Nolan, Michael. Polymorphism of Y-STR haplotypes is governed by patrilineal ancestry combined with effects of male migration. Master of Science (Biomedical Sciences, Molecular Genetics) September, 2015, 40 pp., 23 tables, 4 figures, bibliography, 35 titles.

This study examined geographic origins of Y-haplogroups and effects of migration using Y-STR haplotype databases compiled from the literature. Accuracy of haplogroup prediction was analyzed by varying the number of loci in haplotype definitions and by including rapidly mutating Y-STRs. Lastly, haplogroup diversities of populations were analyzed with respect to evolutionary history/size of populations and effects of admixture.

These analyses demonstrated: a) haplotype definitions with more loci increased haplogroup prediction accuracy; b) older populations did not negatively impact haplogroup prediction; c) including rapidly mutating loci as part of the haplotype-definition had minimal impact on haplogroup prediction and inferring population clustering, but had moderate impact on Network analysis; and d) haplogroup diversities increased with male admixture.

POLYMORPHISM OF Y-STR HAPLOTYPES IN POPULATIONS IS GOVERNED BY PATRILINEAL ANCESTRY COMBINED WITH EFFECTS OF MALE MIGRATION

Michael Robert Nolan, B.S., B.S.

APPROVED:

_____

Major Professor

_____

Committee Member

_____

Committee Member

_____

Committee Member

_____

University Member

_____

Chair, Department of Molecular and Medical Genetics

_____

Dean, Graduate School of Biomedical Sciences

POLYMORPHISM OF Y-STR HAPLOTYPES IN POPULATIONS IS GOVERNED BY PATRILINEAL ANCESTRY COMBINED WITH EFFECTS OF MALE MIGRATION

THESIS

Presented to the Graduate Council of the

Graduate School of Biomedical Sciences

University of North Texas

Health Science Center at Fort Worth

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Michael Robert Nolan, B.S., B.S.

Fort Worth, TX

September, 2015

ACKNOWLEDGEMENTS

Let me express gratitude to my major professor, Dr. Ranajit Chakraborty whose expertise, guidance, patience and time while performing research under his supervision. I would like to thank Dr. John Planz for his open door policy which allowed a lot of time discussing forensic genetics. I would also like to thank both Dr. Clark and Dr. Larue for their insight on how to give scientific presentations. Lastly, I would like to thank Dr. Jack Ballantyne from the University of Central Florida for providing me access to the haplotype data in the U.S. Y-STR database.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION: Specific AIMs, Research Hypotheses and Expected Results

Evolutionary analyses of Y-linked SNPs provide clustering of Y-haplotypes defining haplogroups, whose geographic origins have been studied at least at continental levels[1,2,3]. Haplogroups are defined as groups of haplotypes that share a common ancestral SNP. Lack of recombination along with Bayesian prediction algorithms[4,5] allow haplogroup prediction from Y-STR haplotype data. Thus, the analysis of haplogroup diversity within populations can reveal the effects of admixture, male migration and explain, in part, polymorphism of Y-STR haplotypes. Accuracy of the Haplogroup Predictor software has been demonstrated[6,7]; however more thorough analyses of the factors that dictate the accuracy of haplogroup predictions using Y-STR data was warranted, including the effects of adding more loci in haplotype-definitions and newly discovered rapidly mutating Y-STR (RM Y-STR) loci.

With four specific AIMs of this proposed study, the hypotheses to be tested were: (i) accuracy of haplogroup prediction increases with the number of loci encompassed in the haplotype-definition; (ii) evolutionary age of populations contributes to haplogroup diversity affecting the relationship of haplogroup prediction accuracy with the number of loci comprising the haplotype definition; (iii) admixture due to male migration between populations, increases the haplogroup diversity further, affecting the haplogroup predictability in comparison to that in their corresponding ancestral populations; and (iv) inclusion of RM Y-STRs has a further impact on population assignments and haplogroup prediction.

Published datasets were used for all of these analyses, including haplotypes with known haplogroups from prior data[8], PowerPlex®Y23 (PP®Y23) haplotypes from global populations[9]

and the U.S. Y STR database[10]. Apart from the use of the Haplogroup Predictor, other programs used are: STRUCTURE analysis for population assignments of Y-STR haplotypes and Network for median joining network analysis for congruence (or discordance) of haplotypes and population origin, and haplogroup diversity analyses for estimating the extent of haplogroup diversities[11,12,13,14]. These were prompted by the central hypothesis of this research, namely, polymorphism of Y-STR haplotypes in populations is governed by patrilineal ancestry combined with effects of male migration. In this study, I addressed the above central hypothesis by using the concept of Y-haplogroup as the indicator of patrilineal ancestry, with which the following four specific AIMs were studied:

Specific AIM I: Predictability of Y-haplogroup (i.e., patrilineal ancestry) from Y-STR haplotypes and its relationship with evolutionary age of the haplogroups and the number of STR loci encompassed in the definition of Y-STR haplotypes. This AIM was designed to test the hypothesis: *Accuracy of haplogroup prediction of Y-STR haplotypes increases with the number of loci in the haplotypes as well as increased power of discrimination of haplotypes.*

This AIM was studied with data on 201 Y-STR haplotypes collected from global populations whose haplogroup definitions are available from prior data with primary focus on the 171 Y-STR haplotypes which have supported major-haplogroups by the prediction software. With the use of batch version of Haplogroup Predictor algorithms[15], each Y-STR haplotype was assigned a haplogroup. The haplotypes were subsequently reduced in their size with respect to commonly used haplotype-definitions (Table 1.1). I examined the trend of accuracy of haplogroup assignment with reduction of number of loci as well as power of discrimination of the haplotypes. One previous study found a prediction error of 4.8% using the Haplogroup

Predictor software[7]. Additionally, it was expected that an increase of loci in the Y-STR haplotype will increase the accuracy of haplogroup prediction.

Table 1.1: Haplotype-definitions

Loci set	Number of Loci	List of Loci
Maximum Haplotype Definition (MXHT)	20	Yfiler® + DYS576* + DYS570* + DYS481
Yfiler®	17	PP®Y12 + DYS448 + DYS456 + DYS458 + DYS635 + YGATAH4
PowerPlex®Y (PP®Y12)	12	SWGDAM + DYS437
SWGDAM	11	MHT + DYS438 + DYS439
Minimum Haplotype Definition (MHT)	9	DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a/b

*RM Y-STR locus

Specific AIM II: Confounding effects of evolutionary age of populations on haplogroups diversity and its relationship with number of loci in the Y-STR haplotype. The hypothesis tested by this AIM is: *Populations that are older in evolutionary age are expected to produce increased haplotype diversity influencing the prediction of haplogroups and its relationship with the number of loci comprising the haplotype-definition.*

The PowerPlex®Y23 Y-haplotype data on numerous worldwide populations are available in the literature[9]. From this resource, Y-STR haplotypes were compiled from chosen populations of African, Asian, Caucasian, Native American origin. For each of these populations, haplogroup predictions were made by the same software used for Specific AIM I. Populations of known geographic areas had their haplogroup assigned. Additionally, I reduced the loci in haplotypes (from 20 to 9 by commonly used platforms, similar to the first AIM). It was expected that older populations will have higher haplogroup diversity and that there will be diminished prediction

capability with fewer loci comprising the haplotype definition. Small isolated populations, e.g. Native Americans, were expected to have a lesser haplogroup diversity.

Specific AIM III: Effect of admixture on Haplogroup Diversity and Haplogroup Predictability. The corresponding hypothesis for this AIM is: *Increased haplogroup diversity in admixed populations is the consequence of male migration which is expected to bring in additional haplogroups due to mating between previously isolated populations.*

PowerPlex®Y23 data on Y-STR haplotypes from populations of U.S. Hispanics, Admixed Brazilian and British African were subjected to analyses similar to that of Specific AIM II to examine how admixture alters the extent of haplogroup diversity in comparison to that in the ancestral populations between which admixture occurred. The expectation was to observe a higher haplogroup diversity in known admixed populations and consequently a somewhat compromised predictability of haplogroups in such admixed populations.

Specific AIM IV: The last specific AIM of this project was to **analyze the effect of rapidly mutating Y-STRs (RM Y-STR) on prediction of patrilineal ancestry** to test the hypothesis: *Rapidly mutating Y-STRs are expected to have a negative impact on the prediction of patrilineal ancestry and identifying underlying population stratification.*

This analysis was done in three ways. First, the assignment of haplogroups were made with and without the two RM Y-STRs in the twenty loci haplotype-definition (MXHT definition, see Table 1.1). Second, the haplotypes were subjected to a STRUCTURE analysis under the no-admixture model both with and without the two rapidly mutating[11,12]. Lastly, median joining networks were constructed with an admixed population and their corresponding ancestral populations to demonstrate the effect of including the RM Y-STRs. The difference in the results

between the two haplotype-definitions demonstrated the effect of RM Y-STRs on patrilineal ancestry through haplogroup diversity and population clustering. It was expected that the presence of RM Y-STRs would adversely affect haplogroup prediction and population clustering.

CHAPTER II

BACKGROUND AND SIGNIFICANCE

Polymorphism of human Y-chromosome is generally studied either by Y-linked short tandem repeat (Y-STR) haplotypes or by Y-haplogroups defined by specific single nucleotide polymorphism (SNP) sites of the non-recombining segment of the Y-chromosome (NRY-segment). Y-STR haplotypes are described by the repeat sizes of alleles within each STR locus encompassed in the haplotypes. STRs are comprised of numerous types of repeat motifs; simple, compound, complex and repeats containing non-variable and non-repetitive regions[16]. Both autosomal and Y-STRs have mutation rates of the order of 10^{-3} per locus per generation. However, recently a new class of STRs on the Y-chromosome was discovered which mutate almost an order of magnitude faster (10^{-2} per locus per generation) than the standard STRs. These are named as rapidly mutating Y-STRs (RM Y-STR)[17]. Only three of the originally defined thirteen rapidly mutating loci[17] are compatible with Haplogroup Predictor[4,5]. Yfiler® Plus and PowerPlex®Y23 are two of the new Y-STR kits and both possess RM Y-STRs as part of their haplotype definitions. Yfiler® Plus contains six RM Y-STRs, of which one is a duplicated locus (DYS570, 576, 627, 449 and the duplicated locus DYF387S1)[18] and PowerPlex®Y23 contains two of them (DYS 570 and 576)[19]. The analyses here used the haplotypes as defined in Table 1.1 in Chapter I. The assessment of the accuracy of Haplogroup Predictor was done in a previous study, utilizing a small haplotype-definition consisting of 7 loci, in part concluded "an increase in the number of STRs employed to predict the haplogroup would not enhance accuracy, considering the few reference samples available with the seven standard STRs and associated haplogroups,…"[20]. The author of the Haplogroup Predictor program responded indicating that if more loci had been used the haplogroup prediction program would

have performed better and that additional loci would increase the prediction probability for the correct haplogroup[21].

Y-haplogroups are informative of paternal ancestry. Specific mutations at certain SNP sites that occurred during the history of human evolution define Y-haplogroups. Due to comparatively higher rate of mutations at the STR loci (than the SNP sites), each specific Y-haplogroup generally contains multiple Y-STR haplotypes. However, lack of recombination in the NRY-segment of Y-chromosome produces an association of Y-haplogroups with Y-STR haplotypes, this is what allows for Y-haplogroup prediction from Y-STR haplotypes by using specific computational algorithms[4,5]. Further, the initial geographic association of Y-haplogroups (dictated by the timing and geographic origin of the haplogroup-defining SNP mutations) gets diluted with male migration across populations, so that the self-declared race/ethnicity of modern populations may not always be a reliable indicator of patrilineal genetic ancestry[22].

Recently large amounts of population data have been generated solely from Y-STR-based haplotypes[9,23]. The minimum number of loci needed for accurate haplogroup prediction is important for the use of the Haplogroup Predictor software. Comprehensive population analyses addressing this issue are not yet available. Since haplogroups are groups of haplotypes sharing a common ancestral SNP allele, it is important to determine how the evolutionary age of the haplogroup affects haplogroup prediction. Polymorphism of haplotypes should affect the outcome of the haplogroup prediction. Homogenous and heterogeneous (i.e., admixed) populations can be operationally defined with those of without and with evidence of admixture, respectively. Numerous studies examined the topic of admixture from the perspective of the entire genome [25,26,27]. Relatively, only a few recent studies exist that focused on the diversity of

the ancestral male lineage as defined by the haplogroup, two of those being studies on U.S. and Cuban populations[9,24]. The effect of small population size on haplogroup diversity was analyzed with the populations in Table 2.1.

Table 2.1: Data Selection for Specific AIM II

Population	Sample size
Large Populations from Africa	
Nigeria (Yoruba)	n=81
Benin (Beninese)	n=51
South Africa (Xhosa)	n=114
Large Populations from Asia	
Southern China (Han)	n=30
Xuanwei (Han)	n=145
South Korea (Korean)	n=300
Chengdu (Han)	n=100
Ibaraki (Japanese)	n=163
Large Populations from Europe	
Central England (English)	n=81
Southern England (English)	n=114
Wales (Welsch)	n=118
Ireland (Irish)	n=31
Mecklenburg-Vorpommern, Germany [German]	n=176
Stuttgart, Germany [German]	n=118
Small Populations	
North Alaska (Inupiat)	n=148
West Alaska (Yupik)	n=141
Central Alaska, USA (Athapaskan)	n=152
Total haplotypes	n=2,063

Furthermore, haplogroup diversities of admixed populations were calculated and compared to the diversities in corresponding ancestral populations (Table 2.2), from which the effect of male migration between populations on the polymorphism of the Y chromosome was assessed.

Lastly, the impact of RM Y-STRs on the determination of patrilineal ancestry was examined in this reserach. The effect of RM Y-STRs on haplotype clustering has been studied and demonstrates poor geographic clustering when the haplotypes consisting solely of rapidly mutating loci are used in lieu of standard STRs which resulted in better clustering by geographic region[23]. Although low levels of population structure have been demonstrated with RM Y-STR analysis[23], the higher mutation rates of RM Y-STRs and their impact on the reliability of prediction of patrilineal ancestry warranted this analysis.

Table 2.2: Data Selection for Specific AIM III and AIM IV

Population	Sample size
Admixed populations	
British Africans	n=171
Rio de Janeiro (Admixed Brazilian)	n=123
San Paulo (Admixed Brazilian)	n=120
Corresponding ancestral populations	
Central England (English)	n=81
Southern England (English)	n=114
Mecklenburg-Vorpommern, Germany [German]	n=176
Stuttgart, Germany [German]	n=118
Aragon, Spain [Spanish]	n=200
Asturias, Spain [Spanish]	n=256
Nigeria (Yoruba)	n=81
Benin (Beninese)	n=51
Sao Gabriel de Cachoeira, Brazil [Native American]	n=61
Total haplotypes	n=1,552
U.S. Populations (U.S. Y-STR Database release 4.1)	
U.S. Hispanic	n=952
African Americans	n=1297
U.S. Caucasian	n=1479
U.S. Asian	n=649
Native American	n=882
Total haplotypes	n=5,259

The potential to use haplogroup prediction on the Y-STR profiles may provide insight in investigative leads with respect to criminal or civil concerns. This background information generated the central hypothesis of this research, namely, polymorphism of Y-STR haplotypes in populations is governed by patrilineal ancestry combined with effects of male migration.

CHAPTER III

MATERIALS AND METHODS

Data Compilation

The data utilized in this study were compiled from three sources. The first being PowerPlex®Y23 haplotypes with known haplogroups from prior data available as supplementary material in the publication by Hallast, et al[8]. The second source used was the PowerPlex®Y23 haplotypes from numerous worldwide populations available as supplementary material in the publication by Purps, et al[9]. The last source was the PowerPlex®Y23 haplotypes of major U.S. populations which comprise release 4.1 of the U.S. Y-STR database[10].

Data Analysis

All haplotypes that were subjected to haplogroup prediction with the program Haplogroup Predictor[4,5] used the initial suggested fitness score of 40 and probability of 95%[15]. One locus, DYS549, is not compatible with the program. Two loci, DYS643 and DYS533, are not supported with data in all haplogroups[28]. As a consequence these three loci (DYS549, DYS643 and DYS533) were removed from haplotype-definitions in haplogroup predictions; this resulted in the maximum haplotype definition (MXHT) as identified in Tables 1.1 and 3.1 being comprised of 20 loci. The same samples had their haplotypes reduced to the Yfiler®, PowerPlex®Y (PPY®12), SWGDAM and minimum haplotype definitions (MHT) when addressing the question of haplotype definitions effect on haplogroup predictability.

Table 3.1: Haplotype-definitions for Specific AIM IV and mutation rates

Loci	Mutation rate[29]	Maximum Haplotype Definition sans RM Y-STR (MXHT Minus)	Maximum Haplotype Definition (MXHT)	PowerPlex® Y23 sans RM Y-STR (PP®Y23 Minus)	PowerPlex® Y23 (PP®Y23)
DYS19	$2.20x10^{-3}$	✓	✓	✓	✓
DYS389I	$2.82x10^{-3}$	✓	✓	✓	✓
DYS389II	$3.62x10^{-3}$	✓	✓	✓	✓
DYS390	$2.00x10^{-3}$	✓	✓	✓	✓
DYS391	$2.63x10^{-3}$	✓	✓	✓	✓
DYS392	$0.43x10^{-3}$	✓	✓	✓	✓
DYS393	$1.12x10^{-3}$	✓	✓	✓	✓
DYS385a	$2.20x10^{-3}$	✓	✓	✓*	✓*
DYS385b		✓	✓	✓*	✓*
DYS438	$0.43x10^{-3}$	✓	✓	✓	✓
DYS439	$4.95x10^{-3}$	✓	✓	✓	✓
DYS437	$1.21x10^{-3}$	✓	✓	✓	✓
DYS448	$1.23x10^{-3}$	✓	✓	✓	✓
DYS456	$4.31x10^{-3}$	✓	✓	✓	✓
DYS458	$6.77x10^{-3}$	✓	✓	✓	✓
DYS635	$3.47x10^{-3}$	✓	✓	✓	✓
YGATAH4	$2.51x10^{-3}$	✓	✓	✓	✓
DYS481	$5.12x10^{-3}$	✓	✓	✓	✓
DYS570**	$12.12x10^{-3}$		✓		✓
DYS576**	$14.46x10^{-3}$		✓		✓
DYS643	$0.86x10^{-3}$			✓	✓
DYS549	$3.57x10^{-3}$			✓	✓
DYS533	$5.01x10^{-3}$			✓	✓

*DYS385a/b are excluded from Network analysis
**DYS570 and DYS576 are the rapidly mutating Y-STRs[17]

Haplogroups assigned from the prediction algorithm are reported as the ISOGG 2012 nomenclature[28,30] and were concatenated to their corresponding major-haplogroups according to Table 3.2 for the purpose of analysis. Haplogroup diversity was calculated as $\hat{h} = 1 - \sum_i \hat{p}_i^2$ where $\hat{h}$ is the estimated haplogroup diversity and $\hat{p}_i$ is the observed relative frequency of each

major-haplogroup in a population. STRUCTURE analysis[11,12] used a burnin length of 100,000 was used with 3000 Markov chain Monte Carlo (MCMC) repetitions. In the analysis of the effect of the inclusion of the RM Y-STRs on population clustering the MXHT and MXHT minus haplotype-definitions were used. Network[13,14] analysis utilized the PowerPlex®Y23 (PP®Y23) and PP®Y23 minus haplotype-definitions with DYS385a/b loci excluded due to the inability to assign alleles to loci a or b, as defined in Table 3.1. The networks were drawn with all loci equally weighted and an epsilon value of 0.

Table 3.2: Haplogroups and their Major-haplogroups

Haplogroup	Major Haplogroup
C3	C
E1a E1b1a E1b1b	E
G1 G2a G2c	G
H	H
I1 I2a (xI2a1) I2a1 I2b (xI2b1) I2b1	I
J1 J2a4b J2a4h J2a4 (x bh) J2b	J
L	L
N	N
R1a R1b R2	R
Q	Q
T	T
O2 O3	O

CHAPTER IV

RESULTS

AIM I: Factors affecting Predictability of Haplogroups from Y-STR haplotype data

This AIM was studied by utilizing data from Hallast, et al[8], which contain 201 PP®Y23 haplotypes with known haplogroups from prior data. Thirty haplotypes belonged to major-haplogroups (A, B, D, K, etc) that are not supported by the prediction software were run through the software but not included in any analysis. This left 171 haplotypes with known major-haplogroups supported by Haplogroup Predictor. Assessment of the results was complex given the haplogroup-nomenclature issues surrounding the use of either the 2012 ISOGG nomenclature or the short form that includes a terminal SNP[28,30]. As a consequence, when the assigned haplogroup from the prediction software shared the major-haplogroup of the sample, the assignment was categorized as correct. Haplotypes that were known to be of haplogroups which are not supported did yield some haplogroup predictions. However, the assignments yielded from this analysis could not be correct as a result of the lack of support for the known haplogroups.

Table 4.1 Counts of correctly assigned major-haplogroups

Haplotype Definition	C	E	G	H	I	J	L	N	O	Q	R	T	Total Correct Assignments
MXHT (20 loci)	0	25	13	2	13	6	2	1	14	4	39	4	123
Yfiler® (17 loci)	0	25	13	2	14	6	2	1	14	4	40	4	125
PP®Y12 (12 loci)	1	25	11	2	11	5	1	0	13	4	38	4	115
SWGDAM (11 loci)	2	25	11	2	12	7	2	0	12	4	37	4	118
MHT (9 loci)	1	24	11	2	10	2	2	0	9	2	32	1	96
Total Known	9	29	14	5	15	19	2	5	19	5	45	4	171

Table 4.1 lists the number of correctly assigned major-haplogroups; that is out of 9 haplotypes belonging to haplogroup C one was correctly assigned to C under the MHT definition, two were

correctly assigned to C under the SWGDAM definition, etc. Table 4.2 lists the proportion of properly assigned haplogroups by the haplogroup predictor according to their haplogroup definition; that is out of 29 haplotypes with a known haplogroup E, 25 were correctly assigned that haplogroup under the MXHT definition, resulting in proportion of 0.86. In general, there was an increase in the number of haplotypes that were assigned haplogroups as the haplotype-definition was expanded to include more loci as seen in the total column of Table 4.1.

Table 4.2 Proportions of correctly assigned major-haplogroups

Haplotype Definition	C	E	G	H	I	J	L	N	O	Q	R	T	Total
MXHT	0.00	0.86	0.93	0.40	0.87	0.32	1.00	0.20	0.74	0.80	0.87	1.00	0.72
Yfiler®	0.00	0.86	0.93	0.40	0.93	0.32	1.00	0.20	0.74	0.80	0.89	1.00	0.73
PP®Y12	0.11	0.86	0.79	0.40	0.73	0.26	0.50	0.00	0.68	0.80	0.84	1.00	0.67
SWGDAM	0.22	0.86	0.79	0.40	0.80	0.37	1.00	0.00	0.63	0.80	0.82	1.00	0.69
MHT	0.11	0.83	0.79	0.40	0.67	0.11	1.00	0.00	0.47	0.40	0.71	0.25	0.56
Total Known	9	29	14	5	15	19	2	5	19	5	45	4	171

However predictability, meaning the proportion of correct haplogroup assignment, was dependent on the haplogroup that the haplotype belonged to. Major-haplogroups E, G and R demonstrated high levels of predictability. In contrast, major-haplogroups C and N demonstrated the lowest levels of predictability. Table 4.3 lists the counts of incorrect haplogroup assignments. Five haplotypes out of 171 received incorrect major-haplogroup assignments at the PP®Y12 definition, two haplotypes received incorrect assignments at the SWGDAM and Yfiler® definitions. This resulted in an error rate of 4.17%, 1.67% and 1.57% for the PP®Y12, Yfiler® and SWGDAM definitions, respectively. Note that non-assignments are not included in the calculation for error rate or correct assignment proportions as neither a correct nor incorrect haplogroup assignment was made. Haplotypes with unsupported known haplogroups were not used in calculating the error rate. With these caveats, these analyses showed that even with 12

Y-STR loci in the haplotype-definition (namely the PP®Y12 kit), the haplogroup prediction error rate at the major-haplogroup level does not exceed 5%. With more enhanced haplotype definition (i.e., Yfiler® and SWGDAM definitions) such error rates of haplogroup prediction are below 2%.

Table 4.3: Incorrect haplogroup predictions

	Incorrect Predictions	Total Assignments	Error rate
MXHT	0	123	0
Yfiler®	2	127	0.0157
PP®Y12	5	120	0.0417
SWGDAM	2	120	0.0167
MHT	0	96	0

AIM II: Confounding effects of evolutionary age of populations on haplogroups diversity and its relationship with number of loci in the Y-STR haplotype

The data set used for AIM II consisted of 2,063 haplotypes identified in Table 2.1 acquired from Purps et al[9]. Table 4.4 lists the results from the aggregated populations, meaning the pooled populations listed in Table 2.1, as the number of assigned haplogroups for each major-haplogroup as previously defined in Table 3.2; that is 24 of 2,063 haplotypes were assigned the major-haplogroup C under the MHT definition. The total number of haplotypes which received a haplogroup assignment is listed under each haplotype-definition. In total, there was a substantial increase in the number of haplotypes with assigned haplogroups from a minimum of 1,227 for the MHT definition, up to 1,556 for the Yfiler® definition. There was a slight decrease in assigned haplogroups from the Yfiler® definition to the MXHT definition from 1,556 assignments to 1,532 assignments. These results suggest that haplogroup predictability increased as the number of loci comprising the haplotype-definition increased. However,

increased predictability in relation with the increased number of loci was not equally distributed among all major haplogroups.

Table 4.4: Counts of assigned major-haplogroups from worldwide populations

Aggregate Populations	Haplotype definition				
Major-haplogroup	MHT	SWGDAM	PP®Y12	Yfiler®	MXHT
C	24	42	54	52	56
E	195	219	222	231	225
G	13	11	11	15	15
H	5	1	1	1	1
I	80	88	92	97	112
J	15	17	17	20	17
L	2	1	2	1	0
N	17	17	20	18	16
O	280	342	372	448	438
Q	147	160	160	172	153
R	444	495	500	495	495
T	5	5	5	6	4
Not Predicted	836	665	607	507	531
Total Predicted	1227	1398	1456	1556	1532
Total	2063	2063	2063	2063	2063

Haplogroups C, E, I, O, Q and R all showed an increase in the number of haplotypes assigned with haplogroups. Haplogroup predictability of haplotypes assigned as major haplogroups C and O benefited the most from the increased loci, both in individual populations and in aggregate. At the Yfiler® loci panel 53.8% and 37.5% more haplotypes were predicted with the C and O major haplogroups than at the MHT loci panels, respectively. Major-haplogroup H and L had a decrease in haplogroups assigned, but that may have resulted from the low sample size of assigned haplogroups. Generally, haplogroup assignment in each population increased as the number of loci comprising the haplotype-definition increased. A notable exception to this was the Stuttgart, German population where the predictability of the R major-haplogroup decreased as the number of loci increase with a substantial drop in the number of haplotypes assigned

within the R major-haplogroup in haplotypes larger than the MHT. Table 4.5 lists the number of assigned haplotypes for major-haplogroup R and there was a decrease in this assignment immediately after the MHT definition from 51 to 39 in the SWGDAM definition.

Table 4.5: Counts of assigned haplotypes to the R major-haplogroup from Stuttgart, Germany

	Haplotype definition				
Major-Haplogroup	MHT	SWGDAM	PP®Y12	Yfiler®	MXHT
R	51	39	36	37	37

Table 4.6 lists the haplogroup diversities of the individual populations of Table 2.1 across each haplotype-definition, in which a haplogroup diversity of 0 indicates that all assigned haplogroups were the same.

Table 4.6: Haplogroup diversity of worldwide populations

	Haplotype definition				
Population	MHT	SWGDAM	PP®Y12	Yfiler®	MXHT
Benin (Beninese)	0.000	0.044	0.044	0.000	0.044
Nigeria (Yoruba)	0.000	0.000	0.026	0.000	0.000
South Africa (Xhosa)	0.157	0.058	0.056	0.050	0.074
Central Alaska (Athapaskan)	0.690	0.687	0.693	0.703	0.719
North Alaska (Inupiat)	0.451	0.491	0.489	0.480	0.514
West Alaska (Yupik)	0.454	0.469	0.499	0.463	0.568
Central England (English)	0.364	0.345	0.326	0.383	0.385
Ireland (Irish)	0.165	0.133	0.133	0.137	0.077
Mecklenburg, Germany (German)	0.504	0.496	0.497	0.516	0.516
South England (English)	0.423	0.440	0.444	0.457	0.442
Stuttgart, Germany (German)	0.590	0.637	0.646	0.643	0.667
Wales (Welsch)	0.262	0.247	0.247	0.259	0.263
Chengdu China (Han)	0.230	0.371	0.348	0.223	0.169
Ibaraki, Japanese (Japanese)	0.098	0.172	0.187	0.214	0.207
South China (Han)	0.000	0.142	0.142	0.083	0.087
South Korea (Korean)	0.232	0.244	0.288	0.275	0.276
Xuanwei China (Han)	0.129	0.131	0.208	0.123	0.160

The highest levels of haplogroup diversity were observed in the Athapaskan population across all haplotype definitions and the lowest in the Yoruba and Beninese populations. Table 4.7 is the weighted average of haplogroup diversities for the populations as defined by their geographic location.

Table 4.7: Average of haplogroup diversities by geographic area

	Haplotype definition				
Geographic Area	MHT	SWGDAM	PP®Y12	Yfiler®	MXHT
Africa	0.058	0.032	0.042	0.020	0.040
Alaska	0.538	0.563	0.573	0.554	0.610
Europe	0.420	0.410	0.410	0.430	0.433
Asia	0.186	0.227	0.260	0.221	0.220

The Alaskans showed the trend of largest haplogroup diversity, largely due to patrilineal diversity in them as a result of male migration. The Europeans and Alaskans have diverse patrilineal ancestry when compared to the Africans and Asians. There was considerable heterogeneity in the patrilineal ancestry in Alaska and Europe. In contrast, patrilineal ancestry was observed to be comparatively more homogeneous in Africa and Asia, as reflected in the lower levels of haplogroup diversity in these populations.

AIM III: Effect of admixture on Haplogroup Diversity and Haplogroup Predictability

Populations used in this AIM are identified in Table 2.2, consisting of haplotypes from Purps et al.[9] and the U.S. Y-STR database. Three sets of admixed populations were analyzed for their haplogroup diversity; the British Africans, two Admixed Brazilian populations and U.S. Hispanics, along with their corresponding ancestral populations. The British African population in Table 4.8 had a substantially greater haplogroup diversity than the corresponding ancestral African populations, but had diversity levels similar to the English populations.

Table 4.8: Haplogroup diversities of British Africans and ancestral populations

Admixed Population	British African			
Haplogroup Diversity	0.369			
Corresponding Ancestral Population	Central England (English)	Nigeria (Yoruba)	South England (English)	Benin (Beninese)
Haplogroup Diversity	0.385	0.000	0.442	0.044

The two Admixed Brazilian populations in Table 4.9 both had slightly higher haplogroup diversities than the corresponding ancestral populations with the exception of the Stuttgart, German population.

Table 4.9: Haplogroup diversities of Admixed Brazilians and ancestral populations

Admixed Population	Sao Paulo (Admixed Brazilian)			Rio de Janeiro (Admixed Brazilian)	
Haplogroup Diversity	0.655			0.579	
Corresponding Ancestral Population	Native American (Brazilian)	Argon, Spain (Spanish)	Asturias Spain (Spanish)	Mecklenburg, Germany (German)	Stuttgart, Germany (German)
Haplogroup Diversity	0.567	0.448	0.500	0.516	0.667

Table 4.10 lists the haplogroup diversities of populations in the United States. The haplotypes analyzed from the U.S. Y-STR database indicate that the haplogroup diversity for the U.S. Hispanic population is almost identical to the Native Americans; both are substantially higher than the U.S. Caucasian and African-American population.

Table 4.10: Haplogroup diversities of U.S. Hispanic and ancestral populations

Admixed Population	U.S. Hispanic		
Haplogroup Diversity	0.719		
Corresponding Ancestral Population	Native American	U.S. Caucasian	African American
Haplogroup Diversity	0.711	0.488	0.435

The three analyzed population groups shared a common feature with the admixed populations demonstrating higher haplogroup diversities than at least one of their ancestral populations. When there is an admixture there is an increased haplogroup diversity with respect to at least one of the ancestral populations in the British African population. This is consistent with AIM II where it was suggested the African haplogroup diversity was more homogenous. Similarly, the Admixed Brazilians and U.S. Hispanics had higher haplogroup diversity than more than one of their corresponding ancestral populations. In other words, the general common feature of admixture appears to be that gene migration through males brings in added haplogroup diversity in the admixed population, more conspicuously observed in relation to the haplogroup diversity of the ancestral population(s) that is (are) of relatively more homogeneous with respect to patrilineal ancestry.

AIM IV: Effect of rapidly mutating Y-STRs (RM Y-STR) on prediction of patrilineal ancestry

Populations in Table 2.2 were used in AIM IV when addressing the effect of RM Y-STRs on haplogroup prediction and population clustering. Table 4.11 lists the aggregate number of assigned haplogroups from the PP®Y23 data from the U.S. Y-STR database run as both the MXHT and MXHT Minus haplotype-definitions as defined in Table 3.1. There was a slight decrease in haplogroups with assigned haplotypes when the two RM Y-STRs were not included as part of the haplotype-definition. Only the African Americans had an increase of assigned haplogroups when the two RM Y-STRs were removed from the haplotype definition. The other four populations; U.S. Hispanics, U.S. Caucasian, Native Americans and U.S. Asians showed a slight decrease in haplogroup assignment when the two RM Y-STRs were removed. In

aggregate there was no increase in assignment of haplogroup by Haplogroup Predictor as a consequence of removing the RM Y-STRs from the haplotype.

Table 4.11: Counts of assigned haplotypes from American populations with and without RM Y-STRs

U.S. Y-STR Database	Haplotype definition	
	MXHT	MXHT Minus
C	35	36
E	1023	1025
G	104	106
H	21	21
I	392	385
J	140	134
L	22	22
N	23	23
O	259	254
Q	391	393
R	1870	1856
T	30	28
Not Assigned	949	976
Total Assigned	4310	4283
Total	5259	5259

Aggregate number of assigned major-haplogroups for the U.S. Y-STR database release 4.1 by each haplogroup-definition as defined in Table 2

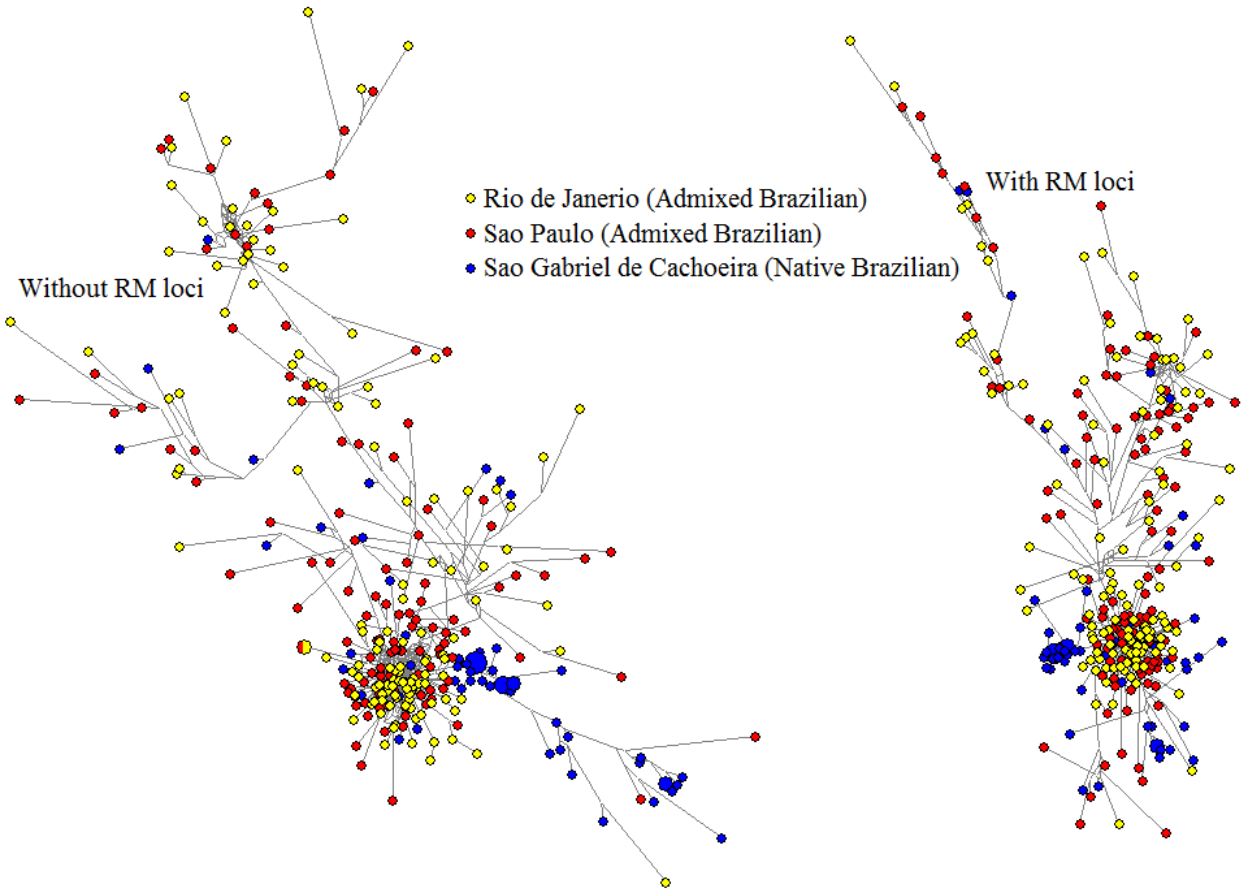
Table 4.12 lists the aggregate number of assigned haplogroups from the worldwide populations listed in Table 2.2. Similar to the U.S. Y-STR database, these populations demonstrated no increase in assignment of haplogroups as a consequence of removing the RM Y-STRs from the haplotype. Only two populations, the admixed Brazilian of Rio de Jaeniro and Yoruba of Nigeria, of the twelve analyzed had an increase in haplogroup predictability as a result of removing the RM Y-STRs from the haplotypes. Median joining network (MJN) analysis used the PowerPlex®Y23 and PowerPlex®Y23 Minus haplotypes defined in Table 3.1.

Table 4.12: Counts of assigned haplotypes from worldwide populations with and without RM Y-STRs

World Wide Populations	Haplotype definition	
	MXHT	MXHT Minus
C	0	0
E	321	322
G	33	31
H	2	2
I	147	141
J	51	51
L	3	3
N	7	7
O	3	3
Q	39	38
R	686	682
T	22	18
Not Assigned	238	254
Total Assigned	1314	1298
Total	1552	1552

Figure 1 illustrates the difference in population clustering among these two haplotype variations for the Native and admixed Brazilian populations. The MJN on the left had the two RM Y-STRs removed from the haplotype definitions and the MJN on the right includes the two RM Y-STRs. The Native Brazilian population of Sao Gabriel de Cachoeira demonstrated improved haplotype clustering when the two RM Y-STRs are removed. The main cluster had little difference, most of the change occurred around the periphery. Figure 2 are the STRUCTURE barplots of inferred population clusters of the three Brazilian populations. The barplot on left includes the two RM Y-STRs and the barplot on the bottom has excluded the two RM Y-STRs using the PP®Y23 and PP®Y23 Minus haplotype definitions as defined in Table 3.1 including DYS385a/b as a single loci.

Figure 1
MJN of PP®Y23 haplotypes from three Brazilian populations with and without rapidly mutating loci



The inferred population clusters show more homogenous clustering with the Native Brazilians and greater heterogeneity with the Admixed Brazilians. There was nearly no difference in the barplots between the two haplotypes. Table 4.13 lists the proportions of assigned population clusters from the barplots seen in Figure 2. These proportions of assigned haplogroups for the nine populations as visually represented in Figure 2 are nearly identical irrespective of the inclusion of the RM Y-STRs in the haplotype-definition.

Figure 2
STRUCTURE barplots PP®Y23 data from nine populations with and without rapidly mutating loci

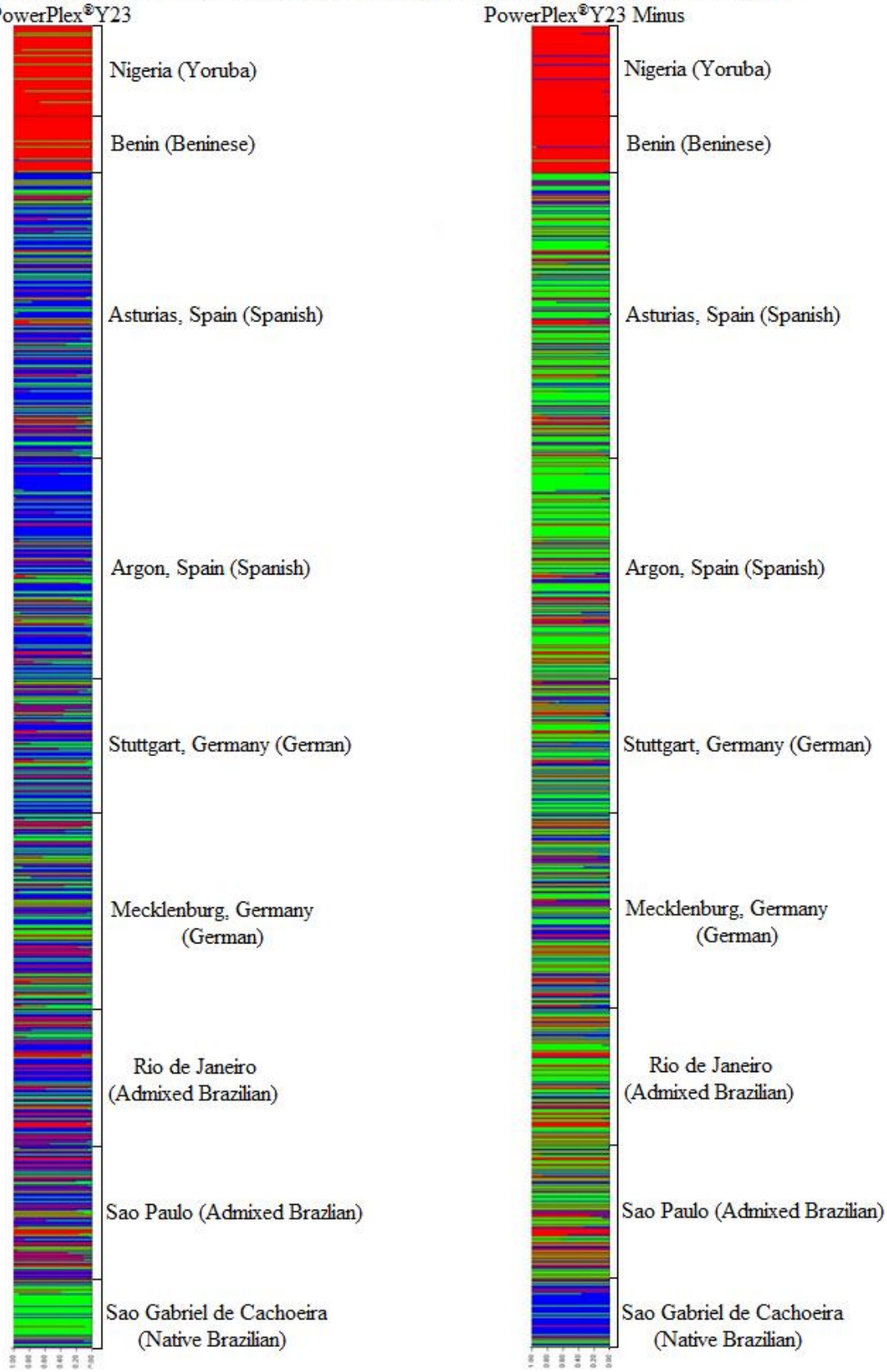


Table 4.13: Inferred population clusters of Admixed Brazilians and corresponding ancestral populations

PP®Y23 Population	Red	Blue	Green	Individuals
Sao Gabriel de Cachoeria (Native American)	0.105	0.179	0.716	61
Sao Paulo, Admixed Brazilian	0.345	0.449	0.206	120
Rio de Janeiro (Admixed Brazilian)	0.299	0.528	0.173	123
Mecklenburg, Germany	0.222	0.43	0.348	176
Stuttgart, Germany	0.194	0.517	0.29	118
Argon, Spain	0.133	0.659	0.209	200
Asturias, Spain	0.157	0.593	0.25	256
Benin (Beninese)	0.905	0.020	0.075	51
Nigeria (Yoruba)	0.909	0.000	0.091	81

PP®Y23 Minus Population	Red	Green	Blue	Individuals
Sao Gabriel de Cachoeria (Native American)	0.115	0.174	0.711	61
Sao Paulo, Admixed Brazilian	0.387	0.449	0.164	120
Rio de Janeiro (Admixed Brazilian)	0.335	0.528	0.137	123
Mecklenburg, Germany	0.265	0.43	0.305	176
Stuttgart, Germany	0.231	0.517	0.253	118
Argon, Spain	0.21	0.653	0.137	200
Asturias, Spain	0.211	0.594	0.194	256
Benin (Beninese)	0.959	0.021	0.021	51
Nigeria (Yoruba)	0.957	0.000	0.043	81

Table 4.14a: Proportion of assigned haplogroups for the admixed Brazilian and corresponding ancestral populations from the MXHT haplotype definition

	Major-Haplogroup									
MXHT	E	G	I	J	N	Q	R	T	Not Assigned	Total
Sao Paulo (Admixed Brazilian)	0.192	0.050	0.075	0.025	0.000	0.017	0.433	0.033	0.175	120
Rio de Janeiro (Admixed Brazilian)	0.154	0.024	0.089	0.041	0.000	0.000	0.472	0.000	0.220	123
Sao Gabriel de Cachoeira (Native Brazilian)	0.066	0.033	0.000	0.049	0.000	0.574	0.164	0.033	0.082	61
Argon, Spain*	0.050	0.020	0.100	0.055	0.000	0.000	0.630	0.005	0.135	200
Asturias, Spain*	0.063	0.023	0.078	0.059	0.000	0.000	0.598	0.039	0.137	256
Mecklenburg, Germany*	0.023	0.028	0.193	0.023	0.040	0.000	0.608	0.000	0.080	176
Stuttgart Germany*	0.076	0.017	0.110	0.059	0.000	0.000	0.314	0.017	0.398	118
Nigeria	0.864	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.136	81
Benin	0.843	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.137	51

*Major Haplogroups H, L and O are excluded from this table. H had one assignment in Argon and Stuttgart, L had one assignment in Asturias and O had one in Mecklenburg

Tables 4.14a and 4.14b list the proportions of assigned major-haplogroups under the MXHT and MXHT Minus haplotypes for the nine populations used in the STRUCTRE analysis. There was minimal change in the haplogroup assignments when the haplotypes lacked the RM Y-STRs. There was an increase in the E major-haplogroup and the main European major-haplogroups R and I in the admixed populations when compared to the Native Brazilian population.

Table 4.14b: Proportion of assigned haplogroups for the admixed Brazilian and corresponding ancestral populations from the MXHT Minus haplotype definition

	Major-Haplogroup									
MXHT minus	E	G	I	J	N	Q	R	T	Not Assigned	Total
Sao Paulo (Admixed Brazilian)	0.183	0.050	0.075	0.033	0.000	0.017	0.425	0.033	0.183	120
Rio de Janeiro (Admixed Brazilian)	0.163	0.024	0.089	0.049	0.000	0.000	0.472	0.000	0.203	123
Sao Gabriel de Cachoeira (Native Brazilian)	0.066	0.033	0.000	0.049	0.000	0.574	0.148	0.033	0.098	61
Argon, Spain*	0.055	0.020	0.100	0.050	0.000	0.000	0.630	0.005	0.135	200
Asturias, Spain*	0.066	0.020	0.078	0.059	0.000	0.000	0.602	0.023	0.148	256
Mecklenburg, Germany*	0.023	0.028	0.188	0.023	0.040	0.000	0.602	0.000	0.091	176
Stuttgart Germany*	0.068	0.008	0.102	0.059	0.000	0.000	0.314	0.017	0.424	118
Nigeria	0.877	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.123	81
Benin	0.824	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.157	51

*Major Haplogroups H, L and O are excluded from this table. H had one assignment in Argon and Stuttgart, L had one assignment in Asturias and O had one in Mecklenburg

Figure 3 illustrates the difference in population clustering among the PP®Y23 and PP®Y23 minus variations of haplotypes from the British African population and its corresponding ancestral populations (South English, Central English, Beninese and Yoruba); the MJN on the left had the two RM Y-STRs removed from the haplotype definitions while that on the right includes the two RM Y-STRs. Both the English and African populations tend to cluster tighter when the RM Y-STRs are deleted from the haplotypes. Similar to Figure 1, changes in

clustering occurred more along the periphery of the network and the major clusters yielded little change.

Figure 3

MJN of PP®Y23 haplotypes from five populations with and without rapidly mutating loci

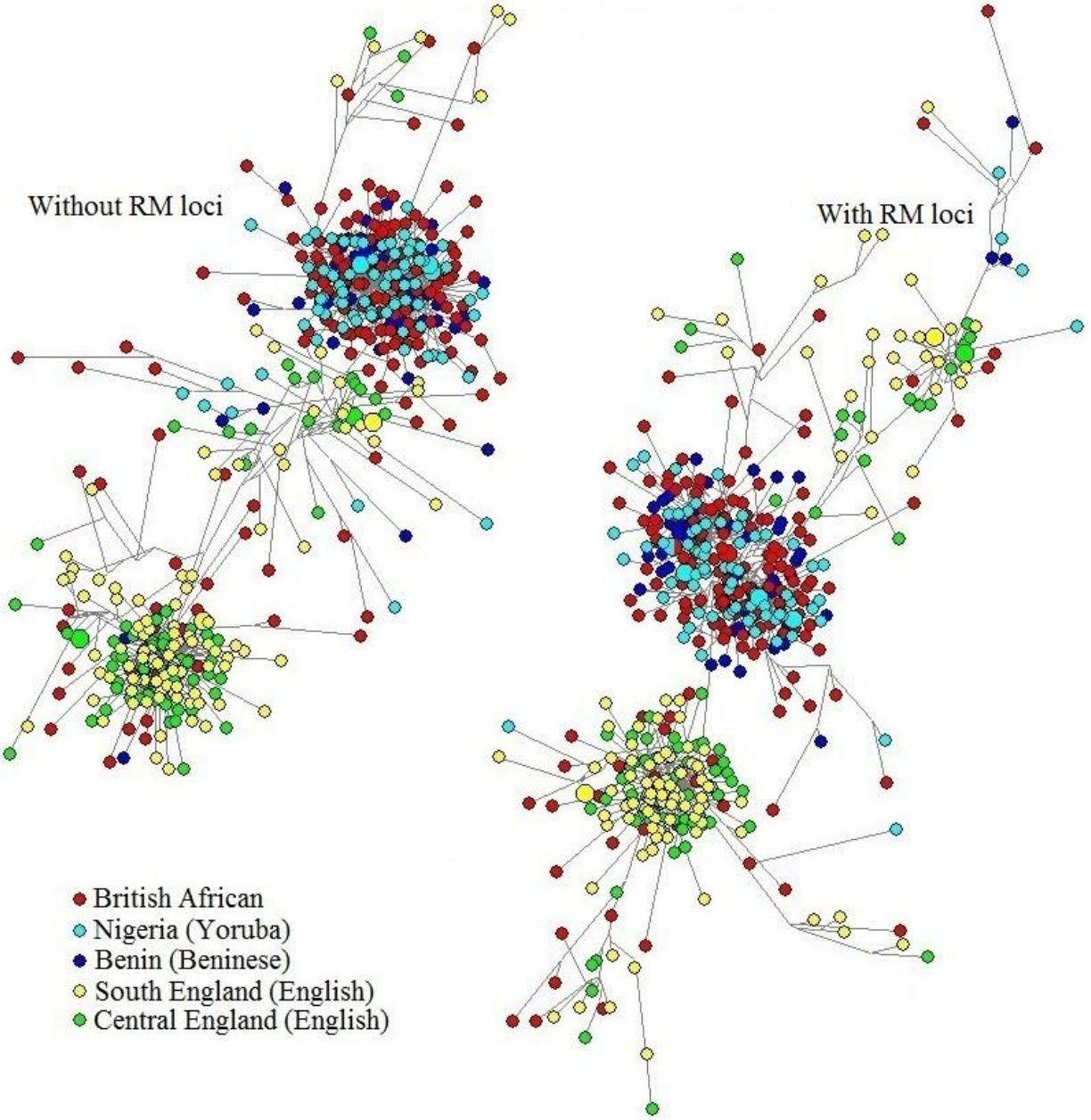


Figure 4 is the barplot of the British African and corresponding ancestral populations, it illustrates minimal difference between inferred population clusters with and without inclusion of

the RM Y-STRs. Table 4.15 lists the proportions of the inferred population clustering in the Figure 4 barplot. There was minimal difference between the inferred population clusters based on the inclusion or exclusion of the RM Y-STRs.

Figure 4
STRUCTURE barplots of PP®Y23 data from five populations with and without rapidly mutating loci

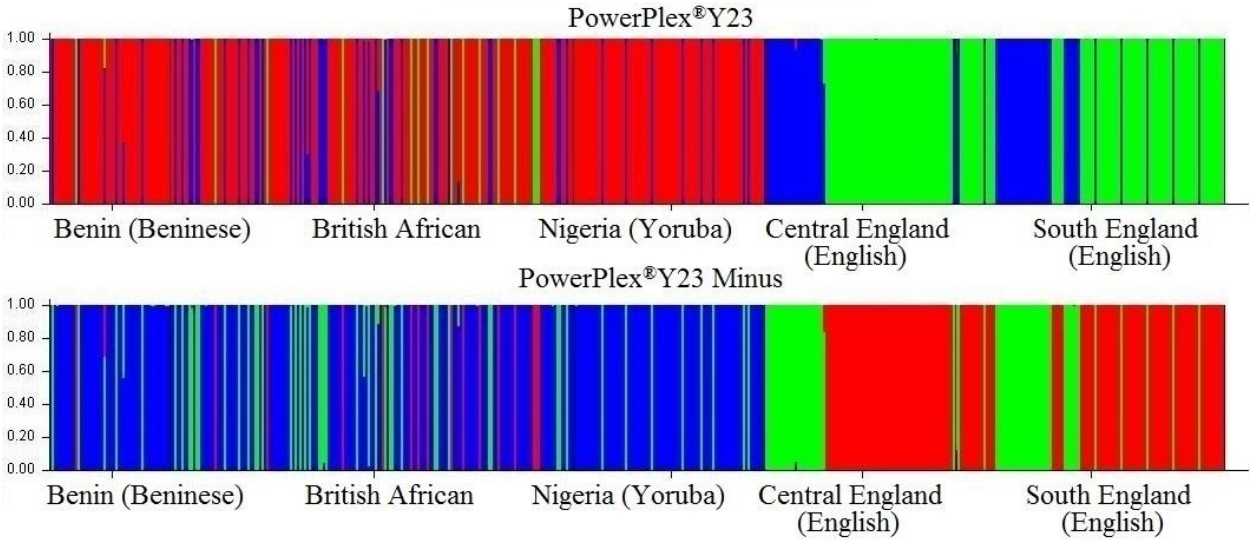


Table 4.15: Inferred population clusters of British Africans and corresponding ancestral populations

PP®Y23	Inferred Population Cluster			
	Red	Green	Blue	Individuals
Benin (Beninese)	0.875	0.023	0.102	51
British African	0.689	0.088	0.223	171
Nigeria (Yoruba)	0.901	0	0.099	81
Central England (English)	0.001	0.67	0.329	81
South England (English)	0.001	0.658	0.341	114
PP®Y23 Minus	Inferred Population Cluster			
	Blue	Red	Green	Individuals
Benin (Beninese)	0.874	0.026	0.101	51
British African	0.698	0.088	0.214	171
Nigeria (Yoruba)	0.901	0	0.099	81
Central England (English)	0.001	0.669	0.331	81
South England (English)	0.001	0.658	0.341	114

Table 4.16 lists the proportions of assigned major-haplogroups with the MXHT and MXHT minus haplotypes. There was minimal change in the haplogroup assignments when the haplotypes lacked the RM Y-STRs.

Table 4.16: Proportion of assigned major-haplogroups of British African and corresponding ancestral populations

	Major-Haplogroup									
MXHT	E	G	I	J	L	Q	R	T	Not Assigned	Total
Benin (Beninese)	0.843	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.137	51
Nigeria (Yoruba)	0.864	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.136	81
British African*	0.702	0.006	0.029	0.006	0.012	0.012	0.105	0.012	0.105	171
Central England (English)	0.000	0.025	0.185	0.012	0.000	0.000	0.691	0.000	0.086	81
South England (English)	0.026	0.018	0.175	0.009	0.000	0.000	0.596	0.009	0.167	114
	Major-Haplogroup									
MXHT Minus	E	G	I	J	L	Q	R	T	Not Assigned	Total
Benin (Beninese)	0.824	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.157	51
Nigeria (Yoruba)	0.877	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.123	81
British African*	0.702	0.006	0.018	0.000	0.012	0.006	0.099	0.012	0.135	171
Central England (English)	0.000	0.025	0.173	0.012	0.000	0.000	0.667	0.000	0.123	81
South England (English)	0.026	0.018	0.167	0.009	0.000	0.000	0.605	0.009	0.167	114

*The British African population had two haplotypes assigned major-haplogroup O in the MXHT and MXHT minus population with the relative frequency of 0.012

The proportion of non-African haplogroups; namely the non-E haplogroups are not the same for British Africans and African populations, which contributes to the haplogroup diversity. The

British African haplotypes demonstrated more concordance with the African populations of Benin and Nigeria than the English in the MJN's (Fig. 3), STRUCTURE (Fig. 4) and haplogroup prediction (Table 4.16).

CHAPTER V

DISCUSSION AND CONCLUSION

Part I: Accuracy

Out of the 171 haplotypes with known major-haplogroups analyzed here, the predictability of a sampled haplotype was dependent on both the known haplogroup and the haplotype-definition. Accuracy of haplogroup assignment was largely dependent on the haplogroup from which the haplotype was derived from. Major-haplogroups E, G, I and R were the best performing haplogroups with respect to proportion predicted (>80%). The major-haplogroups T and L also demonstrated high assignment proportions. However, their sample sizes are too small to draw any general conclusion. Nonetheless, these results are largely concordant with previous findings and statements by the programs author that the main limitation of haplogroup prediction is the availability of adequate allele frequency data[5,7]; suggesting that major-haplogroups with the highest predictability are those the best available allele frequency. Importantly, incorrect haplogroup assignments were low (Table 4.3). Incorrect haplogroup assignments were highest under the PP®Y12 haplotype-definition with 5 haplotypes receiving an incorrect assignment, resulting in 4.17% error. No incorrect assignments were made at either the MXHT or MHT definition. The 4.17% error rate is less than the 4.8% error found in a previous study[7]. However, the error here is not directly comparable as correct classification at the major-haplogroup level was used as the determination of a correct assignment and non-supported haplogroups were not included in the calculations. Of the 30 haplotypes from non-supported major-haplogroups (A, B, D, F, K, M and P) there were 3 to 7 haplotypes which were assigned haplogroups dependent on the haplotype definition. Table 5.1 lists these counts and proportions of assignments by haplotype definition.

Table 5.1: Assignments to haplotypes of non-supported haplogroups

	MHT	SWGDAM	PP®Y12	Yfiler®	MXHT	Total Known
Counts	3	5	7	7	5	30
Proportion	0.100	0.167	0.233	0.233	0.167	

This result indicates that Haplogroup Predictor will assign a supported-haplogroup to a sample from a non-supported-haplogroup. The high proportions of assignments for haplotypes from non-supported haplogroups suggests that caution should be used with prediction algorithms. A note of importance is that haplogroup assignment did not change across any of the haplotype-definitions for any samples in the dataset used for AIM I.

Part II: The effect of haplotype-definition on haplogroup prediction

It has been suggested that the addition of loci to a haplotype will not enhance prediction accuracy[20]. The proportion of assigned haplogroups from the 171 haplotypes with known haplogroups demonstrate greater assignment of haplogroups across haplotype definitions (Table 4.2). The populations used in AIM II (n=2,063) demonstrated an increase in the number of assigned haplogroups as the haplotype-definition included more loci. There was a slight decrease in the number of assigned haplogroups at the MXHT haplotype-definition; this may be a result of the inclusion of the two RM Y-STRs not seen in the other haplotype-definitions; however, this explanation is not concordant with the results discussed in the part III and the MXHT definition also included one standard Y-STR loci not seen in the other haplotype-definitions. Another potential explanation is that some loci may be less informative of haplogroups as a result of mutations which result in a substantial enough divergence from the allele frequency data set which would result in a failure to predict a haplogroup. It has been

found that when Y-STRs mutate large alleles tend to contract and small alleles tend to expand when mutating[31]. That is, when a large allele expands or a small allele contracts, a loci may become less informative with respect to haplogroup assignment. Spearman's rank correlation coefficient of the results in Table 4.4 resulted in a $\rho = 0.900$ with a p-value of 0.05. Based on the analysis here there is a positive correlation of increased predictability as more loci encompassed in the haplotype-definition; although it is not significant at the 5% level.

Part III: The effect of RM Y-STRs on haplogroup predictability and population clustering

The removal of the RM Y-STRs did not result in an increase of assigned haplogroups as initially expected. However, this is concordant with the finding that haplogroup predictability tends to increase as the number of loci encompassed in the haplotype. There was a slight decrease in assigned haplogroups in both the aggregate U.S. populations (n=5,259) and worldwide populations (n=1,552) in Tables 4.11 and 4.12, respectively. The U.S. populations had 81.95% of samples assigned a haplogroup under the MXHT definition and 81.44% of samples assigned a haplogroup with the MXHT minus definition. The worldwide populations had 84.66% of samples assigned a haplogroup under the MXHT definition and 83.63% of samples assigned a haplogroup with the MXHT minus definition. These results suggest that the elimination of RM Y-STRs from a haplotype has minimal impact on haplogroup prediction.

The two admixed Brazilian populations and one of the corresponding ancestral populations, the native Brazilian population, were analyzed in STRUCTURE (Figure 2, Table 4.13). The results demonstrated that the inclusion of two RM Y-STRs had minimal impact on population clustering in STRUCTURE analysis. Simultaneously, haplogroup-assignment of the two admixed Brazilian populations and the corresponding ancestral populations (Table 4.14a and 4.14b) also demonstrated minimal change of haplogroup assignment with removal of the RM

STR loci. However Network analysis (Figure 3) found that the main clusters tended to stay stable relative to the periphery which had the most change in clustering. This is where the native Brazilian population tended to cluster together better when RM Y-STRs were excluded. These results were similar to the admixed British Africans and their corresponding ancestral populations from England and Africa where the main population clusters changed little and most of the change occurred on the periphery. Previous research indicated that haplotypes consisting solely of RM Y-STRs had a substantial impact on population clustering in Network analysis[23]. Yet the effect of two RM Y-STR loci as part of a haplotype consisting primarily of Y-STRs with the standard mutation rate of 10^{-3} had little impact on clustering.

Part IV: The effect of admixture on haplogroup diversity

The British African population (Table 4.8) had a much higher haplogroup diversity than the African component of their ancestral population, but it was slightly lower than the English component of their ancestral population. This suggests that an admixed population can arise from a more homogenous and admixed population, African and English, respectively. However, the contemporaneous sampling of individuals identified as the ancestral populations are not necessarily the same ancestral populations of antiquity in genetic terms. The admixed Brazilian populations (Table 4.9) both have higher haplogroup diversities than four of the five corresponding ancestral populations. In U.S. populations (Table 4.10) there is a similar trend, namely the U.S. Hispanics have a substantially higher haplogroup diversity than two of its ancestral populations, namely African Americans and U.S. Caucasians. The common trend in all the admixed populations is increased haplogroup diversity, also conspicuously seen in some of the ancestral populations, but not all. Which implies that in the admixture process some of the ancestral populations may not be homogenous as a result of early events of admixture.

These results suggest that as distinct populations mate the haplogroup diversity of the resulting population increases.

Part V: Confounding effects of the population size and evolutionary age on haplogroup diversity

Initially it was expected that the diversity of the three Alaskan populations (Athapaskan, Yupik and Inupiat) were going to be the lowest given their small size. However the Alaskan populations demonstrated the highest levels of haplogroup diversities both in individual populations (Table 4.6) and in aggregate (Table 4.7). There are a few possible explanations:

I) The effect of male migration through the Americas resulted in increased haplogroup diversity in these populations.

II) As a consequence of being small populations (130,998 American Indian and Native Alaskans based on the 2010 census[32]), once diversity is introduced it represents a larger proportion of the population than one that is large.

III) Sociocultural norms are lax regarding ethnic identification with respect to paternal ancestry. Of the 130,998 American Indians and Native Alaskans, 20% identify as two races (White; American Indian and Native Alaskan)[32].

Haplotype sharing was seen with the Alaskan populations in the Purps et al (2014) with Estonian, Finnish, Hungarian, Dutch and Maasai populations[9]; this suggests male migration is responsible, in part, for the increased haplogroup diversity in the Alaskan populations. The African populations had the lowest average haplogroup diversities of the four geographic areas tested. This indicates that the male migration into these three African nations did not result in the gene flow required to increase haplogroup diversity as seen Europe, Asia and Alaska populations.

Translational Application

Genetic variation of the Y chromosome has been studied with respect to pathologies including infertility, coronary artery disease and prostate cancer. Of particular interest is the effect of Haplogroup I on coronary artery disease in a study of British men; A 50% higher risk of coronary artery disease was found in men who inherited haplogroup I compared to other Y-haplogroups[33,34]. In the study it was found that haplogroup I is associated with down regulation of the UTY and PRKY genes in marcophages of British men and that there was no evidence of admixture in any of the cohorts[33,34]. Recent studies also suggest the mammalian Y chromosome influences autosomal gene expression[33]. The consequence of the association of a haplogroup with a pathology begs the question: How do haplogroups impact pathologies on individuals who are admixed? The regulation of autosomal gene expression by the Y chromosome needs to take evolutionary history and autosomal admixture into consideration. Since Y-haplogroups co-evolved with the autosome it is reasonable to consider Y-Auto dysregulation as a result of admixture in the genome. Research involving the regulation of the autosome by the Y chromosome should take into account two factors: The haplogroups of the samples and the ancestry of the autosomal region where regulation is suspected.

Forensic Application

In DNA forensics, DNA mixture deconvolution is a frequent issue for evidentiary samples. Haplogroup prediction may be of aid in the deconvolution of mixed Y-haplotype profiles, when multiple males are the contributors of DNA in the evidentiary sample. Ge et al. (2010) discussed the logic of inferring all possible haplotypes that can explain a mixture profile[35]. Due to the linkage of loci in the NRY region and association of haplotypes with haplogroups, not all of these possible haplotypes are biologically probable, as there are strong

associations of specific allele-haplogroups combinations. If all possible haplotypes from a mixture profile are run in the haplogroup prediction software it is expected that some would receive haplogroup assignment and others not. This, together with consideration of higher fitness scores of haplotypes with assigned haplogroups, would reduce the number of biologically probable haplotypes in a Y-STR DNA mixture, improving the statistical strength of interpretation of DNA mixtures involving Y-STR data.

CONCLUSION

The assertion that an increase of loci in a haplotype would not enhance accuracy of predictions[20] appears to be incorrect based on this analysis. The prediction capability of Haplogroup Predictor does increase as the operationally defined haplotypes encompass more loci. The inclusion of RM Y-STRs as a portion of the haplotype, two of the twenty loci in the MXHT haplotype, resulted in no substantial change in proportion of haplogroup prediction nor in any substantial change with STRUCTURE analysis. This suggests that haplotypes partially comprised of RM Y-STRs have minimal impact on haplogroup prediction, although for Network analysis the impact is more pronounced, particularly at the extremities of the haplotype networks, suggesting that the clustering of the recently evolved haplotypes are affected by the RM STRs used in the present analyses. The three small and more isolated Alaskan populations had larger haplogroup diversities than larger populations, likely as a consequence of patrilineal diversity of male migrations in these populations. Admixed populations tend to arise from ancestral populations where at least one more homogenous than the admixed population. This observation was seen in all three separate population groups.

REFERENCES

1. Cherson, AD (2012) Atlas of Genetic Genealogy. Greencore Environmental Information Services. http://atlas.xyvy.info.
2. Jobling, M. A & Tyler-Smith, C. The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* **4,** 598–612 (2003).
3. McDonald Group *Y Haplogroups of the World* http://www.scs.illinois.edu/~mcdonald/ last accessed 5-31-2015
4. Athey, T. W. Haplogroup Prediction from Y-STR Values Using an Allele- Frequency Approach. 1–7 (2005).
5. Athey, T. W. Haplogroup Prediction from Y-STR Values Using a Bayesian-Allele-Frequency Approach. 34–39 (2006).
6. Larmuseau, M. H. D., Vanderheyden, N., Van Geystelen, A. & Decorte, R. A substantially lower frequency of uninformative matches between 23 versus 17 Y-STR haplotypes in north Western Europe. Forensic Sci. Int. Genet. 11, 214–9 (2014).
7. Núñez, C., Geppert, M., Baeta, M., Roewer, L. & Martínez-Jarreta, B. Y chromosome haplogroup diversity in a Mestizo population of Nicaragua. Forensic Sci. Int. Genet. 6, 192–195 (2012).
8. Hallast, P. *et al.* The Y-Chromosome Tree Bursts into Leaf: 13,000 High-Confidence SNPs Covering the Majority of Known Clades. *Mol. Biol. Evol.* **32,** 661–673 (2014).
9. Purps, J. *et al.* A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Sci. Int. Genet.* **12C,** 12–23 (2014).
10. U.S. Y-STR Database release 4.1 https://www.usystrdatabase.org/ received 1/22/2015
11. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945–959 (2000).
12. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Mol. Ecol. Notes 7, 574–578 (2007).
13. Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37-48
14. Fluxus Engineering http://www.fluxus-engineering.com/sharepub.htm#a1 last accessed 5/13/2015
15. Batch Processing Program http://www.hprg.com/hapest5/page5.html last accessed 4/12/2015
16. Butler, J. M., Kline, M. C., Decker, A. E. Addressing Y-Chromosome Short Tandem Repeat Allele Nomenclature. *Journal of Genetic Genealogy.* 4(2): 125-148 (2008).
17. Ballantyne, K. N. *et al.* A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Sci. Int. Genet.* **6,** 208–18 (2012).
18. Yfiler® Plus PCR Amplification Kit, #4485610 revision B. Carlsbad, CA: Life Technologies (2014)
19. PowerPlex®Y23 System, DC2305 and DC2320. Madison, WI: Promega Corportation (2015)
20. Muzzio, M. et al. Software for Y-haplogroup predictions: A word of caution. Int. J. Legal Med. 125, 143–147 (2011).

21. Athey, W. Comments on the article, Software for y Haplogroup Predictions, a Word of Caution. *Int. J. Legal Med.* **125,** 901–903 (2011).
22. Lao, O. et al. Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Hum. Mutat. 31, 1875–1893 (2010).
23. Ballantyne, K. N. *et al.* Toward male individualization with rapidly mutating y-chromosomal short tandem repeats. *Hum. Mutat.* **35,** 1021–32 (2014).
24. Marcheco-Teruel, B. et al. Cuba: Exploring the History of Admixture and the Genetic Basis of Pigmentation Using Autosomal and Uniparental Markers. PLoS Genet. 10, e1004488 (2014).
25. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343,** 747–51 (2014).
26. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
27. Montinaro, F. *et al.* Unravelling the hidden ancestry of American admixed populations. *Nat. Commun.* **6,** 6596 (2015).
28. Y-Haplogroup Predictor Instructions http://www.hprg.com/hapest5/page4.html last accessed 5-19-2015
29. Oh, Y. N. *et al.* Haplotype and mutation analysis for newly suggested Y-STRs in Korean father–son pairs. *Forensic Sci. Int. Genet.* **15,** 64–68 (2015).
30. International Society of Genetic Genealogy (2012). Y-DNA Haplogroup Tree 2012, Version: [7.65], Date: [5 December 2012], http://www.isogg.org/tree/ [17, May, 2015].
31. Ge, J. *et al.* Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic Sci. Int. Genet.* **3,** 179–84 (2009).
32. United States Census Bureau: Profile of General Population and Housing Characteristics: 2010 Demographic Profile Data http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk
33. Case, L. K. & Teuscher, C. Y genetic variation and phenotypic diversity in health and disease. *Biol. Sex Differ.* **6,** 1–9 (2015)
34. Bloomer, L. D. S. *et al.* Male-specific region of the y chromosome and cardiovascular risk phylogenetic analysis and gene expression studies. *Arterioscler. Thromb. Vasc. Biol.* **33,** 1722–1727 (2013).
35. Ge, J., Budowle, B. & Chakraborty, R. Interpreting Y chromosome STR haplotype mixture. Leg. Med. 12, 137–143 (2010).