Research Activities in the Applied Genetics group

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

Applied Genetics Group – Forensic & Clinical Genetics

Peter Vallone (Group Leader), Becky Steffen, Erica Romsos, Katherine Gettings, Kevin Kiesler, Margaret Kline, Lisa Borsuk, Sarah Riman, David Duewer, Megan Cleveland

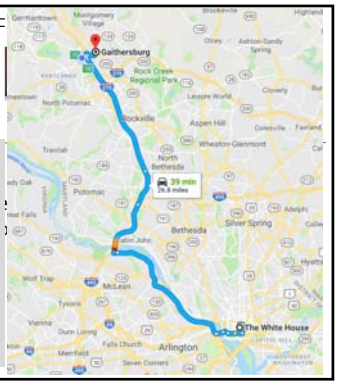
Topics
- NIST Standard Reference Materials (SRMs)
  - 2372a (Human DNA Quantitation), 2391d (PCR-based Profiling)
- STRBase 2.0
- Rapid DNA Interlaboratory Study
- ***Ongoing Sequencing Projects***

Applied Genetics Group – F

Peter Vallone (Group Leader), Becky Steffen, Erica Romsos, Katherine Gettings, Kevin Kiesler

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.



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

Table 1. Certified Values of Number and Mass Concentration for SRM 2372a[a]

The copy number values are metrologically traceable to the natural units count 1 and ratio 1 and International System of Units (SI) derived units of volume. The DNA mass concentration values are metrologically traceable to the natural units count and ratio 1 and SI derived units of mass and volume.

Component	Copy Number[b] (per nL)	DNA[c] (ng/µL)
A (red cap)	15.1 ± 1.5	49.8 ± 5.0
B (white cap)	17.5 ± 1.8	57.8 ± 5.8
C (blue cap)	14.5 ± 1.5	47.9 ± 4.8

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

Digital PCR Platforms at NIST

Fluidigm BioMark (2010)

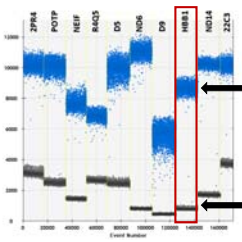
Bio-Rad QX100 (2012) upgraded to QX200 (2014)

Bio-Rad AutoDG (2015)

RainDance (2015)

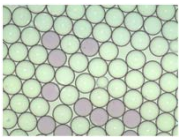
Digital PCR

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

$\lambda = -\ln(N_{neg}/N_{tot})$

$C = \lambda/(FV)$

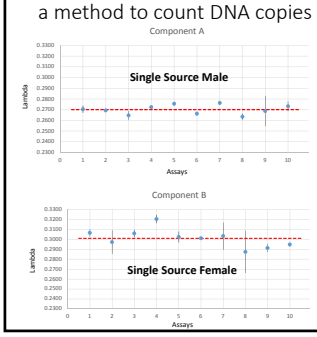
λ = number of targets per partition
V = mean partition volume
F = volume fraction of sample in the reaction mixture
C = target concentration in a sample

- Thermal cycler
- PCR assays – multiple targets
- Partition volume
- Treatment of artifacts
- Pipettes/balance
- Pre-treatment of DNA
- PCR master mix
- Primers/probes
- Lambda range

Not an exhaustive list

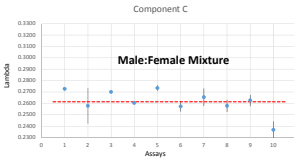
Quantification by dPCR

a method to count DNA copies



BIO-RAD QX200

Values assigned by 10 human custom genomic assays
Probing single copy targets in the human genome



All samples behave as expected with commercial qPCR assays (11 tested)

Converting copies per nanoliter to nanograms nuclear DNA per microliter

$$[\mathrm{nDNA}]\,\frac{\mathrm{ng}}{\mu\mathrm{L}} = \left(\frac{\lambda\ \text{copies of target}}{\text{droplet}}\right)\left(\frac{\mu\mathrm{L}\ \text{mixture}}{F\ \mu\mathrm{L}\ \text{sample}}\right)\left(\frac{\text{droplet}}{V\ \text{mixture}}\right)\left(\frac{\mathrm{HHGE}}{r\ \text{target}}\right)\left(\frac{n\ \text{base pairs}}{\mathrm{HHGE}}\right)\left(\frac{\overline{w}\ \mathrm{g}}{\text{mol base pairs}}\right)\left(\frac{\text{mol base pairs}}{6.022\ 10^{23}\ \text{base pairs}}\right)\left(\frac{10^3\ \mathrm{nL}}{\mu\mathrm{L}}\right)\left(\frac{10^9\ \mathrm{ng}}{\mathrm{g}}\right) \quad (1)$$

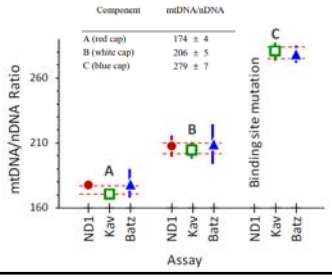
where r is the number of assay targets per human haploid genome equivalents (HHGE), n is the number of nucleotide base pairs (bp) per double-stranded HHGE, and $\overline{w}$ is the average molar mass of a bp in the DNA polymer.

For independent multiplicative factors such as these, the combined relative uncertainty of their product can be estimated from the square root of the sum-of-squares of the individual relative uncertainties [14, Section 5.1.6]:

$$\frac{u([\mathrm{nDNA}])}{[\mathrm{nDNA}]} = \sqrt{\left(\frac{u(\lambda)}{\lambda}\right)^2 + \left(\frac{u(F)}{F}\right)^2 + \left(\frac{u(V)}{V}\right)^2 + \left(\frac{u(r)}{r}\right)^2 + \left(\frac{u(n)}{n}\right)^2 + \left(\frac{u(\overline{w})}{\overline{w}}\right)^2} \quad (2)$$

***Duewer** DL, Kline MC, Romsos EL, Toman B. Anal Bioanal Chem. 2018 410:2879-2887*

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

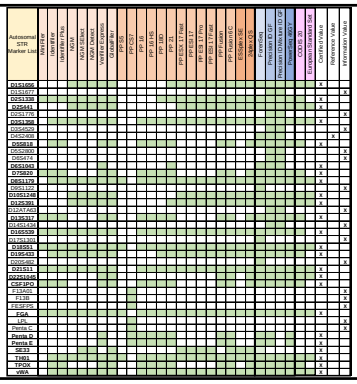
- Successor to SRM 2391c
- Similar format – five tubes
  - A-C three single source components
  - D one mixture; approximately 3:1 (F:M)
  - E one component: cells spotted on FTA paper (from cell lines)
- Components A-D are DNA extracted from blood (not cell lines)
- Certified allele calls for core STR loci
- Characterized by CE- and NGS-based methods

Supports the FBI Quality Assurance Standards

Autosomal STR Markers

Autosomal STR Markers
ThermoFisher CE STR kits
Promega CE STR kits
Qiagen Investigator CE STR kits
Illumina NGS kit
ThermoFisher NGS kits
Promega NGS kits
CODIS 20/ESS 12

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



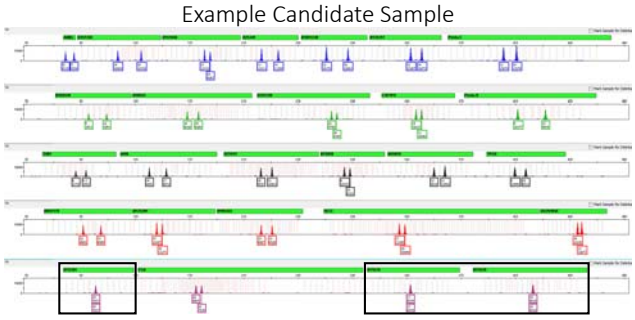
Y-STR Markers

Y-STR Markers
ThermoFisher CE STR kits
Promega CE STR kits
Qiagen Investigator CE STR kits
Illumina NGS kit
ThermoFisher NGS kits
Promega NGS kits

23 Certified Y-STR Markers
0 Reference Y-STR Markers
6 Information Y-STR Markers

Y-STR Marker List	GlobalFiler	Yfiler	Yfiler Plus	PP Fusion	PP Fusion 6C	PowerPlex Y23	Signex QS	ForenSeq	Precision ID GF	Precision ID Mixture ID GF	Rapidhelix	Certified Value	Reference Value	Information Value
DYS19												X		
DYS385a/b												X		
DYS389I/II												X		
DYS390												X		
DYS391												X		
DYS392												X		
DYS393												X		
DYS437												X		
DYS438												X		
DYS439												X		
DYS448												X		
DYS449														X
DYS460												X		
DYS456												X		
DYS458												X		
DYS481												X		
DYS505														X
DYS518														X
DYS522														X
DYS533												X		
DYS549												X		
DYS570												X		
DYS576												X		
DYS612														X
DYS627														X
DYS635												X		
DYS643												X		
Y-GATA-H4												X		
DYS3872I												X		

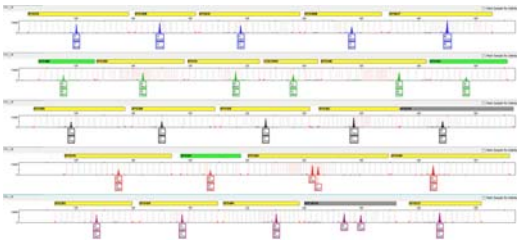
Data Collection for Sample Screening: Autosomal STR
Example Candidate Sample



Fully Heterozygous with PowerPlex Fusion 6C

*Y-STR Markers

Data Collection for Sample Screening: Y-STR



YFiler Plus Profile

YHRD: No matches in 188,209 Haplotypes
(using Minimal Haplotype)
https://yhrd.org

Whit Athey's Haplogroup
Predictor: E1b1a
http://www.hprg.com/hapest5/hapest5a/hapest5.htm?order=num

Results Table

Haplo-group	Fitness score	Proba-bility (%)
E1b1a	58	100.0
E1b1b	18	0.0
G2a	20	0.0
G2c	5	0.0
H	25	0.0
I1	7	0.0
I2a (xI2a1)	20	0.0
I2a1	3	0.0
I2b (xI2b1)	7	0.0
I2b1	14	0.0
J1	11	0.0
J2a1b	5	0.0
J2a1h	6	0.0
J2a1 x J2a1-M	11	0.0
J2b	9	0.0
L	13	0.0
N	2	0.0
Q	17	0.0
R1a	11	0.0
R1b	4	0.0
T	16	0.0

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

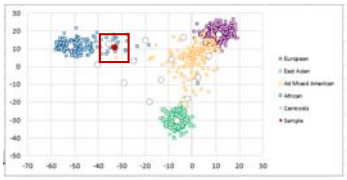
Data Collection for Sample Screening: SNPs

ForenSeq SNP Phenotype and Ancestry Estimation

Hair Color Results	
Brown	0.16
Red	0.00
Black	0.84
Blond	0.00

Eye Color Results	
Intermediate	0.00
Brown	1.00
Blue	0.00

Biogeographical Ancestry Results	
Distance to Nearest Centroid	3.36



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

Data Collection for Sample Screening: mtDNA

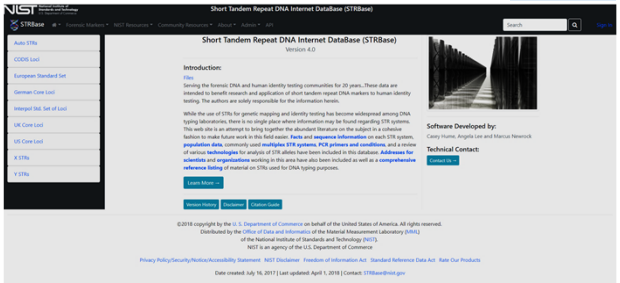
Illumina mtDNA Whole Genome Sequencing protocol with Nextera XT Sample Prep Kit

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

Haplogroup	Ancestry	Match
L1b1a12	African	unique

STRBase 2.0

Draft image



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

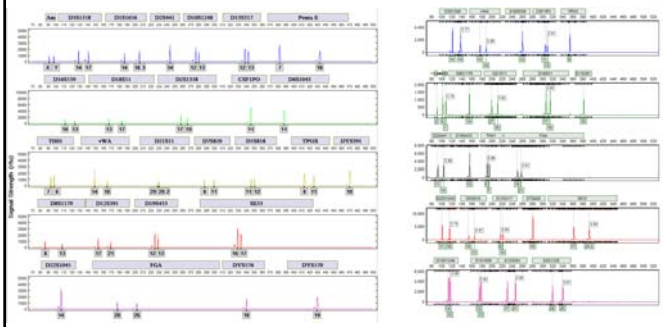
Viewer will link to NCBI Bioproject

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

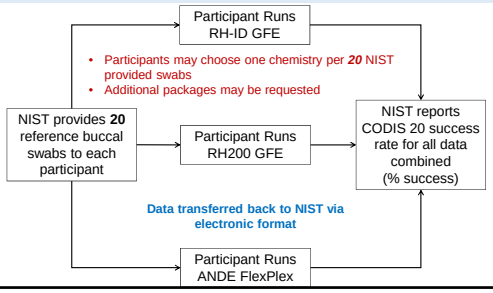
Supporting the use of Rapid DNA in the booking station

Profiles from ***high quality single source*** samples generated by Rapid DNA instruments



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



Sequencing Projects

- FGx and S5 platforms

- **1036 population samples**
- **Highly polymorphic locus SE33**
- **STRSeq resource**
- **Nomenclature support**
- **Sensitivity studies**
- Concordance projects

Applications for Sequencing STRs

- Targeted sequencing of STRs
  - STR motif sequence variation; flanking region variation
  - Further understand simple versus complex repeat motifs
  - Characterize stutter
- Applications
  - One to one matching?
  - With the new U.S. core loci we are already quite high (>10^{-20})
  - Partial profiles
  - Kinship
- Greater degree of multiplexing
  - Not confined by dye colors; smaller PCR amplicons (degraded samples)
  - PCR for sample enrichment
  - ***Still using PCR*** – stochastic effects, stutter
- Mixtures
  - Resolve alleles identical by length, but differ by sequence
  - Separate stutter from low level contributors (based on sequence)
  - A sequenced allele *may* have a lower frequency (higher RMP or LR)

Allele frequencies of sequenced STR alleles are needed to formally apply this gain in information -> **Generate population data!**

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:

D4S2408

Allele	N	Freq	Sequence Allele	N	Freq
7	1	0.6%	[ATCT]7	1	0.6%
8	23	14.4%	[ATCT]8	23	14.4%
9	60	37.5%	[ATCT]9	18	11.3%
			[ATCT] GTCT [ATCT]7	42	26.3%
10	53	33.1%	[ATCT]10	53	33.1%
11	21	13.1%	[ATCT]11	21	13.1%
12	2	1.3%	[ATCT]12	2	1.3%

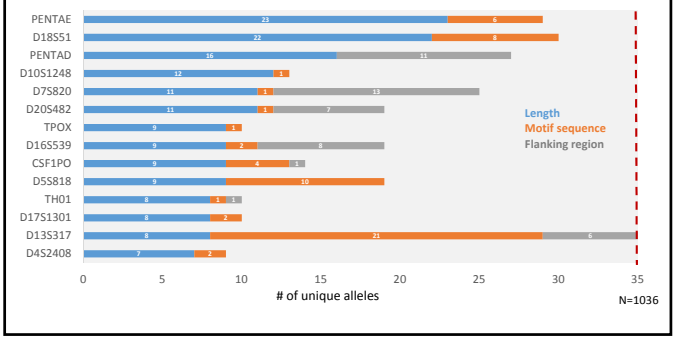
Length	Sequence
8,9	[ATCT]8, [ATCT]9
2pq	2pq
2*0.144*0.375	2*0.144*0.113
0.108	0.033
1 in 9.3	1 in 30.7

Sequencing of 1036 NIST population samples

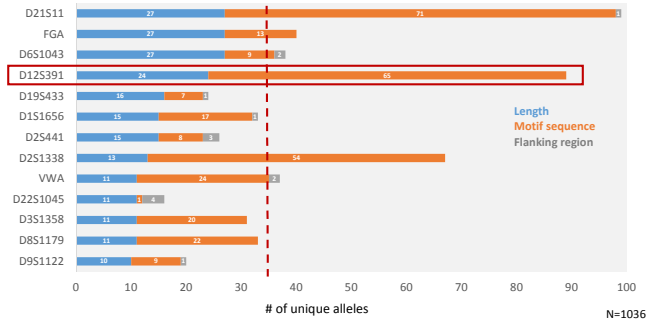
- Work performed on Illumina FGx – ForenSeq kit
- Allele calls were made with Ilumina-UAS and STRait Razor and compared to CE length-based calls (***high confidence***)
- Will include flanking region variation (SNPs, indels)
- Purpose: provide sequence allele frequencies for four U.S. Population groups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

Simple autosomal STRs

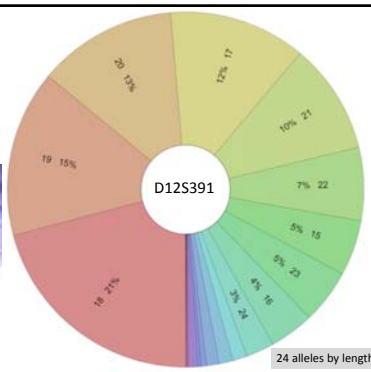
Alleles detected by sequencing



Compound/Complex autosomal STRs

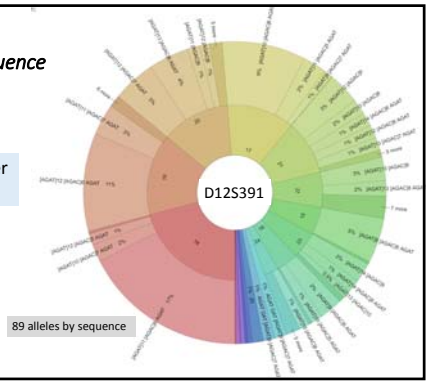


D12S391
Allele frequencies ***by length***
Global n=1036



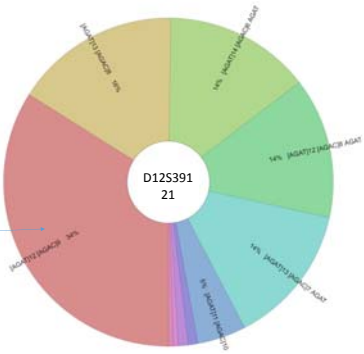
D12S391
Allele frequencies ***by sequence***
Global n=1036

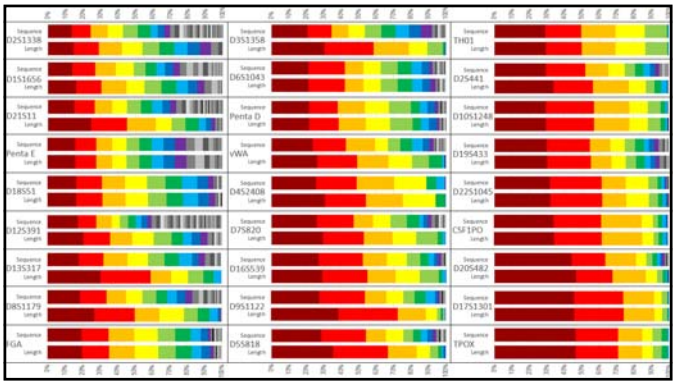
Alleles by length are further resolved by sequencing



D12S391
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
3	[AGAT]11 [AGAC]9 AGAT
4	[AGAT]12 [AGAC]8 AGAT
5	[AGAT]12 [AGAC]9
6	[AGAT]13 [AGAC]4 AGGC [AGAC]2 AGAT
7	[AGAT]13 [AGAC]7 AGAT
8	[AGAT]13 [AGAC]8
9	[AGAT]14 [AGAC]6 AGAT
10	[AGAT]14 [AGAC]7





Allele frequencies

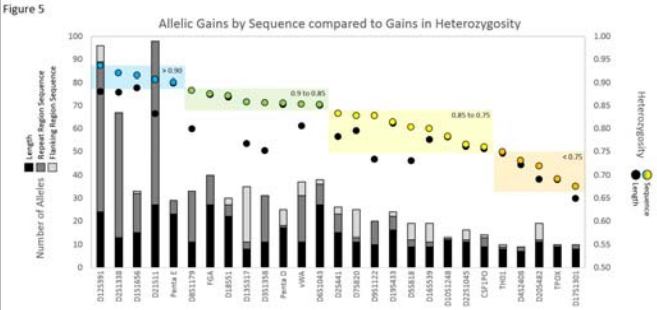
D12S391

Sequence	89 alleles
Length	24 alleles

TH01

Sequence	10 alleles
Length	8 alleles

Figure 5



Gettings et al 2018 submitted

Allele frequency tables
Example D7S820

			Counts					Frequencies				
Locus	Allele	Bracket	All	AfAm	Asian	Cauc	Hisp	All	AfAm	Asian	Cauc	Hisp
D7S820	6	[TATC]6	1	1	0	0	0	0.0005	0.0015	0.0000	0.0000	0.0000
D7S820	7	[TATC]7	25	3	0	17	5	0.0121	0.0044	0.0000	0.0235	0.0106
D7S820	7	[TATC]7	9	5	1	3	0	0.0043	0.0073	0.0052	0.0042	0.0000
D7S820	8	[TATC]8	197	80	2	74	41	0.0951	0.1170	0.0103	0.1025	0.0869
D7S820	8	[TATC]8	146	76	24	30	16	0.0705	0.1111	0.1237	0.0416	0.0339
D7S820	8.1	[TATC]8	1	0	0	1	0	0.0005	0.0000	0.0000	0.0014	0.0000
D7S820	9	[TATC]9	205	52	5	111	37	0.0989	0.0760	0.0258	0.1537	0.0784
D7S820	9	[TATC]9	45	27	4	10	4	0.0217	0.0395	0.0206	0.0139	0.0085
D7S820	9	[TATC]9	2	0	0	0	2	0.0010	0.0000	0.0000	0.0000	0.0042
D7S820	10	[TATC]10										118
D7S820	10	[TATC]10										191
D7S820	10	[TATC]10										064
D7S820	10	[TATC]9 TATT	1	1	0	0	0	0.0005	0.0015	0.0000	0.0000	0.0000
D7S820	10.3	[TATC]11	1	0	0	0	1	0.0005	0.0000	0.0000	0.0000	0.0021
D7S820	11	[TATC]11	443	130	60	132	121	0.2138	0.1901	0.3093	0.1828	0.2564
D7S820	11	[TATC]11	37	7	5	16	9	0.0179	0.0102	0.0258	0.0222	0.0191
D7S820	11	[TATC]11	5	2	3	0	0	0.0024	0.0029	0.0155	0.0000	0.0000
D7S820	11	[TATC]6 TGTC [TATC]4	1	0	0	0	1	0.0005	0.0000	0.0000	0.0000	0.0021
D7S820	12	[TATC]12	230	56	31	86	57	0.1110	0.0819	0.1598	0.1191	0.1208
D7S820	12	[TATC]12	50	3	2	29	16	0.0241	0.0044	0.0103	0.0402	0.0339
D7S820	12	[TATC]12	2	1	1	0	0	0.0010	0.0015	0.0052	0.0000	0.0000
D7S820	13	[TATC]13	49	9	5	21	14	0.0236	0.0132	0.0258	0.0291	0.0297
D7S820	13	[TATC]13	8	1	0	4	3	0.0039	0.0015	0.0000	0.0055	0.0064
D7S820	14	[TATC]14	2	1	0	1	0	0.0010	0.0015	0.0000	0.0014	0.0000
D7S820	14	[TATC]14	2	0	0	2	0	0.0010	0.0000	0.0000	0.0028	0.0000
			2072	684	194	722	472	1	1	1	1	1

Counts and frequencies for each sequenced allele
By population and "Global"

Further extension of table includes repeat region and 5'/3' flanking regions
rs numbers for identified SNPs and/or indels

		Full Sequence		
Bracket	Flanking Region Variants from GRCh38 (5' to 3')	5' Flank	Repeat Region	3' Flank
TATC]6	rs7789995	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]7	rs7789995	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]7	rs7789995, rs16887642	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATC	GTTAGTTCATTCTAAACTAT
TATC]8	rs7789995	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]8	rs7789995, rs16887642	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATC	GTTAGTTCATTCTAAACTAT
TATC]8	rs7789995, GRCh38 Chr7:84160204 insA, rs16887642	TAGTGAGATAAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATC	GTTAGTTCATTCTAAACTAT
TATC]9	rs7789995	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCATTCTAAACTAT
TATC]9	rs7789995, rs16887642	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCATTCTAAACTAT
TATC]9		TATTTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]10	rs7789995	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]10		TATTTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]10	rs7789995, rs16887642	TTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCATTCTAAACTAT
TATC]9 TATT	rs7789995	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATTT	GTTAGTTCGTTCTAAACTAT
TATC]11	GRCh38 Chr7:84160204 delA	TTAGTGAGAT-AAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]11	rs7789995	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]11		TATTTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]11	rs7789995, rs16887642	TTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCATTCTAAACTAT
TATC]6 TGTC [TATC]4	rs7789995	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTgTCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]12	rs7789995	TTAGTGAGATAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]12		TATTTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]12	rs7789995, rs16887642	TTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCATTCTAAACTAT
TATC]13	rs7789995	TTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]13		TATTTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]14	rs7789995	TTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT
TATC]14		TATTTAGTGAGATTAAAAAAACTATCAATCTGTC	TATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATC	GTTAGTTCGTTCTAAACTAT

Sequencing SE33

ELECTROPHORESIS

Research Article | Full Access

Sequence-based U.S. population data for the SE33 locus

Lisa A. Borsuk, Katherine B. Gettings, Carolyn R. Steffen, Kevin M. Kiesler, Peter M. Vallone

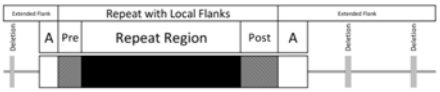
First published: 07 May 2018 | https://doi.org/10.1002/elps.201800091

- 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

	CE	NGS	CE	NGS
	N		H obs	
All	**53**	**238**	**0.935**	**0.965**
Cauc	**39**	**143**	**0.950**	**0.981**
Asian	**22**	**70**	**0.928**	**0.959**
AfAm	**41**	**141**	**0.927**	**0.956**
Hisp	**36**	**121**	**0.928**	**0.958**

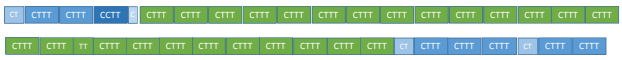
SE33

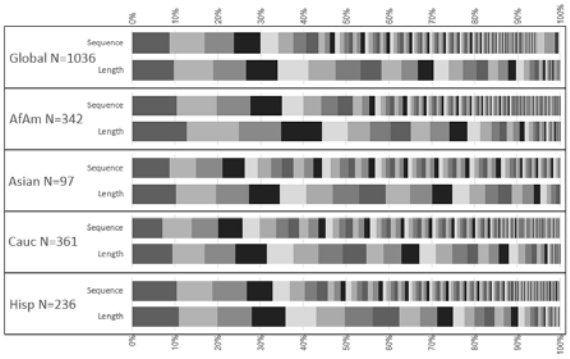


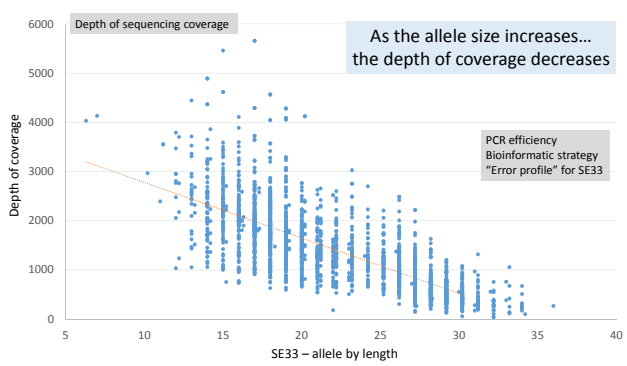
- In the NIST Population Set (n=1036) CE data
- Observed **length-based** range 6.3 to 36 alleles
- 53 unique alleles by length

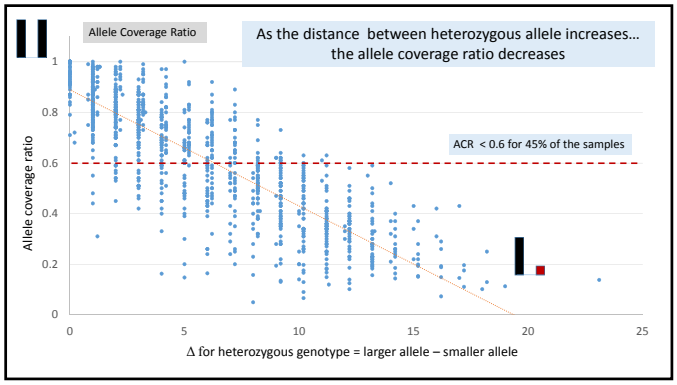
SE33 - 25.2 Allele from the Reference Genome Sequence

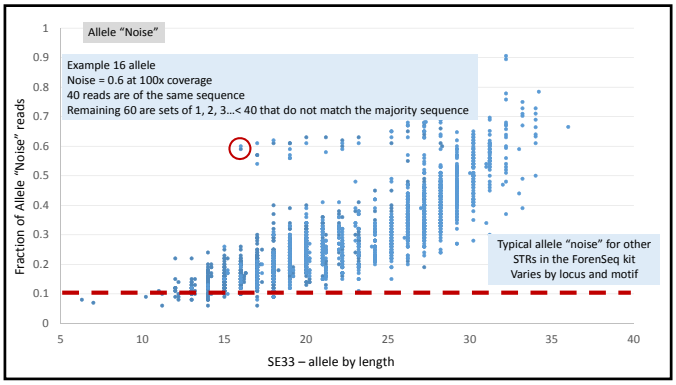
CT [CTTT]2 CCTT C **[CTTT]16 TT [CTTT]9** CT [CTTT]3 CT [CTTT]2











STRSeq

- Bioproject hosted at NCBI to catalog ***unique*** STR alleles
  - https://www.ncbi.nlm.nih.gov/bioproject/380127
- Fully annotated sequence of the STR amplicon
- Each STR allele will have an accession number

https://strider.online/

Collaborative Effort (NIST, KCL, UNT, USC)

1786 + 1043 + 839 + 944 = 4612 samples

Example given for D12S391

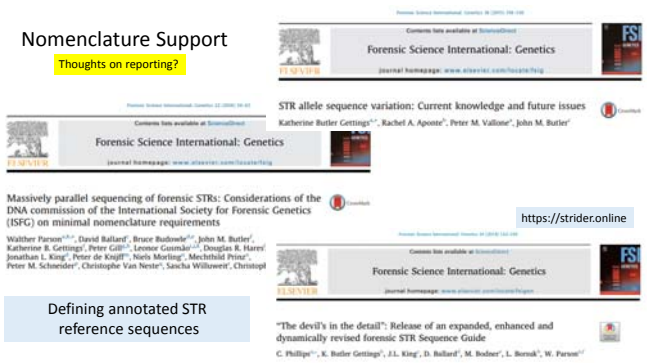
STRSeq

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

- Bioproject hosted at NCBI to cata
  - https://www.ncbi.nlm.nih.gov/biop
- Fully annotated sequence of the
- Each STR allele will have an acces

Nomenclature Support

Thoughts on reporting?



https://strider.online

Defining annotated STR reference sequences

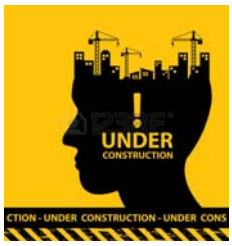
NGS – MPS – Sequencing noise

Thoughts

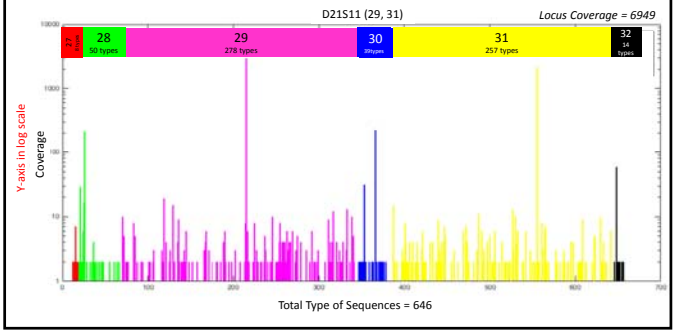
Sarah Riman
Hari Iyer (NIST statistical engineering division)

- 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'
- Zygosisty

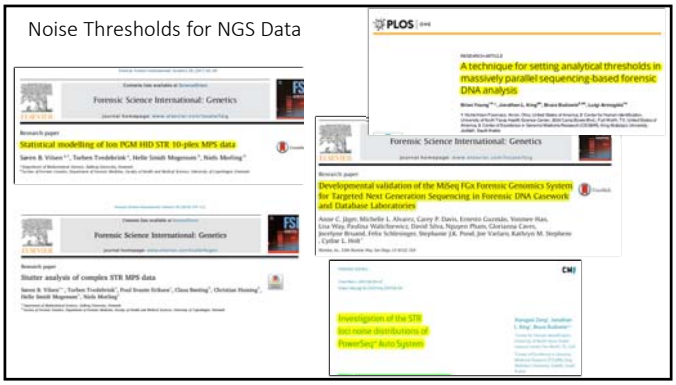
First steps
- Understand the data from sensitivity studies
- Develop tools to assess



Coverage of All Sequences on the D21S11 Locus



Noise Thresholds for NGS Data



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

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

Experiments

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

Dilution Study (pg of genomic DNA)

Sample 1	500	250	125	60	30	15
Sample 2	500	250	125	60	30	15
Sample 3	500	250	125	60	30	15

Single source samples – amplified in triplicate

PowerSeq 46GY: Loci and Amplicon Size

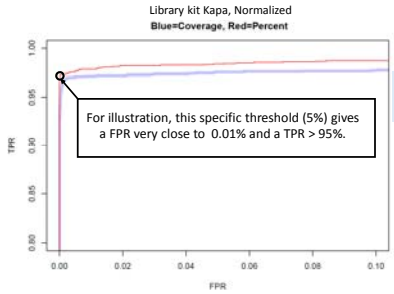
Locus Name	Size (bp)	Locus Name	Size (bp)
D8S1179	203–256	DYS391	147–178
D21S11	203–273	DYS19	168–204
D7S820	211–258	DYS385ab	202–303
CSF1PO	185–229	DYS389I/II	256–294
D3S1358	192–240	DYS390	204–248
TH01	220–254	DYS392	143–184
D13S317	209–257	DYS393	224–256
D16S539	198–246	DYS437	191–197
D2S1338	187–289	DYS438	203–242
D19S433	193–253	DYS439	204–224
vWA	202–262	DYS446	213–255
TPOX	196–244	DYS456	141–185
D18S51	190–277	DYS458	171–199
D5S818	191–239	DYS481	136–184
FGA	176–268	DYS533	242–284
Penta D	160–265	DYS549	199–230
Penta E	179–264	DYS570	157–217
Amelogenin	192, 198	DYS576	155–203
D12S391	161–208	DYS635	155–179
D2S441	188–204	DYS643	150–210
D10S1248	135–179	Y-GATA-H4	231–251
D22S1045	202–254		
D22S1045	129–178		
DYS391	147–178		

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.

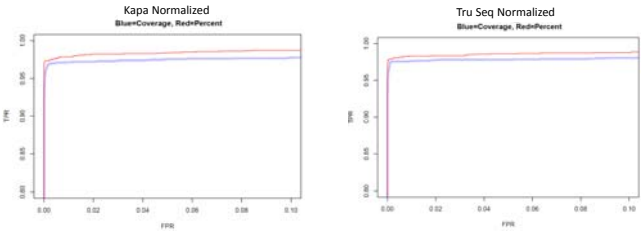
ROCs can be used to determine:
- A global analytical threshold
- A DNA amount-dependent threshold
- A DNA amount and locus-dependent threshold

Comparison of Coverage versus Percent ROC

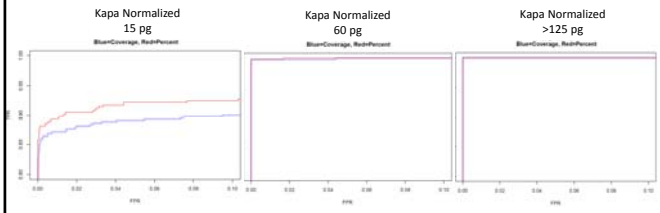


All DNA amounts, samples, replicates, and loci are collapsed into this plot

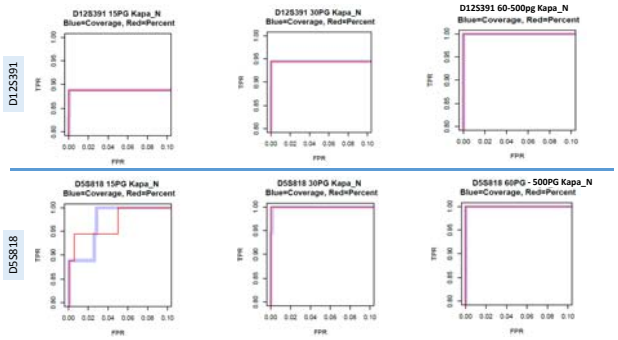
ROCs as a function of library preparation kit



ROCs as a function of DNA input amount



A DNA amount and locus-dependent threshold (all Kapa Normalized)

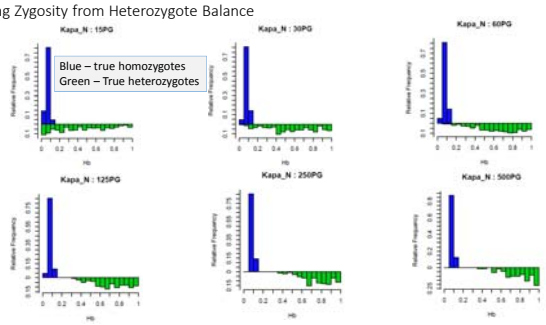


Inferring Zygosity from Heterozygote Balance

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.*
  - NIJ: *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. In no case does such identification imply a 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.