PseudoSeq Update

7/13/11

genome-anno

LH

Pseudogenes in GENCODE v7

- ENSEMBL / HAVANA transcripts (gt & tt: protein_coding): 76006

- HAVANA pgenes (gt: pgene): 12159

- HAVANA pgenes (gt: pgene, tt: processed, transcribed): 141
- HAVANA pgenes (gt: pgene, tt: unprocessed, transcribed): 268
  - **Gold standard (positives): 409**

- HAVANA pgenes (gt: pgene, tt: unprocessed, not transcribed): 8107
- HAVANA pgenes (gt: pgene, tt: processed, not transcribed): 1860
  - **Pseudogene set: 9967**

Alignment of pseudogenes to the reference genome

- Alignment tool: BLAT
- Reference genome: hg19
- Extract all alignment blocks: at least one of the alignment block has to be longer than 75 nucleotides

Gold standard alignments

- Pgenes with (total: 409):
  - Zero alignment blocks: 89 (22%)
  - One alignment blocks: 84 (21%)
  - Multiple (2 - 5) alignment blocks: 105 (26%)
  - Too many (> 5) alignment blocks: 131 (31%)

Pseudogene alignments

- Pgenes with (total: 9967):
  - Zero alignment blocks: 3198 (32 %)
  - One alignment blocks: 1907 (19%)
  - Multiple (2 - 5) alignment blocks: 2150 (22%)
  - Too many (> 5) alignment blocks: 2712 (27%)

RNA-Seq data

- Human Body Map
  - 16 tissues
  - 75 nucleotide single-end reads
  - HiSeq Illumina platform: 1 lane per tissue

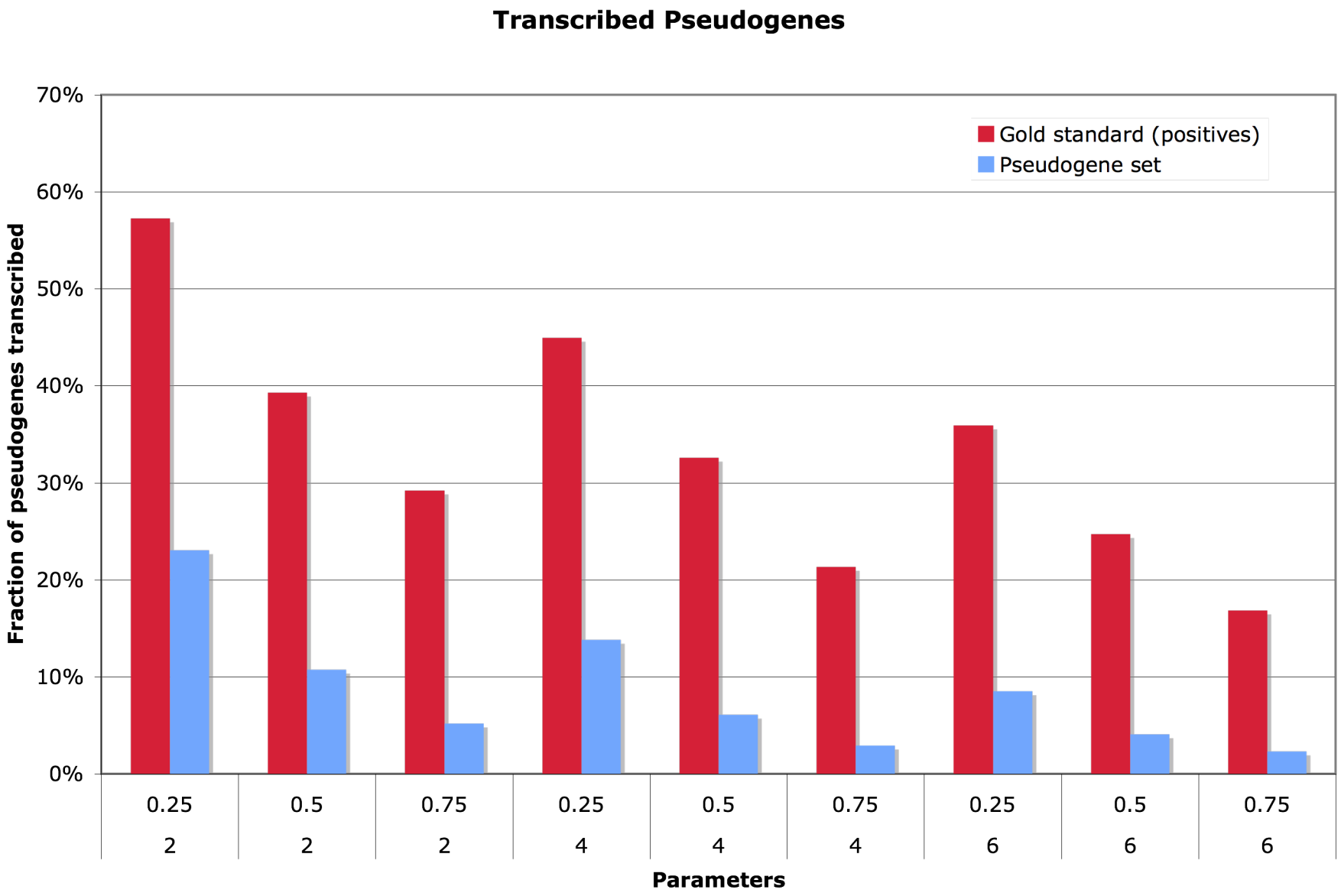
- Alignment tool: bowtie
  - hg19 + gencodeV7 splice junction library (alignment is performed concurrently)
  - Unique mapping
  - 2 mismatches (end-to-end)

- Average number of mapped reads per tissue: ~ 60M

Pseudogene expression Case 1: zero alignment pairs

2 parameters:

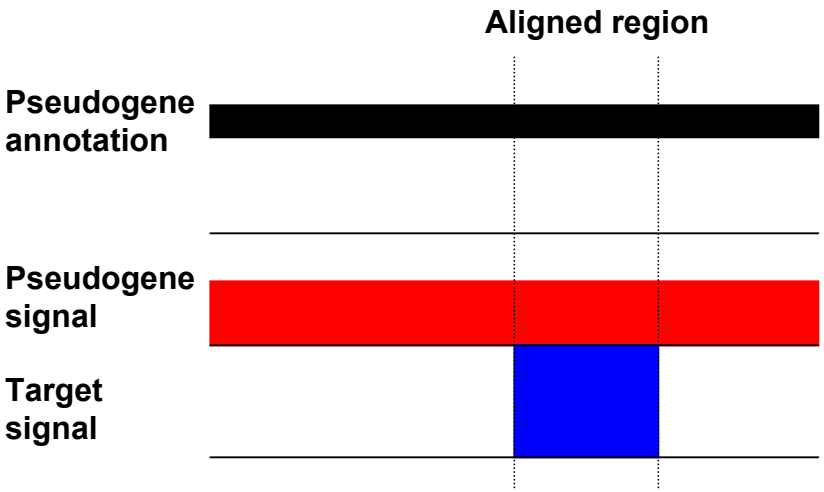
- Minimum read coverage (number of reads per position):
  - 2, 4, 6
- Fraction of pseudogene length with minimum read coverage (mapped reads)
  - 0.25, 0.5, 0.75



Totals: Gold standard = 89, Pseudogene set = 3198

Pseudogene expression
Case 2: one alignment pairs

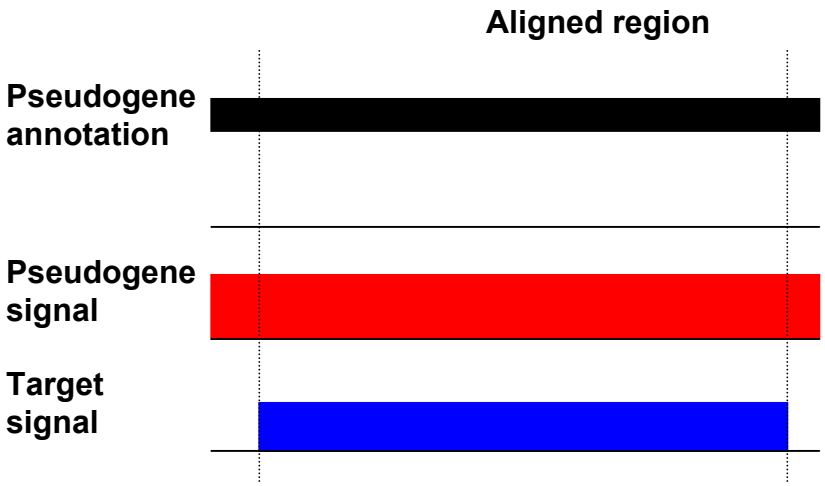
Case 2.1 - aligned region represents minority of pseudogene



Minority of cases

Check if there is significant signal outside the aligned region to determine if pseudogene is transcribed

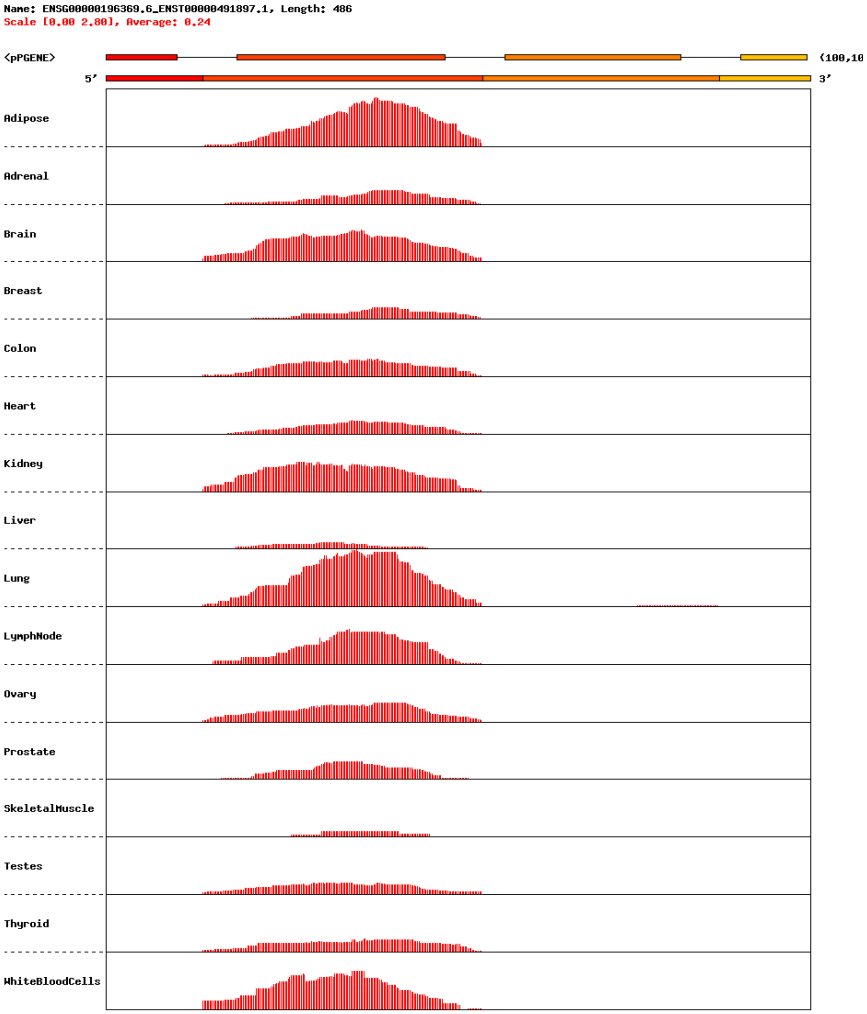
Case 2.2 - aligned region represents majority of pseudogene

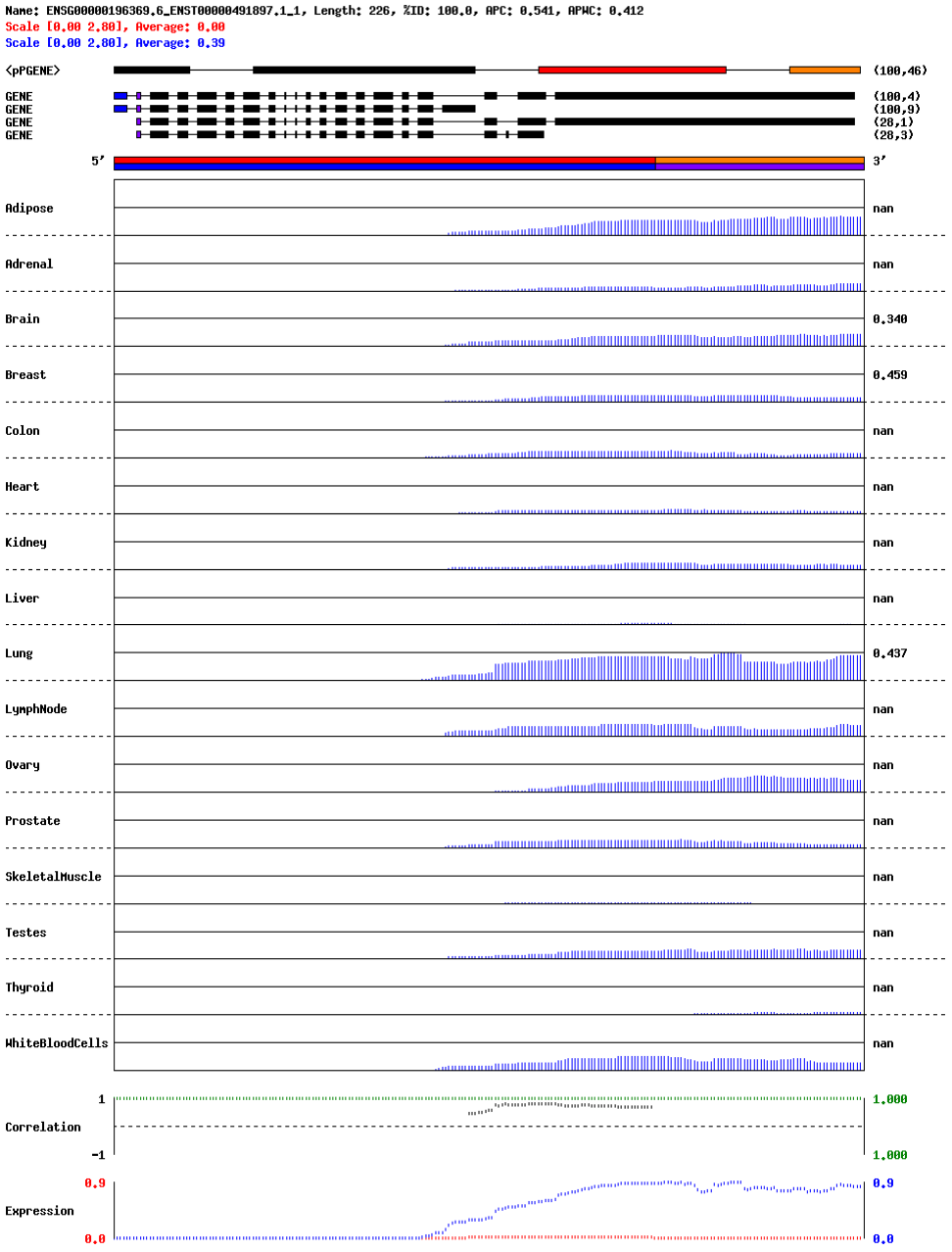


Majority of cases

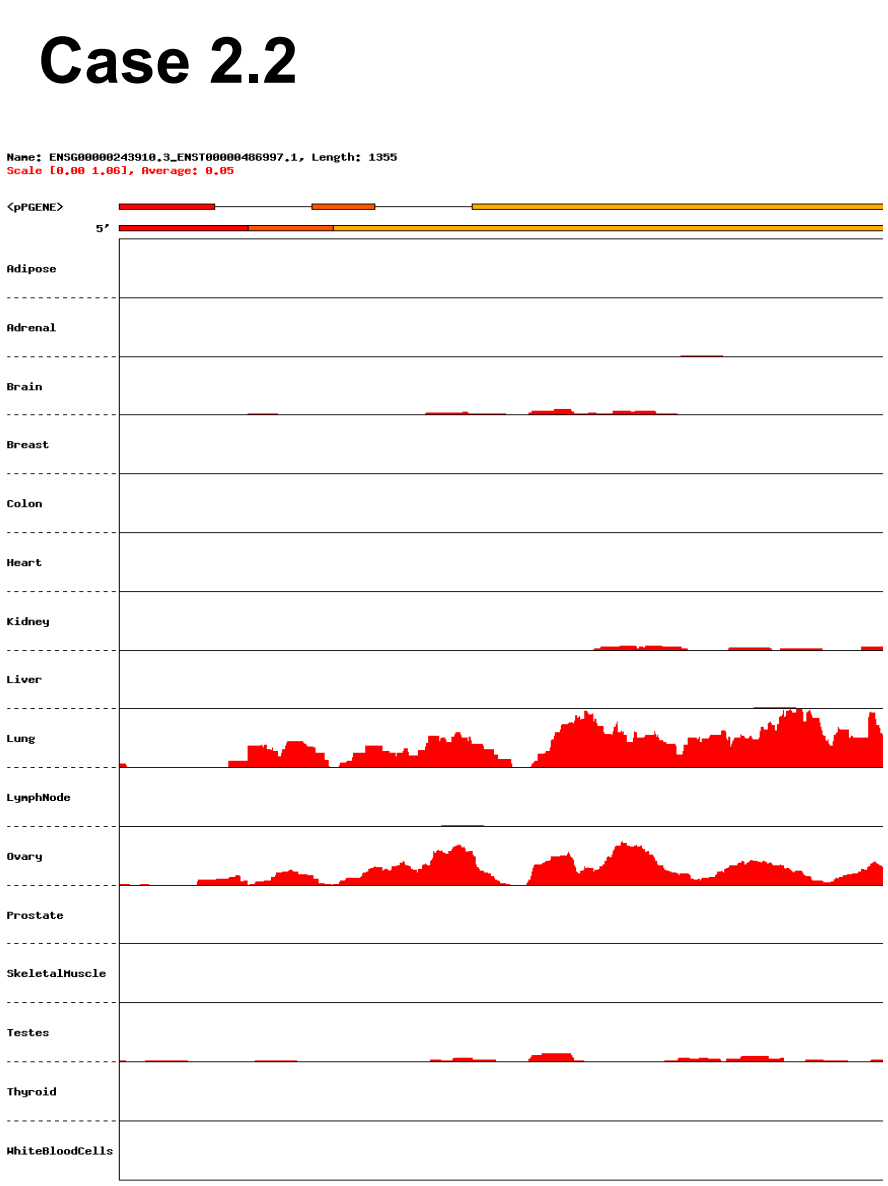
1. **Signal (pgene) > Signal (target) in at least one tissue => transcribed pseudogene**
2. **Signal (pgene) mirrors Signal (target) across all samples => mapping artifact**
3. **Signal (pgene) is independent of Signal (target) across samples => transcribed pseudogene**

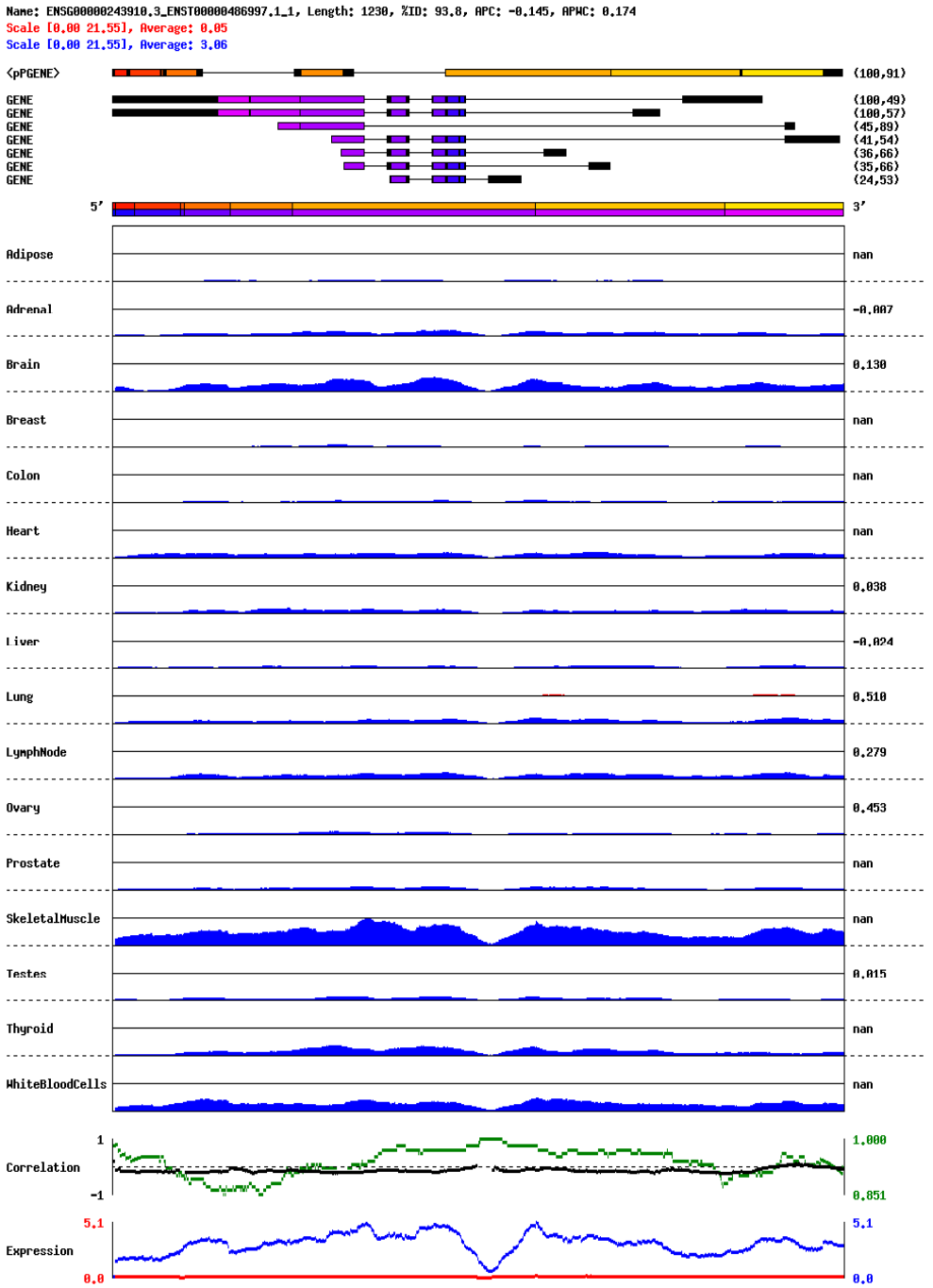
Case 2.1


Name: ENSG00000196369.6_ENST00000491897.1, Length: 486
Scale [0.00 2.80], Average: 0.24
<pPGENE>
(100,10
5'
3'
Adipose
Adrenal
Brain
Breast
Colon
Heart
Kidney
Liver
Lung
LymphNode
Ovary
Prostate
SkeletalMuscle
Testes
Thyroid
WhiteBloodCells


Name: ENSG00000196369.6_ENST00000491897.1_1, Length: 226, %ID: 100.0, APC: 0.541, APWC: 0.412
Scale [0.00 2.80], Average: 0.00
Scale [0.00 2.80], Average: 0.39
<pPGENE>
(100,46)
GENE
(100,4)
GENE
(100,9)
GENE
(28,1)
GENE
(28,3)
5'
3'
Adipose
nan
Adrenal
nan
Brain
0.340
Breast
0.459
Colon
nan
Heart
nan
Kidney
nan
Liver
nan
Lung
0.437
LymphNode
nan
Ovary
nan
Prostate
nan
SkeletalMuscle
nan
Testes
nan
Thyroid
nan
WhiteBloodCells
nan
1
Correlation
-1
1.000
1.000
0.9
Expression
0.0
0.9
0.0

Case 2.2


Name: ENSG00000243910.3_ENST00000486997.1, Length: 1355
Scale [0.00 1.06], Average: 0.05
<pPGENE>
5'
Adipose
Adrenal
Brain
Breast
Colon
Heart
Kidney
Liver
Lung
LymphNode
Ovary
Prostate
SkeletalMuscle
Testes
Thyroid
WhiteBloodCells


Name: ENSG00000243910.3_ENST00000486997.1_1, Length: 1230, %ID: 93.8, APC: -0.145, APWC: 0.174
Scale [0.00 21.55], Average: 0.05
Scale [0.00 21.55], Average: 3.06
<pPGENE>
(100,91)
GENE (100,49)
GENE (100,57)
GENE (45,89)
GENE (41,54)
GENE (36,66)
GENE (35,66)
GENE (24,53)
5'
3'
Adipose nan
Adrenal -0.007
Brain 0.130
Breast nan
Colon nan
Heart nan
Kidney 0.038
Liver -0.024
Lung 0.510
LymphNode 0.279
Ovary 0.453
Prostate nan
SkeletalMuscle nan
Testes 0.015
Thyroid nan
WhiteBloodCells nan
Correlation
1
-1
1.000
0.851
Expression
5.1
0.0
5.1
0.0