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

Yuxiang Tan[#1], Yixia Tian[#1], Junyu Chen[2], Zhinan Yin[1], Hengwen Yang*[1]

[1] The First Affiliated Hospital, Biomedical Translational Research Institute, Guangdong Province Key Laboratory of Molecular Immunology and Antibody Engineering, Jinan University, Guangzhou 510632, China

[2] Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, People's Republic of China

[#] These authors contributed equally

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

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.gdimmunity.com

33 Keywords: database, scientific name, amplicon, primer

34

Background

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

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

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

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

To resolve this problem, we here introduce Database of Primer Scientific Names (DPSN), which is a database that has been manually curated to correct problematic and inconsistent features (SN, FN, position, and length) of primers in probeBase according to an improved convention of naming SNs. The new SNs still correspond to the old SNs and corrected FNs in a one-to-one relation.

Construction and content

Data source

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

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

Derivation of new SN naming convention

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

In the old SN, such as 907r, all reverse primers that start or end from position 907 will have the same name, which leads to ambiguity. Therefore, including additional information of the version, such as 907ar to indicate the version, could help to specify the primer sequence. Consistent with the old SN rule, “f” and “r” denote “forward primer” and “reverse primer,” respectively. Moreover,

because the name of the taxonomic group provides helpful information for users to select appropriate primers, DPSN also includes a shorthand for the name of the taxonomic group ("A" for Archaea, "B" for Bacteria, "U" for universal, and "N" for nano) in front of the SN. Additionally, on account of the need to separate SSU primers from large subunit (LSU) primers, the header represents of target get from the FNs is retained. For instance, the SN 907ar represents the FN "S-D-Bact-0907-a-A-20", which was corrected and recorded as S-B907ar in DPSN.

Amplification range validation of the primers

To validate the amplification location of primers according to the *E. coli* K-12 reference, BLAT of the National Center for Biotechnology Information [6] was employed as the aligner. However, because of the presence of degenerate bases in primer sequences, the primers had to be converted into expanded regular sequences, which was achieved using a customized Python script before BLAT alignment. In particular, the additional parameters "-minMatch = 1-minScore = 8-minIdentity = 70-stepSize =1-tileSize =8" were applied to BLAT, considering the length and mismatch tolerance of primers.

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

in SILVA [5].

Utility and Discussion

How to use DPSN

In order to make the search easy, DPSN supports fuzzy search on all the fields. This means user can use any keyword to find the related primer(s), such as the intended 5' position and the sub-string of the primer sequence. In return, DPSN will present the corrected information of the related primers. As well as the original "Short Name" and "Full Name" from the probeBase, which will help the user to connect the use of the primers in original papers. All the sequence in DPSN are the same as the ones in probeBase.

Summary of Corrections and Discussion

Our careful review of probeBase identified five sequences with multiple primer names, 14 groups of SNs with multiple targets, and 91 SNs inconsistent with their FNs. Of the total 173 primers in probeBase, the SNs for only 63 primers (36.42%) could direct the user to a unique sequence and

be considered as correct. Five sequences were multifold and had multiple primer names (Table 1). Thirty SNs from 14 groups pointed to more than one sequence. The positions in 91 SNs were different from the 5′ position of their FNs. Overall, the positions of 35 primers in probeBase were found to be incorrect.

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

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	AATTGGANTCAACGCCGG	50	49
Arch958Bf	S-D-Arch-0958-a-S-19	958–976	AATTGGABTCAACGCCGGR	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

*TM: dissociation temperature (°C)

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	TGCTGCCTCCCGTAGGAGT	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	AGAGTTTGATCMTGGCTCAG	45	51
63F	S-D-Bact-0043-s-S-21	21–41	CAGGCCTAACACATGCAAGTC	52	52
Arch855R	S-D-Arch-0896-a-A-20	915–934	TCCCCCGCCAATTCCTTTAA	50	52

bio-pJBS-V3.SER	S-D-Bact-0947-a-A-20	946–965	GGTAAGGTTCTTCGCGTTGC	55	53
Primer3	S-D-Bact-0518-c-A-17	340–357	GCCTACGGGAGGCAGCAG	72	57
Primer2	S-D-Bact-0340-a-A-18	518–534	ATTACCGCGGCTGCTGG	65	52

*TM: dissociation temperature (°C)

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	GWATTACCGCGGCKGCTG	61	54	18
Primer2	S-D-Bact-0340-a-A-18	518–534	ATTACCGCGGCTGCTGG	65	52	17
U529r	S-*-Univ-0522-a-A-18	522–536	ACCGCGGCKGCTGGC	80	54.5	15
Primer3	S-D-Bact-0518-c-A-17	340–357	GCCTACGGGAGGCAGCAG	72	57	18
Arch958f	S-D-Arch-0958-a-S-19	958–975	AATTGGANTCAACGCCGG	50	49	18

*TM: dissociation temperature (°C)

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	GCCTACGGGAGGCAGCAG	72	57	S-D-Bact-0340-a-S-18
527f	S-D-Bact-0517-a-S-16	517–532	ACCGCGGCCKGCTGGC	81	66	S-D-Bact-0517-a-A-16
536r	S-D-Bact-0519-a-A-18	519–536	CAGCMGCCGCGGTAATWC	61	54	S-D-Bact-0519-a-S-18

*TM: dissociation temperature (°C)

In DPSN, all of the SNs of the primers in probeBase have been updated according to the new naming rule along with additional version information to provide a more unique identifier that is still convenient to use. Overall, 110 problematic primers were corrected. Using the amended primer name in DPSN, users can simply refer to the SN to specify a primer, because of the one-to-one relation among the SN, FN, and sequence, and without the inconvenience of appending additional information such as author name or sequence in the article.

Conclusion

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

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

Abbreviations

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

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

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

Database URL: http://dpsn.gdimmunity.com

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Science and Technology Department of Guangdong Province of China (grant no. 2017A030310179 to Dr. Yuxiang Tan); The Major International Joint Research

Program of China (grant no. 31420103901), and the "111 project" (grant no. B16021) to Dr. Zhinan Yin and Hengwen Yang. The three grants supported the whole study through design and collection, analysis and interpretation of data and in writing the manuscript.

Authors' contributions

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

Acknowledgements

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

References

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

2. The Human Microbiome Project Consortium. A framework for human microbiome research. Nature 2012;**486**:215-21.

179 3. Greuter D, Loy A, Horn M, Rattei T. probeBase—an online resource for rRNA-targeted
180 oligonucleotide probes and primers: new features. Nucleic Acids Res. 2016;44:D586-9.

181 4. Klindworth A, Pruesse E, Schwwer T, Peplies J, Quast C, Horn M, et al. Evaluation of general
182 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based
183 diversity studies. Nucleic Acids Res. 2013;41:e1.

184 5. Quast C, Pruesse E, Yilmaz P, Gerken J, Schwwer T, Yarza P, Peplies J, Glöckner FO. The
185 SILVA ribosomal RNA gene database project: improved data processing and web-based tools.
186 Nucleic Acids Res. 2013;41:D590-6.

187 6. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656-4.

188

189 **Table 1: Primer groups with the same sequence in probeBase**

Old SN	Old FN	Position	Sequence	GC%	TM*	Corr. SN	Corr. FN
Arch958Bf	S-D-Arch-0938-b-S-19	938–956	AATTGGABTCAACGCCGGR	47	51.5	A958bf	S-D-Arch-0958-b-S-19
Arch958Bf	S-D-Arch-0958-a-S-19	958–976	AATTGGABTCAACGCCGGR	47	51.5		
U519f	S-*-Univ-0519-a-S-18	519–536	CAGCMGCCGCGGTAATWC	61	54	U519bf	S-*-Univ-0519-a-S-18
536r	S-D-Bact-0519-a-A-18	519–536	CAGCMGCCGCGGTAATWC	61	54		
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	B518ar	S-D-Bact-0518-a-A-17
Primer2	S-D-Bact-0340-a-A-18	518–534	ATTACCGCGGCTGCTGG	65	52		
63f	S-D-Bact-	21–41	CAGGCCTAACA	52	52	B43af	S-D-Bact-

	0043-s-S-21		CATGCAAGTC				0043-a-S-21
P63f	S-D-Bact-0042-a-S-21	42–62	CAGGCCTAACA CATGCAAGTC	52	55		
A21f	S-D-Arch-0007-b-S-20	7–26	TTCCGGTTGATC CYGCCGGA	60	57	A6bf	S-D-Arch-0006-b-S-20
A2f	S-D-Arch-0007-a-S-20	7–26	TTCCGGTTGATC CYGCCGGA	60	57		

*TM: dissociation temperature (°C)