Forensic Science International: Genetics 48 (2020) 102342

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Predicting haplogroups using a versatile machine learning program (PredYMaLe) on a new mutationally balanced 32 Y-STR multiplex (CombYplex): Unlocking the full potential of the human STR mutation rate spectrum to estimate forensic parameters

Caroline Bouakaze[a,1,4], Franklin Delehelle[a,p,4], Nancy Saenz-Oyhéréguy[a,4], Andreia Moreira[a,4], Stéphanie Schiavinato[a], Myriam Croze[a,2], Solène Delon[a], Cesar Fortes-Lima[a,3], Morgane Gibert[a], Louis Bujan[b], Eric Huyghe[b], Gil Bellis[c], Rosario Calderon[d], Candela Lucia Hernández[d], Efren Avendaño-Tamayo[e], Gabriel Bedoya[f], Antonio Salas[g], Stéphane Mazières[h], Jacques Charioni[h,i], Florence Migot-Nabias[j], Andres Ruiz-Linares[h,k], Jean-Michel Dugoujon[a], Catherine Thèves[a], Catherine Mollereau-Manaute[a], Camille Noûs[l], Nicolas Poulet[m], Turi King[n], Maria Eugenia D'Amato[o], Patricia Balaresque[a,*]

[a] Laboratoire d´Anthropologie Moléculaire et Imagerie de Synthèse (AMIS), UMR5288 - CNRS & Université Toulouse III, 37 allées Jules Guesde, 31073 Toulouse Cedex 3, France
[b] Equipe d'acceuil EA3694, Hôpital Paule de Viguier, 330 Avenue de Grande Bretagne, TSA 70034, 31059 Toulouse Cedex 9, France
[c] INED Institut National d'Etudes Démographiques, 133 Boulevard Davout, 75980 Paris cedex 20, France
[d] Department of Biodiversity, Ecology and Evolution, Faculty of Biology, Complutense University. 28040 Madrid, Spain
[e] Grupo de Ciencias Básicas Aplicadas del Tecnológico de Antioquia, Tecnológico de Antioquia, Institución Universitaria, Medellín 050034, Colombia
[f] GENMOL (Genética Molecular), Instituto de Biología, Universidad de Antioquia Medellín Colombia, Colombia
[g] Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela, GenPoB Research Group, Instituto de Investigaciones, Sanitarias (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain
[h] Aix Marseille Univ, CNRS, EFS, ADES, Marseille, France
[i] Etablissement Français du Sang PACA Corse, Marseille, France
[j] Université de Paris, MERIT, IRD, F-75006, Paris, France
[k] Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China
[l] Laboratoire Cogitamous, CNRS & Université Toulouse III, 31000 Toulouse, France
[m] Pôle écohydraulique AFB-IMT, allée du Pr Camille Soula, 31400 Toulouse, France
[n] Department of Genetics, University of Leicester, Leicester, United Kingdom
[o] Forensic DNA Laboratory, Department of Biotechnology, Faculty of Natural Sciences, University of Western Cape, Cape Town, South Africa
[p] REVA Unit, UMR 5505 - CNRS & Université de Toulouse, Institut de Recherche en Informatique de Toulouse, 31400 Toulouse, France

ARTICLE INFO

Keywords:
Y-STR
Machine learning
Assignation accuracy and haplogroup prediction (Hg prediction)
Incremental mutation rates

ABSTRACT

We developed a new mutationally well-balanced 32 Y-STR multiplex (**CombYplex**) together with a machine learning (ML) program **PredYMaL**e to assess the impact of STR mutability on haplogourp prediction, while respecting forensic community criteria (high DC/HD). We designed CombYplex around two sub-panels M1 and M2 characterized by average and high-mutation STR panels. Using these two sub-panels, we tested how our program PredYmale reacts to mutability when considering basal branches and, moving down, terminal branches. We tested first the discrimination capacity of CombYplex on 996 human samples using various forensic and statistical parameters and showed that its resolution is sufficient to separate haplogroup classes. In parallel,

* Corresponding author at: CNRS & University of Toulouse III (UMR 5288), Laboratoire d'Anthropobiologie Moléculaire et Imagerie de Synthèse 37, allées Jules Guesde, 31073 Toulouse France.
E-mail address: patricia.balaresque@univ-tlse3.fr (P. Balaresque).
[1] Present address: Institut National de Police Scientifique, Laboratoire de Police Scientifique de Lyon, 31 Avenue Franklin Roosevelt, 69134 Ecully Cedex, France.
[2] Present address: Division of EcoScience, Ewha Womans University, Seoul.
[3] Present address: Sub-department of Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Norbyvagen 18C, SE-752 36 Uppsala, Sweden.
[4] These authors contributed equally to the work.

https://doi.org/10.1016/j.fsigen.2020.102342
Received 19 December 2019; Received in revised form 10 June 2020; Accepted 11 June 2020
Available online 29 June 2020

Pred**YMaL**e was designed and used to test whether a ML approach can predict haplogroup classes from Y-STR profiles. Applied to our kit, SVM and Random Forest classifiers perform very well (average 97 %), better than Neural Network (average 91 %) and Bayesian methods (< 90 %). We observe heterogeneity in haplogroup assignation accuracy among classes, with most haplogroups having high prediction scores (99–100 %) and two (E1b1b and G) having lower scores (67 %). The small sample sizes of these classes explain the high tendency to misclassify the Y-profiles of these haplogroups; results were measurably improved as soon as more training data were added. We provide evidence that our ML approach is a robust method to accurately predict haplogroups when it is combined with a sufficient number of markers, well-balanced mutation rate Y-STR panels, and large ML training sets. Further research on confounding factors (such as CNV-STR or gene conversion) and ideal STR panels in regard to the branches analysed can be developed to help classifiers further optimize prediction scores.

1. Introduction

The Y-chromosome has been extensively used to identify male individuals in forensic communities [1] and to reconstruct the family and evolutionary history of paternal lineages in geneticists [2] and genealogists communities [3]. Questions related to the latter research topic are diverse and to address them on the Y-chromosome which is characterized by a low genetic diversity in human species, it can be advantageous to capture not only long-term but also short-term genomic information. It would help to optimally study not only the biogeographic informativeness of Y-haplotypes [4] but also Y-specific migration paths and social structure, surname diffusion, paternal history of royal family members, and paternal lineage diffusion [3,5–16]. But whatever the objectives and the technics used, the key problem remains the same: finding a good equilibrium between the resolution needs (markers and mutation rates) and the costs involved. Retrieving long-term genomic information has classically been completed using Y-SNaPshot analyses (for a review on Y-SNP typing see [17]), and very recently by using massively parallel sequencing [18]. Retrieving short-term genomic information has mainly consisted in Y-STR profiling in accessing the maximum of STRs variants and polymorphism either by (i) designing Y-STR multiplexes including highly mutable markers to better discriminate closely related individuals [19,20] or (ii) by sequencing and extracting length-based Y-STR polymorphism STR loci from Next Generation Sequencing technologies as implemented in STRait Razor [21] to get rid of the excess of variants. To access short and long-term information while diminishing costs, some studies have chosen to generate high resolution Y-STR data and to use previously developed tools to predict haplogroup classes [22–25]. Among these methods, Neural Network-based models (Felix Immanuel website[55] http://www.y-str.org/)) and Bayesian-allele frequency approaches [26] were the first to have been developed, although ML approaches have been also tested [27] (see Supplementary data 1 for a review). However, the large bias in haplogroup prediction error [25] has urged the development of ready-to-use predictive tools, while considering more carefully the impact of STR mutation rates. The human Y-STR mutation rate spectrum is wide with a 1000 to 10,000-fold of magnitude. Although this represents a powerful source of variation for designing tools in forensic genetics (from molecular to computational-based types), it is currently poorly explored.

Table 1
CombYplex M1 (a) et M2 (b): markers, molecular structures, primers and amplification conditions. *Dyes: Blue: FAM; Green: VIC; Yellow: NED; Red: PET. Nomenclature is given according to the following papers: [52] [53]; [54] and the STRidER Reference database: https://strider.online/).

a) M1

			Primer Forward				Primer Reverse				Litterature		Observed	Male C1+
M1 Markers	Mutation rate	Repeat structure	Dye/Name	Sequence (5'-3')	Tm (°C)	[µM]*	Name	Sequence (5'-3')	Tm (°C)	[µM]*	Allele Range	Expected size range (bp)	Allele Range	Allele
DYS485	4,04E-04	$(TTA)_n$	FAM_DYS485_F3	catataacaaaattgaatgtgtactcc	57,3	0,5	DYS485_R2	agcctgggtgacaagagttatac	58,7	0,5	11-21	109-139	11-20	16
DYS588	3,92E-04	$(GCATT)_n$	FAM_DYS588_F	gaatgcagaaccctcaagga	60,2	0,18	DYS588_R	agcctgggtgacagaaacac	60,2	0,18	9-16	142-170	10-18	12
DYS502	3,85E-04	$(AAT)_4(TGC)_3(CAT)_n$	FAM_DYS502_F	cagcaagccaccataccata	59,6	0,25	DYS502_R	tgtgcttttggagtttggag	58,9	0,25	6-9	205-214	6-10	8
DYS461 / YGATAA7.2	9,89E-04	$(TAGA)_n(CAGA)1$	FAM_DYS461_F	aatacataataaatgataggcagagga	57,9	0,6	DYS461_R	gagagctgaataagttgtatcaggtaa	58,6	0,6	8-13	249-269	7-14	11
DYS638	1,04E-03	$(TTTA)_n$	FAM_DYS638_F	ttctaatttcagtgctttcaattttc	59,3	1,2	DYS638_R	aggtgggtcatgaggtcagt	59,4	1,2	8-13	303-323	7-13	11
DYS643	1,5E-03	$(CTTTT)_n$	VIC_DYS643_F	aagccatgcctggttaaactac	59,6	0,1	DYS643_R	accaaacacccacttcc	60,5	0,1	8-13	132-159	7-16	10
DYS587	2,62E-03	$(CAATA)_n[(CAGTA)(CAATA)]_3$	VIC_DYS587_F2	cttctttggaaagtagcatttcat	58,1	0,8	DYS587_R2q	aaagtctgacaatgagaagggtttctaagttcagg	68,9	0,8	8-16	191-222	8-16	11
DYS575	3,91E-04	$(AAAT)_n$	VIC_DYS575_F2	cagaggttccagtaagcttagatca	60,3	0,3	DYS575_R2	cattgatgggctttaggttga	59,9	0,3	8-12	260-276	8-12	10
DYS578	9,95E-04	$(AAAT)_n$	VIC_DYS578_F	gaggcggaacttttcagtgag	60	0,5	VIC_DYS578_R2	cagaagtccctgtgtttcaa	60,1	0,5	7-10	305-317	6-11	9
DYS632	3,97E-04	$(CATT)_n$	NED_DYS632_F	cacagtttcagtcttgcatgg	59,3	0,09	DYS632_R	tctgggcaacagaaggagac	60,4	0,09	8-10	106-114	7-11	9
DYS508	3,03E-03	$(TATC)_n$	NED_DYS508_F	acaatggcaatcccaaattc	59,6	0,4	DYS508_R	gaacaaataaggtgggatggat	59,1	0,4	8-15	165-193	8-15	11
DYS640	3,98E-04	$(AAAT)_n$	NED_DYS640_F	ggaaaaaccatgagatcctgtc	59,8	0,2	DYS640_R	aagcccgttcatatttttaaagac	57,9	0,2	9-13	252-268	7-13	11
DYS511	1,52E-03	$(GATA)_n$	NED_DYS511_F	tggggtggatgtgtaggtaga	60,2	0,3	DYS511_R	tctggttgtgccttagatttga	59,7	0,3	9-14	307-327	7-14	10
DYS577	4,11E-04	$(ATTC)_n$	PET_DYS577_F	ttttttctacgtgtgtatccactaacc	59,8	0,15	DYS577_R	gtgtccccagccctgtta	59,5	0,15	8-11	100-112	6-12	9
DYS556	1,59E-03	$(AATA)_n$	PET_DYS556_F	tcaccaatgacattttacagca	59,1	0,6	DYS556_R	ttggttagtgtaatgcatccag	57,7	0,6	8-12	156-172	8-13	11
DYS517	3,21E-03	$(AAAG)_nN_{13}(AAAG)_3$	PET_DYS517_F2	aactgaccagcaaaaatgttaaa	57,9	0,5	DYS517_R2	tgtctgagacctacaagattgc	57,1	0,5	10-18	213-245	9-18	13
DYS565	2,09E-03	$(ATAA)_n$	PET_DYS565_F2	ccaggaagcagtgttgcat	59,8	0,3	DYS565_R2	gcagttctctgcctgtatgg	58,5	0,3	9-14	280-300	9-14	12
DYS538	3,94E-04	$(GATA)_n$	PET_DYS538_F	ttggggaaaacagatggtgt	60,2	1,7	DYS538_R	ccaaatacccatcataggaagaa	59,2	1,7	9-13	339-355	8-13	10

b) M2

			Primer Forward				Primer Reverse				Acc. To litterature		Observed	Male con
M2 Markers	Mutation rate	Repeat structure	Name	Sequence (5'-3')	Tm (°C)	[µM]*	Name	Sequence (5'-3')	Tm (°C)	[µM]*	Allele Range	Expected size range (bp)	Allele Range	Allele
SRY			FAM_SRY_F2	gcgaaactcagagatcagcaag	60,1	0,08	SRY_R1	tgtgcctcctggaagaatgg	61,9	0,08				
UTY			FAM_UTXUTY_F1	cagtgttaccagccttaacag	53,7	0,2	UTY_R	ggcaggtctacttttgtagag	52	0,1				
UTX							UTX_R	tctgtggactaggtttgtggt	55,6	0,11				
Y-GATA-A10	3,32E-03	(TCCA)2 (TATC)n	FAM_YGATAA10_F	cctgccatctctatttatcttgcatata	61,9	0,26	YGATAA10_R	ataaatggagatagtgggtggatt	59,1	0,26	9-16	150-178	9-16	13
DYS570	1,24E-02	$(TTTC)_n$	FAM_DYS570_F2	tgtgacatcaaggttatgaaacg	59,9	0,29	DYS570_R2	ggtgaaaattattcagcatagtcaag	59,5	0,29	14-24	214-254	12-24	18
DYS549	4,55E-03	$(GATA)_n$	FAM_DYS549_F	gaaaagaaaagtgtaagccaaacc	59,6	0,95	DYS549_R	tttggtggcataagtggtaatg	59,8	0,95	9-15	193-317	8-16	12
DYS460	6,22E-03	(TATC)n ou (ATAG)n	VIC_DYS460_F	atcctctgcctatcatttattatgtat	57,1	0,4	DYS460_R	gaatacagaggaatctgacacc	59,0	0,4	8-13	99-119	7-13	12
DYS442	9,78E-03	(TATC)2 (TGTC)3 (TATC)n	VIC_DYS442_F2	tgcaaaatcacggaaccaa	61	0,15	DYS442_R2	caagccactgcaaatgtca	59,4	0,15	9-16	173-201	8-16	12
DYS510	5,99E-03	$(GATA)_3N_{12}(GATA)_nN_{13}(GGAT)_4N_9(GATA)_3$	VIC_DYS510_F	ttttttcctcccttaccacaga	58,7	0,48	DYS510_R	tctggagaagacagaacttgtca	59,1	0,48	9-15	245-269	8-15	11
DYS541	3,92E-03	$(TATC)_n(TTC)_1(TATC)_3$	VIC_DYS541_F	catcattaattctatctgttcatccat	58,8	0,45	DYS541_R	tggataaagaacacctttaagaagc	59,3	0,45	10-15	310-330	6-15	12
DYS576	1,43E-02	$(AAAG)_n$	NED_DYS576_F	ccaagcaacatagcaagacct	59,4	0,13	DYS576_R	aagcgtatttgtctttggctttt	59,4	0,13	13-22	108-144	13-25	19
DYS513	6,09E-03	(TATC)n	NED_DYS513_F2	tgttgtaaaaatgactactgtggtatg	58,6	0,22	DYS513_R2	ccacatcagcactattacttaactca	58,9	0,22	9-15	294-318	9-15	12
DYS458	8,36E-03	$(GAAA)_n$	NED_DYS458_F	tgggtggtggaggttactgt	60,3	0,12	DYS458_R	cccaaagttctggcattacaa	60,0	0,12	11-24	183-235	11-24	18
DYS481	4,97E-03	$(CTT)_n$	PET_DYS481_F	aggaatgtggctaacgctgt	59,8	0,2	DYS481_R	accagaaggttgcaagactca	59,9	0,2	18-32	109-151	18-32	22
DYS612	1,45E-02	$(CCT)n(CTT)_1(TCT)_n(CCT)_1(TCT)_n$	PET_DYS612_F	cccccatgccagtaagaata	59,8	1,25	DYS612_R	tgagggaaggcaaaagaaaa	59,8	1,25	19-31	186-222	16-31	26
DYS444	5,45E-03	$(ATAG)_n$	PET_DYS444_F	catagaatgaaaggtgtgaacca	59,0	0,45	DYS444_R	tgccattcaaactcacgttg	60,7	0,45	9-16	264-282	8-16	12
DYS533	5,01E-03	$(ATCT)_n$	PET_DYS533_F	attcatctaacatctttgtcatctacc	58,2	0,95	DYS533_R	ttaacttgccttttgcatcc	59,2	0,95	9-14	334-354	8-14	12

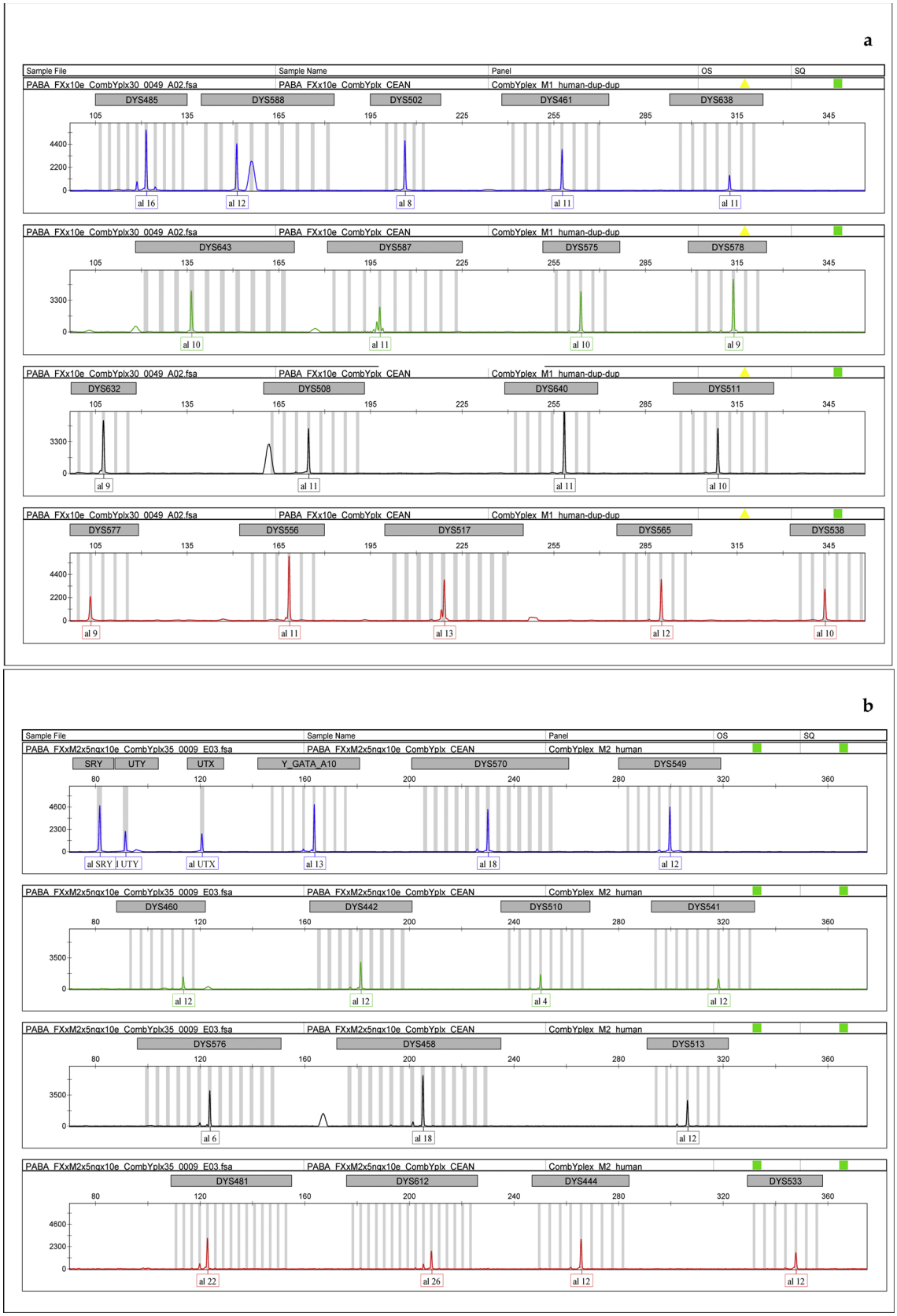

Fig. 1. a CombYplex M1 profile of male control (line 1: blue dye, line 2: green dye, line 3: yellow dye and line 4: red dye); ;two artifacts can occasionally be observed on the M1 electrophergram: in the polymorphism zone of the DYS588 locus (blue dye, line 1) and in the polymorphism zone of the DY508 locus (yellow dye, line 3), as shown here. b CombYplex M2 profile (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

In this paper, we assessed whether a well-balanced STR multiplex, associated with machine learning (ML) approaches can efficiently predict haplogroups, while still providing the high Discrimination Capacity (DC) index required in forensic genetics. We designed a 32 Y-STR-typing kit "CombYplex" around two panels of STRs (M1 and M2) mutating at various rates (selected from 3.85×10^{-04} to 1.45×10^{-02} mutation/locus/generation) to test the impact STR mutability on Hg prediction. Then, we designed "PredYMaLe" (Predicting Y-lineages using ML models), a program that includes various ML approaches to predict Y-haplogroup classes from a set of Y-STR markers.

First, for the CombYplex design, we assembled and typed a panel of 996 male individuals from three continents (Africa, Europe, and South America) available in our collections; we tested the discrimination power of CombYplex by computing both classic forensic and statistics parameters, e.g. Haplotype Diversity (HD), Discrimination Capacity (DC), Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Second, we tested whether the ML approaches implemented in the PredYMaLe program could efficiently predict the haplogroup lineages. We used a sub-panel of 503 chromosomes on four panels of STRs (the full 32-STR CombYPlex, the Y-filer, and the CombYplex_M1 and M2 only) for which haplogroup data were available. We evaluated the impact of STR-assembly on assignment accuracy, by considering first seven main Hg classes (considered as basal Y-tree branches) and then 12 detailed Hg classes (including E-subdivided terminal-like branches, considered as terminal Y-tree branches) to test the impact of Hg subdivision. Although not all haplogroup lineages could be tested in this article, the wide range of coalescence ages associated with the Hgs tested here (from 5 KYA for R1b1a1a2a1a2a1b1a1-M167 to 45 KYA for Hgs I-170 or J-M304 [28]) should give a good preview of the prediction scores for comparable clades existing in the Y-tree and of the associated divergence between the relative haplotypes. Our results showed that: (i) the full and well-balanced STR profiles (CombYplex or Y-filer) give the best prediction scores using the SVM and Random Forest classifiers, whereas Neural Network or Bayesian approaches, the most currently used methods for Hg prediction, fall short; (ii) PredYMaLe and CombYplex can predict haplogroup classes with an average assignment accuracy of 97 % using Support Vector Machines (SVM) and Random forest classifiers, but classifiers are sensitive to STR panel composition, STR number, and training dataset size. These results can be used in the future to design well-balanced STR panels (extracted from whole genome sequencing data) with a high number of markers, featuring high discrimination capacity and accurate predictions of haplogroup lineages with appropriate ML methods.

2. Materials and methods

2.1. Database of Y-STR characteristics

For 220 Y-STRs, we collected information on Y-STR molecular characteristics, mutation rates, and polymorphisms for humans. This database is available in Supplementary data 2.

2.2. Selecting Y-STRs and constructing multiplexes: CombYplex M1 and M2

We selected a set of 32 Y-STRs from our database to construct two complementary multiplexes: one with average-mutating markers (M1) and one with high-mutating markers (M2). These markers were chosen to be polymorphic and to have the simplest molecular structure as possible (see Table 1). M1 includes the following **18 Y-STRs:** DYS485, DYS588, DYS502, DYS461, DYS638, DYS643, DYS587, DYS575, DYS578, DYS632, DYS508, DYS640, DYS511, DYS577, DYS556, DYS517, DYS565 and DYS538. Their mutation rates range from 3.85×10^{-04} to 3.21×10^{-03} mutation/locus/generation. Their molecular structures, primers and conditions are detailed in Table 1 and an example of a M1_CombYplex profile is proposed in Fig. 1a.

M2 includes a **sex-testing assay** (derived from [29]) and the following **14 Y-STRs:** Y-GATA-A10, DYS570, DYS549, DYS460, DYS442, DYS510, DYS541, DYS576, DYS513, DYS458, DYS481, DYS612, DYS444, and DYS533. These markers were chosen to be highly polymorphic and to have the simplest molecular structure as possible; however, when STR with pure molecular structures could not be selected, we compromised between a simple structure and high STR mutation rate (e.g. DYS612 and DYS533). Their mutation rates range from 3.32×10^{-03} to 1.45×10^{-02} mutation/locus/generation. Their molecular structures, primers and conditions are detailed in Table 1 and an example of a M2_CombYplex profile is proposed in Fig. 1b.

The multiplexes were designed using the shortest amplicons as possible, with a maximum size of 356 bp for DYS533 (M2). They were designed to be used independently or combined, according to the degree of resolution required. The cost of a full CombYplex reaction (32 Y-linked STRs + three sex-typing markers) is only 4.3 € (in France and based on public prices for all the reagents), and one of the assets of this multiplex in regard to its resolution. This tool was developed on an ABI Prism 3730 DNA Analyzer 48-capillary array system (Life Technologies), due to contextual and logistic reasons, but its design strategy can be transposed to Next Generation Sequencing systems.

2.3. Population samples

Samples, available from collaborations and internal collections, were obtained from healthy human volunteers with consent forms. They were extracted from various substrates including saliva and whole blood. A total of **996** samples were used in this study (plus one male control, one female control and 1 *AZFc* deleted Y-chromosome male to control for deletion) and genotyped with the CombYplex kit. This dataset includes **six native West African** populations: three populations from Benin: 59 Bariba (Parakou region), 47 Yoruba (Ketou region), and 68 Fon (Cotonou and Ouidah regions), two populations from Ivory Coast: 47 Ahizi (Nigui-Saff region) and 37 Yacouba (Danané region), and one population from Mali: 13 Bwa (Segou region), **three native South African** populations (97 Xhosa, 90 Zulu, and 33 Tswana), **three admixed African-descendant** populations (52 French Guyana and Suriname Noir Marron, 56 Ketou-Yoruba, 35 Brazil - Rio de Janeiro, 20 Colombia), **one native American** population (6 Palikour), and **11 European populations** (30 Spain Barcelona, 19 Spain Galicia, 24 Spain Granada, 25 Spain Huelva, 46 France Loire-Atlantique, 50 France Vendée, 21 France Sarthe, 30 France Maine and Loire, 81 France Ariège-Pyrénées, and 57 France Haute-Garonne).

2.4. Analysis of grouped samples

DNA samples were grouped based on two criteria: geographic ("GEO" sample) and phylogenetic ("HAPLO" sample).

In the **"GEO sample"** the geographic location of individuals is based on two generations of residence. All the 996 male individuals are included in this sample, to evaluate forensic parameters and control the discrimination power of the sample.

The **"HAPLO sample"**, a haplogroup-based sample, is a subset of the GEO sample, used to evaluate haplogroup predictions with ML methods. It includes 503 individuals for whom Y-SNP haplogroup and Y-filer profiles were also available. Since many studies have already tested the added value of PPY23 and Y-filer plus, we did not type these additional products due to the costs involved. We used Y-filer, a mutationally relatively balanced Y-STR kit for which we already had data in our database. We removed DYS385a/b and analysed only 15-STRs from the Y-filer panel since we have found evidence of conversion and outlier alleles in previous work [30]. Eight main Hgs were first considered to calculate forensics parameters (E1a, E1E1a, E1b1b, F, G, I, J, R1b1a1a2). However, haplogroup classes represented by a very low number of individuals were not included in the subsequent ML analyses (7 individuals in Hg F-M213*/F-M89*, and 2 individuals in Hg

E1b1b1b1a-M81 included in E1b1b for 12-classes analyses): 7-Main and 12 detailed classes were considered in ML-analyses. Hg G and E1b1b had the lowest sample sizes, with 9 and 12 individuals respectively; we kept these Hgs in the 7-main classes to test the potential impact of a low number of individuals. The results for these two Hgs will have to be considered carefully due to the effect of small training sets reported in the ML literature.

First, the HAPLO sample was used to test the efficiency of CombYplex using classic forensics parameters (Haplotype diversity, Gene Diversity, Discrimination Capacity and Match Probability) and to test whether CombYplex could discriminate haplogroup classes using discriminant analyses (PCA). Second, it was used to test whether haplogroups could be predicted from the full 32 Y-STR, from the M1 and M2 only (lower number of markers and contrasted mutation rate), or from the Y-filer Y-STR profiles using an ML program. The HAPLO subsample includes six European populations (n = 201; 26 Spain Barcelona, 14 Spain Galicia, 19 Spain Granada, 22 Spain Huelva, 64 France Pyrenees, 56 France Haute-Garonne), five native African populations (n = 191; 52 Benin Parakou Bariba, 60 Benin Cotonou Fon, 36 Ivory-Coast Ahizi, 30 Ivory-Cost Yacouba, 13 Mali Bwa), and five admixed African-descendant populations (n = 111; 8 French Guyana Aluku, 50 Ketou-Yoruba, 27 Noir-Marron, 12 Brazil-Rio de Janeiro, and 14 Colombia).

2.5. DNA extraction

The DNA extraction method was chosen according to the sample substrate. DNA was extracted from: (i) **whole blood**, using the QiaAmp DNA Blood mini-kit (Qiagen), (ii) **serum**, using the i-genomic DNA Blood mini-kit (Euromedex), and (iii) **saliva**, using the OG-300 Oragene DNA Self-Collection Kit (DNA Genotek) following the respective manufacturer's instructions. The quantity and quality of DNA extracted was estimated using a NanoDrop Spectrophotometer 2000C (LabTech).

2.6. PCR amplification conditions: CombYplex M1 and M2

CombYplex M1 and M2 were amplified in a reaction volume of 12.5 µL with 6.25 µL of QIAGEN Multiplex PCR Plus Kit (Qiagen), 1.25 µL Q-solution (Qiagen), 4 µL of the CombYplex M1 or M2 primer mix (see Table 1a and b for concentrations) and 5 ng of DNA template (limit of detection tested: 2–2.5 ng). Thermal cycling was conducted on a GeneAmp PCR System 2700 (Applied Biosystems) using the following conditions: 95 °C for 5 min; 30 cycles: 95 °C for 30 s, 62 °C for 90 s, 72 °C for 30 s; 68 °C for 30 min, 10 °C hold. To ensure that the resultant PCR amplicons were A-tailed (thereby avoiding the split peak phenomenon when visualized), a 2 µL reaction mix incorporating 0.125 U Taq polymerase (Fisher BioReagents) and a 1X PCR buffer system was added to 5 µL of PCR products prior to incubation for a further 45 min at 72 °C.

2.7. Detection and analysis of PCR products

Diluted A-tailed PCR products were mixed to 8.8 µL Hi-Di™ formamide (Applied Biosystems) and 0.2 µL GS600LIZ Size Standard (Applied Biosystems). After incubation at 95 °C for 5 min, the samples were loaded onto an ABI Prism 3730 and a 3500 DNA Analyzer 48-capillary array system (Applied Biosystems). The G5 matrix filter DS-33 was used to detect the five dyes 6-FAM™ (blue), VIC™ (green), NED™ (yellow), PET™ (red) and LIZ™ (orange). The samples were injected for 15 s at 1600 V. Separations were performed at 15,000 V for 30 min with a run temperature of 63 °C using the POP™-7 Polymer for 3730 (Applied Biosystems), run through a 50 cm capillary array (Applied Biosystems). Following data collection, samples were analysed with GeneMapper v4.0 (Applied Biosystems).

2.8. SNP genotyping methods

The populations Fon, Bariba, Yoruba, Ahizi and Yacouba were genotyped using 96 Y-SNPs on a BioMark™ HD system (Fluidigm, USA) as described in [31]. Y-SNP haplogroups were assigned according to ISOGG Y-DNA Haplogroup Tree 2015 (http://www.isogg.org/tree/) updated in April 2015. All other populations were genotyped using classic SNaPshot technics using a hierarchical approach. In total, 14 haplogroup lineages were detected and grouped in 7-Main and 12-Detailed classes for ML-analyses (Supp Data 6); they were used in combination with 4 sets Y-STR profiles (full CombYplex, Y-filer, CombYplex_M1 and CombYplex_M2) in PredYmale program to calculate how accurately a haplogroup lineage could be assigned.

2.9. Sequencing

Each locus was sequenced for the Male 1 control sample. Primers for sequencing are reported in Supplementary data 3. Each PCR product was sequenced in two reactions using forward and reverse PCR primers. The sequence reaction was performed with the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). Sequence products were run on an ABI 3730 DNA Analyzer (Applied Biosystems). Sequences were analysed using Sequence Scanner Software v1.0 (Applied Biosystems) and BioEdit Sequence Alignment Editor version 7.2.5.

2.10. Forensic parameters and discrimination indexes: population grouping and comparative analyses

For GEO and HAPLO grouped samples, the following diversity parameters were calculated: haplotype diversity (HD) was calculated using $HD = \frac{n}{n-1}(1 - \sum xi^2)$, where n = the number of haplotypes in the dataset and xi = the frequency of the ith haplotype [32], gene diversity (GD) was calculated analogously to HD where n and xi denote the total number of samples and the relative frequency of the *ith* allele [33], discrimination capacity (DC) was defined as the ratio between the number of different haplotypes and the total number of haplotypes: $DC = \frac{Ndiff}{N}$ where *Ndiff* was the number of different haplotypes, N was the sample size, and match probability (MP) was calculated as the sum of squared haplotype frequencies $MP = \sum pi^2$ where pi was the frequency of the *ith* haplotype. Haplotype number (n) and haplotype frequencies were estimated using Arlequin v 3.5.2.2 [34]. We represented the distribution of Y-STR haplotypes according to their haplogroup class by PCA: analyses were carried out using R software v 2.15.3 [35] and ade4 packages [36]. In addition, we performed Linear Discriminant Analyse (LDA) using the MASS package [37] to estimate the proportion of haplogroups that were classed to a satisfactory precision. For LDA analysis, about 75 % of individuals per haplogroup class taken randomly are used to train the model, while the remaining 25 % is used to validate the trained classifier by testing its efficiency. This procedure was run 100 times. Given that the ML training and the split between training and validation datasets are heuristic, all the scores are averaged over 100 trials. We tested haplogroup prediction on the most represented haplogroup classes in our sample: E1a, E1a1a, E1a1b, G, I, J, and R1a1a1 (and on the collapsed root E group, including E1a, E1a1a and E1a1b).

2.11. Predicting haplogroups using machine-learning approaches: PredYMaLe

Haplogroups are usually defined by a given set of SNPs, but here, we explore whether they could also be recovered from the phylogenetic information contained in the Y-STR haplotype profiles alone. Different methods have been developed to predict haplogroups based on STRs, such as the Bayesian-based haplogroup predictor (http://www.hprg.com/hapest5/index.html) or Nevgen (https://www.nevgen.org), but neither of these is based on generalized ML models such at those

proposed here. Here, similarly to the work of Schlecht et al. [27], albeit with a higher resolution, we developed a generalist ML-based approach to the problem of haplogroup assignation from Y-STR profiles, then applied it to the particular case of the CombYplex profiles. We also assessed whether it performs better than the more common linear discriminant analysis.

We ran a pre-pilot study to test the efficiency of seven ML models (detailed in [38]) so the fittest ML models could be implemented in PredYmale (details of pre-pilot study in Supplementary Data 4). Three models were eventually selected: Support Vector Machines (SVM), Random Forest Classifiers and k-Nearest Neighbors (kNN). These models follow the same concept: they build a classifier (a function) that maps a point in the problem space (here, a sample defined by its repeat counts for a given set of STRs) to a given class (here, a haplogroup). It should be noted that naive Bayes classifiers, a common method to address the problem of linking a set of STR markers to a haplogroup, and tested in a pilot run, have been constantly outperformed by SVMs and Random Forest Classifiers.

Support Vector Machines (SVM) are classifiers that linearly partition the problem space by determining the frontier of the hyperplane maximizing its distance to the training samples [39]. Although SVMs were originally designed to discriminate between only two classes, they can be used in multi-class classification problematics [40], the problem being then divided in as many one-versus-all sub-problems as there are classes, which are solved independently. These partial classifiers are then merged to define the final classifier. Concretely, each sample in the training set is represented in the problem space by a point whose coordinates are the number of repetitions for each STR. Samples with close characteristics will cluster together. The SVM will determine a set of hyperplanes maximizing the margin between the classes. New points (i.e. unlabelled samples) are classified in either class depending on where they find themselves with regards to these hyperplanes.

Random Forest Classifier decision trees [41] are linear classifiers that partition the problem space by defining a tree of binary conditions based on the features of a sample. Each new sample is then run through this tree of questions until it reaches a leaf, containing its predicted haplotype. Since a decision tree tends to over-fit the dataset it has been trained with, it might encounter difficulties generalizing when confronted with new samples. The random forest model [42] was developed to alleviate this limitation. At first, it trains multiple independent trees on several distinct subsets of the training data. Then, their outputs are averaged to define the final classifier. To improve the efficiency of random forests, we trained them with the AdaBoost boosting algorithm [43]. AdaBoost successively trains several copies of a base classifier (here a random forest) on the same dataset, and the training is adapted over generations to force the classifier to focus on hard to classify samples. Finally, all the generated classifiers over the generations are weighted according to their performances and combined to produce the final classifier. In our case, the learning process generates a decision tree defining questions on the number of repetitions of each STR. Depending on the answer, the sample to be classified will fall in one of the haplogroups. A notable advantage of this method is that its architecture (a sequence of questions) is easy for a human to understand, making the classification process transparent.

The **k-nearest neighbour** algorithm (also known as k-NN) is a non-parametric classification method. To produce a prediction for an unlabelled point, the algorithm combines the labels of the k closest points from the learning dataset according to a voting system. There are many ways to adapt the algorithm to the problem at hand, for instance by choosing the distance used, by applying a preliminary dimension reduction, by weighting the votes and so on. An advantage of the k-NN is that its error rate in a multi-class classification problem is proved to be bounded as an expression of the Bayes error rate, giving it a solid theoretical ground.

2.11.1. Implementing PredYMaLe

We developed PredYMaLe (Predicting Y-lineages using machine learning models), a graphical interface to our automatic labelling solution, available at https://gitlab.com/delehef/predymale/. It is implemented in Python using the scikit-learn machine learning library and the Qt5 GUI library, and is available for GNU/Linux, macOS and Windows. PredYMaLe can be used on any Y-STR dataset where every sample is represented as a set of numerical repeat values (*e.g.* CombYplex, PPY23, etc.). Empty or null values are deliberately not supported in PredYMaLe: to avoid biases stemming from an imperfect dataset, we advise users to remove or insightfully fix erroneous profiles. The predicted labels can be exported to a CSV file for easy interoperability with other programs.

2.11.2. Procedure

We tested whether haplogroups could be predicted using the three selected ML models implemented in the PredYMaLe program, and the

Table 2
Forensic parameter estimates for GEO and HAPLO samples for the full CombYplex, M1 and M2 and Y-filer. Parameters calculated: Genetic Diversity or Haplotype Diversity (GD/HD), Discrimination Capacity (DC), and Match Probability (MP).

Population (Geo sample)		CombYplex total				CombYplex M1				CombYplex M2			
	N	n	GD/HD	DC	MP	n	GD/HD	DC	MP	n	GD/HD	DC	MP
All pop	996	916	0,9998	0,9196	0,0012	607	0,9964	0,6094	0,0053	889	0,9998	0,8926	0,0013
South America : native (Palikur)	6	6	0,9999	1	0,1666	4	0,9630	0,6667	0,2778	6	0,9999	1	0,1667
South America : admixed	107	96	0,9986	0,8972	0,0118	84	0,9921	0,7850	0,0197	92	0,9982	0,8598	0,0127
Africa native	444	391	0,9995	0,8806	0,0029	242	0,9917	0,5450	0,0124	374	0,9994	0,8423	0,0033
Africa admixed	56	52	0,9982	0,9286	0,0210	45	0,9953	0,8036	0,0268	52	0,9981	0,9286	0,0210
Europe	383	368	0,9998	0,9608	0,0030	253	0,9916	0,6606	0,0123	364	0,9998	0,9504	0,0029

Haplogroup (Haplo sample)		CombYplex total				CombYplex M1				CombYplex M2				Y-filer			
Total Hg	503	n	GD/HD	DC	MP	n	GD/HD	DC	MP	n	GD/HD	DC	MP	n	GD/HD	DC	MP
E1a	15	14	0,9956	0,9333	0,0756	12	0,9891	0,8000	0,0933	13	0,9919	0,8667	0,0844	10	0,8889	0,6667	0,2000
E1b1a	275	244	0,9992	0,8873	0,0049	192	0,9958	0,6982	0,0093	238	0,9989	0,8655	0,0053	228	0,9988	0,8291	0,0056
E1b1b	12	12	1	1	0,0833	11	0,9931	0,9166	0,0972	11	0,9931	0,9167	0,0972	10	0,9877	0,8333	0,1111
F	7	7	1	1	0,1429	7	1	1	0,1429	7	1,0000	1	0,1428	7	1	1	0,1429
G	9	9	1	1	0,0987	8	0,9843	0,8750	0,1562	9	1,0000	1	0,0987	9	1	1	0,1250
I	14	13	0,9949	0,9286	0,0816	13	0,9949	0,9285	0,0816	13	0,9949	0,9286	0,0816	14	1	1	0,0714
J	12	12	1	1	0,0833	11	0,9931	0,9167	0,0972	12	1,0000	1	0,0833	11	0,9931	0,9167	0,0972
R1b1a1a2	159	152	0,9997	0,9560	0,0070	97	0,9810	0,6100	0,0291	151	0,9996	0,9497	0,0070	142	0,9989	0,8931	0,0081

N = Number of samples**; n** = number of distinct haplotypes; **HD**: haplotype diversity (gene diversity); **DC**: discrimination capacity; **MP,** match probability.

four four different Y-STR profiles (CombYplex full, CombYplex_M1, CombYplex_M2, and Y-filer). Each model was trained and evaluated using the HAPLO dataset (503 individuals, 7 Main and 12 Detailed Hg classes considered, 19 populations) and according to the same protocol. The dataset was normalized in the [0; 1] range to avoid numerical discrepancies influencing the final result. Similar to the LDA analyses, 75 % of the samples were used to train the model, while the remaining 25 % were used to evaluate the trained classifier by testing its efficiency. Given that ML training, as well as the split between training and validation datasets are heuristic, all the scores are averaged over 100 trials. This also alleviates score outliers and offers a better interpretation of the performances of multiple models on real datasets. For that purpose, we performed two runs of analyses: for the first run, individuals were considered to belong to one of seven major haplogroup classes (E1a-M33, E1b1a-M2*, E1b1b-M215, G-M201, I-M170, J-M304, and R1a1a1-M269 called *MainHg*), and for the second run, to one of twelve more detailed haplogroup classes (E1a-M33, E1b1a1-M2*, E1b1a7-M191, E1b1a7a-U174, E1b1a8a-U209, E1b1a8a1-U290, E1b1b1-M35*, G-M201, I-M170, J-M304, R1b1a1a2-M269, and R1b1a1a2a1a2a1b1a1-M167 called *DetailedHg*). The poorly represented haplogroup classes (e.g. F-189, and E1b1b1b1a-M81) could not be included in the procedure.

2.11.3. Validation

The evaluation process gives a score to a model, reflecting the efficiency of its predictions. We used the standard success score defined as s = nC /nT, where nC is the number of successfully labelled validation samples and nT the total count of validation samples. One success rating noted 'score' considers prediction as correct only if the predicted label of the validation sample matches the expected one.

3. Results

3.1. CombYplex: from polymorphism to discrimination power

The CombYplex polymorphism was assessed based on 996 samples. All CombYplex profiles are available in Supplementary data 5. As expected, we observed an increasing level of polymorphism from the less discriminative set of M1 markers (mean allele number: 6; Table 1) to the most discriminative M2 set (mean allele number: 9; Table 1). Forensic parameters were calculated for the GEO and HAPLO sample groups defined above (Table 2). GD and HD were greater than 0.999 for all GEO and HAPLO sub-groups using full CombYplex profiles. As expected, when M1 and M2 were analysed independently, M2 was always more discriminant than M1, with MP values oscillating from 0.001 (all populations) to 0.003 (Europe) using the GEO sample, and from 0.007 (Hg R) to 0.14 (Hg F) using the HAPLO sample. Indexes of discrimination capacity and match probability were observed in line with these values.

3.2. Inter-haplogroup comparative analyses: PCA and LDA

We tested whether CombYplex and Y-filer profiles could easily discriminate between haplogroup classes using the HAPLO sample (Supplementary 6). For this aim, we performed a PCA with seven haplogroup classes (MainHg) and a LDA (Table 3). PCA results based on CombYplex showed that haplogroup classes are well-discriminated along the two first axes (Fig. 2a, especially R1b1a1 and E1a1a), but also along the second and third axes (Fig. 2b, G, and I). LDA scores reach 94 % in average, and oscillate from excellent (100 for E1a-M33, E1b1a-M2, G-201, J-M304, R1b1a1a2-M269), to very good (95 for I-M170), and correct for the less represented classe (62 % for E1b1b-M35*).

In comparison, discrimination of haplogroup classes appears less efficient using Y-filer profiles, both on F1xF2 and the F2xF3 axes (Fig. 3a, b) but also using LDA (81 % on average).

These results provide evidence of the high resolving power of the 32 Y-STR CombYplex profile, not only for investigating paternal lineages but also for discriminating among haplogroups. Based on these encouraging results, we assessed whether haplogroup classes can be predicted using an ML approach based on the full CombYplex, CombYplex_M1, CombYplex_M2 and Y-filer profiles.

3.3. Haplogroup prediction (HP) using Y-STR profiles and PredYMaLe program

We tested whether haplogroup classes can be predicted using an ML-based approach on CombYplex, CombYplex_M1, CombYplex_M2 and Y-filer profiles. Results from the first run (seven major haplogroup classes :E1a-M33, E1b1a-M2*, E1b1b-M215, G-M201, I-M170, J-M304, and R1a1a1-M269 called *MainHg*) were very informative on the three methods and the four datasets tested. Although HP scores using SVM and Random Forest are similar, SVM performed slightly better than Random Forest (Table 3); on average, these two methods gave much better results than kNN: Random Forest/SVM HP average 3 methods

Table 3
Prediction scores (%) for seven haplogroup classes using three machine learning methods (SVM, Random Forest, k Nearest Neighbors) and LDA on four Y-STR datasets (CombYplex, M1, M2, Y-filer kit). For LDA, 10 individuals have been removed for Y-filer kit due to missing data; DYS502 has been removed from M1 analyses due to the lack of polymorphism.

Haplogroup	N	Method	Prediction score (in %)			
			Full CombYplex	M1	M2	Y-filer
E1a-M33	15	SVM	100	100	100	100
		Random Forest	97	99	83	99
		k Nearest Neighbors (kNN)	67	100	67	67
		LDA	100	100	100	97
E1b1a	275	SVM	100	99	97	99
		Random Forest	100	100	97	100
		k Nearest Neighbors (kNN)	99	100	97	100
		LDA	99	97	98	100
E1b1b	12	SVM	67	33	67	67
		Random Forest	28	28	28	54
		k Nearest Neighbors (kNN)	33	33	33	33
		LDA	62	61	55	75
All E collapsed	*302*	*SVM*	*100*	100	*96*	*96*
		Random Forest	*100*	100	*97*	*100*
		k Nearest Neighbors (kNN)	*99*	100	*93*	*100*
G	9	SVM	67	67	0	67
		Random Forest	71	75	5	69
		k Nearest Neighbors (kNN)	67	67	0	33
		LDA	100	88	67	88
I	14	SVM	100	100	100	75
		Random Forest	99	98	79	74
		k Nearest Neighbors (kNN)	75	100	75	75
		LDA	95	94	81	44
J	12	SVM	100	100	67	67
		Random Forest	98	100	13	39
		k Nearest Neighbors (kNN)	67	100	0	67
		LDA	100	100	14	67
R1b1a1a2-M269	159	SVM	95	98	93	98
		Random Forest	97	95	97	98
		k Nearest Neighbors (kNN)	100	98	95	98
		LDA	100	100	99	96
Average	496	SVM	97	96	92	95
		Random Forest	97	96	90	95
		k Nearest Neighbors (kNN)	73	97	52	68
		LDA	94	91	73	81

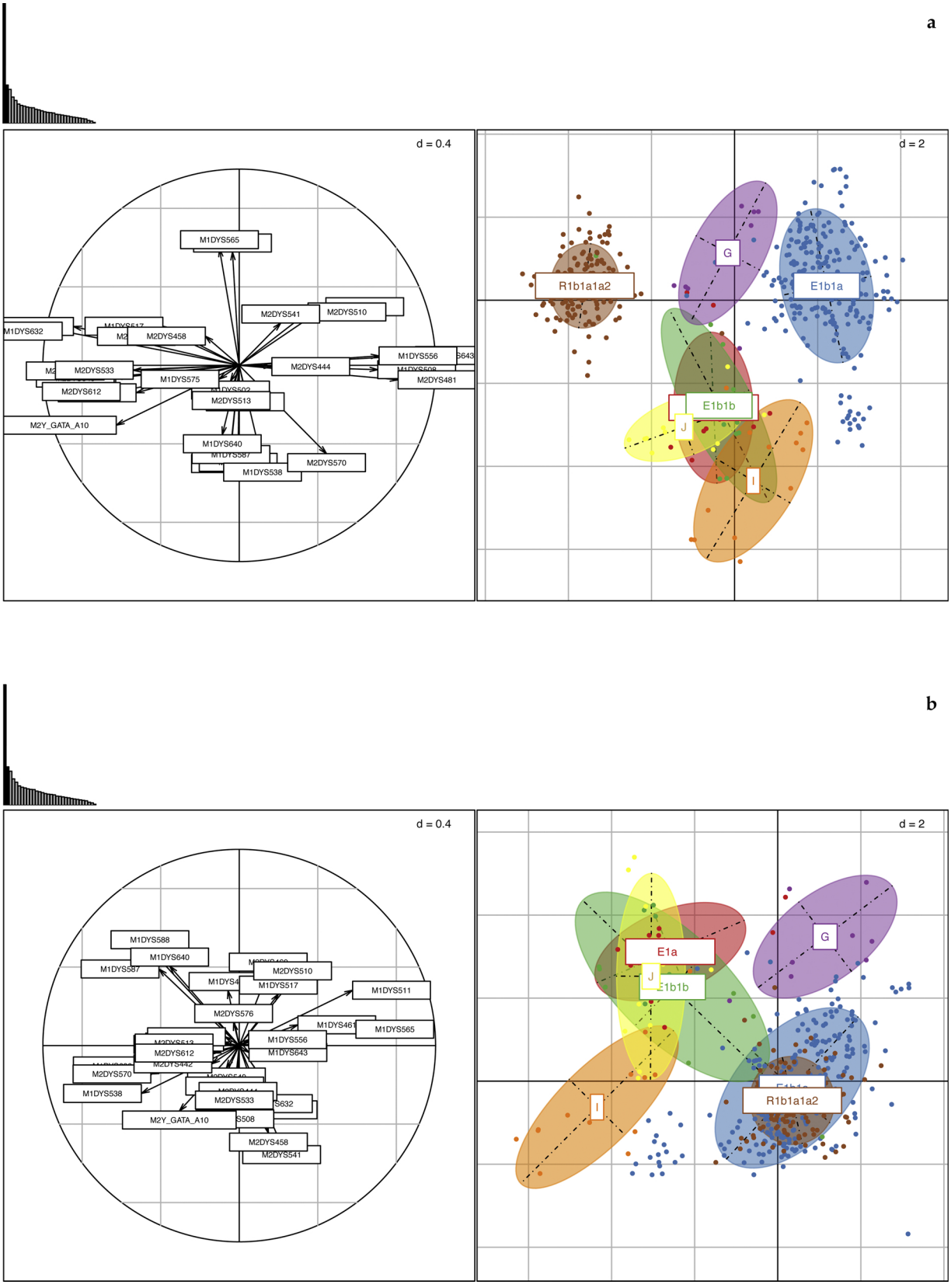

Fig. 2. a PCA for CombYplex F1xF2. b PCA for CombYplex F2xF3.

90–97 %; kNN HP average 3 methods: 52–73 %; Table 3).

Compared to classic LDA (73–94 %), SVM and Random Forest models perform systematically better, whatever the STR dataset, and especially using CombYplex. This result illustrates the combined impact of the marker number and the mutation rate range chosen on assignation accuracy. However, LDA performs better than kNN also for the three methods tested here. From the four STR datasets tested, we noted a noticeable performance of CombYplex (SVM: 97 %) compared with

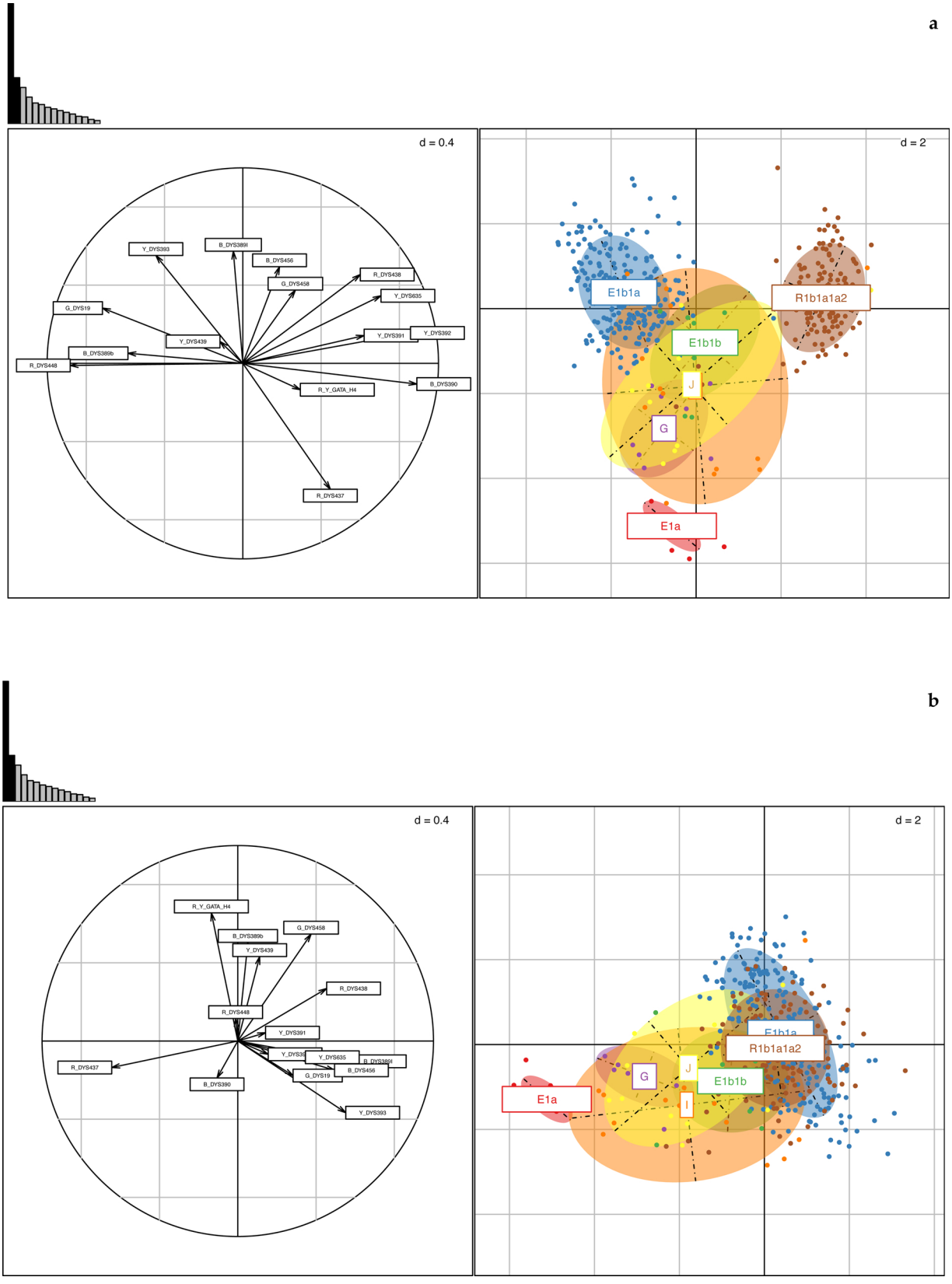

Fig. 3. a PCA for Y-filer F1xF2. b PCA for Y-filer F2xF3.

M1 (SVM: 96 %) and Y-filer (SVM: 95 %), the M2 subset being systematically declassed (SVM: 92 %, RF: 90 %, kNN: 52 %); when all E classes are collapsed, HP scores are very high (SVM et RF: 96–100 %). A strong heterogeneity in HP scores is observed between haplogroup classes, even when the best method (SVM) is considered with the best STR combination (CombYplex): the G (67 %) and E1b1b (67 %) branches give the lowest HP scores compare to all others branches (100 %). These two haplogroup classes represent the least represented ones

(respectively N = 9, 12), thus, suggesting the strong influence of sample size on the efficiency of HP. By analyzing confusion matrices for the best combination SVM/CombYplex and the worst combination kNN/M2, we observe clear differences in misclassification profiles (Fig. 4): for the best combination, only two misclassifications are observed: E1b1a for E1b1b, and R for G. In contrast, 5 miss-targeted classifications are observed for kNN/M2, illustrating the incapability of this model/STR panel to associate an STR profile to a defined haplogroup class (especially for G: 0 % HP).

No classifier exhibits a particularly skewed behavior regarding either of the metrics; all of them, on both datasets, follow the same pattern: F1-score and markedness stay close, while the informedness tends to score lower, denoting conservative classifiers. Therefore, defining the best classifier as the one with the best overall scores is straightforward. For a more detailed insight, Supplementary Data 7 (Supp Tables 7a–7c) contain the per-class, per-dataset and per-classifier precisions, recalls, F1-scores, informednesses and markednesses.

The second run aimed to test the impact of sub-branch on haplogroup assignation accuracy score. We used a maximum resolution by considering the 12 most represented haplogroup branches (E1a-M33, E1b1a1-M2*, E1b1a7-M191, E1b1a7a-U174, E1b1a8a*-U209, E1b1a8a1-U290, E1b1b1-M35*, G-M201, E1b1b1b1a, I J, R1b1a1a2-M269, R1b1a1a2a1a2a1b1a1-M167) and the two best models selected from the first run: SVM and Random Forest (Table 4). Per-class, per-dataset and per-classifier precisions, recalls, F1-scores, informednesses and markednesses are given in Supplementary Data 7 (Supp Tables 7d-7f).

The average HP scores are high for both models and the four datasets, but they are lower than those from the first run, probably due to the smaller sample sizes and the close genetic affinity of the different classes. Better prediction performances are observed for Random Forest, all STR datasets considered, with the highest average HP score obtained for CombYplex. The lowest scores are observed for M2 with an average HP score of 71 % for Random Forest; this Y-STR dataset also has higher heterogeneity in HP scores between classes (from 27 % for E1b1a1 to 100 % E1a-M33; Table 4). By analyzing the confusion matrices for the best combination (Random Forest/CombYplex) and the worse (SVM/M2), we noticed that misclassification profiles are different (Fig. 5). For Random Forest/CombYplex, misclassifications occur mainly across phylogenetically neighbors E1b1a and R1b1a1a2 branches. In contrast, for SVM/M2, misclassifications are associated with very diverse branches on the whole Y-chromosome phylogenetic tree (e.g. hg G), reflecting the impact either of highly mutating markers, the lower number of STR loci in this panel or the lack of association between STR profile and Y-haplogroup due to the impact of additional molecular mechanism as gene conversion.

4. Discussion

In this paper, we assess whether a panel of well-balanced Y-STR mutations, built around two sub-STR panels (from 3.85 × 10^{-04} to 1.45 × 10^{-02} mutation/locus/generation), associated with machine learning (ML) approaches can efficiently predict haplogroups. We developed the 32 Y-STR panel "CombYplex" and genotyped it on 996 male individuals from three continents (West and South Africa, West Europe, South America) to explore and confirm the discrimination capacity of the full, M1 and M2 panels, using classing forensic and statistics parameters. Then, we developed the ML approach PredYMaLe (Predicting Y-lineages using ML models) and tested it on an assembled panel of 503 individuals, for which Hg and Y-filer information were also available in our database allowing a direct comparison of Y-STR assemblies.

4.1. STR panels and ML classifers: an ideal association?

We have demonstrated noticeable differences in prediction scores between STR panels and ML methods. Among all ML classifiers, SVM and Random Forests give better and more homogeneous prediction scores (90–97 %) compared with kNN (52–97 %) for this dataset, independently of the panels analysed.

When performing basal branch analyses (7-classes), the mutationally well-balanced panels (CombYplex, Y-filer and the average-mutating panel (M1)) performed better than the M2 panel, which was systematically outperformed. This result suggests that mutationally well-balanced or average STR panels should be preferred when analysing basal branches. The lower performance of M2 could imply either that assignation accuracy is affected by homoplasia using M2, due to the high mutation rate of the panel, or by the low number of STRs analysed (14 STRs). The latter argument is less probable since the 15 selected STRs of the Y-filer profiles gave better results.

When moving toward terminal branches (12-classes), mutationally

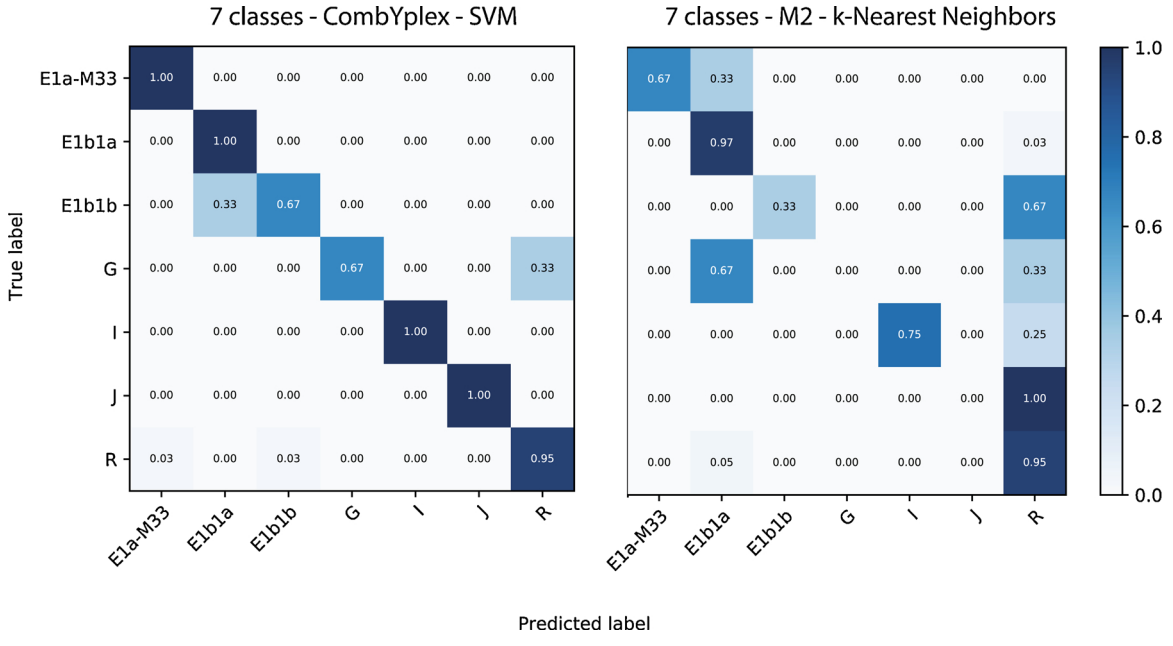

Fig. 4. Confusion matrices for the first run on MainHg (7 haplogroup classes) for CombYplex/SVM and M2/k-Nearest Neighbors.

Table 4
Prediction scores (%) for twelve haplogroup classes using the two best machine learning methods (SVM and Random Forest) on four Y-STR datasets (CombYplex, M1, M2, Y-filer).

Haplogroup	N	Method	Prediction score (in %)		
			CombYplex	M2 only	Y-filer
E1a-M33	15	SVM	100	100	100
		Random Forest	98	90	99
E1b1a1-M2*	44	SVM	45	27	27
		Random Forest	58	46	37
E1b1a7-M191	17	SVM	40	60	80
		Random Forest	40	40	60
E1b1a7a-U174	79	SVM	75	80	90
		Random Forest	81	70	87
E1b1a8a-U209	66	SVM	75	62	56
		Random Forest	72	74	70
E1b1a8a1-U290	69	SVM	35	47	47
		Random Forest	56	59	63
E1b1b1-M35*	10	SVM	100	67	67
		Random Forest	68	32	48
G-M201	9	SVM	67	33	67
		Random Forest	88	28	92
I	14	SVM	100	75	75
		Random Forest	100	83	72
J	12	SVM	100	33	33
		Random Forest	100	32	43
R1b1a1a2-M269	134	SVM	85	85	94
		Random Forest	97	99	91
R1b1a1a2a1a2a1b1a1-M167	25	SVM	86	29	0
		Random Forest	84	60	58
Average	494	SVM	71	64	67
		Random Forest	79	71	74

well-balanced STR panels (CombYplex, Y-filer) performed better than M1 and M2 panels. M1 composed solely of average mutating STRs (18 STRs) were less performant due to its lack of discrimination power, giving equivalent results to M2 with four additional STR loci. Assignation accuracies for M1 and M2 decrease for the less represented classes, reflecting the need for the largest training set possible, and also a well-balanced STR panel with a sufficient number of STR loci when exploring closely related phylogenetic branches.

4.2. Variation in performance accuracies across Hg classes

We showed that some haplogroups (e.g. E1a, I, J) have very distinct and unambiguous Y-STR profiles leading to 100 % assignation accuracy scores, while others haplogroups (e.g. G, E1b1b) are more prone to misclassification within the STR panels and datasets analysed here. The impact of complexifying molecular mechanisms, such as gene conversion [44], CNV-STR [50] which potentially affect these profiles cannot be excluded [30] and could be further investigated. However the

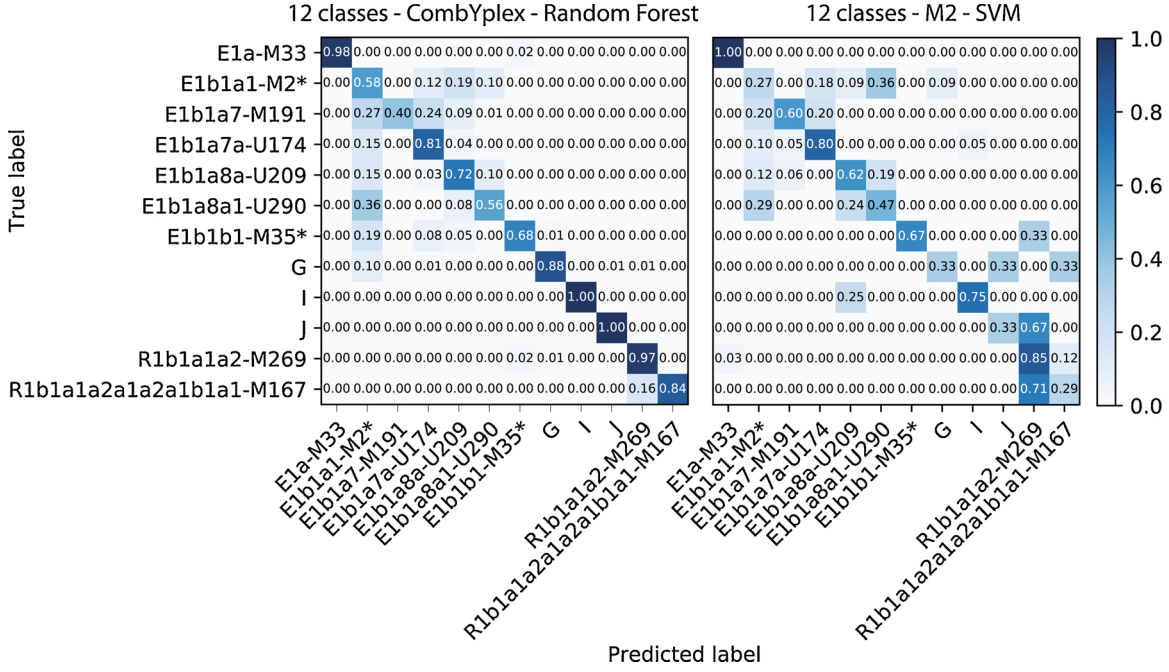

Fig. 5. Confusion matrices for the second run on DetailedHg (12 haplogroup classes) for CombYplex/Random Forest/ and M2/SVM.

consistently worst scores of misclassification for the G and E1b1b haplogroups is likely to be the simple consequence of their small sample size. If the low accuracy of less well represented classes is problematic, empirical trends suggest that results are instantly improved when more training data are available. By running PredYmale with 10 additional G profiles collected recently, we observed that the prediction accuracy score reaches 83 %, illustrating that prediction accuracy is significant improved when more training data are available. We encourage users to train and use PredYmale on their own datasets, to learn about the prediction scores expected for the part of the tree explored. Given that PredYmale computations are rather fast, users should not hesitate to use larger datasets, or to adapt their STR panels to attain the best prediction scores.

4.3. Using PredYMaLe with other STR panels

Our results demonstrate the need to find a good equilibrium between the number of markers, their mutability and the sample size of the training set according to the tree structure considered. When analysing basal branches, well-balanced STR panels or average mutating STR panels can be selected preferably with SVM or Random Forest classifiers to ensure higher prediction scores. The M1 panel, an average mutating STR panel, gives very good results. Since these STRs have generally simpler motifs or low repeat counts, they can be extracted from whole-genome sequencing data using pre-existing tools (STRait Razor, [21]) and used to predict basal branches.

When moving toward terminal branches, mutationally well-balanced STR panels associated with SVM or Random Forest classifiers can be selected. In both cases, a minimal number of markers ($>$ 20–30 STRs) is required to guarantee the best prediction scores possible. In forensic genetics, two commercial kits are commonly used, PPY23 [19] and Y filer ® Plus [20]. We have briefly tested whether our program could be confidently used with these panels by running PredYmale on published data. Based on our previous conclusions, we have only included the most represented classes ($N > 20$). We analysed 451 individuals from five basal branches (E1b1b, G, I, J, R) for PPY23 [45,46], and 282 individuals from four basal branches (G, I, J, R1) for Y filer ® Plus [47]. The average prediction scores obtained with SVM and Random Forest reached 98.5 % for PPY23 and 97 % for Y-filer plus (equiv. sample for CombYplex reaches 98.5 %). These results confirm the high prediction scores obtained with the SVM and Random Forest classifiers, for the three mutationally well balanced panels, for basal branches and sufficently large training sets.

4.4. Predicting Hg using ML approaches: SVM, random forest and nearest neighbours classifiers

By developing an ML program (PredYMaLe), designed to predict haplogroups using any Y-STR profiles, we show that ML models, especially SVM and Random Forest, give much better HP results compared to alternative ML methods, including Bayesian, or Neural Network-based models. Interestingly these two classifiers have been reported to perform quite well for many other biological data [48]. An interesting observation resides in the large variance of scores depending on the algorithm used: naive Bayes methods giving the worst results, while SVMs reach excellent precisions. The low accuracy of naive Bayes-based methods, in this case, can be explained by the fact that these algorithms consider features independently, and so cannot capture the information contained in their covariance patterns. SVMs, on the other hand, by maximizing the margin between the training classes, typically give excellent results as long as first, the problem is linearly well separable, which seems to be the case in this study, and second, that there is no consequent overlap between the different classes. Were it not the case, one can apply the “kernel trick” [49], which uses Mercer’s theorem to computationally cheaply immerse the dataset in a much larger space, where classes that are not linearly separable in the original space might become linearly separable.

In conclusion, support vector machines, random forests and nearest-neighbors classifiers are interesting alternatives to Bayesian or Neural networks classifiers to predict Y-haplogroups. Future users should note that although we developed and mostly used PredYmale with datasets featuring Y-STR profiles sampled with the CombYplex kit, the underlying ML concepts in our tool can be used on any STR panel (using STR repetition counts). We encourage users to train and use PredYmale on their own datasets regardless of the typing method.

Acknowledgments

We thank all DNA donors and volunteers associated with the sampling sessions. We also warmly thanks Prof. Maria Cátira Bortolini for giving us access to Brazilian samples, to Prof. Antoine Gessain for the Guyanese Noir Marrons. This work was supported by a Maturation research grant (CB’s post-doctoral position), Research and Post-graduate Teaching Pole (PRES), the University Toulouse III (11.007), the LABEX DRIIHM (Investing in a future programme, ANR-11-LABX-0010), the OHM Haut Vicdessos, the Spanish Ministry of Economy and Competitiveness’s grants (CGL2010-15191/BOS and CGL2014-53985-R) and the National Research Foundation Grant IFR160623173836 (MED). This work was performed using HPC resources from CALMIP (grant P1434). FD was supported by a PhD studentship (INSA, France), AM by a PhD studentship (Ministry of research, French government), NS by La Estancia de Otoño HOCR Cia. Ltda. (grant number 201509), CLH by a Spanish’s research contract. CFL was supported by the EUROTAST Marie Curie Initial Training Network (EU FP7/2007-2013, grant no. 290344) and the Sven and Lilly Lawski's Foundation (N2019-0040). Ethics approvals were obtained: from the Senate Research Committee of the University of the Western Cape for South African samples under (ethic number 15-4-97, DC-2011-1436), from the Ethics Committee of the Faculty of Health Sciences, University d'Abomey-Cavali, Benin for the Beninese samples (ethic number 07/T4/2015/CE/FSS/UAC, 30th October 2015), from the University Bioethics committee (Sede de Investigación Universitaria, SIU) for the Colombian samples (ethic number 09-12-225 form), from the research ethics committee of the Universidade Federal do Rio Grande do Sul (Resolution no. 98002/1998) for the Brazilian samples Brazilian Ethics Commission, CONEP ethic number 1333/2002). Other samples from Africa were collected in the 80 s or before, and ethics approval were not requested at that time; however, all participants were volunteers with the purpose of collaborating with scientific studies, gave oral consent for the collection, and the confidentiality of their personal information has been preserved, following Helsinki Declaration.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version at doi:https://doi.org/10.1016/j.fsigen.2020.102342.

References

[1] M. Kayser, Forensic use of Y-chromosome DNA: a general overview, Hum. Genet. 136 (5) (2017) 621–635, https://doi.org/10.1007/s00439-017-1776-9.
[2] M.A. Jobling, C. Tyler-Smith, The human Y chromosome: an evolutionary marker comes of age, Nat. Rev. Genet. 4 (2003) 598–612, https://doi.org/10.1038/nrg1124.
[3] F. Calafell, M.H.D. Larmuseau, The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research, Hum. Genet. 136 (5) (2017) 559–573, https://doi.org/10.1007/s00439-016-1740-0.
[4] J. Pardo-Seco, et al., Biogeographical informativeness of Y-STR haplotypes, Sci. Bull. Elsevier 64 (19) (2019) 1381–1384, https://doi.org/10.1016/J.SCIB.2019.07.025.
[5] P. Gill, et al., Identification of the remains of the Romanov family by DNA analysis, Nat. Genet. 6 (2) (1994) 130–135, https://doi.org/10.1038/ng0294-130.
[6] F. Austerlitz, E. Heyer, Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population, Proc. Natl. Acad. Sci. U. S. A. 95 (25) (1998) 15140–15144. The National Academy of Sciences.

[7] T.E. King, et al., "Identification of the Remains of King Richard III", Nature Communications vol. 5, Nature Publishing Group, 2014, pp. 1–56318, https://doi.org/10.1038/ncomms6631 5631.

[8] T.E. King, et al., Thomas Jefferson's Y chromosome belongs to a rare European lineage, Am. J. Phys. Anthropol. 132 (4) (2007) 584–589, https://doi.org/10.1002/ajpa.20557.

[9] G.R. Bowden, et al., Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England, Mol. Biol. Evol. 25 (2) (2008) 301–309, https://doi.org/10.1093/molbev/msm255.

[10] R. Chaix, et al., Genetic traces of east-to-west human expansion waves in Eurasia, Am. J. Phys. Anthropol. 136 (3) (2008) 309–317, https://doi.org/10.1002/ajpa.20813.

[11] E. Heyer, et al., Genetic diversity and the emergence of ethnic groups in Central Asia, BMC Genet. 10 (49) (2009) 1–8, https://doi.org/10.1186/1471-2156-10-49.

[12] E. Heyer, et al., Patrilineal populations show more male transmission of reproductive success than cognatic populations in Central Asia, which reduces their genetic diversity, Am. J. Phys. Anthropol. 157 (4) (2015) 537–543, https://doi.org/10.1002/ajpa.22739.

[13] T.E. King, M.A. Jobling, Founders, drift, and infidelity: the relationship between y chromosome diversity and patrilineal surnames, Mol. Biol. Evol. 26 (5) (2009) 1093–1102, https://doi.org/10.1093/molbev/msp022.

[14] T.E. King, M.A. Jobling, 'What's in a name? Y chromosomes, surnames and the genetic genealogy revolution', Trends Genet. 25 (8) (2009) 351–360, https://doi.org/10.1016/j.tig.2009.06.003.

[15] P. Verdu, et al., Limited dispersal in mobile hunter-gatherer Baka Pygmies, Biol. Lett. 6 (2010) 858–861, https://doi.org/10.1098/rsbl.2010.0192.

[16] C. Martinez-Cadenas, et al., The relationship between surname frequency and Y chromosome variation in Spain, Eur. J. Hum. Genet. 24 (1) (2016) 120–128, https://doi.org/10.1038/ejhg.2015.75.

[17] B. Sobrino, M. Brión, A. Carracedo, SNPs in forensic genetics: a review on SNP typing methodologies, Forensic Sci. Int. 154 (2–3) (2005) 181–194, https://doi.org/10.1016/j.forsciint.2004.10.020.

[18] A. Ralf, et al., Forensic Y-SNP analysis beyond SNaPshot: high-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing, Forensic Sci. Int. Genet. 41 (2019) 93–106, https://doi.org/10.1016/j.fsigen.2019.04.001.

[19] J. Purps, et al., A global analysis of Y-chromosomal haplotype diversity for 23 STR loci, Forensic Sci. Int. Genet. 12 (2014) 12–23, https://doi.org/10.1016/j.fsigen.2014.04.008.

[20] S. Gopinath, et al., Developmental validation of the Yfiler ® plus PCR Amplification Kit: an enhanced Y-STR multiplex for casework and database applications, Forensic Sci. Int. Genet. 24 (2016) 164–175, https://doi.org/10.1016/j.fsigen.2016.07.006.

[21] D.H. Warshauer, et al., STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, Forensic Sci. Int. Genet. 7 (4) (2013) 409–417, https://doi.org/10.1016/j.fsigen.2013.04.005.

[22] K.L. Young, et al., Paternal genetic history of the Basque population of Spain, Hum. Biol. 83 (4) (2011) 455–475.

[23] Mirabal, et al., Human Y-chromosome short tandem repeats: a tale of acculturation and migrations as mechanisms for the diffusion of agriculture in the Balkan Peninsula, Am. J. Phys. Anthropol. 142 (2010) 380–390, https://doi.org/10.1002/ajpa.21235.

[24] Šehović, et al., Network analysis on the in silico assigned Y chromosome haplogroups in Western Balkan populations, Genet. Appl. 1 (2) (2017) 36–43, https://doi.org/10.31383/ga.vol1iss2pp36-43.

[25] J. Jannuzzi, et al., Male lineages in Brazilian populations and performance of haplogroup prediction tools, Forensic Sci. Int. Genet. 44 (2020) 1–7, https://doi.org/10.1016/j.fsigen.2019.102163.

[26] T.W. Athey, Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach, J. Genet. Geneal. 2 (2006) 34–39.

[27] J. Schlecht, et al., Machine-learning approaches for classifying haplogroup from Y chromosome STR data, PLoS Comput. Biol. 4 (6) (2008) e1000093, https://doi.org/10.1371/journal.pcbi.1000093.

[28] T. Kivisild, The study of human Y chromosome variation through ancient DNA, Hum. Genet. 136 (2017) 529–546, https://doi.org/10.1007/s00439-017-1773-z.

[29] V.C. Cadamuro, et al., Determined about sex: sex-testing in 45 primate species using a 2Y/1X sex-typing assay, Forensic Sci. Int. Genet. 14 (2015) 96–107, https://doi.org/10.1016/j.fsigen.2014.09.010.

[30] P. Balaresque, et al., Gene conversion violates the stepwise mutation model for microsatellites in y-chromosomal palindromic repeats, Hum. Mutat. 35 (5) (2014) 609–617, https://doi.org/10.1002/humu.22542.

[31] C. Fortes-Lima, et al., Genetic population study of Y-chromosome markers in Benin and Ivory Coast ethnic groups, Forensic Sci. Int. Genet. 19 (2015) 232–237, https://doi.org/10.1016/j.fsigen.2015.07.021.

[32] M. Nei, et al., Polymorphism and evolution of the Rh blood groups, Jpn. J. Hum. Genet. 26 (1981) 263–278, https://doi.org/10.1007/BF01876357.

[33] M. Nei, Analysis of Gene diversity in subdivided populations, Proc. Natl. Acad. Sci. 70 (12) (1973) 3321–3323, https://doi.org/10.1073/pnas.70.12.3321.

[34] L. Excoffier, H.E.L. Lischer, Arlequin Suite Ver 3.5: a New Series of Programs to Perform Population Genetics Analyses under Linux and Windows', Molecular Ecology Resources, John Wiley & Sons, Ltd, 2010, pp. 564–567, https://doi.org/10.1111/j.1755-0998.2010.02847 (10.1111), 10(3).

[35] R. CoreTeam, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing., Vienna, Austria, 2017.

[36] S. Dray, A.-B. Dufour, The ade4 package: implementing the duality diagram for ecologists, J. Stat. Softw. 22 (4) (2007) 1–20, https://doi.org/10.18637/jss.v022.i04.

[37] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, Springer New York (Statistics and Computing), New York, NY, 2002, https://doi.org/10.1007/978-0-387-21706-2.

[38] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[39] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning vol. 20, Kluwer Academic Publishers-Plenum Publishers, 1995, pp. 273–297, https://doi.org/10.1023/A:1022627411411 (3).

[40] Chih-Wei Hsu, Chih-Jen Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 415–425, https://doi.org/10.1109/72.991427.

[41] L. Breiman, et al., Classification and Regression Trees, Chapman & Hall/CRC, 1984, p. p368.

[42] T.K. Ho, Random decision Forest, Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montréal, 1995, pp. 278–282.

[43] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139, https://doi.org/10.1006/JCSS.1997.1504 Academic Press.

[44] S. Rozen, et al., Abundant gene conversion between arms of palindromes in human and ape Y chromosomes, Nature 423 (6942) (2003) 873–876, https://doi.org/10.1038/nature01723.

[45] H. Pamjav, et al., A study of the Bodrogköz population in North-Eastern Hungary by Y chromosomal haplotypes and haplogroups, Mol. Genet. Genomics 292 (4) (2017) 883–894, https://doi.org/10.1007/s00438-017-1319-z.

[46] A. Heraclides, et al., Y-chromosomal analysis of Greek Cypriots reveals a primarily common pre-ottoman paternal ancestry with Turkish cypriots, PLoS One 12 (6) (2017) e0179474, https://doi.org/10.1371/journal.pone.0179474.

[47] D.S. Lacerenza, et al., Investigation of extended Y chromosome STR haplotypes in Sardinia, Forensic Sci. Int. Genet. 27 (2017) 172–174, https://doi.org/10.1016/j.fsigen.2016.12.009.

[48] M. Fernández-Delgado, et al., Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (2014) 3133–3181.

[49] M. Aizerman, et al., Theoretical foundations of the potential function method in pattern recognition learning, Autom. Remote. Control. 25 (1964) 821–837.

[50] P. Balaresque, et al., Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis, Hum. Mutat. 29 (10) (2008) 1171–1180, https://doi.org/10.1002/humu.20757.

[52] M. Kayser, et al., A comprehensive survey of human Y-chromosomal microsatellites, Am. J. Hum. Genet. 74 (6) (2004) 1183–1197, https://doi.org/10.1086/421531.

[53] W. Parson, et al., Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63, https://doi.org/10.1016/j.fsigen.2016.01.009.

[54] Leonor Gusmão, et al., DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis, DNA Commission of the International Society of Forensic Genetics, Forensic Sci. Int. 10 (2006), https://doi.org/10.1016/j.forsciint.2005.04.002.

[55] Felix Immanuel website. 2013.