Encyclopedia of short RNAs

A. Breschi - R. Murad

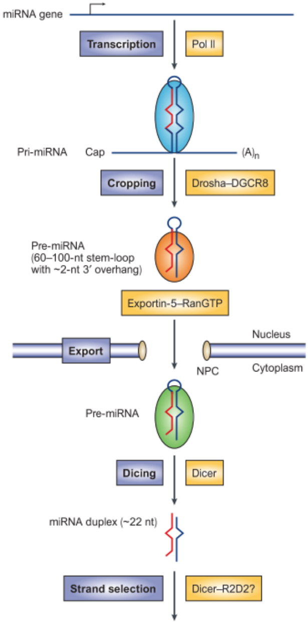
ENCODE AWG call

December 11th, 2015

What are miRNAs?

- non-coding RNAs
- small (approx 22 nt)
- Are derived from a primary transcript which carries one or more hairpin structures which are cleaved to produce the mature miRNA

Kim, 2005, Nat. Rev.



ENCODE3 CSHL shortRNA-seq samples - ENCSR171CXD

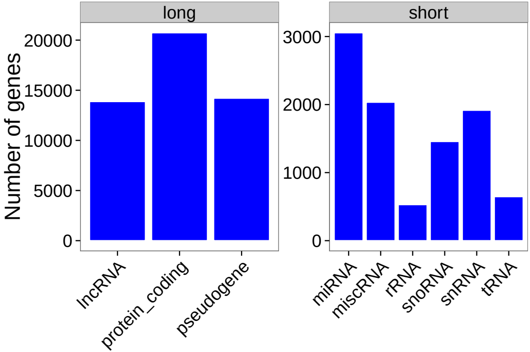
Human Cell Lines		
Cell	**Description**	**Rep**
A375	melanoma	2
GM12878	lymphoblastoid	2
HT1080	fibrosarcoma	2
K562	erythroleukemia	2
Karpas-42	B cell non-Hodgkin lymphoma	2
NCI-H460	lung cancer	4
Oci-Ly-7	B cell non-Hodgkin lymphoma	2
SK-MEL-5	melanoma	2
SK-N-DZ	neuroblastoma	4
bipolar spindle neuron		2
hepatocyte		2
neural progenitor cell		2
smooth muscle cell		2

Human Fetal Tissues			
Neuronal		**Others**	
Tissue	**Rep**	**Tissue**	**Rep**
cerebellum	2	heart	2
diencephalon	2	liver	1
frontal cortex	2	lung	2
occipital lobe	2	metanephros	2
parietal lobe	2	skeletal muscle tissue	2
temporal lobe	2	skin of body	2
spinal cord	2	thyroid gland	2
		tongue	2
		urinary bladder	1
		uterus	2

Annotation

GENCODE v19 (2013-12-05, hg19)

- 3,055 hairpins (pre-miRNAs), 20 on chrY



miRBase v19 (2012-7-23, hg19)

- 1,595 hairpins (pre-miRNAs), 2 on chrY
- 2,233 mature miRNAs
- 638 hairpins can give 2 mature miRNAs

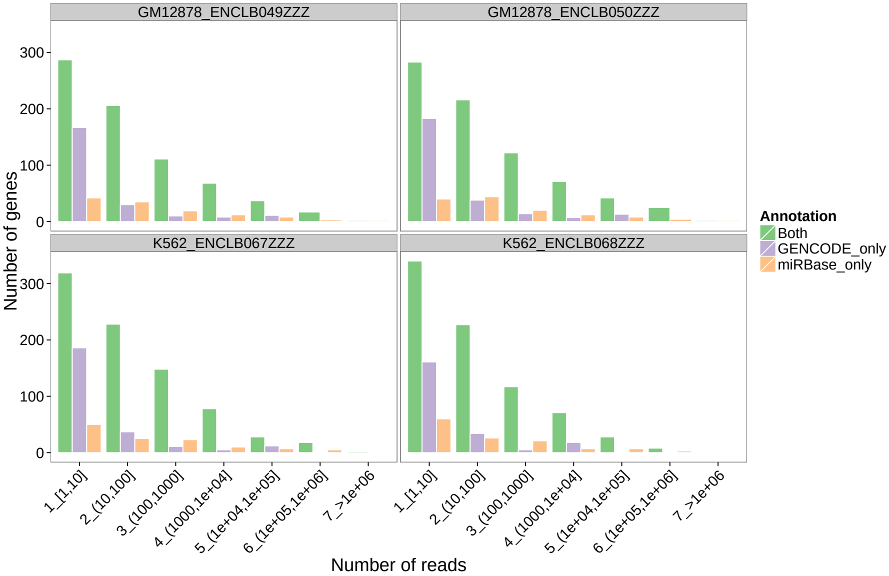
Comparison between GENCODE and miRBase:

- GENCODE does NOT annotate mature miRNAs
- 1,345 hairpins in common
- 250 miRBase only
- 1,710 GENCODE only

GENCODE annotation of miRNAs (`http://uswest.ensembl.org/info/genome/genebuild/ncrna.html`)
miRNAs are **predicted by BLASTN** of genomic sequence slices against miRBase sequences. All species are used. The BLAST hits are clustered and filtered by E value and the aligned genomic sequence is then checked for possible secondary structure using **RNAFold**. If evidence is found that the genomic sequence could form a stable hairpin structure, the locus is used to create a miRNA gene model. The resulting BLAST hit is used as supporting evidence for the miRNA gene.

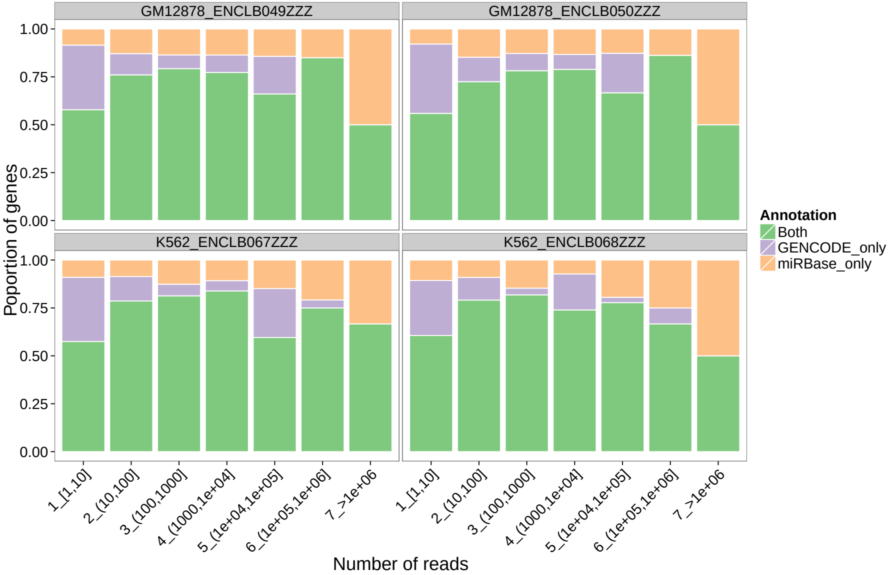
Note: The **miRNA identifier** and name are only associated to the resulting Ensembl miRNA if they are of the same species.

Hairpins common to both annotations are the most detected



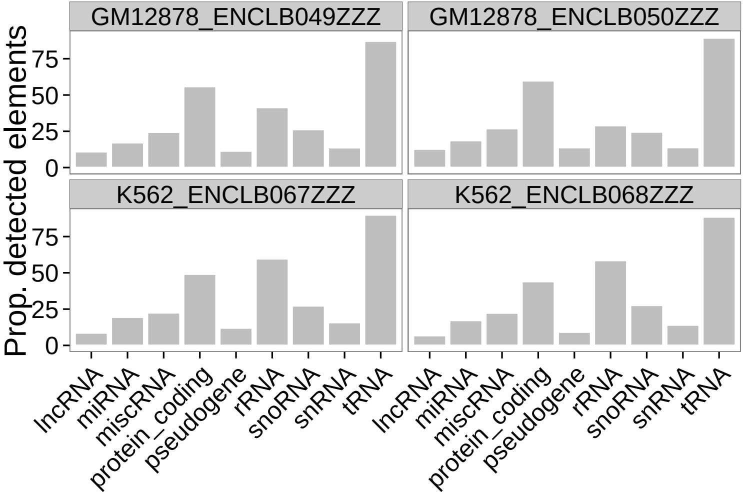
From submitted bam file

Hairpins common to both annotations are the most detected



From submitted bam file

Proportion of detected elements



At least 10 reads to call the element; reads overlapping multiple genes are counted multiple times

Discussion points

- Pre-miRNA (hairpin) annotation: miRBase vs GENCODE
  - Only GENCODE
  - Only miRBase (remove miRNA from GENCODE and concatenate miRBase to the remaining GENCODE annotation)
  - Union of GENCODE and miRBase, ask GENCODE if this is possible, or implemented in ENCODE for each new GENCODE/miRBase release
- Mature miRNAs annotation: not in GENCODE.
- Elements to report in the matrix:
  - All GENCODE elements (+ miRNAs as decided in the point above)
  - Only miRNA matrix as separate matrix
  - Only short RNA matrix as separate matrix
- ...

Moving to the new assembly

UCI-HudsonAlpha miRNA profiling experiments - Human

Human Samples

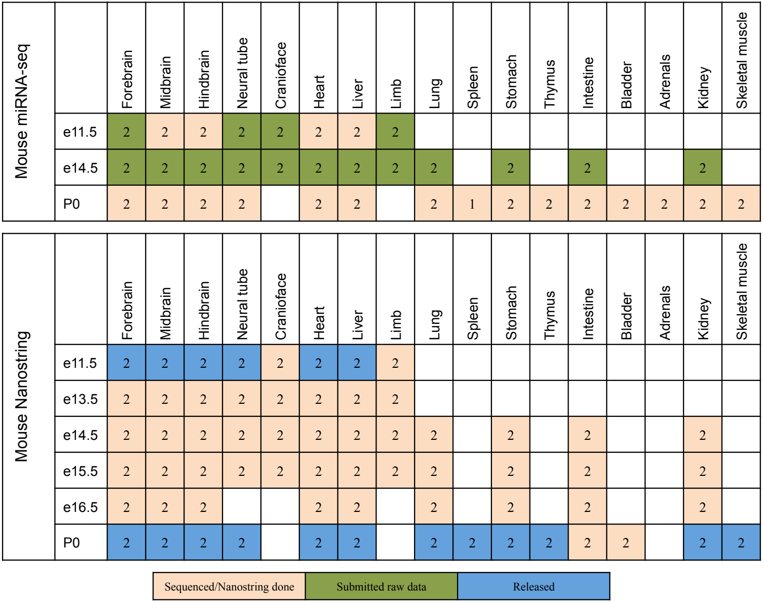
Cell	Nanostring done/ sequenced	Submitted	Released
GM12878			2/2
K562			2/2
H1-hESC	0/2		2/0
HepG2	0/2		2/0
HUVEC	0/2	2	
HCT116	0/2		2/0
LHCN-diff	2/2		
MCF7	0/2	2/0	
A549 EtOH treated	2/1		
HTB11	2		
A549 DXM treated	2/1		
HeLa	0/2		
A549	0/1		
GM12891	1/0		
LHCN cycling	0/1		
NHEK	0/1		
H9 hepatocyte	2/2		
H9 neuron		2/2	
H9 smooth muscle		2/2	
Bipolar spindle neuron		2/2	
Fibroblast of arm		2/2	
Induced pluripotent stem cells	2/2		

ENTEX sample	Nanostring done/ sequenced	Submitted	Released
Thyroid, donor 1	1/0		
Thyroid, donor 2	2/0		
Adrenal gland, donor 1	2/0		
Adrenal gland, donor 2	2/0		
Leg skin, donor 1	2/0		
Leg skin, donor 2	2/0		
Skeletal muscle, donor 1	2/0		
Skeletal muscle, donor 2	2/0		
Adrenal gland (ENTEX 49-50)	2/0		
Adrenal gland (ENTEX 51-52)	2/0		
Leg skin (ENTEX 42-43)	2/0		
Leg skin (ENTEX 44-45)	2/0		

Nanostring done/ sequenced	Submitted raw data	Released

(# of Nanostring reps / # of miRNA-seq reps)

UCI-HudsonAlpha miRNA profiling experiments - Mouse



Mouse miRNA-seq		Forebrain	Midbrain	Hindbrain	Neural tube	Cranioface	Heart	Liver	Limb	Lung	Spleen	Stomach	Thymus	Intestine	Bladder	Adrenals	Kidney	Skeletal muscle
	e11.5	🟩 2	🟧 2	🟧 2	🟩 2	🟩 2	🟧 2	🟧 2	🟩 2									
	e14.5	🟩 2	🟩 2	🟩 2	🟩 2	🟩 2	🟩 2	🟩 2	🟩 2	🟩 2		🟩 2		🟩 2			🟩 2	
	P0	🟧 2	🟧 2	🟧 2	🟧 2		🟧 2	🟧 2		🟧 2	🟧 1	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2

Mouse Nanostring		Forebrain	Midbrain	Hindbrain	Neural tube	Cranioface	Heart	Liver	Limb	Lung	Spleen	Stomach	Thymus	Intestine	Bladder	Adrenals	Kidney	Skeletal muscle
	e11.5	🟦 2	🟦 2	🟦 2	🟦 2	🟧 2	🟦 2	🟦 2	🟧 2									
	e13.5	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2									
	e14.5	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2		🟧 2		🟧 2			🟧 2	
	e15.5	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2		🟧 2		🟧 2			🟧 2	
	e16.5	🟧 2	🟧 2	🟧 2			🟧 2	🟧 2		🟧 2		🟧 2		🟧 2			🟧 2	
	P0	🟦 2	🟦 2	🟦 2	🟦 2		🟦 2	🟦 2		🟦 2	🟦 2	🟦 2	🟦 2	🟧 2	🟧 2		🟦 2	🟦 2

🟧 Sequenced/Nanostring done	🟩 Submitted raw data	🟦 Released

Integrating NanoString annotation

Genome: assembly GRCh38
GCA_000001405.15 (no-alt analysis set)

- chr 1-22, X, Y, M
- 42 unlocalized scaffolds from chr 1-5, 9, 11, 14-17, 22, Y (random)
- 127 unplaced scaffolds (chr unknown)
- EBV genome

Annotation: Gencode v23
4,205 miRNAs (same in Gencode v24)

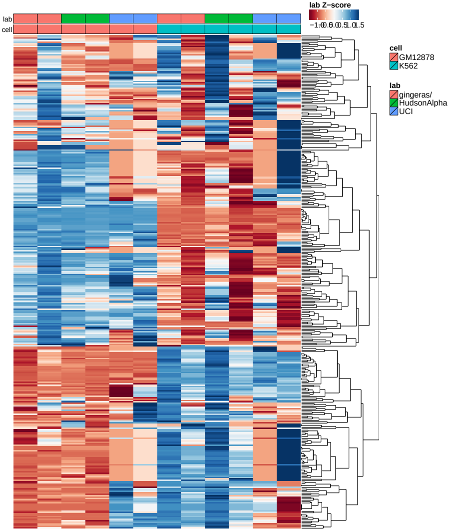
NanoString probes:

- 19 controls
- 743 miRNAs, of which:
- 80 viruses (of which 34 EBV)
- 654 human, of which:
- 17 are withdrawn from miRBase v21 because part of larger RNA molecules (e.g. tRNAs)
- 637 in miRBase v21 (belonging to 666 pre-miRNAs, see suppl. slides for detail)
- 619/637 in Gencode v23 (belonging to **647**/666 pre-miRNAs)

`ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz`

Removing protocol variation highlights cell-specific pre-miRNAs

Scaling within lab



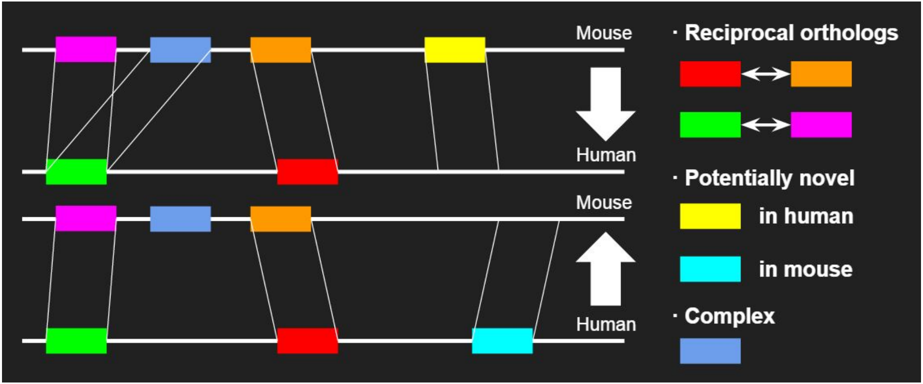
Scaling within each experiment

RNA evaluation samples. Filter for expression –> 353 pre-miRNAs

Comparing miRNAs in human and mouse

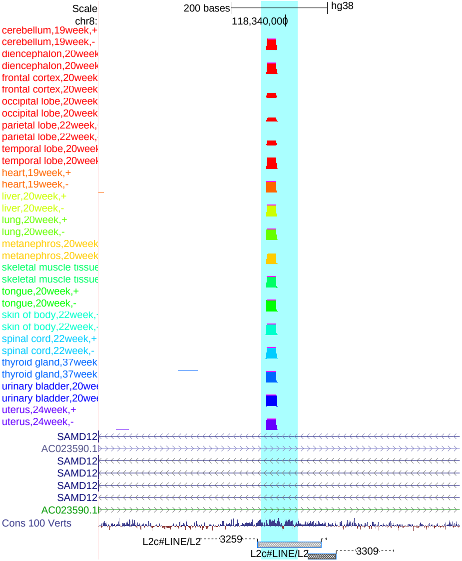
Finding orthologous pre-miRNAs between human and mouse

Human annotation: Gencode v23, hg38 (GRCh38), 4,205 miRNAs
Mouse annotation: Gencode vM7, mm10 (GRCm38), 2,202 miRNAs

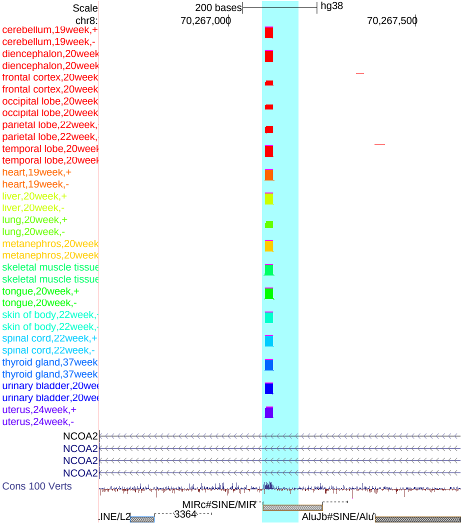


	Reciprocal orth.	Novel (detected)	Complex
Human	504	606 (16)	103
Mouse	504	1081 (?)	109

chr8:118339950-118340025; ENSMUSG00000104794.1; mmu-mir-28c



chr8:70267089-70267186; ENSMUSG00000099183.1; Gm27881



Acknowledgments

Guigó lab

- Roderic Guigó
- Sarah Djebali

Gingeras lab

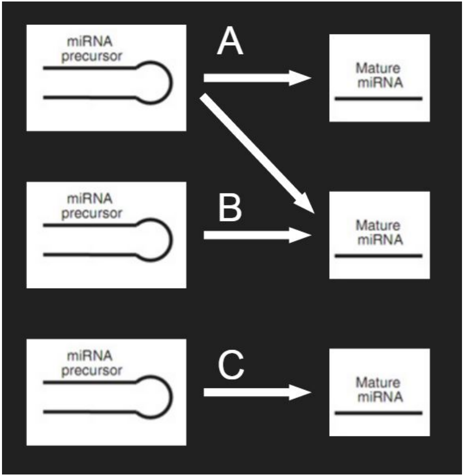
- Tom Gingeras
- Carrie Davis
- Alex Dobin

Mortazavi lab

- Ali Mortazavi
- Rabi Murad

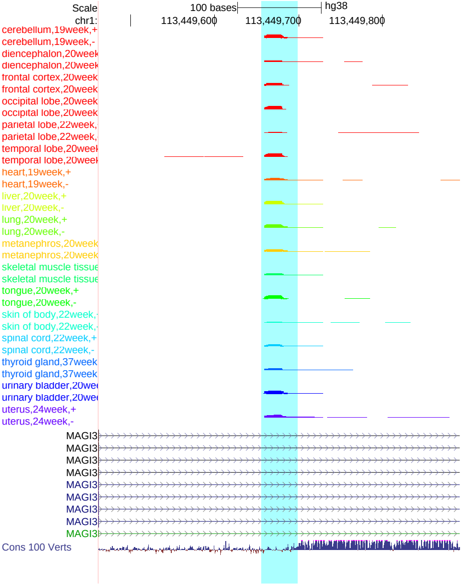
Suppplementary slides

Nanostring probes

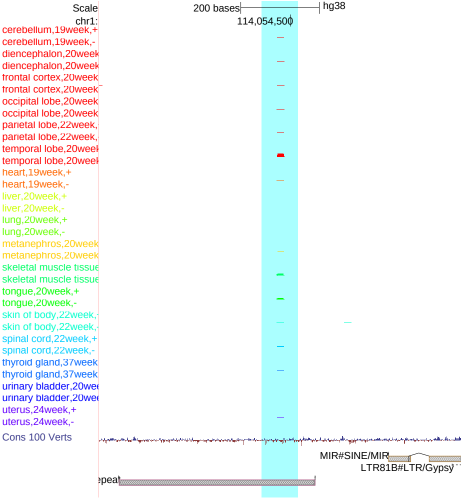


- A: one-to-two
  - 150 cases, involving:
  - 75 primary miRNAs
  - 132 mature miRNAs
  - –> merge and take the max in each sample
- B: many-to-one (up to 10)
  - 176 cases, involving:
  - 162 primary miRNAs
  - 72 mature miRNAs
  - –> duplicate lines
- C: one-to-one
  - 452 cases
- many-to-many
  - 37 cases, involving:
  - 23 primary miRNAs
  - 19 mature miRNAs

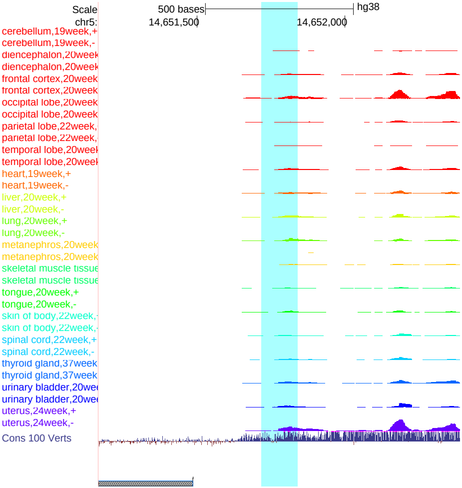
chr1:113449656-113449699; ENSMUSG00000092747.1; Gm22180



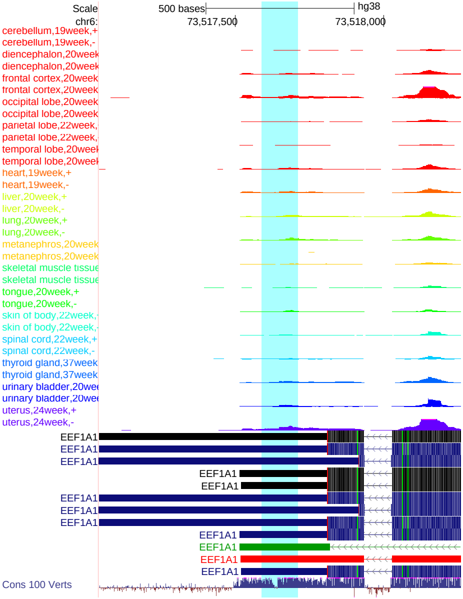
chr1:114054427-114054519; ENSMUSG00000084508.1; Gm25325



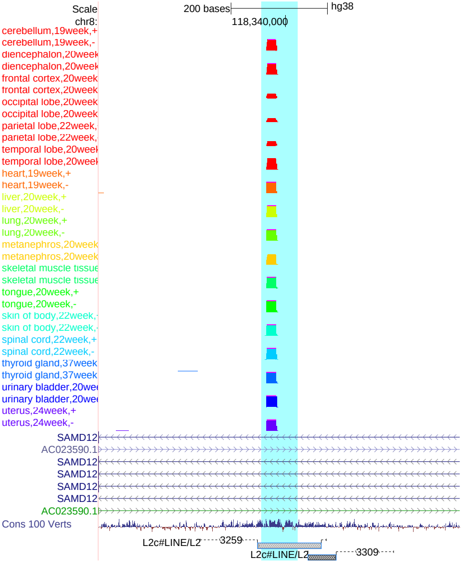
chr5:14651717-14651839; ENSMUSG00000076137.1, ENSMUSG00000105814.1; Gm26384



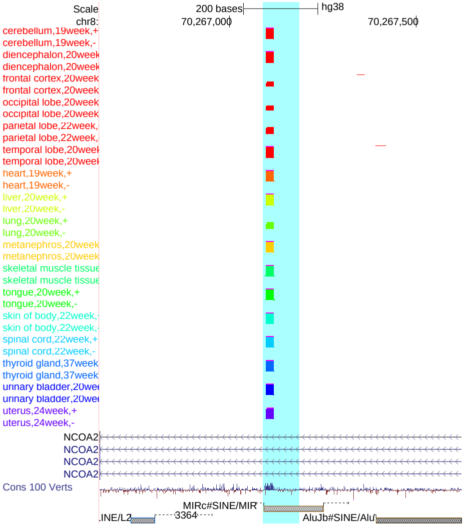
chr6:73517589-73517711; ENSMUSG00000076138.1; Gm26377



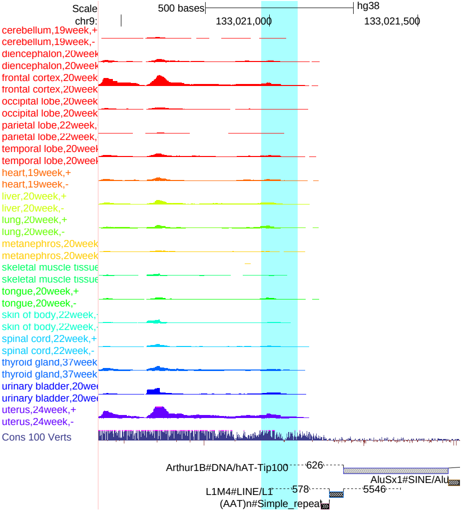
chr8:118339950-118340025; ENSMUSG00000104794.1; mmu-mir-28c



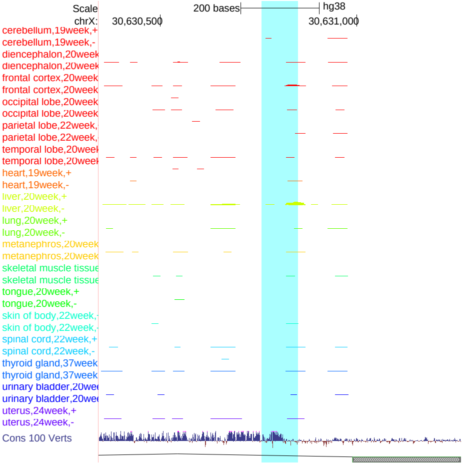
chr8:70267089-70267186; ENSMUSG00000099183.1; Gm27881



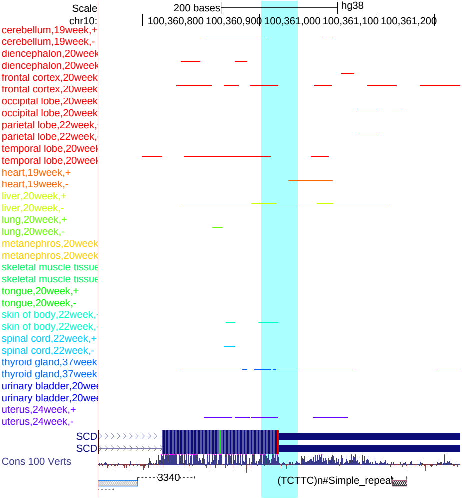
chr9:133020972-133021094;
ENSMUSG00000076135.1,
ENSMUSG00000076136.1; Gm24276



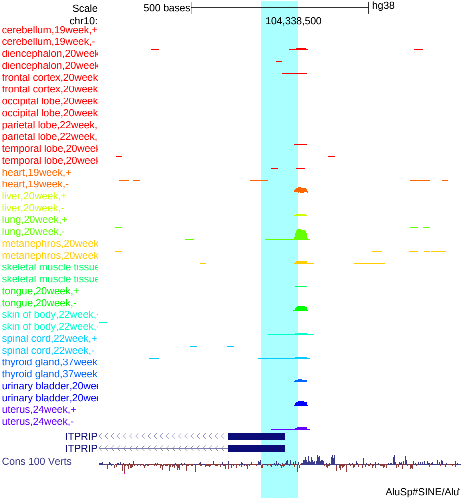
chrX:30630758-30630850;
ENSMUSG00000096567.1; Gm24098



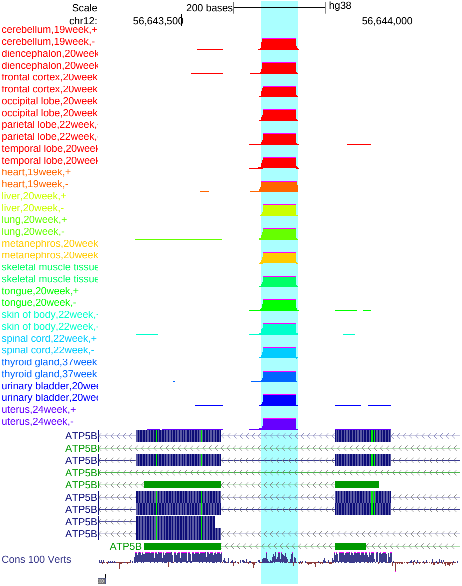
chr10:100360903-100360965; ENSMUSG00000093315.1; Mir5114



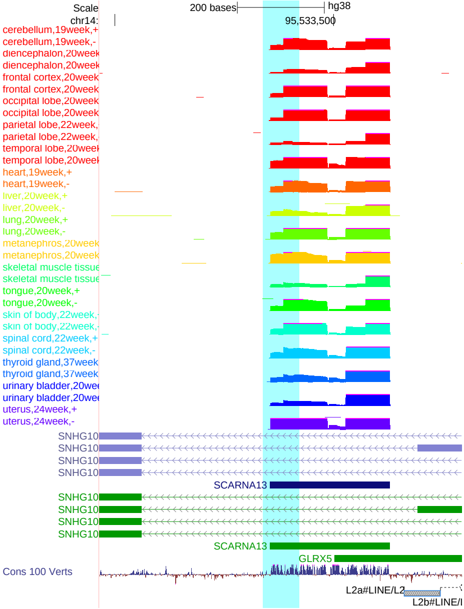
chr10:104338338-104338440; ENSMUSG00000092966.1; Gm22240



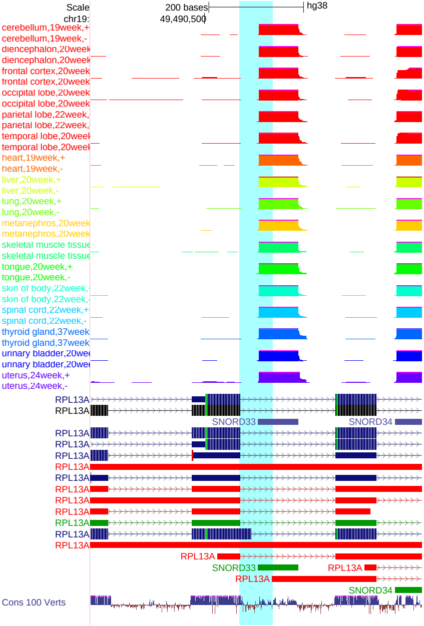
chr12:56643675-56643754; ENSMUSG00000104755.1; mmu-mir-677



chr14:95533338-95533421; ENSMUSG00000106366.1; mmu-mir-3069



chr19:49490575-49490645; ENSMUSG00000105913.1; mmu-mir-5121



chr20:37523524-37523615; ENSMUSG00000077990.1; Gm23134