National Institute of Standards and Technology
Technology Administration,
U.S. Department of Commerce

Instituto de Ciencias Forenses Luis Concheiro. Universidade de Santiago de Compostela. Spain.

Email: mafondevila@hotmail.com

# Characterization of U.S. Population Samples using a 34plex Ancestry Informative SNP Multiplex

P-025

Christopher Phillips[1], Manuel Fondevila[1,2], Peter M. Vallone[2], Carla Santos[1], Ana Freire-Aradas[1], John M. Butler[2],
Maria Victoria Lareu[1] and Angel Carracedo[1].

[1]*Instituto de Ciencias Forenses Luis Concheiro. Santiago de Compostela. Spain*
[2]*U.S. National Institute of Standards and Technology, Biochemical Science Division, Gaithersburg, Maryland, USA*

The number, scope, and ease of typing of single nucleotide polymorphism (SNP) markers makes them ideal supplements to existing forensic markers sets. SNP typing panels offer additional benefits such as the ability to type degraded DNA, paternity analysis, and the opportunity to infer externally visible traits. SNP markers also carry the potential to infer the most likely population of origin of an individual using SNPs with highly differentiated allele frequency distributions. One recent example is a 34plex assay developed by C. Phillips et al. using SNaPshot primer extension reactions[1].

A significant amount of population information has already been generated, published, and uploaded to the SPSmart open access SNP browsers[2]. The original work by Phillips et al. analyzing the CEPH population diversity panel indicated a low error rate of ancestry prediction when confining comparisons to the three major population groups of Europe, Africa and East Asia. However, admixed populations commonly found in the U.S. (e.g. Hispanics and African Americans) represent a significant source of error or at least reduced assignment probabilities when making ancestry predictions. Thorough and wide population surveys in areas where admixture is the predominant pattern, is an important part of the process of assessing this potential source of classification error. In order to contribute to the data already available to end-users, we have generated allele frequencies for a set of U.S. African Americans, Caucasian and Hispanics for 34 ancestry informative SNPs. The data accumulated should contribute to improved characterization of admixed U.S. populations and their analysis through SNP genotyping.

## Introduction

### Geographical and population origin prediction through SNPs

The selection of an array of ancestry informative autosomal SNPs to develop genotyping tests for the forensic prediction of geographical and population origin of an unknown DNA sample has been an area of research for several years.

Our group, at the Institute of Forensic Sciences Luis Concheiro, developed a single 34 SNP multiplex for forensic ancestry inference [1]. We also characterised a number of global populations to generate the *SPSmart* open-access population allele frequency browser [2]. Analysis of the population data indicates the predictive power of the 34 SNP test can infer the most likely population origin of an unknown DNA sample comparing Africans, Europeans, East Asians with an error rate close to zero [1] even with partial profiles.

It has been observed in additional studies [3], that the power of the 34plex test is significantly eroded when there is some degree of admixture in the populations analyzed. The greater the amount of admixture, the greater the erosion of predictive power provided by the likelihoods generated from the statistical analysis. While there are approaches that can be taken to adjust for admixture in the populations analyzed [3], some knowledge of the likely population history and structure is required.

The study described here genotyped a sample of individuals self-identified as belonging to one of the four most frequent population groups representative of the majority of U.S. inhabitants: U.S. Caucasians, African-Americans, U.S. Asians and U.S. Hispanics. **The study therefore has a double purpose: (a) to obtain allele frequency estimates for the 34 SNP markers of the ancestry informative assay, for U.S. populations, available for use by the forensic community and (b) to improve our understanding of the complex admixture patterns of U.S. populations.**

## Materials & Methods

**Amplification and genotyping reaction:**

We performed the 34plex SNP analysis following the guidelines published by Phillips et al. [1] with the following modifications:

- In order to compensate for the variation in the sensitivity of ABI Prism instruments we conducted several experiments to determine the optimum cycle number and DNA input for NIST equipment.
- Several SNaPshot extension primer lengths have been modified to more adequately distribute the signal for each component marker of the 34plex genotyped with POP4 and 36 cm capillary arrays.
- We have substituted one of the less informative markers (rs727811) for a new East Asian informative SNP: rs3827760.

### Primer modifications

| SNP | Variation | sequence |
|---|---|---|
| New P04 | +58bp | $[ct]_{33}$CTCATTAGTCCTTGGCTC |
| New P01 probe | +60bp | $[ct]_{34}$CCACTCCACCGCTAAT |
| New P02 probe | +60bp | $t[ct]_{34}$CAGGATCGATTGGTTCC |
| New P25a | -2bp | $[ct]_{28}$GGTTGGATGTTGGGGCT |
| P28 probe | n/a | $[N]_{74}$CGCCACGTTTTCACA |

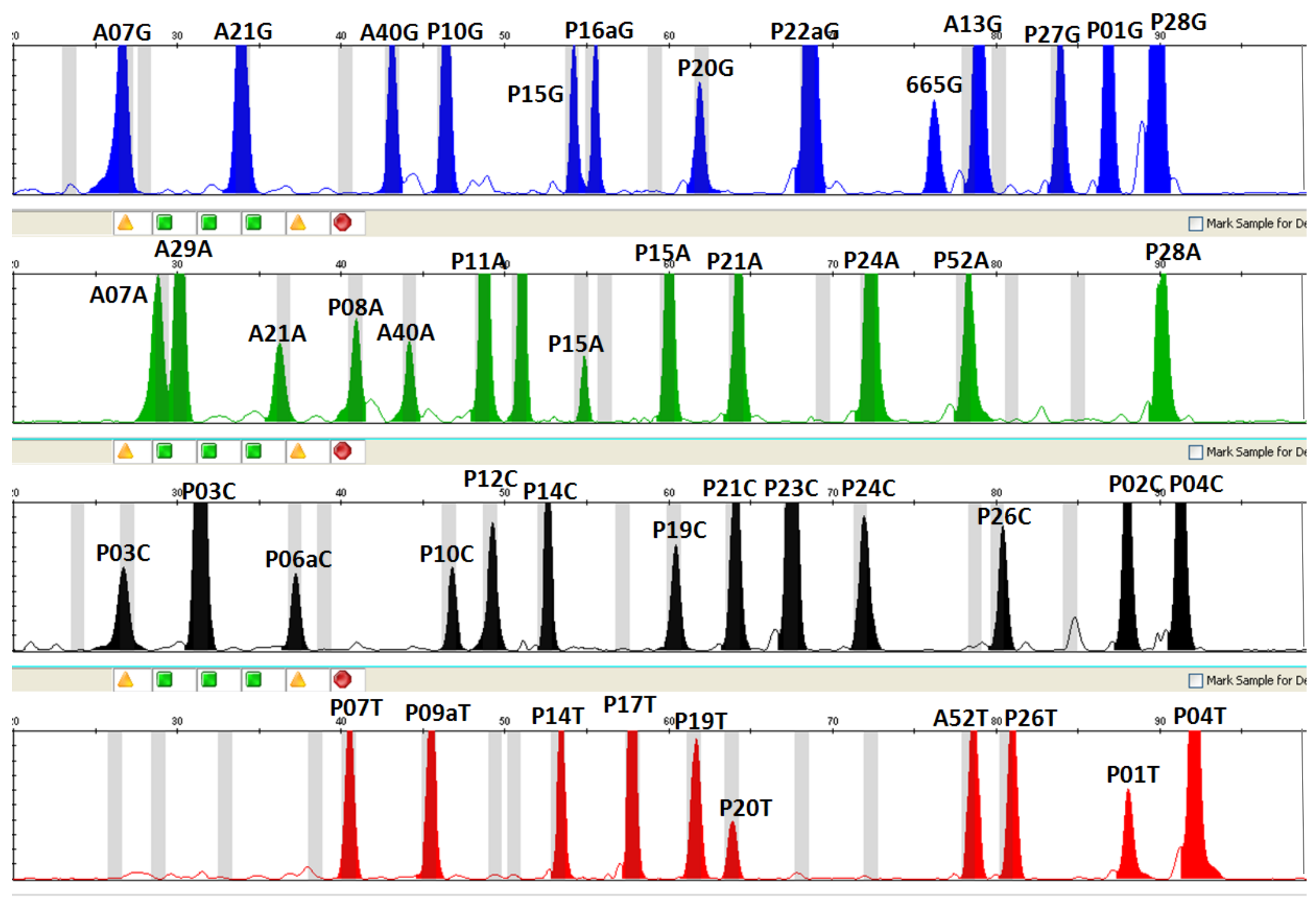
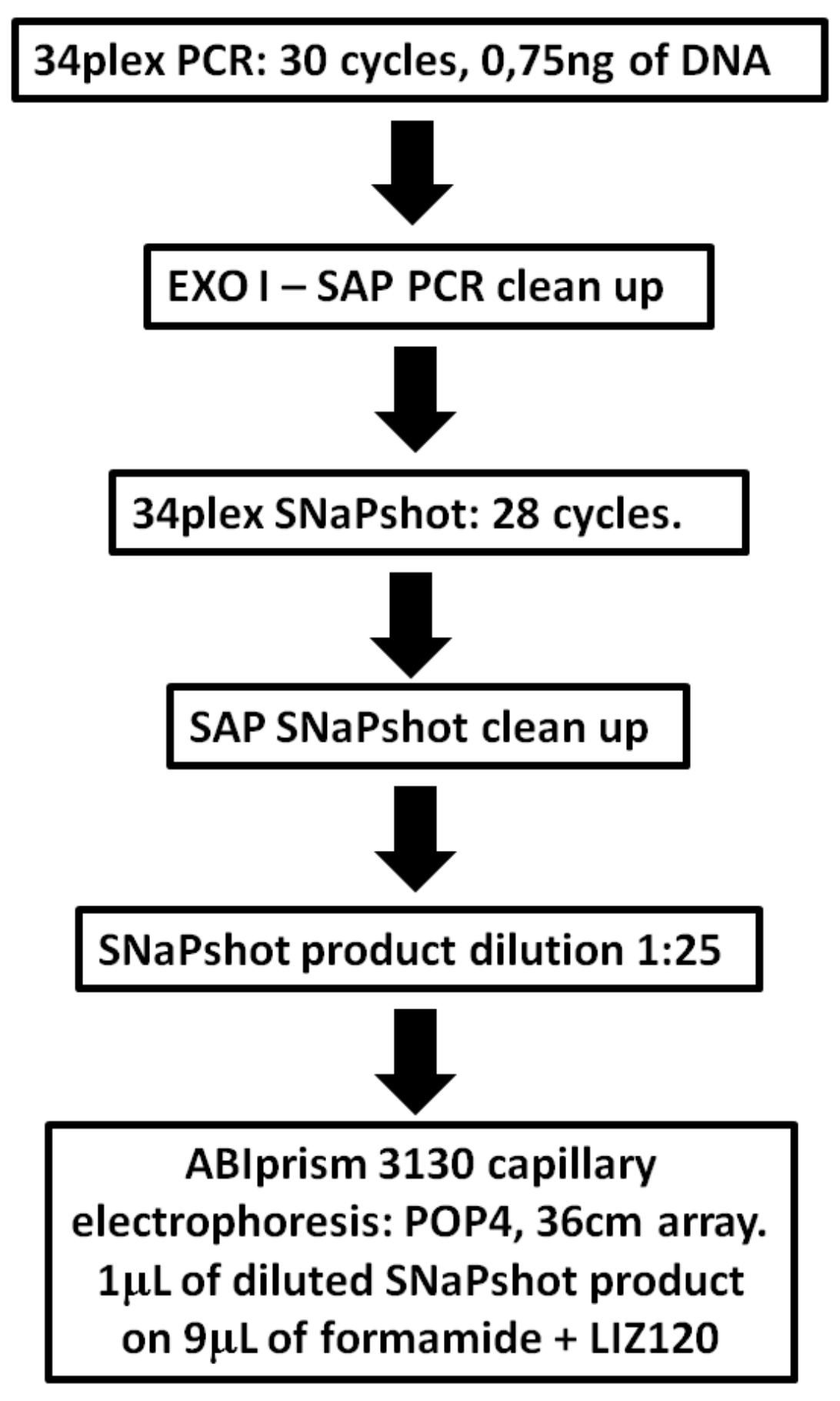


Fig. 1. Typical 34plex electropherogram

### Work flow


34plex PCR: 30 cycles, 0,75ng of DNA
EXO I – SAP PCR clean up
34plex SNaPshot: 28 cycles.
SAP SNaPshot clean up
SNaPshot product dilution 1:25
ABIprism 3130 capillary electrophoresis: POP4, 36cm array. 1µL of diluted SNaPshot product on 9µL of formamide + LIZ120

**Population samples:**

We have used the population samples from NIST, unrelated individuals, all of self-declared ancestry for genotyping [4], comprising:

- **262 U.S. Caucasians**
- **260 U.S. African-Americans**
- **140 U.S. Hispanics**
- **50 U.S. Asians**

For comparison purposes population data from human unadmixed groups, has also been analyzed. Population data was obtained from the *SPSmart* SNP browser online database [2]

**Data analysis:**

We employed the GeneMapper ID-X software from AB for the analysis of electrophoresis results. A new set of Bins and Panels for GeneMapper ID-X using POP4 and 36 cm capillary array is accessible through the NIST STRBase web page [www.cstl.nist.gov/biotech/strbase]. Structure v2.3.2 software has been applied for the allele frequency calculations, population substructure and predictive power of the test. 50,000 length of burning period value, 200,000 MCMC Reps and independent allele frequency, admixture model were used for the Structure calculations.

## Results

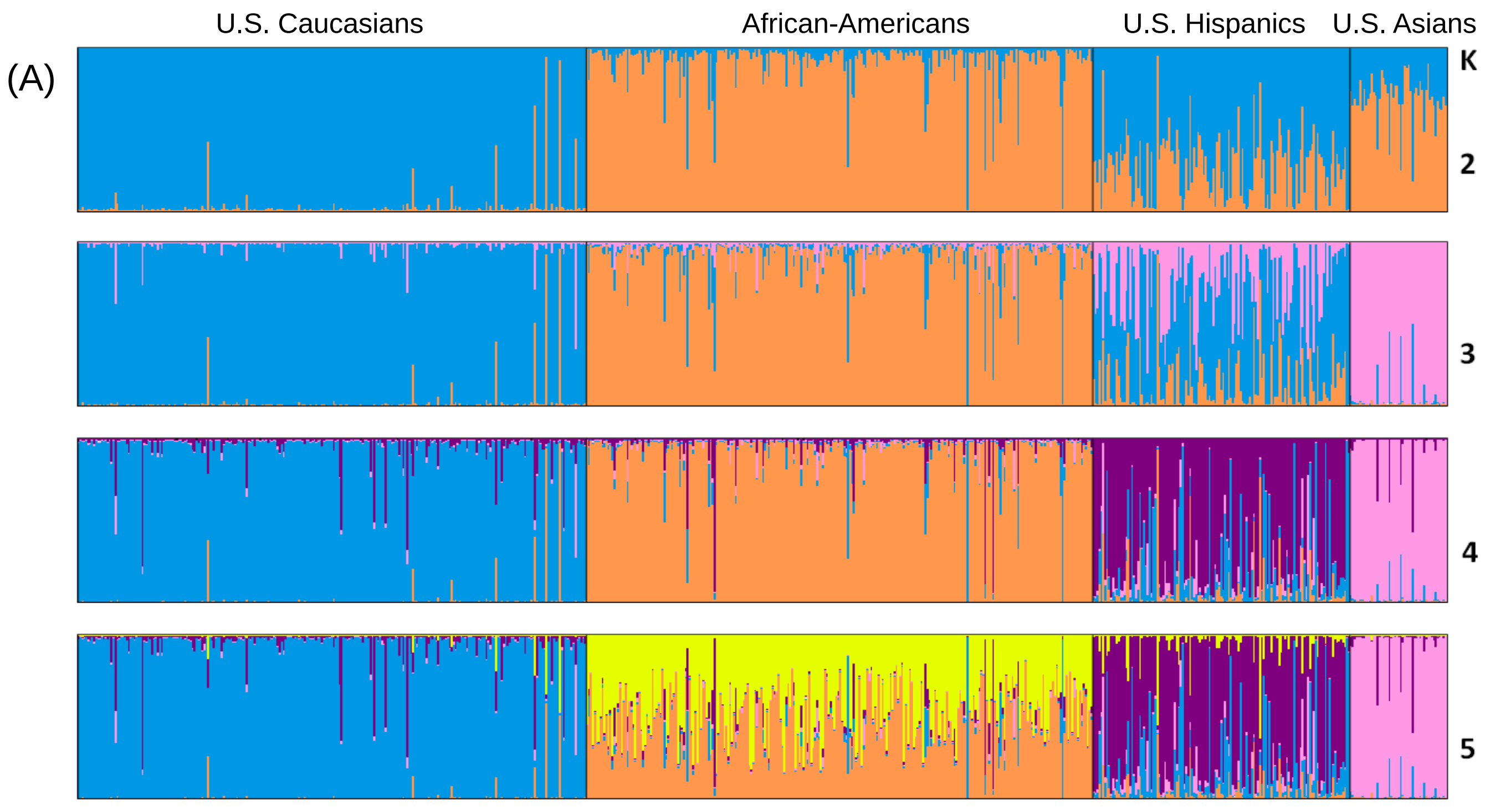
A well optimized reaction was obtained with the modifications implemented to our original 34plex design as shown in Fig. 1. Allele frequencies were calculated for each U.S. population as listed (below):

| Marker | | Population | | | |
|---|---|---|---|---|---|
| | | N=262 | N= 260 | N=50 | N=140 |
| | Allele | U.S. Cauc | Afr Amer | U.S. Asian | U.S. Hisp |
| P01 | G | 0.212 | 0.54 | 0.471 | 0.255 |
| | T | 0.788 | 0.46 | 0.529 | 0.745 |
| P02 | C | 0.994 | 0.554 | 0.961 | 0.936 |
| | A | 0.006 | 0.446 | 0.039 | 0.064 |
| P03 | C | 0.477 | 0.887 | 0.863 | 0.667 |
| | T | 0.523 | 0.113 | 0.137 | 0.333 |
| A07 | A | 0.275 | 0.658 | 0.363 | 0.38 |
| | G | 0.725 | 0.342 | 0.637 | 0.62 |
| A29 | A | 0.532 | 0.357 | 0.115 | 0.486 |
| | G | 0.468 | 0.643 | 0.885 | 0.514 |
| P05 | C | 0.716 | 0.877 | 0.578 | 0.727 |
| | T | 0.284 | 0.123 | 0.422 | 0.273 |
| A21 | A | 0.824 | 0.26 | 0.54 | 0.628 |
| | G | 0.176 | 0.74 | 0.46 | 0.372 |
| P06a | C | 0.657 | 0.937 | 0.872 | 0.592 |
| | T | 0.343 | 0.063 | 0.128 | 0.408 |
| P08 | G | 0.756 | 0.154 | 0.049 | 0.22 |
| | A | 0.244 | 0.846 | 0.951 | 0.78 |
| P07 | T | 0.985 | 0.429 | 0.98 | 0.855 |
| | C | 0.015 | 0.571 | 0.02 | 0.145 |
| A40 | A | 0.649 | 0.765 | 0.373 | 0.706 |
| | G | 0.351 | 0.235 | 0.627 | 0.294 |
| P09a | C | 0.017 | 0.471 | 1 | 0.05 |
| | T | 0.983 | 0.529 | n/a | 0.95 |
| P10 | C | 0.987 | 0.427 | 0.971 | 0.929 |
| | G | 0.013 | 0.573 | 0.029 | 0.071 |
| P11 | A | 0.945 | 0.575 | 0.5 | 0.908 |
| | T | 0.055 | 0.425 | 0.5 | 0.092 |
| P12 | C | 0.323 | 0.906 | 0.98 | 0.713 |
| | T | 0.677 | 0.094 | 0.02 | 0.287 |
| P13 | A | 0.994 | 0.571 | 0.971 | 0.936 |
| | G | 0.006 | 0.429 | 0.029 | 0.064 |
| P14 | C | 0.866 | 0.65 | 0.372 | 0.706 |
| | T | 0.134 | 0.35 | 0.628 | 0.294 |

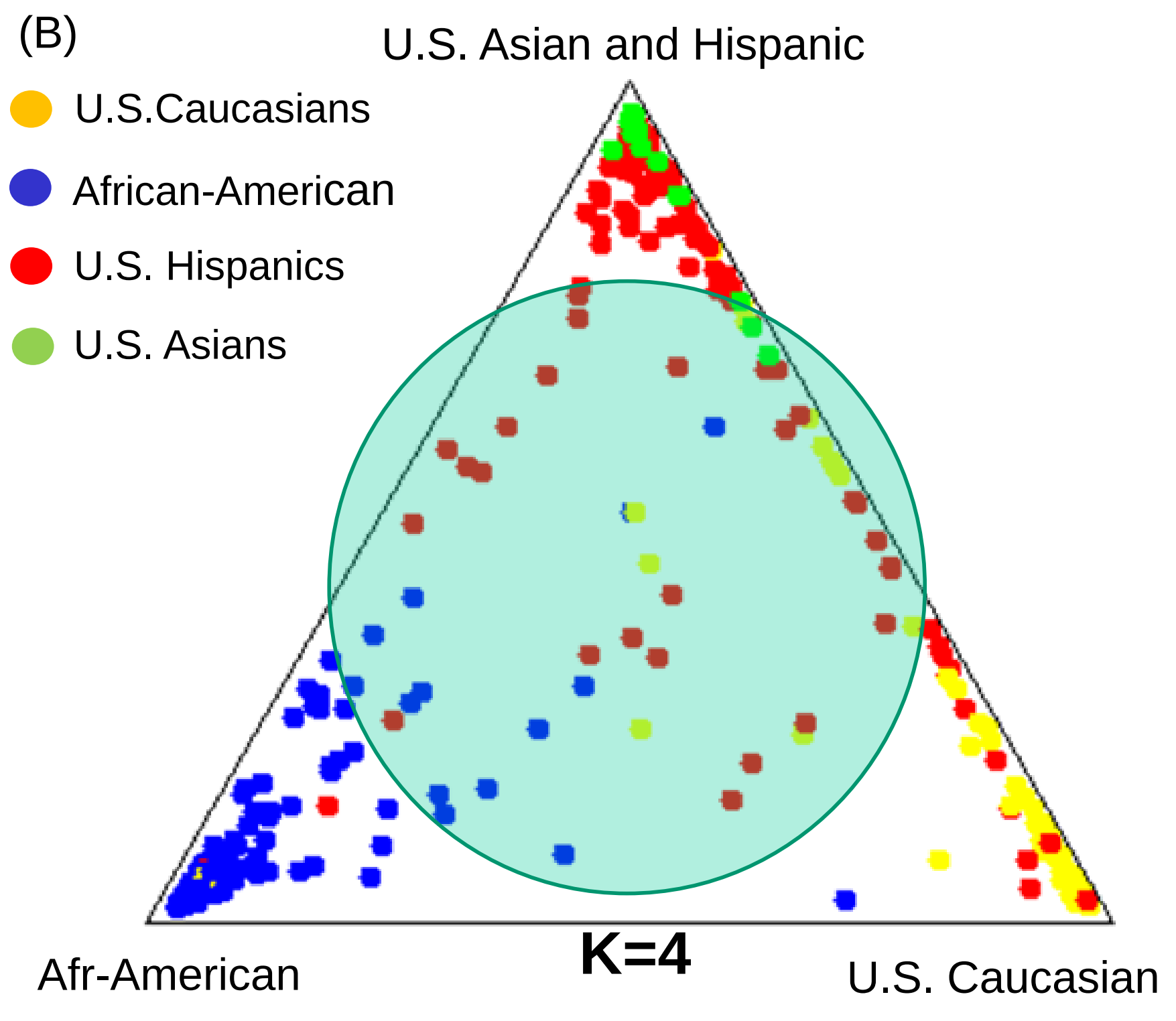
| Marker | | Population | | | |
|---|---|---|---|---|---|
| | | N=262 | N= 260 | N=50 | N=140 |
| | Allele | U.S. Cauc | Afr Amer | U.S. Asian | U.S. Hisp |
| P15 | A | 0.91 | 0.59 | 0.265 | 0.674 |
| | G | 0.09 | 0.41 | 0.735 | 0.326 |
| P16a | G | 0.992 | 0.515 | 1 | 0.961 |
| | A | 0.008 | 0.485 | n/a | 0.039 |
| P17 | C | 0.985 | 0.419 | 0.843 | 0.894 |
| | T | 0.015 | 0.581 | 0.157 | 0.106 |
| P18 | A | 0.954 | 0.875 | 0.225 | 0.78 |
| | G | 0.046 | 0.125 | 0.775 | 0.22 |
| P19 | C | 0.928 | 0.875 | 0.353 | 0.837 |
| | T | 0.072 | 0.125 | 0.647 | 0.163 |
| P20 | G | 0.607 | 0.853 | 0.078 | 0.585 |
| | T | 0.393 | 0.147 | 0.922 | 0.415 |
| P21 | A | 0.563 | 0.85 | 0.765 | 0.635 |
| | C | 0.437 | 0.15 | 0.235 | 0.365 |
| P23 | C | 0.01 | 0.794 | 0.941 | 0.308 |
| | T | 0.99 | 0.206 | 0.059 | 0.692 |
| P22a | G | 0.962 | 0.519 | 1 | 0.897 |
| | A | 0.038 | 0.481 | n/a | 0.103 |
| P24 | T | 0.189 | 0.317 | 0.524 | 0.399 |
| | A | 0.777 | 0.466 | 0.427 | 0.477 |
| | C | 0.034 | 0.217 | 0.049 | 0.124 |
| A52 | A | 0.731 | 0.188 | 0.647 | 0.511 |
| | T | 0.269 | 0.812 | 0.353 | 0.489 |
| P25a | C | 0.95 | 0.151 | 0.059 | 0.536 |
| | G | 0.05 | 0.849 | 0.941 | 0.464 |
| A13 | A | 0.517 | 0.335 | 0.216 | 0.411 |
| | G | 0.483 | 0.665 | 0.784 | 0.589 |
| P26 | C | 0.141 | 0.642 | 0.794 | 0.438 |
| | T | 0.859 | 0.358 | 0.206 | 0.562 |
| P27 | A | 0.091 | 0.351 | 0.136 | 0.127 |
| | C | 0.712 | 0.324 | 0.301 | 0.601 |
| | G | 0.196 | 0.324 | 0.563 | 0.272 |
| P28 | A | 0.982 | 0.983 | 0.235 | 0.791 |
| | G | 0.018 | 0.017 | 0.765 | 0.209 |
| P04 | T | 0.99 | 0.179 | 1 | 0.89 |
| | C | 0.01 | 0.821 | n/a | 0.11 |

### Population substructure analysis

**Fig. 2.** Structure analysis: Cluster plots (A) and triangle plots (B) of four U.S. study populations K = 2 to 5



Above: Despite detecting admixture in African-American, Caucasian and U.S. Asian study populations, the great majority of those samples would be correctly classified by our test. The three groups are well separated in the Structure cluster analysis at K=3. The Hispanic population becomes the fourth separated cluster at K=4. Not all samples are classified within the same cluster, though. The degree of admixture shown in Hispanics is the greatest amongst the analyzed populations.



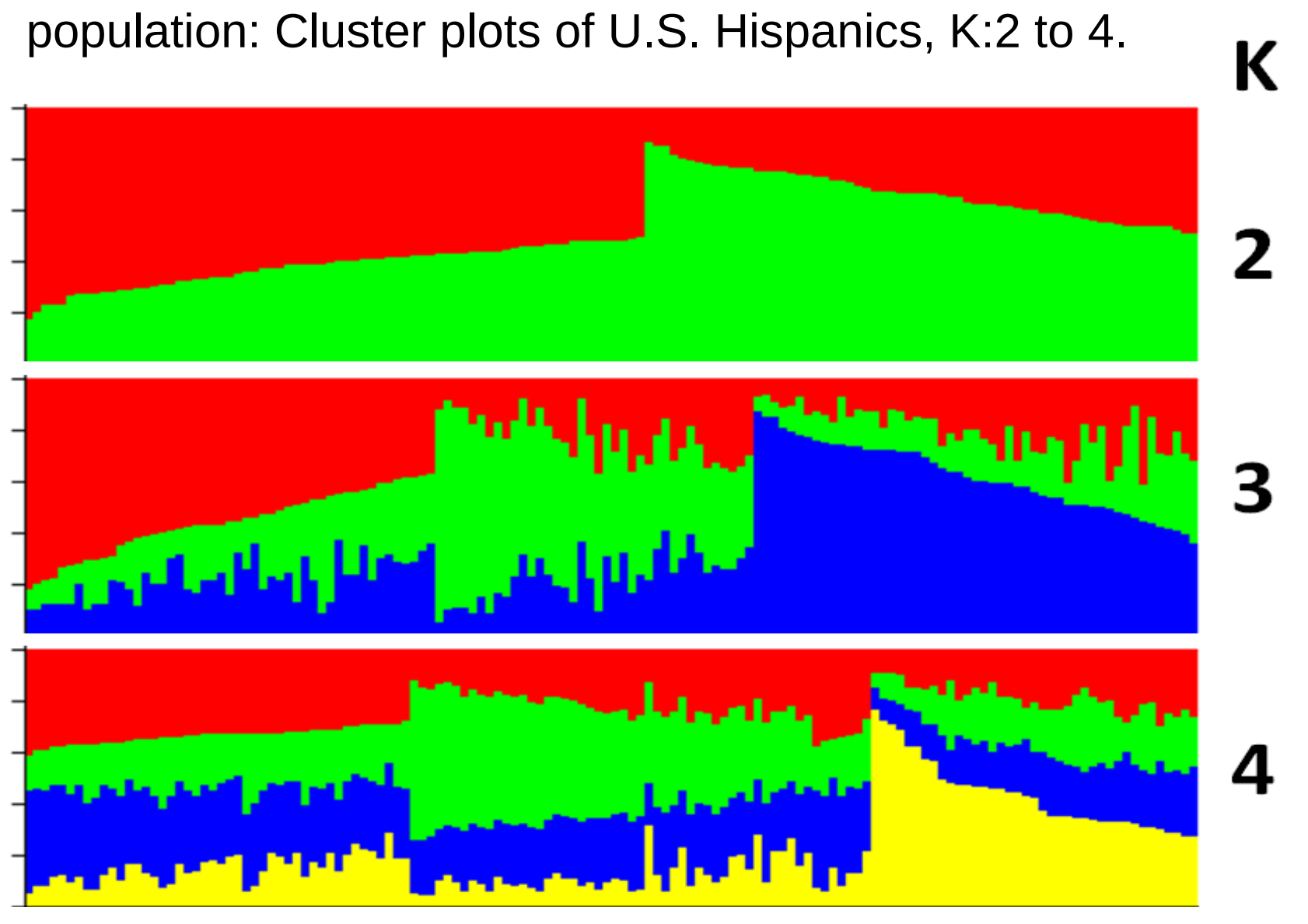
**U.S. population membership proportions**

| Given Pop | Inferred Clusters Ame | Asia | Afr | Eur | N |
|---|---|---|---|---|---|
| Cau : | 0.047 | 0.011 | 0.012 | **0.929** | 262 |
| Af-Am: | 0.036 | 0.017 | **0.915** | 0.032 | 260 |
| Hisp : | **0.680** | 0.055 | 0.069 | 0.195 | 140 |
| Asian : | 0.042 | **0.926** | 0.006 | 0.026 | 50 |

**Unadmixed population membership proportions**

| Given Pop | Inferred Clusters Afr | Ame | Eur | Asia | N |
|---|---|---|---|---|---|
| Afr: | **0.963** | 0.010 | 0.014 | 0.012 | 225 |
| Eur: | 0.005 | 0.019 | **0.962** | 0.015 | 278 |
| Ame: | 0.017 | **0.879** | 0.050 | 0.054 | 209 |
| Asia: | 0.004 | 0.044 | 0.019 | **0.934** | 352 |

**Fig. 3.** Structure analysis of complex U.S. population: Cluster plots of U.S. Hispanics, K:2 to 4.



Left: A triple ancestry substructure is observed within the U.S. Hispanic group, that is expected to correspond to Amerindian and Southern-European influx as well as a more recent African-American ancestry. This observation is also possibly due to high levels of heterogeneity within this population definition, since the "Hispanic" group is based on non-biological characteristics such as language which may also explain the observed substructure. This fact, along with the high degree of admixture observed in this group due to their history, likely also account for the observed decrease in the predictive power of the test.

**Fig. 4.** Correlation between self-declared ancestry and Structure-based genetic ancestry inferred from the 34plex markers and the 24 ancestry SNPs of Lao et al. 2010 [4]

| | Group Inferred from Structure k=4 | | | |
|---|---|---|---|---|
| 24 SNPs of Lao et al. 2010 | U.S. Caucasians | African Americans | U.S. Hispanics | U.S. Asians |
| U.S. Caucasians | 80.6% | 0.4% | 19.0% | 0.0% |
| African Americans | 1.0% | 96.8% | 2.2% | 0.0% |
| U.S. Hispanics | 15.7% | 4.0% | 77.8% | 2.4% |
| U.S. Asians | 0.0% | 0.0% | 0.1% | 99.9% |
| 34plex w/o admixture | U.S. Caucasians | African Americans | U.S. Hispanics | U.S. Asians |
| U.S. Caucasians | 96.6% | 0.4% | 3.1% | 0.0% |
| African Americans | 0.4% | 97.7% | 1.9% | 0.0% |
| U.S. Hispanics | 15.0% | 1.4% | 83.6% | 0.0% |
| U.S. Asians | 0.0% | 0.0% | 8.0% | 92.0% |

Left: Both data sets corresponding to Structure software cluster prediction with a non-admixture model. In both cases with a different set of ancestry informative SNPs we have a similar result for U.S. Hispanic group due to its complex admixture and non-biological grouping. Differences on the classification values for U.S. Caucasian and U.S. Asian groups may be due to a different number of ancestry informative SNPs for each population on both assays.

## Conclusions

- Allele frequency estimates have been made for U.S. representative population groups.
- Good predictive values were observed in spite of high admixture levels amongst the population samples for U.S. Caucasian, African-American, and U.S. Asian groups.
- Lesser values were observed for the Hispanic population. A higher degree of admixture and population substructure due to non-biological factors in the classification of U.S. Hispanics may explain this observation.

References:
1- Phillips C. et al. ( 2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. FSI: Genetics 1: 273-280.
2- Amigo J. et al. (2008) SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. BMC Bioinformatics 9: 428
3- Phillips C. et al. (2009) Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation. PLoS ONE 4(8):e6583. doi:10.1371/journal.pone.0006583
4-Lao O. et al. (2010) Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Hum Mutat. 12: E1875-93.