1 DPSN: standardizing the short names of amplicon-sequencing primers to avoid ambiguity

2

3 Yuxiang Tan^{#1}, Yixia Tian^{#1}, Junyu Chen², Zhinan Yin¹, Hengwen Yang^{*1}

4

5 ¹ The First Affiliated Hospital, Biomedical Translational Research Institute, Guangdong Province

6 Key Laboratory of Molecular Immunology and Antibody Engineering, Jinan University,

7 Guangzhou 510632, China

8 ² Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese

9 Academy of Sciences, Shenzhen 518055, People's Republic of China

10 [#] These authors contributed equally

11

12 ^{*} Corresponding authors: Dr. Hengwen Yang benyang97@gmail.com. Tel/ Fax: +86-020-85222787

13

Abstract

Background: Amplicon sequencing is the most widely used sequencing method to evaluate microbial diversity in virtually all environments. Thus, appropriate and specific primers are needed to amplify amplicon regions in amplicon sequencing. For this purpose, the community currently uses probeBase, which curates rRNA-targeted probes and primers. **Main Body:** We found that 63.58% of the primers in probeBase have problematic issues in the short name, full name, and/or position. Furthermore, the current convention for short names causes ambiguity. We here introduce our new Database of Primer Scientific Names (DPSN), which is a manually curated database for the 173 primers in probeBase complete with a new short name convention. Building on the work of probeBase, we provide a more user-friendly and standardized system. The new short primer naming convention has three basic components: 5' position on the sense strand, version, and direction. An additional character for the name of the taxonomic group is also added in front of the name for convenient use. Furthermore, DPSN contains primers for large subunit as well. In order to separate them from the primers for small subunit, a header character is also recommended. **Conclusion:** All 173 primers in probeBase were corrected according to this new rule, and are stored in DPSN, which

29 is expected to facilitate accurate primer selection and better standardized communication in this

30 field.

31 **Database URL:** The DPSN database is available in a user-interactive website at

32 <http://dpsn.leidailab.cn/>

33 Keywords: database, scientific name, amplicon, primer

34

35 Background

36 Amplicon sequencing is a common sequencing method for microbial research from diverse
37 environmental or clinical samples [1, 2]. Amplicon sequencing is dependent on the choice of
38 primers for carrying out the amplification step. Thus, selection of the most appropriate primers is
39 the foundation of successful amplicon sequencing.

40 probeBase [3] is the only database currently available with updated lists of probes and
41 primers, along with links to other databases providing related information. At present, there is a
42 total of 173 primers recorded in probeBase. In general, a primer is defined according to its short
43 name (SN), full name (FN), and sequence. However, in many cases, only the SN is used for the
44 sake of convenience. There are seven components of an FN [4]. Taking S-D-Bact-0338-a-A-18 as
45 an example: “S” stands for the target gene (Small Sub-Unit (SSU)), “D” represents the largest
46 taxonomic group targeted (Domain), “Bact” is the name of the taxonomic group (Bacteria), “0338”
47 is the 5' position of the sense strand, “a” presents the version, “A” denotes the identical strand
48 (“S” for sense; “A” for antisense), and “18” is the length of the primer. To avoid ambiguity, each
49 primer should have a unique SN; however, this is not the case. Different from FN, there is no

50 guideline for how an SN should be. Therefore, SNs were named in a few different ways, such as
 51 Primer3, Bac927, 926r, and 934mcr. The most common ones were composited by the position and
 52 direction (for example, 926r), or with an additional string for the name of the taxonomic group
 53 (for example, Arch 915r). The lack of clear rules and sufficient information for accuracy leads to
 54 ambiguity of SNs.

55 In fact, there are 14 SNs that refer to multiple primer sequences, which could lead to
 56 confusion and cause several problems in application for users. For example, in the earth
 57 microbiome project website [1], the author of the citation for a given primer is used along with the
 58 SN to better specify the primer. Furthermore, the SN itself could be misleading. For example,
 59 primers 907r and 926r are actually from the same region of the genome but with a difference of
 60 two bases in the sequence. However, based on their SNs alone, a user would misinterpret these
 61 primers as being derived from two different regions.

62 To resolve this problem, we here introduce Database of Primer Scientific Names (DPSN),
 63 which is a database that has been manually curated to correct problematic and inconsistent
 64 features (SN, FN, position, and length) of primers in probeBase according to an improved

65 convention of naming SNs. The new SNs still correspond to the old SNs and corrected FNs in a
66 one-to-one relation.

67 Construction and content

68 Data source

69 Information of all 173 primers in the probeBase dataset was manually extracted [3],
70 including the SN, FN, position, sequence, length, G+C content, and dissociation temperature.

71 The corresponding regions on the reference sequence of *Escherichia coli* K-12 substrain
72 MG1655 was extracted from the SILVA database [5] and served as the reference for confirming
73 the sequence position.

74 Derivation of new SN naming convention

75 A unique SN should have at least three basic components to provide sufficient identifying
76 information: 5' position on the sense strand, version, and direction.

77 In the old SN, such as 907r, all reverse primers that start or end from position 907 will have
78 the same name, which leads to ambiguity. Therefore, including additional information of the

79 version, such as 907ar to indicate the version, could help to specify the primer sequence.

80 Consistent with the old SN rule, “f” and “r” denote “forward primer” and “reverse primer,”

81 respectively. Moreover, because the name of the taxonomic group provides helpful information for

82 users to select appropriate primers, DPSN also includes a shorthand for the name of the taxonomic

83 group (“A” for Archaea, “B” for Bacteria, “U” for universal, and “N” for nano) in front of the SN.

84 Additionally, on account of the need to separate SSU primers from large subunit (LSU) primers,

85 the header represents of target get from the FNs is retained. For instance, the SN 907ar represents

86 the FN “S-D-Bact-0907-a-A-20”, which was corrected and recorded as S-B907ar in DPSN.

87 **Amplification range validation of the primers**

88 To validate the amplification location of primers according to the *E. coli* K-12 reference,

89 BLAT of the National Center for Biotechnology Information [6] was employed as the aligner.

90 However, because of the presence of degenerate bases in primer sequences, the primers had to be

91 converted into expanded regular sequences, which was achieved using a customized Python script

92 before BLAT alignment. In particular, the additional parameters “-minMatch = 1-minScore =

93 8-minIdentity = 70-stepSize =1-tileSize =8” were applied to BLAT, considering the length and

94 mismatch tolerance of primers.

95 To confirm the amplification location, the primers were also checked by the TestProbe

96 function in SILVA [5].

97

98 **Utility and Discussion**

99 **How to use DPSN**

100 In order to make the search easy, DPSN supports fuzzy search on all the fields. This means

101 user can use any keyword to find the related primer(s), such as the intended 5' position and the

102 sub-string of the primer sequence. In return, DPSN will present the corrected information of the

103 related primers. As well as the original "Short Name" and "Full Name" from the probeBase,

104 which will help the user to connect the use of the primers in original papers. All the sequence in

105 DPSN are the same as the ones in probeBase.

106 **Summary of Corrections and Discussion**

107 Our careful review of probeBase identified five sequences with multiple primer names, 14

108 groups of SNs with multiple targets, and 91 SNs inconsistent with their FNs. Of the total 173
109 primers in probeBase, the SNs for only 63 primers (36.42%) could direct the user to a unique
110 sequence and be considered as correct. Five sequences were multifold and had multiple primer
111 names (Table 1). Thirty SNs from 14 groups pointed to more than one sequence. The positions in
112 91 SNs were different from the 5' position of their FNs. Overall, the positions of 35 primers in
113 probeBase were found to be incorrect.

114 In addition, a few FNs were found to be incorrect in probeBase, which have been manually
115 corrected in DPSN. Theoretically, the FNs of primers should be unique, since a single FN
116 represents a unique primer sequence. However, in probeBase, three FNs were duplicated and even
117 represented more than one sequence (Table 2). In the naming rule, the position in an FN is based
118 on the 5' position; however, eight of the primers in probeBase violated this rule (Table 3). Even
119 more importantly, the length information of five FNs did not match the actual lengths of their
120 sequences (Table 4), and the directions of three primers were opposite to the actual direction of
121 their sequences (Table 5).

122 **Table 2: Primer groups with the same FN in probeBase**

Old SN	Old FN	Position	Sequence	GC%	TM*
Arch958f	S-D-Arch-0958-a-S-19	958–975	AATTGGANTCAACGCCCGG	50	49
Arch958Bf	S-D-Arch-0958-a-S-19	958–976	AATTGGABTCAACGCCCGGR	47	51.5
b785	S-D-Bact-0785-a-A-19	785–803	CTACCAGGGTATCTAATCC	47	49
803r	S-D-Bact-0785-a-A-19	785–803	CTACCRGGGTATCTAATCC	47	50
518r	S-D-Bact-0518-a-A-17	518–534	ATTACCGCGGCTGCTGG	65	52
P518r	S-D-Bact-0518-a-A-17	518–534	ATTACCGCGGCTGCTGG	65	52

123 *TM: dissociation temperature (°C)

124 **Table 3: FNs of primers with inconsistent positions in probeBase**

Old SN	Old FN	Position	Sequence	GC%	TM*
338	S-D-Bact-0338-a-A-19	337–355	TGCTGCCTCCCGTAGGA GT	63	58
1114mcr	S-P-Nano-1082-a-A-17	915–931	GGGTCTCGCCTGTTTCC	65	52
27F	S-D-Bact-0008-d-S-20	6–25	AGAGTTTGATCMTGGCT CAG	45	51

63F	S-D-Bact-0043-s- S-21	21–41	CAGGCCTAACACATGCA AGTC	52	52
Arch855R	S-D-Arch-0896-a -A-20	915–934	TCCCCCGCCAATTCCTTT AA	50	52
bio-pJBS-V3.SER	S-D-Bact-0947-a- A-20	946–965	GGTAAGGTTCTTCGCGT TGC	55	53
Primer3	S-D-Bact-0518-c- A-17	340–357	GCCTACGGGAGGCAGCA G	72	57
Primer2	S-D-Bact-0340-a- A-18	518–534	ATTACCGCGGCTGCTGG	65	52

125 *TM: dissociation temperature (°C)

126

127 **Table 4: FNs of primers with the wrong length in probeBase**

Old SN	Old FN	Position	Sequence	GC%	TM*	Corrected length
Uni522r	S-*-Univ-0517-a-A- 15	517–534	GWATTACCGCG GCKGCTG	61	54	18
Primer2	S-D-Bact-0340-a-A- 18	518–534	ATTACCGCGGC TGCTGG	65	52	17
U529r	S-*-Univ-0522-a-A- 18	522–536	ACCGCGGCKGC TGGC	80	54.5	15

Primer3	S-D-Bact-0518-c-A-17	340–357	GCCTACGGGAG GCAGCAG	72	57	18
Arch958f	S-D-Arch-0958-a-S-19	958–975	AATTGGANTCA ACGCCGG	50	49	18

128 *TM: dissociation temperature (°C)

129

130

131 **Table 5: FNs of primers with the wrong strand in probeBase**

Old SN	Old FN	Position	Sequence	GC%	TM *	Corrected FN
Primer3	S-D-Bact-0518-c-A-17	340–357	GCCTACGGGAGG CAGCAG	72	57	S-D-Bact-0340-a-S-18
527f	S-D-Bact-0517-a-S-16	517–532	ACCGCGGCCKGC TGGC	81	66	S-D-Bact-0517-a-A-16
536r	S-D-Bact-0519-a-A-18	519–536	CAGCMGCCGCG GTAATWC	61	54	S-D-Bact-0519-a-S-18

132 *TM: dissociation temperature (°C)

133 In DPSN, all of the SNs of the primers in probeBase have been updated according to the new

134 naming rule along with additional version information to provide a more unique identifier that is

135 still convenient to use. Overall, 110 problematic primers were corrected. Using the amended

136 primer name in DPSN, users can simply refer to the SN to specify a primer, because of the
137 one-to-one relation among the SN, FN, and sequence, and without the inconvenience of appending
138 additional information such as author name or sequence in the article.

139 Conclusion

140 In conclusion, because it is crucial to avoid vagueness in scientific research, the old SN
141 system of primers is problematic and should be replaced by the new naming rule as proposed
142 herein. All of these corrections have been curated in DPSN to improve searching convenience and
143 accuracy. Therefore, with DPSN, users can easily search an old name from probeBase or articles
144 for its amended name. For new articles, it is recommended that authors use the amended name to
145 accurately describe a primer.

146 DPSN currently focuses on only primers for amplicon sequencing on SSU and LSU, and thus
147 it can be assumed that the ambiguity problem still exists for primers in other amplicon regions,
148 such as ITS. Because the primers for these regions were not found in probeBase, we can collect
149 and import these primers into the naming system in DPSN in the future. To keep the database up
150 to date, DPSN accepts data submission of primers from researchers as well.

151 **Abbreviations**

152 **DPSN:** Database of Primers' Scientific Names **SN:** short name **FN:** full name

153 **Declarations**

154 **Ethics approval and consent to participate**

155 Not applicable

156 **Consent for publication**

157 Not applicable

158 **Availability of data and material**

159 Database Name: Database of Primers' Scientific Names (DPSN)

160 Database URL: <http://dpsn.leidailab.cn/>

161 **Competing interests**

162 The authors declare that they have no competing interests.

163 **Funding**

164 This work was supported by the Science and Technology Department of Guangdong Province of
165 China (grant no. 2017A030310179 to Dr. Yuxiang Tan); The Major International Joint Research
166 Program of China (grant no. 31420103901), and the “111 project” (grant no. B16021) to Dr.
167 Zhinan Yin and Hengwen Yang. The three grants supported the whole study through design and
168 collection, analysis and interpretation of data and in writing the manuscript.

169 **Authors’ contributions**

170 YuT conceived of the idea, conducted the analysis, and wrote the manuscript. YiT performed the
171 data collection. JC provided data of LSU primers. HY and ZY supervised the project and
172 participated in the revision of the manuscript. All authors read and approved the final manuscript.

173 **Acknowledgements**

174 We would like to thank Becky Kusko for editing suggestions and Editage (www.editage.com) for
175 English language editing.

176 **References**

- 177 1. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations.
178 BMC Biol. 2014;12:69.

- 179 2. The Human Microbiome Project Consortium. A framework for human microbiome research.

180 Nature 2012;**486**:215-21.
- 181 3. Greuter D, Loy A, Horn M, Rattei T. probeBase—an online resource for rRNA-targeted

182 oligonucleotide probes and primers: new features. Nucleic Acids Res. 2016;44:D586-9.
- 183 4. Klindworth A, Pruesse E, Schwwer T, Peplies J, Quast C, Horn M, et al. Evaluation of general

184 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based

185 diversity studies. Nucleic Acids Res. 2013;41:e1.
- 186 5. Quast C, Pruesse E, Yilmaz P, Gerken J, Schwwer T, Yarza P, Peplies J, Glöckner FO. The

187 SILVA ribosomal RNA gene database project: improved data processing and web-based tools.

188 Nucleic Acids Res. 2013;41:D590-6.
- 189 6. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656-4.

190

191

Table 1: Primer groups with the same sequence in probeBase

Old SN	Old FN	Position	Sequence	GC%	TM*	Corr. SN	Corr. FN
Arch95	S-D-Arch-09	938–956	AATTGGABTCA	47	51.5	A958bf	S-D-Arch -0958-b-S
8Bf	38-b-S-19		ACGCCGGR				
Arch95	S-D-Arch-09	958–976	AATTGGABTCA	47	51.5		-19
8Bf	58-a-S-19		ACGCCGGR				
U519f	S-*-Univ-051	519–536	CAGCMGCCGCG	61	54	U519bf	S-*-Univ- 0519-a-S-
	9-a-S-18		GTAATWC				
536r	S-D-Bact-051	519–536	CAGCMGCCGCG	61	54		18
	9-a-A-18		GTAATWC				
518r	S-D-Bact-051	518–534	ATTACCGCGGCT	65	52	B518ar	S-D-Bact- 0518-a-A-
	8-a-A-17		GCTGG				
P518r	S-D-Bact-051	518–534	ATTACCGCGGCT	65	52		17
	8-a-A-17		GCTGG				
Primer2	S-D-Bact-034	518–534	ATTACCGCGGCT	65	52		
	0-a-A-18		GCTGG				
63f	S-D-Bact-004	21–41	CAGGCCTAACA	52	52	B43af	S-D-Bact-

	3-s-S-21		CATGCAAGTC			0043-a-S-
						21
P63f	S-D-Bact-004	42–62	CAGGCCTAACA	52	55	
	2-a-S-21		CATGCAAGTC			
A21f	S-D-Arch-00	7–26	TTCCGGTTGATC	60	57	
	07-b-S-20		CYGCCGGA			S-D-Arch
						A6bf
						-0006-b-S
A2f	S-D-Arch-00	7–26	TTCCGGTTGATC	60	57	-20
	07-a-S-20		CYGCCGGA			

192

193 *TM: dissociation temperature (°C)

194