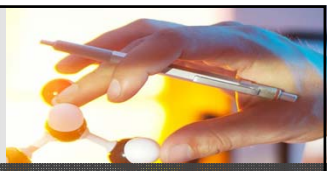
Mark Wilson
Research Leader
Applied Genomics
Battelle, Inc.



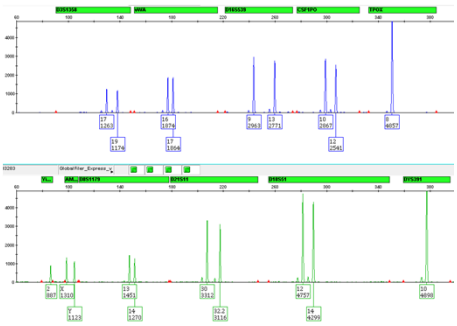
Validation challenges for implementing massively parallel sequencing into forensic DNA casework

Green Mountain DNA Conference
Burlington, VT
August 2, 2016

Topics

1. Patterns in MPS data from STRs
2. Controls and Bioinformatics
3. Allele is the Sequence
4. Probabilistic Modeling and MPS

Current Generation



Other core repeat numbers are *inferred* from the original reference conversion…

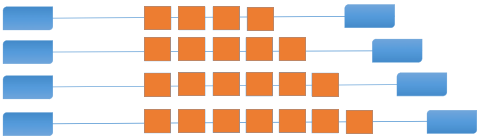
If 162 bp = 4 repeats, then:

166 bp = 5

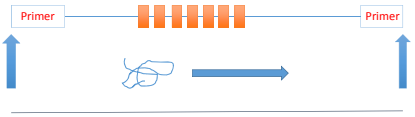
170 bp = 6

174 bp = 7

etc.

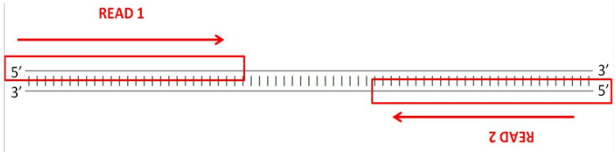


Typing by size



Typing by DNA sequence

Paired-end Sequencing



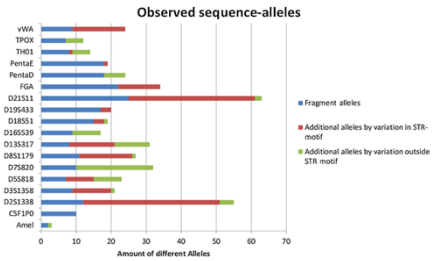
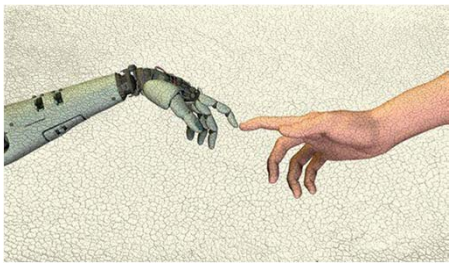


Fig. 5: Kristiaan J. van der Gaag, et.al. Forensic Science International: Genetics 24 (2016) 86–96

Massively Parallel

FGA	[TTTC]3TTTTTTCT[CTTT]16CTCC[TTCC]2	96	1008	24	2439	294	0.1205	
FGA	[TTTC]3TTTTTTCT[CTTT]16CTCC[TTCC]2	104	806	26				
FGA	[TTTC]3TTTTTTCT[CTTT]15CTCC[TTCC]2	92	171	23			0.1696	
FGA	[TTTC]3TTTTTTCT[CTTT]13CTCC[TTCC]2	100	137	25			0.1700	
FGA	[TTTC]3TTTTTTCT[CTTT]14CTCC[TTCC]2	88	13	22			0.0129	n-8
FGA	[TTTC]3TTTTTTCT[CTTT]15CTCC[TTCC]2	108	10	27			0.0124	n+4
FGA	GTT[CTTT]3TTTCT[CTTT]16CTCC[TTCC]2	96	7	24				
FGA	[TTTC]3TTTTTTCT[CTTT]6CCTT[CTTT]9CTCC[TTCC]2	104	6	26				
FGA	[TTTC]3TTTTTTCT[CTTT]6CCTT[CTTT]7CTCC[TTCC]2	96	5	24				
FGA	[TTTC]3TTTTTTCT[CTTT]6CCTT[CTTT]13CTCC[TTCC]2	96	5	24				
FGA	[TTTC]3TTTTTTCTCTCT[CTTT]13CTCC[TTCC]2	104	5	26				
FGA	[TTTC]3TTTTTTCT[CTTT]15ATTTCTCC[TTCC]2	96	4	24				
FGA	[TTTC]3TTTTTTCT[CTTT]6CCTT[CTTT]9CTCC[TTCC]2	96	4	24				
FGA	[TTTC]3TTTTTTCT[CTTT]5CCTT[CTTT]12CTCC[TTCC]2	104	4	26				
FGA	[TTTC]3TTTTTTCTCCTT[CTTT]15CTCC[TTCC]2	96	4	24				
FGA	[TTTC]3TTTTTTCT[CTTT]16[TTCC]3	96	3	24				
FGA	[TTTC]3TTTTTTCT[CTTT]16CTCCTTCCCTCC	104	3	26				
FGA	[TTTC]3TTTTTTCT[CTTT]17ATTTCTCC[TTCC]2	104	3	26				
FGA	[TTTC]3TTTTTTCT[CTTT]16CTTCCTTTCTCC[TTCC]2	104	3	26				
FGA	[TTTC]3TTTTTTCT[CTTT]15CCTT[CTTT]2CTCC[TTCC]2	104	3	26				
FGA	[TTTC]3TTTTTTCT[CTTT]14CCTT[CTTT]3CTCC[TTCC]2	104	3	26				
FGA	[TTTC]3TTTTTTCT[CTTT]11CCTT[CTTT]6CTCC[TTCC]2	104	3	26				
FGA	[TTTC]3TTTTTTCT[CTTT]10CCTT[CTTT]7CTCC[TTCC]2	104	3	26				
FGA	[TTTC]3TTTTTTCT[CTTT]9CTTCC[TTTC]8TCC[TTCC]2	104	3	26				
FGA	[TTTC]3TTTTTTCT[CTTT]9CTTC[TTTC]8TCC[TTCC]2	96	3	24				
FGA	[TTTC]3TTTTTTCT[CTTT]9CCTT[CTTT]6CTCC[TTCC]2	96	3	24				
FGA	[TTTC]3TTTTTTCT[CTTT]3CTTCC[TTTC]14TCC[TTCC]2	104	3	26				
FGA	[TTTC]3TTTTTTCT[CTTT]3CCTT[CTTT]14CTCC[TTCC]2	104	3	26				

Data Analysis



Locus	Target	Length	Abundance	Repeat Number
D3S1358	TCTA[TCTG]3[TCTA]12	64	2,196	16
D3S1358	TCTA[TCTG]3[TCTA]14	72	2,122	18
D3S1358	TCTA[TCTG]3[TCTA]13	68	311	17
D3S1358	TCTA[TCTG]3[TCTA]11	60	242	15
D3S1358	TCTA[TCTG]2[TCTA]13	64	45	16
D3S1358	TCTA[TCTG]2[TCTA]15	72	25	18
D3S1358	TCTA[TCTG]3[TCTA]10	56	12	14
D3S1358	TCTA[TCTG]3[TCTA]10CCTA[TCTA]3	72	11	18
D3S1358	TCTA[TCTG]3[TCTA]15	76	11	19
D3S1358	[TCTA]2[TCTG]2[TCTA]12	64	10	16
D3S1358	TCTA[TCTG]2[TCTA]14	68	10	17
D3S1358	TCTA[TCTG]4[TCTA]13	72	10	18
D3S1358	TCTA[TCTG]3TCTATCCA[TCTA]12	72	10	18
D3S1358	TCTATCTGCCTGTCTG[TCTA]12	64	7	16
D3S1358	TCTA[TCTG]2[TCTA]12	60	6	15
D3S1358	TCTA[TCTG]4[TCTA]11	64	6	16
D3S1358	[TCTA]2[TCTG]2[TCTA]14	72	6	18
D3S1358	TCTA[TCTG]2TCTT[TCTA]12	64	5	16
D3S1358	TCTA[TCTG]3CCTA[TCTA]13	72	5	18

'Identify Reads' Application

Locus	Target	Length	Abundance	Allele Size	Call
D3S1358	TCTA[TCTG]3[TCTA]12	64	1132	16	allele
D3S1358	TCTA[TCTG]3[TCTA]14	72	1104	18	allele
D3S1358	TCTA[TCTG]3[TCTA]13	68	159	17	stutter
D3S1358	TCTA[TCTG]3[TCTA]11	60	132	15	stutter
D3S1358	TCTA[TCTG]2[TCTA]15	72	13	18	stutter
D3S1358	TCTA[TCTG]2[TCTA]13	64	9	16	stutter
D3S1358	TCTA[TCTG]3[TCTA]10	56	5	14	stutter
D3S1358	[TCTA]2[TCTG]2[TCTA]14	72	5	18	error
D3S1358	TCTA[TCTG]4[TCTA]13	72	4	18	stutter
D3S1358	TCTA[TCTG]4[TCTA]11	64	4	16	stutter
D3S1358	TCTA[TCTG]2[TCTA]14	68	4	17	stutter
D3S1358	TCTA[TCTG]2[TCTA]12	60	4	15	stutter
D3S1358	TCTA[TCTG]2TCCG[TCTA]14	72	4	18	error
D3S1358	TCTACCTG[TCTG]2[TCTA]14	72	4	18	error
D3S1358	TCTA[TCTG]3[TCTA]2TCTG[TCTA]9	64	3	16	error
D3S1358	TCTA[TCTG]3[TCTA]15	76	3	19	stutter

'Identify Reads' Application

Allele Size	Call	Stutter Position	Allele 1 stutter percent	Allele 2 stutter percent
16	allele			
18	allele			
17	stutter	'1/-1	14.0459364	14.40217391
15	stutter	'-1/-3	11.66077739	11.95652174
18	stutter	'2/0	1.148409894	1.177536232
16	stutter	'0/-2	0.795053004	0.815217391
14	stutter	'-2/-4	0.441696113	0.452898551
18	error			
18	stutter	'2/0	0.35335689	0.362318841
16	stutter	'0/-2	0.35335689	0.362318841
17	stutter	'1/-1	0.35335689	0.362318841
15	stutter	'-1/-3	0.35335689	0.362318841
18	error			
18	error			
16	error			
19	stutter	'3/1	0.265017668	0.27173913

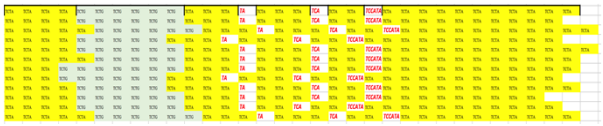
Locus D3S1358 Sample XYZ
Comprehensive Read Analysis

Alleles	Size in bp	Abundance	Repeats
TCTA[TCTG]3[TCTA]12	64	2,196	16
TCTA[TCTG]3[TCTA]14	72	2,122	18

Stutter Products	Size in bp	Abundance	Repeats	Explanation
TCTA[TCTG]3[TCTA]13	68	311	17	N-4, 72
TCTA[TCTG]3[TCTA]11	60	242	15	N-4, 64
TCTA[TCTG]3[TCTA]10	56	12	14	N-8, 64
TCTA[TCTG]3[TCTA]15	76	11	19	N+4, 72
TCTA[TCTG]2[TCTA]14	68	10	17	N-4(2), 72
TCTA[TCTG]2[TCTA]12	60	6	15	N-4(2), 64

Variants	Size in bp	Abundance	Repeats	Explanation
TCTA[TCTG]2[TCTA]13	64	45	16	P1,P2,etc
TCTA[TCTG]2[TCTA]15	72	25	18	
[TCTA]2[TCTG]2[TCTA]12	64	10	16	
TCTA[TCTG]4[TCTA]13	72	10	18	
TCTA[TCTG]3TCTATCCA[TCTA]12	72	10	18	
Etc				

Complex stutter patterns at complex loci D21S11

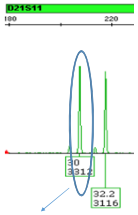


Repeat motif stutter or SNP?

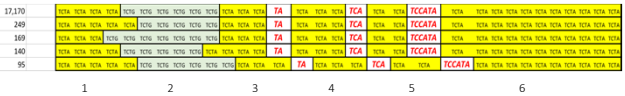
TCTA	TCTA	TCTA	TCTG	TCTG	TCTG
TCTA	TCTA	TCTG	TCTG	TCTG	TCTG
TCTA	TCTA	TCTG	TCTG	TCTG	TCTG
TCTA	TCTA	TCTA	TCTG	TCTG	TCTG

Multiple hypotheses require probability modeling

Different sequences of the same size at D21



size	abundance	possible explanation
127	249	+4 (1) '-4 (2)
127	169	-8 (1) '+8 (2)
127	140	-4 (2) +4 (3)
127	95	+4 (1) -4 (6)



Contents lists available at ScienceDirect

Forensic Science International: Genetics

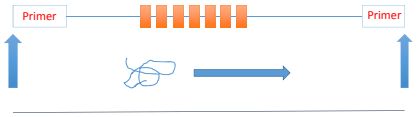
journal homepage: www.elsevier.com/locate/fsig

Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements

Walther Parson[a,b,*], David Ballard[c], Bruce Budowle[d,e], John M. Butler[f], Katherine B. Gettings[g], Peter Gill[h,i], Leonor Gusmão[j,k], Douglas R. Hares[l], Jodi A. Irwin[l], Jonathan L. King[g], Peter de Knijff[m], Niels Morling[n], Mechthild Prinz[o], Peter M. Schneider[p], Christophe Van Neste[q], Sascha Willuweit[r], Christopher Phillips[s]

"it is imperative that repeat region start and end locations be strictly defined for all STR loci employed in MPS."

Typing by size

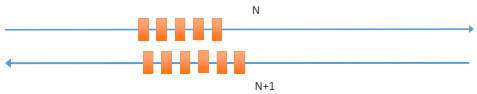


Typing by DNA sequence

Number of Apparent Repeats

"Out of 58 STR loci for which MPS designs have become available at the time of this writing (listed in Tables 2–4 of [35]), 23 have been designated historically **on the reverse strand**. In 17 of these loci, the change to the forward strand for repeat region designation results in a **potential shift of the reading frame** (Table 1). This shift of reading frame would be consistent with the earlier ISFG recommendations [7] that **the repeat region begins with the first possible repeat motif**. This change can cause a shift in the position of features within the motif and/or an increase in the number of apparent repeats."

Which Strand?



Forward Strand Reporting Issues: At the DYS389I/II loci, the potential exists for a two nucleotide shift, which would result in the appearance of one extra repeat in the larger allele.

Previously reported reverse strand: [TCTG]5 [TCTA]12 - 48 nt. - [TCTG]3 [TCTA]9 (29)
Forward strand, no frame shift: [TAGA]9 [CAGA]3 - 48 nt. - [TAGA]12 [CAGA]5 (29)
Forward strand, frame shift: [GATA]9 [GACA]3 48 nt. [GATA]12 **[GACA]6 (30)**

Retaining the repeat nomenclature requires defining an appropriate strand and a consistent definition of what constitutes a repeat.

Alternative alignments can alter the type mtDNA example

```
TTAACCCCCTCCCCCATGCT   - case
TCAAAACCCCCTCCCCATGCT  - rCRS

TTAA--CCCCCTCCCCCATGCT          16178-16198
TCAAAACCCCCTCCCC-ATGCT          16179T, 16182D, 16183D, 16193.1C

TT--AACCCCCTCCCCCATGCT           16178-16198
TCAAAACCCCCTCCCC-ATGCT           16179T, 16180D, 16181D, 16193.1C
```

String type:

```
16178-16198
TTAACCCCCTCCCCCATGCT
```

String Approach

- The allele is the sequence
- Repeat motif is irrelevant for typing
  - If different repeat motif, then different sequence
  - All repeat motifs are subsumed under the sequence
  - There are multiple ways to express repeat motifs
- Repeat motif may be relevant for reporting only.
- Primer-to-primer sequence can be converted to size in bp, which then can be converted to number of repeats, provided the start and stop points are retained, and a standard length to repeat size conversion is applied for each locus.

Simple stutter patterns

Genotype 14,15

Locus	Sequence	Repeat size	abundance
D8S1179	TCTATCTG[TCTA]12	14	18,605
D8S1179	[TCTA]2TCTG[TCTA]12	15	17,880
D8S1179	TCTATCTG[TCTA]11	13	3,060
D8S1179	[TCTA]2TCTG[TCTA]11	14	2,890

Positive stutter

Locus	Sequence	Repeat size	abundance
D5S818	[AGAT]12	13	47,317
D5S818	[AGAT]11	12	5,065
D5S818	[AGAT]13	14	714
D5S818	[AGAT]10	11	320

Stutter patterns can be complex

Locus	Sequence	Length	Abundance	Repeat #
D3S1358	TCTA[TCTG]3[TCTA]13	68	13,047	17
D3S1358	TCTA[TCTG]3[TCTA]14	72	10,199	18
D3S1358	TCTA[TCTG]3[TCTA]12	64	2,006	16
D3S1358	TCTA[TCTG]3[TCTA]11	60	175	15
D3S1358	TCTA[TCTG]2[TCTA]14	68	144	17
D3S1358	TCTA[TCTG]2[TCTA]15	72	101	18

Singletons

TCTA TCTG TCTG TCTG TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA TCTA = 16

TCTATCTGTCTTTCTGTCTA[TCTA]$_{11}$

TCTA[TCTG]$_3$TCTT[TCTA]$_{11}$

TCTA[TCTG]$_3$TCTATCTT[TCTA]$_{10}$

TCTA[TCTG]$_3$[TCTA]$_4$TCTG[TCTA]$_7$

TCTA[TCTG]$_3$[TCTA]$_5$TCTC[TCTA]$_6$

TCTA[TCTG]$_3$[TCTA]$_6$TCTG[TCTA]$_5$

TCTA[TCTG]$_3$[TCTA]$_7$TCTC[TCTA]$_4$

TCTA[TCTG]$_3$[TCTA]$_8$TCTG[TCTA]$_3$

TCTA[TCTG]$_3$[TCTA]$_9$TCAA[TCTA]$_2$

TCTA[TCTG]$_3$[TCTA]$_9$TCTG[TCTA]$_2$

TCTA[TCTG]$_3$[TCTA]$_{11}$TCCA

Many sequences occur once and are one base different from an allele

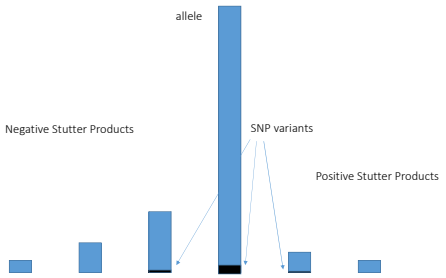
Data Analysis – Negative Controls

	MB		MB (2nd)	
Locus	**Length**	***Abundance***	**Length**	***Abundance***
D3S1358	68	*7*	*68*	*2*
D3S1358	64	*4*	*68*	*2*
D3S1358	72	*1*	*64*	*1*
D3S1358	64	*1*	*64*	*1*
			60	*1*
TH01	28	*24*	*28*	*4*
TH01	39	*1*	*39*	*3*
TH01	32	*1*	*39*	*1*

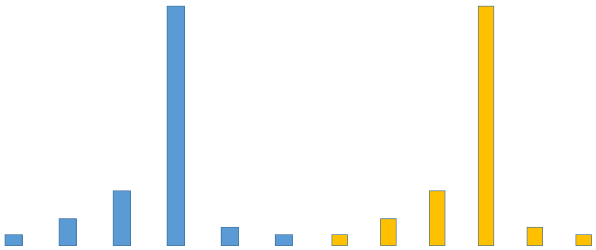
Data Analysis – Negative Controls

	MB		MB (2nd)	
Locus	**Length**	***Abundance***	**Length**	***Abundance***
D18S51	52	*13*	*48*	*3*
D18S51	76	*6*	*40*	*1*
D18S51	48	*5*		
D18S51	40	*3*		
D18S51	72	*2*		
D18S51	72	*1*		
D18S51	56	*1*		
D18S51	76	*1*		
PentaE	60	*10*	*85*	*2*
PentaE	80	*4*	*75*	*1*
PentaE	75	*3*	*75*	*1*
PentaE	80	*1*	*70*	*1*
PentaE	55	*1*		

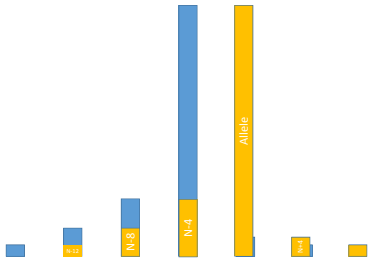
Sequence variants distribution revealed by MPS



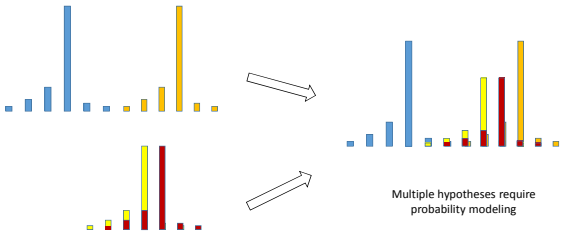
Stutter pattern distributions revealed by MPS



Stutter pattern distributions revealed by MPS



Mixtures! – Taking full advantage of the depth of data requires probabilistic modeling



Summary

- Patterns in MPS data from STRs
  - Stutter and SNP complexity could be used to support genotyping
  - Potential utility in mixture separation
- Controls and Bioinformatics
  - Bioinformatic processing of instrument output needs further development
- Typing by Sequence
  - Back compatibility calculations to repeat numbers are necessary
  - Structural repeat motif is irrelevant to sequence-based typing
- Probabilistic Modeling and MPS
  - Complex data interpretation requires complex models

The End