The Numbers behind DNA Analysis:
How do you get 1 in a trillion from only testing a few hundred people?

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

Lockheed Martin BEACON Lecture
Rockville, MD
August 19, 2009

Group Mission Statement

Advancing technology and traceability through quality genetic measurements to aid work in
- forensic DNA testing,
- clinical genetics,
- agricultural biotechnology, and
- DNA biometrics.

We are finding new ways to use DNA…
DNA Testing Requires a Reference Sample

A DNA profile by itself is fairly useless because it has no context...

DNA analysis for identity only works by comparison – you need a reference sample

Crime Scene Evidence compared to Suspect(s) (Forensic Case)
Child compared to Alleged Father (Paternity Case)
Victim’s Remains compared to Biological Relative (Mass Disaster ID)
Soldier’s Remains compared to Direct Reference Sample (Armed Forces ID)

DNA Profile
Rarity estimate
of the specific
DNA profile

Elements Going into the Calculation of a Rarity Estimate for a DNA Sample

1
Population allele frequencies

2
DNA Profile (with specific alleles)

3
Appropriate genetic formulas

There are different ways to express the profile rarity

Generating a DNA Profile

DNA Profile (with specific alleles)

Population allele frequencies

Rarity estimate of the specific DNA profile

Appropriate genetic formulas

Table 11.3 Random match probability for a 13-loci STR profile using the U.S. Caucasian allele frequencies listed in Table 11.1.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allele 1</th>
<th>Allele 2</th>
<th>Race 1</th>
<th>Race 2</th>
<th>Formula</th>
<th>Expected Genotype Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1S8099</td>
<td>91</td>
<td>11</td>
<td>0.13206</td>
<td>0.02719</td>
<td>z = 0.20347</td>
<td>0.2880</td>
</tr>
<tr>
<td>D8S1179</td>
<td>12</td>
<td>14</td>
<td>0.27459</td>
<td>0.10850</td>
<td>z = 0.08405</td>
<td>0.1128</td>
</tr>
<tr>
<td>D3S1358</td>
<td>9</td>
<td>9</td>
<td>0.161</td>
<td>0.161</td>
<td>z = 0.074</td>
<td>1 in 837 trillion</td>
</tr>
<tr>
<td>D16S539</td>
<td>9</td>
<td>9</td>
<td>0.15542</td>
<td>0.25564</td>
<td>z = 0.2903</td>
<td>0.0614</td>
</tr>
<tr>
<td>D13S317</td>
<td>13</td>
<td>13</td>
<td>0.1181</td>
<td>0.1181</td>
<td>z = 0.057</td>
<td>0.0470</td>
</tr>
<tr>
<td>D19S433</td>
<td>10</td>
<td>10</td>
<td>0.197</td>
<td>0.197</td>
<td>z = 0.090</td>
<td>0.037</td>
</tr>
<tr>
<td>D21S11</td>
<td>21</td>
<td>21</td>
<td>0.197</td>
<td>0.197</td>
<td>z = 0.090</td>
<td>0.037</td>
</tr>
<tr>
<td>D7S820</td>
<td>9</td>
<td>9</td>
<td>0.15744</td>
<td>0.15744</td>
<td>z = 0.057</td>
<td>0.0470</td>
</tr>
<tr>
<td>D2S1338</td>
<td>9</td>
<td>9</td>
<td>0.15744</td>
<td>0.15744</td>
<td>z = 0.057</td>
<td>0.0470</td>
</tr>
<tr>
<td>D18S51</td>
<td>11</td>
<td>11</td>
<td>0.13206</td>
<td>0.02719</td>
<td>z = 0.20347</td>
<td>0.2880</td>
</tr>
<tr>
<td>D19S433</td>
<td>11</td>
<td>11</td>
<td>0.13206</td>
<td>0.02719</td>
<td>z = 0.20347</td>
<td>0.2880</td>
</tr>
<tr>
<td>D13S317</td>
<td>13</td>
<td>13</td>
<td>0.1181</td>
<td>0.1181</td>
<td>z = 0.057</td>
<td>0.0470</td>
</tr>
<tr>
<td>D16S539</td>
<td>9</td>
<td>9</td>
<td>0.161</td>
<td>0.161</td>
<td>z = 0.074</td>
<td>1 in 837 trillion</td>
</tr>
</tbody>
</table>

1 Random match probability for a 13-loci STR profile using the U.S. Caucasian allele frequencies listed in Table 11.1.

DNA in the Cell

The vast majority of DNA is the same from person to person

22 pairs + XX or XY

Only a Small Varying Region is Targeted and Probed for Each DNA Marker Examined

Organization of Information

Printed

Volumes in a Set of Encyclopedias

Genetic

23 Pairs of Chromosomes in a Cell

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

You know that no two people share the same fingerprint, but did you know that the cells that make up your body also have a unique fingerprint unlike anyone else’s? Your cells contain a complex molecule that we call DNA. Unless you have an identical twin, no one else has DNA just like yours.

Scientists can analyze DNA. If a criminal leaves DNA at a crime scene, police can use it to prove who committed the crime. At NIST, we help crime labs analyze DNA accurately. We make DNA standards so crime labs can tell if their results are right.

DNA Storage is by the order of nucleotides, genes and chromosomes

Text Storage is by the order of letters, words and paragraphs

Identification of Information

Printed Information
- Library
- Book
- Chapter
- Page Number

Genetic Information
- Body
- Cell
- Nucleus
- Chromosome
- Locus (part of chromosome)

Characteristics of DNA

- Each person has a unique DNA profile (except identical twins).
- Each person’s DNA is the same in every cell.
- An individual’s DNA profile remains the same throughout life.
- Half of your DNA comes from your mother and half from your father.

The Human DNA Genome within a Cell

- Nuclear DNA (3.2 billion bp)
  - Inherited from only your mother
- Mitochondrial DNA (16,569 bp)
  - Inherited from only your mother
  - Inherited from both your mother and your father

Genetic Inheritance

- Father’s Sperm
- Mother’s Egg

- Father contributes: 22 autosomes (1 of each pair), X or Y
- Mother contributes: 22 autosomes (1 of each pair), X and mtDNA

Human Genome and Inheritance

- Nuclear DNA
- Mitochondrial DNA

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Basis of DNA Profiling

The genome of each individual is unique (with the exception of identical twins) and is inherited from parents.

Probe subsets of genetic variation in order to differentiate between individuals (statistical probabilities of a random match are used).

DNA typing must be performed efficiently and reproducibly (information must hold up in court).

Current standard DNA tests DO NOT look at genes – little/no information about race, predisposal to disease, or phenotypical information (eye color, height, hair color) is obtained.

Short Tandem Repeat (STR) Markers

An accordion-like DNA sequence that occurs between genes:

- The FBI has selected 13 core STR loci that must be run in all DNA tests in order to provide a common currency with DNA profiles.

DNA Marker Nomenclature

TH01
Tyrosine Hydoxylase gene, intron 01

D16S539
D: DNA
16: chromosome 16
S: single copy sequence
539: 539th locus described on chromosome 16
Short Tandem Repeat (STR) Markers

PCR primers anneal to unique sequences bracketing the variable STR repeat region. The overall PCR product size is measured, the DNA template containing the STR marker is amplified, and a fluorescent dye is added to the PCR mixture. The PCR product is separated (based on size) and detected on a gel or capillary electrophoresis instrument. A DNA profile is produced by separating DNA molecules by size and dye color. A sample’s profile is compared to the allelic ladder to determine the number of repeats for each allele.

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

Individuals will differ from one another in terms of their STR profile. STR genotype can then be put into an alpha numeric form for search on a DNA database.

STR Results

- Individuals will differ from one another in terms of their STR profile.
- STR genotype can then be put into an alpha numeric form for search on a DNA database.

<table>
<thead>
<tr>
<th>Locus</th>
<th>AMEL</th>
<th>D8S1179</th>
<th>D21S11</th>
<th>D18S51</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual #1</td>
<td>X,Y 11,13</td>
<td>28,32,2</td>
<td>17,18</td>
<td></td>
</tr>
<tr>
<td>Individual #2</td>
<td>XX 11,14</td>
<td>30,31 12,15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What would be entered into a DNA database for searching:

- AMEL
- D8S1179
- D21S11
- D18S51

PATERNITY TESTING

A DNA Profile is Produced by Separating DNA Molecules by Size and Dye Color

The labeled fragments are separated (based on size) and detected on a gel or capillary electrophoresis instrument. A DNA profile is produced by separating DNA molecules by size and dye color. The labeled fragments are separated (based on size) and detected on a gel or capillary electrophoresis instrument. A DNA profile is produced by separating DNA molecules by size and dye color.

Paternity testing is used to determine if a parent is the biological father of a child. Alleged Father(s) is asked to donate DNA sample.
Population Allele Frequencies

DNA Profile frequencies

Rarity estimate of the specific DNA profile

Appropriate genetic formulas

Decide on Number of Samples and Ethnic/Racial Grouping

Usually = 100 per group

Gather Samples

Often anonymous samples from a blood bank

Analyze Samples at Desired Genetic Loci

Summarize DNA types

Determine Allele Frequencies for Each Locus

Perform Statistical Tests on Data

Hardy-Weinberg equilibrium for allele independence

Linkage equilibrium for locus independence

Examination of genetic distance between populations

Use Database(s) to Estimate an Observed DNA Profile Frequency

A Computer Program Used to Perform Statistical Analysis of STR Allele Frequencies

Figure 20.2, J.M. Butler (2005)

Comparison of Allele Frequencies Measured with Different Studies

<table>
<thead>
<tr>
<th>D13S317</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>Frequency</td>
</tr>
<tr>
<td>7</td>
<td>0.0001</td>
</tr>
<tr>
<td>8</td>
<td>0.0290</td>
</tr>
<tr>
<td>9</td>
<td>0.0300</td>
</tr>
<tr>
<td>10</td>
<td>0.0076</td>
</tr>
<tr>
<td>11</td>
<td>0.0340</td>
</tr>
<tr>
<td>12</td>
<td>0.0400</td>
</tr>
<tr>
<td>13</td>
<td>0.0520</td>
</tr>
<tr>
<td>14</td>
<td>0.0480</td>
</tr>
<tr>
<td>15</td>
<td>0.0960</td>
</tr>
</tbody>
</table>

Minimum allele frequency = 0.0003

30 times more samples in the larger study yet the allele frequencies are fairly similar.

Comparison of Allele Frequencies Between Different Population Groups

D13S317 Allele Frequencies from NIST U.S. Population Data

<table>
<thead>
<tr>
<th>Allele</th>
<th>Caucasian</th>
<th>African-American</th>
<th>Hispanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.11258</td>
<td>0.0295</td>
<td>0.1218</td>
</tr>
<tr>
<td>9</td>
<td>0.09749</td>
<td>0.0215</td>
<td>0.1258</td>
</tr>
<tr>
<td>10</td>
<td>0.01512</td>
<td>0.0125</td>
<td>0.1600</td>
</tr>
<tr>
<td>11</td>
<td>0.03940</td>
<td>0.0260</td>
<td>0.1257</td>
</tr>
<tr>
<td>12</td>
<td>0.04851</td>
<td>0.0164</td>
<td>0.2113</td>
</tr>
<tr>
<td>13</td>
<td>0.13447</td>
<td>0.1455</td>
<td>0.1179</td>
</tr>
<tr>
<td>14</td>
<td>0.04801</td>
<td>0.0348</td>
<td>0.0648</td>
</tr>
<tr>
<td>15</td>
<td>0.00002</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Minimum Allele Frequency = 5/(2N)

Data behind FBI PopStats Program


Brendan Budowle, Ph.D., Brendan Sheu, M.S.; Stephen NCIgelski, M.B.A.; and Ramji Chintalapudi, Ph.D.

CODIS STR Loci Data from 41 Sample Populations*

There was little evidence for departures from Hardy-Weinberg expectations (HWE) in any of the populations.

The Fst estimates over all thirteen STR loci are 0.0006 for African Americans, 0.0005 for Caucasians, 0.0021 for Hispanics, 0.0039 for Asians, and 0.0282 for Native Americans.

The Same 13 Locus STR Profile in Different Populations

1 in 837 trillion in U.S. Caucasian population (NIST)
1 in 2.46 quadrillion (10^15) in U.S. Caucasian population (FBI)*
1 in 1.86 quadrillion (10^15) in Canadian Caucasian population*
1 in 16.6 quadrillion (10^15) in African American population (NIST)
1 in 17.6 quadrillion (10^15) in African American population (FBI)*
1 in 18.0 quadrillion (10^15) in U.S. Hispanic population (NIST)

These values are for unrelated individuals assuming no population substructure (using only p^2 and 2pq).


*http://www.cfsa.ca/pplus/profiler.htm

STR Cumulative Profile Frequency with Multiple Population Databases

<table>
<thead>
<tr>
<th>STR Locus</th>
<th>Profile Count</th>
<th>Number of Populations Used</th>
<th>Cumulative Profile Frequency Range (1E^x)</th>
<th>Cumulative Profile Frequency against U.S. Caucasians (Adjusted R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3S1358</td>
<td>16, 17</td>
<td>164</td>
<td>5.34E+10 to 6.3E+10</td>
<td>9.1E+10</td>
</tr>
<tr>
<td>VV4</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>M6A</td>
<td>12, 13</td>
<td>164</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D17S128</td>
<td>13, 14</td>
<td>164</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D20S800</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D21S11</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D30S129</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D31S13</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D31S14</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D4S221</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D7S1829</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D10S95</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
<tr>
<td>D16S247</td>
<td>15, 16</td>
<td>146</td>
<td>2.46E+10 to 5.8E+10</td>
<td>10.1E+10</td>
</tr>
</tbody>
</table>

10^14 to 10^21
How Statistical Calculations are Made

- Generate data with set(s) of samples from desired population group(s)
  - Generally only 100-150 samples are needed to obtain reliable allele frequency estimates
- Determine allele frequencies at each locus
  - Count number of each allele seen
- Allele frequency information is used to estimate the rarity of a particular DNA profile
  - Homozygotes (p²), Heterozygotes (2pq)
  - Product rule used (multiply locus frequency estimates)

Applying Genetic Models and Formulas

DNA Profile (with specific alleles)
Rarity estimate of the specific DNA profile
Appropriate genetic formulas

Hardy–Weinberg Equilibrium (HWE)

- Godfrey Hardy (1877–1947) and Wilhelm Weinberg (1862–1937) both independently discovered the mathematics for independent assortment that is now associated with their names as the Hardy–Weinberg principle.
- HWE proportions of genotype frequencies can be reached in a single generation of random mating. HWE is simply a way to relate allele frequencies to genotype frequencies.

Punnett Square

A Three-Generation Family Pedigree with Genetic Results from a Single STR Marker (FGA)

The Second National Research Council Report (NRC II) Published in 1996

- Recommends various formulas to use to correct for inbreeding (subpopulation structure)
- Theta (θ) is a measure of the average level of co-ancestry (i.e., inbreeding)
  - Usually <0.01 with normal groups
  - Usually <0.03 with closed populations (e.g., Native American tribes)

Inbreeding means mating of two persons who are more closely related than if they were chosen at random” (NRC II, p. 98).
Comparison of STR Genotype Frequencies with Different Correction Factors

Table 11.2 Comparison of statistical treatment for homozygotes and heterozygotes under different circumstances.

<table>
<thead>
<tr>
<th>Correction Method</th>
<th>Unconditional (NCRI) Recommendation 4.1</th>
<th>Conditional with Substructure Adjustment</th>
<th>Conditional with Population Substructure Adjustments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homozygote</td>
<td>$p_0^2 = \frac{1}{1 + \frac{1}{n}}$</td>
<td>$p_0^2 = \frac{1}{1 + \frac{1}{n}}$</td>
<td>$p_0^2 = \frac{1}{1 + \frac{1}{n}}$</td>
</tr>
<tr>
<td>Heterozygote</td>
<td>$2p_0p_1 = \frac{2}{1 + \frac{1}{n}}$</td>
<td>$2p_0p_1 = \frac{2}{1 + \frac{1}{n}}$</td>
<td>$2p_0p_1 = \frac{2}{1 + \frac{1}{n}}$</td>
</tr>
</tbody>
</table>

Note: Allele frequency values ($p_j$, $p_k$) for the Table and Table 12 are sample data from Table 1.1 (U.S. Caucasoid). Note that if $n > 10^6$, their unconditional and conditional formulas can be the same (i.e., they are ‘exact’).

Example Calculations with Population Substructure Adjustments

Table 11.3 Example calculations with population substructure adjustment (as Appendix 10). Assume allelic data in Table 1.1 and $n = 10^6$.

How Are Such Large Numbers Generated with Random Match Probabilities?

- Each allele is sampled multiple times to produce a statistically stable allele frequency
- Using theoretical models from genetics, multiple loci are multiplied together to produce an estimate of the rarity of a particular DNA profile (combination of STR alleles based on individual allele frequencies)
- Remember that relatives will share genetic characteristics and thus have STR profiles that are more similar to one another than unrelated individuals
- We are not looking at every person on the planet nor are we looking at every nucleotide in the suspect's genome

Example Calculations with Corrections for Relatives

Example DNA Forensic Categories Typically Faced

- **Single Source**: DNA profile of the evidence sample providing indications of it being of a single source origin
- **Mixture of DNA**: Evidence sample DNA profile suggests it being a mixture of DNA from multiple (more than one) individuals
- **Kinship Determination**: Evidence sample DNA profile compared with that of one or more reference profiles is to be used to determine the validity of stated biological relatedness among individuals

Three DNA Forensic Categories

- **Exclusion** (no match)
  - **Known** (K) Sample
  - **Evidence** (E) Sample
- **Non-exclusion** – “Match” or “inclusion”
- **Inconclusive result** (no decision as there is insufficient data to support a conclusion)

The Three Possible Outcomes of Evidence Examination (Q-K Comparison)

Three DNA Forensic Categories

- **Single Source**: DNA profile of the evidence sample providing indications of it being of a single source origin
- **Mixture of DNA**: Evidence sample DNA profile suggests it being a mixture of DNA from multiple (more than one) individuals
- **Kinship Determination**: Evidence sample DNA profile compared with that of one or more reference profiles is to be used to determine the validity of stated biological relatedness among individuals


http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm
Inclusions (Matches) Require Statistics

- It would not be scientifically justifiable to speak of a match as proof of identity in the absence of underlying data that permit some reasonable estimate of how rare the matching characteristics actually are.

-- NRC II, p. 192

The Statistic (Determining the Weight of the Evidence) Should Be Calculated from the Evidence

### Evidence (partial profile):

<table>
<thead>
<tr>
<th>Type</th>
<th>Locus</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus 1</td>
<td>16,17</td>
<td>1 in 9</td>
</tr>
<tr>
<td>Locus 2</td>
<td>17,18</td>
<td>1 in 9</td>
</tr>
<tr>
<td>Locus 3</td>
<td>21,22</td>
<td>1 in 12</td>
</tr>
<tr>
<td>Locus 4</td>
<td>12,14</td>
<td>1 in 16</td>
</tr>
<tr>
<td>Locus 5</td>
<td>28,30</td>
<td>1 in 11</td>
</tr>
</tbody>
</table>

Product = 1 in 171,000

### Reference (full profile):

<table>
<thead>
<tr>
<th>Type</th>
<th>Locus</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus 1</td>
<td>16,17</td>
<td>1 in 9</td>
</tr>
<tr>
<td>Locus 2</td>
<td>17,18</td>
<td>1 in 9</td>
</tr>
<tr>
<td>Locus 3</td>
<td>21,22</td>
<td>1 in 12</td>
</tr>
<tr>
<td>Locus 4</td>
<td>12,14</td>
<td>1 in 16</td>
</tr>
<tr>
<td>Locus 5</td>
<td>28,30</td>
<td>1 in 11</td>
</tr>
<tr>
<td>Locus 6</td>
<td>14,16</td>
<td>1 in 26</td>
</tr>
<tr>
<td>Locus 7</td>
<td>12,13</td>
<td>1 in 9</td>
</tr>
<tr>
<td>Locus 8</td>
<td>11,14</td>
<td>1 in 31</td>
</tr>
<tr>
<td>Locus 9</td>
<td>9,9</td>
<td>1 in 32</td>
</tr>
<tr>
<td>Locus 10</td>
<td>9,11</td>
<td>1 in 14</td>
</tr>
<tr>
<td>Locus 11</td>
<td>6,6</td>
<td>1 in 19</td>
</tr>
<tr>
<td>Locus 12</td>
<td>8,8</td>
<td>1 in 3</td>
</tr>
<tr>
<td>Locus 13</td>
<td>10,10</td>
<td>1 in 21</td>
</tr>
</tbody>
</table>

Product = 1 in 665 trillion

The reference sample is still a "match" – just not as much information is available from the evidence for comparison.

Single Source vs. Mixture Samples

**Single Source Sample**

Locus 1: 16,16
Locus 2: 9,9
Locus 3: 8,12
Locus 4: 9,9
Locus 5: 17,19

One or two peaks observed at each locus (tested DNA region)

**Mixture Sample**

Locus 1: 16,16
Locus 2: 9,9
Locus 3: 8,12
Locus 4: 9,9
Locus 5: 17,19

More than two peaks observed at more than two loci (tested DNA regions)

Different possible combinations could have given rise to the particular mixture observed.

Thank you for your attention...

Our team publications and presentations are available at:

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

See also http://www.dna.gov/research/nist

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

john.butler@nist.gov