Nomenclature Issues and the Y-Chromosome

John M. Butler, Ph.D.
National Institute of Standards and Technology

4th International Conference on Genetic Genealogy
Houston, TX
October 20, 2007

The Problem

• STR marker and allele nomenclatures differ between DNA testing laboratories that make data comparisons frustrating at best

http://www.ysearch.org/conversion_page.asp

The Solution

• Standardize the nomenclature and encourage (or require) all testing laboratories to calibrate their results to this standard nomenclature

Questions to Address Regarding STR Marker and Allele Nomenclature

• Where do STR marker names come from?
• Who decides on STR allele nomenclature?
• How is calibration of STR allele calls performed in the forensic DNA community?
• What are some potential solutions to aid the genetic genealogy community in marker standardization and STR allele calibration?

Presentation Outline

• Introduction to myself and NIST
• Overview of DNA Typing Process
• STR Allele Repeat Nomenclature
• NIST Standard Reference Materials

NIST and NIJ Disclaimer

Funding: Interagency Agreement 2003-IJ-R-029 between the National Institute of Justice and NIST Office of Law Enforcement Standards

Points of view are mine and do not necessarily represent the official position or policies of the US Department of Justice or the National Institute of Standards and Technology.

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.

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Nomenclature Issues and the Y-Chromosome
J.M. Butler (Genetic Genealogy Conference, Houston, TX) October 20, 2007

Our Team Mission Statement

• The NIST Human Identity Project Team is trying to lead the way in forensic DNA… through research that helps bring traceability and technology to the scales of justice.

NIST History and Mission

• National Institute of Standards and Technology (NIST) was created in 1901 as the National Bureau of Standards (NBS). The name was changed to NIST in 1988.

• NIST is part of the U.S. Department of Commerce with a mission to develop and promote measurement, standards, and technology to enhance productivity, facilitate trade, and improve the quality of life.

• NIST supplies over 1,300 Standard Reference Materials (SRMs) for industry, academia, and government use in calibration of measurements.

• NIST defines time for the U.S.

Location of NIST (near Washington, DC)

NIST Gaithersburg Campus

Located in Gaithersburg, Maryland, on approximately 234 hectares (578 acres) just off Interstate 270 about 25 miles northwest of Washington, D.C.

NIST Human Identity Project Team

All NIST publications and presentations available on STRBase:
http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm

• 22 publications since 2002 on Y-chromosome work
• >40 presentations and 10 training workshops to the community in the past year

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Overview of DNA Typing Process

Short Tandem Repeat (STR) Markers

PCR primers anneal to unique sequences bracketing the variable STR repeat region

DNA template containing STR marker

Fluorescent dye

PCR product size generated

The overall PCR product size is measured

Sample #1

Sample #2

11 GATA repeats ("11" is all that is reported)

PCR Primers in an STR Kit
Measurement (genotype determination) is performed by comparing allele size (relative to an internal size standard) to a commercially provided STR kit allelic ladder with calibrated repeat numbers (sized according to the same internal size standard).

Companies Supply Allelic Ladders in STR Kits to Aid Interlaboratory Consistency

Comparison of Allelic Ladder to Samples to Convert Size into Allele Repeat Number

STR genotyping is performed by comparison of sample data to allelic ladders

Methods for Typing STR Samples

- Allelic Ladders
  - Available in commercial kits
- Precise Sizing (and a sequenced sample)
- Locus-specific brackets (LSBs)

Concordance Studies Reveal Potential Primer Binding Site Mutations with Different Primer Sets

Appears to be an allele 11 dropout/reduction due to primer binding site mutation in MiniFiler
How Forensic DNA Typing Differs from Genetic Genealogy

- **Common set of core STR loci** are used to enable DNA database compatibility
- **Commercially available STR kits** are used with allelic ladders to ensure consistency between laboratories
- **Required calibration of STR allele calls** to available NIST SRMs to ensure accuracy

## Types of STR Repeat Units

Requires size based DNA separation to resolve different alleles from one another

- **Di**nucleotide (CA)(CA)(CA)
- **Tri**nucleotide (GCC)(GCC)(GCC)
- **Tetra**nucleotide (AATG)(AATG)(AATG)
- **Penta**nucleotide (AGAAAA)(AGAAAA)
- **Hexa**nucleotide (AGTACA)(AGTACA)

Short tandem repeat (STR) = microsatellite = simple sequence repeat (SSR)

### Equivalent Tetranucleotide STR Motifs

<table>
<thead>
<tr>
<th>Top Strand</th>
<th>Complementary (Bottom Strand)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGAT</td>
<td>TCTA</td>
</tr>
<tr>
<td>GATA</td>
<td>CTAT</td>
</tr>
<tr>
<td>ATAG</td>
<td>TATC</td>
</tr>
<tr>
<td>TAGA</td>
<td>ATCT</td>
</tr>
</tbody>
</table>

Same DNA sequence with different starting points for naming the repeat:

- ...AGATAGATAGATAGATAGATAG...
- ...AGATAAGATAGATAGATAGATAG...
- ...AGATAGATAGATAGATAGATAG...
- ...AGATAAGATAGATAGATAGATAG...

## STR Repeat Motifs

From D.N.A. Box 5.1 – J.M. Butler (2005) Forensic DNA Typing, 2nd Ed.

- Theoretically, there are 256 possible motifs for tetranucleotide repeats (e.g., GATA). However, because STRs are tandemly repeated, some motifs are actually equivalent to others.

- Two rules can be used to identify whether motif A is equivalent to motif B. Motif A is considered equivalent to motif B when (1) motif A is inversely complementary to motif B or (2) motif A is different from motif B or the inversely complementary sequence of motif B by frameshift.


http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Nomenclature Issues and the Y-Chromosome
J.M. Butler (Genetic Genealogy Conference, Houston, TX) October 20, 2007

Categories for STR Markers

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Repeat Structure</th>
<th>13 CODIS Loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple repeats – contain units of identical length and sequence</td>
<td>(GATA)(GATA)(GATA)</td>
<td>TPOX, CSF1PO, D5S818, D13S317, D16S539</td>
</tr>
<tr>
<td>Simple repeats with non-consensus alleles (e.g., TH01 9.3)</td>
<td>(GATA)(GATA)(GACA)</td>
<td>TH01, D18S51, D7S820</td>
</tr>
<tr>
<td>Compound repeats – comprise two or more adjacent simple repeats</td>
<td>(GATA)(GACA)(CA)(CATA)</td>
<td>VWA, FGA, D3S1358, D8S1179</td>
</tr>
<tr>
<td>Complex repeats – contain several repeat blocks of variable unit length</td>
<td>(GATA)(GACA)(CA)(CATA)</td>
<td>D21S11</td>
</tr>
</tbody>
</table>

These categories were first described by Urquhart et al. (1994) Int. J. Legal Med. 107:13-20

STR Repeat Nomenclature Rules


- For sequences within genes, use the coding strand
- For other sequences, select the first GenBank database entry or original literature description
- Define the repeat sequence which will provide the largest number of consecutive repeats (as far as possible on the 5'end)
- If two sequences are repeated (i.e., compound repeats), include both motifs in determining the repeat number
- Microvariants: should be designated by the number of complete repeats and the number of base pairs of the partial repeat separated by a decimal point (Int. J. Legal Med. 1994, 107:159-160) e.g. TH01 allele 9.3

Microvariant Alleles

- Defined as alleles that are not exact multiples of the basic repeat motif or sequence variants of the repeat motif or both
- Alleles with partial repeat units are designated by the number of full repeats and then a decimal point followed by the number of bases in the partial repeat
- Example: TH01 9.3 allele: [TCAT]4 -CAT [TCAT]5

Use of “TAGA” vs “GATA” results in a single repeat difference (Y-GATA/H4)

Reference sequence: GenBank accession G42676 (submitted May 1999 by White et al.)

NIST SRM 2395 follows ISFG guidelines (for our primer pair): first adjacent repeat starting from 5'end in TAGA

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Nomenclature Issues and the Y-Chromosome
J.M. Butler (Genetic Genealogy Conference, Houston, TX) October 20, 2007

DYS714 Reference Sequence

Alternate: \([\text{TTCCT}]_n[\text{CTTCT}]_2[\text{TTCCT}]_2[\text{CTTCT}]_2[\text{TTCCT}]_2\]

Comment: would add 8 repeats

\(19 + 8 = 27\)

ISFG Guidelines for Y STR Alleles

• Locus nomenclature should be DYS number if possible
• Allelic ladders should be used
• Allele nomenclature discussed…

ISFG Guidelines for Y STR Allele Nomenclature


• Number of complete repeats
• A partial repeat (variant allele) is designated by number of complete repeats separated by a dot followed by the number of bases in the incomplete repeat (e.g., 17.3)
• Some locus nomenclatures take into account the total number of repeat units (non-variant plus variant) while others have taken into account only the variable repetitive stretches
  - “If 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 repeat units of the DNA, that comprise both variant and non-variant repeats”
• Duplicated systems such as DYS385 have to be treated as genotypes and alleles should be separated by a hyphen (e.g., “11-14”)

DYS389I/II

Original paper (Kayser et al. (1997) Int. J. Legal Med. 110:141-149) defines allele nomenclature without repeat segment “C”; it has now been added in more recent nomenclatures thus making alleles +3 repeats larger
Nomenclature Issues and the Y-Chromosome

J.M. Butler (Genetic Genealogy Conference, Houston, TX) October 20, 2007

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Nomenclature Issues and the Y-Chromosome
J.M. Butler (Genetic Genealogy Conference, Houston, TX) October 20, 2007

A Nomenclature Comparison Chart

<table>
<thead>
<tr>
<th>Marker</th>
<th>DNA Fragment</th>
<th>DNA Length</th>
<th>Family</th>
<th>Type</th>
<th>Geographic</th>
<th>Defined</th>
<th>Allelic</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYS385</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>DYS388</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>DYS389</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>DYS390</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>DYS391</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>DYS392</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>DYS393</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>DYS394</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>?</td>
</tr>
<tr>
<td>DYS395</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>?</td>
<td>a</td>
<td>?</td>
</tr>
</tbody>
</table>

Note: Differences in standards are due to different methods of counting STR values.

Attempts to Resolve Nomenclature Issues


Yfiler Kit H4 Nomenclature Resolution

- Establishing a consensus nomenclature can facilitate data comparison for proficiency testing, quality assurance, and casework results. Efforts into nomenclature standardization should be supported and lauded…

- There are differences in allele designations at the GATA H4 marker between those recommended in the Applied Biosystems AmpFISTR Yfiler™ polymerase chain reaction amplification kit (Applied Biosystems, Foster City, CA) and the ISFG recommendations. The nomenclature for the GATA H4 marker in the Yfiler kit is based on the allele repeat structure defined by the National Institute of Standards and Technology Standard reference material (SRM) 2395 and the work of Butler et al.

- Those who choose to follow the allele nomenclature recommendations of the ISFG Commission should add a correction factor of nine to the Yfiler allele number, and they should refer to this marker as GATA H4.1.


ISFG DNA Commission

- DNA polymorphisms (1989)
- PCR based polymorphisms (1992)
- Naming variant alleles (1994)
- Repeat nomenclature (1997)
- Mitochondrial DNA (2000)
- Y-STR use in forensic analysis (2001)
- Mixture Interpretation (2006)
- Disaster Victim Identification (2007)

International Society of Forensic Genetics
http://www.isfg.org/

- An international organization responsible for the promotion of scientific knowledge in the field of genetic markers analyzed with forensic purposes.

- Founded in 1968 and represents more than 800 members from over 50 countries.

- A DNA Commission regularly offers recommendations on forensic genetic analysis.

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Nomenclature Issues and the Y-Chromosome
J.M. Butler (Genetic Genealogy Conference, Houston, TX) October 20, 2007

ISFG DNA Commission on Y-STRs

DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis


Subject Matter Experts

Leonor Gusmão (Portugal)
John Butler (USA)
Peter Gill (UK)
Manfred Kayser (Netherlands)
Lutz Roewer (Germany)
Chris Tyler-Smith (UK)

ISFG Board Members

Angel Carracedo (Spain)
Wolfgang Mayr (Austria)
Niels Morling (Denmark)
Mecki Prinz (USA)
Peter Schneider (Germany)

Nomenclature Covered by ISFG (2006)

- Nomenclature of 11 core Y-STRs (Table 1)
  - Some nomenclature is historical and not optimal
- “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…”
- Coverage of 63 additional loci that were known at the time (Table 2)

Genetic genealogy has gone beyond these defined loci...

NIST Standard Reference Materials (SRMs)

Y-Chromosome Standard NIST SRM 2395

Human Y-Chromosome DNA Profiling Standard

- 5 male samples + 1 female sample (neg. control)
- 100 ng of each (50 µL at ~2 ng/µL)
- 22 Y STR markers sequenced
- 9 additional Y STR markers typed
- 42 Y SNPs typed with Marligen kit

Certified for all loci in commercial Y-STR kits:

- Y-PLEX 6
- Y-PLEX 5
- Y-PLEX 12
- PowerPlex Y
- Y-filer

Sequence Determination of Y STR Repeat Region for Each Component

5 male components in SRM 2395 have 5 different Y-SNP backgrounds: R1b, J2, E3a, G, and I

Y-SNP Results on SRM 2395

from Marligen Signet™ Multiplexes (Luminex bead assay)

42 Y-SNPs measured across all samples

SRM 2395

<table>
<thead>
<tr>
<th>Component</th>
<th>Component A</th>
<th>Component B</th>
<th>Component C</th>
<th>Component D</th>
<th>Component E</th>
<th>Component F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A/AG</td>
<td>A/AC</td>
<td>A/AC</td>
<td>A/AT</td>
<td>A/AC</td>
<td>A/TT</td>
</tr>
<tr>
<td>Component A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>Component B</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>Component C</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>Component D</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>Component E</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>Component F</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

SRM components are all distinguishable from one another with these Y SNPs

Y-SNP Results on SRM 2395

5 male components in SRM 2395 have 5 different Y-SNP backgrounds: R1b, J2, E3a, G, and I

Sequencing Performed

DYS392 (forward) A: 13 TAT repeats
DYS392 (forward) B: 11 TAT repeats
DYS392 (forward) C: 11 TAT repeats
DYS392 (forward) D: 11 TAT repeats
DYS392 (forward) E: 12 TAT repeats

SWGDAM recommended loci:

DYS19, DYS385 a/b, DYS389 I/II, DYS390, DYS391, DYS392, DYS393, DYS438, DYS439

Helps meet FBI Standard 9.5 (and ISO 17025)...traceability to a national standard

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Nomenclature Issues and the Y-Chromosome
J.M. Butler (Genetic Genealogy Conference, Houston, TX) October 20, 2007

Sequencing Individual DYS385 Alleles

Sequence Summaries for SRM 2395

Y-STR Kits Used in Forensic Testing

Y-STR Types Being Added to NIST SRM 2395 Certificate

Y-STR Loci in Use for Genetic Genealogy

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Nomenclature Issues and the Y-Chromosome
J.M. Butler (Genetic Genealogy Conference, Houston, TX) October 20, 2007

67 Y-STR Loci in Use by FamilyTree DNA

| Loci in bold and underlined have certified values in current NIST SRM 2395 |

Family Tree DNA (12, 37, or 67 loci) – DYS19, DYS385 a/b, DYS388, DYS389I, DYS390, DYS391, DYS392, DYS393, DYS426, DYS439, DYS437, DYS447, DYS448, DYS449, DYS454, DYS455, DYS456, DYS459 a/b, DYS464 a/b/c/d, DYS438, DYS442, DYS443, DYS460, YCAII a/b, DYS456, DYS470, DYS471, DYS472, DYS481, DYS482, DYS490, DYS492, DYS511, DYS520, DYS531, DYS534, DYS537, DYS557, DYS565, DYS568, DYS572, DYS576, DYS590, DYS594, DYS617, DYS640, DYS641

43 Y-STR Loci in Use by Relative Genetics

| Loci in bold and underlined have certified values in current NIST SRM 2395 |


New Y-STR Loci Offer a Great View… but without the underlying support of standardization which takes time

Points to Keep in Mind…

- Initial selection of material (SRM components) was for a specific purpose usually and may not address every need in the future (a new locus may not exhibit a diverse set of alleles)
- The forensic community uses commercial STR typing kits – and only wants a confirmation of the allele calls against an allelic ladder
- Some duplicated Y-STR loci (e.g., DYS464) will not be able to have every allele sequenced
- There are lots of loci that could be “certified” – how do we decide which ones to include in future certificate updates?

STRBase Resources for Genetic Genealogists

http://www.cstl.nist.gov/biotech/strbase/

- PowerPoint slides providing background information on Y-chromosome loci and testing – http://www.cstl.nist.gov/biotech/strbase/YmtDNAworkshop.htm

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm