ANALYTICAL IMPROVEMENTS IN BIOGEOGRAPHIC ANCESTRY INFERENCE

M. Gabriel[1], C. Buchanan-Wright[1], A. Kumar[1], J. Lim[1], CW. Chang[1], J. Deng.[1], R. Lagacé[1] , S. Wootton[1]
[1]Thermo Fisher Scientific, South San Francisco, CA 94080

ABSTRACT

Ancestry-informative markers (AIMs) can be useful alongside phenotype informative markers to limit a suspect pool or provide investigative leads when STR profiles are either incomplete or fail to provide a database match. On next-generation sequencing (NGS) platforms, these markers can be multiplexed by the hundreds or thousands and present a comprehensive depiction of the unknown individual's ancestry and lineage using a consolidation of both autosomal and haploid markers. Panels including the Precision ID Ancestry Panel[1,2] as well as consortia grown panels such as MAPlex and VISAGE tools have motivated a recent software effort to allow for customizable marker sets and allele frequencies and to improve biogeographic prediction accuracy and error estimation. We propose a bootstrapped maximum likelihood approach to ancestry admixture prediction and assess its ability to assign biogeographical ancestry to samples from the 1000 Genomes Project and from samples typed for the ALFRED database[3]. Admixed profiles were then created by simulating inheritance from randomly selected profiles, and predictions were run on the simulated offspring. Using a higher confidence interval, we demonstrate correlation between reported error and uncertainty in the prediction due to lack of differentiation. We conclude that overall, predictions are generally consistent with self-reported ancestry, and for populations with predictions of higher uncertainty, we propose inclusion of region-specific markers that can further discriminate these populations.

INTRODUCTION

With rising interest in the application of forensic phenotyping and biogeographic ancestry, the interpretation of genotypes to ancestry result has become one of the key considerations[6]. While STRUCTURE[7] has been regarded one of the gold standards in interpretation, there has been discussion around the use of alternative methods as a complement. Here, we discuss a bootstrapped method of predicting admixture ratios in Converge™. The Precision ID Ancestry Panel was used as an example here, but the software enable custom panel use along with user defined genotype frequencies for user defined population groups.

MATERIALS AND METHODS

The below seven root populations (Table 1) for the Precision ID Ancestry Panel were created by hierarchically clustering 66 populations from ALFRED (Kidd) using the below similarity metric applied to reference allele frequencies.

Table 1.

Root populations
Africa
America
Southwest Asia (Middle East)
Oceania
East Asia
South Asia

$$S(A,B) = \sum_{i=1}^{N} \left| p_i^A - p_i^B \right|$$

where p^A_i = reference allele frequency of i^{th} SNP

A maximum likelihood approach is used to calculate the most likely admixture proportions across the root populations. In Converge, these predictions are bootstrapped across a subset of the SNPs in order to capture uncertainty in the predictions (Figure 1). Each bootstrap sampling is run through the core admixture algorithm, and this is repeated N times.

Figure 2 shows an example prediction for a profile extracted from an individual from the CEPH Utah dataset, where each column represents one iteration. The dotted yellow line shows the mean percentage of the European component across all iterations, and the yellow arrows indicate the variation across the predictions, which would provide the probable range.

Figure 1. Bootstrapped ancestry prediction

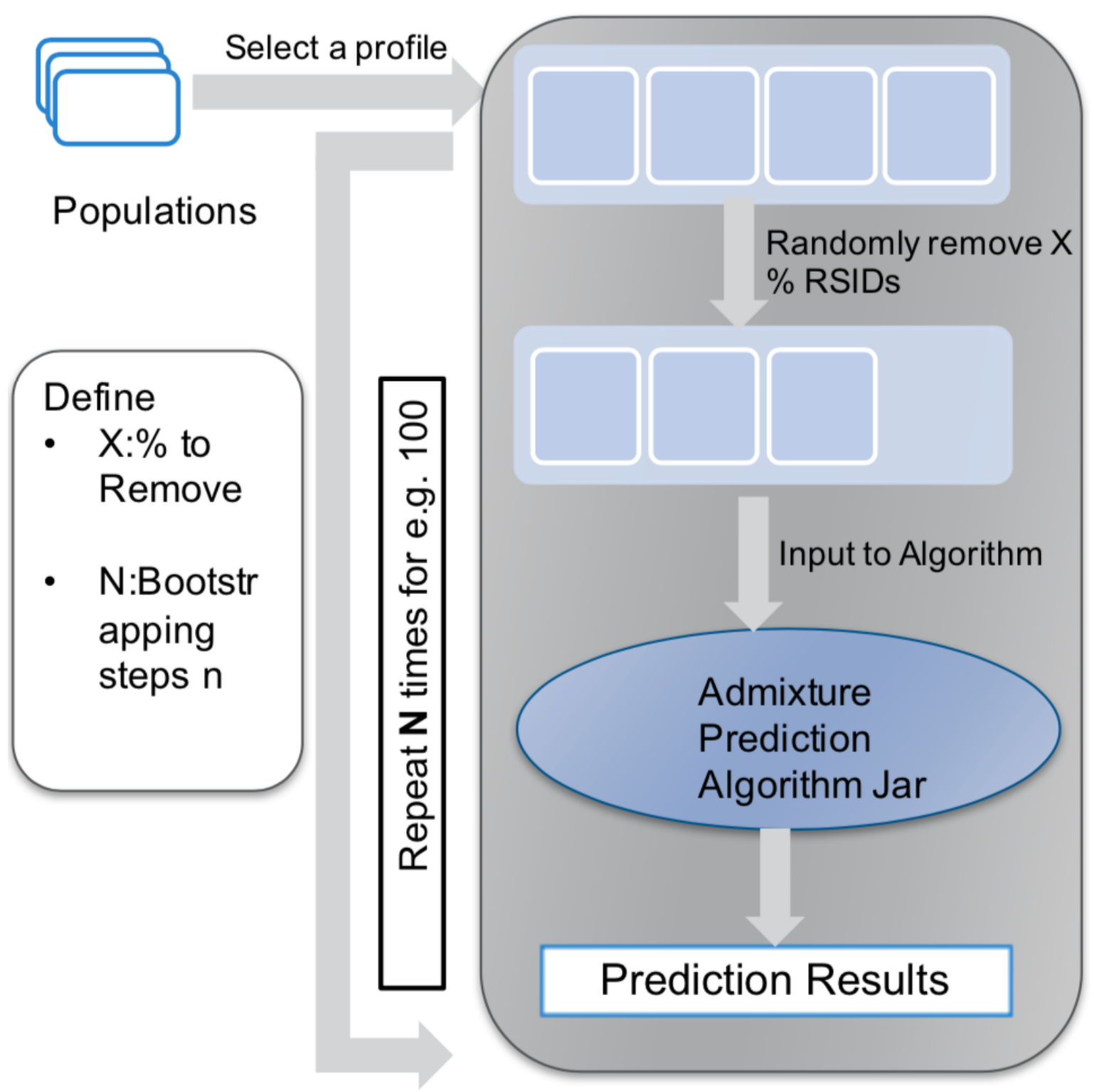
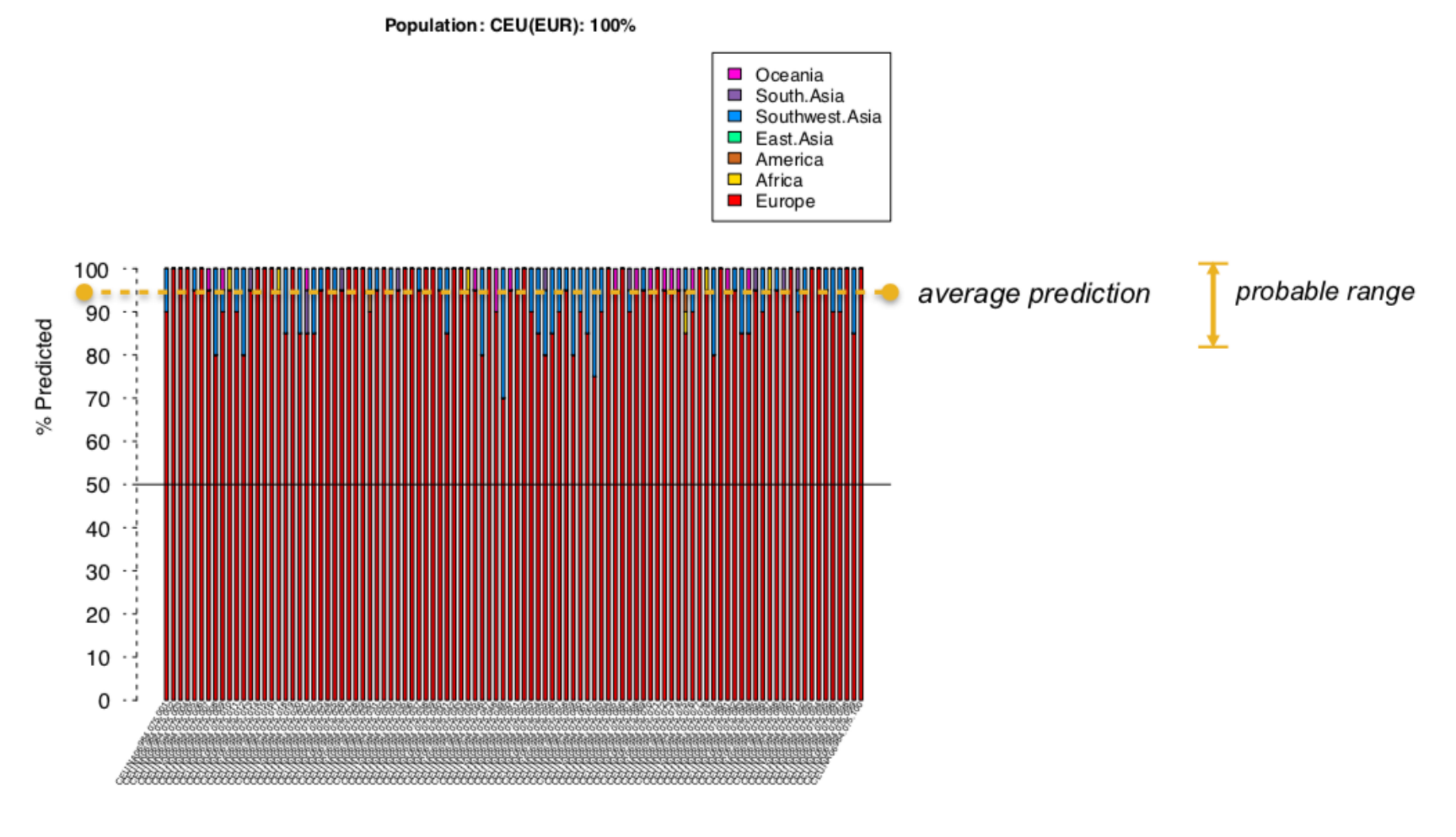

Select a profile
Populations
Define
• X:% to Remove
• N:Bootstrapping steps n
Repeat N times for e.g. 100
Randomly remove X % RSIDs
Input to Algorithm
Admixture Prediction Algorithm Jar
Prediction Results

Figure 2. Bootstrap iteration ancestry results



CEU individual (CEPH Utah) run through algorithm with 50% of SNPs sampled 100 times. Each bar represents results from one prediction.

To assess the accuracy of this method, we have run the ancestry module in silico on individual profiles from the below IGSR populations. Profiles for the Precision ID Ancestry Panel were extracted from the 1000 Genomes FTP site using tabix. The population groups represent both individuals who would likely assign to a single ancestry as well as those who would likely appear admixed (e.g. PUR).

Predictions for all individuals were performed by sampling 50% of SNPs 100 times with replacement (due to small panel size). A 99% confidence interval was used to allow for a more conservative estimate.

Table 2: 1000 Genomes Populations Used and Number of Profiles Extracted

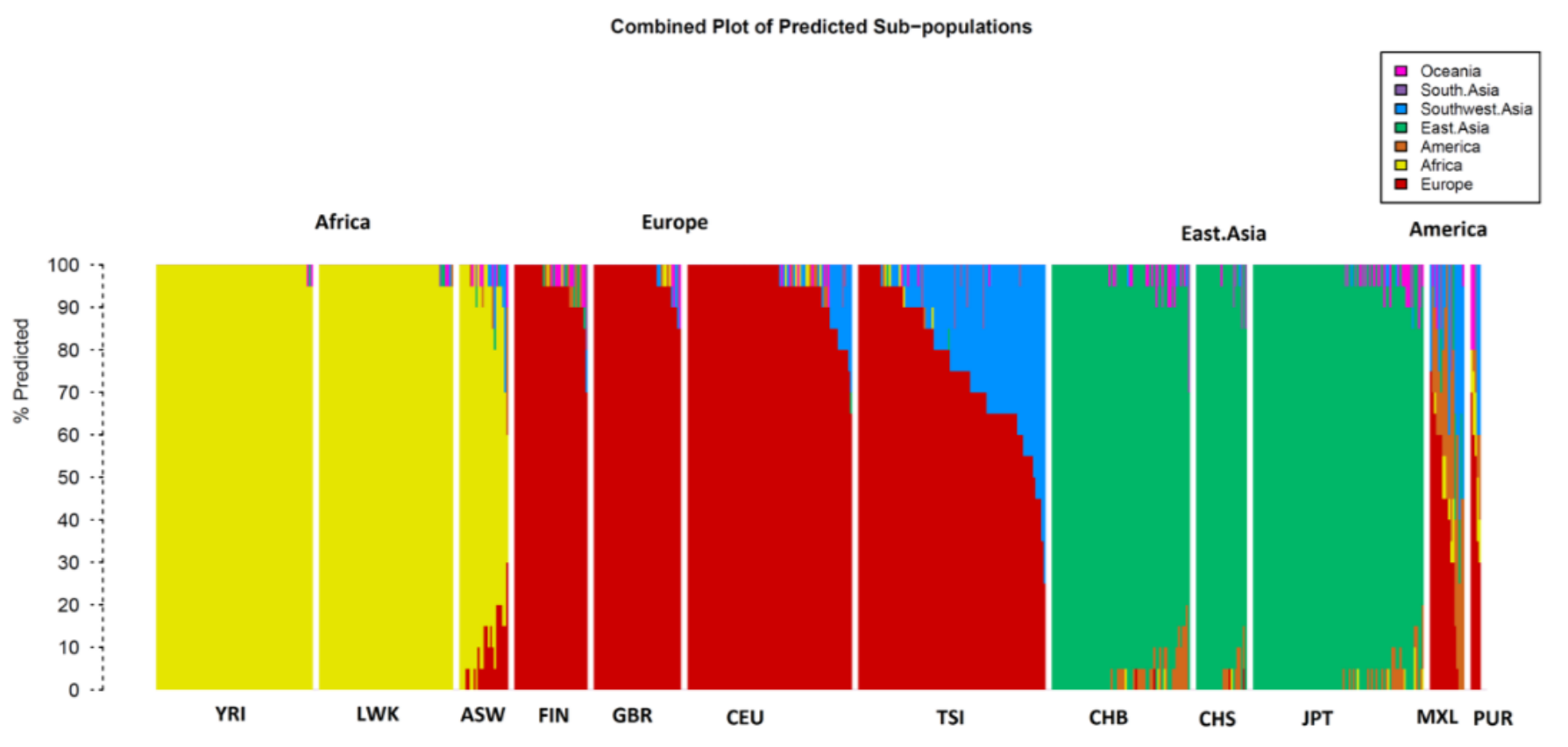
IGSR: The International Genome Sample Resource

No	Population Code	Population Name	# Profiles	1000G Group
1	ASW	Americans of African Ancestry in SW USA	24	AFR
2	LWK	Luhya in Webuye, Kenya	66	AFR
3	YRI	Yoruba in Ibadan, Nigeria	77	AFR
4	PUR	Puerto Ricans from Puerto Rico	5	AMR
5	MXL	Mexican Ancestry from Los Angeles USA	17	AMR
6	CHB	Han Chinese in Bejing, China	68	EAS
7	CHS	Southern Han Chinese	25	EAS
8	JPT	Japanese in Tokyo, Japan	84	EAS
9	CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	81	EUR
10	FIN	Finnish in Finland	36	EUR
11	TSI	Toscani in Italia	92	EUR
12	GBR	British in England and Scotland	43	EUR

RESULTS

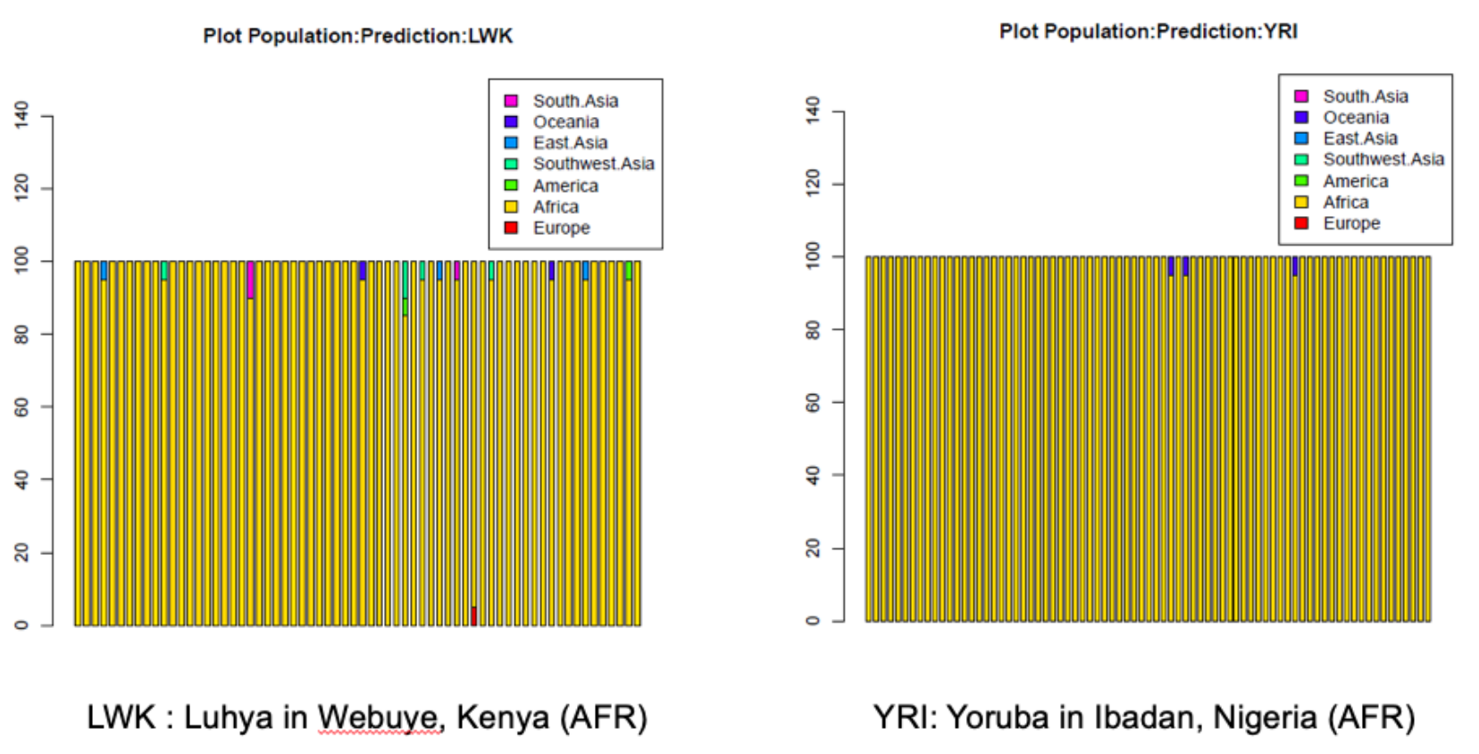
Means of the bootstrapped ancestry predictions were consolidated to display results in a STRUCTURE-like chart.

Figure 3. Mean ancestry compositions shown for each of the IGSR individual profiles, grouped by population group, and then by similarity in ancestry composition. Each column indicates one individual. Population codes are labeled below, and the population names can be cross referenced in Table 2.



Profiles from the Kenyan and Nigerian groups showed very little variability in predictions, with ancestry averaging close to 100% African.

Figure 4. Mean of bootstrapped ancestry predictions for Kenyan and Nigerian individuals. Each bar represents a single individual.

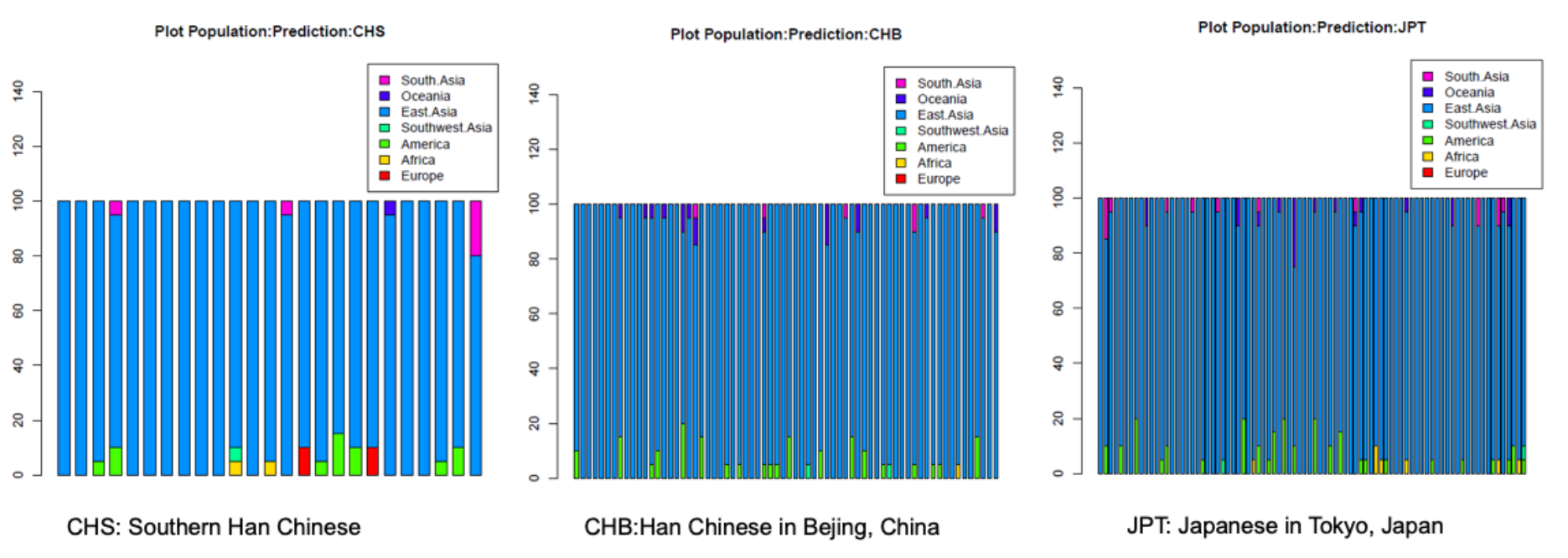


LWK : Luhya in Webuye, Kenya (AFR)

YRI: Yoruba in Ibadan, Nigeria (AFR)

Likewise, Southern Han Chinese, Han Chinese in Beijing, and Japanese populations showed near full attribution to East Asia, but as shown in Figure 5, a small percentage of Americas is also bleeding into the prediction. This variability is captured in the probable range.

Figure 5. Mean of bootstrapped ancestry predictions for East Asian populations. Each bar represents a single individual.



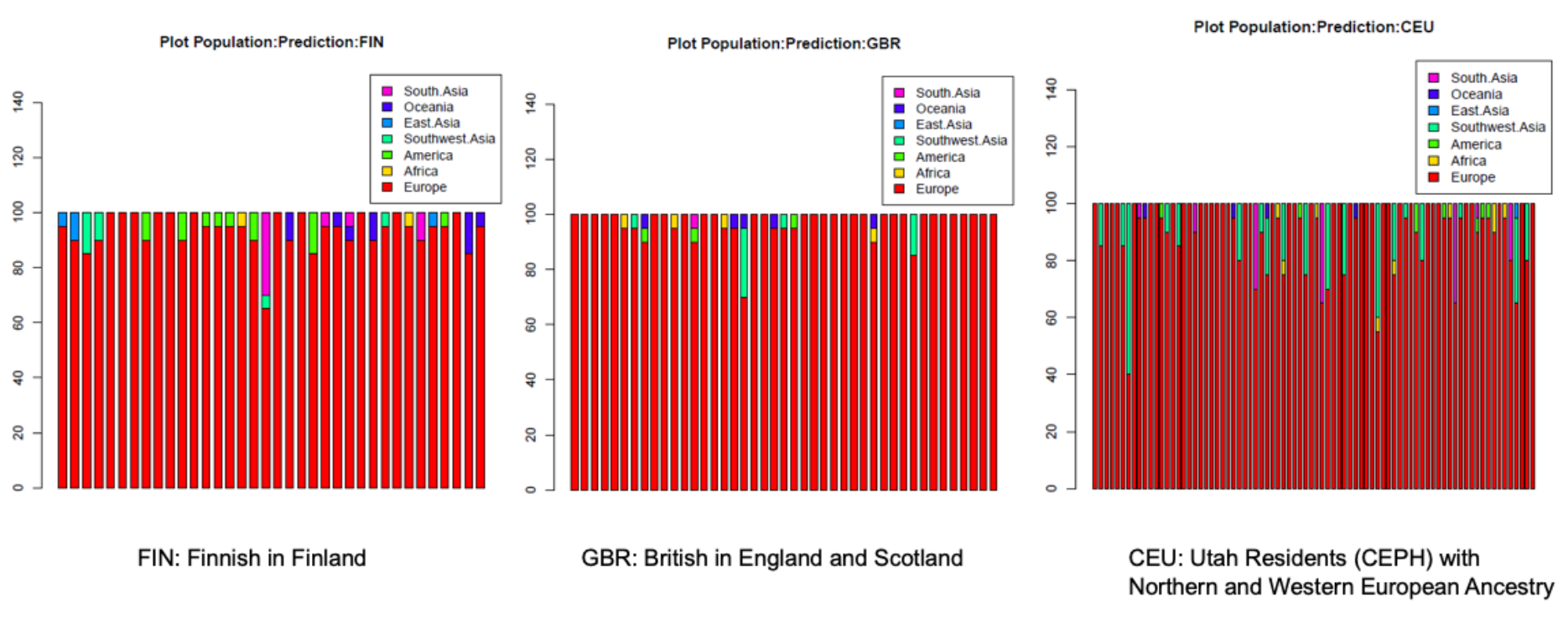
CHS: Southern Han Chinese

CHB:Han Chinese in Bejing, China

JPT: Japanese in Tokyo, Japan

Individuals of Finland, Great Britain, and Utah were mostly predicted to be European, with a base line noise around 10% or less.

Figure 6. Mean of bootstrapped ancestry predictions for Finnish, British, and Utah individuals. Each bar represents a single individual.



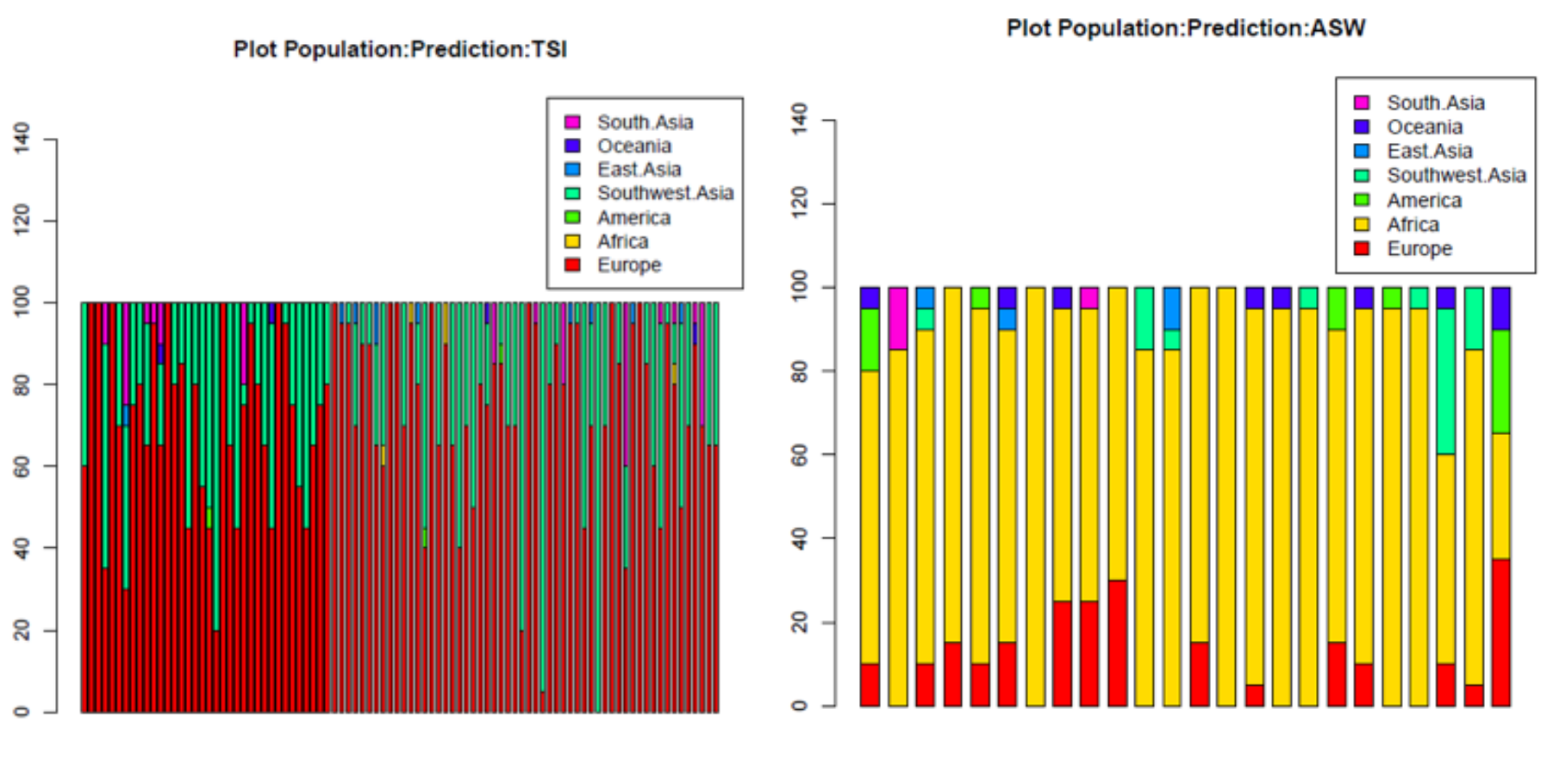
FIN: Finnish in Finland

GBR: British in England and Scotland

CEU: Utah Residents (CEPH) with Northern and Western European Ancestry

More interesting results were found in the Tuscan group and the admixed Americans of African Ancestry populations. The Tuscan individuals, while showning correctly as mostly European, also showed a high percentage of Southwest Asian. As shown in the TSI column of Figure 9, the error of the mean with a confidence interval of 99% is higher compared to other populations groups, indicating a lack of confidence in the mean prediction. The ASW group from the United States represents a admixed group, which individuals are mostly African with varying degrees of European percentages.

Figure 7. Mean of bootstrapped ancestry predictions for TSI and ASW. Each bar represents a single individual.



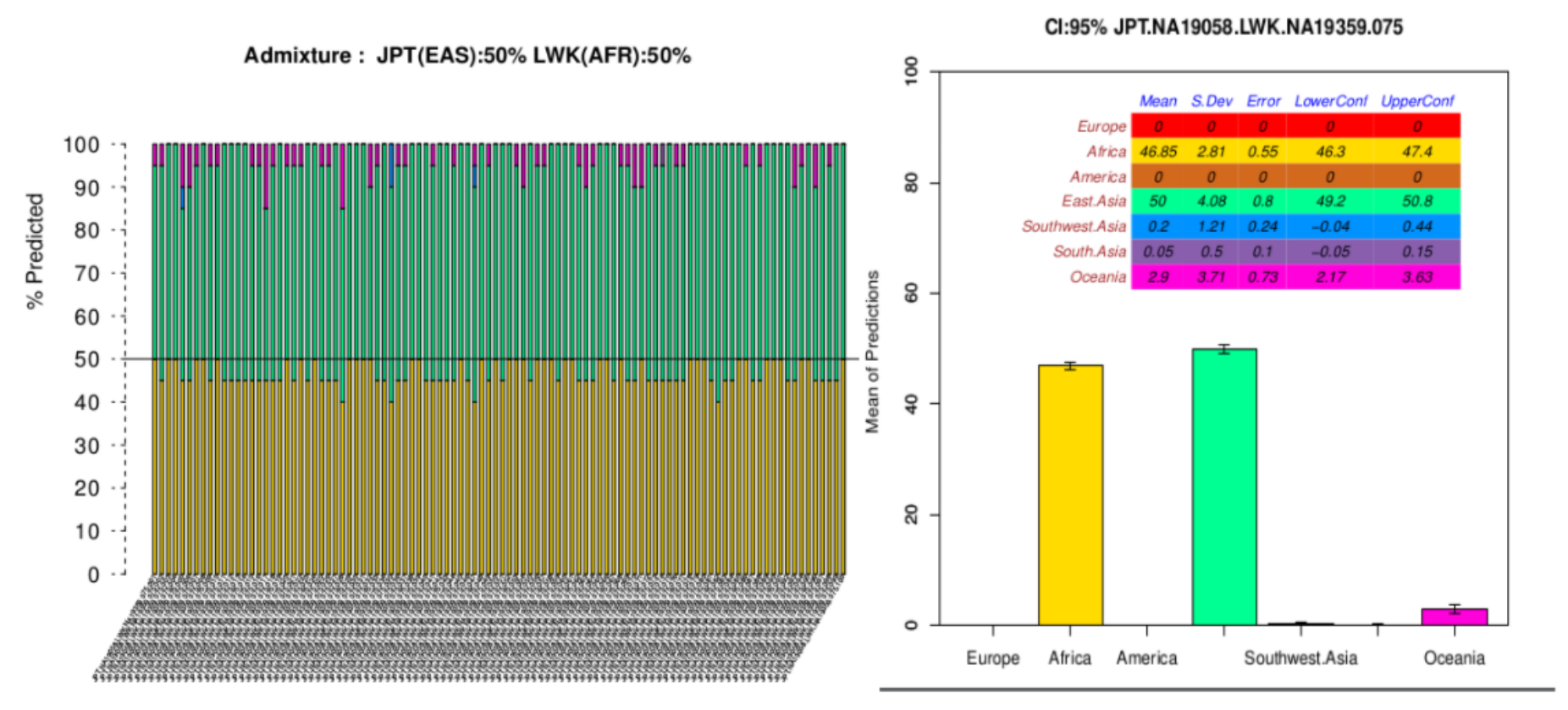
TSI: Toscani in Italia (EUR)

ASW: Americans of African Ancestry in SW USA (AFR)

Finally, a number of simulated admixtures were run through the classifier, with Figure 8 showing a individual with half Japanese ancestry and half Kenyan ancestry.

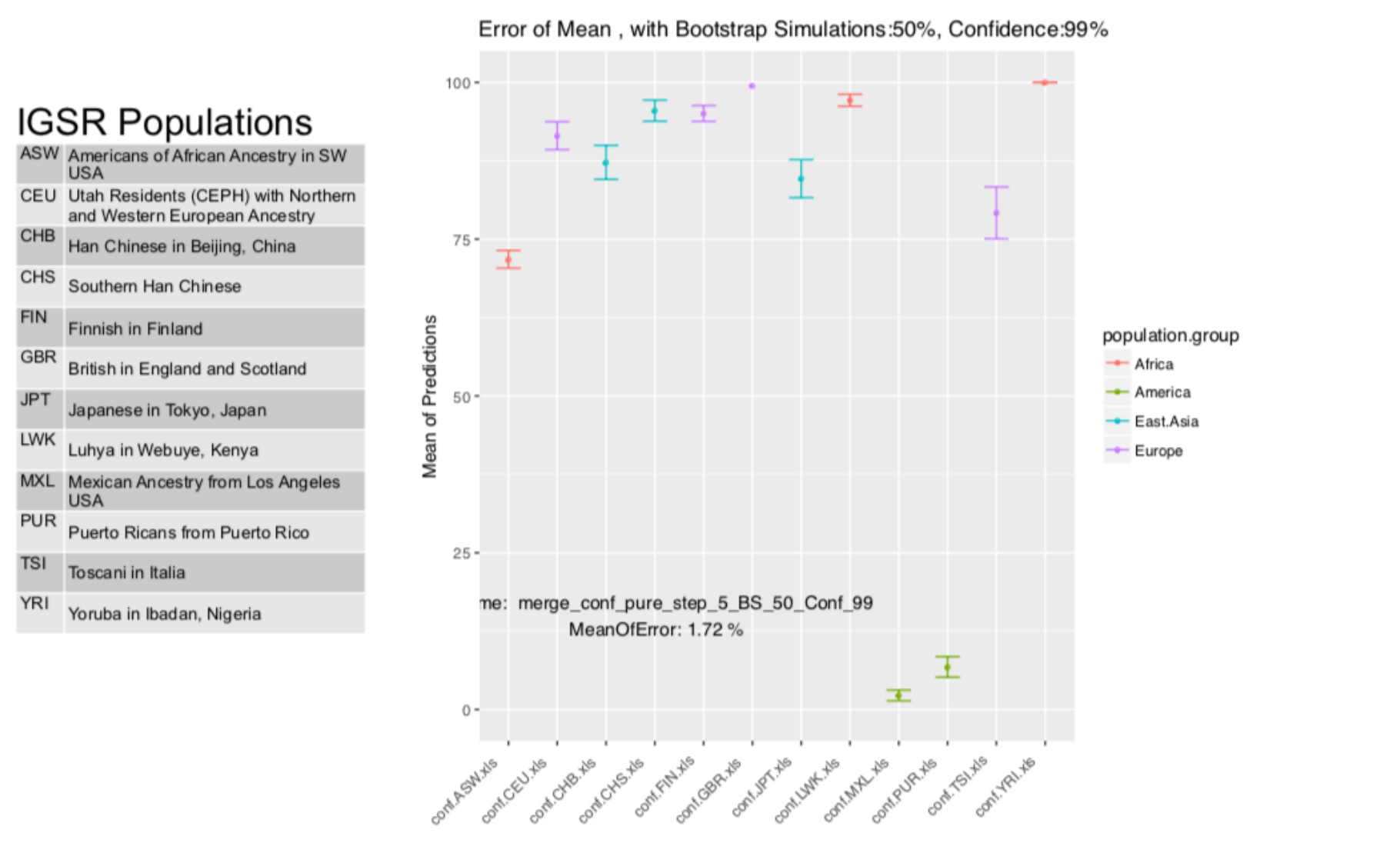
Figure 8. Simulated admixture between a Japanese individual and a Kenyan individual. Each bar represents a single bootstrap prediction. The graph to the right shows the mean for each root group along with the probable range.

Admixture : JPT(EAS):50% LWK(AFR):50%



	Mean	S.Dev	Error	LowerConf	UpperConf
Europe	0	0	0	0	0
Africa	46.85	2.81	0.55	46.3	47.4
America	0	0	0	0	0
East.Asia	50	4.08	0.8	49.2	50.8
Southwest.Asia	0.2	1.21	0.24	-0.04	0.44
South.Asia	0.05	0.5	0.1	-0.05	0.15
Oceania	2.9	3.71	0.73	2.17	3.63

Figure 9. Error of Mean for each population group using a sampling size of 50% of SNPs, a 99% confidence interval, and 100 iterations of bootstrapping.



IGSR Populations

ASW	Americans of African Ancestry in SW USA
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
CHB	Han Chinese in Beijing, China
CHS	Southern Han Chinese
FIN	Finnish in Finland
GBR	British in England and Scotland
JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MXL	Mexican Ancestry from Los Angeles USA
PUR	Puerto Ricans from Puerto Rico
TSI	Toscani in Italia
YRI	Yoruba in Ibadan, Nigeria

CONCLUSIONS

In summary, the bootstrapped predictions across the IGSR individuals show strength in some prediction capabilities (Africa, Northern Europeans) as well as some populations where there may be less clarity between the reported ancestries (i.e. Southern Europeans). In the latter situations, we rely on the variability between iterations of bootstrapped predictions to indicate less confidence in the reported ancestry means. As shown in Figure 9, some populations have higher error of the mean, so users can adjust expectations accordingly.

REFERENCES

1. Kidd K., et al., Progress toward an efficient panel of SNPs for ancestry inference, Forensic Science International: Genetics, Volume 10, Pages 23-32. (2014)
2. Kosoy R, et al., Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 30:69-78. (2009)
3. C. Phillips, et al., MAPlex - A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations, Forensic Science International: Genetics, Volume 42, Pages 213-226,
4. Rajeevan H, et al., ALFRED: an allele frequency resource for research and teaching. Nucleic Acids Research..(2011)
5. 4 Consortium, T. 1. G. P. (n.d.). A global reference for human genetic variation. Nature, 526, 68 EP –. http://doi.org/10.1038/nature15393
6. Cheung, E.Y.Y., Gahan, M.E. & McNevin, D. Int J Legal Med (2017) 131: 901. https://doi.org/10.1007/s00414-016-1504-3
7. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000 Jun;155(2):945-59. PubMed PMID: 10835412; PubMed Central PMCID: PMC1461096.

TRADEMARKS/LICENSING

For Research, Forensic, or Paternity Use Only. Not for use in diagnostic procedures.