DNAmod: the DNA modification database

Ankur Jai Sood[1,2,†], Coby Viner[2,3,†], and Michael M. Hoffman[1–4,*]

[1]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
[2]Princess Margaret Cancer Centre, Toronto, ON, Canada
[3]Department of Computer Science, University of Toronto, Toronto, ON, Canada
[4]Vector Institute, Toronto, ON, Canada

September 7, 2018

Abstract

Covalent DNA modifications, such as 5-methylcytosine (5mC), are increasingly the focus of numerous research programs. In eukaryotes, both 5mC and 5-hydroxymethylcytosine (5hmC) are now recognized as stable epigenetic marks, with diverse functions. Bacteria, archaea, and viruses contain various other modified DNA nucleobases. Numerous databases describe RNA and histone modifications, but no database specifically catalogues DNA modifications, despite their broad importance in epigenetic regulation. To address this need, we have developed DNAmod: the DNA modification database.

DNAmod is an open-source database (https://dnamod.hoffmanlab.org) that catalogues DNA modifications and provides a single source to learn about their properties. DNAmod provides a web interface to easily browse and search through these modifications. The database annotates the chemical properties and structures of all curated modified DNA bases, and a much larger list of candidate chemical entities. DNAmod includes manual annotations of available sequencing methods, descriptions of their occurrence in nature, and provides existing and suggested nomenclature. DNAmod enables researchers to rapidly review previous work, select mapping techniques, and track recent developments concerning modified bases of interest.

Introduction

A rapidly growing body of research is continuing to reveal numerous gene-regulatory effects of covalent DNA modifications, such as 5-methylcytosine (5mC). We now recognize 5mC as a stable epigenetic mark and as having diverse functions beyond transcriptional repression[11]. An increasing number of studies demonstrate the importance of other cytosine modifications, such as 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC)[2,8,25,41,44]. More recently, three analogous modifications of thymine were found to occur in mammals[36,51] and can now largely be sequenced[18]. N^6-methyladenine, previously thought to mainly occur as an RNA modification in eukaryotes, has now been found in the DNA of multiple eukaryotes[23]. Bacteria, archaea, and especially bacteriophages have long been known to harbor a diverse array of modified bases[17,49]. Their genomes can also have hypermodified bases—modified DNA bases that substitute for the unmodified base in many positions genome-wide[16,49].

Multiple databases profile RNA modifications[3,7,52] and human histone modifications[54], but no database catalogues DNA modifications systematically. Some databases include particular classes of DNA modifications[42]. These include restriction endonucleases and DNA methyltransferases in

*Correspondence: michael.hoffman@utoronto.ca
†Ankur Jai Sood and Coby Viner contributed equally to this work

REBASE[39]; methylation databases, like MethDB[1]; databases including DNA metabolic pathways, such as KEGG[26]; and those focused on DNA damage and repair, like REPAIRtoire[29].

Since DNA modifications are a key aspect of epigenetic regulation, there is a pressing need to organize them in a single location. We have accordingly created DNAmod: the DNA modification database (https://dnamod.hoffmanlab.org). DNAmod is the first database to comprehensively catalogue DNA modifications and provides a single resource to launch an investigation of their properties.

Database construction and visualization

DNAmod consists of two components: a relational database back-end and a web interface front-end. We used the Chemical Entities of Biological Interest (ChEBI) database[12,21] to seed the DNAmod database. We imported a nucleobase-related subset of ChEBI, consisting of chemical entities and related annotations. We performed queries against the entities to construct a set of candidate DNA modifications for DNAmod, retaining most of these as a separate *unverified* set. Then, we filtered candidate entities into a manually curated set of *verified* DNA modifications, augmenting them with modification-specific annotations.

The web interface front-end allows users to either search or browse through the catalogue of DNA modifications, integrating ChEBI's information with our own.

Identifying candidate DNA modifications from ChEBI

DNAmod leverages ChEBI[21] to define a set of modified DNA candidates for inclusion and to add preliminary information for each candidate. ChEBI is a database of small biologically relevant molecules, which affect living organisms. We queried ChEBI via ChEBI Web Services[21]. We used Biopython[9] and the Python Simple Object Access Protocol (SOAP) client, suds[33], to query ChEBI and construct the DNAmod database.

ChEBI provides an ontology which encodes the relationships between its compounds. We used this ontology to precisely define the notion of parents and children, which we used to hierarchically retrieve and display modifications. We used two kinds of relationships for this purpose, both of which have associated symbols, defined by ChEBI[12]: $\mathcal{F}$ *has functional parent* and $\triangle$ *is a*. We used these relationships to find candidate DNA modifications, by identifying entities related to the core nucleobases, which we represent by their symbols: {`A`, `C`, `G`, `T`, `U`}. We included uracil, since many of its descendants in the ontology are modifications of thymine (`CHEBI:17821`, which is equivalent to 5-methyluracil), and are not annotated as descendants of thymine itself. For each of these bases, we imported all entities that are annotated in the ontology as a child of one of these bases, via the $\mathcal{F}$ *has functional parent* relationship. ChEBI ranks entities based on their degree of curation. We only imported entities with the highest rating—three stars—indicating manual curation by ChEBI. Whenever possible, we only included entities as nitrogenous bases (nucleobases). If ChEBI did not have the nucleobase, we then selected the nucleoside form and finally, if necessary, the nucleotide. These imported bases formed the candidate set of modifications (the unverified set), from which we created a curated set of DNA modifications (the verified set).

The ChEBI ontology does not generally encode $\mathcal{F}$ *has functional parent* relationships for nucleobases beyond the children of the unmodified nucleobases. It instead encodes modified nucleobases with an $\triangle$ *is a* relationship to their parent base. This is because descendant entities of specific modifications are generally subtypes of the class of modifications from which they originate. For example, 3-methyladenine $\triangle$ *is a* methyladenine. Methyladenine, however, $\mathcal{F}$ *has functional parent* adenine, since it is conceived of as possessing adenine as a characteristic group and as being derived via functional modification[12]. We therefore need to use both of these relationships, within the ChEBI ontology,

to accurately capture the full nucleobase hierarchy.

ChEBI also provides selected citations, associated with some of its entities. We retrieved the citations from ChEBI as PubMed IDs[30]. We used the Biopython[9] package `Bio.Entrez` to query the PubMed citation database, using NCBI's Entrez Programming Utilities[30]. We retrieved the details of each citation, and use them to construct a formatted citation. We currently support only publications indexed in PubMed.

Manual curation and annotation

We manually created a *whitelist*, which contains our curated (or verified) set of candidates that we deem DNA modifications. For each of these bases, we also imported all descendants with an eventual $\boxed{\mathcal{F}}$ *has functional parent* or $\triangle$ *is a* relationship with any of the members of the verified set. We expanded the verified set to include any bases recursively imported in this manner, since they were children of verified DNA nucleobases. This rule had one exception: we excluded any bases that possess an ancestor in our *blacklist* of non–DNA modifications.

We can formalize the above description of bases imported from the ChEBI ontology[12] and subsequent filtering as follows. Let $a \boxed{\mathcal{F}} b$ specify that a has the $\boxed{\mathcal{F}}$ *has functional parent* relationship with b. The definition of $\boxed{\mathcal{F}}$ is transitive: for all n entities, l_i, for $i = 0$ to $n-1$, between a and b,

$$a \boxed{\mathcal{F}} b \Leftrightarrow (a \boxed{\mathcal{F}} l_{n-1}) \wedge (l_i \boxed{\mathcal{F}} l_{i-1} \forall i \in (0, n)) \wedge (l_0 \boxed{\mathcal{F}} b)\,.$$

The analogous definitions hold for $\triangle$.

We call each l_i a *child* of l_{i-1} and call each l_{i-1} a *parent* of l_i. We refer to a as a *descendant* of b and refer to b as an *ancestor* of a. Let $\mathcal{C}$ represent the first level of children of the unmodified nucleobases, such that $\mathcal{C} = \{x \mid x \boxed{\mathcal{F}} y, y \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{U}\}\}$. Let $\mathcal{V} \subset \mathcal{C}$ represent the manually-annotated, verified proper subset of $\mathcal{C}$.

We manually curated a blacklist of excluded entities, $\mathcal{B}$, satisfying: $\mathcal{B} \subseteq \{b \mid (b \boxed{\mathcal{F}} p \vee b \triangle p)\,, p \in \mathcal{V}\}$. We imported the set of verified DNA modifications, $\mathcal{M}$, defined in set-builder notation with predicates, as:

$$\mathcal{M} = \mathcal{V} \cup \left\{z \;\middle|\; (\exists v \in \mathcal{V})\,(\forall b \in \mathcal{B})\left[\left(z \boxed{\mathcal{F}} v \vee z \triangle v\right) \wedge \neg\left(z \boxed{\mathcal{F}} b \vee z \triangle b\right)\right]\right\}.$$

Finally, we added a small number of bases manually, that do not have any of the DNA bases or uracil as a parent in their ontology, but are nonetheless notable modified bases, such as 2′-deoxyinosine.

We additionally provided two kinds of manual annotations: sequencing techniques and occurrence in nature, for each modified DNA base. We surveyed the literature of sequencing methods for covalent DNA modifications[5,28,35,37,43], and annotated the available methods for each base, providing curated citations. These annotations include the method's name, our categorizations of the basis for the method (such as chemical conversion), its resolution, and any further qualifier (Table 1A). Qualifiers include limitations (such as applicability to only some genomic regions), enrichment methods, and advantages (such as optimization for single-cell sequencing). We considered any method which involves affinity-based recognition of targets to be of "low" resolution[4]. These methods can also suffer from low specificity or antibody cross-reactivity[5]. Conversely, we annotated any methods based principally upon the detection of a chemically converted modification as "high" resolution. This generally reflects the resulting resolution of the method's output data and often corresponds to the necessity to bin genomic regions during downstream analyses of the detected analyte.

For each modified base, we investigated if it had been previously reported to occur *in vivo*. This included any endogenous occurrences, as well as those stimulated exogenously, such as from exposure to an environmental toxin. We annotated any modification observed *in vivo* as "natural". We additionally provided non-exhaustive examples of some organisms in which the modifications have been

Table 1. Possible annotations within DNAmod's curated (A) sequencing method data and (B) natural occurrence information. Each row lists a field and all terms ever used to annotate it. [square brackets]: optional prefixes. ⟨ angle brackets ⟩: description of term, rather than the complete enumeration provided for other terms.

(A) Sequencing method annotations

Field	Terms
Mapping method	⟨ method abbreviation ⟩
Method detail	affinity-based, chemical conversion, chemical conversion and immunoprecipitation, chemical tagging, direct detection, DNMT1 conversion, enzyme-mediated chemical tagging, excision repair enzyme-based, restriction endonuclease
Resolution	low, high, single-base
Qualifier	5hmU:G mismatch only, CpG contexts only, [low-input or] single-cell, [methylation-insensitive] restriction digestion, microarray probes, salt gradient stratification, specific fragments, strand-specific, target sequences

(B) Natural occurrence annotations

Field	Terms
Function[a]	damage, demethylation intermediate, [possible] epigenetic mark, hypermodified nucleobase, restriction-modification
Functional detail	[highly] cytotoxic, mutagenic, reactive oxygen species, specific transcriptional roles, transcription terminator
Origin	natural, synthetic, synthetic and RNA
Organism	⟨ binomial name ⟩

[a] Each row contains all possible instantiations of the field on the left, except that terms within the "Function" field are often combined, as conjunctions.

reported. We based these annotations on our ability to find evidence of *in vivo* occurrence, as opposed to publications describing only the synthesis or physicochemical properties of a nucleobase. For each of these annotations, we also briefly annotated a primary biological function, if known (Table 1B). For any modification not observed *in vivo*, we annotated it as "synthetic" and listed a reference pertaining to its synthesis or in which the synthetic base was used.

We entered these annotations in two annotation source files (Table 1), which we later imported into our database. This decoupled them from the rest of our pipeline and allows outside experts to submit additions without requiring knowledge of our pipeline or programming workflow.

DNAmod integrates manually-curated nomenclature, including the name and abbreviation deemed most consistent and in common use[8,10,27]. We additionally provide recommendations for one-letter symbols of selected modified bases, and in some instances for their base-pairing complements, as previously described[47]. The DNAmod web interface displays recommended notation in an organized table (Figure 1).

Recommended notation

Name	5-formylcytosine
Abbreviation	5fC
Symbol	f
Complement	guanine:5-formylcytosine
Symbol	3

Mapping techniques

Method	Method detail	Resolution	Qualifier	References
CLEVER-seq	chemical tagging	single-base	single-cell	Zhu, C, et al. 2017. Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell*. 20(5).
MAB-seq	chemical conversion	single-base		Wu, H, et al. 2016. Base-resolution profiling of active DNA demethylation using MAB-seq and caMAB-seq. *Nature Protocols*. 11(6).
MAB-seq	chemical conversion	single-base	low-input or single-cell	Wu, X, et al. 2017. Simultaneous mapping of active DNA demethylation and sister chromatid exchange in single cells. *Genes & Development*. 31(5).
Pvu-seal-seq	enzyme-mediated chemical tagging	single-base		Sun, Z, et al. 2015. A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Molecular Cell*. 57(4).
fC-CET	chemical conversion	single-base		Xia, B, et al. 2015. Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nature Methods*. 12(11).
fCAB-seq	chemical conversion	single-base		Song, CX, et al. 2013. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*. 153(3).
fluorogenic labeling	chemical tagging	single-base		Liu, C, et al. 2017. Fluorogenic labeling and single-base resolution analysis of 5-formylcytosine in DNA. *Chemical Science*. 8(11).
redBS-seq	chemical conversion	single-base		Booth, MJ, et al. 2014. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nature Chemistry*. 6(5).

Nature

Origin	Function	Functional detail	Organisms	References
natural	demethylation intermediate and epigenetic mark		*Homo sapiens* *Mus musculus*	Song, CX, et al. 2013. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*. 153(3). Song, CX, et al. 2013. Potential functional roles of DNA demethylation intermediates. *Trends in Biochemical Sciences*. 38(10). Booth, MJ, et al. 2014. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nature Chemistry*. 6(5). Lu, X, et al. 2015. Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Research*. 25(3). Bachman, M, et al. 2015. 5-Formylcytosine can be a stable DNA modification in mammals. *Nature Chemical Biology*. 11(8). Iurlaro, M, et al. 2016. In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biology*. 17(1).

Figure 1. Manually-curated recommended notation, mapping techniques, and natural occurrence data for 5-formylcytosine (5fC). See Table 1 for an explanation of the mapping and natural occurrence table headers.

We store all data, either imported from ChEBI or from our manual annotations, within a SQLite[24] database, used via the Python `sqlite3` package[15].

Website generation

We created a static website to display and provide navigation for the information contained within the database. We generated it by formatting the database content using the templating engine Jinja2[40]. Two templates were sufficient to generate all HTML files. We used a single template for all modification pages and another for the homepage. We also recorded the date of the most recent update to the database. All web pages use the Bootstrap[34] framework, which provides a standardized, portable, and mobile-compatible viewing format. We visualized the chemical structure of each compound from its Simplified Molecular-Input Line-Entry System (SMILES)[50] data, if available from ChEBI, as a vector graphic. We did this using the cheminformatics toolkit Open Babel[32], via its Python wrapper Pybel[31].

Searching and navigation

DNAmod makes modifications accessible via three main navigation options, each provided on a tab of the DNAmod homepage. First, users may search for modifications by several fields. Second, users may find curated DNA modifications via a pie menu[6]. Third, users may find candidate entities as a list, categorized by their parent unmodified nucleobases.

Client-side search functionality provides a means of rapidly finding bases with differing nomenclature (Figure 2A), while maintaining a static web page. This functionality relies on the elasticlunr.js JavaScript module[45]. Searches match to multiple fields: common or International Union of Pure and Applied Chemistry (IUPAC) names, all synonyms, any assigned abbreviation, and recommended notation symbol, when available. DNAmod displays curated DNA modifications in green, and others in magenta. The search results provide the field matched by the query, such as "abbreviation", along with the common name of the associated hit.

Alternatively, users may browse the modifications in DNAmod through a pie menu[6] interface (Figure 2B). This interface hierarchically arranges the bases according to their structure within the ChEBI ontology. The innermost ring consists of the four unmodified DNA bases, with an additional "other" category. This category encapsulates modified bases found in DNA, but which are not modifications of one of the four DNA bases. Consecutive outer rings represent children of the previous base or category. We demarcated natural versus synthetic bases by colouring natural bases in teal and synthetic bases in grey.

DNAmod structure and content

Individual modification pages visually represent the data contained within the backing database. We standardize and display all modifications in an identical format. DNAmod may omit some information, however, depending upon the extent of ChEBI's annotations and whether the page describes a verified DNA modification or merely a candidate entry.

Modification pages begin with a header displaying the DNA modification's ChEBI name. The top-right corner of the page lists the unmodified ancestor of the modification. For example, 5-hydroxymethyluracil is a modification of thymine (Figure 3), whereas 6-dimethyladenine is a modification of adenine.

(A)

(B)

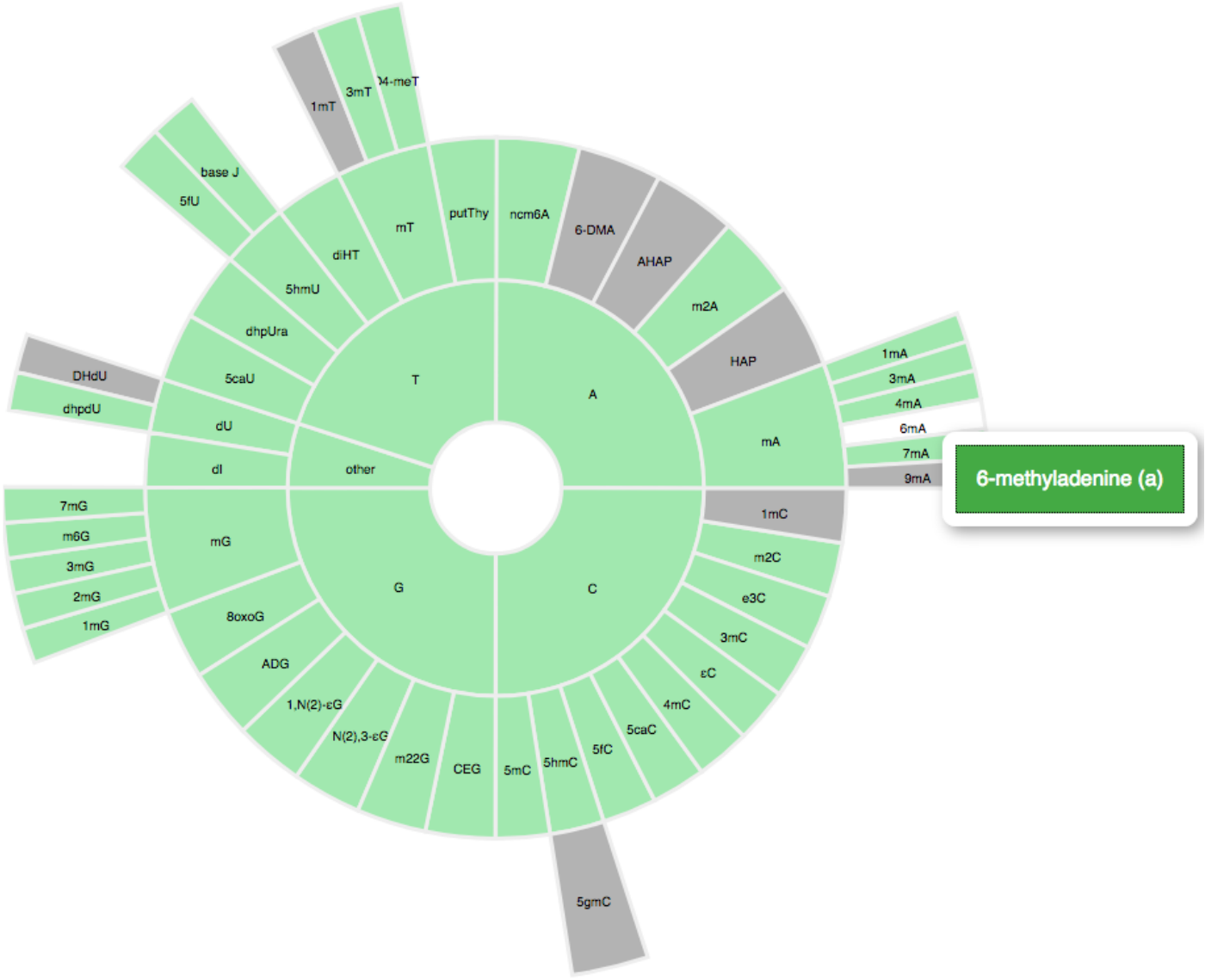

Figure 2. Finding 6-methyladenine by (A) searching for its abbreviation "6mA" or (B) via the pie menu.

Each modification begins with a short textual description of its chemistry, followed by a table containing its chemical properties. We import these from ChEBI, which provides their chemical formula, net charge, and average mass.

We annotate entities with all names available from ChEBI, including: their IUPAC name, SMILES[50] string, International Chemical Identifier (InChI) and hashed InChIKey[22] strings, and common synonyms. We also provide a recommended abbreviation and in some instances a suggested single-letter symbol for bioinformatic purposes, from our proposed expanded alphabet[47] (Figure 3).

We provide literature annotations for many DNA modifications, focusing upon those observed *in vivo*. We provide a list of methods that have been used to map the genomic locations of a modification (see above). We additionally provide information on a modification's occurrence, either naturally or only synthetically, where applicable, including some organisms in which it has been observed *in vivo* (see above). Finally, each page ends with the ChEBI database reference and a ChEBI-derived list of related literature citations (Figure 3). Our website has semantic web support, making use of the Resource Description Framework in Attributes (RDFa)[38] technique, augmented by Chemical Information Ontology (CHEMINF)[19] and PubChemRDF[14] Semanticscience Integrated Ontology (SIO)[13] annotations—providing machine-readable descriptions of key website features.

Discussion

DNAmod enables researchers to rapidly obtain information on covalently modified DNA nucleobases and assist those interested in profiling a modification. It additionally provides a reference toward standardization of modified base nomenclature and offers the potential to track recent developments within the field. We have kept DNAmod up to date for 3 yr and expect to continue to maintain it, particularly as new discoveries about DNA modifications are made. We also hope that DNAmod will serve to highlight underappreciated modifications that may have substantial biological importance.

The nomenclature used to describe a particular DNA modification is often inconsistent, with some early efforts toward standardization of particular classes[10,27]. The ChEBI name, for instance, often corresponds to the common chemical name of the compound, which is occasionally distinct from its common name within the biological literature, in the context of a DNA modification. We address this and attempt to encourage standardization by endeavouring to ensure that other names are annotated, while providing specific nomenclature recommendations. In particular, the suggested name of verified DNA modifications, as displayed on the homepage and within the recommended notation section, is always manually-curated and sometimes differs from the name assigned by ChEBI.

Our database, like many others, relies upon the ChEBI ontology. Like any large and complex endeavour, curating ChEBI is a substantial undertaking, requiring protracted deployment of expertise and effort. While ChEBI has a dedicated team of expert curators, who assiduously and continually improve ChEBI, their resources are naturally limited. Accordingly, while ChEBI has an issue tracker where we and others can suggest changes, revisions to ChEBI are highly dependent on user reports and the team's available bandwidth. A recent study found that ChEBI contains a non-negligible fraction of errors and omissions, across most entity categories[53]. Such errors naturally propagate to its downstream databases, including our own. While we have made efforts to further curate data and report relevant issues back upstream, we do inherit some errors and limitations. As in any project of this nature, we surely have our own errors and omissions. We lack a dedicated curator; accordingly, we curate this data on a best-effort basis. DNAmod has its own issue tracker, and we would appreciate if users could report any of our own errors or omissions, so that we can address them or facilitate reporting them upstream.

The inclusion of assays available to sequence different DNA modifications provides a means of assessing and selecting a sequencing method. It additionally attempts to track sequencing methods over

5-hydroxymethyluracil: DNAmod — A modification of thymine

Description

A primary alcohol that is uracil bearing a hydroxymethyl substituent at the 5-position.

Chemical properties

Chemical formula	Net charge	Average mass
C5H6N2O3	0	142.11282

Recommended notation

Name	5-hydroxymethyluracil
Abbreviation	5hmU
Symbol	g

Nomenclature

IUPAC	SMILES	InChI	InChIKey	Synonyms
5-(hydroxymethyl)pyrimidine-2,4(1H,3H)-dione	OCc1c[nH]c(=O)[nH]c1=O	InChI=1S/C5H6N2O3/c8-2-3-1-6-5(10)7-4(3)9/h1,8H,2H2,(H2,6,7,9,10)	JDBGXEHEIRGOBU-UHFFFAOYSA-N	5-(hydroxymethyl)uracil 5-hydroxymethyl uracil 5-(hydroxymethyl)-2,4(1h,3h)-pyrimidinedione

Mapping techniques

Method	Method detail	Resolution	Qualifier	References
Hardisty-labelling	chemical tagging	single-base	target sequences	Hardisty, RE, et al. 2015. Selective Chemical Labeling of Natural T Modifications in DNA. *Journal of the American Chemical Society*. 137(29).
SMRT	direct detection	single-base	target sequences	Clark, TA, et al. 2011. Direct detection and sequencing of damaged DNA bases. *Genome Integrity*. 2.
Yu-labelling	chemical tagging	single-base	5hmU:G mismatch only	Yu, M, et al. 2014. Detection of mismatched 5-hydroxymethyluracil in DNA by selective chemical labeling. *Methods (San Diego, Calif.)*. 72.

Nature

Origin	Function	Functional detail	Organisms	References
natural	damage, demethylation intermediate, and possible epigenetic mark		*Gyrodinium cohnii* *Homo sapiens* *Mus musculus* *Rattus norvegicus*	Rae, PM, et al. 1973. 5-Hydroxymethyluracil in the DNA of a dinoflagellate. *Proceedings of the National Academy of Sciences of the United States of America*. 70(4). Cathcart, R, et al. 1984. Thymine glycol and thymidine glycol in human and rat urine: a possible assay for oxidative DNA damage. *Proceedings of the National Academy of Sciences of the United States of America*. 81(18). Ravanat, JL, et al. 1999. Simultaneous determination of five oxidative DNA lesions in human urine. *Chemical Research in Toxicology*. 12(9). Pfaffeneder, T, et al. 2014. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nature Chemical Biology*. 10(7).

Database references: CHEBI:16964

Citations

Bianchini, F, et al. 1998. Urinary excretion of 5-(hydroxymethyl) uracil in healthy volunteers: effect of active and passive tobacco smoke. *International Journal of Cancer*. 77(1).
Cathcart, R, et al. 1984. Thymine glycol and thymidine glycol in human and rat urine: a possible assay for oxidative DNA damage. *Proceedings of the National Academy of Sciences of the United States of America*. 81(18).
Chen, HJ, et al. 2005. Measurement of urinary excretion of 5-hydroxymethyluracil in human by GC/NICI/MS: correlation with cigarette smoking, urinary TBARS and etheno DNA adduct. *Toxicology Letters*. 155(3).
Clark, TA, et al. 2011. Direct detection and sequencing of damaged DNA bases. *Genome Integrity*. 2.
Djuric, Z, et al. 1991. Quantitation of 5-(hydroxymethyl)uracil in DNA by gas chromatography with mass spectral detection. *Chemical Research in Toxicology*. 4(6).
Frenkel, K, et al. 1985. Quantitative determination of the 5-(hydroxymethyl)uracil moiety in the DNA of gamma-irradiated cells. *Biochemistry*. 24(17).
Hardisty, RE, et al. 2015. Selective Chemical Labeling of Natural T Modifications in DNA. *Journal of the American Chemical Society*. 137(29).
Klungland, A, et al. 2001. 5-Formyluracil and its nucleoside derivatives confer toxicity and mutagenicity to mammalian cells by interfering with normal RNA and DNA metabolism. *Toxicology Letters*. 119(1).
Kow, YW, et al. 2002. Repair of deaminated bases in DNA. *Free Radical Biology & Medicine*. 33(7).
LaFrancois, CJ, et al. 1998. Quantification of 5-(hydroxymethyl)uracil in DNA by gas chromatography/mass spectrometry: problems and solutions. *Chemical Research in Toxicology*. 11(7).
Pfaffeneder, T, et al. 2014. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nature Chemical Biology*. 10(7).
Rae, PM, et al. 1973. 5-Hydroxymethyluracil in the DNA of a dinoflagellate. *Proceedings of the National Academy of Sciences of the United States of America*. 70(4).
Ravanat, JL, et al. 1999. Simultaneous determination of five oxidative DNA lesions in human urine. *Chemical Research in Toxicology*. 12(9).
Yu, M, et al. 2014. Detection of mismatched 5-hydroxymethyluracil in DNA by selective chemical labeling. *Methods (San Diego, Calif.)*. 72.

About Home Contact

Figure 3. The full modification page for 5-hydroxymethyluracil (5hmU).

time, as resolution improves, and especially to highlight recent developments, like direct-detection of various modifications via nanopore sequencing[48]. The sequencing annotations we provide annotate nucleobases which are directly elucidated by the method and only for the base or set of bases which the method independently maps. This includes those that are obtained in addition to another nucleobase. For instance, confounded mixtures are often obtained. For example, 5mC and 5hmC cannot be distinguished with only conventional bisulfite sequencing. Alternatively, some methods have the capacity to independently resolve between modifications, such as various nanopore-based methods. Therefore, while many use oxidative bisulfite sequencing (oxBS-seq) in combination with conventional bisulfite sequencing to elucidate 5hmC via subtraction, we only annotated it as a sequencing method for 5mC, which it directly elucidates[5]. Conversely, we only annotate TET-assisted bisulfite sequencing (TAB-seq) under 5hmC, which it directly elucidates[5], although many use it to also detect 5mC.

We demarcated bases found to occur *in vivo*, providing examples of organisms in which a modification has been found, along with associated citations. This merely substantiates its *in vivo* presence, however. We did not attempt to comprehensively list the organisms which contain any particular modification. Finally, we expect our brief annotations of the biological roles of various DNA modifications to change as further research is conducted.

Future work

We plan to keep DNAmod updated continuously, manually reviewing newly added ChEBI compounds, requesting appropriate additions to ChEBI, and curating any improvements. We also endeavour to annotate recently developed sequencing methods as we come across them.

Integrating additional external databases will further increase DNAmod's utility. In particular, we envision potential integration with domain-specific DNA modification databases, such as those cataloguing compounds formed from the operation of particular biological pathways. For instance, modifications involved in DNA damage and repair could be linked to REPAIRtoire[29] data.

We used ChEBI Web Services[21] to obtain information from their database. ChEBI has, however, recently released a Python application programming interface (API), permitting us to directly access their data[46]. Switching from our current web-based queries to use of their API would likely result in a more robust system and expedite the database-building process.

Availability of data and materials

The DNAmod website, including a description and contact information, as well as the backing SQLite database, are freely available at: https://dnamod.hoffmanlab.org. Python source code, web assets, and an issue tracker for this project are available at: https://bitbucket.org/hoffmanlab/dnamod. Persistent availability is ensured by Zenodo, in which we have deposited the current version of our code (https://doi.org/10.5281/zenodo.640631) and SQLite database (https://doi.org/10.5281/zenodo.640561). All source code and web assets are licensed under a GNU General Public License, version 2 (GPLv2). DNAmod's data is licensed under a Creative Commons Attribution 4.0 International license (CC BY 4.0).

List of abbreviations

5caC 5-carboxylcytosine

5fC 5-formylcytosine

5hmC 5-hydroxymethylcytosine

5hmU 5-hydroxymethyluracil

5mC 5-methylcytosine

6mA 6-methyladenine

API application programming interface

ChEBI Chemical Entities of Biological Interest

CHEMINF Chemical Information Ontology

DNMT DNA methyltransferase

InChI International Chemical Identifier

IUPAC International Union of Pure and Applied Chemistry

oxBS-seq oxidative bisulfite sequencing

RDFa Resource Description Framework in Attributes

SIO Semanticscience Integrated Ontology

SMILES Simplified Molecular-Input Line-Entry System

TAB-seq TET-assisted bisulfite sequencing

TET Ten-eleven translocation enzyme

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceptualization, M.M.H; Methodology, A.J.S., C.V., and M.M.H; Software, A.J.S. and C.V.; Resources, M.M.H; Data Curation, A.J.S. and C.V.; Writing — Original Draft, A.J.S. and C.V.; Writing — Review & Editing, A.J.S., C.V., and M.M.H; Visualization, A.J.S., C.V., and M.M.H; Funding Acquisition, M.M.H; Supervision, C.V. and M.M.H.

Acknowledgments

We thank Daniel D. De Carvalho and Christopher E. Mason for helpful feedback on early versions of DNAmod. We thank the creators of ChEBI[12], and all those who have worked to improve it[20,21,46]. In particular, we thank Gareth Owen, Steve Turner, and Marcus Ennis for actively responding to curation requests and Venkatesh Muthukrishnan for managing ChEBI issues. We thank Egon L. Willighagen for useful suggestions in a PubPeer review of an early version of this work. We thank Carl Virtanen, Qun Jin, and Zhibin Lu for technical assistance.

Funding

This work was supported by the University of Toronto Undergraduate Research Opportunities Program (to A.J.S.), the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03948 to M.M.H. and Alexander Graham Bell Canada Graduate Scholarships to C.V.), the Canadian Institutes of Health Research (201512MSH-360970 to M.M.H.), the Canadian Cancer Society (703827 to M.M.H.), the Ontario Ministry of Training, Colleges and Universities (Ontario Graduate Scholarships to C.V.), the Ontario Institute for Cancer Research through funding provided by the Government of Ontario (CSC-FR-UHN to John E. Dick), the Ontario Ministry of Research, Innovation and Science (ER-15-11-223 to M.M.H.), the University of Toronto McLaughlin Centre (MC-2015-16 to M.M.H.), and the Princess Margaret Cancer Foundation.

References

[1] Amoreira C, Hindermann W, Grunau C (2003) An improved version of the DNA methylation database (MethDB). Nucleic Acids Res 31:75–77, https://doi.org/10.1093/nar/gkg093

[2] Bachman M, Uribe-Lewis S, Yang X, Burgess HE, Iurlaro M, Reik W, Murrell A, Balasubramanian S (2015) 5-formylcytosine can be a stable DNA modification in mammals. Nat Chem Biol 11:555–557, https://doi.org/10.1038/nchembio.1848

[3] Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, Wirecki TK, de Crécy-Lagard V, Ross R, Limbach PA, Kotter A, Helm M, Bujnicki JM (2018) MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res 46:D303–D307, https://doi.org/10.1093/nar/gkx1030

[4] Booth MJ, Marsico G, Bachman M, Beraldi D, Balasubramanian S (2014) Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. Nat Chem 6:435–440, https://doi.org/10.1038/nchem.1893

[5] Booth MJ, Raiber EA, Balasubramanian S (2015) Chemical methods for decoding cytosine modifications in DNA. Chem Rev 115:2240–2254, https://doi.org/10.1021/cr5002904

[6] Callahan J, Hopkins D, Weiser M, Shneiderman B (1988) An empirical comparison of pie vs. linear menus. In: O'Hare JJ (ed) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp 95–100, https://doi.org/10.1145/57167.57182

[7] Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, Zhang X, Vendeix FAP, Fabris D, Agris PF (2011) The RNA modification database, RNAMDB: 2011 update. Nucleic Acids Res 39:D195–D201, https://doi.org/10.1093/nar/gkq1028

[8] Chen K, Zhao BS, He C (2016) Nucleic acid modifications in regulation of gene expression. Cell Chem Biol 23:74–85, https://doi.org/10.1016/j.chembiol.2015.11.007

[9] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423, https://doi.org/10.1093/bioinformatics/btp163

[10] Cooke MS, Loft S, Olinski R, Evans MD, Bialkowski K, Wagner JR, Dedon PC, Møller P, Greenberg MM, Cadet J (2010) Recommendations for standardized description of and nomenclature concerning oxidatively damaged nucleobases in DNA. Chem Res Toxicol 23:705–707, https://doi.org/10.1021/tx1000706

[11] Dantas Machado AC, Zhou T, Rao S, Goel P, Rastogi C, Lazarovici A, Bussemaker HJ, Rohs R (2014) Evolving insights on how cytosine methylation affects protein-DNA binding. Brief Funct Genomics 14:61–73, https://doi.org/10.1093/bfgp/elu040

[12] Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36:D344–D350, https://doi.org/10.1093/nar/gkm791

[13] Dumontier M, Baker CJ, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, Del Rio NR, Duck G, Furlong LI, Keath N, Klassen D, McCusker JP, Queralt-Rosinach N, Samwald M, Villanueva-Rosales N, Wilkinson MD, Hoehndorf R (2014) The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semantics 5:14, https://doi.org/10.1186/2041-1480-5-14

[14] Fu G, Batchelor C, Dumontier M, Hastings J, Willighagen E, Bolton E (2015) PubChemRDF: Towards the semantic annotation of PubChem compound and substance databases. J Cheminf 7:34, https://doi.org/10.1186/s13321-015-0084-4

[15] Gerhard H (2016) sqlite3. https://docs.python.org/2/library/sqlite3.html

[16] Gommers-Ampt JH, Borst P (1995) Hypermodified bases in DNA. FASEB J 9:1034–1042, https://doi.org/10.1096/fasebj.9.11.7649402

[17] Grosjean H (2009) Nucleic acids are not boring long polymers of only four types of nucleotides: a guided tour. In: Grosjean H (ed) DNA and RNA Modification Enzymes: Structure, Mechanism, Function and Evolution, Landes Bioscience, Austin, TX, pp 1–18

[18] Hardisty RE, Kawasaki F, Sahakyan AB, Balasubramanian S (2015) Selective chemical labeling of natural T modifications in DNA. J Am Chem Soc 137:9270–9272, https://doi.org/10.1021/jacs.5b03730

[19] Hastings J, Chepelev L, Willighagen E, Adams N, Steinbeck C, Dumontier M (2011) The Chemical Information Ontology: provenance and disambiguation for chemical data on the biological semantic web. PLOS One 6(10):e25,513, https://doi.org/10.1371/journal.pone.0025513

[20] Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res 41:456–463, https://doi.org/10.1093/nar/gks1146

[21] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res 44:D1214–D1219, https://doi.org/10.1093/nar/gkv1031

[22] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) InChI, the IUPAC international chemical identifier. J Cheminf 7:23, https://doi.org/10.1186/s13321-015-0068-4

[23] Heyn H, Esteller M (2015) An adenine code for DNA: a second life for N6-methyladenine. Cell 161:710–713, https://doi.org/10.1016/j.cell.2015.04.021

[24] Hipp DR, Kennedy D, Mistachkin J (2000–2018) SQLite. https://www.sqlite.org

[25] Iurlaro M, McInroy GR, Burgess HE, Dean W, Raiber EA, Bachman M, Beraldi D, Balasubramanian S, Reik W (2016) In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. Genome Biol 17:141, https://doi.org/10.1186/s13059-016-1001-5

[26] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462, https://doi.org/10.1093/nar/gkv1070

[27] Khromov-Borisov NN (1997) Naming the mutagenic nucleic acid base analogs: the Galatea syndrome. Mutat Res 379:95–103, https://doi.org/10.1016/S0027-5107(97)00112-7

[28] Korlach J, Turner SW (2012) Going beyond five bases in DNA sequencing. Curr Opin Struct Biol 22:251–261, https://doi.org/10.1016/j.sbi.2012.04.002

[29] Milanowska K, Krwawicz J, Papaj G, Kosiński J, Poleszak K, Lesiak J, Osińska E, Rother K, Bujnicki JM (2011) REPAIRtoire—a database of DNA repair pathways. Nucleic Acids Res 39:D788–D792, https://doi.org/10.1093/nar/gkq1087

[30] NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 46:D8–D13, https://doi.org/10.1093/nar/gkx1095

[31] O'Boyle NM, Morley C, Hutchison GR (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. Chem Cent J 2:5, https://doi.org/10.1186/1752-153X-2-5

[32] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. J Cheminf 3:33, https://doi.org/10.1186/1758-2946-3-33

[33] Ortel J, Noehr J, van Gheem N (2011) suds. https://pypi.org/project/suds

[34] Otto M, Thornton J, Rebert C, Thilo J, XhmikosR, Fenkart H, Lauke PH, et al (2011–2018) Bootstrap. http://getbootstrap.com

[35] Pachter L (2013) *Seq. https://liorpachter.wordpress.com/seq/

[36] Pfaffeneder T, Spada F, Wagner M, Brandmayr C, Laube SK, Eisen D, Truss M, Steinbacher J, Hackner B, Kotljarova O, Schuermann D, Michalakis S, Kosmatchev O, Schiesser S, Steigenberger B, Raddaoui N, Kashiwazaki G, Müller U, Spruijt CG, Vermeulen M, Leonhardt H, Schär P, Müller M, Carell T (2014) Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. Nat Chem Biol 10:574–581, https://doi.org/10.1038/nchembio.1532

[37] Plongthongkum N, Diep DH, Zhang K (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet 15:647–661, https://doi.org/10.1038/nrg3772

[38] RDFa Working Group (2015) RDFa 1.1 primer - third edition. W3C Working Group Note, URL http://www.w3.org/TR/2015/NOTE-rdfa-primer-20150317/

[39] Roberts RJ, Vincze T, Posfai J, Macelis D (2015) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 43:D298–D299, https://doi.org/10.1093/nar/gku1046

[40] Ronacher A (2008) Jinja2 (the Python template engine). http://jinja.pocoo.org/

[41] Rothbart SB, Strahl BD (2014) Interpreting the language of histone and DNA modifications. Biochim Biophys Acta, Gene Regul Mech 1839:627–643, https://doi.org/10.1016/j.bbagrm.2014.03.001

[42] Rother K, Papaj G, Bujnicki JM (2009) Databases of DNA modifications. In: Grosjean H (ed) DNA and RNA Modification Enzymes: Structure, Mechanism, Function and Evolution, Landes Bioscience, Austin, TX, pp 622–623

[43] Song CX, Yi C, He C (2012) Mapping recently identified nucleotide variants in the genome and transcriptome. Nat Biotechnol 30:1107–1116, https://doi.org/10.1038/nbt.2398

[44] Song CX, Szulwach KE, Dai Q, Fu Y, Mao SQ, Lin L, Street C, Li Y, Poidevin M, Wu H, Gao J, Liu P, Li L, Xu GL, Jin P, He C (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. Cell 153:678–691, https://doi.org/10.1016/j.cell.2013.04.001

[45] Song W (2012–2018) Elasticlunr.js. http://elasticlunr.com

[46] Swainston N, Hastings J, Dekker A, Muthukrishnan V, May J, Steinbeck C, Mendes P (2016) libChEBI: an API for accessing the ChEBI database. J Cheminf 8:11, https://doi.org/10.1186/s13321-016-0123-9

[47] Viner C, Johnson J, Walker N, Shi H, Sjöberg M, Adams DJ, Ferguson-Smith AC, Bailey TL, Hoffman MM (2016) Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. bioRxiv 043794, https://doi.org/10.1101/043794

[48] Wallace EVB, Stoddart D, Heron AJ, Mikhailova E, Maglia G, Donohoe TJ, Bayley H (2010) Identification of epigenetic DNA modifications with a protein nanopore. Chem Commun 46:8195–8197, https://doi.org/10.1039/c0cc02864a

[49] Weigele P, Raleigh EA (2016) Biosynthesis and function of modified bases in bacteria and their viruses. Chem Rev 116:12,655–12,687, https://doi.org/10.1021/acs.chemrev.6b00114

[50] Weininger D (1988) SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. J Chem Inf Model 28(1):31–36, https://doi.org/10.1021/ci00057a005

[51] Wu H, Zhang Y (2014) Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell 156:45–68, https://doi.org/10.1016/j.cell.2013.12.019

[52] Xuan JJ, Sun WJ, Lin PH, Zhou KR, Liu S, Zheng LL, Qu LH, Yang JH (2018) RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res 46:D327–D334, https://doi.org/10.1093/nar/gkx934

[53] Yumak H, Chen L, Halper M, Zheng L, Perl Y, Elhanan G (2016) A quality-assurance study of ChEBI. In: Jaiswal P, Hoehndorf R, Arighi CN, Meier A (eds) Proceedings of the Joint International Conference on Biological Ontology and BioCreative, CEUR-WS.org, Corvallis, Oregon, USA, vol 1747, URL http://ceur-ws.org/Vol-1747/IT701_ICBO2016.pdf

[54] Zhang Y, Lv J, Liu H, Zhu J, Su J, Wu Q, Qi Y, Wang F, Li X (2010) HHMD: the human histone modification database. Nucleic Acids Res 38:D149–D154, https://doi.org/10.1093/nar/gkp968