Emerging Topics: Rapid PCR/DNA Typing and Ultra High-Throughput Sequencing for HID applications

Peter M. Vallone, PhD
Applied Genetics Group
National Institute of Standards and Technology

Life Technologies HID Professional Services Meeting

February 26, 2013
Frederick, MD

Outline

- Rapid PCR Protocols
- Rapid STR Typing Workflows
- Integrated Rapid DNA Instruments
- Ultra High-Throughput Sequencing (NGS)

Rapid PCR Applications

- Faster sample-to-answer
  - Successful rapid PCR cycling reduces STR workflow times (**less than 2 hours in the laboratory**)
  - Single source reference and databasing samples
- Increased throughput (more runs per day)
- Integrated platforms for forensics and biometrics
  - Rapid DNA instruments (swab in → answer out)

Thermal Cycling Times for Commercial STR Kits

Year	Run on a 9700 thermal cycler	Hot start	Time per cycle	Cycles	Post soak	Total time
1997/98	Profiler Plus/Cofiler	11 min	3 min	28	60 min	2:52
1999	SGM Plus	11 min	3 min	28	45 min	2:53
2000	PowerPlex 16	12 min	1 min 45 s	32	30 min	3:00
2001	Identifiler	11 min	3 min	28	60 min	2:58
2003	PowerPlex Y	12 min	1 min 45 s	32	30 min	3:18
2004	Yfiler	11 min	3 min	30	80 min	2:45
2007	PowerPlex S5	2 min	4 min	30	45 min	3:21
2007	minifiler	11 min	3 min 20 s	30	45 min	3:16
2009	ESI 16, 17 ESX 16,17	2 min	4 min	30	45 min	3:22
2009	PowerPlex 16 HS	2 min	1 min 45 s	32	30 min	2:42
2009	NGM	11 min	3 min 20 s	29	10 min	2:33
2009	Identifler Direct	11 min	3 min	26	25 min	2:34
2010	Idenfiler Plus	11 min	3 min 20 s	28	10 min	2:18
2011	PowerPlex 18D	2 min	1 min 10s	27	20 min	1:25
2012	NGM Express (direct)	1 min	48 s	26	5 min	0:45
2012	PowerPlex 21	1 min	1 min 40 s	30	10 min	1:23
2012	PowerPlex Y23	2 min	1 min 40 s	30	20 min	1:33
2012	PowerPlex Fusion	1 min	1 min 40 s	30	10 min	1:24
2012	GlobalFiler Express (direct)	1 min	33 s	26	8 min	0:40

DNA Polymerases

- AmpliTaq Gold® is typically used
  - Heat activated (avoid non-specific PCR products)
- Takara SpeedSTAR™ HS DNA Polymerase
  - Extension times of 100 bp/s are possible (compared to 20 bp/s for other polymerases)
  - Hot-start formulation is antibody mediated
- Qiagen
  - QIAGEN Fast Cycling PCR Kit
- New England Biolabs/Finnzymes
  - Phusion and Phire DNA Polymerases
  - Q5
- KAPA Biosystems
  - KAPA2G Fast PCR Kits
- Biotium
  - CheetahTM Taq
- Fermentas
  - ~~PyroStart Master Mix~~
- EMD Millipore
  - KOD DNA Polymerse

DNA Polymerase Characteristics

Why are cycling times decreasing?

- Accuracy (geometric selection)
- Proofreading (3' – 5' exonuclease activity)
- **Processivity:** the number of nucleotides incorporated before dissociation

E.Coli polymerase III

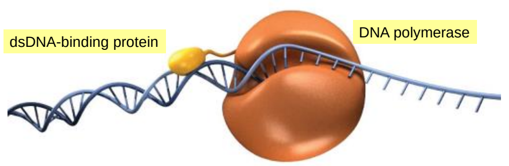
- Example: E.Coli polymerase III subunit (alone)
  - **P**rocessivity = 10 nt
  - **S**peed = 20 nt/s
- If associated with **sliding clamp** (and replisome subunits)
  - **P**rocessivity = 50 kb
  - **S**peed = 1000 nt/s

Taq Polymerase: P = 50 nt, S = 20 nt/s

Pomerantz, R. T. & O'Donnell, M. (2007) *Trends Microbiol. 15, 156–164*

Chimeric Polymerases

- Based on work Wang et al 2004 NAR 32: 1197-1207
  - Enhance polymerase **processivity** by covalently linking a non-specific ds binding protein to the polymerase domain
  - 16 to 32 fold increase in polymerase efficacy

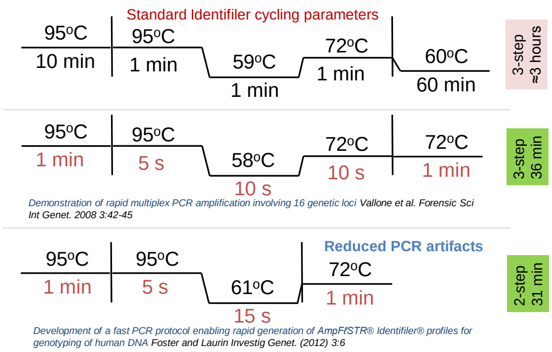


http://www.thermoscientificbio.com/pcr-enzymes-master-mixes-and-reagents/phire-hot-start-ii-dna-polymerase/
https://www.neb.com/tools-and-resources/feature-articles/anatomy-of-a-polymerase-how-structure-effects-function

Developing Protocols for 8 Thermal Cyclers

Cycling times given for a rapid 3-step 28 cycle protocol

PCR Thermal Cycling Profile



28 cycles on GeneAmp 9700

Effective heating/cooling rate

min	Cycler	Effective Heating/Cooling deg/s
36	GeneAmp 9700	1.6
19	Mastercycler Pro S	6.8
36	Rotor-Gene Q	1.6
22	SmartCycler	4.4
17	Philisa	10.9
30	Piko	2.2
22	SpeedCycler^2	4.4
17	Palm PCR	10.9

Rates for heating/cooling were estimated from the total cycling time

Comparative Throughput (Cycling)

		3 step	2 step		3 step	2 step
Cycler	# samples	Fastest Cycling Time (min)	Fastest Cycling Time (min)	Runs to complete 96 samples	Total min	Total min
GeneAmp PCR System 9700	96	36	31	1	36	31
Mastercycler Pro S	96	19	17	1	19	17
Rotor-Gene Q	72	36	32	2	72	64
SmartCycler	16	22	18	6	132	108
Philisa	8	17	14	12	204	168
Piko	96	30	26	1	30	26
SpeedCycler^2	96	22	19	1	22	19
Palm PCR	12	17	17	8	136	136

- Varying characteristics of heating/cooling and tube (reaction vessel)
- Rotor-Gene Q and SmartCycler are real-time PCR instruments

While cycling times may be rapid, the throughput in some cases is reduced from the standard 96-well format

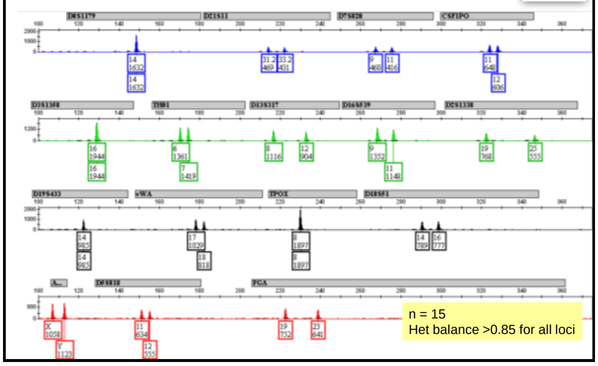
Experiments and PCR Conditions

- Develop a successful PCR protocols for each cycler with 2- and 3-step cycling conditions
- Sensitivity study: 1 sample, 7 concentrations in duplicate; compare 2- and 3-step PCR protocols
- Rapid STR typing workflow example (less than 2 hours)
- 95 samples amplified on a 9700 cycler → compare 2- and 3-step PCR protocols

- 1 X Takara PCR mastermix, 1 U SpeedStar polymerase *Premix Ex Taq*™ (Perfect Real Time)
- 10 μL total reaction volume in a thin walled tube (8-strip) or proprietary tube
- 2 μL of Identifiler PCR primer mix
- ~1 ng of template DNA
- 2- and 3-step cycling conditions
- Separation and detection on a 3130xl or 3500/3500xl

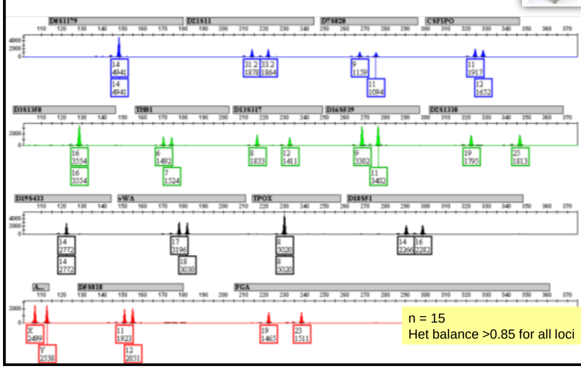
1 ng of DNA template
3130xl; 2 kV 5 s

AB 9700 3-Step (36 min)



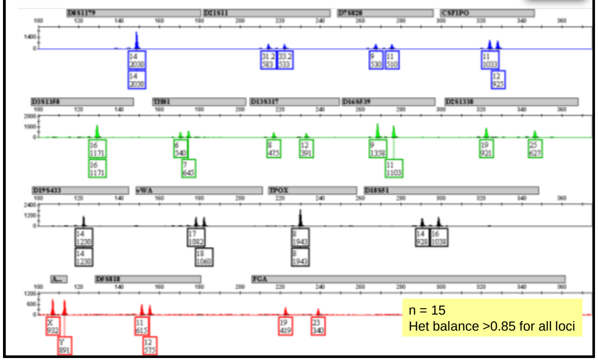
1 ng of DNA template
3130xl; 2 kV 5 s

AB 9700 2-Step (31 min)



1 ng of DNA template
3130xl; 2 kV 5 s

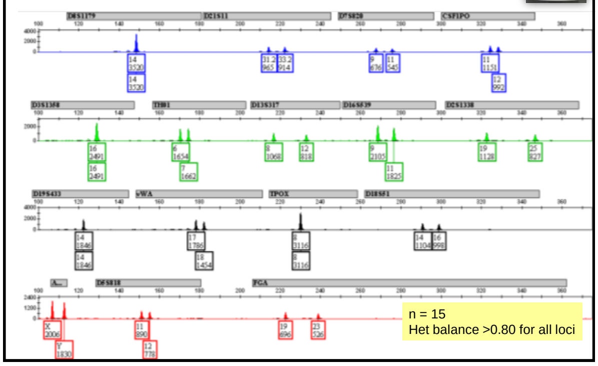
MasterCycler Pro 3-Step (19 min)



n = 15
Het balance >0.85 for all loci

1 ng of DNA template
3130xl; 2 kV 5 s

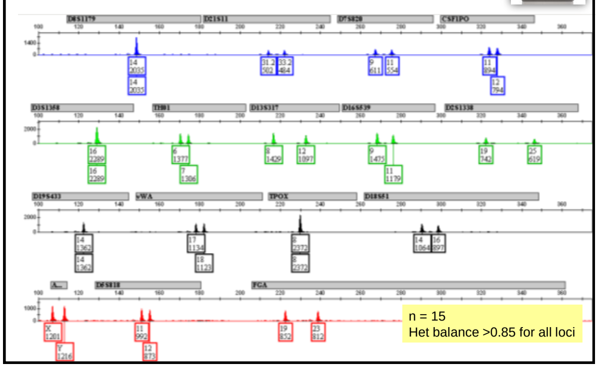
MasterCycler Pro 2-Step (17 min)



n = 15
Het balance >0.80 for all loci

1 ng of DNA template
3130xl; 2 kV 5 s

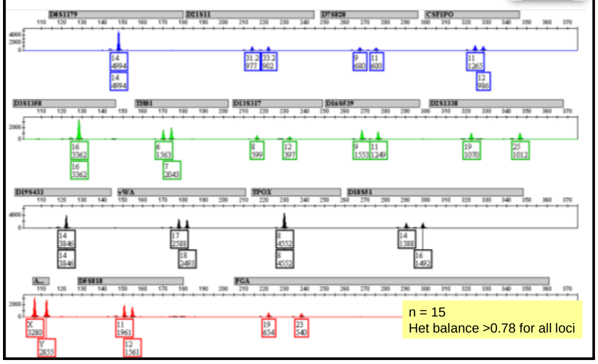
SmartCycler 3-Step (22 min)



n = 15
Het balance >0.85 for all loci

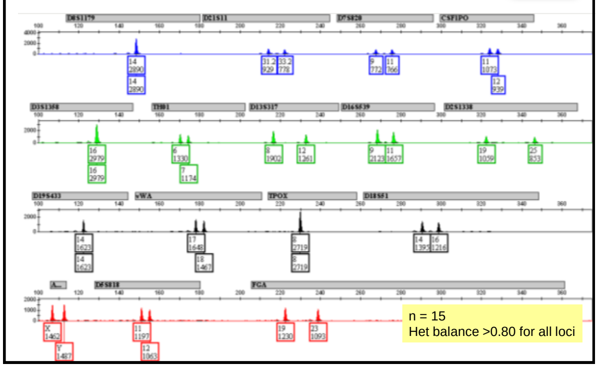
1 ng of DNA template
3130xl; 2 kV 5 s

SmartCycler 2-Step (18 min)



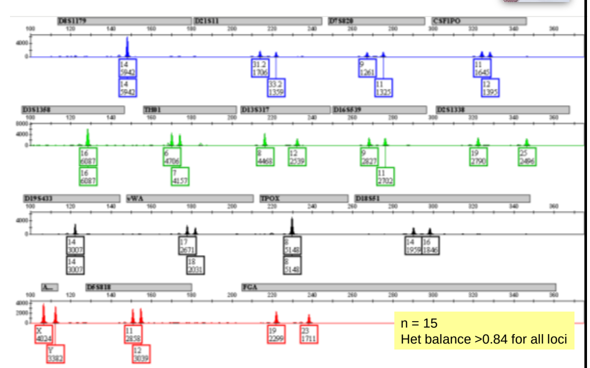
1 ng of DNA template
3130xl; 2 kV 5 s

RotorGene 3-Step (36 min)



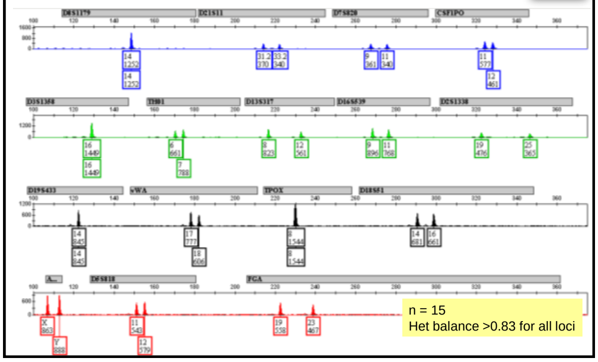
1 ng of DNA template
3130xl; 2 kV 5 s

RotorGene 2-Step (32 min)



1 ng of DNA template
3130xl; 2 kV 5 s

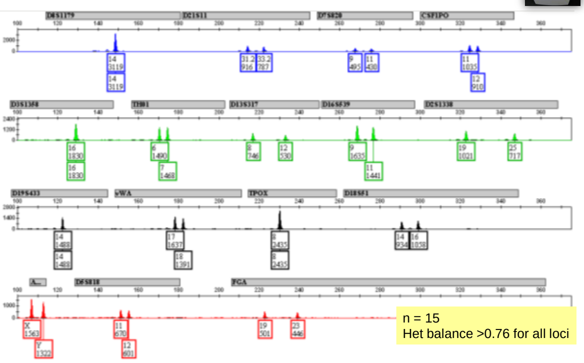
Piko 3-Step (30 min)



n = 15
Het balance >0.83 for all loci

1 ng of DNA template
3130xl; 2 kV 5 s

Piko 2-Step (26 min)

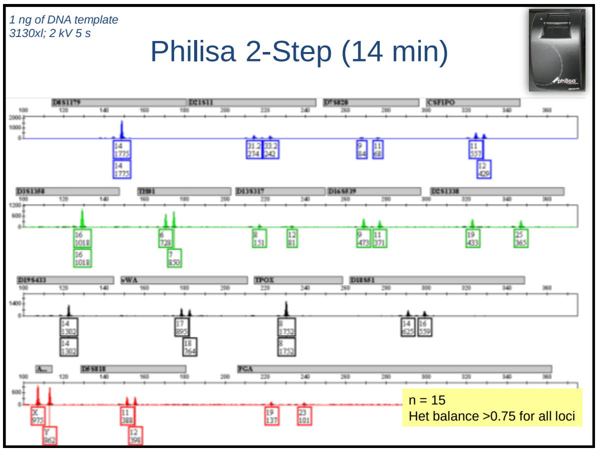


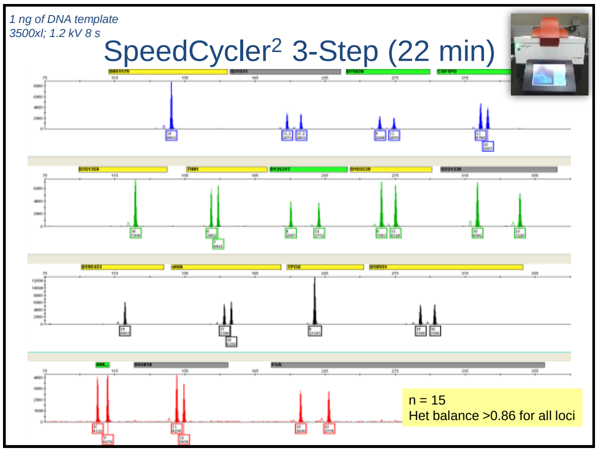
n = 15
Het balance >0.76 for all loci

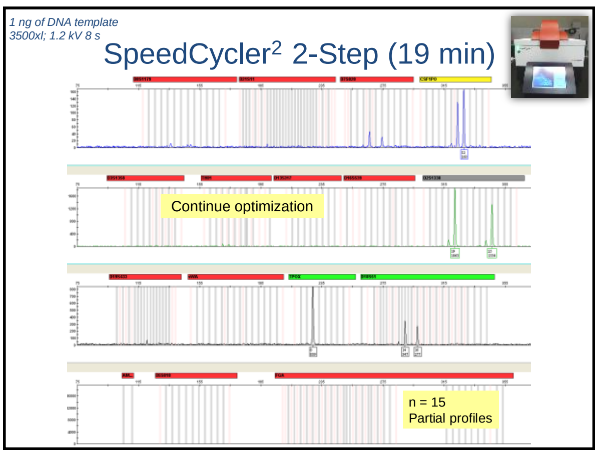
1 ng of DNA template
3130xl; 2 kV 5 s

Philisa 3-Step (17 min)

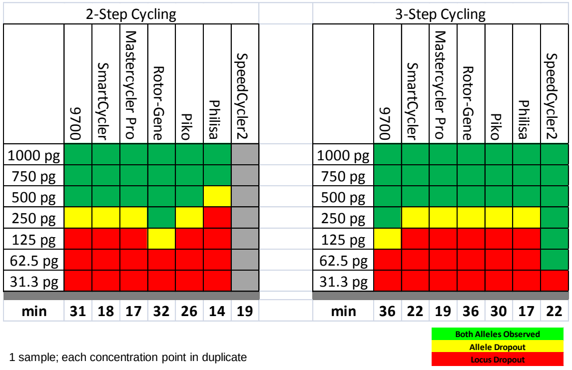
n = 15
Het balance >0.85 for all loci


1 ng of DNA template
3130xl; 2 kV 5 s
Philisa 2-Step (14 min)
n = 15
Het balance >0.75 for all loci


1 ng of DNA template
3500xl; 1.2 kV 8 s
SpeedCycler^2 3-Step (22 min)
n = 15
Het balance >0.86 for all loci


1 ng of DNA template
3500xl; 1.2 kV 8 s
SpeedCycler^2 2-Step (19 min)
Continue optimization
n = 15
Partial profiles

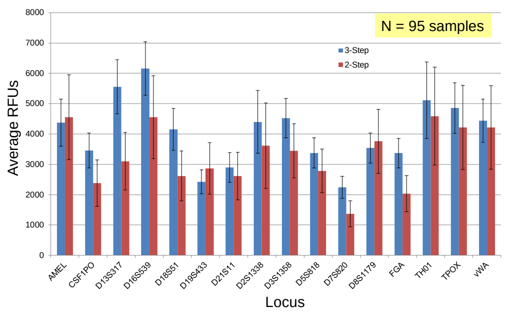
Sensitivity



1 sample; each concentration point in duplicate

RFU comparison for 9700 data

2-step versus 3-step



Peak Height Balance and Stutter

9700 data 2-step versus 3-step

N = 95 samples

n		Median PHR 2-step	Median PHR 3-step
95	**AMEL**	0.937	0.935
78	**CSF1PO**	0.871	0.904
71	**D13S317**	0.873	0.917
76	**D16S539**	0.875	0.909
82	**D18S51**	0.887	0.899
72	**D19S433**	0.914	0.929
71	**D21S11**	0.893	0.894
85	**D2S1338**	0.900	0.889
78	**D3S1358**	0.863	0.900
63	**D5S818**	0.850	0.897
78	**D7S820**	0.903	0.907
73	**D8S1179**	0.910	0.905
80	**FGA**	0.884	0.908
60	**TH01**	0.940	0.910
62	**TPOX**	0.919	0.930
77	**vWA**	0.880	0.893

error = 0.05 - 0.09

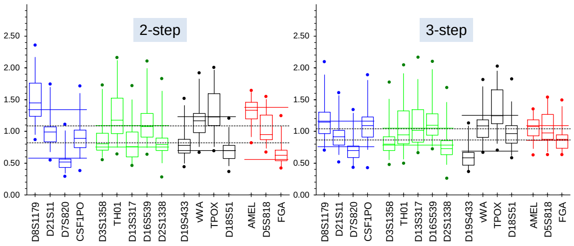
% stutter

		2-step		3 step	
n	Locus	Median	MADe	Median	MADe
99	TH01	2.2	1.1	2.1	1.0
115	TPOX	3.2	1.5	3.0	1.4
109	D7S820	5.4	2.2	5.0	2.3
96	D13S317	5.6	2.0	5.3	1.8
75	CSF1PO	6.3	1.9	5.5	1.5
107	D16S539	7.1	2.5	6.8	2.2
92	D5S818	8.0	1.6	7.2	1.1
93	D8S1179	8.4	1.6	7.5	1.6
115	vWA	8.8	2.1	8.1	2.0
133	D21S11	9.2	1.6	8.6	1.4
130	D18S51	9.3	2.8	8.6	2.2
129	FGA	9.5	1.5	8.7	2.4
101	D19S433	9.6	2.4	8.7	1.5
85	D3S1358	10.6	2.1	9.4	1.9
112	D2S1338	10.9	2.3	10.8	2.1

Intra-color balance

9700 data 2-step versus 3-step

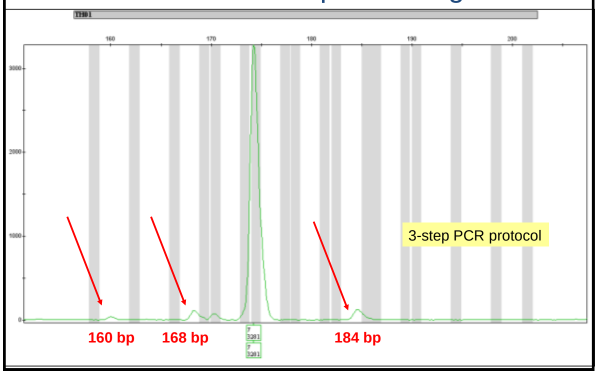
N = 95 samples



Y-axis = Locus RFU / Dye Group Average

Plots created by Dave Duewer's (NIST) excel-based software

PCR Artifacts in the TH01 amplicon range



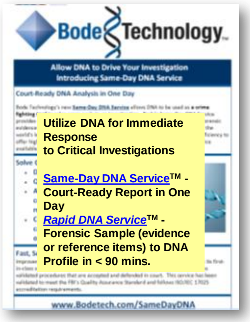
PCR Artifacts (3-step only)

Artifacts Observed	9700	Smart Cycler	Master Cycler Pro	Rotor-Gene	Streck Philisa
D16 @ 287 bp	35	4	1	6	0
D8 @ 121 bp	6	0	1	3	0
D8 @ 174 bp	14	10	1	7	0
TH01 @ 160 bp	28	2	1	11	0
TH01 @ 168 bp	83	32	1	40	0
TH01 @ 184 bp	59	19	0	25	0
TPOX @ 219 bp	77	13	2	22	2
Total # Artifacts	302	80	7	114	2

N = 95 samples

- TH01 @ 184 is often covered/disguised by the base of the 9.3 allele peak
- 50 RFU threshold used when identifying artifacts

Rapid STR profiling in a lab setting



NIST presentations on Rapid STR Typing

MAAFS 2012
"Rapid DNA Testing Approaches for Reference Samples"
http://www.cstl.nist.gov/strbase/pub_pres/Butts-MAAFS2012-rapid-DNA-testing.pdf

Promega 2012
"Rapid DNA Testing Approaches for Reference Samples"
http://www.cstl.nist.gov/strbase/pub_pres/Butts-ISHI2012-Rapid-DNA.pdf

Example Rapid Work Flow in Lab Setting

single source reference samples

DNA extraction → Rapid PCR → CE separation and detection

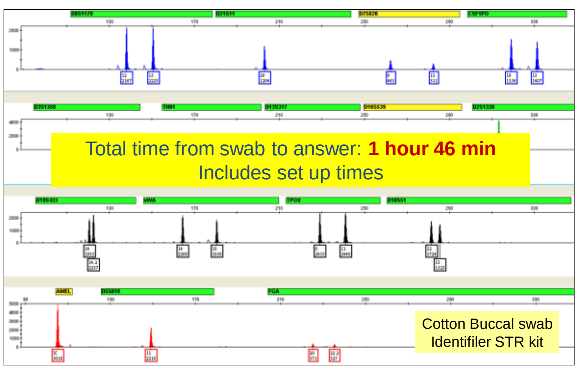
- Prep-N-Go extraction (cotton buccal)
- Rapid Identifiler (9700 & Philisa cyclers)
- Separations on an 8 capillary 3500

Direct PCR → CE separation and detection

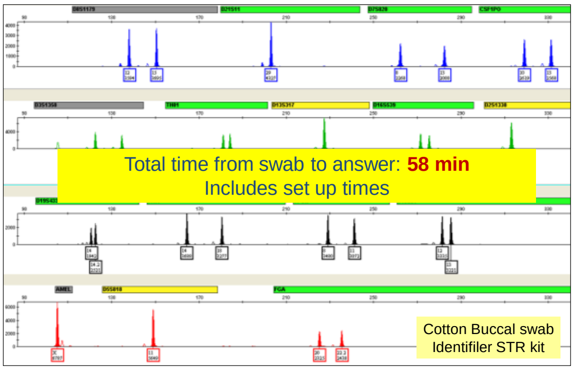
- 1.2 mm blood punch
- GlobalFiler Express (9700)
- Separations on an 8 capillary 3500

8 unique samples were typed in parallel

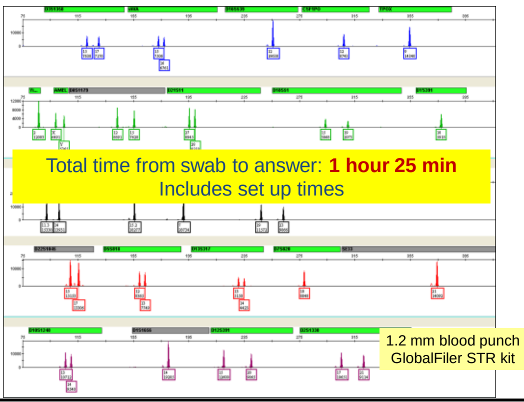
Prep-N-Go → 9700 (2-step) → 3500



Prep-N-Go → Philisa (2-step) → 3500



FTA → 9700 → 3500



Conclusions

- Successful protocols developed for 7/8 cyclers tested
  - 14 min PCR on Philisa cycler
- Continue work on Palm PCR and SpeedCycler^2
- Under the stated conditions sensitivity is around 250-500 pg of template DNA
- 2-step PCR protocol:
  - Faster
  - Similar sensitivity compared to 3-step
  - Comparable RFUs; peak height balance and stutter
  - Fewer PCR artifacts
- Complete STR profiling in < 2 h (swab-to-answer)

Rapid DNA

Rapid DNA (R-DNA) describes the fully automated (hands free) process of developing a CODIS Core STR profile from a reference sample buccal swab

D8S1179	{15,16}
D21S11	{29,29}
D7S820	{9,11}
CSF1PO	{10,11}
D3S1358	{16,17}
TH01	{6,7}
D13S317	{8,12}
D16S539	{10,11}
VWA	{15,17}
TPOX	{8,12}
D18S51	{11,15}
D5S818	{9,11}
FGA	{19,22}
Amel	{X,Y}

Performance Testing Goals

- Testing of R-DNA platforms for baseline performance of **concordance**, **reproducibility**, and **reliability**

- Type similar sample sets on multiple instruments and from multiple vendors
- Results will help guide platform improvements and additional testing

Carry this out through an **inter-laboratory study**

Initial Goals

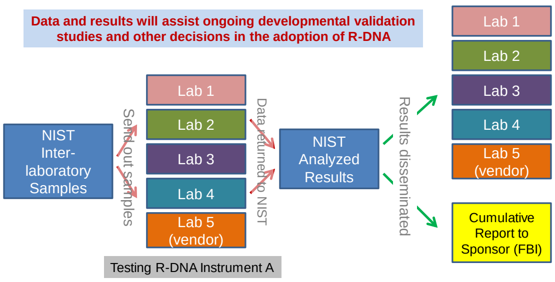
Performance Asessment

- Run 5-10 cartridges for baseline performance
  - Confirm that the instrument is operating
  - General level of genotyping success (hoping for greater than 80%)
- Run 50-80 samples for a performance testing (part of an inter-lab study)
  - Assess genotyping success
  - Additional CE metrics (peak balance, stutter, etc)

Running single-source reference samples

Inter-laboratory Testing Results

- Provide participants and sponsor with data and feedback
  - ✓ Each **participant** and will receive their specific performance feedback
  - ✓ The **sponsor** (FBI) will get a cumulative report for dissemination

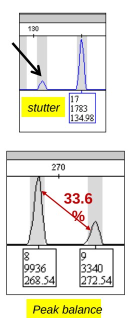


What will this data provide?
High level

- Is the correct profile obtained?
- Typing success
  - Per lane, chip, overall
- Incorrect profiles
- Partial profiles
- Allele drop out
- Contamination
- General operational issues
  - Instrument/chip failures
  - Hardware and software

What will this data provide?
Detailed-expert user; developer

- Electropherogram characteristics
  - Signal intensity
  - Peak balance (inter- and intra locus)
  - Stutter, PCR artifacts, adenylation
  - Sizing precision of peaks
- Manual versus automated allele calls
  - Confirm optimal software allele calling parameters



Ultra high-throughput sequencing

Next Generation Sequencing
Ultra High-Throughput Sequencing

- Going in depth into STR loci and beyond
  - STRs are useful for legacy (databases)
  - Millions of bases of sequence variants (SNPs)
- Opens up new human identity applications: complex kinship, biogeographical ancestry, externally visible traits, degraded samples?, mixtures?, other applications

Applications are currently being addressed by the forensic genetics community (*Kayser and deKnijff 2011)*

Next Generation Sequencing
Ultra High-Throughput Sequencing

- Challenges
  - Repeating sequences (STRs) and read lengths
  - Sample requirements (10 ng to 5 µg)
  - Cost and time per unit of information
  - Data analysis (storage, assembly, interpretation)
  - Policy, privacy, disease related markers
  - Validation
  - Standards/reference materials
    - Accuracy of sequence information
    - Errors, platform and bioinformatics-based bias

Multiplexing samples and reduce data set while maintaining quality coverage

A single NGS experiment

Single sample – full genome coverage

Multiplexing samples and reduce data set while maintaining quality coverage

A single NGS experiment

1	9	17	25	33	41	49	57	65	73	81	89
2	10	18	26	34	42	50	58	66	74	82	90
3	11	19	27	35	43	51	59	67	75	83	91
4	12	20	28	36	44	52	60	68	76	84	92
5	13	21	29	37	45	53	61	69	77	85	93
6	14	22	30	38	46	54	62	70	78	86	94
7	15	23	31	39	47	55	63	71	79	87	95
8	16	24	32	40	48	56	64	72	80	88	96

96 samples, high depth coverage of the **forensically relevant markers**

100s, 1000s, 500k, 1M per sample

- STRs and SNPs for one-to-one matching
- Ancestry markers (X, Y, mito, autosomal)
- Phenotypic markers (eye color, hair color, etc)
- Kinship (linked and unlinked markers)
- Other
- If possible, avoid disease related markers

Mitigate costs by multiplexing samples and sequencing forensically relevant information

NIST assessment of NGS

- Starting with mitochondrial DNA analysis
  - Simple 16.5 kb genome, compare to Sanger data
- **SRM 2392 (CHR, 9947A) and 2392-I (HL-60)**
- Plus 4 additional ‘challenging’ NIST pop samples (maximum differences from rCRS)

Initial Approaches

- To guide future purchasing decisions pilot studies were performed on 3 platforms
  - Life Technologies PGM (April – May 2012)
    - **Edge Biosystems** – outsourced library prep and sequencing
  - Illumina HiSeq (June – August 2012)
    - **Beckman-Coulter Genomics** – outsourced library prep and sequencing
  - Life Technologies SOLiD (June – July 2012)
    - **In-house collaboration with NGS group at NIST**

Niels Morling's lab in Copenhagen has shared 454 results for SRM 2392

Initial Approaches

- **Life Technologies PGM** instrument was installed at NIST in September 2012
  - Completed instrument training – Sept 2012
  - Initial run on PGM – Nov 2012
  - Second run – just completed
- Plan to obtain an Illumina MiSeq in 2013

Goals

- To compare the sequencing results of each platform to Sanger sequence (as found in the SRM certificate)
  - Identify any errors, low level heteroplasmy (mixtures), platform specific bias
- Obtain experience with:
  - NGS library preparation
  - Data analysis (general workflow and mito specific)
  - Outsourcing sequencing
- Compare data from the platforms
  - Which is the most accurate (for mito)
  - Platform and informatics bias

Variant Calls - Concordance

- Three platforms tested
  - Ion Torrent
    - Edge Biosystems
    - NIST
  - Illumina HiSeq
    - Beckman Genomics
  - Life Technologies SOLiD
    - NIST
- Compare to Sanger data
  - Insertions at 309, 315
    - All variant callers missed
  - Heteroplasmy at 1393
    - Not detected by Sanger
    - Approx. 17 - 18% of reads
  - All other sites in agreement
    - One missed call at 13759

SRM 2392 - component B - 9947a			EdgeBio	NIST	NIST	Beckman Genomics	NIST
Position	Reference	Sanger Call	PGM	PGM run 1	PGM run 2	Illumina	SOLiD
93	A	G	G	G	G	G	G
195	T	C	C	C	C	C	C
214	A	G	G	G	G	G	G
263	A	G	G	G	G	G	G
309.1	:	C					
309.2	:	C					
315.1	:	C					
750	A	G	G	G	G	G	G
1393	G	G	G/A	G/A	G/A	G/A	G/A
1438	A	G	G	G	G	G	G
4135	T	C	C	C	C	C	C
4769	A	G	G	G	G	G	G
7645	T	C	C	C	C	C	C
7861	T	C	C	C	C	C	C
8448	T	C	C	C	C	C	C
8860	A	G	G	G	G	G	G
9315	T	C	C	C	C	C	C
13572	T	C	C	C	C	C	C
13759	G	A	A		A	A	A
15326	A	G	G	G	G	G	G
16311	T	C	C	C	C	C	C
16519	T	C	C	C	C	C	C

Future Directions

- Expand work with PGM and MiSeq platforms
- Sequence SRM 2391c alleles?
- Assess NGS software packages
  - GATK, NextGene, CLC bio, Sequencher, ?

Thank you for your attention!

Questions?

Peter.Vallone@nist.gov (1-301-975-4872)

Acknowledgements

Erica Butts - Rapid STR Typing Protocols
Kevin Kiesler – NGS work
Dave Duewer – Data analysis software (stutter, peak height ratios, multiplex balance)

Outside funding agencies:
FBI - Evaluation of Forensic DNA Typing as a Biometric Tool
NIJ – Interagency Agreement with the Office of Law Enforcement Standards