Characterization of ancestry informative markers in the Tigray population of Ethiopia: a contribution to the identification process of dead migrants in the Mediterranean Sea

H.R.S. Kumar, K. Haddish, D. Lacerenza, S. Aneli, C. Di Gaetano, G. Tewelemedhin, R.V. Manukonda, N. Futwi, V. Alvarez-Iglesias, M. de la Puente, M. Fondevila, M.V. Lareu, C. Phillips, C. Robino

PII: S1872-4973(19)30305-9

DOI: https://doi.org/10.1016/j.fsigen.2019.102207

Reference: FSIGEN 102207

To appear in: *Forensic Science International: Genetics*

Received Date: 24 June 2019

Revised Date: 16 November 2019

Accepted Date: 19 November 2019

Please cite this article as: Kumar HRS, Haddish K, Lacerenza D, Aneli S, Di Gaetano C, Tewelemedhin G, Manukonda RV, Futwi N, Alvarez-Iglesias V, de la Puente M, Fondevila M, Lareu MV, Phillips C, Robino C, Characterization of ancestry informative markers in the Tigray population of Ethiopia: a contribution to the identification process of dead migrants in the Mediterranean Sea, *Forensic Science International: Genetics* (2019), doi: https://doi.org/10.1016/j.fsigen.2019.102207

Characterization of ancestry informative markers in the Tigray population of Ethiopia: a contribution to the identification process of dead migrants in the Mediterranean Sea

H.R.S. Kumar[a], K. Haddish[a], D. Lacerenza[b,c], S. Aneli[d,e], C. Di Gaetano[d,e], G. Tewelemedhin[a], R.V. Manukonda[a], N. Futwi[f], V. Alvarez-Iglesias[g], M. de la Puente[g], M. Fondevila[g], M.V. Lareu[g], C. Phillips[g], C. Robino[b*]

[a] Department of Forensic Medicine, University of Mekelle, Mekelle, Ethiopia

[b] Department of Public Health Sciences and Pediatrics, University of Turin, Turin, Italy

[c] Health Statistics and Biometry Residency Program, University of Turin, Turin, Italy

[d] Department of Medical Sciences, University of Turin, Turin, Italy

[e] Italian Institute for Genomic Medicine, Turin, Italy

[f] Tigray Health Research Institute, Mekelle, Ethiopia

[g] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain

[*]**Corresponding author:** Department of Public Health Sciences and Pediatrics, University of Turin, Corso Galileo Galilei 22, 10126 Turin, Italy

e-mail: carlo.robino@unito.it (C. Robino)

Highlights

- A Tigray (Ethiopia) population sample was typed with 46 AIM-indel and 31 AIM-SNPs
- A ~50% non-African genetic component was seen in Tigray by STRUCTURE analysis
- AIMs provide differentiation between Tigray and other sub-Saharan African populations
- Limited differentiation was possible between Tigray and Middle Eastern populations
- A reference AIMs dataset can help the identification of Eastern African dead migrants

Abstract

Determination of bio-geographical ancestry by means of DNA ancestry informative markers (AIMs) can contribute to the identification of human remains in missing person cases and mass disasters. While the presence of Eastern Africans among the migrant victims of trafficking accidents in the Mediterranean Sea is often suspected, few studies have addressed the ability of autosomal AIM panels in current use in forensic laboratories to provide differentiation of populations within the African continent. In this study, two assays consisting of 46 AIM-Indels and 31 AIM-SNPs were typed in a Tigray population sample from Northern Ethiopia. STRUCTURE analysis showed that the Tigray population is characterized by a strong (~50%) non-African genetic component shared with European and Middle Eastern populations. The intermediate position of the Tigray sample between sub-Saharan African and European / Middle Eastern reference population samples was confirmed by principal component analysis. Both AIM panels provided effective differentiation between Tigray and sub-Saharan African populations. Classification accuracy of other populations involved in the current Mediterranean migrant crisis, like South Asians, was superior with the AIM-SNP panel compared to the AIM-Indel panel. Misclassification of Middle Eastern samples as Tigray was frequent with both AIM-indel (~30% misclassified) and AIM-SNPs (~20%). However, with AIM-SNPs, error rates were reduced to acceptable levels by applying cautionary minimum thresholds to assignment likelihoods. Establishment of an Eastern African reference database of AIMs that can be genotyped by means of low cost, small-scale assays compatible with capillary electrophoresis, sets a balance between the need for ancestry inference tools and the budget limitations faced by Italian laboratories engaged in the humanitarian identification of dead migrants recovered from the Mediterranean Sea.

Keywords

Ancestry Informative Markers (AIMs); Indel, SNP; Ethiopia; Humanitarian investigations; Mediterranean migration crisis

1. Introduction

In recent years, analysis of ancestry informative markers (AIMs) has emerged as an important complementary tool in forensic DNA testing [1]. Although the most straightforward application of AIMs is as a genetic-based substitute for eyewitness testimony, ancestry inference of unidentified human remains can provide helpful investigative leads in missing persons cases [2] as well as aiding historical investigations [3]. Disaster victim identification (DVI) through DNA analysis of candidate family members also benefits from ancestry information, since AIMs can guide the choice of the most appropriate genotype and haplotype frequencies to be used in kinship calculations [4,5]. In cases involving individuals with genetic ancestry outside of well-studied geographic areas, AIMs can help identify genetically homogeneous groups of victims that can then be used themselves as genotype and haplotype reference datasets in matching probability estimates [6]. Finally, together with other types of post-mortem data available (age, stature, etc.), ancestry information can be used to adjust prior probabilities in likelihood ratio (LR) calculations [7].

All the aforementioned applications of AIMs in DVI cases extend to humanitarian investigations of mass disasters caused by the trafficking of migrants across the Mediterranean Sea to European countries. It has been estimated that just between January 2014 and June 2017 nearly 14,500 persons died in the attempt to reach Europe [8], a number equaling the total reported deaths for the whole period of 1988-2013 [9]. The Central Mediterranean route (from Libya to Italy across the Straits of Sicily) represents by far the most dangerous migration route, with 208 sunken ship incidents recorded between 2014 and 2017, and an average number of 44.8 fatalities per incident [8]. The geographical origin of migrants travelling the Central Mediterranean route, while constantly evolving over the years due to shifting socio-economic and political factors, spans a wide geographic area from Bangladesh to Western Africa, encompassing the Middle East, Eastern Africa and Northern Africa [10].

Efforts to identify deceased migrants are hampered by several factors, including data management issues [11] and budget constraints, as exemplified by the case of the 18th April 2015 shipwreck. This

was the deadliest accident of its kind ever recorded in the Straits of Sicily; with over 500 recovered bodies. The ongoing identification process of the 18th April 2015 victims would have not been possible, so far, without voluntary personnel and free equipment supplied by a consortium of Italian universities prompted by the National Office of the Commissioner for Missing Persons [12]. Moreover, despite the need for appropriate reference populations to accurately infer ancestry in an operational context, some of the main regions feeding the current migration flow to Europe, like Eastern Africa, are not represented in worldwide reference population datasets commonly employed in the development and validation of forensic ancestry panels, such as the Human Genome Diversity Project - Centre d'Étude du Polymorphisme Humain (HGDP-CEPH) [13] and the 1000 Genomes project [14].

With this in mind, we set out to establish a reference AIMs database representative of Eastern Africa by analyzing the Tigray population of Ethiopia. According to the latest Ethiopian census, Tigray people are the fourth largest ethnic group in Ethiopia, reaching up to 4.5 million and representing over 95% of the population in the regional state of Tigray (Northern Ethiopia) [15]. Tigray is also the major ethnic group of neighboring Eritrea [16], amongst the most common countries of origin for sub-Saharan African refugees crossing the Mediterranean Sea to reach Europe [10].

2. Materials and methods

2.1. Samples

Buccal swabs were collected from 252 consenting adult donors (Mekelle University students and staff) self-reported as having four grandparents of Tigray origin. The study was authorized by Mekelle University Research Ethics Review Committee (ERC 0841/2016).

2.2. Amplification reactions

DNA was isolated from buccal swabs with the ChargeSwitch® gDNA Normalized Buccal Cell Kit (Invitrogen) and 1 µl of each DNA extract (1-3 ng) was used in the following PCR experiments. A total of 77 AIMs were tested comprising: 46 AIM-Indels (herein, "46-I"), amplified according to the

multiplex PCR protocol previously described by Pereira et al. [17]; 31 SNPs ("Global AIMs Nano set", herein, "31-N") amplified according to the multiplex PCR and single base extension (SBE) protocols described by de la Puente et al. [18]. Detection and separation of PCR and SBE products were carried out using the ABI Prism 3500 Genetic Analyzer and GeneMapper ID software v5.1 (Thermo Fisher Scientific).

2.3. Statistical analyses

Tests of Hardy–Weinberg equilibrium (HWE), pairwise tests of linkage disequilibrium within and across the two AIM panels, as well as pairwise genetic distances (F_{ST}) were calculated with Arlequin software version 3.5 [19].

Ancestry inference was performed in comparison to reference population 46-I and 31-N genotypes derived from 1000 Genomes phase III data [14], including: sub-Saharan Africans (AFR: Esan n=99; Gambian n=113; Luyha n=99; Mende n=85; Yoruba n=108); East Asians (EAS: Han Chinese n=103; Dai n=93; Japanese n=104; Southern Han n=105; Vietnamese n=99); Europeans (EUR: British n=90; Finnish n=99; Iberian n=106; Toscani n=107; Utah n=99); South Asians (SAS: Bengali n=86; Gujarati n=103; Punjabi n=96; Tamil n=102; Telugu n=102). It should be noted that genotypes of two Indels in 46-I (rs3031979 and rs4183) are not listed in the 1000 Genomes population data. Reference samples were complemented with a Middle Eastern (MEA) reference population derived from HGDP-CEPH (Druze from Israel n=42; Bedouin from Israel n=46; Palestinians from Israel n=46; Algerian Mozabite n=29). 46-I genotypes of MEA populations were obtained from the literature [17,20]. Genotyping of MEA HGDP-CEPH samples with the 31-N set was performed in house, as previously described [18]. Although the 46-I and 31-N panels were shown to reliably distinguish other major population groups such as Native Americans and Oceanians [17,18], no reference populations from these geographic areas were included in the present study, as they lack relevance in the ongoing Mediterranean migration crisis.

Population analyses with STRUCTURE v. 2.3.4 [21] were performed with the following parameters: three replicates (for inferred clusters K:2 to K:7) of 100,000 burnin steps and 100,000 MCMC

iterations; correlated allele frequencies under the Admixture model (POPFLAG, both with and without LOCPRIOR option). The estimated $\log_n$ probability of data (-lnP(D)) values were plotted using STRUCTURE HARVESTER [22]. Ancestry membership plots were constructed using a combination of CLUMPP v. 1.1.2 [23] and distruct v. 1.1 [24].

Principal Component Analysis (PCA) was performed using the *princomp* function in R programming language environment [25] and an in-house developed script. The naïve Bayes classifier implemented in the Snipper 2.0 app suite (http://mathgene.usc.es/snipper/) was used to evaluate the classification success of AIMs in the tested populations through cross-validation, and to estimate log LRs of individual assignment probabilities [26]. Snipper cross-validation comparisons were also used to obtain Shannon's Divergence measures for each AIM [26]. Rosenberg's informativeness-for-assignment metric (I_n) [27] was then derived from divergence values by converting the natural log to log(2).

3. Results

3.1. Genetic characterization of the Tigray population

Genotypes of 252 Tigray individuals for the tested 46-I and 31-N markers are reported in Supplementary Table S1. 31-N genotypes of reference MEA HGDP-CEPH population samples are given in Supplementary Table S2. Tigray population allele frequency estimates are shown in Supplementary Table S3 together with the results of HWE test. No deviation from HWE was observed after Bonferroni correction for multiple testing ($\alpha = 0.0006$). Pairwise test of LD in the Tigray sample indicated significant LD between Indel marker rs16384 and SNP marker rs8137373 ($p < 0.00001$) after Bonferroni correction for multiple testing ($\alpha = 0.000016$). This is the closest pair of AIMs in the two sets, located on chromosome 22 and separated by 316 kilobases (dbSNP build 151).

Supplementary Table S4 reports for each AIM, the population specific divergence value obtained by comparing the Tigray sample with five population reference groups combined (I_nTIG), plus pairwise

divergence values calculated by comparing the Tigray sample with each population reference group. It can be seen that the markers displaying higher divergence values between Tigray and combined reference populations were mainly SNPs from the 31-N set, with the top Indel rs25630 ranking only 10th in the general list of I_nTIG values.

Pairwise F_{ST} values between Tigray and reference populations are shown in Table S5. All pairwise comparisons were statistically significant ($p < 0.00001$). It is also evident that F_{ST} values obtained with the 31-N panel were consistently higher than those from the 46-I set.

3.2. Ancestry analysis

STRUCTURE ancestry estimates with both admixture and admixture LOCPRIOR models showed a plateau in -ln P(D) values was reached after K=4, whether 46-I or 31-N panels were analyzed separately or in combination (Supplementary Figure S1). Ancestry membership proportions applying a four-group clustering (admixture LOCPRIOR) are plotted in Fig. 1A. It was evident that the Tigray population could not be separated using the present AIM panels, and it could be described as a balanced combination of African and non-African (European) components.

The same pattern was seen in MEA reference populations, which showed a prevalent European component, with an increase in African ancestral proportions in the Mozabite population sample from Algeria. The non-African component in the Tigray population was 58% measured by the 46-I panel, and 45% with 31-N (51%, when applying the combined AIM sets).

These results are mirrored in the 2D PCA plots (PC1 vs PC2) in Fig. 1B, with Tigray samples occupying an intermediate position between the AFR and EUR population clusters. Slightly improved clustering of Tigray, SAS and MEA populations was evident for the 31-N panel when compared to 46-I.

3.3. Classification success of AIM panels

The Snipper cross validation success rates of the Tigray study sample and reference population groups are shown in Table 1. This data indicates that 100% correct classification of Tigray samples could be achieved using the 46-I set alone. In contrast, 2.4% of Tigray samples were wrongly classified with the 31-N set, half of these erroneous samples being assigned to the MEA population group. In the reference population groups, a large proportion of MEA samples was misclassified as Tigray; with error rates ranged between 31.9%, when applying the 46-In set, to 17.8% for the 31-N set. In other reference population groups, samples wrongly classified as Tigray were always <1% with 31-N, while >3% of AFR and SAS samples were assigned to the Tigray population with 46-I. Combining the two AIM panels, while eliminating classification error in the Tigray sample, increased misclassification rates of SAS and MEA samples compared to 31-N alone.

To evaluate if the high sensitivity but low specificity displayed by the two AIM panels when used as tools to identify Tigray samples, could be explained by the difference in sample size between the study sample and reference MEA population samples, the Tigray sample was randomly split into six subsets (n=42) of a size comparable to MEA population subgroups. These subsets were then used separately in replications of the cross validation test. Average classification success of Tigray samples in replicates was 99.2% (±1.2% SD) with 46-I, and 96.8% (±1.2% SD) with 31-N, while 100% correct classification was confirmed when using the combination of 46-I and 31-N markers. Average misclassification rates of MEA samples as Tigray remained the same: 28.3% (±3.8% SD) for 46-I; 15.5% (±2.3% SD) for 31-N; 19.8% (±2.4% SD) for the combined AIM panels.

In the identification of human remains, DNA degradation can interfere with genotyping success of AIMs and consequently reduce the number of markers available for ancestry inference. The effect on classification accuracy as determined by cross validation using a reduced set of AIMs is shown in Supplementary Fig. S2. Markers were removed one by one from each panel in order of decreasing I_nTIG value. Elimination of the five Indels with the top I_nTIG values had minimal effect on classification rates, whereas a modest reduction of classification success of Tigray samples (from 96.8% to 93.6%) was seen in 31-N after exclusion of rs1871534, which ranked as the 5th most informative SNP in terms of I_nTIG, but which was also one of the most divergent loci in pairwise comparisons between Tigray and EUR, SAS and MEA reference population samples. Classification success rates with the combined 46-I and 31-N sets after removing from each panel the five markers with highest I_nTIG values matched those observed for the reduced 31-N panel alone, except for a slight increase in classification accuracy of Tigray samples from 93.7% to 97.6%. In contrast, the combined use of the top ten markers in terms of I_nTIG from the two AIM panels was necessary in order to achieve classification success rates equivalent to those attained with the 31-N complete set alone.

To improve balance between Snipper classification success and error rates, different thresholds were applied to LR values obtained through cross validation. Only samples with pairwise LR assignment probabilities that exceeded the set LR threshold were classified. The effect of thresholds was investigated only for population groups primarily involved in the Mediterranean migration crisis, i.e. AFR, SAS, MEA and Tigray representing Eastern Africa. Results are shown in Table 2. The error rates shown in Table 2 indicate that, leaving aside the MEA population, an LR threshold of 10^3 reduced error rates to <1% in both AIM panels. A striking difference between the two AIM panels was seen in the magnitude of LR assignment probability values obtained for SAS samples. With the 46-I panel, 93.9% of SAS individuals could not achieve LR assignment probabilities >10^3. Using the 31-N set, the ratio of unclassified SAS samples when applying the same LR threshold was reduced to 18.4%. Always assuming an LR threshold of 10^3, the misclassification rate of MEA samples was

1.8% for the 31-N panel compared to 11% for 46-I. The percentage of unclassified MEA samples was also smaller in 31-N (65.6%) than in 46-I (76.1%). Combining the two AIM panels generally reduced the ratio of unclassified samples observed at different LR thresholds and allowed for marginal improvement of classification success in Tigray samples. However, this came at the cost of an evident increase in error rates for SAS and MEA samples, in particular, compared to the 31-N panel alone.

Fig. 2 shows ranked pairwise (Tigray vs MEA) LR assignment probabilities obtained from cross validation in the four MEA population subgroups. It can be seen that, with a few exceptions in Palestinian and Bedouin samples, misclassification with high LR assignment probabilities (e.g. LR > 10^3) always involved individuals from the Mozabite population from Algeria.

4. Discussion

The 41-I AIM-Indel and 31-N AIM-SNP panels employed in this study were primarily designed to achieve ancestry inference at intercontinental level, separating sub-Saharan Africa, East Asia, Europe and Native America (plus Oceania for 31-N) [17,18]. Evaluation of a population sample from Eastern Africa (Tigray) confirmed that both AIM panels could effectively discriminate between Tigray and the non-African populations of Europeans and East Asians. In particular, of the top three AIMs in terms of I_nTIG, two (rs9809818 and rs17822931) were previously identified to be among the best markers separating East Asian and non-East Asian populations [1]. The 46-I panel was also previously shown to be able to differentiate Central South Asian populations in the CEPH-HGDP panel [21]. However, our results indicated an increased ability of the 31-N panel to distinguish between the South Asian reference populations from 1000 Genomes project and other populations (including Tigray) compared to 46-I, especially when cautionary thresholds based on LR assignment probabilities were applied.

Notably, both the 46-I and 31-N panels allowed for ancestry inference within Africa, differentiating with accuracy between Eastern Africans (Tigray) and other sub-Saharan African populations. This is in accordance with previous data showing that Tigray and, in general, Semitic and Cushitic speaking populations from the Horn of Africa region stand out among sub-Saharan African populations as being characterized by a strong (40%–50%) non-African component [28], probably reflecting ancient Eurasian backflow into Eastern Africa [29,30]. A clear non-African component in populations from the Horn of Africa is also evident in previous studies conducted with compact AIM panels built for forensic purposes. Similar STRUCTURE patterns were seen for Somalis, Ethiopian Jews and admixed African populations (African Americans) when applying a global 55-SNP panel [31]. Somalis displayed almost equal African and European cluster proportions when tested with a 126-SNP panel (EUROFORGEN Global ancestry panel) [32] and shared a minor membership proportion (averaging 26%) with Eurasian populations, if analyzed with a custom-built 111-SNP ancestry panel developed to analyze North African and Middle Eastern populations (EUROFORGEN NAME ancestry panel) [33]. The single most divergent AIM between Tigray and other sub-Saharan African population was rs2414778 located in the Duffy antigen receptor for chemokines locus, known to be the most differentiated marker between African and non-African populations [1]. The T>C mutation at rs2414778 leads to failure of Duffy antigen expression on the surface of red blood cells in humans, conferring resistance to *Plasmodium vivax* malaria, and is therefore under strong natural selection in malaria-endemic regions [34]. The C variant, while fixed in sub-Saharan African reference samples, had a reduced frequency of 0.687 in Tigray. Nevertheless, single removal of the rs2414778 SNP from the 31-N panel did not affect overall classification accuracy of the assay in Tigray and sub-Saharan African samples.

Limited specificity was observed when trying to differentiate between Tigray and Middle Eastern populations, with ~30% (46-I) to ~20% (31-N) of Middle Eastern reference samples identified as Tigray. Also in this case, the 31-N panel gave an overall better classification performance compared to 46-I. The 31-N error rates could be reduced to a much more acceptable ~5%, by applying a

cautionary LR threshold of 10^2, while retaining the ability to correctly identify ~50% of the Middle Eastern samples. In contrast, for the 46-I set the same error rate of ~5% was associated with over 90% of samples falling below the LR classification threshold. Misclassification rates and LR assignment probabilities in pairwise Tigray / Middle East comparisons of Middle Eastern samples increased according to an East-West gradient, reaching a maximum in the Mozabite population sample. This reflects a documented cline of sub-Saharan African ancestry decreasing from Western Sahara eastward [35]. Notably, no misclassification was observed between the Israel Druze population sample and Tigray, in agreement with a subdivision of Levantine populations in two main branches: one, including the Druze, genetically close to European and Central Asian populations; the second, including Palestinians and Bedouins, with stronger affinities with North African and Ethiopian populations [36]. Recently, it was shown that limitations in the ability to classify Middle Eastern samples, as observed in the present study, could not be completely overcome even by adopting an extended 237 AIM-SNP assay (EUROFORGEN Global and NAME panels combined) with a reported assignment error rate of 22.4% in cross-validation studies including six continental populations plus Northern Africa (but not Eastern Africa) [33]. In the context of the present European refugee crisis and of the efforts being made to identify deceased migrants, discrimination between Middle Eastern and Eastern African individuals can also rely on additional genetic data, including differential distribution of Y-chromosomal [37,38] and mitochondrial DNA variation [39], and to some extent on additional non-genetic tools such as cranial morphometric analysis [40]. However, identification by high-density SNP genotyping of putative autochthonous ancestral components in Eastern and Northern Africa [28,35] suggests the future possibility to implement current worldwide AIM panels with complementary assays improving ancestry inference within this specific geographical area, in a similar way to what was previously achieved for the Mediterranean basin [41], Eurasia [33,42] and the Pacific region [43].

In general, combined use of the two AIM panels did not lead to a significant improvement in overall classification accuracy, compared to that provided by the 31-N set alone. A further factor was the

presence of a pair of closely linked markers in the two AIMs sets (rs16384 and rs8137373), shown to be in strong LD in the Tigray population. Use of multiple markers located on the same genomic segment may affect likelihood calculations and can bias co-ancestry proportion estimates in admixed individuals [44]. The 46-I, therefore, could play the role of a complementary set of AIMs when several 31-N markers fail to amplify due to DNA degradation. A further reason to include 46-I in the analysis of challenging samples is to screen for DNA mixtures and contamination, given the ability of AIM-indel assays to readily identify unbalanced heterozygous peaks in genotypes [45].

Markers included in the 31-N set are a selection of the 128 AIMs previously identified as those displaying maximum divergence at inter-continental level after thorough bioinformatic interrogation of public genomic data [44]. Besides providing basic bio-geographical ancestry inference between the main areas of origin of migrants travelling the Central Mediterranean route, the 31-N SNaPshot assay possesses additional features making it suitable for the humanitarian identification of migrant victims drowned in the Straits of Sicily. First of all, it was specifically designed to target short-amplicon markers, thus enabling the genotyping of low-level DNA samples [18,46]. Recovery of low-level DNA is expected in the case of human remains submerged in sea water for long periods of time [47], such as those retrieved from the 18th April 2015 shipwreck [12]. Moreover, it represents a validated, low cost and time-effective method compatible with standard capillary electrophoresis (CE) equipment, thus fitting in to the current budget limitations imposed on Italian laboratories involved in the identification efforts of migrant victims [12].

5. Concluding remarks

A reference database of 77 AIMs was established in the Tigray population of Ethiopia. Markers consisted of 46 Indels and 31 SNPs from two small-scale CE genotyping assays that can be easily integrated in operational forensic laboratories. The provided dataset can act as an Eastern African reference for general investigative purposes and in missing persons or DVI cases. The availability of new AIMs data will hopefully contribute to identification procedures in migration accidents occurring

in the Mediterranean Sea, in which the identification of victims of Eastern African ancestry is frequently required.

Finally, even though only eight SNP markers in the 31-N AIM set overlap with those included in commercially available AIM assays recently developed for massive parallel sequencing (MPS) platforms [48,49], it is expected that some of the caveats indicated by the present study will also apply to MPS-based AIM-SNP panels. In particular, the non-African genetic component observed in Eastern Africa, while allowing accurate differentiation from other sub-Saharan African populations, can lead to possible misclassification with admixed individuals from populations displaying variable proportions of African ancestry. Although specific data for the Ethiopian Tigray population is not presently available, it was shown that, when analyzing a neighboring Eastern African population (Somalia) with the 165-SNP Precision ID Ancestry Panel AIM, 5% of the samples were erroneously reported as African American and 3% as Middle Eastern in origin [50]. For the same reason, it is advisable in the future to include Eastern African reference samples in validation studies of custom-built MPS assays aimed at the differentiation of Eurasian sub-population groups.

Acknowledgments

This work was supported by: WWS Project mobility grant to C.R; Fondi di Ricerca locale (ex 60%) Department of Medical Sciences Grant 2018 to C.D.G.

References

[1] C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. 18 (2015) 49–65.

[2] C. Hollard, C. Keyser, T. Delabarde, A. Gonzalez, C. Vilela Lamego, et al., Case report: on the use of the HID-Ion AmpliSeq™ Ancestry Panel in a real forensic case, Int. J. Legal Med. 131 (2017) 351-358.

[3] A.D. Ambers, J.D. Churchill, J.L. King, M. Stoljarova, H. Gill-King, et al., More comprehensive forensic genetic marker analyses for accurate human remains identification using massively parallel DNA sequencing, BMC Genomics 17 (2016) 750.

[4] M. Prinz, A. Carracedo, W.R. Mayr, N. Morling, T.J. Parsons, et al., DNA Commission of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role of forensic genetics for disaster victim identification (DVI), Forensic Sci. Int. Genet. 1 (2007) 3-12.

[5] R.V. Rohlfs, S.M. Fullerton, B.S. Weir, Familial identification: population structure and relationship distinguishability, PLoS Genet. 8 (2012) e1002469.

[6] L. Olivieri, D. Mazzarelli, B. Bertoglio, D. De Angelis, C. Previderè, et al., Challenges in the identification of dead migrants in the Mediterranean: The case study of the Lampedusa shipwreck of October 3rd, 2013, Forensic Sci. Int. 285 (2018) 121-128.

[7] W.H. Goodwin, The use of forensic DNA analysis in humanitarian forensic action: the development of a set of international standards, Forensic Sci. Int. 278 (2017) 221-227.

[8] F. Laczko, A. Singleton, J. Black (Eds.), Fatal Journeys Volume 3 – Part 1: Improving Data on Missing Migrants, IOM's Global Migration Data Analysis Centre, Berlin, 2017.

[9] S.T.D. Ellingham, P. Perich, M. Tidball-Binz, The fate of human remains in a maritime context and feasibility for forensic humanitarian action to assist in their recovery and identification, Forensic Sci. Int. 279 (2017) 229-234.

[10] https://data2.unhcr.org/en/situations/mediterranean

[11] U. Hofmeister, S.S. Martin, C. Villalobos, J. Padilla, O. Finegan, The ICRC AM/PM Database: Challenges in forensic data management in the humanitarian sphere, Forensic Sci. Int. 279 (2017) 1-7.

[12] V. Piscitelli, A. Iadicicco, D. De Angelis, D. Porta, C. Cattaneo, Italy's battle to identify dead migrants. Lancet Glob. Health. 4 (2016) e512-513.

[13] H.M. Cann, C. de Toma, L. Cazes, M.F. Legrand, V. Morel, et al., A human genome diversity cell line panel, Science. 296 (2002) 261–262.

[14] 1000 Genomes Project Consortium, G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, et al., An integrated map of genetic variation from 1092 human genomes, Nature 491 (2012) 56–65.

[15] http://www.csa.gov.et/census-report/complete-report/census-2007

[16] https://www.easo.europa.eu/sites/default/files/public/Eritrea-Report-Final.pdf

[17] R. Pereira, C. Phillips, N. Pinto, C. Santos, S.E. dos Santos, et al., Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing, PLoS One 7 (2012) e29684.

[18] M. de la Puente, C. Santos, M. Fondevila, L. Manzo; EUROFORGEN-NoE Consortium, et al., The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs, Forensic Sci. Int. Genet. 22 (2016) 81-88.

[19] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (2010) 564-567.

[20] C. Santos, C. Phillips, F. Oldoni, J. Amigo, M. Fondevila, et al, Completion of a worldwide reference panel of samples for an ancestry informative Indel assay, Forensic Sci. Int. Genet. 17 (2015) 75–80.

[21] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.

[22] D. Earl, B. von Holdt, STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method, Conserv. Genet. Res. 4 (2012) 359–361.

[23] M. Jakobsson, N.A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, Bioinformatics 23 (2007) 1801–1806.

[24] N.A. Rosenberg, distruct: a program for the graphical display of population structure, Mol. Ecol. Notes 4 (2004) 137–138.

[25] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[26] C. Phillips, A. Salas, J.J. Sanchez, M. Fondevila, A. Gomez-Tato, et al., Inferring ancestral origin using a single multiplex assay of autosomal ancestry informative marker SNPs, Forensic Sci Int Genet 1 (2007) 273–280.

[27] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet. 73 (2003) 1402-1422.

[28] L. Pagani, T. Kivisild, A. Tarekegn, R. Ekong, C. Plaster, et al., Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool, Am. J. Hum. Genet. 91 (2012) 83-96.

[29] M. Gallego Llorente, E.R. Jones, A. Eriksson, V. Siska V, K.W. Arthur, et al., Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent, Science 350 (2015) 820-822.

[30] J.K. Pickrell, N. Patterson, P.R. Loh, M. Lipson, B. Berger, et al., Ancient west Eurasian ancestry in southern and eastern Africa, Proc. Natl. Acad. Sci. U.S.A. 111 (2014) 2632-2637.

[31] K.K. Kidd, W.C. Speed, A.J. Pakstis, M.R. Furtado, R. Fang, A. et al., Progress toward an efficient panel of SNPs for ancestry inference, Forensic Sci. Int. Genet. 10 (2014) 23-32.

[32] M. Eduardoff, T.E. Gross, C. Santos, M. de la Puente, D. Ballard, et al., Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™, Forensic Sci. Int. Genet. 23 (2016) 178-189.

[33] V. Pereira, A. Freire-Aradas, D. Ballard, C. Børsting, V. Diez et al., Development and validation of the EUROFORGEN NAME (North African and Middle Eastern) ancestry panel, Forensic Sci. Int. Genet. 42 (2019) 260-267.

[34] R.E. Howes, A.P. Patil, F.B. Piel , O.A. Nyangiri, C.W. Kabaria, et al., The global distribution of the Duffy blood group, Nat. Commun. 2 (2011) 266.

[35] B.M. Henn, L.R. Botigué, S. Gravel, W. Wang, A. Brisbin, et al., Genomic ancestry of North Africans supports back-to-Africa migrations, PLoS Genet. 8 (2012) e1002397.

[36] M. Haber, D. Gauguier, S. Youhanna, N. Patterson, P. Moorjani, et al., Genome-wide diversity in the levant reveals recent structuring by culture, PLoS Genet. 9 (2013) e1003316.

[37] E. D'Atanasio, G. Iacovacci, R. Pistillo, M. Bonito, J.M. Dugoujon, et al., Rapidly mutating Y-STRs in rapidly expanding populations: Discrimination power of the Yfiler Plus multiplex in northern Africa, Forensic Sci. Int. Genet. 38 (2019) 185-194.

[38] G. Iacovacci, E. D'Atanasio, O. Marini, A. Coppa, D. Sellitto, et al., Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries, Forensic Sci. Int. Genet. 27 (2017) 123-131.

[39] A. Boattini, L. Castrì, S. Sarno, A. Useli, M. Cioffi, et al., mtDNA variation in East Africa unravels the history of Afro-Asiatic groups, Am J Phys Anthropol. 150 (2013) 375-385.

[40] J.T. Hefner, S.D. Ousley, Statistical classification methods for estimating ancestry using morphoscopic traits, J. Forensic Sci. 59 (2014) 883-890.

[41] O. Bulbul, L. Cherni, H. Khodjet-El-Khil, H. Rajeevan, K.K. Kidd, Evaluating a subset of ancestry informative SNPs for discriminating among Southwest Asian and circum-Mediterranean populations, Forensic Sci. Int. Genet. 23 (2016) 153-158.

[42] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, et al., Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359-366.

[43] C. Santos, C. Phillips, M. Fondevila, R. Daniel, R.A.H. van Oorschot, et al., Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region, Forensic Sci. Int. Genet. 20 (2016) 71-80.

[44] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, et al., Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set Forensic Sci. Int. Genet. 11 (2014) 13-25.

[45] C. Santos, M. Fondevila, D. Ballard, R. Banemann, A.M. Bento, et al., Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: Results of a collaborative EDNAP exercise, Forensic Sci. Int. Genet. 19 (2015) 56-67.

[46] J.J. Sanchez, P. Endicott, Developing multiplexed SNP assays with special reference to degraded DNA templates, Nat. Protoc. 1 (2006) 1370-1378.

[47] A. Mameli, G. Piras, G. Delogu, The successful recovery of low copy number and degraded DNA from bones exposed to seawater suitable for generating a DNA STR profile, J. Forensic Sci. 59 (2014) 470-473.

[48] A.C. Jäger, M.L. Alvarez, C.P. Davis, E. Guzmán, Y. Han, Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories, Forensic Sci. Int. Genet. 28 (2017) 52-70.

[49] M. Al-Asfi, D. McNevin, B. Mehta, D. Power, M.E. Gahan, et al., Assessment of the Precision ID Ancestry panel, Int. J. Legal Med. 132 (2018) 1581-1594.

[50] V. Pereira, H.S. Mogensen, C. Børsting, N. Morling, Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers, Forensic Sci. Int. Genet. 28 (2017) 138-145.

Figure captions

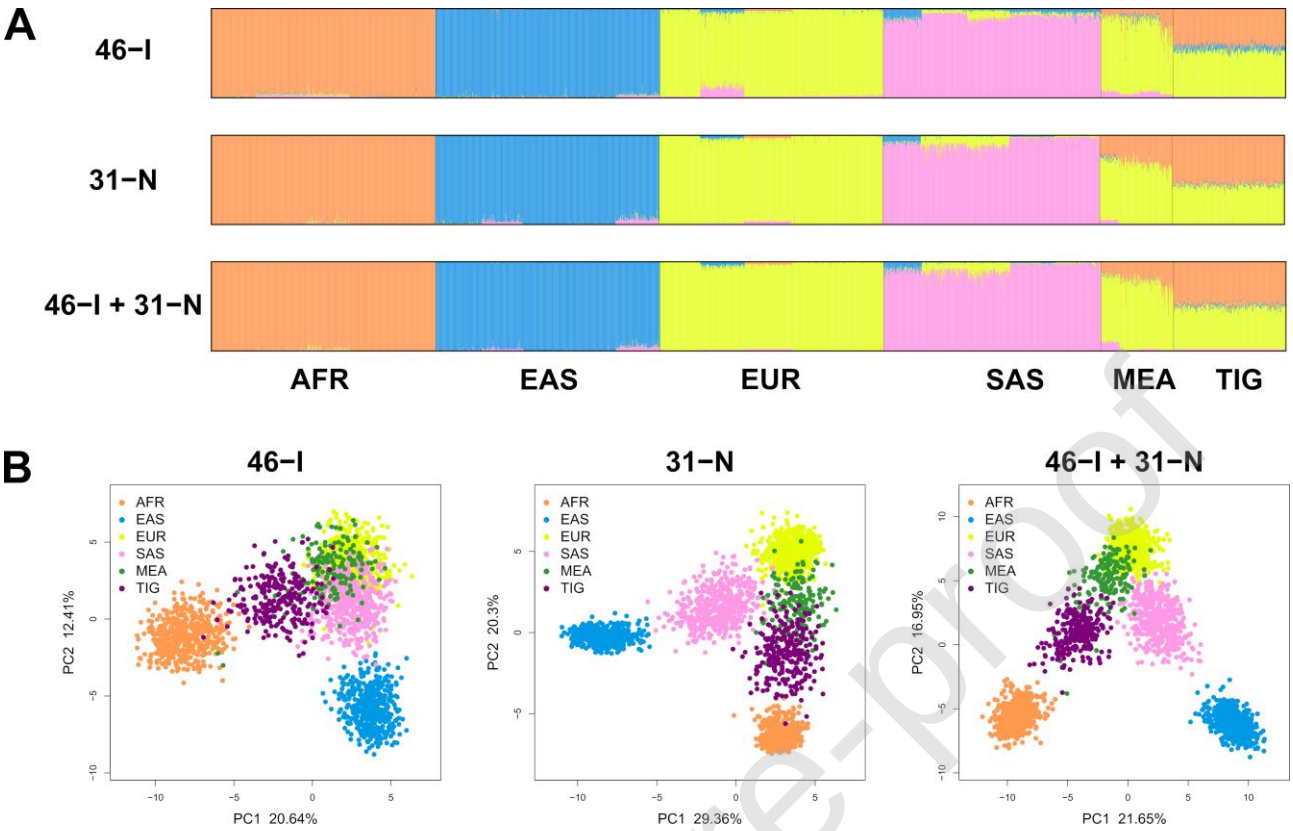

Figure 1. (A) STRUCTURE ancestry analysis of the Tigray study sample (TIG) and reference population groups (K:4, admixture LOCPRIOR ancestry model) using 46-I (top), 31-N (middle) and combined 46-I and 31-N AIM panels (bottom). Order of population samples within reference population groups is as specified in section 2.3. (B) Principal component analysis (PC1 vs PC2) of the Tigray study sample (TIG) and reference population groups. Left to right: 46-I; 31-N; combined 46-I and 31-N AIM panels.

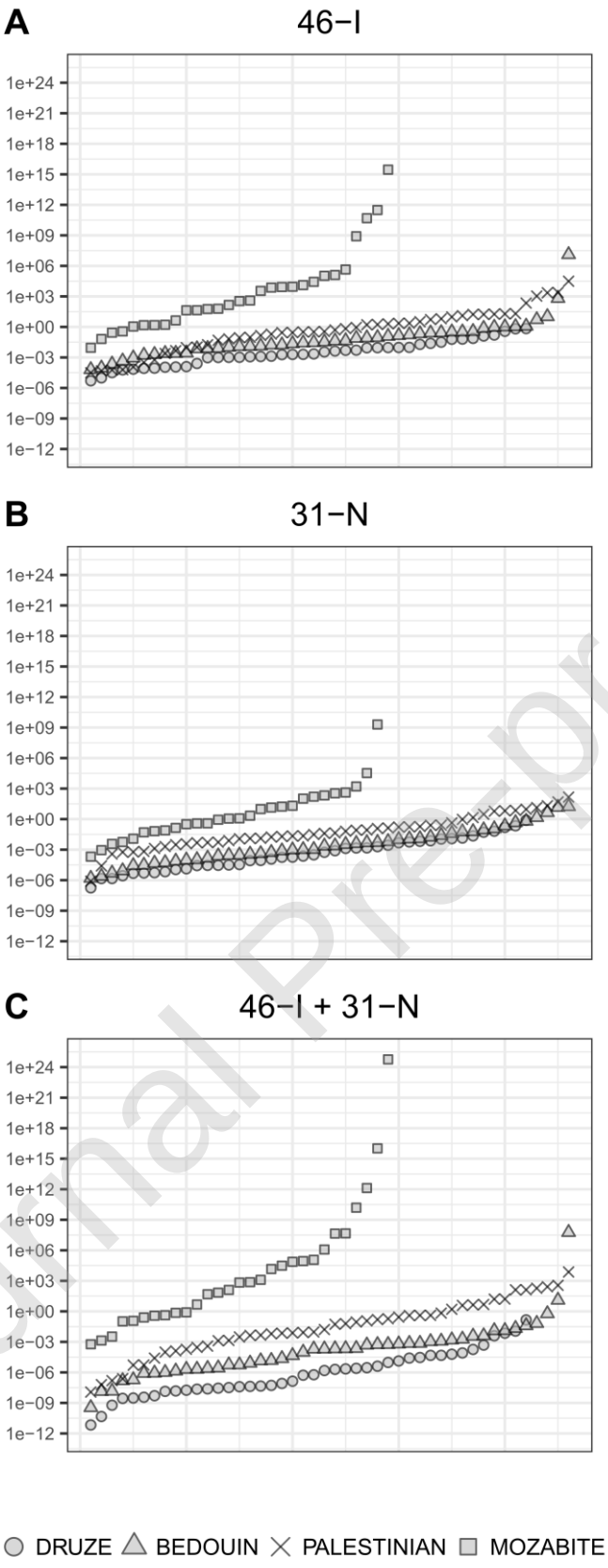

Figure 2. Plot of ranked pairwise assignment probabilities in MEA population subgroups using: (A) 46-I; (B) 31-N; (C) combined 46-I and 31-N AIM panels. The Y axis is a log scale of the LR of Tigray classification probability / MEA classification probability.

Table

46-I	AFR	EAS	EUR	SAS	MEA	TIG
AFR	**96.03 %**	0.00 %	0.00 %	0.00 %	0.00 %	3.97 %
EAS	0.00 %	**98.41 %**	0.00 %	1.59 %	0.00 %	0.00 %
EUR	0.00 %	0.00 %	**45.92 %**	13.52 %	40.16 %	0.40 %
SAS	0.00 %	0.00 %	0.00 %	**62.99 %**	33.54 %	3.48 %
MEA	0.00 %	0.00 %	0.00 %	0.00 %	**68.10 %**	31.90 %
TIG	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	**100.00 %**
31-N	**AFR**	**EAS**	**EUR**	**SAS**	**MEA**	**TIG**
AFR	**99.21 %**	0.00 %	0.00 %	0.00 %	0.00 %	0.79 %
EAS	0.00 %	**99.80 %**	0.00 %	0.20 %	0.00 %	0.00 %
EUR	0.00 %	0.00 %	**92.84 %**	1.19 %	5.96 %	0.00 %
SAS	0.00 %	0.00 %	0.00 %	**97.96 %**	1.43 %	0.61 %
MEA	0.00 %	0.00 %	0.00 %	0.00 %	**82.21 %**	17.79 %
TIG	0.79 %	0.00 %	0.00 %	0.40 %	1.19 %	**97.62 %**
46-I+31-N	**AFR**	**EAS**	**EUR**	**SAS**	**MEA**	**TIG**
AFR	**99.60 %**	0.00 %	0.00 %	0.00 %	0.00 %	0.40 %
EAS	0.00 %	**100.00 %**	0.00 %	0.00 %	0.00 %	0.00 %
EUR	0.00 %	0.00 %	**81.31 %**	0.80 %	17.89 %	0.00 %
SAS	0.00 %	0.00 %	0.00 %	**92.64 %**	6.54 %	0.82 %
MEA	0.00 %	0.00 %	0.00 %	0.00 %	**79.75 %**	20.25 %
TIG	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	**100.00 %**

Table 1 Comparison of classification success (shown in bold) estimated from Snipper cross-validation of the Tigray study sample (TIG) and reference population groups using 46-I (top), 31-N (middle) and combined 46-I and 31-N AIM panels (bottom).

	LR ≥10^1			LR≥10^2			LR≥10^3			LR≥10^4			LR≥10^5			LR≥10^6		
	N	C	W	N	C	W	N	C	W	N	C	W	N	C	W	N	C	W
46-I																		
AFR	.069	.911	.020[a]	.163	.833	.004[a]	.278	.720	.002[a]	.448	.552	.000	.653	.347	.000	.821	.179	.000
SAS	.444	.394	.162[b]	.759	.200	.041[c]	.939	.059	.002[c]	.992	.008	.000	1.0	.000	.000	1.0	.000	.000
MEA	.282	.509	.209[a]	.540	.319	.141[a]	.761	.129	.110[a]	.896	.043	.061[a]	.957	.006	.039[a]	.982	.000	.018[a]
TIG	.008	.992	.000	.028	.972	.000	.068	.932	.000	.139	.861	.000	.238	.762	.000	.425	.575	.000
31-N																		
AFR	.024	.976	.000	.079	.921	.000	.238	.762	.000	.448	.552	.000	.732	.268	.000	.966	.034	.000
SAS	.045	.947	.008[d]	.092	.906	.002[c]	.184	.816	.000	.341	.659	.000	.523	.477	.000	.689	.311	.000
MEA	.202	.694	.104[a]	.448	.497	.055[a]	.656	.326	.018[a]	.804	.184	.012[a]	.908	.086	.006[a]	.975	.019	.006[a]
TIG	.075	.909	.016[e]	.206	.790	.004[f]	.294	.702	.004[f]	.472	.524	.004[f]	.663	.337	.000	.837	.163	.000
46-I+31-N																		
AFR	.002	.996	.002[a]	.016	.982	.002[a]	.032	.968	.000	.062	.938	.000	.137	.863	.000	.224	.776	.000
SAS	.075	.892	.033[c]	.127	.855	.018[c]	.201	.797	.002[c]	.315	.683	.002[c]	.421	.579	.000	.546	.454	.000
MEA	.135	.693	.172[a]	.227	.632	.141[a]	.399	.509	.092[a]	.521	.399	.080[a]	.632	.313	.055[a]	.748	.203	.049[a]
TIG	.000	1	.000	.000	1	.000	.000	1	.000	.028	.972	.000	.036	.964	.000	.075	.925	.000

[a] All classified as Tigray
[b] 96.6% classified as MEA; 0.4% classified as Tigray
[c] All classified as MEA
[d] 75.0% classified as MEA; 25.0% classified as Tigray
[e] 50.0% classified as AFR; 25.0% classified as SAS; 25.0% classified as MEA
[f] All classified as AFR

Table 2. Classification success (cross validation) of the Tigray study sample (TIG) and reference population groups using 46-I (top), 31-N (middle) and combined 46-I and 31-N AIM panels (bottom) when applying different LR thresholds. N indicates samples with assignment probability below LR threshold, which could not be classified. C indicates samples with assignment probability above LR threshold, which were correctly classified. W indicates samples with assignment probability above LR threshold, which were wrongly classified.