How a Modified Algorithm and Small SNP Set can be used for Fast and Accurate Extended Kinship Estimation

Melissa Kotkin

Sr. Global Product Manager

Workflow for Kinship Estimation

Sample

Investigator Quantiplex Pro
EZ2 Connect Fx
TissueLyser
QIAamplifier
QIAgility

Prepare

ForenSeq® Kits

MainstAY
DNA Signature Prep
Kintelligence
mtDNA Control Region
mtDNA Whole Genome

Sequence

MiSeq FGx® System

Analyze

Universal Analysis Software
GEDmatch and GEDmatch PRO

What is Forensic Investigative Genetic Genealogy (FIGG)?

Application of genetic genealogy to investigations to generate leads in criminal cases or identify remains

Gained notoriety in 2018 with the arrest and subsequent conviction of the Golden State Killer – Joseph DeAngelo

Verogen acquired GEDmatch in December 2019

- To secure its future and further empower its use for public genealogy research
- To improve utility, oversight and governance of forensic and missing persons comparisons
- To create a workflow specifically designed for forensic samples and forensic laboratories

GEDmatch users choose when uploading their profiles whether or not to permit their data to be used for criminal case comparisons

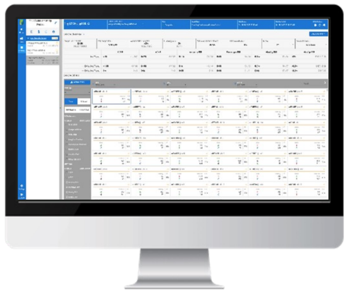
Forensic/Law Enforcement Comparisons governed by strict terms of use and privacy policies

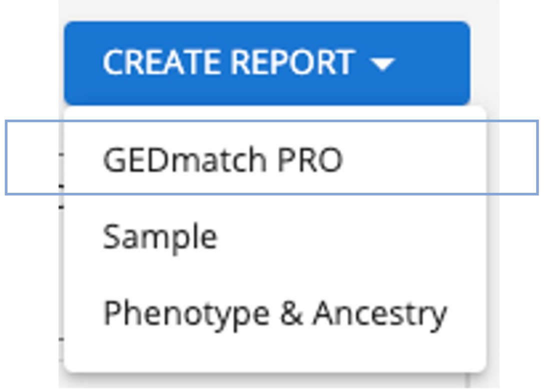
- All forensic and missing persons comparisons conducted via a dedicated portal – GEDmatch PRO
- Forensic case comparisons permitted only for violent crime and evidence profiles can be compared against opted in data only
- Comparisons to identify human remains for missing persons cases can be compared against the entire GEDmatch database

FIGG Workflow Overview

Kintelligence

Universal Analysis Software



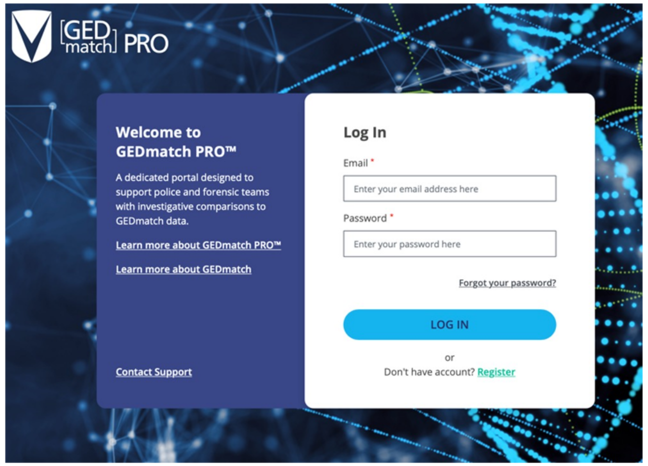


CREATE REPORT ▾

GEDmatch PRO

Sample

Phenotype & Ancestry



[GED match] PRO

Welcome to GEDmatch PRO™

A dedicated portal designed to support police and forensic teams with investigative comparisons to GEDmatch data.

Learn more about GEDmatch PRO™

Learn more about GEDmatch

Contact Support

Log In

Email *

Enter your email address here

Password *

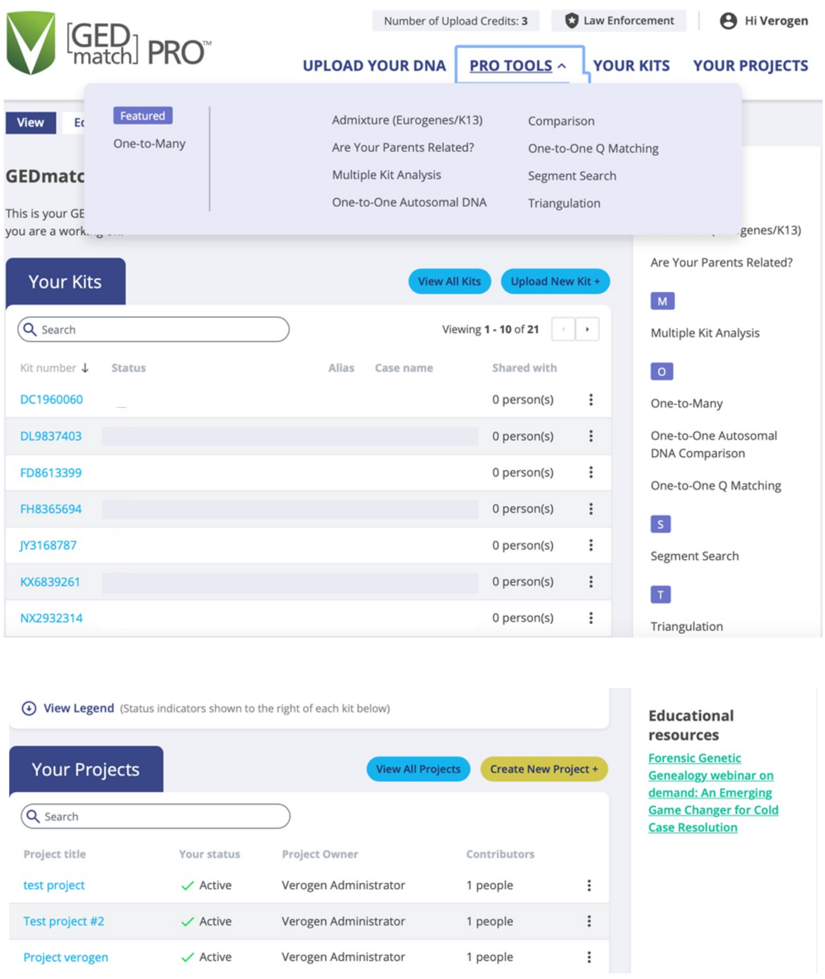
Enter your password here

Forgot your password?

LOG IN

or

Don't have account? Register



[GED match] PRO™

Number of Upload Credits: **3** | Law Enforcement | Hi Verogen

UPLOAD YOUR DNA | PRO TOOLS ^ | YOUR KITS | YOUR PROJECTS

View | Ed

Featured

One-to-Many

Admixture (Eurogenes/K13)

Are Your Parents Related?

Multiple Kit Analysis

One-to-One Autosomal DNA

Comparison

One-to-One Q Matching

Segment Search

Triangulation

GEDmatc

This is your GE you are a work...

Your Kits

View All Kits | Upload New Kit +

Search

Viewing **1 - 10** of **21**

Kit number ↓	Status	Alias	Case name	Shared with
DC1960060				0 person(s)
DL9837403				0 person(s)
FD8613399				0 person(s)
FH8365694				0 person(s)
JY3168787				0 person(s)
KX6839261				0 person(s)
NX2932314				0 person(s)

...genes/K13)

Are Your Parents Related?

M Multiple Kit Analysis

O One-to-Many

One-to-One Autosomal DNA Comparison

One-to-One Q Matching

S Segment Search

T Triangulation

View Legend (Status indicators shown to the right of each kit below)

Your Projects

View All Projects | Create New Project +

Search

Project title	Your status	Project Owner	Contributors
test project	✓ Active	Verogen Administrator	1 people
Test project #2	✓ Active	Verogen Administrator	1 people
Project verogen	✓ Active	Verogen Administrator	1 people

Educational resources

Forensic Genetic Genealogy webinar on demand: An Emerging Game Changer for Cold Case Resolution

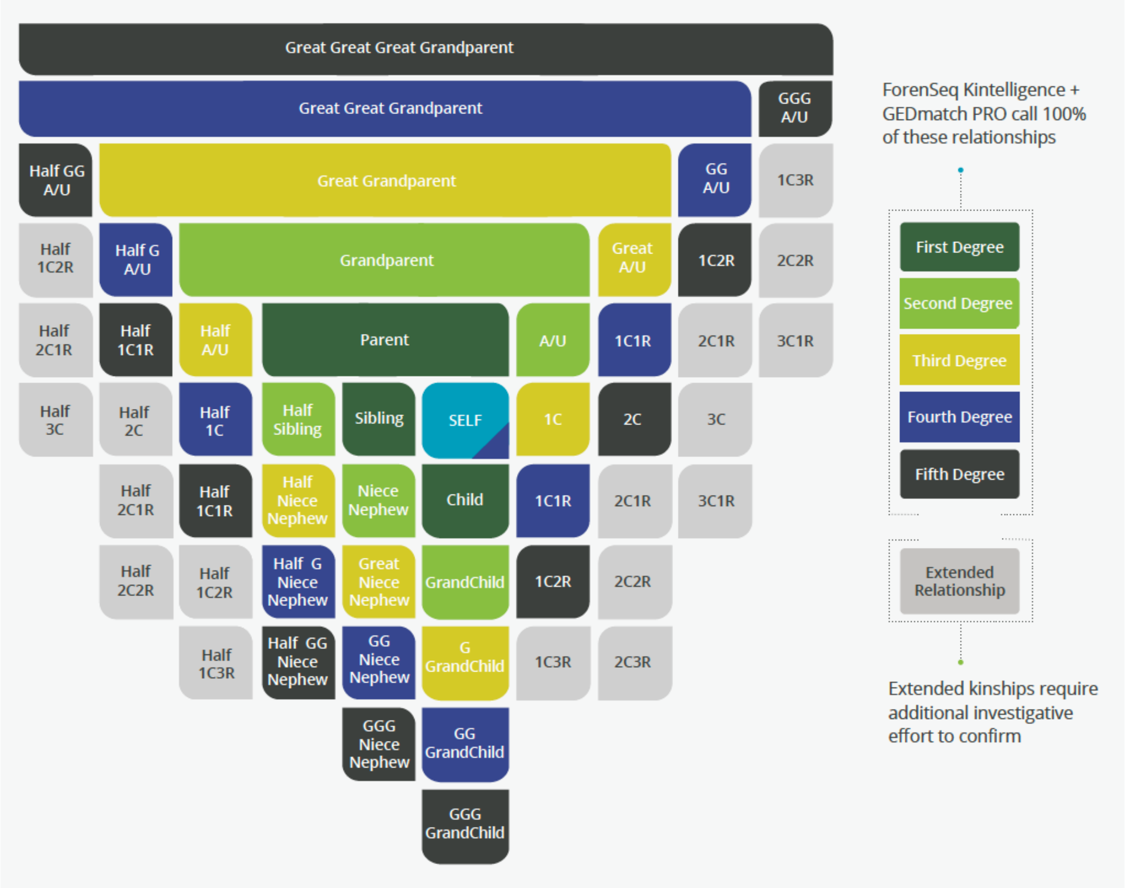
Choosing the Right Approach

STRs alone allow traditional kinship analysis that only reach direct (1st degree) relatives with any certainty

STRs and identity SNPs can reach 2nd and some 3rd degree relatives

Higher density, targeted SNP sets offer the chance to reach 3rd, 4th, and 5th degree relatives where direct references are no longer accessible or available

Mitochondrial DNA offers maternal lineage information to supplement autosomal data or as a lower power option



Great Great Great Grandparent

Great Great Grandparent | GGG A/U

Half GG A/U | Great Grandparent | GG A/U | 1C3R

Half 1C2R | Half G A/U | Grandparent | Great A/U | 1C2R | 2C2R

Half 2C1R | Half 1C1R | Half A/U | Parent | A/U | 1C1R | 2C1R | 3C1R

Half 3C | Half 2C | Half 1C | Half Sibling | Sibling | SELF | 1C | 2C | 3C

Half 2C1R | Half 1C1R | Half Niece Nephew | Niece Nephew | Child | 1C1R | 2C1R | 3C1R

Half 2C2R | Half 1C2R | Half G Niece Nephew | Great Niece Nephew | GrandChild | 1C2R | 2C2R

Half 1C3R | Half GG Niece Nephew | GG Niece Nephew | G GrandChild | 1C3R | 2C3R

GGG Niece Nephew | GG GrandChild

GGG GrandChild

ForenSeq Kintelligence + GEDmatch PRO call 100% of these relationships

First Degree

Second Degree

Third Degree

Fourth Degree

Fifth Degree

Extended Relationship

Extended kinships require additional investigative effort to confirm

Simplified DNA Painter structure

Kintelligence | Explore and establish genetic connections

ForenSeq MainstAY Kit

Autosomal & Y STRs

ForenSeq DNA Signature Kit

STRs & SNPs

For short-range and complex STR-based kinship analysis

A choice of STR kits to meet specific application requirements

Compatible with traditional reference databases

ForenSeq Kintelligence Kit

X SNPs (106)

Y SNPs (85)

Kinship SNPs (9867)

Identity SNPs* (94)

Phenotypic SNPs* (22)

Ancestry SNPs* (56**)

End-to-end workflow designed specifically for long range kinship analysis including Forensic Genetic Genealogy

Works in harmony with the GEDmatch database to reach 2nd cousins

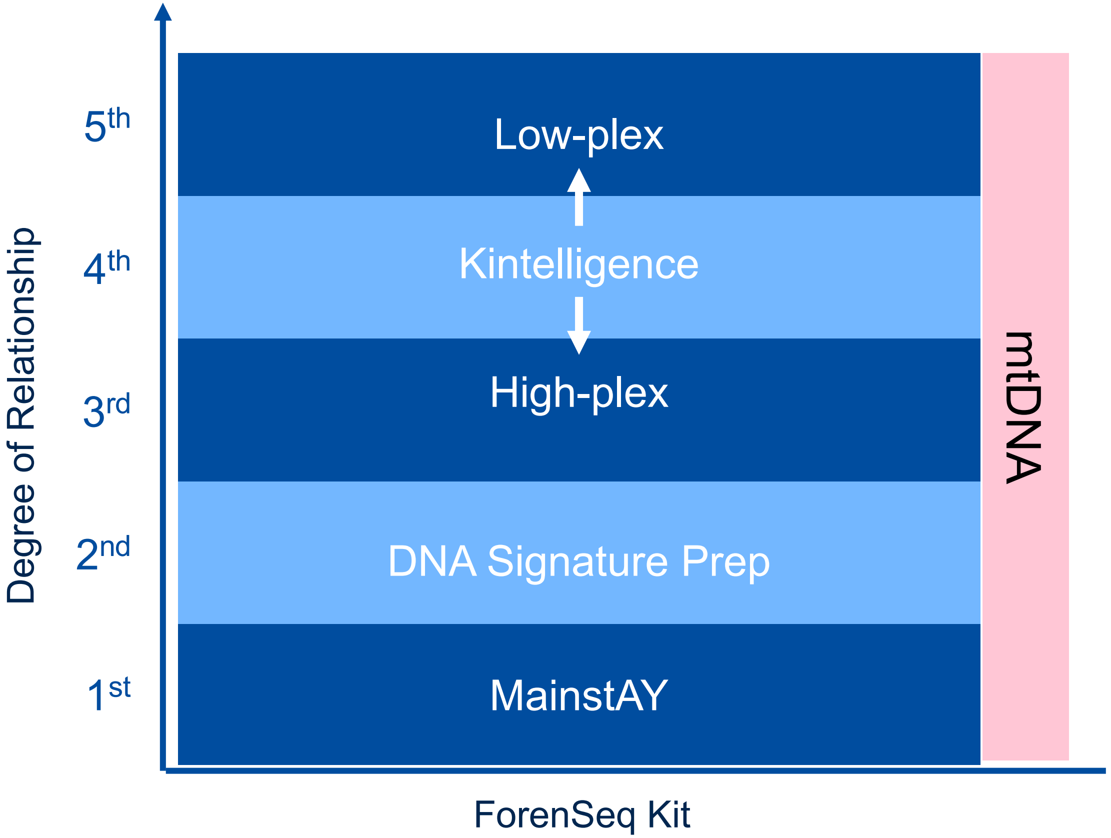
Minimizes data privacy concerns by excluding medically relevant targets

Optimized for performance on low input and degraded samples

UAS: Kintelligence Analysis Module

GEDmatch PRO

Choosing the Right Kit



Key Considerations

Is nuclear DNA available?

What relationship do you need to reach?

How many samples do you need to run?

How many kits and data types do you want to manage?

What in-house data storage and kinship analysis capability do you already have?

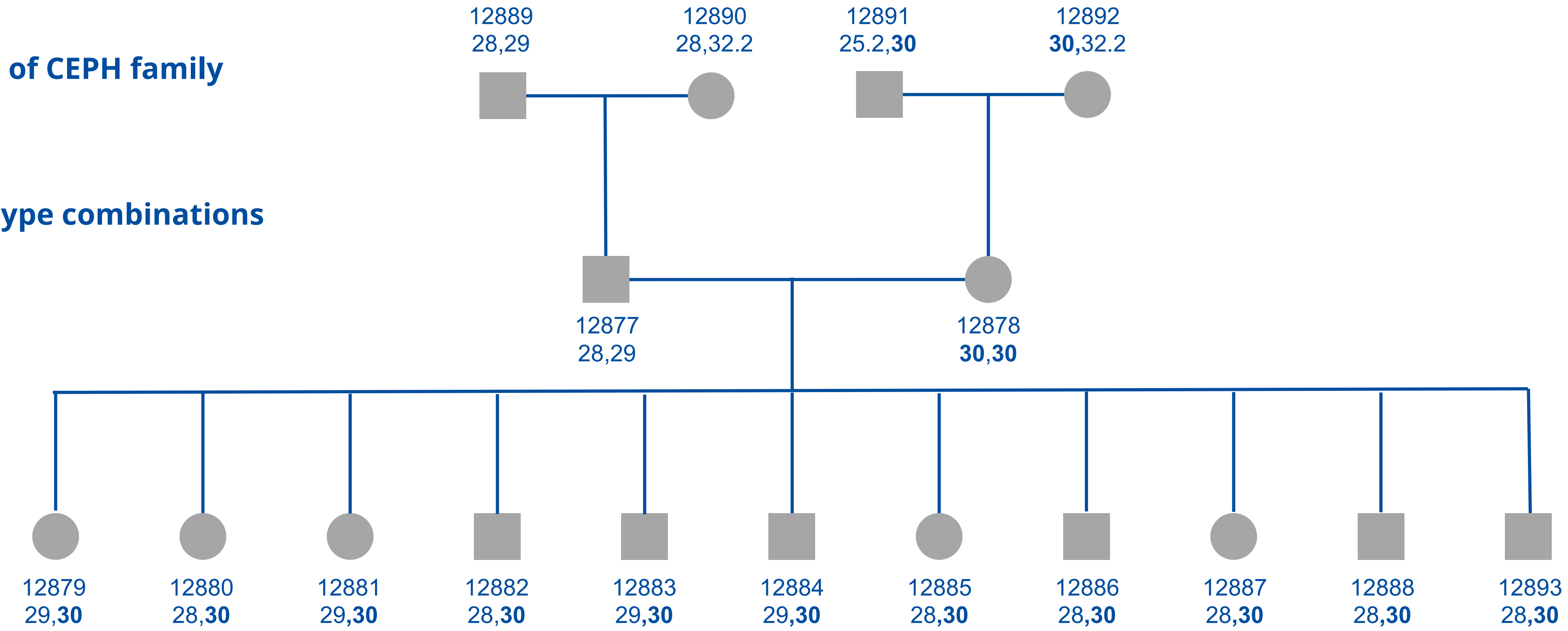
Example Data | Pedigree Analysis – CE vs MainstAY

CE

Length-based analysis of CEPH family 1463

Locus D21S11

4/11 unique sex/genotype combinations in the 3rd generation



12889
28,29

12890
28,32.2

12891
25.2,**30**

12892
30,32.2

12877
28,29

12878
30,30

12879
29,**30**

12880
28,**30**

12881
29,**30**

12882
28,**30**

12883
29,**30**

12884
29,**30**

12885
28,**30**

12886
28,**30**

12887
28,**30**

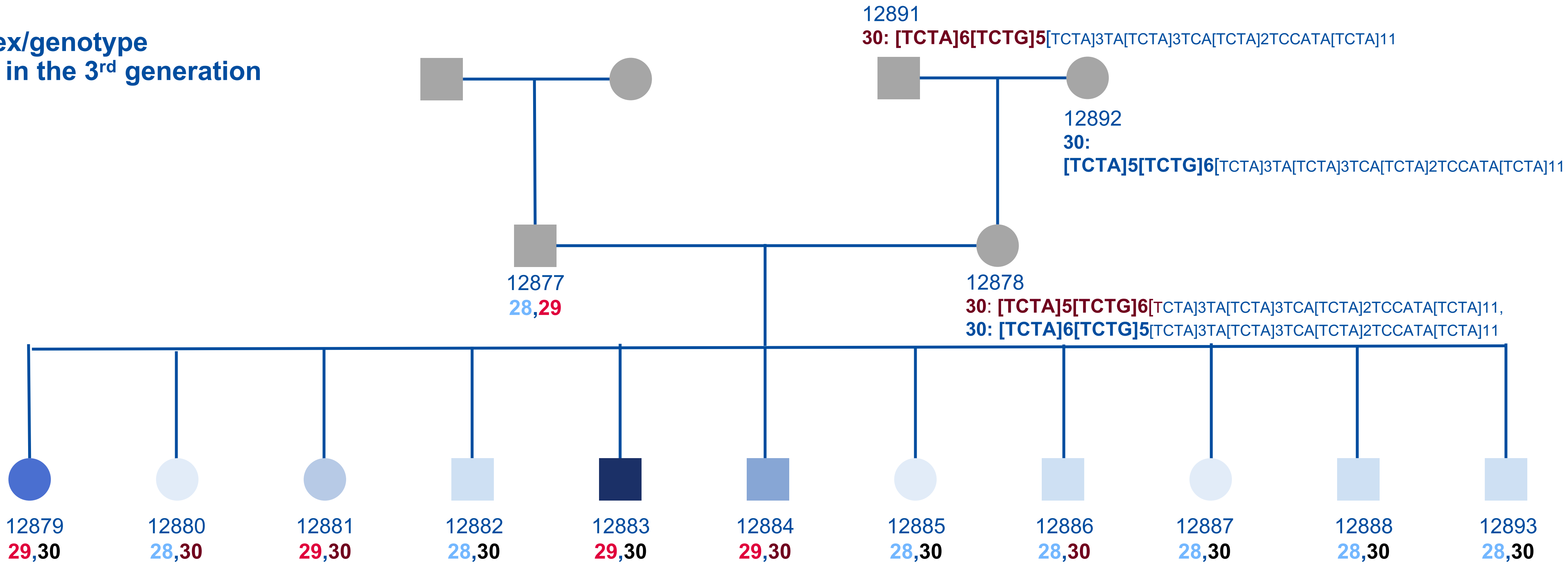
12888
28,**30**

12893
28,**30**

Example Data | Pedigree Analysis – CE vs MainstAY

NGS

8/11 unique sex/genotype combinations in the 3rd generation



12891
30: [TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11

12892
30:
[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11

12877
28,29

12878
30: **[TCTA]5[TCTG]6**[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11,
30: [TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11

12879
29,30

12880
28,30

12881
29,30

12882
28,30

12883
29,30

12884
29,30

12885
28,30

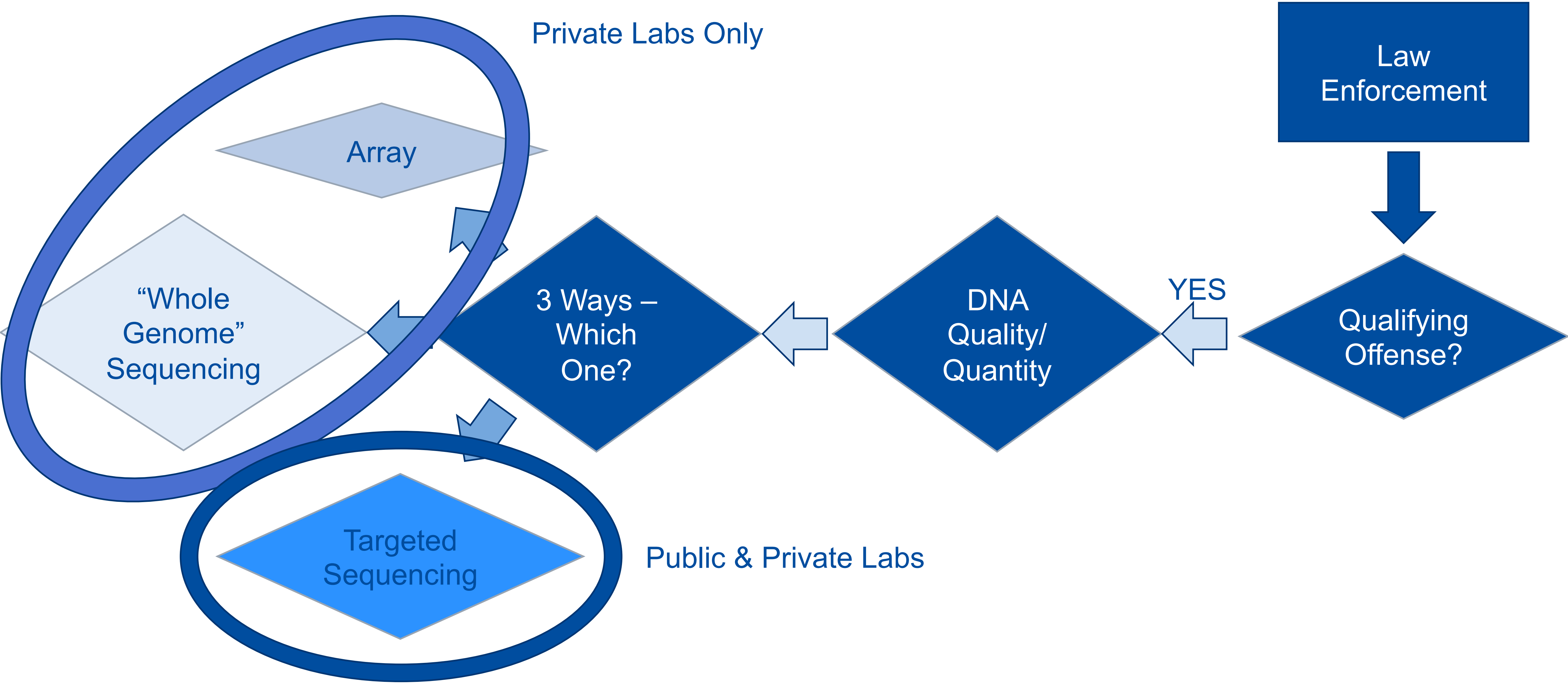
12886
28,30

12887
28,30

12888
28,30

12893
28,30

To FIGG or not to FIGG? (And how…?)



Which Method, When & Why?

	Arrays	Whole Genome Sequencing	Targeted Sequencing Kintelligence
PROs	• Most of the public data in genetic genealogy databases is array data • Fast • Least expensive	• Generates the most information • Potential to reach more distant relatives	• Made for forensic samples • Targets only the DNA required to support most identifications • Easy and cost effective
CONs	• Need lots of DNA • Doesn't perform well on DNA that is degraded or contaminated with microbial material	• Most expensive • Reveals sensitive info • Difficult to perform • Poor quality samples cause 'holes' in the data	• May not reach the most distant relatives for samples with very little DNA available
When to use	For the genetic tests that populate public databases	For more data when Kintelligence isn't enough	As the primary test for all forensic unidentified & missing persons samples

Comparison of Array-Based and Sequencing-Based Methods

Study Design

DNA degradation series on a blood sample

DI range:* 1 to 460

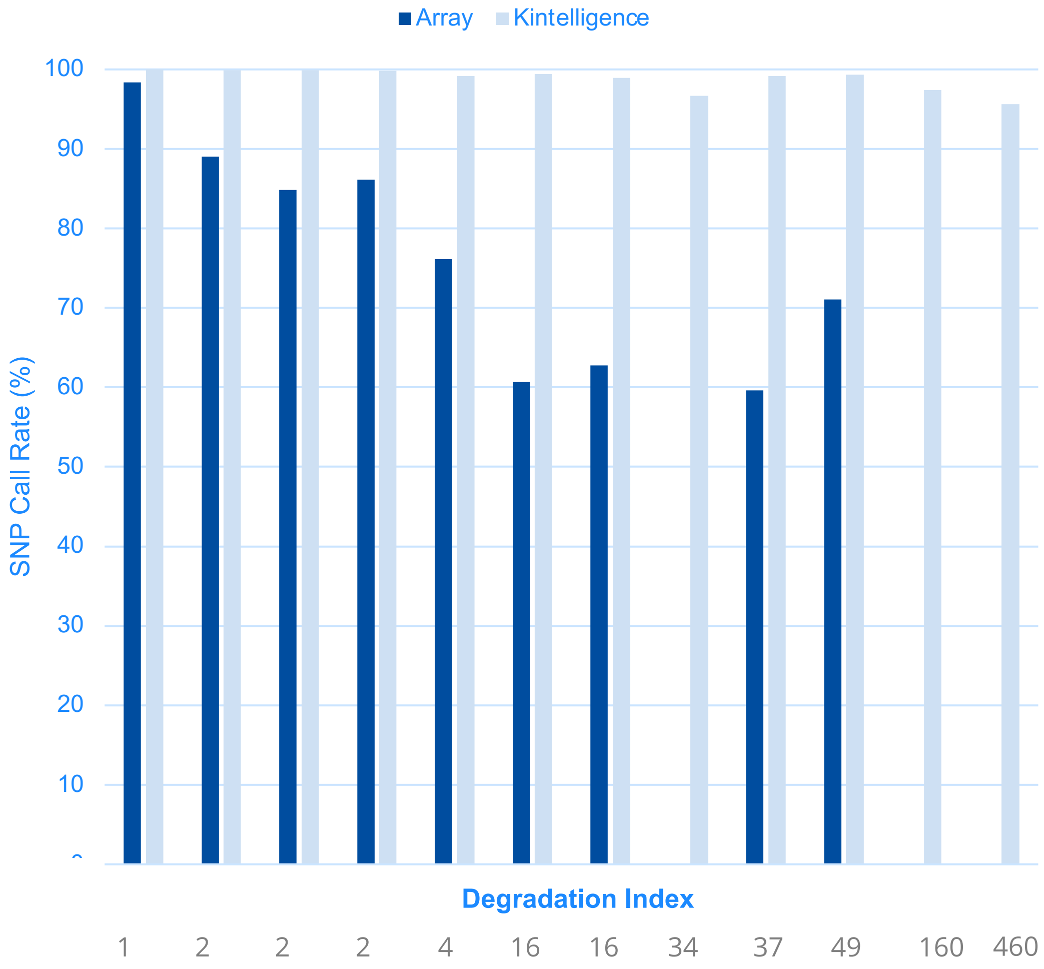
1 ng DNA input DNA for Sequencing

0.7 to 3.5ng input DNA for Microarrays (4ul input volume)

Results

Array - steep decline in SNP call rates at relatively low levels of degradation. Call rates below 75% typically lead to uninformative genealogy matches in operational settings

Kintelligence – 98.8% average call rate with most degraded sample (DI: 460) giving a call rate of 95.6



* DI: InnoQuant® Human DNA Quantification & Degradation Assessment Kit

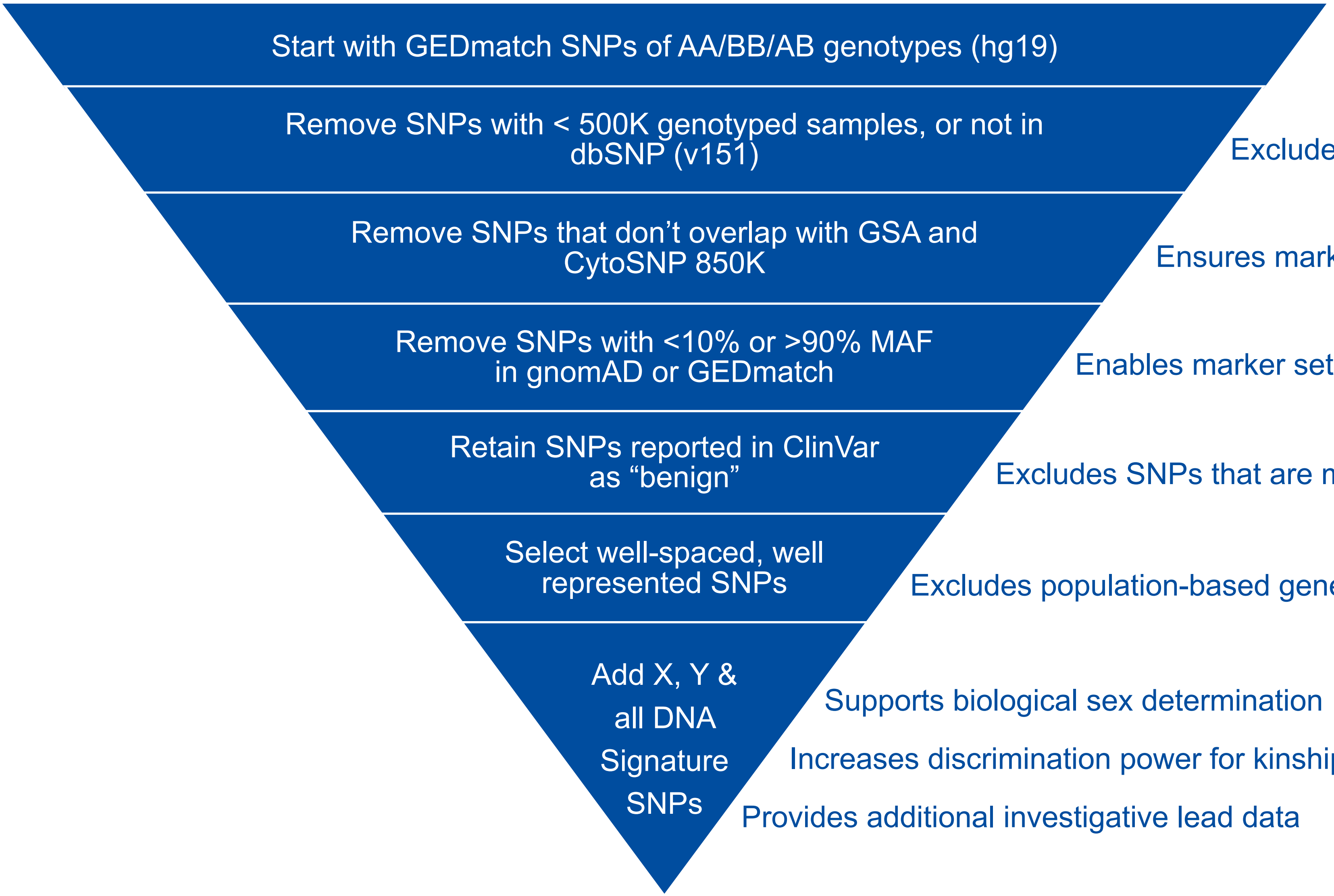
Why a 10K panel?

- **How do we improve performance for degraded/low input samples?**
  - Targeted panel with small amplicon insert sites means that the sensitivity per SNP will be significantly higher than WGS and microarray

- **How can standard Forensic Labs to process their own samples?**
  - Keep number of SNPs targeted small, thus validated MiSeq FGX system can be used
  - Costs lower than WGS

- **How can data be compared to existing microarray/WGS databases?**
  - Build a new algorithm based on existing methods for calculating kinship coefficients

Kintelligence | Forensic SNP Selection Methodology



Start with GEDmatch SNPs of AA/BB/AB genotypes (hg19)

Remove SNPs with < 500K genotyped samples, or not in dbSNP (v151)

Excludes less informative and less discriminatory SNPs

Remove SNPs that don’t overlap with GSA and CytoSNP 850K

Ensures marker overlap with on-market methods

Remove SNPs with <10% or >90% MAF in gnomAD or GEDmatch

Enables marker set to work robustly across global populations

Retain SNPs reported in ClinVar as “benign”

Excludes SNPs that are medically relevant

Select well-spaced, well represented SNPs

Excludes population-based genetic linkages

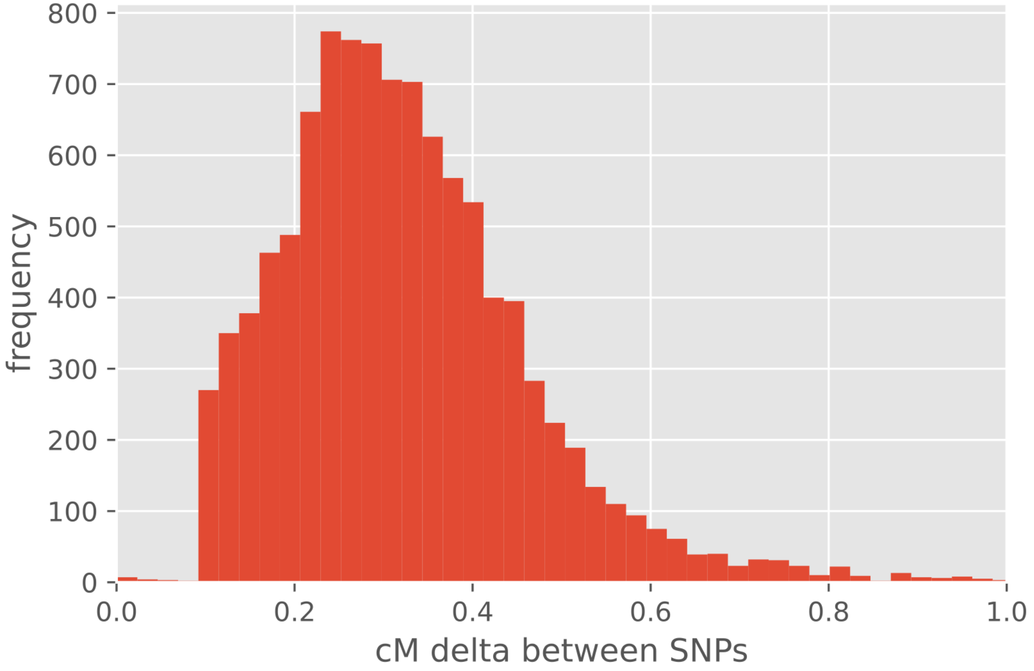
Add X, Y & all DNA Signature SNPs

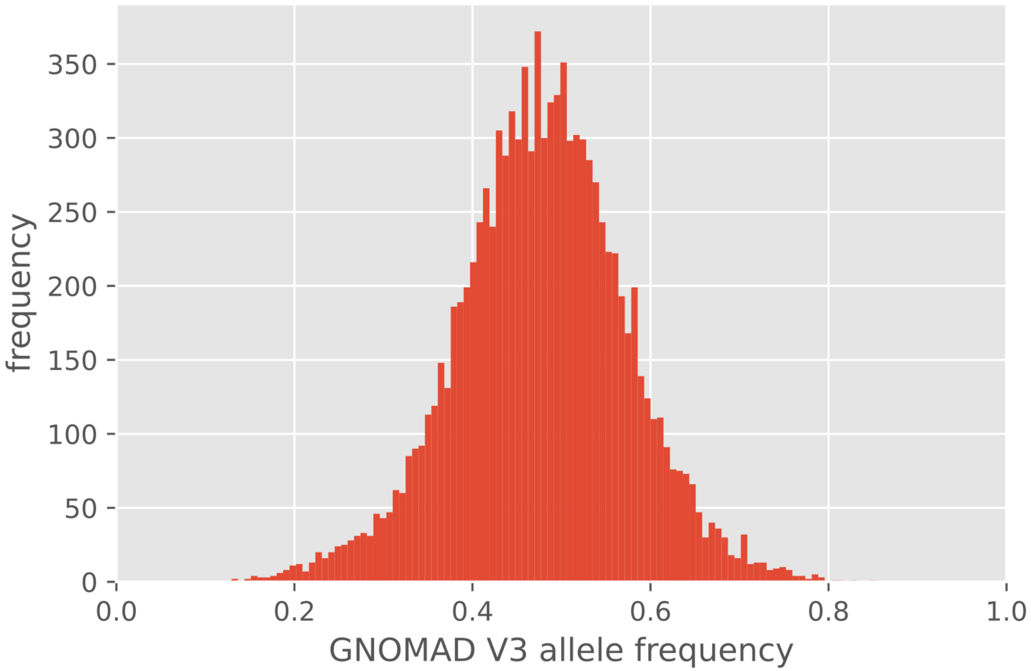
Supports biological sex determination

Increases discrimination power for kinship

Provides additional investigative lead data

Kintelligence Panel Design





- Able to amplify large set of SNPs with extremely degraded samples
- Maximize panel value by using SNPs with minimal linkage and high variability in the populations

Segment vs. Non-segment based Comparisons

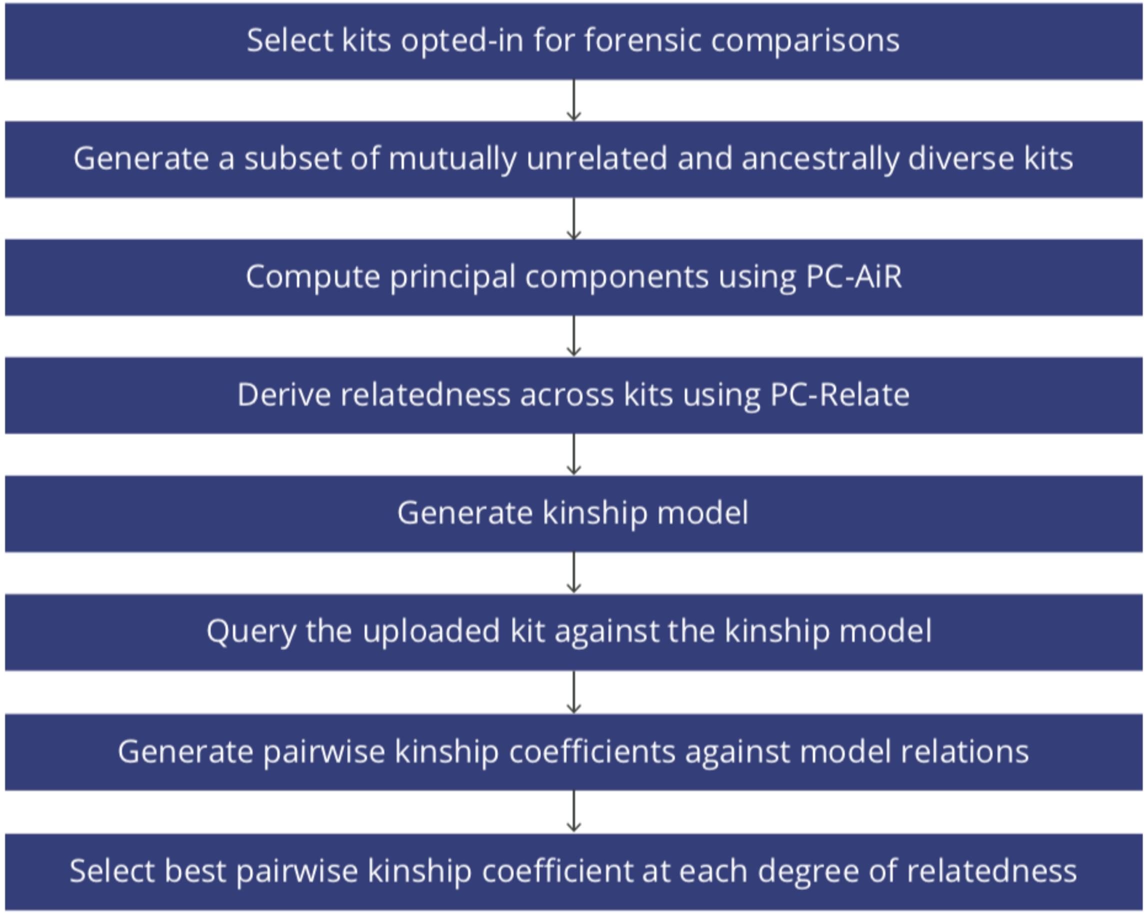
Segment (Traditional)

- Relies on DNA between SNP locations on a chromosome
- Uses data from WGS (Billions of SNPs) and Microarray (650,000 - 2,500,000 SNPs)
- To understand most likely relationships, evaluates:
  - Number of shared centimorgans – Genetic distance; size of matching DNA segments in autosomal DNA tests
  - Average segment length and longest segment across matches

Non-segment (Kintelligence)

- Does NOT rely on overlap between long, contiguous segments to compare kits and generate match
- One-to-Many Kinship tool

Verogen Method I Kinship estimation



Select kits opted-in for forensic comparisons

Generate a subset of mutually unrelated and ancestrally diverse kits

Compute principal components using PC-AiR

Derive relatedness across kits using PC-Relate

Generate kinship model

Query the uploaded kit against the kinship model

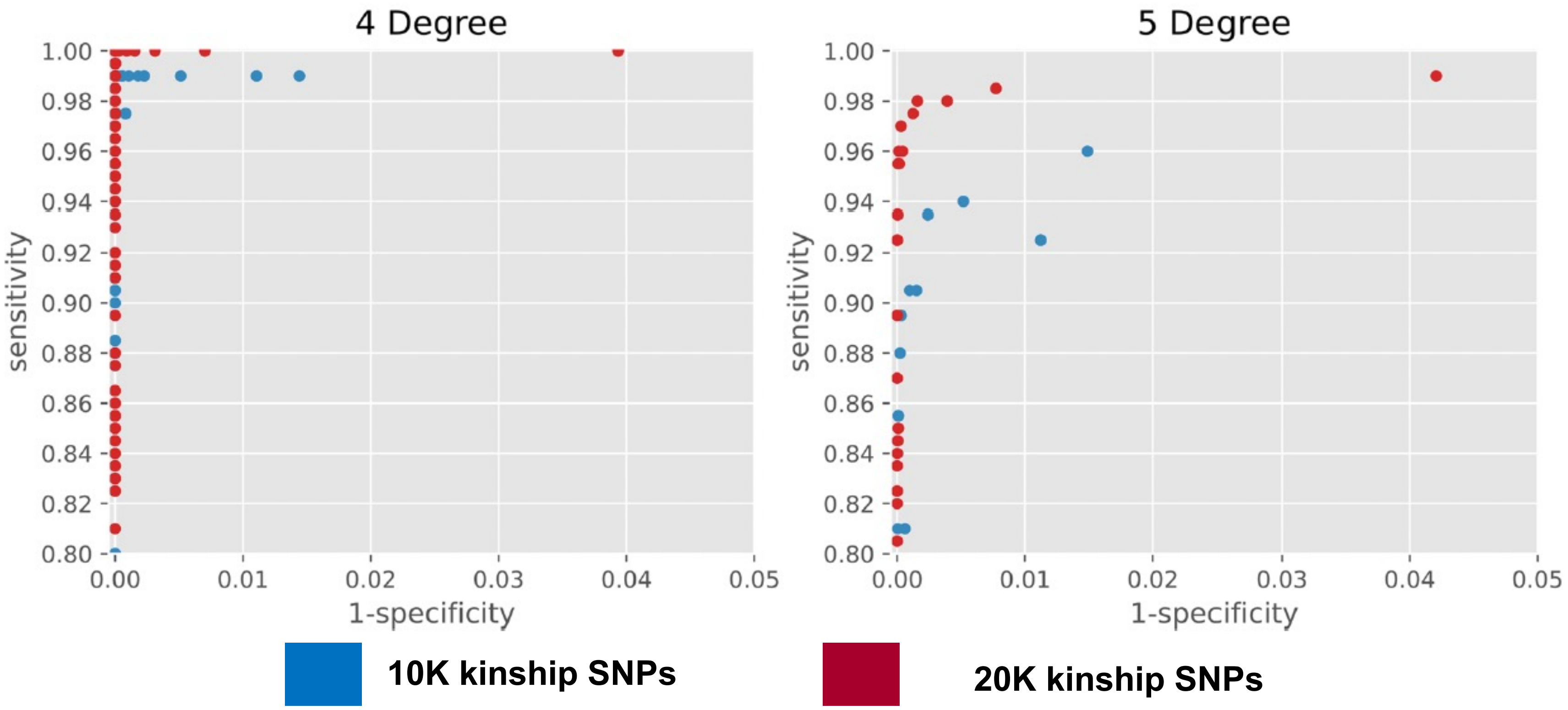
Generate pairwise kinship coefficients against model relations

Select best pairwise kinship coefficient at each degree of relatedness

Figure 2: The Verogen method of kinship estimation yields a simple measure of relatedness.

- **PC-Air**
  - Accounts for relatedness in population to provide ancestry estimations
  - Identifies mutually unrelated kits that are maximally ancestrally diverse
- **PC-Relate**
  - Estimate measures of recent genetic relatedness in samples with an unknown or unspecified population structure without reference population allele frequencies, even when endogamy or consanguinity are present.
  - Identifies ancestry-representative PCs that adjust for family structure and generate relatedness estimates as kinship coefficients in the presence of population structure, admixture, and departures from the Hardy-Weinberg equilibrium.
- **Windowed kinship**
  - Calculates kinship Coefficient across regions of the genome in order to identify related segments

Why 10k SNPs?

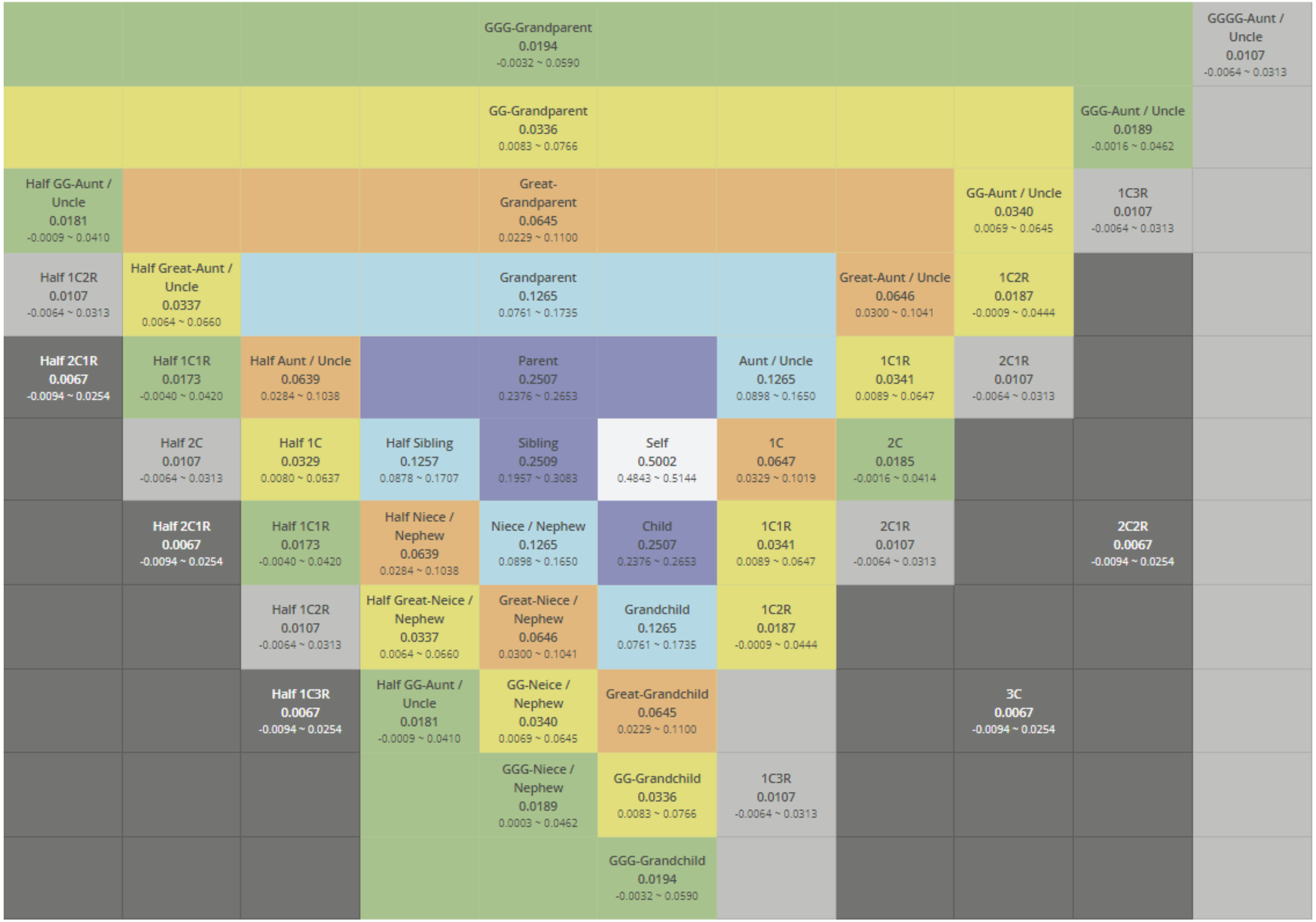


Pedsim simulated relationships from 1000 genomes using windowed kinship

- More SNPs means more read coverage to call them reliably
- Maximize results while minimizing chances of drop out with reasonable plexity on the MiSeq FGx

Kinship Performance

- Assume GEDMatch segment matching is the “gold standard”
- Create “Kintelligence” kits using subset of SNPs from GEDMatch
- Create random dropouts in order to simulate performance with missing data



				GGG-Grandparent 0.0194 -0.0032 ~ 0.0590						GGGG-Aunt / Uncle 0.0107 -0.0064 ~ 0.0313
				GG-Grandparent 0.0336 0.0083 ~ 0.0766					GGG-Aunt / Uncle 0.0189 -0.0016 ~ 0.0462	
Half GG-Aunt / Uncle 0.0181 -0.0009 ~ 0.0410				Great-Grandparent 0.0645 0.0229 ~ 0.1100				GG-Aunt / Uncle 0.0340 0.0069 ~ 0.0645	1C3R 0.0107 -0.0064 ~ 0.0313	
Half 1C2R 0.0107 -0.0064 ~ 0.0313	Half Great-Aunt / Uncle 0.0337 0.0064 ~ 0.0660			Grandparent 0.1265 0.0761 ~ 0.1735			Great-Aunt / Uncle 0.0646 0.0300 ~ 0.1041	1C2R 0.0187 -0.0009 ~ 0.0444		
Half 2C1R **0.0067** **-0.0094 ~ 0.0254**	Half 1C1R 0.0173 -0.0040 ~ 0.0420	Half Aunt / Uncle 0.0639 0.0284 ~ 0.1038		Parent 0.2507 0.2376 ~ 0.2653		Aunt / Uncle 0.1265 0.0898 ~ 0.1650	1C1R 0.0341 0.0089 ~ 0.0647	2C1R 0.0107 -0.0064 ~ 0.0313		
	Half 2C 0.0107 -0.0064 ~ 0.0313	Half 1C 0.0329 0.0080 ~ 0.0637	Half Sibling 0.1257 0.0878 ~ 0.1707	Sibling 0.2509 0.1957 ~ 0.3083	Self 0.5002 0.4843 ~ 0.5144	1C 0.0647 0.0329 ~ 0.1019	2C 0.0185 -0.0016 ~ 0.0414			
	Half 2C1R **0.0067** **-0.0094 ~ 0.0254**	Half 1C1R 0.0173 -0.0040 ~ 0.0420	Half Niece / Nephew 0.0639 0.0284 ~ 0.1038	Niece / Nephew 0.1265 0.0898 ~ 0.1650	Child 0.2507 0.2376 ~ 0.2653	1C1R 0.0341 0.0089 ~ 0.0647	2C1R 0.0107 -0.0064 ~ 0.0313		**2C2R** **0.0067** **-0.0094 ~ 0.0254**	
		Half 1C2R 0.0107 -0.0064 ~ 0.0313	Half Great-Neice / Nephew 0.0337 0.0064 ~ 0.0660	Great-Niece / Nephew 0.0646 0.0300 ~ 0.1041	Grandchild 0.1265 0.0761 ~ 0.1735	1C2R 0.0187 -0.0009 ~ 0.0444				
		Half 1C3R **0.0067** **-0.0094 ~ 0.0254**	Half GG-Aunt / Uncle 0.0181 -0.0009 ~ 0.0410	GG-Neice / Nephew 0.0340 0.0069 ~ 0.0645	Great-Grandchild 0.0645 0.0229 ~ 0.1100			**3C** **0.0067** **-0.0094 ~ 0.0254**		
				GGG-Niece / Nephew 0.0189 0.0003 ~ 0.0462	GG-Grandchild 0.0336 0.0083 ~ 0.0766	1C3R 0.0107 -0.0064 ~ 0.0313				
					GGG-Grandchild 0.0194 -0.0032 ~ 0.0590					

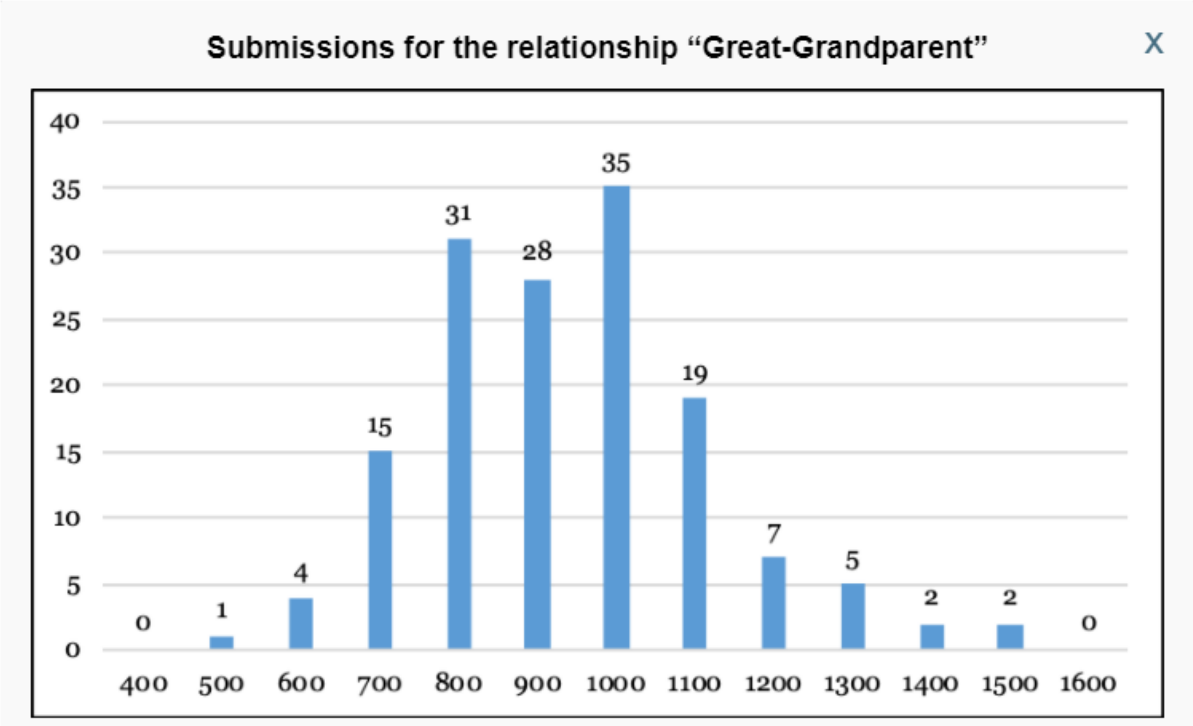
Each value above represents a **Kinship coefficient**; it's **mean** then **min-max** range.

● 1st degree relation ● 2nd degree relation ● 3rd degree relation ● 4th degree relation ● 5th degree relation ● 6th degree relation ● 7th degree relation

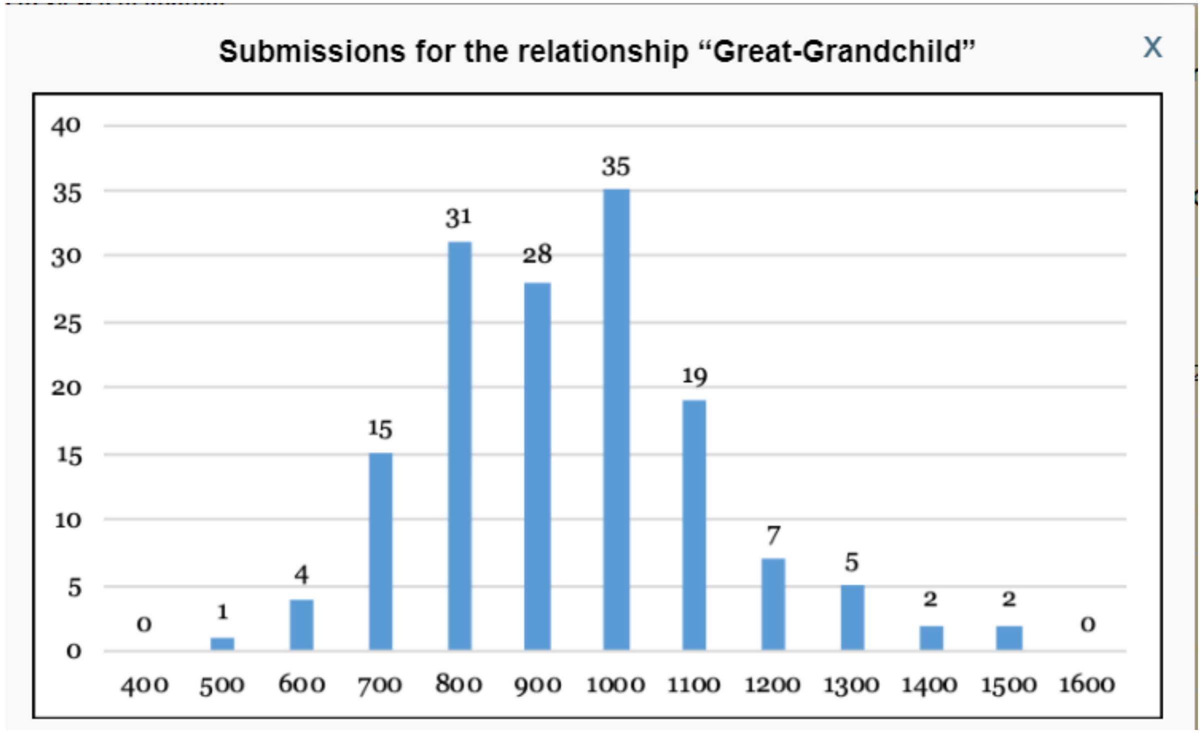
Close relatives

Distant relatives

Shared cM ranges, 3rd degree example



Submissions for the relationship “Great-Grandparent”



Submissions for the relationship “Great-Grandchild”

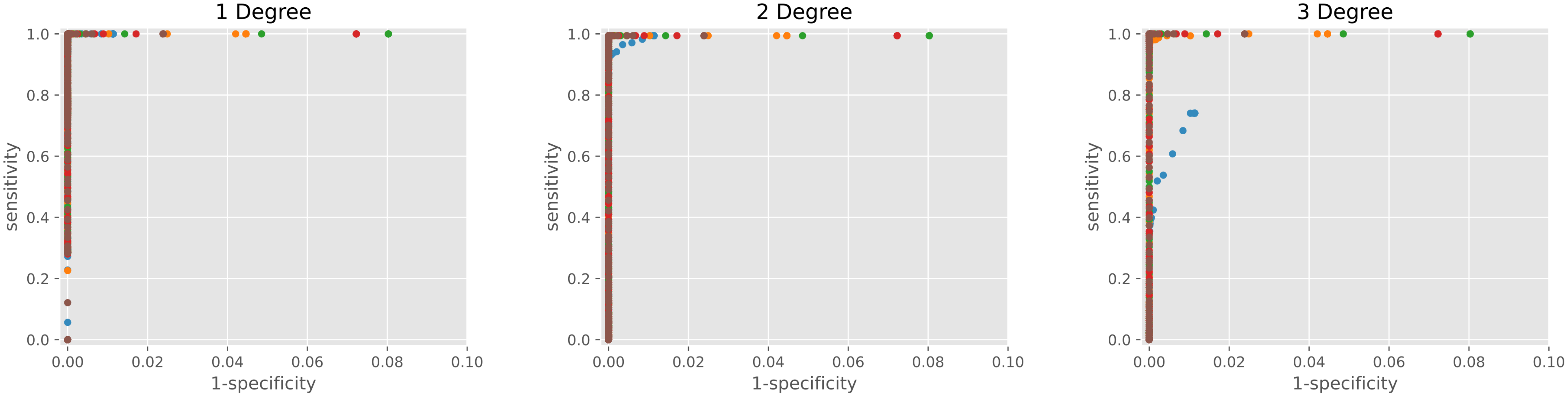
Large range of shared cM even for same degree.

Close relationships will be 3500 – 700 shared cM

Distant relationship will be 700 – 100 shared cM

https://dnapainter.com/tools/sharedcmv4

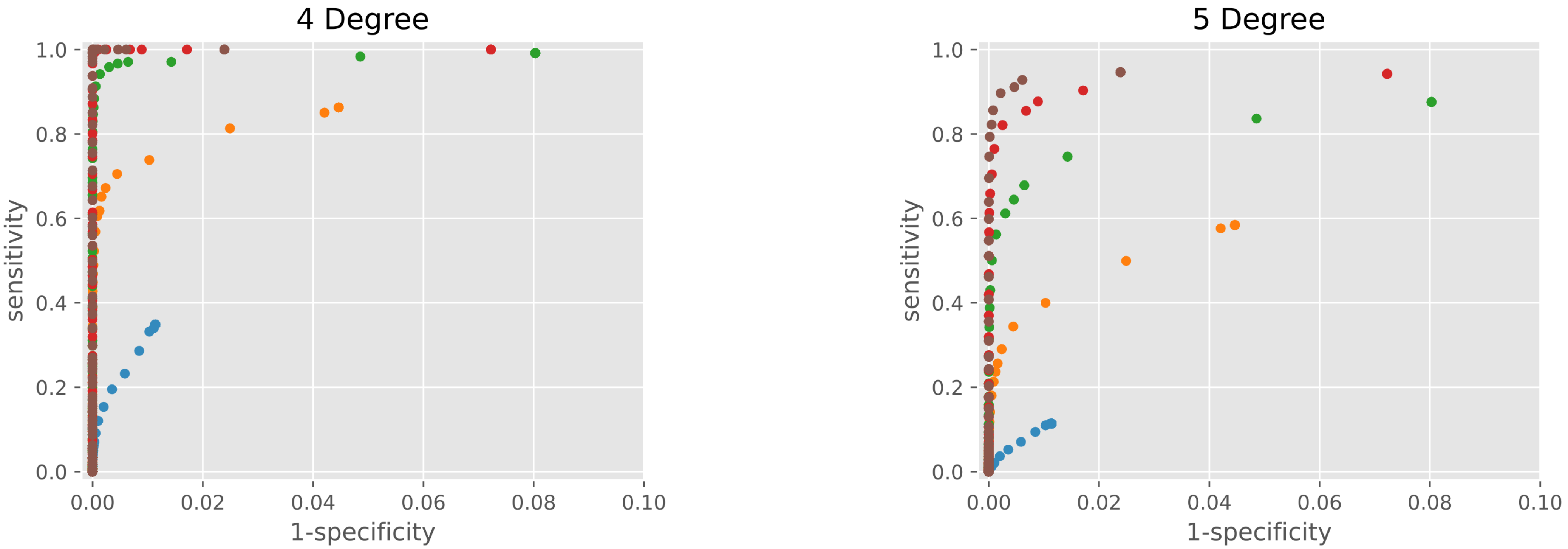
Kintelligence Algorithm Performance – Close Relatives



Number of SNPs used for Kinship

🟦 2000 🟧 4000 🟩 6000 🟥 8000 🟫 10000

Kintelligence Algorithm Performance – Distant Relatives



Number of SNPs used for Kinship

🟦 2000 🟨 4000 🟩 6000 🟥 8000 🟫 10000

Close relationship summation

Near perfect sensitivity and specificity with >=2000 SNPs for 1st and 2nd degree

Near perfect sensitivity and specificity with >=4000 SNPs for 3rd degree

Distant relationship summation

Near perfect sensitivity and specificity with >= 8000 SNPs for 4th degree

Can’t achieve perfect sensitivity for 5th degree

8000 SNPs = 94.2% max sensitivity

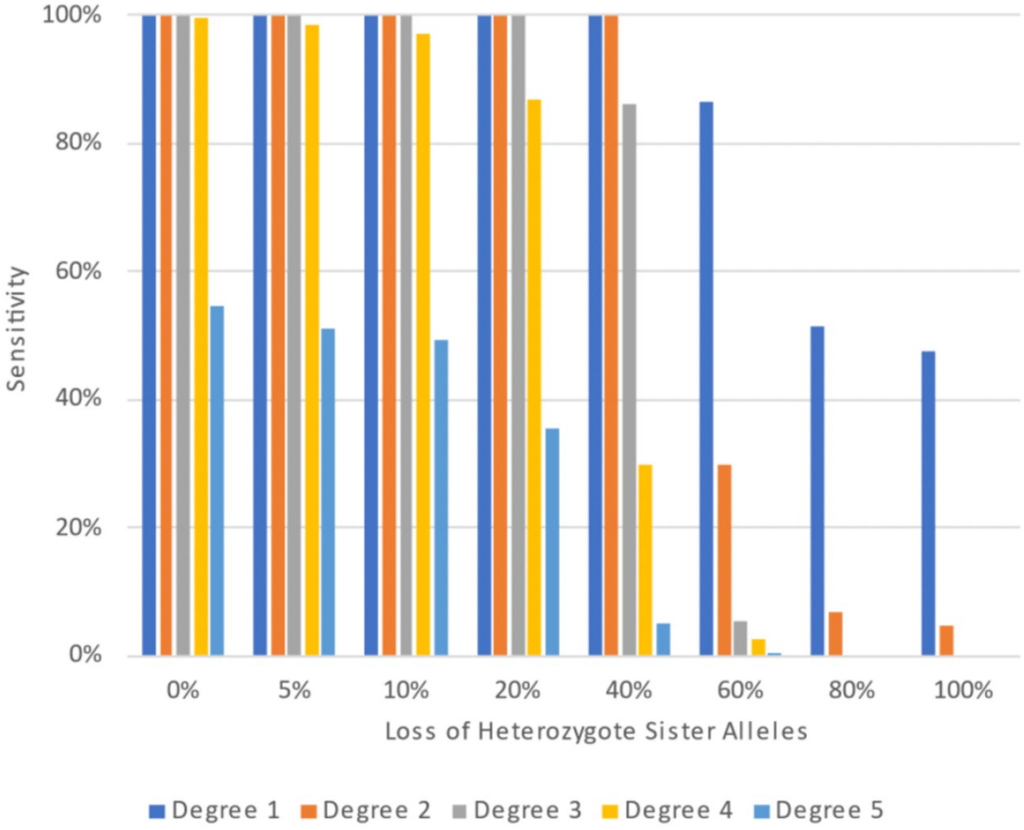
10000 SNPs = 94.6% max sensitivity

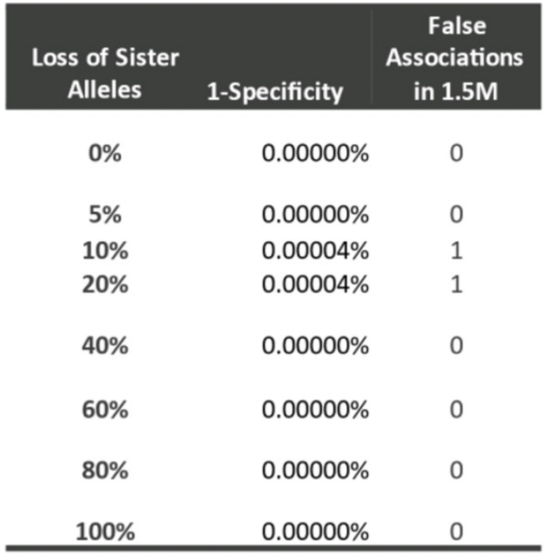
Considerations for 5th degree results:

We’re using GEDMatch as our gold standard and considering a 100 cM hit as real without any manual curating. There may be “5th degree” hits that aren’t actual relationships

.

Loss of Heterozygosity





Loss of Sister Alleles	1-Specificity	False Associations in 1.5M
0%	0.00000%	0
5%	0.00000%	0
10%	0.00004%	1
20%	0.00004%	1
40%	0.00000%	0
60%	0.00000%	0
80%	0.00000%	0
100%	0.00000%	0

- Create “loss of sister alleles” by switching het calls to hom calls
  - Thus, at 5%, 2.5% of the het calls go to hom alt and 2.5% go to hom ref
- Using GEDmatch Pro high confidence thresholds

Summary

- Kintelligence targets only the DNA required to support most identifications
  - Made for forensic samples
  - Can be performed in-house

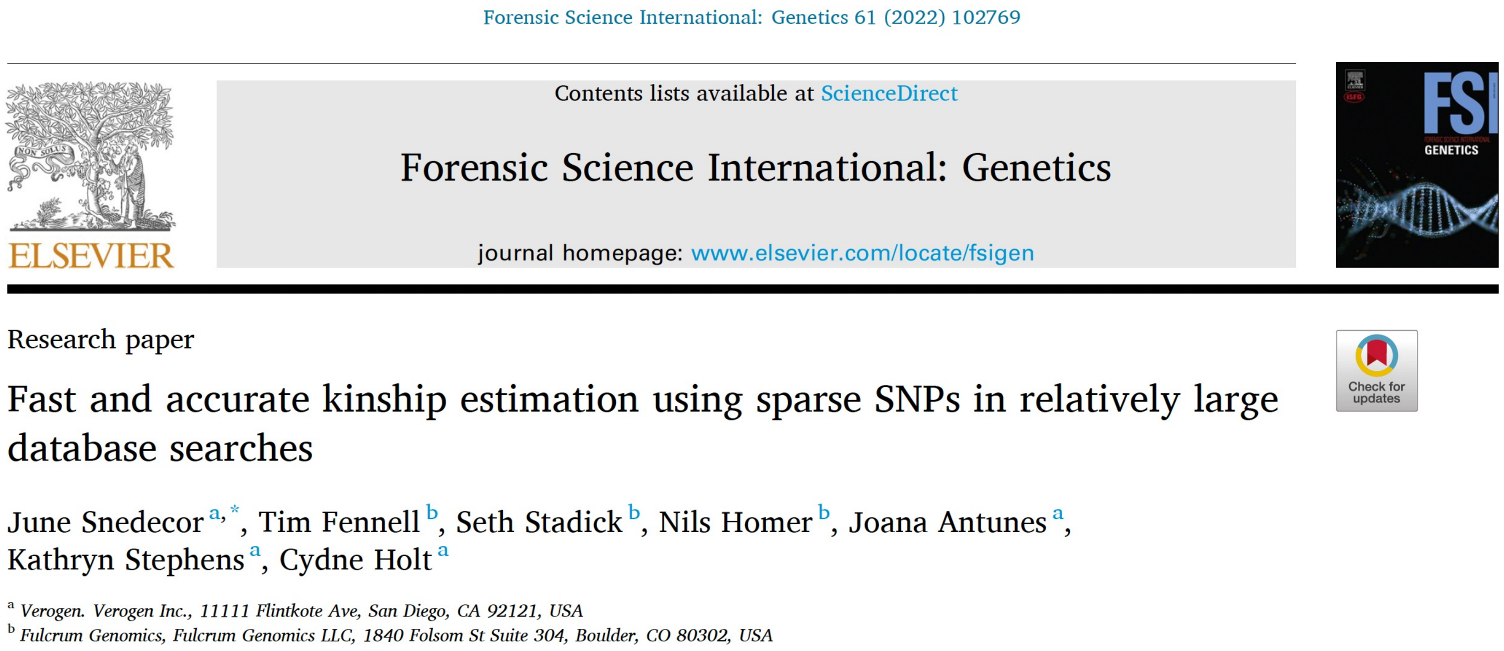
- The kinship algorithm applied to data generated using 10K SNP multiplex supports near perfect detection of relations extending to 3rd degree with a high degree a specificity even with reduced locus call rates and sister allele dropout.

Thank You

melissa.kotkin@qiagen.com

Special thanks to:

- **June Snedecor**
- **Tim Fennell**
- **Seth Stadick**
- **Nils Homer**
- **Joana Antunes**
- **Kathryn Stephens**
- **Cydne Holt**



Forensic Science International: Genetics 61 (2022) 102769

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Fast and accurate kinship estimation using sparse SNPs in relatively large database searches

June Snedecor[a,*], Tim Fennell[b], Seth Stadick[b], Nils Homer[b], Joana Antunes[a], Kathryn Stephens[a], Cydne Holt[a]

[a] Verogen. Verogen Inc., 11111 Flintkote Ave, San Diego, CA 92121, USA
[b] Fulcrum Genomics, Fulcrum Genomics LLC, 1840 Folsom St Suite 304, Boulder, CO 80302, USA