Research Activities in the Applied Genetics group

Peter M. Vallone, Ph.D.
Leader, Applied Genetics Group NIST
May 31, 2018
University of Copenhagen

The National Institute of Standards and Technology (NIST) was founded in 1901 and is now part of the U.S. Department of Commerce. NIST is one of the nation’s oldest physical science laboratories.

Congress established the agency to remove a major challenge to U.S. industrial competitiveness at the time—a second-rate measurement infrastructure that lagged behind the world.

Applied Genetics Group – Forensic & Clinical Genetics

Topics

• NIST Standard Reference Materials (SRMs)
  • 2372a (Human DNA Quantitation), 2391d (PCR-based Profiling)
• STRBase 2.0
• Rapid DNA Interlaboratory Study
• Ongoing Sequencing Projects

SRM 2372a - Human DNA Quantitation Standard

• On sale March 26, 2018
  • https://www.s.nist.gov/srmors/view_detail.cfm?srm=2372a
• Certified by digital PCR measurements

To be used as a dPCR calibrant
OR to assign a value to in house or commercial DNAs

Digital PCR Platforms at NIST

Fluidigm BioMark (2010)
Chamber digital
Bio-Rad QX200 (2012)
upgraded to QX200 (2014)
Roche (2010)
Bio-Rad AutoDG (2015)

Digital PCR
Partitioning of DNA targets into individual chambers or droplets

No need for a standard curve
Positives
(1 copy of the target)
Negatives
(0 copies of the target)
dPCR is counting accessible amplifiable targets
Factors that affect quantification

\[ \lambda = \ln(\frac{N_{\text{tot}}}{N_{\text{neg}}}) \]

\[ C = \lambda(FV) \]

- \( \lambda \) = number of targets per partition
- \( V \) = mean partition volume
- \( F \) = volume fraction of sample in the reaction mixture
- \( C \) = target concentration in a sample

Not an exhaustive list.

Quantification by dPCR

A method to count DNA copies

SRM 2372a includes the ratio of mitochondrial to nuclear haploid genomes

mtDNA/nDNA ratio for three mitochondrial quantification assays optimized for dPCR.

SRM 2372a provides the ratio of mtDNA to gDNA, which bridges the gap between well characterized mtDNA quantification assays and availability of a commercial standard.

Candidate SRM 2391d

PCR-based DNA profiling

- Successor to SRM 2391c
- Similar format – five tubes
  - A-C: three single source components
  - D: one mixture; approximately 3.3 (F:M)
  - E: one component: cells spotted on FTA paper (from cell lines)
- Components A-D are DNA extracted from blood (not cell lines)
- Certified allele calls for core STR loci
- Characterized by CE- and NGS-based methods

Autosomal STR Markers

- 24 Certified Autosomal STR Markers
- 1 Reference Autosomal STR Marker
- 15 Information Autosomal STR Markers
Y-STR Markers

Data Collection for Sample Screening: Y-STR

Information for additional marker systems
Support the adoption of new markers and technology platforms

- Mitochondrial genome sequence
- Identity SNPs – for degraded samples
- Ancestry SNPs – biogeographical ancestry prediction
- Phenotype SNPs – eye and hair color prediction

Data Collection for Sample Screening: SNPs

ForenSeq SNP Phenotype and Ancestry Estimation

Additional markers to be characterized:
- X-STRs, Indels, INNULS, other SNP Panels, and Microhaplotypes

Data Collection for Sample Screening: mtDNA

EMPOP results:
https://empop.online/haplotypes#matches_details

<table>
<thead>
<tr>
<th>Haplogroup</th>
<th>Ancestry</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1b1a12</td>
<td>African</td>
<td>unique</td>
</tr>
</tbody>
</table>
**STRBase 2.0**

- First round of development
  - STR fact sheets (for 24 loci)
  - Variant allele reporting
  - Goal is to have a beta site up this summer

- Provide search, sort, and download functionalities
- Automated submission of variant alleles
- Embedded viewer for STR sequence and presentations
- Other ideas or functionalities – let me know

**STR Fact sheet: example D1S1656**

*Visualization of the Sequence*

Identify surrounding sequence and provide observed SNPs

**Rapid DNA Assessment III**

- Summer 2018
- Core 20 STR markers
- Projected: 8 labs and 2 vendors
  - ANDE and IXI platforms *(new kits, configurations)*
- 20 samples per lab (single source swab)
  - Currently collecting single source swabs

*Supporting the use of Rapid DNA in the booking station*

**NIST Applied Genetics Group**

Rapid DNA Maturity Assessment 2018

To measure the status of rapid DNA typing technology for the 20 CODIS core loci
In support of lab and booking station Rapid DNA implementation

- Participants may choose one chemistry per 20 NIST provided swabs
- Additional packages may be requested
- NIST provides 20 reference buccal swabs to each participant
- NIST reports CODIS 20 success rate for all data combined (% success)

Data transferred back to NIST via electronic format
Sequencing Projects
- FGx and SS platforms
- 1036 population samples
- Highly polymorphic locus SE33
- STRSeq resource
- Nomenclature support
- Sensitivity studies
- Concordance projects

Sequencing Forensic STRs in Population Samples
When a match is made in a forensic case, allele frequencies are used to calculate how common or rare the DNA profile is in a given population

Example of length versus sequence-based frequency calculation:

<table>
<thead>
<tr>
<th>Allele</th>
<th>N</th>
<th>Pct</th>
<th>Sequence</th>
<th>N</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
<td>0.2%</td>
<td>ATCT170</td>
<td>1</td>
<td>0.2%</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>3.9%</td>
<td>ATCT175</td>
<td>1</td>
<td>0.2%</td>
</tr>
<tr>
<td>9</td>
<td>45</td>
<td>17.5%</td>
<td>ATCT190</td>
<td>10</td>
<td>11.8%</td>
</tr>
<tr>
<td>10</td>
<td>53</td>
<td>20.1%</td>
<td>ATCT180</td>
<td>53</td>
<td>53.3%</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>14.3%</td>
<td>ATCT176</td>
<td>21</td>
<td>13.9%</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>1.3%</td>
<td>ATCT152</td>
<td>2</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

Sequencing of 1036 NIST population samples
- Work performed on Illumina FGx – ForenSeq kit
- Allele calls were made with Illumina-UAS and STRait Razor and compared to CE length-based calls (high confidence)
- Will include flanking region variation (SNPs, indels)
- Purpose: provide sequence allele frequencies for U.S. Population groups
  - U.S.: Caucasian, African American, Hispanic, Asian
- The manuscript was submitted in late May
  - Focus on the autosomal loci

Allele frequencies of sequenced STR alleles are needed to formally apply this gain in information to generate population data!
Allele frequencies by length
Global n=1036

Move to the next dimension of the data:
Sequence
D12S391

Alleles by length are further resolved by sequencing

Allele "21" frequencies by sequence 10 "flavors"
Global n=1036

"Flavors of D12S391 - 21 allele"
1 [AGAT]10 [AGAC]10 AGAT
2 [AGAT]11 [AGAC]10 AGAT
3 [AGAT]11 [AGAC]9 AGAT
4 [AGAT]12 [AGAC]8 AGAT
5 [AGAT]12 [AGAC]7 AGAT
6 [AGAT]13 [AGAC]6 AGAT
7 [AGAT]13 [AGAC]5 AGAT
8 [AGAT]13 [AGAC]4 AGAT
9 [AGAT]14 [AGAC]3 AGAT
10 [AGAT]14 [AGAC]2 AGAT

Allele frequencies

TH01
Sequence
Length
10 alleles
8 alleles

Figure 5
Allelic Gains by Sequence compared to Gains in Heterozygosity

Gettings et al 2018 submitted
Counts Frequencies By population and “Global”

<table>
<thead>
<tr>
<th>Allele</th>
<th>Population</th>
<th>CE</th>
<th>NGS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Sequencing SE33

- Focus solely on SE33
- Required further bioinformatic curation of the FGx sequence files
- 53 alleles by length
- 238 allele by sequence
- Data supported by CE-based allele calls

SE33

- In the NIST Population Set (n=1036) CE data
- Observed length-based range 6.3 to 36 alleles
- 53 unique alleles by length

As the allele size increases... the depth of coverage decreases
As the distance between heterozygous allele increases... the allele coverage ratio decreases.

Typical allele "noise" for other STRs in the ForenSeq kit
Varies by locus and motif

Collaborative Effort (NIST, KCL, UNT, USC)
1786 + 1043 + 839 + 944 = 4612 samples

Example given for D12S391

STRSeq

• Bioproject hosted at NCBI to catalog unique STR alleles
• Fully annotated sequence of the STR amplicon
• Each STR allele will have an accession number

STRSeq

• A unique record for each observed allele
• CE-based allele call
• Genomic coordinates
• Platform/kit used to generate sequence

Nomenclature Support

Defining annotated STR reference sequences

Thoughts on reporting?
NGS – MPS – Sequencing noise

Thoughts

- How should sequence ‘noise’ be characterized?
- Coverage or normalized coverage?
- Evaluate locus specific performance
- Setting thresholds?
- Need to understand data first
- True allele versus noise
- for now: including artifacts and stutter as ‘noise’
- Zygosity

First steps
- Understand the data from sensitivity studies
- Develop tools to assess

Coverage of All Sequences on the D21S11 Locus

Understand the characteristics of single source DNA profiles generated by the NGS system by evaluating...

- Receiver Operating Characteristic (ROC) curves to define where alleles can be clearly separated from noise (attributed to either stutter or random causes)
- Zygosity to minimize the risks of misidentifying a heterozygote as a homozygote or a homozygote as heterozygote

Experiments

- Promega PowerSeq Auto 46GY
- Illumina MiSeq (v3 and 92 samples/run)
- Library preparation; TruSeq and Kapa
- Normalized and non-normalized

<table>
<thead>
<tr>
<th>Dilution Study (pg of genomic DNA)</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>125</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

In this study we evaluated reference samples ONLY
We did not test any unknown (crime-stain), mixtures or degraded samples.
**ROC-defined analytical threshold**

- Is a two dimensional chart which plot the **true positive** versus the **false positive** rates for a given parameter
- Is performed to determine which AT would lead to optimal levels of detection where error rates are minimized
- The true positive rate represents the proportion of true allele sequences known to be present at a specific locus/method/DNA amount
- The false positive rate generated at a locus of interest/method/DNA amount represents the proportion of noise sequences falsely classified as true sequences
- Assumption of a certain distribution is not required with ROCs
- Measures performance of different ATs and gives proportion of false positives and negatives.

**ROCs as a function of library preparation kit**

**ROCs as a function of DNA input amount**

**Inferring Zygosity from Heterozygote Balance**

Two primary risks associated with the process of allele designation:

- A heterozygote is inaccurately assigned as a homozygote due to:
  - Allele drop-out has occurred
  - Heterozygote imbalance has resulted in one of the alleles being interpreted as a stutter
- A homozygote is inaccurately called as a heterozygote due to:
  - A large stutter band is within range to be designated as an allele
  - A drop-in event occurs
Inferring Zygosity from Heterozygote Balance

A comparison of the distribution of the homozygotes and heterozygotes showed marked differences associated with the differences in the DNA input.

Further work

• Incorporate stutter and accountable artifacts into the ROCs
• Perform further sensitivity experiments as needed
• Create mock casework type samples/mixtures
  • Derive and test thresholds

Thank you for your attention! Questions?

Contact: Peter.Vallone@nist.gov

• Funding
  • NIST Special Programs Office: Forensic DNA
  • FBI Biometrics Center of Excellence: Forensic DNA Typing as a Biometric tool.
  • NII: STRSeq and Nomenclature
  • DHS S&T: Rapid DNA for Kinship

• Disclaimer – 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 Commerce or the Department of Justice. Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. The inclusion of such identification does not imply recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

• All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.