Investigative Genetic Genealogy from Highly Mixed Samples

Ellen Greytak, PhD
Director of Bioinformatics
Parabon NanoLabs

Snapshot DNA ANALYSIS

PARABON NANOLABS

© 2023 Parabon NanoLabs, Inc. All rights reserved.

1

Takeaway Message

PARABON

- A SNP profile that can be uploaded to genetic genealogy (GG) databases is **NOT** enough for a GG case to be workable
- The data has to be of **high enough quality** that the GG database matches are reliable (i.e. the amount of shared DNA reflects the true relationship)
- Low-quality data (e.g. mixtures) can result in missing matches, spurious matches, and/or incorrect DNA sharing amounts, which make GG extremely difficult

2

Data for Genetic Genealogy

PARABON

- Genome-wide SNP data:

# rsid	chromosome	position	genotype
rs3094315	1	752566	AA
rs3131972	1	752721	GG
rs12124819	1	776546	AA
rs11240777	1	798959	AG
rs6681049	1	800007	CC
rs4970383	1	838555	AC
rs4475691	1	846808	CT
rs7537756	1	854250	AG
rs13302982	1	861808	GG
rs1110052	1	873558	GG
rs2272756	1	882033	AA
rs3748597	1	888659	CC
rs13303106	1	891945	AA
rs28415373	1	893981	CC
rs13303010	1	894573	AA
rs6696281	1	903104	CC
rs28391282	1	904165	GG
...850,000 more rows...			

3

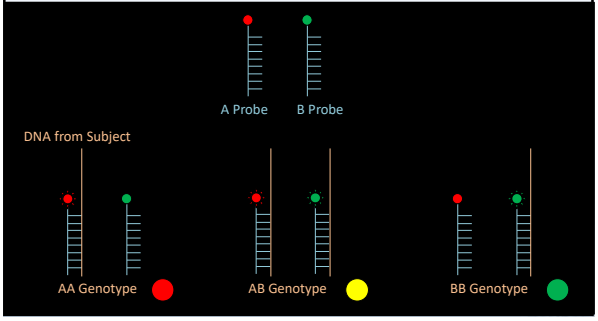
Generate SNP Data: Microarray

- Microarray Genotyping
  - Cost effective
  - CytoSNP-850K chip designed for lower-quality samples
  - Appropriate for most samples (≥1 ng)
  - GREAT for mixtures!

4

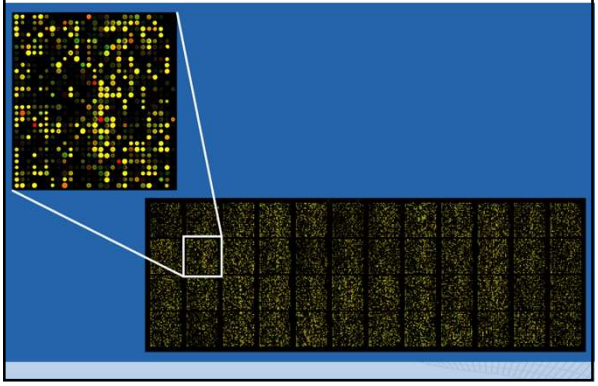
Generate SNP Data: Microarray

- 50-base pair fluorescent probes test for the presence of each allele



5

Generate SNP Data: Microarray



6

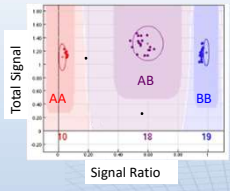
Generate SNP Data: Microarray

PARABON

- Scanner measures the amount of fluorescent signal of each color at each probe
- Software clusters signals to call genotypes

Colored Points: Samples with called genotypes at this SNP

Black Points: Samples with missing genotypes at this SNP



- Assess quality using **call rate** (proportion of targeted SNPs with a genotype call)

7

Data for Genetic Genealogy

PARABON

# rsid	chromosome	position	genotype
rs3094315	1	752566	AA
rs3131972	1	752721	GG
rs12124819	1	776546	AA
rs11240777	1	798959	AG
rs6681049	1	800007	CC
rs4970383	1	838555	AC
rs4475691	1	846808	CT
rs7537756	1	854250	AG
rs13302982	1	861808	GG
rs1110052	1	873558	GG
rs2272756	1	882033	AA
rs3748597	1	888659	CC
rs13303106	1	891945	AA
rs28415373	1	893981	CC
rs13303010	1	894573	AA
rs6696281	1	903104	CC
rs28391282	1	904165	GG

8

Data for Genetic Genealogy

PARABON

- Upload forensic SNP data to a GG database

# rsid	chromosome	position	forensic
rs3094315	1	752566	AA
rs3131972	1	752721	AG
rs12124819	1	776546	GG
rs11240777	1	798959	AA
rs6681049	1	800007	CT
rs4970383	1	838555	AC
rs4475691	1	846808	TT
rs7537756	1	854250	GG
rs13302982	1	861808	AA
rs1110052	1	873558	GG
rs2272756	1	882033	AA
rs3748597	1	888659	CC
rs13303106	1	891945	GG
rs28415373	1	893981	CT
rs13303010	1	894573	AA
rs6696281	1	903104	CT
rs28391282	1	904165	AG

9

IBD Segment Detection

PARABON

› How does the GG database determine whether they're related?

# rsid	chromosome	position	database	forensic
rs3094315	1	752566	AA	AA
rs3131972	1	752721	GG	AG
rs12124819	1	776546	AA	GG
rs11240777	1	798959	AG	AA
rs6681049	1	800007	CC	CT
rs4970383	1	838555	AC	AC
rs4475691	1	846808	CT	TT
rs7537756	1	854250	AG	GG
rs13302982	1	861808	GG	AA
rs1110052	1	873558	GG	GG
rs2272756	1	882033	AA	AA
rs3748597	1	888659	CC	CC
rs13303106	1	891945	AA	GG
rs28415373	1	893981	CC	CT
rs13303010	1	894573	AA	AA
rs6696281	1	903104	CC	CT
rs28391282	1	904165	GG	AG

10

IBD Segment Detection

PARABON

Find the shared Identical-By-Descent (IBD) segments

# rsid	chromosome	position	database	forensic
rs3094315	1	752566	AA	AA
rs3131972	1	752721	GG	AG
rs12124819	1	776546	AA	GG
rs11240777	1	798959	AG	AA
rs6681049	1	800007	CC	CT
rs4970383	1	838555	AC	AC
rs4475691	1	846808	CT	TT
rs7537756	1	854250	AG	GG
rs13302982	1	861808	GG	AA
rs1110052	1	873558	GG	GG
rs2272756	1	882033	AA	AA
rs3748597	1	888659	CC	CC
rs13303106	1	891945	AA	GG
rs28415373	1	893981	CC	CT
rs13303010	1	894573	AA	AA
rs6696281	1	903104	CC	CT
rs28391282	1	904165	GG	AG

11

IBD Segment Detection

PARABON

Identical-By-State (IBS) 0 = Definitely not IBD

# rsid	chromosome	position	database	forensic
rs3094315	1	752566	AA	AA
rs3131972	1	752721	GG	AG
rs12124819	1	776546	AA ⟷	GG
rs11240777	1	798959	AG	AA
rs6681049	1	800007	CC	CT
rs4970383	1	838555	AC	AC
rs4475691	1	846808	CT	TT
rs7537756	1	854250	AG	GG
rs13302982	1	861808	GG ⟷	AA
rs1110052	1	873558	GG	GG
rs2272756	1	882033	AA	AA
rs3748597	1	888659	CC	CC
rs13303106	1	891945	AA ⟷	GG
rs28415373	1	893981	CC	CT
rs13303010	1	894573	AA	AA
rs6696281	1	903104	CC	CT
rs28391282	1	904165	GG	AG

12

IBD Segment Detection

PARABON

IBS1 or IBS2 = Maybe IBD

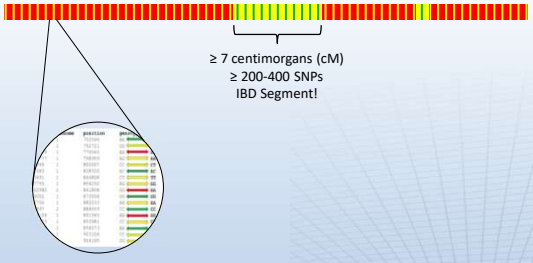
# rsid	chromosome	position	database	forensic
rs3094315	1	752566	AA	AA
rs3131972	1	752721	GG	AG
rs12124819	1	776546	AA	GG
rs11240777	1	798959	AG	AA
rs6681049	1	800007	CC	CT
rs4970383	1	838555	AC	AC
rs4475691	1	846808	CT	TT
rs7537756	1	854250	AG	GG
rs13302982	1	861808	GG	AA
rs1110052	1	873558	GG	GG
rs2272756	1	882033	AA	AA
rs3748597	1	888659	CC	CC
rs13303106	1	891945	AA	GG
rs28415373	1	893981	CC	CT
rs13303010	1	894573	AA	AA
rs6696281	1	903104	CC	CT
rs28391282	1	904165	GG	AG

13

IBD Segment Detection

PARABON

Chromosome 1



≥ 7 centimorgans (cM)
≥ 200-400 SNPs
IBD Segment!

14

Mixed Samples

PARABON

- DNA from the person-of-interest (POI) is mixed with DNA from another person, usually the victim
- A single-source sample from the victim is usually available
- What effect does this have on the data and the GG database matches?

15

Mixed Sample Genotypes: 20%

- 80% POI + 20% Victim → "20% Mixture"
- AA + AA → AA — Correct
- AB + AB → AB — Correct
- BB + BB → BB — Correct
- AA/BB + AB → AA/BB — Correct or Missing
- AB + AA/BB → AB — Correct or Missing
- AA + BB → NN — Missing
- SNPs are **correct** or **missing**

16

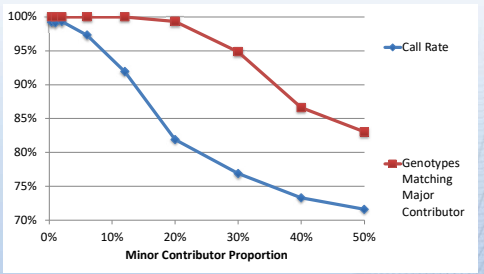
Mixed Sample Genotypes: 50%

- 50% POI + 50% Victim → "50% Mixture"
- AA + AA → AA — Correct
- AB + AB → AB — Correct
- BB + BB → BB — Correct
- AA/BB + AB → NN — Missing
- AB + AA/BB → NN — Missing
- AA + BB → AB — Incorrect
- Called SNPs are either **incorrect heterozygotes** or **shared**

17

Microarray for Mixed Samples

- Laboratory-created mixtures → genotyping
- Mixing decreases call rate and accuracy



18

Mixtures

PARABON

› Increased heterozygosity → increased sharing

# rsid	chromosome	position	database	forensic
rs3094315	1	752566	AA	AA
rs3131972	1	752721	GG	AG
rs12124819	1	776546	AA	GG
rs11240777	1	798959	AG	AA
rs6681049	1	800007	CC	CT
rs4970383	1	838555	AC	AC
rs4475691	1	846808	CT	TT
rs7537756	1	854250	AG	GG
rs13302982	1	861808	GG	AA
rs1110052	1	873558	GG	GG
rs2272756	1	882033	AA	AA
rs3748597	1	888659	CC	CC
rs13303106	1	891945	AA	GG
rs28415373	1	893981	CC	CT
rs13303010	1	894573	AA	AA
rs6696281	1	903104	CC	CT
rs28391282	1	904165	GG	AG

19

Mixtures

PARABON

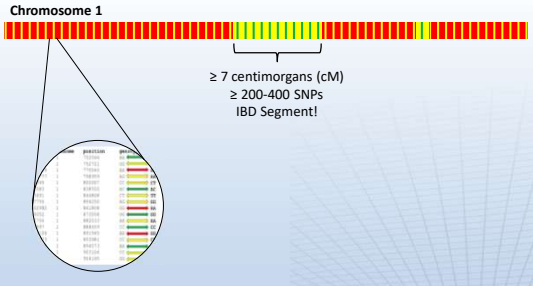
› Increased heterozygosity → increased sharing

# rsid	chromosome	position	database	forensic	
rs3094315	1	752566	AA	AA	
rs3131972	1	752721	GG	AG	
rs12124819	1	776546	AA	GA	IBS0→1
rs11240777	1	798959	AG	AA	
rs6681049	1	800007	CC	CT	
rs4970383	1	838555	AC	AC	
rs4475691	1	846808	CT	TC	IBS1→2
rs7537756	1	854250	AG	GG	
rs13302982	1	861808	GG	AA	
rs1110052	1	873558	GG	GA	IBS2→1
rs2272756	1	882033	AA	AA	
rs3748597	1	888659	CC	CC	
rs13303106	1	891945	AA	GA	IBS0→1
rs28415373	1	893981	CC	CT	
rs13303010	1	894573	AA	AA	
rs6696281	1	903104	CC	CT	
rs28391282	1	904165	GG	AG	

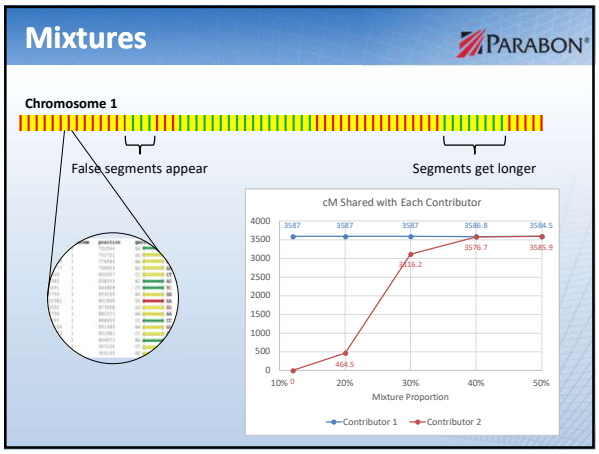
20

Single Source

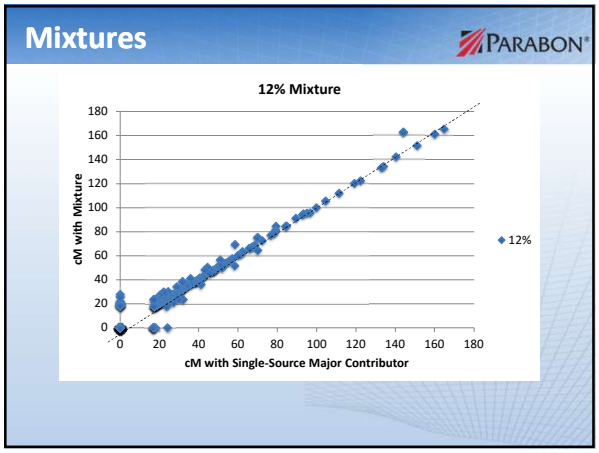
PARABON



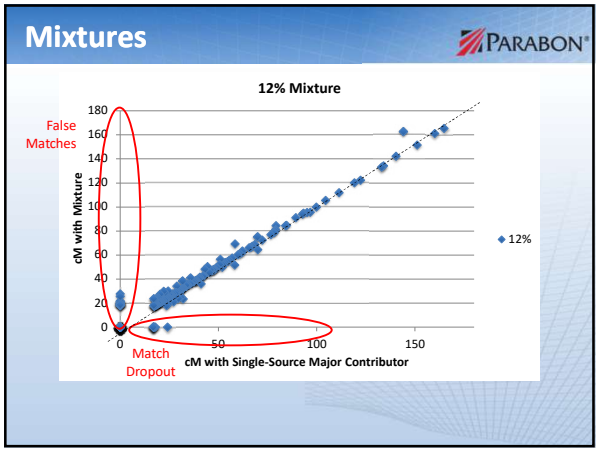
21


Mixtures
PARABON
Chromosome 1
False segments appear
Segments get longer
cM Shared with Each Contributor
Mixture Proportion
Contributor 1
Contributor 2

22


Mixtures
PARABON
12% Mixture
cM with Mixture
cM with Single-Source Major Contributor
12%

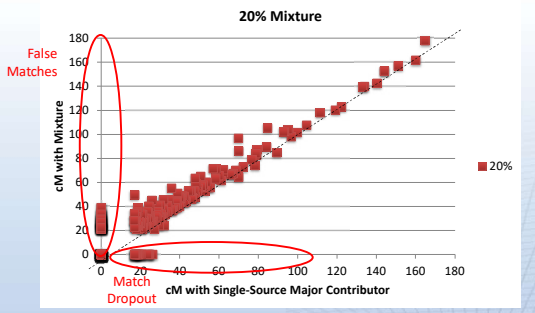
23


Mixtures
PARABON
12% Mixture
False Matches
Match Dropout
cM with Mixture
cM with Single-Source Major Contributor
12%

24

Mixtures

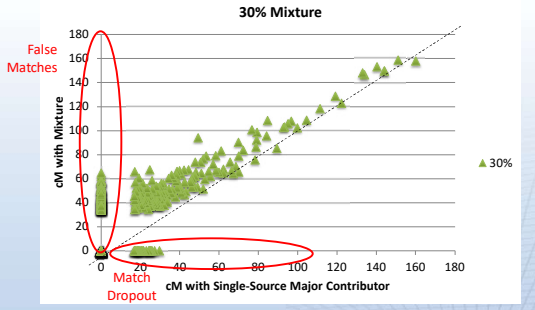
PARABON



25

Mixtures

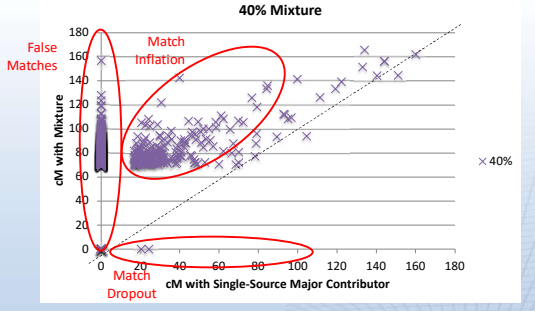
PARABON



26

Mixtures

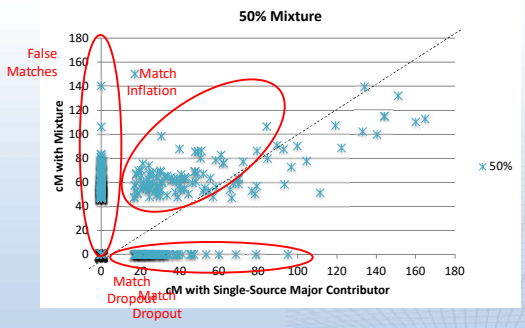
PARABON



27

Mixtures

PARABON

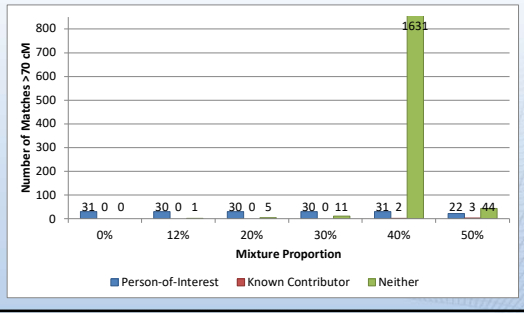


28

Mixtures

PARABON

- A mixture's match list **IS NOT** just a mixture of the contributors' matches!



29

Mixtures

PARABON

- Now when you upload a mixture to GEDmatch…

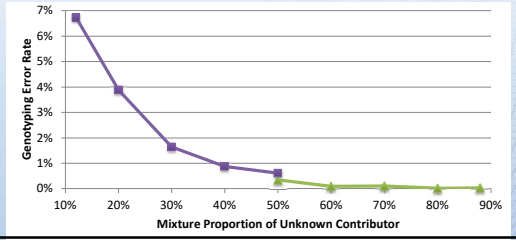
Fails HTZ ratio test

- Ouch.

30

Microarray for Mixed Samples

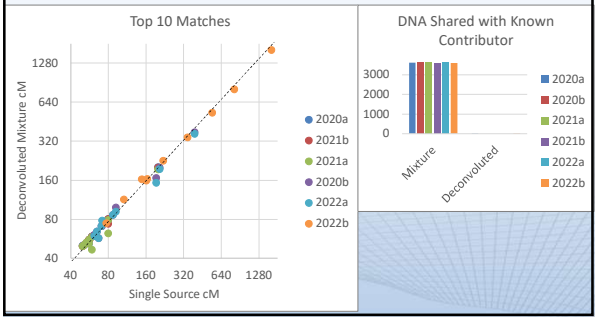
- Mixture deconvolution
- Trained a machine learning model to predict genotypes of unknown contributor from mixed sample and single source known contributor



31

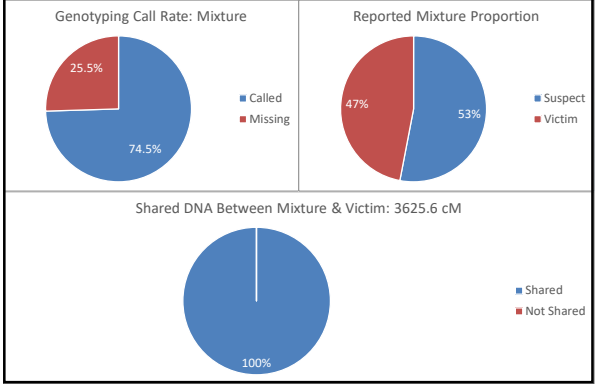
Mixtures for GG

- Biannual proficiency test results: deconvoluted data from 30-40% mixtures



32

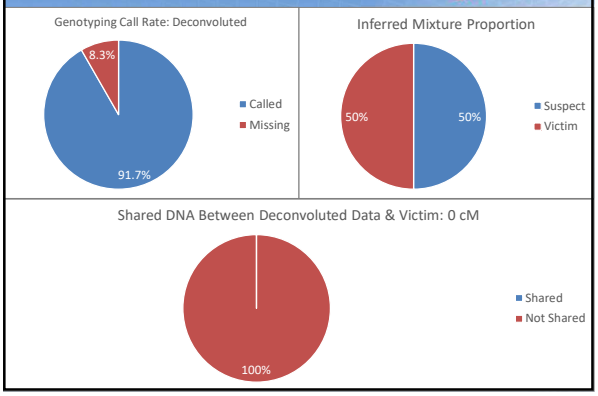
Mixture Case Study #1



33

Mixture Case Study #1

PARABON



34

Mixture Case Study #1

PARABON

- First ever high-level mixture used for GG (June 2018)
- Top match: **>3100 cM**
- Chromosomes all yellow
- Parent-child match!
- BUT…

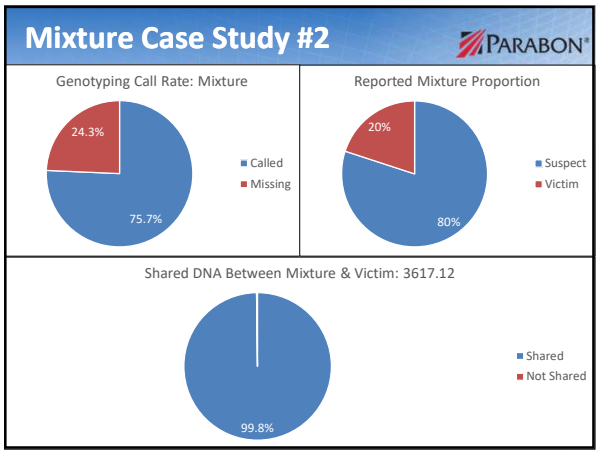
35

Mixture Case Study #1

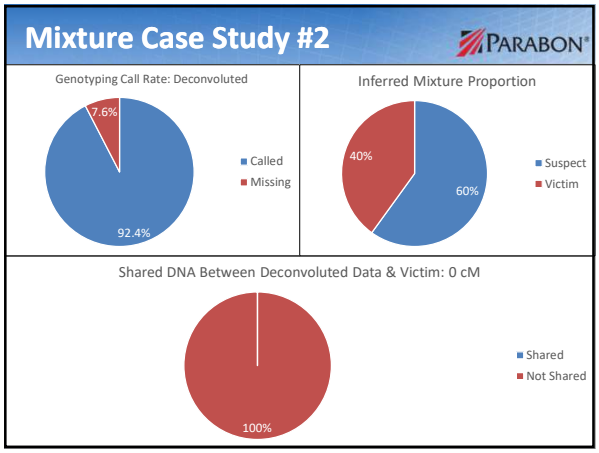
PARABON

- First ever high-level mixture used for GG (June 2018)
- Parent-child match
- BUT…
- Match has the **same last name** as the victim – Oh No! Did something go wrong with the deconvolution?
- *PANIC*
- Genealogy determined…just a weird coincidence!
- (Phew!)

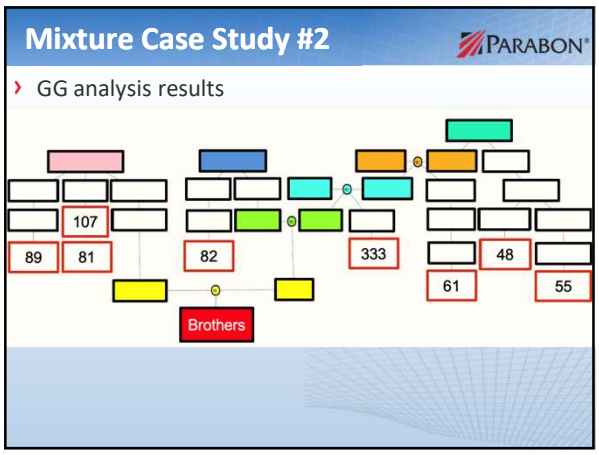
36


Mixture Case Study #2
PARABON
Genotyping Call Rate: Mixture
24.3%
75.7%
Called
Missing
Reported Mixture Proportion
20%
80%
Suspect
Victim
Shared DNA Between Mixture & Victim: 3617.12
99.8%
Shared
Not Shared

37


Mixture Case Study #2
PARABON
Genotyping Call Rate: Deconvoluted
7.6%
92.4%
Called
Missing
Inferred Mixture Proportion
40%
60%
Suspect
Victim
Shared DNA Between Deconvoluted Data & Victim: 0 cM
100%
Shared
Not Shared

38

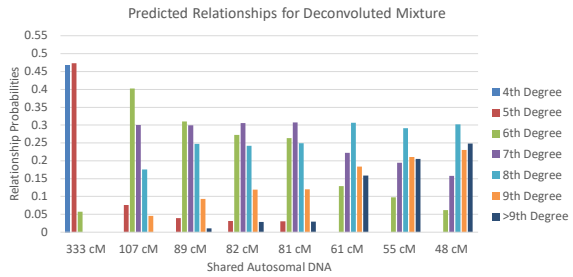

Mixture Case Study #2
PARABON
GG analysis results
107
89
81
82
333
48
61
55
Brothers

39

Mixture Case Study #2

PARABON

- Relationships they *should* have based on IBD sharing

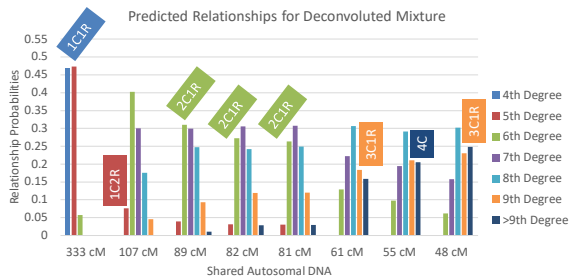


40

Mixture Case Study #2

PARABON

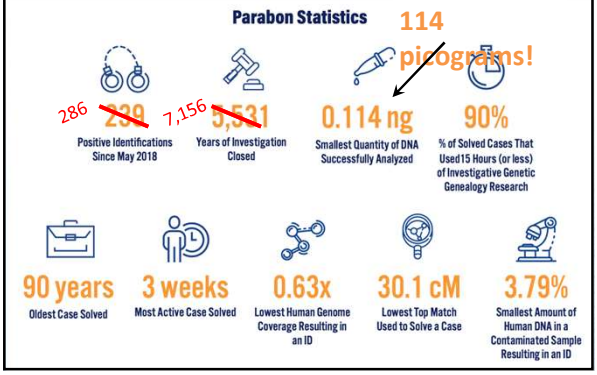
- Relationships they *actually* have based on genealogy



41

Questions?? ellen@parabon.com

PARABON



42