



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

Yuxiang Tan<sup>1,2#</sup>, Yixia Tian<sup>1#</sup>, Junyu Chen<sup>2</sup>, Zhinan Yin<sup>1</sup>, Hengwen Yang<sup>1\*</sup>

<sup>1</sup> The First Affiliated Hospital, Biomedical Translational Research Institute, Jinan University, Guangzhou 510632, Guangdong Province, China

<sup>2</sup> CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, Guangdong Province, China

**Abstract:** 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. However 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 from probeBase and 42 new added primers 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. All 173 primers in probeBase were corrected according to this new rule, and are stored in DPSN, which is expected to facilitate accurate primer selection and better standardized communication in this field.

**Keywords:** database, scientific name, amplicon, sequencing, primer

Amplicon sequencing is a common sequencing method for microbial research from diverse environmental or clinical samples<sup>[1-2]</sup>. 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<sup>[3]</sup> is the only database currently

Supported by the National Natural Science Foundation of China (32070121), by the Science and Technology Department of Guangdong Province of China (2017A030310179) and by the “111 Project” (B16021)

<sup>#</sup>These authors contributed equally to this study.

\*Corresponding author. Tel/Fax: +86-20-85222787; E-mail: [benyang97@gmail.com](mailto:benyang97@gmail.com)

Received: 12 December 2020; Revised: 10 March 2021; Published online: 27 May 2021

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<sup>[4]</sup>. 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 (Supplemental Table 1), which could lead to confusion and cause several problems in application for users. For example, in the earth microbiome project website (<https://earthmicrobiome.org/protocols-and-standards/16s/>)<sup>[1]</sup>, the author of the citation for a given primer was used along with the SN (515F (Parada)–806R (Aprill)) to better specify the primer, which could be avoided by a better nomenclature system. 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.

## 1 Construction and content

### 1.1 Data source

Information of all 173 primers in the probeBase dataset was manually extracted<sup>[3]</sup>, 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<sup>[5]</sup> and served as the reference for confirming the sequence position.

### 1.2 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 27f, which is commonly used in full length 16S amplification, all forward primers that start or end from position 27 will have the same name, which leads to ambiguity and misunderstanding of the true target region. Therefore, including additional information of the version, such as 27ar 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 old SN 27f represents three different FN “S-D-Bact-0009-b-S-19”, “S-D-Bact-0008-d-S-20” and “S-D-Bact-0008-c-S-20” (Supplemental Table 1), which was corrected and recorded as S-B9bf, S-B8df and S-B8cf respectively in DPSN. Then users will know S-B9bf is the forward primer starts from position 9 on 16S mainly for Bacteria detection, while S-B8df and S-B8cf starts from position 8 and their sequence must be different because of the different version ID.

### 1.3 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<sup>[6]</sup> was employed as the aligner.

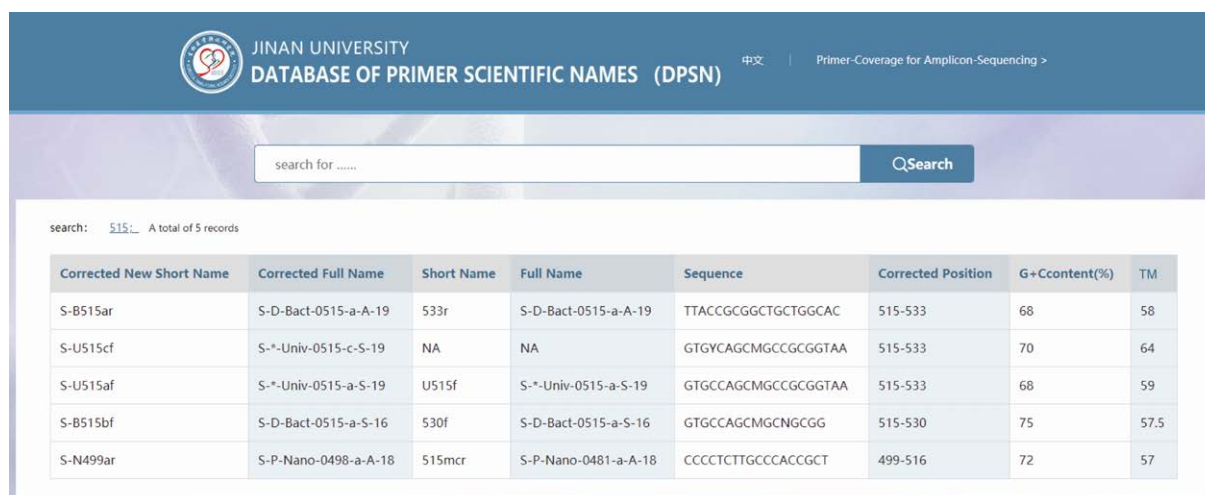
However, because of the presence of degenerate bases in primer sequences, the primers had to be converted into expanded regular sequences, which was achieved using a customized Python script before BLAT alignment. In particular, the additional parameters “-minMatch =1 -minScore =8 -minIdentity =70 -stepSize =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<sup>[5]</sup>.

## 2 Utility and Discussion

### 2.1 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 (Figure 1). In



Corrected New Short Name	Corrected Full Name	Short Name	Full Name	Sequence	Corrected Position	G+Ccontent(%)	TM
S-B515ar	S-D-Bact-0515-a-A-19	533r	S-D-Bact-0515-a-A-19	TTACCGCGGCTGCTGGCAC	515-533	68	58
S-U515cf	S-*-Univ-0515-c-S-19	NA	NA	GTGYCAGCMGCCCGGTAA	515-533	70	64
S-U515af	S-*-Univ-0515-a-S-19	U515f	S-*-Univ-0515-a-S-19	GTGCCAGCMGCCCGGTAA	515-533	68	59
S-B515bf	S-D-Bact-0515-a-S-16	530f	S-D-Bact-0515-a-S-16	GTGCCAGCMGCNCGG	515-530	75	57.5
S-N499ar	S-P-Nano-0498-a-A-18	515mcr	S-P-Nano-0481-a-A-18	CCCCTCTGCCACCGCT	499-516	72	57

Figure 1. An example of using DPSN. By searching for 515, all the primers (both forward and reverse) contain this position information will be listed. Users can find the “Corrected New Short Name” as the SN in the new naming convention. The “Corrected Full Name” might be different from “Full Name”, if the FN in probeBase was wrong (such as the fifth primer “S-P-Nano-0481-a-A-18” was corrected as “S-P-Nano-0498-a-A-18”). “Short Name”, “Full Name”, “Sequence”, “G+C content(%)” and “TM” (dissociation temperature) were from probeBase. The “Corrected Position” was checked by both BLAT and SILVA using the primer sequence.

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.

## 2.2 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).

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.

Table 1. Primer groups with the same sequence in probeBase

Old SN	Old FN	Position	Sequence	GC%	TM/°C*	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	B518ar	S-D-Bact-0518-a-A-17
P518r	S-D-Bact-0518-a-A-17	518–534	ATTACCGCGGCTGCTGG	65	52		
Primer2	S-D-Bact-0340-a-A-18	518–534	ATTACCGCGGCTGCTGG	65	52		
63f	S-D-Bact-0043-s-S-21	21–41	CAGGCCTAACACATGCAAGTC	52	52	B43af	S-D-Bact-0043-a-S-21
P63f	S-D-Bact-0042-a-S-21	42–62	CAGGCCTAACACATGCAAGTC	52	55		
A21f	S-D-Arch-0007-b-S-20	7–26	TTCCGGTTGATCCYGCCGGA	60	57	A6bf	S-D-Arch-0006-b-S-20
A2f	S-D-Arch-0007-a-S-20	7–26	TTCCGGTTGATCCYGCCGGA	60	57		

\*TM: dissociation temperature (°C).

Table 2. Primer groups with the same FN in probeBase

Old SN	Old FN	Position	Sequence	GC%	TM/°C*
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/°C*
338	S-D-Bact-0338-a-A-19	337–355	TGCTGCCTCCCGTAGGAGT	63	58
1114mcr	S-P-Nano-1082-a-A-17	915–931	GGGTCTCGCTGTTTCC	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/°C*	Corrected length/bp
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/°C*	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	ACCGCGGCKGCTGGC	81	66	S-D-Bact-0517-a-A-16
536r	S-D-Bact-0519-a-A-18	519–536	CAGCMGCCGCGTAATWC	61	54	S-D-Bact-0519-a-S-18

\*TM: dissociation temperature (°C).

### 3 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.

### Authors' contributions

Yuxiang Tan conceived of the idea, conducted the analysis, and wrote the manuscript. Yixia Tan performed the data collection. Junyu Chen provided

data of LSU primers. Hengwen Yang and Zhina Yin 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](http://www.editage.com)) for English language editing.

### References

- [1] Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. *BMC Biology*, 2014, 12: 69.
- [2] The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 2012, 486(7402): 215–221.
- [3] Greuter D, Loy A, Horn M, Rattei T. probeBase—an online resource for rRNA-targeted oligonucleotide probes and primers: new features 2016. *Nucleic Acids Research*, 2016, 44(D1): D586–D589.
- [4] Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, 2013, 41(1): e1.
- [5] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 2013, 41(Database issue): D590–D596
- [6] Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Research*. 2002, 12: 656–664.

# DPSN: 扩增子测序引物标准化命名数据库

谭宇翔<sup>1,2#</sup>, 田义霞<sup>1#</sup>, 陈俊宇<sup>2</sup>, 尹芝南<sup>1</sup>, 杨恒文<sup>1\*</sup>

<sup>1</sup>暨南大学附属第一医院, 生物医学转化研究院, 广东 广州 510632

<sup>2</sup>中国科学院深圳先进技术研究院, 深圳合成生物学创新研究院, 中国科学院定量工程生物学重点实验室, 广东 深圳 518055

**摘要:** 扩增子测序是目前用来衡量微生物多样性时使用最广泛的测序手段。因为其扩增的需要, 所以选择合适的特定引物是必需的, 且其对结果影响甚大。probeBase 是目前最常用的记录了人工校正后的 rRNA 探针和引物的数据库。然而, 我们发现 probeBase 中 63.58% 的引物存在不同程度的注释错误, 包括命名重复、命名无规律以及匹配位置错误等。更严重的是, 目前主流的短命名方式不具有唯一性, 导致对应关系模糊不清。因此, 我们定义了更简单可行的短命名标准并开发了新的引物科学命名数据库 (DPSN), 并对 probeBase 里的所有 173 个引物进行了校正, 并新加入了 4 个新的改良引物以及 38 个针对大亚基的新引物。新的短命名规则包含 3 个基本要素: 在正链 5' 端的位置、版本号和方向。此外在前面加入了识别大/小亚基以及主要针对的菌界的标志。使用 DPSN, 可以快速查找感兴趣区域对应的引物并比较不同版本的差异选择合适的引物。同时还能建立命名与序列的明确一一对应关系, 避免歧义。

**关键词:** 数据库, 命名, 扩增子, 测序, 引物

(本文责编: 张晓丽)

基金项目: 国家自然科学基金(32070121); 广东省自然科学基金(2017A030310179); “111”项目(B16021)

#共同第一作者。

\*通信作者。Tel/Fax: +86-20-85222787; E-mail: benyang97@gmail.com

收稿日期: 2020-12-12; 修回日期: 2021-03-10; 网络出版日期: 2021-05-27

**谭宇翔**, 博士, 中国科学院深圳先进技术研究院助理研究员。广东省“珠江人才计划”海外青年引进计划和深圳市海外高层次人才“孔雀计划”获得者。博士毕业于波士顿大学生物信息学系, 主要从事转录组数据分析。博士后于暨南大学生物医学转化研究院开始从事肠道微生物的扩增子测序数据分析。当前研究主要针对肠道微生物组, 结合宏基因组测序和培养组技术, 进行深入至菌株层级的生物信息学分析。主要致力于炎症性肠病的肠道紊乱与发病机制的关系, 以及婴幼儿肠道微生物组的发育评估和菌群失调与疾病的关系。实验室已建立了一系列针对肠道微生物的生物信息分析流程, 并为了提高生物信息分析的精度及可重复性, 自主开发了相关数据库和新的分析算法。



## 补充材料

Supplemental table 1. Primer groups with the same short name in probeBase

本文补充材料见网络版 <http://journals.im.ac.cn/actamicrocn>。补充材料为作者提供的原始数据, 作者对其学术质量和内容负责。

<http://journals.im.ac.cn/actamicrocn>