bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DNAmod: the DNA modification database

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

[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

August 25, 2016

Abstract

Covalent DNA modifications, such as 5-methylcytosine (5mC), are increasingly the focus of numerous research programs. In eukaryotes, both 5mC and 5-hydroxymethylcytosine are now recognized as stable epigenetic marks, with diverse functions. Bacteria, archaea, and viruses contain various modified DNA nucleobases, including several in which one base is largely or entirely replaced by a particular covalent modification. Numerous databases describe RNA and histone modifications, but no database specifically catalogues DNA modifications, despite their broad importance as an element of epigenetic regulation. To address this need, we have developed DNAmod: the DNA modification database. DNAmod is an open-source database (http://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 its 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[1]. An increasing number of studies demonstrate the importance of other cytosine modifications, such as 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC)[2–6]. More recently, three analogous modifications of thymine were found to occur in mammals[7,8] and can now largely be sequenced[9]. N^6-methyladenine, previously thought to mainly occur as a RNA modification, has now been found in the DNA of multiple eukaryotes[10]. Bacteria, archaea, and especially bacteriophages have long been known to have a diverse array of modified[11] and hypermodified bases—modified DNA bases that largely or completely replace an unmodified base[12].

RNA modifications are profiled across multiple databases, including RNAMDB[13], MODOMICS[14], and RMBase[15]. Furthermore, histone modifications in humans are catalogued in HHMD[16]. Despite widespread recognition of DNA modifications as an important element of epigenetic regulation, no

*To whom correspondence should be addressed (michael.hoffman@utoronto.ca).
†These authors contributed equally to this work as first authors.

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

database exists to catalogue them. Some databases include particular classes of DNA modifications[17], such as restriction endonucleases and DNA methyltransferases in REBASE[18]; methylation databases, like MethDB[19]; databases including DNA metabolic pathways, such as KEGG[20]; and those focused on DNA damage and repair, like REPAIRtoire[21]. There is, however, a pressing need to focus upon DNA modifications from a broad perspective and organize them in a single location. In order to address this, we have created DNAmod: the DNA modification database (http://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 use the Chemical Entities of Biological Interest (ChEBI) database[22,23] to seed DNAmod, importing a nucleobase-related subset of its contents, consisting of chemical entities and related annotations. We perform 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 filter candidate entities into a curated set of verified DNA modifications, augmenting them with modification-specific annotations. Finally, we provide the ability 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 the ChEBI database[23] 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 query ChEBI via ChEBI Web Services[23]. We use Biopython[24] and the Python Simple Object Access Protocol client, suds[25], to query ChEBI and construct the DNAmod database.

ChEBI provides an ontology which encodes the relationships between its compounds. We use this ontology to define precisely the notion of parents and children, which we use to hierarchically retrieve and display modifications. We use two kinds of relationships for this purpose, each of which can also be represented by their associated symbols, defined by ChEBI[22]: $\mathcal{F}$ *has functional parent* and $\triangle$ *is a*. We use these relationships to find candidate DNA modifications, by identifying entities related to the core nucleobases, which we represent by their symbols: $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{U}\}$. We include uracil, since many of its descendents in the ontology are modifications of thymine (`CHEBI:17821`, which is equivalent to 5-methyluracil), and are not annotated as descendents of thymine itself. For each of these bases, we import 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 import entities with the highest rating—three stars—indicating manual curation by ChEBI. Whenever possible, we only include entities as nitrogenous bases (nucleobases). If not available, we then select their nucleoside form and finally, if necessary, the nucleotide. These imported bases form the candidate set of modifications (the "unverified" set), from which we create 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 descendent 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,

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

since it is conceived of as possessing adenine as a characteristic group and as being derived via functional modification[22]. We therefore need to make use of both of these two relationships, within the ChEBI ontology, to accurately capture the desired nucleobase hierarchy.

ChEBI also provides selected citations, associated with some of its entities. We query ChEBI for its citations, via their PubMed IDs[26]. We use the Biopython[24] package `Bio.Entrez` to query the PubMed citation database, using NCBI's Entrez Programming Utilities[26]. We retrieve the details of each citation, and use them to construct a formatted citation. At this time, we only support publications that are indexed in PubMed.

Manual curation and annotation

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

We proceed to formalize the above description, of bases imported based upon the ChEBI ontology[22] and their filtering, as follows. Let $a \boxed{\mathcal{F}} b$ specify that a has the $\boxed{\mathcal{F}}$ *has functional parent* relationship with b. Similarly, let $a \triangle b$ specify that a has the $\triangle$ *is a* 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 definition holds 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 *descendent* 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 $\mathcal{C}$. Finally, we manually curate 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 import the set of verified DNA modifications, $\mathcal{M}$, defined in set-builder notation with predicates, as:

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

We additionally provide 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[27–30], and annotated the available methods for each base. These annotations include the method's name, our categorizations of the basis for the method (such as chemical conversion), its resolution, limited genome-wide applicability or use of an enrichment method, and the citation for the method (Table 1A). We consider any method which involves affinity-based recognition of targets to be of "low" resolution[31]. These methods can also suffer from low specificity or cross-reactivity of the antibody[27]. Conversely, we annotate 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*, either as an endogenously-generated modification or those that have been observed to occur as a result of exogenous stimuli, such as exposure to an environmental toxin. We annotate any modification observed *in vivo* merely as "natural". We additionally provide non-exhaustive *examples* of some organisms in which the modifications have been reported. We annotate any modification not observed *in vivo* as "synthetic", and list a reference in which it was synthesized or in which the synthetic base was used. For each of these annotations, we also briefly annotate a primary biological function, if known (Table 1B).

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Recommended notation

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

Mapping techniques

Method	Method detail	Resolution	Enrichment or limitation	References
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).
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).
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). Refer to Table 1 for an explanation of the mapping and natural occurrence table headers.

We enter these annotations in two annotation source files (Table 1), which we later import into our database. This decouples them from the rest of our pipeline and allows experts to submit additions from their domain of expertise, 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[2,32,33]. We additionally provide recommendations for one-letter symbols of selected modified bases, and in some instances for their base-pairing complements. We have previously described these, as part of our expanded epigenetic alphabet, which we currently use to model modification-sensitive transcription factor binding sites[34]. We provide an example of these tables for 5-formylcytosine in Figure 1.

We store all data, either imported from ChEBI or from our manual annotations, within a SQLite[35] database, used via the Python `sqlite3` package[36].

Website generation

We use a static website to display and provide navigation for the information contained within the database. We generate it by formatting the content of the database using Jinja2[37], a static Python templating engine. Two templates are sufficient to generate all HTML files. We use a single template for all modification pages and another for the homepage. We also record the date of the most recent update to the database. All webpages use the Bootstrap[38] framework, which provides a standardized, portable, and mobile-compatible viewing format. An image of the chemical structure of each compound is created by converting Simplified Molecular-Input Line-Entry System (SMILES) data, if available from ChEBI, into a vector graphic, using the cheminformatics toolkit Open Babel[39], via its Python wrapper Pybel[40].

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Possible annotations within DNAmod's curated (A) sequencing method data and (B) natural occurrence information. We list all terms currently used to annotate their respective fields, contained within annotation source files. Each row contains all possible instantiations of the field on the left, except that terms within the "Function" field in (B) are often combined, as conjunctions. Terms within square brackets indicate optional prefixes, that are occasionally used. Terms whose instantiation is within angle brackets denote a description of the term, as opposed to the complete enumeration provided for other terms.

(A). Sequencing method annotations

Field	Term instantiations
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
Enrichment or limitation	5hmU:G mismatch only, CpG contexts only, [methylation-insensitive] restriction digestion, microarray probes, specific fragments, target sequences, gradient stratification

(B). Natural occurrence annotations

Field	Term instantiations
Function	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 ⟩

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Searching and navigation

The modifications contained within DNAmod are accessible via either a search input field or by selecting them from a visual representation of curated modified DNA bases or a separate list of candidate entities. Three tabs on the DNAmod homepage provide these navigation options. The first tab provides the ability to search for a DNA modification, the second tab contains the curated DNA modifications displayed as a pie menu, and the third tab lists all other 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 webpage. We use the elasticlunr.js JavaScript module[41] to implement this. Searching allows matching to multiple fields: the common or International Union of Pure and Applied Chemistry (IUPAC) names, all synonyms, any assigned abbreviation, and a symbol, if available. DNAmod returns 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, the modifications in DNAmod can be browsed through a pie menu[42] interface (Figure 2B), which hierarchically arranges the bases according to their structure within the ChEBI ontology. The innermost ring consists of the four unmodified DNA bases and consecutive outer rings represent children of the previous base. We demarcate 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 is for 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.

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 available names available from ChEBI, including: their IUPAC name, SMILES string, 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[34] (Figure 3).

We provide literature annotations for many modifications, including all DNA modifications 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).

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(A)

(B)

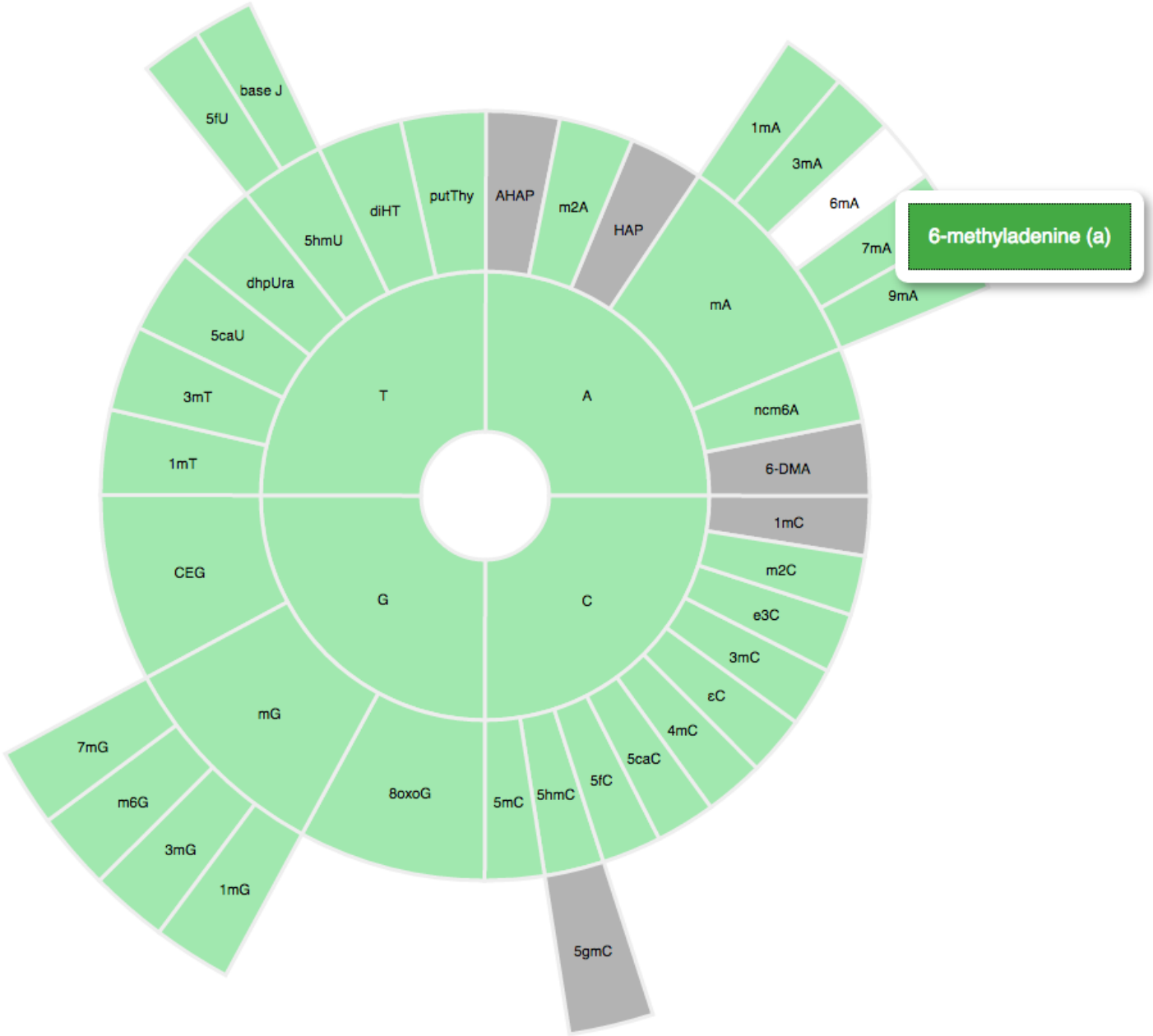


Figure 2. Finding 6-methyladenine by (A) searching for its abbreviation "6mA" or (B) via the pie menu.

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

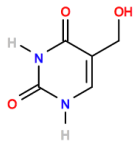
5-hydroxymethyluracil: DNAmod — A modification of thymine

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	Synonyms
5-(hydroxymethyl)pyrimidine-2,4(1H,3H)-dione	OCc1c[nH]c(=O)[nH]c1=O	5-(hydroxymethyl)uracil 5-hydroxymethyl uracil 5-(hydroxymethyl)-2,4(1h,3h)-pyrimidinedione

Mapping techniques

Method	Method detail	Resolution	Enrichment or limitation	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

Ravanat, JL, et al. (1999) Simultaneous determination of five oxidative DNA lesions in human urine. *Chemical research in toxicology*, 12(9).
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).
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).
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).
Clark, TA, et al. (2011) Direct detection and sequencing of damaged DNA bases. *Genome integrity*, 2.
Pfaffeneder, T, et al. (2014) Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nature chemical biology*, 10(7).
Yu, M, et al. (2014) Detection of mismatched 5-hydroxymethyluracil in DNA by selective chemical labeling. *Methods (San Diego, Calif.)*, 72.
Hardisty, RE, et al. (2015) Selective Chemical Labeling of Natural T Modifications in DNA. *Journal of the American Chemical Society*, 137(29).
Frenkel, K, et al. (1985) Quantitative determination of the 5-(hydroxymethyl)uracil moiety in the DNA of gamma-irradiated cells. *Biochemistry*, 24(17).
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).
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).
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).

Home About Contact

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

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 expect DNAmod to continue to grow, particularly as new discoveries about DNA modifications are made. We also hope that DNAmod will serve to highlight modifications that have received inadequate attention, but may be of substantial biological importance.

The nomenclature used to describe a particular DNA modification is often inconsistent, with some early efforts toward standardization of particular classes[32,33]. 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.

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 time, as resolution improves, and especially to highlight recent developments, like direct-detection of various modifications via nanopore sequencing[43]. 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. 5mC and 5hmC, for example, 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, oxidative bisulfite sequencing (oxBS-seq), often used in combination with conventional bisulfite sequencing to elucidate 5hmC via subtraction, is only annotated as a sequencing method for 5mC, which it directly elucidates. Conversely, TET-assisted bisulfite sequencing (TAB-seq), also used for 5hmC detection, is only annotated under 5hmC, which it directly elucidates[27].

We demarcate bases that have been found to occur *in vivo*, providing examples of organisms in which a modification has been found, along with associated citations. This is merely to substantiate its *in vivo* presence, however, and does not comprehensively list organisms which contain that particular modification. Finally, our brief annotations of the biological roles of various DNA modifications are expected to change as further research is conducted.

Future work

We plan to keep DNAmod updated continuously, manually reviewing newly added ChEBI compounds, continuing to request that missing DNA modifications be added to ChEBI (which we then automatically import), and curating any additions. We also add new sequencing annotations, as we come across them, and plan to continue to do so.

Integrating additional external databases will further increase DNAmod's utility. In particular, we envision potential integration with domain-specific DNA modification databases. For instance, modifications involved in DNA damage and repair could be linked to REPAIRtoire[21] data.

We used ChEBI Web Services[23] to obtain information from their database. ChEBI has, however, recently released a Python application programming interface (API), permitting us to directly access their data[44]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Availability

The DNAmod website and its backing SQLite database are freely available at: http://dnamod.hoffmanlab.org. Python source code and web assets for this project and an issue tracker are available at: http://bitbucket.org/hoffmanlab/dnamod. To ensure persistent availability, we have deposited in Zenodo the current version of our code (doi:10.5281/zenodo.60827) and SQLite database (doi:10.5281/zenodo.60824). 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).

Acknowledgements

We thank Daniel D. De Carvalho and Christopher E. Mason for helpful feedback on early versions of DNAmod. We thank the creators of ChEBI[22], and all those who have worked to improve it[23,44,45]. In particular, we thank Gareth Owen, Steve Turner, and Marcus Ennis for actively responding to curation requests and Venkatesh Muthukrishnan for managing ChEBI issues. The authors thank Carl Virtanen, Qun Jin, and Zhibin Lu for technical assistance. Thanks to Gabriel Moreno-Hagelsieb for revising the *NAR* LaTeX template for use with an external BibTeX file. Thanks to Thomas D. Schneider for providing a *NAR* BibTeX template. We thank Casey M. Bergman for collating and distributing these TeX files.

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 an Alexander Graham Bell Canada Graduate Scholarship to C.V.), the Canadian Cancer Society (703827 to M.M.H.), the Ontario Ministry of Training, Colleges and Universities (Ontario Graduate Scholarship 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 University of Toronto McLaughlin Centre (MC-2015-16 to M.M.H.), and the Princess Margaret Cancer Foundation.

Conflict of interest statement. None declared.

References

[1] Dantas Machado,A.C., Zhou,T., Rao,S., Goel,P., Rastogi,C., Lazarovici,A., Bussemaker,H.J., and Rohs,R. (2014) Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief. Funct. Genomics*, **14**, 61–73.

[2] Chen,K., Zhao,B.S., and He,C. (2016) Nucleic acid modifications in regulation of gene expression. *Cell Chem. Biol.*, **23**, 74–85.

[3] Bachman,M., Uribe-Lewis,S., Yang,X., Burgess,H.E., Iurlaro,M., Reik,W., Murrell,A., and Balasubramanian,S. (2015) 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.*, **11**, 555–557.

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[4] Iurlaro,M., McInroy,G.R., Burgess,H.E., Dean,W., Raiber,E.A., Bachman,M., Beraldi,D., Balasubramanian,S., and Reik,W. (2016) In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biol.,* **17**, 141.

[5] Song,C.X., Szulwach,K.E., Dai,Q., Fu,Y., Mao,S.Q., Lin,L., Street,C., Li,Y., Poidevin,M., Wu,H., Gao,J., Liu,P., Li,L., Xu,G.L., Jin,P., and He,C. (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell,* **153**, 678–691.

[6] Rothbart,S.B. and Strahl,B.D. (2014) Interpreting the language of histone and DNA modifications. *Biochim. Biophys. Acta, Gene Regul. Mech.,* **1839**, 627–643.

[7] Pfaffeneder,T., Spada,F., Wagner,M., Brandmayr,C., Laube,S.K., 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,C.G., Vermeulen,M., Leonhardt,H., Schär,P., Müller,M., and Carell,T. (2014) Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat. Chem. Biol.,* **10**, 574–581.

[8] Wu,H. and Zhang,Y. (2014) Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell,* **156**, 45–68.

[9] Hardisty,R.E., Kawasaki,F., Sahakyan,A.B., and Balasubramanian,S. (2015) Selective chemical labeling of natural T modifications in DNA. *J. Am. Chem. Soc.,* **137**, 9270–9272.

[10] Heyn,H. and Esteller,M. (2015) An adenine code for DNA: a second life for N6-methyladenine. *Cell,* **161**, 710–713.

[11] 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.

[12] Gommers-Ampt,J.H. and Borst,P. (1995) Hypermodified bases in DNA. *FASEB J.,* **9**, 1034–1042.

[13] Cantara,W.A., Crain,P.F., Rozenski,J., McCloskey,J.A., Harris,K.A., Zhang,X., Vendeix,F.A.P., Fabris,D., and Agris,P.F. (2011) The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res.,* **39**, D195–D201.

[14] Machnicka,M.A., Milanowska,K., Osman Oglou,O., Purta,E., Kurkowska,M., Olchowik,A., Januszewski,W., Kalinowski,S., Dunin-Horkawicz,S., Rother,K.M., Helm,M., Bujnicki,J.M., and Grosjean,H. (2013) MODOMICS: a database of RNA modification pathways—2013 update. *Nucleic Acids Res.,* **41**, D262–D267.

[15] Sun,W.J., Li,J.H., Liu,S., Wu,J., Zhou,H., Qu,L.H., and Yang,J.H. (2016) RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res.,* **44**, D259–D265.

[16] Zhang,Y., Lv,J., Liu,H., Zhu,J., Su,J., Wu,Q., Qi,Y., Wang,F., and Li,X. (2010) HHMD: the human histone modification database. *Nucleic Acids Res.,* **38**, D149–D154.

[17] Rother,K., Papaj,G., and Bujnicki,J.M. (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.

[18] Roberts,R.J., Vincze,T., Posfai,J., and Macelis,D. (2015) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.,* **43**, D298–D299.

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[19] Amoreira,C., Hindermann,W., and Grunau,C. (2003) An improved version of the DNA methylation database (MethDB). *Nucleic Acids Res.,* **31**, 75–77.

[20] Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M., and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.,* **44**, D457–D462.

[21] Milanowska,K., Krwawicz,J., Papaj,G., Kosiński,J., Poleszak,K., Lesiak,J., Osińska,E., Rother,K., and Bujnicki,J.M. (2011) REPAIRtoire—a database of DNA repair pathways. *Nucleic Acids Res.,* **39**, D788–D792.

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

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

[24] Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B., and de Hoon,M.J.L. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* **25**, 1422–1423.

[25] Ortel,J., Noehr,J., and van Gheem,N. (2011) suds. https://fedorahosted.org/suds/

[26] NCBI Resource Coordinators. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.,* **44**, D7–19.

[27] Booth,M.J., Raiber,E.A., and Balasubramanian,S. (2015) Chemical methods for decoding cytosine modifications in DNA. *Chem. Rev.,* **115**, 2240–2254.

[28] Plongthongkum,N., Diep,D.H., and Zhang,K. (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat. Rev. Genet.,* **15**, 647–661.

[29] Song,C.X., Yi,C., and He,C. (2012) Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat. Biotechnol.,* **30**, 1107–1116.

[30] Korlach,J. and Turner,S.W. (2012) Going beyond five bases in DNA sequencing. *Curr. Opin. Struct. Biol.,* **22**, 251–261.

[31] Booth,M.J., Marsico,G., Bachman,M., Beraldi,D., and Balasubramanian,S. (2014) Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat. Chem.,* **6**, 435–440.

[32] Khromov-Borisov,N.N. (1997) Naming the mutagenic nucleic acid base analogs: the Galatea syndrome. *Mutat. Res.,* **379**, 95–103.

[33] Cooke,M.S., Loft,S., Olinski,R., Evans,M.D., Bialkowski,K., Wagner,J.R., Dedon,P.C., Møller,P., Greenberg,M.M., and Cadet,J. (2010) Recommendations for standardized description of and nomenclature concerning oxidatively damaged nucleobases in DNA. *Chem. Res. Toxicol.,* **23**, 705–707.

[34] Viner,C., Johnson,J., Walker,N., Shi,H., Sjöberg,M., Adams,D.J., Ferguson-Smith,A.C., Bailey,T.L., and Hoffman,M.M. (2016) Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. *bioRxiv,* **043794**.

bioRxiv preprint doi: https://doi.org/10.1101/071712; this version posted August 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[35] Hipp,D.R., Kennedy,D., and Mistachkin,J. (2000–2016) SQLite. https://www.sqlite.org

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

[37] Ronacher,A. (2008) Jinja2 (The Python Template Engine). http://jinja.pocoo.org/

[38] Otto,M., Thornton,J., Rebert,C., Thilo,J., XhmikosR, Fenkart,H., Lauke,P.H., et al. (2011–2016) Bootstrap. http://getbootstrap.com/

[39] O'Boyle,N.M., Banck,M., James,C.A., Morley,C., Vandermeersch,T., and Hutchison,G.R. (2011) Open Babel: an open chemical toolbox. *J. Cheminf.*, **3**, 33.

[40] O'Boyle,N.M., Morley,C., and Hutchison,G.R. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 5.

[41] Song,W. (2012–2016) Elasticlunr.js. http://elasticlunr.com/

[42] Callahan,J., Hopkins,D., Weiser,M., and Shneiderman,B. (1988) In O'Hare,J.J., (ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* An empirical comparison of pie vs. linear menus. pp. 95–100.

[43] Wallace,E.V.B., Stoddart,D., Heron,A.J., Mikhailova,E., Maglia,G., Donohoe,T.J., and Bayley,H. (2010) Identification of epigenetic DNA modifications with a protein nanopore. *Chem. Commun.*, **46**, 8195–8197.

[44] Swainston,N., Hastings,J., Dekker,A., Muthukrishnan,V., May,J., Steinbeck,C., and Mendes,P. (2016) libChEBI: an API for accessing the ChEBI database. *J. Cheminf.*, **8**, 11.

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