CAGE support for genes extended by RACE454 (1st set)

Sarah Djebali, CRG, Barcelona

11/27/2012 - Gencode experimental call

CAGE data

- Jan11 freeze ENCODE CAGE (12 cell lines, whole cell, polya+):
  - clusters made in each cell line and further filtered by:
    - idr <= 0.01,
    - tss prediction strength >= 0.5 (Timo's HMM classifier).
  - 12,000-28,000 clusters depending on cell line.
  - Strandedly merged => 51,037 cage clusters (avg lg 125nt).
- Fantom5 CAGE (~1,000 samples) (helicosCAGE):
  - Input is ~ 1 million permissive DPI* clusters (Kawaji),
  - classified as TSS based in Timo's improved classifier,
  - tss prediction strength >= 0.228 -> 217,572 cage clusters.

* DPI = Decomposition-based Peak Identification

Gencode annotation

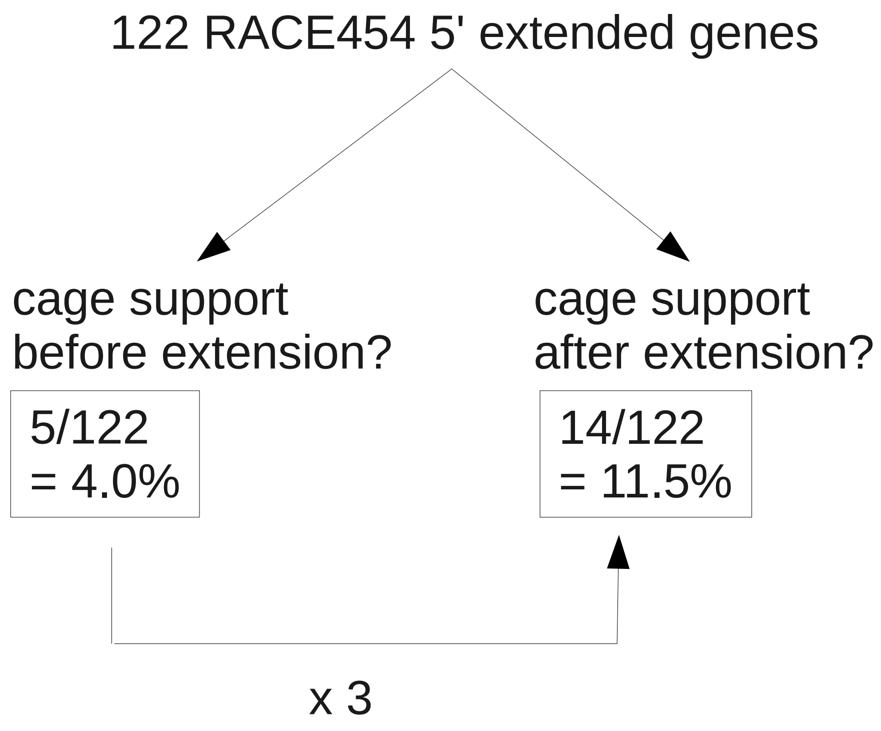
- 400 Gencode v7 long non-coding transcripts selected for RACE454 based on lack of 5' (cage) and 3' (ditag) support (cage: jan11 encode, ditag: 16 experiments):
  - RACE was performed in two tissues (brain and testis),
  - RACE products were 454 sequenced,
  - 454 reads were aligned (Blat) and alignments with highest % similarity were sent to Havana for manual curation.
- Gencode v15 file of the subset of 371 genes (2248 transcripts) manually curated based on 454 read alignments (sent By Electra):
  - 122 were 5' extended by RACE454.
  - I will focus on them when looking for cage support.

Criteria to consider a Gencode TSS as supported by CAGE

- Given a CAGE cluster set and a Gencode annotation file:
  - Get Gencode TSS for each gene (distinct most 5' bp of each tr),
  - Extend by 50bp on each side,
  - Strandedly merge → Gencode TSS clusters,
  - Strandedly merge Gencode TSS clusters with CAGE clusters →Gencode/CAGE TSS clusters,
  - Divide them into:
    - GencOnly,
    - CageOnly,
    - GencCage → the ones I am interested in,
  - On each Gencode/CAGE TSS cluster, report information about initial Gencode TSS cluster(s) and associated gene(s).

Results for the 122 RACE454 5' extended genes

ENCODE cage (Jan11, 14 cell lines, idr 0.01)



122 RACE454 5' extended genes

cage support
before extension?

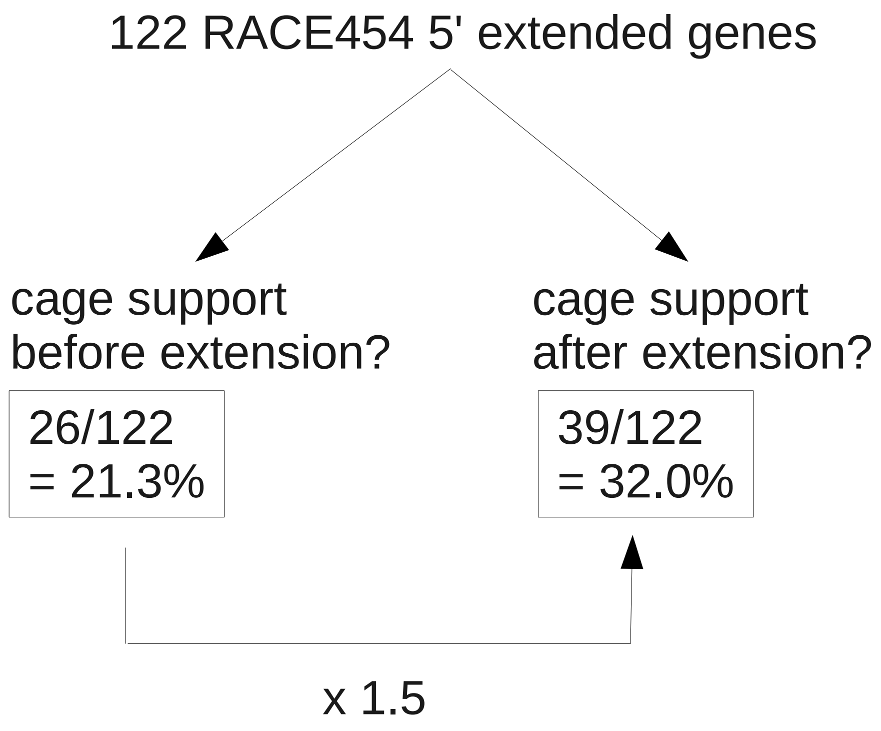
5/122
= 4.0%

cage support
after extension?

14/122
= 11.5%

x 3

Fantom5 cage (sep12, ~1,000 samples)



122 RACE454 5' extended genes

cage support
before extension?

26/122
= 21.3%

cage support
after extension?

39/122
= 32.0%

x 1.5

At the transcript level, and for the 55 5' extended transcripts, results are not as good:
- from 0 to 1 case for encode,
- from 8 to 9 cases for fantom5.

Results for the 400 genes selected for RACE454

ENCODE cage (Jan11, 14 cell lines, idr 0.01)

400 selected genes

cage support
before extension?

18/400
= 4.5%

cage support
after extension?

38/371
= 10.2%

x 2.3

~ 64% for protein coding genes
(70% for protein coding genes
detected by RNAseq in the
same 14 samples)

Fantom5 cage (sep12, ~1,000 samples)

400 selected genes

cage support
before extension?

100/400
= 25.0%

cage support
after extension?

121/371
= 32.6%

x 1.3

~ 79% for protein coding genes