CAGE as a tool for cancer research and biomarker discovery

CAGE symposium

"New Topics of cancer research with CAGE method"

Bogumil Kaczkowski
Foreign Postdoctoral Researcher
RIKEN CLST, Yokohama, Japan

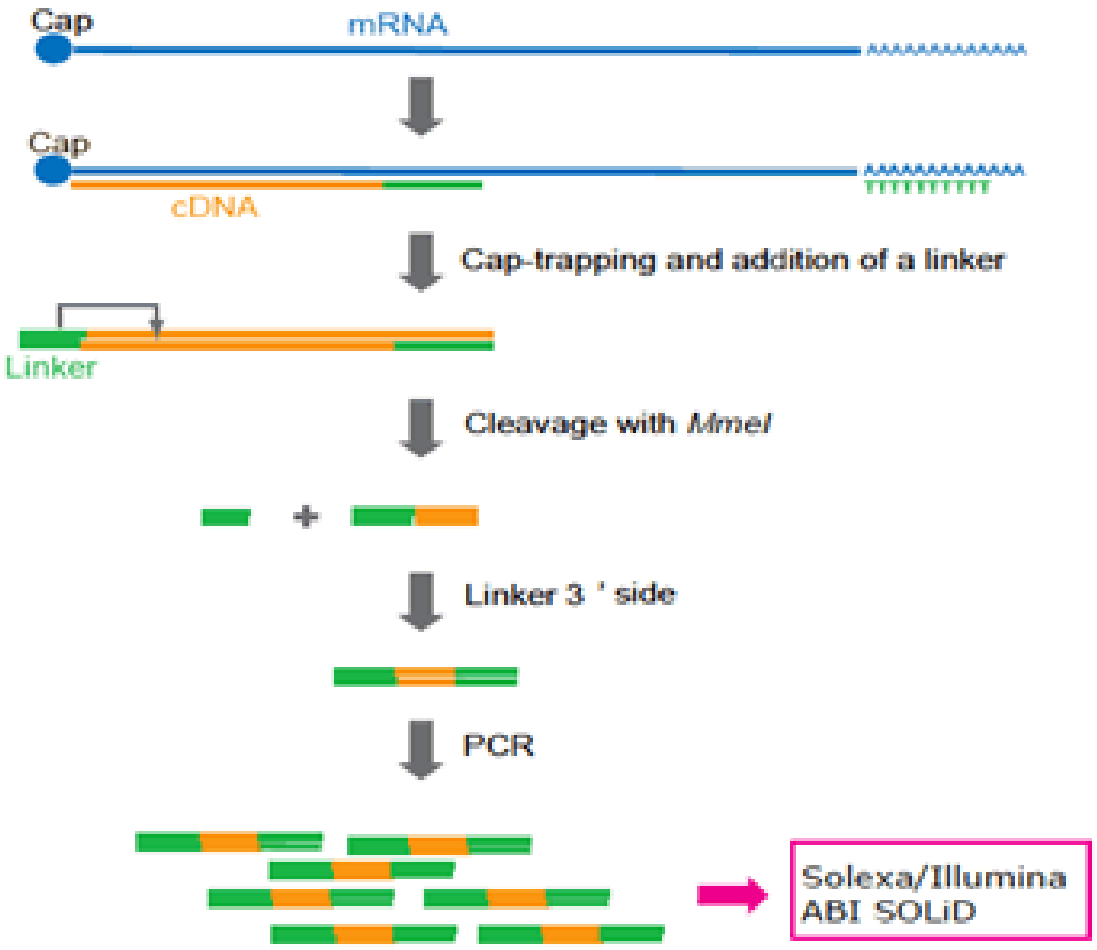
Sept 13, 2016

Agenda

- Introduction to CAGE
- CAGE and biomarker discovery
- CAGE and cancer lncRNAs
- CAGE and enhancerRNAs
- CAGE and repetitive elements

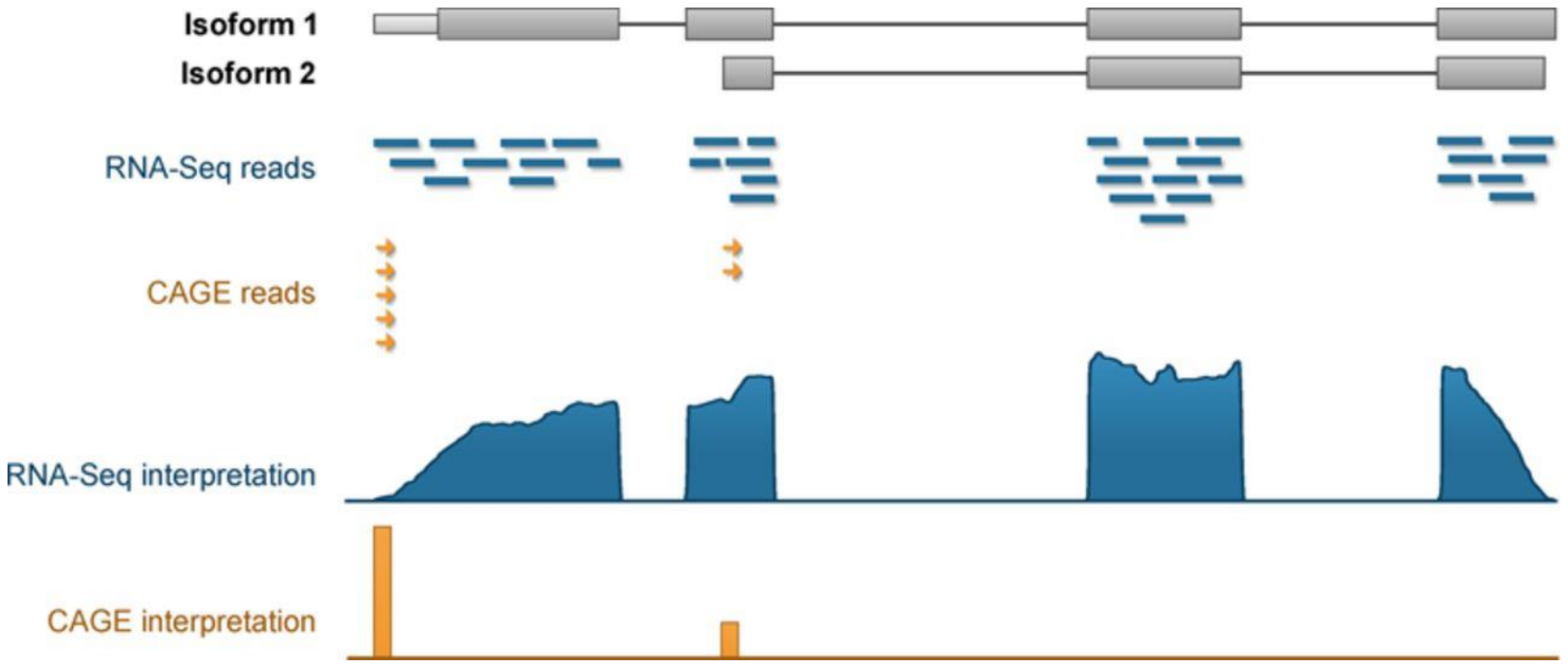
Quick intro: Cap Analysis Gene Expression (CAGE)

- **Capture of 5’ end of capped RNAs as a short sequence tags**
- **followed by high-throughput sequencing**



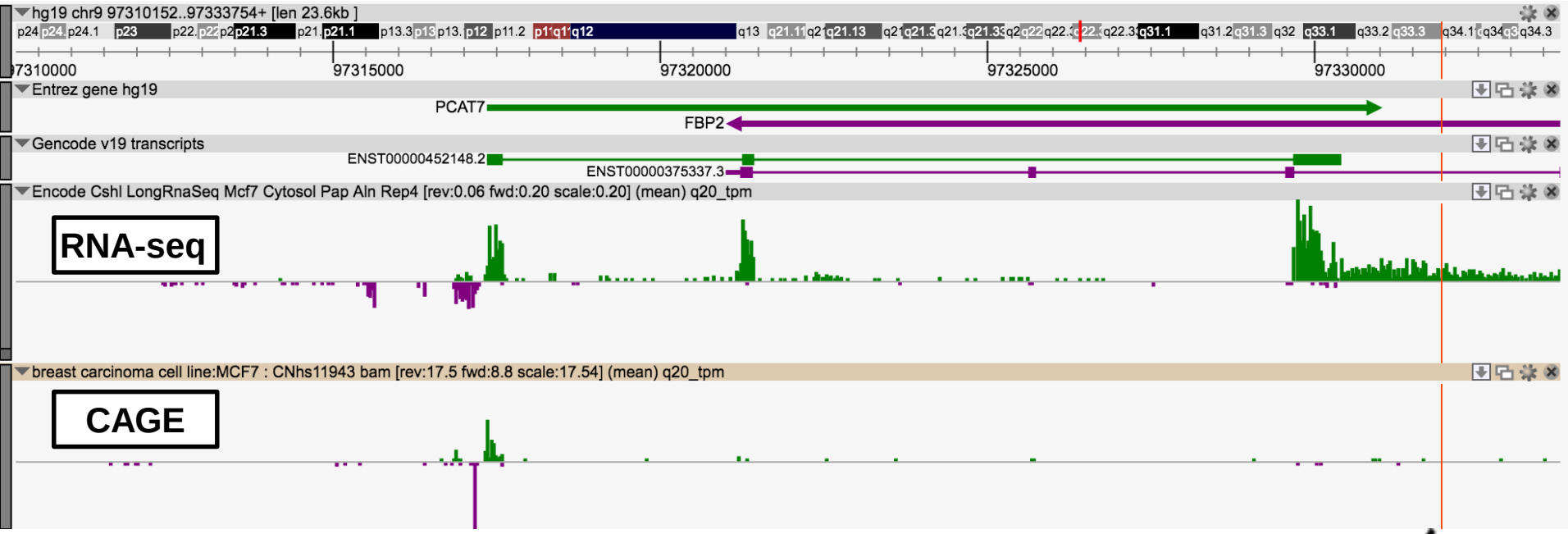
http://fantom.gsc.riken.jp/protocols/basic.html

CAGE Cap Analysis Gene Expression vs RNA-Sequencing



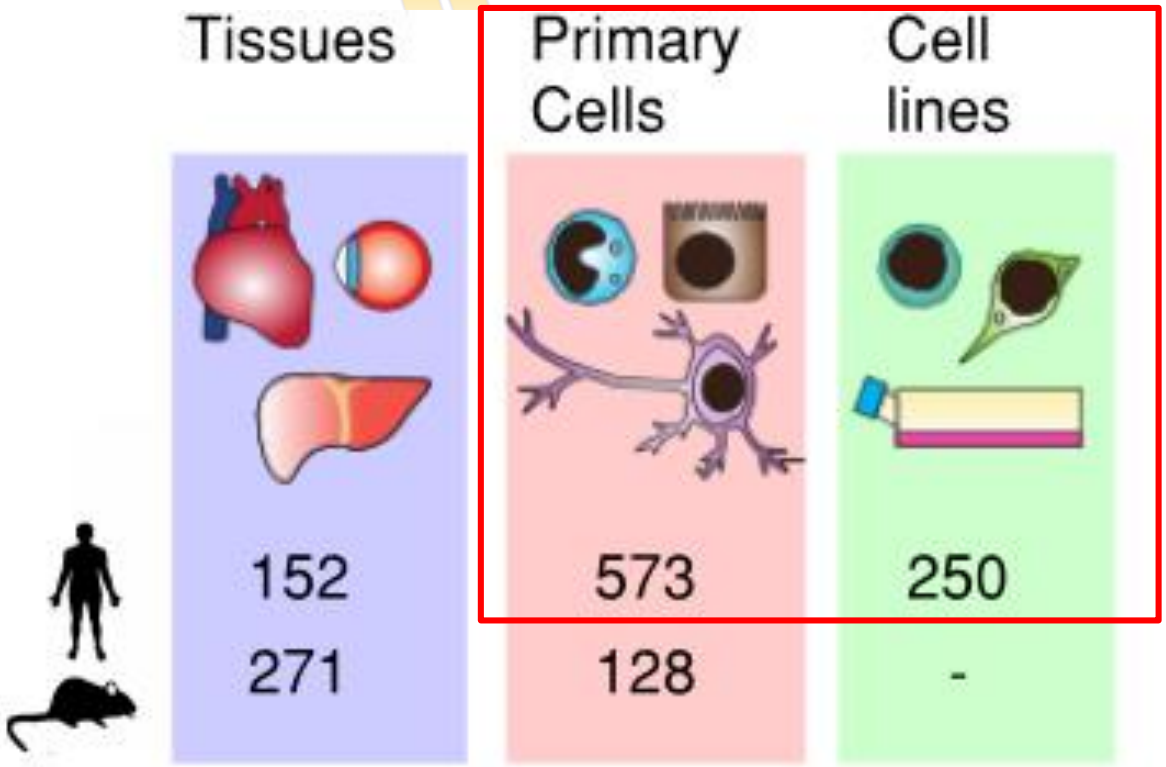
Nancy Yiu-Lin Yu et al. Nucl. Acids Res. 2015;nar.gkv608

CAGE Cap Analysis Gene Expression vs RNA-Sequencing a genome browser example



ZENBU genome browser
http://fantom.gsc.riken.jp/zenbu/

Severin, J. et al. (2014). Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nature Biotechnology*, *32*(3), 217–219.



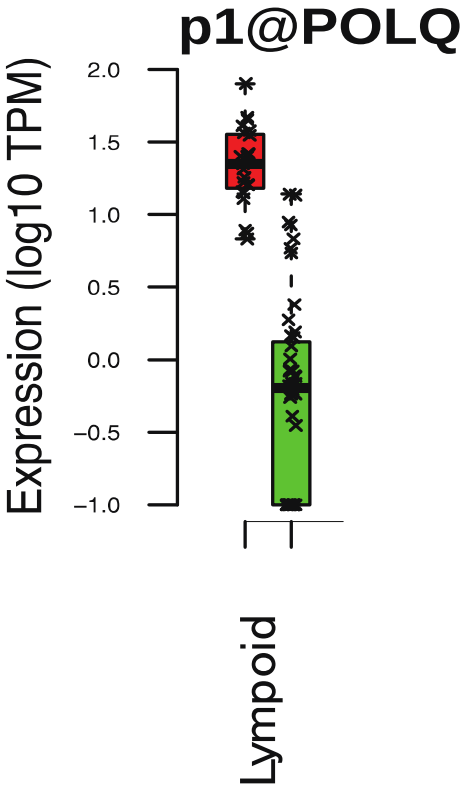
Cancer vs. Normal Analysis

Cap Analysis of Gene Expression

- **CAGE is 5' sequence tag technology that globally determines transcription start sites (TSS) in the genome and their expression levels.**
- **CAGE was applied to a broad selection of cancer cell lines and primary cells**

FANTOM Consortium and the RIKEN PMI and CLST (DGT). (2014).
A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470.

Finding differentially expressed genes in cancer

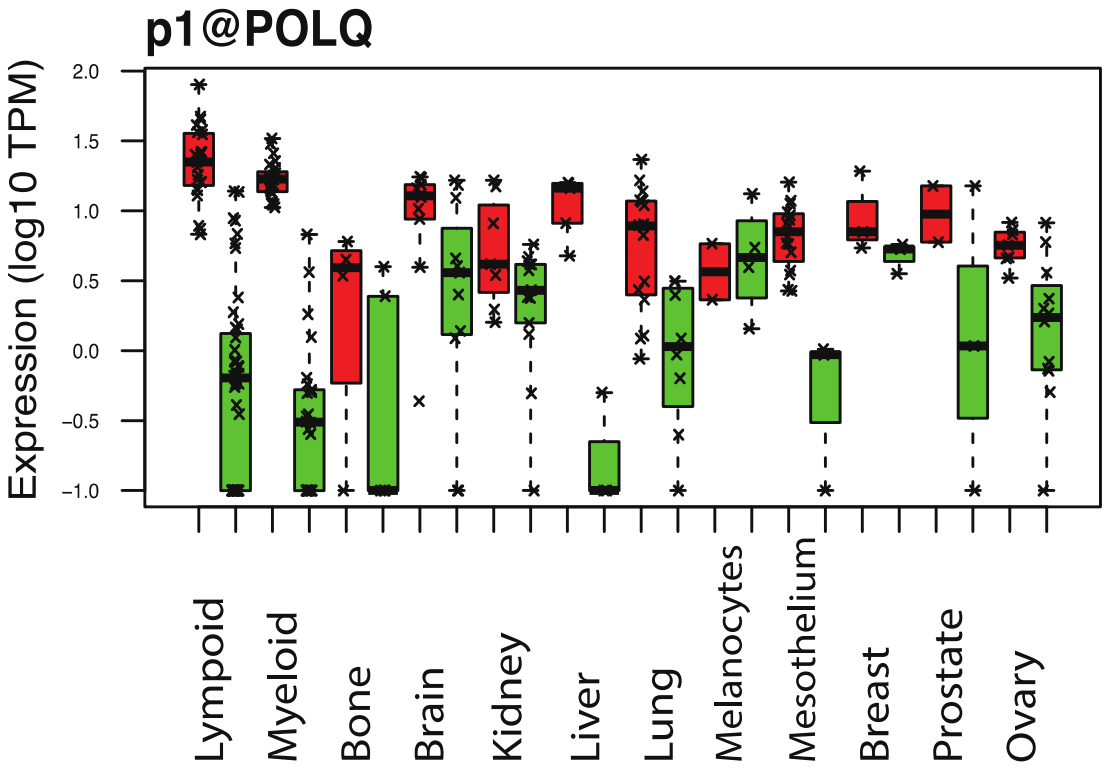


CANCER

NORMAL

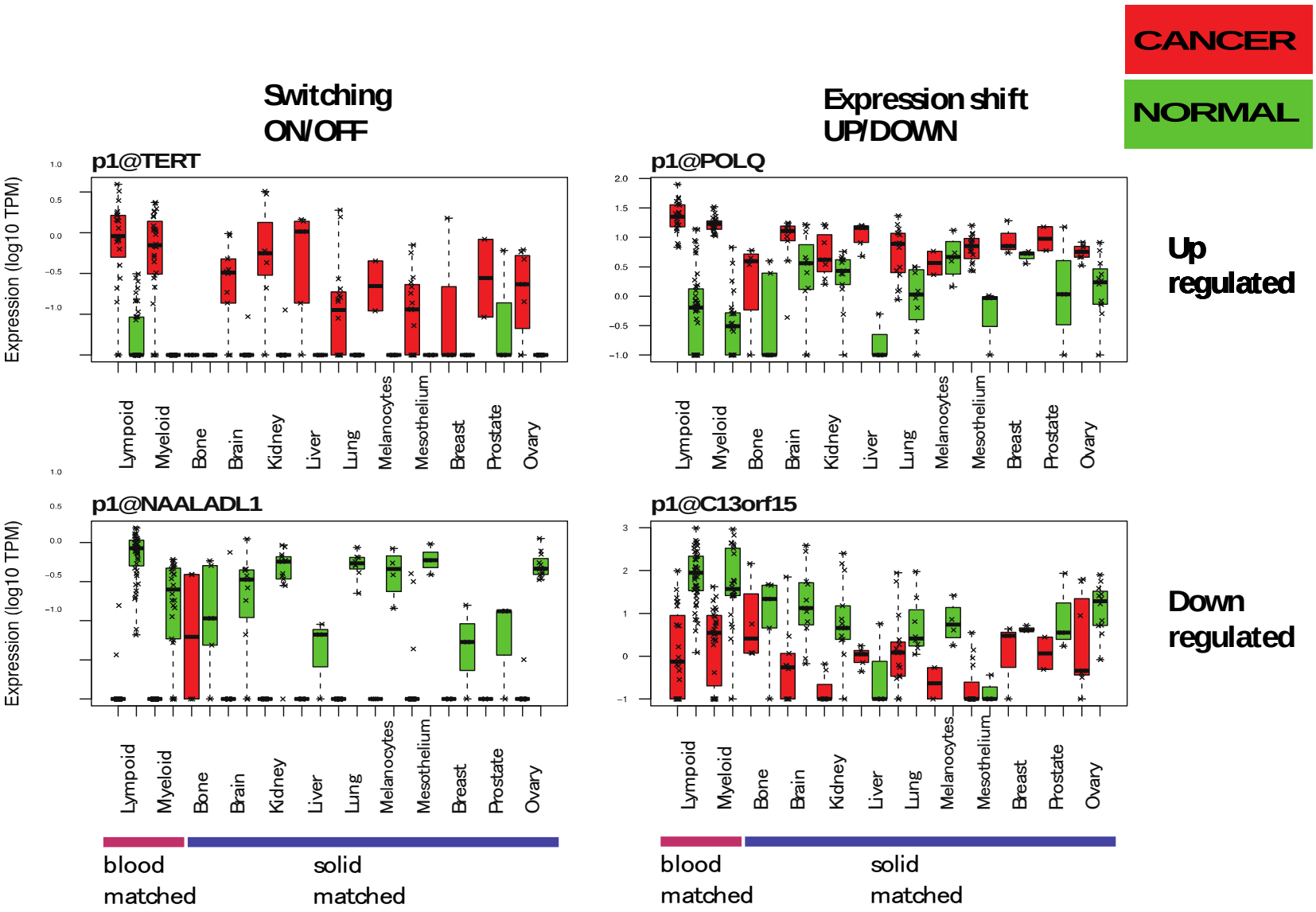
- We compare cancer cell lines to normal primary cells
- But we also want to compare matching cancer-normal pairs

Pan Cancer Differential Expression



- **Genewise Negative Binomial Generalized Linear Models** as implemented in **edgeR**
- Design matrix accounted for origin
- Contrast matrix to extract global cancer vs normal change
- FDR < 0.01
- Fold change > 2

Pan Cancer Differential Expression



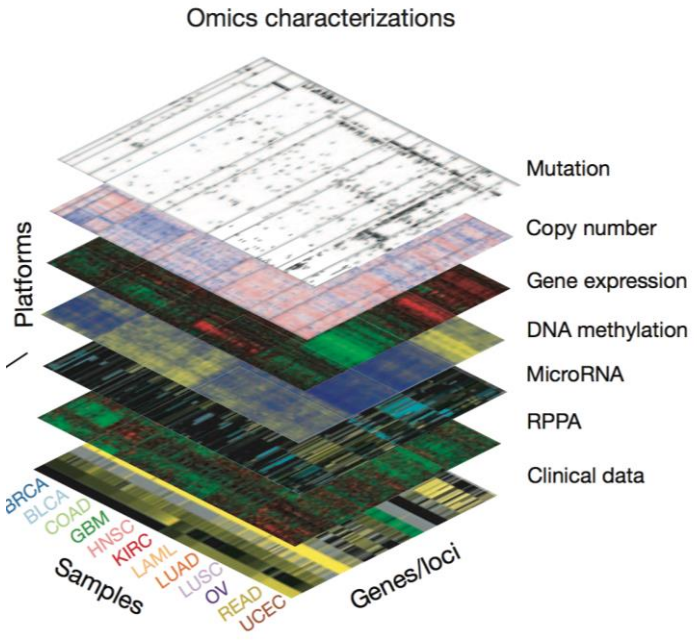
Annotation of differentially expressed promoters

	UP-regulated		DOWN-regulated	
	# promoter	# genes	# promoter	# genes
protein coding	575	430	371	295
protein coding gene body	347	234	77	19
lncRNAs	209	148	18	14
antisense	65	52	4	2
pseudogene	34	29	6	6
small ncRNAs	14	13	0	0
not_annotated	351		37	
Sum	1595	906	513	336

We work with cancer cell lines, are the results relevant to clinical tumors?

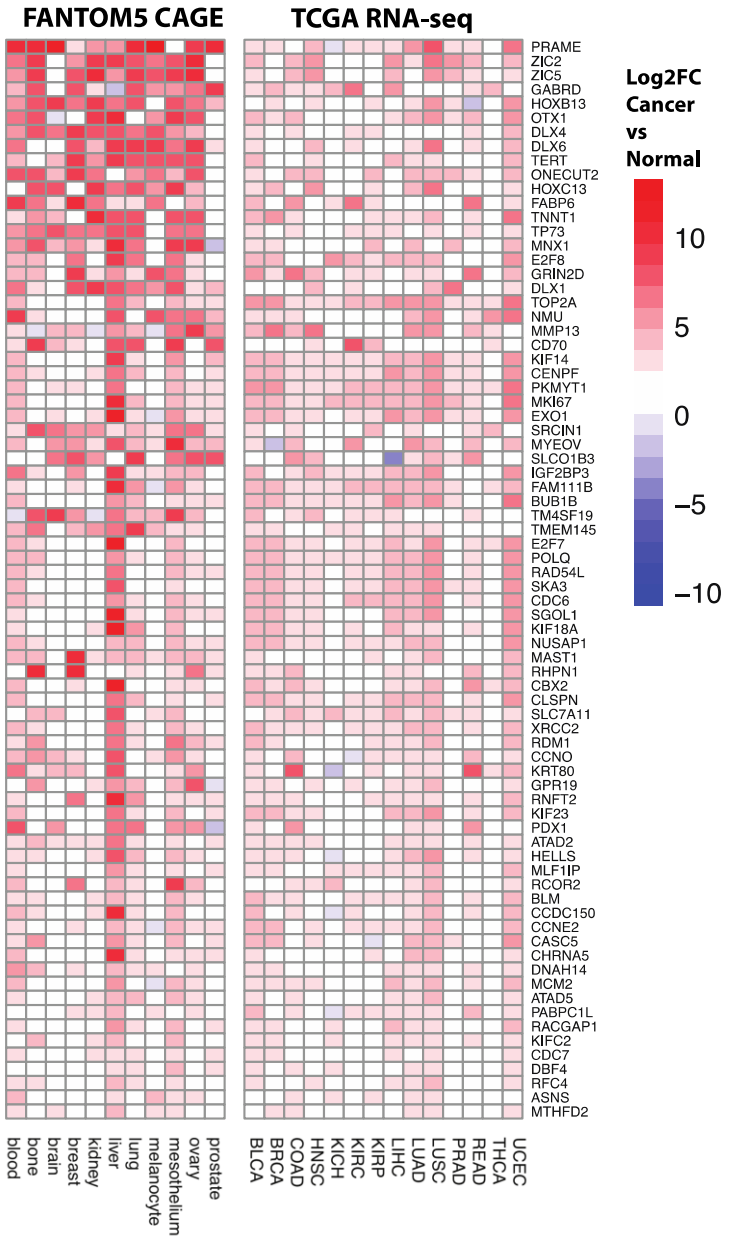
TCGA The Cancer Genome Atlas

- Available RNA-seq, gene expression profiles for:
  - ~ **5000 clinical tumor** samples
  - ~ 500 normal tissues
  - **14 cancer types**



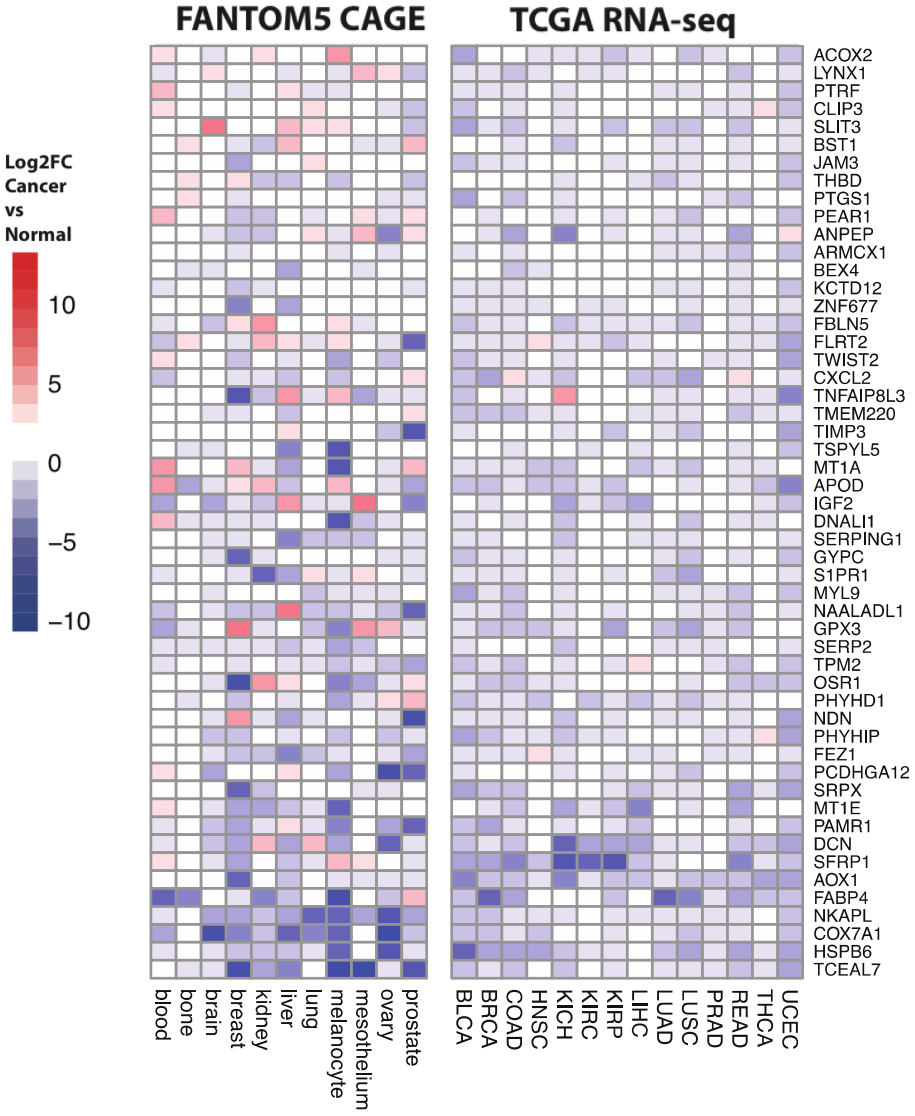
THE CANCER GENOME ATLAS

Figure 2



Up regulated in both F5 and TCGA

- 76 genes (~17%)
- enriched in genes involved in:
  - cell cycle
  - DNA metabolism,
  - biopolymer metabolism
  - homeobox genes (developmental)



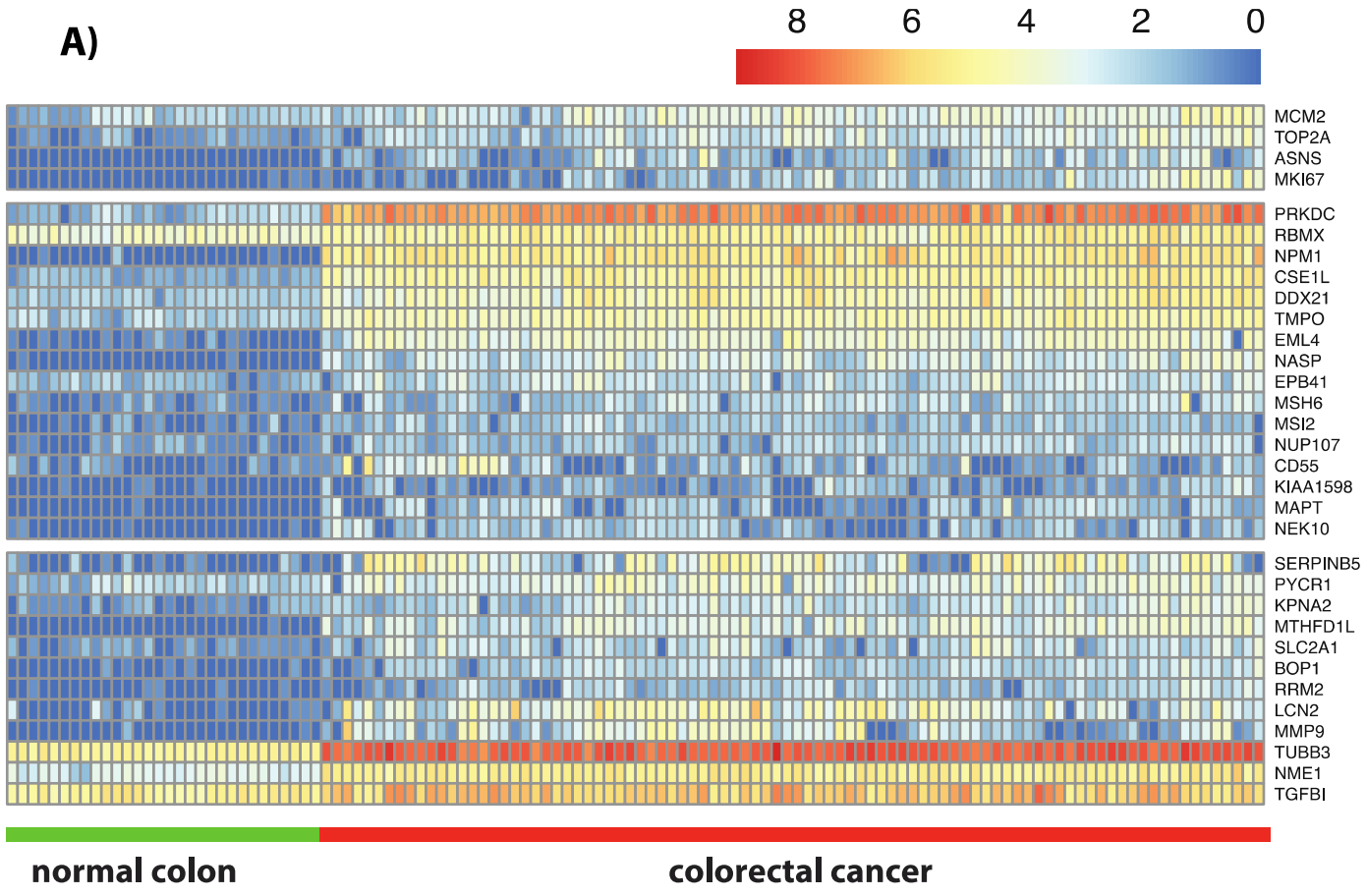
Down regulated in both F5 and TCGA

- 54 genes (~20%)
- enriched in genes involved in:
  - oxidoreductase activity

Do RNA translate into protein?

Are the biomarker candidates de-regulated
at protein level?

Protein level conformation of cancer up-regulated PC genes



Spectral count data from 90 colorectal cancers and 30 normal (mass spec from Zhang *et al.* 2015)

Long non coding RNAs?

Annotation of differentially expressed promoters

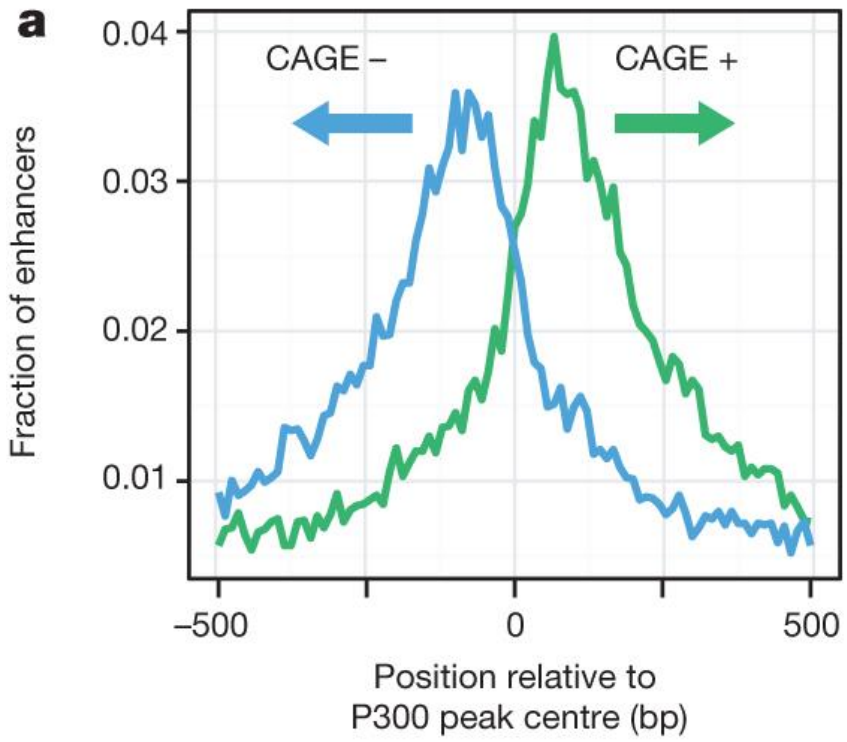
	UP-regulated		DOWN-regulated	
	# promoter	# genes	# promoter	# genes
protein coding	575	430	371	295
protein coding gene body	347	234	77	19
lncRNAs	209	148	18	14
antisense	65	52	4	2
pseudogene	34	29	6	6
small ncRNAs	14	13	0	0
not_annotated	351		37	
Sum	1595	906	513	336

lncRNA confirmed in TCGA tumors (25 up, 3 down)

func_name miTranscriptome	gene_name (GENCODE19)	lncRNA type	mean tpm	FANTOM5 Pan-Cancer DE	# of TCGA cancers UP (TCGA)	# of TCGA cancers DOWN (TCGA)
	PVT1	lncrna	0.407	UP	9	0
CAT647	CTD-2023M8.1	linc	1.046	UP	5	0
CAT122	**RP11-284F21.7****	antisense	3.201	UP	5	0
CAT2260	RP11-191L9.4	linc	0.404	UP	4	0
MNX1-AS1	MNX1-AS1	antisense	0.198	UP	4	0
CAT1138	G084254	linc	2.695	UP	3	0
CAT266	G044387	linc	0.103	UP	3	0
CAT62	RP4-792G4.2	antisense	0.178	UP	3	0
LINC00898	LINC00898	linc	0.104	UP	3	0
DGCR5	*DGCR5*	antisense	3.738	UP	3	0
CAT2039	AC009005.2	antisense	0.417	UP	3	0
CAT1022	G080198	linc	2.452	UP	3	0
CAT1167	**PCAT7**	antisense	0.086	UP	2	0
FEZF1-AS1	FEZF1-AS1	antisense	0.958	UP	2	0
CAT1572	RP11-438N16.1	linc	1.574	UP	2	0
DLX6-AS1	DLX6-AS2	antisense	0.259	UP	2	0
CAT1833	RP11-57A19.2	linc	0.528	UP	2	0
CAT615	G064032	linc	0.048	UP	2	0
LINC00669	LINC00669	linc	4.096	UP	2	0
CAT800	**RP11-328M4.2****	antisense	0.287	UP	2	0
MFI2-AS1	MFI2-AS1	antisense	1.301	UP	2	0
CAT219	AC074117.10	antisense	1.447	UP	2	0
	LINC00511	linc	0.256	UP	3	0
	AC006262.5	linc	0.223	UP	2	0
	RP11-435O5.2	linc	0.475	UP	2	0
CAT1235	RP11-124N14.3*	antisense	7.352	DOWN	2	-4
MEG3	***MEG3***	linc	182.23	DOWN	0	-2
	MT1L	pseudogene	116.644	DOWN	0	-9

Enhancer RNAs

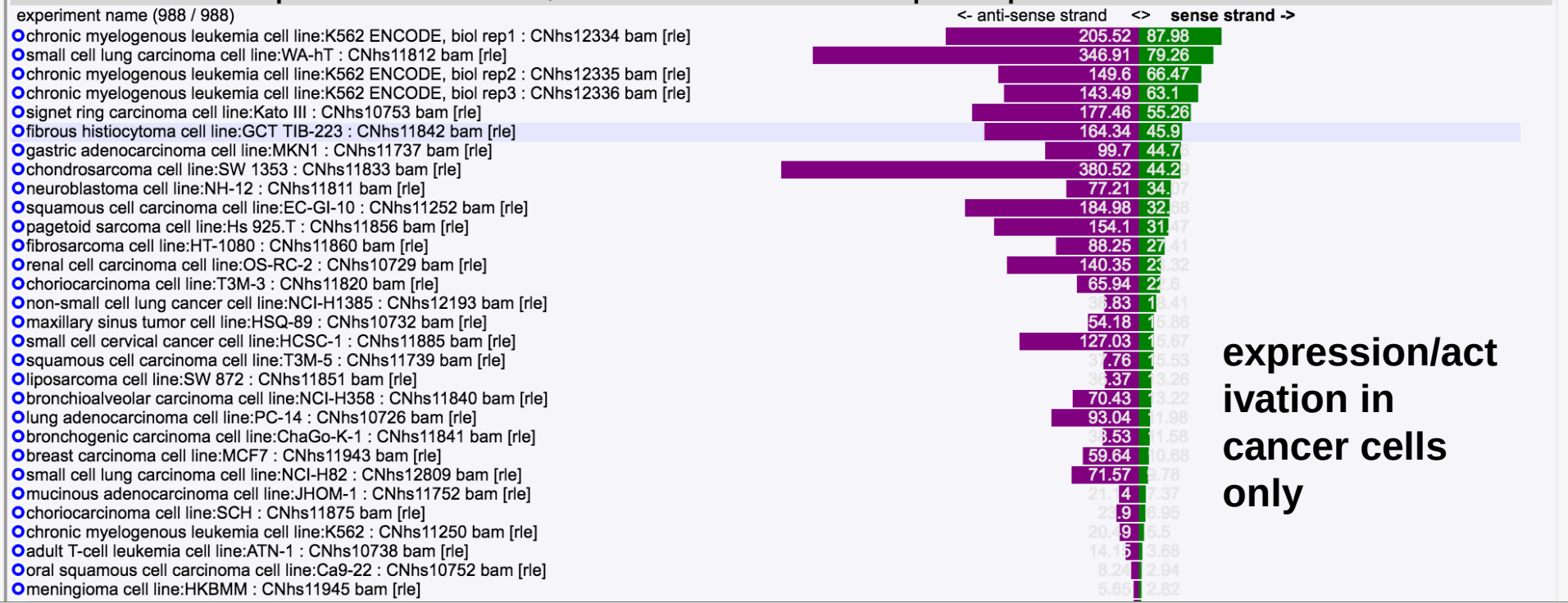
- CAGE data can be used to estimate the activity of enhancers
- balanced bidirectional capped transcription)



R Andersson *et al. Nature* **507**, 455-461 (2014) doi:10.1038/nature12787

Example of “cancer enhancer”

balanced bidirectional transcription



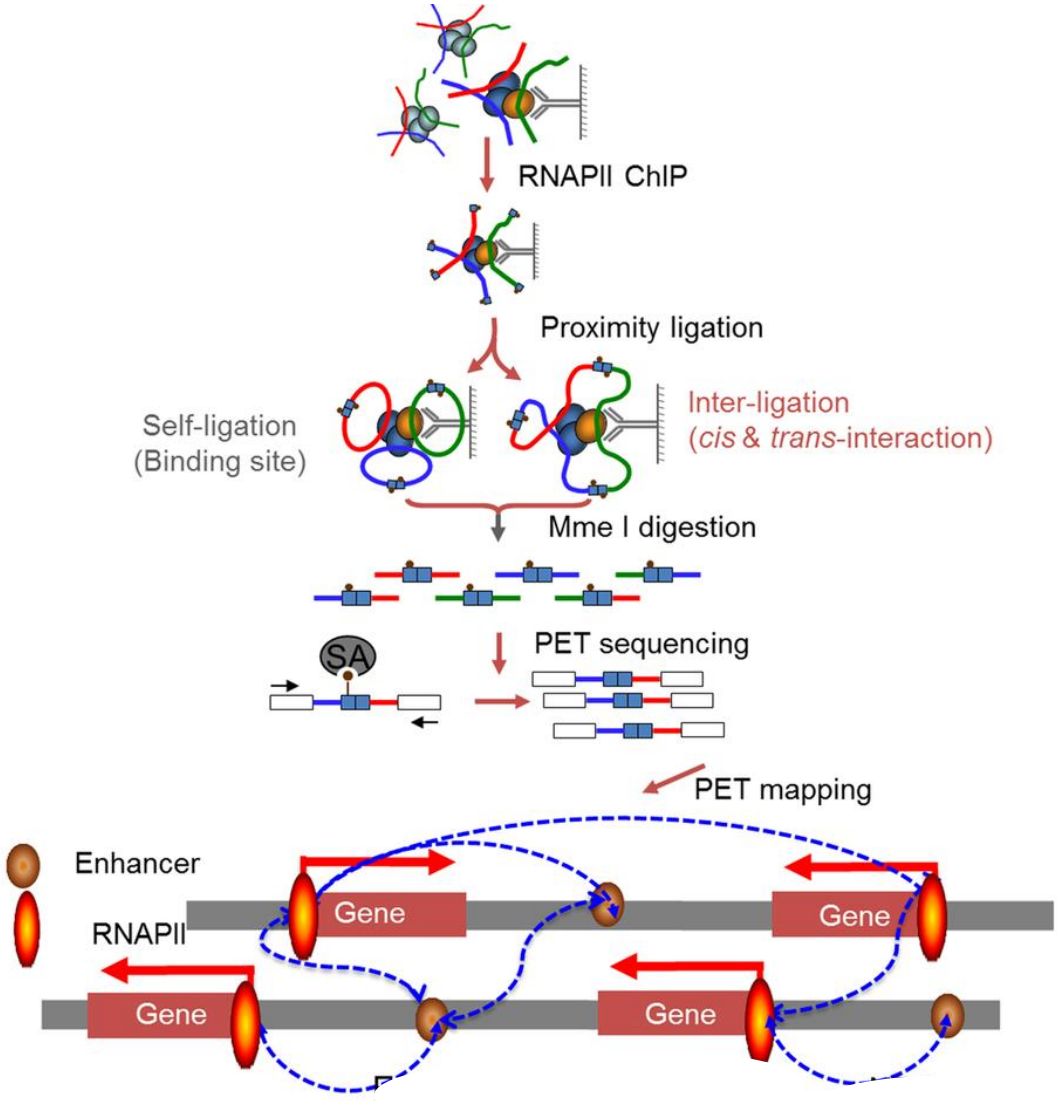
expression/activation in cancer cells only

Enhancer RNAs

- differential expression analysis on CAGE tags counts under 43,011 enhancers
- We discovered 90 enhancer up regulated in cancer

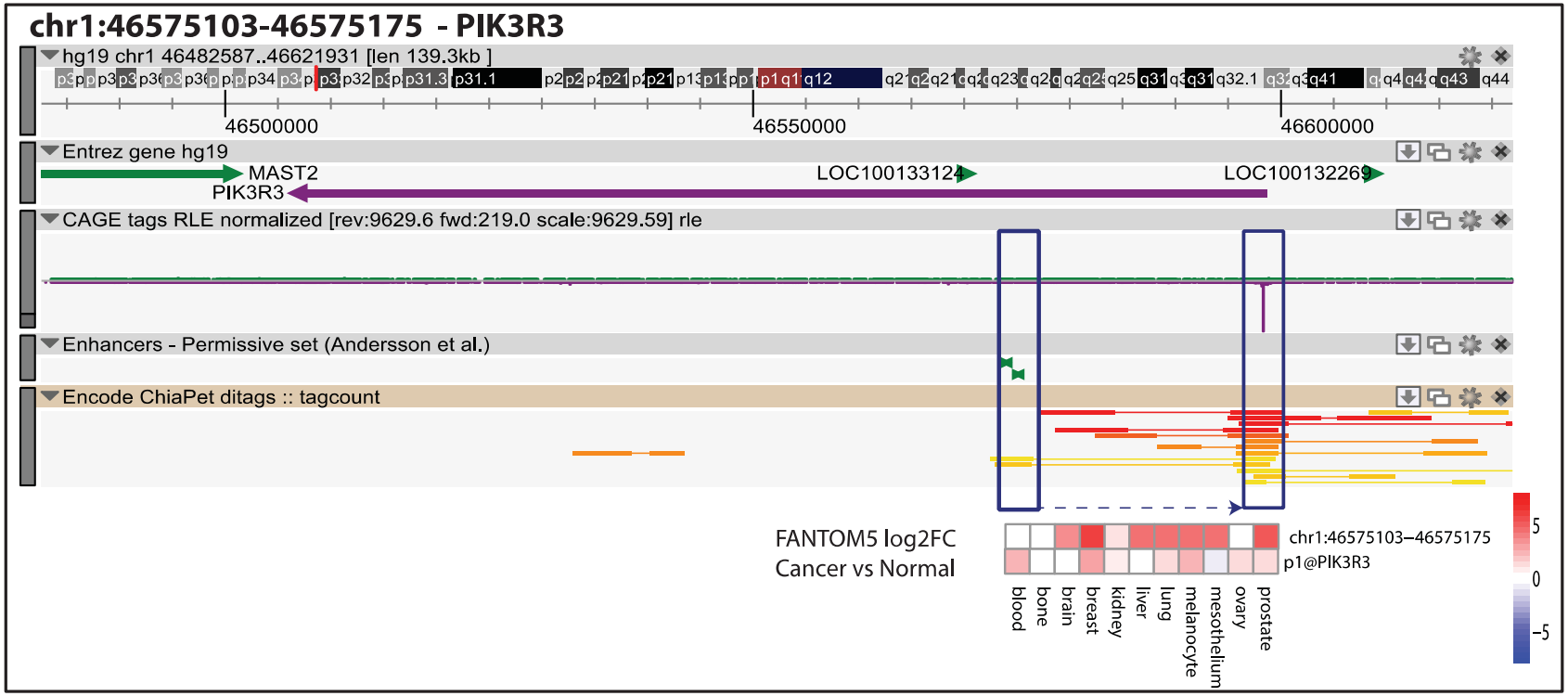
	Up-regulated		Down-regulated	
	Pan-Cancer	**Solid-Cancer**	**Pan-Cancer**	**Solid-Cancer**
Enhancers	28	62	0	0

Pan-cancer enhancers physically associated with genes implicated in cancer.



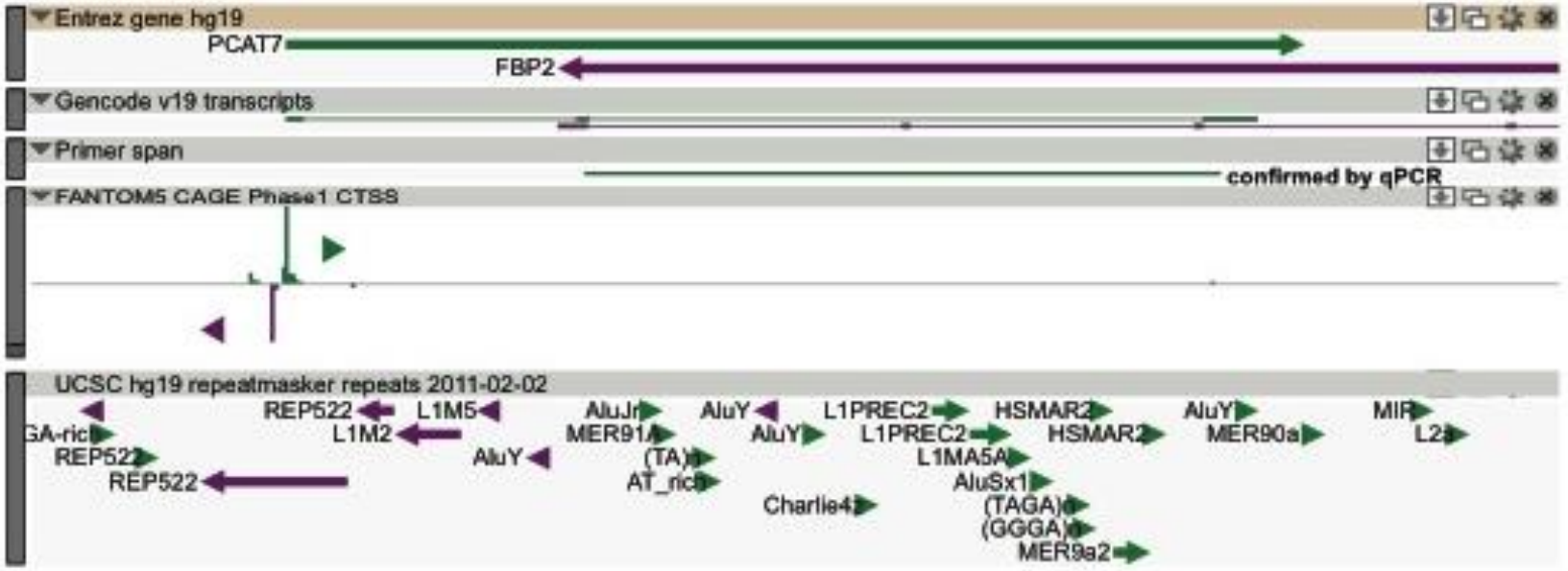
- Chia-PET data
- (Chromatin Interaction Analysis by Paired-End Tag Sequencing)
- from Encode project
- Physical chromatin associations of enhancers to promoters of genes related to cancer.

Pan-cancer enhancers physically associated with genes implicated in cancer.



- Enhancer at chr1:46575103-46575175 is shown to physically associate with he promoter for the PIK3R3 gene.
- PIK3R3- phosphoinositide-3-kinase, regulatory subunit 3
- PIK3R3 was reported increase proliferation in colorectal, lung, gastric cancer, leukemia and glioma.

CAGE and repetitive elements



CAGE let's us see if promoter overlaps repeat elements

Kaczkowski et al. (2016). *Cancer Research*, *76*(2), 216–226.

Promoters overlapping repetitive elements.

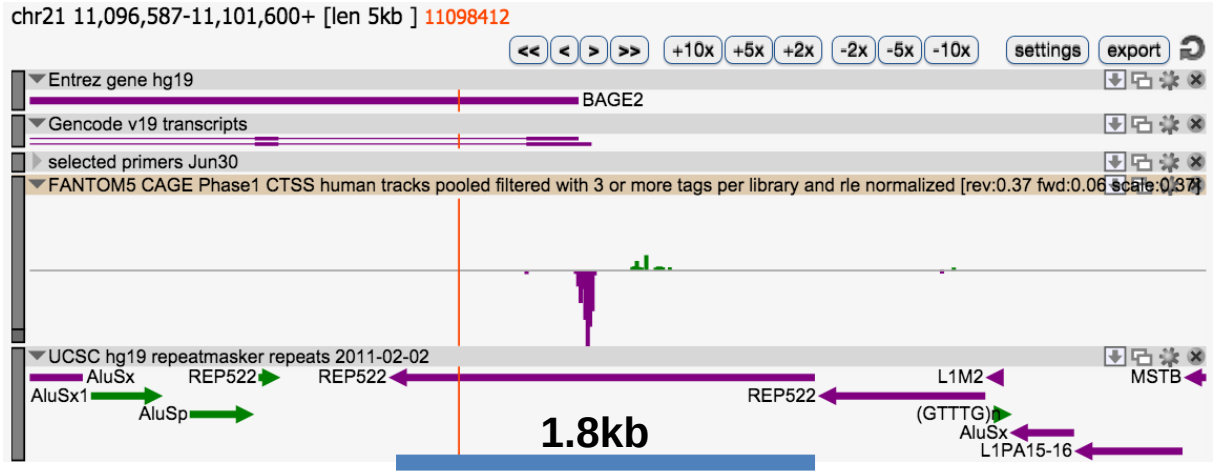
	# all overlapping promoters	Up-regulated			Protein coding						
		# promoters	Odd ratio	p-value	5' transcript	intronic	exon	3UTR	lncRNAs	Pseudogene	non annotated
REP522	72	**25**	**62.05**	**2.20E-16**	1	0	0	0	**9**	3	12
SINE/Alu	3961	**138**	**4.44**	**2.20E-16**	5	**67**	1	1	**11**	3	**50**
LTR/ERV1	3932	**133**	**4.30**	**2.20E-16**	7	**12**	0	0	**31**	2	**83**
LINE/L1	3426	**67**	**2.35**	**1.75E-09**	**2**	**22**	0	0	12	0	32
LTR/ERVL	1488	20	1.57	0.049	2	2	0	0	8	0	8
LINE/L2	3220	25	0.90	0.70	2	**4**	0	0	**4**	0	**17**
LTR/ERVL-MaLR	3613	31	0.99	1	6	**4**	0	0	**10**	0	**11**
Simple_repeat	11982	204	2.13	2.20E-16	86	70	4	7	17	1	63
Low_complexity	2013	18	1.04	0.81	**15**	2	2	6	2	0	4

Gencode v19 annotation

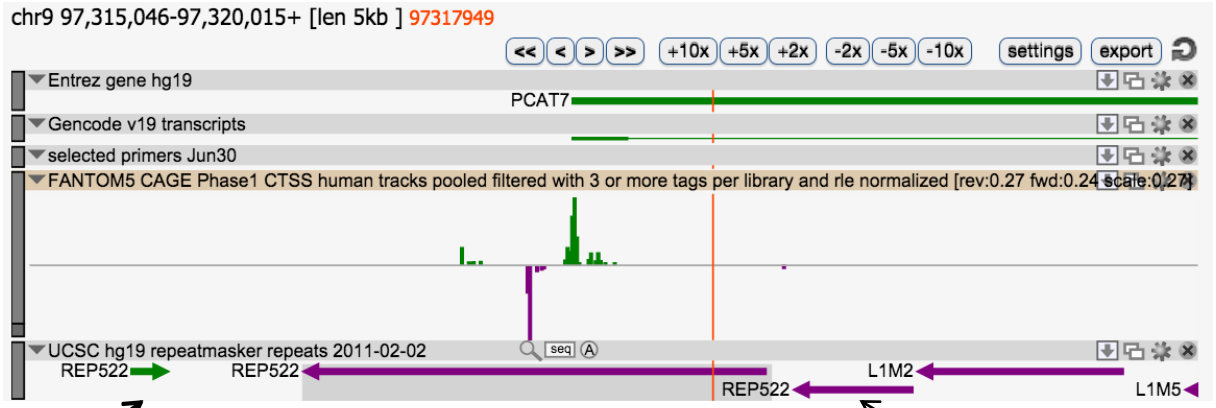
Kaczkowski et al. (2016). *Cancer Research*, *76*(2), 216–226.

REP522 story

BAGE2 - B melanoma antigen family, member 2



PCAT7 - prostate cancer associated transcript 7 (non-protein coding)



REP522 without promoter

REP522 with promoter

REP522 without promoter

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

Summary

- Pan cancer biomarkers
  - protein coding
  - lncRNAs
  - enhancers
- Activation of transcription from repeats in cancer
- Specific activation of REP522 element

FANTOM5
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

Acknowledgements

Pan cancer paper
Yuji Tanaka
Hideya Kawaji
Albin Sandelin
Robin Andersson
Masayoshi Itoh
Timo Lassmann
Yoshihide Hayashizaki
Piero Carninci
Alistair Forrest

Chung-Chau Hon
Michiel de Hoon
Jessica Severin
Giovanni Pascarella
Kosuke Hashimoto
Jordan Ramilowski
Erik Arner

LINC02021 KD
Jay Shin
Yuji Tanaka
Jasmine Ooi
Yuri Ishizu
Alistair Forrest
Piero Carninci

FANTOM5 Consortium

Funding:

1. **Japan Society for the Promotion of Science (JSPS Fellowship)**
2. **Foreign Postdoctoral Researcher program**
3. Research Grant for RIKEN Omics Science Center from MEXT to YH.
4. Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to YH.
5. Research Grant from MEXT to the RIKEN Center for Life Science Technologies

All DGT members for valuable discussions

RIKEN's Programs for Junior Scientists

Research Article

Cancer Research

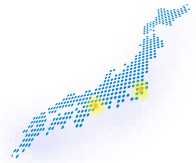
Transcriptome Analysis of Recurrently Deregulated Genes Across Multiple Cancers Identifies New Pan-Cancer Biomarkers

Bogumil Kaczkowski[1], Yuji Tanaka[1,2], Hideya Kawaji[1,2,3], Albin Sandelin[4], Robin Andersson[4], Masayoshi Itoh[1,3], Timo Lassmann[1,5], the FANTOM5 consortium, Yoshihide Hayashizaki[3], Piero Carninci[1], and Alistair R.R. Forrest[1,6]

Abstract

Genes that are commonly deregulated in cancer are clinically attractive as candidate pan-diagnostic markers and therapeutic targets. To globally identify such targets, we compared Cap Analysis of Gene Expression (CAGE) profiles from 225 different cancer cell lines and 339 corresponding primary cell samples to identify transcripts that are deregulated recurrently in a broad range of cancer types. Comparing RNA-seq data from 4,055 tumors and 563 normal tissues profiled in the The Cancer Genome Atlas and FANTOM5 datasets, we identified a core transcript set with theranostic potential. Our analyses also revealed enhancer RNAs which are upregulated in cancer, defining promoters which overlap with repetitive elements (especially SINE/Alu and LTR/ERV1 elements) that are often upregulated in cancer. Lastly, we documented for the first time upregulation of multiple copies of the REP522 interspersed repeat in cancer. Overall, our genome-wide expression profiling approach identified a comprehensive set of candidate biomarkers with pan-cancer potential, and extended the perspective and pathogenic significance of repetitive elements which are frequently activated during cancer progression. *Cancer Res; 1–11. ©2015 AACR.*

Kaczkowski, B., et al. (2015). *Cancer Research*, ©2015 AACR



Thank you for your attention!

REP522 -- largely palindromic, unclassified interspersed repeat

Description

REP522 was originally called a telomeric satellite, even though in human, this 1.8Kb sequence is not found in tandem arrays, nor is it particularly restricted to telomeric regions. The central 1/3rd forms a >600bp imperfect palindrome. Other parts of these element display similarity to both LINE and LTR retrotransposons; to cut-and-paste DNA transposons; even to a Helitron! It is unclear how much of this is chimeric homology vs false positive.

The model is 1817 long.
The average hit is 431.