A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes

Nucleic Acids Research, 2003, Vol. 31, No. 7 ***1805–1812***
DOI: 10.1093/nar/gkg274

SURVEY AND SUMMARY

A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes

Richard J. Roberts*, Marlene Belfort[1], Timothy Bestor[2], Ashok S. Bhagwat[3], Thomas A. Bickle[4], Jurate Bitinaite, Robert M. Blumenthal[5], Sergey Kh. Degtyarev[6], David T. F. Dryden[7], Kevin Dybvig[8], Keith Firman[9], Elizaveta S. Gromova[10], Richard I. Gumport[11], Stephen E. Halford[12], Stanley Hattman[13], Joseph Heitman[14], David P. Hornby[15], Arvydas Janulaitis[16], Albert Jeltsch[17], Jytte Josephsen[18], Antal Kiss[19], Todd R. Klaenhammer[20], Ichizo Kobayashi[21], Huimin Kong, Detlev H. Krüger[22], Sanford Lacks[23], Martin G. Marinus[24], Michiko Miyahara[25], Richard D. Morgan, Noreen E. Murray[26], Valakunja Nagaraja[27], Andrzej Piekarowicz[28], Alfred Pingoud[17], Elisabeth Raleigh, Desirazu N. Rao[27], Norbert Reich[29], Vladimir E. Repin[30], Eric U. Selker[31], Pang-Chui Shaw[32], Daniel C. Stein[33], Barry L. Stoddard[34], Waclaw Szybalski[35], Thomas A. Trautner[36], James L. Van Etten[37], Jorge M. B. Vitor[38], Geoffrey G. Wilson and Shuang-yong Xu

New England Biolabs, Beverly, MA 01915, USA, [1]Molecular Genetics Program, New York State Department of Health, Albany, NY 12201-2002, USA, [2]Genetics and Development, Columbia University, New York, NY 10032, USA, [3]Department of Chemistry, Wayne State University, Detroit, MI 48202, USA, [4]Department of Microbiology, Biozentrum, Universitat Basel, CH-4056 Basel, Switzerland, [5]Program in Bioinformatics and Proteomics/Genomics, Medical College of Ohio, Toledo, OH 43699-0008, USA, [6]SibEnzyme, 630090 Novosibirsk, Russia, [7]School of Chemistry, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JJ, UK, [8]Department of Genetics, University of Alabama at Birmingham, Birmingham, AL 35294, USA, [9]Biophysics Laboratories, School of Biological Sciences, University of Portsmouth, Portsmouth PO1 2DT, UK, [10]A.N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, 119992 Moscow, Russia, [11]The University of Illinois College of Medicine, Urbana, IL 61801-3602, USA, [12]Department of Biochemistry, University of Bristol Medical School, Bristol BS8 1TD, UK, [13]Department of Biology, University of Rochester, Rochester, NY 14627-0211, USA, [14]Howard Hughes Medical Institute, Duke University Medical Center, Durham, NC 27710, USA, [15]Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, UK, [16]Institute of Biotechnology, LT-2028 Vilnius, Lithuania, [17]Institut für Biochemie, Justus-Liebig-Universität, D-35392 Giessen, Germany, [18]Department of Dairy and Food Science, Royal Veterinary and Agricultural University, DK-1958 Frederiksberg C, Denmark, [19]Institute of Biochemistry, BRC, H-6701 Szeged, Hungary, [20]Departments of Food Science and Microbiology, North Carolina State University, Raleigh, NC 27695-7624, USA, [21]Department of Molecular Biology, Institute of Medical Science, University of Tokyo, 4-6-1, Shirokanedai, Minato-ku, Tokyo 108-8639, Japan, [22]Institut für Virologie-Charite, Humboldt Universitat, D-10098 Berlin, Germany, [23]Brookhaven National Laboratory, Upton, NY 11973-5000, USA, [24]Department of Pharmacology, University of Massachusetts Medical School, Worcester, MA 01655, USA, [25]National Institute of Health Sciences, 1-18-1, Kamiyoga, Setagaya-ku, Tokyo 158-8501, Japan, [26]Institute of Cell and Molecular Biology, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JR, UK, [27]Department of Microbiology and Cell Biology, Indian Institute of Science, IN-560012 Bangalore, India, [28]Institute of Microbiology, Warsaw University, Miecznikowa 1, 02-096 Warsaw, Poland, [29]University of California, Santa Barbara, Santa Barbara, CA 93106-0001, USA, [30]State Research Center of Virology and Biotechnology 'Vektor', Koltsovo, Novosibirsk region 633059, Russia, [31]Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA, [32]Department of Biochemistry, The Chinese University of Hong Kong, Hong Kong, [33]Department of Microbiology, University of Maryland, College Park, MD 20742, USA, [34]Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, [35]McArdle Laboratory,

*To whom correspondence should be addressed. Tel: +1 978 927 3382; Fax: +1 978 921 1527; Email: roberts@neb.com

University of Wisconsin, Madison, WI 53706, USA, [36]MPI für Molekulare Genetik, Ihnestrasse 73, D-14195 Berlin, Germany, [37]Department of Plant Pathology, University of Nebraska-Lincoln, Lincoln, NE 68583, USA and [38]Faculdade de Farmacia de Lisboa, 1600-083 Lisbon, Portugal

Received December 4, 2002; Revised and Accepted February 3, 2003

ABSTRACT

A nomenclature is described for restriction endonucleases, DNA methyltransferases, homing endonucleases and related genes and gene products. It provides explicit categories for the many different Type II enzymes now identified and provides a system for naming the putative genes found by sequence analysis of microbial genomes.

INTRODUCTION

There are three main groups of restriction endonucleases (REases) called Types I, II and III (1,2). Since 1973, REases and DNA methyltransferases (MTases) have been named based on an original suggestion by Smith and Nathans (3). They proposed that the enzyme names should begin with a three-letter acronym in which the first letter was the first letter of the genus from which the enzyme was isolated and the next two letters were the first two letters of the species name. Extra letters or numbers could be added to indicate individual strains or serotypes. Thus, the enzyme *Hin*dII was one of four enzymes isolated from *Haemophilus influenzae* serotype d. The first three letters of the name were italicized. Later, a formal proposition for naming the genes encoding REases and MTases was adopted (4). When there were only a handful of enzymes known, these schemes were very useful, but as more enzymes have been found, often from different genera and species with names whose three-letter acronyms would be identical, considerable laxity in naming conventions has appeared. In addition, we now know that each major type of enzyme can contain sub-types. This especially applies to the Type II enzymes, of which more than 3500 have been characterized (5). In this paper we revisit the naming conventions and outline an updated scheme that incorporates current knowledge about the complexities of these enzymes. We describe a set of naming conventions for REases and their associated MTases. Since the homing endonucleases (6) have been named in an analogous fashion, we propose that similar guidelines be applied to that group of enzymes. Finally, it is important to realize that the aim of this document is to provide a nomenclature for these enzymes, not to provide a rigorous classification.

GENERAL RULES

First, we introduce a number of general changes, standard abbreviations and definitions that are recommended for use.

1. 'Restriction enzyme' and 'restriction endonuclease' should be regarded as synonymous and the abbreviation REase (or in some cases, R) is preferred. However, the abbreviation ENase, which has been used extensively, may also be used. Alternative names such as restrictases should be avoided. The abbreviation R-M should be used for restriction-modification. Homing endonuclease should be abbreviated HEase.

2. Methyltransferase is the preferred name, since it correctly describes the activity. Methylase, while in common use, is not strictly accurate and should be avoided in print. The abbreviation MTase (or in some cases, M) should be the standard.

3. Italics will no longer be used for the first three-letter acronym of the REase or MTase name. Many journals already avoid italics and retaining the italic convention is not easily translated to computers and serves no essential purpose. The convention of naming different enzymes from the same isolate of the same organism with increasing Roman numerals will continue.

4. Restriction enzyme names should not include a space between the main acronym and the Roman numeral. This practice, which has been employed to avoid the inelegant look caused when characters in italic fonts are juxtaposed next to characters in a regular font, is incorrect. Now that italics will no longer be used in names there is no reason to continue this practice. The previous scheme of using a raised dot after the prefix will be abandoned and a normal dot (period) should be used. Furthermore, except for the single period or hyphen (in homing endonucleases) that is used to separate the prefix from the main part of the name, no punctuation marks, such as parentheses, periods, commas or slashes, should be used in REase or MTase names. Only alphanumeric characters should appear. Already the enzymes from *Nostoc* species C have been changed from their original Nsp(7524)I to NspI and many others have also changed. The most recent is Bst4.4I, which has changed to Bst44I.

5. The designation of the three main types of REases as Type I, Type II and Type III will continue, with the capital 'T' preferred. However, they will be divided into subtypes as indicated below. One new type of REase will be added. This is Type IV, which will include those systems that cleave only methylated DNA as their substrate and show only weak specificity, such as the McrA, McrBC and Mrr systems of *Escherichia coli.*

6. The sequence databases contain many genes that are excellent candidates to encode DNA MTases and REases, based on sequence similarity. These will be named according to the same guidelines as are used for biochemically characterized enzymes, but will carry the suffix P to indicate their putative nature. Once they have been characterized biochemically and shown to be active, the P will be dropped and their names will be changed to a regular name with the next Roman numeral that is appropriate.

7. The current convention of naming R-M enzymes with a prefix M, R, etc. will be expanded to include the protein products of related genes such as the controlling proteins (e.g. C.BamHI) and the nicking enzymes that cleave G/T mismatches (e.g. V.HpaII for the vsr-like enzyme associated with the HpaII system) and N.BstNBI for the regular nicking

enzymes. In addition, up to two characters will be allowed in the prefix. This will enable enzymes, such as Eco57I, with both REase activity and MTase activity fused in a single protein to be designated RM.Eco57I. Its accompanying MTase would remain as M.Eco57I. Note that the current convention of permitting the REase to be named either with or without the 'R' prefix will be continued. Thus, R.EcoRI and EcoRI will be considered synonymous as will RM.Eco57I and Eco57I. For certain nicking enzymes that have been obtained from the Type IIT enzymes, where one of the two heterodimeric subunits has been inactivated the resultant mutant nicking enzymes should be called Nt.BbvCI or Nb.BbvCI, where the 't' and the 'b' indicate cleavage of the top or bottom strand of the normal recognition sequence.

8. When two REase or MTase genes are present and associated with a single R-M system they should be referred to with the second character of the prefix being an Arabic 1 or 2. Thus, the two M gene products of the HphI R-M system would be M1.HphI and M2.HphI.

9. The standard abbreviations for methylated bases should be 5-methylcytosine (m5C), N4-methylcytosine (m4C) and N6-methyladenine (m6A). It is not necessary to use a superscript for the number.

10. Isoschizomers are REases that recognize the same sequence. The first example discovered is called a prototype and all subsequent enzymes that recognize the same sequence are isoschizomers of the prototype. Neoschizomers are that subset of isoschizomers that recognize the same sequence, but cleave at different positions from the prototype. Thus, AatII (recognition sequence: GACGT↓C) and ZraI (recognition sequence: GAC↓GTC) are neoschizomers of one another, while HpaII (recognition sequence: C↓CGG) and MspI (recognition sequence: C↓CGG) are isoschizomers, but not neoschizomers. Analogous designations are not appropriate for MTases, where the differences between enzymes are not so easily defined and usually have not been well characterized.

11. The solitary MTases (i.e. not associated with an REase) such as the Dam and Dcm MTases of *E.coli* and the eukaryotic MTases such as Dnmt1 and Dnmt3a will be named systematically in accordance with the general rules established for the prokaryotic enzymes. Thus, the systematic name for the Dam MTase of *E.coli* K12 will be M.EcoKDam and the murine maintenance MTase will be M.MmuDnmt1. However, it will be acceptable to refer to them by their more commonly used trivial names, Dam, Dcm, Dnmt1, etc., but it will simplify automated searching and cross-referencing of the literature if the systematic name, including the M prefix, also appears at least initially in a publication. Solitary MTases that are phage or virus borne are also named with the prefix M and the name of the phage or virus that carries them. Optionally, the host name may be included. Thus, the MTase encoded by phage SPR of *Bacillus subtilis* is named M.SPRI (7) and the MTase encoded by the archaeal virus ϕCh1 of *Natrialba magadii* is named M.NmaPhiCh1I (8).

12. The rules for naming genes of Type II R-M systems should adhere to the proposals of Szybalski *et al.* (4) with the obvious extensions to accommodate C, V and N genes. Thus, the entire name should be italicized, the first letter will be lower case and the capital letter(s) used as a prefix for the protein will become the suffix for the gene. The gene for EcoRI thus becomes *ecoRIR* and that for its MTase is *ecoRIM*. In the case of genes with two prefixes in the protein name the gene name would incorporate both letters of the prefix. Thus, the gene for RM.Eco57I would become *eco57IRM*. In the case of Type I enzymes, an acronym for the source organism should be followed by the traditional gene designations, hsdS, hsdR and hsdM. Thus, the three genes of the EcoKI restriction system would be *ecoKIhsdM*, *ecoKIhsdR* and *ecoKIhsdS*. However, it will be acceptable to omit the *ecoKI* where appropriate.

13. It sometimes happens that two genes are required for a single enzyme activity, effectively encoding two subunits. In these situations the two genes and their products should carry a suffix A and B. For example, BbvCI is a heterodimeric REase. The two gene products would be called R.BbvCIA (or just BbvCIA) and R.BbvCIB (or just BbvCIB), and the active holoenzyme would be BbvCI. Note that the two separate MTases of this system would be M1.BbvCI and M2.BbvCI. For enzymes like Eco57I, which have both endonuclease and MTase activity in the same polypeptide chain, the endonuclease would be referred to as RM.Eco57I, but the second MTase activity associated with this system would be called M.Eco57I. For MTases, an example is M.AquI, which has one gene encoding the N-terminal region up to the middle of the variable region of this m5C MTase and a second gene encoding the remaining C-terminal region (9). In this case, the two parts of this protein should be referred to as M.AquIA and M.AquIB and the genes as *aquIAM* and *aquIBM*.

DETAILS OF TYPES AND SUBTYPES

Types I, II, III and IV

The original subdivision of Types I, II and III will be maintained and a new Type IV added to accommodate a class of methyl-dependent restriction enzymes. The previously proposed candidates for new types, such as Eco57I and GsuI, will be incorporated as subtypes of existing Type II enzymes.

Type I

The key characteristics of the Type I R-M systems are that these enzymes are multisubunit proteins that function as a single protein complex and usually contain two R subunits, two M subunits and one S subunit (10). The symbol for Type I systems is *hsd*, thus the genes are *hsdR*, *hsdM* and *hsdS*, and their protein products are HsdR, HsdM and HsdS, respectively. The protein products can be abbreviated by omitting Hsd. The S subunit is the specificity subunit that determines which DNA sequence is recognized. The R subunit is essential for cleavage (restriction) and the M subunit catalyzes the methylation reaction: in all known cases the methylated base formed is m6A. When Type I enzymes act on unmethylated substrates, they function mainly as REases (they may also methylate unmodified sites with a low probability) and have an absolute requirement for ATP during cleavage. They cleave the DNA at variable positions away from their recognition sequence. The location of the cleavage sites is determined by either the collision and stalling of two such complexes during translocation along a DNA chain, or the stalling of a single enzyme on a single-site circular substrate following DNA translocation. The biochemical nature of the termini produced upon cleavage is unknown and the enzymes do not turn over in

the cleavage reaction. In contrast, when these complexes encounter a hemimethylated substrate, in which one strand of the recognition sequence is methylated, as would occur immediately after DNA replication of a fully methylated substrate, then the complex functions as a DNA MTase, using S-adenosylmethionine (AdoMet) as the donor of the methyl group. A complex of two M subunits and one S subunit is fully functional as an MTase. Probably the best known Type I enzyme is EcoKI (11). The REase is referred to as R.EcoKI or EcoKI, but it is important to remember that it is also an MTase. The MTase complex of two HsdM and one HsdS is referred to as M.EcoKI. When referring to phenotypes the preferred convention is r_{KI}^+ m_{KI}^+ etc.

Four sub-categories of Type I enzymes (A, B, C and D) are in common use (12). These are based on genetic complementation and their use will be continued. If experimental evidence defines new subtypes, then additional letters may be used as suffixes to describe them. A number of artificially created hybrid enzymes have been described (13), which often include those with new specificities. These should be named as deemed appropriate, but without a Roman numeral at the end.

Type II

The Type II REases recognize specific DNA sequences and cleave at constant positions at or close to that sequence to produce 5′-phosphates and 3′-hydroxyls. Usually they require Mg^{2+} ions as a cofactor, although some have more exotic requirements (see below). They may act as monomers, dimers or even tetramers and usually act independently of their companion MTase. The MTases usually act as monomers and transfer a methyl group from the donor S-adenosyl-L-methionine directly to double-stranded DNA and form m4C, m5C or m6A. Because of the interest in these Type II REases for recombinant DNA technology, more than 3500 have been characterized (5). Given the assay that is used to find them, which detects any activity yielding a consistent DNA fragmentation pattern, it is no surprise that they come in a large variety of 'flavors'. Early on it was recognized that while then-normal Type II enzymes recognized palindromic sequences and cleaved symmetrically within them, the Type IIS enzymes cut outside their normally asymmetric sequences and differed in other interesting ways (14). We now know of additional enzymes that cleave on both sides of their recognition sequence (e.g. BcgI), are activated by AdoMet (e.g. Eco57I), interact with two copies of their recognition sequence (e.g. EcoRII) or have unusual subunit structures (e.g. BbvCI).

These additional kinds of enzymes will be considered subdivisions of Type II. It should be recognized that for the purposes of nomenclature some enzymes would fall into more than one subdivision. Specifically, some of the criteria are based on the sequence cleaved and others on the structure of the enzymes themselves, so not all subdivisions are mutually exclusive, e.g. BcgI is both Type IIB and IIH. Type IIS enzymes, originally designated as enzymes with cleavage sites shifted away from their recognition sequence (4), will be retained, but a new Type IIA will be defined that includes all Type II REases that recognize asymmetric sequences. A new Type IIP will be used to designate the enzymes that recognize symmetric sequences (palindromes).

The overriding criterion for inclusion as a Type II enzyme would be that it yields a defined fragmentation pattern and cleaves either within or close to its recognition sequence at a fixed site or with known and limited variability. In general, the Type II REases and their associated MTases are separate, independent enzymes, but in several classes (e.g. IIB, IIG and IIH) the R and M genes are fused into a single composite gene. The nomenclature for the subtypes of the Type II enzymes currently known is shown below. It should be noted that these designations are not intended to be exclusive, but rather to permit enzymes with common characteristics to be referred to as a group. Conservation of structural domains with associated enzymatic activities is observed between different classes of Type II enzymes and also between other types of R-M enzymes.

The Type II subdivisions are summarized in Table 1 and described in more detail below.

Type IIP

This would be used as a generic description for all enzymes that recognize symmetric sequences, often termed palindromes, and cleave at fixed symmetrical locations either within the sequence or immediately adjacent to it. The recognition sequences and cleavage sites of these enzymes should be represented as in the following example: EcoRI: G↓AATTC. In full double-stranded form this corresponds to:

```
5′ G↓A A T T C
3′ C T T A A↑G
```

Note that enzymes such as SinI (recognition sequence: GGWCC), BglI (recognition sequence: GCCNNNN↓NGGC) and HindII (recognition sequence: GTYRAC) belong to Type IIP because the recognition mechanism still involves a symmetric homodimer.

Type IIA

This would be used as a generic designation for any Type II enzymes that recognize asymmetric sequences irrespective of whether they cleave away from the sequence or within the sequence. Typically these systems have one REase gene and two MTase genes, one to modify each strand of the asymmetric recognition sequence. However, occasionally two R genes are found as with Bpu10I (15), or both M genes are fused as with M.FokI (16). When more than one R or M gene is present the genes and their protein products should be named with either an Arabic 1 or 2 in the prefix of the name. Thus, the two MTases of the SapI system would be named M1.SapI and M2.SapI if the proteins are being referred to, or *sapIM1* and *sapIM2* for the genes. However, the two subunits of the Bpu10I REase would be designated R.Bpu10IA and R.Bpu10IB and their genes *bpu10IAR* and *bpu10IBR*. The recognition sequences and cleavage sites of the Type IIS REases should be represented as in the following example:

HphI: GGTGA(8/7) where the first numeral in the parentheses indicates the position of cleavage on the strand written and the second numeral indicates the cleavage position on the complementary strand. In full double-stranded form this corresponds to:

```
5′ GGTGANNNNNNNN↓
3′ CCACTNNNNNNN↑
```

Note that when recognition sequences are assigned, the convention is to write the single-stranded sequence such that

Table 1. Subtypes of Type II REases

Subtype[a]	Defining feature	Examples	Recognition sequence
A	Asymmetric recognition sequence	FokI	GGATG (9/13)
		AciI	CCGC (–3/–1)
B	Cleaves both sides of target on both strands	BcgI	(10/12) CGANNNNNNTGC (12/10)
C	Symmetric or asymmetric target. R and M functions in one polypeptide	GsuI	CTGGAG (16/14)
		HaeIV	(7/13) GAYNNNNNRTC (14/9)
		BcgI	(10/12) CGANNNNNNTGC (12/10)
E	Two targets; one cleaved, one an effector	EcoRII	↓CCWGG
		NaeI	GCC↓GGC
F	Two targets, both cleaved coordinately	SfiI	GGCCNNNN↓NGGCC
		SgrAI	CR↓CCGGYG
G	Symmetric or asymmetric target. Affected by AdoMet	BsgI	GTGCAG (16/14)
		Eco57I	CTGAAG (16/14)
H	Symmetric or asymmetric target. Similar to Type I gene structure	BcgI	(10/12) CGANNNNNNTGC (12/10)
		AhdI	GACNNN↓NNGTC
M	Subtype IIP or IIA. Require methylated target	DpnI	Gm6 A↓TC
P	Symmetric target and cleavage sites	EcoRI	G↓AATTC
		PpuMI	RG↓GWCCY
		BslI	CCNNNNNN↓NNGG
S	Asymmetric target and cleavage sites	FokI	GGATG (9/13)
		MmeI	TCCRAC (20/18)
T	Symmetric or asymmetric target. R genes are heterodimers	Bpu10I	CCTNAGC (–5/–2)[b]
		BslI	CCNNNNNN↓NNGG

[a]Note that not all subtypes are mutually exclusive. E.g. BslI is of subtype P and T.
[b]The abbreviation indicates double strand cleavage as shown below:
5′ C C↓T N A G C
3′ G G A N T↑C G

cleavage lies downstream of the sequence. If cleavage takes place within the sequence, then the single-strand designation is always written so that the sequence of the strand is first alphabetically.

Type IIB

This would be used for enzymes that cleave on both sides of the recognition sequence. At present there are many well defined members of this class (AloI, BplI, Bst44I, BaeI, BcgI, BsaXI, Bsp24I, CjeI, CjePI, HaeIV, Hin4I and PpiI). In this case the recognition sequence and cleavage sites should be represented as exemplified for BcgI:

BcgI—recognition sequence:
(10/12)CGANNNNNNTGC(12/10)

Here the (10/12) preceding the recognition sequence indicates that cleavage occurs 10 bases in front of the sequence on the strand written and 12 bases before the sequence on the complementary strand. The (12/10) following the recognition sequence indicates cleavage 12 bases after the recognition sequence on the strand written and 10 bases after the sequence on the complementary strand. In double-stranded form this would be written:

↓NNNNNNNNNNCGANNNNNNTGCNNNNNNNNNNNN↓
↑NNNNNNNNNNNNGCTNNNNNNACGNNNNNNNNNN↑

Type IIC

This would be used as a generic term for all enzymes that have a hybrid structure containing both cleavage and modification domains within a single polypeptide. Examples include all of the Type IIB, IIG and some Type IIH enzymes.

Type IIE

This would be used for enzymes that interact with two copies of the recognition sequence, one being the actual target of cleavage, the other being the allosteric effector. The best studied examples are EcoRII (17) and NaeI (18). FokI, MboII and Sau3AI were found to exhibit similar properties. Other enzymes such as Acc36I, AtuBI, BsgI, BpmI, Cfr9I, Eco57I, HpaII, Ksp632I, NarI, SacII and SauBMKI are likely to be members of this group because they are reported to be stimulated by oligonucleotide duplexes containing the specific recognition site.

Type IIF

This would be used for enzymes that interact with, and cleave coordinately, two copies of their recognition sequence. Examples include BspMI, Cfr10I, NgoMIV, SfiI and SgrAI.

Type IIG

This would be used for enzymes that have both R and M domains fused to form single polypeptides and that may be stimulated or inhibited by AdoMet, but otherwise resemble Type II enzymes. These include Bce83I, BseMII, BseRI, BsgI, BspLU11III, Eco57I, GsuI, MmeI and Tth111II. The recognition sequences may or may not be asymmetric. Thus, both Type IIA and Type IIP enzymes may be of Type IIG.

Type IIH

This would be used for enzymes that contain genetic features resembling Type I enzymes, but biochemically behave as Type II enzymes. At present three examples have been characterized: AhdI and PshAI, both of which comprise a three gene system akin to that of a typical Type I enzyme (G.G.Wilson, unpublished results), and BcgI, which is a two gene system. Several hypothetical systems have gene organizations that resemble that of BcgI.

Type IIM

This would be used for DpnI and similar enzymes that recognize a specific methylated sequence in DNA and cleave at a fixed site. Note that the methyl-dependent enzymes such as McrA, McrBC are not considered members of this subclass, because they do not have well defined recognition sequences and cleavage sites. They are included within the Type IV enzymes.

Type IIS

This would be used for Type IIA enzymes that cleave at least one strand of the DNA duplex outside of the recognition sequence (i.e. cleavage is shifted relative to the recognition sequence). Note that for some enzymes, such as BsmI (recognition sequence: GAATGC), cleavage of the strand written takes place outside of the recognition sequence, whereas cleavage of the complementary strand takes place within the recognition sequence. This is still considered a Type IIS enzyme. However, in most cases both strands are cleaved away from the recognition sequence, which therefore remains intact. These were the earliest sub-classes of the Type II restriction enzymes to be recognized (14).

Type IIT

This would be used for enzymes that are composed of heterodimeric subunits. This subtype includes enzymes like BbvCI, Bpu10I and BslI.

Nicking enzymes

Two types of nicking enzymes are known. One type includes those that behave functionally like REases, but cleave only one strand of the DNA substrate. These enzymes should be named with the prefix N and their recognition sequences should be written such that the strand displayed is the strand nicked. Thus, N.BstSEI has the recognition sequence: GAGTCNNNN↓ which is abbreviated to GAGTC(4). Similarly, the mutants of AlwI and MlyI that have interrupted the dimerization function, and which have become nicking enzymes, are named N.AlwI (19) and N.MlyI (20). For enzymes such as Bpu10I, where the wild-type REase has two subunits, each of which nicks a different strand, the mutant nicking enzymes made by inactivating one or the other subunit should be named Nt.Bpu10I for the enzyme that nicks the top strand of the normal recognition sequence and Nb.Bpu10I for the enzyme that nicks the bottom strand.

In full double-stranded format Nt.Bpu10I would recognize

```
5′ C C↓T N A G C
3′ G G  A N T C G
```

while Nb.Bpu10I would recognize

```
5′ C C T N A  G C
3′ G G A N T↑C G
```

Alternatively this may be written

```
5′ G C↓T N A G G
3′ C G  A N T C C
```

A single-stranded representation of their recognition sites would be Nt.Bpu10I (recognition sequence: CC↓TNAGC) and Nb.Bpu10I (recognition sequence: GC↓TNAGG or CCTNA↑GC). Note that the use of ↑ always denotes cleavage of the lower strand.

A second type of nicking enzyme is found exclusively in association with m5C-MTases, where it serves to nick the G/T mismatches that can result from deamination of m5C within the recognition sequence of the MTase. The best studied of these is the Vsr protein that accompanies the Dcm MTase of *E.coli* K-12, M.EcoKDcm. Vsr recognizes the specific G/T mismatch that occurs if there is deamination of the methylated cytosine residue within the context of the CCWGG recognition sequence of M.EcoKDcm (21). These kinds of mismatch nicking enzymes are named with the prefix V and should be given the acronym of the MTase gene with which they are associated. Thus, Vsr, the product of the V gene that overlaps with the gene for M.EcoKDcm, is systematically named V.EcoKDcm. However, the trivial name Vsr, which was originally designated for this protein, is an acceptable synonym. For other V genes and their products the systematic names are preferred. Thus, V.HpaII is the preferred name for the mismatch nicking endonuclease that accompanies M.HpaII.

Control proteins

Some R-M systems are found to have an additional gene that encodes a protein involved in the control of expression of the R gene. The best studied examples are the PvuII and BamHI systems, where the products of the C genes, C.PvuII (22) and C.BamHI (23), serve as transcriptional activators; this prevents the expression of the R genes following transfer of the systems into naive hosts, until such time as C protein has accumulated and methylation is sufficient to provide protection against what would otherwise be the deleterious action of the REase.

Type III

These systems are composed of two genes (*mod* and *res*) encoding protein subunits that function either in DNA recognition and modification (Mod) or restriction (Res) (10,24,25). Both subunits are required for restriction, which also has an absolute requirement for ATP hydrolysis. For DNA cleavage, the enzyme must interact with two copies of a non-palindromic recognition sequence and the sites must be in an inverse orientation in the substrate DNA molecule. Cleavage is preceded by ATP-dependent DNA translocation as with the Type I REases. The enzymes cleave at a specific distance away from one of the two copies of their recognition sequence. The Mod subunit can function independently of the Res subunit to methylate DNA: in all known cases the methylated base formed is m6A and full modification is actually hemimethylation. This is not deleterious because of the requirement for two unmodified sites in inverse repeat orientation for cleavage. DNA replication puts all of the unmodified sites in the same orientation. The best-known examples of Type III enzymes are EcoP1I and EcoP15I. Putative Type III R-M systems are easily recognized because of their similarity at the sequence level. When naming the genes for these enzymes the *mod* gene of EcoP1I would be systematically named *ecoP1Imod*, but the abbreviation *mod* is acceptable when it does not result in confusion.

Type IV

These systems are composed of one or two genes encoding proteins that cleave only modified DNA, including

methylated, hydroxymethylated and glucosyl-hydroxymethylated bases. Their recognition sequences have usually not been well defined except for EcoKMcrBC, which recognizes two dinucleotides of the general form RmC (a purine followed by a methylated cytosine—either m4C or m5C) and which are separated by anywhere from 40 to 3000 bases. Cleavage takes place ~30 bp away from one of the sites. The best studied example at both the genetic and biochemical level is EcoKMcrBC of *E.coli* (26,27), but on the basis of sequence similarity it is likely that there are many such systems in other bacteria and archaea. As with the genes of the Type I and Type III systems, the abbreviations McrBC for the enzyme and *mcrBC* for the gene are acceptable.

Hypothetical enzymes

Hypothetical REases and DNA MTases can often be found by similarity searching in DNA sequences or their presence may be inferred when specific sequences in plasmid or bacterial DNAs are found to be methylated. It is convenient and useful to be able to refer to such hypothetical enzymes by name. The following convention for naming these enzymes is proposed. They should be named as though they were normal R-M systems, but should carry the suffix 'P' to indicate their putative nature. Once biochemical or unequivocal genetic activity, such as phage restriction, is demonstrated the suffix 'P' and any open reading frame (ORF) designations can be dropped allowing the main element of the name to be retained. Furthermore a Roman numeral should be included to indicate whether it is the first, second, third, etc. enzyme to be found in that organism. Note that the P extension should remain with the gene until such time as a gene product has been demonstrated to be functional.

This 'P' convention is illustrated with genes from *H.influenzae* serotype d. Two Type II REases, HindII and HindIII, and their associated MTases had been characterized biochemically (28–31). One Type I system had been demonstrated genetically (32) and the MTase, presumably associated with this system, had been partially characterized biochemically (30,31). In the genome there are two putative Type I systems, although only one has a complete set of intact genes (33). The intact system therefore carries the designation HindI. In addition to these three systems, there was also known to be a Dam-like MTase, now called M.HindDam. However, also in the genome are putative m5C-MTase and REase genes (genes HI1040 and HI1041) that show high similarity to the known R-M system, HgiDI (34). The MTase encoded by HI1041 leads to a functional protein with specificity identical to that of M.HgiDI (R.D.Morgan, J.Patti and R.J.Roberts, unpublished results). It is therefore named M.HindV. However, the adjacent gene for the putative endonuclease is inactive and so it is named HindVP. One other R-M system can also be seen in the genome, this time encoding a Type III system. Neither the R nor the M gene have yet been demonstrated to be active and so these are named HindORF1056P and M.HindORF1056P. If they are shown to be active they would be renamed HindVI and M.HindVI. The convention here is to name the system after the ORF encoding the MTase gene. This is to ensure that the two genes are given names that indicate they are part of the same R-M system.

Homing endonucleases

Homing endonucleases have been classified into four families according to conserved sequence motifs. These are the LAGLIDADG, GIY-YIG, H-N-H and His-Cys box families (35). Nomenclature of the homing endonucleases is patterned after that of REases, with a three-letter genus-species designation, followed by a Roman numeral (6). Whereas intron endonucleases are characterized by the prefix I- (for intron), the intein endonucleases are characterized by the prefix PI- (for protein insert), and where the endonuclease is not intron- or intein-encoded, the prefix is F- (for freestanding). The systematic nomenclature does not preclude maintaining historic names. Counter to the original conventions proposed (6), the above nomenclature will extend to putative homing endonucleases without demonstrated catalytic activity. As with hypothetical REases, the suffix P will be used to denote the putative nature of the assignment, and the P will be dropped once nuclease activity has been confirmed. Hybrid homing endonucleases will be preceded by the prefix H-, followed by the authors' designation, e.g., an I-DmoI/I-CreI chimera could be H-DreI, or an I-TevI/I-BmoI hybrid could be H-TevBmo. Those homing endonucleases that have been characterized biochemically will continue to be listed within REBASE (5).

Adherence to these conventions and updates

The authors of this proposal have all agreed to follow these recommendations and it is hoped that other authors and journals will also adhere to these conventions. If further changes become appropriate, then REBASE (5) should be consulted for the latest modifications and practices.

REFERENCES

1. Boyer,H.W. (1971) DNA restriction and modification mechanisms in bacteria. *Annu. Rev. Microbiol.*, **25**, 153–176.
2. Yuan,R. (1981) Structure and mechanism of multifunctional restriction endonucleases. *Annu. Rev. Biochem.*, **50**, 285–315.
3. Smith,H.O. and Nathans,D. (1973) A suggested nomenclature for bacterial host modification and restriction systems and their enzymes. *J. Mol. Biol.*, **81**, 419–423.
4. Szybalski,W., Blumenthal,R.M., Brooks,J.E., Hattman,S. and Raleigh,E.A. (1988) Nomenclature for bacterial genes coding for class-II restriction endonucleases and modification methyltransferases. *Gene*, **74**, 279–280.
5. Roberts,R.J. and Macelis,D. (2003) REBASE—restriction enzymes and methylases. *Nucleic Acids Res.*, **31**, 418–420.
6. Belfort,M. and Roberts,R.J. (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res.*, **25**, 3379–3388.
7. Noyer-Weidner,M., Jentsch,S., Pawlek,B., Gunthert,U. and Trautner,T.A. (1983) Restriction and modification in *Bacillus subtilis*: DNA methylation potential of the related bacteriophages Z, SPR, SPβ, ϕ3T, and ρ11. *J. Virol.*, **46**, 446–453.
8. Klein,R., Baranyi,U., Rossler,N., Greineder,B., Scholz,H. and Witte,A. (2002) *Natrialba magadii* virus ϕCh1: first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. *Mol. Microbiol.*, **45**, 851–863.
9. Karreman,C. and de Waard,A. (1990) *Agmenellum quadruplicatum* M.*Aqu*I, a novel modification methylase. *J. Bacteriol.*, **172**, 266–272.
10. Dryden,D.T., Murray,N.E. and Rao,D.N. (2001) Nucleoside triphosphate-dependent restriction enzymes. *Nucleic Acids Res.*, **29**, 3728–3741.
11. Murray,N.E. (2000) Type I restriction systems: sophisticated molecular machines. *Microbiol. Mol. Biol. Rev.*, **64**, 412–434.
12. Titheradge,A.J., King,J., Ryu,J. and Murray,N.E. (2001) Families of restriction enzymes: an analysis prompted by molecular and genetic data

for type ID restriction and modification systems. *Nucleic Acids Res.*, **29**, 4195–4205.
13. Gubler,M., Braguglia,D., Meyer,J., Piekarowicz,A. and Bickle,T.A. (1992) Recombination of constant and variable modules alters DNA sequence recognition by type IC restriction-modification enzymes. *EMBO J.*, **11**, 233–240.
14. Szybalski,W., Kim,S.C., Hasan,N. and Podhajska,A.J. (1991) Class-IIS restriction enzymes—a review. *Gene*, **100**, 13–26.
15. Stankevicius,K., Lubys,A., Timinskas,A., Vaitkevicius,D. and Janulaitis,A. (1998) Cloning and analysis of the four genes coding for *Bpu*10I restriction-modification enzymes. *Nucleic Acids Res.*, **26**, 1084–1091.
16. Looney,M.C., Moran,L.S., Jack,W.E., Feehery,G.R., Benner,J.S., Slatko,B.E. and Wilson,G.G. (1989) Nucleotide sequence of the *Fok*I restriction-modification system: separate strand-specificity domains in the methyltransferase. *Gene*, **80**, 193–208.
17. Reuter,M., Kupper,D., Meisel,A., Schroeder,C. and Kruger,D.H. (1998) Cooperative binding properties of restriction endonuclease *Eco*RII with DNA recognition sites. *J. Biol. Chem.*, **273**, 8294–8300.
18. Huai,Q., Colandene,J.D., Topal,M.D. and Ke,H. (2001) Structure of *Nae*I-DNA complex reveals dual-mode DNA recognition and complete dimer rearrangement. *Nature Struct. Biol.*, **8**, 665–669.
19. Xu,Y., Lunnen,K.D. and Kong,H. (2001) Engineering a nicking endonuclease N.*Alw*I by domain swapping. *Proc. Natl Acad. Sci. USA*, **98**, 12990–12995.
20. Besnier,C.E. and Kong,H. (2001) Converting *Mly*I endonuclease into a nicking enzyme by changing its oligomerization state. *EMBO Rep.*, **2**, 782–786.
21. Hennecke,F., Kolmar,H., Brundl,K. and Fritz,H.-J. (1991) The *vsr* gene product of *E. coli* K-12 is a strand- and sequence-specific DNA mismatch endonuclease. *Nature*, **253**, 776–778.
22. Tao,T., Bourne,J.C. and Blumenthal,R.M. (1991) A family of regulatory genes associated with Type II restriction-modification systems. *J. Bacteriol.*, **173**, 1367–1375.
23. Sohail,A., Ives,C.L. and Brooks,J.E. (1995) Purification and characterization of C.*Bam*HI, a regulator of the *Bam*HI restriction-modification system. *Gene*, **157**, 227–228.
24. Mucke,M., Reich,S., Moncke-Buchner,E., Reuter,M. and Kruger,D.H. (2001) DNA cleavage by type III restriction-modification enzyme *Eco*P15I is independent of spacer distance between two head to head oriented recognition sites. *J. Mol. Biol.*, **312**, 687–698.
25. Janscak,P., Sandmeier,U., Szczelkun,M.D. and Bickle,T.A. (2001) Subunit assembly and mode of DNA cleavage of the type III restriction endonucleases *Eco*P1I and *Eco*P15I. *J. Mol. Biol.*, **306**, 417–431.
26. Raleigh,E.A. and Wilson,G. (1986) *Escherichia coli* K-12 restricts DNA containing 5-methylcytosine. *Proc. Natl Acad. Sci. USA*, **83**, 9070–9074.
27. Stewart,F.J., Panne,D., Bickle,T.A. and Raleigh,E.A. (2000) Methyl-specific DNA binding by McrBC, a modification-dependent restriction enzyme. *J. Mol. Biol.*, **298**, 611–622.
28. Smith,H.O. and Wilcox,K.W. (1970) A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties. *J. Mol. Biol.*, **51**, 379–391.
29. Kelly,T.J.,Jr and Smith,H.O. (1970) A restriction enzyme from *Hemophilus influenzae.* II. Base sequence of the recognition site. *J. Mol. Biol.*, **51**, 393–409.
30. Roy,P.H. and Smith,H.O. (1973) DNA methylases of *Hemophilus influenzae* Rd. I. Purification and properties. *J. Mol. Biol.*, **81**, 427–444.
31. Roy,P.H. and Smith,H.O. (1973) DNA methylases of *Hemophilus influenzae* Rd. II. Partial recognition site base sequences. *J. Mol. Biol.*, **81**, 445–459.
32. Gromkova,R., Bendler,J. and Goodgal,S. (1973) Restriction and modification of bacteriophage S2 in *Haemophilus influenzae*. *J. Bacteriol.*, **114**, 1151–1157.
33. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J., Dougherty,B.A. Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
34. Dusterhoft,A., Erdmann,D. and Kroger,M. (1991) Stepwise cloning and molecular characterization of the *Hgi*DI restriction-modification system from *Herpetosiphon giganteus* Hpa2. *Nucleic Acids Res.*, **19**, 1049–1056.
35. Belfort,M., Derbyshire,V., Cousineau,B. and Lambowitz,A. (2002) Mobile introns: pathways and proteins. In Craig,N., Craigie,R., Gellert,M. and Lambowitz,A. (eds), *Mobile DNA II*. ASM Press, Washington, DC, pp. 761–783.