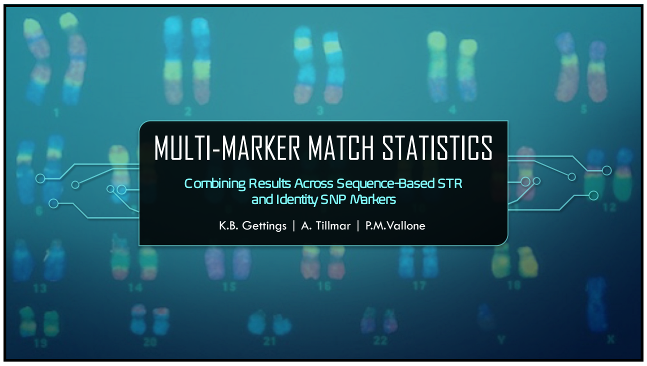
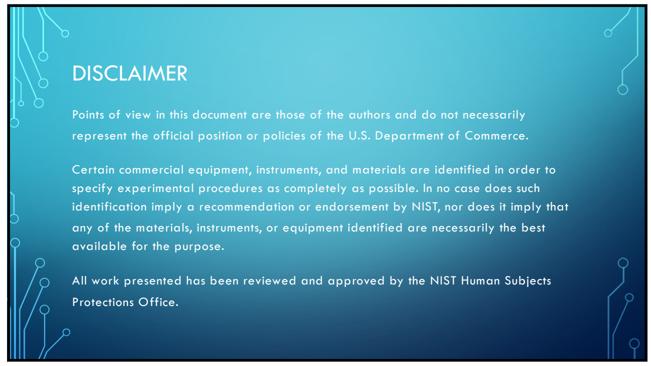
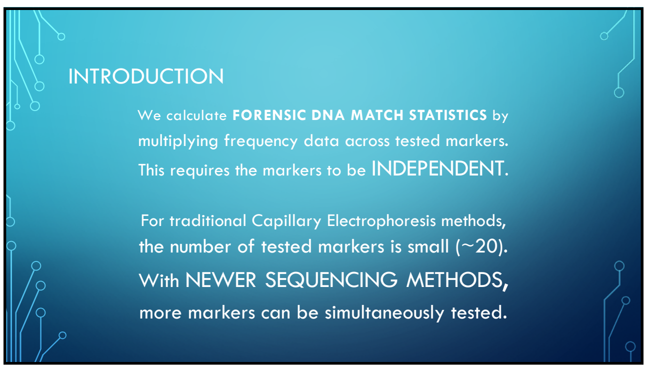

MULTI-MARKER MATCH STATISTICS
Combining Results Across Sequence-Based STR and Identity SNP Markers
K.B. Gettings | A. Tillmar | P.M.Vallone

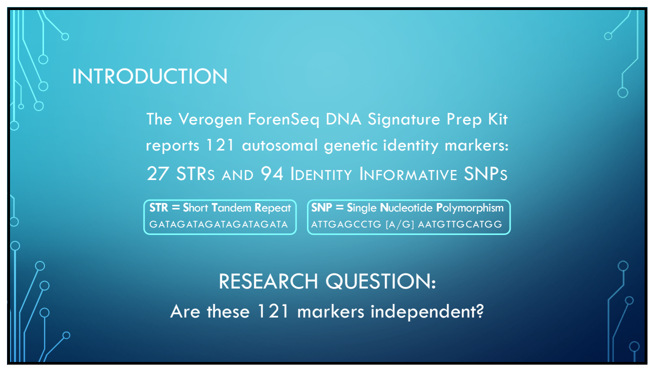
1


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.
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.

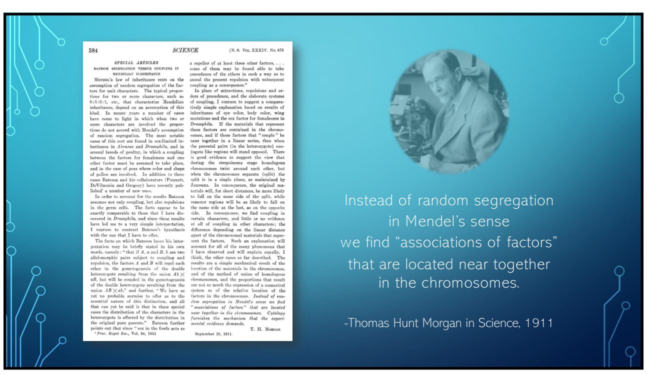
2


INTRODUCTION
We calculate FORENSIC DNA MATCH STATISTICS by multiplying frequency data across tested markers.
This requires the markers to be INDEPENDENT.
For traditional Capillary Electrophoresis methods, the number of tested markers is small (~20).
With NEWER SEQUENCING METHODS, more markers can be simultaneously tested.

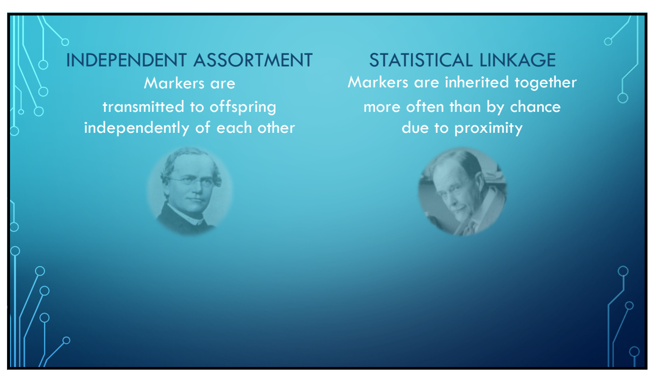
3


INTRODUCTION
The Verogen ForenSeq DNA Signature Prep Kit reports 121 autosomal genetic identity markers:
27 STRs AND 94 IDENTITY INFORMATIVE SNPs
STR = Short Tandem Repeat
GATAGATAGATAGATAGATA
SNP = Single Nucleotide Polymorphism
ATTGAGCCTG [A/G] AATGTTGCATGG
RESEARCH QUESTION:
Are these 121 markers independent?

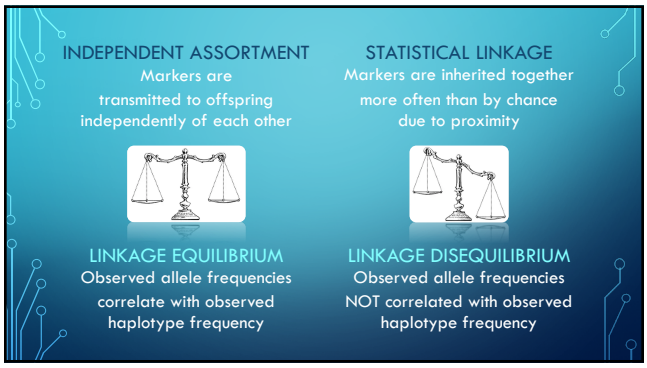
4


Instead of random segregation in Mendel's sense we find "associations of factors" that are located near together in the chromosomes.
-Thomas Hunt Morgan in Science, 1911

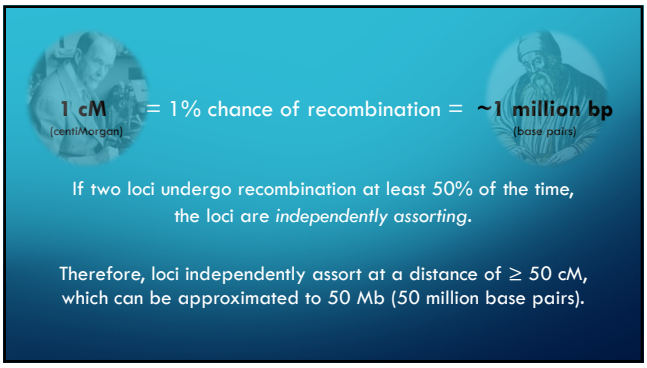
5


INDEPENDENT ASSORTMENT
Markers are transmitted to offspring independently of each other
STATISTICAL LINKAGE
Markers are inherited together more often than by chance due to proximity

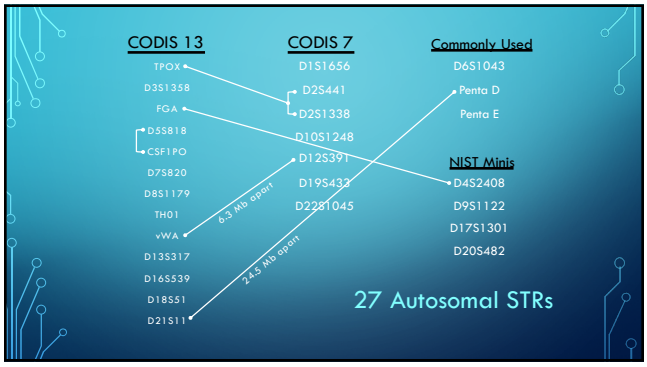
6


INDEPENDENT ASSORTMENT
Markers are transmitted to offspring independently of each other
STATISTICAL LINKAGE
Markers are inherited together more often than by chance due to proximity
LINKAGE EQUILIBRIUM
Observed allele frequencies correlate with observed haplotype frequency
LINKAGE DISEQUILIBRIUM
Observed allele frequencies NOT correlated with observed haplotype frequency

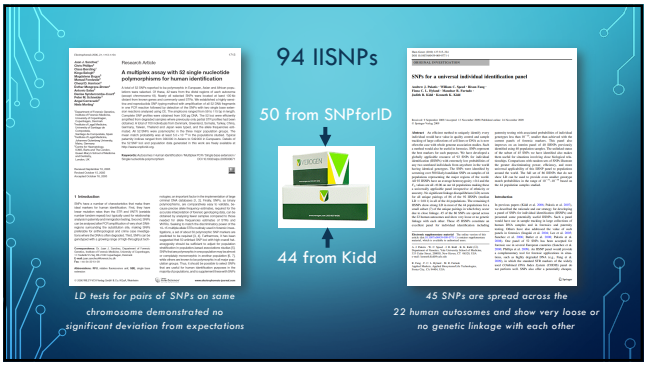
7


1 cM
(centiMorgan)
= 1% chance of recombination =
~1 million bp
(base pairs)
If two loci undergo recombination at least 50% of the time, the loci are independently assorting.
Therefore, loci independently assort at a distance of ≥ 50 cM, which can be approximated to 50 Mb (50 million base pairs).

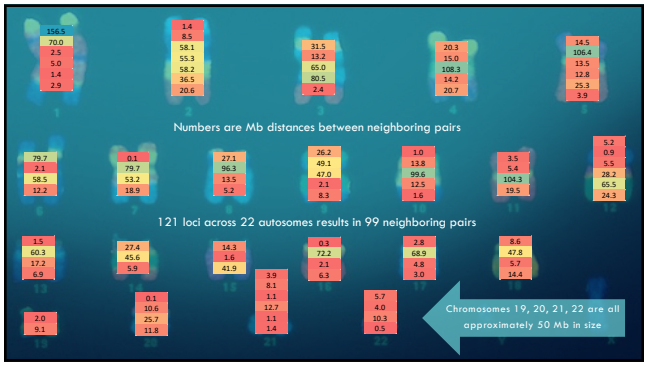
8


CODIS 13
TPOX
D3S1358
FGA
D5S818
CSF1PO
D7S820
D8S1179
TH01
vWA
D13S317
D16S539
D18S51
D21S11
CODIS 7
D1S1656
D2S441
D2S1338
D10S1248
D12S391
D19S433
D22S1045
Commonly Used
D6S1043
Penta D
Penta E
NIST Minis
D4S2408
D9S1122
D17S1301
D20S482
6.3 Mb apart
24.5 Mb apart
27 Autosomal STRs

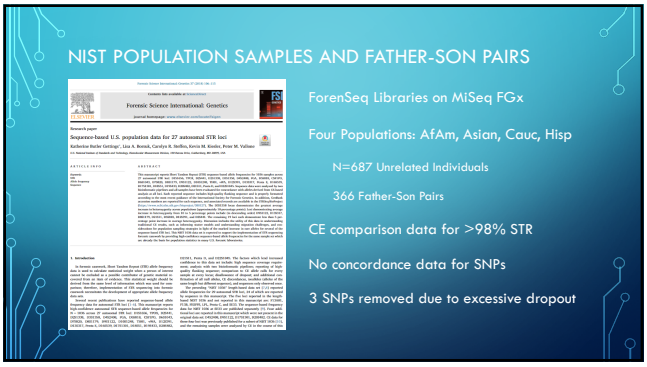
9


94 IISNPs
50 from SNPforID
44 from Kidd
LD tests for pairs of SNPs on same chromosome demonstrated no significant deviation from expectations
45 SNPs are spread across the 22 human autosomes and show very loose or no genetic linkage with each other

10


Numbers are Mb distances between neighboring pairs
121 loci across 22 autosomes results in 99 neighboring pairs
Chromosomes 19, 20, 21, 22 are all approximately 50 Mb in size

11


NIST POPULATION SAMPLES AND FATHER-SON PAIRS
ForenSeq Libraries on MiSeq FGx
Four Populations: AfAm, Asian, Cauc, Hisp
N=687 Unrelated Individuals
366 Father-Son Pairs
CE comparison data for >98% STR
No concordance data for SNPs
3 SNPs removed due to excessive dropout

12

TESTS FOR LINKAGE DISEQUILIBRIUM

Gametic Phase Unknown N=687 Population Samples

Likelihood Ratio test

Numerator - expected haplotype frequencies based on allele frequencies (H^*)

Denominator - observed haplotype frequencies (H)

D1S1656	rs1294331	H^*	H
25	A	0.1	0.3

$$LR = \frac{L_{H^*}}{L_H}$$

13

TESTS FOR LINKAGE DISEQUILIBRIUM

Gametic Phase Known 366 Father-Son Pairs

Exact Test

Build tables of observed haplotype frequencies

Determine statistical probability of finding tables with same haplotype frequencies

	D1S1656	rs1294331
M-01C	25	A
P-01C	6	G

14

RESULTS

15

Tested for all locus combinations (6903 pairings)

Chromosome 1

rs1490413
rs560681
D1S1656
rs1294331
rs10495407
rs891700
rs1413212

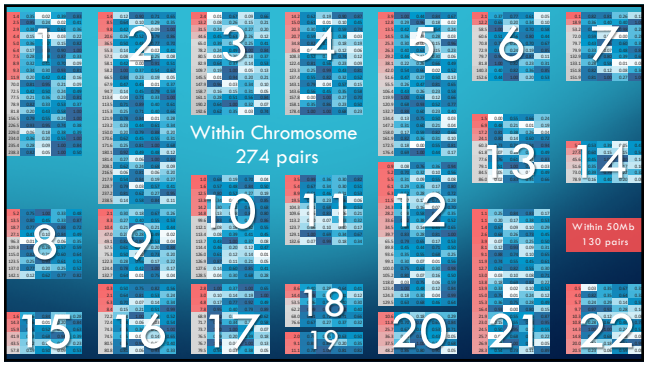
21 Pairings

Father-Son Samples
uncorrected p-values

Sorted by Mb distance between pairs

ΔMb	AfAm	Asian	Cauc	Hisp
1.4	0.35	0.02	0.39	0.83
2.5	0.95	0.24	0.02	0.41
2.9	0.38	0.71	0.61	0.36
4.4	0.15	0.02	0.03	0.22
5.0	0.36	0.47	0.15	0.82
6.4	0.17	0.83	0.90	1.00
7.5	0.28	0.58	0.87	0.33
8.9	0.32	0.41	0.70	0.09
9.3	0.34	0.30	0.93	0.81
11.9	0.20	0.42	0.42	0.16
70.0	0.81	0.95	0.21	0.09
72.5	0.42	0.50	0.24	0.49
77.5	0.21	0.36	0.23	0.81
78.9	0.82	0.33	0.53	0.37
81.8	0.20	0.43	0.58	1.00
156.5	0.79	0.55	0.24	1.00
226.5	0.93	0.95	0.74	0.38
229.0	0.06	0.18	0.38	0.39
234.0	0.36	0.20	0.55	1.00
235.4	0.28	0.09	1.00	0.84
238.3	0.82	0.05	1.00	0.50

16



17

CONSIDERATIONS FOR FURTHER ANALYSIS

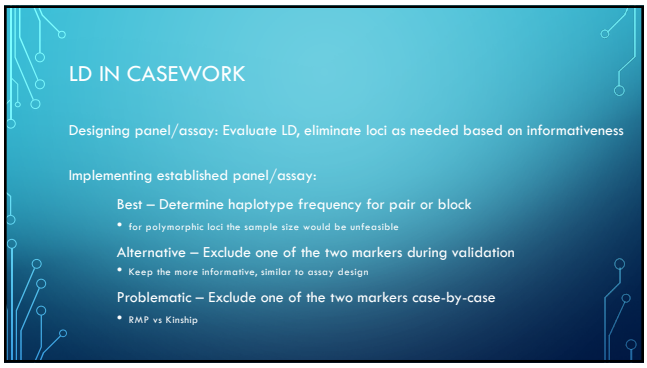
When LD is detected:

- Ideally rule out technical issues by testing on different platforms/assays
- Ideally confirm with multiple sample sets from same population, and multiple test methods
- Polymorphic STRs with p-value = 0 may be due to high number of alleles
  - Increase permutations in analysis
  - Bin STR alleles by length or motif
  - Sequence more samples

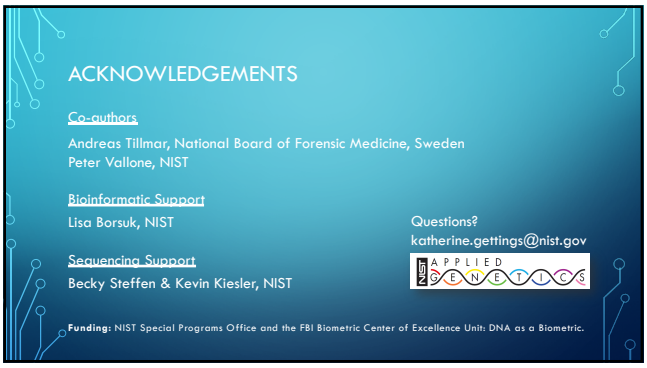
If no p-values are significant after correction:

- Could limit testing to syntenic pairs, pairs within 50Mb, LD blocks
- Even if p-value is not significant, explore data for pairs causing lower p-values

18


LD IN CASEWORK
Designing panel/assay: Evaluate LD, eliminate loci as needed based on informativeness
Implementing established panel/assay:
Best – Determine haplotype frequency for pair or block
• for polymorphic loci the sample size would be unfeasible
Alternative – Exclude one of the two markers during validation
• Keep the more informative, similar to assay design
Problematic – Exclude one of the two markers case-by-case
• RMP vs Kinship

19


ACKNOWLEDGEMENTS
Co-authors
Andreas Tillmar, National Board of Forensic Medicine, Sweden
Peter Vallone, NIST
Bioinformatic Support
Lisa Borsuk, NIST
Questions?
katherine.gettings@nist.gov
Sequencing Support
Becky Steffen & Kevin Kiesler, NIST
Funding: NIST Special Programs Office and the FBI Biometric Center of Excellence Unit: DNA as a Biometric.

20