This article was downloaded by you on: Mar 19, 2016

A Journal of the Bangladesh Pharmacological Society (BDPS)
Journal homepage: www.banglajol.info
Abstracted/indexed in Academic Search Complete, Asia Journals Online, Bangladesh Journals Online, Biological Abstracts, BIOSIS Previews, CAB Abstracts, Current Abstracts, Directory of Open Access Journals, EMBASE/Excerpta Medica, Google Scholar, HINARI (WHO), International Pharmaceutical Abstracts, Open J-gate, Science Citation Index Expanded, SCOPUS and Social Sciences Citation Index;

Bangladesh J Pharmacol 2016; 11: S154-S160

ISSN: 1991-0088

Cross-cell DNA methylation annotation and analysis for pan-cancer study

Binhua Tang[1,2,3], Weiliang Zhu[4] and Changping Wu[4]

[1]Epigenetics and Function Group, College of the Internet of Things, Hohai University, Jiangsu 213 022, China; [2]School of Public Health, Shanghai Jiao Tong University, Shanghai 200 225, China; [3]CBMI, Harvard Medical School, Boston, MA 02 115, USA; [4]The Third Affiliated Hospital, Soochow University, Jiangsu 213 003, China.

Article Info

Received:	12 January 2016
Accepted:	27 February 2016
Available Online:	11 March 2016

DOI: 10.3329/bjp.v11iS1.26851

Cite this article:
Tang B, Zhu W, Wu C. Cross-cell DNA methylation annotation and analysis for pan-cancer study. Bangladesh J Pharmacol. 2016; 11: S154-S160.

Abstract

Pan-cancer study can uncover cell- and tissue-specific genomic loci and regions with underlying biological functions, as one of fundamental procedures toward precision medicine. We utilized the online curated resource of DNA methylation annotation knowledgebase, to implement the cross-cell interrogation of pan-cancer study of breast cancer. The study revealed genome-wide differentially-methylated loci and regions by the reduced representation bisulfite sequencing profiling. The knowledgebase contains three level of curated information across multiple cancer and normal cells from the ENCODE Consortium. The reference base covers all identified differentially-methylation CpG sites and regions of interest, further annotated gene information, together with tumor suppressor gene and methylation level. Lastly, it includes the inferred functional association network and related Gene Ontology analysis results based on all the tumor suppressor genes identified from the differentially-methylated regions of interest. Our knowledgebase and analysis results provide a thorough reference source for biomedical researchers and clinicians. The cross-cell analysis results are deposited at: http://github.com/gladex/DMAK.

Introduction

Pan-cancer study can uncover most of cell- and tissue-specific genomic loci and regions with underlying biological functions (Kristensen et al., 2014; Leiserson et al., 2015; The Cancer Genome Atlas Research Network et al., 2013; Witte et al., 2014). Meanwhile, it provides meaningful insights from the genome-wide interrogation of cross-cell analysis and annotation.

While till now, to biomedical researchers and clinicians, there is no systematic reference source of functional association between DNA methylation and transcriptional regulation for wet-lab experiment design and post-experiment validation. Thus, this is an imperative for most biologists and biomedical researchers to improve their research outcomes and efficiency (Bock and Lengauer, 2008; Roadmap Epigenomics Consortium et al., 2015).

Here, we utilized our online curated reference source for DNA Methylation Annotation Knowledgebase (DMAK) and implemented the cross-cell analysis in pan -cancer study. The knowledgebase provides multiple read-to-use analysis results and annotation information for the pan-cancer interrogation and cross-validation.

We deposited the curated information knowledgebase and related analysis results on GitHub for direct download and usage for free. Proper citation is suggested for any usage, possible reanalysis or refinement.

Structure and Purpose of DMAK

DMAK contains three levels of curated information across multiple cell types from ENCODE Consortium portal (de Souza, 2012; Pennisi, 2012; Tang and Wang, 2015; The ENCODE Project Consortium, 2012). The cell types investigated as below include breast cancer (T-47D and MCF-7), cervical cancer (HeLa-S3), endometrial cancer (ECC-1), blood cancer (GM12878, GM12891, GM12892, HL-60 and K562), brain cancer (SK -N-MC, SK-N-SH, SK-N-SH_RA, PFSK-1 and U87), liver cancer (HepG2), colon cancer (HCT-116), pancreas cancer (PANC-1), lung cancer (A549), and human embryonic stem cell (H1-hESC).

As depicted in Figure 1, the first level of DMAK was the curation of raw data sources from ENCODE Consortium portal; for the study case in our work, we emphasized on cross-cell DNA methylation profiling information for detecting differentially-methylated features and patterns within breast cancer T-47D cell type.

This level content includes the summary for the analysis procedure and fundamental functions as discussed in the following sections.

The second level mainly focuses on integrative analysis on the curated DNA methylation data in RRBS format (Blattler et al., 2014; Ziller et al., 2013), we implemented function annotation for methylated CpG sites, identified differentially-methylated regions (DMR), and classified the hyper- and hypo-methylated regions or differential DMR candidates (Kemp Christopher et al., 2014). The detailed analysis procedure and results are given in the following section.

The third level analysis mainly includes the visualization and function analysis for the annotated results, which include the functional association network for tumor suppressor genes identified from the hyper- or hypo-DMRs detected from the above analysis.

We curated information and constructed the comprehensive knowledgebase using data sources mainly from ENCODE Consortium portal, together with other commonly-used tools, and the self-compiled scripts and programs.

Annotation and Analysis Procedure in DMAK

This section mainly discusses the functions and analysis procedure for DMAK, which covers fundamental functions of DMAK reference source, listed as,

1) Statistical information detected for sequencing read coverage, number of Cs and Ts of for the 688,445 CpG sites across all cancer and normal cell lines listed above. For consistence, all DNA methylation data sets from ENCODE Consortium are based on the RRBS platform. The output format is given Figure 2.

2) Analysis and annotation results for the methylated CpG sites (mCpG), which provide the methylation percentage value for all mCpG sites across the

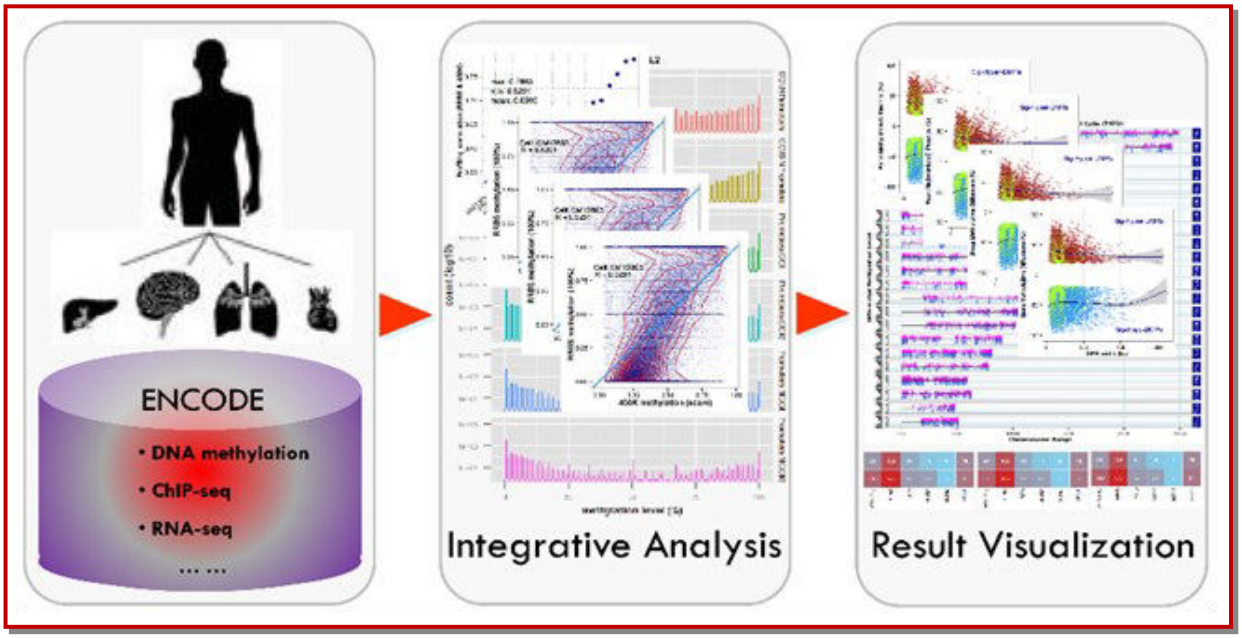
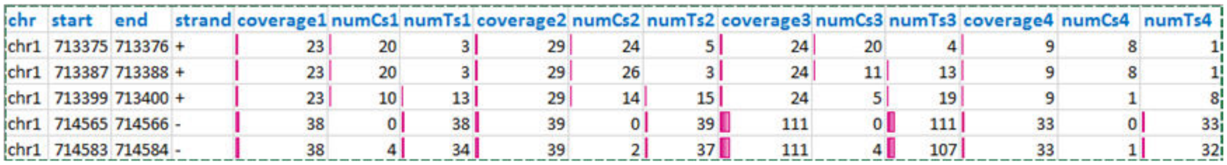
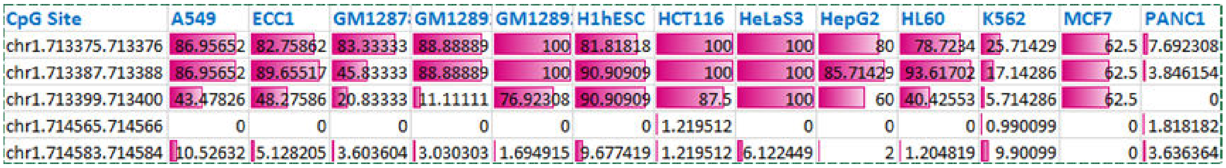


Figure 1: Schematic illustration for the DMAK structure. The first level contains ENCODE data preprocess (namely, cell curation and data format process); the second level includes integrative analysis on the ENCODE data, namely DNA methylation CpGs annotation, identification of differentially-methylated CpGs and regions; the third level covers result visualization and further-multi-scale interrogation of biological functions



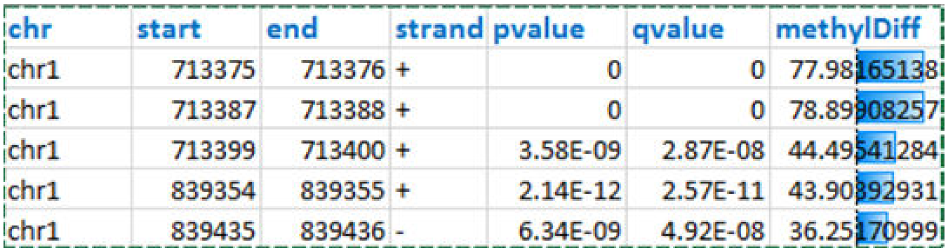
chr	start	end	strand	coverage1	numCs1	numTs1	coverage2	numCs2	numTs2	coverage3	numCs3	numTs3	coverage4	numCs4	numTs4
chr1	713375	713376	+	23	20	3	29	24	5	24	20	4	9	8	1
chr1	713387	713388	+	23	20	3	29	26	3	24	11	13	9	8	1
chr1	713399	713400	+	23	10	13	29	14	15	24	5	19	9	1	8
chr1	714565	714566	-	38	0	38	39	0	39	111	0	111	33	0	33
chr1	714583	714584	-	38	4	34	39	2	37	111	4	107	33	1	32

Figure 2: Schematic illustration of statistical information detected from RRBS sequencing read coverage, number of Cs and Ts



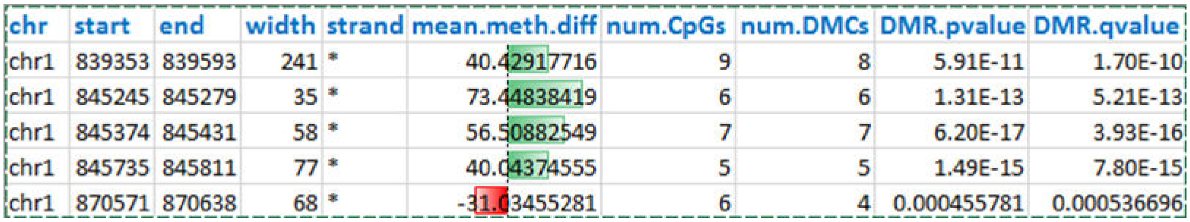
CpG Site	A549	ECC1	GM1287	GM1289	GM1289	H1hESC	HCT116	HeLaS3	HepG2	HL60	K562	MCF7	PANC1
chr1.713375.713376	86.95652	82.75862	83.33333	88.88889	100	81.81818	100	100	80	78.7234	25.71429	62.5	7.692308
chr1.713387.713388	86.95652	89.65517	45.83333	88.88889	100	90.90909	100	100	85.71429	93.61702	17.14286	62.5	3.846154
chr1.713399.713400	43.47826	48.27586	20.83333	11.11111	76.92308	90.90909	87.5	100	60	40.42553	5.714286	62.5	0
chr1.714565.714566	0	0	0	0	0	0	1.219512	0	0	0	0.990099	0	1.818182
chr1.714583.714584	10.52632	5.128205	3.603604	3.030303	1.694915	9.677419	1.219512	6.122449	2	1.204819	9.90099	0	3.636364

Figure 3: Schematic illustration of methylated CpG sites across all listed cell types. The methylation value is of percentage format, annotated with purple bars for each cell line



chr	start	end	strand	pvalue	qvalue	methylDiff
chr1	713375	713376	+	0	0	77.98165138
chr1	713387	713388	+	0	0	78.89908257
chr1	713399	713400	+	3.58E-09	2.87E-08	44.49541284
chr1	839354	839355	+	2.14E-12	2.57E-11	43.90892931
chr1	839435	839436	-	6.34E-09	4.92E-08	36.25170999

Figure 4: Schematic illustration of significantly DMC (SDMC). The methylation difference is annotated with blue bars for each CpG site in T-47D cell line



chr	start	end	width	strand	mean.meth.diff	num.CpGs	num.DMCs	DMR.pvalue	DMR.qvalue
chr1	839353	839593	241	*	40.42917716	9	8	5.91E-11	1.70E-10
chr1	845245	845279	35	*	73.44838419	6	6	1.31E-13	5.21E-13
chr1	845374	845431	58	*	56.50882549	7	7	6.20E-17	3.93E-16
chr1	845735	845811	77	*	40.04374555	5	5	1.49E-15	7.80E-15
chr1	870571	870638	68	*	-31.03455281	6	4	0.000455781	0.000536696

Figure 5: Schematic illustration for the identified DMRs with reference to T-47D cell type. The table provides the width, mean methylation difference, the corresponding p-value and adjusted q-value for each DMR entry. The mean methylation difference is annotated with green or red bars for each DMR

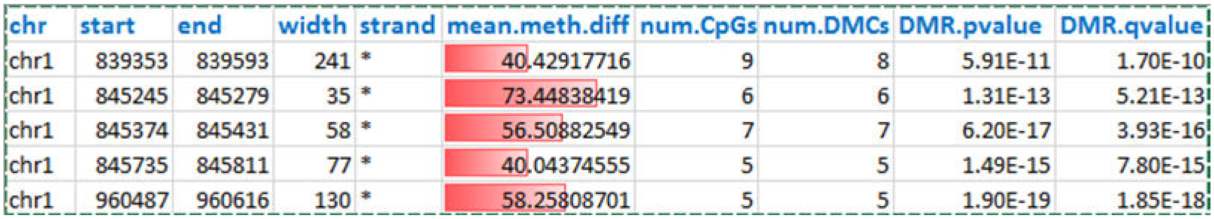
mentioned cell lines. The higher percentage value indicates the higher methylation status, and vice versa (Figure 3).

3) Analysis and annotation results for the significant differentially-methylated CpG sites (SDMC) with reference to one cell type, here we selected T-47D as the study case. The results are further filtered based on the lifted methylation difference threshold (at least 25% methylation difference for the paired groups). And the SDMC list contains 106,252 DMCs (Akalin et al., 2012a; Akalin et al., 2012b), together the related statistical p-value and adjusted q-value are also provided in Figure 4.

4) Statistical analysis and annotation results for the differentially-methylated regions (DMR) with reference to one cell type, for consistence we selected T-47D as the case. We identified 16,277 DMR candidates from all the DMCs, with the adjusted q-value ≤0.01, CpG base methylation difference cut off, 25, and DMR mean methylation difference cut off, 20. Within those candidates, 8,936 entries present hyper-methylated and 7,341 with hypo-methylated status. With the lifted thresholds, namely adjusted q-value ≤0.001, differentially-methylated CpG base count ≥5, we further detected 7,537 significant DMRs (Sig-DMRs), where 3,512 entries are significantly hypermethylated-DMRs (Sig-Hyper-DMRs), and 4,025 significantly hypomethylated-DMRs (Sig-Hypo-DMRs). The output format is shown in Figure 5.

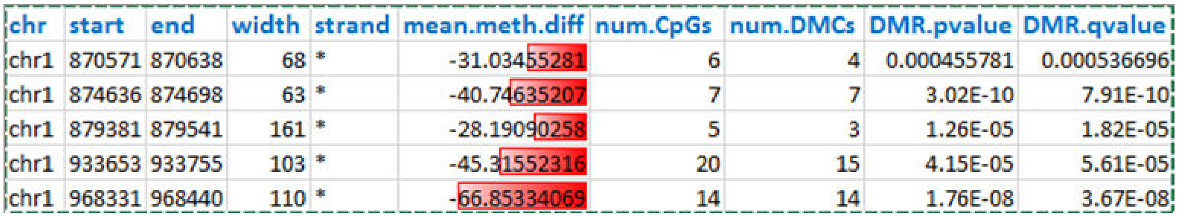
5) Statistical analysis and annotation results for the significantly hypermethylated-DMRs (Sig-Hyper-DMRs) with reference to T-47D cell type as shown in the output format (Figure 6).

6) Statistical analysis and annotation results for the significantly hypomethylated-DMRs (Sig-Hypo-DMRs) with reference to T-47D cell type (Figure 7).



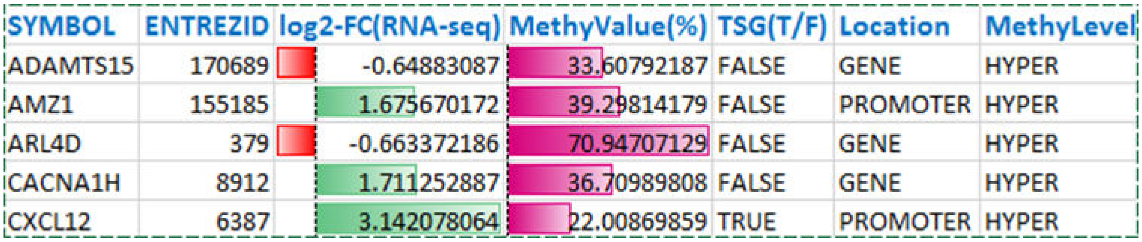
chr	start	end	width	strand	mean.meth.diff	num.CpGs	num.DMCs	DMR.pvalue	DMR.qvalue
chr1	839353	839593	241	*	40.42917716	9	8	5.91E-11	1.70E-10
chr1	845245	845279	35	*	73.44838419	6	6	1.31E-13	5.21E-13
chr1	845374	845431	58	*	56.50882549	7	7	6.20E-17	3.93E-16
chr1	845735	845811	77	*	40.04374555	5	5	1.49E-15	7.80E-15
chr1	960487	960616	130	*	58.25808701	5	5	1.90E-19	1.85E-18

Figure 6: Schematic illustration for the identified hyper-DMR with reference to T-47D cell type. The mean methylation difference is annotated with red bars for each DMR



chr	start	end	width	strand	mean.meth.diff	num.CpGs	num.DMCs	DMR.pvalue	DMR.qvalue
chr1	870571	870638	68	*	-31.03455281	6	4	0.000455781	0.000536696
chr1	874636	874698	63	*	-40.74635207	7	7	3.02E-10	7.91E-10
chr1	879381	879541	161	*	-28.19090258	5	3	1.26E-05	1.82E-05
chr1	933653	933755	103	*	-45.31552316	20	15	4.15E-05	5.61E-05
chr1	968331	968440	110	*	-66.85334069	14	14	1.76E-08	3.67E-08

Figure 7: Schematic illustration for the identified hypo-DMR with reference to T-47D cell type. The mean methylation difference is annotated with red bars for each DMR



SYMBOL	ENTREZID	log2-FC(RNA-seq)	MethyValue(%)	TSG(T/F)	Location	MethyLevel
ADAMTS15	170689	-0.64883087	33.60792187	FALSE	GENE	HYPER
AMZ1	155185	1.675670172	39.29814179	FALSE	PROMOTER	HYPER
ARL4D	379	-0.663372186	70.94707129	FALSE	GENE	HYPER
CACNA1H	8912	1.711252887	36.70989808	FALSE	GENE	HYPER
CXCL12	6387	3.142078064	22.00869859	TRUE	PROMOTER	HYPER

Figure 8: Schematic illustration for the identified gene information (SYMBOL and ENTREZ ID), log2 fold change, methylation percentage, tumor suppressor gene category (TRUE/FALSE), location (Promoter, CDS, Gene, 5'UTR, 3'UTR and Intron) and related methylation level (HYPER/HYPO) from DMRs with reference to T-47D cell type. The log2 fold change and methylation percentage is annotated with red bars for each annotated gene

7) Statistical analysis and annotation results for the identified genes from all DMRs (hyper-DMRs and hypo-DMRs) with reference to T-47D cell type (Figure 8).

Visualization and Function Analysis for the Annotation Results

This section discusses the visualization and function analysis for the annotated results. Together we seek to detect whether there exists any functional association (Szklarczyk et al., 2015) between those identified genes from hyper-DMRs and hypo-DMRs, which can explain the differential expression between those genes qualitatively, especially for the genes belonging to tumor suppressor genes (TSG) (Bedi et al., 2014; Blattler et al., 2014; Zhao et al., 2013).

Thus we annotated the genes identified from DMRs with TSG information, filtered out those from unknown sources, and constructed the TSG functional association networks for hyper-DMR and hypo-DMR, respectively.

For illustration and space limitation, Figure 8 depicts the 20-TSG functional association structures for hyper- and hypo-DMRs, respectively. For validating the high fidelity of the analysis results, those 20 TSGs are randomly selected from the TSG list for each case.

And interestingly, we found most of those TSG nodes are functionally associated to form clusters. In hyper-DMR case, Figure 9A, only 4 out of 20 TSGs are dissociated from the TSG cluster; for hypo-DMR case, Figure 9B, it is comparatively loosely-connected and 10 out of 20 TSGs are not linked to the TSG cluster.

The complete TSG functional association network structures for hyper-DMR and hypo-DMR are provided in DMAK package deposited at GitHub. The TSGs in those structures are highly physically connected and functional associated in DMRs for our T-47D breast cancer case.

Figure 10 depicts the Gene Ontology (Sherman et al., 2007)analysis results for the two functional protein association network inferred for the TSGs. The upper (A) is for the hyper-DMRs, and the corresponding GO terms clearly prove such processes as transcription regulation, differentiation, mutation, activator, pathway in cancer and tumor suppressor, which are closed related to the hypermethylation outcomes of tumor suppressor genes.

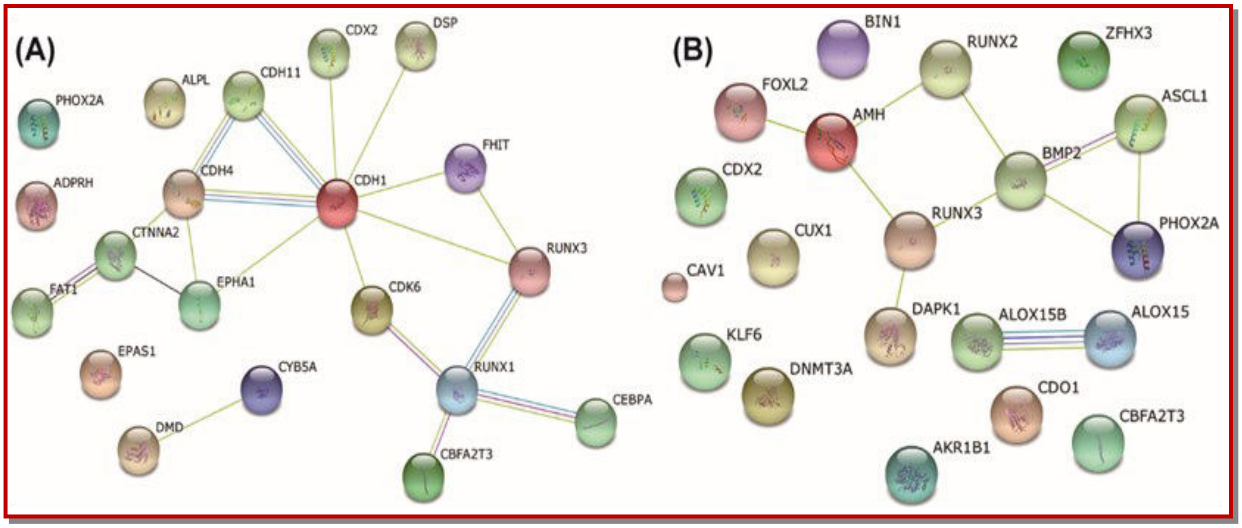


Figure 9: Illustrative diagram for functional protein association network inferred for the tumor suppressor genes (TSG), identified from hyper-DMRs (A) and hypo-DMRs (B). The nodes represent the TSGs detected from DMRs, and links represent the association evidences. The full landscape for the TSG association networks in hyper-DMR and hypo-DMR are provided in the DMAK package deposited at GitHub

And the bottom (B) for the hypo-DMRs, and its GO terms present positive regulation of transcription and gene expression, differentiation, which to a certain extent confirm its connectivity to the hypomethylation outcomes of TSGs.

Conclusion

Our cross-cell DNA methylation annotation and analysis provide the systematic information knowledgebase for pan-cancer study. It contains curated reference results for ready-to-use information for sharing and rapid reanalysis.

The first level of the knowledgebase is about raw data preprocess, we collected the data from the ENCODE Consortium portal. The second level is for annotation and function analysis; in this study case, we focused on DNA methylation in breast cancer cell, T-47D, annotated and identified the differentially-methylated sites and regions, and further identified the underlying tumor suppressor genes within the regions. The third level is for visualization and validation procedures. We further constructed the functional association network for the identified tumor suppressor genes, and further annotated the networks with Gene Ontology information, which can provide statistically significant evidences for the hyper-methylated and hypo-methylated processes in the breast cancer context.

Our work provides a versatile and comprehensive platform for all biomedical researchers, especially for the genome-wide biomedical analysts, to interrogate and validate their hypothesis in an efficient and uniform way.

In coming days, further annotation and analysis results concerning pan-cancer analysis will be updated into the knowledgebase, especially we seek to provide an interactive environment for biomedical researchers to fetch and utilize this knowledgebase.

Acknowledgements

The work was supported by the Research Fellowship from Harvard Medical School, the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (2nd Phase), the Fundamental Research Funds for China Central Universities under Grant Nos. 2012B03914 and XZX10B007-02, Changzhou Science & Technology Program under Grant No. CE20155050. This work made use of the Open Cloud Consortium (OCC)-sponsored project resource, which supported in part by grants from Gordon and Betty Moore Foundation and the National Science Foundation (USA) and major contributions from OCC members.

Conflict of Interest

The authors declare no conflict of interest.

References

Akalin A, Garrett-Bakelman FE, Kormaksson M, Busuttil J, Zhang L, Khrebtukova I, Milne TA, Huang Y, Biswas D, Hess JL, Allis CD, Roeder RG, Valk PJM, Löwenberg B, Delwel R, Fernandez HF, Paietta E, Tallman MS, Schroth GP, Mason CE, Melnick A, Figueroa ME. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. PLoS Genet 2012a; 8: e1002781.

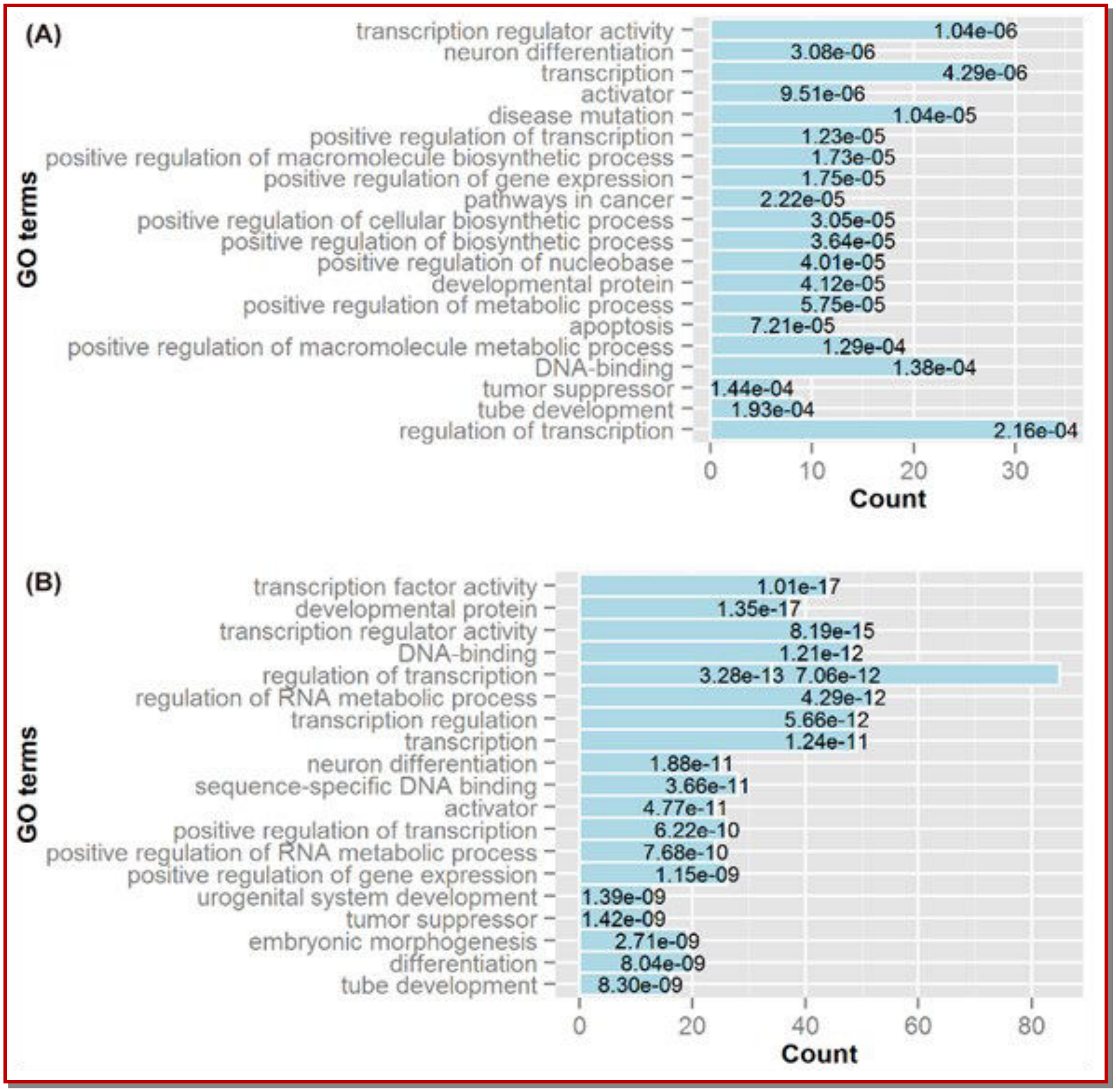


Figure 10: The Gene Ontology (GO) analysis for the two functional protein association network inferred for the TSGs, the upper (A) for hyper-DMRs and the bottom (B) for the hypo-DMRs, sorted by the p-value

Akalin A, Kormaksson M, Li S, Garrett-Bakelman F, Figueroa M, Melnick A, Mason C. Methylkit. A comprehensive r package for the analysis of genome-wide DNA methylation profiles. Genome Biology 2012b; 13: R87.

Bedi U, Mishra VK, Wasilewski D, Scheel C, Johnsen SA. Epigenetic plasticity: A central regulator of epithelial-to-mesenchymal transition in cancer. Oncotarget 2014; 5: 2016-29.

Blattler A, Yao L, Witt H, Guo Y, Nicolet C, Berman B, Farnham P. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. Genome Biology 2014; 15: 469.

Bock C, Lengauer T: Computational epigenetics. Bioinformatics 2008; 24: 1-10.

de Souza N. Genomics: The encode project. Nat Meth 2012; 9: 1046-1046.

Kemp Christopher J, Moore James M, Moser R, Bernard B, Teater M, Smith Leslie E, Rabaia Natalia A, Gurley Kay E, Guinney J, Busch Stephanie E, Shaknovich R, Lobanenkov Victor V, Liggitt D, Shmulevich I, Melnick A, Filippova Galina N. Ctcf haploinsufficiency destabilizes DNA methylation and predisposes to cancer. Cell Reports 2014; 7: 1020-29.

Kristensen VN, Lingjaerde OC, Russnes HG, Vollan HKM, Frigessi A, Borresen-Dale A-L. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer 2014; 14: 299-313.

Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, Ding L, Raphael BJ. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 2015; 47: 106-14.

Pennisi E. Encode project writes eulogy for junk DNA. Sci 2012; 337: 1159-61.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Roadmap Epigenomics C: Integrative analysis of 111 reference human epigenomes. Nature 2015; 518: 317-30.

Sherman B, Huang D, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler M, Lane HC, Lempicki R. David knowledgebase: A gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics. 2007; 8: 426.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. String v10: Protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research 2015; 43: D447-52.

Tang B, Wang X. Inferring genome-wide interplay landscape between DNA methylation and transcriptional regulation P J Pharm Sci 2015; 28: 349-52.

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013; 45: 1113-20.

The ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature 2012; 489: 57 -74.

Witte T, Plass C, Gerhauser C. Pan-cancer patterns of DNA methylation. Genome Med. 2014; 6: 1-18.

Zhao M, Sun J, Zhao Z: Tsgene. A web resource for tumor suppressor genes. Nucleic Acids Research 2013; 41: D970-76.

Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LTY, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A. Charting a dynamic DNA methylation landscape of the human genome. Nature 2013; 500: 477-81.

Author Info
Binhua Tang (Principal contact)
e-mail: bh.tang@outlook.com; Mobile: +86-18861231716

First two authors contributed equally

Your feedback about this paper

1. Number of times you have read this paper 0

2. Quality of paper

○ Excellent ○ Good ○ Moderate ○ Not good

3. Your comments