Advancing Forensic DNA Analysis in Support of DoD Missions

PRESENTED BY:

Dr. Ellen Greytak

Director of Bioinformatics, Parabon NanoLabs

MODERATED BY:

Steve Redifer

2020-05-21

Homeland Defense & Security Information Analysis Center

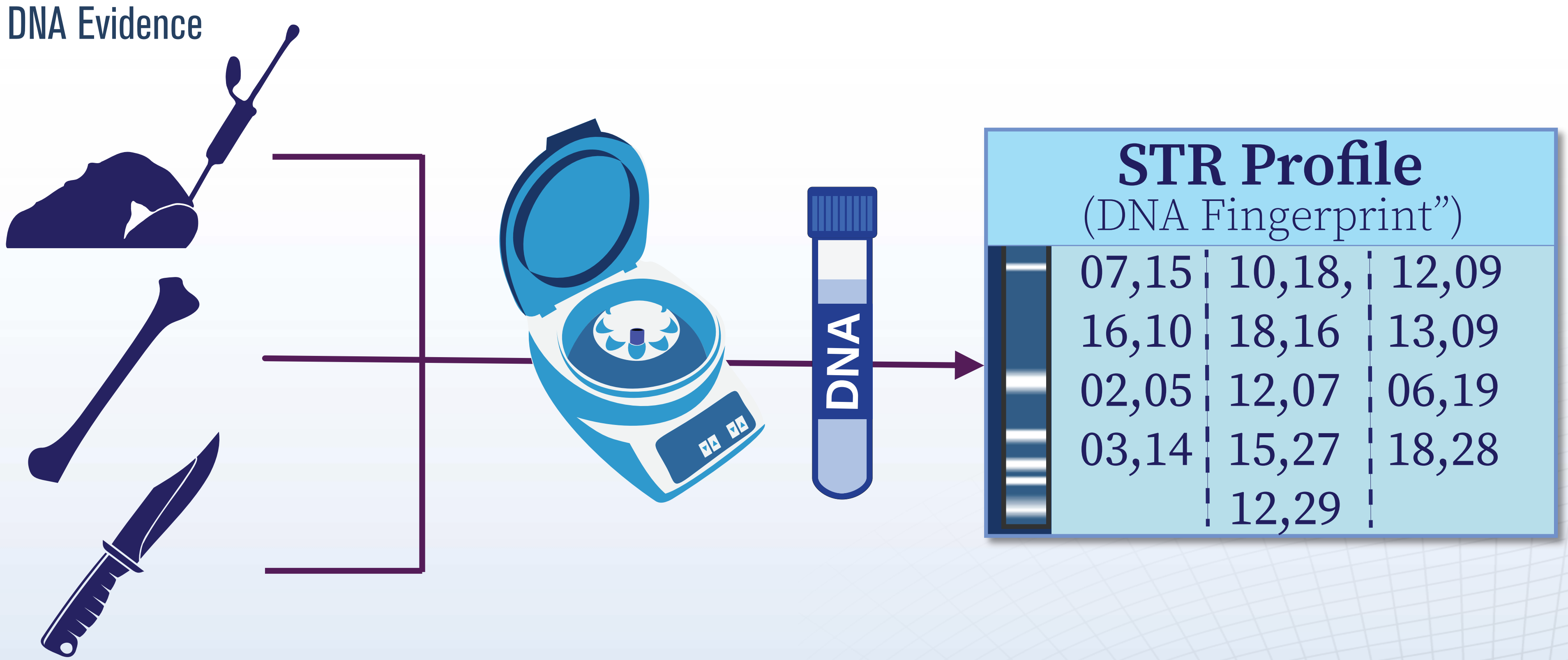
info@hdiac.org
https://www.hdiac.org

Advancing Forensic DNA Analysis in Support of DoD Missions

Ellen Greytak, PhD
Director of Bioinformatics
Parabon NanoLabs, Inc.

PARABON®
NANOLABS

Traditional Forensic DNA Analysis

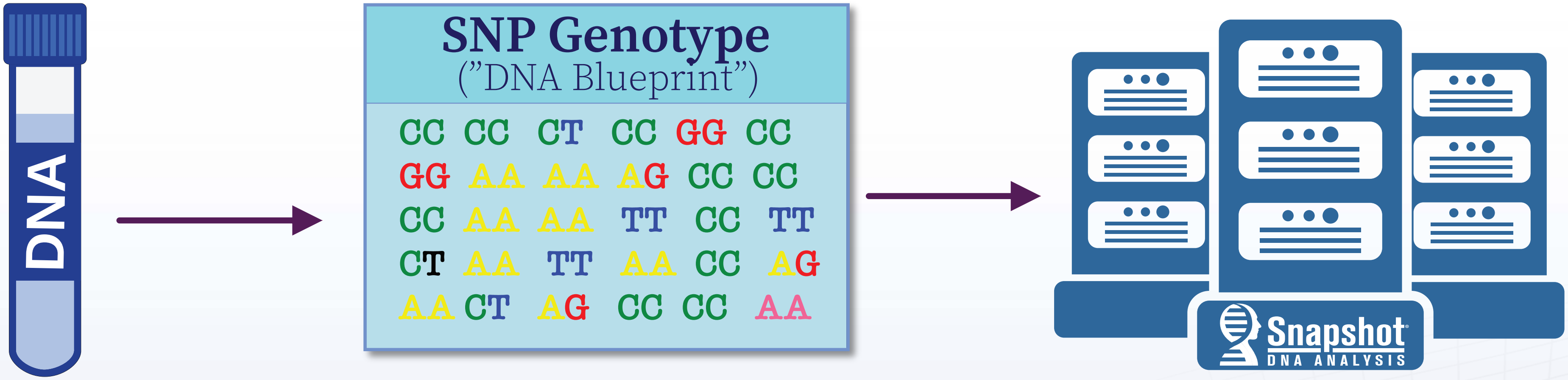


STR Profile
(DNA Fingerprint”)

07,15	10,18,	12,09
16,10	18,16	13,09
02,05	12,07	06,19
03,14	15,27	18,28
	12,29	

- Short Tandem Repeats (STRs)
- Identity questions ONLY
  - Is this person in a database (e.g., CODIS)?
  - Is this person a suspect we’ve already identified?
  - Definitively establishes identity

Advanced Forensic DNA Analysis



SNP Genotype
(”DNA Blueprint”)

CC CC CT CC GG CC
GG AA AA AG CC CC
CC AA AA TT CC TT
CT AA TT AA CC AG
AA CT AG CC CC AA

- Single Nucleotide Polymorphisms (SNPs)
- Generate NEW leads
  - What does this unknown person look like?
  - Is this unknown person related to someone we DO know? If so, how closely?
  - Is this unknown person related to volunteers in genetic genealogy databases?
  - Generates leads, but identity is established through traditional methods

DoD-Funded Projects

DNA Phenotyping	• DTRA Phase I and Phase II SBIRs • Sequential Phase II SBIR
Distant Kinship Analysis	• DTRA SBIR Modification • ARO Contract
Genetic Genealogy	• Direct result of Phenotyping and Kinship projects
Next-Generation Sequencing (NGS) Analysis	• ARO Contract

DoD-Funded Project Success

DNA Phenotyping

- Commercialized for law enforcement

Distant Kinship Analysis

- Commercialized for law enforcement
- Transitioning to AFDIL

Genetic Genealogy

- Commercialized for law enforcement
- 111 solved cases (and counting!)

Next-Generation Sequencing (NGS) Analysis

- Transitioned to DoD forensics labs

DNA Phenotyping

Snapshot®: Predict the physical appearance of an unknown person from DNA

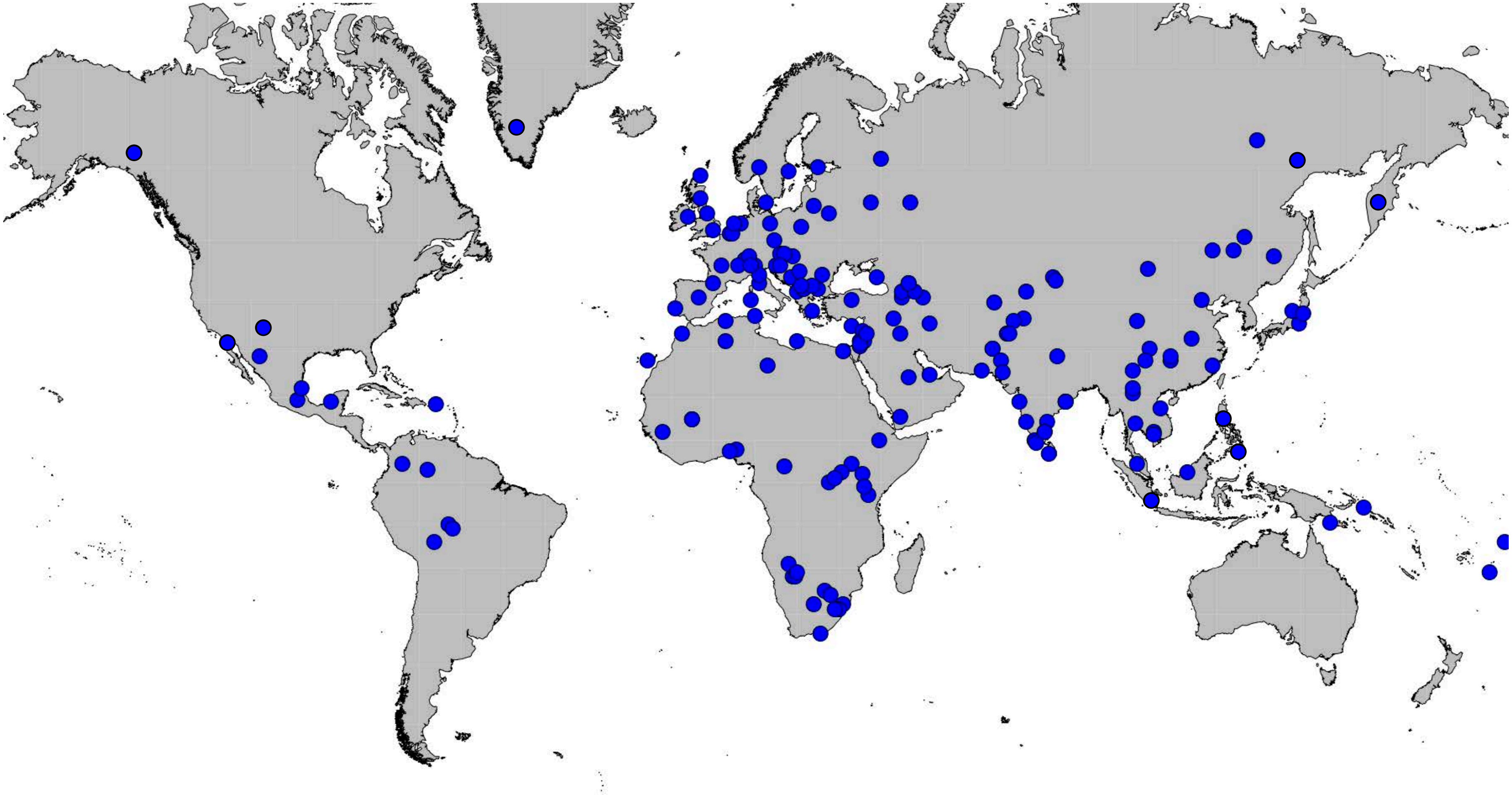
PARABON® NANOLABS

DNA Phenotyping

- **DoD Problem**: in theaters of war, a DNA database may not be available for traditional forensic DNA comparisons
- **DTRA SBIR solicitation goal**: given a DNA sample from an unexploded IED, predict the appearance of the person who planted it

Ancestry Inference

- >13,000 background subjects from >150 populations (aggregated from multiple studies for worldwide coverage)

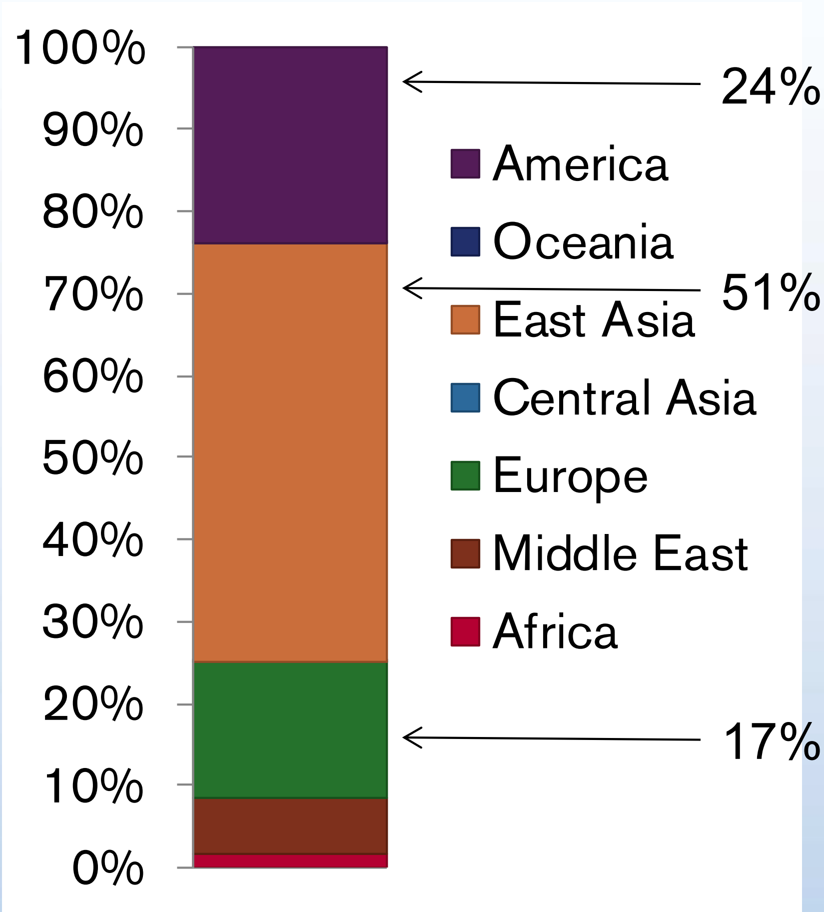


Ancestry Inference

- Tens of thousands of SNPs across the genome
- Very precise estimates of ancestry across as many populations as are defined in the background
- Can detect even low levels of admixture (even when the mixture is between populations we haven't previously observed)

Global Ancestry: Admixed Individual

- Real results from an unknown individual given to us during a blind evaluation test

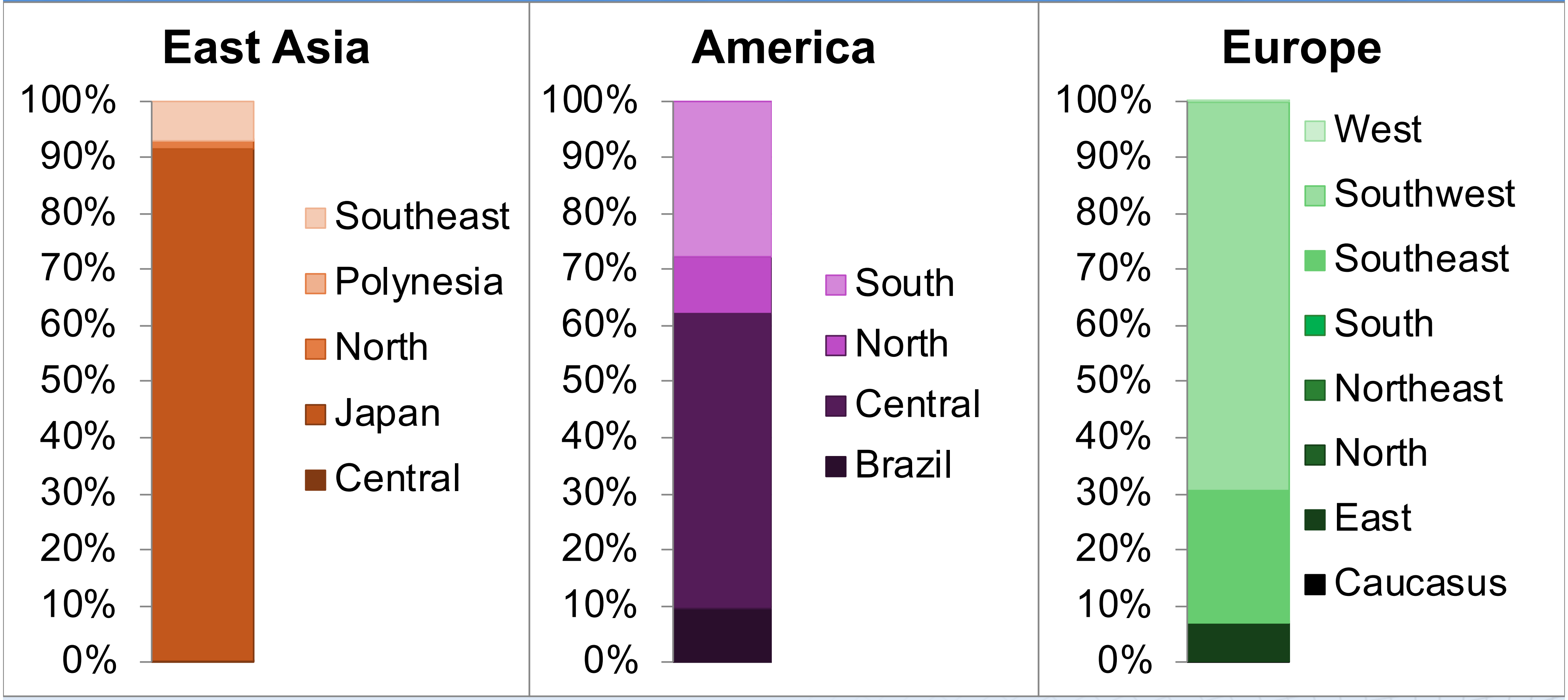


24% Native American ancestry

51% East Asian ancestry

17% European ancestry

Regional Ancestry: Admixed Individual



- Conclusion: This individual is half Japanese and half Latino
- Y chromosome haplogroup = O1b2a (most common in Japan and Korea)
- Therefore, this individual most likely has a Japanese father and a Central American Latino mother

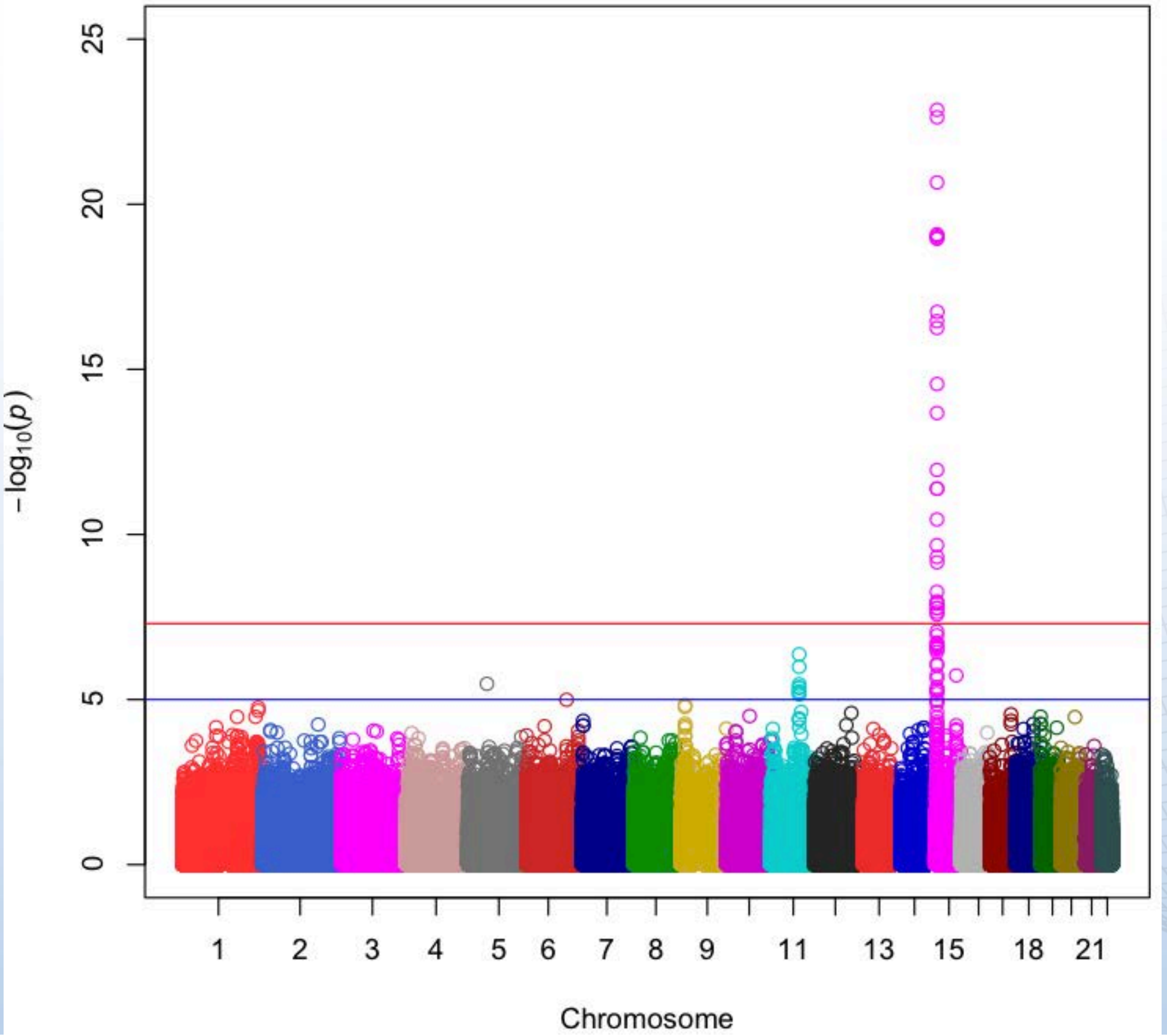
Genome-Wide Association

- Detect statistical associations between SNPs and a heritable phenotype
- Requires a large genotype+phenotype (G+P) database of subjects with known phenotype and genome-wide data

Subject	Eye Color	SNP1 Genotype	SNP2 Genotype	SNP3 Genotype	SNP4 Genotype	…	SNP1,000,000 Genotype
1	Blue	A/G	C/C	G/G	G/G	…	T/T
2	Green	A/A	T/T	A/G	A/G	…	T/T
3	Hazel	G/G	C/T	A/G	A/A	…	T/T
4	Brown	A/A	T/T	A/A	G/G	…	C/T
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3,000	Blue	G/G	T/T	G/G	G/G	…	C/T

Genome-Wide Association

- A strong statistical association suggests the SNP might be biologically involved in the phenotype

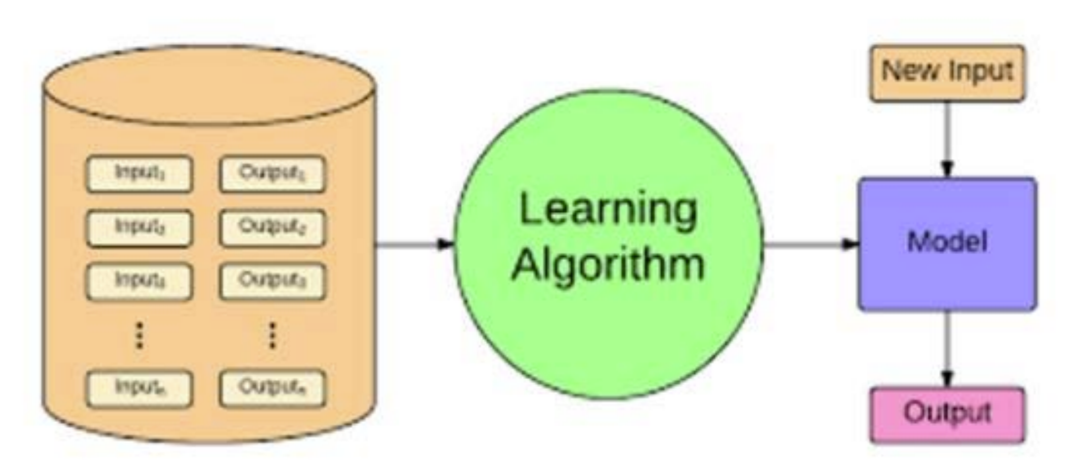


Interaction Analysis

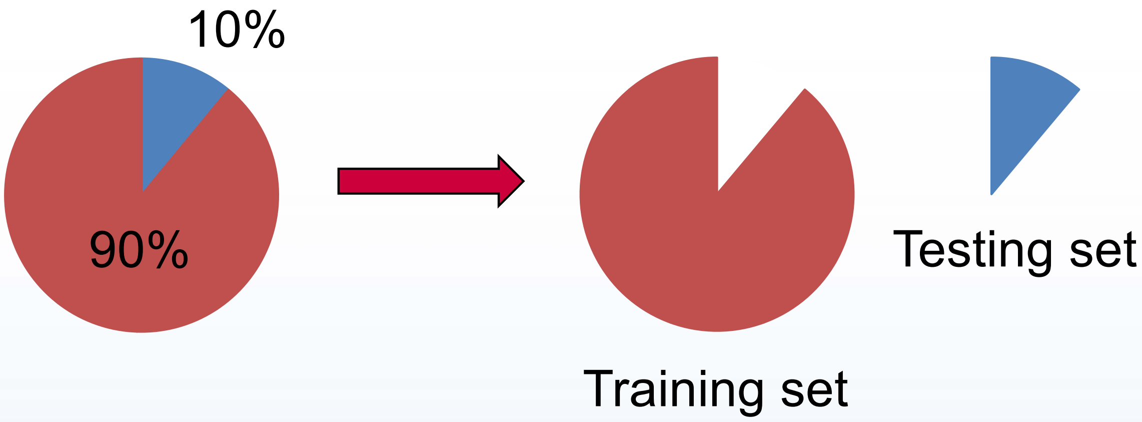
- Single SNP association testing may not capture the whole story
- Many traits are influenced not only by individual SNPs but by non-additive (epistatic) *interactions* among several SNPs
- However, looking for high-order interactions (e.g., 3, 4, and 5 factors) on a genome-wide scale results in a combinatorial explosion of possible models
  - 8,333,250,000,000,000,000,000,000,000 (10^{27}) possible 5-way interactions among 1 million SNPs
- Parabon has developed software (Crush-MDR) that uses a distributed evolutionary search algorithm to explore the massive space of possible interactions

Genotype-to-Phenotype (G2P) Modeling

- Supervised machine learning (ML)
- Train the model using:
  - Known phenotype values
  - Extracted features (selected SNPs, covariates such as ancestry and sex)
- Learn a function to predict outcome from features
- Evaluate accuracy



Cross-Validation



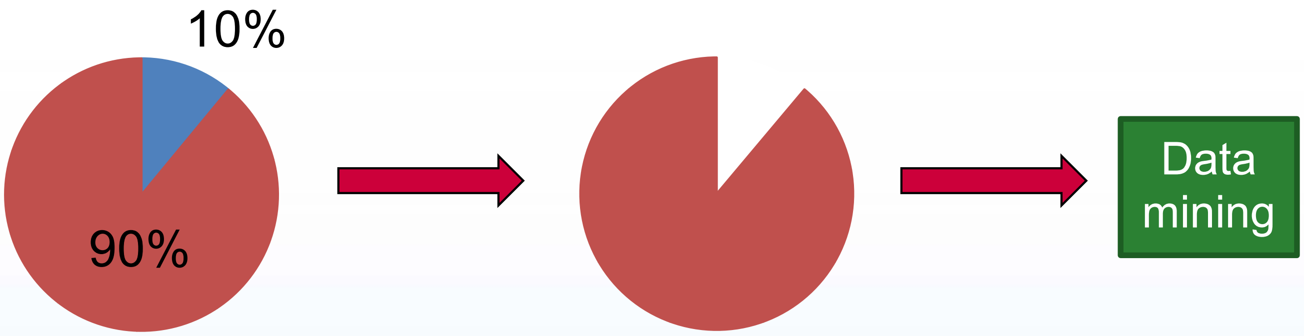
10%

90%

Testing set

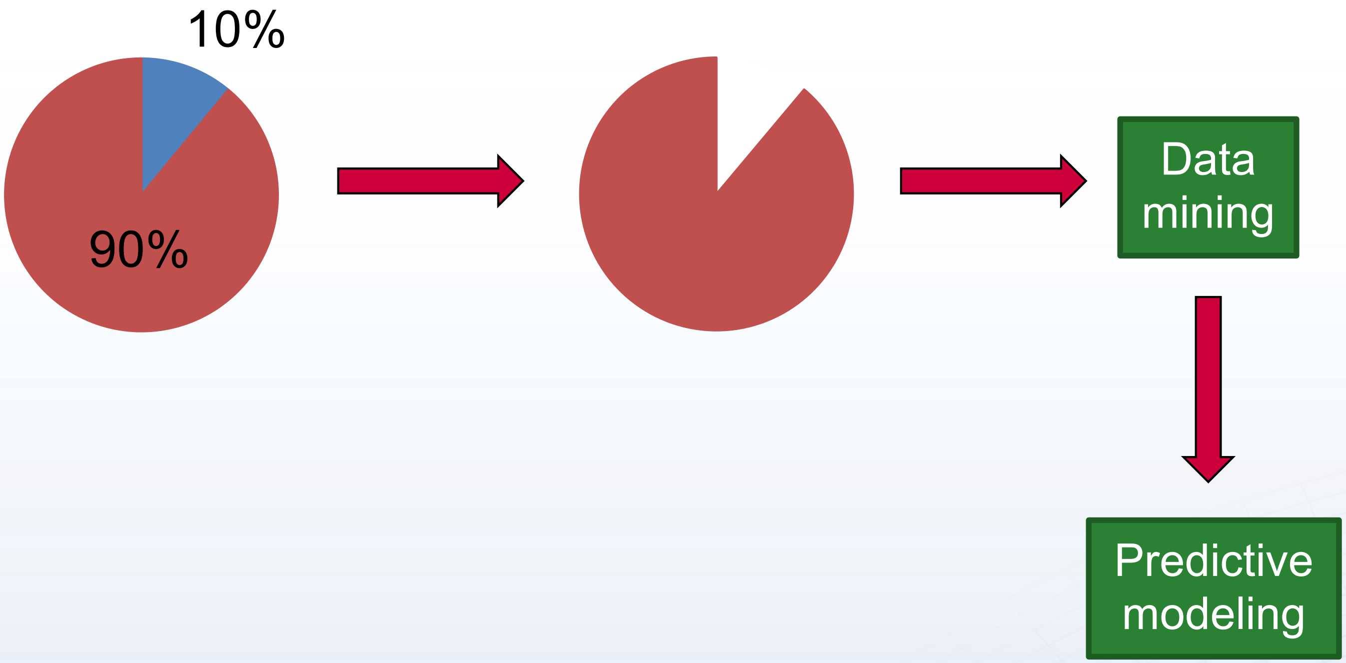
Training set

Cross-Validation



Data mining

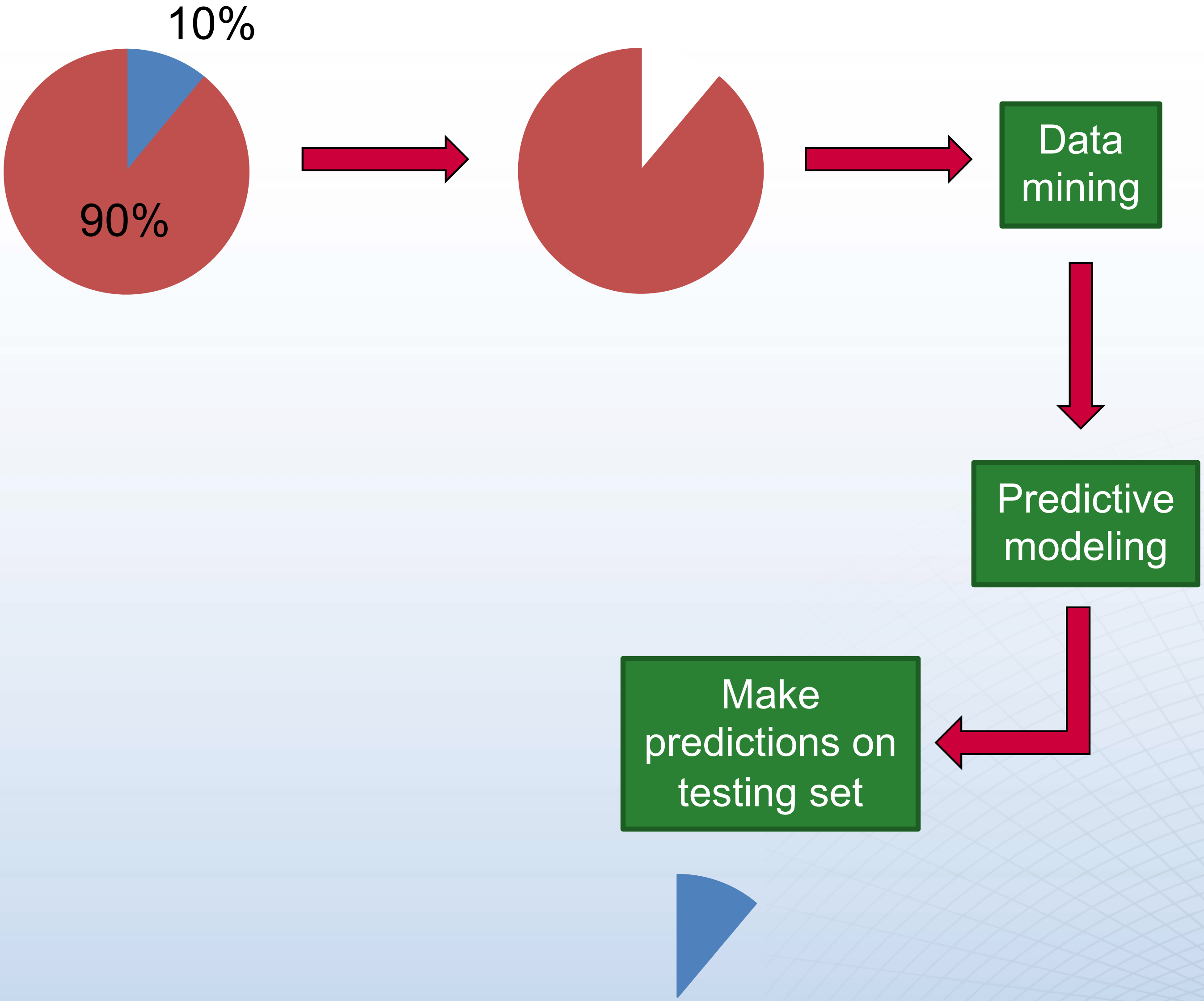
Cross-Validation



Data mining

Predictive modeling

Cross-Validation

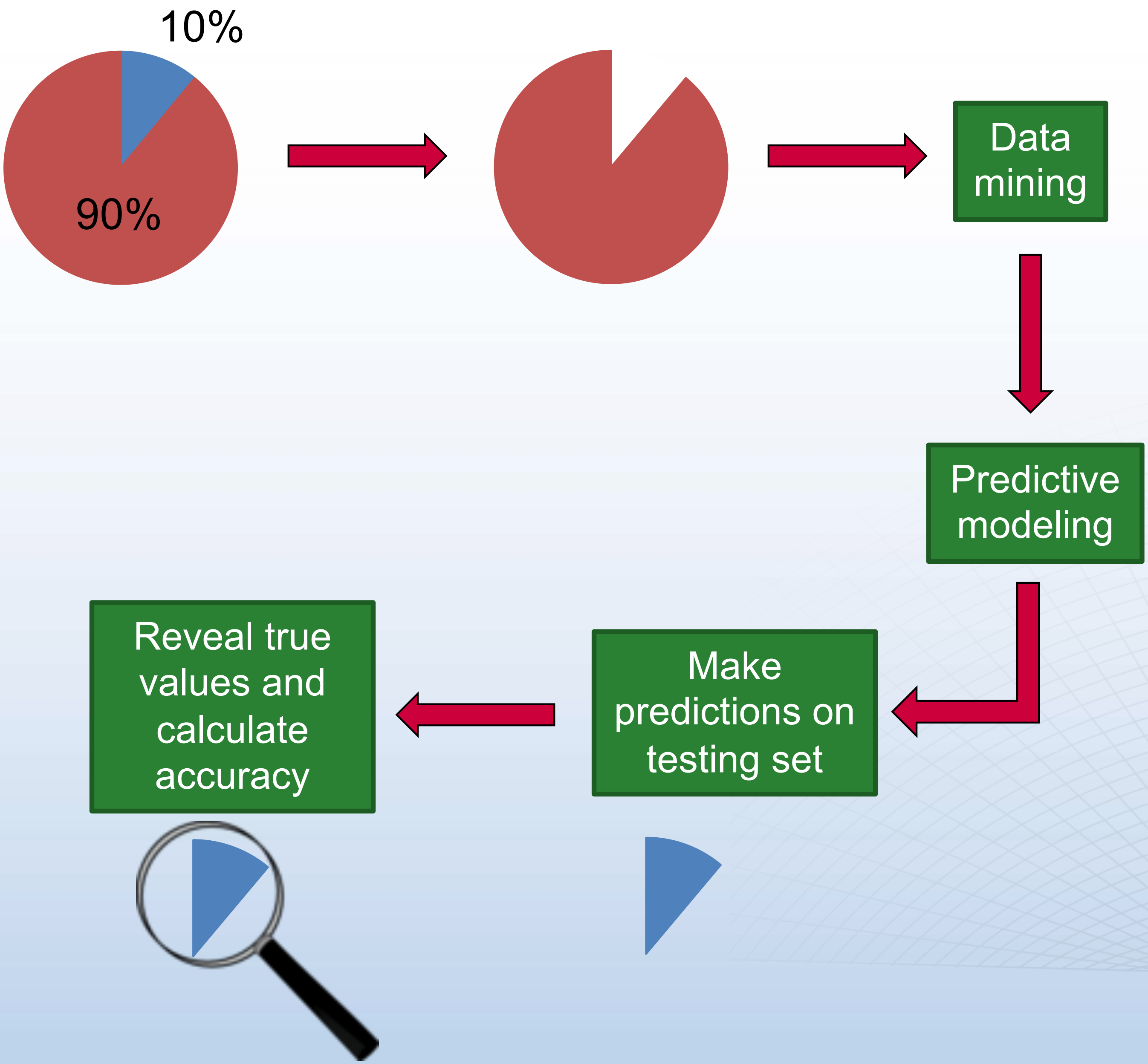


Data mining

Predictive modeling

Make predictions on testing set

Cross-Validation



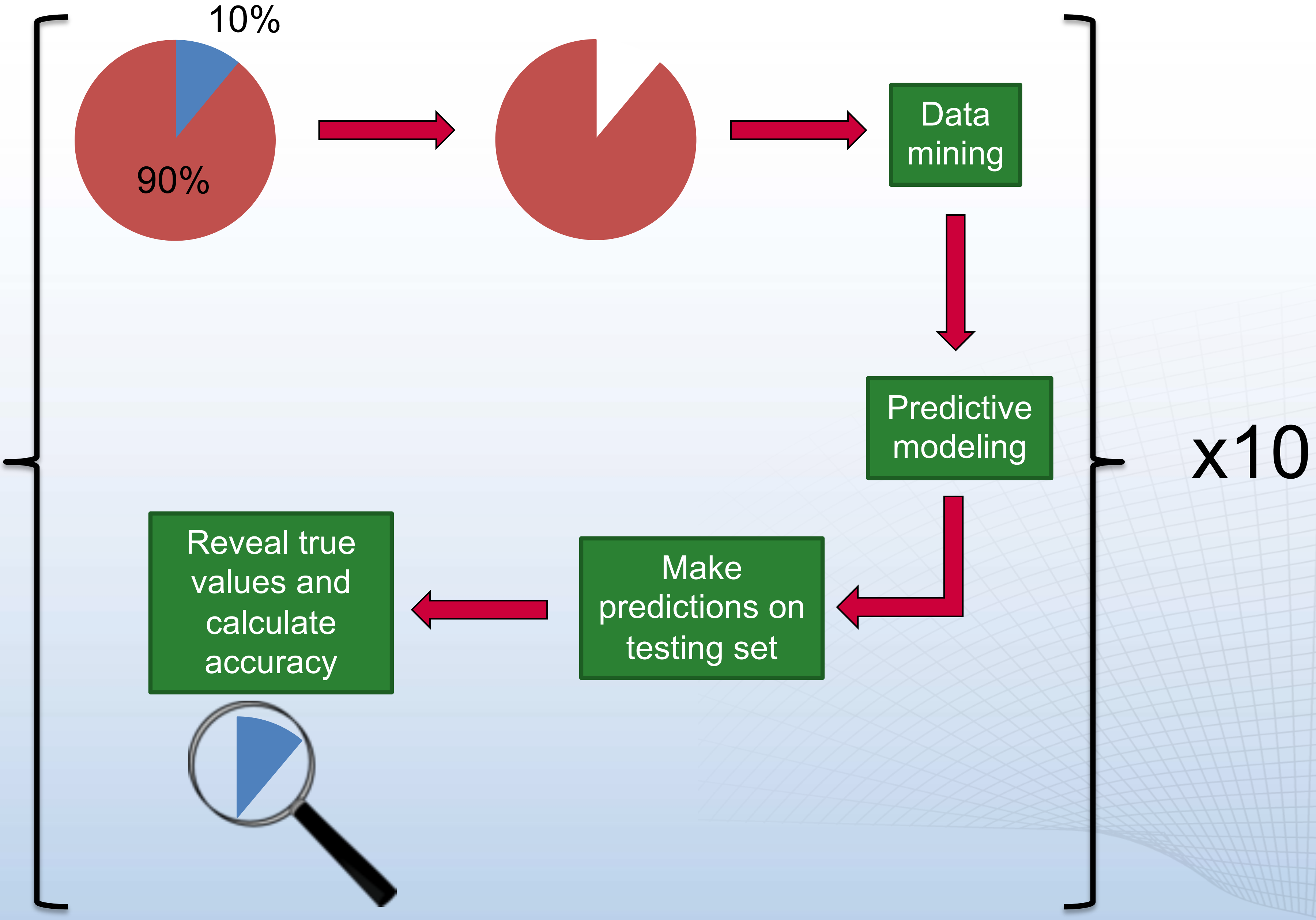
Data mining

Predictive modeling

Make predictions on testing set

Reveal true values and calculate accuracy

Cross-Validation



Data mining

Predictive modeling

Make predictions on testing set

Reveal true values and calculate accuracy

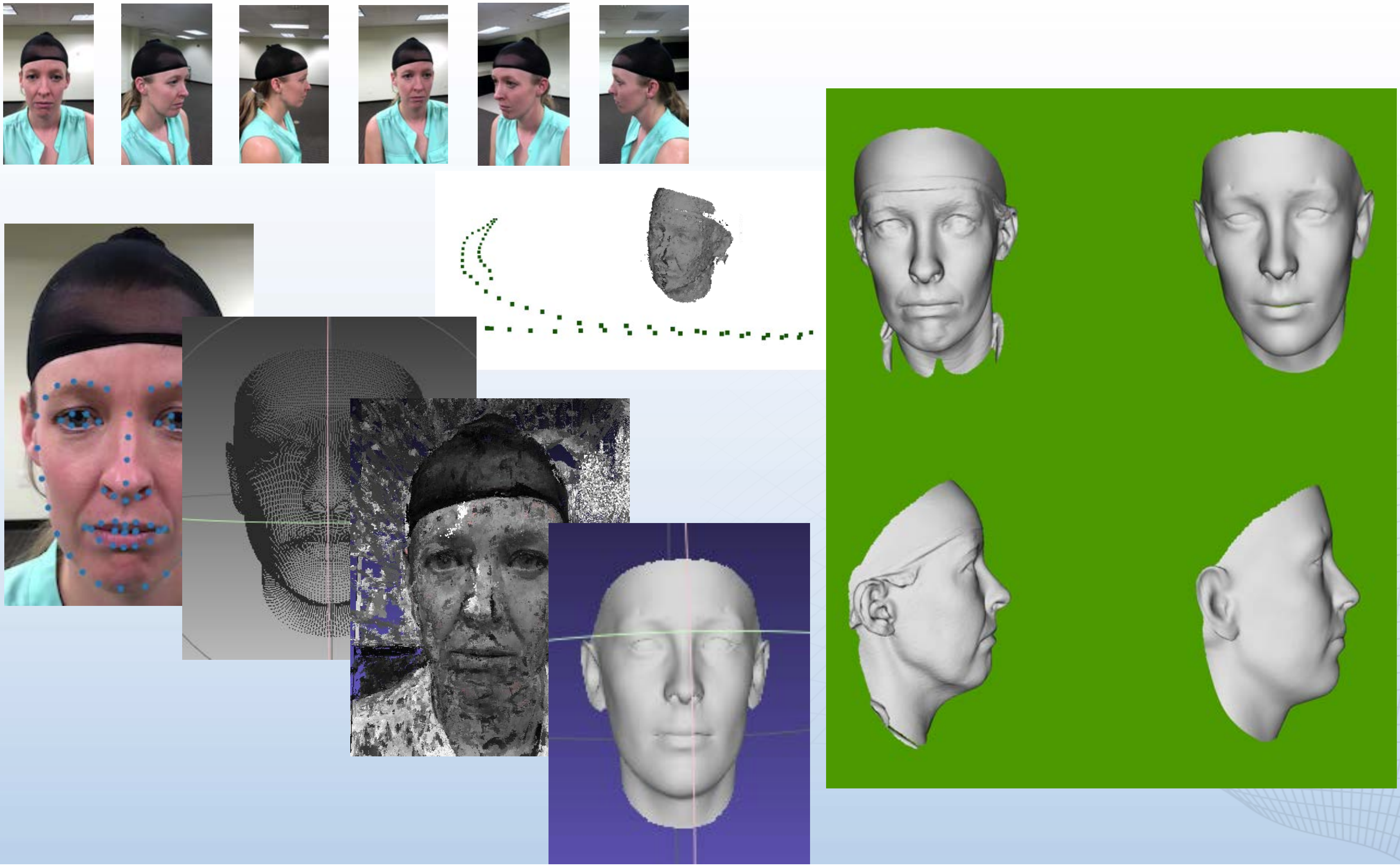
x10

Cross-Validation

- This means we are performing data mining and predictive modeling 10 times for each phenotype
- However, we now have out-of-sample predictions on every single subject in our dataset
- At the end, we build a final model using all of the data, and the cross-validation accuracy approximates the accuracy of this final model
- Use the cross-validation results to make confidence statements about new predictions and *exclude* trait categories that are highly unlikely

Face Morphology

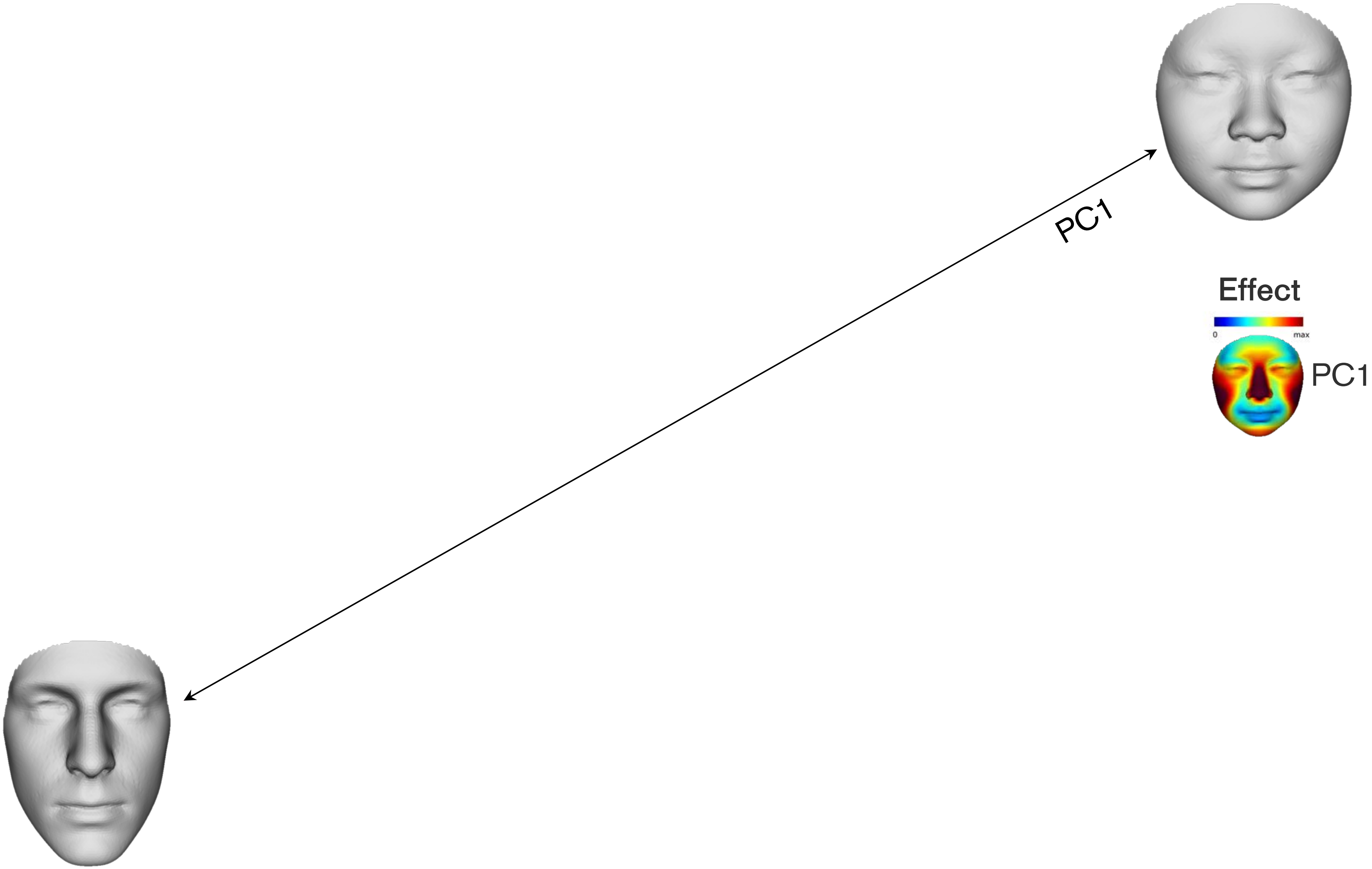
- Can we apply this same process to a complex phenotype, such as face morphology?



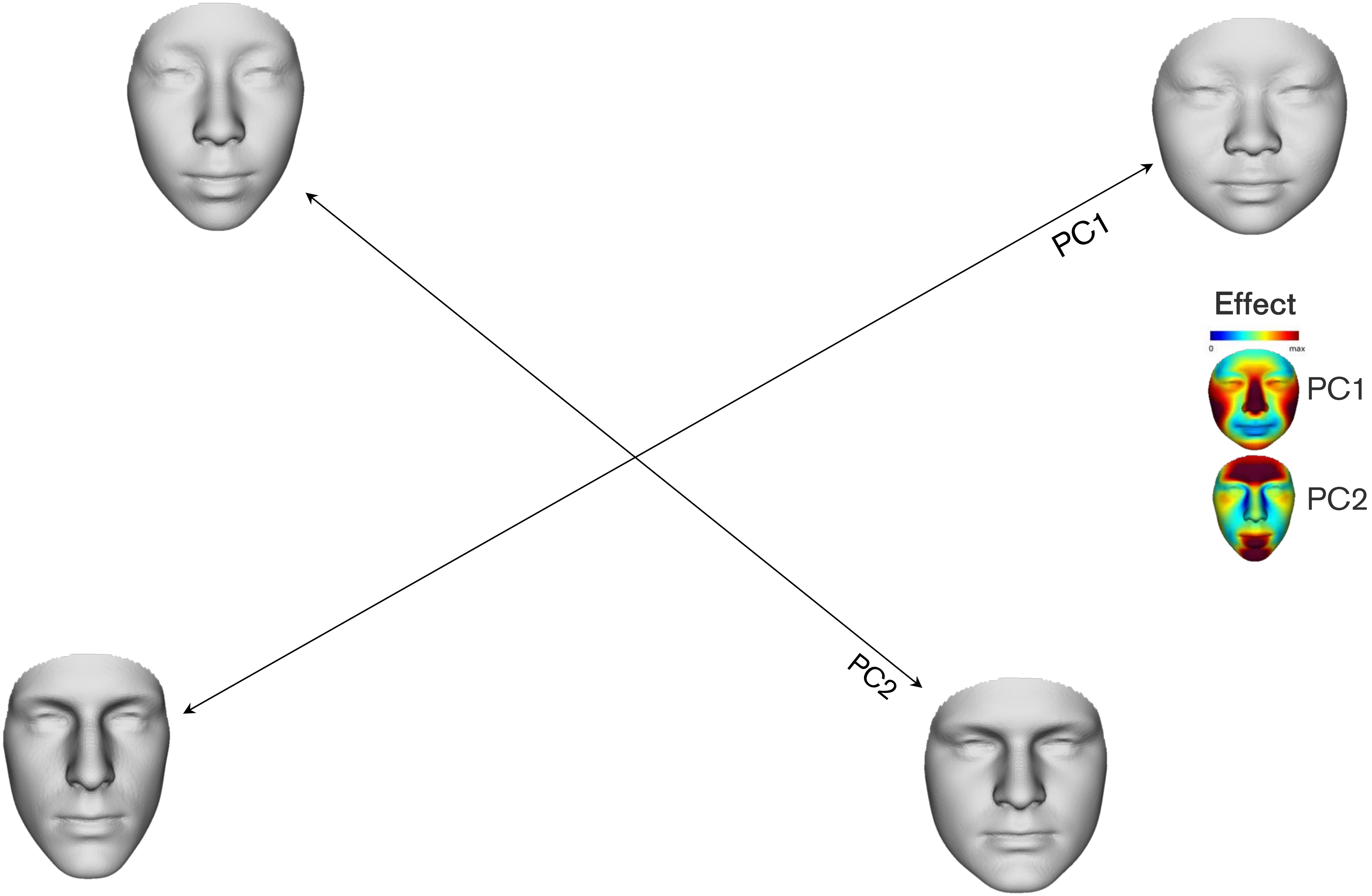
Face Morphology

- Face morphology is captured as a 3D face scan using a smartphone app
- The scan's shape is converted into thousands of quasi-landmarks (QLs), which are homologous across subjects
- Each QL has 3 coordinates (x,y,z), so faces are described by tens of thousands of numerical variables, many of which are correlated with one another
- Principal component analysis (PCA) reduces the number of variables to a reasonable number while still retaining most of the variation

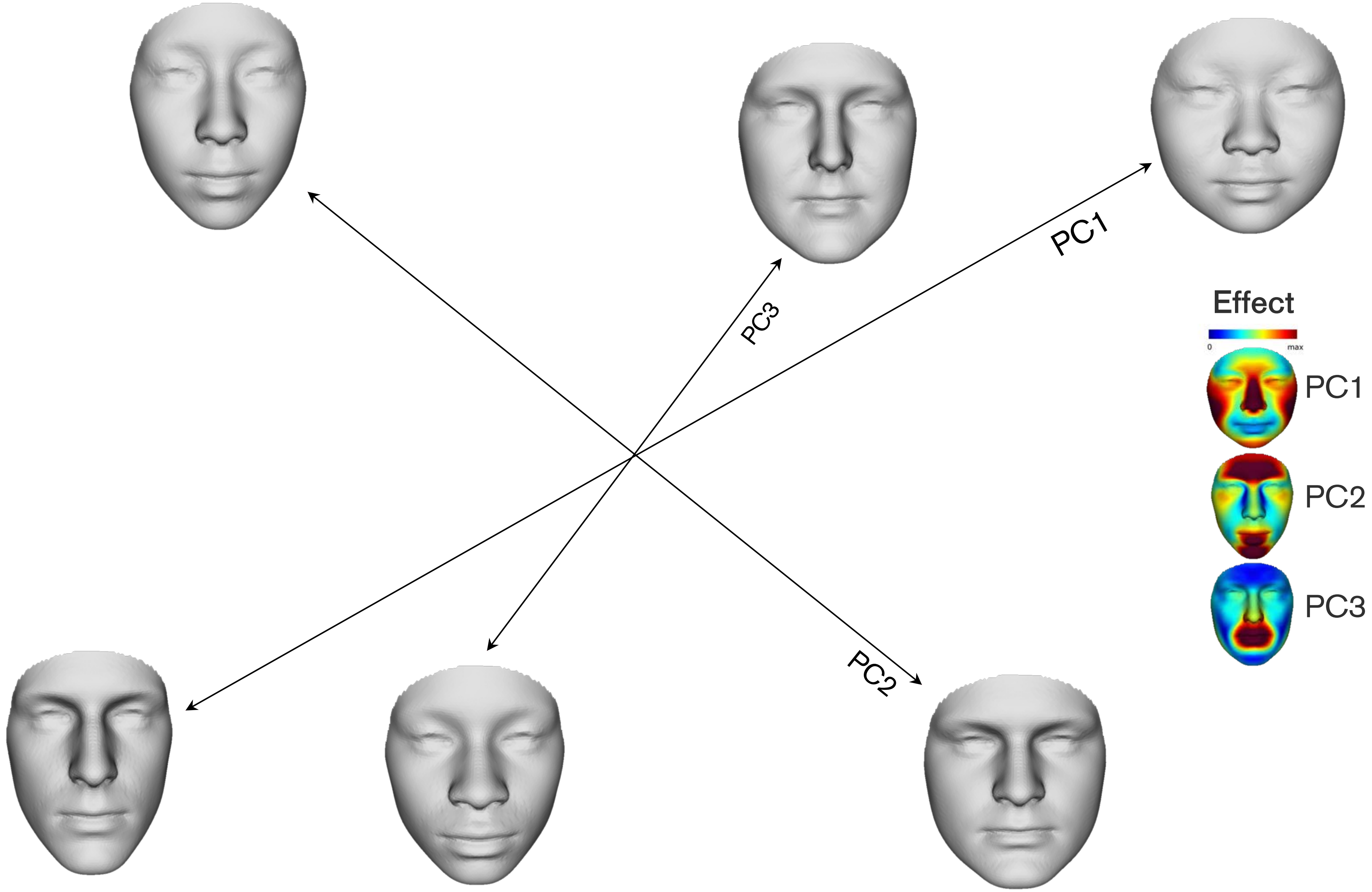
Face Space



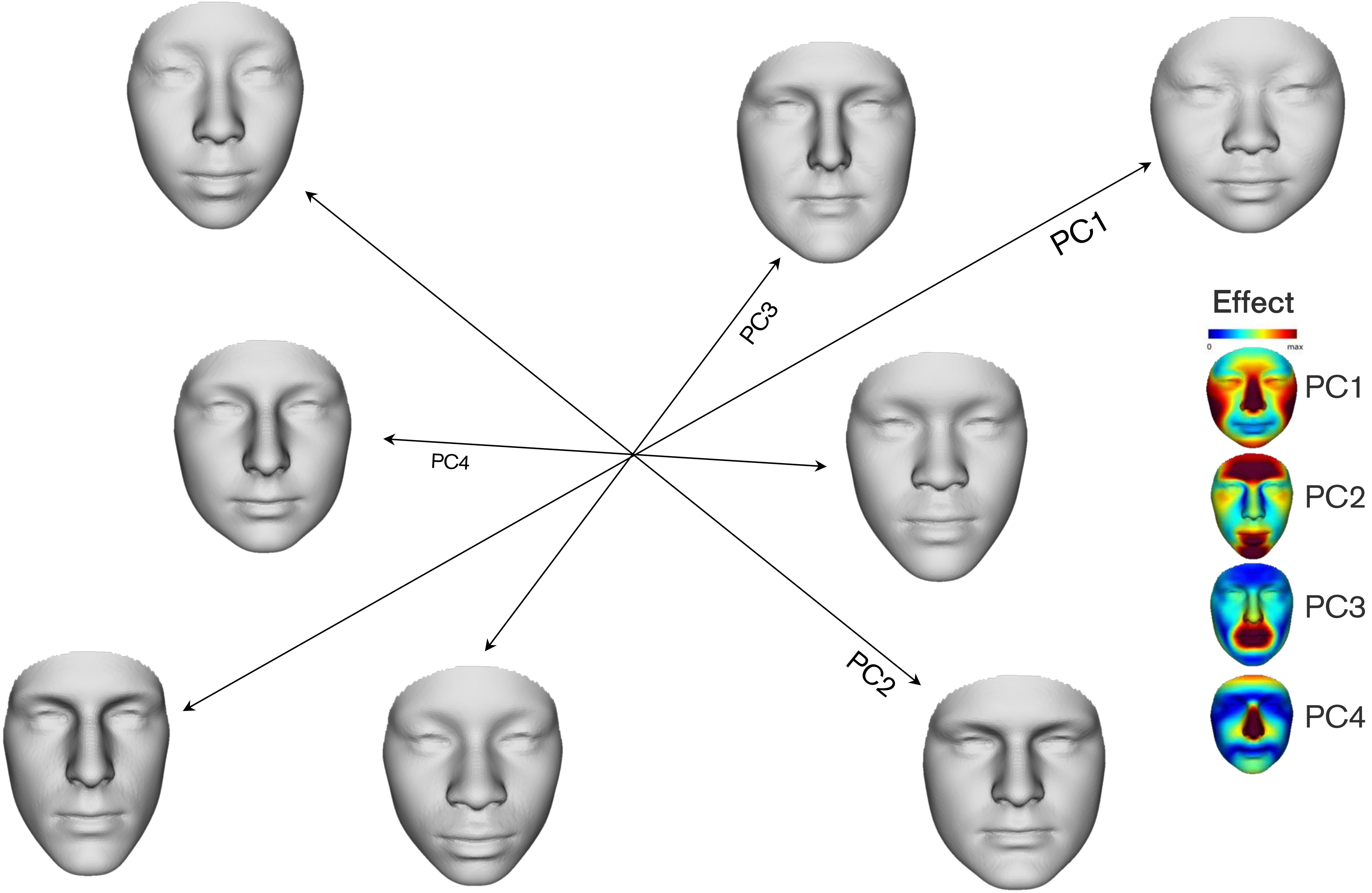
Face Space



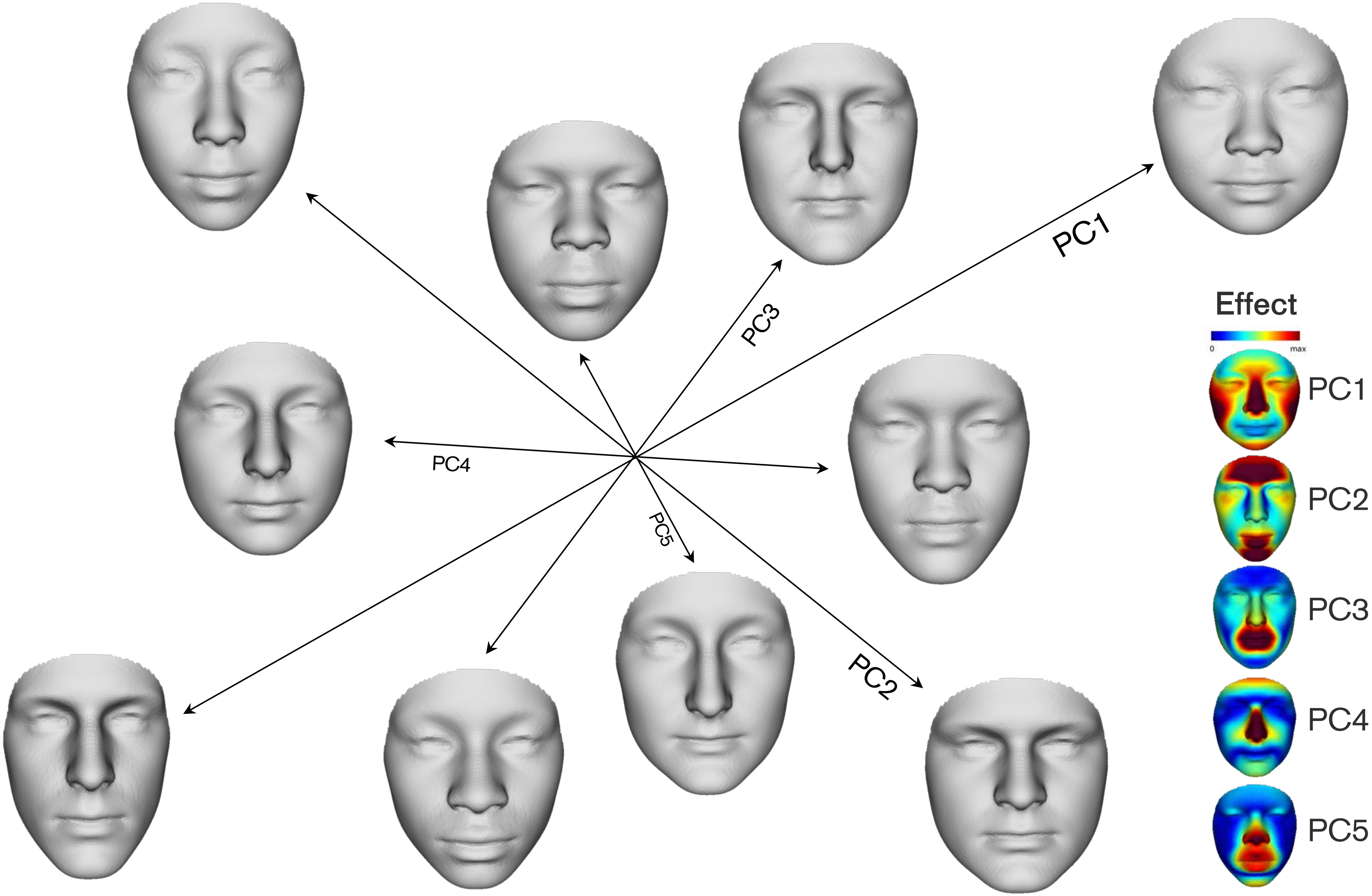
Face Space



Face Space



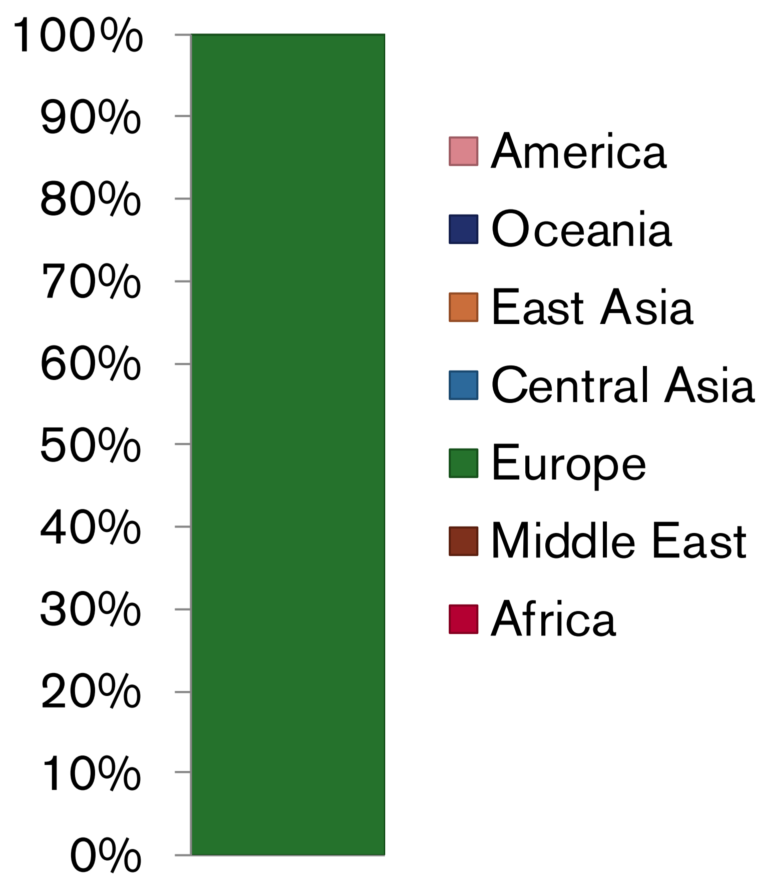
Face Space



Snapshot: Making Predictions

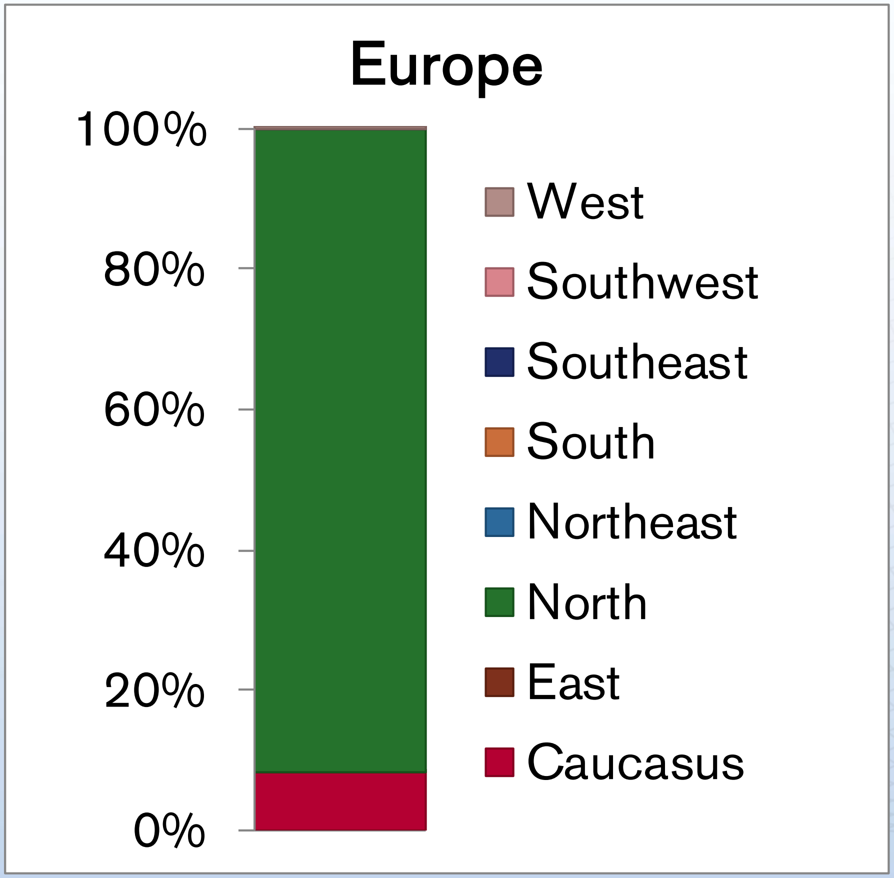
Global Ancestry

- DNA from an unknown subject is partitioned into 7 parts according to its proportional similarity to each of the continental groups



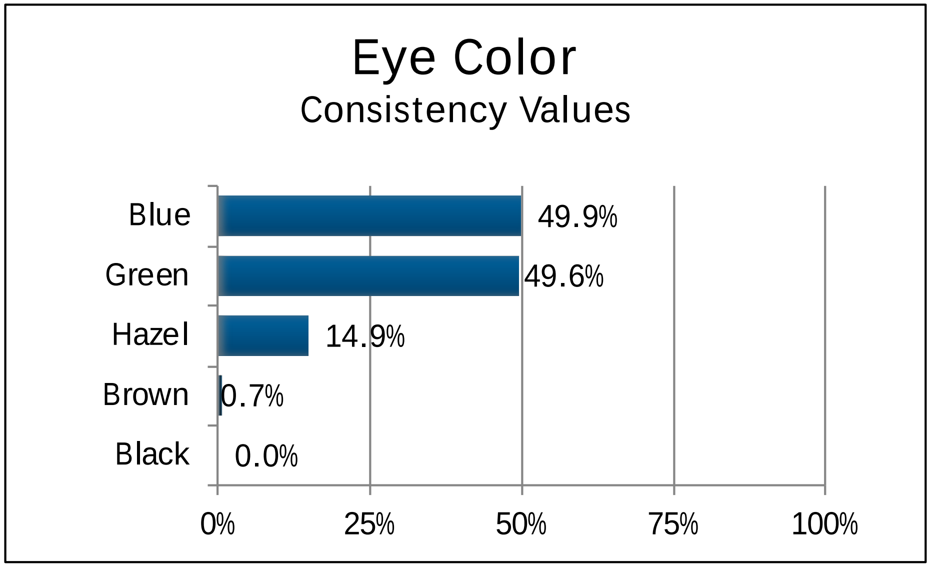
Regional Ancestry

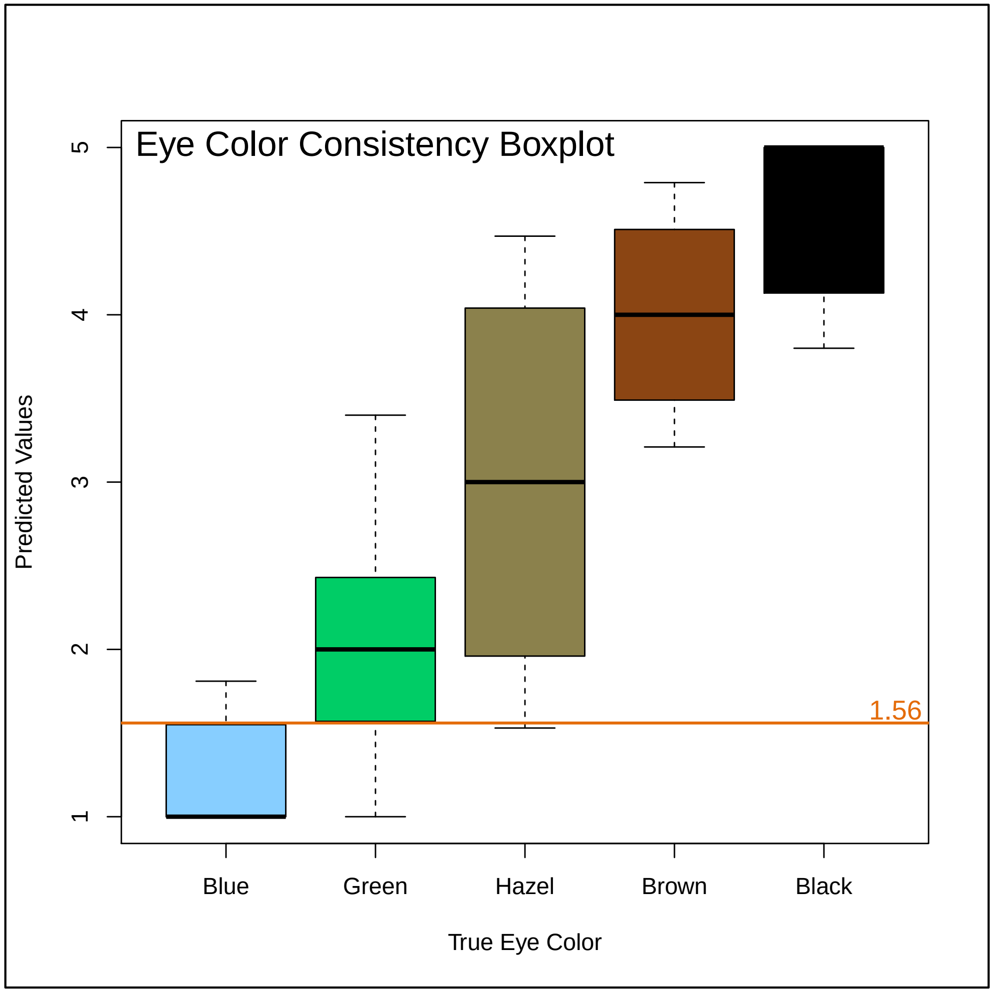
- Within each of the 7 continental populations, we can narrow the source of an individual's ancestry using the same statistical approach with different defined populations



G2P Prediction

- Predicted Value = 1.560

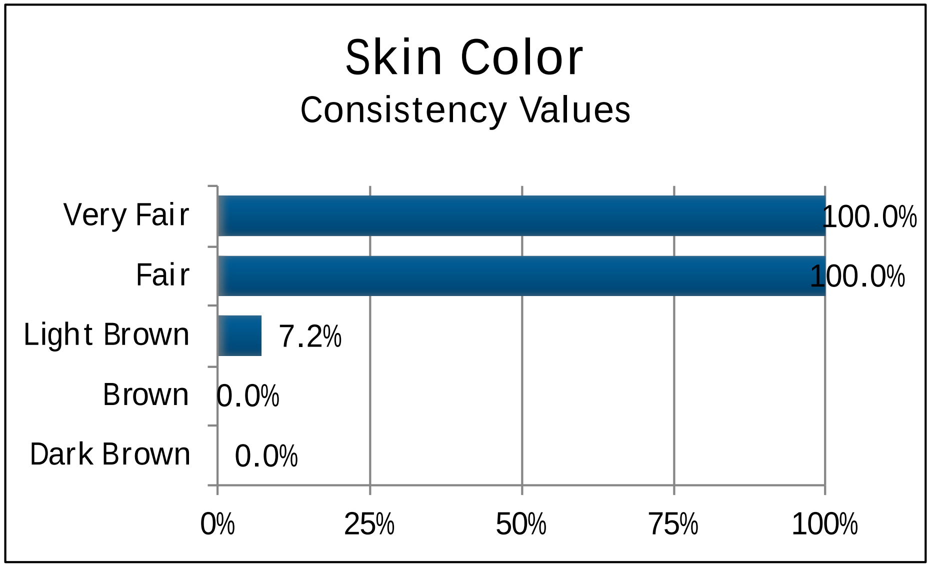


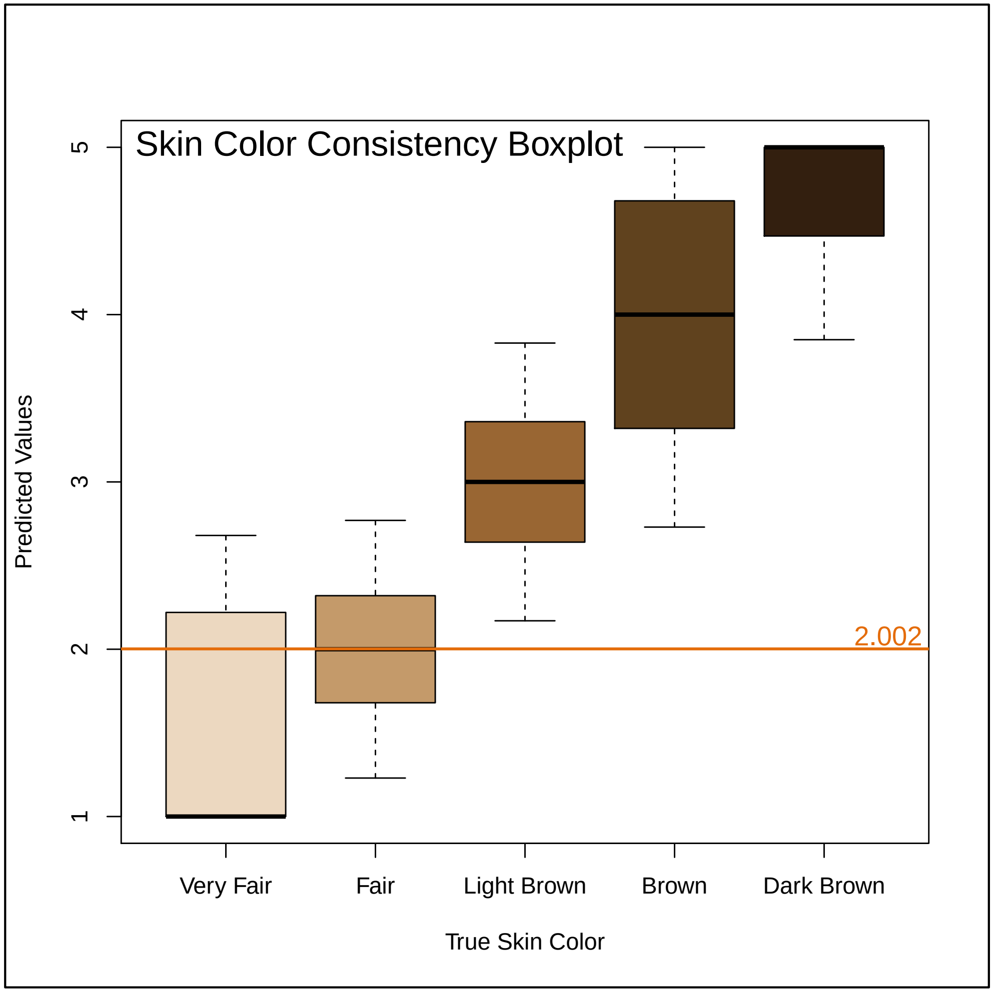


- Blue (50.4% confidence)
- Blue or Green (85.1% confidence)
- NOT Brown or Black (99.3% confidence)

Sample Results – Skin Color

- Predicted Value = 2.002

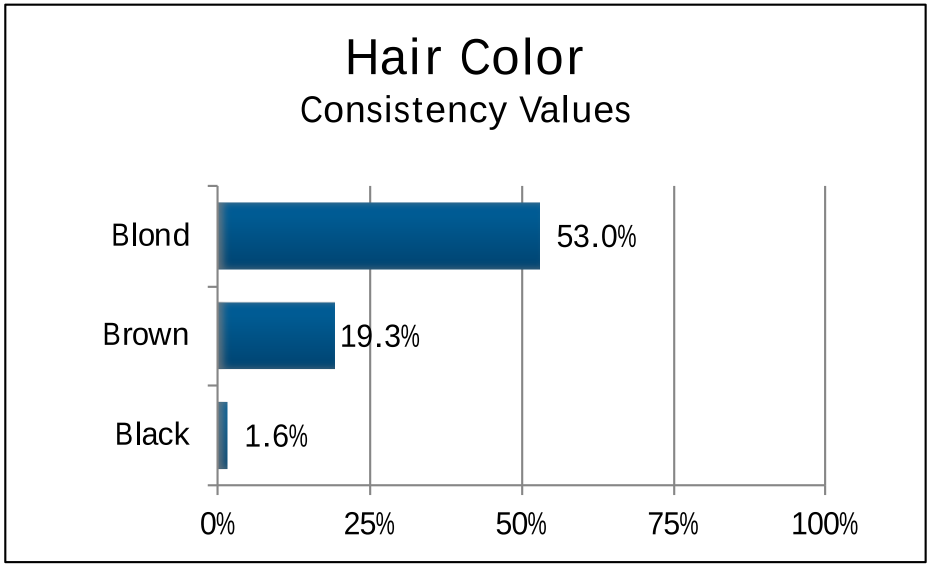


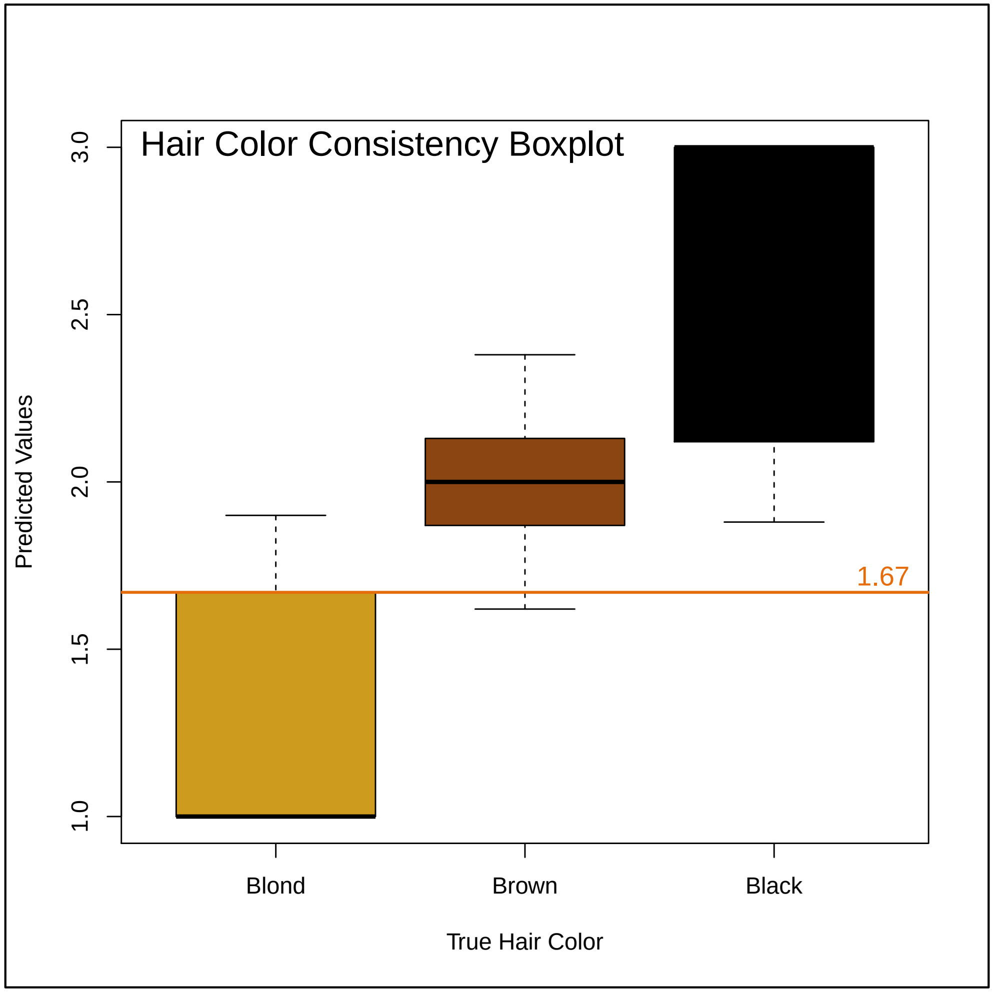


- Very Fair or Fair (92.9% confidence)
- NOT Light Brown, Brown, or Dark Brown (92.9% confidence)

Sample Results – Hair Color

- Predicted Value = 1.670

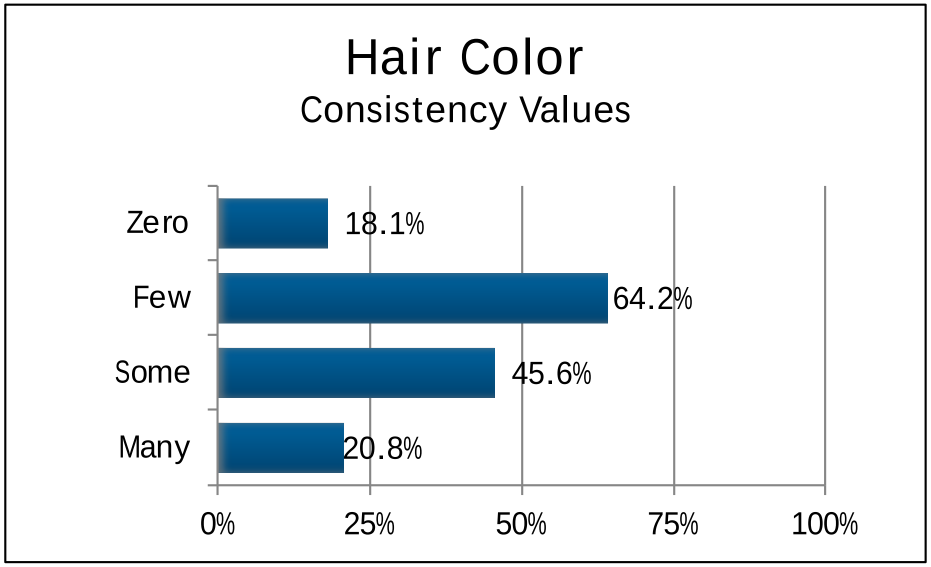


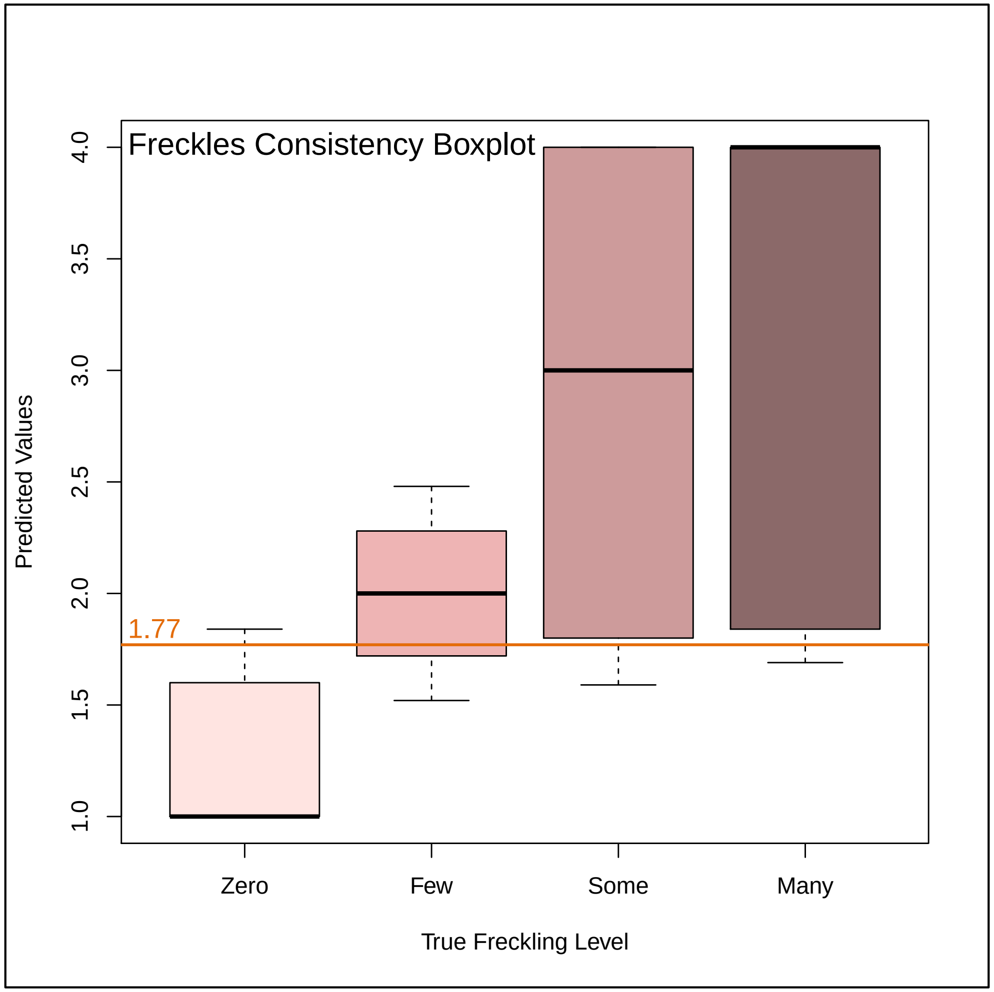


- Blond (80.7% confidence)
- Blond or Brown (98.3% confidence)
- <u>NOT</u> Black (98.4% confidence)

Sample Results – Freckling

- Predicted Value = 1.770

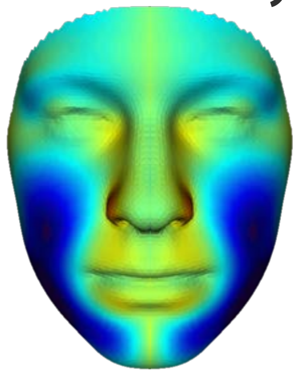




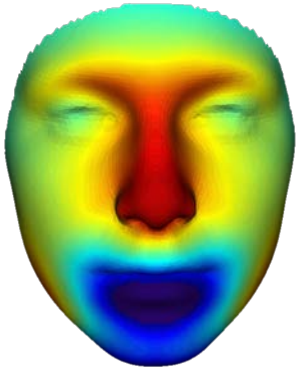
- Few (54.3% confidence)
- Few or Some (79.2% confidence)

Sample Results – Face Shape

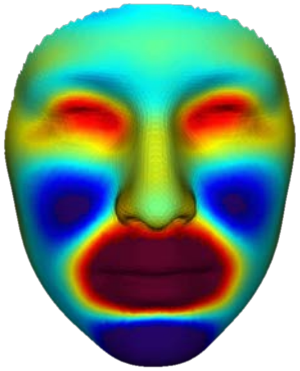
- Predict the face and compare it to a face predicted using only sex and ancestry; heat maps show the differences



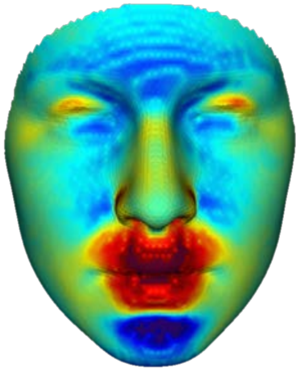
x: narrower jaw & chin; slightly wider nose & mouth



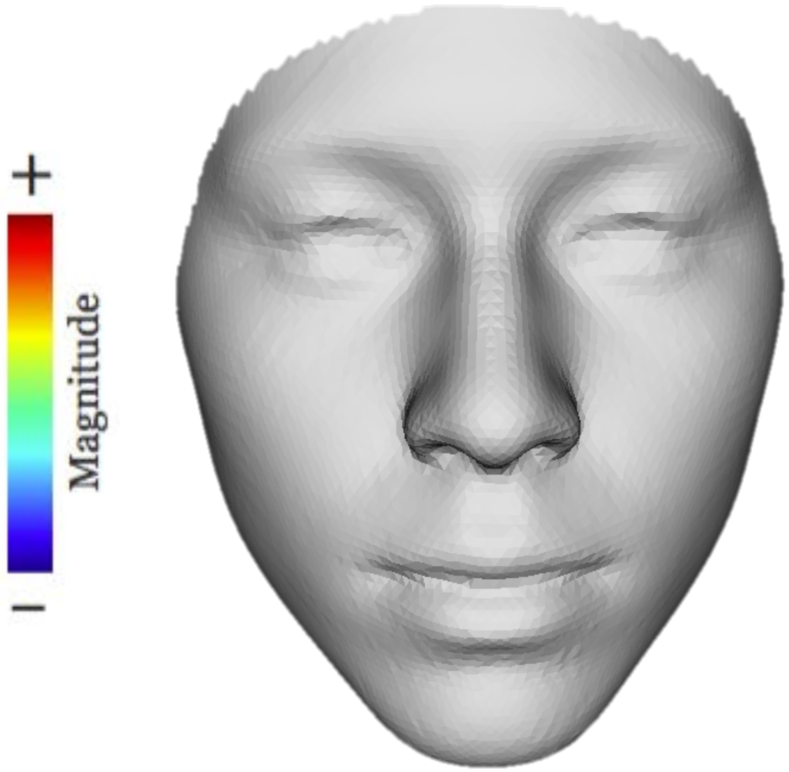
y: higher brow, nose & jaw; lower mouth & chin



z: more protruding eyes & mouth; more recessed cheeks & chin

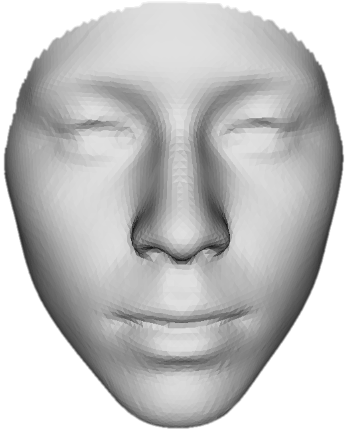


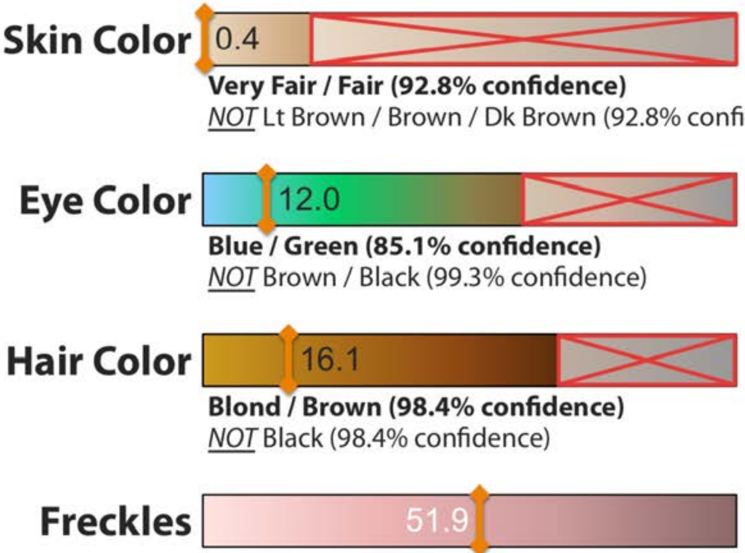
area: larger mouth; smaller brow, cheeks & chin



Sample Results – Face Shape

- Caricaturize the face according to the heat maps, fit a head to it, and apply the predicted pigmentation
  - Default age (25) and BMI (22)





Very Fair / Fair (92.8% confidence)
NOT Lt Brown / Brown / Dk Brown (92.8% confidence)

Blue / Green (85.1% confidence)
NOT Brown / Black (99.3% confidence)

Blond / Brown (98.4% confidence)
NOT Black (98.4% confidence)

Few / Some (79.2% confidence)

Sex:	Female ♀
Age:	Unknown (Composite shown at age 25)
Body Mass:	Unknown (Composite shown at BMI 22, Normal)
Ancestry:	Northern European

Law Enforcement Casework

- Predictive models are built using high-quality data, so they assume that the data they receive when making a prediction is also high-quality
- In casework, the DNA is of varying quality and quantity, and we need to take this into account when making predictions

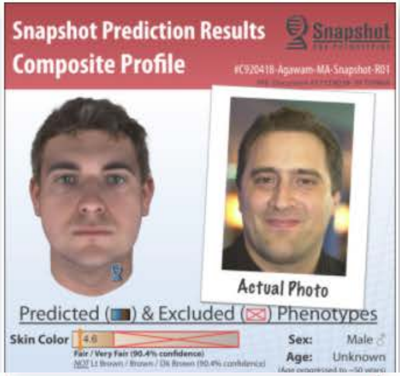
Source		Type		Quantity		Call Rate	
Semen	48.3%	**Single Source**	80.2%	**≤ 2.5 ng**	22.0%	**> 95%**	47.9%
Blood	24.0%	**Low Mixture**	15.5%	**2.5-5 ng**	12.9%	**90-95%**	11.3%
Tissue	10.3%	**High Mixture (Deconvoluted)**	4.3%	**5-10 ng**	13.3%	**80-90%**	17.9%
Saliva	7.9%			**10-20 ng**	17.8%	**70-80%**	6.2%
Bone	5.0%			**20-40 ng**	27.0%	**60-70%**	12.1%
Touch	4.5%			**40-80 ng**	3.3%	**<60%**	4.7%
				>80 ng	3.7%		

Law Enforcement Casework

- Genotyping call rates are never 100% with casework samples because of the small DNA quantities
  - We use a machine learning algorithm that allows for missing data (many do not)
  - For each sample, we recalculate the cross-validation results and confidence intervals for that set of SNPs
- DNA degradation results in short fragment lengths
  - Optimized genotyping protocol for forensic samples
- Mixtures confuse both genotype calling and phenotype prediction
  - We developed new computational methods to deconvolute mixtures

Law Enforcement Casework


Snapshot Prediction Results
Composite Profile
#OPD-FL-2001-380051-Snapshot
Actual Photo
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 98.0
Dark / Dk Olive (99.99% confidence)
Sex: Male ♂
Age: Unknown

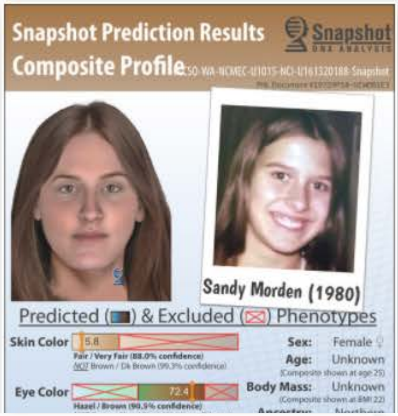

Snapshot Prediction Results
Composite Profile
#C920418-Agawam-MA-Snapshot-R01
Actual Photo
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 4.6
Fair / Very Fair (90.4% confidence)
Sex: Male ♂
Age: Unknown

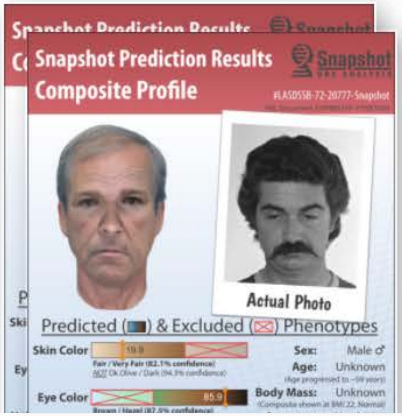

Snapshot Prediction Results
Composite Profile
Snapshot #RM13053293
Actual Photo
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 7.3
Fair / Very Fair (84.4% confidence)
Sex: Male ♂
Age: Unknown


Snapshot Prediction Results
Composite Profile
Arrest Photo
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 2.7
Very Fair / Fair (98.2% confidence)
Sex: Male ♂
Age: Unknown

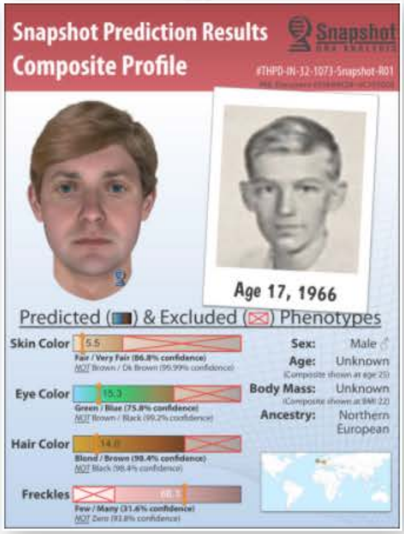

Snapshot Prediction Results
Composite Profile
#DKSO-IL-2016-02305-Snapshot
Actual Photo
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 5.1
Eye Color 10.2
Sex: Male ♂
Age: Unknown
Body Mass: Unknown

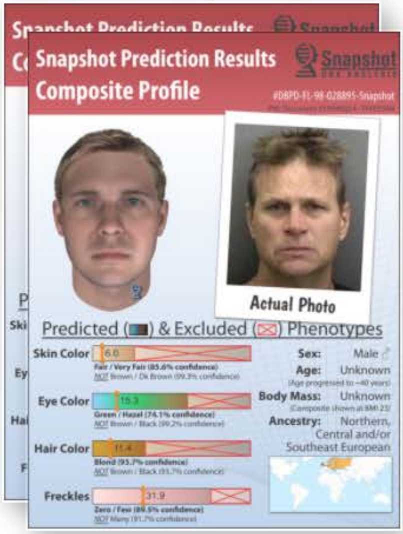

Snapshot Prediction Results
Composite Profile
#LPD-IL-76-00714-Snapshot
Actual Photo
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 18.0
Lt Brown / Fair (77.4% confidence)
Eye Color 22.3
Sex: Male ♂
Age: Unknown
Body Mass: Unknown

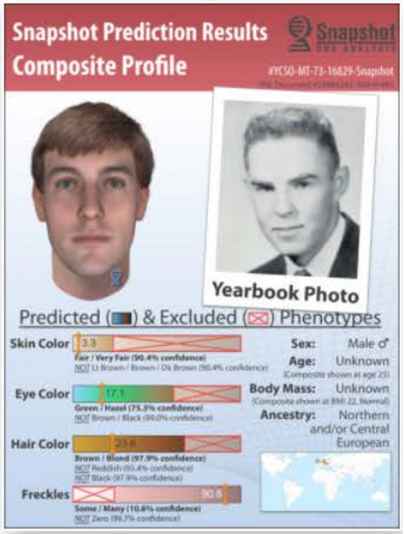

Snapshot Prediction Results
Composite Profile
Sandy Morden (1980)
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 5.8
Fair / Very Fair (88.0% confidence)
Eye Color 72.4
Hazel / Brown (90.5% confidence)
Sex: Female ♀
Age: Unknown
Body Mass: Unknown


Snapshot Prediction Results
Composite Profile
#LASDSSB-72-20777-Snapshot
Actual Photo
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 19.9
Fair / Very Fair (82.1% confidence)
Eye Color 85.9
Sex: Male ♂
Age: Unknown
Body Mass: Unknown

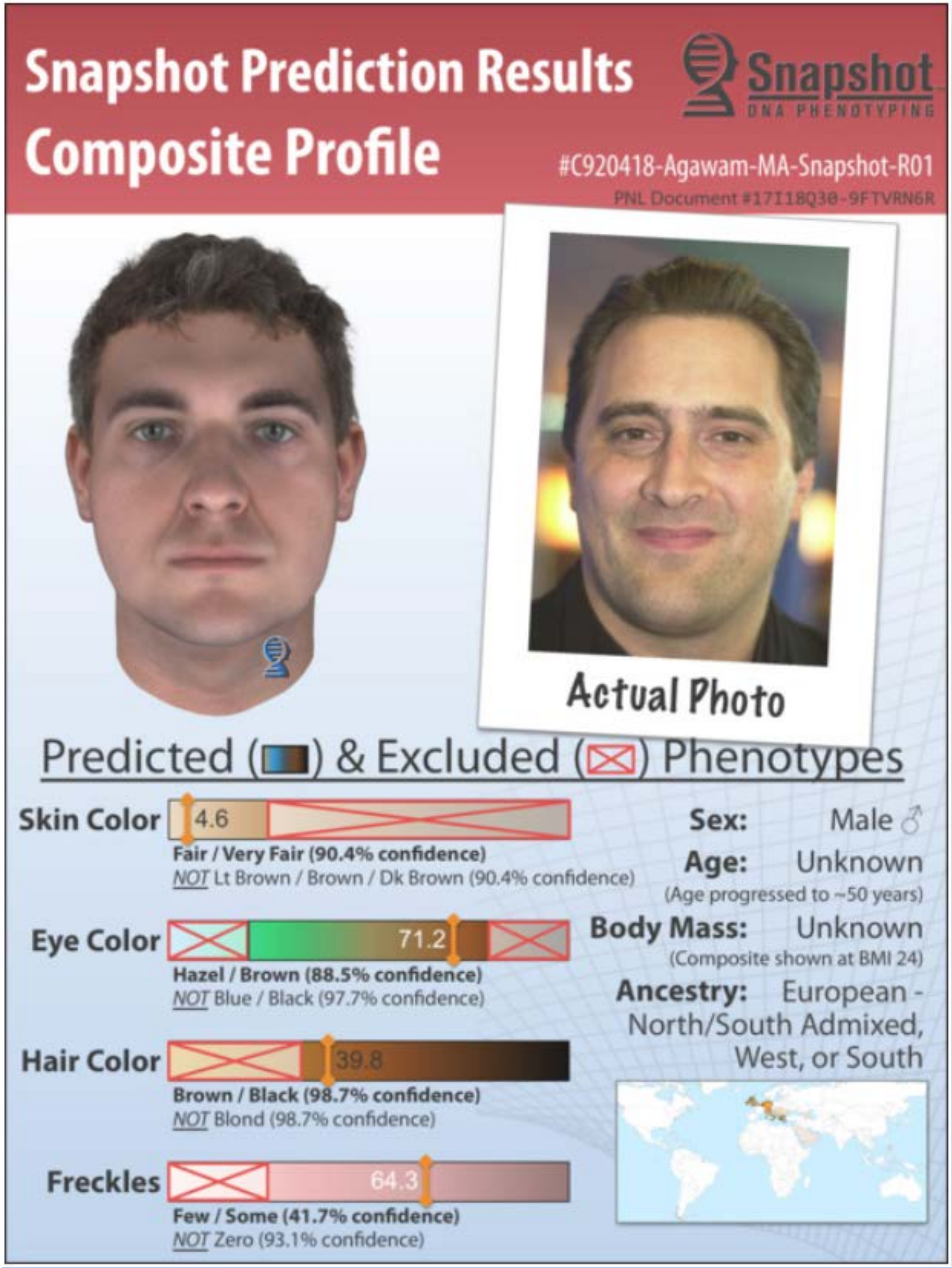

Snapshot Prediction Results
Composite Profile
#SCSO-FL-99-29040-Snapshot
Photo at Age 26
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 4.7
Fair / Very Fair (91.0% confidence)
Eye Color 71.3
Hair Color
Freckles
Sex: Male ♂
Age: Unknown
Body Mass: Unknown
Ancestry: Admixed European (West/South) & North Middle Eastern


Snapshot Prediction Results
Composite Profile
#THPD-IN-32-1073-Snapshot-R01
Age 17, 1966
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 5.5
Fair / Very Fair (86.8% confidence)
Eye Color 15.3
Green / Blue (75.8% confidence)
Hair Color 14.0
Blond / Brown (98.4% confidence)
Freckles 68.3
Few / Many (31.6% confidence)
Sex: Male
Age: Unknown
Body Mass: Unknown
Ancestry: Northern European


Snapshot Prediction Results
Composite Profile
#DBPD-FL-98-028891-Snapshot
Actual Photo
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 6.0
Fair / Very Fair (85.6% confidence)
Eye Color 15.3
Green / Hazel (74.1% confidence)
Hair Color 11.4
Blond (95.7% confidence)
Freckles 31.9
Zero / Few (88.5% confidence)
Sex: Male ♂
Age: Unknown
Body Mass: Unknown
Ancestry: Northern, Central and/or Southeast European


Snapshot Prediction Results
Composite Profile
#YCSO-MT-73-16829-Snapshot
Yearbook Photo
Predicted (■) & Excluded (☒) Phenotypes
Skin Color 3.3
Fair / Very Fair (90.4% confidence)
Eye Color 17.1
Green / Hazel (75.3% confidence)
Hair Color 23.6
Brown / Blond (97.9% confidence)
Freckles 90.8
Some / Many (10.8% confidence)
Sex: Male ♂
Age: Unknown
Body Mass: Unknown
Ancestry: Northern and/or Central European

Law Enforcement Casework



Snapshot Prediction Results
Composite Profile

Snapshot DNA PHENOTYPING

#C920418-Agawam-MA-Snapshot-R01

PNL Document #17I18Q30-9FTVRN6R

Actual Photo

Predicted (▇) & Excluded (☒) Phenotypes

Skin Color 4.6

Fair / Very Fair (90.4% confidence)

NOT Lt Brown / Brown / Dk Brown (90.4% confidence)

Eye Color 71.2

Hazel / Brown (88.5% confidence)

NOT Blue / Black (97.7% confidence)

Hair Color 39.8

Brown / Black (98.7% confidence)

NOT Blond (98.7% confidence)

Freckles 64.3

Few / Some (41.7% confidence)

NOT Zero (93.1% confidence)

Sex: Male ♂

Age: Unknown

(Age progressed to ~50 years)

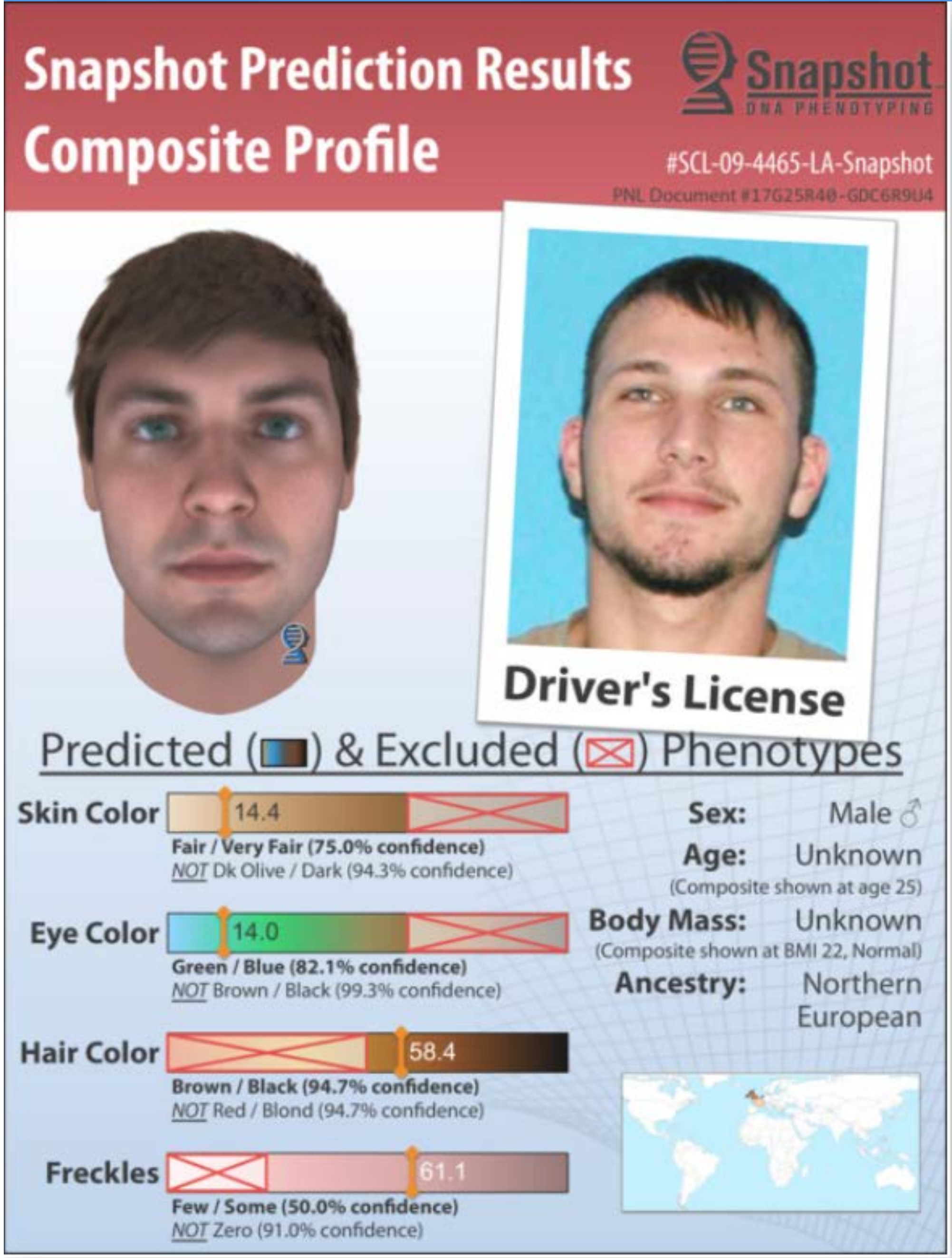
Body Mass: Unknown

(Composite shown at BMI 24)

Ancestry: European - North/South Admixed, West, or South

- 1992 homicide of Lisa Ziegert in Agawam, MA
- Thousands of men listed in the case file
- Detectives used phenotyping results to prioritize who to talk to, starting with the few who most closely matched the predictions
- One of those men fled, eventually confessing to the crime
- His DNA matched the crime scene sample

Law Enforcement Casework



Snapshot Prediction Results
Composite Profile

Snapshot DNA PHENOTYPING

#SCL-09-4465-LA-Snapshot
PNL Document #17G25R40-GDC6R9U4

Driver's License

Predicted (▨) & Excluded (☒) Phenotypes

Skin Color 14.4
Fair / Very Fair (75.0% confidence)
NOT Dk Olive / Dark (94.3% confidence)

Eye Color 14.0
Green / Blue (82.1% confidence)
NOT Brown / Black (99.3% confidence)

Hair Color 58.4
Brown / Black (94.7% confidence)
NOT Red / Blond (94.7% confidence)

Freckles 61.1
Few / Some (50.0% confidence)
NOT Zero (91.0% confidence)

Sex: Male ♂
Age: Unknown
(Composite shown at age 25)
Body Mass: Unknown
(Composite shown at BMI 22, Normal)
Ancestry: Northern European

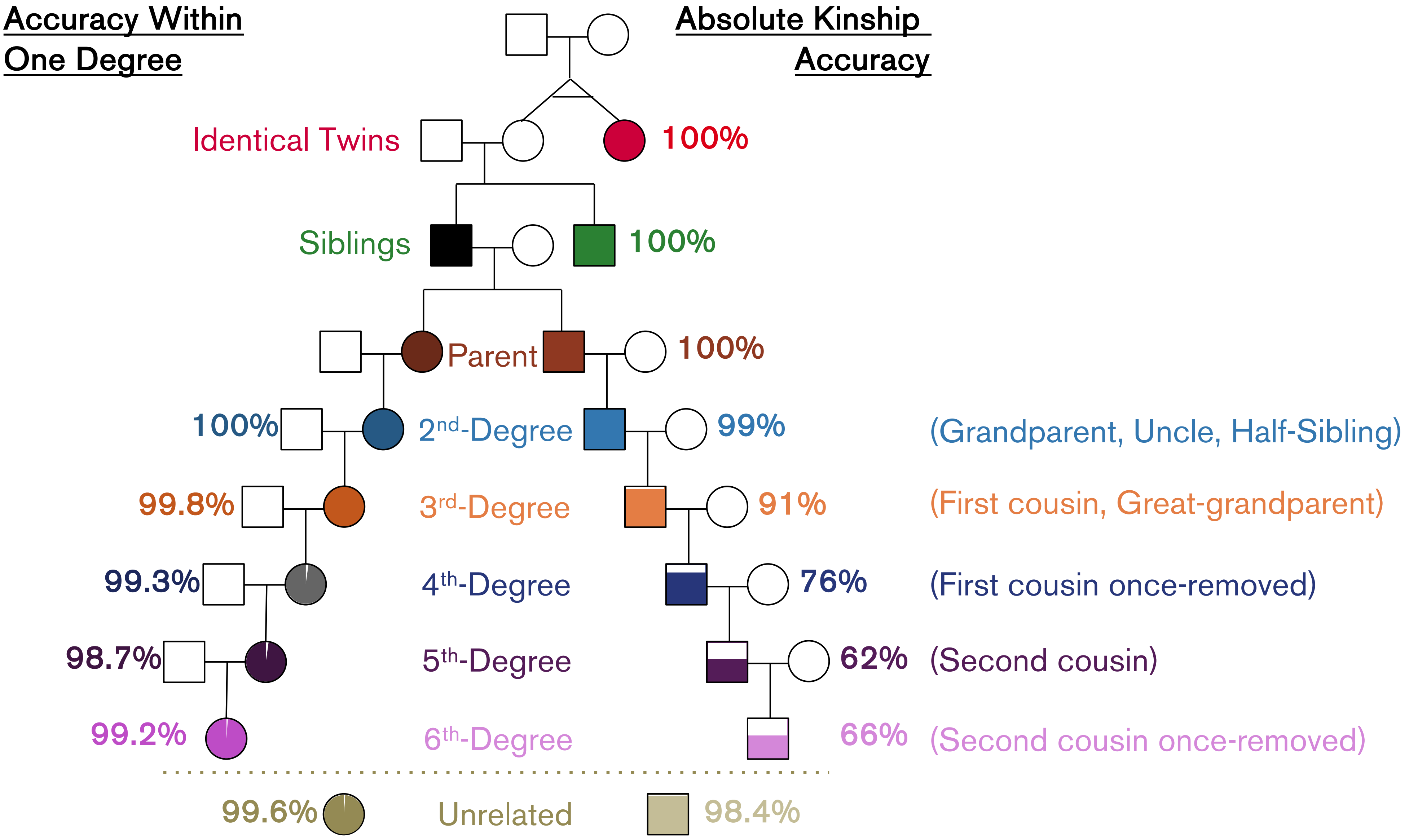
- 2009 homicide of Sierra Bouzigard in Lake Charles, LA
- For 6 years, detectives had been looking for a Hispanic male based on Sierra's cell phone history
- DNA phenotyping results were completely different
- Detectives released the composite to the public
- A tip led to a suspect whose DNA matched the crime scene sample

Distant Kinship Inference

Kinship Inference

- **DoD problem**: phenotype information isn’t always enough
    - Enemy groups sometimes have a familial connection
    - If a DNA sample doesn’t directly match known individuals in a database, might it instead be related to someone in that database?
- **SBIR Modification**: Given DNA from two subjects, determine their degree of relatedness
- Snapshot predictive models
    - Outcome is relationship between a <u>pair</u> of genomes
    - Supervised machine learning using thousands of related and unrelated pairs, including families with members of different ethnic backgrounds

Kinship Inference Accuracy



	Accuracy Within One Degree	Absolute Kinship Accuracy	
Identical Twins		100%	
Siblings		100%	
Parent		100%	
2nd-Degree	100%	99%	(Grandparent, Uncle, Half-Sibling)
3rd-Degree	99.8%	91%	(First cousin, Great-grandparent)
4th-Degree	99.3%	76%	(First cousin once-removed)
5th-Degree	98.7%	62%	(Second cousin)
6th-Degree	99.2%	66%	(Second cousin once-removed)
Unrelated	99.6%	98.4%	

Kinship from Highly Degraded DNA

- Snapshot Kinship requires high-quality data for tens of thousands of SNPs
- Past accounting mission: identify missing soldiers from past conflicts by matching their DNA to family references
    - Autosomal and Y STRs: limited success on highly degraded samples, require family references with specific relationships
    - Mitochondrial DNA (mtDNA): requires family references with specific relationships (direct maternal relatives)
- **DoD Problem**: Not all missing soldiers can be matched to their available family references using existing techniques
- **Collaboration with AFMES-AFDIL**: autosomal kinship from highly degraded DNA
    - DoD Office of the Deputy Assistant Secretary of Defense for Emerging Capabilities and Prototyping

Kinship from Highly Degraded DNA

- Innovation:
  - Laboratory: new methods to obtain thousands of autosomal SNPs from highly degraded bone samples
  - Bioinformatics: determine kinship from sparse, low-quality data

Family Reference	A	A	T	G	C	A	T	G	G	T	A	C
	A	G	T	A	C	A	C	G	A	T	G	C
Unidentified	A	?	?	G	?	?	?	G	?	T	?	?
	A	?	C	?	?	?	?	?	A	C	?	?

- Blind test: successfully matched samples from as far back as World War II to family references
  - Transitioning to AFDIL

Genetic Genealogy

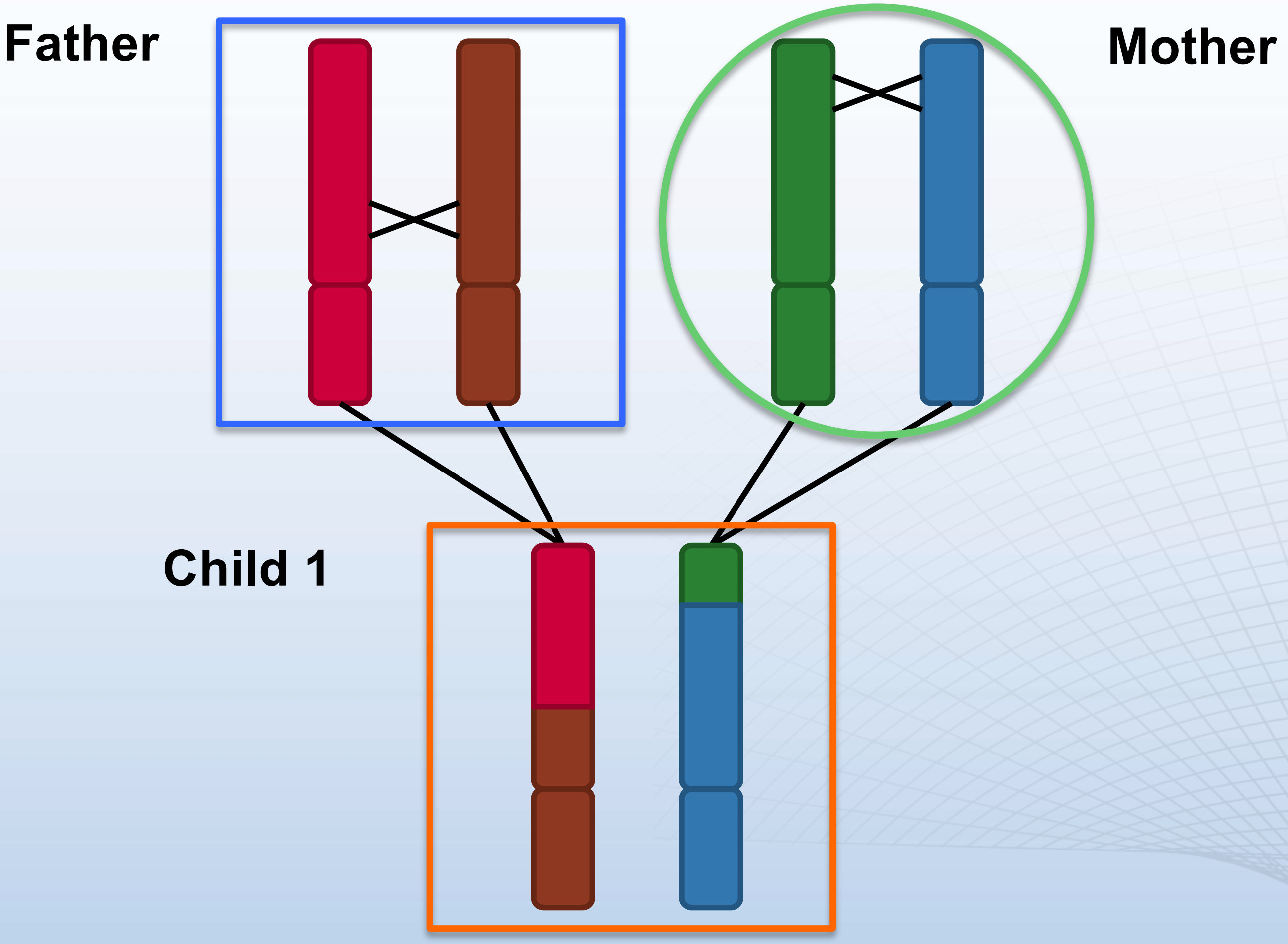
PARABON®
NANOLABS

Procedure

1. Generate DNA data from a forensic sample from an unknown subject
2. Upload DNA data to a database to find other individuals that share DNA with the subject

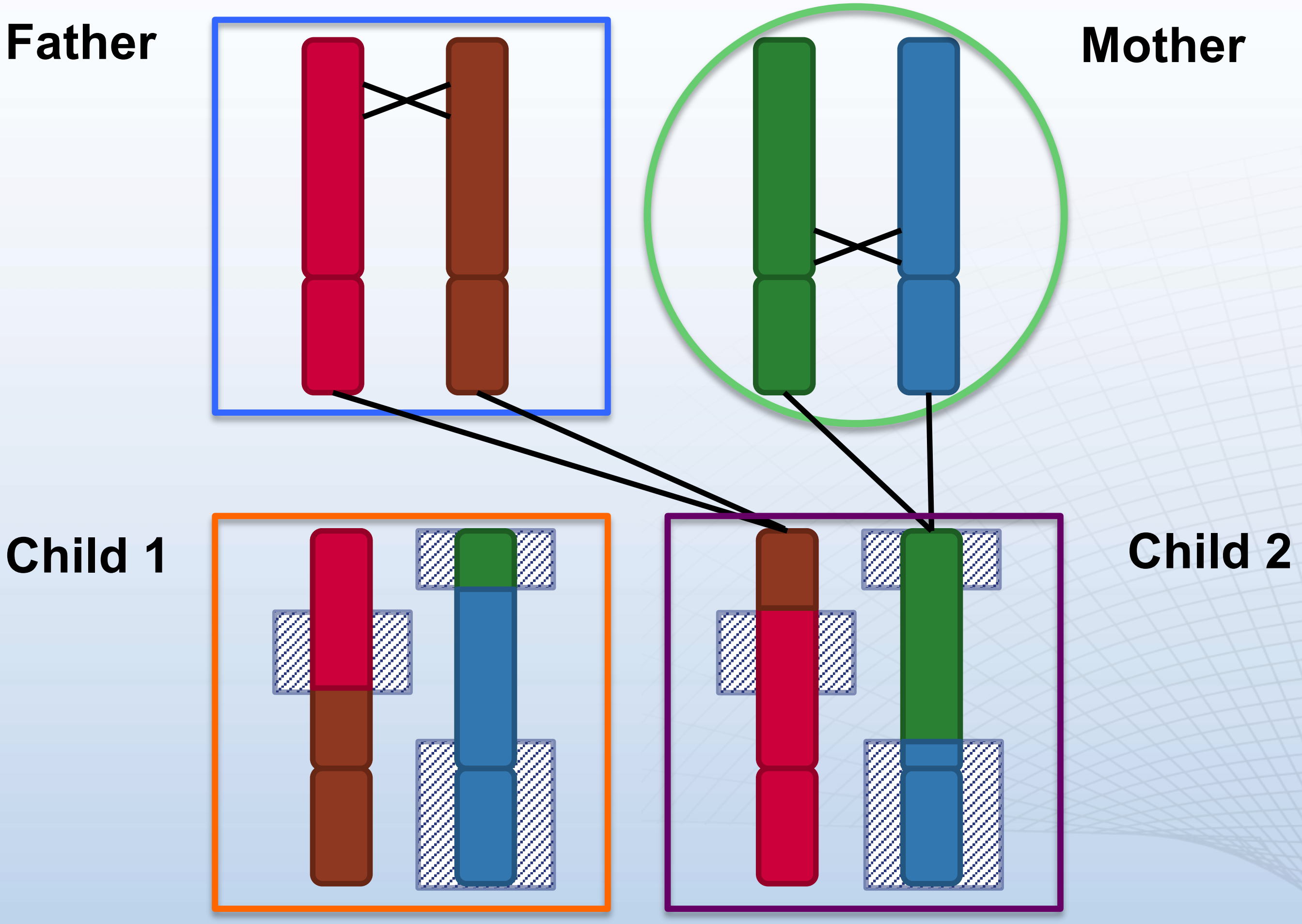
How does it work?

- A child inherits one copy of each chromosome from their father and one from their mother
- The chromosomes recombine, so each parent passes on a version that is a *combination* of their own two copies



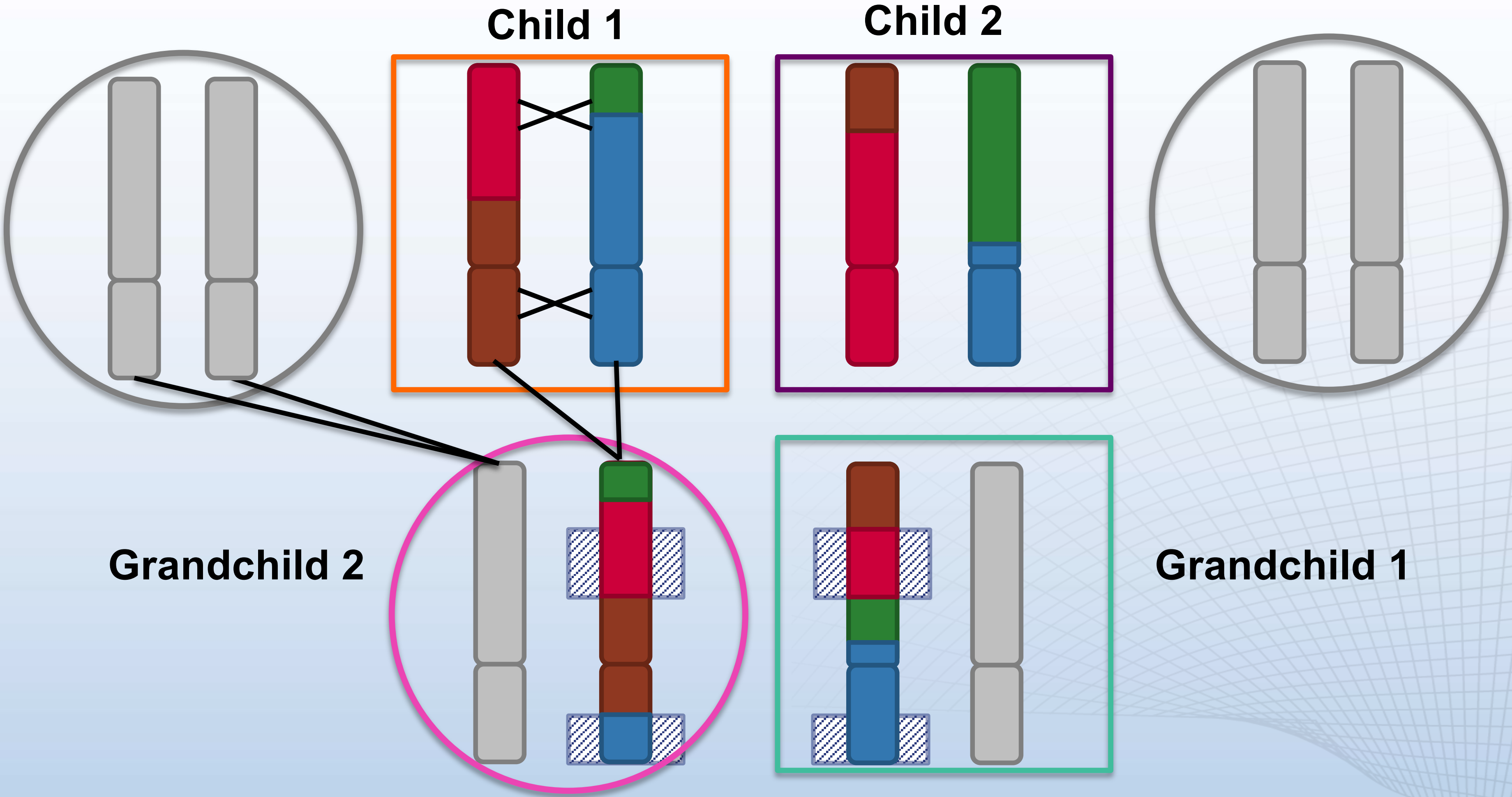
How does it work?

- Recombination happens randomly for each child, so full siblings do <u>not</u> inherit the same copies



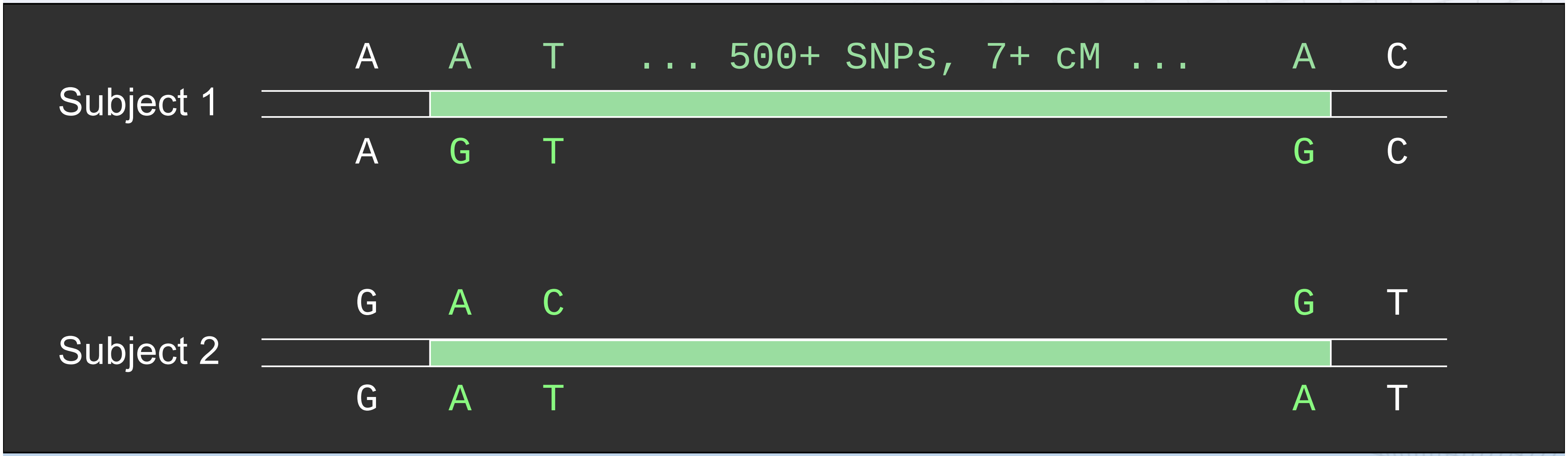
How does it work?

- The more distantly related two subjects are, the smaller the shared segments become as they are broken up by recombination



Detecting Shared Segments

- Common to share at least one allele at a SNP, but not common to have long uninterrupted stretches of shared alleles unless the two individuals inherited that DNA from a common ancestor
- Look for shared segments that are long enough to be highly unlikely to be shared by chance



A A T ... 500+ SNPs, 7+ cM ... A C

Subject 1

A G T G C

G A C G T

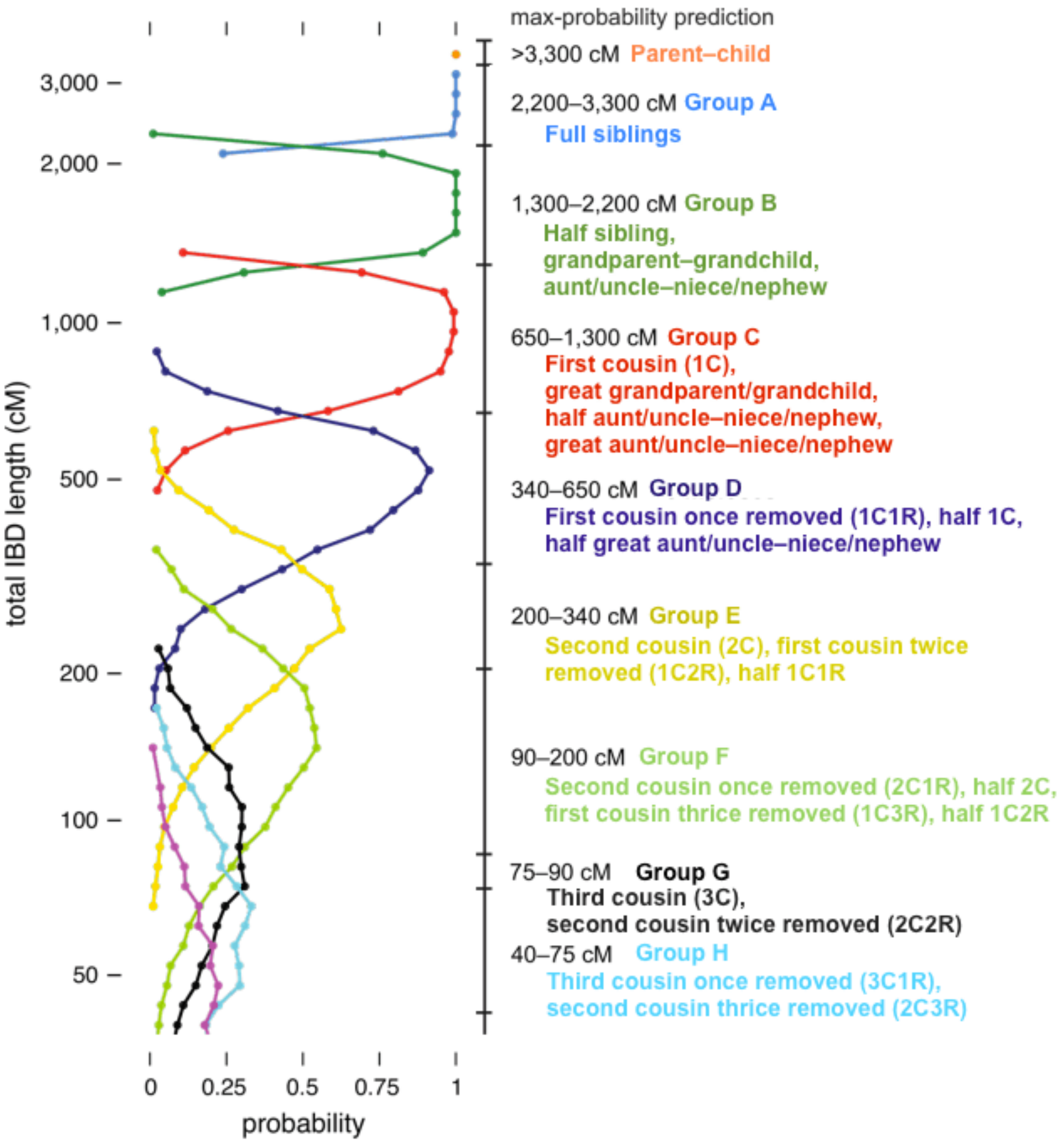
Subject 2

G A T A T

Relatedness Using SNPs

Amount of shared DNA (in cM) across the genome correlates with relatedness

Closer relatives share more DNA



Ball, et al. *Ancestry DNA Matching* (2016)

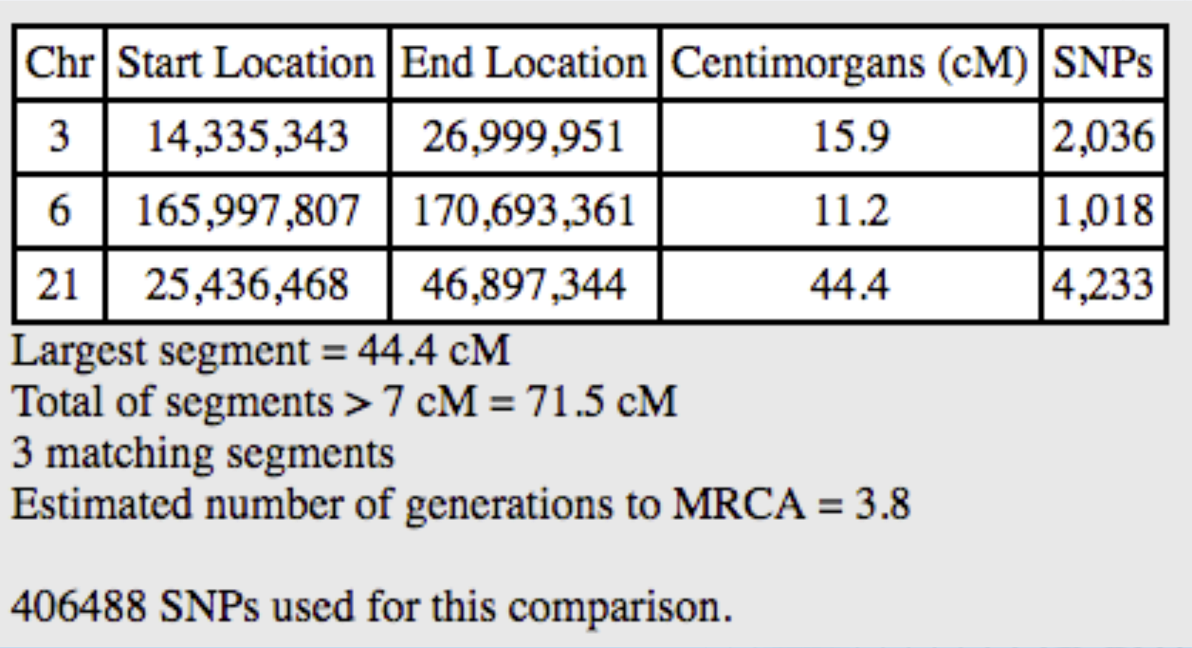
Edited by Leah Larkin, Ph.D.

Genetic Genealogy Databases

- Law enforcement usage allowed for unidentified remains and violent crimes
  - GEDmatch
  - FamilyTreeDNA (FTDNA)
- No law enforcement usage
  - 23andMe
  - AncestryDNA
  - MyHeritage

Database Searching

- Compare queried file to all opted-in users' files
- Only the amount and location of shared DNA is shown, **not the DNA itself**



Chr	Start Location	End Location	Centimorgans (cM)	SNPs
3	14,335,343	26,999,951	15.9	2,036
6	165,997,807	170,693,361	11.2	1,018
21	25,436,468	46,897,344	44.4	4,233

Largest segment = 44.4 cM
Total of segments > 7 cM = 71.5 cM
3 matching segments
Estimated number of generations to MRCA = 3.8

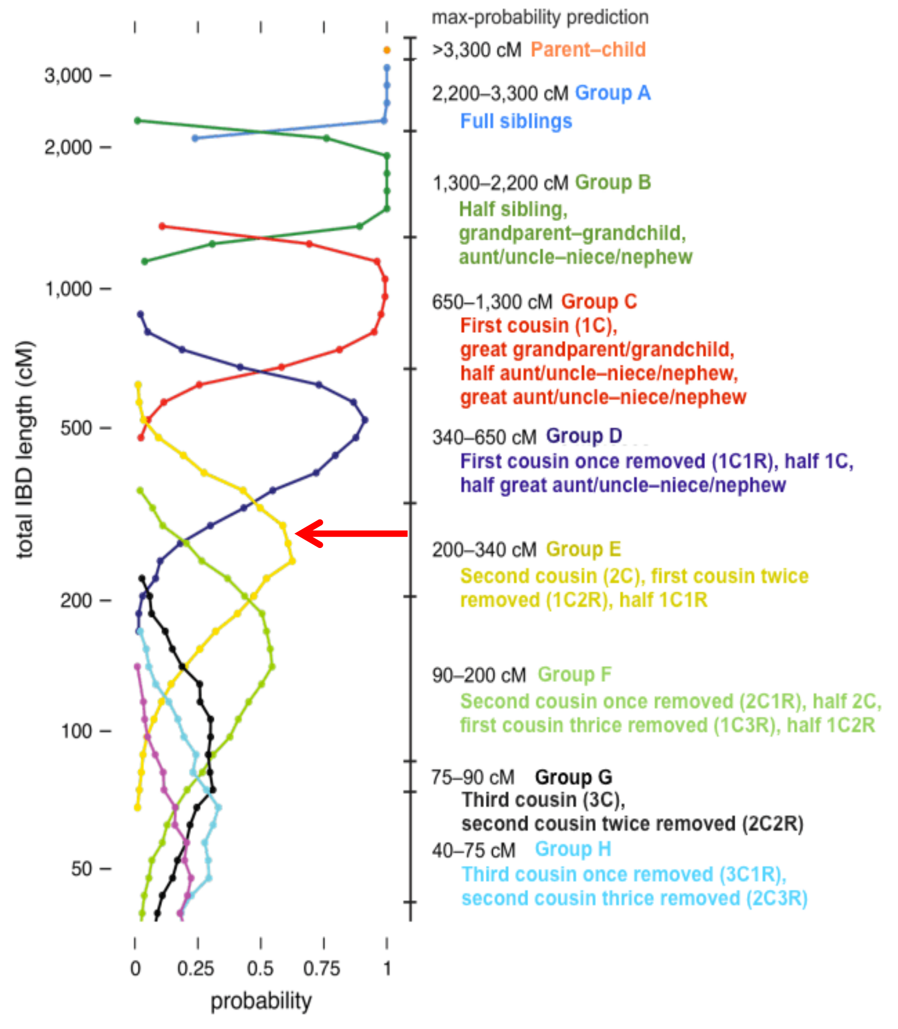
406488 SNPs used for this comparison.

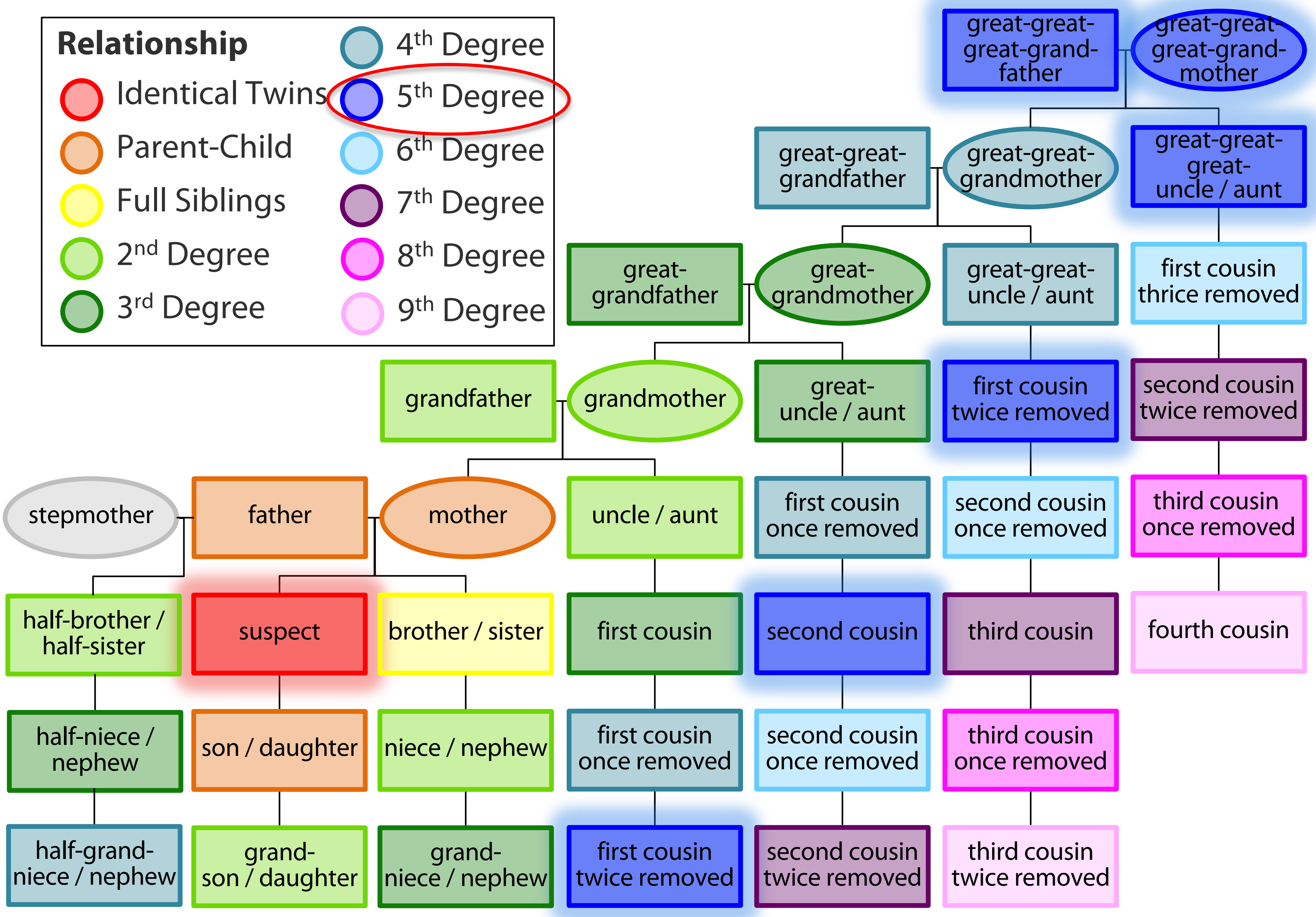
Procedure

1. Generate DNA data from a forensic sample from an unknown subject
2. Upload DNA data to a database to find other individuals that share DNA with the subject
3. Build family trees back in time to possible common ancestors using public records

Genetic Genealogy

- Total shared DNA: 300 cM
- The amount of sharing can be consistent with multiple possible degrees of relatedness
- Each degree of relatedness has multiple possible relationship types





Relationship

Identical Twins

Parent-Child

Full Siblings

2nd Degree

3rd Degree

4th Degree

5th Degree

6th Degree

7th Degree

8th Degree

9th Degree

great-great-great-grand-father

great-great-great-grand-mother

great-great-grandfather

great-great-grandmother

great-great-great-uncle / aunt

great-grandfather

great-grandmother

great-great-uncle / aunt

first cousin thrice removed

grandfather

grandmother

great-uncle / aunt

first cousin twice removed

second cousin twice removed

stepmother

father

mother

uncle / aunt

first cousin once removed

second cousin once removed

third cousin once removed

half-brother / half-sister

suspect

brother / sister

first cousin

second cousin

third cousin

fourth cousin

half-niece / nephew

son / daughter

niece / nephew

first cousin once removed

second cousin once removed

third cousin once removed

half-grand-niece / nephew

grand-son / daughter

grand-niece / nephew

first cousin twice removed

second cousin twice removed

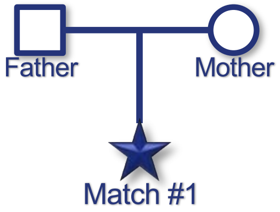
third cousin twice removed

Family Tree Building

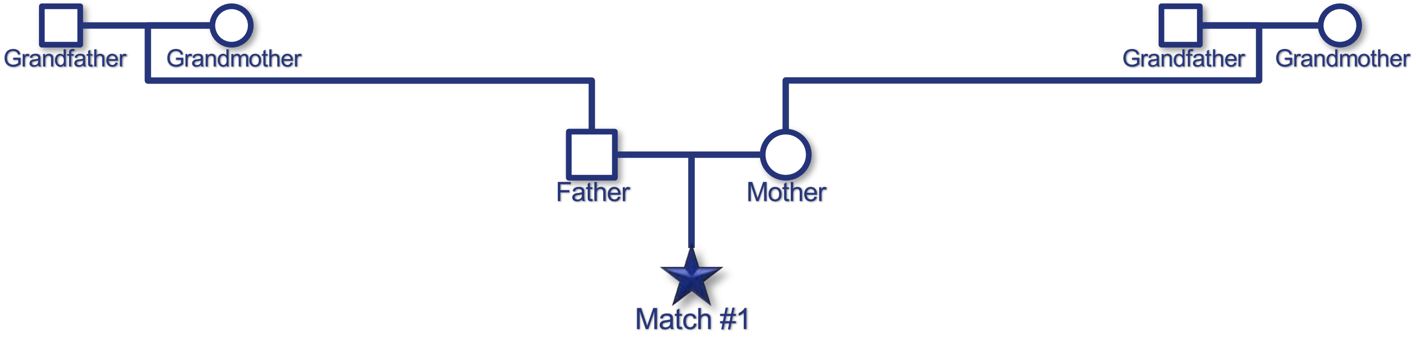
★

Match #1

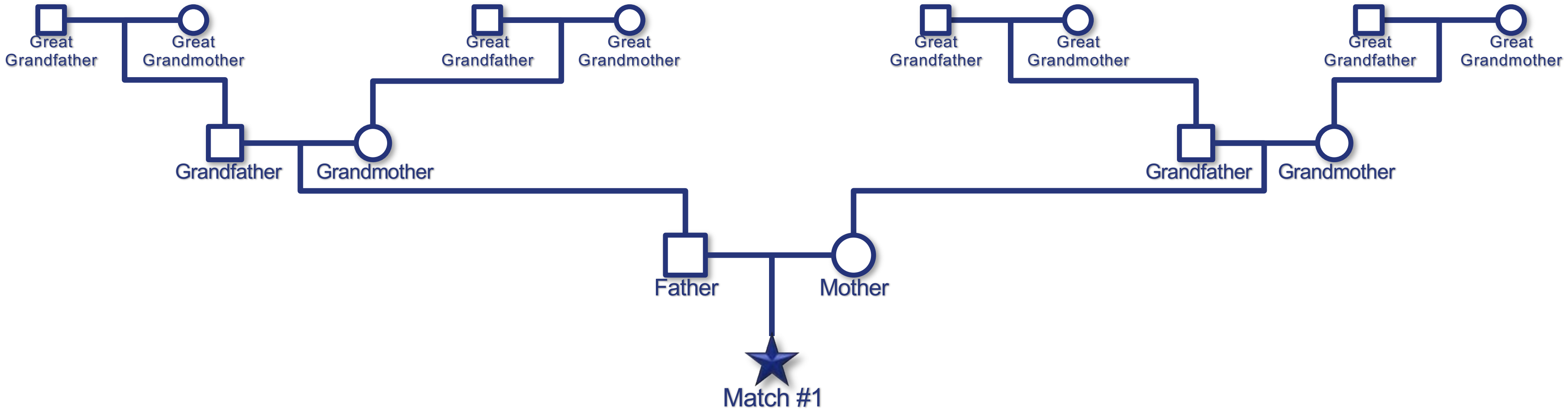
Family Tree Building



Family Tree Building



Family Tree Building



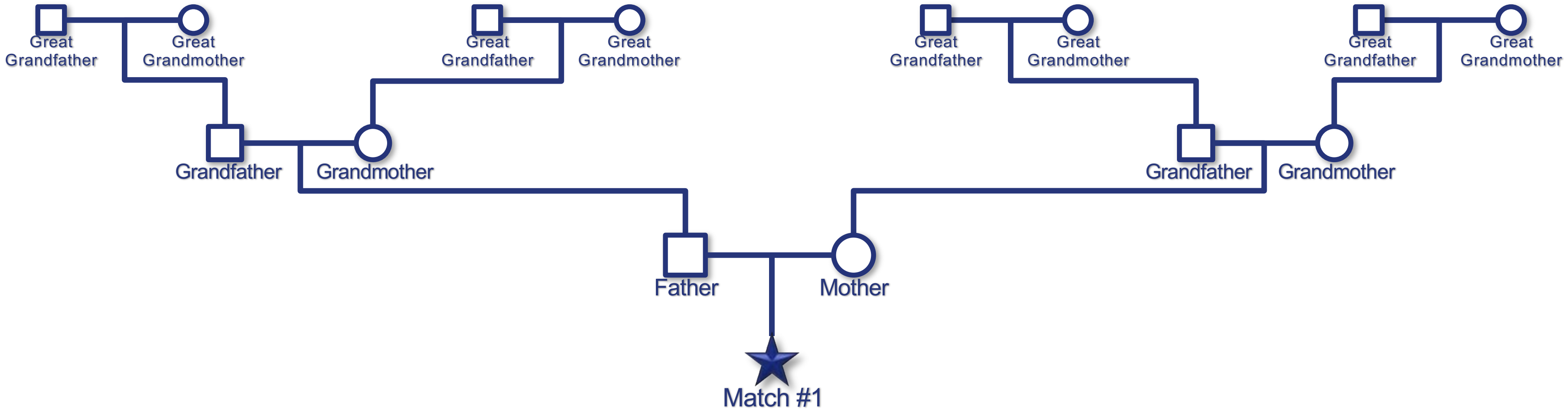
Family Tree Building

- Records are not always available
- Biological family trees don’t always match family trees on paper
  - Misattributed paternity
  - Unknown parentage
  - Adoption

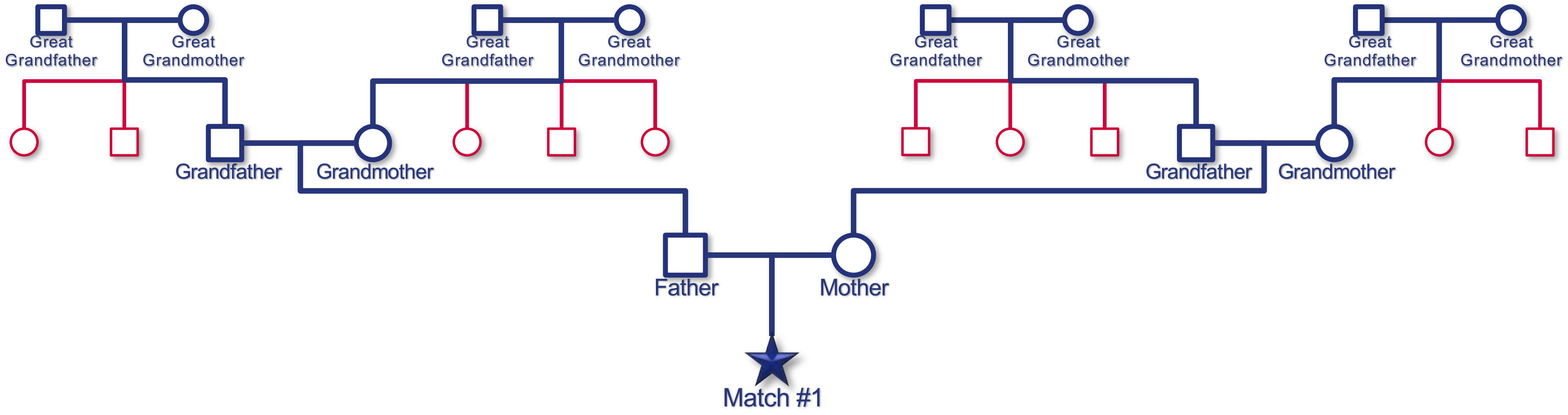
Procedure

1. Generate DNA data from a forensic sample from an unknown subject
2. Upload DNA data to a database to find other individuals that share DNA with the subject
3. Build family trees back in time
4. Identify all of the other descendants of those common ancestors at the right genetic distance to be the Subject

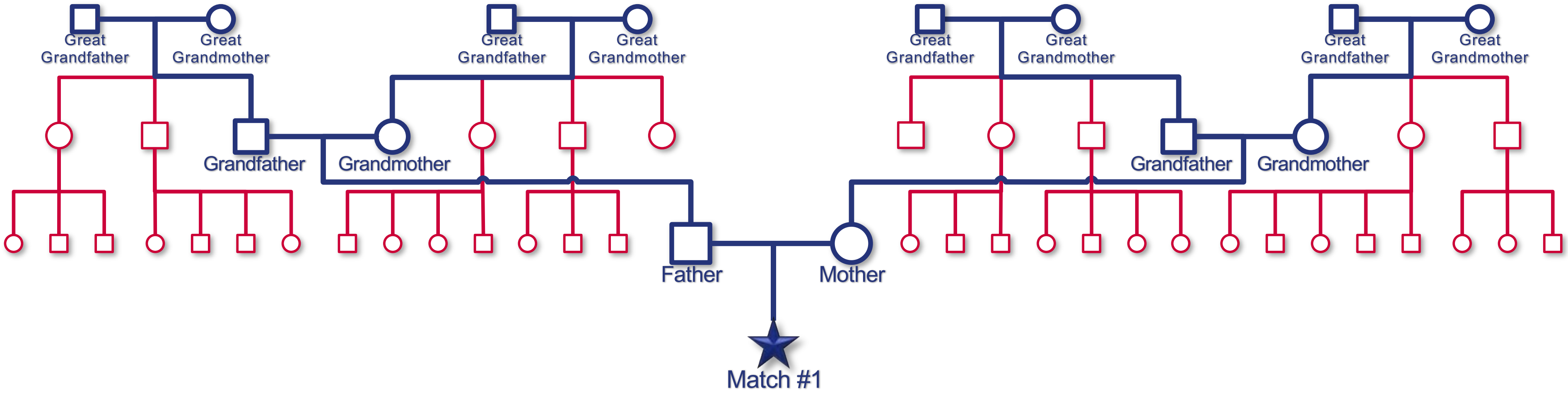
Descendancy Research



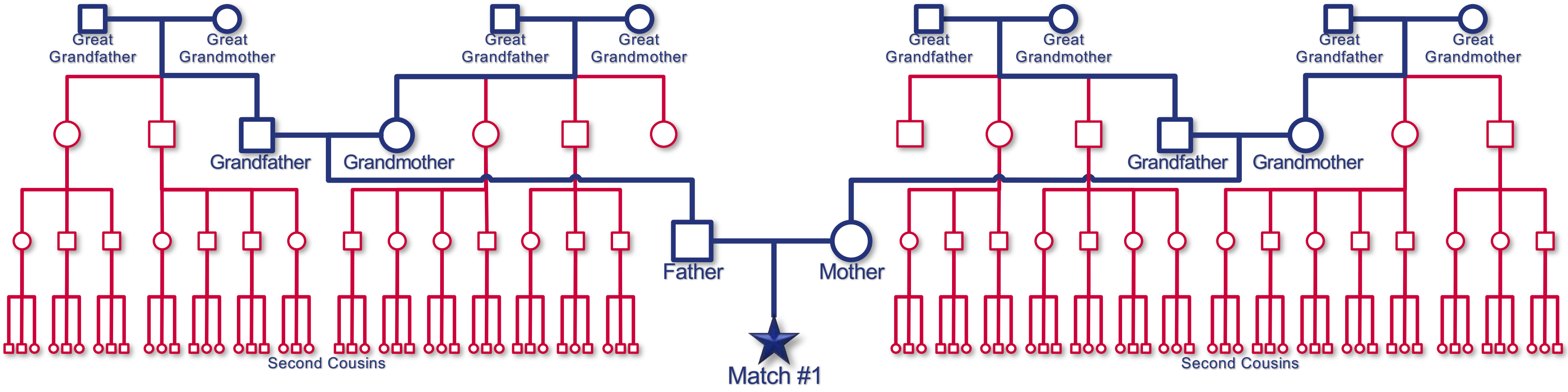
Descendancy Research



Descendancy Research



Descendancy Research



Descendancy Research

- Not all descendants can be identified
  - Out of wedlock births
  - Misattributed parentage
  - Families can be very large

Procedure

1. Generate DNA data from a forensic sample from an unknown subject
2. Upload DNA data to a database to find other individuals that share DNA with the subject
3. Build family trees back in time
4. Identify all of the other descendants of those common ancestors at the right genetic distance
5. Narrow down the possible identities

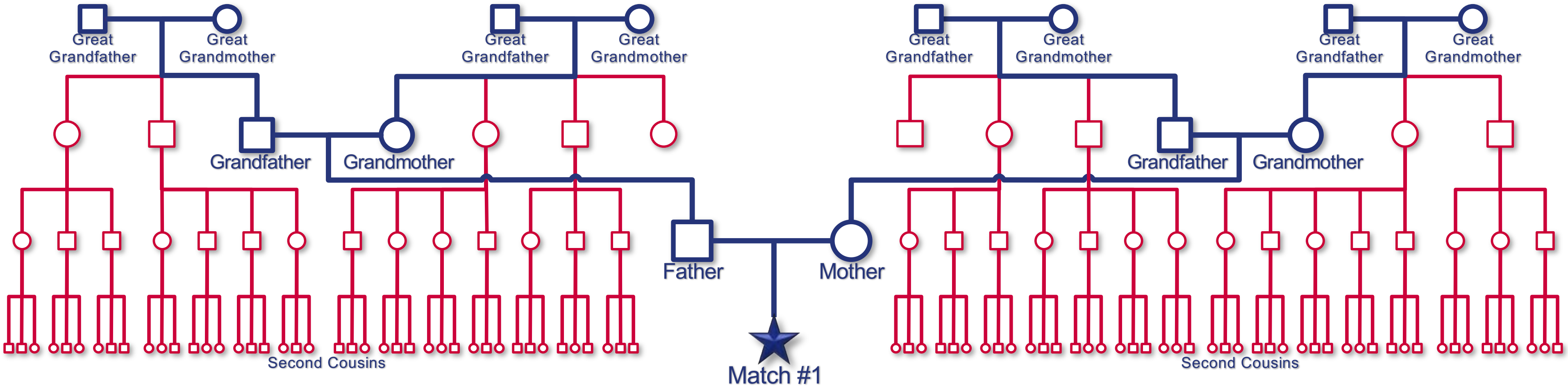
Narrowing the search

- Demographic criteria
  - Sex of the subject
  - Date of the crime
  - Location of the crime
- Genetic sharing between matches (triangulation)
- Ancestry and Phenotype predictions
- Targeted Kinship testing

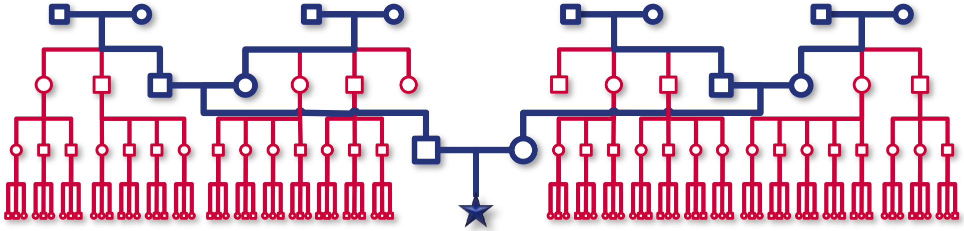
Triangulation

- If there are multiple matches to the subject that don’t share DNA with each other, they represent different branches of the subject’s family tree
- The goal is to find an intersection between their family trees (a marriage or out-of-wedlock birth)

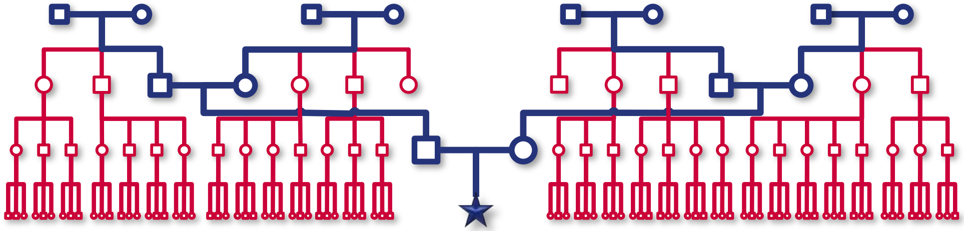
Triangulation

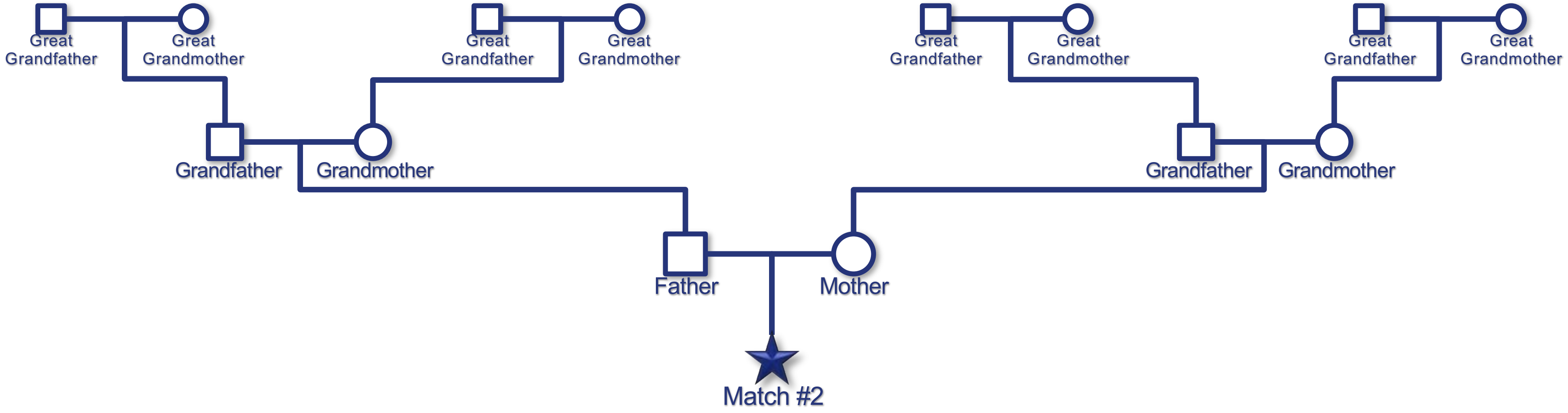


Triangulation



Triangulation





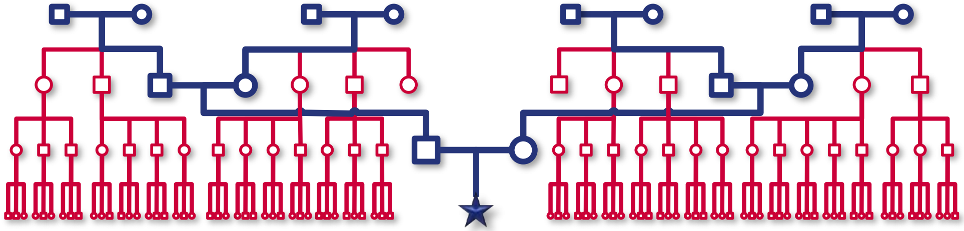
Great Grandfather
Great Grandmother
Great Grandfather
Great Grandmother
Great Grandfather
Great Grandmother
Great Grandfather
Great Grandmother

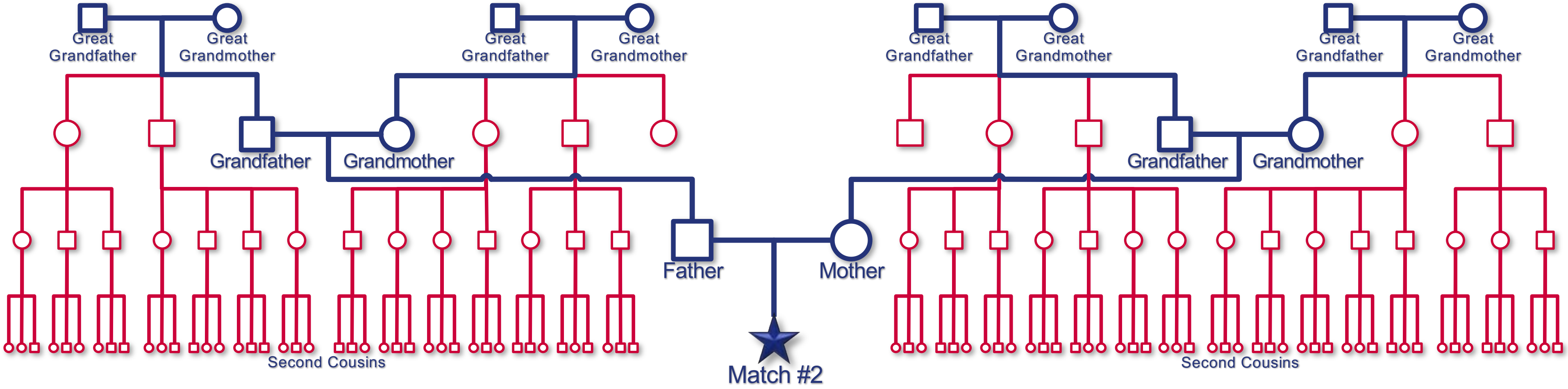
Grandfather
Grandmother
Grandfather
Grandmother

Father
Mother

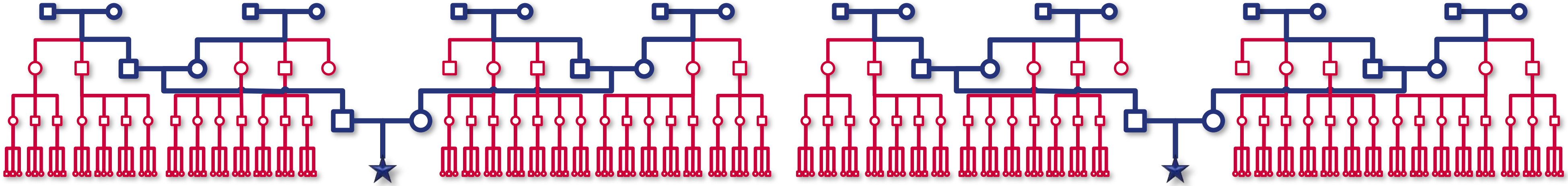
Match #2

Triangulation

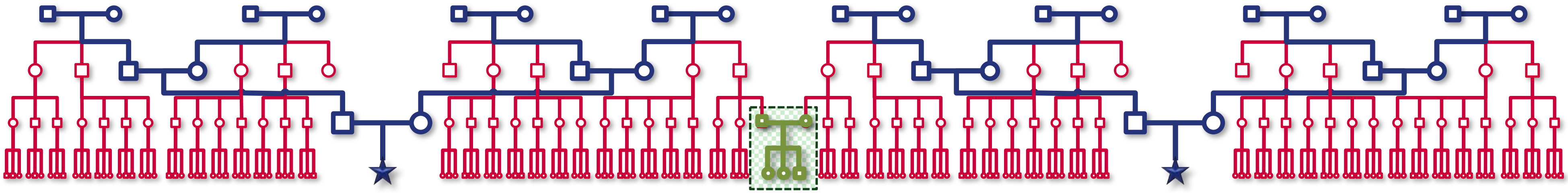




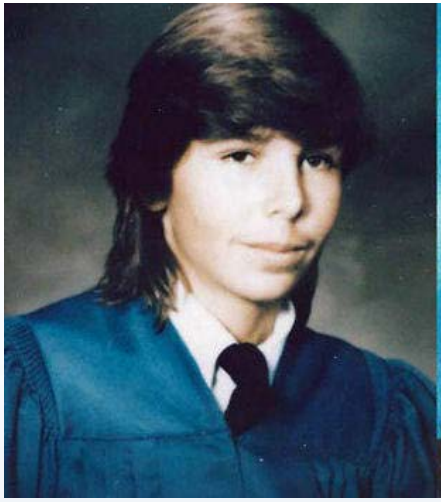
Triangulation



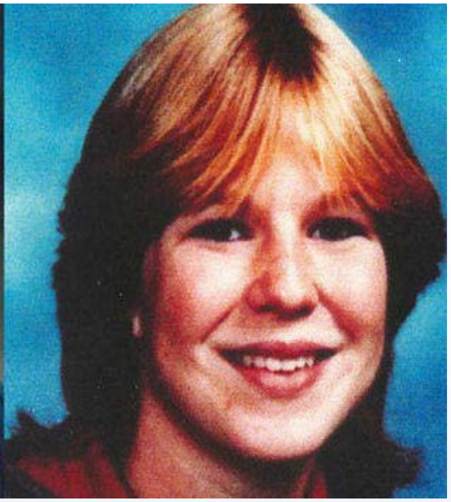
Triangulation



Case Study #1

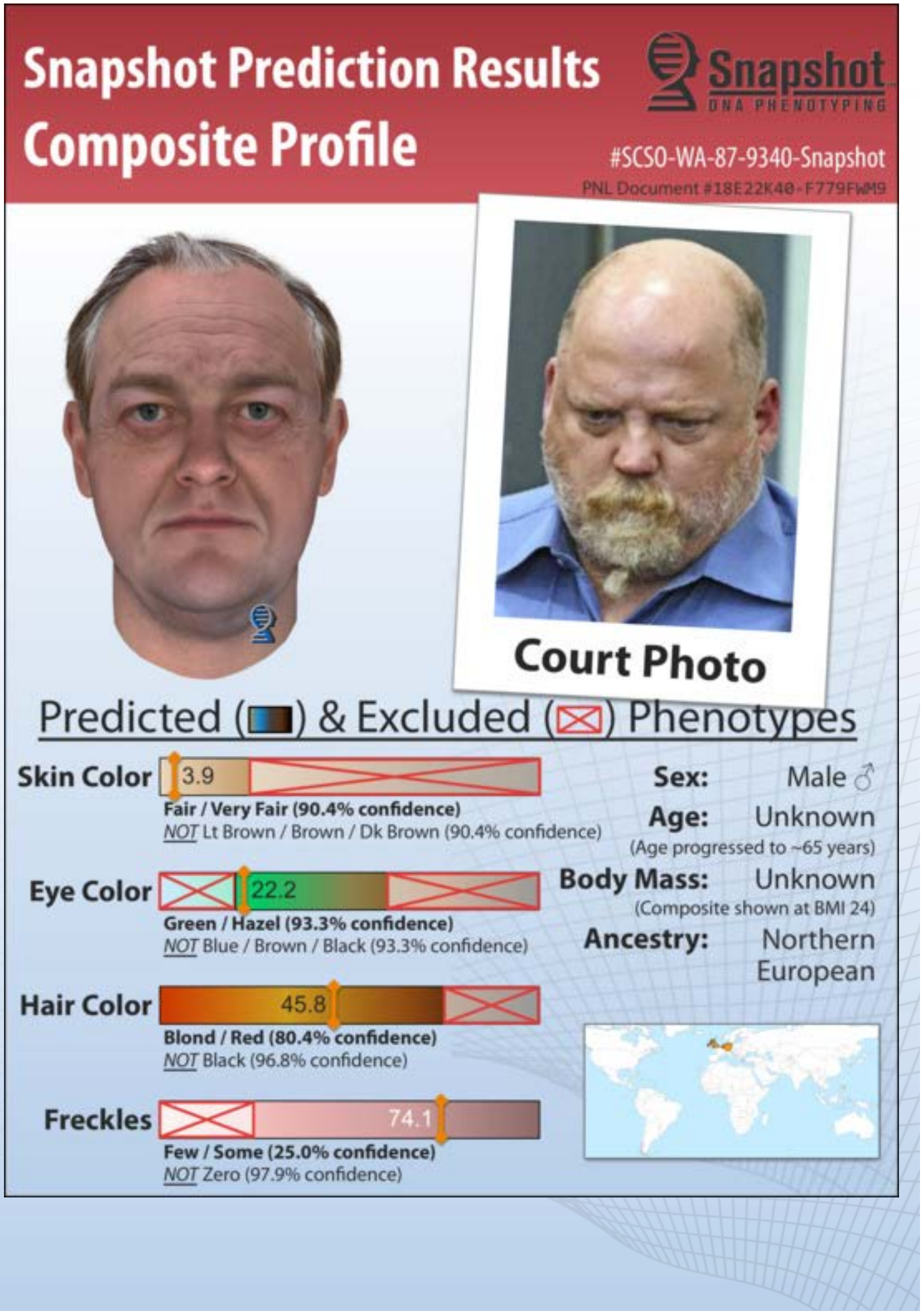


Jay Cook



Tanya Van Cuylenborg

- 1987 double murder of Jay Cook (20) and Tanya Van Cuylenborg (18)
- Snohomish County, WA



Snapshot Prediction Results
Composite Profile

Snapshot DNA PHENOTYPING

#SCSO-WA-87-9340-Snapshot
PNL Document #18E22K40-F779FWM9

Court Photo

Predicted (▬) & Excluded (☒) Phenotypes

Skin Color
Fair / Very Fair (90.4% confidence)
NOT Lt Brown / Brown / Dk Brown (90.4% confidence)

Eye Color
Green / Hazel (93.3% confidence)
NOT Blue / Brown / Black (93.3% confidence)

Hair Color
Blond / Red (80.4% confidence)
NOT Black (96.8% confidence)

Freckles
Few / Some (25.0% confidence)
NOT Zero (97.9% confidence)

Sex: Male ♂
Age: Unknown
(Age progressed to ~65 years)
Body Mass: Unknown
(Composite shown at BMI 24)
Ancestry: Northern European

Case Study #1

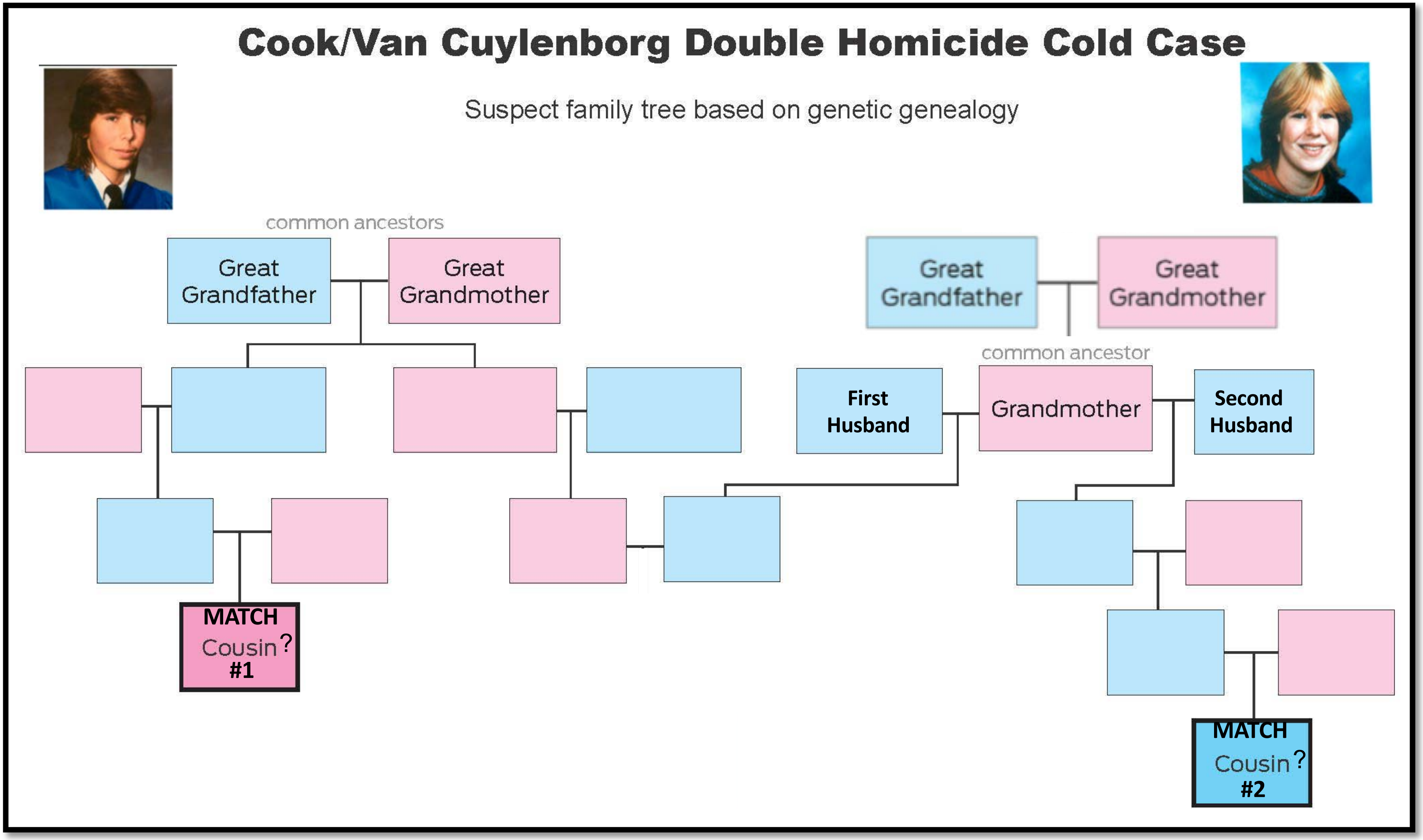
Cook/Van Cuylenborg Double Homicide Cold Case

Suspect family tree based on genetic genealogy

MATCH
Cousin ?
#1

MATCH
Cousin ?
#2

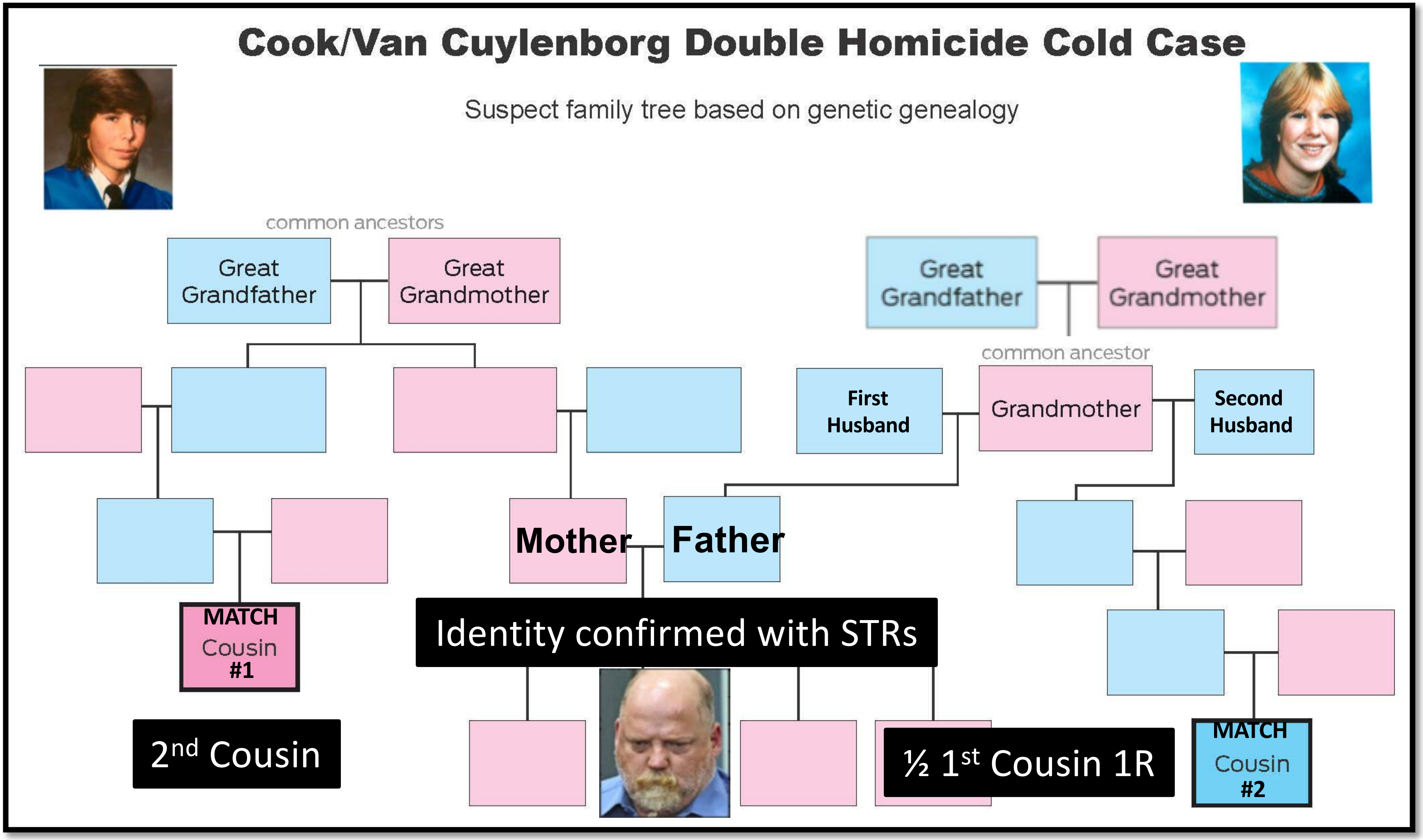
Case Study #1



Cook/Van Cuylenborg Double Homicide Cold Case

Suspect family tree based on genetic genealogy

Case Study #1



Cook/Van Cuylenborg Double Homicide Cold Case

Suspect family tree based on genetic genealogy

common ancestors

Great Grandfather

Great Grandmother

Great Grandfather

Great Grandmother

common ancestor

First Husband

Grandmother

Second Husband

Mother

Father

Identity confirmed with STRs

MATCH
Cousin
#1

2nd Cousin

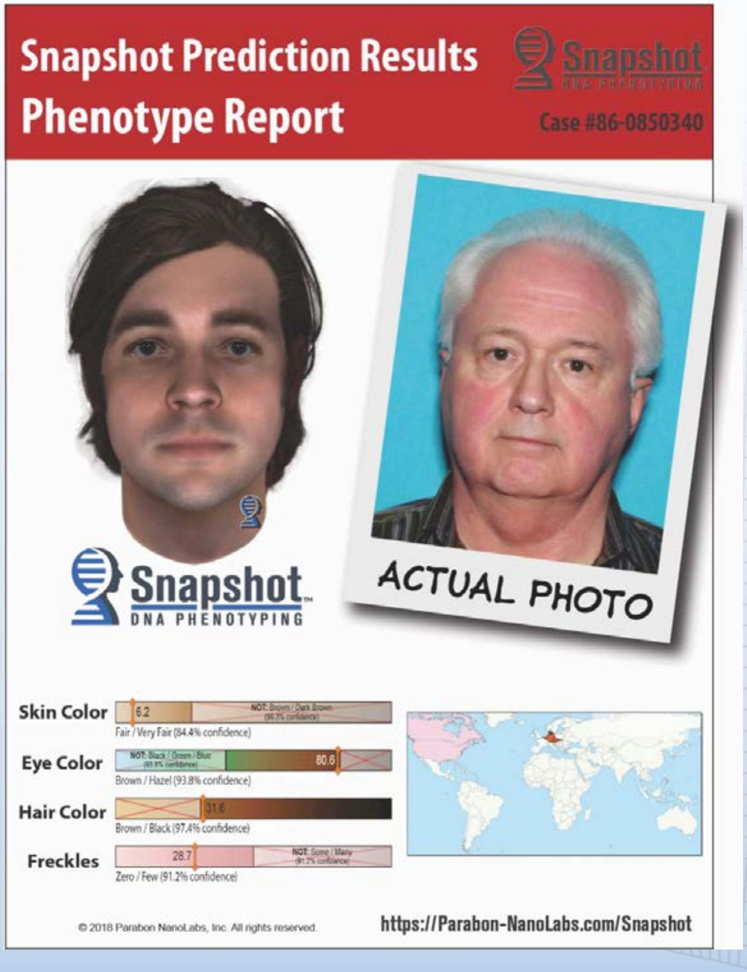
½ 1st Cousin 1R

MATCH
Cousin
#2

Case Study #2

- 1986 homicide of 12-year-old Michella Welch
- Tacoma, Washington

Michella Welch, 12



Snapshot Prediction Results
Phenotype Report

Snapshot DNA PHENOTYPING

Case #86-0850340

Snapshot DNA PHENOTYPING

ACTUAL PHOTO

Skin Color 6.2 — NOT: Brown / Dark Brown (99.3% confidence)
Fair / Very Fair (84.4% confidence)

Eye Color NOT: Black / Green / Blue — 80.6
Brown / Hazel (93.8% confidence)

Hair Color 31.6
Brown / Black (97.4% confidence)

Freckles 28.7 — NOT: Some / Many (91.2% confidence)
Zero / Few (91.2% confidence)

© 2018 Parabon NanoLabs, Inc. All rights reserved.

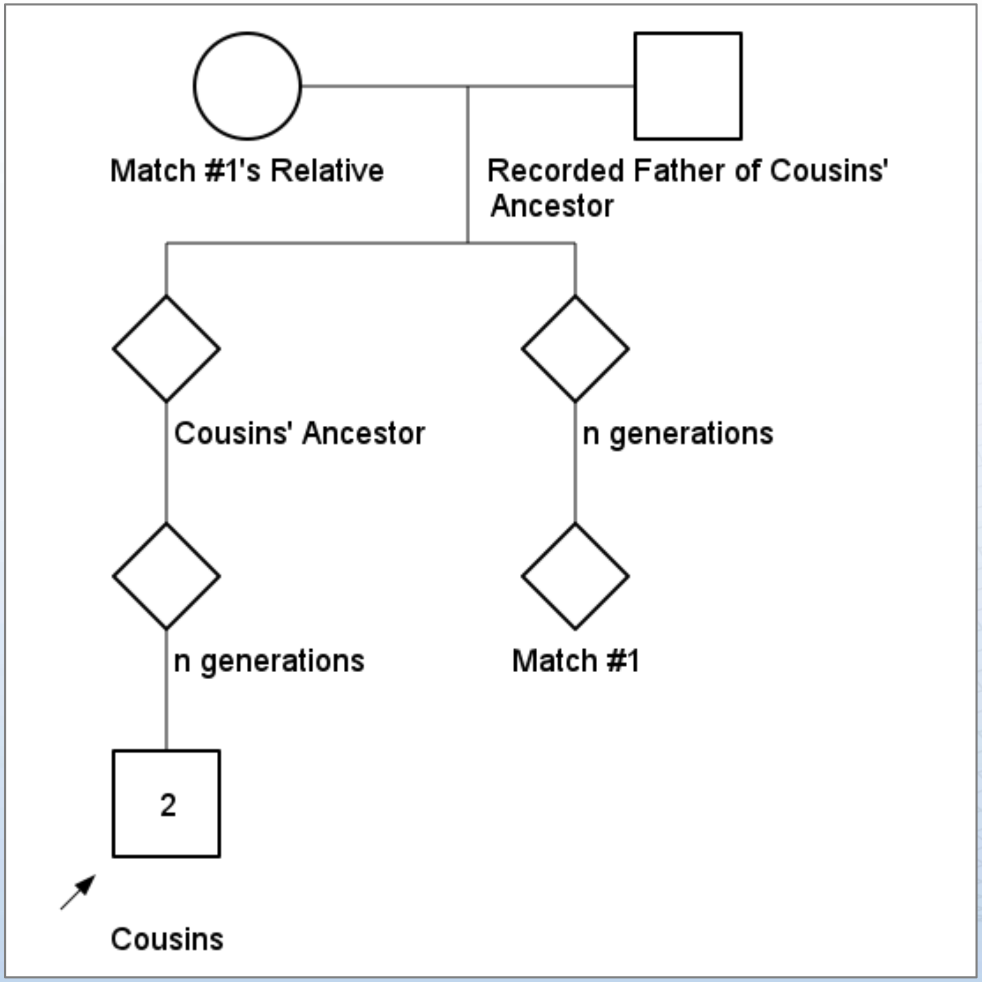
https://Parabon-NanoLabs.com/Snapshot

Case Study #2

- Predicted genetic ancestry: 90% Northern European + 10% North Native American
- Top two matches in GEDmatch shared no DNA
- Pair of brothers of possible interest
    - Cousins to Match #1
    - Lived in Tacoma in 1986
    - Had two Native American great-great-grandmothers (~1/8 Native American)
    - But ...

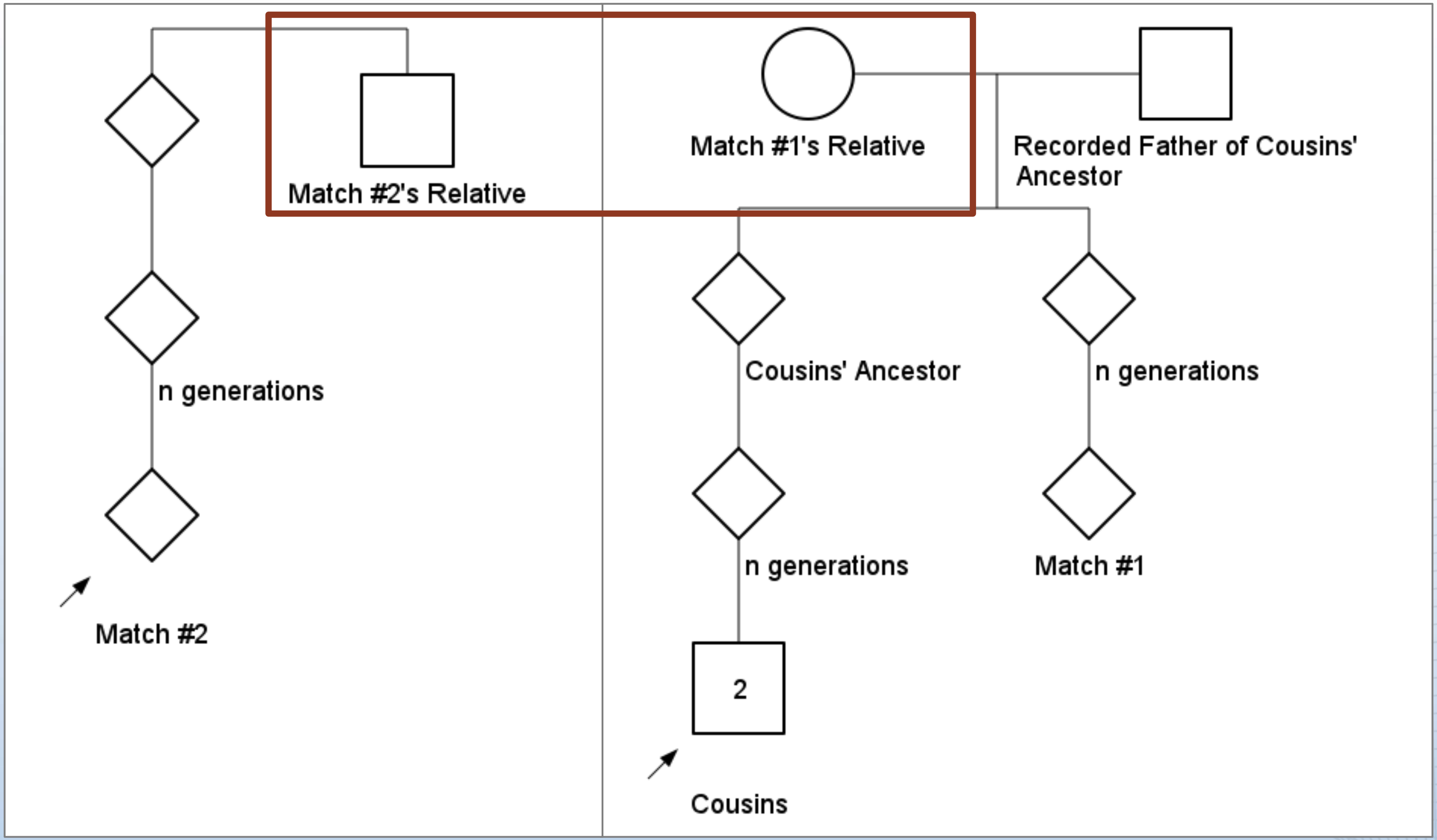
Case Study #2

- One degree closer than expected given amount of shared DNA; no connection to match #2



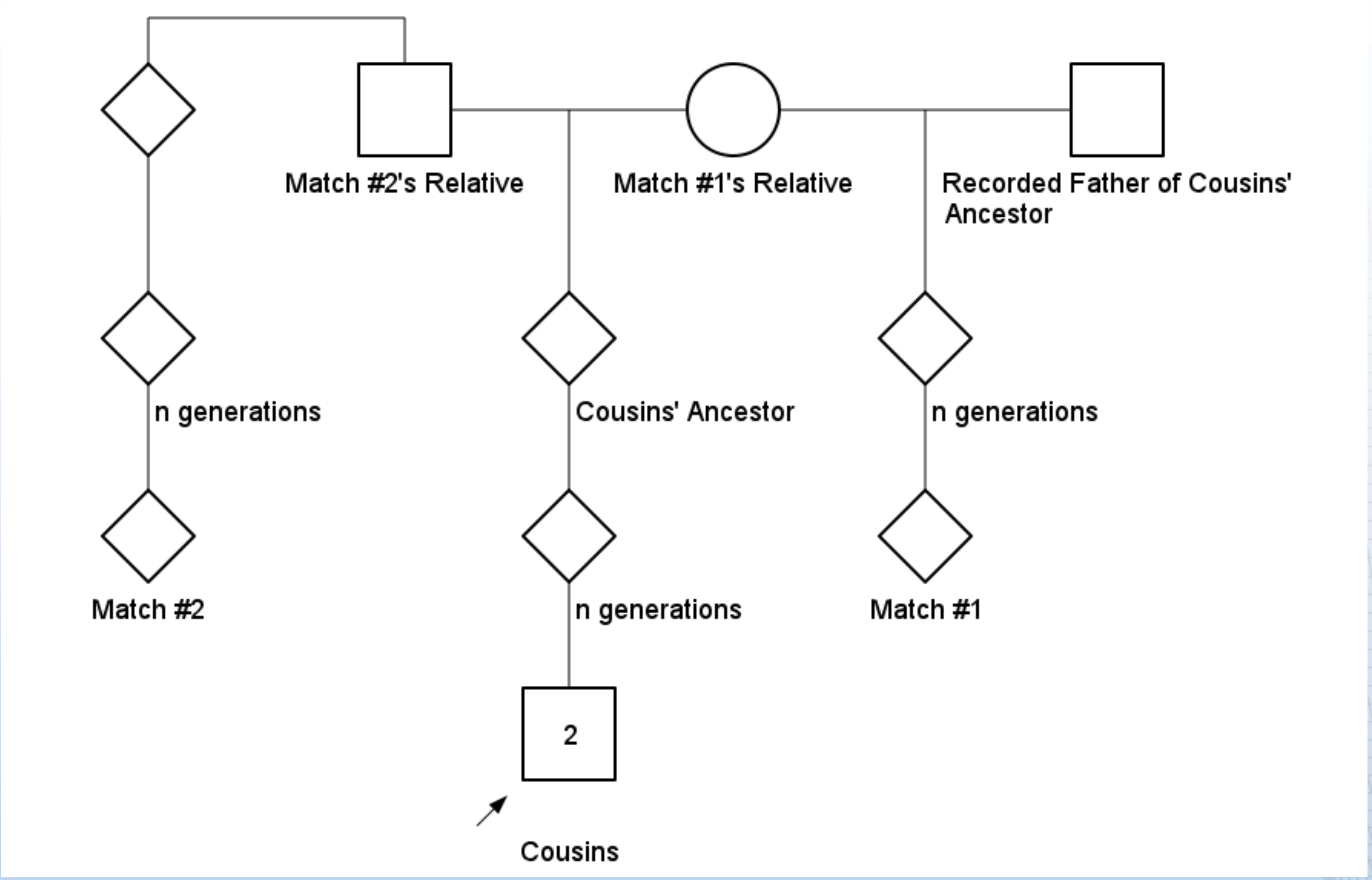
Case Study #2

- Match #1's Relative and Match #2's Relative lived in the same small town in the year before the Cousins' Ancestor was born



Case Study #2

- Misattributed paternity can account for the amount of shared DNA and the connection to Match #2



Forensic Analysis of Next-Generation Sequencing (NGS) Data

PARABON® Fx™

Forensic Analysis Platform

PARABON®
NANOLABS

Forensic NGS

- NGS is a powerful new tool for forensics
  - STRs and SNPs in a single assay
  - Sequencing STRs yields additional alleles that are helpful for answering identity questions and analyzing mixed samples

Forensic NGS

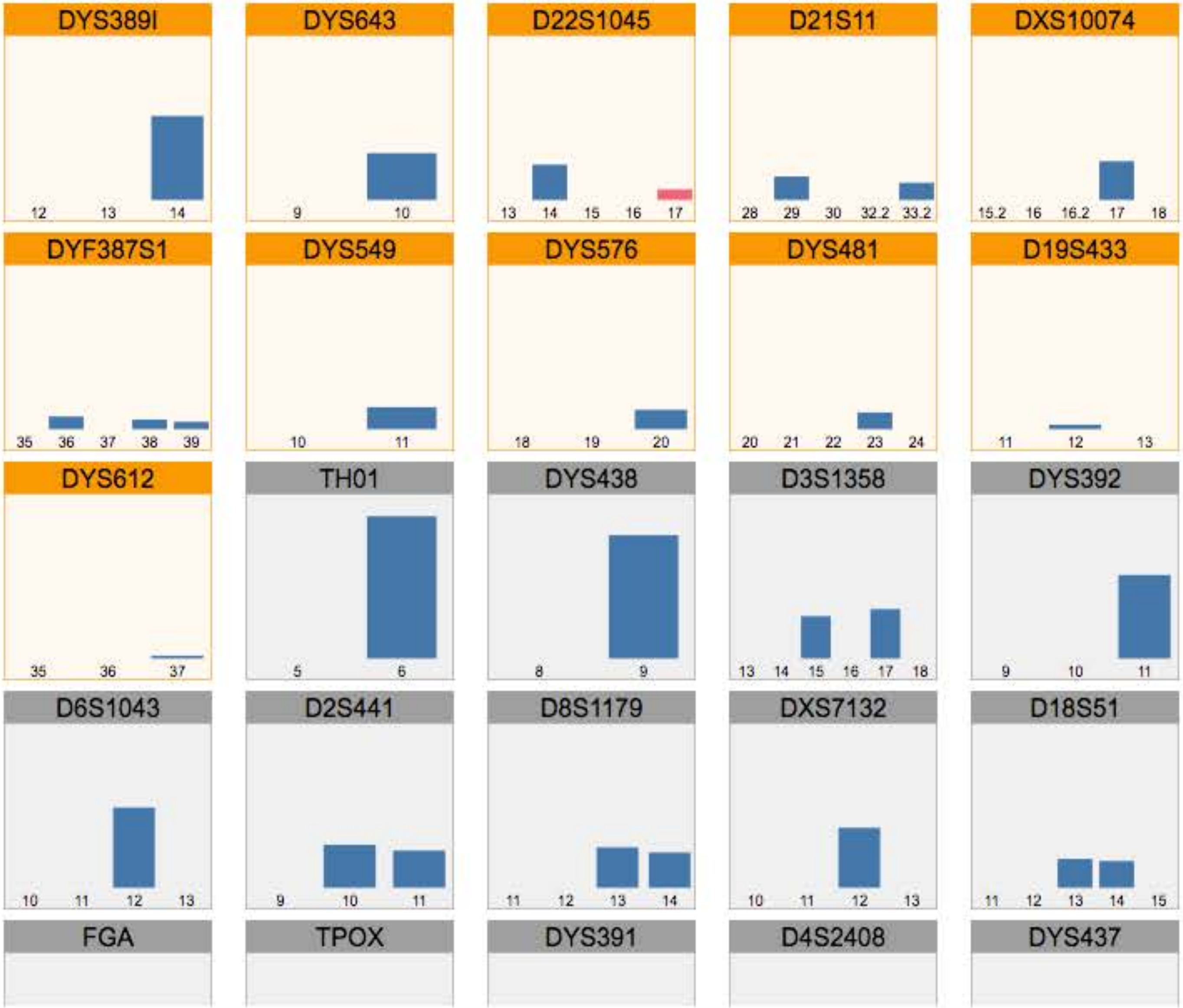
- **DoD problem**: multiple NGS system manufacturers, each with their own proprietary software
  - Can’t compare samples across NGS kits
  - Analytical capabilities are limited to those tools included in manufacturers’ software
- **Solicitation**: platform-agnostic software that can analyze STRs, SNPs, and mitochondrial DNA from NGS data

Forensic NGS

- Parabon Fx™ Forensic Analysis Platform
  - Analyze raw data from any NGS system and compare across kit types
  - Extensible platform with a plugin architecture
  - Dozens of built-in plugins provide extensive analytical capabilities
  - Can be installed on a local server or accessed over the internet
  - CAC card access, STIG-compliant

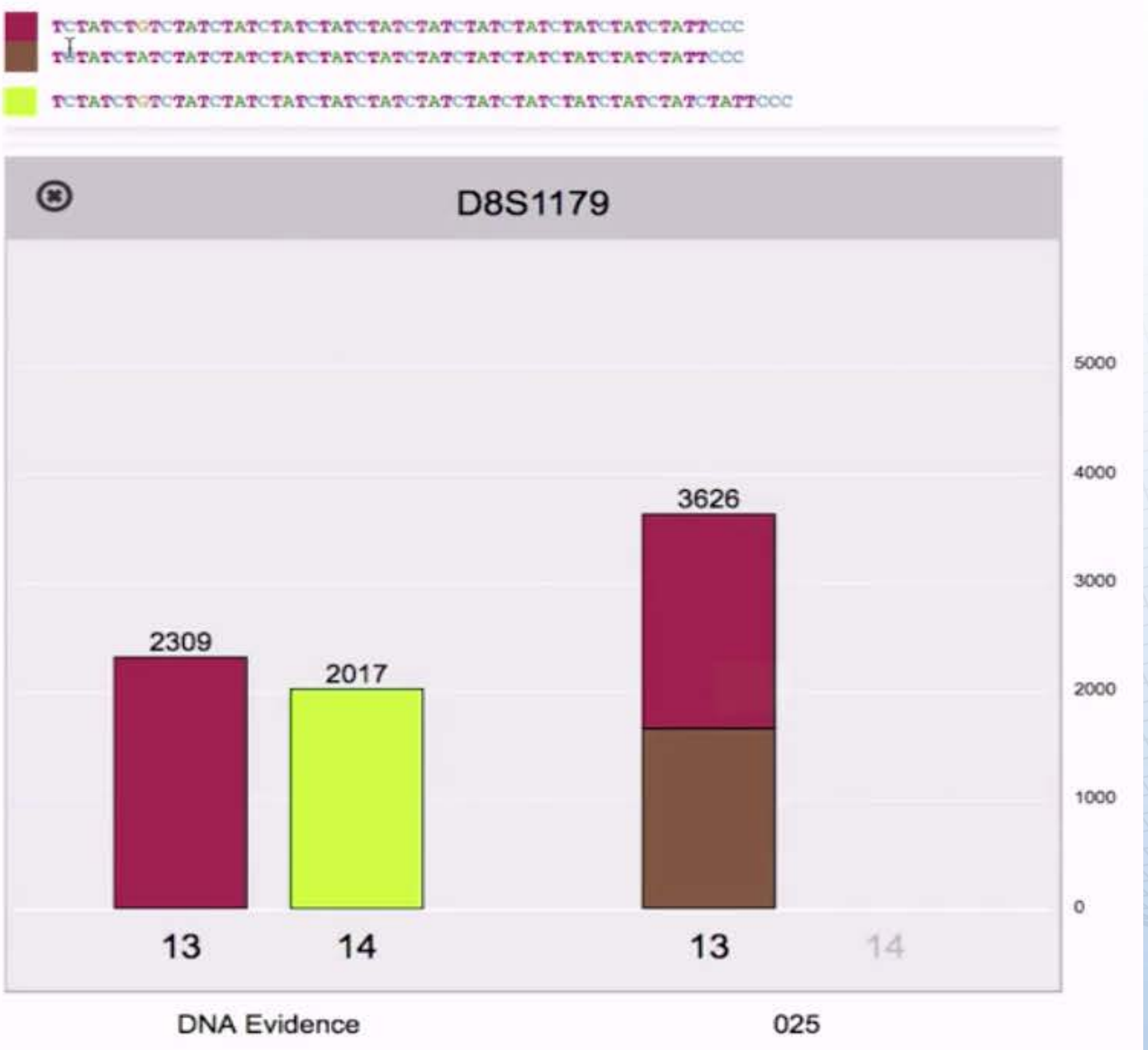
NGS STRs

- Create, edit, and approve an STR profile using allele length or sequence



NGS STRs

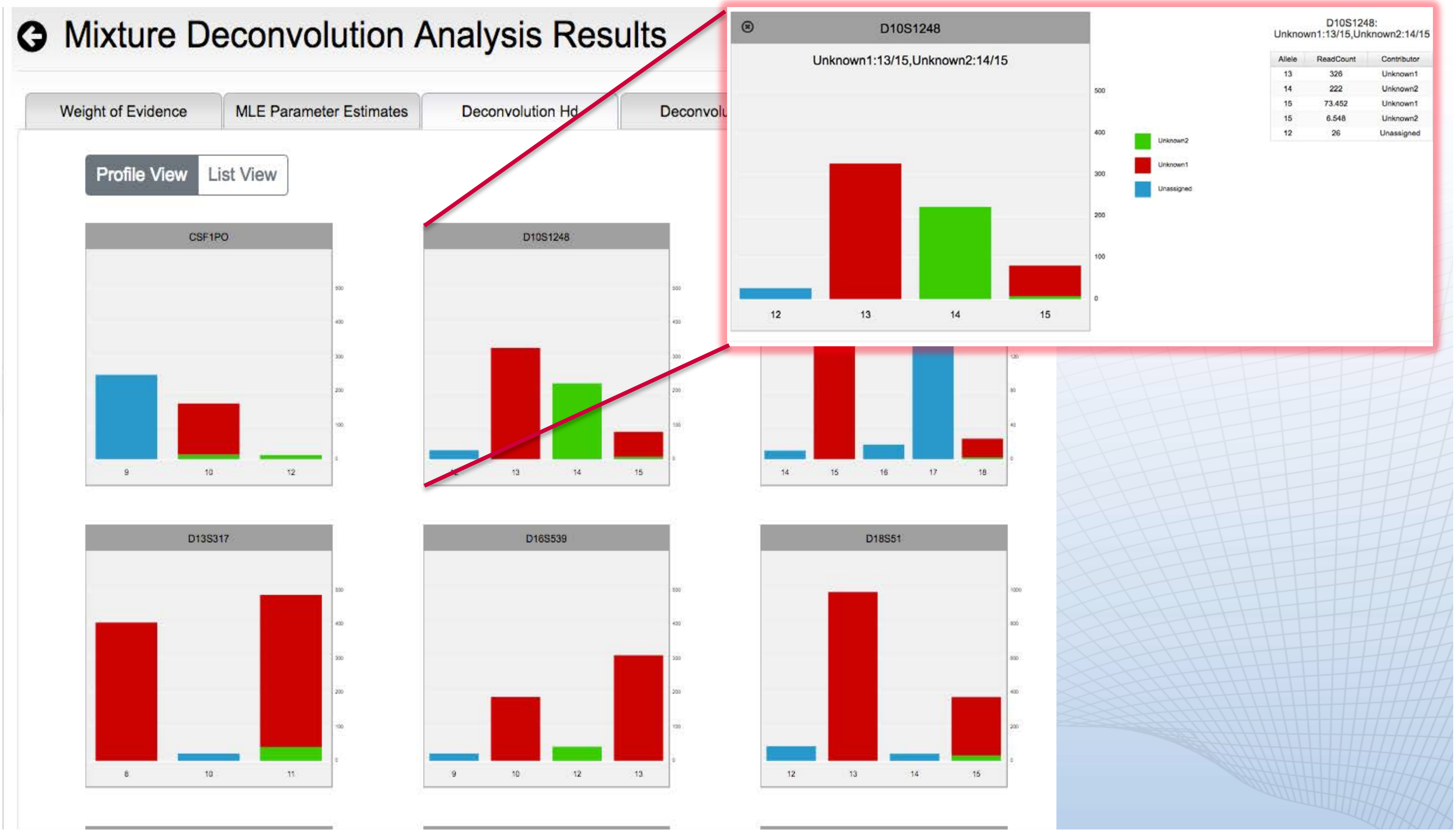
- Compare STR profiles for identity, Y-chromosome lineage, parentage, or sibship using length or sequence alleles



🟥 TCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATTCCC
🟫 TCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATTCCC
🟩 TCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATTCCC

NGS STRs

- Deconvolute mixed STR profiles using length or sequence alleles



Mixture Deconvolution Analysis Results

D10S1248:
Unknown1:13/15,Unknown2:14/15

Allele	ReadCount	Contributor
13	326	Unknown1
14	222	Unknown2
15	73.452	Unknown1
15	6.548	Unknown2
12	26	Unassigned

NGS SNPs

- Align sequence reads to the reference genome and call a SNP profile

chr5:179,263,724

Total Count 235

A 172 (73%)

C 0

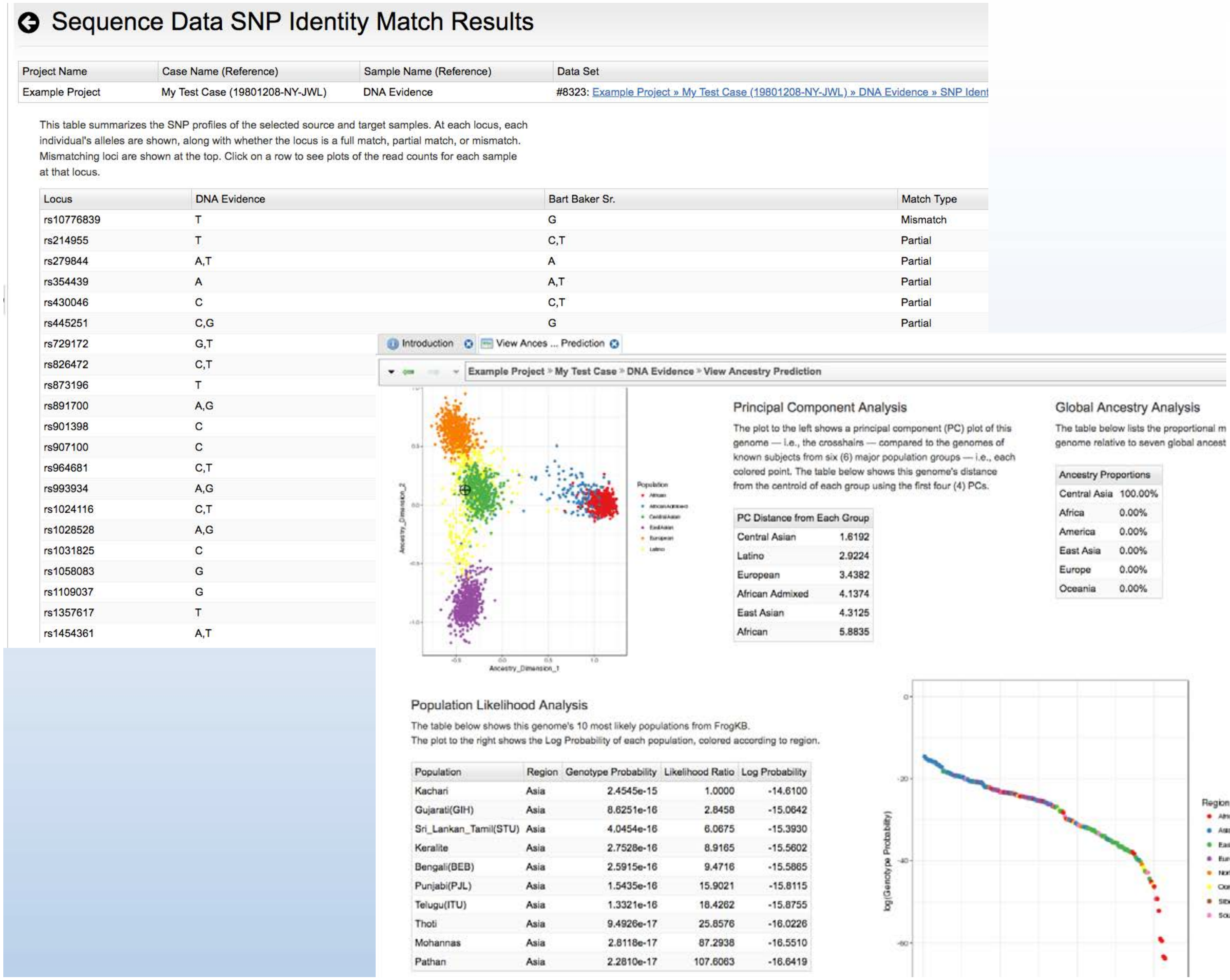
G 63 (26%)

T 0

N 0

NGS SNPs

- Perform identity matching, infer ancestry, predict phenotype



Sequence Data SNP Identity Match Results

Project Name	Case Name (Reference)	Sample Name (Reference)	Data Set
Example Project	My Test Case (19801208-NY-JWL)	DNA Evidence	#8323: Example Project » My Test Case (19801208-NY-JWL) » DNA Evidence » SNP Ident

This table summarizes the SNP profiles of the selected source and target samples. At each locus, each individual's alleles are shown, along with whether the locus is a full match, partial match, or mismatch. Mismatching loci are shown at the top. Click on a row to see plots of the read counts for each sample at that locus.

Locus	DNA Evidence	Bart Baker Sr.	Match Type
rs10776839	T	G	Mismatch
rs214955	T	C,T	Partial
rs279844	A,T	A	Partial
rs354439	A	A,T	Partial
rs430046	C	C,T	Partial
rs445251	C,G	G	Partial
rs729172	G,T		
rs826472	C,T		
rs873196	T		
rs891700	A,G		
rs901398	C		
rs907100	C		
rs964681	C,T		
rs993934	A,G		
rs1024116	C,T		
rs1028528	A,G		
rs1031825	C		
rs1058083	G		
rs1109037	G		
rs1357617	T		
rs1454361	A,T		

Principal Component Analysis

The plot to the left shows a principal component (PC) plot of this genome — i.e., the crosshairs — compared to the genomes of known subjects from six (6) major population groups — i.e., each colored point. The table below shows this genome's distance from the centroid of each group using the first four (4) PCs.

PC Distance from Each Group	
Central Asian	1.6192
Latino	2.9224
European	3.4382
African Admixed	4.1374
East Asian	4.3125
African	5.8835

Global Ancestry Analysis

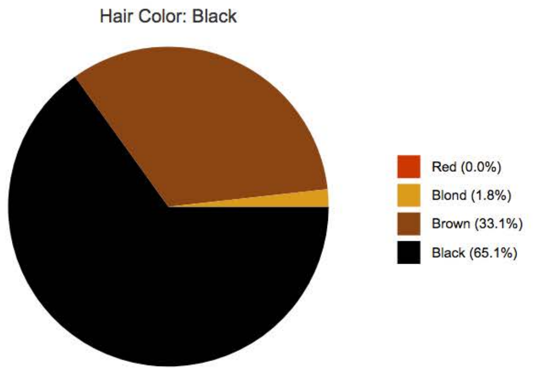
The table below lists the proportional m genome relative to seven global ancest

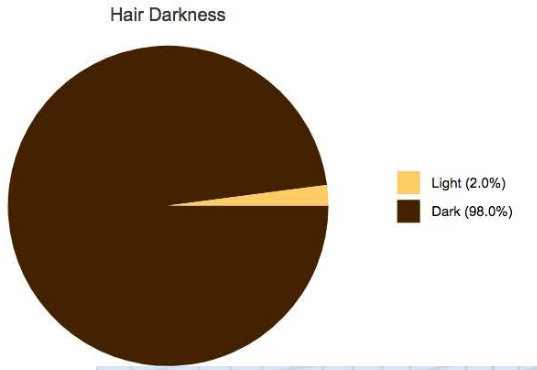
Ancestry Proportions	
Central Asia	100.00%
Africa	0.00%
America	0.00%
East Asia	0.00%
Europe	0.00%
Oceania	0.00%

Population Likelihood Analysis

The table below shows this genome's 10 most likely populations from FrogKB.
The plot to the right shows the Log Probability of each population, colored according to region.

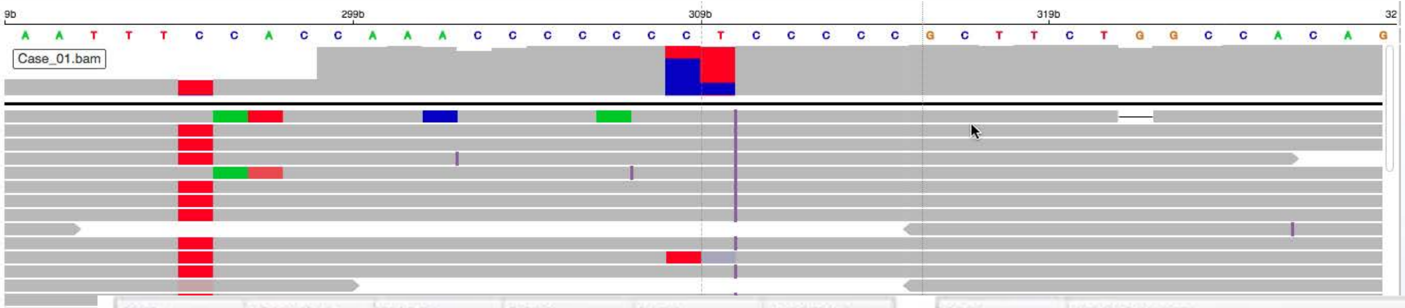
Population	Region	Genotype Probability	Likelihood Ratio	Log Probability
Kachari	Asia	2.4545e-15	1.0000	-14.6100
Gujarati(GIH)	Asia	8.6251e-16	2.8458	-15.0642
Sri_Lankan_Tamil(STU)	Asia	4.0454e-16	6.0675	-15.3930
Keralite	Asia	2.7528e-16	8.9165	-15.5602
Bengali(BEB)	Asia	2.5915e-16	9.4716	-15.5865
Punjabi(PJL)	Asia	1.5435e-16	15.9021	-15.8115
Telugu(ITU)	Asia	1.3321e-16	18.4262	-15.8755
Thoti	Asia	9.4926e-17	25.8576	-16.0226
Mohannas	Asia	2.8118e-17	87.2938	-16.5510
Pathan	Asia	2.2810e-17	107.6063	-16.6419

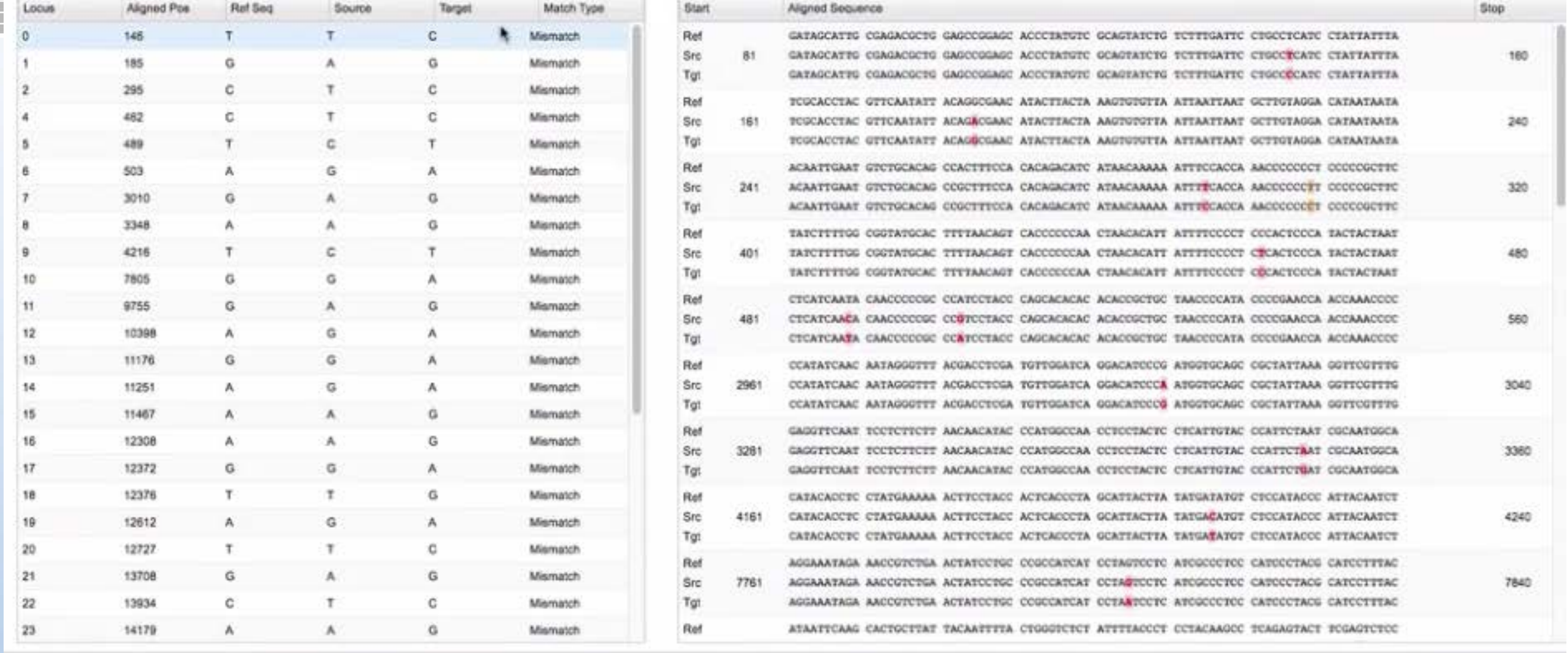




NGS mtDNA

- Create a mitochondrial DNA sequence profile and perform haplotype-based or string-based lineage matching



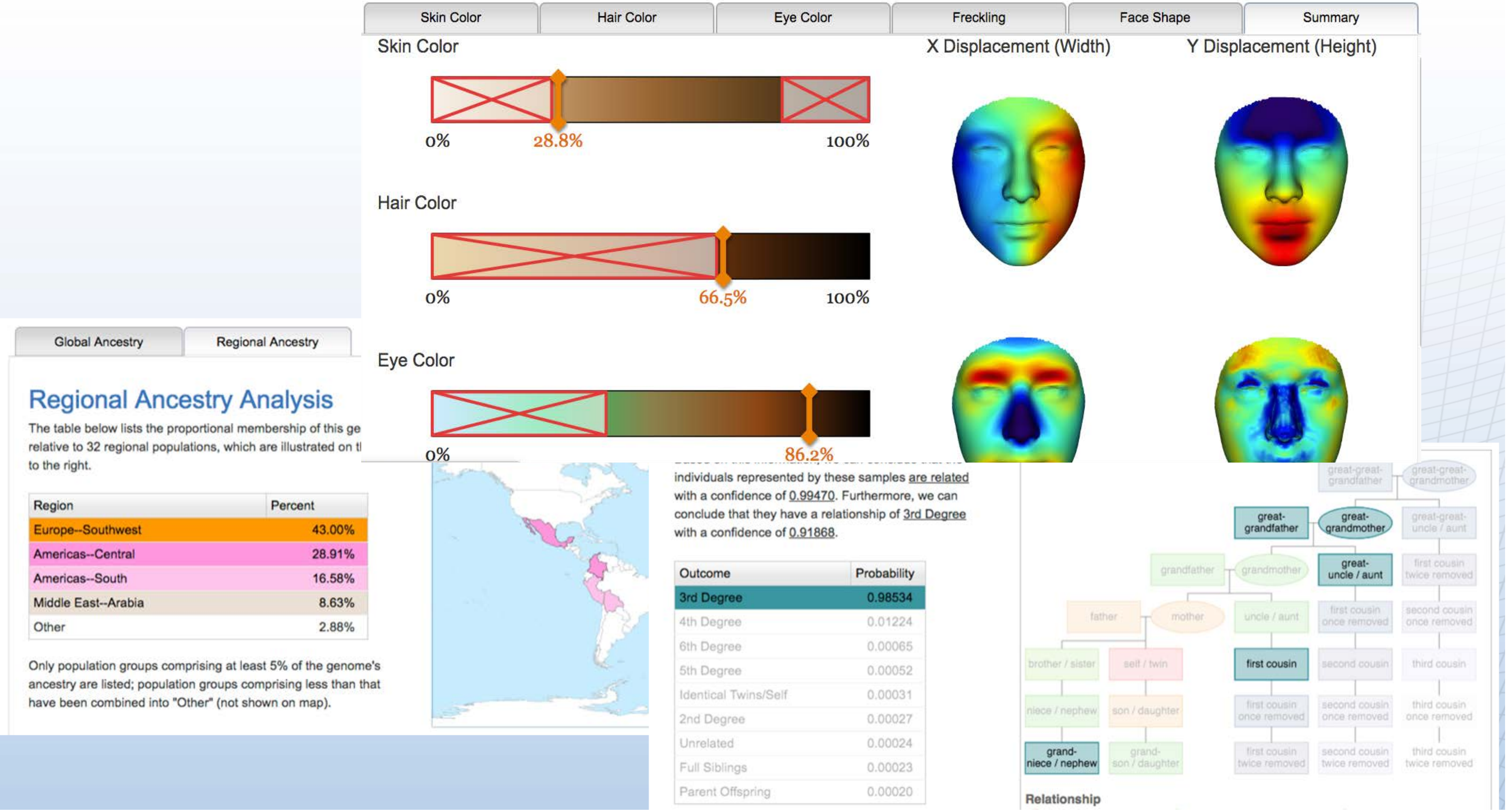


Locus	Aligned Pos	Ref Seq	Source	Target	Match Type
0	146	T	T	C	Mismatch
1	185	G	A	G	Mismatch
2	295	C	T	C	Mismatch
4	462	C	T	C	Mismatch
5	489	T	C	T	Mismatch
6	503	A	G	A	Mismatch
7	3010	G	A	G	Mismatch
8	3348	A	A	G	Mismatch
9	4216	T	C	T	Mismatch
10	7805	G	G	A	Mismatch
11	9755	G	A	G	Mismatch
12	10398	A	G	A	Mismatch
13	11176	G	G	A	Mismatch
14	11251	A	G	A	Mismatch
15	11467	A	A	G	Mismatch
16	12308	A	A	G	Mismatch
17	12372	G	G	A	Mismatch
18	12376	T	T	G	Mismatch
19	12612	A	G	A	Mismatch
20	12727	T	T	C	Mismatch
21	13708	G	A	G	Mismatch
22	13934	C	T	C	Mismatch
23	14179	A	A	G	Mismatch

```
Start     Aligned Sequence                                                                                 Stop
Ref        GATAGCATTG CGAGACGCTG GAGCCGGAGC ACCCTATGTC GCAGTATCTG TCTTTGATTC CTGCCTCATC CTATTATTTA
Src   81   GATAGCATTG CGAGACGCTG GAGCCGGAGC ACCCTATGTC GCAGTATCTG TCTTTGATTC CTGCCTCATC CTATTATTTA   160
Tgt        GATAGCATTG CGAGACGCTG GAGCCGGAGC ACCCTATGTC GCAGTATCTG TCTTTGATTC CTGCCCCATC CTATTATTTA
Ref        TCGCACCTAC GTTCAATATT ACAGGCGAAC ATACTTACTA AAGTGTGTTA ATTAATTAAT GCTTGTAGGA CATAATAATA
Src   161  TCGCACCTAC GTTCAATATT ACAGACGAAC ATACTTACTA AAGTGTGTTA ATTAATTAAT GCTTGTAGGA CATAATAATA   240
Tgt        TCGCACCTAC GTTCAATATT ACAGGCGAAC ATACTTACTA AAGTGTGTTA ATTAATTAAT GCTTGTAGGA CATAATAATA
Ref        ACAATTGAAT GTCTGCACAG CCACTTTCCA CACAGACATC ATAACAAAAA ATTTCCACCA AACCCCCCCT CCCCCGCTTC
Src   241  ACAATTGAAT GTCTGCACAG CCGCTTTCCA CACAGACATC ATAACAAAAA ATTTTCACCA AACCCCCCTT CCCCCGCTTC   320
Tgt        ACAATTGAAT GTCTGCACAG CCGCTTTCCA CACAGACATC ATAACAAAAA ATTTCCACCA AACCCCCCCT CCCCCGCTTC
Ref        TATCTTTTGG CGGTATGCAC TTTTAACAGT CACCCCCCAA CTAACACATT ATTTTCCCCT CCCACTCCCA TACTACTAAT
Src   401  TATCTTTTGG CGGTATGCAC TTTTAACAGT CACCCCCCAA CTAACACATT ATTTTCCCCT CTCACTCCCA TACTACTAAT   480
Tgt        TATCTTTTGG CGGTATGCAC TTTTAACAGT CACCCCCCAA CTAACACATT ATTTTCCCCT CCCACTCCCA TACTACTAAT
Ref        CTCATCAATA CAACCCCCGC CCATCCTACC CAGCACACAC ACACCGCTGC TAACCCCATA CCCCGAACCA ACCAAACCCC
Src   481  CTCATCAACA CAACCCCCGC CCGTCCTACC CAGCACACAC ACACCGCTGC TAACCCCATA CCCCGAACCA ACCAAACCCC   560
Tgt        CTCATCAATA CAACCCCCGC CCATCCTACC CAGCACACAC ACACCGCTGC TAACCCCATA CCCCGAACCA ACCAAACCCC
Ref        CCATATCAAC AATAGGGTTT ACGACCTCGA TGTTGGATCA GGACATCCCG ATGGTGCAGC CGCTATTAAA GGTTCGTTTG
Src   2961 CCATATCAAC AATAGGGTTT ACGACCTCGA TGTTGGATCA GGACATCCCA ATGGTGCAGC CGCTATTAAA GGTTCGTTTG   3040
Tgt        CCATATCAAC AATAGGGTTT ACGACCTCGA TGTTGGATCA GGACATCCCG ATGGTGCAGC CGCTATTAAA GGTTCGTTTG
Ref        GAGGTTCAAT TCCTCTTCTT AACAACATAC CCATGGCCAA CCTCCTACTC CTCATTGTAC CCATTCTAAT CGCAATGGCA
Src   3281 GAGGTTCAAT TCCTCTTCTT AACAACATAC CCATGGCCAA CCTCCTACTC CTCATTGTAC CCATTCTAAT CGCAATGGCA   3360
Tgt        GAGGTTCAAT TCCTCTTCTT AACAACATAC CCATGGCCAA CCTCCTACTC CTCATTGTAC CCATTCTGAT CGCAATGGCA
Ref        CATACACCTC CTATGAAAAA ACTTCCTACC ACTCACCCTA GCATTACTTA TATGATATGT CTCCATACCC ATTACAATCT
Src   4161 CATACACCTC CTATGAAAAA ACTTCCTACC ACTCACCCTA GCATTACTTA TATGACATGT CTCCATACCC ATTACAATCT   4240
Tgt        CATACACCTC CTATGAAAAA ACTTCCTACC ACTCACCCTA GCATTACTTA TATGATATGT CTCCATACCC ATTACAATCT
Ref        AGGAAATAGA AACCGTCTGA ACTATCCTGC CCGCCATCAT CCTAGTCCTC ATCGCCCTCC CATCCCTACG CATCCTTTAC
Src   7761 AGGAAATAGA AACCGTCTGA ACTATCCTGC CCGCCATCAT CCTAGTCCTC ATCGCCCTCC CATCCCTACG CATCCTTTAC   7840
Tgt        AGGAAATAGA AACCGTCTGA ACTATCCTGC CCGCCATCAT CCTAATCCTC ATCGCCCTCC CATCCCTACG CATCCTTTAC
Ref        ATAATTCAAG CACTGCTTAT TACAATTTTA CTGGGTCTCT ATTTTACCCT CCTACAAGCC TCAGAGTACT TCGAGTCTCC
```

Microarray Plugins

- The DoD distribution of Fx also includes plugins for ancestry, phenotype, and kinship analysis of SNP microarray data using Snapshot



Regional Ancestry Analysis

The table below lists the proportional membership of this ge
relative to 32 regional populations, which are illustrated on t
to the right.

Region	Percent
Europe--Southwest	43.00%
Americas--Central	28.91%
Americas--South	16.58%
Middle East--Arabia	8.63%
Other	2.88%

Only population groups comprising at least 5% of the genome's ancestry are listed; population groups comprising less than that have been combined into "Other" (not shown on map).

individuals represented by these samples are related with a confidence of 0.99470. Furthermore, we can conclude that they have a relationship of 3rd Degree with a confidence of 0.91868.

Outcome	Probability
3rd Degree	0.98534
4th Degree	0.01224
6th Degree	0.00065
5th Degree	0.00052
Identical Twins/Self	0.00031
2nd Degree	0.00027
Unrelated	0.00024
Full Siblings	0.00023
Parent Offspring	0.00020

Fx Extensions

- New analyses developed by Parabon can quickly be distributed to DoD users as Fx plugins
  - Kinship analysis of highly degraded DNA
  - Ancestry analysis of highly degraded DNA
  - Automated pedigree reconstruction
  - Automated mixture deconvolution
  - Etc.

Questions?

ellen@parabon.com

PARABON®
NANOLABS