DNA Phenotyping and Kinship Determination

Ellen McRae Greytak, PhD
Director of Bioinformatics
Parabon NanoLabs, Inc.
Forensic Applications of DNA Phenotyping

› Predict a person’s ancestry and/or appearance (“phenotype”) from his or her DNA

› Generate investigative leads when DNA doesn’t match a database (e.g., CODIS)

› Gain additional information (e.g., pigmentation, detailed ancestry) about unidentified remains

› Main value is in excluding non-matching individuals to help narrow a suspect list
  - Without information on age, weight, lifestyle, etc., phenotyping currently is not targeted toward individual identification
Workflow of a Parabon® Snapshot™ Investigation

1. DNA Evidence Is Collected and Sent to Crime Lab

2. Crime Lab Extracts DNA and Produces STR Profile (a.k.a. “DNA Fingerprint”)

3. STR Profile Checked Against DNA Database(s)

4. Snapshot Composite Ordered

5. Unidentified DNA Is Sent to Service Lab (DNA Extracted If Needed)

6. Genotyping Lab Produces SNP Profile (a.k.a. “DNA Blueprint”)

7. Genotype Data Is Sent to Parabon

8. Parabon Analyzes Genotype Data


10. Investigator Uses Snapshot Report To:
    - Generate Leads
    - Exclude Suspects
    - Identify Remains

NOTE: SNPs Profiles Do Not Contain Sufficient Genetic Information to Produce a SNP Genotype

DNA Evidence

Unidentified Remains

DNA Crime Lab

DNA Service Labs

Parabon NanoLabs

© 2015 Parabon NanoLabs, Inc. All Rights Reserved.

Ready To Learn More? Contact Us For A Free Consultation: http://Parabon-NanoLabs.com/Snapshot
Ancestry Inference

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

- 23% Native American ancestry
- 52% East Asian ancestry
- 17% European ancestry
Conclusion: This individual is half Japanese and half Latino.
Data Mining: Association Testing

- Use data mining of genotype+phenotype (G+P) data to identify those SNPs that have the strongest predictive power for phenotype

<table>
<thead>
<tr>
<th>Subject</th>
<th>Eye Color</th>
<th>SNP1 Genotype</th>
<th>SNP2 Genotype</th>
<th>SNP3 Genotype</th>
<th>SNP4 Genotype</th>
<th>...</th>
<th>SNP1,000,000 Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Blue</td>
<td>A/G</td>
<td>C/C</td>
<td>A/A</td>
<td>G/G</td>
<td>...</td>
<td>T/T</td>
</tr>
<tr>
<td>2</td>
<td>Green</td>
<td>A/A</td>
<td>T/T</td>
<td>A/G</td>
<td>A/G</td>
<td>...</td>
<td>T/T</td>
</tr>
<tr>
<td>3</td>
<td>Hazel</td>
<td>G/G</td>
<td>C/T</td>
<td>G/G</td>
<td>A/A</td>
<td>...</td>
<td>T/T</td>
</tr>
<tr>
<td>4</td>
<td>Brown</td>
<td>A/A</td>
<td>T/T</td>
<td>G/G</td>
<td>G/G</td>
<td>...</td>
<td>C/T</td>
</tr>
<tr>
<td>3,000</td>
<td>Blue</td>
<td>G/G</td>
<td>T/T</td>
<td>A/A</td>
<td>G/G</td>
<td>...</td>
<td>C/T</td>
</tr>
</tbody>
</table>
Data Mining: 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 SNPs

Ritchie et al. 2001 – AJHG 69:138-147
Looking for high-order interactions (>2 SNPs) 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
- Currently impossible to exhaust this search space

Parabon has developed software that uses a distributed evolutionary search algorithm to explore the massive space of possible interactions

Allows us to discover previously-unknown SNP associations
Predictive Modeling

› Combine discovered SNPs into a machine learning model that can be used for prediction on new samples

› Entire mining and modeling process is wrapped in a cross-validation framework, which allows direct calculation of *out-of-sample* accuracy
Prediction Results – Eye Color

- Predicted Value = 1.638
- Green (61% confidence)
- Green or Blue (79.4% confidence)
- NOT Brown or Black (99.3% confidence)
Prediction Results – Skin Color

- Predicted Value = 1.687

- Very Fair (78% confidence)
- Very Fair or Fair (97.1% confidence)
- NOT Light Olive, Dark Olive, or Dark (97.1% confidence)
Prediction Results – Hair Color

- Predicted Value = 2.560
- Blond (15.2% confidence)
- Blond or Red (99.6% confidence)
- NOT Brown or Black (99.6% confidence)
Compare the Predicted Face Shape for this individual to the Average Predicted Face Shape for subjects with the same sex and ancestry.

- **X:** narrower jaw and chin; wider brow
- **Y:** long face – higher eyes, nose, and jaw; longer chin
- **Z:** more prominent nose, mouth, and upper cheekbones; less prominent chin and cheeks
Apply predicted pigmentation to predicted face shape
**Snapshot Prediction Results**

**Composite Profile**

**DNA #NBC_01**

**SNL Document #13H24KR52. NBC_01**

**Predicted (■) & Excluded (□) Phenotypes**

**Skin Color**

- Very Fair / Fair (97.6% confidence)
- NOT Light Olive / Dark Olive / Dark (97.6% confidence)

**Eye Color**

- Brown / Hazel (97.8% confidence)
- NOT Green / Blue / Black (97.8% confidence)

**Hair Color**

- Brown / Black (84.1% confidence)

**Freckles**

- Few / Some (75.0% confidence)

**Sex:** Female ♀

**Age:** Unknown

(Composite shown at age 25)

**Ancestry:** Northern/Eastern European

---

**Predicted (■) & Excluded (□) Phenotypes**

**Skin Color**

- 27.8 Light Olive / Fair (80.4% confidence)
- NOT Very Fair / Dark (91.0% confidence)

**Eye Color**

- Hazel / Brown (95.5% confidence)
- NOT Green / Blue / Black (95.5% confidence)

**Hair Color**

- Blond / Red (85.8% confidence)
- NOT Black (94.9% confidence)

**Freckles**

- Zero (95.8% confidence)
- NOT Few / Some / Many (95.8% confidence)

**Sex:** Female ♀

**Age:** Unknown

(Composite shown at age 25)

**Ancestry:** European and Latino
Additional Blind Predictions

- Additional blind predictions vs. actual photographs are available for government and law enforcement personnel.
Cases That Have Gone Public

Miami-Dade Police Department

Fort Wayne Police Department

Calcasieou Parish Sheriff’s Office
Developed a new method that uses machine learning to predict relatedness between two genomes

**Accuracy Within One Degree**
- Siblings: 100%
- Parent: 100%
- 2nd-Degree: 100%
- 3rd-Degree: 94%
- 4th-Degree: 77%
- 5th-Degree: 98%
- 6th-Degree: 93%
- Unrelated: 99.5%

**Absolute Kinship Accuracy**
- Grandparent, Uncle, Half-Sibling: 100%
- First cousin, Great-grandparent: 97.5%
- First cousin once-removed: 95%
- Second cousin: 77%
- Second cousin once-removed: 45%

**Diagram**
- Tree structure representing kinship relationships with accuracy percentages for each degree of kinship.