Sequencing of Full Mitochondrial Genomes for NIST Population Samples

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

Email: Kevin.Kiesler@nist.gov

Kevin M. Kiesler[1], K.S. Andreaggi[2,3], J.D. Ring[2,3], C.R. Taylor[2,3], C.K. Marshall[2,3], and P.M. Vallone[1]

[1]U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA
[2]SNA International, Alexandria, VA, USA
[3]Armed Forces Medical Examiner System's Armed Forces DNA Identification Lab (AFMES-AFDIL), Dover, DE, USA

Poster # P146

The U.S. National Institute of Standards and Technology (NIST) has undertaken a population sequencing study of 607 mitochondrial genomes (mtGenomes). The goal of the study is to increase the availability of high-quality whole mtGenome haplotypes for haplotype frequency estimations, allowing the increased discrimination power of the full mtGenome to be harnessed. Whole mtGenome sequencing is enabled by Next Generation Sequencing (NGS) workflows which are vastly simplified compared to Sanger-type sequencing, which is typically limited to the mtDNA Control Region, or portions thereof, due to the labor required to produce mtDNA sequence data. Samples used in this study were grouped into categories of self-reported race: African American, U.S. Caucasian, and U.S. Hispanic individuals, representing three major U.S. populations. Assessment of resolution for mtGenome relative to control region data are presented. Previously characterized markers for this sample set include Y-chromosome Short Tandem Repeats (Y-STRs) and Ancestry Informative Single Nucleotide Polymorphisms (AISNPs). Comparisons of ancestry inference from mtGenome, AISNP, Y-STR data, and self-reporting are presented.

Methods – mtGenome Sequencing

- **Full mitochondrial genome (mtGenome) sequences obtained for 607 NIST population samples**
  - **Self-reported: African American (n = 236), U.S. Caucasians (n = 247), and U.S. Hispanics (n = 124)**
- **Long PCR amplification of full mitochondrial genome (mtGenome)**
  - **Primers & PCR conditions from Fendt *et al.* [1]**
- **Sequenced on next generation instrument**
  - **KAPA Hyperplus library procedure from Ring *et al.* [2]**
  - **MiSeq FGx instrument using MiSeq Reagent Kit v3-600**
- **mtGenome data curated using CLC Bio Genomics Workbench – AQME tool [3]**
  - **Data review and curation performed by AFMES-AFDIL staff**
  - **Minimum heteroplasmy variant frequency 5 %**

Figure 1: Mitochondrial genome showing control region with amplification strategy.

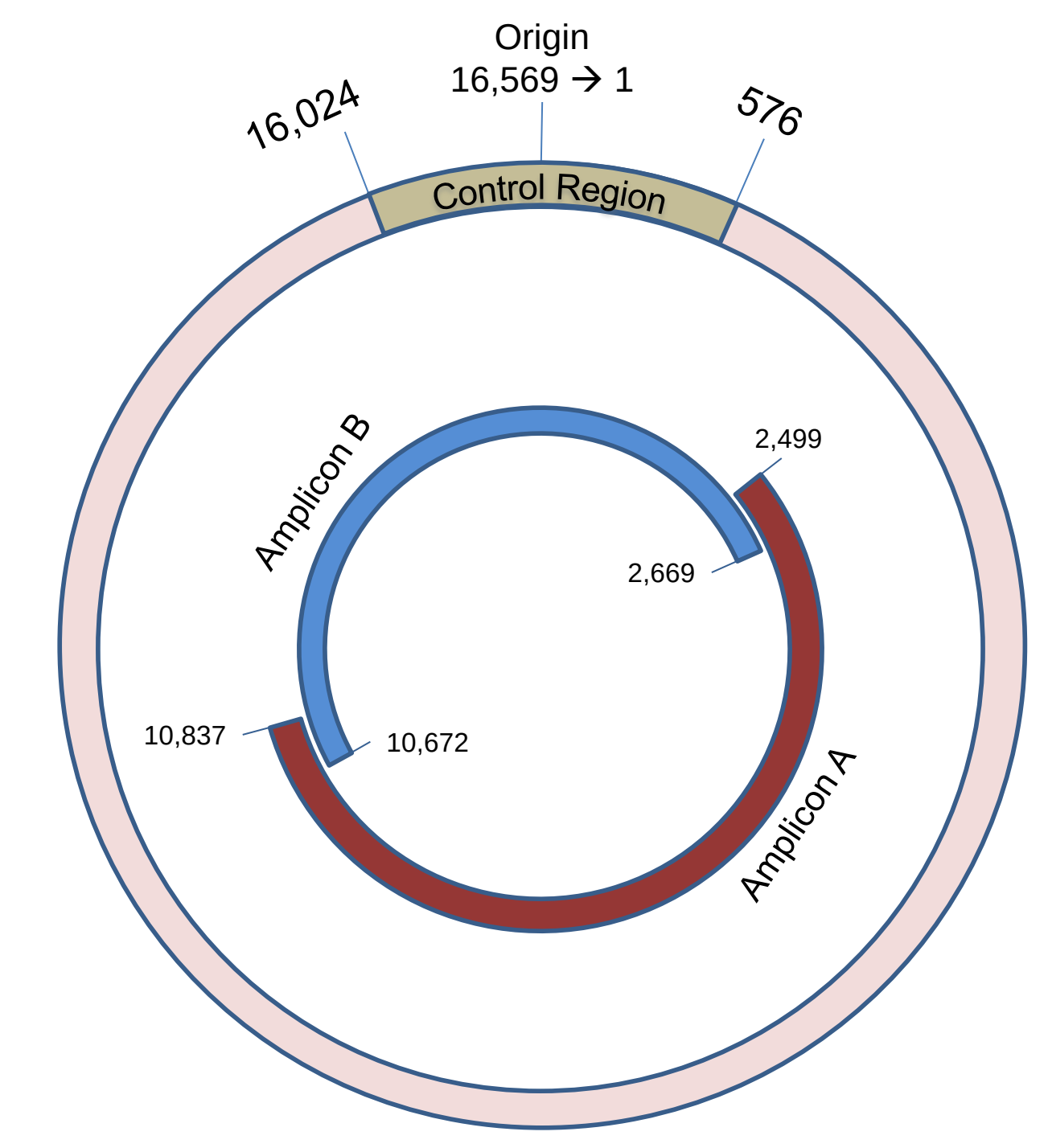


Figure 2: Library construction process.

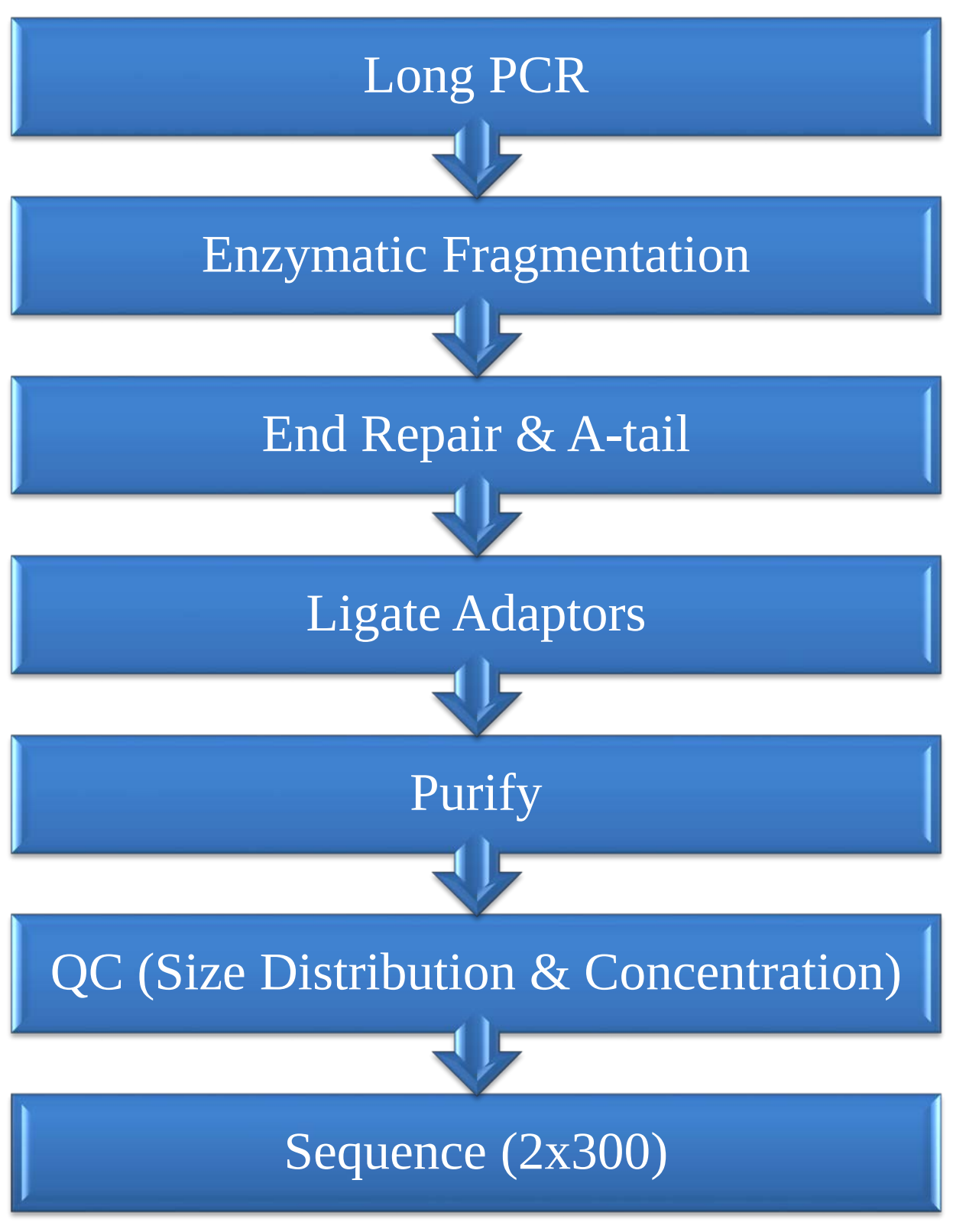
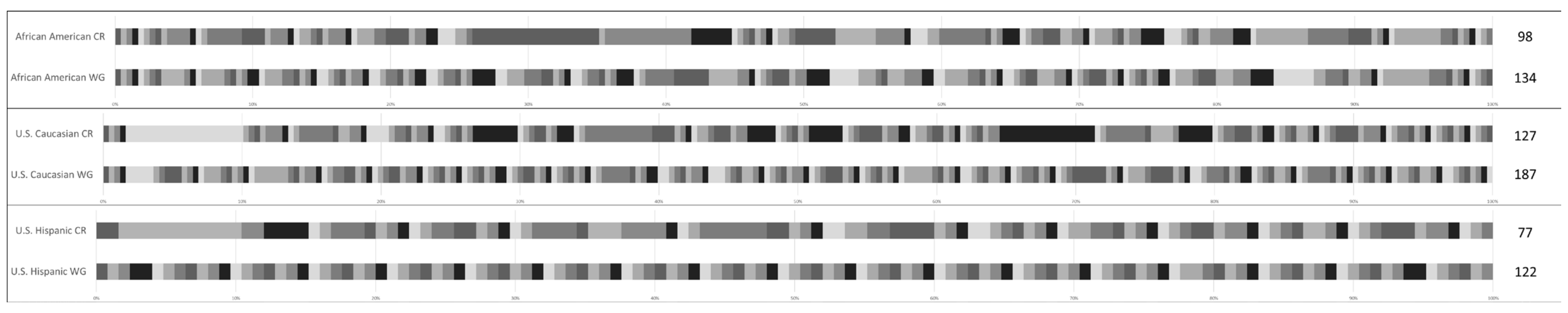

Long PCR
Enzymatic Fragmentation
End Repair & A-tail
Ligate Adaptors
Purify
QC (Size Distribution & Concentration)
Sequence (2x300)

Table 1: Comparison of shared profiles from Control Region (nt 16,024 to 16,569 & 1 to 576) vs whole mtGenome.

	All Samples		African American		U.S. Caucasian		U.S. Hispanic	
Number of Profiles Matching	Control Region	Whole Genome	Control Region	Whole Genome	Control Region	Whole Genome	Control Region	Whole Genome
11	1	0			1	0		
6	1	0					1	0
5	1	1			1	0	0	1
4	5	0	1	0	4	0		
3	7	3	4	1	3	2		
2	28	11	13	4	8	2	7	5
Unique	488 (88 %)	571 (94 %)	194 (82 %)	225 (95 %)	190 (77 %)	237 (96 %)	104 (84 %)	109 (88 %)
Total #	607	607	236	236	247	247	124	124

Figure 2: Comparison of haplogroups derived from Control Region vs whole mtGenome for three U.S. populations with number of haplogroups shown at end of each bar graph.



Discussion - Whole mtGenome vs Control Region

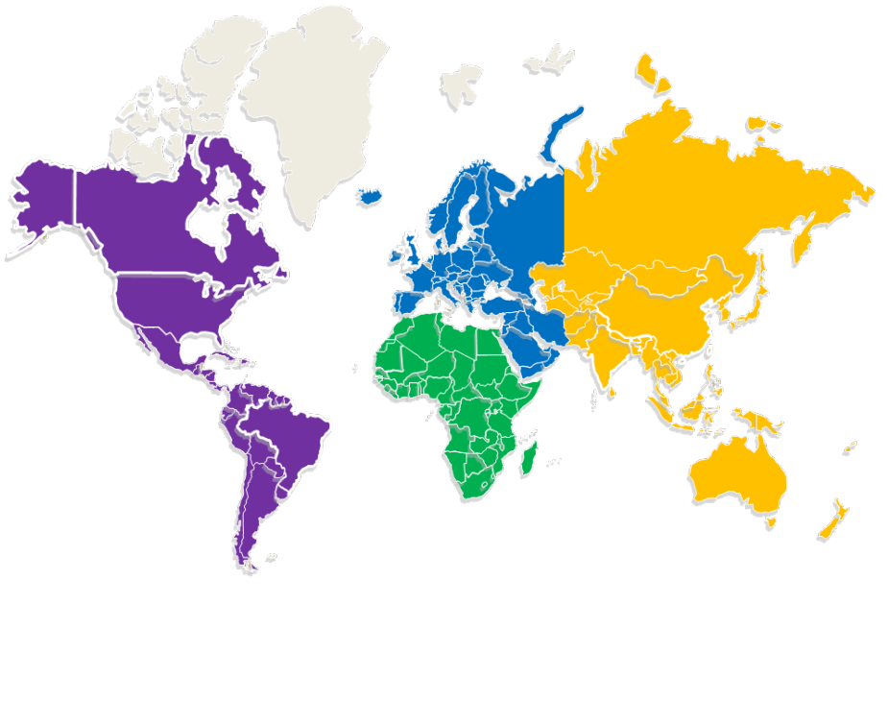
- Indels at positions 309, 315, 573, and 16,193 were not considered differences in profiles.
- Point heteroplasmy was considered a difference between two profiles.
  - Heteroplasmy was observed 209 times in these 607 mtGenomes.
- Overall 83 additional profiles could be uniquely identified in this population set when using whole mtGenome (WG) analysis vs control region (CR), an increase of 17 % within the study.
- The Caucasian population benefitted the most from whole mtGenome analysis with an increase in unique profiles of 25 %, followed by African Americans with a 16 % increase, and Hispanics with a 5 % increase.
- What happened to the most common shared CR profiles when WG is used?
  - African American group of 4 → group of 3 + 1 unique profile
  - U.S. Caucasian group of 11 → 11 unique profiles
  - U.S. Hispanic group of 6 → group of 5 + 1 unique profile
- Whole mtGenome sequencing enables deeper subclade classification vs control region (see Figure 2).

Ancestry inference using mtGenome, autosomal SNPs, and Y STRs

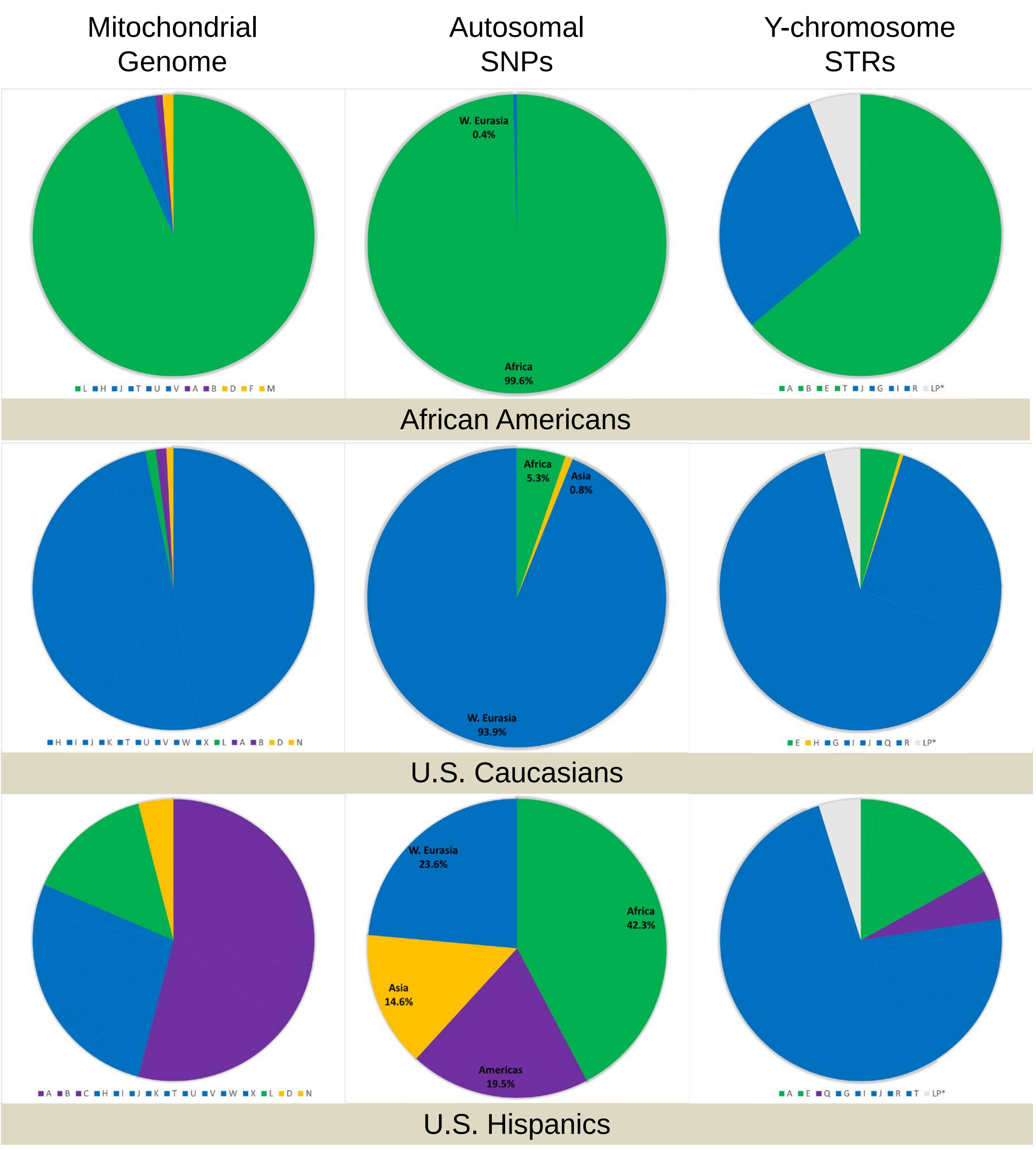
- **Full mitochondrial haplogroups assigned by CLC Bio Genomics Workbench – AQME Tool using EMPOP database values**
  - **Self-reported: African American (n = 236), U.S. Caucasians (n = 247), and U.S. Hispanics (n = 124)**
- **Autosomal ancestry-informative SNPs from ForenSeq**
  - **Libraries prepared with Verogen ForenSeq Signature DNA Prep Kit with DNA Primer Mix B (DPMB)**
  - **Sequenced on MiSeq FGx using MiSeq FGx Reagent Kit**
  - **Ancestry informative SNPs from Kidd et. al. [4]**
  - **Ancestry predictions from SNP data made using FROG-kb [5]**
- **Y-STR profiles from ForenSeq and Promega PPY-23**
  - **From NIST 1036 population sequencing [6] and ForenSeq DNA Signature Kit (unpublished)**
  - **Y Hgs were estimated using the NevGen Haplogroup Predictor (desktop version)**
    - https://www.nevgen.org/
  - **Y haplogroups categorized by continent according to Kivisild [8] and Lao [9]**

Figure 3: Ancestry prediction for three marker types for the three populations studied. Predictions were grouped into broad continental origins: African (green), European (blue), Asian (Orange), and Americas (purple); for mitochondrial haplogroups according to Rishishwar [7] and for Y-STR haplogroups according to Kivisild [8] and Lao [9]. For SNP data the top metapopulation predicted with The Standalone FROG-kb Ancestry Inference Batch Likelihood Computation Tool found on GitHub: https://github.com/haseenaR/FROGAncestryCalc was used. Location of the metapopulation was converted into continental ancestry groups. Mitochondrial and Y-chromosome subclades were reduced to highest “single letter” level for simplicity of display in this figure.

Color code Key:
Ancestry Prediction by Continent



		Count	Percent
African Americans	L	220	93.2%
	H	2	0.8%
	J	2	0.8%
	T	1	0.4%
	U	5	2.1%
	V	1	0.4%
	A	1	0.4%
	B	1	0.4%
	D	1	0.4%
	F	1	0.4%
	M	1	0.4%
	Total	236	
U.S. Caucasians	H	116	47.0%
	I	2	0.8%
	J	29	11.7%
	K	18	7.3%
	T	33	13.4%
	U	28	11.3%
	V	7	2.8%
	W	4	1.6%
	X	2	0.8%
	L	3	1.2%
	A	2	0.8%
	B	1	0.4%
	D	1	0.4%
	N	1	0.4%
	Total	247	
U.S. Hispanics	A	35	28.2%
	B	9	7.3%
	C	23	18.5%
	H	8	6.5%
	I	2	1.6%
	J	2	1.6%
	K	2	1.6%
	T	5	4.0%
	U	9	7.3%
	V	2	1.6%
	W	1	0.8%
	X	3	2.4%
	L	18	14.5%
	D	4	3.2%
	N	1	0.8%
	Total	124	



Y-STR Haplogroups			
		Count	Percent
African Americans	A	4	1.7%
	B	5	2.1%
	E	141	59.7%
	T	1	0.4%
	G	3	1.3%
	I	8	3.4%
	J	1	0.4%
	R	59	25.0%
	LP*	14	5.9%
	Total =	236	
U.S. Caucasians	E	11	4.5%
	H	1	0.4%
	G	8	3.3%
	I	40	16.4%
	J	15	6.1%
	Q	1	0.4%
	R	158	64.8%
	LP*	10	4.1%
	Total =	244	
U.S. Hispanics	A	2	1.6%
	E	19	15.3%
	Q	7	5.6%
	G	6	4.8%
	I	9	7.3%
	J	8	6.5%
	R	65	52.4%
	T	2	1.6%
	LP*	6	4.8%
	Total =	124	

*LP = low probability Y-haplogroup prediction from STR data. Minimum probability threshold of 0.10 used for NevGen haplogroup predictor.
**See poster P145 for a larger population scale Y-STR analysis.

Discussion – Ancestry Inference relative to self-identified ancestry

- For the U.S. Caucasian samples all three marker systems gave similar predictions.
  - Categorized as “West Eurasian”: mtGenome = 96.8 %, AISNP = 93.8 %, Y-STR = 91.0 %
- For the U.S. African American samples, mtDNA and AISNP inference performed similarly while Y-STR inference performed differently.
  - Categorized as “African”: mtGenome = 93.3 %, AISNP = 99.6 %, Y-STR = 64.0 %
- For U.S. Hispanic samples continental categorizations indicate weaker correlation to self identified ancestry.
  - Categorized as “European” or “Americas”: mtGenome = 81.5 %, AISNP = 43.1 %, Y-STR = 78.2 %

Conclusions:

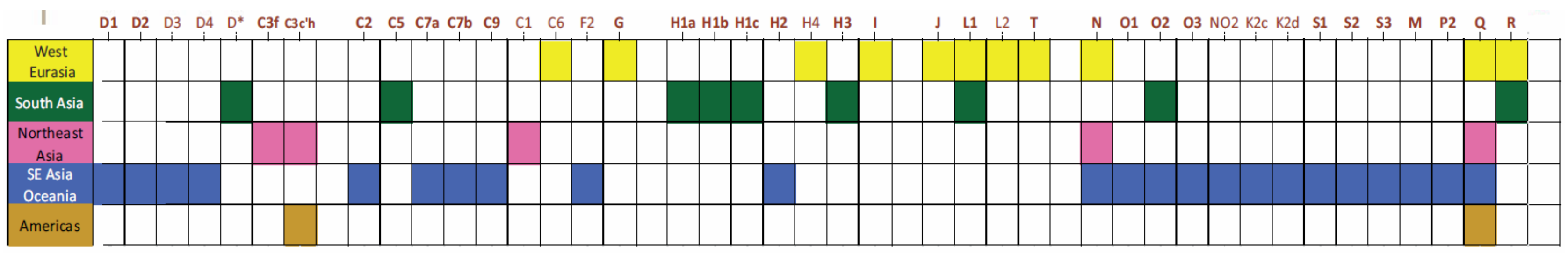
- **As expected, greater discrimination is observed from Whole mtGenome sequence compared to the Control Region sequence.**
  - **Gains in unique profiles: African American = 16 %, U.S. Caucasian = 25 %, and U.S. Hispanics = 5 %**
- **Whole mtGenome sequence may be used to differentiate some individuals when control region mtDNA profiles do not.**
- **Ancestry prediction with mitochondrial DNA works similarly to Y-STR or nuclear DNA SNPs.**
  - **Differences in mtDNA and Y-chromosome ancestry proportions may be due to bias in admixture history.**
  - **Y STR haplogroup inferences from the NevGen Haplogroup Predictor are informative, but additional Y-STR or Y-SNP typing may be needed to support improved paternal lineage predictions.**
  - **Ancestry inference in groups such as U.S. Hispanics are more variable, due to complex admixture and natural history of such populations.**
- **There is a limitation in the Bayesian Y haplogroup predictor, NevGen, in that it always predicts a result, even when a subclade may not be present in the reference database or support for a prediction is weak. This could result in false positives.**
- **Ancestry is affected by human migration, both ancient and modern, with no clearly defined boundaries for haplogroups. Many haplogroups are represented on multiple continents at varying frequencies. This could lead to misclassification at the “continent” level.**
- **Future: finish the mitochondrial for the remaining NIST population samples (n ≈ 662) and submit to EMPOP to increase the number whole mitochondrial genomes in the database.**

Figure 4: Haplogroups ascribed to continent for (A) mitochondrial DNA according to [7] and (B) Y-STR haplogroup predicted by NEVGEN Y-haplogroup prediction tool according to [8].

A.

		African						East Asian						Euro-Indian									Asian		American		
Cont	Pop	L0	L1	L2	L3	L4	L5	D	F	G	N	Y	Z	H	I	J	K	T	V	W	X	U	M	R	A	B	C
African	ESN	7	20	27	43	2																					
	GWD		15	48	42	2																6					
	LWK	18	8	12	47	5	10																1				
	MSL	2	17	39	24	1																2					
	YRI	5	17	38	47	1																					
East Asian	CDX							12	25														24	14		19	5
	CHB							23	16	5	8		1				1						19	6	7	12	5
	CHS							23	16	2	10		3										19	11	6	16	2
	JPT							39	6	11	9	1	4										14		6	14	
	KHV							2	27		4	1											33	11	1	21	1
European	CEU													51	1	8	3	10	3	5		18					
	FIN													37	2	7	6	3	4	2	2	36					
	GBR													39	3	10	5	10		2	5	18					
	IBS		1		1						2			58	1	4	7	8	6			18		1			
	TSI		1					1						54		8	9	13	2	4	1	15					
Indian	BEB							2	1					1		2				2		11	58	8	1		
	GIH			1	2									7		1	1	2		2	2	15	41	32			
	ITU										1			2	1	1	1	4		5		14	61	13			
	PJL								3	2				7		1		3		2	1	11	55	11			
	STU										3			12		1						14	50	22	1		
AdmixedAmerican	ACB	4	21	38	26	1								1		1						1			1	1	1
	ASW	7	16	14	24			1														1	1		1		1
	CLM	1	2	2	4			2						1				2					1		40	33	6
	MXL							9						5					1	1		2			25	15	9
	PEL			1	2			13						1											14	40	15
	PUR	2	5	4	10									2		6	1	1				4			38	7	25
	World	46	123	224	272	12	10	127	94	20	37	2	8	278	8	50	34	56	16	25	11	186	377	129	141	178	70

B.


D1 D2 D3 D4 D* C3f C3c'h C2 C5 C7a C7b C9 C1 C6 F2 G H1a H1b H1c H2 H4 H3 I J L1 L2 T N O1 O2 O3 NO2 K2c K2d S1 S2 S3 M P2 Q R
West Eurasia
South Asia
Northeast Asia
SE Asia Oceania
Americas

Download a PDF copy of this poster:

Funding: NIST Special Programs Office: Forensic DNA and the FBI Biometric Center of Excellence Unit: DNA as a Biometric

A copy of this presentation is available at: http://strbase.nist.gov/NISTpub.htm

References:
1. Fendt L.F., Zimmermann B., Daniaux M., and Parson W.: Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. BMC Genomics. 2009;10:139.
2. Ring J.D., Sturk-Andreaggi K., Peck M.A., and Marshall C.: A performance evaluation of Nextera XT and KAPA Hyper Plus for rapid Illumina library preparation of long-range mitogenome amplicons. Forensic Sci. Int. Genet. 2017;29:174-180.
3. Sturk-Andreaggi K., Peck M.A., Boysen C., Dekker P., McMahon T.P., and Marshall C.K.: AQME: A forensic mitochondrial DNA analysis tool for next-generation sequencing data. Forensic Sci. Int. Genet. 2017;31:189-97.
4. Kidd K.K., Speed W.C., Pakstis A.J., Furtado M.R., Fang R., Madbouly A. , Maiers M., Middha M., Friedlaender F.R., Kidd J.R.: Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci. Int. Genet. 2014;10:23-32.
5. Rajeevan H., Soundararajan U., Pakstis A.J., and Kidd K.K.: Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb. Investigative Genetics. 2012;3:18.
6. Coble, M.D., Hill, C.R., Butler J.M.: Haplotype data for 23 Y-chromosome markers in four U.S. population groups. Forensic Sci. Int. Genet. 2013;7:e66-e68.
7. Rishishwar L, Jordan IK.: Implications of human evolution and admixture for mitochondrial replacement therapy. BMC Genomics. 2017;18:140.
8. Kivisild T.: The study of human Y chromosome variation through ancient DNA. Human Genetics. 2017;136:529-546.
9. Lao O., Vallone P.M., Coble M.D., et al.: Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Human Mutation. 2010;31:E1875-E1893.

Disclaimer: Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce. Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose. This work was approved by the NIST Human Subjects Protection Office.