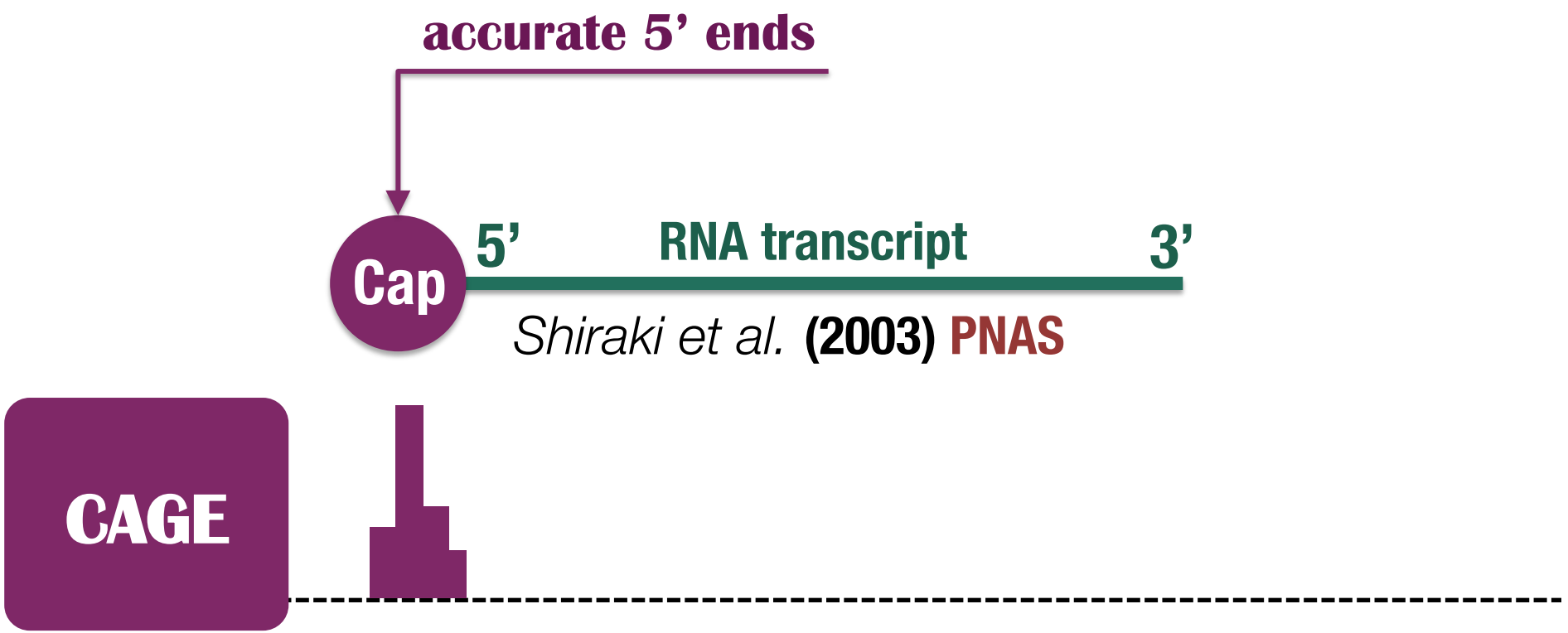
Integrating FANTOM CAGE into GENCODE

Chung-Chau, HON

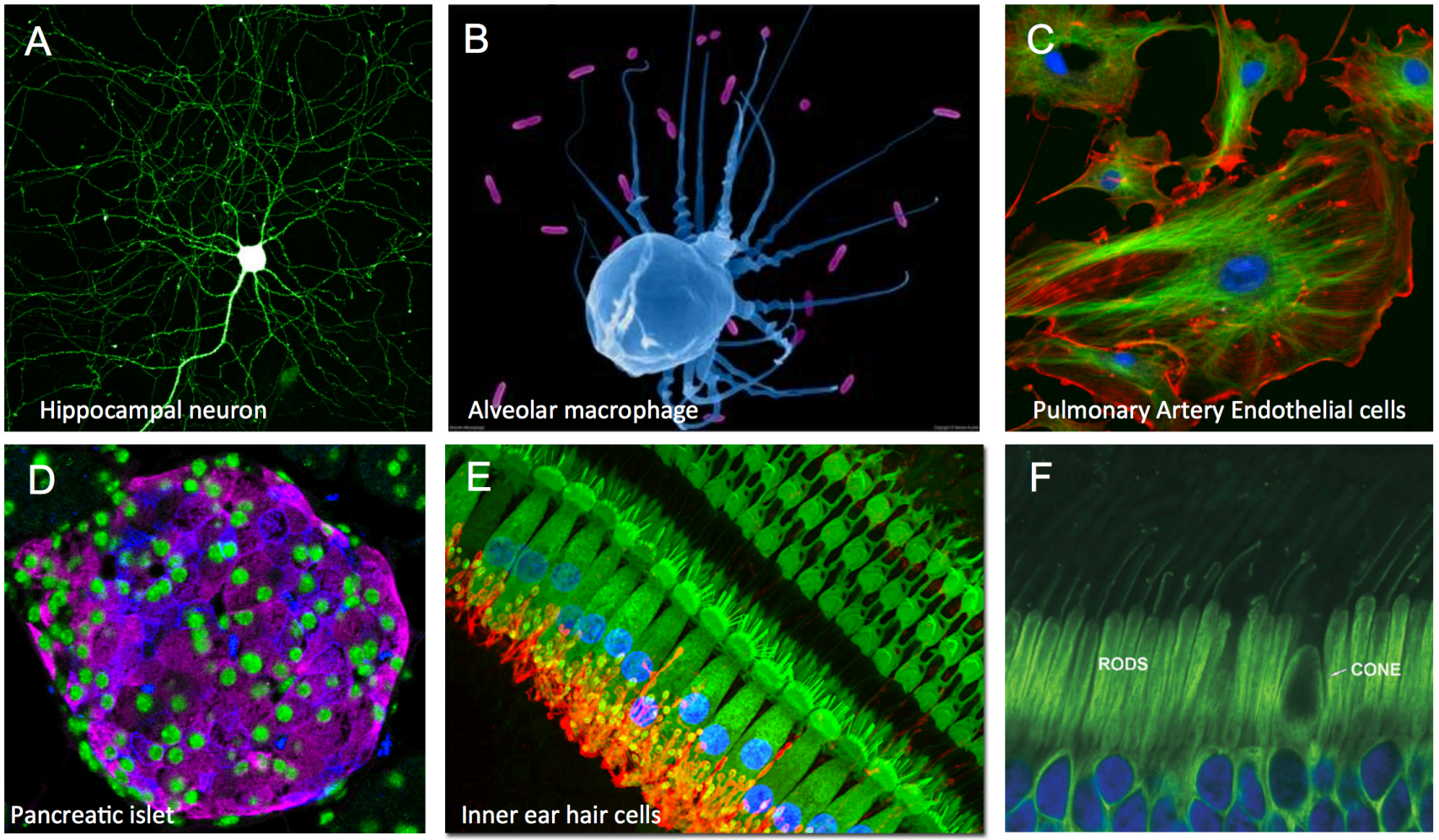
Yokohama, Division of Genome Technologies, FANTOM Consortium

FANTOM CAGE : *Cap Analysis of Gene Expression*

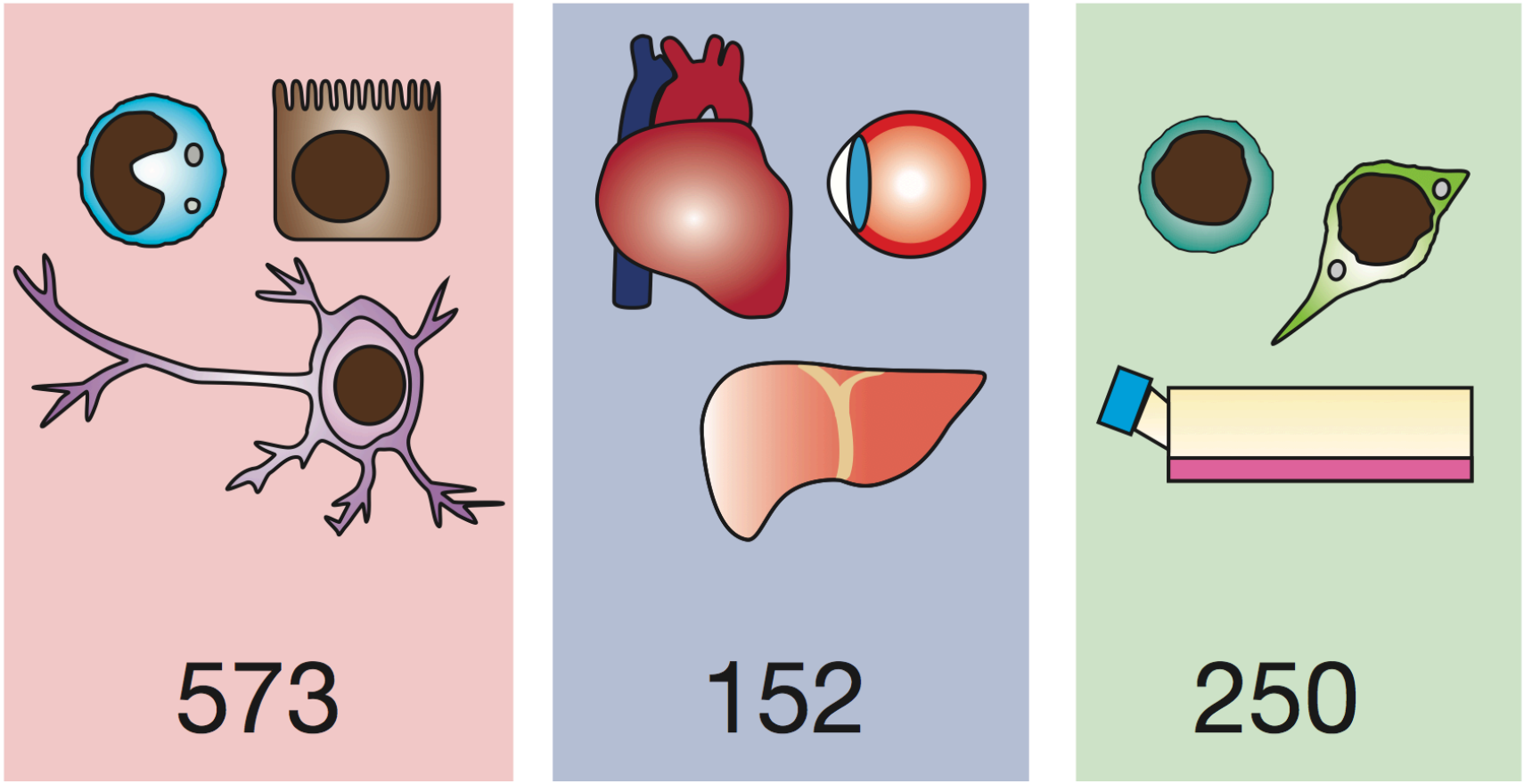


FANTOM 5

“The Complete Transcriptional Landscape of Cell Types”

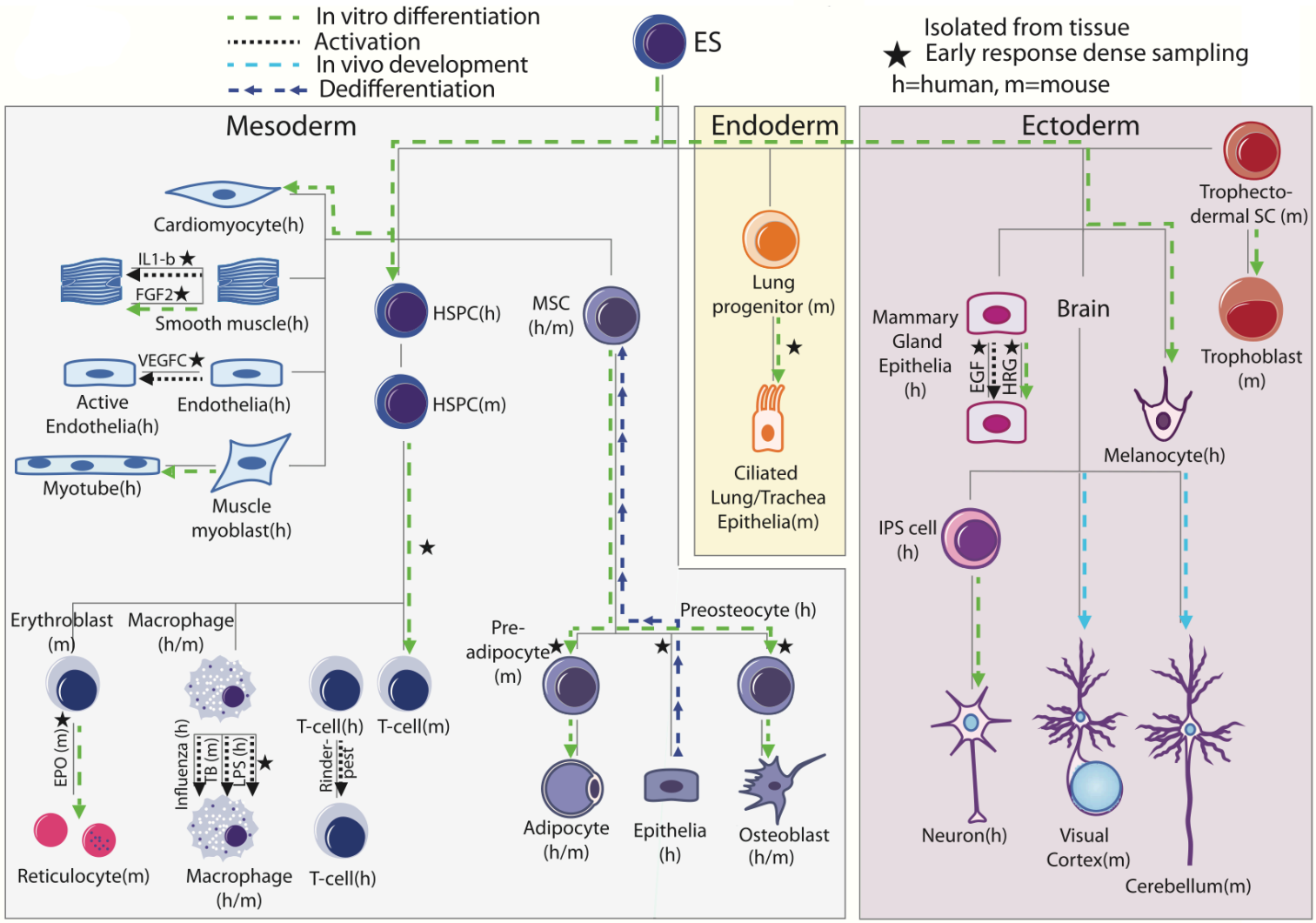


FANTOM5 Phase 1 : “Static State Snapshot”



n=975

FANTOM5 Phase 2 : “Dynamic Time Courses”



n=854

Atlas of Human Promoters & Enhancers

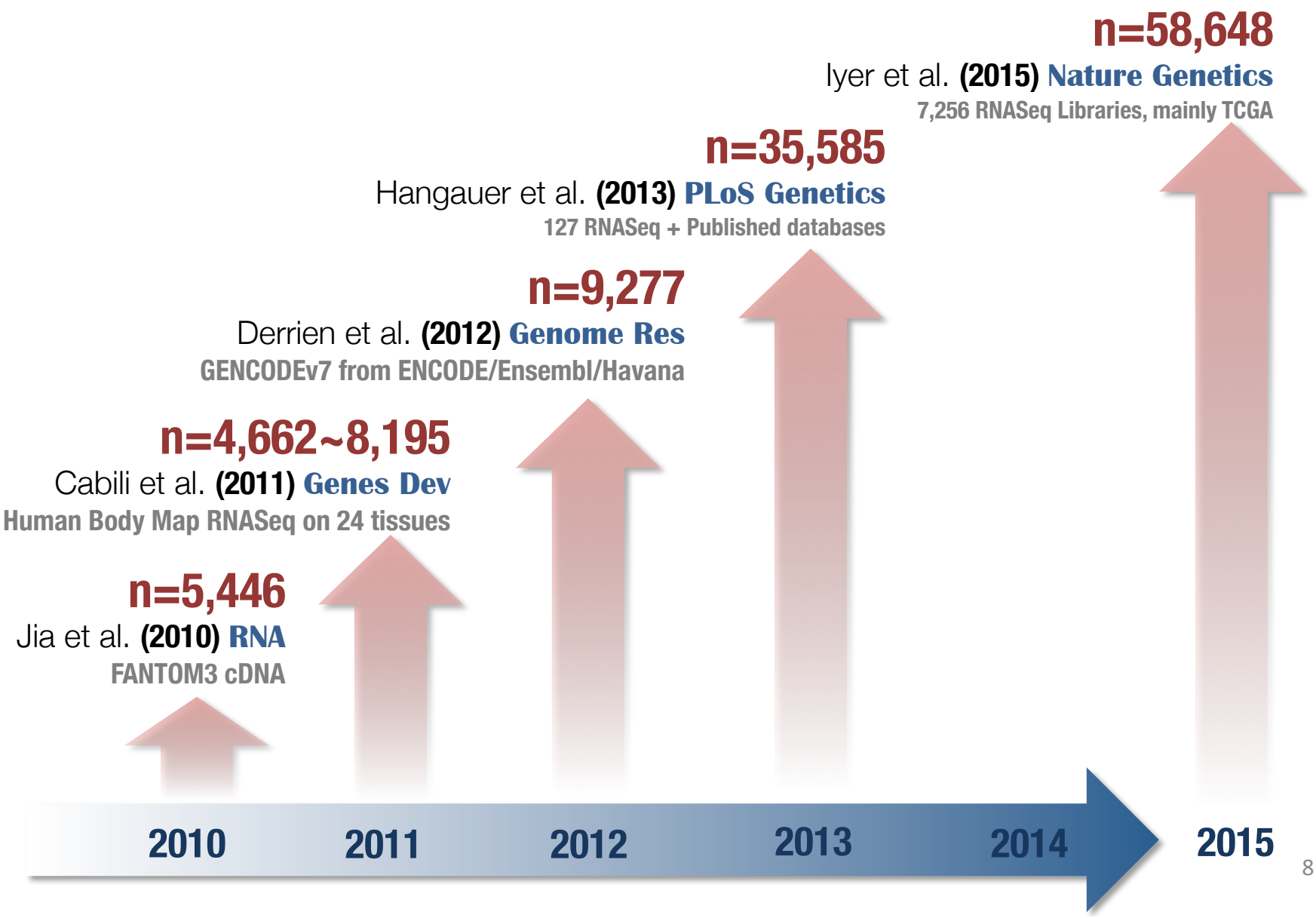
An atlas of active enhancers across human cell types and tissues. *Andersson et al.* **Nature** (2014)

A promoter-level mammalian expression atlas. *Forrest et al.* **Nature** (2014)

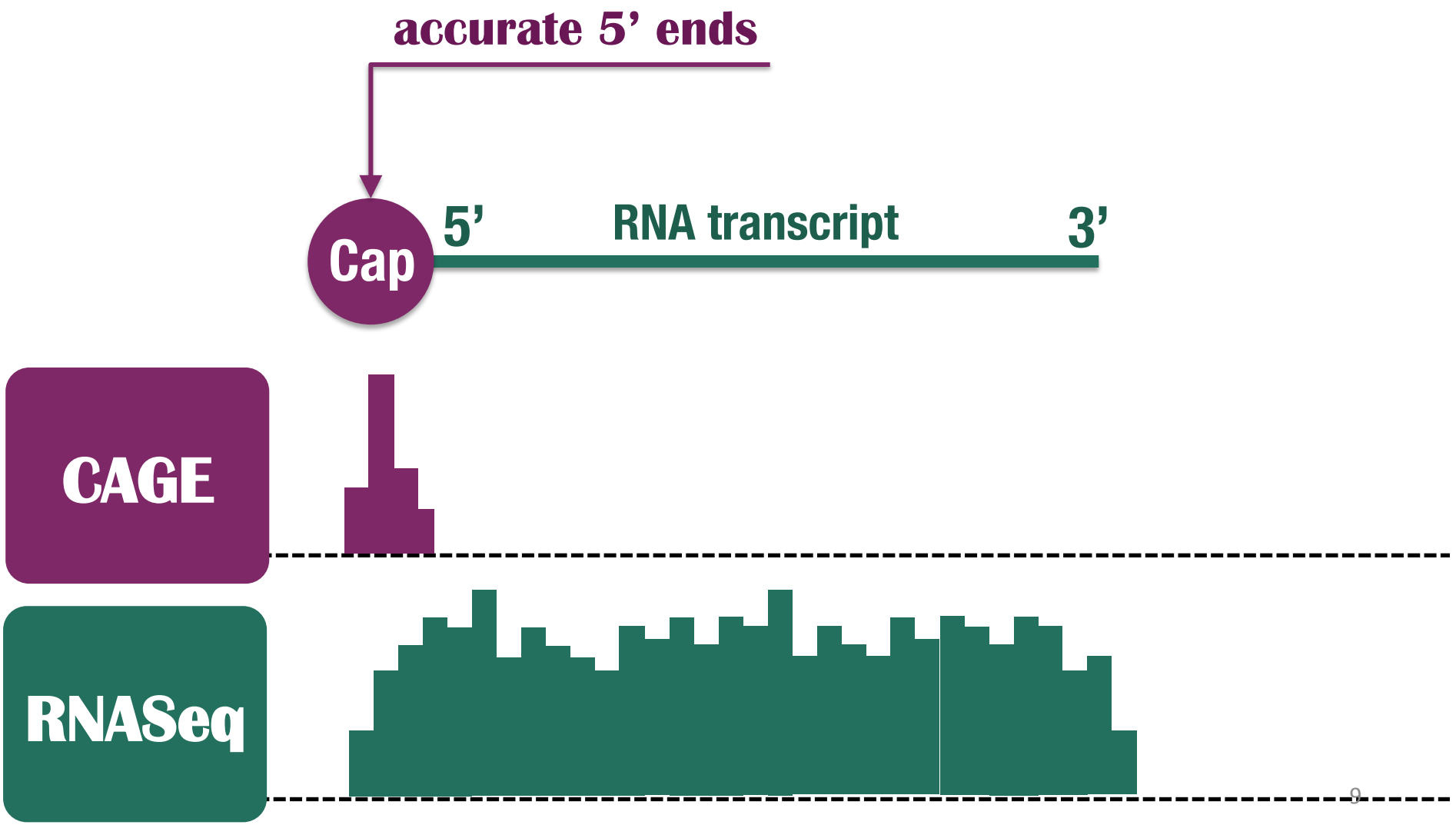
The Missing Genes

How many non-coding RNAs are there?

Unfinished Story : *Growing numbers, growing uncertainties*



Combining FANTOM CAGE with RNASeq

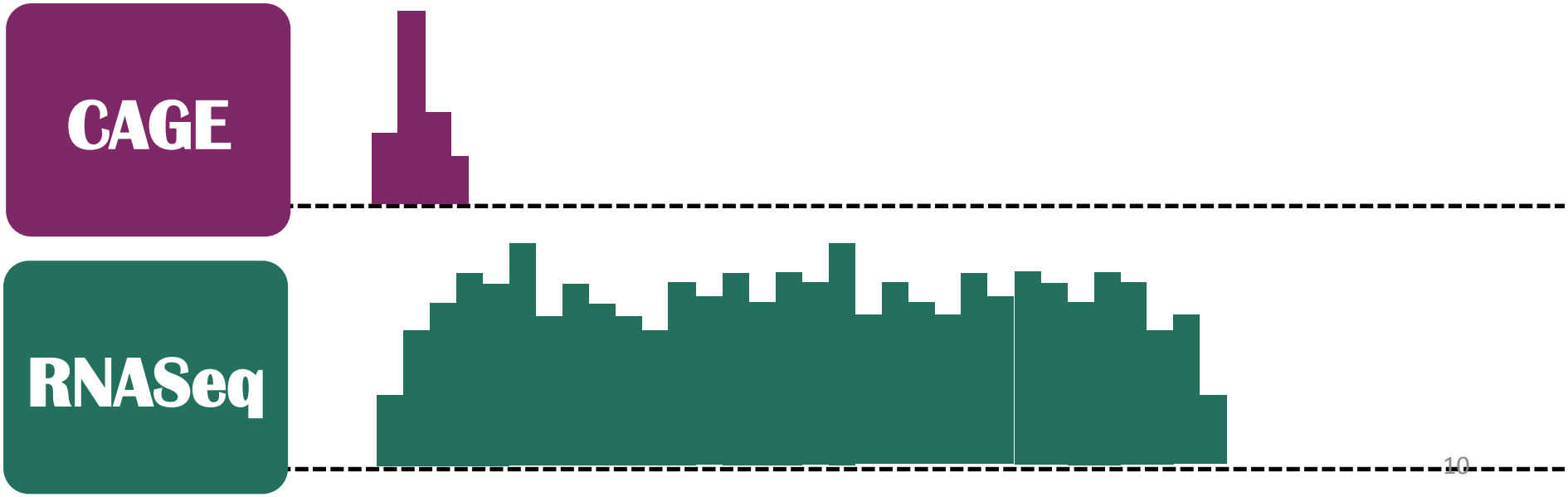


ARTICLE

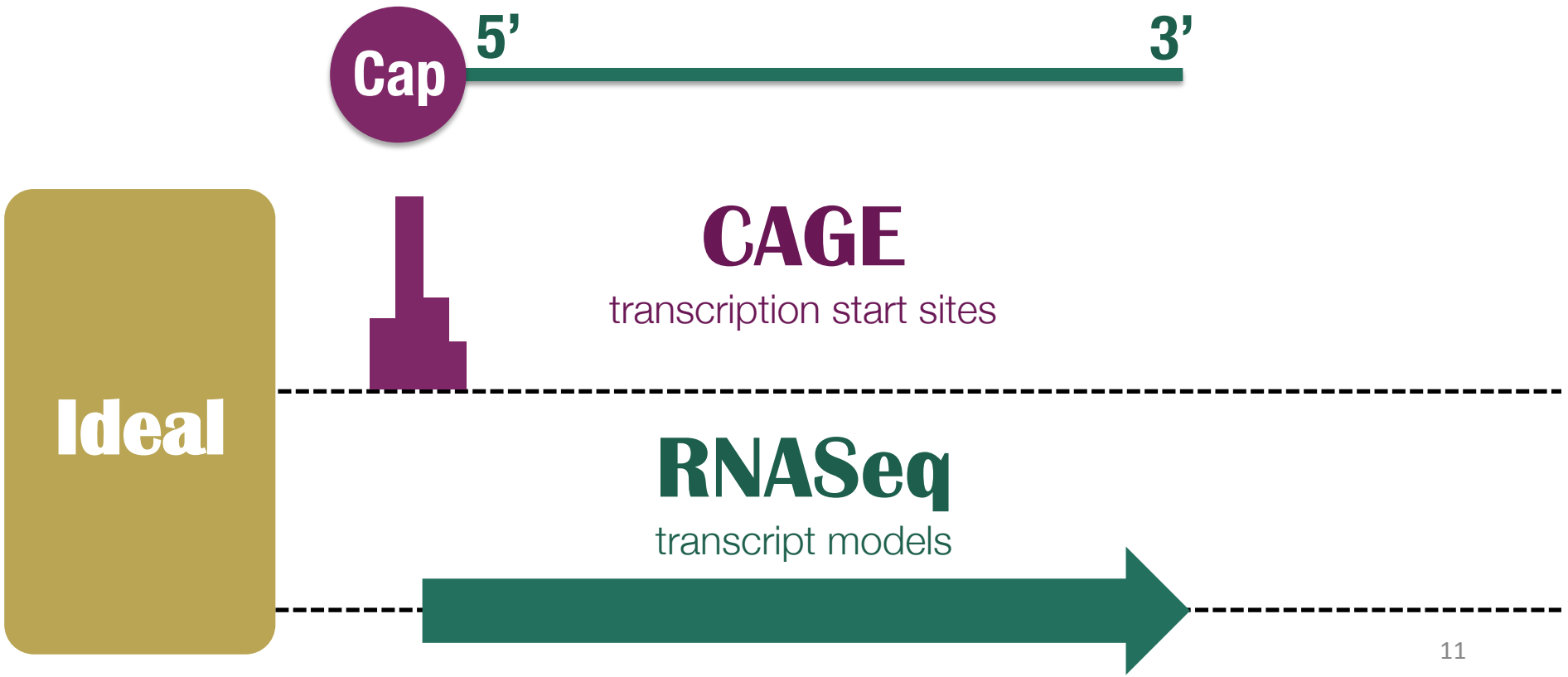
doi:10.1038/nature21374

An atlas of human long non-coding RNAs with accurate 5′ ends

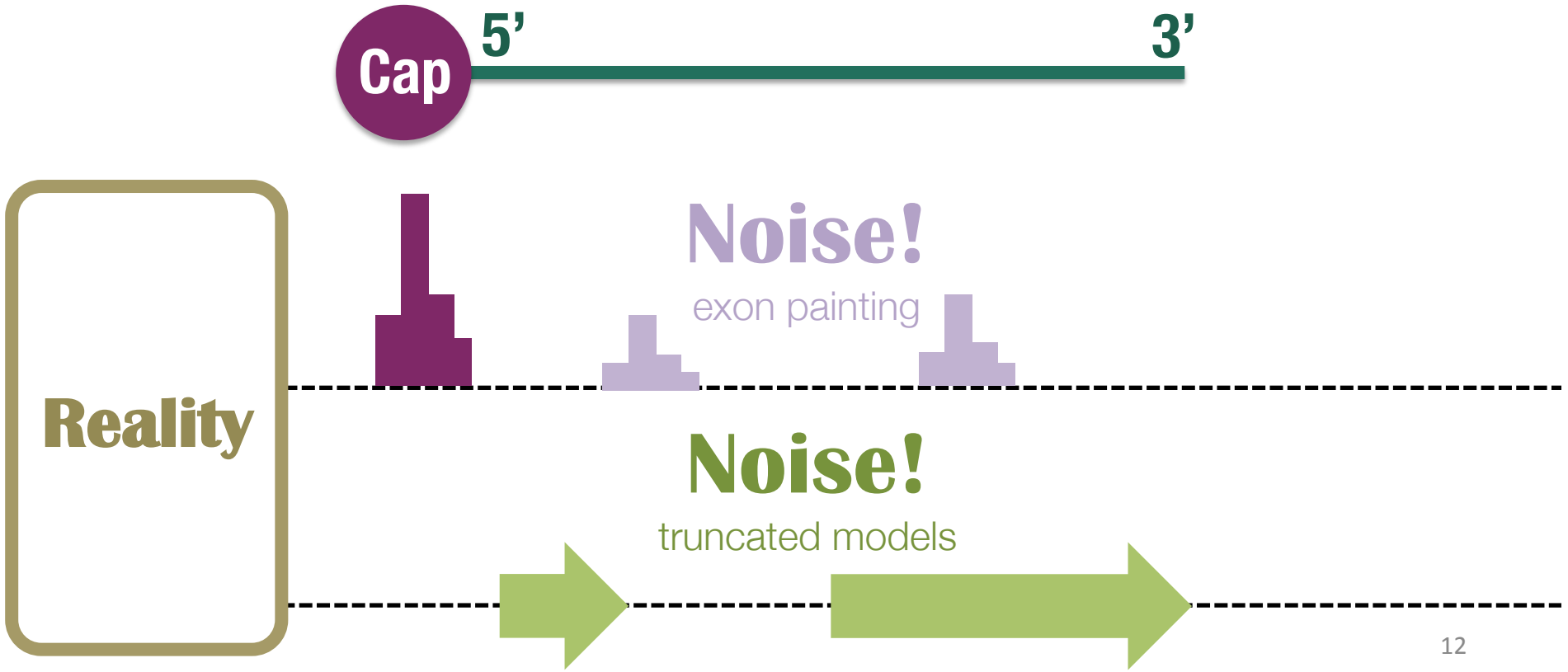
Chung-Chau Hon[1], Jordan A. Ramilowski[1,2], Jayson Harshbarger[1,2], Nicolas Bertin[2,3]†, Owen J. L. Rackham[4,5], Julian Gough[4], Elena Denisenko[6], Sebastian Schmeier[6], Thomas M. Poulsen[7], Jessica Severin[1,2], Marina Lizio[1,2], Hideya Kawaji[1,2,8], Takeya Kasukawa[1], Masayoshi Itoh[1,2,8], A. Maxwell Burroughs[1,2,9], Shohei Noma[1,2], Sarah Djebali[10,11]†, Tanvir Alam[12], Yulia A. Medvedeva[13,14], Alison C. Testa[15], Leonard Lipovich[16,17], Chi-Wai Yip[1], Imad Abugessaisa[1], Mickaël Mendez[1,2]†, Akira Hasegawa[1,2], Dave Tang[1,2,18], Timo Lassmann[1,2,18], Peter Heutink[1,19], Magda Babina[20], Christine A. Wells[21,22], Soichi Kojima[23], Yukio Nakamura[24,25], Harukazu Suzuki[1,2], Carsten O. Daub[1,2,26], Michiel J. L. de Hoon[1,2], Erik Arner[1,2], Yoshihide Hayashizaki[2,8], Piero Carninci[1,2] & Alistair R. R. Forrest[1,2,15]



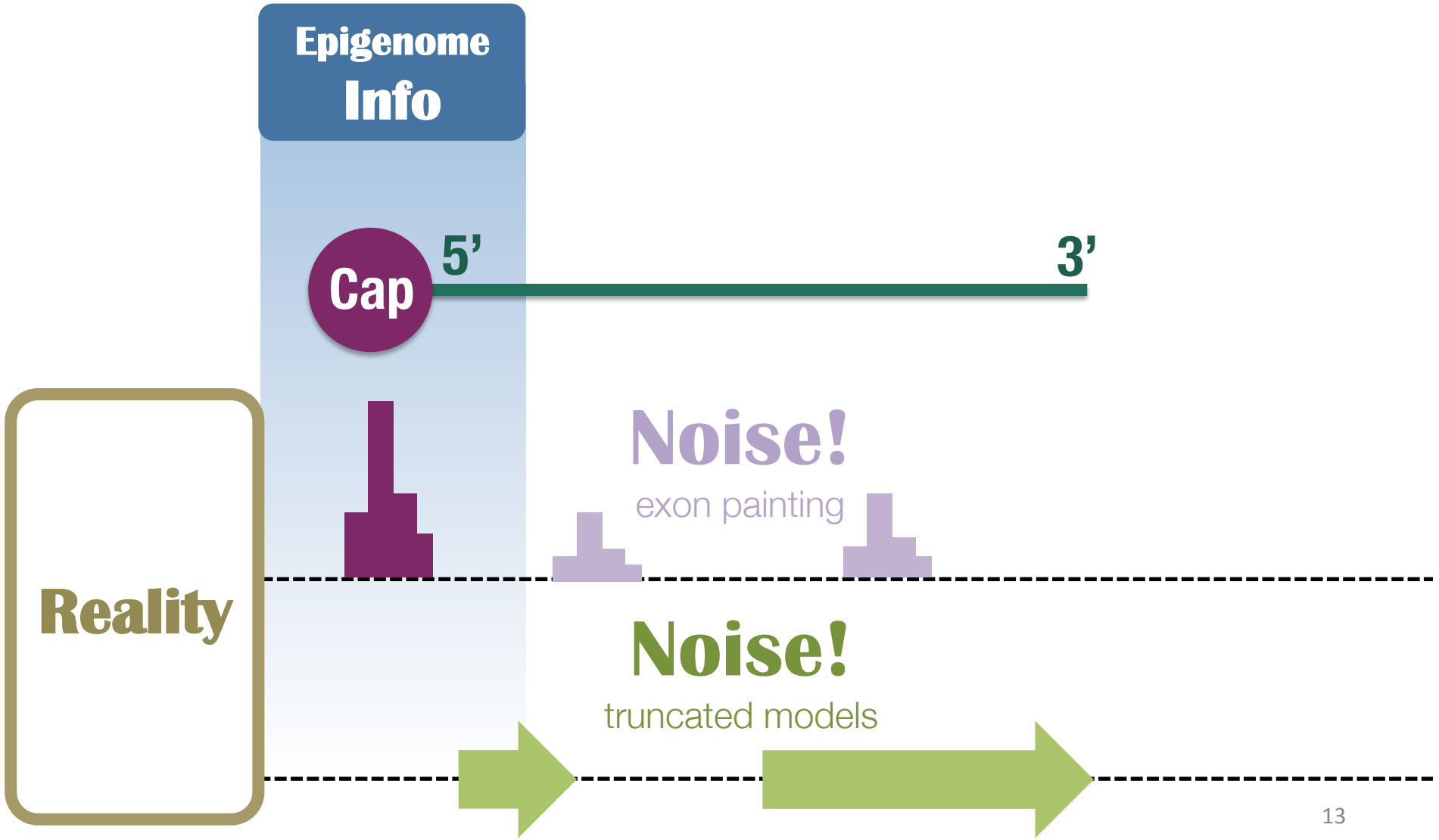
Building a transcriptome with accurate 5’ end



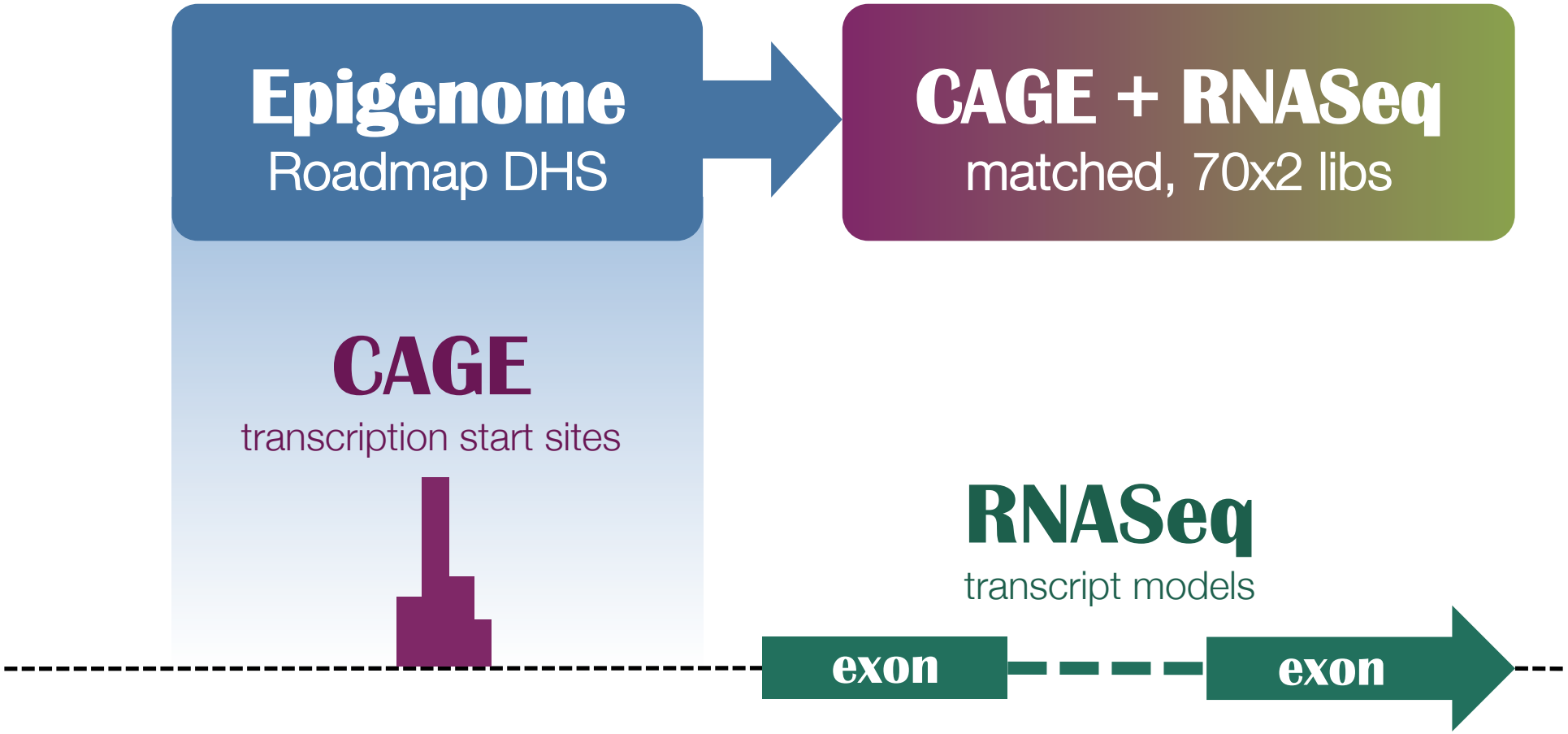
Building a transcriptome with accurate 5’ end



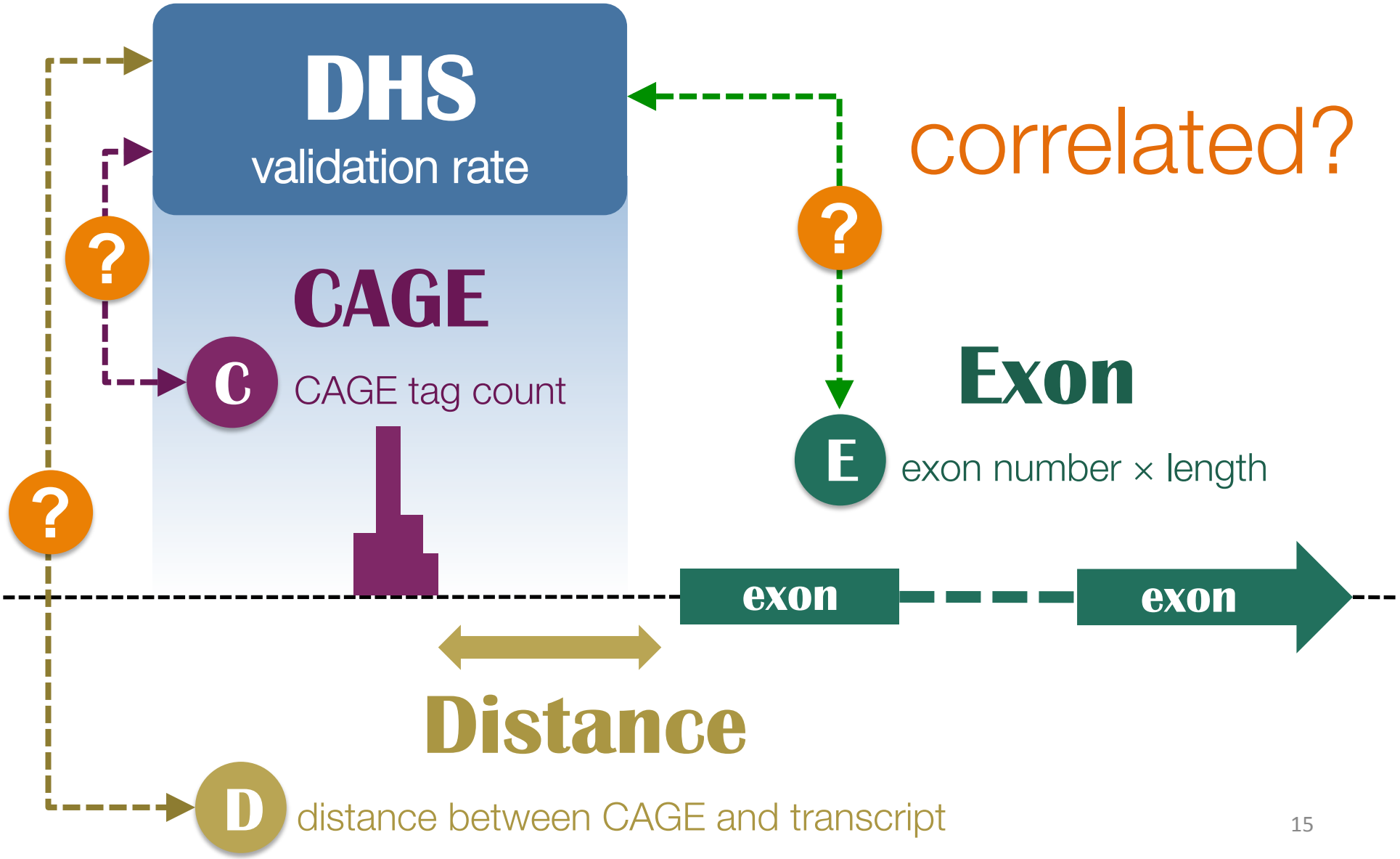
Building a transcriptome with accurate 5’ end



TIEScore : *Transcription Initiation Evidence Score*



TIEScore : *Transcription Initiation Evidence Score*

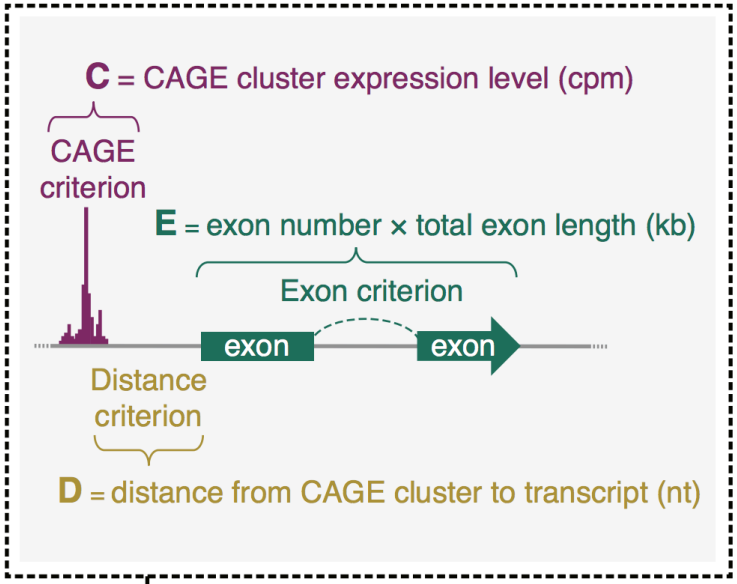


TIEScore : *Transcription Initiation Evidence Score*

1

Input

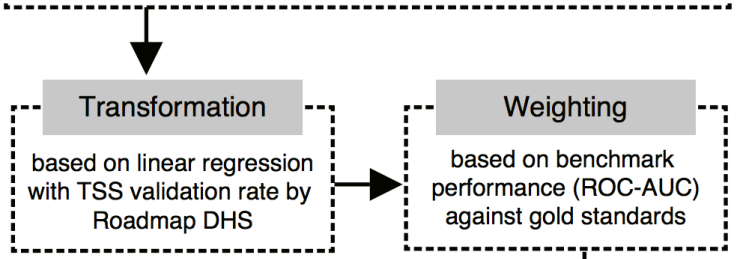
CAGE + Transcript Properties



2

Process

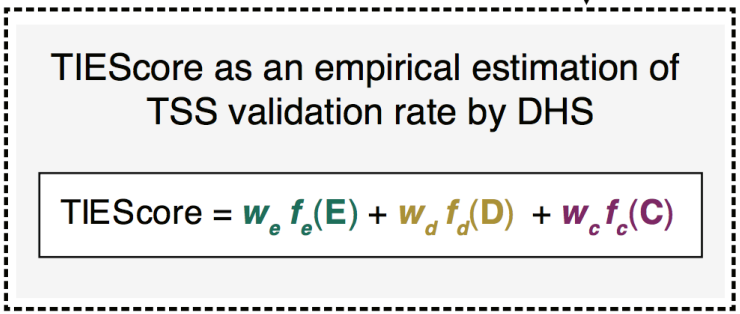
Transform, Weight and Sum

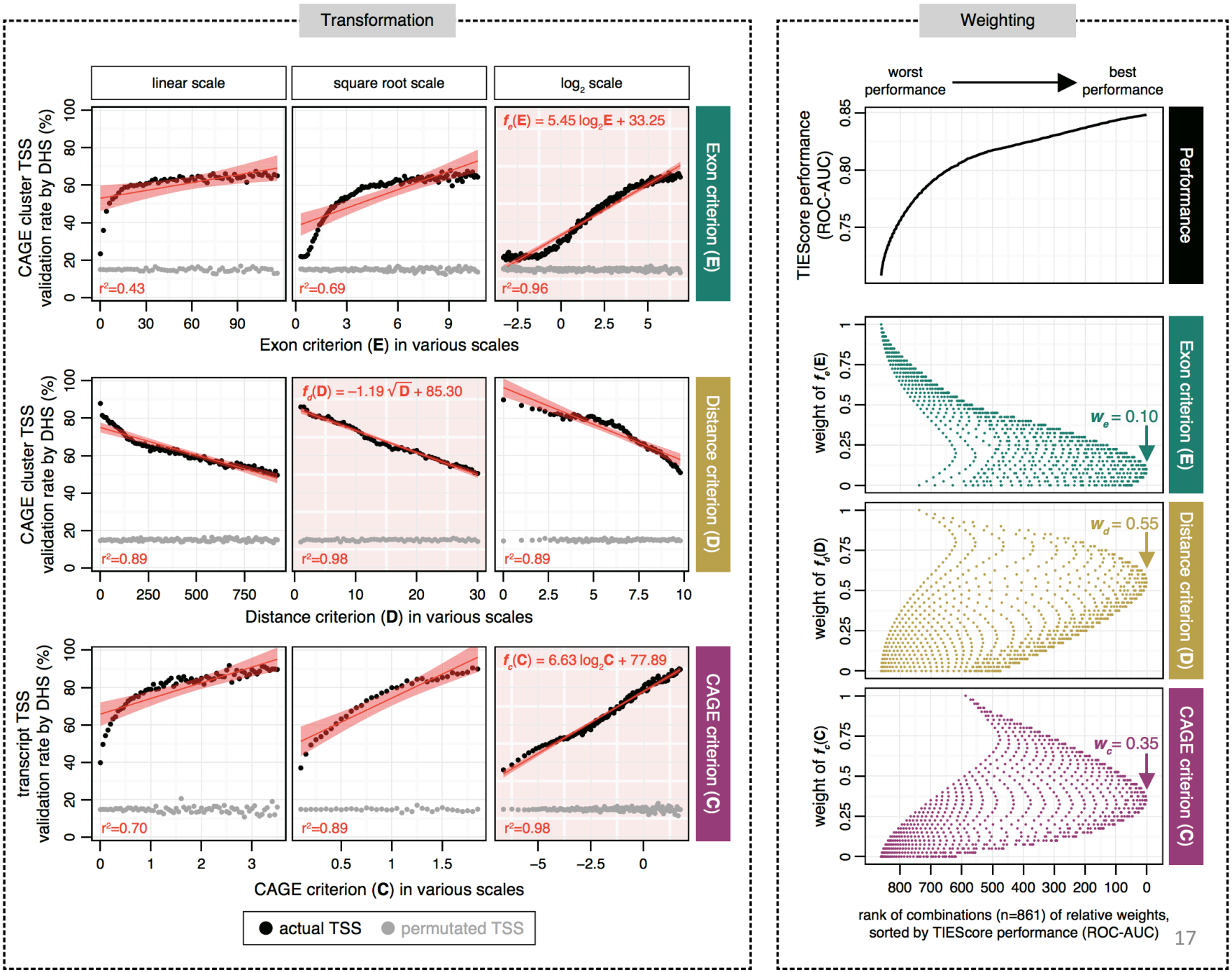


3

Output

TIEScore




Transformation
linear scale
square root scale
log2 scale
CAGE cluster TSS validation rate by DHS (%)
fe(E) = 5.45 log2E + 33.25
r²=0.43
r²=0.69
r²=0.96
Exon criterion (E)
Exon criterion (E) in various scales
fd(D) = −1.19 √D + 85.30
r²=0.89
r²=0.98
r²=0.89
Distance criterion (D)
Distance criterion (D) in various scales
transcript TSS validation rate by DHS (%)
fc(C) = 6.63 log2C + 77.89
r²=0.70
r²=0.89
r²=0.98
CAGE criterion (C)
CAGE criterion (C) in various scales
actual TSS
permutated TSS
Weighting
worst performance
best performance
TIEScore performance (ROC-AUC)
Performance
weight of fe(E)
we = 0.10
Exon criterion (E)
weight of fd(D)
wd = 0.55
Distance criterion (D)
weight of fc(C)
wc = 0.35
CAGE criterion (C)
rank of combinations (n=861) of relative weights, sorted by TIEScore performance (ROC-AUC)

TIEScore : *Transcription Initiation Evidence Score*

TIEScore as an empirical estimation of
TSS validation rate by DHS

$$\text{TIEScore} = w_e f_e(\mathbf{E}) + w_d f_d(\mathbf{D}) + w_c f_c(\mathbf{C})$$

$w_e = 0.10$
$f_e(\mathbf{E}) = 5.45 \log_2 \mathbf{E} + 33.25$

Exon criterion (**E**)

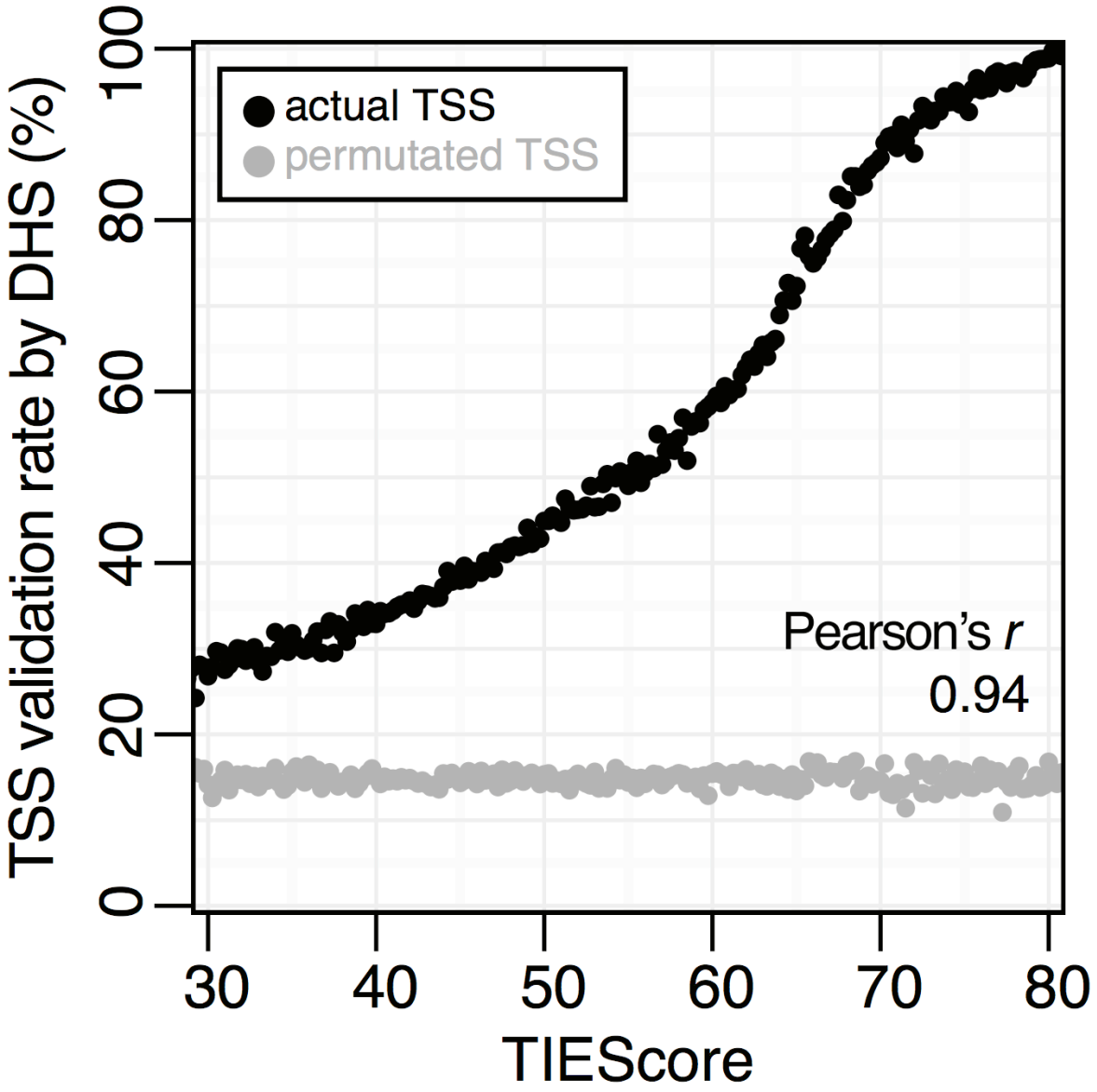
$w_d = 0.55$
$f_d(\mathbf{D}) = -1.19 \sqrt{\mathbf{D}} + 85.30$

Distance criterion (**D**)

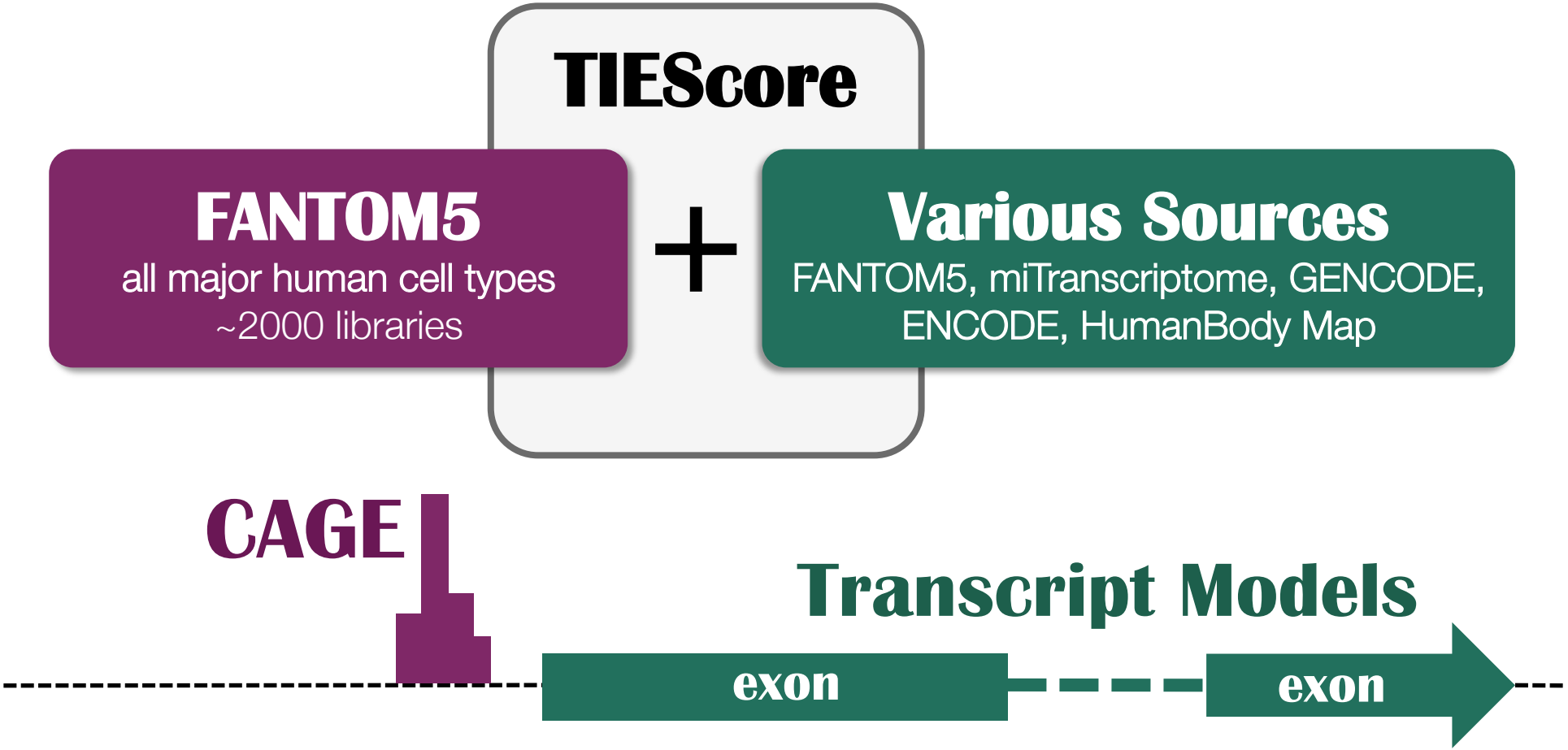
$w_c = 0.35$
$f_c(\mathbf{C}) = 6.63 \log_2 \mathbf{C} + 77.89$

CAGE criterion (**C**)

TIEScore Predicts TSS Validation Rate by DHS



TIEScore : *Combining CAGE and Transcript Models*

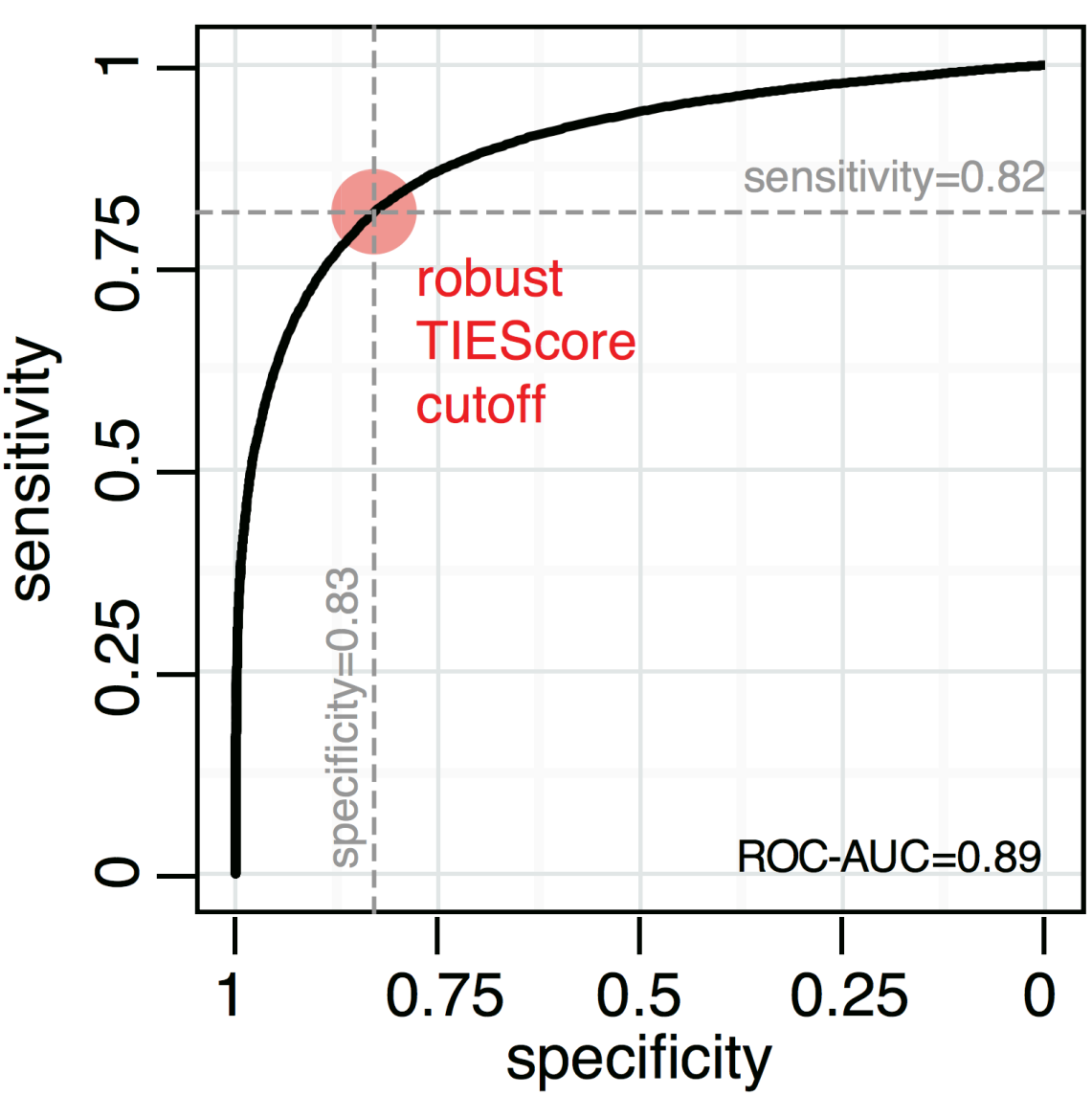


TIEScore : *Transcription Initiation Evidence Score*

FANTOM CAT

FANTOM CAGE Associated Transcriptome

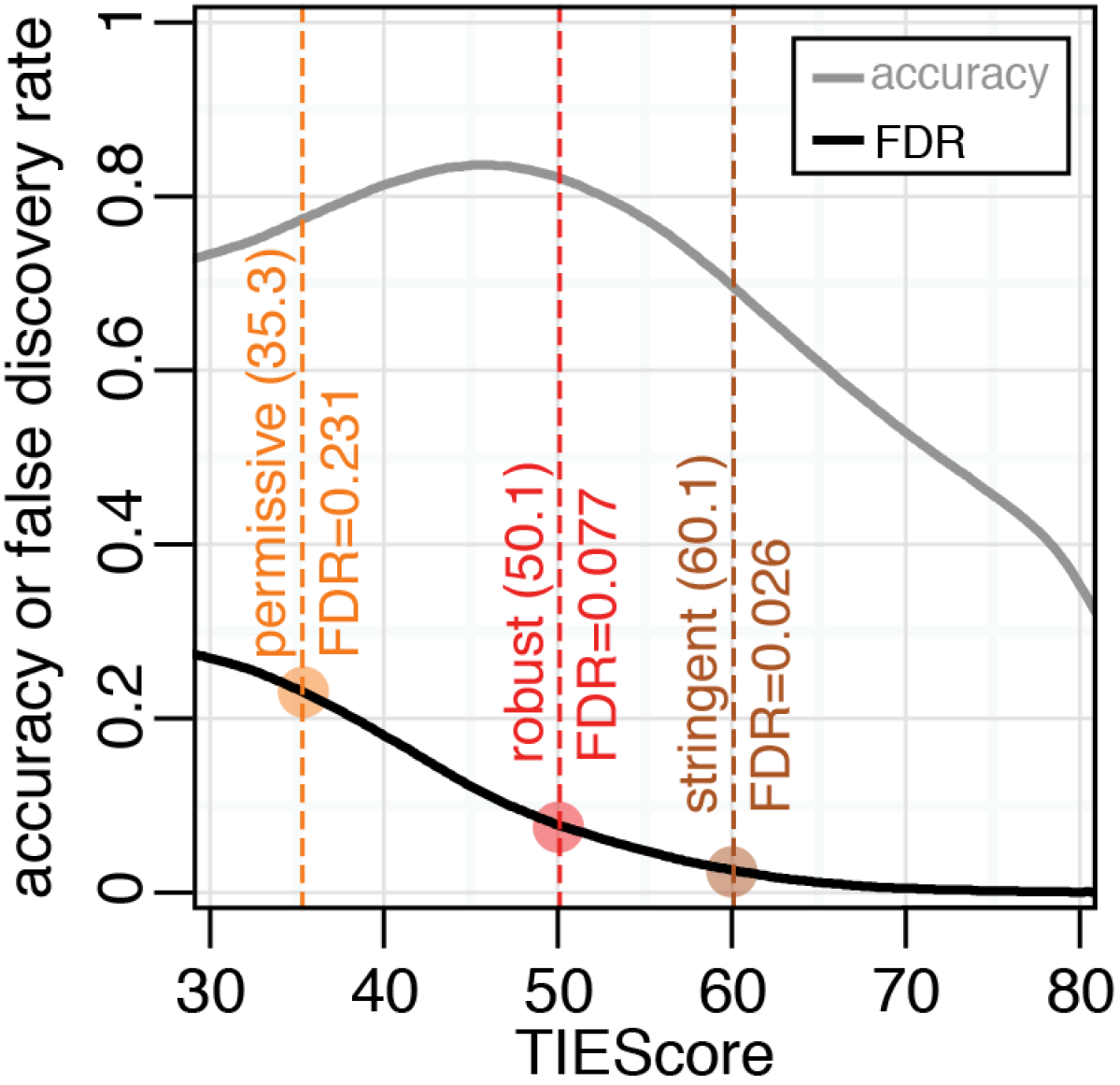
FANTOM CAT : *TIEScore Cutoffs*



Golden Standard

TSS & non-TSS based on Roadmap

FANTOM CAT : *TIEScore Cutoffs*



number of lncRNA loci

63,132
permissive

23,887
robust

13,213
stringent

FANTOM CAT : *versus other lncRNA catalogs*

miTranscriptome

Iyer et al. **(2015)** **Nature Genetics**

7,256 RNASeq Libraries, mainly TCGA

GENCODE

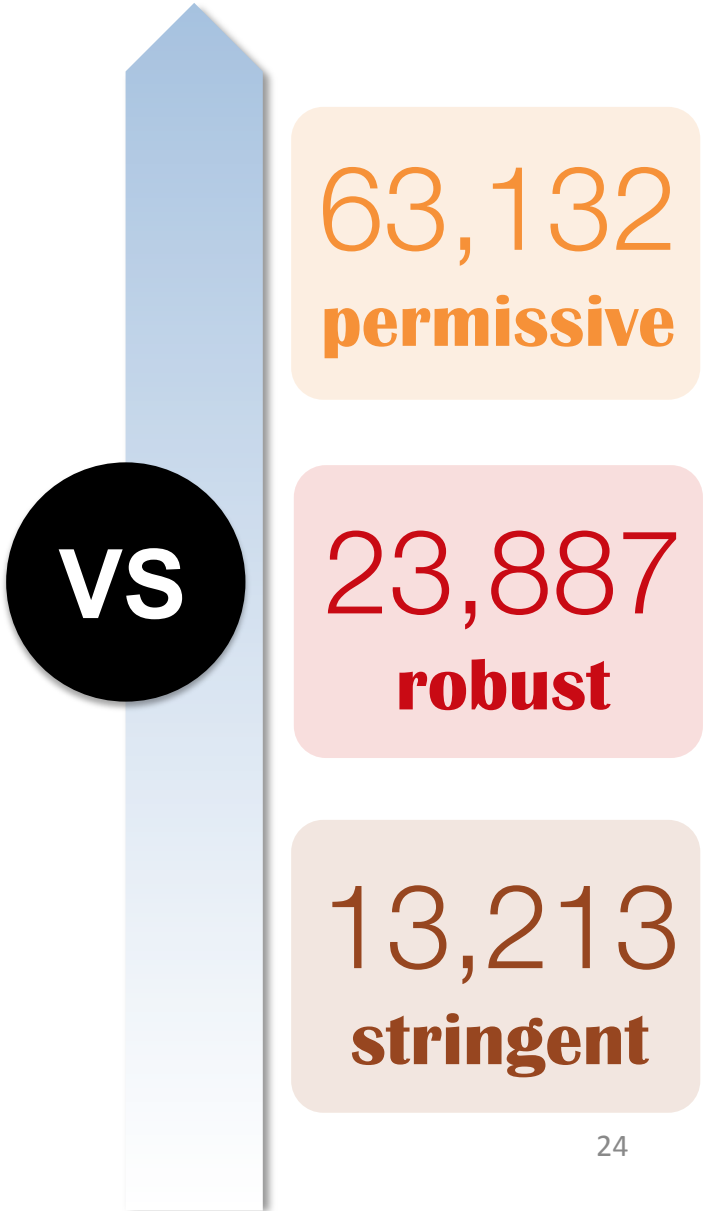
Derrien et al. **(2012)** **Genome Res**

ENCODE/Ensembl/Havana

Human BodyMap

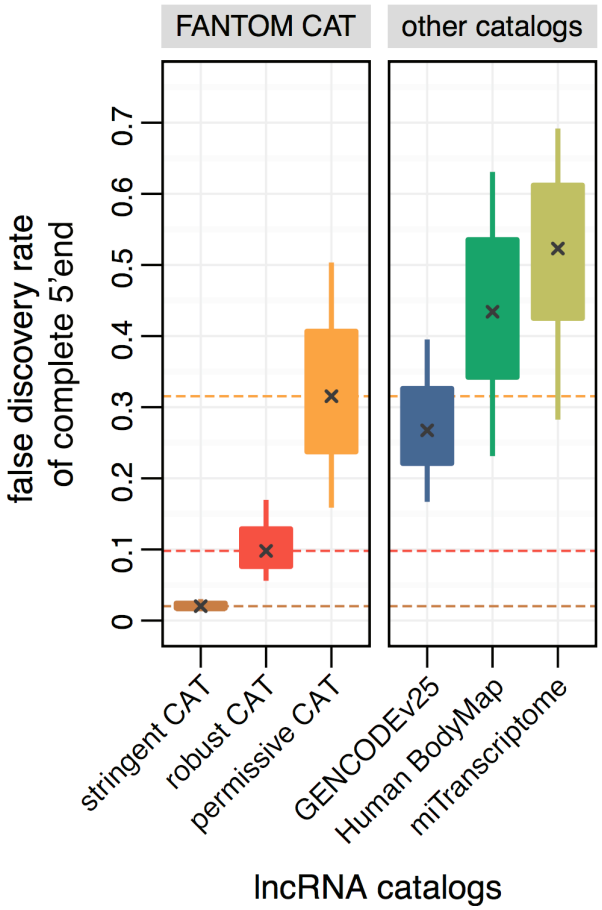
Cabili et al. **(2011)** **Genes Dev**

RNASeq on 24 tissues

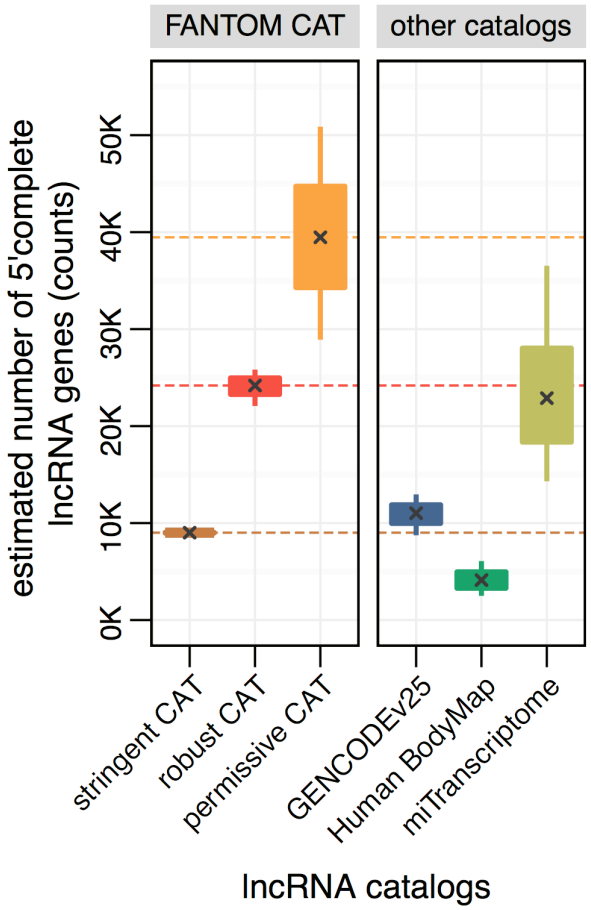


FANTOM CAT : *versus other lncRNA catalogs*

Accuracy



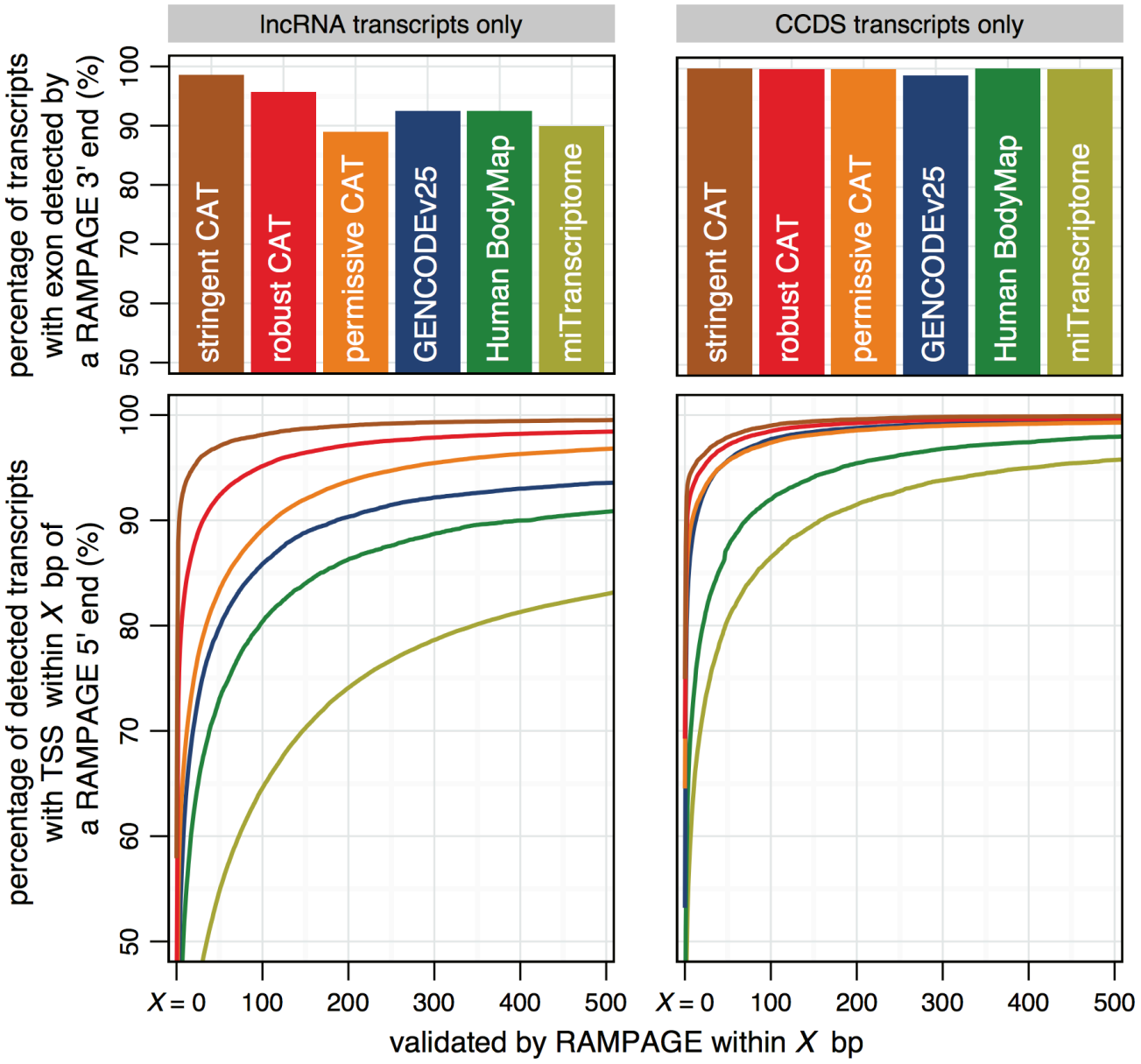
Scope



Golden Standard

TSS & non-TSS based on Roadmap

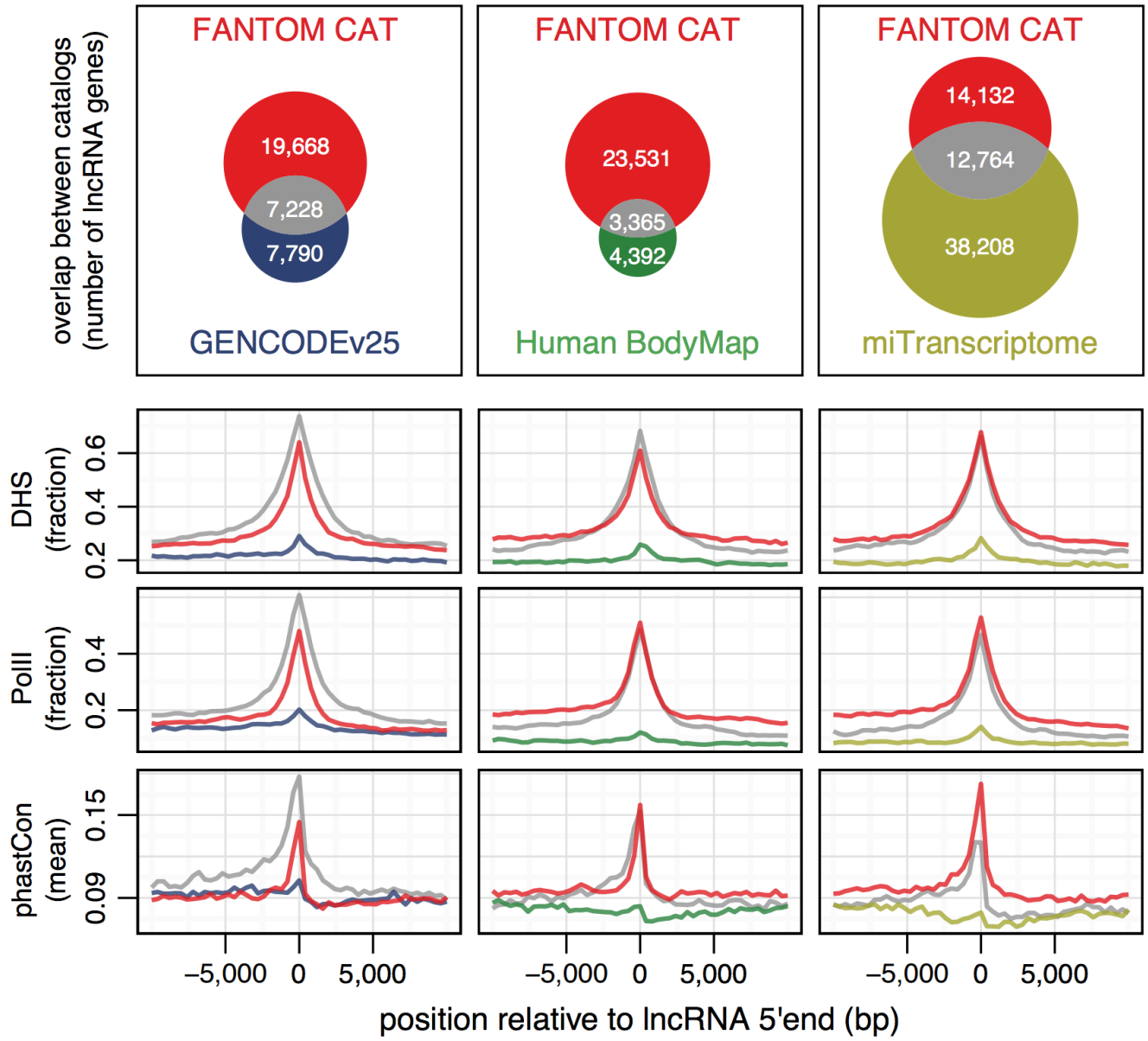
FANTOM CAT : *versus other lncRNA catalogs*



ENCODE RAMPAGE

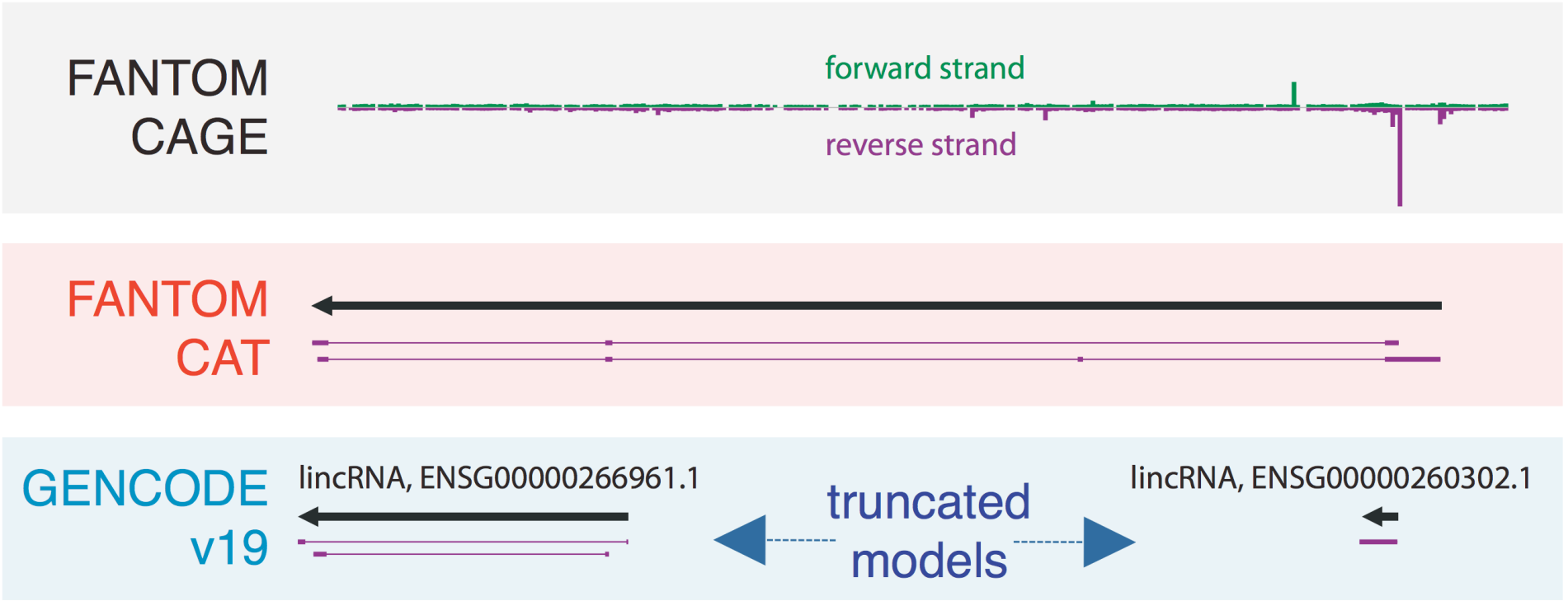
Paired-end read linking 5'end to exon

FANTOM CAT : *versus other lncRNA catalogs*



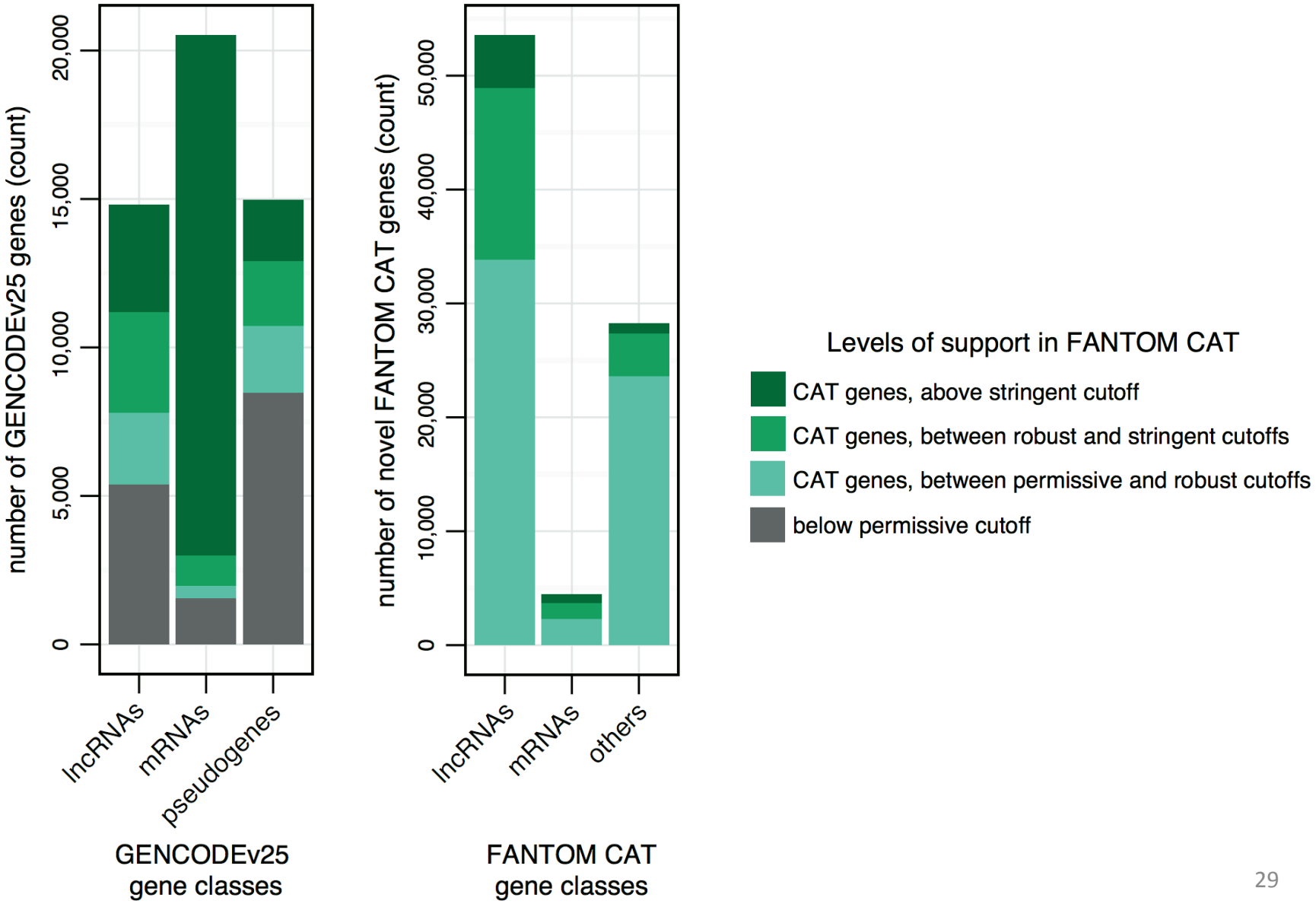
Genomic + Epigenomic
Conservation + Roadmap epigenome

FANTOM CAT : *versus other lncRNA catalogs*



lack of CAGE support

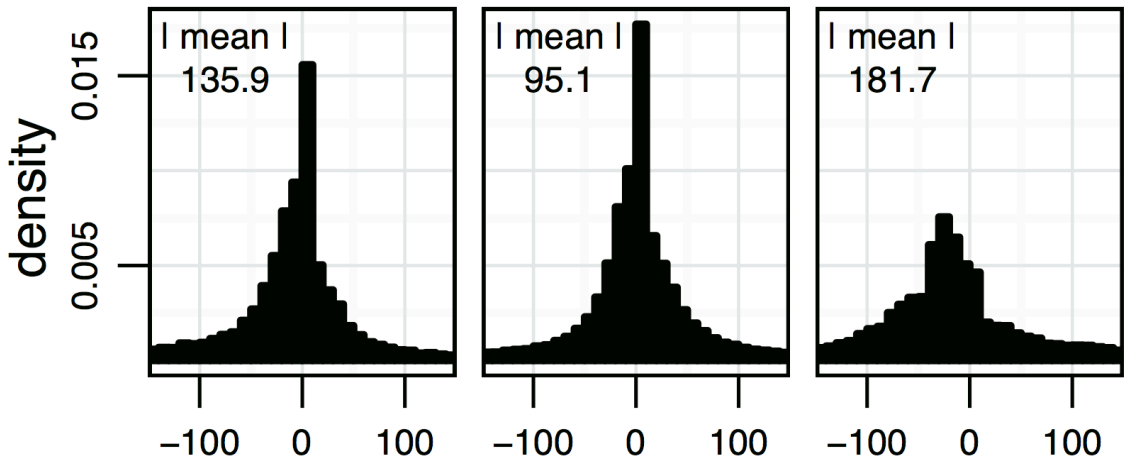
FANTOM CAT : *versus GENCODEv25*



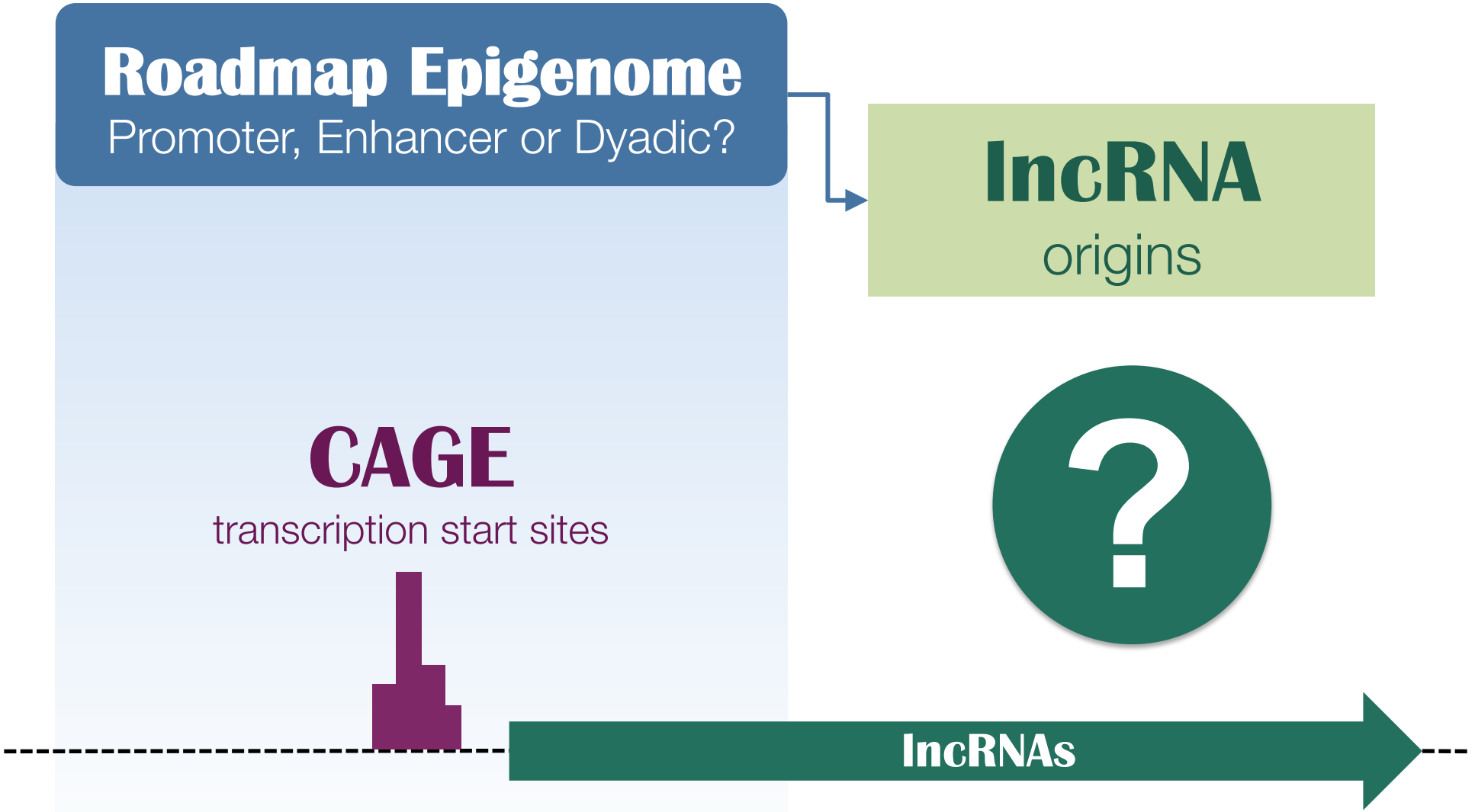
FANTOM CAT : *versus GENCODEv25*

GENCODEv25 transcripts with
5'end revised by FANTOM CAT

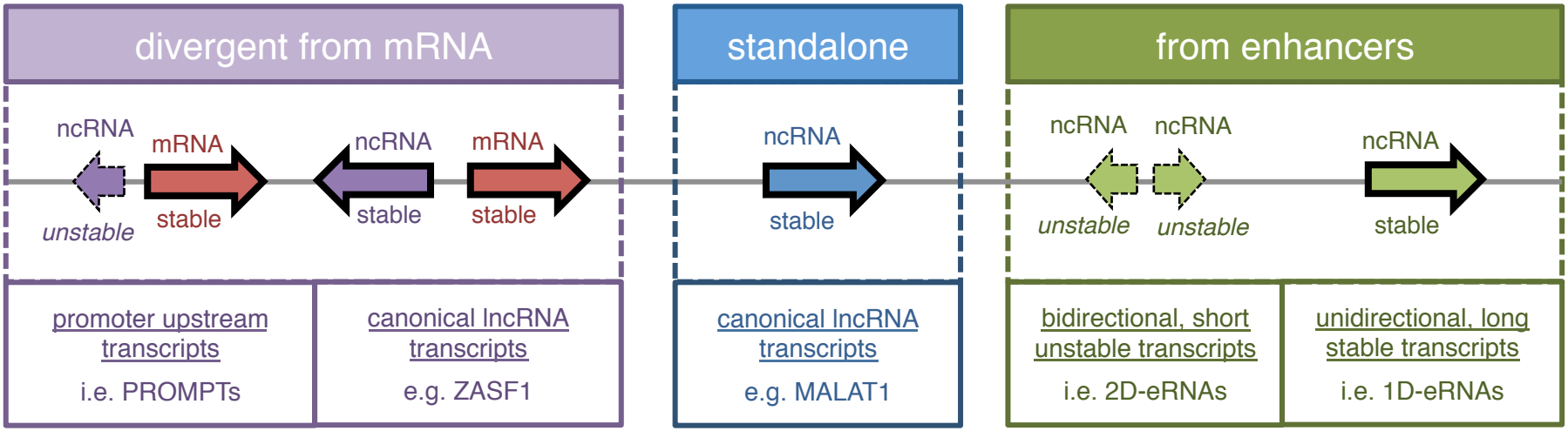
	lncRNA	protein coding	pseudogene
total	26,578 100%	145,926 100%	17,616 100%
≥ 0 bp	16,914 64%	128,207 88%	8,111 46%
≥ 10 bp	12,812 48%	93,788 64%	7,376 42%
≥ 50 bp	7,061 27%	45,343 31%	4,867 28%
≥ 100 bp	5,090 19%	29,713 20%	3,456 20%
≥ 250 bp	3,033 11%	14,963 10%	1,970 11%



FANTOM CAT : 5'completeness

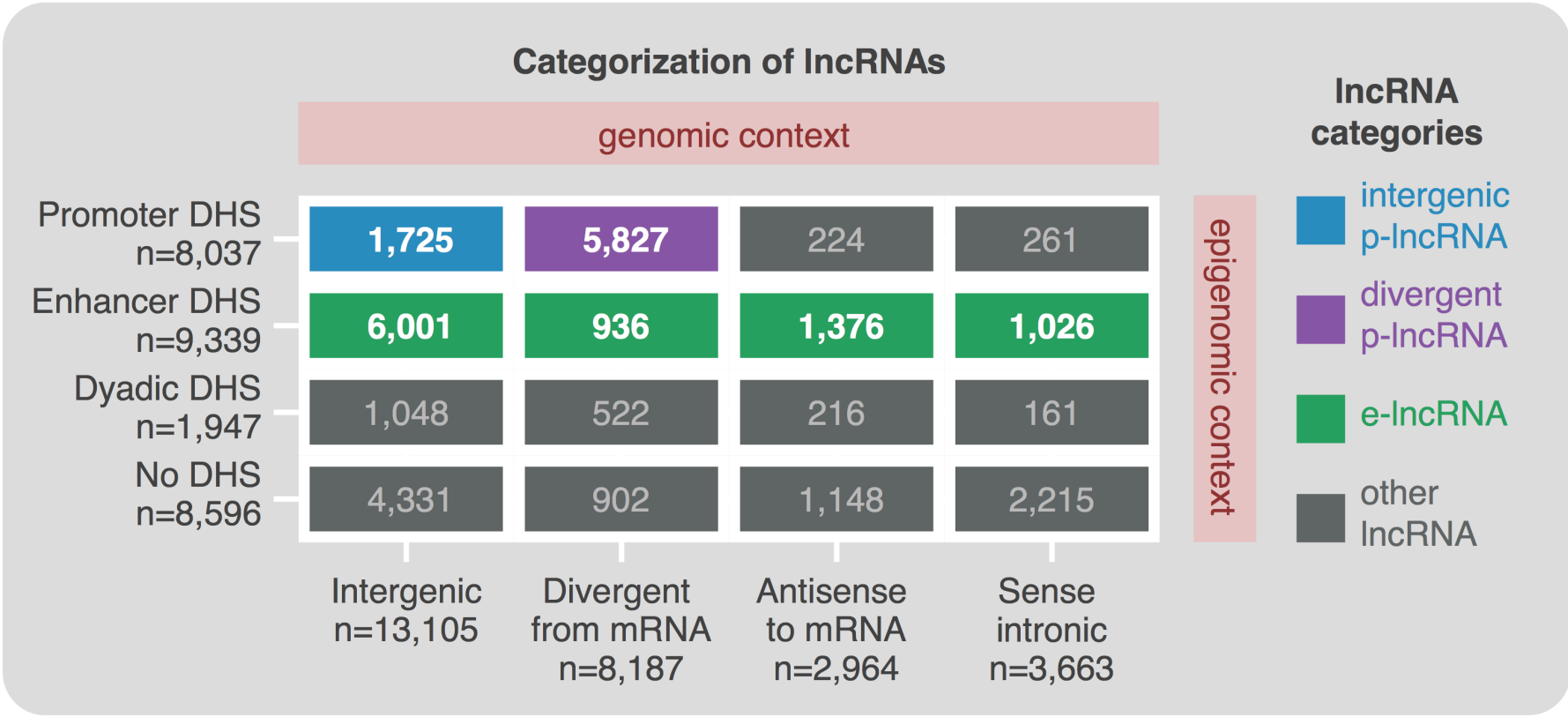


lncRNA: *origin based on regulatory regions*



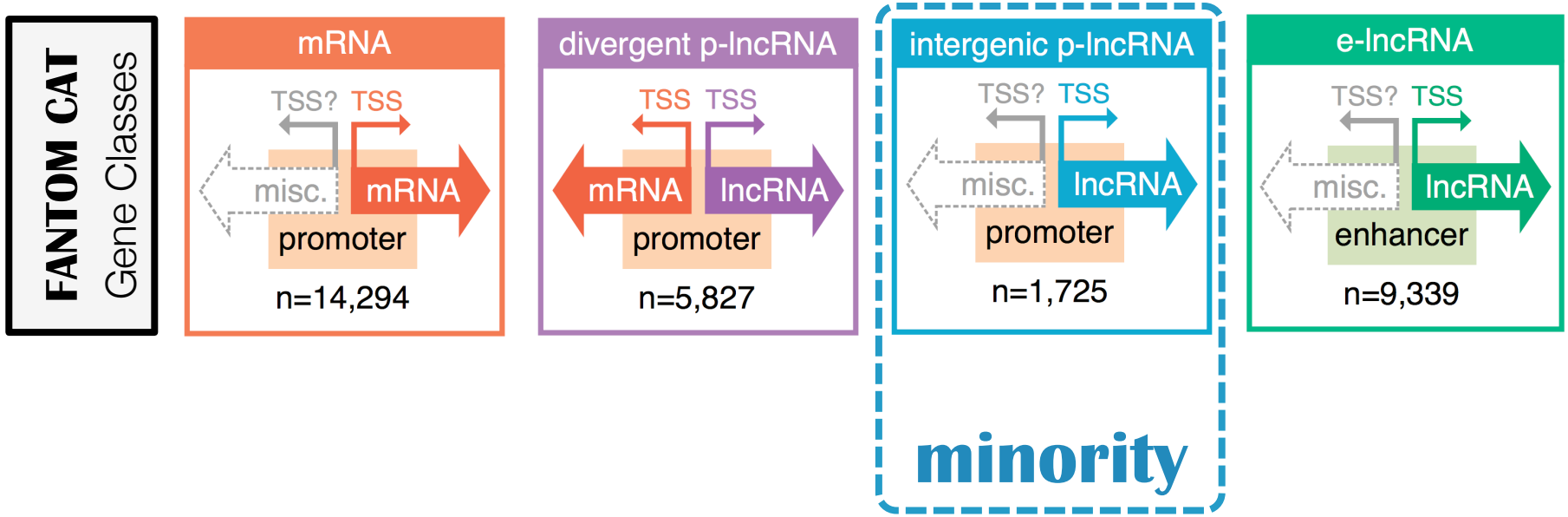
What is a gene?

lncRNA: *origin based on regulatory regions*



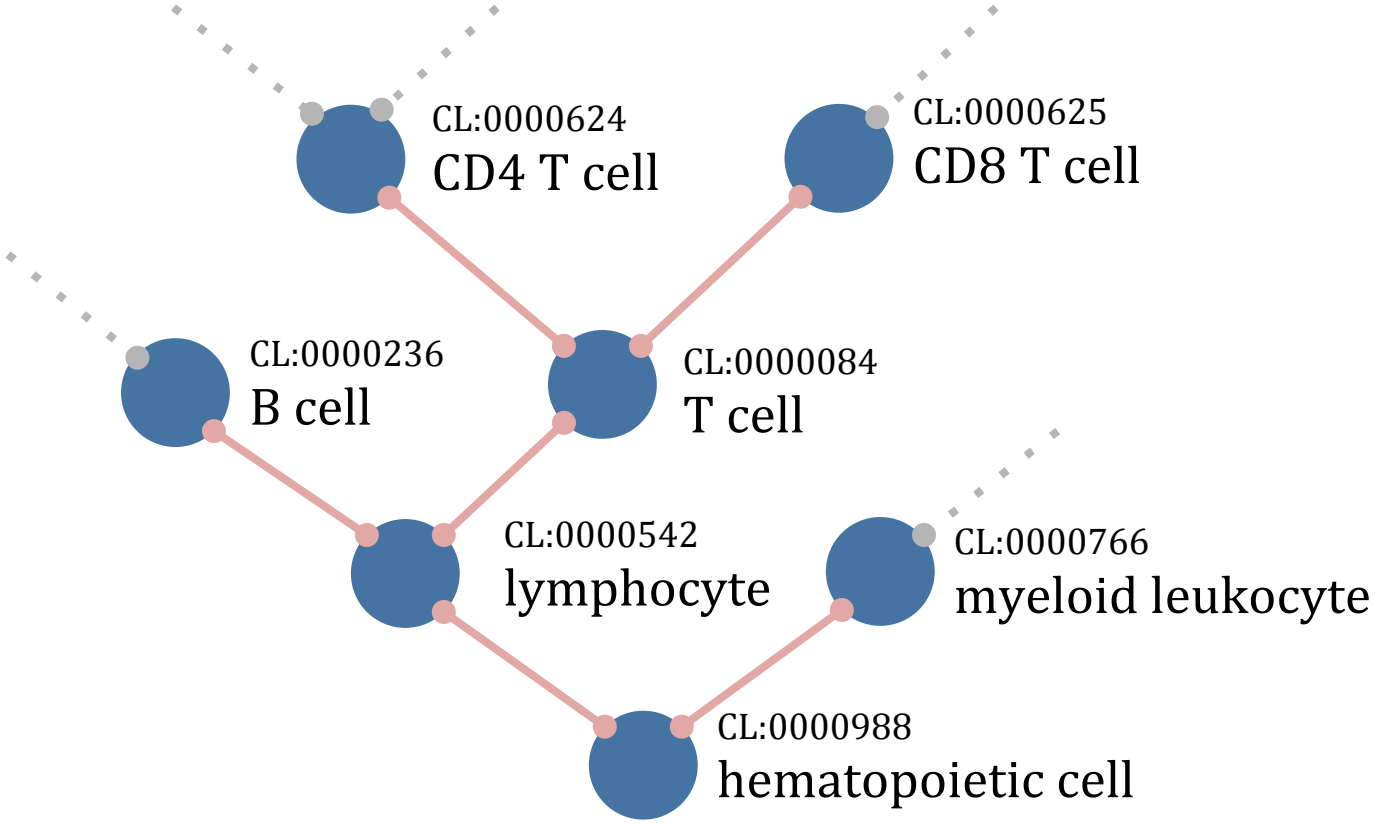
Genomic + Epigenomic Context

lncRNA: *origin based on Roadmap DHS*



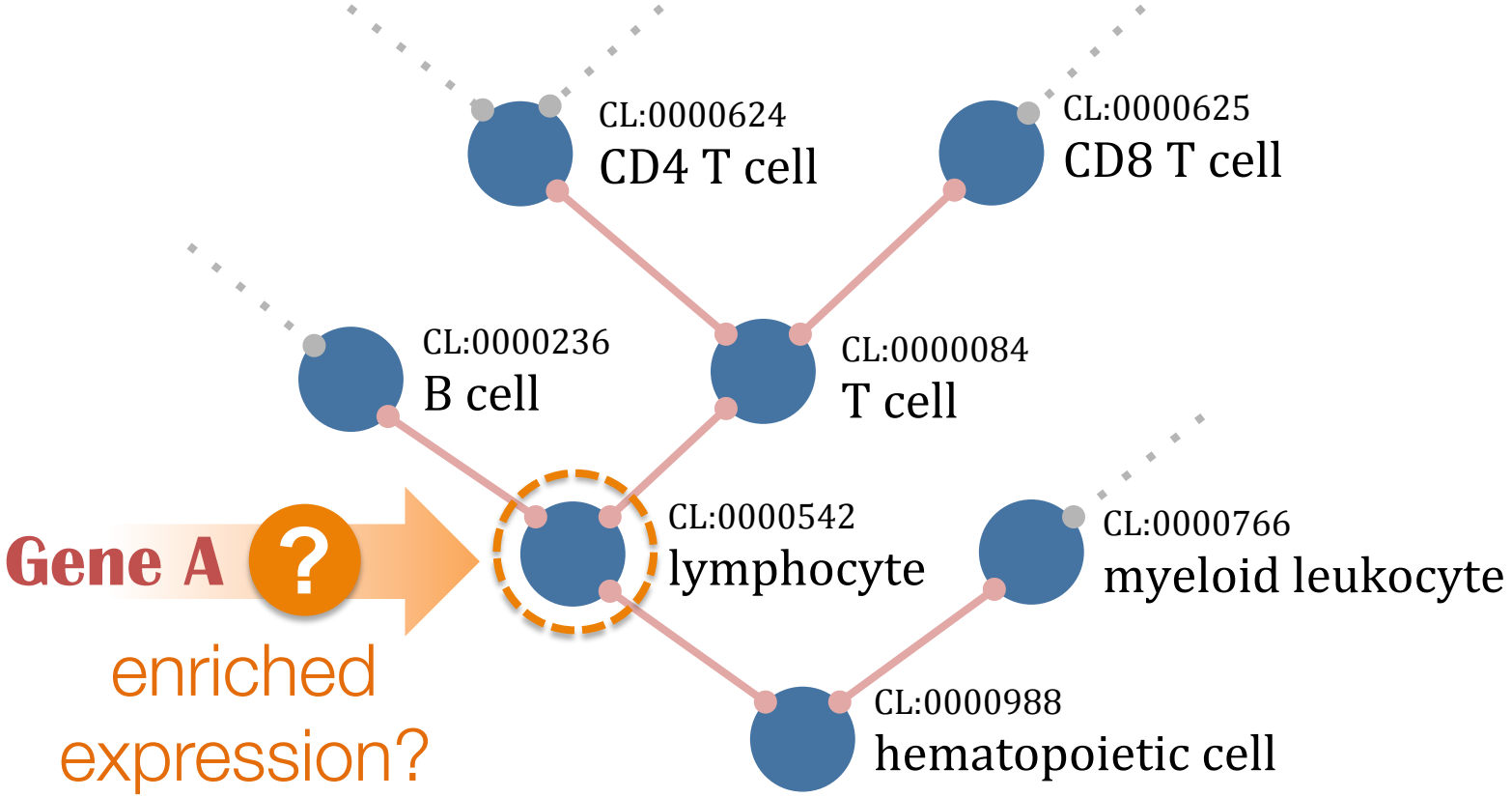
What is a gene?

Expression atlas : *Sample Ontology Enrichment Analysis*



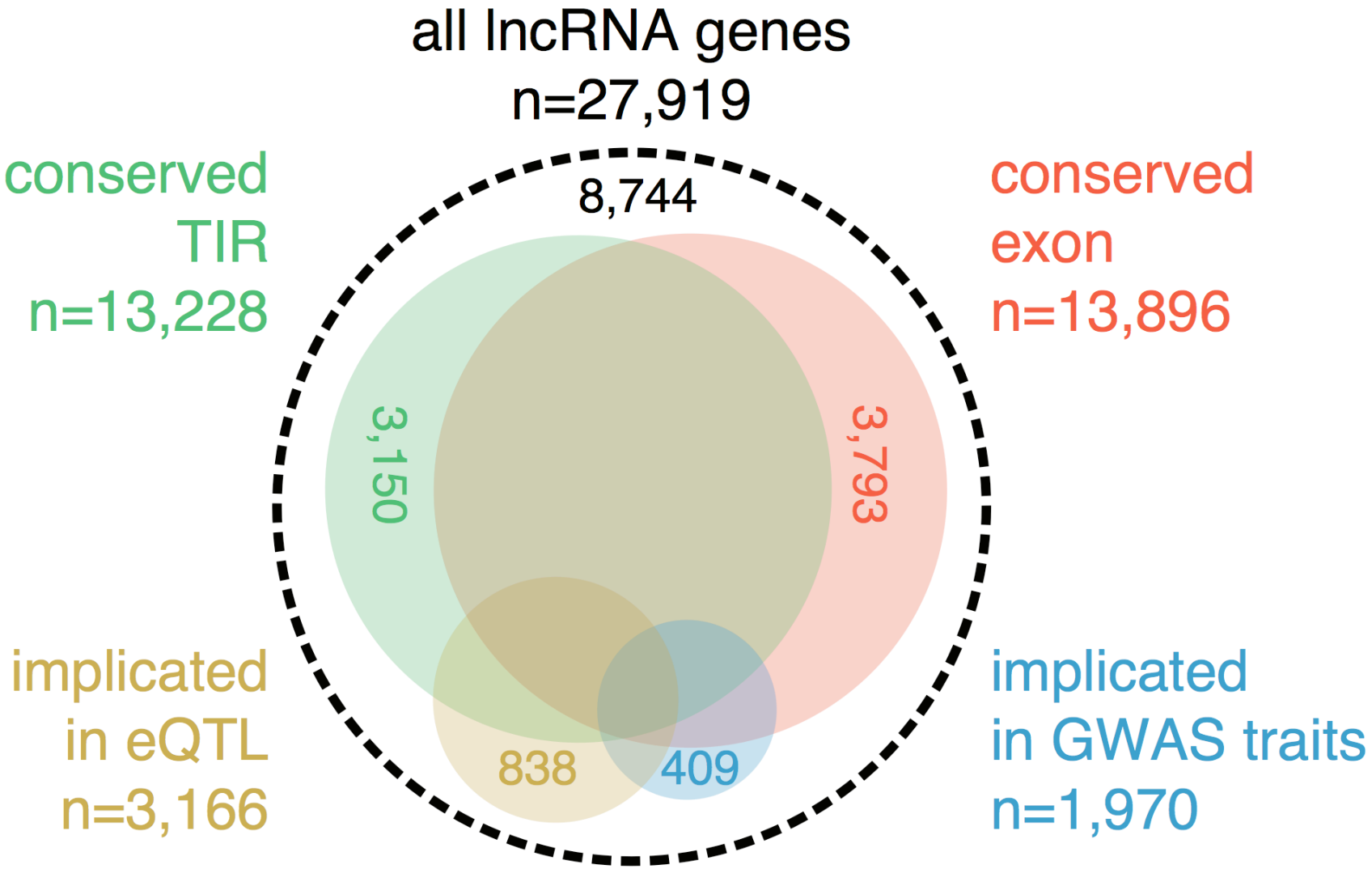
~2000 samples annotated with ~600 sample ontologies

Expression atlas : *Sample Ontology Enrichment Analysis*



every gene in every ontology

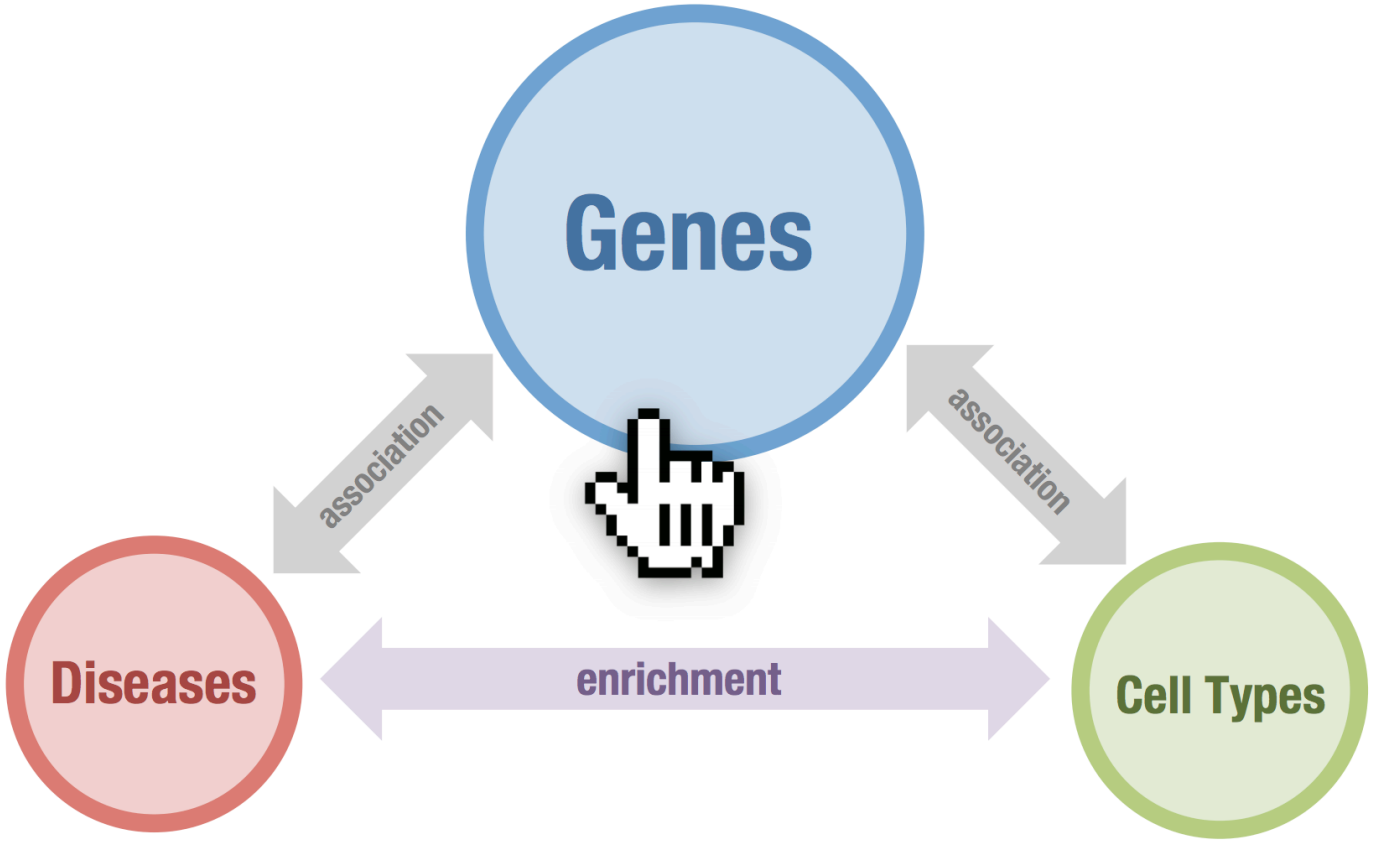
Functional evidences : *bring it altogether*



Web Resource : *FANTOM CAT Browser*

Online Interactive Resources

http://fantom.gsc.riken.jp/cat/



FANTOM CAT : *continuously maintained*

1. Regularly Updated
2. More CAGE (RIKEN)
3. More Transcript Models (Public)
4. Poly-A Tail Data Mining From Public RNASeq
5. Extension to Mouse Genome (MGI)

FANTOM CAT : *Integration with GENCODE*

1. Direct Integration of FANTOM CAGE peaks

- Run TIEScore with raw CAGE peaks to existing GENCODE models?
- Improves existing GENCODE models 5'end

2. Integration of FANTOM CAT models

- Three levels:
  - 1) Novel loci only; 2) Novel isoforms of GENCODE loci; 3) Revised GENCODE transcripts;
- Compatibility of definition of a "gene" in FANTOM CAT vs GENCODE
- Splicing junctions are only as good as the models themselves → filter on splicing junction support from public RNASeq?
- Compatibility of identifiers
- FANTOM CAT is on hg19 → liftover to hg38?

3. Integration of FANTOM CAT annotations

- Genomic context: divergent, antisense, intron etc
- Epigenomic context: promoter, enhancer etc
- Expression Atlas: specificity to certain cell-types

4. Mode of collaboration

- Personnel exchange?
- Update cycle?