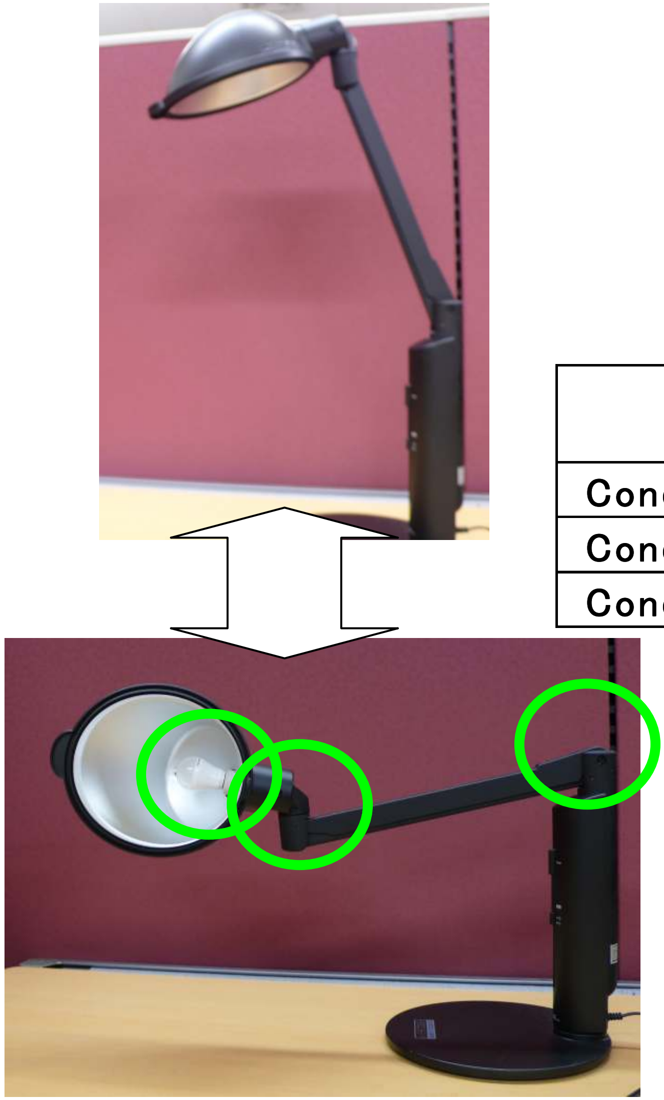
Extracting "articulated transcription starting sites (ATSS)" with Factor Analysis on CAGE tag data.

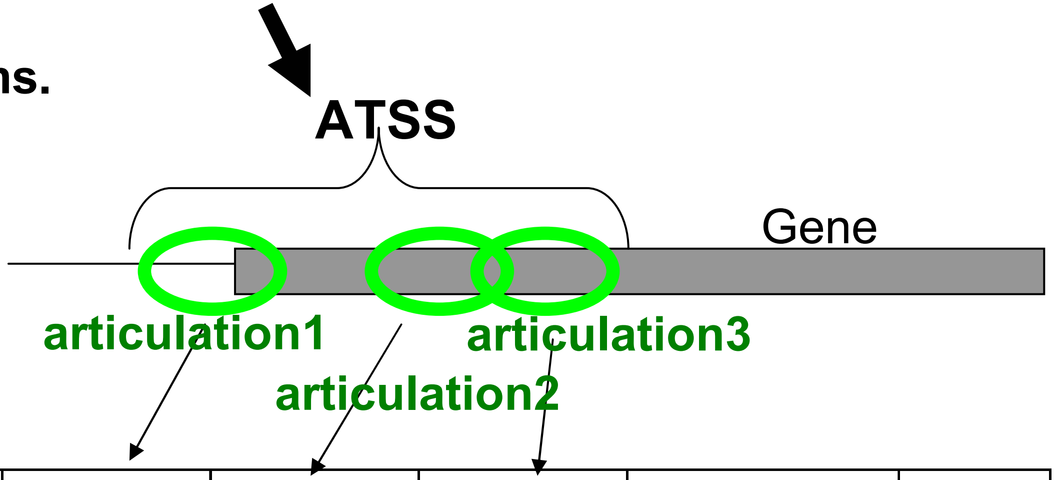
17th Oct, 2011

Shuji Kawaguchi, Kei Iida, Tetsuro Toyoda
BASE, RIKEN, JAPAN

What is articulated transcription starting sites (ATSS) ?

Articulated arm can take various forms.



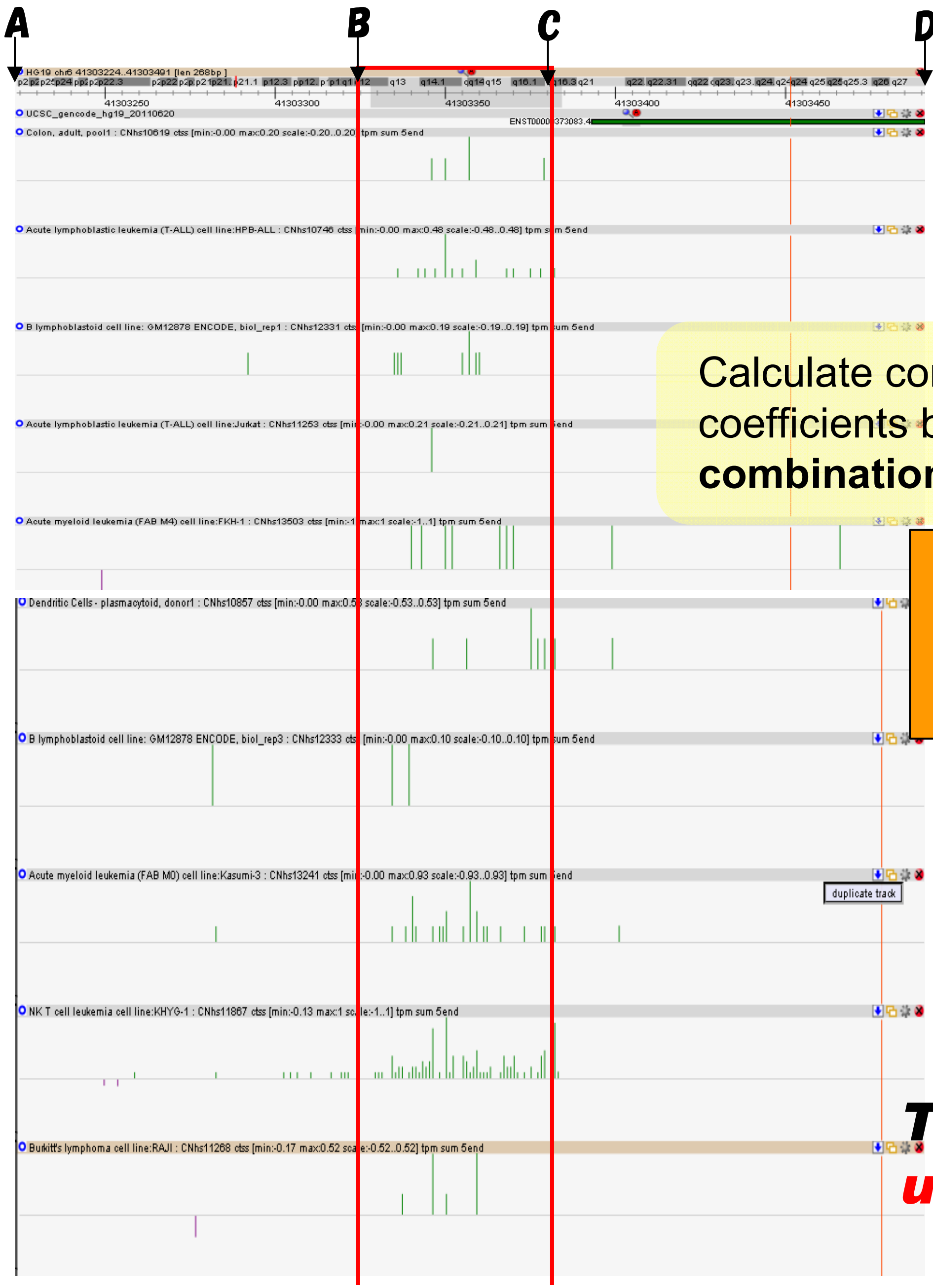


	aticulation 1	aticulation 2	aticulation 3	expression level	signal peptide
Condition1	✓			Middle	on
Condition2		✓		Low	off
Condition3	✓		✓	High	both

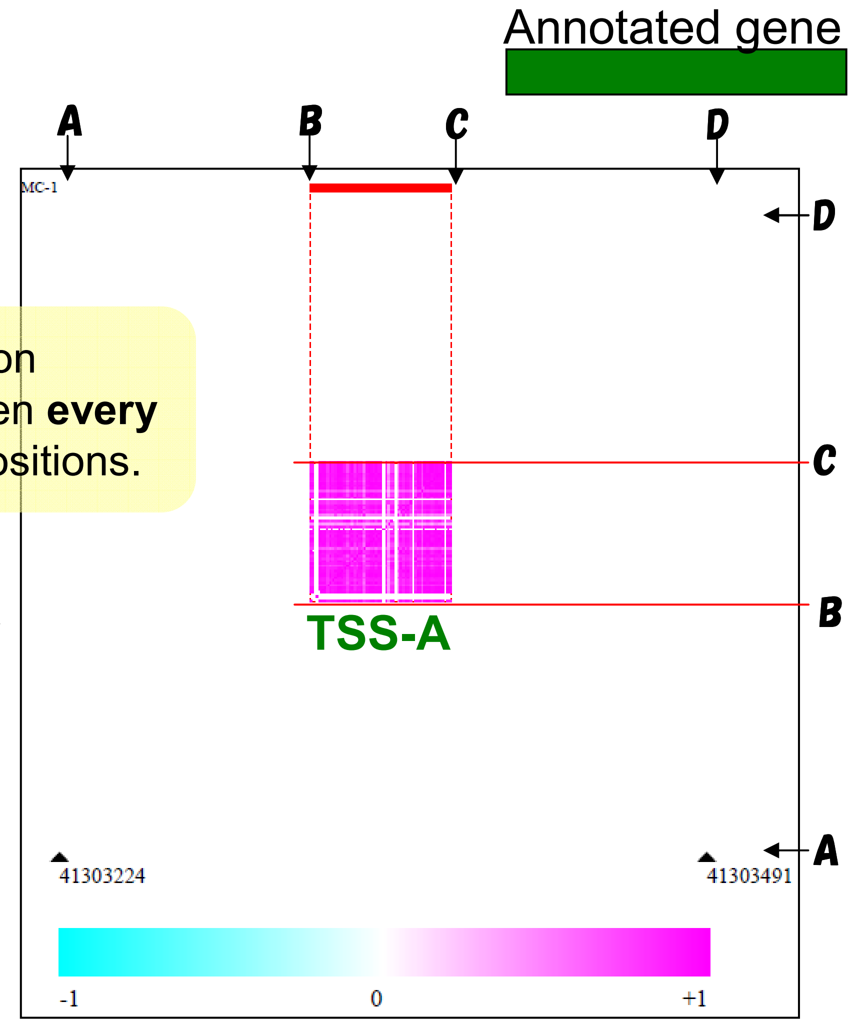
ATSS can take various patterns of expressions by separation or combined using of articulations

Transcription-Starting-Site-Articulations: **TSS-A**

How can we find TSS-A? → Use correlation info.



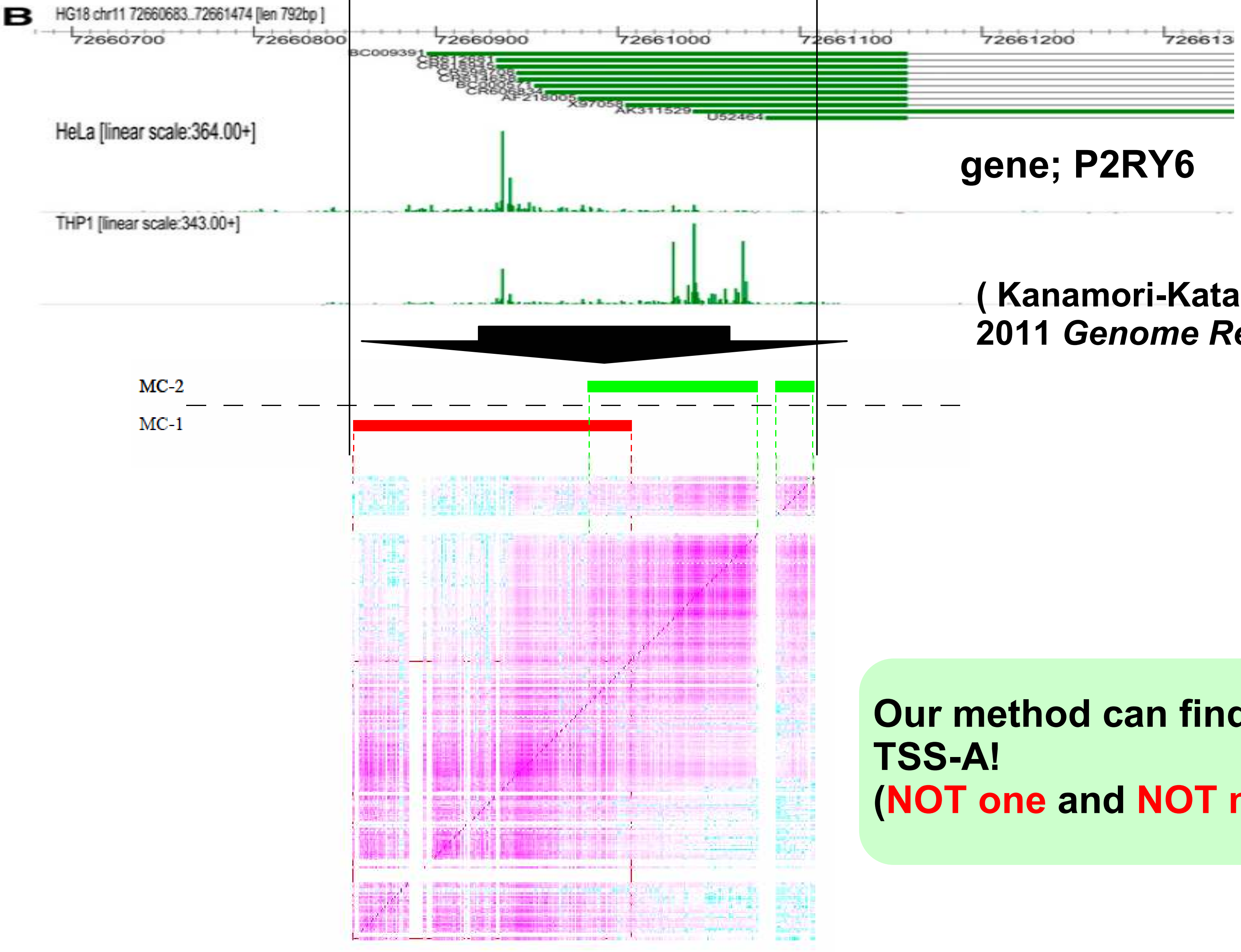
Calculate correlation coefficients between **every combination** of positions.



TSS-A is defined as group of TSSs under the same regulation.

What does the TSS-A show us?

Example; HeliscopeCAGE shows two peaks, active in different states.



gene; P2RY6

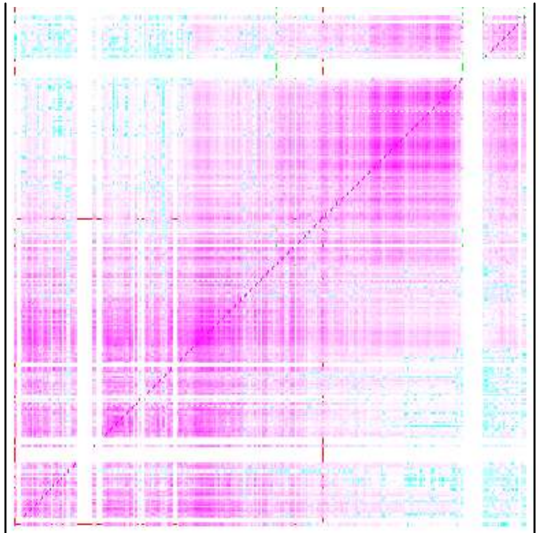
(Kanamori-Katayama *et al.* 2011 *Genome Res.*)

Our method can find “two” TSS-A!
(NOT one and NOT many)

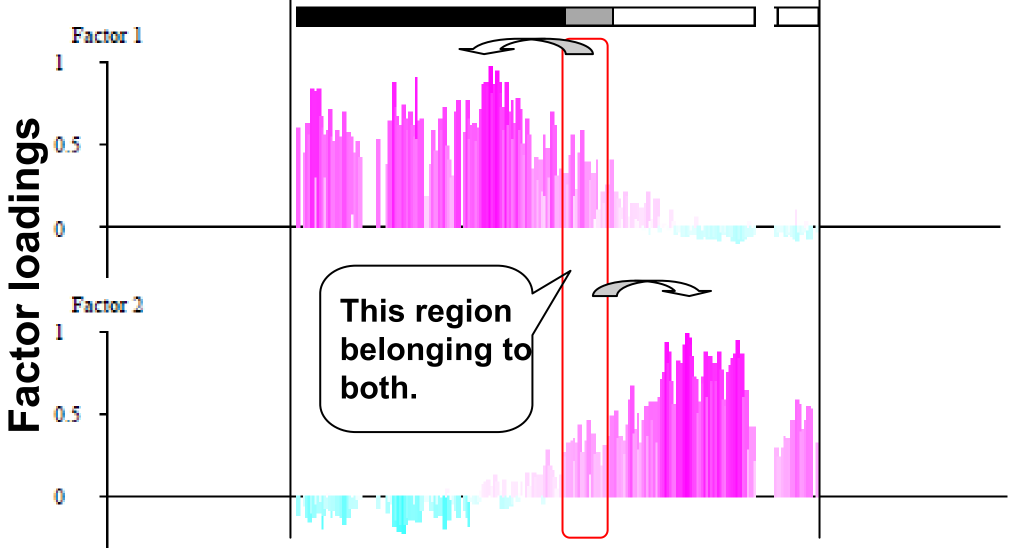
How is the method working?

Our method splits the regions by the "Factor Analysis (Promax method)".

1. We calculate correlation coefficients.

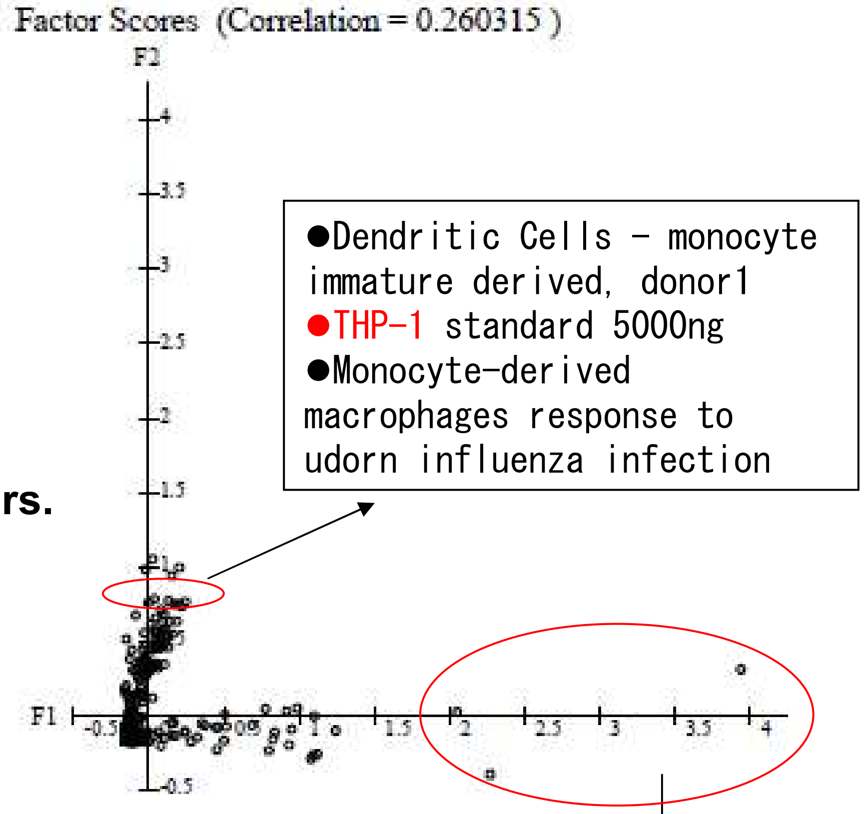


2. We perform factor analysis, then define clusters.



This region belonging to both.

3. Each cluster are characterized with factor scores.



●Dendritic Cells - monocyte immature derived, donor1
●THP-1 standard 5000ng
●Monocyte-derived macrophages response to udorn influenza infection

●cervical cancer cell line:D98-AH2
●mucinous adenocarcinoma cell line:JHOM-1
●ntestinal epithelial cells, donor1

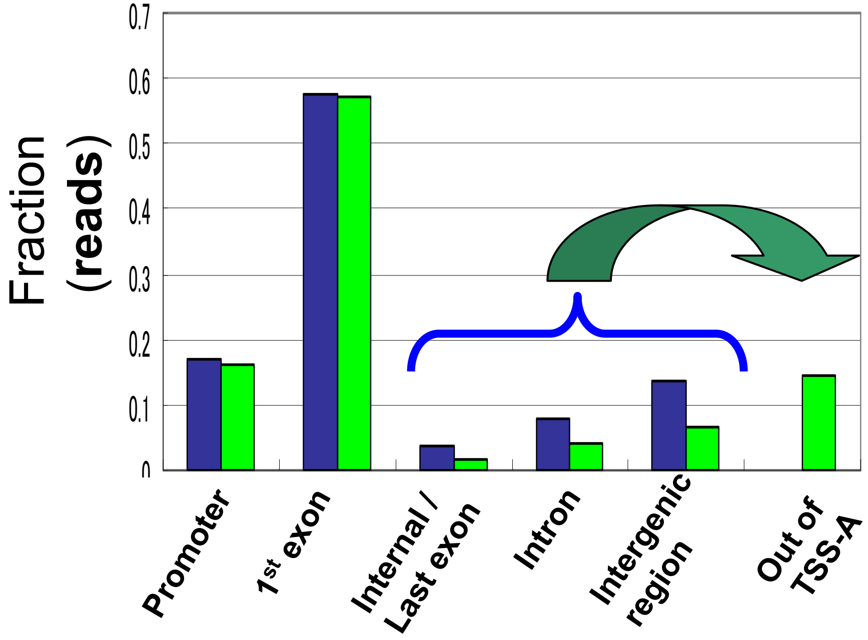
Many TSS-A are defined on 1st Exons / Promoter regions

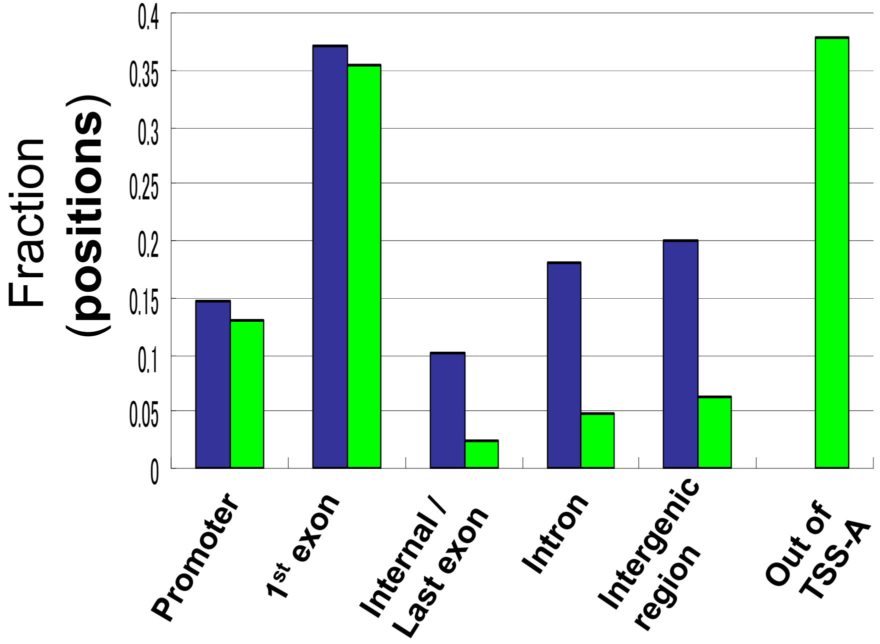
CNhs10722

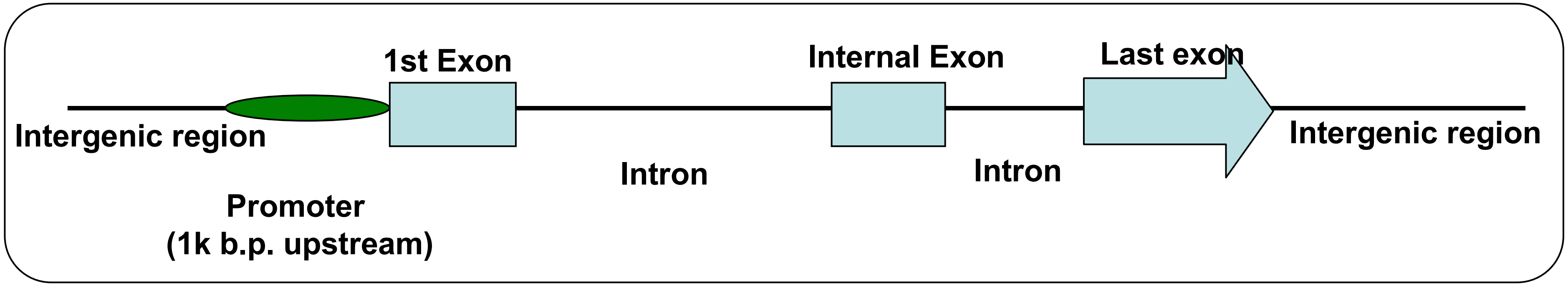
acute myeloid leukemia (FAB M5) cell line:THP-1 (fresh)

🟦 :ALL CAGE tags

🟩 :CAGE tags within TSS-A







1st Exon

Internal Exon

Last exon

Intergenic region

Intergenic region

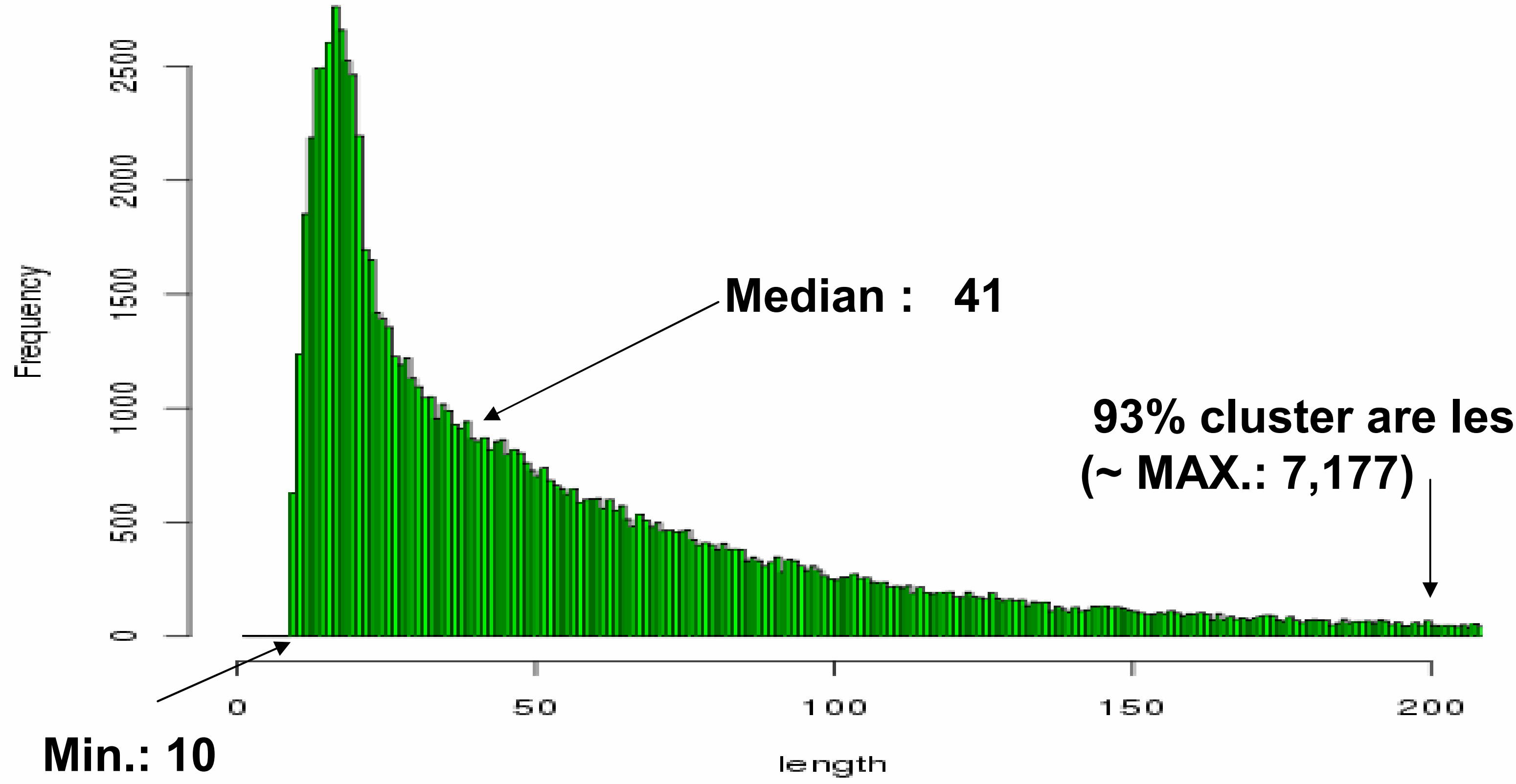
Intron

Intron

Promoter
(1k b.p. upstream)

Most TSS-A are short: ~200 b.p.

…may be easy to use on following analyses.



Median : 41

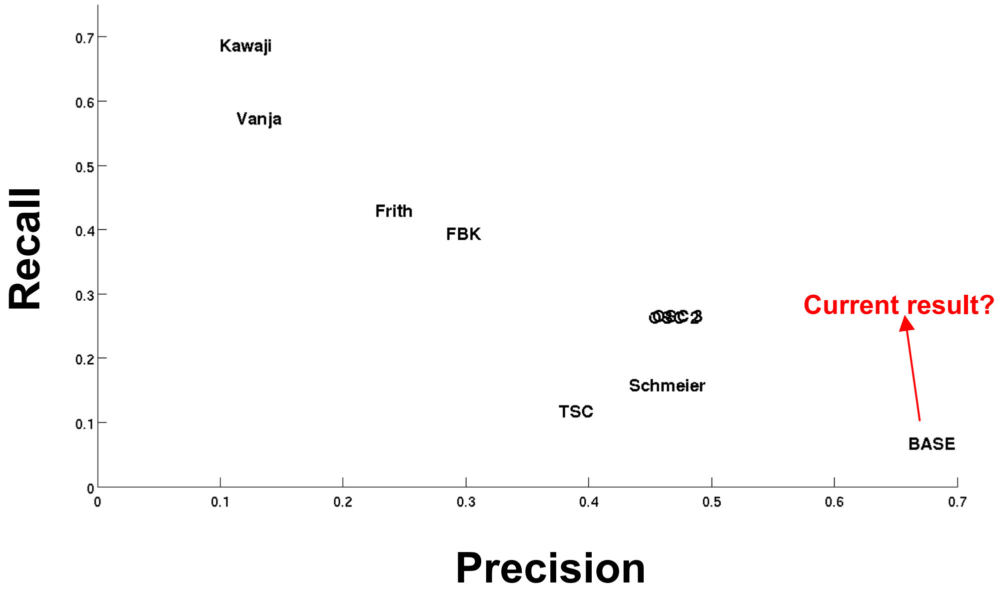
93% cluster are less than 200
(~ MAX.: 7,177)

Min.: 10

Problem: TSS-A fail to describe some known promoters?

Version	# of TSS-A
UPDATE011 @clustering competition	38,416
UPDATE011, after improbement	79,906
UPDATE013	97,941

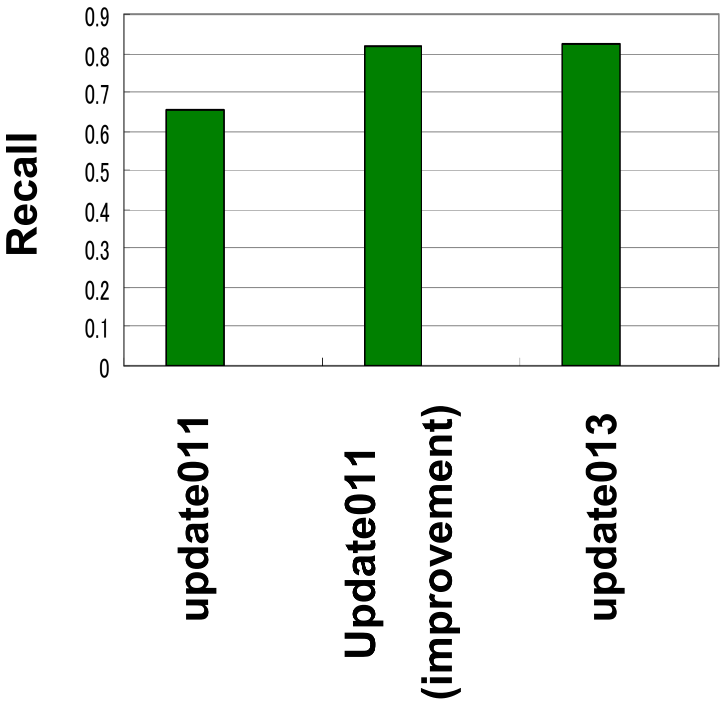
[Based on 203,514 ENCODO promoters]



Current result?

(A figure from FANTOM mailing list posted by Dr. Ben Brown)

[Based on 21,188 refgene promoters]



Summary

1. We define 97,941 TSS-A with using correlation of CAGE tag expression.
2. TSS-A are defines as differently regulated regions. We showed that TSS-A can be overlapped.
3. Performance of our method to describe TSS-A are still under improvements.

Acknowledgment

Shuji Kawaguchi
Development the method.

Tetsuro Toyoda
General management of our project.