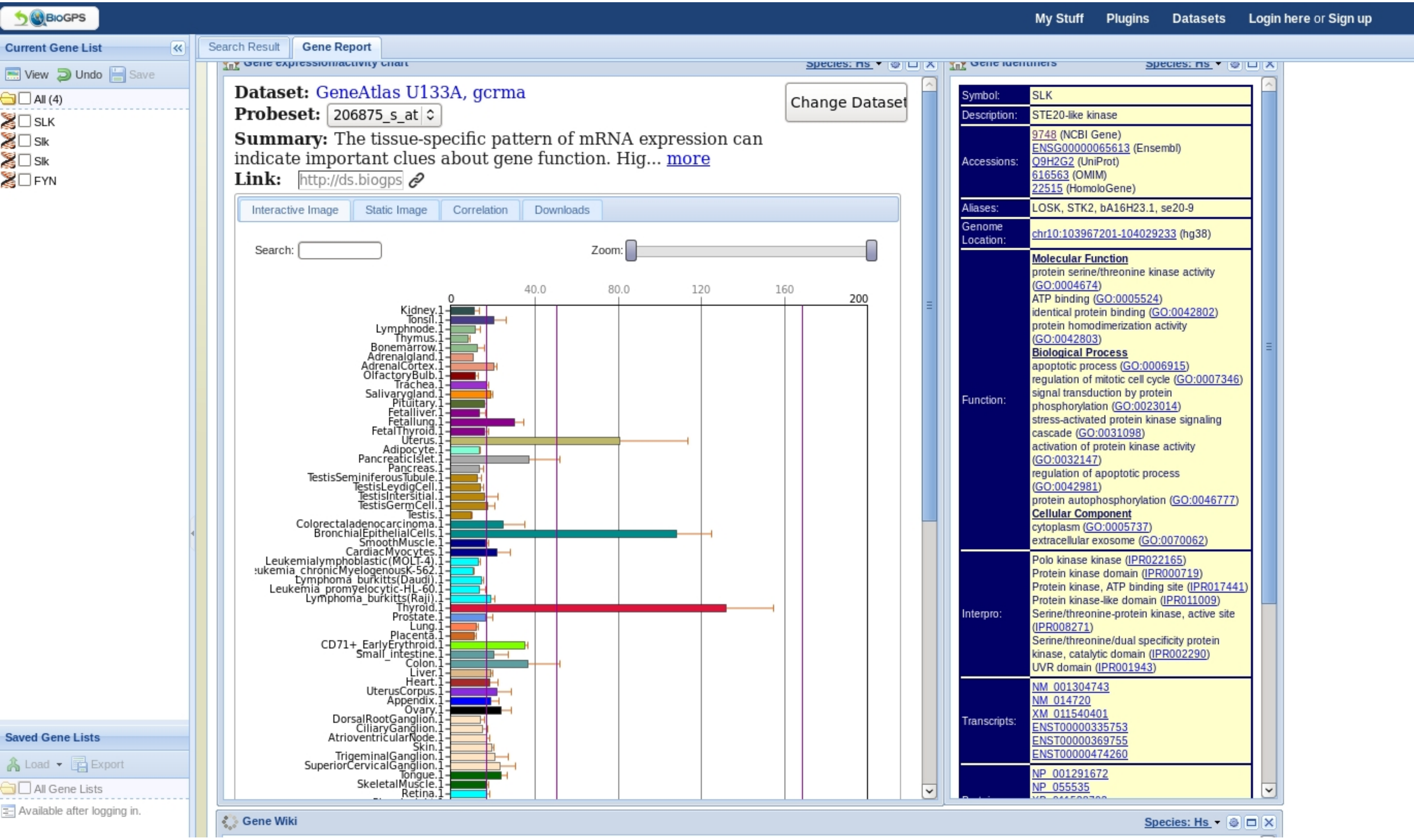
Gene **SOX2** [NM_003106, ENSG00000181449, ENST00000325404]

A549_cell_longNonPolyA	0.495852
A549_cell_longPolyA	2.30451
A549_cytosol_longPolyA	10.8769
A549_nucleus_longPolyA	14.639
AG04450_cell_longNonPolyA	0
AG04450_cell_longPolyA	0
BJ_cell_longNonPolyA	0
BJ_cell_longPolyA	0
CD20_Plus_cell_longNonPolyA	0
CD20_Plus_cell_longPolyA	0
GM12878_cell_longNonPolyA	0
GM12878_cell_longPolyA	0.045262
GM12878_cytosol_longNonPolyA	0
GM12878_cytosol_longPolyA	0.028214
GM12878_nucleolus_total	0.095877
GM12878_nucleus_longNonPolyA	0.113634
GM12878_nucleus_longPolyA	0.05033
H1_hESC_cell_longNonPolyA	37.3365

http://promoter.bx.psu.edu/ENCODE/get_human_expr.php?assembly=hg19&gene=Sox2

BioGPS

http://biogps.org/#goto=genereport&id=9748



□ All (4)
□ SLK
□ Slk
□ Slk
□ FYN

Dataset: GeneAtlas U133A, gcrma

Probeset: 206875_s_at

Summary: The tissue-specific pattern of mRNA expression can indicate important clues about gene function. Hig... more

Link: http://ds.biogps

Change Dataset

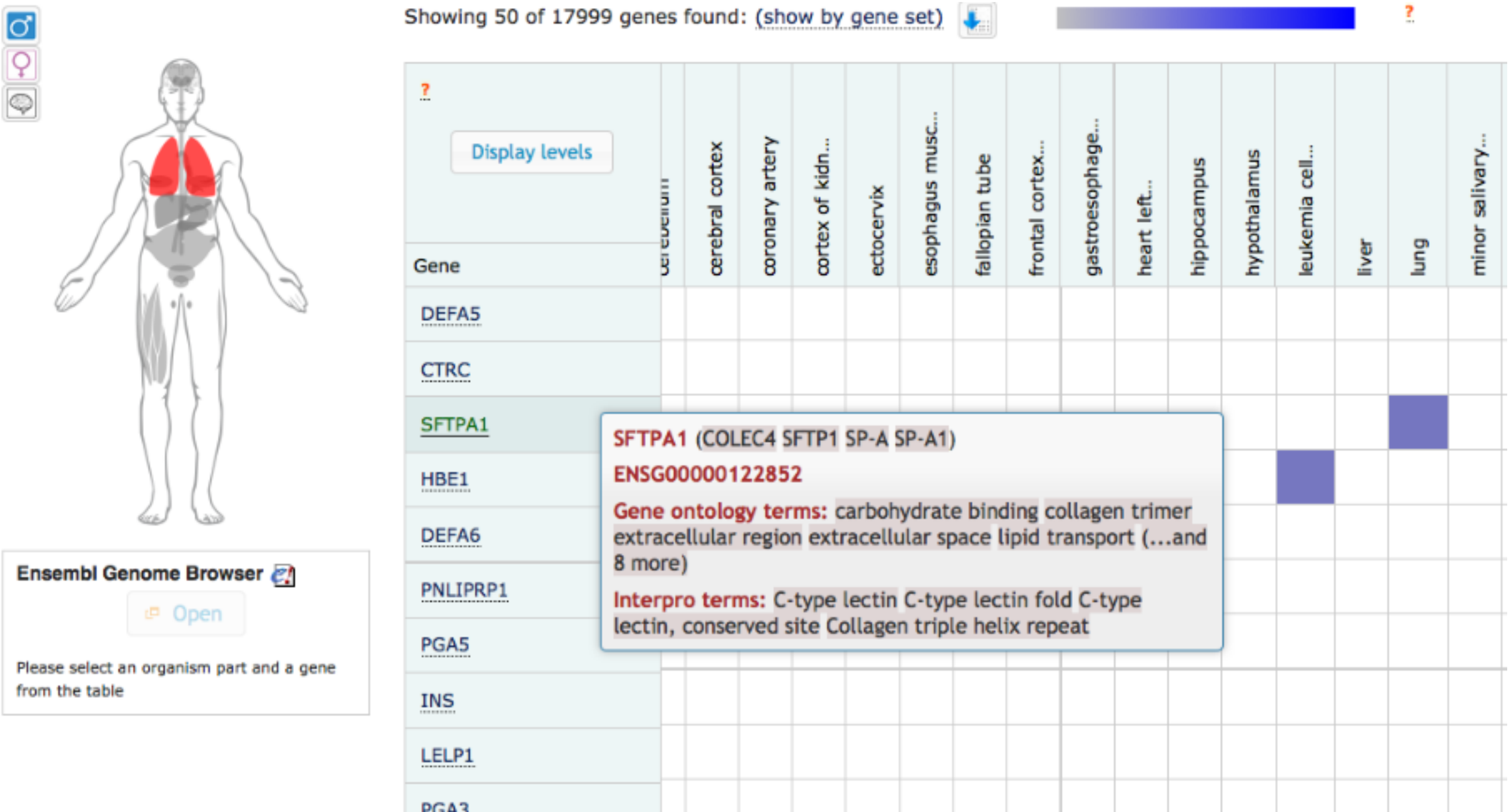
Symbol:	SLK
Description:	STE20-like kinase
Accessions:	9748 (NCBI Gene) ENSG00000065613 (Ensembl) Q9H2G2 (UniProt) 616563 (OMIM) 22515 (HomoloGene)
Aliases:	LOSK, STK2, bA16H23.1, se20-9
Genome Location:	chr10:103967201-104029233 (hg38)
Function:	**Molecular Function** protein serine/threonine kinase activity (GO:0004674) ATP binding (GO:0005524) identical protein binding (GO:0042802) protein homodimerization activity (GO:0042803) **Biological Process** apoptotic process (GO:0006915) regulation of mitotic cell cycle (GO:0007346) signal transduction by protein phosphorylation (GO:0023014) stress-activated protein kinase signaling cascade (GO:0031098) activation of protein kinase activity (GO:0032147) regulation of apoptotic process (GO:0042981) protein autophosphorylation (GO:0046777) **Cellular Component** cytoplasm (GO:0005737) extracellular exosome (GO:0070062)
Interpro:	Polo kinase kinase (IPR022165) Protein kinase domain (IPR000719) Protein kinase, ATP binding site (IPR017441) Protein kinase-like domain (IPR011009) Serine/threonine-protein kinase, active site (IPR008271) Serine/threonine/dual specificity protein kinase, catalytic domain (IPR002290) UVR domain (IPR001943)
Transcripts:	NM_001304743 NM_014720 XM_011540401 ENST00000335753 ENST00000369755 ENST00000474260
	NP_001291672 NP_055535

All Gene Lists

Available after logging in.

Expression Atlas

https://www.ebi.ac.uk/gxa/experiments/E-MTAB-2706



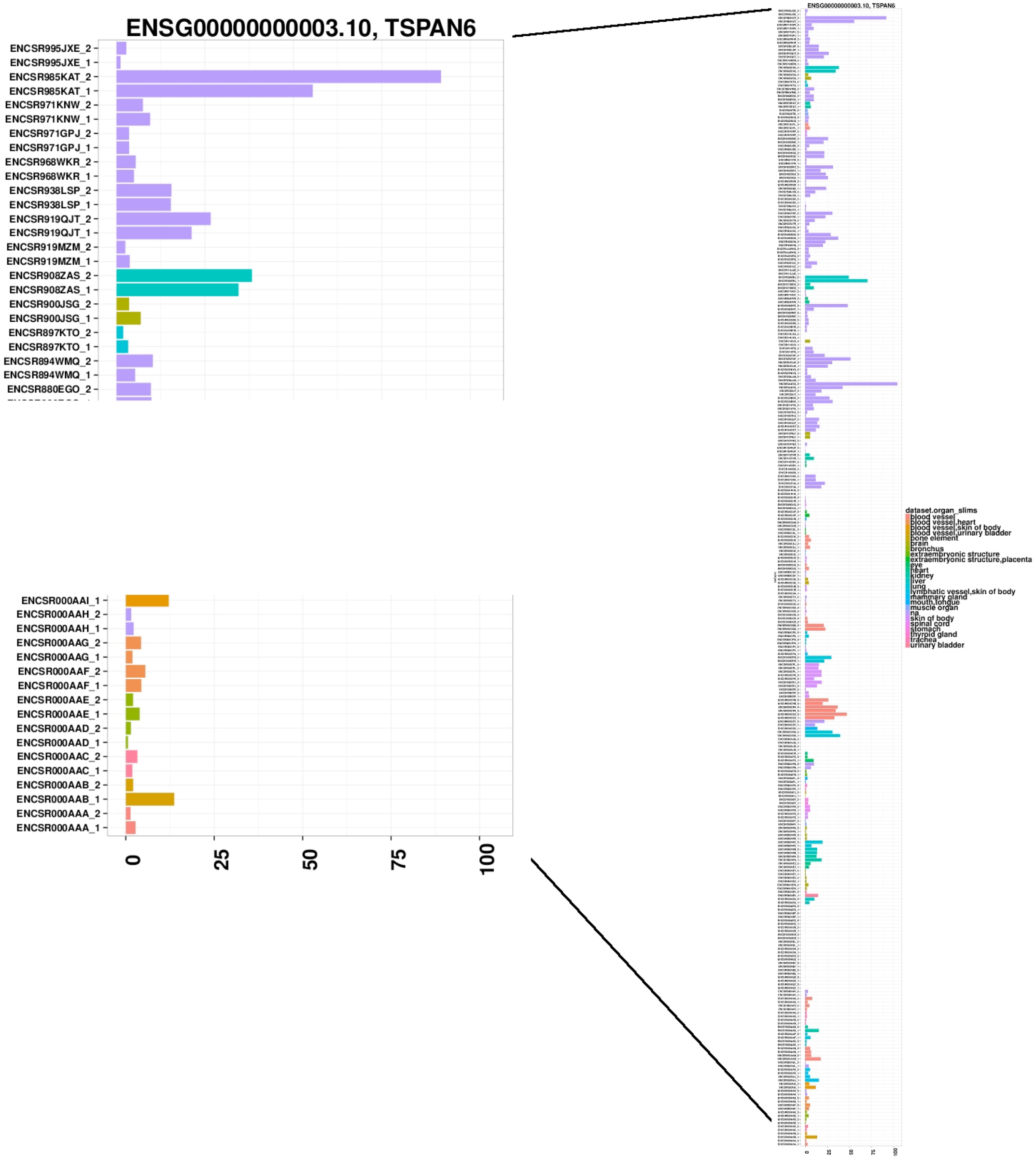
Showing 50 of 17999 genes found: (show by gene set)

Gene	cerebellum	cerebral cortex	coronary artery	cortex of kidn...	ectocervix	esophagus musc...	fallopian tube	frontal cortex...	gastroesophage...	heart left...	hippocampus	hypothalamus	leukemia cell...	liver	lung	minor salivary...
DEFA5																
CTRC																
SFTPA1															■	
HBE1													■			
DEFA6																
PNLIPRP1																
PGA5																
INS																
LELP1																
PGA3																

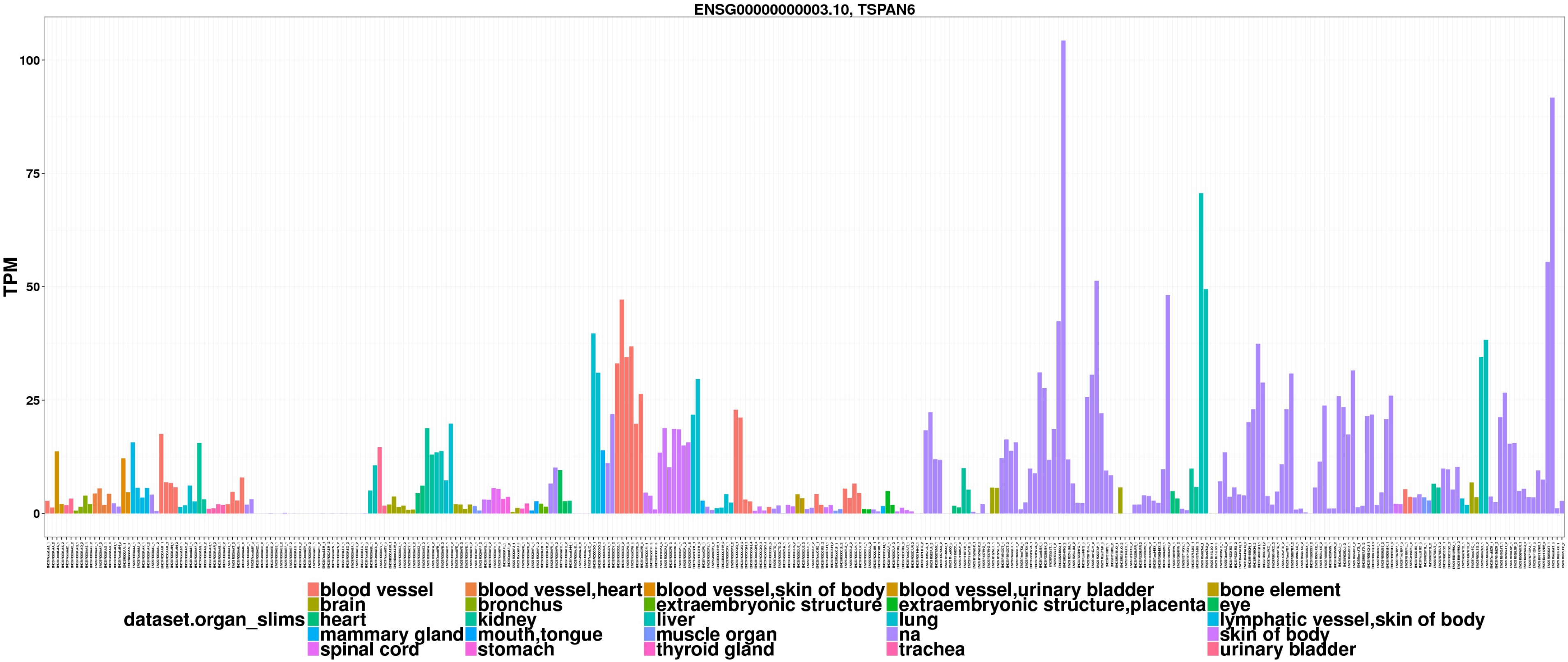
SFTPA1 (COLEC4 SFTP1 SP-A SP-A1)
ENSG00000122852
Gene ontology terms: carbohydrate binding collagen trimer extracellular region extracellular space lipid transport (...and 8 more)
Interpro terms: C-type lectin C-type lectin fold C-type lectin, conserved site Collagen triple helix repeat

Ensembl Genome Browser
Open
Please select an organism part and a gene from the table

All samples. Absolute TPMs
Sorted: by sample+bioreplicate name. **Colored**: by organ



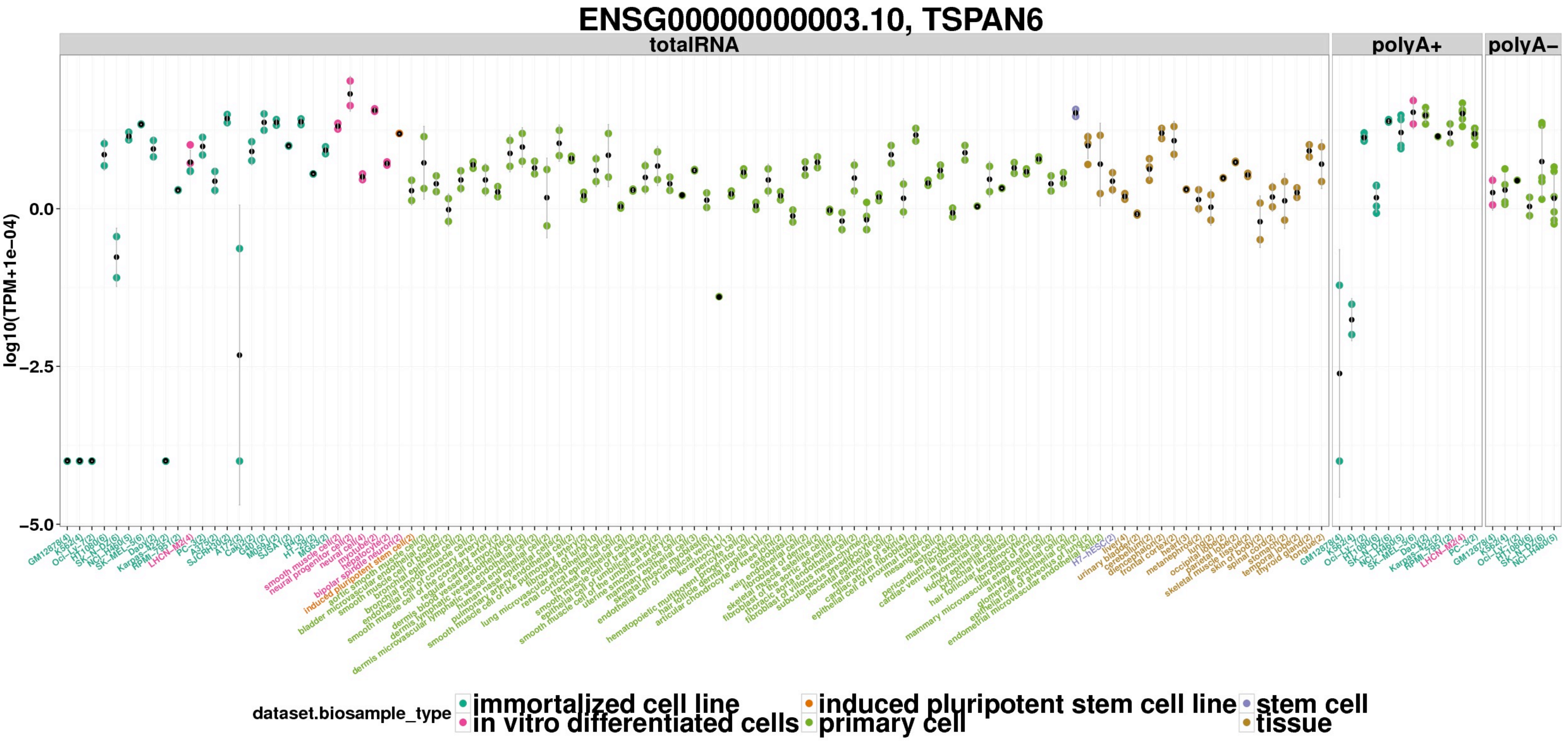
All samples. Absolute TPMs
Sorted: by sample+bioreplicate name. **Colored**: by organ



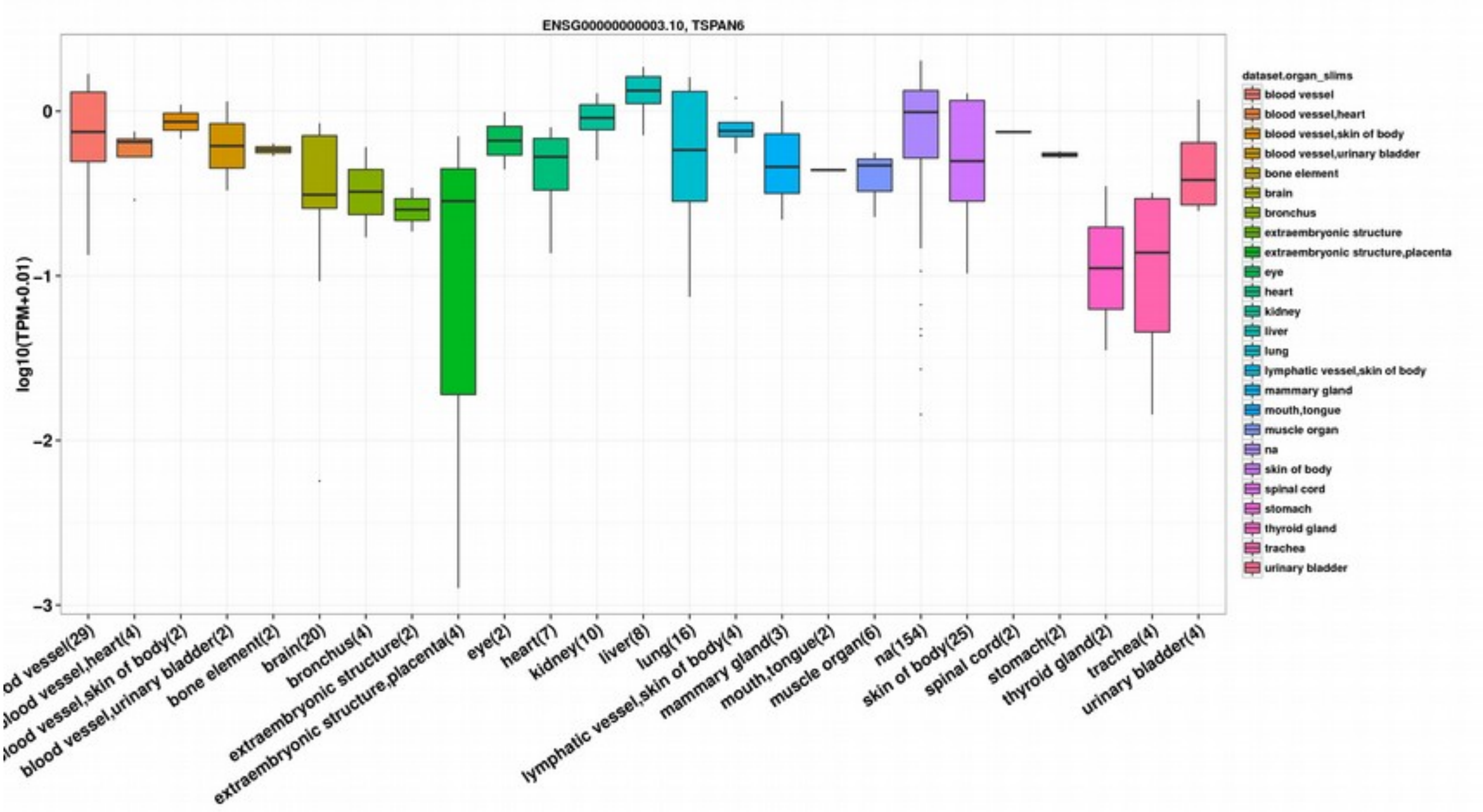
Gingeras samples. Log10 TPMs

Sorted: by sample+bioreplicate name.

Colored: by biosample type. **Faceted**: by preparation protocol

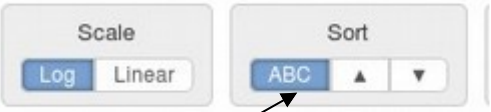


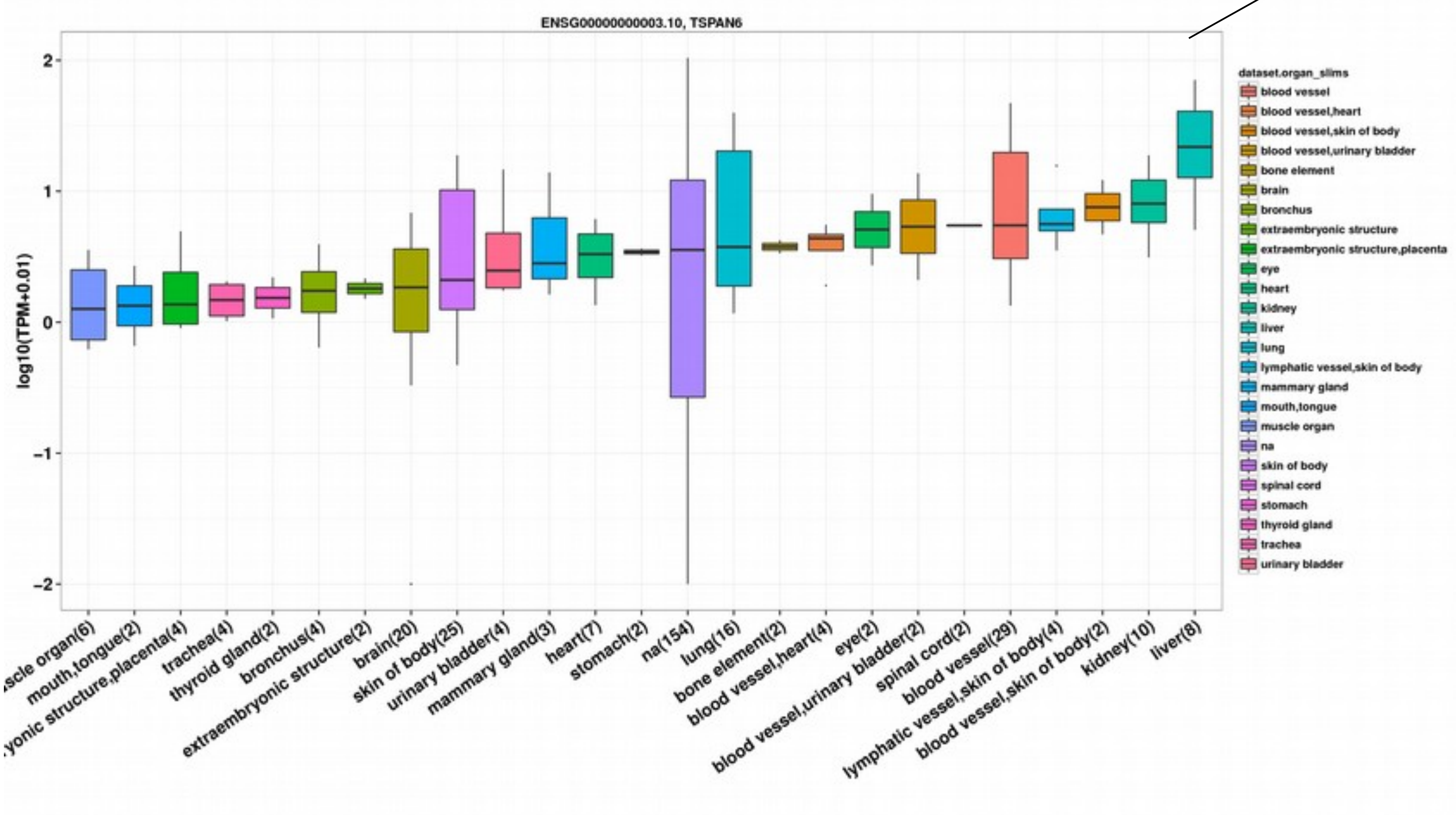
All samples. Log10 TPMs
Sorted: by organ. **Grouped**: by organ. **Colored**: by organ



All samples. Log10 TPMs
Sorted: by median gene expression in organs.
Grouped: by organ. **Colored**: by organ

GTEX portal



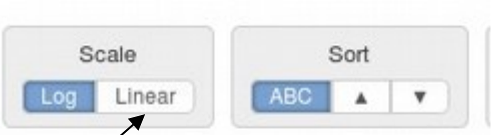


All samples Absolute TPMs

Sorted: by meidan gene expression in system.

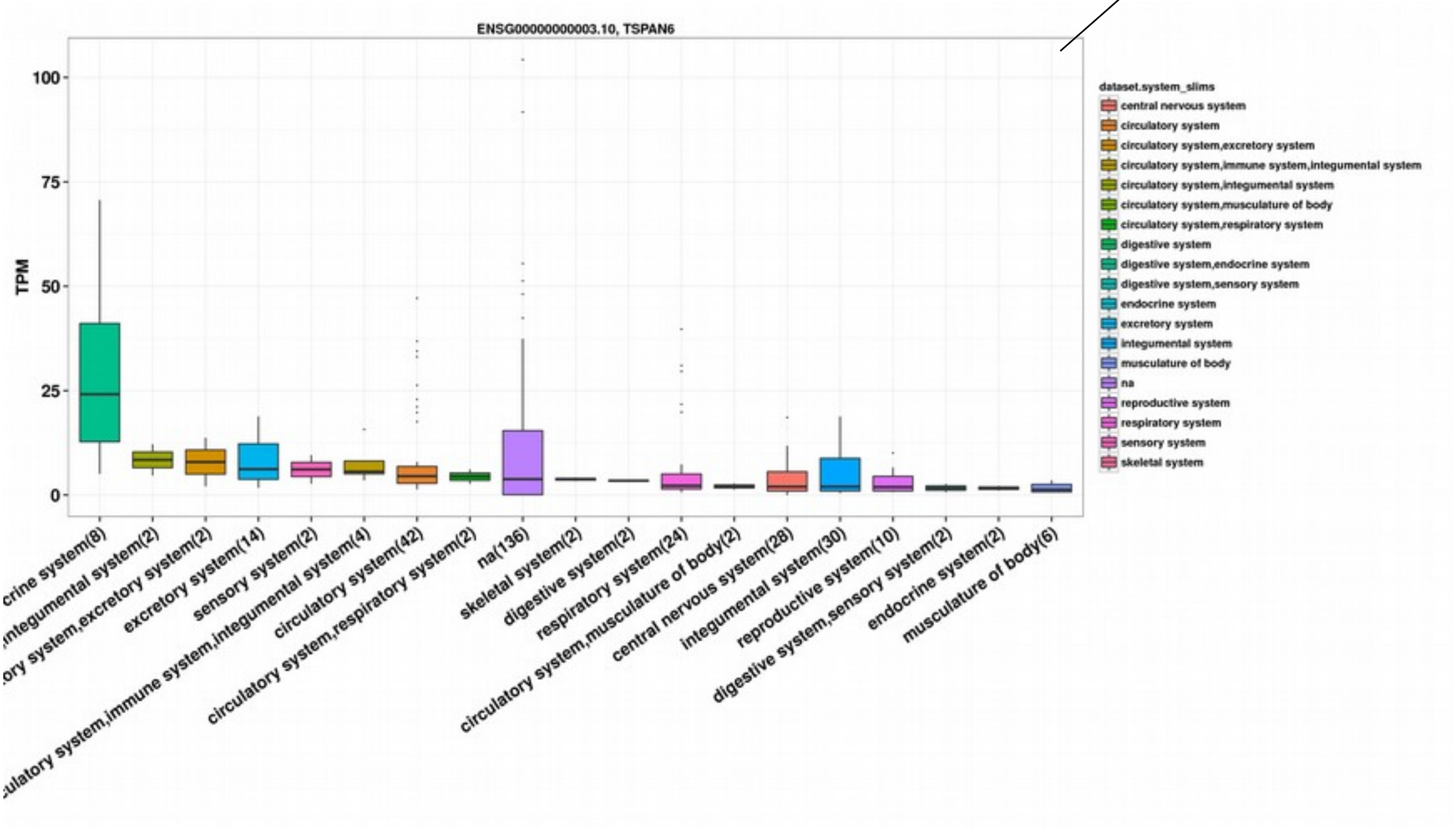
Grouped: by system. **Colored**: by system

GTEX portal

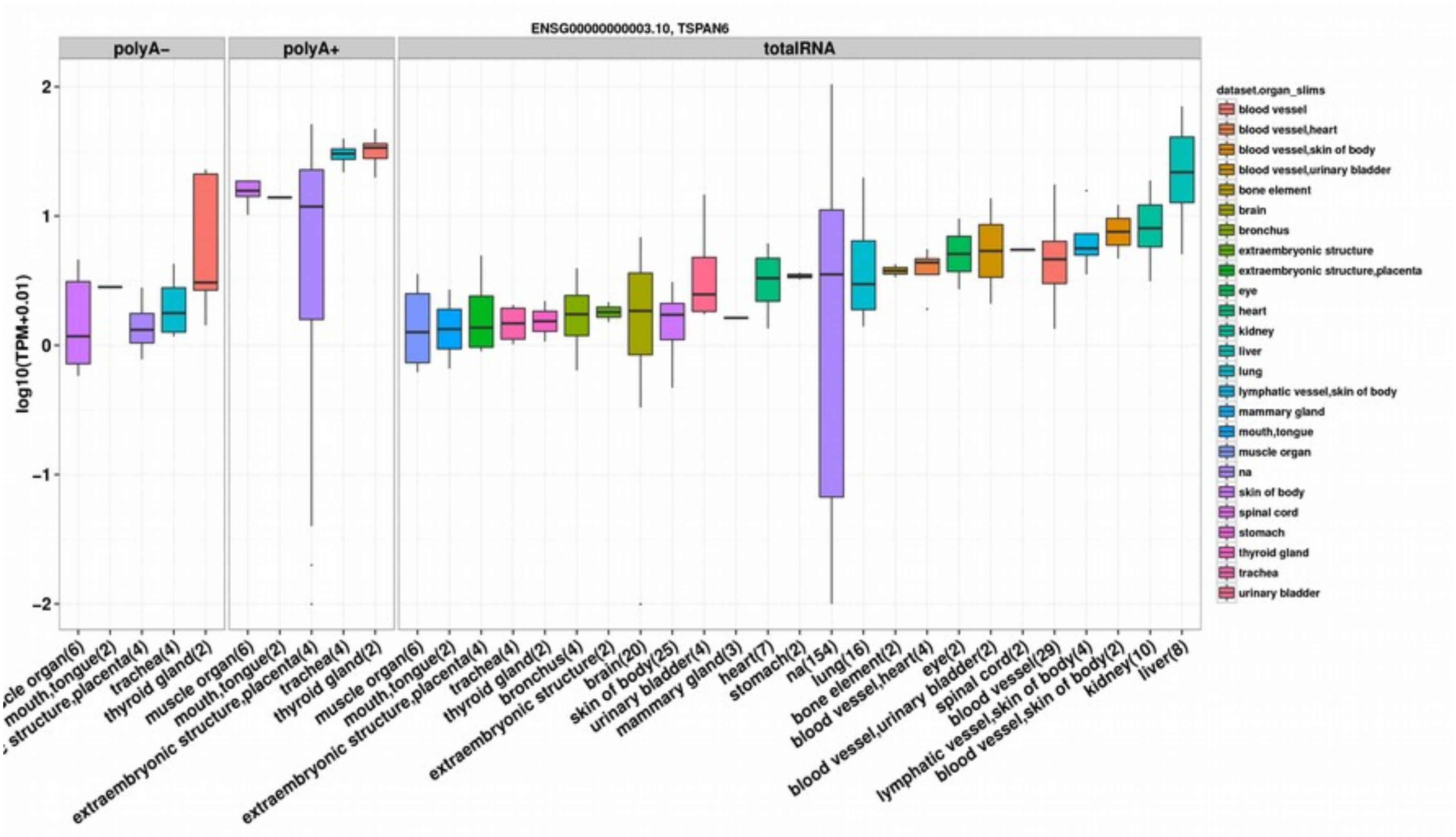


Scale: [✓ Log] Linear

Sort: [✓ ABC] ▲ ▼



All samples. Log10 TPMs
Sorted: by median gene expression by organ.
Grouped: by organ. **Colored**: by organ.
Faceted: by extraction protocol

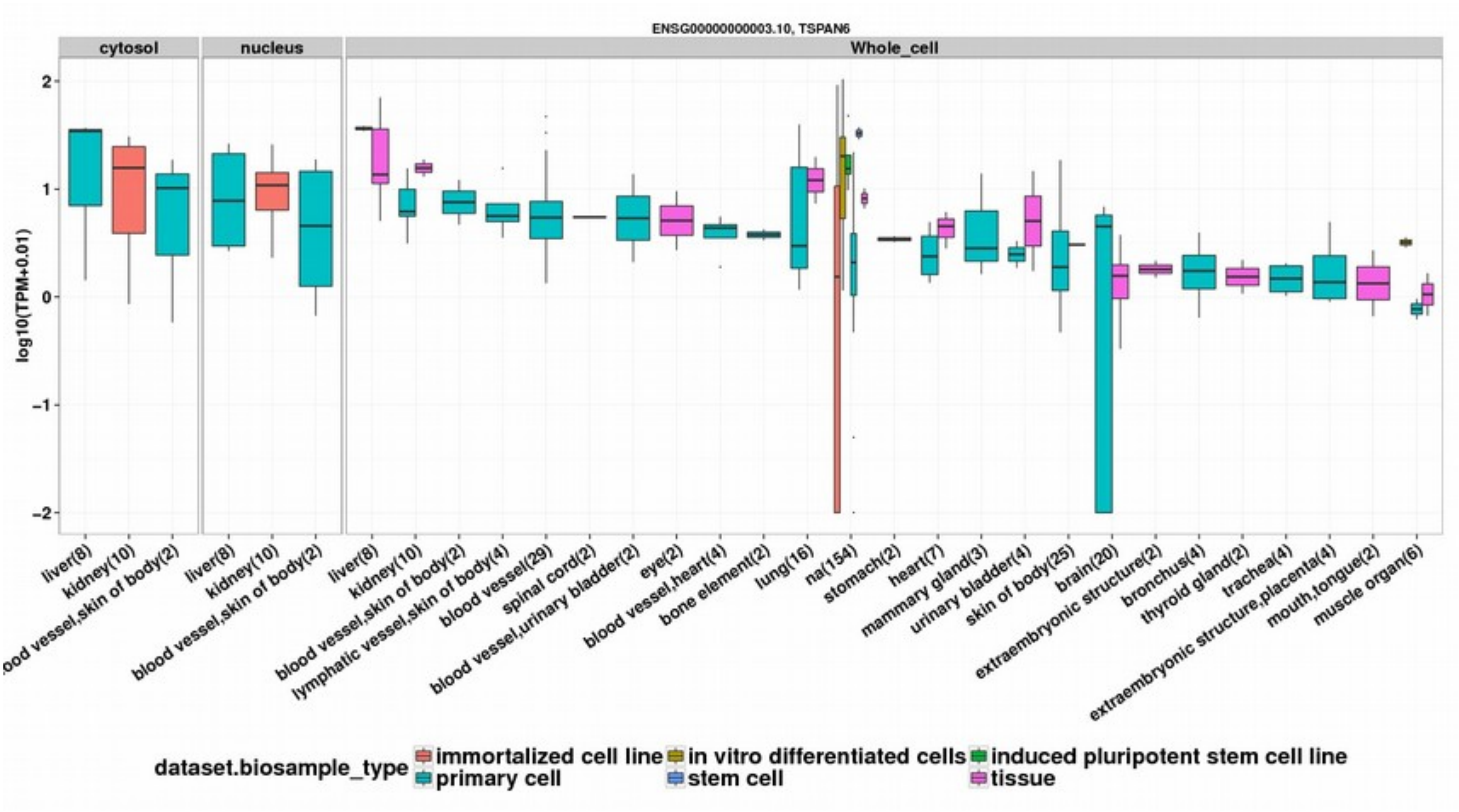


All samples. Log10 TPMs

Sorted: by median gene expression by organ.

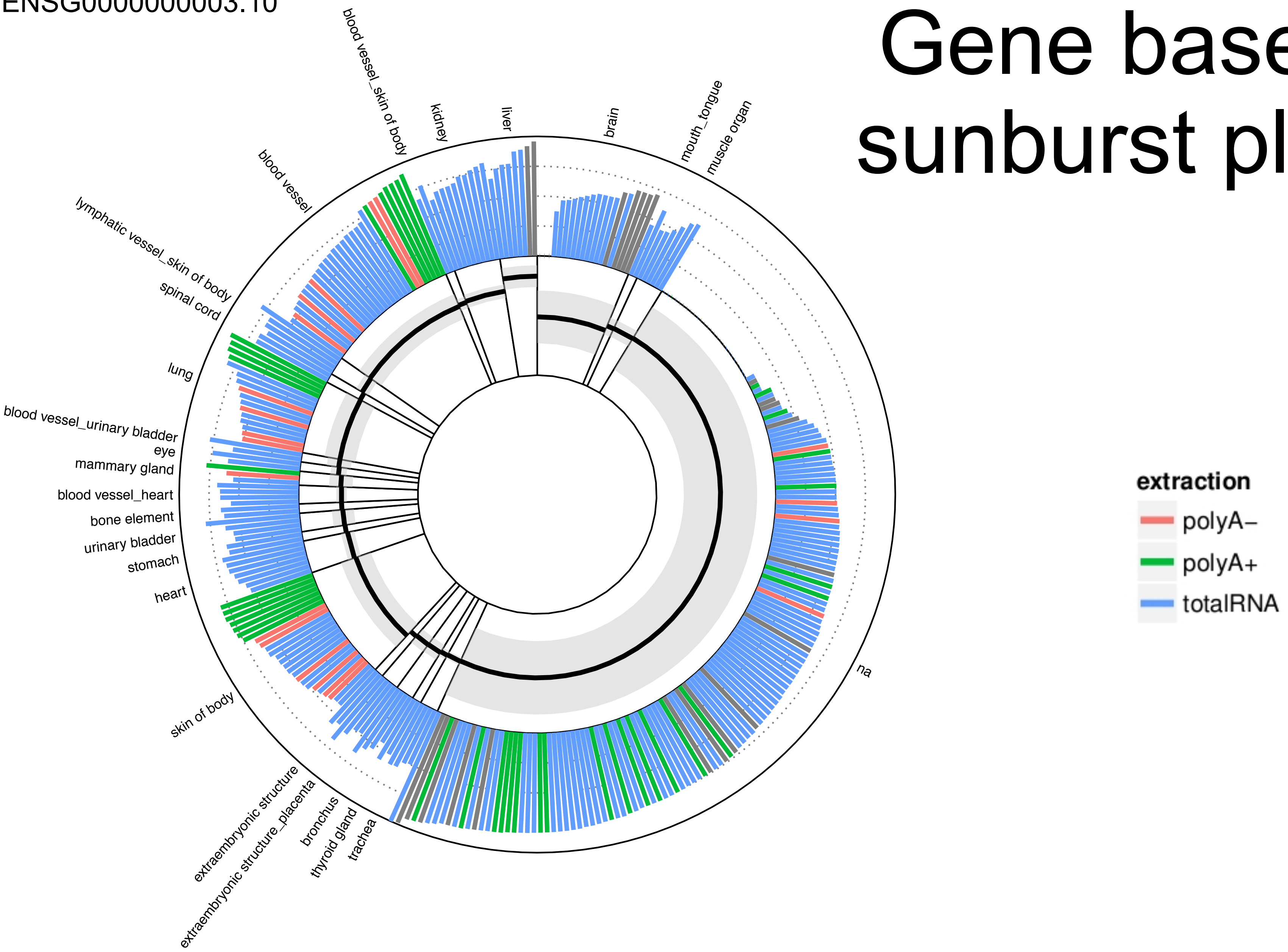
Grouped: by organ. **Colored**: by biosample type.

Faceted: by RNA fraction

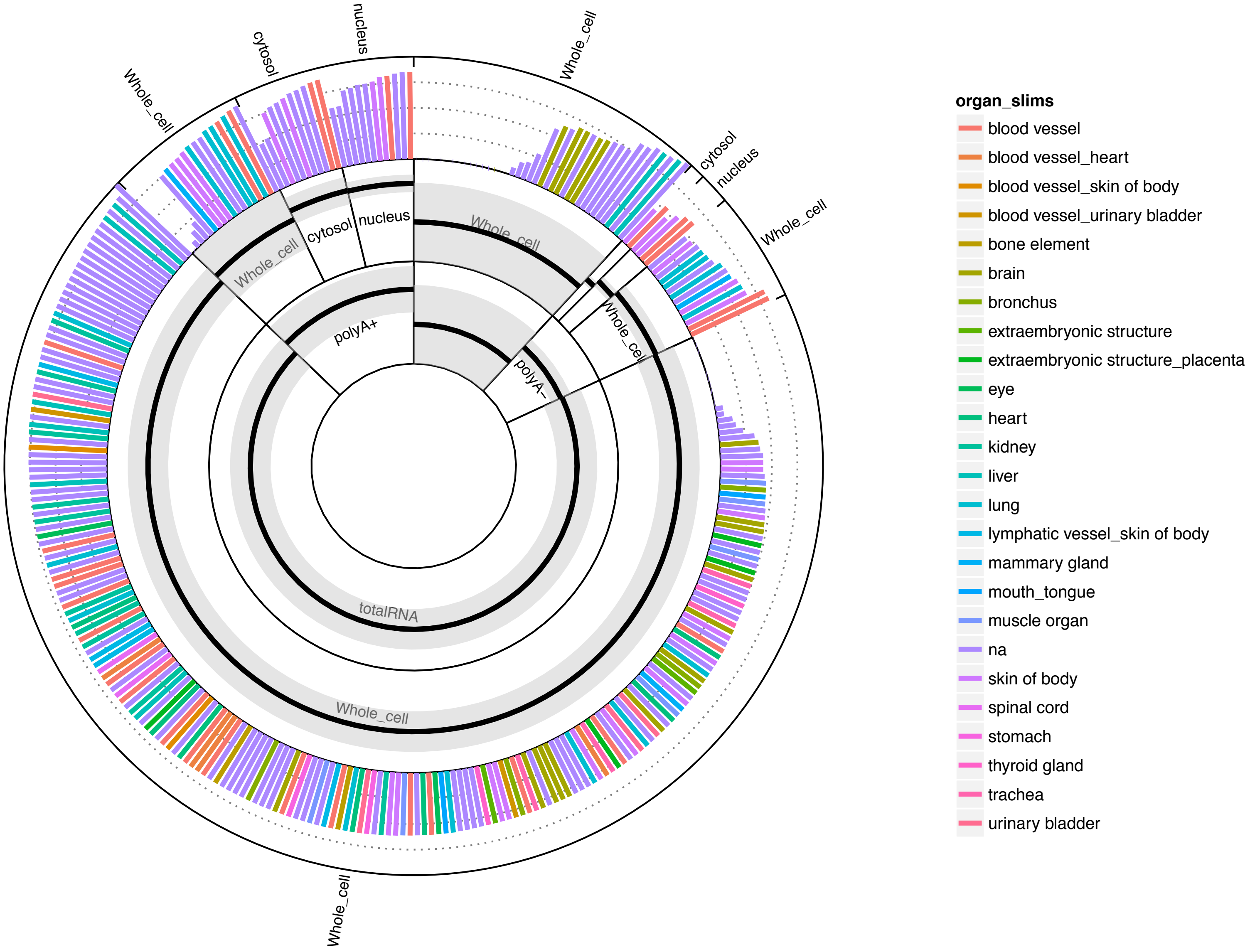


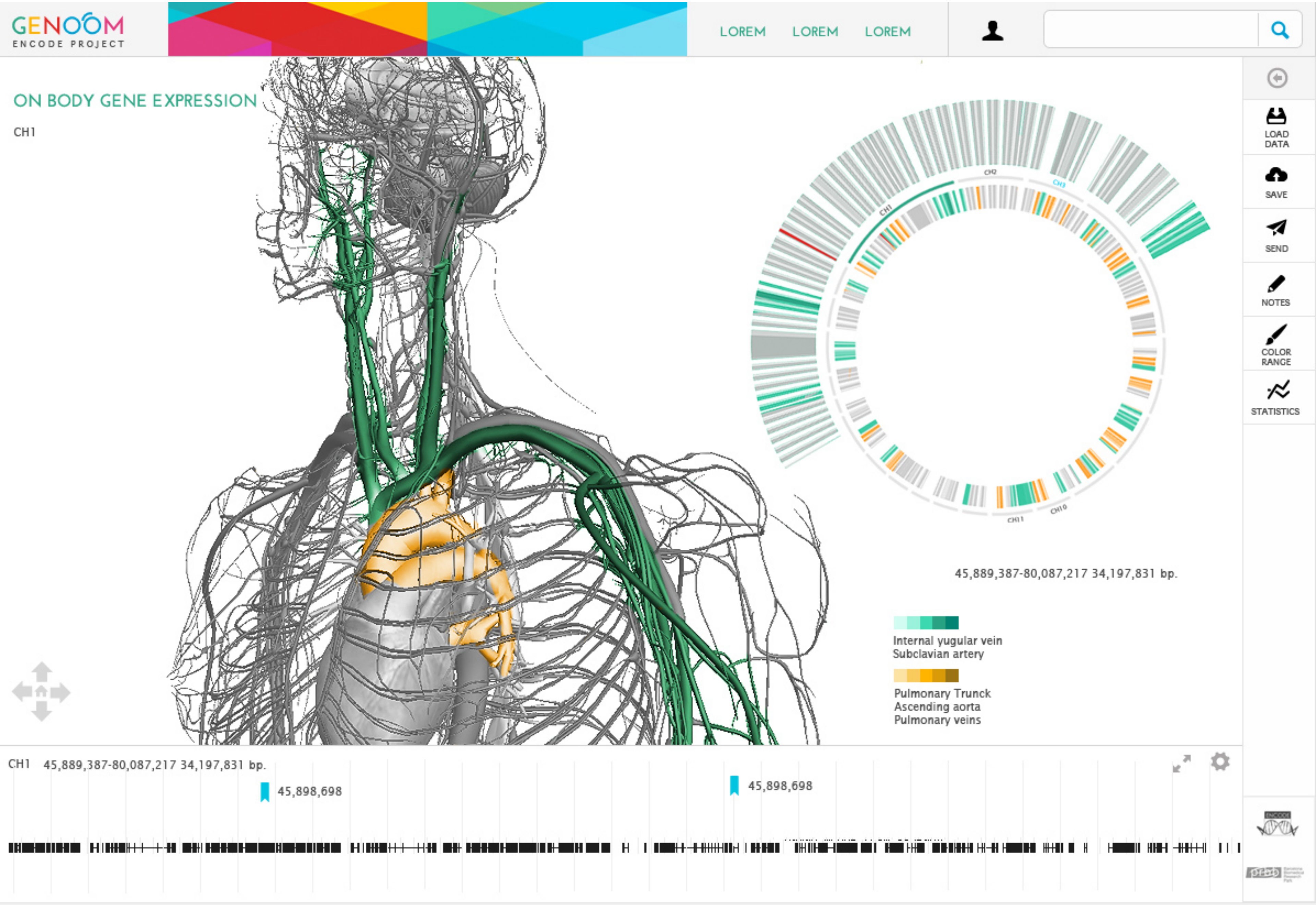
ENSG00000000003.10

Gene based sunburst plots



ENSG00000000003.10





ON BODY GENE EXPRESSION

CH1

45,889,387-80,087,217 34,197,831 bp.

🟩 Internal yugular vein
Subclavian artery

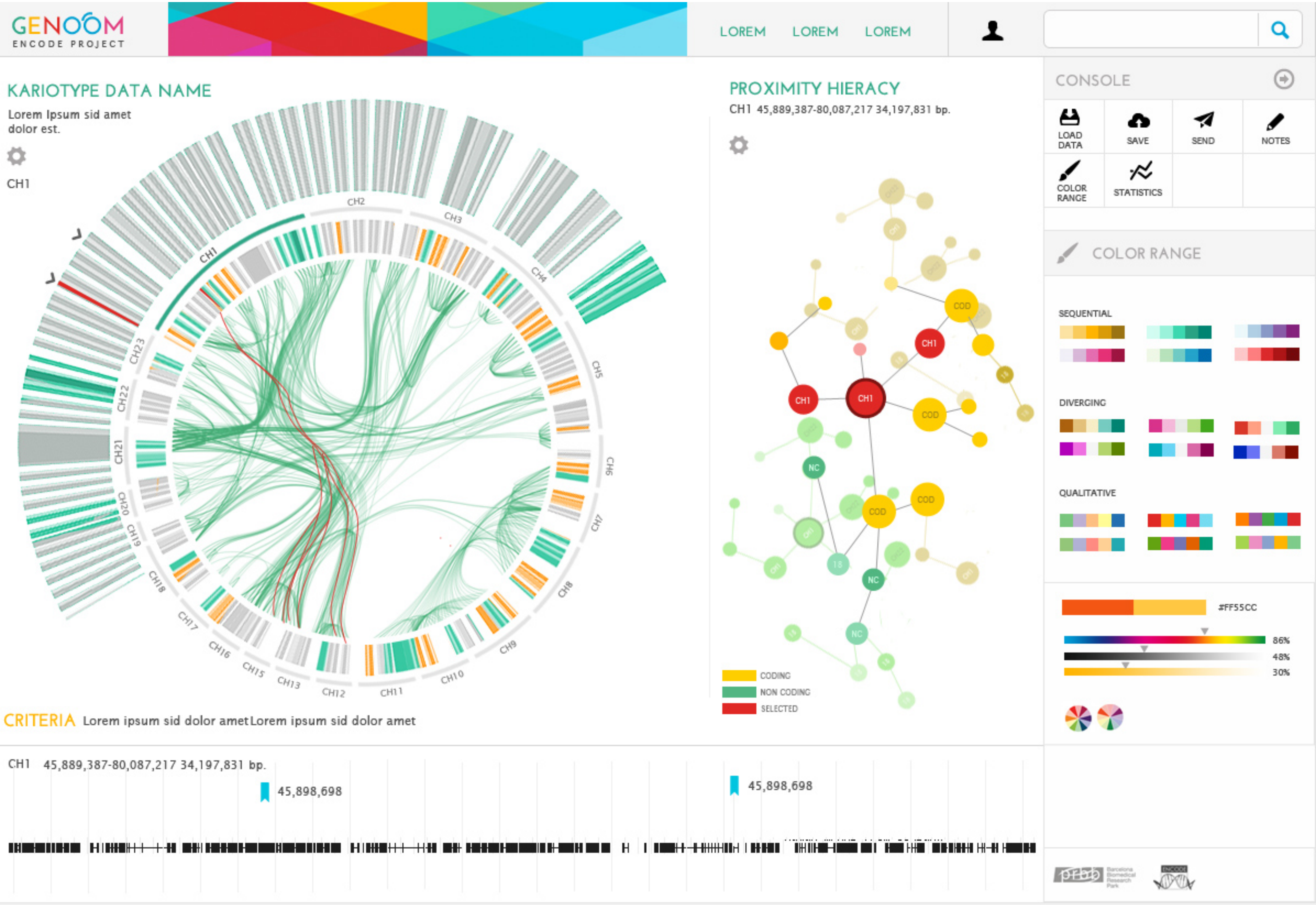
🟧 Pulmonary Trunck
Ascending aorta
Pulmonary veins

CH1 45,889,387-80,087,217 34,197,831 bp.

45,898,698

45,898,698

OneBigRobot



KARIOTYPE DATA NAME

Lorem Ipsum sid amet dolor est.

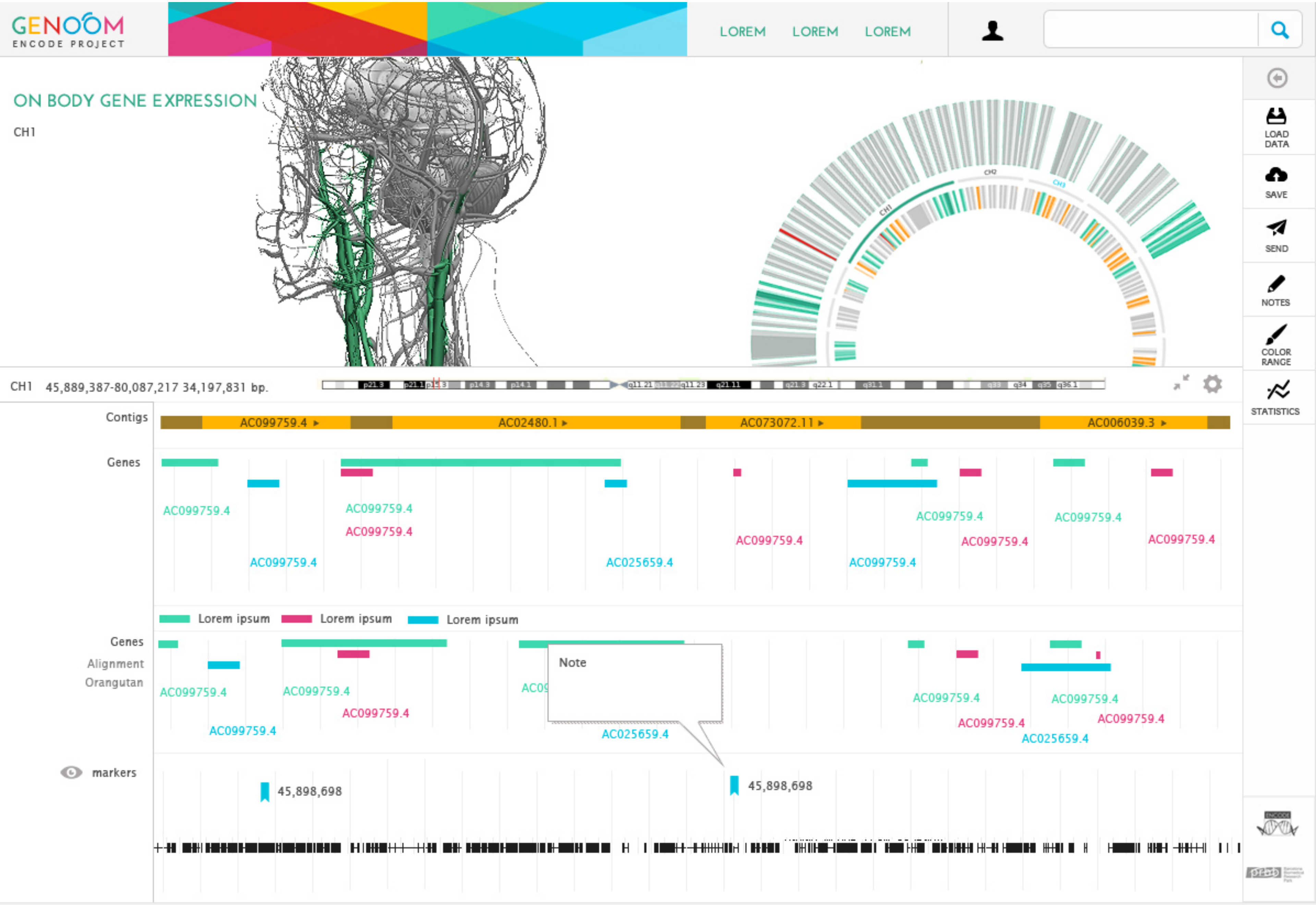
PROXIMITY HIERACY

CH1 45,889,387-80,087,217 34,197,831 bp.

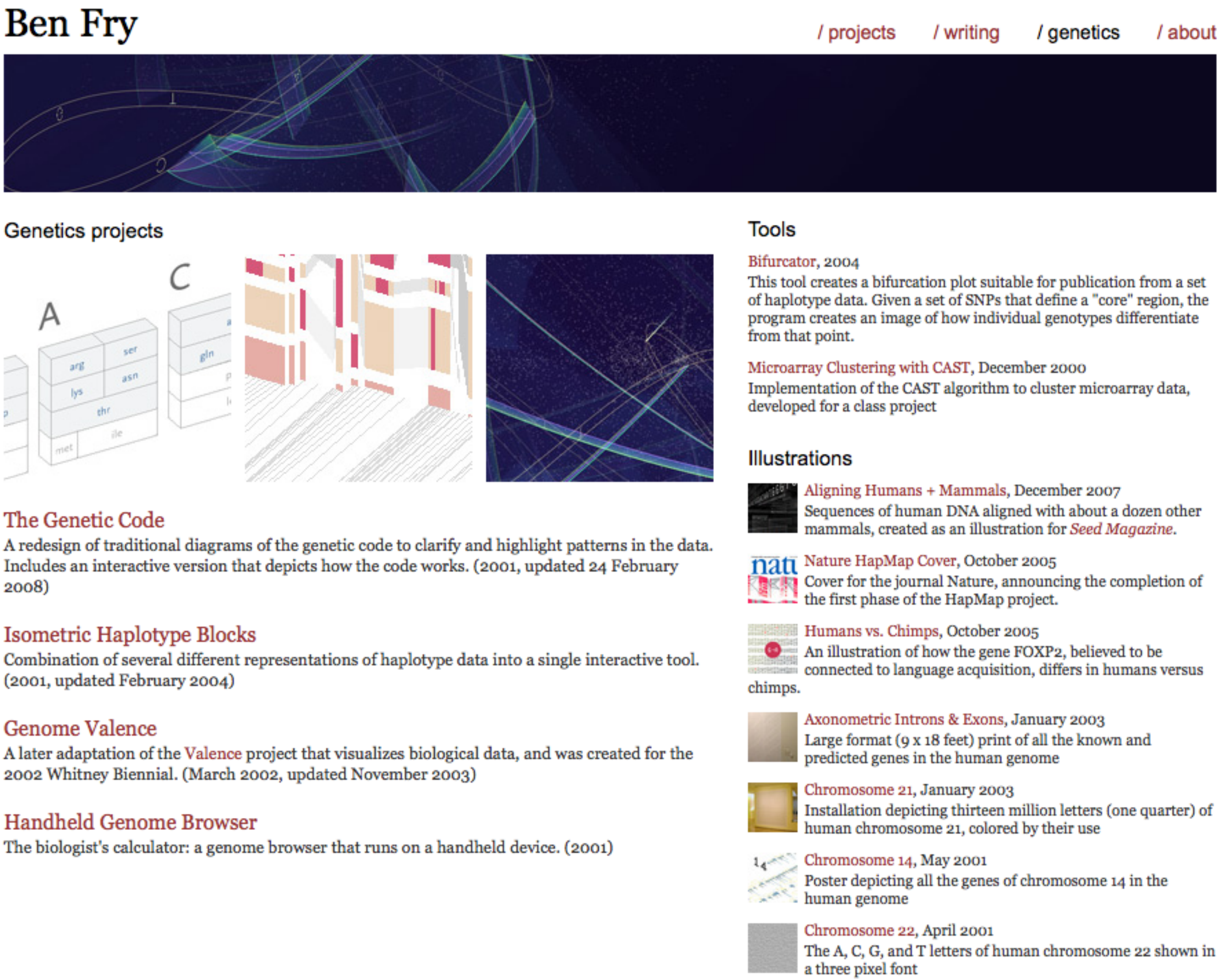
CRITERIA Lorem ipsum sid dolor ametLorem ipsum sid dolor amet

CH1 45,889,387-80,087,217 34,197,831 bp.

OneBigRobot



OneBigRobot



Ben Fry

/ projects / writing / genetics / about

Genetics projects

The Genetic Code

A redesign of traditional diagrams of the genetic code to clarify and highlight patterns in the data. Includes an interactive version that depicts how the code works. (2001, updated 24 February 2008)

Isometric Haplotype Blocks

Combination of several different representations of haplotype data into a single interactive tool. (2001, updated February 2004)

Genome Valence

A later adaptation of the Valence project that visualizes biological data, and was created for the 2002 Whitney Biennial. (March 2002, updated November 2003)

Handheld Genome Browser

The biologist's calculator: a genome browser that runs on a handheld device. (2001)

Tools

Bifurcator, 2004
This tool creates a bifurcation plot suitable for publication from a set of haplotype data. Given a set of SNPs that define a "core" region, the program creates an image of how individual genotypes differentiate from that point.

Microarray Clustering with CAST, December 2000
Implementation of the CAST algorithm to cluster microarray data, developed for a class project

Illustrations

Aligning Humans + Mammals, December 2007
Sequences of human DNA aligned with about a dozen other mammals, created as an illustration for *Seed Magazine*.

Nature HapMap Cover, October 2005
Cover for the journal Nature, announcing the completion of the first phase of the HapMap project.

Humans vs. Chimps, October 2005
An illustration of how the gene FOXP2, believed to be connected to language acquisition, differs in humans versus chimps.

Axonometric Introns & Exons, January 2003
Large format (9 x 18 feet) print of all the known and predicted genes in the human genome

Chromosome 21, January 2003
Installation depicting thirteen million letters (one quarter) of human chromosome 21, colored by their use

Chromosome 14, May 2001
Poster depicting all the genes of chromosome 14 in the human genome

Chromosome 22, April 2001
The A, C, G, and T letters of human chromosome 22 shown in a three pixel font

acknowledgements

- Sarah Djebali
- Anna Vlasova
- Julien Lagarde
- Didac Santesmasses (sunburst plots)
- Josep F. Abril (genome plots, University of Barcelona)
- Griselda Serra (human body browser, OneBigRobot)