Closing Cases with a Single SNP Array: Integrated Genetic Genealogy, DNA Phenotyping, and Kinship Analyses

PARABON® NANOLABS

Ellen Greytak and CeCe Moore
Parabon NanoLabs, Inc

Introduction

Genetic Genealogy (GG) is revolutionizing forensic investigations by generating leads as to the possible identity of unknown victims and suspects in violent crimes. Parabon's GG process is as follows:

1. Generate genome-wide SNP genotypes using microarray genotyping
2. Perform basic phenotyping to determine ancestry and pigmentation
3. Upload the SNP data to GEDmatch and search their public database of ~1 million individuals to determine how much DNA (in cM) each individual shares with the unknown subject; use the amount of shared DNA to determine the approximate degree of relatedness
4. Identify the GEDmatch matches and construct their family trees back to the set of possible common ancestors using online genealogy databases, newspaper archives, obituaries, and other public records
5. Employ descendancy research to enumerate the (potentially thousands of) possible identities of the unknown subject
6. Narrow down the possible identities using triangulation among matches, known sex, possible age, possible location, predicted ancestry, predicted phenotypes, and/or targeted kinship testing on family members

More than a dozen cases have been solved thus far using this new technology, and more than 70 cases are being analyzed.

Privacy

Much has been written about the impact of genetic genealogy on genetic privacy, but a few critical points temper many of these concerns[1]:

- Private databases run by direct-to-consumer (DTC) companies such as Ancestry.com and 23andMe are not searched.
- Only individuals who choose to download their raw data from a DTC company, upload it to GEDmatch, and set it to be publicly searchable can be searched. Users can make their data private or use an alias.
- GEDmatch's Terms of Service explicitly state that law enforcement can and is using the site to identify victims and perpetrators of violent crimes. All users (new and existing) are presented with the full text of this agreement and must agree before using the site.
- GEDmatch users' raw data files (including the unknown individual's) are not accessible to other users. Only the amount and location of shared DNA is disclosed. All data for the unknown is kept private.
- Genetic genealogy results are treated as leads; traditional STR matching is used to confirm identity.

[1] Greytak EM, Kaye DH, Budowle B, Moore C, & Armentrout SL (2018). Privacy and genetic genealogy data. *Science*, **361**(6405), 857.

[2] Erlich Y, Shor T, Carmi S, & Pe I (2018). Re-identification of genomic data using long range familial searches. *BioRxiv* 350231.

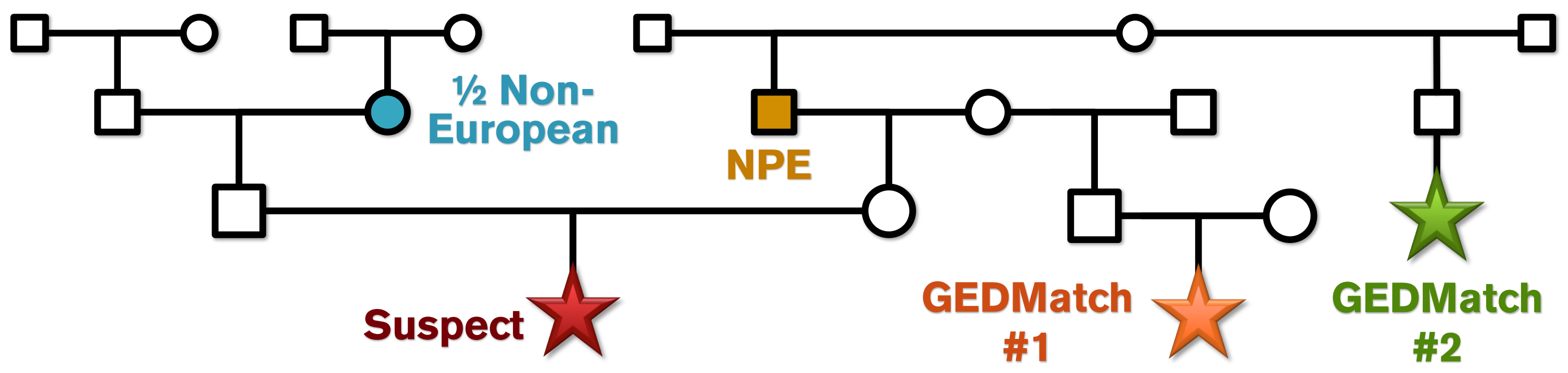
Example Case #1

GEDmatch: Matches at ~400 cM and ~200 cM, no shared DNA

Ancestry: ~90% N European, ~10% non-European (~1/8)

Genealogy: The matches' family trees do not intersect on paper, but match #2's half-uncle lived in the same town as match #1's grandparents when match #1's aunt was conceived, suggesting a non-paternity event (NPE) between these families. That (half-)aunt has a grandson with 1/8 non-European ancestry who is a half-1st cousin to match #1 and a half-1st cousin once-removed to match #2.

Outcome: Abandoned DNA matched to crime scene DNA



Example Case #2

GEDmatch: Matches at ~185 cM and ~80 cM, no shared DNA

Phenotypes: Very fair skin, light brown eyes & hair, no freckles

Genealogy: Built family trees back to match #1's grandparents and match #2's great-great-great grandparents

Descendancy: Found one marriage connecting the family trees, which produced a set of brothers who are 1st cousins twice-removed to match #1 and 4th cousins once-removed to match #2; all but one brother have dark brown eyes and dark brown hair

Outcome: Abandoned DNA collected from the brother with light brown eyes and light brown hair was matched to crime scene DNA

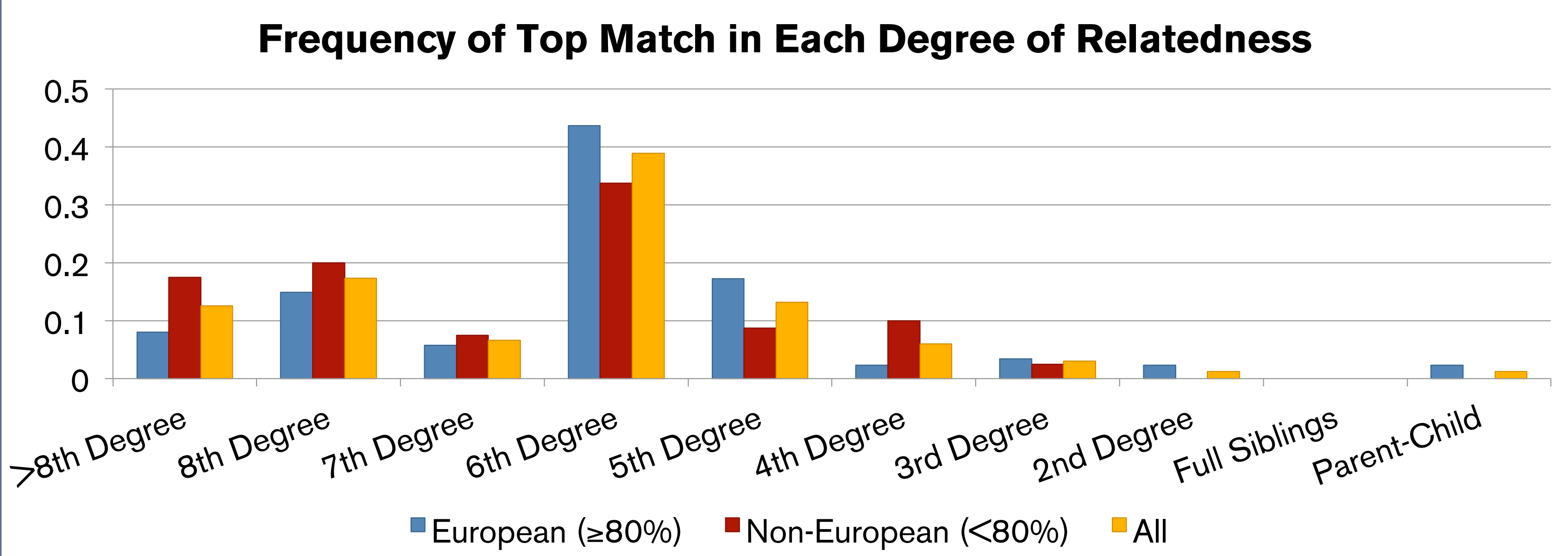
Challenges in Genealogy Research

While it may be theoretically possible to identify any European person from DNA[2], real forensic casework requires significant expertise:

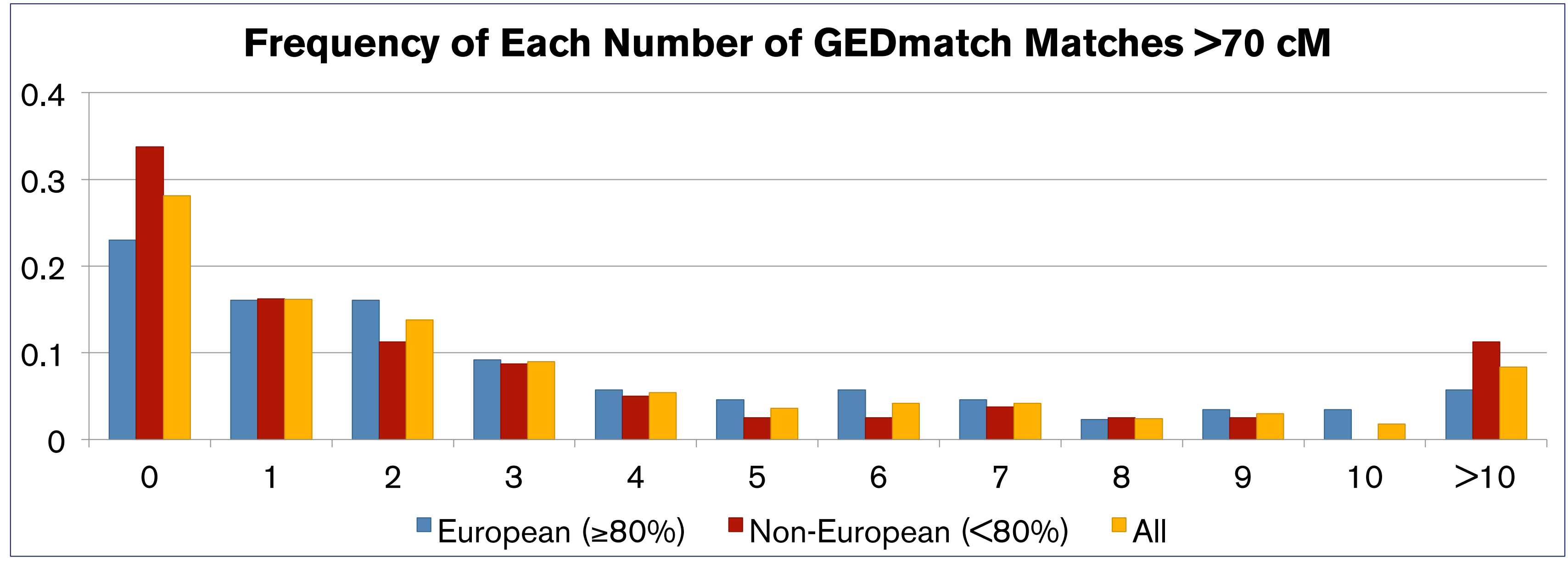
- Shared cM does not perfectly correlate with relationship, especially for forensic samples that may have missing SNPs, so many possible relationships must be considered for each match.
- Simply identifying each match and their family members can be a challenge. Records from other countries cannot be accessed and some states have limited or incomplete information. Many family trees contain misstated paternities or unrecorded adoptions.
- Endogamy (high background relatedness in certain populations) and pedigree collapse (the same families intermarrying over many generations) can vastly inflate the amount of shared DNA, resulting in many matches distantly related to one another in complex ways.
- The age of a perpetrator is often not known, even within 10 years.

Probability of Finding a Match

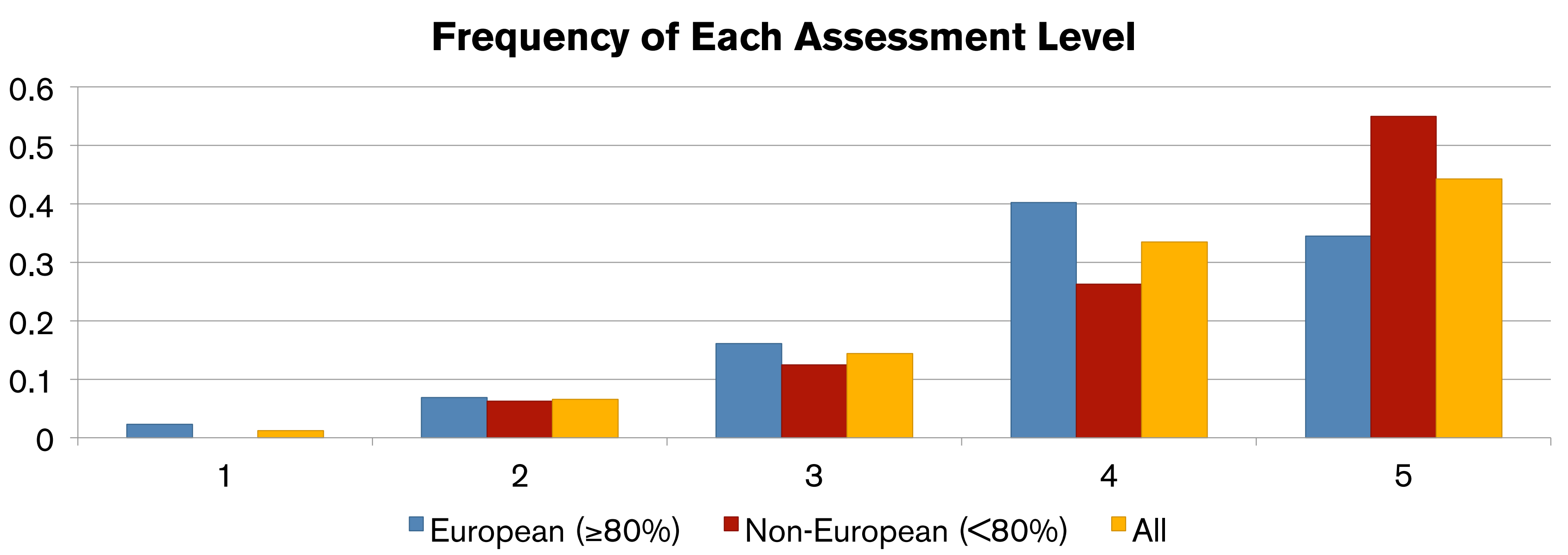
Parabon has evaluated >170 forensic casework samples for GG suitability. First, each sample's top match is assessed and the amount of shared DNA is translated into approximate degree of relatedness. GG databases have a high proportion of participants of European descent, so the graphs below are broken out by European (52% of cases) vs. non-European using a cutoff of 80% European ancestry. European samples tend to have closer matches and a slightly higher probability of success, but non-Europeans often succeed as well.


Frequency of Top Match in Each Degree of Relatedness

Each sample is also evaluated for the number of matches above 70 cM (~3rd cousins). If an individual has >10 matches, they are most likely from an endogamous population.


Frequency of Each Number of GEDmatch Matches >70 cM

Each case is then given an assessment of its probability of success, with 1-3 indicating a high likelihood of being solved using GG alone, 4 indicating GG is likely to generate actionable information that could lead to a solve, and 5 indicating there are insufficient matches for GG to be fruitful at this time. 5's are monitored regularly for new matches.


Frequency of Each Assessment Level