Annotation Pipeline

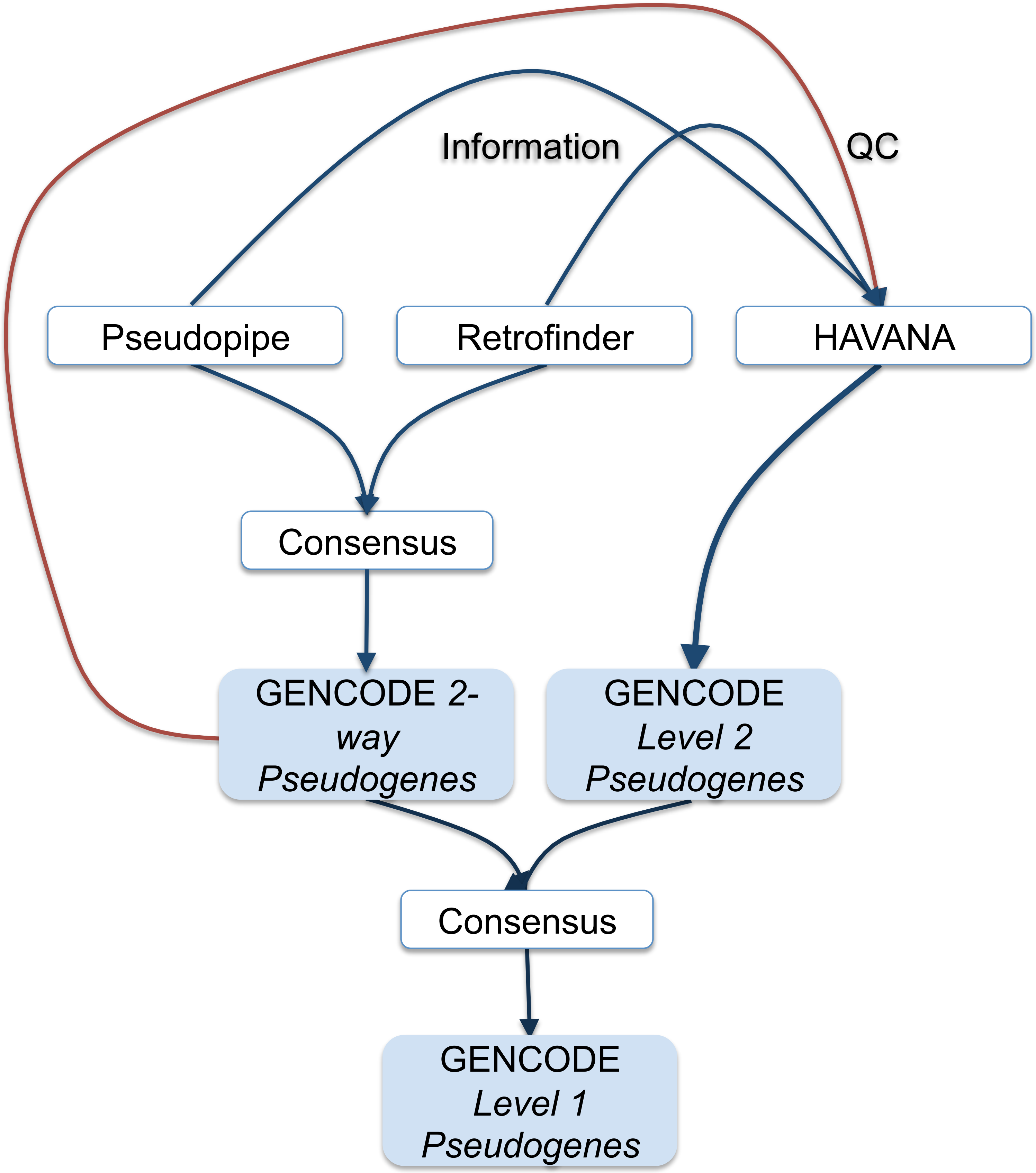

Information
QC
Pseudopipe
Retrofinder
HAVANA
Consensus
GENCODE 2-way Pseudogenes
GENCODE Level 2 Pseudogenes
Consensus
GENCODE Level 1 Pseudogenes

Figure 1. Pipeline for pseudogene annotation

GENCODE Pseudogene Annotation Growth

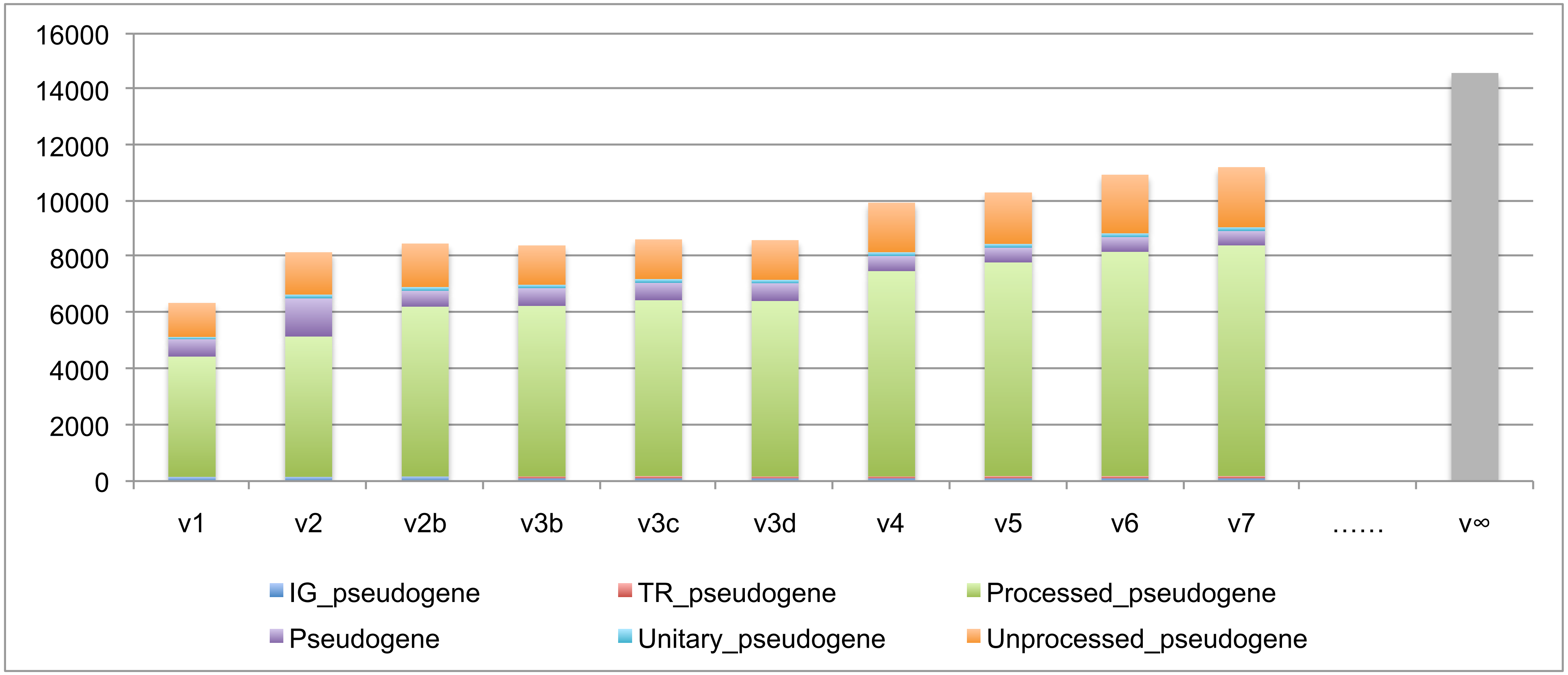


Figure 2. Growth of pseudogene annotation. Number of pseudogene in the whole genome was extrapolated.

GENCODE Pseudogene Parents

Figure 3. Identification of GENCODE pseudogene parents.

- UniProt proteins used by HAVANA annotation;

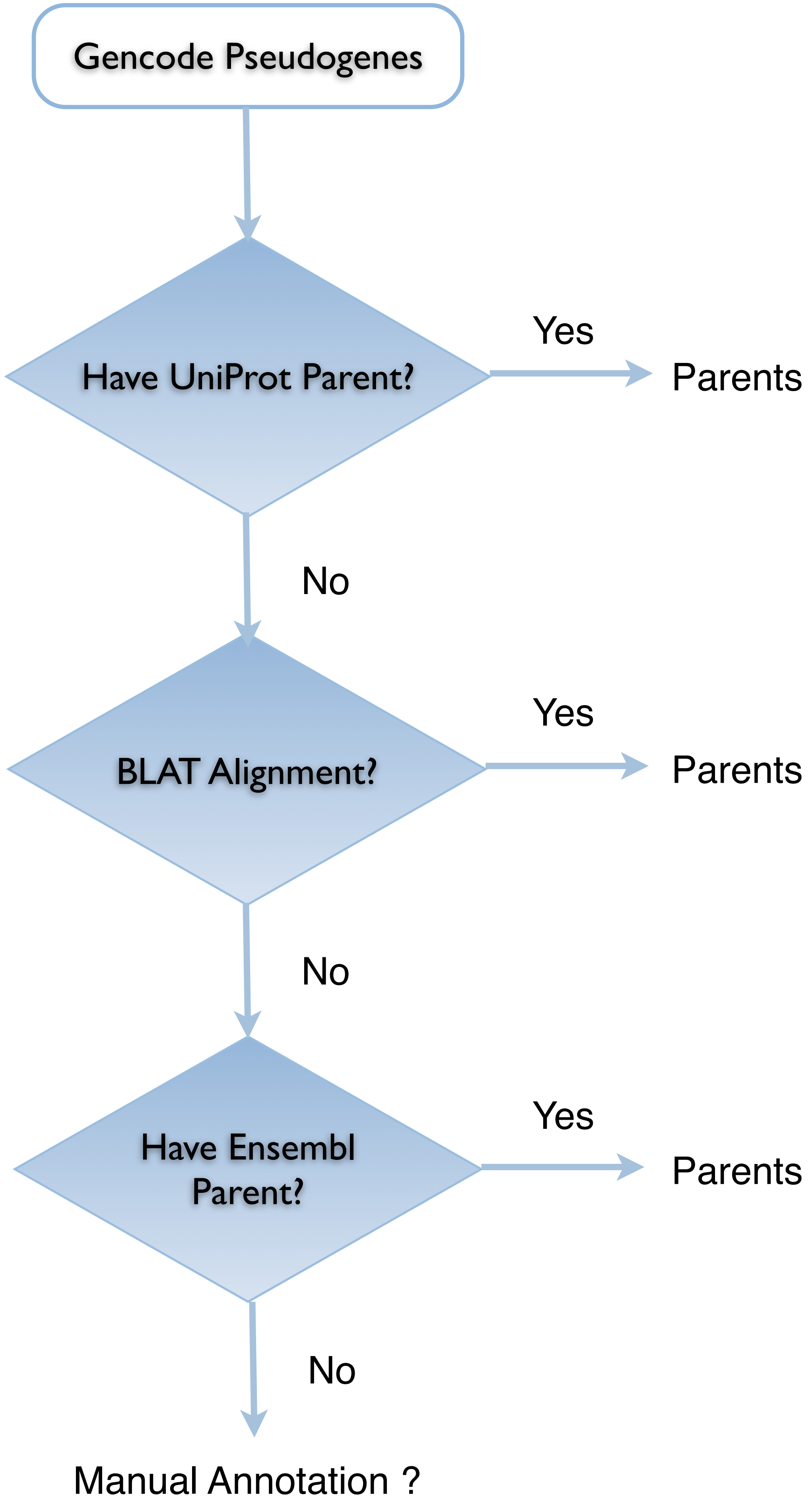
  Parent gene symbols were parsed and mapped to Ensembl 62

- Alignment of pseudogenes to genome

  BLAT pseudogene exons against genome, and identify one-to-one matching regions which overlap coding sequences

- Ensemble peptides used by PseudoPipe

  Find the PseudoPipe output consensus to GENCODE pseudogene, and get PseudoPipe parent information.



Resurrected Unprocessed Pseudogene

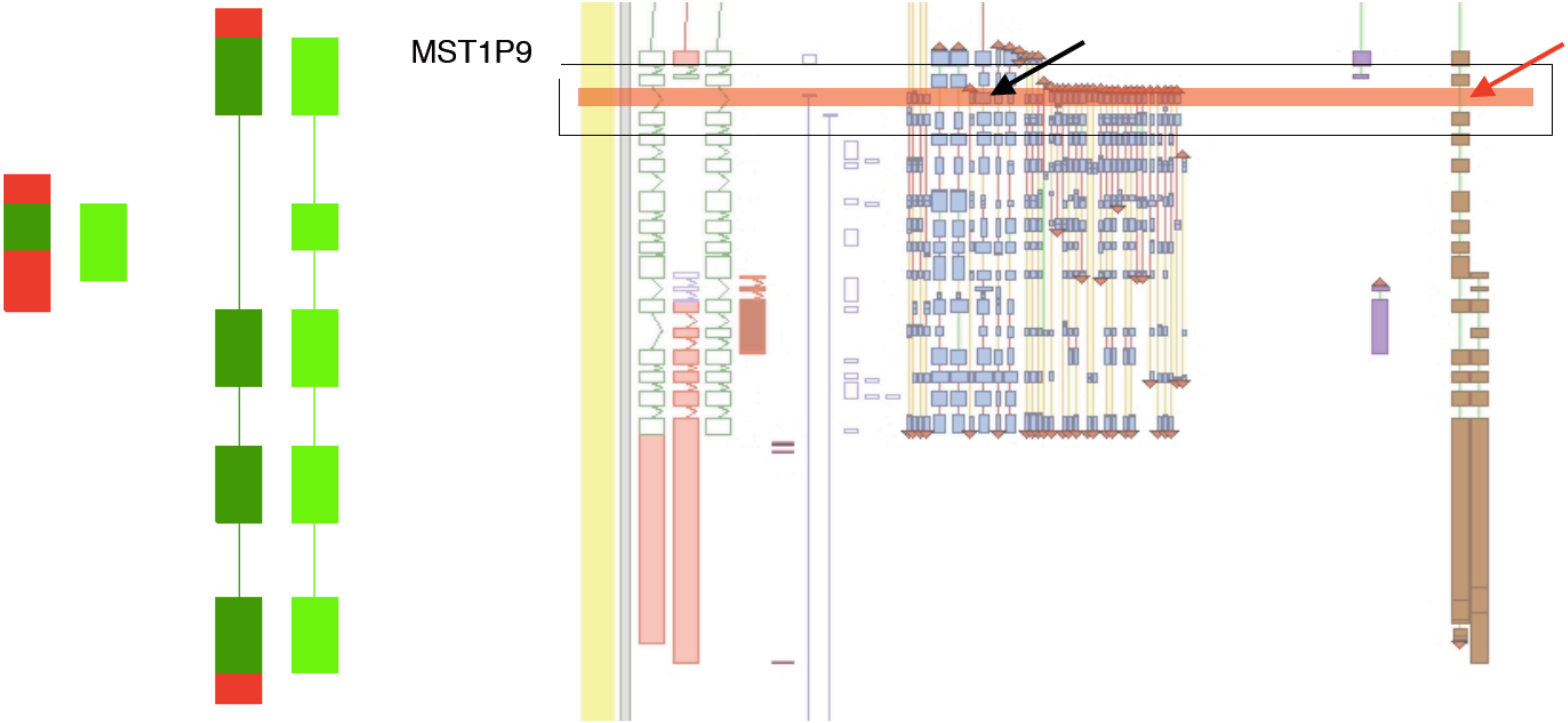
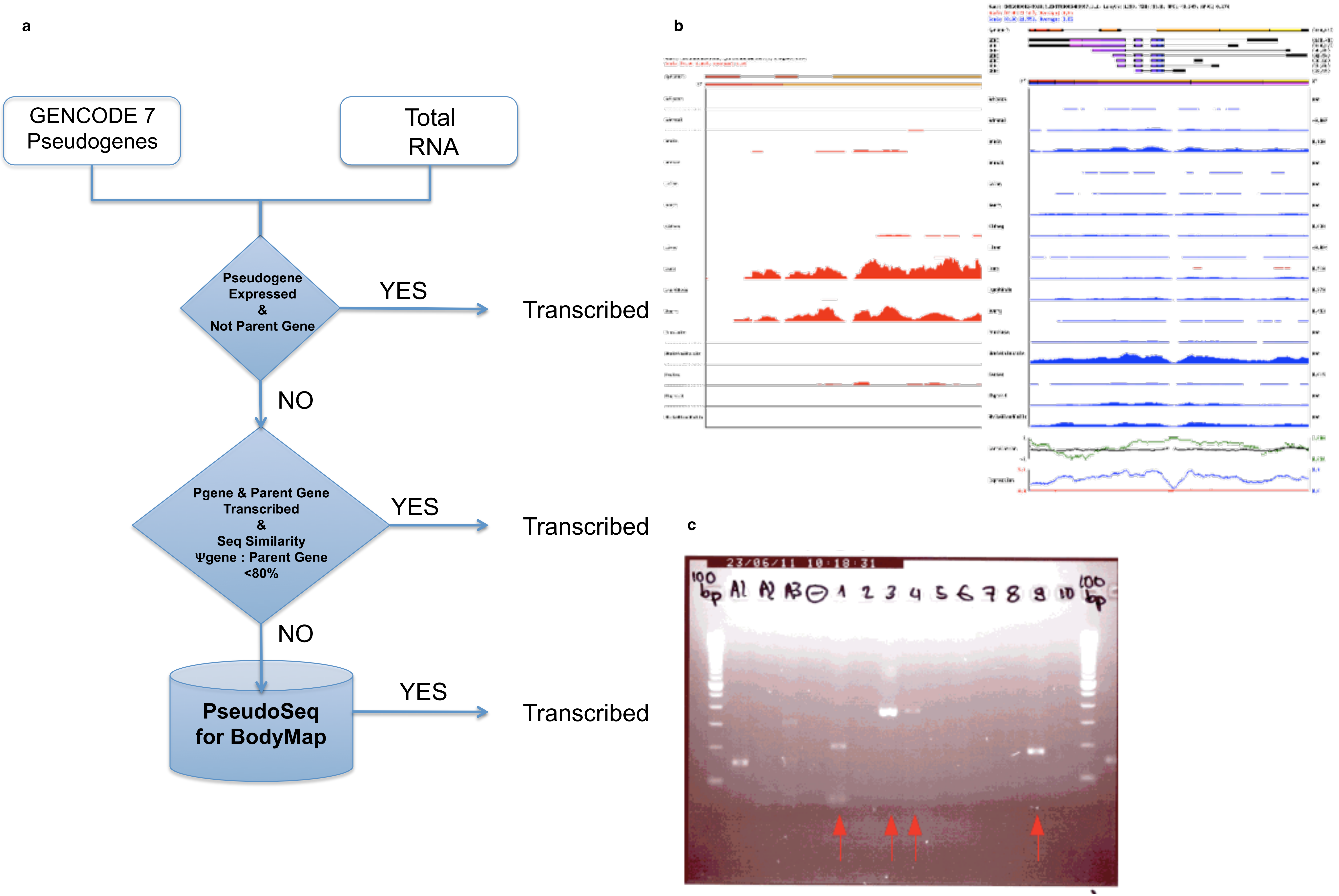


Figure 4. Coding locus resurrected from pseudogene



GENCODE 7 Pseudogenes

Total RNA

Pseudogene Expressed & Not Parent Gene

YES

Transcribed

NO

Pgene & Parent Gene Transcribed & Seq Similarity Ψgene : Parent Gene <80%

YES

Transcribed

NO

PseudoSeq for BodyMap

YES

Transcribed

Figure 5. Transcription of pseudogene. a. Flowchart to pick transcribed pseudogene from total RNA RNA-seq data and BodyMao RNAseq data. b. User Interface of PseudoSeq for transcription from a pseudogene and its parent side by side. c. RT-PCR validation of transcribed pseudogenes.

Average chromatin signatures

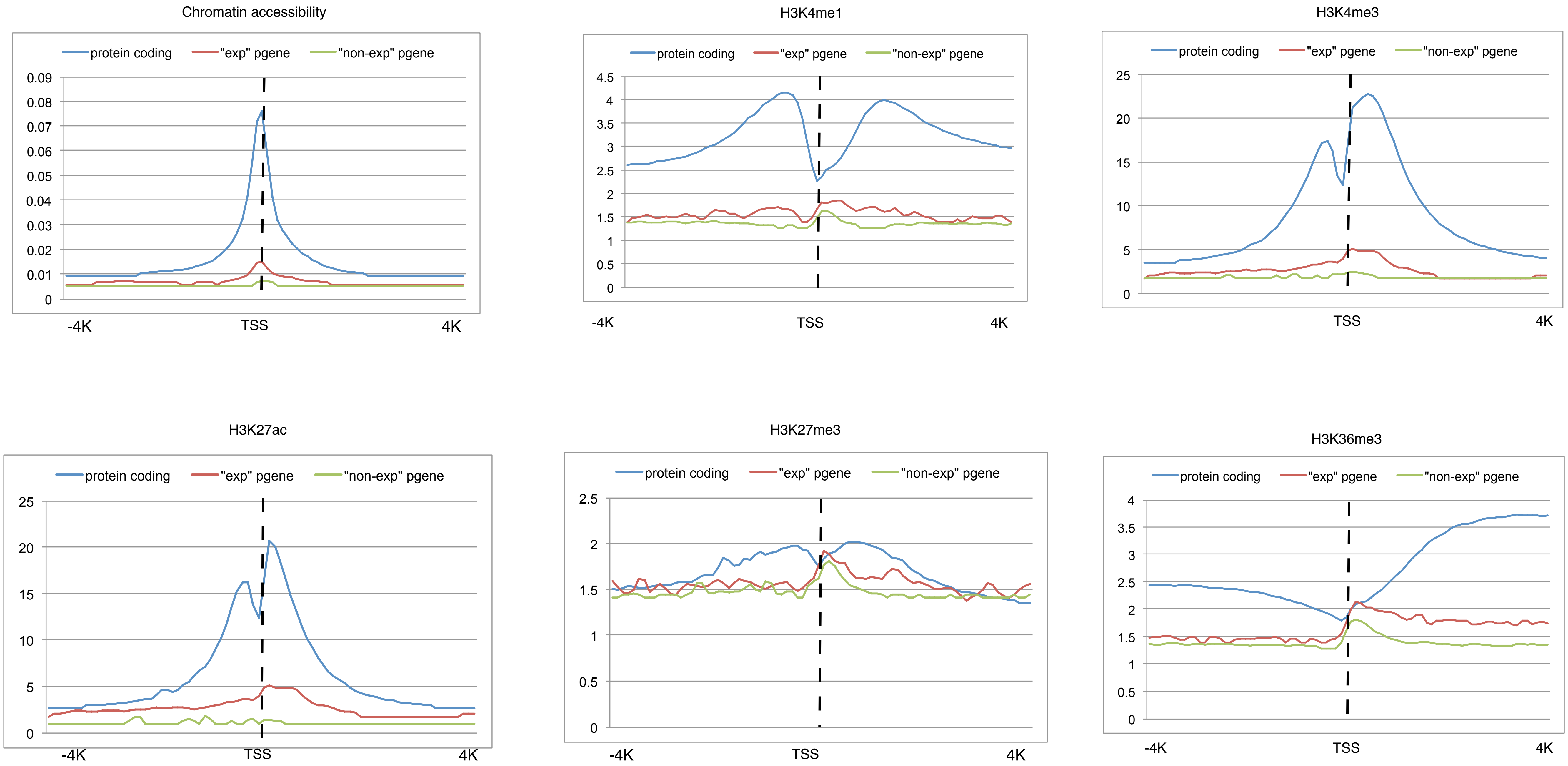
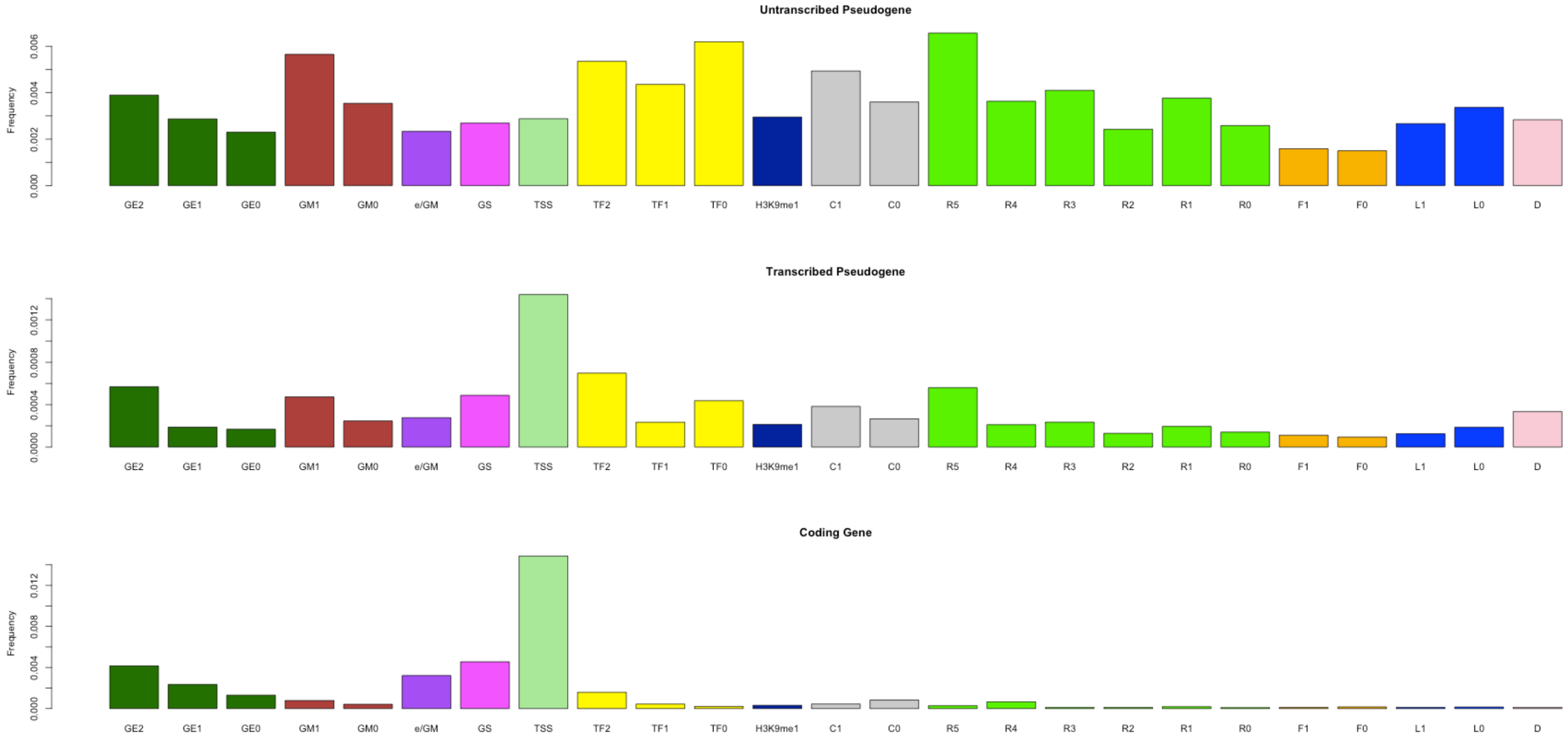
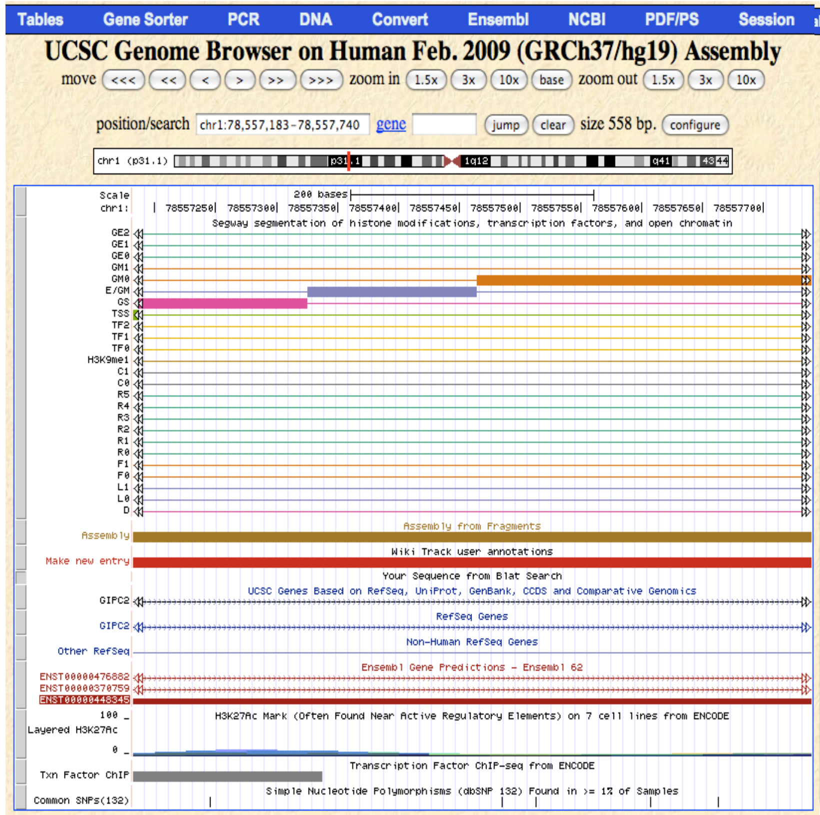


Figure 6. DNaseI hypersensitivity and histone marks of pseudogenes. Data from Chip-Seq results of GM12878.

Segmentation





UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

position/search chr1:78,557,183-78,557,740 size 558 bp.

Segway segmentation of histone modifications, transcription factors, and open chromatin

Assembly from Fragments

Wiki Track user annotations

Your Sequence from Blat Search

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

RefSeq Genes

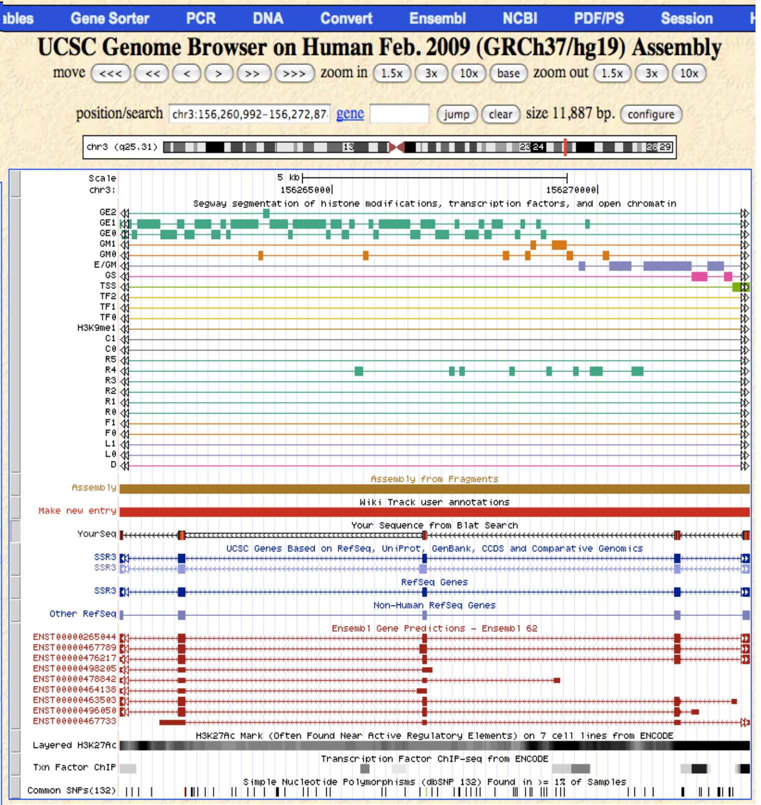
Non-Human RefSeq Genes

Ensembl Gene Predictions - Ensembl 62

H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

Transcription Factor ChIP-seq from ENCODE

Simple Nucleotide Polymorphisms (dbSNP 132) Found in >= 1% of Samples



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

position/search chr3:156,260,992-156,272,87 size 11,887 bp.

Segway segmentation of histone modifications, transcription factors, and open chromatin

Assembly from Fragments

Wiki Track user annotations

Your Sequence from Blat Search

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

RefSeq Genes

Non-Human RefSeq Genes

Ensembl Gene Predictions - Ensembl 62

H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

Transcription Factor ChIP-seq from ENCODE

Simple Nucleotide Polymorphisms (dbSNP 132) Found in >= 1% of Samples

Figure 7. Segway segmentation of pseudogene.

TFBS of Pseudogenes

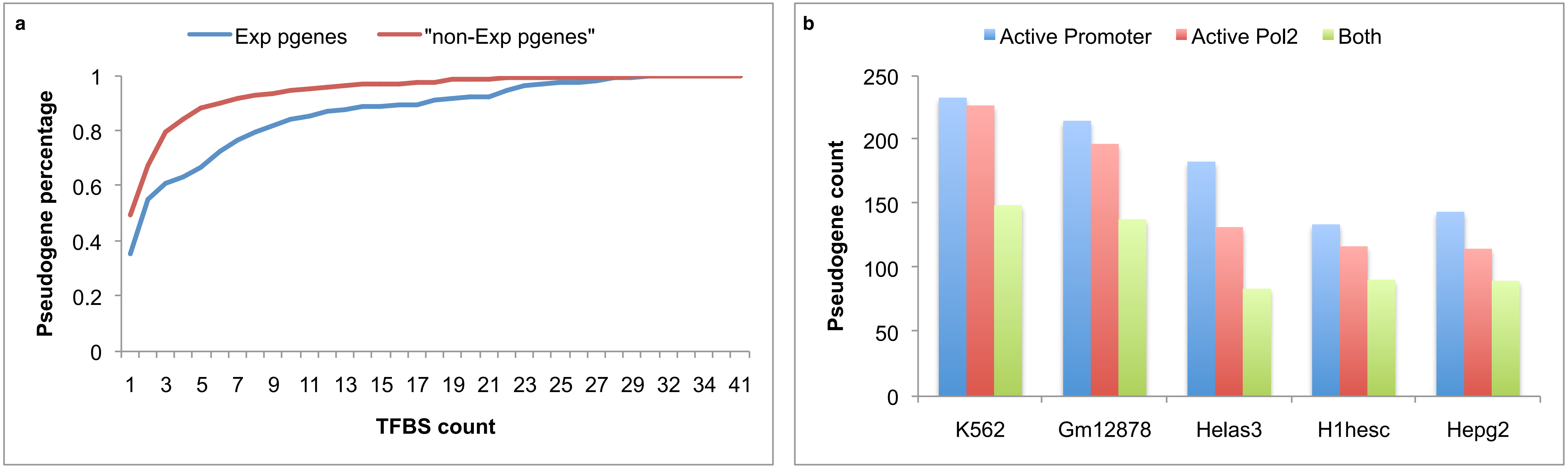


Figure 8. TFBS in the upstream of pseudogenes. a. Distribution of pseudogenes with different number of TFBS in their upstream sequences. Data is for K562. b. Number of pseudogenes with active promoter, active pol2 binding sites or both in different cell lines.

Sequence Preservation of GENCODE Pseudogenes

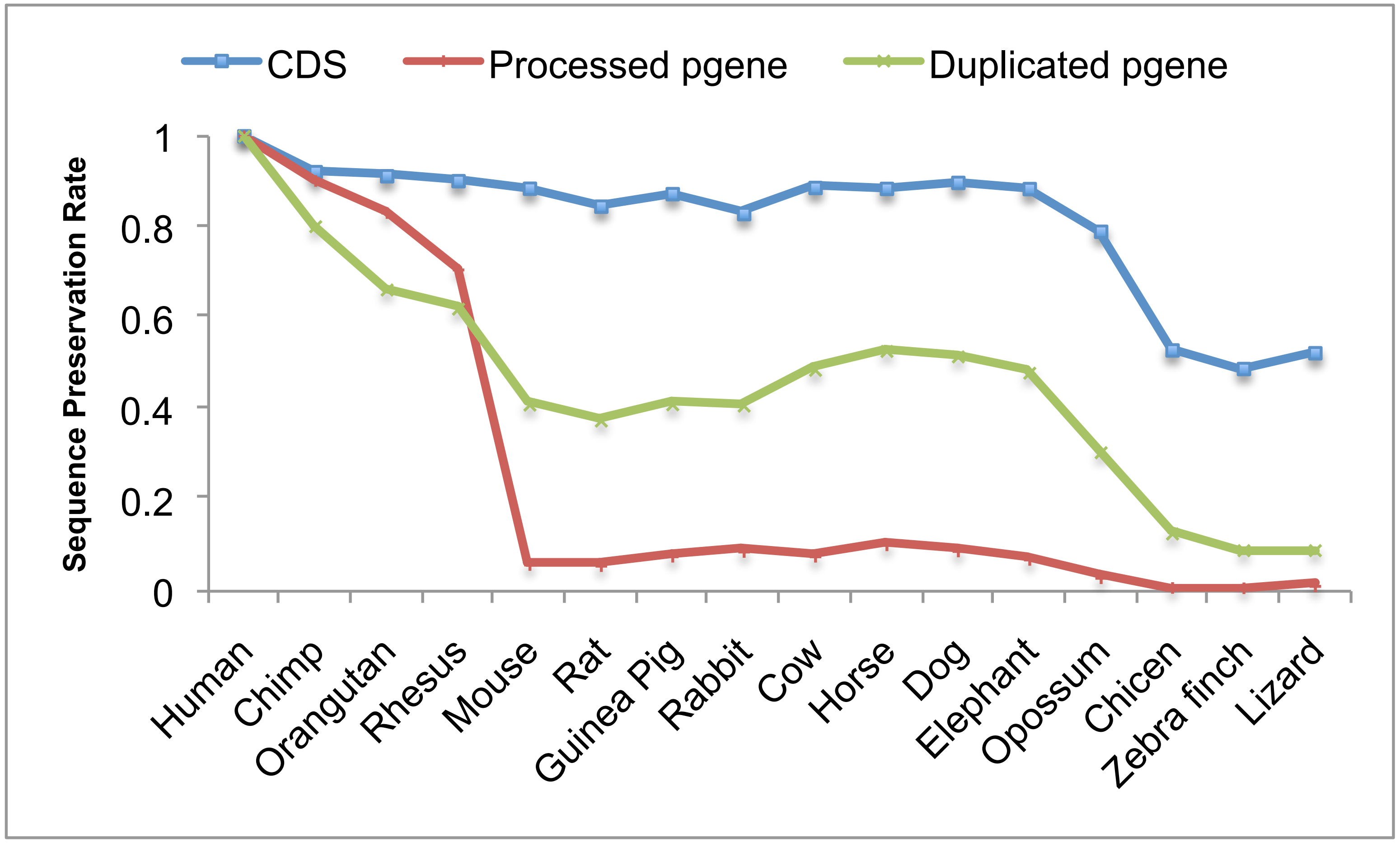


Figure 9. Preservation of human coding sequences, processed pseudogenes and duplicated pseudogenes. Orthologous to human genomic regions from different species were studied. The sequence preservation rate was calculated as percentage of sequences aligned to human sequence from each species. The calculation was based on MultiZ sequence alignment.

Pseudogene Conservation

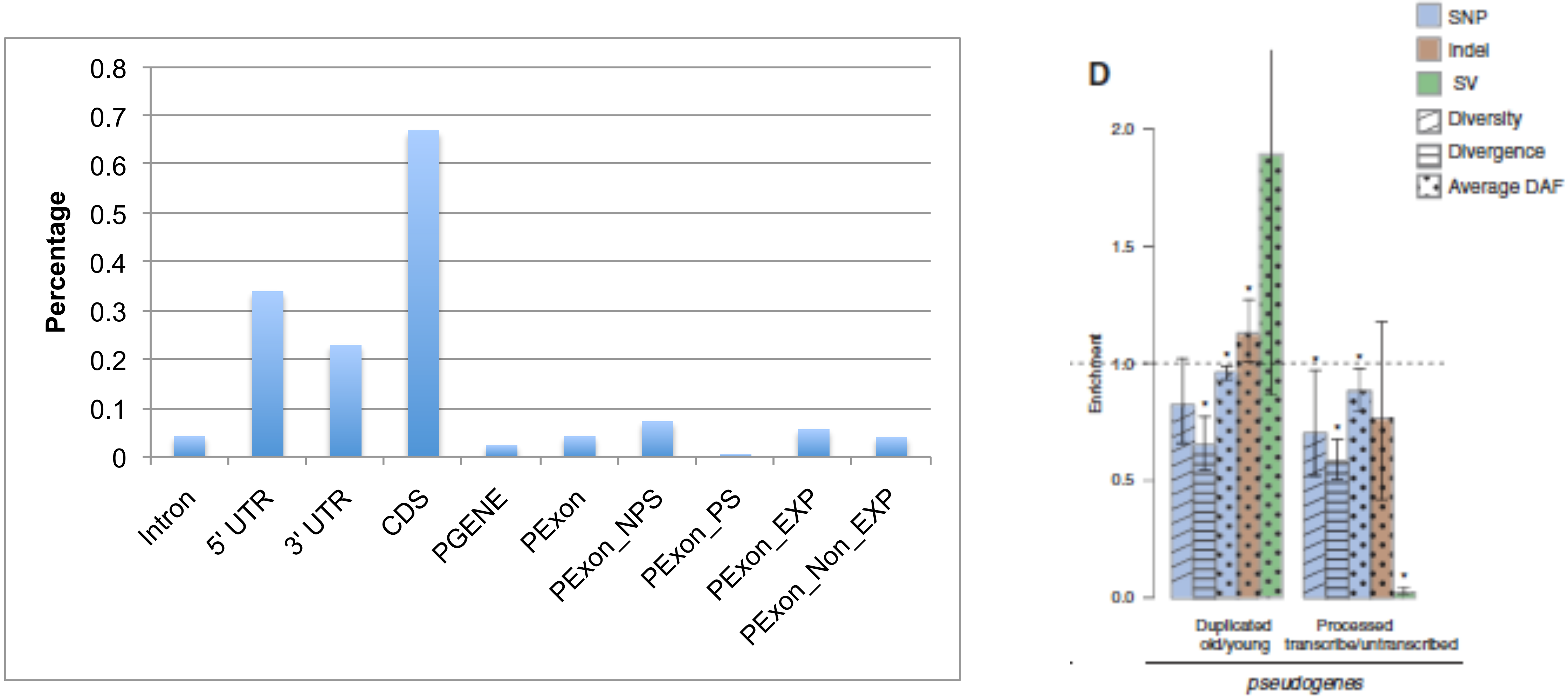
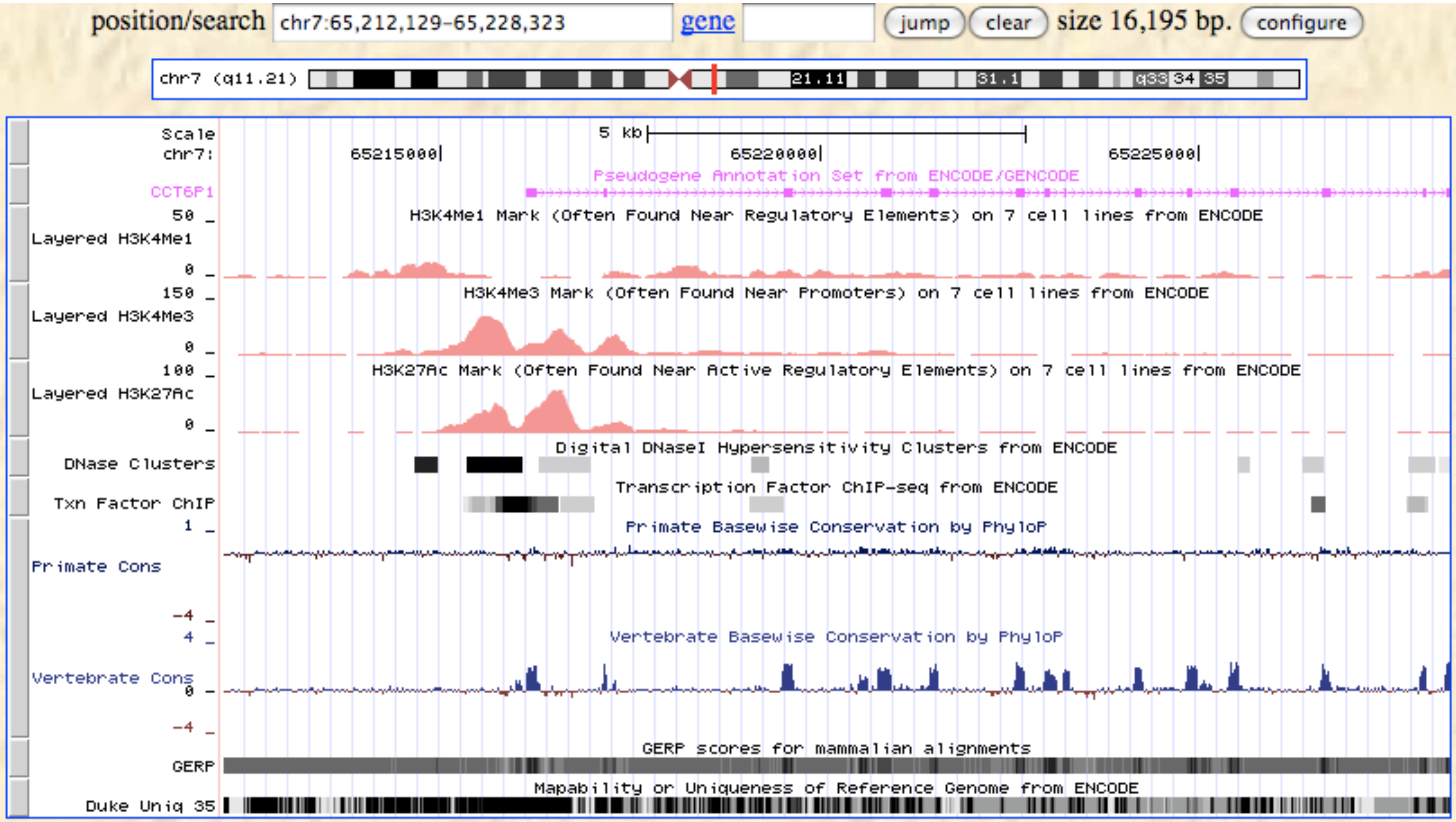


Figure 10. Overlap between GERP constraint elements and human genomic regions. Pseudgenes were divided to different classes, as nonprocessed pseudogene, processed pseudogene, transcribed pseudogenes and non-transcribed pseudogenes.

Case 1: transcribed pseudogene with active chromatin states



position/search chr7:65,212,129-65,228,323 gene jump clear size 16,195 bp. configure

chr7 (q11.21)
21.11
31.1
q33 34 35
Scale
chr7:
5 kb
65215000
65220000
65225000
Pseudogene Annotation Set from ENCODE/GENCODE
CCT6P1
50
Layered H3K4Me1
0
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
150
Layered H3K4Me3
0
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
1
Primate Cons
-4
Primate Basewise Conservation by PhyloP
4
Vertebrate Cons
0
-4
Vertebrate Basewise Conservation by PhyloP
GERP
GERP scores for mammalian alignments
Duke Uniq 35
Mapability or Uniqueness of Reference Genome from ENCODE

HM

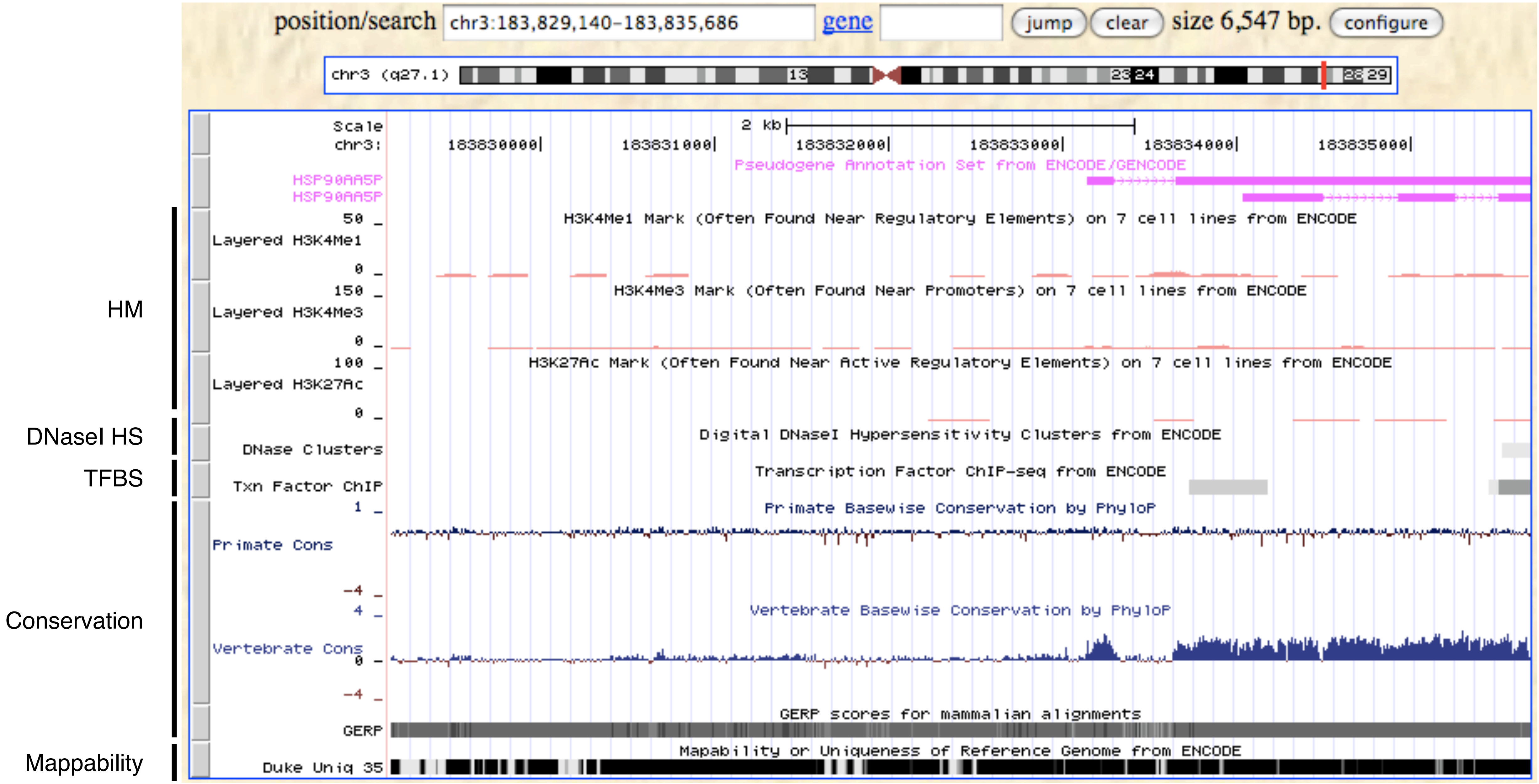
DNaseI HS

TFBS

Conservation

Mappability

Case 2: transcribed pseudogene with inactive chromatin states



position/search chr3:183,829,140-183,835,686 gene jump clear size 6,547 bp. configure

chr3 (q27.1)

Scale 2 kb

chr3: 183830000| 183831000| 183832000| 183833000| 183834000| 183835000|

Pseudogene Annotation Set from ENCODE/GENCODE

HSP90AA5P

HSP90AA5P

Layered H3K4Me1: H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE

Layered H3K4Me3: H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE

Layered H3K27Ac: H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

DNase Clusters: Digital DNaseI Hypersensitivity Clusters from ENCODE

Txn Factor ChIP: Transcription Factor ChIP-seq from ENCODE

Primate Cons: Primate Basewise Conservation by PhyloP

Vertebrate Cons: Vertebrate Basewise Conservation by PhyloP

GERP: GERP scores for mammalian alignments

Duke Uniq 35: Mapability or Uniqueness of Reference Genome from ENCODE