

STRBase: a short tandem repeat DNA database for the human identity testing community

Christian M. Ruitberg, Dennis J. Reeder and John M. Butler*

Biotechnology Division, National Institute of Standards and Technology, 100 Bureau Drive, Mail Stop 8311, Gaithersburg, MD 20899-8311, USA

Received August 31, 2000; Accepted September 17, 2000

ABSTRACT

The National Institute of Standards and Technology (NIST) has compiled and maintained a Short Tandem Repeat DNA Internet Database (<http://www.cstl.nist.gov/biotech/strbase/>) since 1997 commonly referred to as STRBase. This database is an information resource for the forensic DNA typing community with details on commonly used short tandem repeat (STR) DNA markers. STRBase consolidates and organizes the abundant literature on this subject to facilitate on-going efforts in DNA typing. Observed alleles and annotated sequence for each STR locus are described along with a review of STR analysis technologies. Additionally, commercially available STR multiplex kits are described, published polymerase chain reaction (PCR) primer sequences are reported, and validation studies conducted by a number of forensic laboratories are listed. To supplement the technical information, addresses for scientists and hyperlinks to organizations working in this area are available, along with the comprehensive reference list of over 1300 publications on STRs used for DNA typing purposes.

INTRODUCTION

Tandemly repeated DNA sequences, which are widespread throughout the human genome, are polymorphic in nature, making them important genetic markers for mapping studies, disease diagnosis, and human identity testing (1). Short tandem repeats (STRs) contain repeat units that are 2–6 bp in length and can be readily amplified with the polymerase chain reaction (PCR). STRs have become popular in forensic laboratories because low amounts of DNA, even in a degraded form, can be successfully typed. Sample mixtures can be more readily resolved with STR results than with previously used DNA typing technologies (2).

In the United States, a core set of 13 STR markers are being used to generate a nationwide DNA database called the FBI Combined DNA Index System (CODIS). The CODIS Database and similar DNA databases around the world have been successful at linking DNA profiles from repeat offenders and

crime scene evidence. STR typing results are also used to aid hundreds of thousands of paternity testing cases each year.

Forensic DNA testing requires stringent guidelines for DNA sample processing and data analysis. New STR kits are validated by conducting a number of tests to verify that results are reliable and robust. A set of quality assurance standards issued by the DNA Advisory Board must be followed by forensic laboratories in order to submit STR profiles to the national CODIS database (3).

CONTENT OF STRBase

The primary content of STRBase is shown in Figure 1. The information is broken into three sections: general, forensic and supplemental information. An introductory PowerPoint presentation in the 'STRs 101' section explains STRs and their use in forensics to help familiarize people with this field.

General information


Information describing each commonly used STR marker forms the core of STRBase in a format we call STR fact sheets (Fig. 2). These fact sheets are composed of four sections: general information, PCR primers used, PCR product sizes and additional information. The general information section describes other names commonly used for the STR locus, its chromosomal location, the sequence of the core STR repeat unit, the GenBank accession number and the number of repeat units in the reference sequence. Underlined words shown in Figure 2 are hyperlinked to further information. The second section of these fact sheets, PCR primers used, lists the sequence for published primers, or the amplification kit(s) in which the primer set is available for commercial primers (sequences for primers in commercial kits have not been released at this time) (4). Each primer set is referenced to the published paper from our reference list or the commercial source of the kit. Each reference is assigned a number when it is entered into the database, and is referred to in the 'Ref.' column of Figure 2 (see Supplemental information). The PCR product section gives the length and sequence of an amplicon generated using each primer set listed in the PCR primer section for reported alleles. Each sequence is also referenced to a published paper in our reference listing. Links are provided to annotated GenBank sequence information for commonly used STR loci, which include both repeat and flanking regions

*To whom correspondence should be addressed. Tel: +1 301 975 4049; Fax: +1 301 975 8505; Email: john.butler@nist.gov

Present address:


Dennis J. Reeder, Human Identity Group, Applied Biosystems, Inc., 850 Lincoln Center Drive, Foster City, CA 94404, USA.

STRBase



Short Tandem Repeat DNA

Internet DataBase



General Information

- STRs101: Brief Introduction to STRs (downloadable PowerPoint presentation)
- STR Fact Sheets (observed alleles and PCR product sizes)
- Sequence Information (annotated)
- Multiplex STR Sets
- Non-published Variant Allele Reports

Forensic Interest Data

- FBI CODIS Core STR Loci
- DNA Advisory Board Quality Assurance Standards
- NIST Standard Reference Material for PCR-Based Testing
- Chromosomal Locations
- Mutation Rates for Common Loci
- Published PCR Primers
- Validation Studies
- Population Data
- Y-Chromosome STRs
- Sex-Typing Markers

Supplemental Information

- Reference List
- Technology for Resolving STR Alleles
- Addresses for Scientists Working with STRs
- Links to Other Web Sites

Figure 1. Overview of topics covered in STRBase.

with reported primer sequences. The additional information section lists commercial sources for allelic ladders, common multiplexes the locus is in and any available population studies of the locus.

Commercial multiplex STR kits have become widely used by laboratories worldwide because of their ease of use and high discriminatory power. Charts showing the allele product size ranges of commercially available multiplex STR kits, and hyperlinks to vendors are available on STRBase. Information on published multiplexes not commercially available is also provided. Examples from two kits are shown in Figure 3. The PowerPlex™ 16 kit simultaneously amplifies 15 different STR loci, including the two pentanucleotide repeat loci Penta D and Penta E, as well as the sex-typing marker amelogenin. A DNA ladder (ILS-600) labeled with the dye CXR (Promega Corporation, Madison, WI) is used as an internal sizing standard. The Profiler Plus™ kit amplifies 9 STRs and amelogenin, and uses ROX (Applied Biosystems, Foster City, CA) to label its internal sizing standard (GS-500).

Verification of microvariant alleles is becoming very important as STR typing expands, since they are being observed on a regular basis as more samples are studied. STRBase provides a publication venue by listing new alleles as they are found. We invite scientists to submit information on these new and rare alleles so they may be recognized.

Forensic interest information

The FBI has selected thirteen core loci for the CODIS database, which will be used for linking serial crimes and unsolved

D8S1179

General Information

Other Names D6S502, D8	Chromosomal Location 8q	GenBank Accession G08710; has 12 repeats
----------------------------------	--------------------------------------------	-------------------------------------------------------------

Repeat: [TCTA] = GenBank top strand

PCR Primer Information

Reported Primers	Ref.	PCR Primer Sequences
Set 1	369	5' - TTTTGTATTTCATGTGACATTCG - 3' 5' - CGTAGCTATAATTAGTTCATTTTCA - 3'
Set 2	PE ABI	AmpFISTR® Profiler Plus™
Set 3	Promega	GenePrint® PowerPlex™ 2.1, GenePrint® PowerPlex™ 16

PCR Product Sizes of Observed Alleles

Allele (Repeat #)	Set 1	Set 2	Set 3	Repeat Structure	Ref.
7	157 bp	123 bp	203 bp	[TCTA] ₇	716
8	161 bp	127 bp	207 bp	[TCTA] ₈	369
9	165 bp	131 bp	211 bp	[TCTA] ₉	369
10	169 bp	135 bp	215 bp	[TCTA] ₁₀	369
11	173 bp	139 bp	219 bp	[TCTA] ₁₁	369
12	177 bp	143 bp	223 bp	[TCTA] ₁₂	369
13	181 bp	147 bp	227 bp	[TCTA] ₁ [TCTG] ₁ [TCTA] ₁₁	369
14	185 bp	151 bp	231 bp	[TCTA] ₁ [TCTG] ₁ [TCTA] ₁₂	369
15	189 bp	155 bp	235 bp	[TCTA] ₁ [TCTG] ₁ [TCTA] ₁₃	369
16	193 bp	159 bp	239 bp	[TCTA] ₂ [TCTG] ₁ [TCTA] ₁₃	369
17	197 bp	163 bp	243 bp	[TCTA] ₂ [TCTG] ₂ [TCTA] ₁₃	369
18	201 bp	167 bp	247 bp	[TCTA] ₂ [TCTG] ₁ [TCTA] ₁₅	369
19	205 bp	171 bp	251 bp	[TCTA] ₂ [TCTG] ₂ [TCTA] ₁₅	716

Additional Information

Allelic Ladders: Commercially available from [Applied Biosystems](#), [Promega](#)

Common Multiplexes: Profiler Plus, PowerPlex 2.1, PowerPlex 16

Figure 2. An example STR fact sheet for the marker D8S1179.

cases with repeat offenders nationwide. A chromosomal map in STRBase shows the location of each core locus, with links to STR fact sheets, which are available for all thirteen CODIS core loci.

NIST ensures exact and compatible measurements by generating, certifying and issuing Standard Reference Materials (SRMs) to various laboratories. NIST SRM 2391a is designed for PCR-based testing, and contains genotypes for 21 STR loci and other common forensic DNA markers. A chromosome listing is included with information on these markers and restriction fragment length polymorphism (RFLP) based DNA markers.

STR systems and multiplexes routinely used in human identity testing have been extensively validated by many groups, as any new system or multiplex must be in the future. Summaries of published validation studies and the Scientific Working Group on DNA Analysis Methods' (SWGDM, formerly TWGDAM) guidelines for validation of PCR-based DNA typing markers are available on STRBase (5). Researchers can use these summaries to design their own validation studies and scientists using common STR systems can review published validation studies.

DNA typing laboratories spend a substantial amount of time and effort studying allele frequencies for various STR systems and populations. Over 750 published population studies are summarized in STRBase, with a listing of the population examined, the number of unrelated individuals tested and the published reference.

Y-chromosome STRs are raising interest in DNA forensics, as they may be valuable in sexual assault cases, paternity

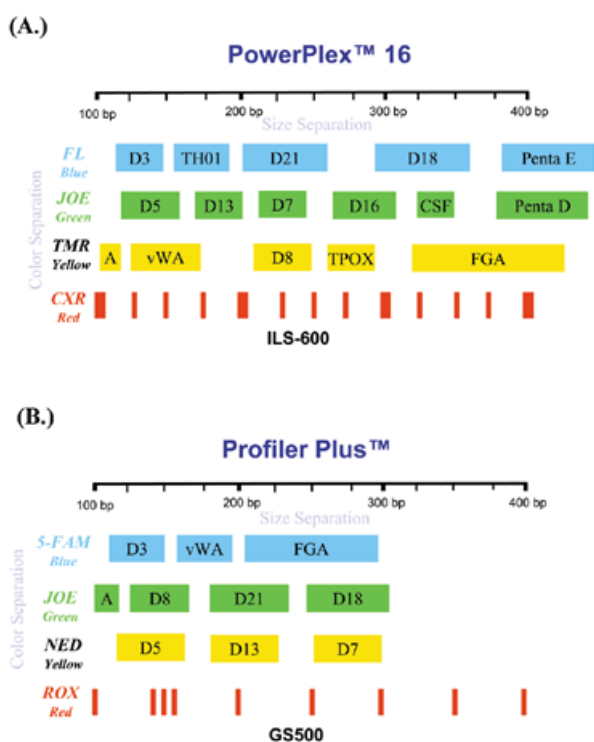


Figure 3. Schematic representation of the STR loci included in two commercially available multiplex amplification kits: (A) PowerPlex™ 16 kit from Promega Corporation and (B) Profiler Plus™ kit from Applied Biosystems. The size range for known alleles is represented by the length of the box containing the locus name. The vertical position of the box reflects the dye color used to label the amplification products. The two kits shown use three different fluorescent dyes that are spectrally distinguishable and colored blue, green and yellow for detection. Different internal sizing standards are available and labeled with a red-colored dye. Each locus box is hyperlinked to the appropriate STR fact sheet.

testing or other cases of mixed male/female DNA. STR fact sheets are available for several Y-chromosome STRs as well as relevant links, references and a PowerPoint presentation explaining the use of Y-chromosome STRs for forensic purposes.

Numerous PCR-based sex-typing assays have been reported in the literature. The most commonly used is amelogenin, which differentiates a 6 bp deletion on the X-chromosome from the Y-chromosome. Primer sequences, PCR product sizes and references for amelogenin and three other sex-typing markers are all listed in STRBase.

Supplemental information

Over 1300 references pertaining to STRs and their application to forensic DNA typing have been gathered from journals, conference proceedings, book chapters and other sources. They come from almost 120 sources, and over 800 are from peer-reviewed journals. The abundance of literature available on the use of STRs for forensic DNA typing shows that it has become an established technology worthy of being used as court evidence.

The rapid pace of developing technologies for DNA analysis can make it difficult to keep track of and understand all methodologies. STRBase includes a brief review of techniques that have been successfully implemented for resolving and detecting STR alleles. Relevant references and hyperlinks to groups working in each area are also provided.

STRBase has more than 60 hyperlinks to organizations involved in DNA typing, commercial sources of instrumentation or DNA testing kits, paternity testing laboratories, electronic journals where STR publications have been found and other useful DNA databases.

Addresses for scientists working with STR markers are listed in STRBase with email links, phone and fax numbers. All scientists working with STR markers are invited to add their information to aid in cooperation of DNA typing laboratories around the world. To have your name listed, send an email message with the above information to john.butler@nist.gov

STRBase ACCESS AND DATA ACQUISITION

The short tandem repeat DNA database is available throughout the world at <http://www.cstl.nist.gov/biotech/strbase/>. When using information from STRBase, please cite this paper and the date which the information was gathered from STRBase.

The information contained in STRBase is taken from published works on short tandem repeats used for DNA typing purposes. The literature is regularly searched for new publications and updates are periodically made. Comments on the database, suggestions for further improvements or submissions should be sent to the corresponding author, attn.: STRBase, or john.butler@nist.gov

NOTE

Certain commercial equipment, instruments or materials are identified in this report to adequately specify an available source of information. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

- Edwards, A., Civitello, A., Hammond, H.A. and Caskey, C.T. (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet.*, **49**, 746–756.
- Sparkes, R., Kimpton, C.P., Watson, S., Oldroyd, N.J., Clayton, T.M., Barnett, L., Arnold, J., Thompson, C., Hale, R., Chapman, J. *et al.* (1996) The validation of a 7-locus multiplex STR test for use in forensic casework: (I) Mixtures, ageing, degradation and species studies. *Int. J. Legal Med.*, **109**, 186–194.
- Reeder, D.J. (1999) Impact of DNA typing on standards and practice in the forensic community. *Arch. Pathol. Lab Med.*, **123**, 1063–1065.
- Smaglik, P. (2000) Legal protests prompt DNA primer release. *Nature*, **406**, 366.
- Technical Working Group on DNA Methods. (1995) Guidelines for a quality assurance program for DNA analysis. *Crime Laboratory Digest*, **22**, 18–43.