Journal: www.jogg.info
Originally Published: Volume 4, Number 2 (Fall 2008)
Reference Number: 42.007

ADDRESSING Y-CHROMOSOME SHORT TANDEM REPEAT (Y-STR) ALLELE NOMENCLATURE

Author(s): John M. Butler, Margaret C. Kline, and Amy E. Decker

Journal of Genetic Genealogy, 4(2):125-148, 2008

Addressing Y-Chromosome Short Tandem Repeat Allele Nomenclature

John M. Butler, Margaret C. Kline, and Amy E. Decker

Abstract

A total of about 120 different Y-chromosome short tandem repeat (Y-STR) markers are currently used by different genetic genealogy testing laboratories. In some cases, different laboratories may designate the same Y-STR allele with two different nomenclatures, making data comparison difficult and frustrating due to needed conversion factors. This article explains how STR allele nomenclatures are typically determined, where ambiguity may exist, and how measurement accuracy and consistency can be promoted through use of common reference materials across all genetic genealogy testing laboratories. Comparisons are made to forensic DNA testing and the benefits of a common set of loci and STR allele nomenclatures.

Introduction

Y-chromosome DNA testing is important for a number of different applications of human genetics (Butler, 2003) including forensic evidence examination (Butler, 2005, pp 201-239), paternity testing (Rolf et al., 2001), historical investigations (Foster et al., 1998), studying human migration patterns throughout history (Stix, 2008), and genealogical research (Brown, 2002).

The genetic markers (loci) most commonly used as part of Y-chromosome DNA analysis include short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs). Since Y-STRs change more rapidly (mutation rate $\approx$ 1 in 10^3 (Dupuy et al., 2004) compared to Y-SNPs (mutation rate $\approx$ 1 in 10^9 (Shen et al., 2000), Y-STR results are preferred for providing an assessment of genetic similarity or difference for potentially related people on a time-scale helpful in genealogical research.

Over the past decade as Y-chromosome testing has grown in popularity, different Y-STR markers have been selected for various uses and by marker availability. In the year 2000 when the field of genetic genealogy was born, there were only about 20 Y-STR markers known to exist on the Y-chromosome (Butler, 2003). Now, in large measure thanks to the efforts of the Human Genome Project (International Human Genome Sequencing Consortium 2004, Skaletsky et al., 2003), over 400 Y-STRs have been characterized on the human Y-chromosome (Redd et al., 2002; Kayser et al., 2004; Hanson and Ballantyne, 2006). However, not all of these Y-STRs are male-specific or sufficiently polymorphic to be helpful in forensic or genetic genealogy applications.

Address for correspondence: Amy Decker, amy.decker@nist.gov. The authors are with the U.S. National Institute of Standards and Technology, Human Identity Project Team.

Received: October 3, 2008; accepted: October 30, 2008.

The various companies providing Y-STR results to the genetic genealogy community currently use about 120 different loci—many of which overlap between test providers—as noted in Table 1. While it would perhaps be convenient for data comparison purposes to have everyone in the genetic genealogy community using the same Y-STR markers, this ideal situation will probably never exist in a consumer-driven, unregulated environment where additional testing information is constantly desired.

A bigger problem for the genetic genealogy community is that different DNA test providers may have different nomenclatures for calling the same Y-STR allele. It is important for users of these DNA test results to appreciate that these differences arise in how a STR repeat sequence is denoted by the laboratory and not because of some measurement mistake. For example, a DNA sequence containing "AGATAGATAGAT" could be considered to have three "AGAT" repeats or two "GATA" repeats depending on how the core repeat unit is designated. Thus, Y-STR results, which are only described as the number of repeats present, may not be fully comparable when the same DNA sample is tested by multiple laboratories. Without appreciating why a conversion factor is needed between specific Y-STR laboratory results, genetic genealogists may come away confused or frustrated when trying to compare their results with others. The purpose of this article is to explain how STR allele nomenclatures are determined, where ambiguity may exist, and how measurement accuracy and consistency can be promoted through use of common reference materials across all genetic genealogy testing laboratories.

Comparison to the Forensic DNA Testing Community

Before we begin a discussion of STR allele nomenclature, which we approach from the perspective of working with the forensic DNA testing community for almost two decades, it is worth discussing measurement quality

Table 1
A Total of 120 Y-STRs Were in Use in August, 2008 by Genetic Genealogy Test Providers. Loci in bold italic font are included in NIST SRM 2395.

Family Tree DNA (12, 25, 37, 67 markers)		DNA Ancestry (33 or 46), DNA Heritage (23 or 43), SMGF	Ethnoancestry (18, 27, 45 markers)	Oxford Ancestors (10)	Genebase		
DYS19	DYS490	***DYS19***	***DYS19***	***DYS19***	***DYS19***	DYS487	DYS644
DYS385a/b	***DYS492***	***DYS385a/b***	***DYS385 a/b***	***DYS385 a/b***	***DYS385a/b***	DYS490	***DYS710***
DYS388	DYS495	***DYS388***	***DYS388***	***DYS388***	***DYS388***	***DYS492***	DYS711
DYS389I	DYS511	***DYS389I***	***DYS389I***	***DYS389I***	***DYS389I***	DYS494	***DYS712***
DYS389II	DYS520	***DYS389II***	***DYS389II***	***DYS389II***	***DYS389II***	DYS495	DYS713
DYS390	DYS531	***DYS390***	***DYS390***	***DYS390***	***DYS390***	DYS504	DYS714
DYS391	***DYS534***	***DYS391***	***DYS391***	***DYS391***	***DYS391***	DYS505	DYS716
DYS392	DYS537	***DYS392***	***DYS392***	***DYS392***	***DYS392***	DYS508	***DYS717***
DYS393	DYS557	***DYS393***	***DYS393***	***DYS393***	***DYS393***	DYS511	DYS724a/b
DYS413	DYS565	***DYS426***	DYS425	DYS425	DYS413a/b	DYS518	Y-GATA-A10
DYS425	DYS568	***DYS437***	***DYS426***	***DYS426***	***DYS426***	DYS520	***Y-GATA-H4***
DYS426	***DYS570***	***DYS438***	DYS434	***DYS437***	DYS434	***DYS522***	YCAII a/b
DYS434	***DYS572***	***DYS439***	***DYS435***	***DYS438***	***DYS435***	DYS525	
DYS435	***DYS576***	DYS441	***DYS436***	***DYS439***	***DYS436***	***DYS527a/b***	
DYS436	DYS578	DYS442	***DYS437***		***DYS437/ DYS457***	DYS531	
DYS437	DYS590	DYS444	***DYS438***		***DYS438***	***DYS532***	
DYS438	DYS594	DYS445	***DYS439***		***DYS439***	DYS533	
DYS439	***DYS607***	DYS446	***DYS449***		DYS441	***DYS534***	
DYS441	DYS617	***DYS447***	***DYS458***		DYS442	DYS537	
DYS442	***DYS635***	***DYS448***	***DYS460***		DYS444	DYS540	
DYS444	DYS640	***DYS449***	***DYS461***		DYS445	DYS549	
DYS445	DYS641	DYS452	DYS462		DYS446	DYS556	
DYS446	DYS643	DYS454	YCAII a/b		***DYS447***	DYS557	
DYS447	***DYS710***	DYS455	***DYS635***		***DYS448***	DYS565	
DYS448	DYS714	***DYS456***	***Y-GATA-H4***		***DYS449***	DYS568	
DYS449	DYS716	***DYS458***	DYS555		DYS450	***DYS570***	
DYS450	***DYS717***	DYS459 a/b	DYS481		DYS452	***DYS572***	
DYS452	DYS724	***DYS460***	DYS487		DYS453	DYS575	
DYS454	DYS725	***DYS461***	DYS490		DYS454	***DYS576***	
DYS455	DYS726	DYS462	DYS494		DYS455	DYS578	
DYS456	YCAIIa/b	DYS463	DYS505		***DYS456***	DYS588	
DYS458	GATA-A10	DYS464 a/b/c/d	***DYS522***		***DYS458***	DYS590	
DYS459 a/b	***GATA-H4***	***DYS635***	DYS531		DYS459 a/b	DYS594	
DYS460	GGAAT-1B07	YCAIIa/b	DYS533		***DYS460***	DYS607	
DYS461	DYF371	Y-GATA-A10	DYS594		***DYS461***	DYS612	
DYS462	DYF385	***Y-GATA-H4***	DYS556		DYS462	DYS614	
DYS463	DYF395	Y-GGAAT-1B07	DYS575		DYS463	DYS617	
DYS464 a/b/c/d	DYF397		DYS578		DYS464 a/b/c/d	DYS626	
DYS472	DYF399		DYS589		DYS468	DYS632	
DYS481	DYF401		DYS549		DYS472	***DYS635***	
DYS485	DYF406S1		DYS636		***DYS481***	DYS640	
DYS487	DYF408		DYS638		DYS484	DYS641	
	DYF411		DYS641		DYS485	DYS643	

assurance and controls used in forensic DNA analysis to help produce accurate results. Quality results are paramount when processing biological evidence from crime scenes and reporting those results in court because a suspect's liberty is at stake. Forensic DNA laboratories in the United States are mandated by Congress to follow strict quality assurance standards (see Butler, 2005, pp. 389-412). In October 1998, the FBI Laboratory's DNA Advisory Board issued Quality Assurance Standards that define how forensic laboratories are required to conduct business (Butler, 2005, pp. 593-611). These Quality Assurance Standards (QAS) were recently revised and will go into effect in July 2009 (CODIS Quality Assurance, 2008). Thus, the forensic DNA community is governed by formal quality assurance standards and individual laboratories are regularly audited for their compliance to these standards.

In order to be able to compare results between the almost 200 public and private forensic DNA laboratories in the United States, a common set of core STR markers are used to enable a common currency of data exchange and DNA database compatibility (Budowle et al., 1998). The U.S. core 13 autosomal STR loci enable the Combined DNA Index System (CODIS) to operate and many other countries have adopted these 13 core STRs in their entirety or as subsets with some additional STR loci (Butler, 2006).

Commercially available STR typing kits are used by all forensic laboratories to maintain a high level of quality assurance in results and to ensure consistency in nomenclature between laboratories. Use of commercial kits does increase the cost of DNA testing but aids in overall quality assurance due to compatibility and consistency of results (both in terms of loci examined and STR allele nomenclature used). These commercial kits come with company-supplied allelic ladders, which are composed of common alleles and used in sample data interpretation to make the specific STR allele designations. While slight differences may exist in alleles present between the various kit allelic ladders as well as the polymerase chain reaction (PCR) primers used to target the STR locus, concordance studies have shown that equivalent results may be obtained (Budowle et al., 2001, Gross et al., 2006).

The current commercially available Y-STR kits, which examine only a modest number of loci (SWGDAM, 2004) compared to what is now available with routine genetic genealogy work, include PowerPlex Y (Promega Corporation, Madison, WI) and Yfiler (Applied Biosystems, Foster City, CA). PowerPlex Y examines 12 Y-STRs (Krenke et al., 2005): DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, and DYS385 a/b. Y-filer types 17 Y-STRs (Mulero et al., 2006a): DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, GATA-H4, and DYS385 a/b.

Another layer of quality assurance is provided by a required calibration of STR allele designations to Standard Reference Materials (SRMs) available from the National Institute of Standards and Technology (NIST). QAS Standard 9.5 states: "The laboratory shall check its DNA procedures annually or whenever substantial changes are made to the protocol(s) against an appropriate and available NIST standard reference material or standard traceable to a NIST standard" (Butler, 2005, p. 606). This external calibration helps ensure consistent performance and STR allele designation of commercial allelic ladders and genotyping software programs. Companies also use the NIST reference materials to ensure consistent and accurate allele calls prior to release of their commercial STR typing kits.

The genetic genealogy community does not have the same level of oversight as forensic laboratories—nor does it have the same need since genealogy results are more for satisfying a curiosity than a court mandated test that could impact someone's liberty. In order to keep operating costs lower, genetic genealogy testing laboratories typically use assays developed in-house and unique combinations of genetic markers, rather than commercially available Y-STR typing kits. In addition, the PCR primer sequences and reaction conditions for these Y-STR assays may be considered proprietary to the laboratories.

The preferred measurement technique in genetic genealogy testing laboratories is PCR product sizing (with an internal size standard for electrophoretic calibration) relative to a few control samples that have usually been sequenced (Butler, 2003). For example, a sequenced control sample for DYS391 containing 10 repeats might produce a PCR product size of 160.23 bp with a specific multiplex assay, and thus a test sample with a PCR product size of 164.35 bp would be designated as having 11 repeats since it is 4 bp larger (and therefore one tetranucleotide repeat unit beyond the 10 repeat reference allele). Note that while PCR products are necessarily integers (e.g., 160 or 161 base pairs) their measurement against an electrophoretic internal size standard results in sizes that are fractions of integers, such as 160.23 bp, when calculated by the genotyping software.

This essentially single-point calibration approach can work very well and generate consistent results within a single laboratory. However, in-house produced control samples are typically available only to the specific testing laboratory, and thus STR allele nomenclatures decided upon by an individual laboratory are not vetted by other laboratories or independent groups. As will be seen in

specific examples below, this divergence in nomenclature opinion has given rise to various ways to describe identical STR allele sequences. In addition, interlaboratory studies have shown that comparing STR typing results between laboratories is best accomplished with common reference materials and methods (Kline et al., 1997).

STR Allele Nomenclature

When reporting the results from an STR allele, the goal of a testing laboratory is to accurately reflect the number of repeat units that exist in the tested DNA sequence. However, different approaches to counting the number of STR repeat units present can result in a different outcome for the same DNA sequence. To aid with inter-laboratory reproducibility and comparison of STR data—especially with DNA databases, a common nomenclature scheme has been developed in the forensic DNA community. The potential for STR allele nomenclature differences has been recognized as an issue for many years and efforts have been made to formalize allele nomenclature rules. The recognized leader in this area has been the International Society for Forensic Genetics (ISFG).[1]

The ISFG DNA Commission Recommendations

The ISFG, which was founded in 1968 and formerly known as the International Society of Forensic Haemogenetics (ISFH), today represents a group of approximately 1100 scientists from more than 60 countries. Meetings are held biannually to discuss the latest topics in forensic genetics. Every few years, as a specific need arises, a DNA Commission of the ISFG is formed and makes recommendations on the use of genetic markers. Publications from these meetings are available[2] and include the following topics (with their publication year):

- DNA polymorphisms (1989)
- PCR based polymorphisms (1992)
- **Naming variant alleles (1994)**
- **Repeat nomenclature (1997)**
- Mitochondrial DNA (2000)
- **Y-STR use in forensic analysis (2001)**
- **Additional Y-STRs - nomenclature (2006)**
- Mixture interpretation (2006)
- Disaster victim identification (2007)
- Biostatistics for paternity testing (2008)

The four sets of DNA Commission recommendations most pertinent to this discussion on Y-STR allele nomenclature were those published in 1994, 1997, 2001, and 2006, and are shown in bold font.

The 1994 ISFG DNA Commission publication addressed designations of alleles containing partial repeat sequences: "When an allele does not conform to the standard repeat motif of the system in question it should be designated by the number of complete repeat units and the number of base pairs of the partial repeat. These two values should be separated by a decimal point" (Bär et al., 1994). For example, an allele with $[AATG]_5ATG[AATG]_4$ is designated as a "9.3" since it contains nine full AATG repeats plus three additional nucleotides. Thus, tetranucleotide repeats (i.e., those containing four nucleotides in the repeat motif) could have x.1, x.2, and x.3 variant alleles that exhibit one, two, or three additional nucleotides beyond the number of complete repeat units found in the allele.

An STR repeat sequence is named by the structure (base composition) of the core repeat unit and the number of repeat units. However, because DNA has two strands, either of which may be used to designate the repeat unit for a particular STR marker, more than one choice is available and confusion can arise without a standard format. The 1997 ISFG DNA Commission recommendations describe how to best handle the choice of the DNA strand and the repeat motif and allele designation (Bär et al., 1997):

Choice of the Strand

- For STRs within protein coding regions (as well as in the intron of the genes), the coding strand should be used.
- For repetitive sequences without any connection to protein coding genes like many of the D#S### loci, the sequence originally described in the literature of the first public database entry shall become the standard reference (and strand) for nomenclature.
- If the nomenclature is already established in the forensic field but not in accordance with the aforementioned guideline, the established nomenclature shall be maintained to avoid unnecessary confusion.

Choice of the Motif and Allele Designation

- **The repeat sequence motif should be defined so that the first 5'-nucleotides that can define a repeat motif are used** (when reading from the 5' end). For example, 5'-GG TCA TCA TCA TGG-3' could be seen as having 3 x TCA repeats or 3 x CAT repeats. However, under the recommendations of the ISFG committee only

1 See the ISFG web site: http://www.isfg.org.

2 See: http://www.isfg.org/Publications/DNA+Commission.

the first one (3 x TCA) is correct because it defines the first possible repeat motif.

- Designation of incomplete repeat motifs should include the number of complete repeats and, separated by a decimal point, the number of base pairs in the incomplete repeat.

- For some highly variable systems, the repetitive structure can be very complex and the definition of a consensus repeat structure can be difficult. In such cases, alleles should be identified according to their size in bp, by comparison with a sequenced [allelic] ladder.

This article further notes: "For those situations where two or more nomenclatures already exist, priority should be given to the nomenclature that more closely adheres to the [1997 ISFG] guidelines. If this is not possible, priority shall be given to the nomenclature that was documented first" (Bär et al., 1997).

Figure 1 illustrates the application of these recommendations with a hypothetical STR sequence. In the upper portion (Figure 1A), the complementary top and bottom strands of a DNA sequence are shown. A few flanking nucleotides are included around the six AGAT repeats shown in bold font. PCR primers illustrated with the arrows anneal to the stable flanking region sequences and enable the specific STR repeat region to be copied from genomic DNA. Note that if the bottom strand was used instead of the top strand, then the repeat motif (read from the 5'-to-3' direction) would be ATCT. In either case, there would be six repeats. However, as illustrated in Figure 1B, if the repeat motif designation is not all the way to the 5' end but instead was called a GATA, ATAG, or TAGA repeat, then there would be one less repeat unit (5 rather than the 6 AGAT repeats) for this particular STR allele.

For Y-STRs there have been two ISFG DNA Commissions addressing confusion with Y-STR allele nomenclature. The 2001 ISFG DNA Commission noted that "the nomenclature of some loci has been based on the total number of repetitive units (non-variant plus variant; e.g., DYS19) whilst others have taken into account only the repetitive stretches of DNA that are variant (e.g., DYS391)" (Gill et al., 2001). This article continues, "If

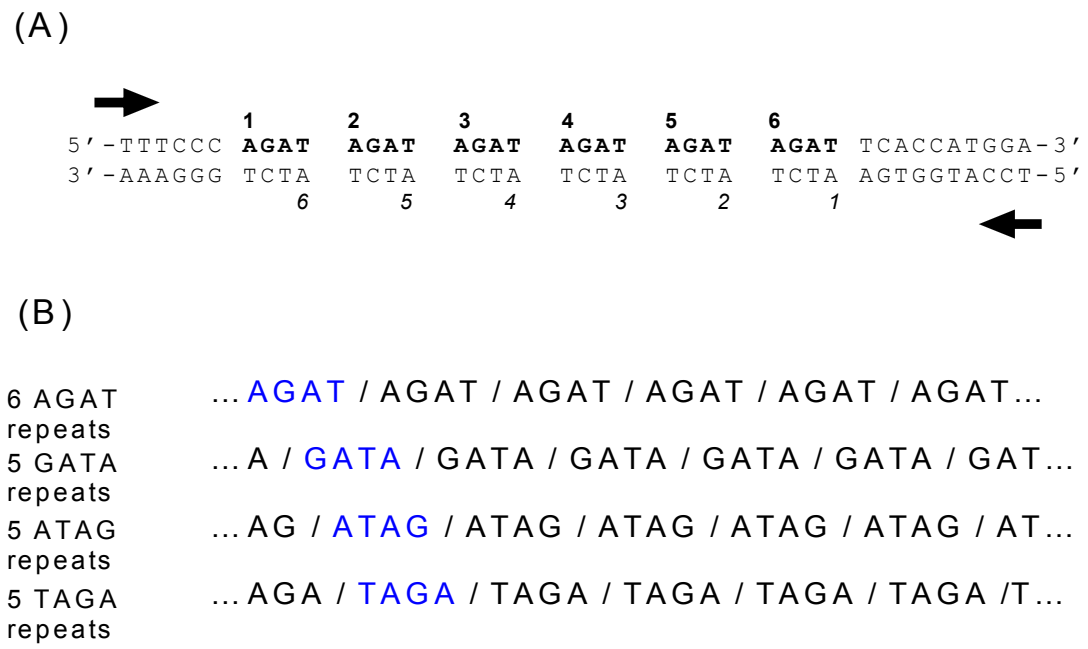


Figure 1. Illustration of different repeat nomenclatures from the same STR sequence and potential impact on overall allele nomenclature. (A) Both strands of DNA sequence shown with top strand motif AGAT occurring 6 times and bottom strand motif of ATCT (5'-to-3') also occurring 6 times. Arrows indicate potential PCR primer positions. (B) Starting the repeat motif in different places can lead to 5 versus 6 full repeats.

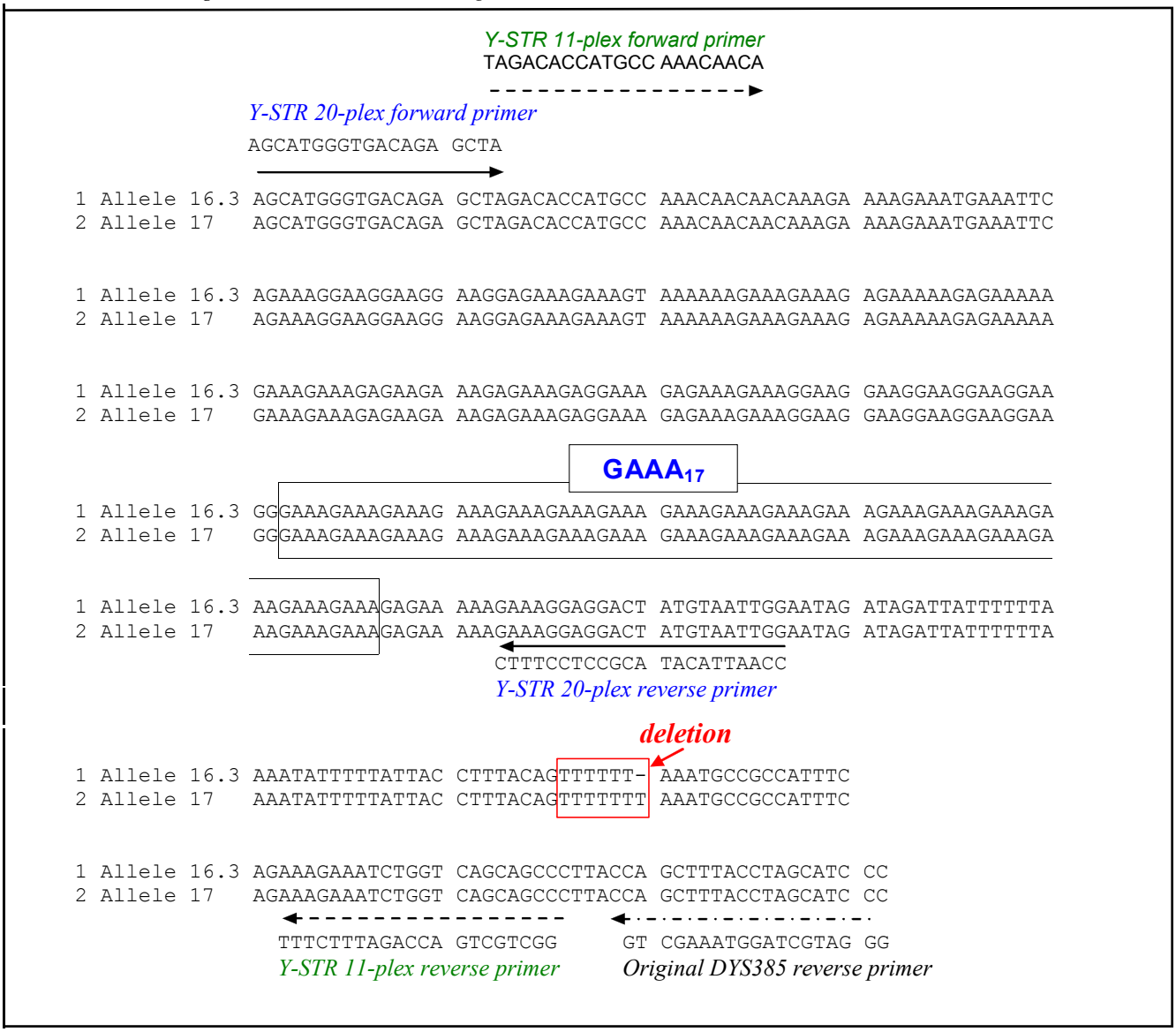


Figure 2. Alignment of top strands from two different male samples for the multi-copy DYS385 locus. Both samples contain 17 GAAA tetranucleotide repeats yet one is typed as a "17" allele and the other as a "16.3" with primers that encompass a polyT stretch (boxed region) 74-80 bases downstream of the repeat region (see Furedi et al., 1999). DYS385 primers internal to the poly-T stretch (e.g., 20-plex primers, Butler et al., 2002) will not be able to measure the deletion that gives rise to this microvariant allele. Modified from original figure published previously by our group (Schoske et al.,. 2004).

a nomenclature is already in use, it is recommended that it should be continued. However, to encourage consistency for newly reported STRs, it is recommended that alleles should be named according to the total number of the repeat units of the DNA that comprises both variant and non-variant repeats." Furthermore, the 2001 DNA Commission recognizes that "For very complex STRs . . . that comprise multiple repeats of different sizes, the designation of alleles is not as easy . . . . In this case, provided that the nomenclature follows ISFG guidelines, the default standard nomenclature will follow from the first publication or the first public database entry" (Gill et al., 2001).

As noted in the 2001 ISFG guidelines, another complication that can arise with some Y-STR loci is that "intermediate alleles can appear due to a single base insertion or deletion in the flanking region" (Gill et al., 2001). Different PCR primers, depending on whether or not they encompass the flanking region variation, can therefore give rise to different results from the same allele (Schoske et al., 2004; Gusmão et al., 2006). Figure

2 shows how a PCR primer amplifying shorter fragments (e.g., the Y-STR 20-plex reverse primer) can be inside the DYS385 flanking region deletion compared to other primers (e.g., the Y-STR 11-plex reverse primer or the original DYS385 reverse primer). In this example, if two DNA test providers used different DYS385 primers to examine the "16.3" allele, one might return a "17" allele call while the other could denote the allele a "16.3". Likewise, if both testing laboratories used the primer pair creating the smaller PCR product, they would be unable to distinguish a true "17" from a true "16.3" allele at DYS385. Note that if following the 2006 ISFG DNA Commission recommendations, this "16.3" allele would have a different designation (see below).

The most comprehensive examination of Y-STR allele nomenclature came with the 2006 ISFG DNA Commission recommendations (Gusmão et al., 2006). This article reviews the historical nomenclature for 11 core Y-STRs widely used in the forensic DNA community: DYS19 (DYS394), DYS385 a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393 (DYS395), DYS438, and DYS439 (GATA-A4). While some of the widely used nomenclatures for these 11 Y-STRs are not ideal, the 2006 ISFG DNA Commission encouraged their continued use because this information is in well-known databases and widely-used commercial kits: "To avoid further confusion due to nomenclature changes, the nomenclature of widely used Y-STRs should not be altered, even if the present guidelines are not followed" (Gusmão et al., 2006). The nomenclatures for 63 additional loci, that were known and characterized at the time, were also covered in this article. However, as can be seen by comparing the information in the 2006 ISFG DNA Commission article to that found in Table 1, genetic genealogy test providers have gone beyond these previously defined loci in an effort to capture greater variation along the Y-chromosome.

In developing STR allele nomenclatures, it is helpful to have information from multiple alleles instead of just a single reference sequence in order to make decisions regarding the total number of repeats that are varying between individuals. The 2006 ISFG DNA Commission recommended that, if possible, Y-STR alleles be sequenced from multiple individuals coming from different Y-SNP-defined haplogroups in order to increase the genetic distance between the sequences. Ideally chimpanzee alleles for these Y-STR alleles should be studied as well in order to determine which portions of an STR repeat region are varying over a large genetic distance (Gusmão et al., 2002).

The eight nomenclature recommendations of the 2006 ISFG DNA Commission are summarized below:

1) Alleles should be named according to the total number of contiguous variant and non-variant repeats determined from sequence data. Single repeat units located adjacent to the main repeat array and consisting of the same sequence as the main variable repeat should be considered as part of the repeat motif. For example, a hypothetical STR allele with the sequence …$(GATA)_n(GACA)_2(GATA)$… should be considered to have *n*+2+1 repeats.

2) Repetitive motifs that are not adjacent to the variable stretch and have three or less units and show no size variation within humans or between humans and chimpanzees should not be included in the allele nomenclature. For example, a hypothetical STR with the sequence …$(GATA)_n(GACA)_2N_8(GATA)_3$…, where N contains eight nucleotides that are not part of the repeat motif, should be called *n*+2, which excludes the non-adjacent $(GATA)_3$ repetitive stretch from the allele nomenclature. If the number of interrupting nucleotides in (N) is similar to or less than the number of nucleotides in the repeat motif, then the region is considered as one repeat unit with a length corresponding to the total number of nucleotides. Thus, …$(GATA)_n(GACA)_2N_4(GATA)_3$… is considered as one complex locus with *n*+2+1+3 units, while …$(GATA)_n(GACA)_2N_5(GATA)_3$… is considered to be two loci with *n*+2 and 3 units, respectively, of which *n*+2 would be included in the primary STR allele nomenclature.

3) Intermediate alleles (e.g., 11.1) fall into two classes: an insertion/deletion either (a) within the repeat motif or (b) in the flanking region encompassed by the PCR primer positions. If the partial repeat is found within the repeat motif, such as …$(GATA)_nT(GATA)_m$, alleles should be called as noted in the 1994 ISFG recommendations: ". . . by the number of complete repeat units and the number of base pairs of the partial repeat separated by a decimal point" (Bär et al., 1994).

4) Intermediate alleles arising due to mutations in the flanking sequences that alter the length or electrophoretic migration of a PCR product should be designated by additional information indicated after the number of complete STR repeat units. For example, an allele with 11 repeats and a T insertion at nucleotides 40 upstream from the repeat is not named "11.1" but rather "11(U40Tins)" where 11 stands for the number of complete repeats, U40 indicates the direction and position of the mutation relative to the STR repeat block (i.e., the mutation is

located 40 bases upstream of the repeat), and "Tins" indicates that a T nucleotide has been inserted. If the exact position of the deletion or insertion cannot be determined because it is part of a homopolymeric tract (i.e., a stretch of the same nucleotides such as TTTTT), then the deletion or insertion should be assigned to the highest numbered end of the homopolymeric stretch. Using **Figure 2** as an example, the deletion that gives rise to the "16.3" allele should more appropriately be referred to as a "17D80Tdel" allele since the single T deletion occurs at the end of a polymeric T stretch that is 80 nucleotides downstream of the repeat region.

5) Point mutations in a PCR primer binding region may prevent sufficient annealing of this primer and result in a "null" or "silent" allele due to failure to generate a detectable amount of PCR product (see Butler, 2005, pp. 133-138 for more information). It is recommended that point mutations which impact primer annealing be verified by DNA sequence analysis and published using a designation as in recommendation #4. For example, DYS438 (D7A→C) would indicate that the "A" nucleotide 7 bases downstream of the DYS438 repeat has changed into a "C" nucleotide in the tested STR allele.

6) If no additional sequence variation is found in the 166 Y-STR markers described by Kayser et al. (2004), then these authors' locus delimitation criteria should be adopted.

7) Journal editors, reviewers, and organizers of quality assurance schemes should focus on the use of standardized nomenclatures in order to obtain uniformity and avoid the spread of confusing nomenclatures.

8) Commercial Y-STR kits should follow the nomenclature recommendations so that direct comparisons between results obtained with different kits are possible.

While these guidelines provide a framework for STR allele nomenclature designation, they do not capture every possible permutation that exists, particularly with complex repeats. Following recommendations #1 and #2 described above, we have devised what we term the "one-change-rule" in that a single change to the repeat motif can be allowed in deciding what to include or not in an STR repeat block. However, when the single change in the repeat motif creates an adjacent homopolymeric stretch, we have decided not to include it in the repeat count. For example, with the repeat motif of CTT, if an adjacent sequence of TTT occurs (e.g., DYS481), then we only count the CTT. On the other hand, with a repeat structure of $(GATA)_n(GACA)$, our repeat count would be n+1.

It is challenging to designate the allele nomenclature for a particular STR marker definitively without extensive sequence characterization and analysis of population variation. It is worth noting that not all loci will be equally well characterized when they are initially used—particularly in the genetic genealogy community where the barrier to adding new Y-STR markers is not as high as in forensic casework. Unfortunately, not every variant allele that has been detected in forensic or genetic genealogy applications has been sequenced and thus the specific nature of intermediate alleles cannot easily be distinguished between recommendations #3 and #4. Thus, in most Y-STR databases today, it is more common to have variant alleles listed according to recommendation #3 (e.g., as *x*.3 allele) rather than according to recommendation #4 as the exact reason for the variant (e.g., *x*(D80Tdel)).

Use of Common Reference Materials

One of the primary ways to support a consistent and calibrated STR allele nomenclature is to use common reference materials between DNA testing laboratories. The National Institute of Standards and Technology (NIST; see http://www.nist.gov), which is part of the U.S. Department of Commerce, provides reference materials for a variety of fields to enable accurate and compatible measurements. NIST supplies over 1300 reference materials to industry, academia, and government laboratories to facilitate quality assurance and support measurement traceability. These Standard Reference Materials (SRMs) are certified through carefully characterizing the properties of supplied components.

In July 2003, NIST released SRM 2395, Human Y-Chromosome DNA Profiling Standard, for use in the standardization of forensic and paternity quality assurance procedures involving Y-STR testing (at the initial time of its release, the PowerPlex Y and Yfiler kits, now commonly used by the forensic community, were in development and not yet available). SRM 2395 includes six components: five male genomic DNA extracts designated as components A-E and one female genomic extract labeled component F. The female DNA sample will, of course, not work with male-specific Y-STR assays and can thus serve as a negative control. The five male DNA samples were originally characterized through DNA sequencing of 22 Y-STR loci and typing an additional 9 Y-STR loci along with 42 Y-SNPs. The sequencing and typing results for these Y-STRs and Y-SNPs are described in the SRM 2395 Certificate of Analysis (SRM 2395, 2008).

The components of SRM 2395 were chosen due to their genetic diversity to represent alleles present in the three

largest U.S. ethnic groups: components A, B and F are from anonymous Caucasian individuals, components C and D are African American in origin, and component E is Hispanic in origin. The original samples were purchased from a commercial blood bank and screened for variation across commonly used Y-STR loci. The five male components in SRM 2395 have five different Y-SNP backgrounds: R-M207, J2-M172, E3a-M2, G-M201, and I-M170 (SRM 2395, 2008).

In September 2008, an update was made to the Certificate of Analysis providing additional information to the already available DNA samples (see also Kline et al., 2006). The revised certificate now has certified and reference values for 41 Y-STR markers that have been confirmed through DNA sequencing performed at NIST. In addition, informational values (without sequence characterization) are available for DYS450, DYS464 a/b/c/d, and YCAII a/b along with the 42 Y-SNP values obtained through use of the Marligen Biosystem's Signet Y-SNP Identification System assay.

There are three levels of confidence in characterized values provided with a NIST SRM: certified, reference, and informational (May et al., 2000). A certified value indicates the highest confidence in the accuracy of the value provided because all known sources of bias have been investigated. Certified values have generally been characterized by two or more independent means. In the case of certified Y-STR values, the individual allele has been sequenced and PCR product sizes determined and genotyped. To be an SRM certified value, the measurement must be run at NIST. However, the nominal values for candidate materials can be corroborated by interlaboratory comparisons involving independent typing and/or sequence analysis. Reference and informational values, which may be defined by only a single method, can be of interest and use, but there is insufficient information available to fully assess uncertainty in the measurement. For SRM 2395 components, reference values have been assigned when sequencing has not been performed on every allele although multiple alleles within the same locus have been sequenced to anchor the base pair genotyping data. Information values have been assigned when fewer alleles of the locus have been sequenced, and thus there is less confidence associated with the allele call.

Certified Y-STR allele designations added to the Certificate of Analysis for SRM 2395 were confirmed using two independent methods, which included PCR product size analysis (relative to sequenced control alleles) and direct DNA sequence analysis of each allele. Size analysis and genotyping includes the electrophoretic separation and sizing of the PCR product compared to an internal size standard followed by a comparison to the sizes of one or more sequenced alleles, such as might be present in a commercially available allelic ladder. The

Table 2

Summarized DYS715 Results from Analysis of NIST SRM 2395 Components that Compare PCR Product Sizes to DNA Sequence Information. Note that PCR product size and initial allele assignments do not match for components C, D, and E (shown in *bold italics* font). The addition of a second repeat block (TGGA), located 20 nucleotides downstream, to the final repeat motif allows the PCR product size and allele assignment to correctly correspond (shown in bold font).

SRM 2395 Component	Initial Repeat Motif and Allele Assignment	Size (bp)	Final Repeat Motif and Allele Assignment
A	$[TAGA]_{14}$ - 14	195.4	$[TAGA]_{14}$ N_{20} $[TGGA]_{10}$ - **24**
B	$[TAGA]_{11}$ - 11	183.5	$[TAGA]_{11}$ N_{20} $[TGGA]_{10}$ - **21**
C	$[TAGA]_{12}$ - ***12***	187.4	$[TAGA]_{12}$ N_{20} $[TGGA]_{10}$ - **22**
D	$[TAGA]_{13}$ - ***13***	191.3	$[TAGA]_{13}$ N_{20} $[TGGA]_{10}$ - **23**
E	$[TAGA]_{12}$ - ***12***	191.4	$[TAGA]_{12}$ N_{20} $[TGGA]_{11}$ - **23**

tested samples are run in-house with the same conditions, instrument and internal size standard. DNA sequence analysis involves the isolation of each individual allele and sequence analysis in order to directly count the number of repeat units. Finally, the repeat designation is correlated to the size variation observed during PCR product analysis.

To illustrate the importance of correlating PCR product size information with DNA sequence, consider our characterization of allele nomenclature for the new Y-STR marker DYS715. As noted in Table 2, initial characterization of the primary TAGA repeat motif found SRM 2395 components D and E with similar sizes (191.3 bp and 191.4 bp) but different numbers of repeats (13 TAGA vs 12 TAGA). Component C also had 12 TAGA repeats but sized at 187.4 bp. This example is evidence that a more complex repeat nomenclature is necessary for the PCR product size and DNA sequencing results to agree. Upon closer examination of the full DNA sequence for each DYS715 allele (Table 2, right column), a second TGGA repeat motif was observed 20 bp downstream of the first repeat. Component D contains 10 TGGA repeats whereas component E contains 11 repeats. Thus, both repeat blocks are variable similar to DYS449 (Redd et al., 2002). When this second repeat block is included in the overall allele nomenclature, the allele types for components D and E both become "23" (13+10 and 12+11) so that the overall allele nomenclature matches with the observed PCR product sizes for DYS715. This example illustrates the importance of having DNA sequencing information on each allele in order to fully certify STR allele designations particularly for loci where internal sequence variability is possible.

Generally speaking STR markers can be classified into several categories based on their repeat pattern as previously described by Urquhart et al. (1994) as shown in Table 3. *Simple repeats* contain units of identical length and sequence, *compound repeats* comprise two or more adjacent simple repeats (typically with a single nucleotide difference between the repeat motifs), and *complex repeats* may contain several repeat blocks of variable unit length as well as variable intervening sequences. As has been noted previously, not all alleles for an STR locus may contain complete repeat units. Some simple repeats may possess non-consensus or variant alleles (e.g., 9.3). In Table 3, we list another category of repeats containing a non-variable non-repetitive region. The DYS715 example shown in Table 2 falls into this category. Example Y-STR markers, based on the newly characterized loci added to the NIST SRM 2395 Certificate of Analysis in September 2008, are separated into the various categories in Table 3.

Table 3
Categories of STR Markers Based on Repeat Patterns Originally Described by Urquhart, et al. (1994)

Category	Example Repeat Structure	Example Y-STR Markers (from recent SRM 2395 additions)
simple repeats	(GATA)(GATA)(GATA)	DYS456, DYS458, DYS481, DYS492, DYS522, DYS532, DYS534, DYS570, DYS572, DYS576
simple repeats with non-consensus alleles	(GATA)(GAT-)(GATA)	DYS712
Compound repeats	(GATA)(GACA)(GATA)	DYS527, DYS607, DYS635, DYS650, DYS652, DYS717
complex repeats	(GATA)(GACA)(CA)(CATA)	DYS710
repeats containing non-variable non-repetitive region	(GATA)N_n(GATA)	DYS449, DYS715

Journal of Genetic Genealogy, 4(2):125-148, 2008

Table 4
Y-STR allele Nomenclature Conversions Relative to NIST Recommendations as of Early August, 2008 for 10 Markers in Order to Obtain Equivalent Results from the Various Genetic Genealogy Test Providers (Listed Anonymously as Providers A-H). Figures 3-11 illustrate likely reasons for these nomenclature differences with NIST recommendations for nomenclature standardization. "=" means equivalent nomenclature; "NT" means not tested; "?" means unable to determine difference relative to NIST recommendations.

	Genetic Genealogy Test Providers Result Conversions Needed							
Marker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
DYS389I	=	=	=	=	=	=	+3	=
DYS389II	=	=	=	=	=	Add DYS389I value	Add DYS389I value +3	=
DYS441	=	=	=	+1	+2	NT	NT	+1
DYS442	=	=	=	+5	+5	NT	NT	+5
DYS454	=	=	=	=	+1	NT	NT	=
DYS458	=	=	=	=	+2	NT	NT	=
DYS481	NT	NT	NT	?	NT	NT	NT	?
DYS594	NT	NT	NT	?	NT	NT	NT	?
GATA-A10	=	=	=	NT	=	NT	NT	+2
GATA-H4	-10	-9	-10	+1	=	NT	NT	+1

Some Specific Examples

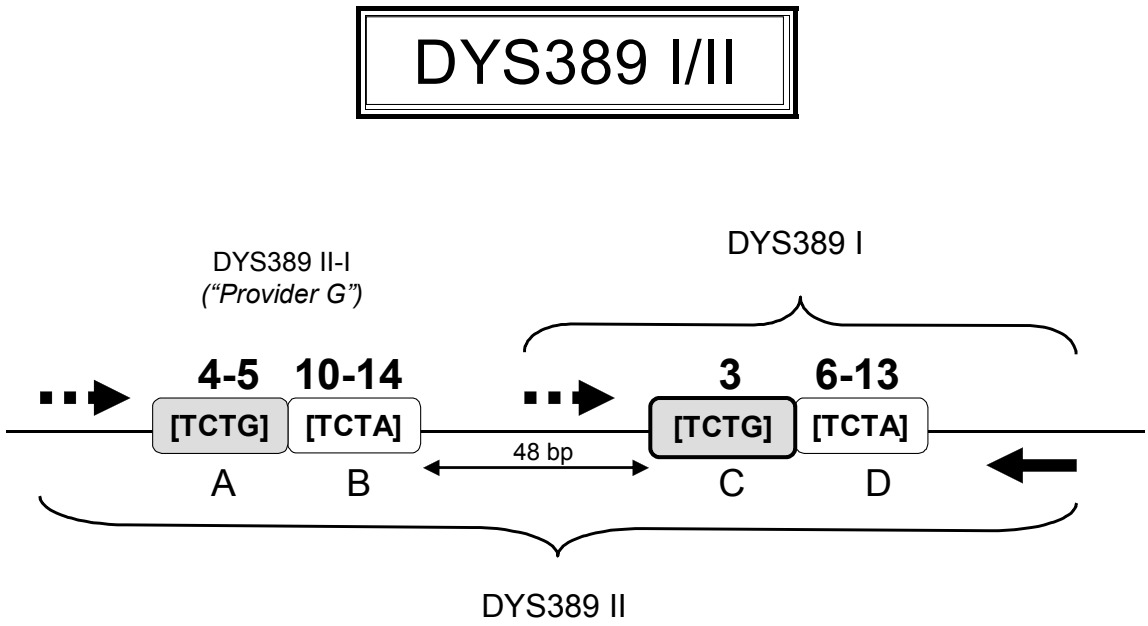
Reviewing some specific examples may help those interested in this topic better understand the challenges that exist with STR allele nomenclature designation. Table 4 lists allele nomenclature conversions required when Y-STR results from different genetic genealogy DNA test providers are compared with NIST nomenclature recommendations. Below, for each marker where differences between companies have been observed, we have tried to describe the likely reasons for each nomenclature difference along with an illustration of the STR repeat sequence and its various interpretations. We also provide our recommendations for the appropriate allele nomenclature in these specific instances.

DYS389I/II

Figure 3 schematically represents the four repeat blocks present at the DYS389 locus, which are designated here as "A", "B", "C", and "D" (see Rolf et al., 1998). Segments "A" and "C" are TCTG repeats that almost never vary while segments "B" and "D" contain TCTA repeat motifs that provide the bulk of the variation at this Y-STR locus. Due to sequence similarity near repeat blocks "A" and "C", the forward PCR primers shown as a dotted arrow in Figure 3, binds twice thus giving rise to two PCR products with a single forward (dotted arrow) and a single reverse (solid arrow) PCR primer. Repeat blocks "B" and "C" are separated by 48 bp (Rolf et al., 1998). The DYS389I PCR product is actually a

subset of the DYS389II amplified product since the forward primer binds to nearly identical flanking region sequences that are approximately 120 bp apart. Some analyses, such as those performed by Redd et al. (2002) or for Provider F (Table 4), treat the larger PCR product as DYS389II-I to better understand the variation occurring in regions "A" and "B" independent of "C" and "D."

One of the first articles on DYS389I/II (Kayser et al., 1997) defined this marker's allele nomenclature without the monomorphic TCTG denoted as segment "C" in Figure 3. Provider G (Table 4) appears to have adopted (or never changed from) the early approach and is thus leaving out segment "C" (and its constant three TCTG repeats), which has now been added by all other laboratories and publications since the late 1990s. Note that this impacts both DYS389I and DYS389II. The Y-Chromosome Haplotype Reference Database (YHRD) and all commercial Y-STR kits include segment "C" in their nomenclatures. NIST supports the inclusion of segment "C" in the nomenclature provided with the SRM 2395 Certificate.



DYS389I
"Provider F": segment D
All others: segments C+D

DYS389II
"Provider F": segments A+B+D
"Provider G": segments A+B only
All others: segments A+B+C+D

Figure 3. Schematic of the two TCTG and two TCTA repeat regions, labeled as segments A-D, present at the DYS389I/II Y-STR marker. Since the forward primer (dashed arrow) anneals twice, two PCR products are generated: DYS389I copies segments C and D while DYS389II encompassed segments A-D. Early nomenclatures did not include the three TCTG repeats in segment C, which impacts both DYS389I and DYS389II. The NIST SRM 2395 certified values support nomenclatures utilizing all four segments.

DYS441

The Y-STR marker DYS441 was first described by Iida et al., 2001. In the original article, the repeat motif was designated as [CCTT], which did not follow the 1997 ISFG recommendations (Bär et al., 1997) of moving the repeat motif as far as possible to the 5'end of the counted strand. As noted in the 2006 ISFG recommendations (Gusmão et al., 2006), DYS441 should more appropriately be designated with a [TTCC] motif, which leads to one extra repeat unit as illustrated in Figure 4. This is likely the reason that results from Providers D and H (Table 4) at DYS441, following the original Iida CCTT motif, are one repeat less than results from Providers A, B, and C (Table 4), which follow the 2006 ISFG recommended nomenclature. SRM 2395 does not currently contain information on DYS441 but NIST supports the ISFG use of the $(TTCC)_n$ motif shown in Figure 4A.

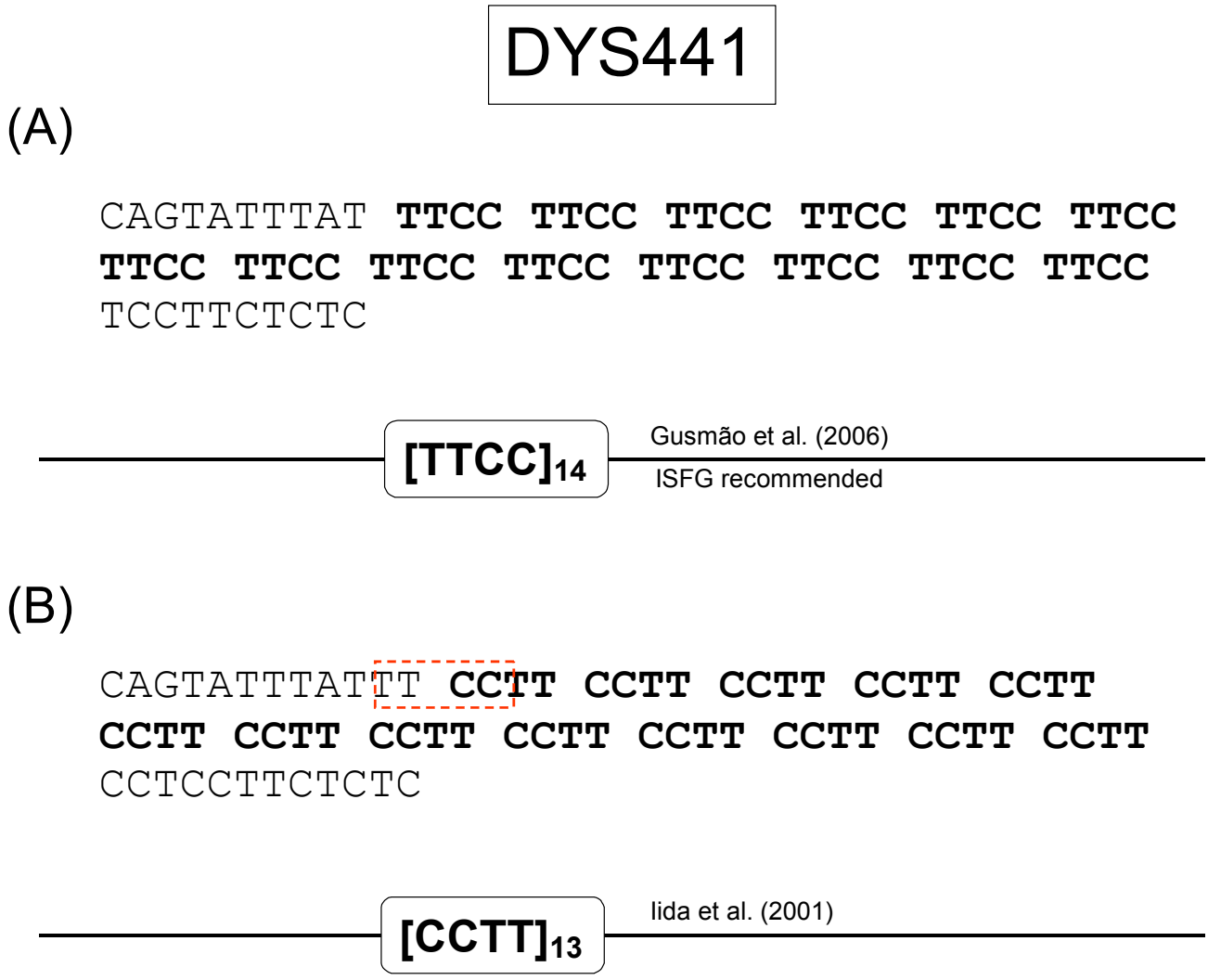


Figure 4. The repeat region and a few flanking nucleotides for DYS441 are compared with two different approaches to defining the nomenclature. (A) The 2006 ISFG recommended motif of TTCC utilizes the furthest repeat to the 5' end, which results in one extra repeat difference compared to (B) the original CCTT motif cited by Iida et al. (2001). Although SRM 2395 does not have certified values for DYS441, NIST supports the ISFG recommendations shown in (A).

DYS442

DYS442 is a compound repeat first described by Iida et al., (2001). In the original article, the two TATC and three TGTC repeat blocks were not included in the nomenclature as illustrated in Figure 5B. The 2006 ISFG recommendations favor including the adjacent repeat blocks in this compound repeat. SRM 2395 does not currently contain information on DYS442 but NIST supports the ISFG use of the $(TATC)_2(TGTC)_3(TATC)_n$ motif shown in Figure 5A, and this results in calling this marker as five repeats greater than in the Iida, et al. (2001) approach.

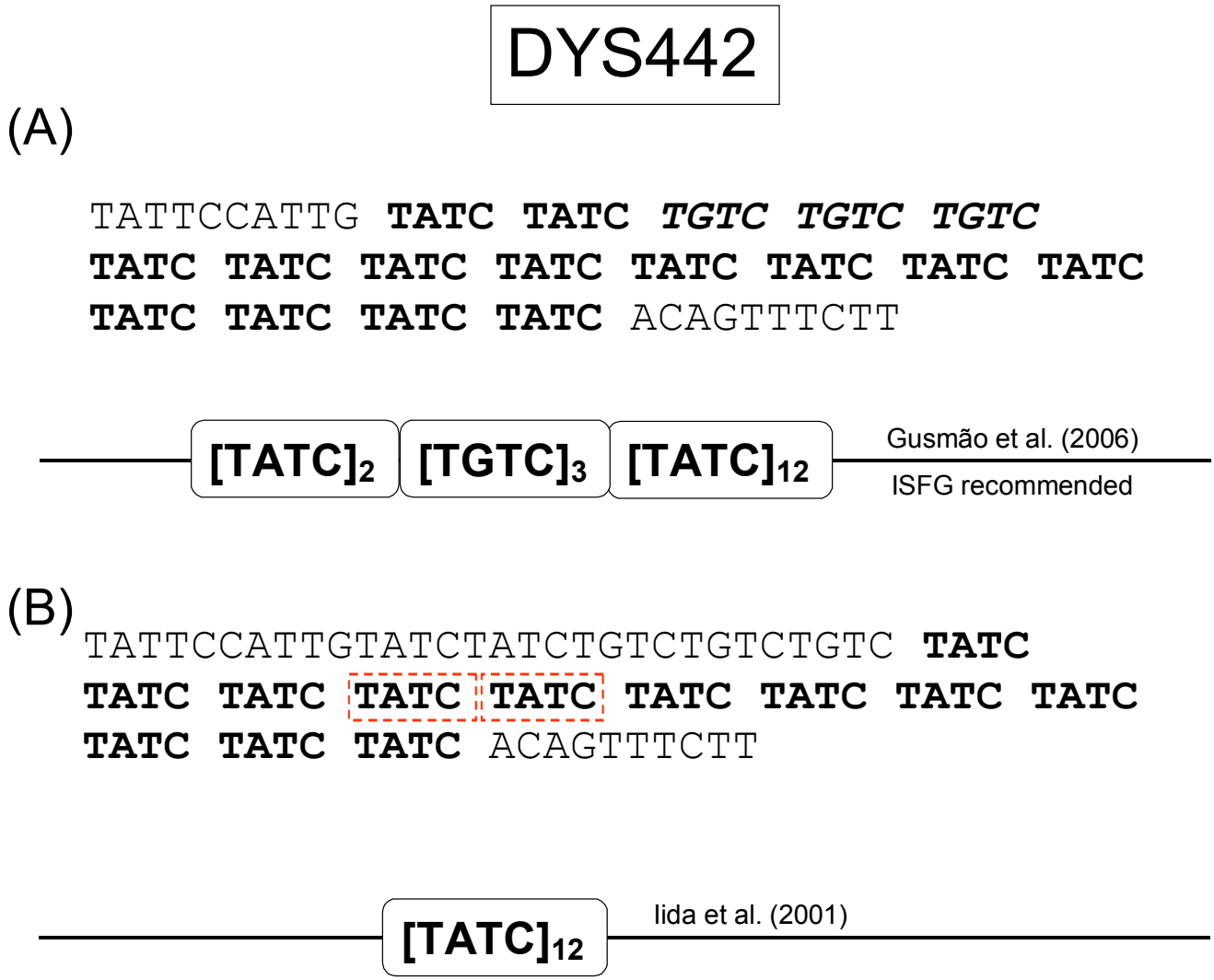


Figure 5. The repeat region and a few flanking nucleotides for DYS442 are compared with two different approaches to defining the nomenclature. (A) The 2006 ISFG recommended motif of $(TATC)_2(TGTC)_3(TATC)_n$ utilizes two earlier TATC repeats and intervening three TGTC repeats to increase the repeat count by five compared to (B) the original TATC motif cited by Iida et al. (2001). Although SRM 2395 does not have certified values for DYS442, NIST supports the ISFG recommendations shown in (A).

DYS454

DYS454 was first described by Redd et al. (2002) and their original nomenclature was advocated by the 2006 ISFG recommendations. It is unclear why any additional nomenclatures, such as the addition of a single repeat, might be considered for DYS454. SRM 2395 does not currently contain information on DYS454 but NIST supports the original nomenclature of Redd et al. (2002) and ISFG use of the $(AAAT)_n$ motif shown in Figure 6.

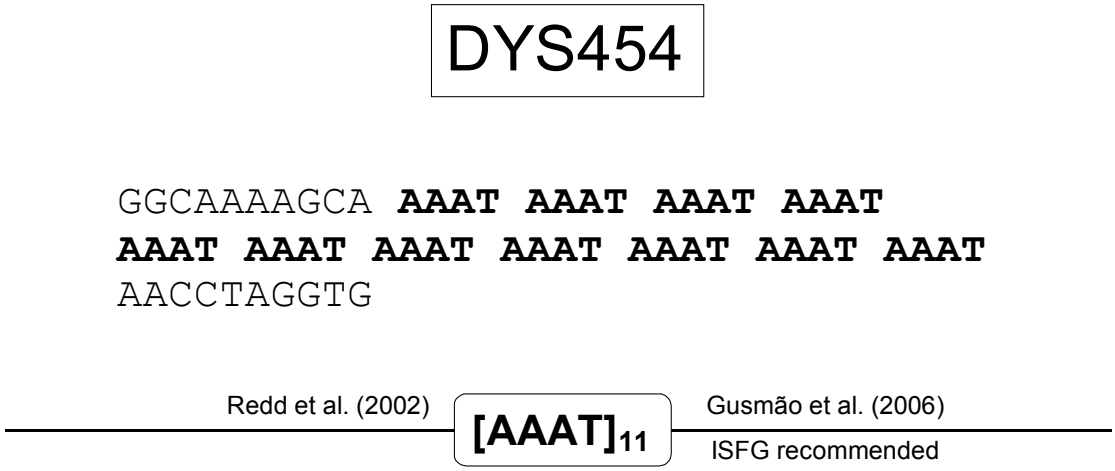


Figure 6. The DYS454 repeat motif that was originally reported by Redd et al. (2002) was advocated by the 2006 ISFG DNA Commission article. It is unclear why an additional nomenclature with one extra repeat might be considered for DYS454. Although SRM 2395 does not have certified values for DYS454, NIST supports the ISFG recommendations.

DYS458

DYS458 was first described by Redd et al. (2002) and their original nomenclature was advocated by the 2006 ISFG recommendations (Figure 7A). This nomenclature is also used in the commercial Y-STR kit Yfiler from Applied Biosystems. Although there are three GAAA repeats which occur six nucleotides upstream of the core GAAA repeat (Figure 7B), the spacing is not correct to connect them to the larger (main) block of GAAA repeats, as previously described in 2006 ISFG recommendation #2. NIST supports the original DYS458 nomenclature of Redd et al. 2002 and ISFG use of the $(GAAA)_n$ motif shown in Figure 7A. SRM 2395 has certified values for its components with DYS458 using this nomenclature.

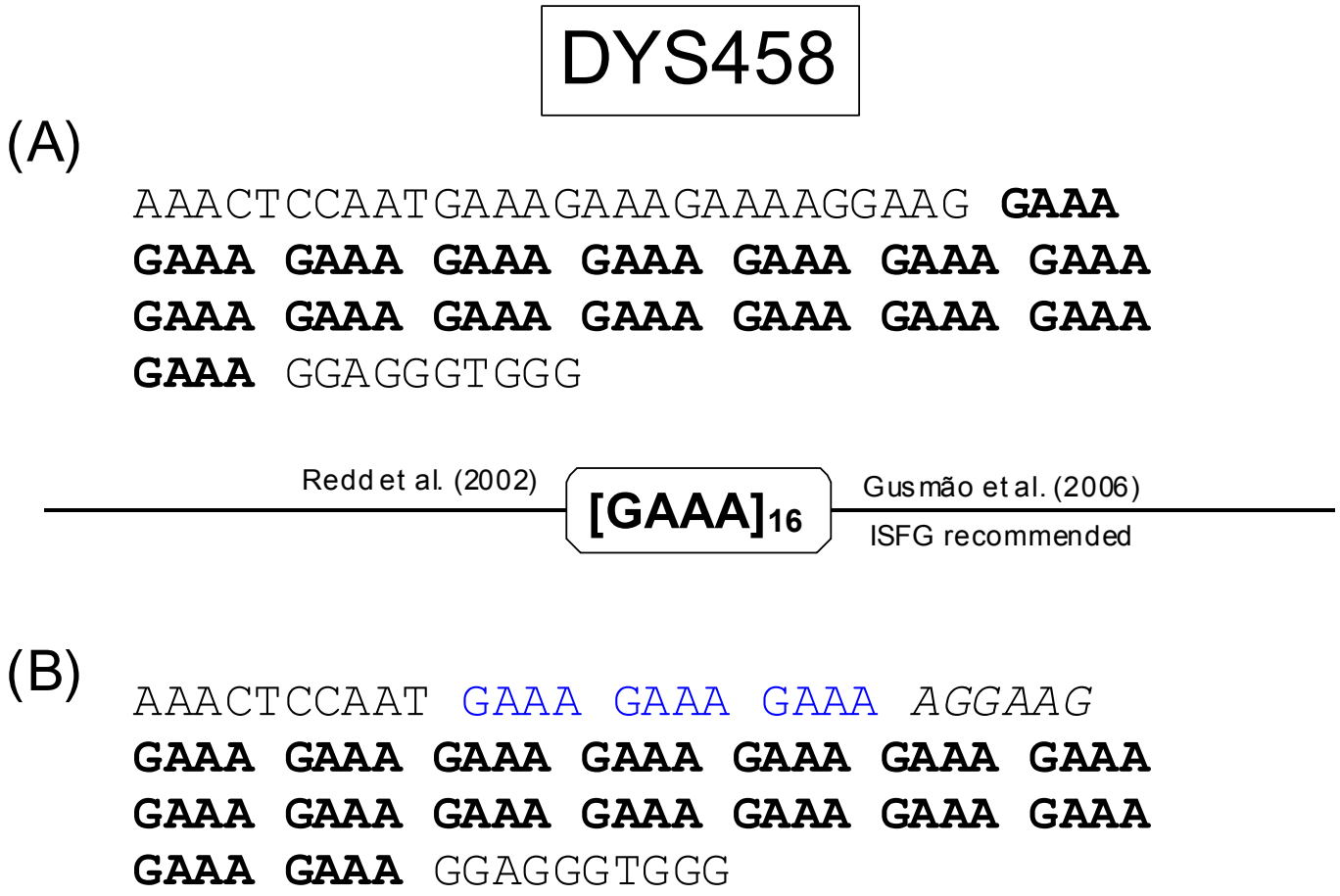


Figure 7. (A) The original DYS458 repeat motif of GAAA described by Redd et al. (2002) was advocated by the 2006 ISFG DNA Commission. This nomenclature is also used in the commercial Y-STR kit Yfiler from Applied Biosystems. (B) Although there are three GAAA repeats (in blue font) that occur six nucleotides upstream of the core GAAA repeat, the spacing is not correct to connect them to the larger block of GAAA repeats (see 2006 ISFG recommendation #2). It is unclear how only an additional two repeats might be considered here. The NIST SRM 2395 certified values support the ISFG recommendations shown in (A).

DYS481

DYS481 was first described by Kayser et al. (2004) with further population data noted in Lim et al. (2007). The CTT simple repeat motif originally described has been certified with NIST SRM 2395 (Figure 8A). While the addition of the adjacent TTT may be considered to qualify under the "one-change-rule" (Figure 8B), the presence of a homopolymeric stretch, rather than a true repeat unit, leads us to favor a nomenclature that only utilizes the CTT repeat. NIST SRM 2395 has certified values for its components with DYS481 using the (CTT)$_n$ motif shown in Figure 8A.

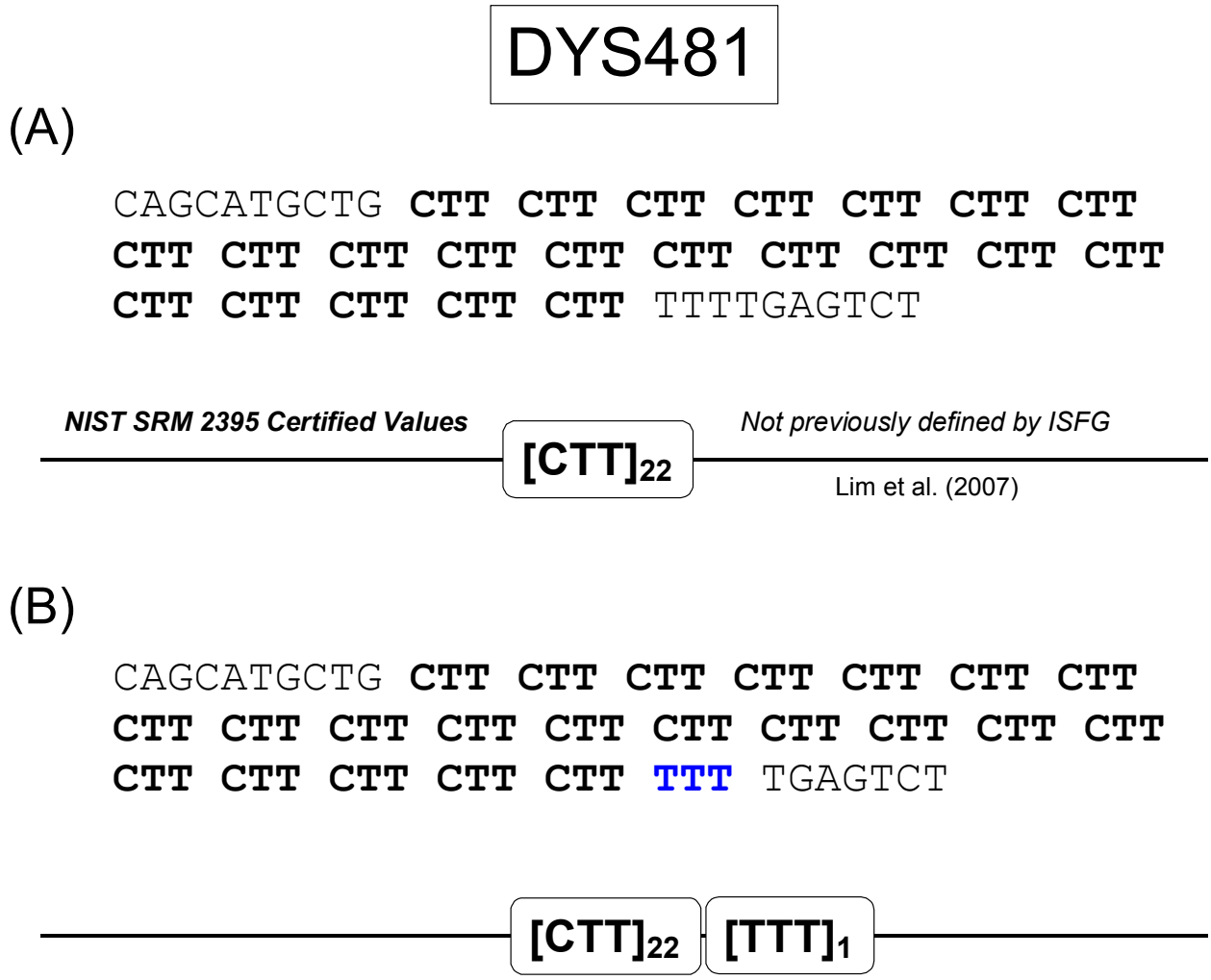


Figure 8. The repeat region and a few flanking nucleotides for DYS481 are compared with two different approaches to defining the nomenclature: (A) a simple CTT motif and (B) the CTT motif plus TTT. While the addition of the TTT may be considered to qualify under the "one-change-rule", the presence of a homopolymeric stretch rather than a true repeat unit leads us to favor the nomenclature shown in (A). DYS481 was not included in the 2006 ISFG recommendations but is in Lim et al. (2007). The NIST SRM 2395 certified values support the nomenclature shown in (A).

DYS594

DYS594 was first described by Kayser et al. (2004) with further population data noted in Butler et al. (2006) and Lim et al. (2007). Although the 2006 ISFG recommended motif for DYS594 was described as TAAAA (Gusmão et al., 2006; see also Butler et al., 2006) as shown in Figure 9(A), it could more appropriately be described as AAATA as shown in Figure 9B. While the addition of the AAAAA may be considered to qualify under the "one-change-rule," the presence of a homopolymeric stretch, rather than a true repeat unit leads us to favor not including it in the final nomenclature. Although SRM 2395 does not have certified values for DYS594, NIST supports the use of the just the AAATA repeat motif without the AAAAA, as shown in Figure 9B.

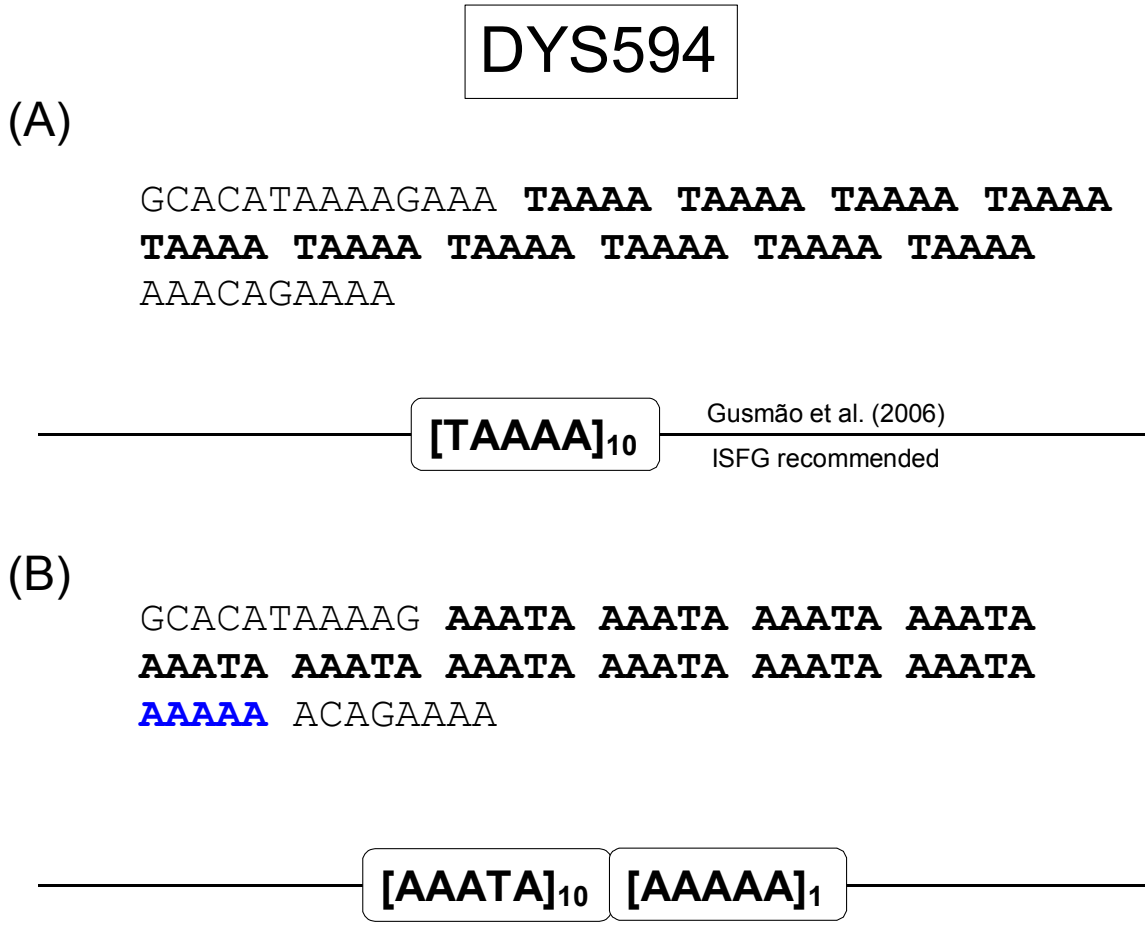


Figure 9. The repeat region and a few flanking nucleotides for DYS594 are compared with two different approaches to defining the nomenclature. Although the 2006 ISFG recommended motif for DYS594 was described as TAAAA (Gusmão et al., 2006; see also Butler et al., 2006) as shown in (A), it could more appropriately be described as AAATA as shown in (B). While the addition of the AAAAA may be considered to qualify under the "one-change-rule," the presence of a homopolymeric stretch rather than a true repeat unit leads us to favor not including it in the final nomenclature. Although SRM 2395 does not have certified values for DYS594, NIST supports the use of the just the AAATA repeat motif without the AAAAA shown in (B).

Y-GATA-A10

Y-GATA-A10 was first described by White et al. (1999) although the allele nomenclature was not clearly defined in the original work. Additional population studies and comparative sequence analysis with chimpanzees (Gusmão et al., 2002) led to inclusion of two TCCA repeats adjacent to the primary TATC repeat motif (Figure 10A). This approach was advocated by the 2006 ISFG recommendations (Gusmão et al., 2006). Some laboratories have apparently decided to count only the TATC repeat block, leading to a repeat count that is two less than the ISFG recommendations (Figure 10B). Although SRM 2395 does not currently have certified values for GATA-A10, NIST supports the ISFG recommendations shown in Figure 10A.

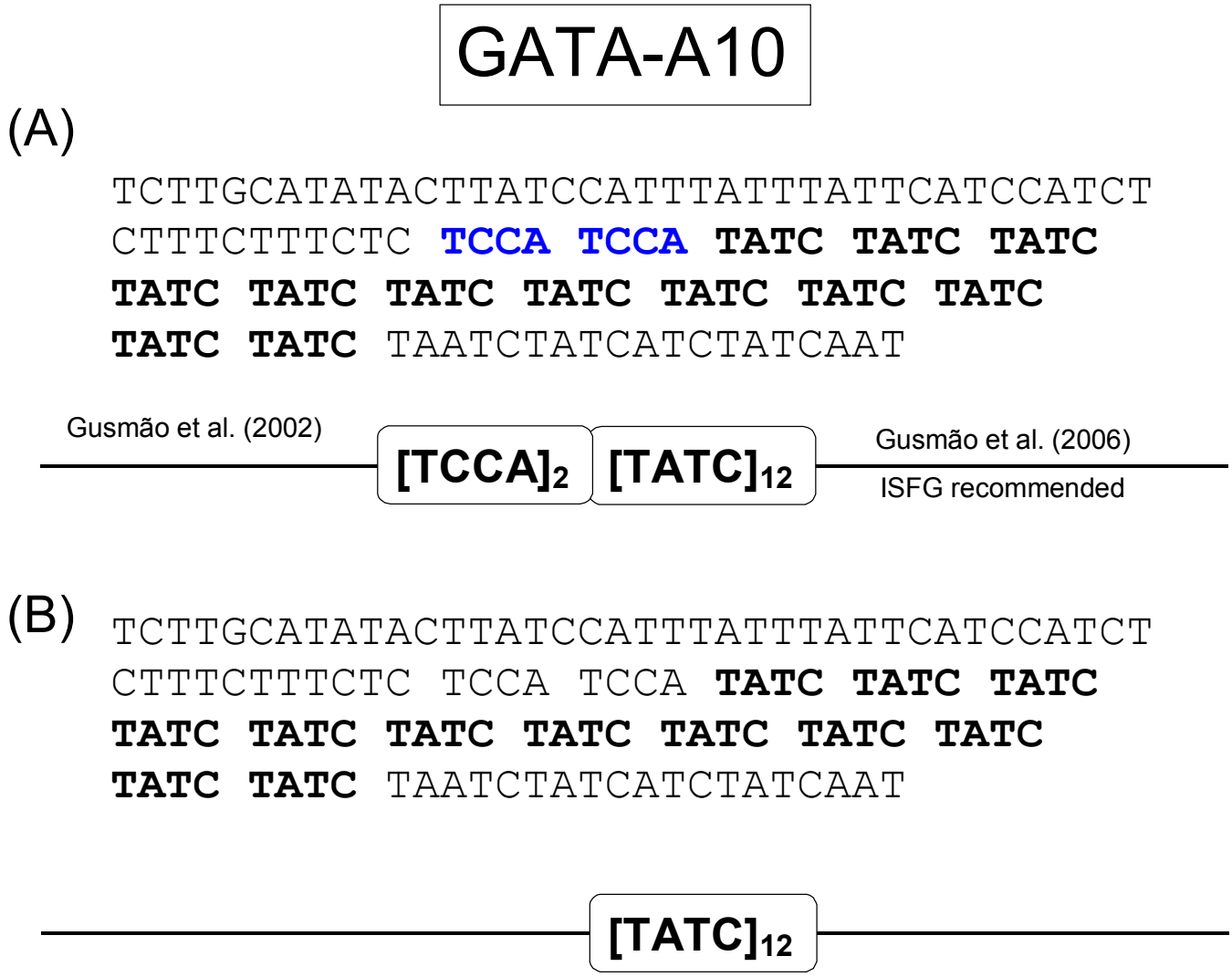


Figure 10. The repeat region and a few flanking nucleotides for GATA-A10 are compared with two different approaches to defining the nomenclature. (A) The 2006 ISFG recommends including two adjacent TCCA repeats to the primary TATC motif due to previous work with chimpanzee sequences (Gusmão et al., 2002). (B) Some laboratories have apparently decided to only count the TATC repeat block leading to a repeat count that is two less than the ISFG recommendations. Although SRM 2395 does not currently have certified values for GATA-A10, NIST supports the ISFG recommendations shown in (A).

Y-GATA-H4

Y-GATA-H4 was first described by White et al. (1999), although the allele nomenclature was not clearly defined in the original work. This marker and GATA-A10 were originally named "GATA" repeats because this was the probe sequence used to locate these markers on the Y-chromosome. The "A10" and "H4" designations came from the 96-well plate position of the specific clone containing the newly discovered Y-STR marker in the probe screen. Figure 11 maps the PCR primer sequences and various nomenclatures on the original GenBank reference sequence submitted in May 1999 by White et al. (1999).

Gonzalez-Neira et al. (2001), the first major forensic group working with this marker, originally proposed 28 repeats as their reference allele using a convoluted $[AGAT]_4$-N_2-$[ATAG]_3$-$[GTAG]_3$-$\mathbf{[ATAG]_{10}}$-N_{13}-$[GATG]_2$-N_1-$[ATAG]_4$-N_4-$[ATAG]_2$ repeat motif.

Additional work by Gusmão et al. (2002), which included some of the same scientists as the Gonzalez-Neira et al. (2001) effort, changed the designated repeat block to $[AGAT]_4$-N_4-$[AGAT]_2$-$[AGGT]_3$-$\mathbf{[AGAT]_{10}}$-N_{24}-$[ATAG]_4$-$[ATAC]_1$-$[ATAG]_2$ and then based on comparative chimpanzee sequence information, decided to break the GATA-H4 repeat into two sections: $[AGAT]_4$-N_4-$[AGAT]_2$-$[AGGT]_3$-$\mathbf{[AGAT]_{10}}$ ("H4.1 locus") and $[ATAG]_4$-$[ATAC]_1$-$[ATAG]_2$ ("H4.2 locus", which is an invariant seven-repeat block in humans).

About this same time, our group at NIST had developed a new assay for detecting the primary variable portion

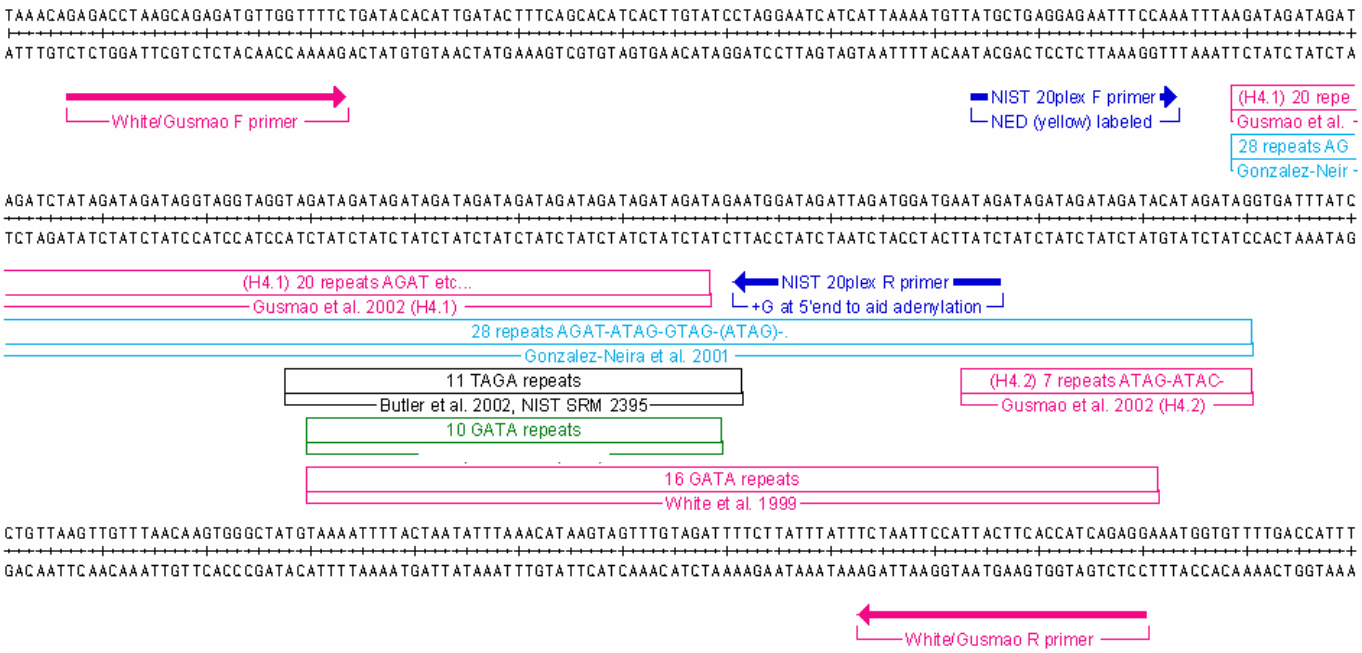


Figure 11. A map of the Y-STR marker GATA-H4 GenBank reference sequence illustrating a number of PCR primers (arrows) and repeat regions (boxes) described in the literature. The NIST SRM 2395 certified values support the nomenclature for the 11 TAGA repeats targeted by the 20-plex primers as described by Butler et al. (2002) rather than the 2006 ISFG recommendations. However, as noted in Mulero et al. (2006b), conversion factors to different nomenclatures can easily be performed.

of the GATA-H4 locus and these PCR primers were published as part of our Y-STR 20-plex assay (Butler et al., 2002). The PCR primer sequences used by White et al. (1999) and Gusmão et al., (2002) are illustrated relative to those employed with the NIST 20-plex assay (Butler et al., 2002). Since our primers only targeted the variable portion of the repeat, we settled on use of the TAGA motif as this is the first adjacent repeat starting from the 5' end of the reference sequence. In the case of the GenBank reference sequence shown in **Figure 11**, there are 11 TAGA repeats.

The 2003 release of NIST SRM 2395 included certified values based on DNA sequencing and Y-STR typing using the NIST 20-plex assay. Unfortunately, at about this same time, Provider H started reporting values for GATA-H4 but with a "GATA" motif rather than the 5'-maximized "TAGA" motif. As can be seen in **Figure 11**, this is the reason for the one repeat difference in nomenclature.

Later, when the 2006 ISFG recommendations were published (Gusmão et al., 2006), they included citation to the Gusmão et al., 2002 approach for GATA-H4. The release of the commercial kit Yfiler prompted the publication of conversion factors between the SRM 2395 values used by Y-filer and the ISFG recommendations used by some laboratories in Europe (Mulero et al. 2006b). We have perpetuated the original SRM 2395 nomenclature in our updated certificate with a citation to the possibility of using conversion factors. Therefore, those who choose to follow the allele nomenclature recommendations of the 2006 ISFG DNA Commission should add a correction factor of nine to the SRM 2395 allele number, and they should refer to this marker as GATA H4.1. Alternatively, those who amplify the entire GATA-H4 region (GATA-H4.1 and GATA-H4.2) should add a correction factor of 16 to the SRM 2395 allele number (see also H4 Nomenclature 2008).

Additional Y-STR Work

Our project team at NIST has been actively involved since 2000 in improving knowledge about the Y-chromosome and its genetic variation. In the past eight years, we have published more than 20 articles on various Y-STR assays (Butler et al., 2002; Schoske et al., 2004), developed NIST SRM 2395 and characterized its components at a number of loci, examined Y-STR duplication events (Butler et al., 2005), studied mutation rates in father/son pairs (Decker et al., 2008), and conducted numerous studies on the genetic diversity of Y-STR and Y-SNP markers in U.S. populations (Vallone and Butler, 2004; Butler et al., 2006; Decker et al., 2007; Butler et al., 2007).

One of the primary drivers for this effort has been to better understand the impact of additional Y-STR loci in resolving common haplotypes and lineages (Butler et al., 2007; Hanson and Ballantyne, 2007; Rodig et al., 2008). In our studies at NIST, we have measured genetic diversity of 82 Y-STR loci in a set of 31 Caucasian, 32 African American, and 32 Hispanic samples (**Table 5**). Understanding this genetic diversity can be helpful as specific markers are selected for potential future applications that may benefit from faster or slower Y-STR variability/mutation rates.

Conclusions

The adoption of Y-STR markers beyond those available in commercial kits has been especially rapid within the genetic genealogy community over the past few years. Differences in allele nomenclature between the various genetic genealogy DNA test providers have lead to frustration and confusion on the part of many users. This article describes the issues behind STR allele nomenclature designation and provides some specific examples. NIST has developed a Standard Reference Material (SRM 2395) that has certified values at many of the Y-STR markers used by the genetic genealogy community. We strongly encourage its use to enable compatible and calibrated measurements to be made between different Y-STR testing laboratories. With Y-STR markers that go beyond those currently characterized in SRM 2395, we encourage DNA test providers to supply their results back to NIST so that we can track the usage of different Y-STRs. "New" markers showing high usage can then be considered for inclusion in future SRM 2395 certificate updates.

Acknowledgments

This work was funded in part by the National Institute of Justice (NIJ) through interagency agreement 2008-DN-R-121 with the NIST Office of Law Enforcement Standards. The early efforts of Richard Schoske and Jill Appleby with sequence analysis on NIST SRM 2395 components are greatly appreciated. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

References

Bär W, Brinkmann B, Lincoln P, Mayr WR, Rossi U (1994) DNA recommendations – 1994 report concerning further recommendations of the DNA Commission of the ISFH regarding PCR-based polymorphisms in STR (short tandem repeat) systems. *Int J Legal Med,* 107: 159-160.

Table 5
Diversity values for 82 Y-STR markers in a common set of DNA samples. These loci were selected based on best candidates from previous studies (Redd et al., 2002; Kayser et al., 2004; Leat et al., 2007; Lim et al., 2007) and examined in a U.S. population screen consisting of 31 Caucasians, 32 African Americans, and 32 Hispanics. Values in parentheses next to the locus name indicate the number of samples used in the calculations of number of alleles and diversity if fewer than the full set of 95 samples. The number of alleles observed and calculated gene diversity values were determined as previously described in Schoske et al. (2004) and Butler et al. (2006). Those loci highlighted in yellow have been sequenced and are included on the updated SRM 2395 certificate.

Locus	# Alleles	Diversity	Locus	# Alleles	Diversity	Locus	# Alleles	Diversity
DYS724 a/b (CDY) (93)	36	0.9691	DYS456 (94)	5	0.7355	DYS462	6	0.5669
DYS464 a/b/c/d (91)	42	0.9646	DYS607	7	0.7355	DYS537	3	0.5648
DYS527 a/b (93)	32	0.9388	DYS438 (94)	5	0.7211	DYS594 (93)	5	0.5617
DYS710 (93)	17	0.9236	DYS19 (94)	5	0.7113	DYS391 (94)	4	0.5502
DYS385 a/b (94)	29	0.9179	DYS508 (93)	7	0.7106	DYS531	6	0.5357
DYS481 (93)	11	0.8359	DYS446 (94)	7	0.7014	DYS556 (93)	4	0.5346
DYS449 (90)	12	0.8345	DYS448 (94)	6	0.6937	DYS721	4	0.5234
DYS712	12	0.834	DYS723 (94)	4	0.6891	DYS426 (91)	3	0.5221
DYS490 (92)	18	0.8201	DYS485 (93)	8	0.6821	DYS565	3	0.5165
DYS504 (94)	9	0.8101	DYS522 (94)	4	0.6792	DYS578	3	0.5165
DYS576 (93)	8	0.8046	DYS495 (94)	5	0.6747	DYS525 (93)	7	0.5157
DYS570 (94)	10	0.8042	DYS716	4	0.6524	DYS450 (91)	3	0.5070
YCAII a/b (91)	13	0.7993	DYS452 (93)	7	0.6487	DYS632 (94)	2	0.5017
DYS557 (93)	7	0.7887	Y-GATA-H4 (94)	5	0.6461	DYS726 (94)	4	0.4907
DYS534 (93)	9	0.7882	DYS505 (93)	5	0.6454	DYS540 (94)	4	0.4871
DYS643 (92)	7	0.7862	DYF406S1 (DYS555)	5	0.6421	DYS393 (94)	4	0.4770
DYS458 (94)	8	0.7808	DYS437 (94)	5	0.6417	DYS717	7	0.4531
DYS635 (94)	8	0.7779	DYS439 (94)	4	0.6388	DYS388 (91)	8	0.4498
DYS652	10	0.7742	DYS520(94)	6	0.6381	DYS719 (94)	6	0.3606
DYS650	10	0.774	Y-GATA-A10	4	0.6336	DYS425	3	0.2278
DYS459 a/b	6	0.768	DYS492 (93)	5	0.6335	DYS454	5	0.1957
DYS463	9	0.768	DYS444 (88)	6	0.6264	DYS645	3	0.1820
DYS447 (91)	9	0.7636	DYS533 (94)	6	0.6264	DYS455	5	0.1781
DYS390 (94)	6	0.7632	DYS460 (91)	4	0.5973	DYS641 (94)	3	0.1219
DYS715 (94)	7	0.7628	DYS392 (94)	7	0.5962	DYS434	3	0.0824
DYS532 (94)	7	0.7541	DYS389I (94)	3	0.5692	DYS575 (94)	2	0.0213
DYS389II (94)	5	0.7447	DYS572 (93)	4	0.5676	DYS472	1	0.0000
DYS709	8	0.7402						

Bär W, Brinkmann B, Budowle B, Carracedo A, Gill P, Lincoln P, Mayr W, Olaisen B (1997) DNA recommendations – further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. *Int J Legal Med*, 110:175-176.

Brown K (2002) Tangled roots? Genetics meets genealogy. *Science*, 295:1634-1635.

Budowle B, Moretti TR, Niezgoda SJ, Brown BL (1998) CODIS and PCR-based short tandem repeat loci:law enforcement tools. *Proceedings of the Second European Symposium on Human Identification* Madison, WI: Promega Corporation, 1998, 73-88; http://www.promega.com/geneticidproc/eusymp2proc/17.pdf

Budowle B, Masibay A, Anderson SJ, Barna C, Biega L, Brenneke S, Brown BL, Cramer J, DeGroot GA, Douglas D, Duceman B, Eastman A, Giles R, Hamill J, Haase DJ, Janssen DW, Kupferschmid TD, Lawton T, Lemire C, Llewellyn B, Moretti T, Neves J, Palaski C, Schueler S, Sgueglia J, Sprecher C, Tomsey C, Yet D (2001) STR primer concordance study. *Forensic Sci Int*, 124: 47-54.

Butler JM, Schoske R, Vallone PM, Kline MC, Redd AJ, Hammer MF (2002) A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. *Forensic Sci Int*, 129:10-24.

Butler JM (2003) Recent developments in Y-short tandem repeat and Y-single nucleotide polymorphism analysis. *Forensic Sci Rev*, 15:91-111.

Butler JM (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2nd Edition). Elsevier Academic Press, New York.

Butler JM, Decker AE, Kline MC, Vallone PM (2005) Chromosomal duplications along the Y-chromosome and their potential impact on Y-STR interpretation. *J Forensic Sci*, 50:853-859.

Butler JM (2006) Genetics and genomics of core STR loci used in human identity testing. *J Forensic Sci*, 51:253-265.

Butler JM, Decker AE, Vallone PM, Kline MC (2006) Allele frequencies for 27 Y-STR loci with U.S. Caucasian, African American, and Hispanic samples. *Forensic Sci Int*, 156:250-260.

Butler JM, Hill CR, Decker AE, Kline MC, Reid TM, Vallone PM (2007) New autosomal and Y-chromosome STR loci: characterization and potential uses. *Proceedings of the Eighteenth International Symposium on Human Identification*. See http://www.promega.com/geneticidproc/

CODIS (FBI's Combined DNA Index System): http://www.fbi.gov/hq/lab/html/codis1.htm

CODIS Quality Assurance (2008): http://www.fbi.gov/hq/lab/html/codis5.htm

Decker AE, Kline MC, Vallone PM, Butler JM (2007) The impact of additional Y-STR loci on resolving common haplotypes and closely related individuals. *FSI Genetics*, 1:215-217.

Decker AE, Kline MC, Redman JW, Reid TM, Butler JM (2008) Analysis of mutations in father-son pairs with 17 Y-STR loci. *FSI Genetics, 2(3):e31-e35.*

Dupuy BM, Stenersen M, Egeland T, Olaisen B (2004) Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Human Mutation*, 23:117-124.

Foster EA, Jobling MA, Taylor PG, Donnelly P, de Knijff P, Mieremet R, Zerjal T, Tyler-Smith C (1998) Jefferson fathered slave's last child. *Nature*, 396:27-28.

Furedi S, Woller J, Padar Z, Angyal M (1999) Y-STR haplotyping in two Hungarian populations. *Int J Legal Med*, 113:38-42.

Gill P, Brenner C, Brinkmann B, Budowle B, Carracedo A, Jobling MA, de Knijff P, Kayser M, Krawczak M, Mayr WR, Morling N, Olaisen B, Pascali V, Prinz M, Roewer L, Schneider PM, Sajantila A, Tyler-Smith C (2001) DNA Commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome STRs. *Forensic Sci Int*, 124:5-10.

Gonzalez-Neira A, Elmoznino M, Lareu MV, Sanchez-Diz P, Gusmão L, Prinz M, Carracedo A (2001) Sequence structure of 12 novel Y chromosome microsatellites and PCR amplification strategies. *Forensic Sci Int*, 122:19-26.

Gross AM, Berdos P, Ballantyne J (2006) Y-STR concordance study between Y-Plex5, Y-Plex6, Y-Plex12, PowerplexY, Y-Filer, MPI, and MPII. *J Forensic Sci*, 51:1423-1428.

Gusmão L, Gonzalez-Neira A, Alves C, Lareu M, Costa S, Amorim A, Carracedo A (2002) Chimpanzee homologous of human Y specific STRs. A comparative study and a proposal for nomenclature. *Forensic Sci Int*, 126:129-136.

Gusmão L, Butler JM, Carracedo A, Gill P, Kayser M, Mayr WR, Morling N, Prinz M, Roewer L, Tyler-Smith C, Schneider PM (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic Sci Int*, 157:187-197.

H4 Nomenclature (2008): http://www.cstl.nist.gov/biotech/strbase/YSTRs/H4_nomenclature.htm

Hanson EK, Ballantyne J (2006) Comprehensive annotated STR physical map of the human Y chromosome: forensic implications. *Legal Med*, 8:110-120; see also http://ncfs.ucf.edu/ystar/ystar.html

Hanson EK, Ballantyne J (2007) An ultra-high discrimination Y chromosome short tandem repeat multiplex DNA typing system. *PLoSONE* 8:e688.

Iida R, Tsubota E, Matsuki T (2001) Identification and characterization of two novel human polymorphic STRs on the Y chromosome. *Int J Legal Med*, 115:54-56.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431:931-945.

Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M, Honda K, Jobling M, Krawczak M, Leim K, Meuser S, Meyer E, Oesterreich W, Pandya A, Parson W, Penacino G, Perez-Lezaun A, Piccinini A, Prinz M, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Weichhold GM, de Knijff P, Roewer L (1997) Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med*, 110(3):125-133 (Appendix 141-149).

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Krüger C, Krawczak M, Nagy M, Dobosz T, Szibor R, de Knijff P, Stoneking M, Sajantila A (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet*, 66:1580–1588.

Kayser M, Kittler R, Ralf A, Hedman M, Lee AC, Mohyuddin A, Mehdi SQ, Rosser Z, Stoneking M, Jobling MA, Sajantila A, Tyler-Smith C (2004) A comprehensive survey of human Y-chromosomal microsatellites. *Am J Hum Genet*, 74(6):1183-1197.

Kline MC, Duewer DL, Newall P, Redman JW, Reeder DJ, Richard M (1997) Interlaboratory evaluation of short tandem repeat triplex CTT. *J Forensic Sci*, 42(5):897-906.

Kline MC, Decker AE, Hill CR, Butler JM (2006) NIST SRM Updates: Value-added to the Current Materials in SRM 2391b and SRM 2395. Poster at 17th International Symposium on Human Identification (Nashville, TN), October 10-12, 2006; available at http://www.cstl.nist.gov/biotech/strbase/pub_pres/Promega2006_Kline.pdf

Krenke BE, Viculis L, Richard ML, Prinz M, Milne SC, Ladd C, Gross AM, Gornall T, Frappier JR, Eisenberg AJ, Barna C, Aranda XG, Adamowicz MS, Budowle B (2005) Validation of male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. *Forensic Sci Int*, 151:111-124.

Leat N, Ehrenreich L, Benjeddou M, Cloete K, Davison S (2007) Properties of novel and widely studied Y-STR loci in three South African populations. *Forensic Sci Int*, 168:154-161.

Lim S-K, Xue Y, Parkin EJ, Tyler-Smith C (2007) Variation of 52 new Y-STR loci in the Y Chromosome Consortium worldwide panel of 76 diverse individuals. *Int J Legal Med*, 121:124-127.

May WE, Parris RM, Beck CM, Fassett JD, Greenberg RR, Guenther FR, Kramer GW, Wise SA, Gills TE, Colbert JC, Gettings RJ, MacDonald BR (2000) Definitions of terms and modes used at NIST for value-assignment of reference materials for Chemical Measurements. *NIST Special Publication* 260-136.

Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK (2006a) Development and validation of the AmpFlSTR Yfiler PCR Amplification Kit: a male specific, single amplification 17 Y-STR multiplex system. *J Forensic Sci*, 51:64-75.

Mulero JJ, Budowle B, Butler JM, Gusmão L (2006b) Letter to the Editor-Nomenclature and allele repeat structure update for the Y-STR locus GATA H4. *J Forensic Sci*, 51:694.

Redd AJ, Agellon AB, Kearney VA, Contreras VA, Karafet T, Park H, de Knijff P, Butler JM, Hammer MF (2002) Forensic value of 14 novel STRs on the human Y chromosome. *Forensic Sci Int*, 130:97-111.

Rodig, H, Roewer, L, Gross, A, Richter, T, de Knijff, P, Kayser, M, Brabetz, W (2008) Evaluation of haplotype discrimination capacity of 35 Y-chromosomal short tandem repeat loci. *Forensic Sci Int*, 174:182-188.

Rolf B, Meyer E, Brinkmann B, de Knijff P (1998) Polymorphism at the tetranucleotide repeat locus DYS389 in 10 populations reveals strong geographic clustering. *Eur J Hum Genet*, 6:583-588.

Rolf B, Keil W, Brinkmann B, Roewer L, Fimmer R (2001) Paternity testing using Y-STR haplotypes: assigning a probability for paternity in case of mutations. *Int J Legal Med,* 115:12-15.

Schoske R, Vallone PM, Kline MC, Redman JW, Butler JM (2004) High-throughput Y-STR typing of U.S. populations with 27 regions of the Y chromosome using two multiplex PCR assays. *Forensic Sci Int*, 139:107-121.

Shen P, Wang F, Underhill PA, Franco C, Yang WH, Roxas A, Sung R, Lin AA, Hyman RW, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (2000) Population genetic implications from sequence variation in four Y chromosome genes. *Proc Natl Acad Sci USA*, 97:7354-7359.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423:825-837.

SRM 2395 (2008): http://www.cstl.nist.gov/biotech/strbase/srm2395.htm and https://srmors.nist.gov/view_detail.cfm?srm=2395.

Stix, G (2008) Traces of the distant past. *Scientific American*, 299:56-63.

SWGDAM (2004) Report on the Current Activities of the Scientific Working Group on DNA Analysis Methods Y-STR Subcommittee. *Forensic Sci Comm*, 6(3).

Urquhart A, Kimpton CP, Downes TJ, Gill P (1994) Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers. *Int J Legal Med*, 107:13-20.

Vallone PM, Butler JM (2004) Y-SNP typing of U.S. African American and Caucasian samples using allele-specific hybridization and primer extension. *J Forensic Sci*, 49:723-732.

White PS, Tatum OL, Deaven LL, Longmire JL (1999) New male-specific microsatellite markers from the human Y chromosome. *Genomics*, 57:433-437.

Y-Chromosome Haplotype Reference Database (YHRD): http://www.yhrd.org/

YHRD Mutation page: http://www.yhrd.org/YSTR%20Loci/Mutations