DMAK: A curated pan-cancer DNA methylation annotation knowledgebase

Binhua Tang

Epigenetics and Function Group, School of Internet of Things, Hohai University, Jiangsu, China; School of Public Health, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Pan-cancer analysis can identify cell- and tissue-specific genomic loci and regions with underlying biological functions. Here we present an online curated DNA Methylation Annotation Knowledgebase, DMAK, which includes the pan-cancer analysis results for differentially-methylated loci and regions by the Reduced Representation Bisulfite Sequencing profiling technology. DMAK contains 3 modules of curated information and analysis results on 688,445 CpG sites across 19 cancer and embryonic stem cell lines from ENCODE, and further analysis of survival associations with clinical sources retrieved from TCGA. The knowledgebase covers all identified differentially-methylated CpG sites and regions of interest, further annotated genomic information, together with tumor suppressor genes information and calculated methylation level. DMAK provides meaningful clues for deriving functional association network and related clinical association results based on protein-coding genes, including tumor suppressor genes, identified from differentially methylated regions of interest. Thus DMAK constitutes a comprehensive reference source for the current epigenetic research and clinical study.

Keywords: differential analysis; DNA methylation; genomic annotation; pan-cancer

Introduction

Pan-cancer analysis can uncover cell- and tissue-specific genomic loci and regions with underlying biological functions of interest. Meanwhile, it can provide meaningful insights by genome-wide interrogation and cross-cell genetic annotation. Especially for the topics of public research consortiums, ENCODE (Encyclopedia Of DNA Elements), focusing on identifying all functional elements in the human genome sequence; and TCGA (The Cancer Genome Atlas), providing comprehensive and multi-dimensional maps of the key genomic changes in 33 types of cancer. Pan-cancer analysis on the consortium resources can unveil the molecular basis of cancer through genome-wide interrogation and deep learning.

While till now, due to data size and technique barrier, there is no comprehensive reference source for wet-lab experiment design and post-experiment validation purposes. Thus, this is an imperative for most biologists and biomedical researchers to improve their research output and efficiency.

Here we present an online curated reference source for DNA methylation annotation and analysis purposes. The information knowledgebase provides multiple read-to-use analysis results and annotation information for pan-cancer interrogation and cross-validation usages.

For the first time, our work attempts to provide a rapid but thorough reference to the epigenetic research fields. Thus we deposit the curated information knowledgebase online for direct and interactive usage.

Structure and function of DMAK

In summary, DMAK contains 3 modules of curated information across 19 cell types retrieved from ENCODE Consortium portal. The cell types analyzed 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 Fig. 1, the first module of DMAK is the curation of raw data sources from ENCODE, including DNA methylation, RNA-seq, Tumor Suppressor Gene (TSG) and corresponding genetic annotation and analysis; within the work, we emphasize on cross-cell DNA methylation profiling information for detecting differentially-methylated loci and regions with the 3 benchmark cell lines, lung cancer A549, breast cancer MCF-7 and T-47D.

The second module mainly focuses on genomic annotation and cross-cell function analysis on the curated DNA methylation data in RRBS format\textsuperscript{17,18}, where 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).\textsuperscript{19} The detailed analysis procedure and results are given in the following section.

The third module includes the function integration and visualization for the annotated results, which includes the functional association network for tumor suppressor genes identified from the hyper- or hypo-DMRs derived from the above analysis, Gene Ontology and corresponding clinical outcome analysis.

We curated information and constructed the comprehensive knowledgebase using NGS data sources (namely RRBS, ChIP-seq and RNA-seq) mainly from ENCODE and TCGA, and the clinical survival resources retrieved from TCGA, together with other commonly-used tools, and the self-compiled programs.

**Figure 1.** Schematic illustration for DMAK structure and function. The left panel covers data preprocess for pan-cancer cell lines (namely, cell line curation and data format process); the middle panel includes annotation and integrative analysis on the curated ENCODE data, namely DNA methylation CpGs annotation, identification of differentially-methylated CpGs and regions; the right panel covers function integration and visualization, which provides clues for further multi-scale validation.

**Annotation and analysis procedure in DMAK**

This section discusses the functions and analysis procedure in DMAK. As depicted in Fig. 2, the panel provides the 4 categories of annotation information, namely, the site methylation levels for all 19 cell lines, individual DMR-Genes (hyper- and hypo-methylated cases for each cell line), specific DMR-Genes (cell-specific hyper- and hypo-methylated cases) and common DMR-Genes (common hyper- and hypo-methylated cases) for the 3 selected benchmark cell lines (A549, MCF-7 and T-47D), respectively. Right after those radio options are the corresponding drop-down selection; together with Segment to View, Rows to View and Download options, which constitute an integrative operation panel for DMAK (Fig. 2).

Thus, corresponding to the panel, the annotation and analysis mainly covers the following sections,

**Statistical information detected for sequencing read coverage**

We performed statistical calculation for the sequencing reads coverage counts (Cs and Ts) for the 688,445 CpG sites across all 19 cell lines listed above. For consistency, all DNA methylation data from ENCODE are based on the RRBS platform. The illustrative output is given in Table 1,
Statistical analysis and annotation for the identified genes from DMRs

We identified genes overlapping with all DMRs (hyper-DMRs and hypo-DMRs) with reference to each cell type (A549, MCF-7 and T-47D), respectively; then we further annotated those gene candidates with other information (symbol, log2 fold change of RNA-seq expression profile, TSG, genomic location and methylation level), thus it provides a thorough overview for those DMRs, depicted in below Table 2.

Summary panel of genome-wide DNA methylation for the 19 cell lines

This panel gives the statistical summary for the genome-wide methylation level for the 19 cell lines, which provides a general guide for comparing DNA methylation status across multiple cells, pairwise comparison or cross-cell analysis in Table 3.

Function integration and visualization

This section discusses function integration and visualization for the analysis results, including pairwise DNA methylation, differential DNA methylation and corresponding differential RNA expression comparison between cell lines, and identified genes of interest that are overlapped with the hyper- and hypo-DMRs, respectively (Fig. 3).

Furthermore we attempt to detect whether there exist any functional association between those identified genes from hyper-DMRs and hypo-DMRs, from protein level we can determine whether or not there is any potential functional link among those identified protein-coding genes, which can further explain the differential expression between those genes qualitatively, especially for the genes belonging to tumor suppressor genes (TSG).18,21,22

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.

Due to space limitation, Fig. 4 depicts the 20-TSG functional association structures for hyper-DMRs. For validating the high fidelity of the analysis results, those 20 TSGs are randomly selected from the TSG list for each case.

Table 1. Schematic illustration of statistical information for calculated methylation level (in percentage) from RRBS profiling technology.

Table 2. Schematic panel of annotation and analysis procedure in DMAK. It provides the 4 categories of annotation information, together with genomic Segment to View, Rows to View and Download options.

Figure 2.
And interestingly, we find most of those TSGs are functionally associated to form clusters. In Fig. 4, only 4 out of 20 TSGs are dissociated from the TSG cluster. Those structures further confirm TSGs are highly physically connected and functional associated in DMRs for the T-47D breast cancer case.

Next, we attempt to identify the clinical association with those gene candidates from DMRs, here for demonstration purpose, we resort to lung carcinoma study (A549 cell) in Fig. 3.

In lung carcinoma, it is recently reported that a long non-coding RNA, UCA1 (Urothelial cancer associated

Table 2. Schematic illustration for the identified gene information (SYMBOL and ENTREZ ID), log2 fold change, methylation percentage, tumor suppressor gene category (TRUE/FALSE), loci (Promoter, CDS, Gene, 5’UTR, 3’UTR and Intronic) and related methylation level (HYPER/HYPO) from DMRs with reference to T-47D cell type.

| Chr | Strand | Start     | End        | SYMBOL | ENTREZID | logFC | Methy | TSGi | Loci       | MethyLevel |
|-----|--------|-----------|------------|---------|----------|-------|-------|------|-----------|------------|
| 117 | chr2−  | 38294745  | 38303323   | CYP1B1  | 1545     | 0.40  | 37.78 | FALSE| PROMOTER  | HYPER      |
| 118 | chr2−  | 38294745  | 38303323   | CYP1B1  | 1545     | 0.40  | 37.78 | FALSE| 5’UTR     | HYPER      |
| 119 | chr2−  | 38294745  | 38303323   | CYP1B1  | 1545     | 0.40  | 37.78 | FALSE| INTRON    | HYPER      |
| 120 | chr2−  | 38294745  | 38303323   | CYP1B1  | 1545     | 0.40  | 23.81 | FALSE| CDS       | HYPER      |
| 200 | chr2−  | 31133330  | 31361592   | GALNT14 | 79623    | -0.44 | 26.02 | FALSE| INTRON    | HYPER      |
| 400 | chr2+  | 42275160  | 42286686   | PKDCC   | 91461    | -0.26 | 28.16 | FALSE| CDS       | HYPER      |
| 437 | chr2+  | 37571752  | 37600465   | QPCT    | 25797    | -0.79 | 37.46 | FALSE| 5’UTR     | HYPER      |

Table 3. Summary panel of statistical information calculated from the RRBS profiling data across the 19 ENCODE cell lines.
1) can up-regulate a potent oncogene ERBB4 (Erb-B2 receptor tyrosine kinase 4) by binding a microRNA, miR-193-3p, during transcriptional regulation.23 We use ERBB4 (2q33.3-q34) and UCA1 (19p13.12) as the study case for lung cancer, together with Kaplan-Meier probability analysis on the RNA-seq data and clinical survival information (lung adenocarcinoma, LUAD) retrieved from TCGA.1,5 Thus, based on the total clinical trial enrolment of 3,568 patients (LUAD), we calculate the clinical association anchored with the 2 candidate genes; as illustrated in Fig. 5, the results validate the 2 genes as the promising biomarkers or potential therapy targets in lung carcinoma.

Materials and methods

DNA methylation arrays 450K

The HumanMethylation 450K Beadchip assay is a CpG-specific array technology and allows for the high-resolution, genome-wide DNA methylation profiling with over 450,000 CpGs covering 99% of all RefSeq genes.24-26

Reduced representation bisulfite sequencing (RRBS)

Reduced representation bisulfite sequencing, or RRBS, is a large-scale random approach for analyzing and comparing genomic methylation patterns. BglII restriction fragments of 500–600 bp sized
selected, together with adapters assembled, were further treated with bisulfite, PCR amplification and clone, and finally sequenced to target methylated CpG sites. From the converted and unconverted read counts at each CpG, the sample coverage and methylation level (in percentage) can be acquired.\textsuperscript{11,27,28}

Annotation for the significant differentially-methylated CpG sites (SDMC)

Here we select one cell line (A549, MCF-7 and T-47D) as the reference cell type, and the annotation 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,\textsuperscript{29,30} together the related statistical p-value and adjusted q-value are also provided.

\textbf{Statistical analysis for the differentially-methylated regions}

We identified 16,277 DMR candidates from all the DMCs, with the adjusted q-value \( \leq 0.01 \), CpG base methylation difference cutoff, 25, and DMR mean methylation difference cutoff, 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 \( \leq 0.001 \), differentially-methylated CpG base count \( \geq 5 \), 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).

**Tools used in the curation and analysis**

Bowtie\(^2\)\(^{31}\) was used to align sequencing reads; SAMtools\(^2\)\(^{32}\) and BAMTools\(^2\)\(^{33}\) were used to process the aligned sequencing reads; methylKit\(^2\)\(^{30}\) was used to analyze part of RRBS data, and DEseq\(^2\)\(^{34}\) was used to analyze RNA-seq data.

**Conclusion**

DMAK provides a comprehensive annotation and analysis knowledgebase for pan-cancer study. It contains 3 modules of curated reference results for ready-to-use information sharing and rapid reanalysis.

The first module of the knowledgebase is about raw data preprocess, and we retrieved DNA methylation data from ENCODE and clinical resources from TCGA. The second 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 is for function integration and visualization procedures. We further constructed the functional association network for the identified tumor suppressor genes, and further performed the clinical association study with the DMR genes of interest, which can provide statistically significant evidences for the hyper-methylated and hypo-methylated processes in the transcriptional regulation context.

Our work provides a versatile and comprehensive platform for the corresponding biomedical research, especially for the genome-wide study, 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, thus it constitutes an interactive and efficient approach for biologists to carry out their research with knowledgebase.

**Availability**

DMAK is deployed at gladex.shinyapps.io/DMAK/ and dma2.hhuc.edu.cn/DMAK/.

**Disclosure of potential conflicts of interest**

No potential conflicts of interest were disclosed.

**Funding**

This work was supported by the Natural Science Foundation of Jiangsu, China (Nos: BE2016655 and BK20161196), Fundamental Research Funds for China Central Universities (No. 2016B08914) and Changzhou Science & Technology Program (No. CE20155050). This work made use of the resources supported by the NSFC-Guangdong Mutual Funds for Super Computing Program (2nd Phase), and 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.
[25] Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schrot GP, Gunderson KL, et al. High density DNA methylation array with single CpG site resolution. Genomics 2011; 98(4):288-295; PMID:21839163; http://dx.doi.org/10.1016/j.ygeno.2011.07.007

[26] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009; 462 (7271):315-322; PMID:19829295; http://dx.doi.org/10.1038/nature08514

[27] Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Research 2005; 33(18):5868-5877; PMID:16224102; http://dx.doi.org/10.1093/nar/gki901

[28] Guo H, Zhu P, Yan L, Li R, Hu B, Lian Y, Yan J, Ren X, Lin S, Li J, et al. The DNA methylation landscape of human early embryos. Nature 2014; 511(7511):606-610; PMID:25079557; http://dx.doi.org/10.1038/nature13544

[29] Akalin A, Garrett-Bakelman FE, Kormaksson M, Busuttil J, Zhang L, Khrebtukova I, Milne TA, Huang Y, Biswas D, Hess JL, et al. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. PLoS Genet 2012; 8(6): e1002781; PMID:22737091; http://dx.doi.org/10.1371/journal.pgen.1002781

[30] Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biology 2012; 13 (10):R87; PMID:23034086; http://dx.doi.org/10.1186/gb-2012-13-10-r87

[31] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth 2012; 9(4):357-359; http://dx.doi.org/10.1038/nmeth.1923

[32] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics 2009; 25(16):2078-2079; PMID:19505943; http://dx.doi.org/10.1093/bioinformatics/btp352

[33] Barnett D, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics 2011; 27 (12):1691-1692; PMID:21493652; http://dx.doi.org/10.1093/bioinformatics/btr174

[34] Anders S, Huber W. Differential expression analysis for sequence count data. Gen Biol 2010; 11(10):R106; http://dx.doi.org/10.1186/gb-2010-11-10-r106