PseudoSeq: Identification of transcribed pseudogenes using multiple RNA-Seq samples

Lukas Habegger
07/07/2011
Gerstein Lab
Yale University

Objective

- Identification of transcribed pseudogenes
- Compare the expression patterns of the pseudogene to its corresponding parent gene
  - Use of multiple RNA-Seq samples
  - Distinguish between potentially transcribed pseudogenes and mapping artifacts

Pseudogenes in GENCODE v6

- ENSEMBL / HAVANA transcripts (gt & tt: protein_coding): 74,226

- ENSEMBL pgenes (gt: pgene): 2,252

- HAVANA pgenes (gt: pgene): 10,343
- HAVANA pgenes (gt: pgene, tt: processed, not transcribed): **7,887**
- HAVANA pgenes (gt: pgene, tt: processed, transcribed): 24
- HAVANA pgenes (gt: pgene, tt: unprocessed, not transcribed): 1,793
- HAVANA pgenes (gt: pgene, tt: unprocessed, transcribed): 20

Alignment of pseudogenes to the reference genome

- Input: HAVANA pgenes (gt: pgene, tt: processed, not transcribed): **7,887**
- Alignment tool: BLAT
- Reference genome: hg19
- Extract all alignment blocks: at least one of the alignment block has to be longer than 75 nucleotides

Characterization of pseudogene alignments

- Pgenes with:
  - Zero alignment blocks: 2,670
  - One alignment blocks: 1,538
  - Multiple (2 - 5) alignment blocks: 1,593
  - Too many (> 5) alignment blocks: 2,086

RNA-Seq data

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

- Alignment tool: bowtie
  - hg19 + gencodeV6 splice junction library (alignment is performed concurrently)
  - Unique mapping
  - 2 mismatches (end-to-end)

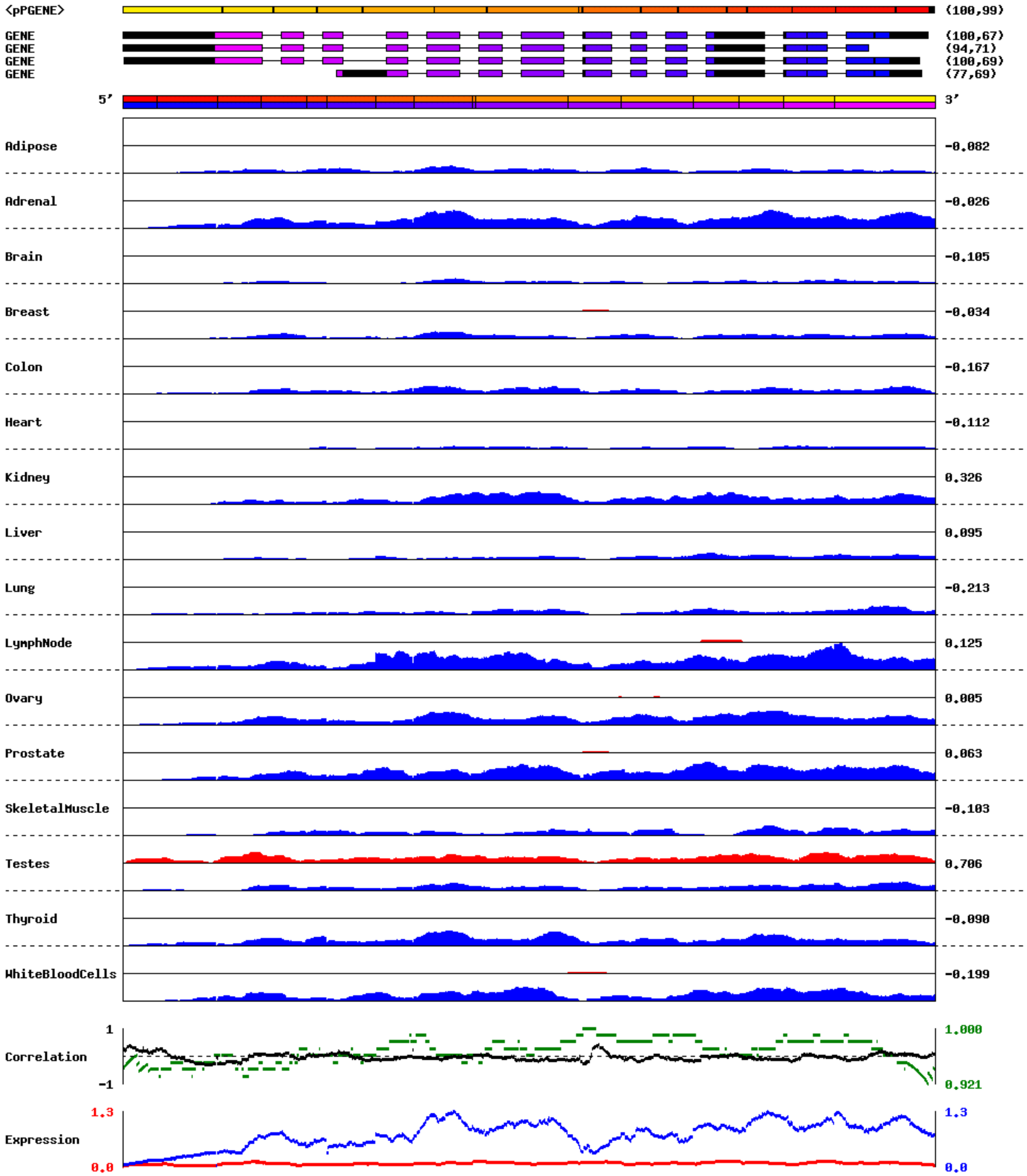
- Average number of mapped reads per tissue: ~ 60M

Results

Category	**Total number of pgenes**	**Number of pgenes with non-zero expression values**	**Number of pgenes with zero expression values**	**Number of pgenes with >0.5 expression values**
Zero alignment pairs	2670	1530	1140	16
One alignment pair	1538	1165	373	33
2 - 5 alignment pairs	1593	1108	485	61
Too many alignment pairs	2086	N/A	N/A	N/A

Example

Name: ENSG00000232553_ENST00000416636_1, Length: 1461, %ID: 96.6, APC: -0.052, APWC: 0.012
Scale [0.00 4.02], Average: 0.06
Scale [0.00 4.02], Average: 0.73


<pPGENE>
(100,99)
GENE
(100,67)
GENE
(94,71)
GENE
(100,69)
GENE
(77,69)
5’
3’
Adipose
Adrenal
Brain
Breast
Colon
Heart
Kidney
Liver
Lung
LymphNode
Ovary
Prostate
SkeletalMuscle
Testes
Thyroid
WhiteBloodCells
Correlation
Expression