

Bioinformatic tool to integrate and understand
aberrant epigenomic and genomic changes
associated with cancer

Methods, development and analysis

Tiago Chedraoui Silva

University of São Paulo (USP)

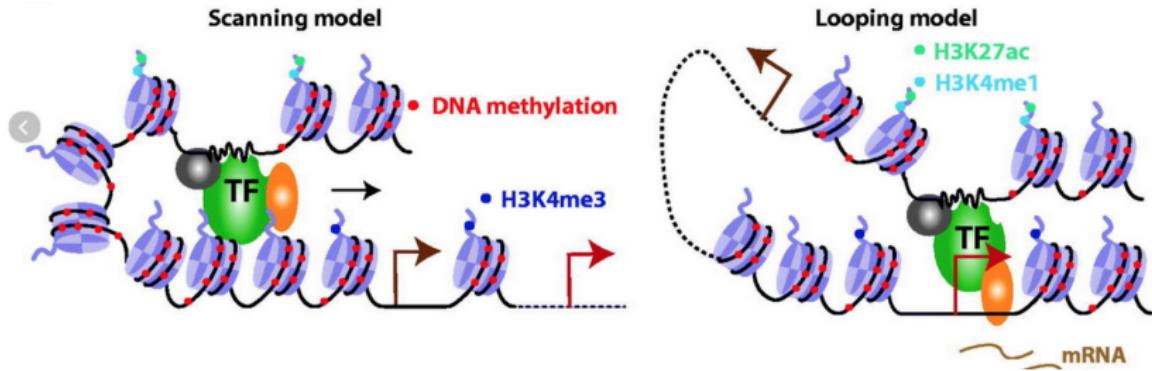
2017-12-22



Overview

- 1 Introduction
 - 2 ELMER
 - 3 Comparison between versions
 - 4 Case of study - TCGA BRCA

Enhancer-mediated gene regulation



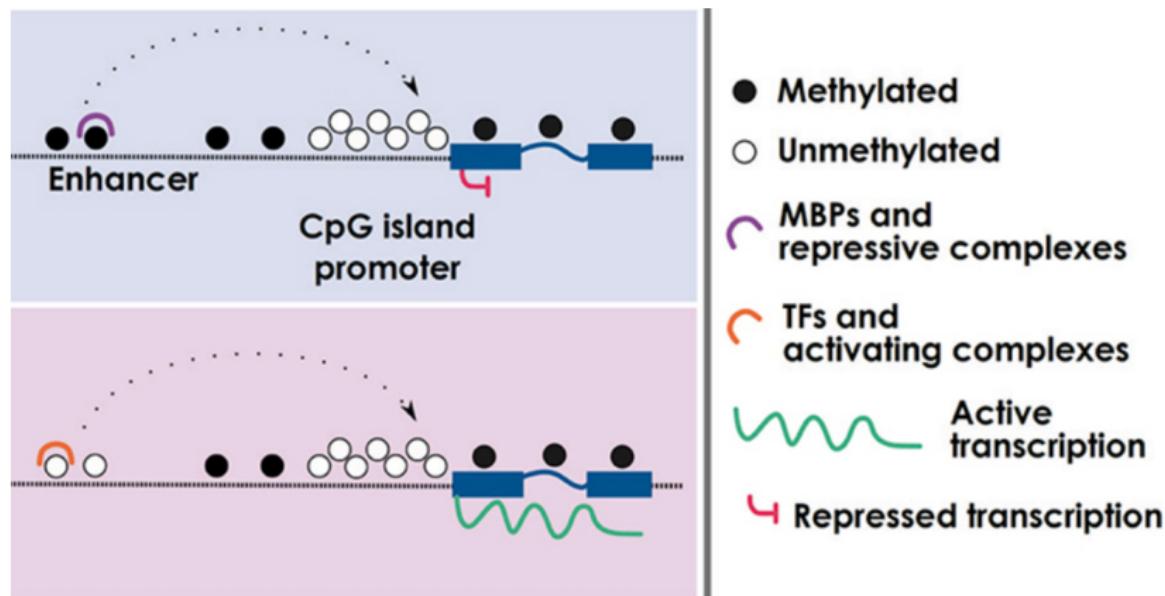
Yao et al. Genome Biology (2015) 16:105.



Enhancer-mediated gene regulation

- 73% of the tested distal elements do not link to the nearest gene (Sanyal et al., 2012)
 - 40% of the enhancers involved in loops do not interact with the TSS of the nearest gene (Li et al., 2012),
 - one-third of the distal interactions were not directed to the promoter of the nearest gene (Mifsud et al., 2015),
 - 85% of tumor-specific enhancers that could be linked to the expression of a nearby gene skipped the nearest gene (Yao et al., 2015).

Enhancer-mediated gene regulation



Source: Front. Aging Neurosci., 05 March 2015 <http://dx.doi.org/10.3389/fnagi.2015.00019>.

Enhancer Linking by Methylation/Expression Relationship

ELMER



Yao et al. *Genome Biology* (2015) 16:105
DOI 10.1186/s13059-015-0668-3



METHOD

Open Access

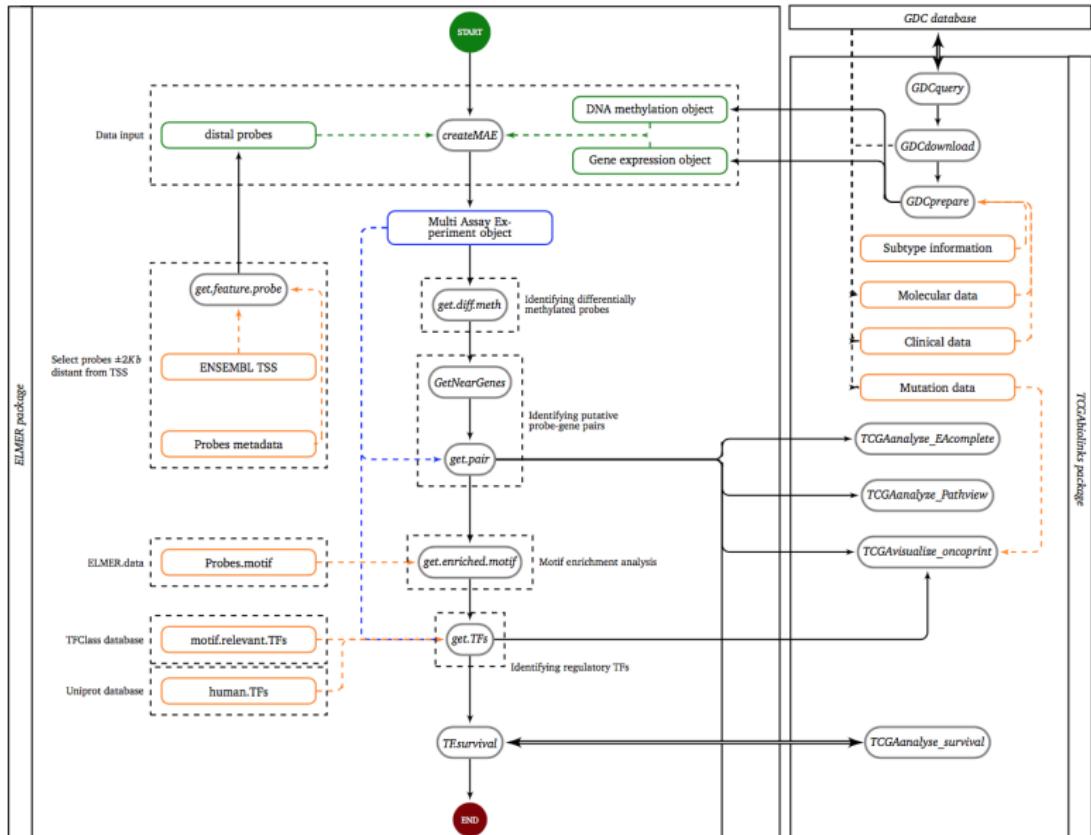
Inferring regulatory element landscapes and transcription factor networks from cancer methylomes



Lijing Yao¹, Hui Shen², Peter W Laird², Peggy J Farnham^{1*} and Benjamin P Berman^{1,3*}

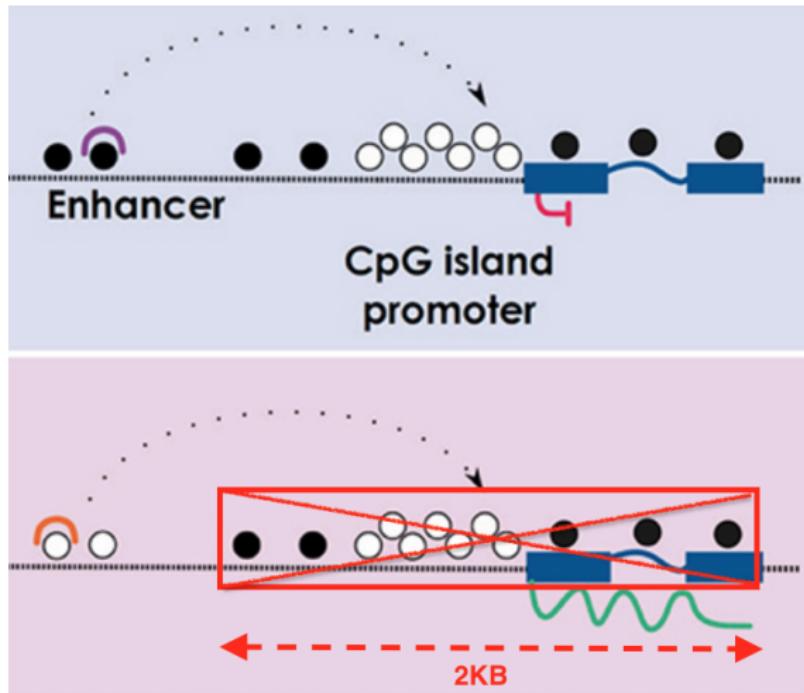
Abstract

Recent studies indicate that DNA methylation can be used to identify transcriptional enhancers, but no systematic approach has been developed for genome-wide identification and analysis of enhancers based on DNA methylation. We describe ELMER (Enhancer Linking by Methylation/Expression Relationships), an R-based tool that uses DNA methylation to identify enhancers and correlates enhancer state with expression of nearby genes to identify transcriptional targets. Transcription factor motif analysis of enhancers is coupled with expression analysis of transcription factors to infer upstream regulators. Using ELMER, we investigated more than 2,000 tumor samples from The Cancer Genome Atlas. We identified networks regulated by known cancer drivers such as GATA3 and FOXA1 (breast cancer), SOX17 and FOXA2 (endometrial cancer), and NFE2L2, SOX2, and TP63 (squamous cell lung cancer). We also identified novel networks with prognostic associations, including RUNX1 in kidney cancer. We propose ELMER as a powerful new paradigm for understanding the *cis*-regulatory interface between cancer-associated transcription factors and their functional target genes.

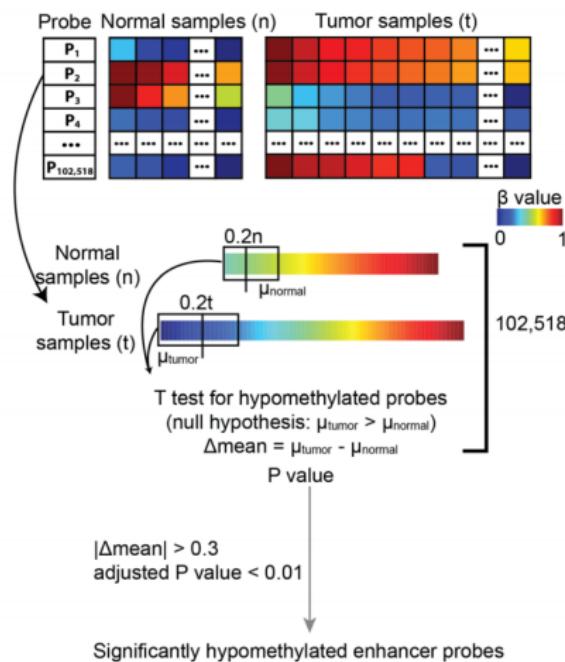


StateHub/StatePaintR/juncivar package

Step 1: Identify distal probes



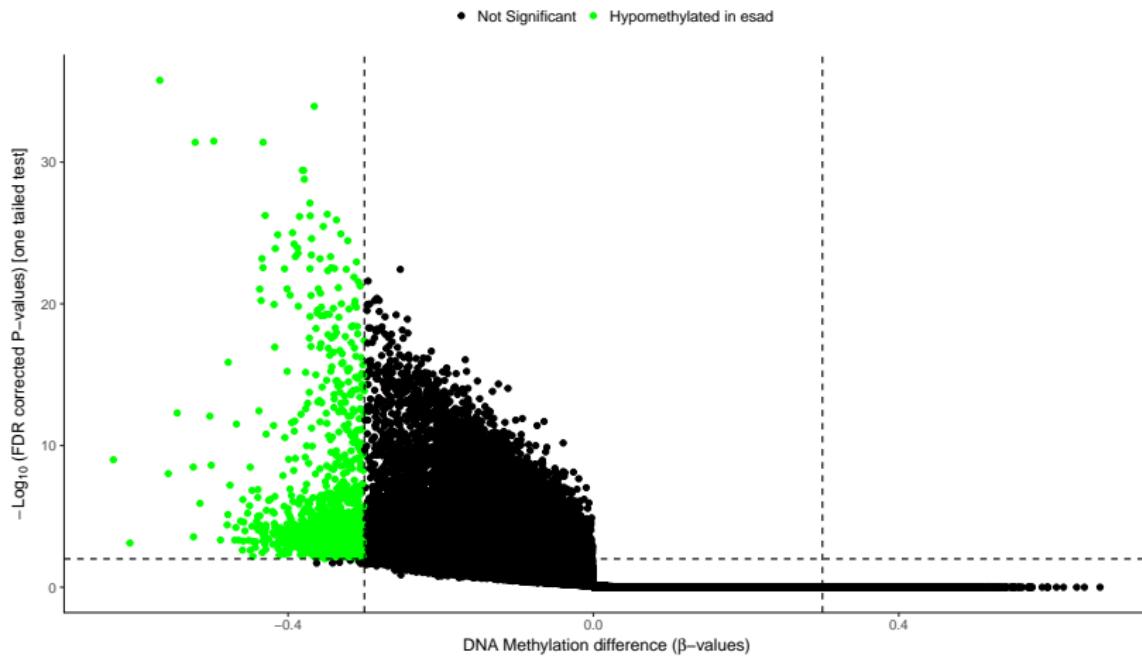
Step 2: Differentially methylated distal probes



Yao et al. Genome Biology (2015) 16:105.

Step 2: Differentially methylated distal probes

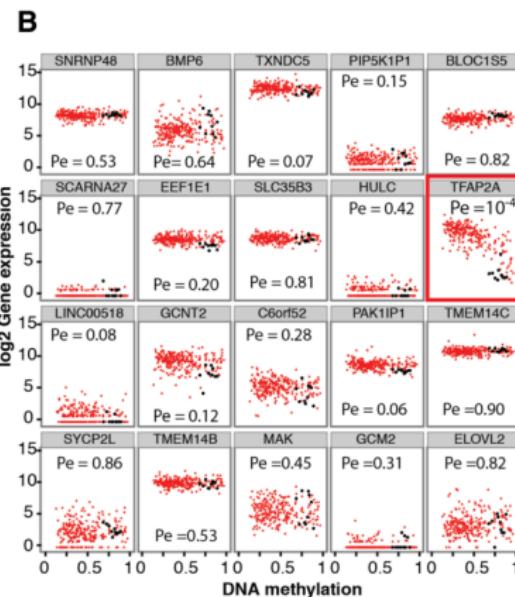
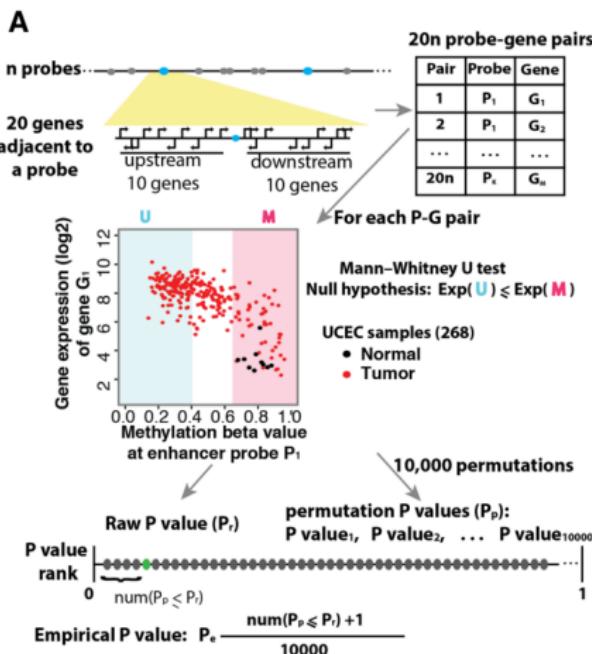
Volcano plot – Probes hypomethylated in esad vs normal



Step 2: Differentially methylated distal probes

probe <chr>	pvalue <dbl>	Cancer_Minus_normal <dbl>	adjust.p <dbl>
cg00019654	8.786907e-05	-0.3171058	8.997079e-04
cg00021978	3.615952e-06	-0.3489671	6.730660e-05
cg00029246	5.053095e-09	-0.3484794	2.415897e-07
cg00069391	3.021990e-06	-0.3026002	5.795861e-05
cg00075975	3.506343e-13	-0.3125811	5.634483e-11
cg00086488	3.139093e-04	-0.4000719	2.416091e-03
cg00094303	7.767398e-07	-0.3139641	1.846501e-05
cg00098799	1.117552e-08	-0.3239755	4.803548e-07
cg00102093	2.417422e-16	-0.3247405	8.931561e-14
cg00108282	2.247518e-06	-0.3832163	4.531800e-05
cg00125601	1.089598e-03	-0.3014288	6.283534e-03
cg00135293	2.310803e-24	-0.4352582	5.976346e-21
cg00146437	9.719259e-08	-0.3541456	3.088890e-06
cg00147248	3.297106e-06	-0.3144497	6.236248e-05
cg00152878	1.698146e-14	-0.4189015	3.873016e-12

Step 3: Get pairs



Yao et al. Genome Biology (2015) 16:105.

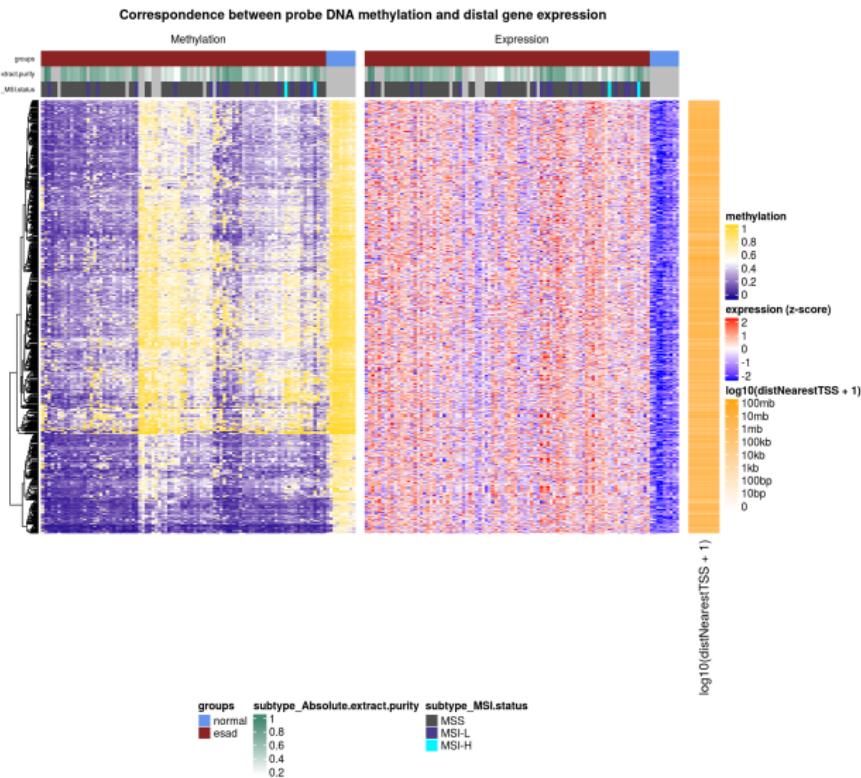
Step 3: Pairs inferred

Probe <chr>	GenID <chr>	Symbol <chr>	Sides <chr>	Raw.p <dbl>	Pe <dbl>	distNearestTSS <int>
cg00002749	ENSG00000068489	PRR11	L7	8.808330e-07	9.999e-05	564624
cg00002749	ENSG00000108423	TUBD1	R2	2.674608e-04	9.999e-05	116585
cg00002749	ENSG00000141378	PTRH2	L1	5.337531e-05	9.999e-05	54550
cg00002749	ENSG00000182628	SKA2	L8	5.978567e-06	9.999e-05	606907
cg00013804	ENSG00000119203	CPSF3	R3	2.039634e-06	9.999e-05	205119
cg00013804	ENSG00000151694	ADAM17	R5	1.307986e-06	9.999e-05	337326
cg00042263	ENSG00000088205	DDX18	R1	2.476222e-05	9.999e-05	68829
cg00042263	ENSG00000115091	ACTR3	L2	4.416642e-04	9.999e-05	3794995
cg00042263	ENSG00000163064	EN1	R4	4.948909e-06	9.999e-05	1098532
cg00050692	ENSG00000138018	EPT1	R9	2.674608e-04	9.999e-05	1006537

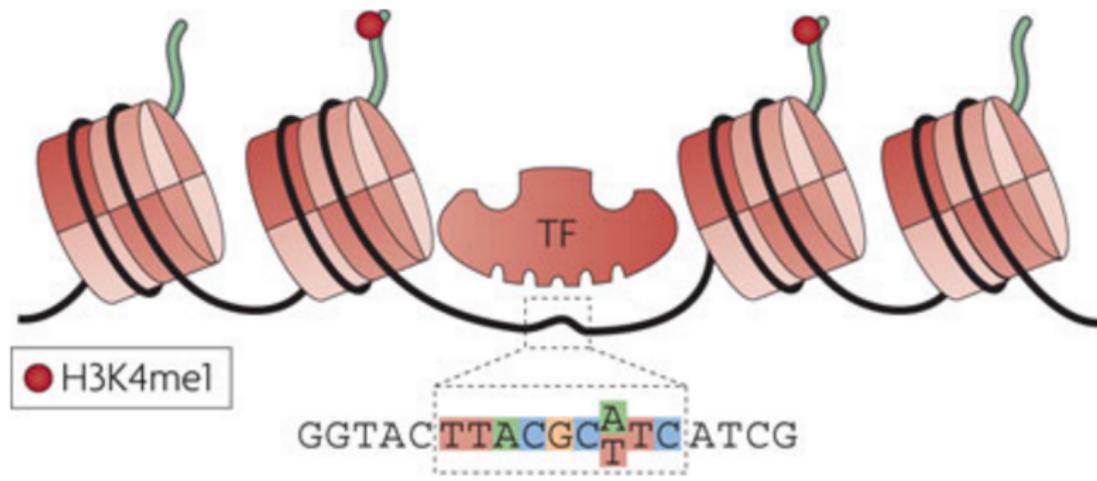
1-10 of 2,106 rows

Previous [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) ... [211](#) [Next](#)

Step 3: Pairs inferred



Step 4: Enriched motifs



Nature Reviews | Genetics

Next-generation genomics: an integrative approach R. David Hawkins et al. Nature Reviews Genetics 11, 476-486

Step 4: Enriched motifs

HOCOMOCO Home Human TFs ▾ Mouse TFs ▾ Downloads ▾ Help Search:

Reset Select Columns Get CSV HUMAN_mono_motifs.tsv PWMS for HUMAN transcription factors

Model	LOGO	Transcription factor	TF family	TF subfamily
AP2C_HUMAN.H10MO.A		TFAP2C (GeneCards)	AP-2[1.3.1]	AP-2gamma[1.3.1.0.3]
BATF_HUMAN.H10MO.A		BATF (GeneCards)	B-ATF-related factors[1.1.4]	B-ATF[1.1.4.0.1]
MAX_HUMAN.H10MO.A		MAX (GeneCards)	bHLH-ZIP factors[1.2.6]	Myc / Max factors[1.2.6.5]
MYC_HUMAN.H10MO.A		MYC (GeneCards)	bHLH-ZIP factors[1.2.6]	Myc / Max factors[1.2.6.5]
USF1_HUMAN.H10MO.A		USF1 (GeneCards)	bHLH-ZIP factors[1.2.6]	USF factors[1.2.6.2]
USF2_HUMAN.H10MO.A		USF2 (GeneCards)	bHLH-ZIP factors[1.2.6]	USF factors[1.2.6.2]
CEBPA_HUMAN.H10MO.A		CEBPA (GeneCards)	C/EBP-related[1.1.8]	C/EBP[1.1.8.1]
CEBPB_HUMAN.H10MO.A		CEBPB (GeneCards)	C/EBP-related[1.1.8]	C/EBP[1.1.8.1]
CEBPE_HUMAN.H10MO.A		CEBPE (GeneCards)	C/EBP-related[1.1.8]	C/EBP[1.1.8.1]
CREB1_HUMAN.H10MO.A		CREB1 (GeneCards)	CREB-related factors[1.1.7]	CREB-like factors[1.1.7.1]
E2F1_HUMAN.H10MO.A		E2F1 (GeneCards)	E2F-related factors[3.3.2]	E2F[3.3.2.1]

Step 4: Motif Enrichment

Objective

Evaluate the enrichment of transcription factors in certain genomic regions.

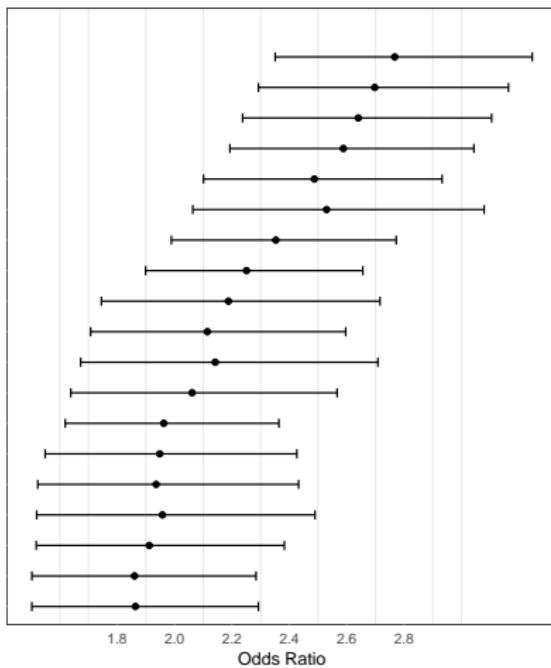
- 1 Perform motif matching of transcription factors in probes regions (window $\pm 250\text{bp}$)
- 2 Evaluate which transcription factors are more likely to occur in those regions than in background regions using Fisher's exact test with FDR correction.

Fisher's exact test

- a: nb of input regions with match for TF motif.
- b: nb of input regions with no match for TF motif.
- c: nb of background regions with match for TF motif.
- d: nb of background regions with no match for TF motif.

Step 4: Motif enrichment

Motif	Odds ratio (95% CI)	# probes (% of paired)
FOSL2	2.77 (2.35–3.25)	201 (0.23%)
FOSL1	2.7 (2.29–3.16)	202 (0.23%)
FOSB	2.64 (2.24–3.1)	193 (0.22%)
FOS	2.59 (2.19–3.04)	193 (0.22%)
JUN	2.49 (2.1–2.93)	184 (0.21%)
BATF	2.53 (2.06–3.08)	118 (0.13%)
JUND	2.35 (1.99–2.77)	186 (0.21%)
JUNB	2.25 (1.9–2.66)	181 (0.2%)
HXB13	2.19 (1.74–2.72)	94 (0.11%)
PIT1	2.11 (1.71–2.6)	106 (0.12%)
PRRX1	2.14 (1.67–2.71)	78 (0.09%)
CDX1	2.06 (1.64–2.57)	91 (0.1%)
LMX1A	1.96 (1.62–2.36)	134 (0.15%)
BATF	1.95 (1.55–2.43)	91 (0.1%)
NKX32	1.94 (1.52–2.43)	83 (0.09%)
HME1	1.96 (1.52–2.49)	74 (0.08%)
IRX3	1.91 (1.52–2.38)	90 (0.1%)
PO4F3	1.86 (1.5–2.28)	106 (0.12%)
PO4F1	1.86 (1.5–2.29)	104 (0.12%)



Step 4: Motif enrichment

motif	NumOfProbes	PercentageOfProbes	lowerOR	upperOR	OR	p.value	FDR	TF family	TF subfamily
FOSL2_HUMAN.H11MO.0.A	201	0.226	2.35	3.25	2.77	2.63e-30	1.45e-28	Fos-related factors{1.1.2}	Fos factors{1.1.2.1}
FOSL1_HUMAN.H11MO.0.A	202	0.227	2.29	3.16	2.70	3.73e-29	1.92e-27	Fos-related factors{1.1.2}	Fos factors{1.1.2.1}
POSB_HUMAN.H11MO.0.A	193	0.217	2.24	3.10	2.64	4.79e-27	1.93e-25	Fos-related factors{1.1.2}	Fos factors{1.1.2.1}
FOS_HUMAN.H11MO.0.A	193	0.217	2.19	3.04	2.59	3.01e-26	1.10e-24	Fos-related factors{1.1.2}	Fos factors{1.1.2.1}
JUN_HUMAN.H11MO.0.A	184	0.207	2.10	2.93	2.49	2.23e-23	6.36e-22	Jun-related factors{1.1.1}	Jun factors{1.1.1.1}
BATF_HUMAN.H11MO.1.A	118	0.133	2.06	3.08	2.53	5.29e-17	7.85e-16	B-ATF-related factors{1.1.4}	B-ATF{1.1.4.0.1}
JUND_HUMAN.H11MO.0.A	186	0.209	1.99	2.77	2.35	2.47e-21	5.77e-20	Jun-related factors{1.1.1}	Jun factors{1.1.1.1}
JUNB_HUMAN.H11MO.0.A	181	0.204	1.90	2.66	2.25	4.78e-19	8.78e-18	Jun-related factors{1.1.1}	Jun factors{1.1.1.1}
HXB13_HUMAN.H11MO.0.A	94	0.106	1.74	2.72	2.19	8.40e-11	7.20e-10	HOX-related factors{3.1.1}	HOX9-13{3.1.1.8}
PIT1_HUMAN.H11MO.0.C	106	0.119	1.71	2.60	2.11	4.53e-11	3.97e-10	POU domain factors{3.1.10}	POU1 (Pit-1-like factors){3.1.10.1}
PRRX1_HUMAN.H11MO.0.D	78	0.0878	1.67	2.71	2.14	6.55e-9	4.55e-8	Paired-related HD factors{3.1.3}	PRRX{3.1.3.21}
CDX1_HUMAN.H11MO.0.C	91	0.102	1.64	2.57	2.06	2.64e-9	1.94e-8	HOX-related factors{3.1.1}	CDX (Caudal type homeobox){3.1.1.9}
LMX1A_HUMAN.H11MO.0.D	134	0.151	1.62	2.36	1.96	3.06e-11	2.74e-10	HD-LIM factors{3.1.5}	Lmx{3.1.5.6}

Step 5: Human TFs

UniProtKB - "transcription factor" AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]" Advanced Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

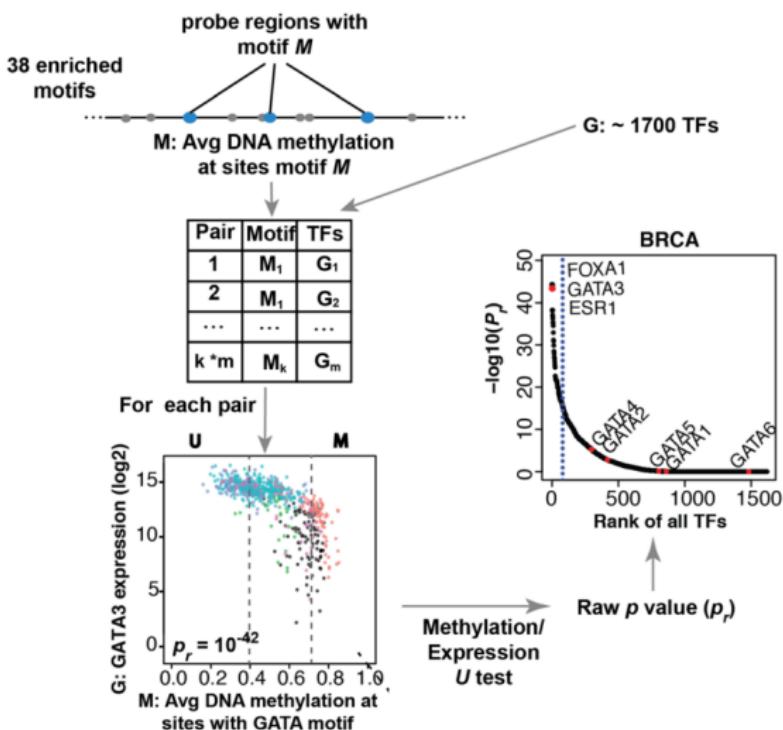
UniProtKB results

Filter by:

- Reviewed (2,017) Swiss-Prot
- Popular organisms Human (2,017)
- Proteomes UP000005640 (2,017)
- Search terms Filter "transcription factor" as: gene ontology (2,060) protein family (25) protein name (409)
- View by Results table Taxonomy Keywords Gene Ontology Enzyme class

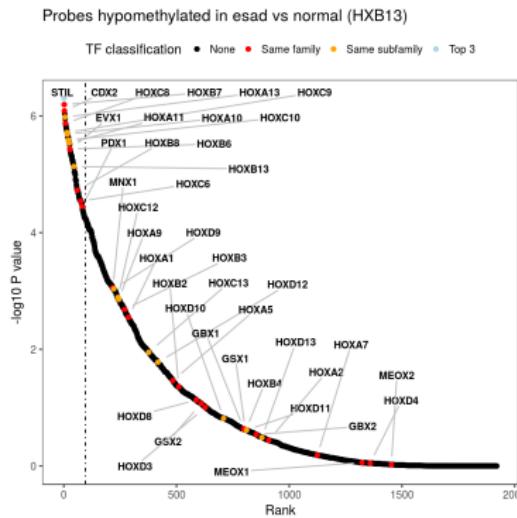
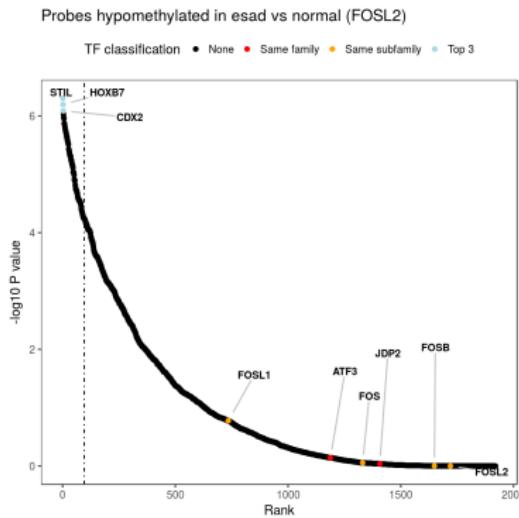
Entry	Entry name	Protein names	Gene names	Organism
P04637	P53_HUMAN	Cellular tumor antigen p53	TP53 P53	Homo sapiens (Human)
P15884	ITF2_HUMAN	Transcription factor 4	TCF4 BHLB19, ITF2, SEF2	Homo sapiens (Human)
Q04206	TF65_HUMAN	Transcription factor p65	RELA NFKB3	Homo sapiens (Human)
Q00613	HSF1_HUMAN	Heat shock factor protein 1	HSF1 HSTF1	Homo sapiens (Human)
Q9NQB0	TF7L2_HUMAN	Transcription factor 7-like 2	TCF7L2 TCF4	Homo sapiens (Human)
P18074	ERCC2_HUMAN	TFIIE basal transcription factor co... protein	ERCC2 XPD, XPDC	Homo sapiens (Human)
P10275	ANDR_HUMAN	Androgen receptor	AR DHTR, NR3C4	Homo sapiens (Human)
P08047	SP1_HUMAN	Transcription factor Sp1	SP1 TSFP1	Homo sapiens (Human)
P05412	JUN_HUMAN	Transcription factor AP-1	JUN	Homo sapiens (Human)
P35222	CTNNB1_HUMAN	Catenin beta-1	CTNNB1 CTNNB, OK/SW-cl.35, PRO2286	Homo sapiens (Human)
P35555	FBN1_HUMAN	Fibrillin-1	FBN1 FBN	Homo sapiens (Human)

Step 5: Identification of master regulator TF

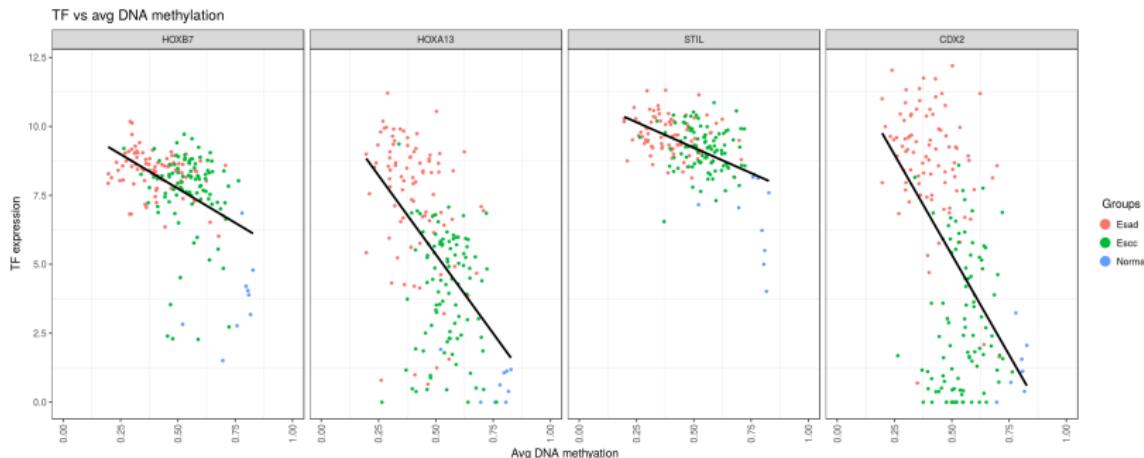


Yao et al. Genome Biology (2015) 16:105.

Step 5: TF ranking plot



Step 5: TF scatter plot



Step 5: TF master regulator table

motif	OR	top.potential.TF.family	pvalue.TF.family	top.potential.TF.subfamily	pvalue.TF.subfamily
All		All	All	All	All
HXB13_HUMAN.H11MO.0.A	2.19	HOXB7	6.39e-7	HOXA13	0.00000105
CDX1_HUMAN.H11MO.0.C	2.06	HOXB7	6.39e-7	CDX2	8.20e-7
HXD9_HUMAN.H11MO.0.D	1.98	HOXB7	6.39e-7	HOXA13	0.00000105
PDX1_HUMAN.H11MO.1.A	1.89	HOXB7	6.39e-7	PDX1	0.0000355
HXC11_HUMAN.H11MO.0.D	1.84	HOXB7	6.39e-7	HOXA13	0.00000105
HXB6_HUMAN.H11MO.0.D	1.84	HOXB7	6.39e-7	HOXB7	6.39e-7
HXD8_HUMAN.H11MO.0.D	1.84	HOXB7	6.39e-7	HOXC8	0.00000134
CDX2_HUMAN.H11MO.0.A	1.83	HOXB7	6.39e-7	CDX2	8.20e-7
HXD12_HUMAN.H11MO.0.D	1.77	HOXB7	6.39e-7	HOXA13	0.00000105
HXC9_HUMAN.H11MO.0.C	1.74	HOXB7	6.39e-7	HOXA13	0.00000105

Showing 1 to 10 of 31 entries

Previous

1

2

3

4

Next

Differences between versions

Table: Main differences between ELMER version 1 and 2)

Features	ELMER Version 1	ELMER Version 2
Main data structure	mee object (own data structure)	MAE object (Bioconductor data structure)
Auxiliary data	Manually created	Programmatically created
Number of human TF	1,982	1,987 (Uniprot database)
Number of TF motifs	91	771 (HOCOMOCO v11 database)
TF classification	78 families	80 families and 308 subfamilies (TFCClass database)
Analysis performed	Normal tumor samples vs experiment	Group 1 vs group 2
TCGA samples source	The Cancer Genome Atlas (TCGA)	The NCI's Genomic Data Commons (GDC)
Genome of reference	GRCh37 (hg19)	GRCh37 (hg19)/GRCh38 (hg38)
DNA methylation platforms	HumanMethylation450	HumanMethylationEPIC / HumanMethylation450

TCGA Breast Invasive Carcinoma (BRCA) Samples

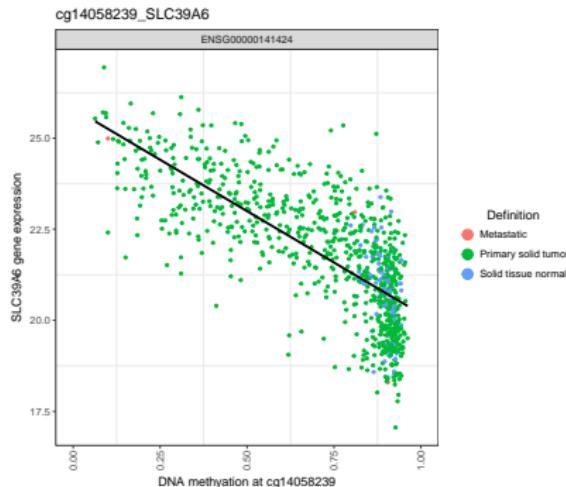
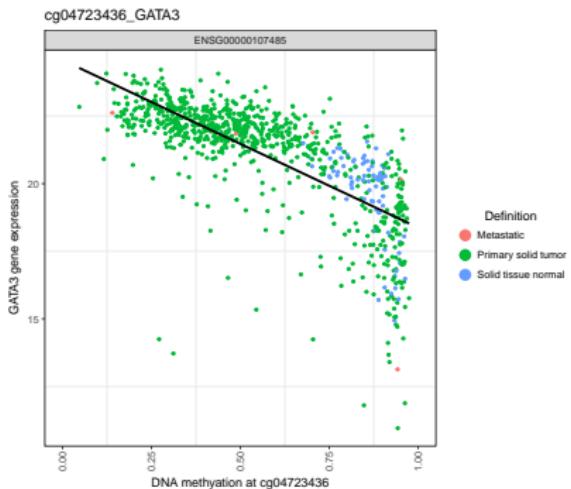
Table: Summary of the available samples in TCGA for BRCA

Group	Samples w/ methylation (450K)	Samples w/ gene ex- pression (FPKM-UQ)	Samples w/ both
Primary solid Tumor	791	1102	778
Solid Tissue Normal	96	113	83

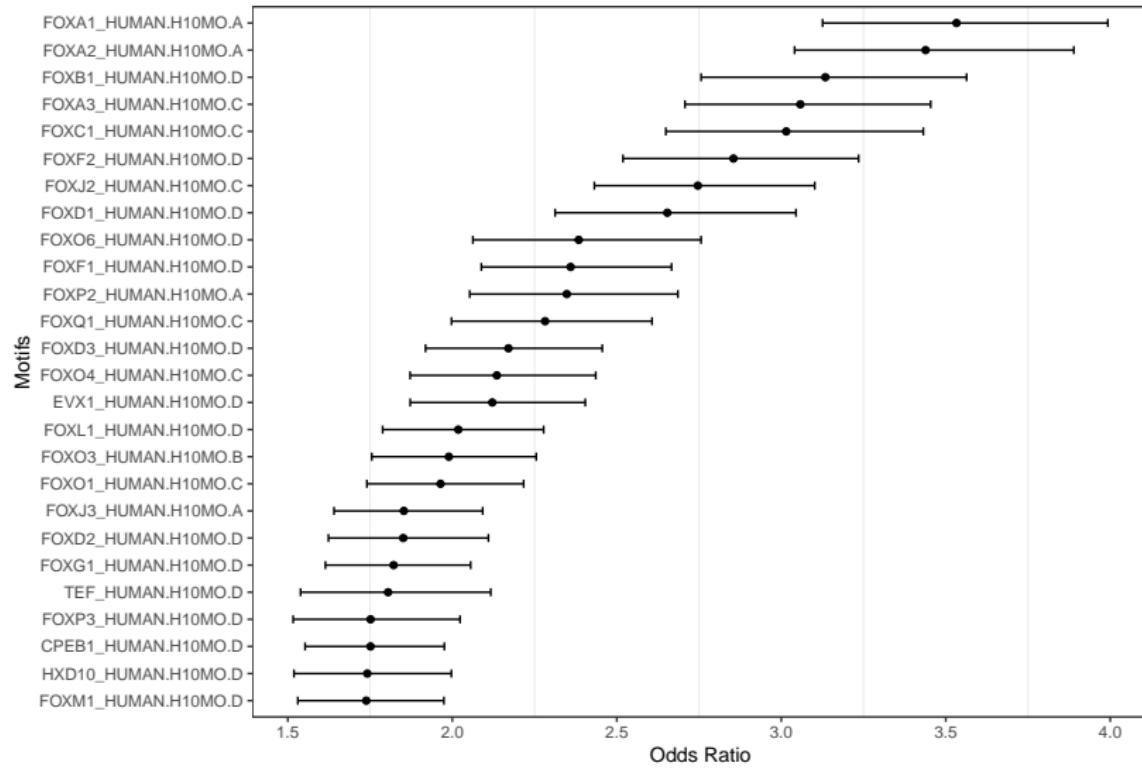
Table: Results

Inferred gene-probe pairs	2167
Enriched motifs	312
Regulatory TF factors	17

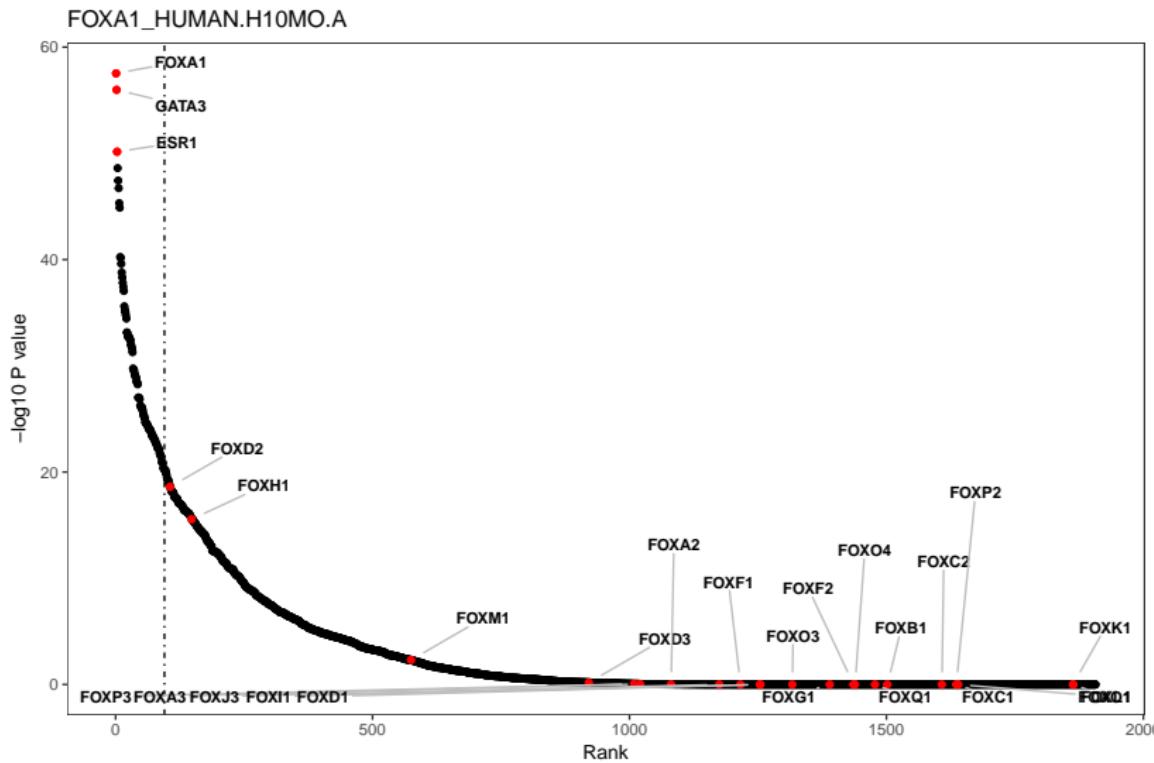
Step 3: Pairs inferred



Top enriched motifs



TF ranking



Regulatory TF

RESEARCH ARTICLE | OPEN ACCESS

Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours

André Albergaria, Joana Paredes, Bárbara Sousa, Fernanda Milanezi, Vítor Carneiro, Joana Bastos, Sandra Costa, Daniella Vieira, Nair Lopes, Eric W Lam, Nuno Lunet and Fernando Schmitt 

Breast Cancer Research 2009, 11:R40 | DOI: 10.1186/bcr2327 | © Albergaria et al.; licensee BioMed Central Ltd. 2009

Received: 5 January 2009 | Accepted: 23 June 2009 | Published: 23 June 2009

Oncotarget, 2015 Sep 8;6(26):21878-91

The pioneer factor PBX1 is a novel driver of metastatic progression in ER α -positive breast cancer.

Magnani L¹, Patten DK¹, Nguyen VT¹, Hong SP¹, Steel JH¹, Patel N¹, Lombardo Y¹, Faronato M¹, Gomes AR¹, Woodley L¹, Page K², Guttery D², Primrose L², Fernandez Garcia D², Shaw J², Viola P³, Green A⁴, Nolan C⁴, Ellis IO⁴, Rakha EA⁴, Shousha S¹, Lam EW¹, Györffy B⁵, Lupien M^{6,7}, Coombes RC¹,

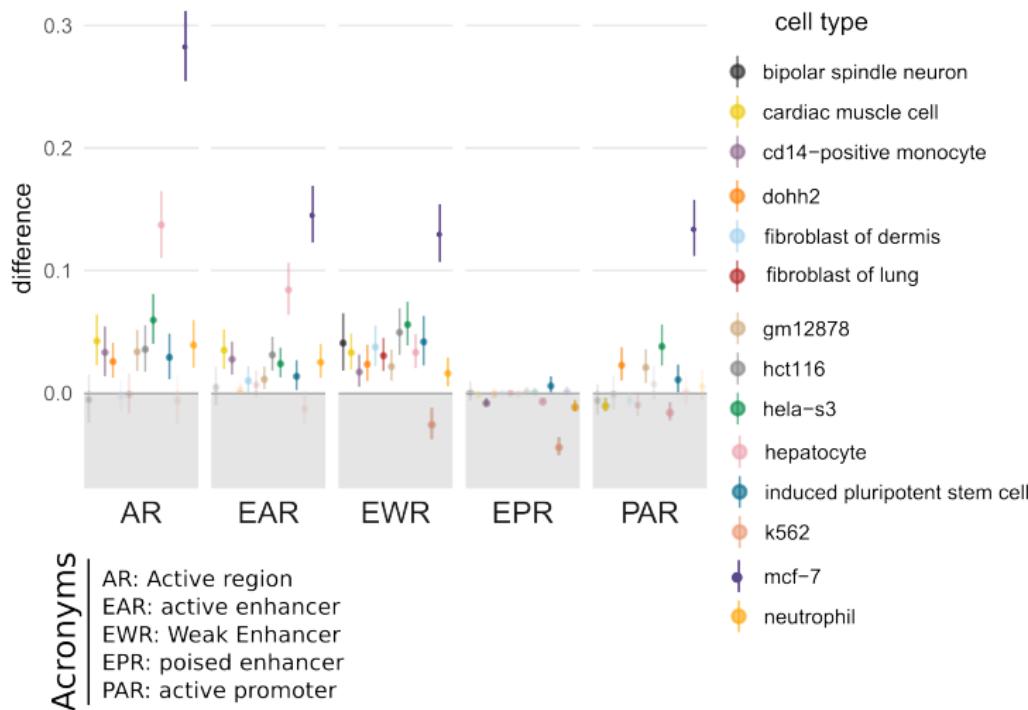
Author information

Article | OPEN

Retinoic acid receptor alpha is associated with tamoxifen resistance in breast cancer

Henrik J. Johansson, Betzabe C. Sanchez, Filip Mundt, Jenny Forshed, Aniko Kovacs, Elena Panizza, Lina Hultin-Rosenberg, Bo Lundgren, Ulf Martens, Gyöngyvér Máthé, Zohar Yakhini, Khalil Helou, Kamilla Krawiec, Lena Kanter, Anders Hierpe, Olle Stål, Barbro K. Linderholm & Janne Lehtio

Annotating chromatin state and verifying enrichment



Comparing inferred results with MCF-7 chIA-PET

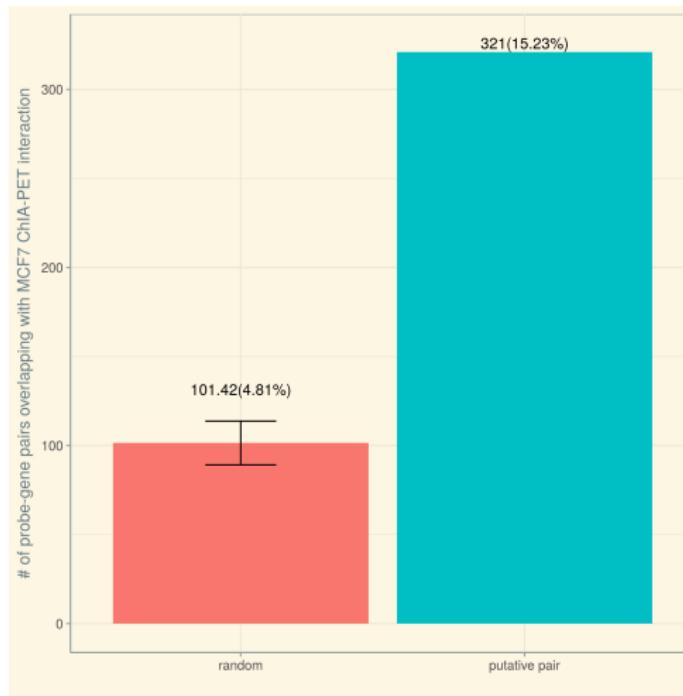


Figure: The graph shows the comparison of the number of probe-gene pairs

Acknowledgments

University of São Paulo

- Houtan Noushmerh



Cedars-Sinai

- Ben Berman
 - Simon Coetzee
 - Dennis Hazelett
 - Nicole Yeager
 - Huy Dinh
 - Michelle Jones
 - Alberto Reyes
 - Iveth Corona



CEDARS-SINAI®



References



Ginsburg, Geoffrey S., and Huntington F. Willard.
Genomic and personalized medicine.
Vol. 1. Academic Press, 2008.



Silva TC, Colaprico A, Olsen C et al.
TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages
F1000Research 2016, 5:1542



H. Noushmehr, D. J. Weisenberger, et al.
Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma.
Cancer Cell, 17(5):510(522, May 2010.



S. Sharma, T. K. Kelly, and P. A. Jones.
Epigenetics in cancer.
Carcinogenesis, 31(1):27(36, Jan 2010.



Tiago Chedraoui Silva*, Antonio Colaprico*, et al.
TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res.,
Nucleic acids research, p.gkv1507, 2015



Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, et al. . Cell,
Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma
Cell, 164(3):550-563, 2016

Thank you for your attention!

Any questions?