Mining methylome databases

Colm J. O’Rourke¹, Therese M. Murphy ², Donal Hollywood¹, and Antoinette S. Perry¹

¹Prostate Molecular Oncology, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland
²Department of Psychiatry and Mental Health Research, St. Vincent’s University Hospital, and School of Medicine & Medical Science, University College Dublin, Dublin, Ireland

Teaser: As methyomics technologies progress, the internet provides a dynamic and under-used resource to distribute epigenetic data to scientists and clinicians.

Following publication of the DNA ‘hardware’ in the Human Genome Project in 2003, large-scale efforts have begun to elucidate the epigenomic ‘software’, held to underpin eukaryotic complexity. As the technology to catalog epigenetics advances, a pertinent question has emerged: how should such a vast wealth of data be organized and distributed to the scientific community? To this end, numerous freely available online methylation databases have been developed, encompassing normal health and development, cancer, and neuropsychiatric disorders, among others.

DNA methylation constitutes arguably the most fundamental epigenetic modification. The fifth base, 5-methylcytosine (5mC), blocks transcription factor binding both directly and by recruiting methyl-binding proteins, which interact with histone deacetylases to induce chromatin condensation. Hence, DNA methylation is a highly efficient transcriptional silencing mechanism. Methylphony patterns are cell type specific and stably maintained, providing cells with a mechanism to respond to specific environmental triggers. Given that each cell type in the human body has a unique epigenome, the breadth of comprehensive methylome mapping throughout populations of cells and individuals is daunting. In recent years, however, the alliance of techniques for traditional methylation analysis (e.g., bisulfite modification, restriction endonucleases, and methyl-specific antibodies) with next-generation sequencing and genome-wide arrays has made methylome mapping a reality.

‘The Internet? Is that thing still around?!”
– Homer J. Simpson

Already, vast amounts of raw data have been generated by individual studies and large-scale initiatives, such as the Roadmap Epigenomics of The National Institutes of Health (NIH), the Encyclopedia of DNA Elements (ENCODE), International Human Epigenome Consortium (IHEC), and International Cancer Genome Consortium (ICGC). However, disseminating such information to the scientific and medical community represents a key challenge. The first database to make methylation data publicly available on the internet, MethDB, emerged in 2001 [1] with subsequent reconfigurations optimizing its efficacy at maintaining and disseminating up-to-date data [2] (Table 1). Last updated in 2009, the database contains information on 20,236 submitted 5mC data sets and 6,312 submitted whole-genome profiles. Its ease of use, multivariate search capabilities, and ability to cross-reference other data sets, make this a powerful resource to interrogate fundamental methylation patterns.

However, with the rapid progress of epigenome-wide technology, the demands of methylation databases have evolved such that they must now accommodate vast influxes of methylomic data rather than simply housing results from a range of individual experiments. National Center for Biotechnology Information (NCBI) Epigenomics, powered by the Roadmap Epigenomics initiative, sought to identify existing Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) data relevant to epigenetics [3]. These data were reorganized into an epigenetics-centered database, comprising ‘studies’ (involving results from experiments with a shared goal, typically whole data sets) and ‘samples’ (results from experiments using the same source of biological materials).

Similarly, the Next-Generation Sequencing Methylation DataBase (NGSmethDB) offers a repository for the collection and retrieval of methylomic data [4]. Intriguingly, this repository has organized its user interface based on the practical appeals of epigenome-wide analysis: namely, one can obtain methylation values for particular chromosomal regions or tissues, analyze methylation at the level of the promoter for specific tissues, and compare methylation patterns across sets of tissues. Although currently only including information on human, mouse, and Arabidopsis, NGSmethDB appears capable of accommodating future influxes of large data sets.

Online oncology

It is now apparent that the methylome suffers numerous ‘hits’ during tumor evolution, leading to hypermethylation (at tumor suppressor promoters), hypomethylation (e.g., noncoding elements, retroviral oncogenes, protooncogene promoters, and pericentromeric regions), and a genome-wide decline in 5mC. Development of cancer-specific methylation databases would be of enormous benefit in elucidating epigenetic aberrations in tumorigenesis, providing potential new avenues for diagnosis, prognosis, and therapies.

Although repositories such as MethDB contain some information pertaining to malignant states, it is not possible to query multiple genes or cancer subtypes. Furthermore, no ranking annotation is assigned to genes or cancer subtypes. Hence, cancer methylation databases have been specifically developed.

Corresponding author: O’Rourke, C.J. (oroarkco@tcd.ie)
Spotlight

Table 1. Overview of query options, return options, and resource outsourcing available across different methylation databases*

<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
<th>Input</th>
<th>Output</th>
<th>Cross-references</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>MethDB</td>
<td><a href="http://www.methdb.net">http://www.methdb.net</a></td>
<td>Species; sex; tissue; gene/locus; sequence ID; exp ID; phenotype; method; environment; data type</td>
<td>Species; sex; tissue; phenotype; method; methylation in DNA type; gene/Locus; analyzed sequence/array; expression; 5mC content</td>
<td>Human Genome Nomenclature Committee; Entrez; PubMed; EMBL; DDBJ; GenBank; OMIM; GDB; MGI; LocusLink; Rat Genome Database; any valid URL</td>
<td>[1,2]</td>
</tr>
<tr>
<td>NGSmeth</td>
<td><a href="http://bioinfo2.ugr.es/NGSmethDB/gbrowse">http://bioinfo2.ugr.es/NGSmethDB/gbrowse</a></td>
<td>Methylation type; data source; experiment; sample; position; gene; cancer; clone; repeat</td>
<td>Unmethylated contexts; differentially methylated contexts; promoter region methylation states; chromosomal region methylation states</td>
<td>PubMed; GEO; WikiGenes; dbSNP; GIRO</td>
<td>[4]</td>
</tr>
<tr>
<td>PubMeth</td>
<td><a href="http://www.pubmeth.org">http://www.pubmeth.org</a></td>
<td>Gene; cancer type and/or subtype; gene and cancer</td>
<td>PubMed ID; cancer type; method; methylation frequency (primary, cell lines, normal); number of references; number of samples; overall methylation frequency</td>
<td>Gene Cards; PubMed; DBTSS</td>
<td>[5]</td>
</tr>
<tr>
<td>MethyCancer</td>
<td><a href="http://methycancer.psych.ac.cn/">http://methycancer.psych.ac.cn/</a></td>
<td>Methylation type; data source; experiment; sample; position; gene; cancer; clone; repeat</td>
<td>Methylation type; methylation status; sample information; related gene and repeat</td>
<td>PubMed; Ensembl; COSMIC</td>
<td>[6]</td>
</tr>
<tr>
<td>MethyloMeDB</td>
<td><a href="http://epigenomics.columbia.edu/methylomedb/">http://epigenomics.columbia.edu/methylomedb/</a></td>
<td>Organism; anatomical region; chromosomal region</td>
<td>Methylation probability; sequence read coverage; chromosomal mapping</td>
<td>PubMed; UCSC Genome Browser</td>
<td>[8]</td>
</tr>
<tr>
<td>DiseaseMeth</td>
<td><a href="http://bioinfo.hrbmu.edu.cn/diseasemeth">http://bioinfo.hrbmu.edu.cn/diseasemeth</a></td>
<td>Genes; chromosome; disease; experimental platform; sample type; cell line; data type</td>
<td>Methylation level; no. relevant data in database; significance of methylation difference between disease and normal data sets</td>
<td>PubMed; GeneCards</td>
<td>[11]</td>
</tr>
</tbody>
</table>

*Abbreviations: COSMIC, Catalogue of Somatic Mutations in Cancer; dbSNP, Single Nucleotide Polymorphism Database; DBTSS, DataBase of Transcriptional Start Sites; DDBJ, DNA DataBank of Japan; EMBL, European Molecular Biology Laboratory; GDB, Gene DataBase; GIRO, Genetic Information Research Institute; MGI, Mouse Genome Informatics; OMIM, Online Mendelian Inheritance in Man.

One relatively simple approach has been adopted by PubMeth, which uses text mining coupled to information-retrieval software to allow rapid literature interrogation [5]. Facilitating gene-centric, cancer-centric, and combination searches, PubMeth provides an impressive resource spanning 43 cancer types and returning information on methylation frequency, sample size, and number of references.

Although text-mining approaches are undoubtedly useful, substantial variables exist between studies (i.e., sample source, methodology, and techniques) that make statistical analyses difficult. Further efforts are needed to integrate methylation data with genetic and pathophysiological contexts. MethyCancer integrates data on methylation, cancer genes, mutations, and cancer pathology with CpG island (CGI) sequencing data [6]. Apart from reconciling current methylation data with genetic and clinical correlates, the developers also hope to use the database as a platform for participants in the Cancer Epigenomics Project in China to share data.

**Neuropsychiatric epigenetics and methyloMe DB**

Mounting evidence now suggests that environmentally induced changes to methylation patterns result in behavioral abnormalities [7]. Moreover, studies have identified a prominent role for aberrant promoter methylation-driven gene inactivation in a range of neuropsychiatric disorders [8]. Neuropsychiatric epigenetic research would benefit from both a text-mining database, similar to PubMeth, and the development of a comprehensive methyloMe repository for psychiatric disorders to facilitate elucidation of epigenetic abnormalities and the identification of novel biomarkers and therapeutic targets. MethyloMeDB is currently the only psychiatric-specific methylation database, containing post-mortem human and mouse genome-wide brain DNA methylation profiles (at single-CpG resolution), relating to depression and schizophrenia with nonpsychiatric controls [9]. Interestingly, it also enables comparisons across different regions of the brain.

**All for one and one for all...?**

Given its pervasive nature across different systems and states and its role in disease, methylation has come into focus within essentially all medical fields.

Although individual studies within a given field, such as mapping the breast cancer methyloMe [10], are undoubtedly promising, current costs of such analyses represent a significant limiting factor for the standard-sized research
group. With rapidly advancing technologies, however, and concomitant falling prices in deep sequencing and array-based technologies, this is likely to change in the near future. As an interim solution, an alternative approach is to integrate independent results into a central repository. In publicizing data, costs per team could be minimized by rapid, shared in-depth data mining and validation (or invalidation). This raises the questions of when and, more importantly, how and where should data sets be made publicly available?

DiseaseMeth has been developed with such an integrated approach in mind, fluidly combining small-scale experimental studies with large-scale sequencing and microarray data into one central database subdivided by disease etiology [11]. This repository encompasses 72 diseases stored across six classifications (‘cancer’, ‘metabolic’, ‘genetic’, ‘autoimmune’, ‘neurological’, and ‘other’) from a variety of biological sources (disease tissue and cell lines), including normal biomaterial for reference purposes. Developed in 2011, DiseaseMeth is still in an early stage, and significant amounts of additional raw data are required for each disease group, which relies on researchers to endorse this integrative, collaborative approach.

The methylome mountain and beyond
A confounding issue in deciphering the methylome is the recently observed phenomenon of non-CpG methylation (i.e., methylation at CpA, CpT, and CpC). Intriguingly, evidence suggests that methylation of up to 25% of cytosines occurs in a non-CpG context [12]. Novel detection methods are required to evaluate non-CpG methylation to obtain a clear and comprehensive picture of the methylome. Techniques such as single-molecule real-time (SMRT) sequencing have been developed for this purpose. In addition to non-CpG methylation, the recent development of a method to distinguish 5-hydroxymethylcytosine (5hmC) from 5mC (ameliorating a substantial shortcoming of previous bisulphite-based technologies) will undoubtedly add another layer of complexity to the methylome [13].

Online databases provide an excellent resource for researchers to retrieve and share large amounts of raw data easily. However, these repositories must be adaptable to encompass novel facets of the methylome and accommodate growing amounts of data as technology progresses. Although the internet represents a constantly evolving resource that aims to meet the ever-changing needs of its users, integrating data derived from multiple sample types and sources generated through varying experimental techniques presents a significant challenge.

A system (utilized by certain databases, such as methDB, NCBI Epigenomics, NGSmethDB, MethylomeDB, and DiseaseMeth) whereby the researcher uploads his/her own raw and normalized data sets offers the advantage that explicit experimental details can be compiled from independent studies, which would not be possible by simple text mining. Nonetheless, substantial variables still exist between studies and there is no general consensus on how to assign rank or weighting to minimize this.

Two new virtual tools have recently been designed to facilitate the sharing and interpretation of the contents of online repositories. EpíExplorer combines data on DNA methylation, histone modifications, and 5’ regulatory regions among others [14], whereas methyliKit is an R software package for DNA methylation analysis from high-throughput sequencing [15]. These are likely to be the first in a long line of in silico analysis packages as the epigenetics community struggles to make sense of the fruits of their labor.

Ultimately, the methylome is only one piece of the epigenome puzzle. Following suit, online databases for histone modifications are beginning to emerge, such as the Compendium of Protein Lysine Acetylation (CPLA) 1.0 and PhosphoSite Plus. Once established, the information contained within these databases may then be merged together, revealing the completed epigenomic jigsaw.

References