POPULATION GENOMICS OF THE MITOCHONDRIAL GENOME SEGMENTS AND THE PREDICTION OF NEUTRAL AND SELECTIVE TRENDS FOR IDENTIFICATION AND ASSOCIATION STUDIES

Iannacone GC[1,2]

1. Instituto de Medicina Legal y Ciencias Forenses
2. Laboratorio de Microbiología Molecular y Biotecnología de la Facultad de Ciencias Biológicas de la Universidad Nacional Mayor de San Marcos.

I. INTRODUCTION

The DNA identification of missing persons in open or closed contexts, as in the case of the 15,000 disappeared between 1980 and 2000 in Peru, presents the risk of generating DNA identifications mistake by two factors:

1. Degraded DNA in the Samples of missing persons (eg bones) with low DNA quantity and/or inhibitors. It produce partial genetic profile information for the markers used.
2. Substructure in the population of relatives of the missing person. It generate an increase in genetic resemblance among individuals. Even more, in populations where the effective population number (Ne) is reduced as in geographically isolated and / or small populations.

In these circumstances, since there is not enough genetic information in nuclear DNA (nDNA) for STR, SNPs and / or INS-DEL markers, it is recommendable to use the genetic information of the hypervariable regions (HV1, HV2 and HV3) in the mitochondrial genome (mtDNA), **(Butler 2015)**. However, it is difficult to reach an identification due the existence of mtDNA haplotypes of HV1, HV2 and HV3 shared among unrelated individuals. Therefore, it is important to increase polymorphism in order to differentiate haplotypes from unrelated individuals who share a founding ancestor (Coancestry).

In this context, we can find two types of polymorphic segments in the coding and hypervariable region of the mtDNA. 1) Segments probably neutral character due it allow greater accumulation of population polymorphism. 2) Segments probably selective where the accumulation is restricted for a smaller amount of population polymorphisms.

In the case of the probably neutral segments, it increase the success of identification by add new polymorphism and / or complementing the lack of information in the hypervariable regions of the mtDNA in case of LCN, **(Just et al 2014)**. Likewise, in the context of low concentrations of mtDNA, sequencing by Next Generation Sequencing (NGS) for whole genome of mtDNA will produce a differential amount of contings that it will impact the reliability of mtDNA sequences by reducing the coverage of reliable sequencing of the hypervariable regions. **(Chaitanya et al 2015)**

On the other hand, in the case of the segments with greater selective probability (population restricted polymorphism), it will allow genetic association in groups of individuals with a certain phenotype among populations and when the association is found, it will be added to support the DNA identification tools.

It is for this reason, the objective of the study to determine at a population genomic level selective and neutral segments in the control region of the mtDNA in order to apply them as tools in the forensic process with DNA.

II. MATERIALS AND METHODS:

1. Study population and alignment of mtDNA sequences:

3295 mitochondrial genomes corresponding to 47 populations worldwide were analyzed. The data was obtained from the "1000 Genomes" database for 2534 mtDNA that includes a population of Peru (86 mtDNA) and from the "Human Genome Diversity Project-HGDP" database with 761 mtDNA.

The ALIVIEW V.1.26 program was used to sequence alignment with the Revised Cambridge Reference Sequence (rCRS) of the Human Mitochondrial DNA (16569bp) as a reference sequence.

2. Numerical matrix of mtDNA sequences:

The following conversion rules were established for each nucleotide as constants **G = 4, C = 5, A = 2, T = 3, N = 10** (Heteroplasmia or doubt in the determination of the base) and **GAP = 15.5** (Selection or insertion). The conserved sequences are replaced by 0. Thus, each nucleotide position among individuals was represented by numerical values .

3. Determination of limits among polymorphic segments in the mtDNA:

The limits are segments where there are no change with respect to the reference sequence, probably of high selective effect. The limit is separated by two polymorphic segment. **(Figure 1)**

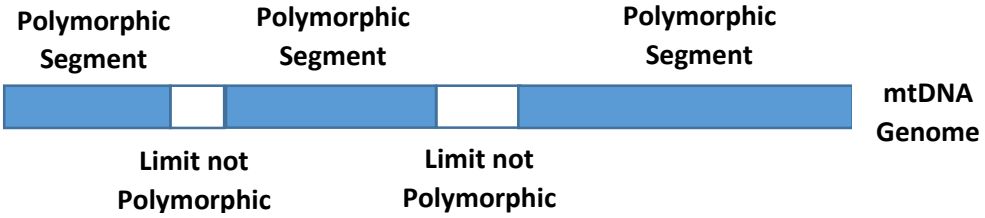


Figure 1: Scheme of polymorphic segments and non-polymorphic limits

For this purpose, the limits were calculated by average, median, mode, standard deviation and the normal distribution analysis. Thus, we determinate the minimum limit to be considered in the numerical matrix of the mtDNA sequences with the aim to obtain the total of polymorphic segements.

4. Population Genomics Algorithm for polymorphic segments determination:

The algorithm is based on the criterion of the equilibrium of the selective, neutral and almost neutral evolutionary theories for DNA sequences in a genome in relation to the probability of supporting, in a certain segment of the mtDNA, greater amount of polymorphisms (Probably Neutral) or an accumulation more restrictive for less polymorphisms (Probably Selective).

This algorithm consists of the following parts:

• Determination of mtDNA nucleotide polymorphisms among individuals in populations with respect to a reference sequence.

• Conversion of the nucleotide polymorphisms to a numerical matrix through indexs.

• Determination of the limits between polymorphic segments. The Conserved segments correspond to segments of zero variations between populations, which allow determining the number of polymorphic segments and their distribution in the mtDNA.

• Analysis of the genetic distances between populations, in order to determine the population groups (Macropopulations) which it allow population genomics comparisons of the polymorphic segments.

• Calculation of the probability of inter and intra population comparison of the mtDNA haplotypes without considering the hypervariable regions. This being an indirect way of measuring the effect of Ne on the density of variations of each polymorphic segment.

• Determination by indexs in the polymorphic segments, the probability of being Selective (S), Neutral (N) and their intermediates as Slightly Selective (SS), Slightly Neutral (SN).

• Inclusion of characteristic factors of polymorphic segments (eg, type of genes, polymorphic density, disease-associated mutations, etc.)

• Multivariate analysis for association of the values of the combined indixes of the polymorphic segments by macropopulation and factors in the mtDNA genome.

Below there is the process flow of the algorithm that we called GenoPopulation. **(Figure 1)**

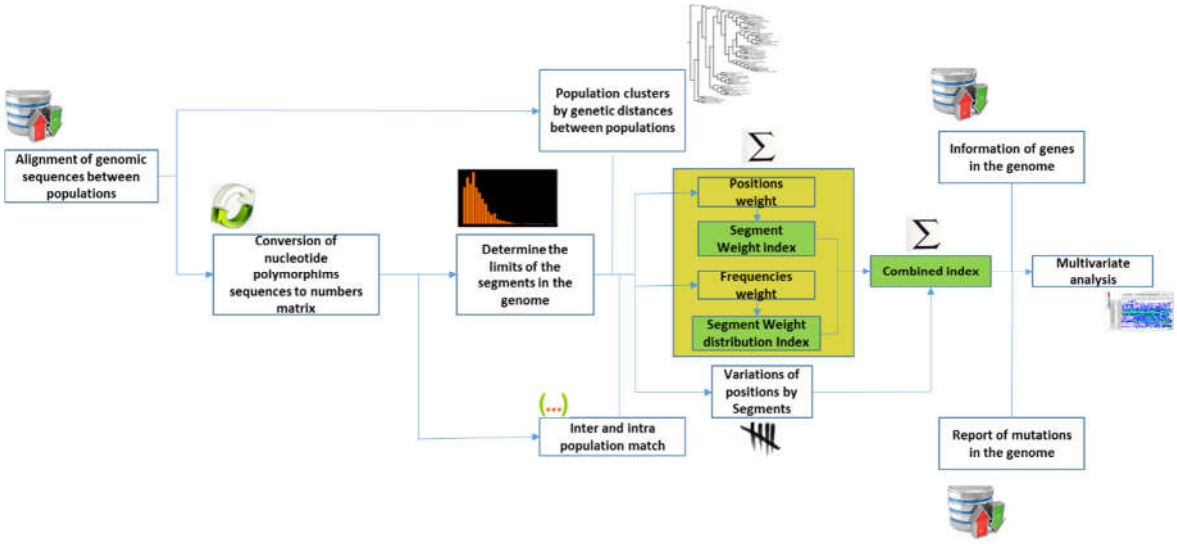


Figure 2: Algorithm Population Genomic Analysis Flow – GenoPopulation.

5. Numerical differentiation of polymorphic segments in mtDNA:

In order to locate the polymorphic segments in the mtDNA among populations, three indexs were developed with the aim to classified these segments according to the polymorphic content:

Segment Weight Index (SWI):

It is defined by quantifying each nucleotide position in the genome by $\left[[\sum_{x=1}^{n}(v_n)]x\ n\right]$, Where "Vn" corresponds to the value of each variants with respect to a reference sequence in "n" number of those variants in a nucleotide position. In the case that is equal to the reference sequence they take the value equal to 0. These values are then unified for each segment by:

$SWI = \left[\frac{\sum_{x=1}^{z}(Ln\ [(\sum_{x=1}^{n}(v_n))x\ n])_z}{z}\right]^{z^{-1}}$, where "Z" is the number of nucleotide positions with variation in a segment. The conserved nucleotide positions are considered as 0, because they have no variation with respect to the reference sequence. Therefore, (SWI) is equal to the sum of the natural logarithms (Ln) of each nucleotide position in a given segment, which it is divided and inverse exponent by the number of nucleotide positions in a segment. Allowing thus, differentiating segments where there are restrictions of variation with respect to those that have more variations by positions.

Segment Weight distribution Index (SWDI):

It is determined by each nucleotide position of the genome by $\prod_{y=1}^{w}\left[\sum_{x=1}^{p}\left(\frac{v_p}{N}\right)\right]$, where "Vp" is the value of the variants in "p" number different variants in a nucleotide position of "N" number sequences. In addition, "w" is the frequency value of other variants in the same nucleotide position. If there are two different variants, the value of both is multiplied. Thus. (SWDI) is determined by:

$SWDI = \sum_{x=1}^{z}\left[\prod_{y=1}^{w}\left[\sum_{x=1}^{p}\left(\frac{v_p}{N}\right)\right]\right]$, where "z" is the number of polymorphic nucleotide positions and therefore are different from 0, since a nucleotide position without variation is equal to 0.

Combined Segment Index (CSI):

In order to be able to join the two previous indices and classify the polymorphic segments. We developed the follow index:

CSI = Ln (SWI X SWDI X # VARIANT BY SEGMENT),

The values range are from negative to positive depending on the degree of greater or lesser conservation. Each polymorphic segments being classified as: Probably selective (S), Probably Slightly selective (SS), Probably Slightly neutral (SN) and Probably Neutral (N).

6. Genetic distances trees of the mtDNA:

The MEGA-X v.10.0.5 program was used for the distance matrix and the MEGA 4.0.2 program was used for the trees analysis by the methodology of Neighbor-Joining, Minimum Evolution and UPGMA.

7. Inter and intra population comparisons of mitochondrial genomes

The comparisons were made at the intrapopulation and interpopulation levels within and between groups formed by the Neighbor-Joining trees using the formulas of haplotypic diversity and population match described by Brinkmann et al 1999. Likewise, we developed for the haplotype comparisons a numerical weight equal which it is unique for each haplotype. The criteria follow the hash-type encryption

applied to biological information as described by Iannacone 2015. Likewise, for the comparison Hypervariable regions were excluded in the analysis.

Then based on the mtDNA transformed sequence matrix, each haplotype was defined as $\prod_{x=1}^{z} \boldsymbol{LN}(\boldsymbol{V_x} * \boldsymbol{Z_x})$ **,** where "V" is the value in a given position "x" of the profile for an individual and "Z" is the nucleotide number in that position with respect to the mtDNA. The values are ranked with the same value corresponding to the match of the same haplotype.

8. <u>Multivariate analysis by neighborhood limits and ranges of type S, SS, SN and N</u>

We used the PRIMER V.7.0 program to analyse the macropopulations with respect to the CSI value of each polymorphic segment in the mtDNA. In this analysis we consider factors such as:

1. Type of polymorphic segment probably Selective, Neutral and its intermediates.
2. Type of sequence where the polymorphic segment is included.

In the polymorphic segments analysis we used the criteria by limits and by neighborhood ranges for the probability of type S, SS, SN and N. In the case of neighborhood ranges each segment includes more than one limit, that is to say in a same neighborhood range, includes polymorphic segments contiguous of the same type probability that it follow the forward direction on the mtDNA genome (from 0 to 16569). The value of a range corresponds to the average of the normalized CSI values (CSIn) for the polymorphic segments included. This normalization of the CSI values was carried out because the polymorphic segments comprise negative values to positive ones, for that reason all the values were converted to positive with the relation $\boldsymbol{e^x}$, where “e” is the natural number and “x” is equal to CSI value in a given polymorphic segment.

RESULTS

a. **<u>Genetic relationships between populations studied:</u>**

With the 3295 mtDNA genome sequences aligned, the genetic distance matrix was obtained with the Maximun Composite Likehood Model method. We observed a similar distribution in the 47 populations for the eight macropopulations found in three considered method (UPGM, Neighbor Joining and Minimum Evolution). **(Figure 3A).** The eight macropupolations correspond to Europeans (EUE), Indo-European (CAU), Continental Asia (ASS3), Maritime Asia (ASF2), Indico (IND1), Latin American (LAT1), East Africa (AFE) and South Africa (AFS), **(Figure 3B).** The Peruvian population (PEL) is located in LAT1. Also is interesting the case of the population of Brazil (BRZ), which is an Amazonian population (Surui) that is grouped with ASF2 by the three methods used.

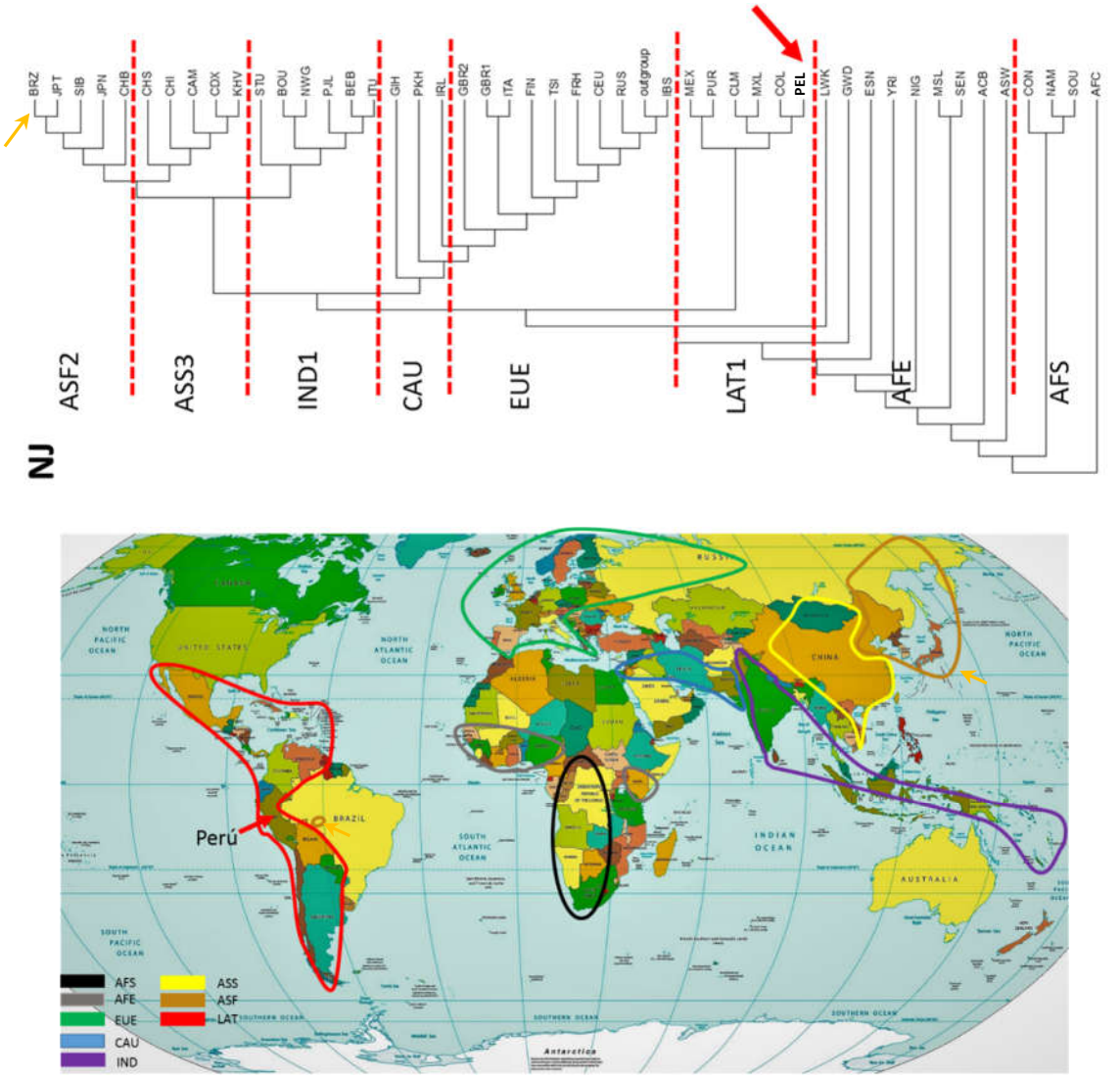


Figure 3. Tree (A) and map (B) of mitochondrial genome relationships among the 47 populations. The names of the macropopulations found in the NJ tree correspond to the colors assigned on the map. Peru (PEL) on the tree..**POPULATIONS:** ACB (Barbados), ASW (USA-African), ESN (Nigeria-Esan), GWD (Gambia), LWK (Kenya-Luhya), MSL (Sierra Leone - Mende), NIG (Nigeria), SEN (Senegal), YRI (Nigeria-Yoruba), AFC (Central African Republic), CON (Congo), NAM (Namibia), SOU (South Africa), CHB (Beijing), JPT (Tokyo), JPN (Japan) SIB (Siberia), CAM (Cambodia), CDX (China-Dai), CHI (China), CHS (China-Han), KHV (Vietnam), BRZ (Brazil-Surui), GIH (India-Guajarati), IRL (Israel), PKH (Pakistan) , CEU (USA-Utha), FIN (Finland), FRH (France), GBR (United Kingdom), GBR2 (United Kingdom-Orkney Islands), IBS (Spain), ITA (Italy), RUS (Russia), TSI (Italy-Tuscany), BEB (Bangladesh-Bengali), BOU (Bougainville), ITU (India-Telugu), NWG (New Guinea), PJL (Pakistan-Punjabi), STU (Sri Lanka), CLM (Colombia-Medellin), COL (Colombia), MEX (Mexico), MXL (USA-Mexicans) PEL (Perú), PUR (Puerto Rico).

b. Limits of polymorphic segments:

Through a normal distribution analysis of nucleotide positions without polymorphic variation (limits) across the populations studied. A total of 33 types of limits were observed and it correspond to a sizes ≥6 bp without variation. The highest frequency of limits were between 6bp – 14bp, which corresponds to approximately 30% of the total limits found.

The limits based on the 47 populations analyzed together generated 576 polymorphic segments in the mtDNA. When analyzing these limits at the level of the eight macropopulations, each group had a certain type and position of limit in the mtDNA genome based on the accumulation of polymorphisms.

The distribution of these limits is influenced in time by the effective population number (genetic drift) and / or Selection of certain polymorphisms in each populations. Then, the pressure on a polymorphism will affects the entire mtDNA haplotype among macropopulations.

Therefore, according to human migration history, the groups of the African continent have the largest number of polymorphic segments and in the most recent populations it have the least number of these segments. Both situations have been observed between the polymorphic segments and the limits found in the populations analyzed.

c. <u>Haplotype match in the populations studied:</u>

There are a great diversity in the control region of the mtDNA as in the case of LAT1 for the population of Peru (PEL) = 98.7% and Mexico (MXL) = 98.4% have the highest haplotypic diversity. Being similar situation in populations where the effective population numbers (Ne) are greater and allow greater variability in the mtDNA genome that it does not include the variability of the hypervariable regions. In the case of populations with lower haplotypic diversity as Mexico (MEX) = 85.3% and Puerto Rico (PUR) = 81.7% in LAT1 have lower Ne.

Also, a high cumulative diversity of segment polymorphisms was observed in the mtDNA coding region that contributes to the diversity. For this raison, we observed a total of forty and six haplotypes matches within and between macropopulations respectively. In the comparision between LAT1 only have one matches with EUE. It is probably due to admixture by Spanish colonization in the XVI century, **(Figure 4).**

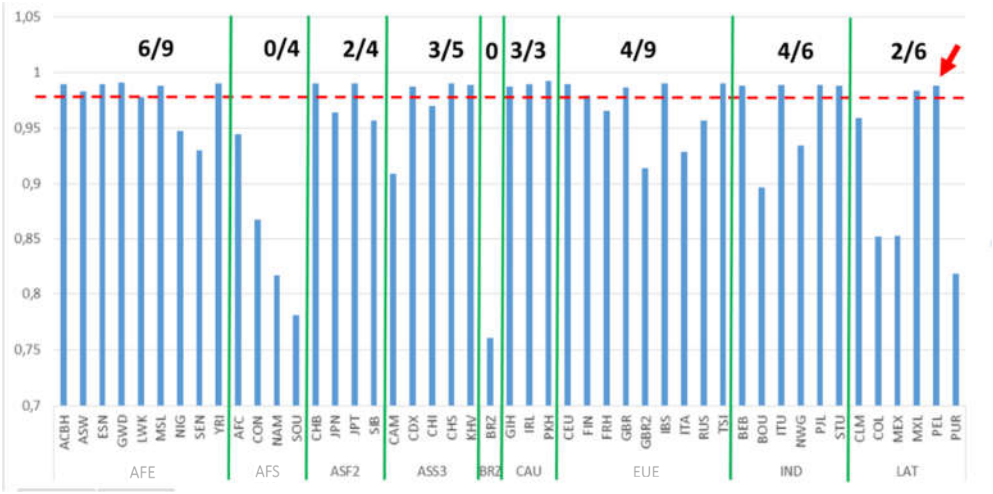


Figure 4. Distribution of haplotypic diversity by populations in each macropopulation found in the Neighbor Joining tree. The fractions correspond to the haplotypic diversities of mtDNA in the coding region that are above 98% vs. the total populations of that macropopulation. The red arrow indicates the population of Peru**. <u>POPULATIONS:</u>** ACB (Barbados), ASW (USA-African), ESN (Nigeria-Esan), GWD (Gambia), LWK (Kenya-Luhya), MSL (Sierra Leone - Mende), NIG (Nigeria), SEN (Senegal), YRI (Nigeria-Yoruba), AFC (Central African Republic), CON (Congo), NAM (Namibia), SOU (South Africa), CHB (Beijing), JPT (Tokyo), JPN (Japan) SIB (Siberia), CAM (Cambodia), CDX (China-Dai), CHI (China), CHS (China-Han), KHV (Vietnam), BRZ (Brazil-Surui), GIH (India-Guajarati), IRL (Israel), PKH (Pakistan) , CEU (USA-Utha), FIN (Finland), FRH (France), GBR (United Kingdom), GBR2 (United Kingdom-Orkney Islands), IBS (Spain), ITA (Italy), RUS (Russia), TSI (Italy-Tuscany), BEB (Bangladesh-Bengali), BOU (Bougainville), ITU (India-Telugu), NWG (New Guinea), PJL (Pakistan-Punjabi), STU (Sri Lanka), CLM (Colombia-Medellin), COL (Colombia), MEX (Mexico), MXL (USA-Mexicans) PEL (Perú), PUR (Puerto Rico).

d. <u>Distribution of the CSI values in the eight macropopulations:</u>

The distribution of CSI values shows variability corresponding to the nucleotide polymorphism in each position of the mtDNA that contribute to the polymorphism. Therefore, these polymorphic segments can be classified, according to the accumulation of variation by positions within polymorphic segments for the probability of a selective or neutral type.
In this context, the type of segment was classify by CSI ≤ -1.5 (Probably Selective - S), -1.5 <CSI <0 (Probably Slightly Selective - SS), 0 <CSI <1.5 (Probably Slightly Neutral - SN) and CSI ≥ 1.5 (Probably Neutral - N).

In addition, the length in base pairs (bp) of the polymorphic segments of type SN and N is longer in base pairs (bp) than those observed in the type SS and S (average of 104 bp vs 54 bp respectively). However, the greatest diversity in the number of polymorphic segments is observed in the type SS and S.

When considering all populations, we have the Selective type - S (49%), Slightly Selective - LS (10%), Slightly Neutral - LN (25%) and Neutral - N (15%). **(Figure 5)**

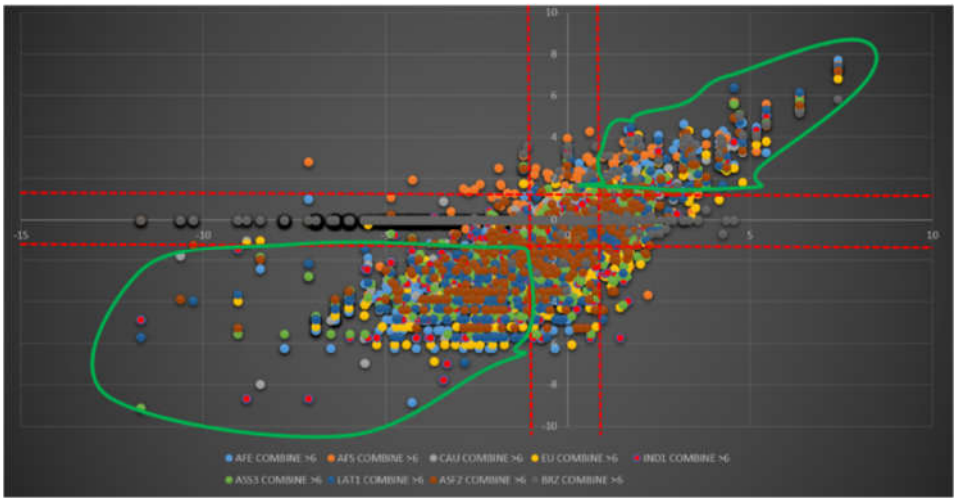


Figure 5. Distribution of the Combine Segment Index (CSI) among the 8 macro populations and the BRZ population. The dashed lines in red delimit the points that fall in areas of slightly selective type (SS) and slightly neutral (SN). The circles in green correspond to those of selective type (S) and neural (N).

By grouping the data by probability neighborhood ranges of polymorphic Segment type along the mtDNA, a greater accumulation of S + N type segments is observed in AFE, EUE and LAT1.

Likewise, considering all populations outside of Africa, LAT1 is the second with the highest number of segments of type S + N.

In the case of S + SS, the highest accumulation correspond to LAT1 and EUE. LAT1 being the one with the highest accumulation of this type of polymorphic segments.

In the case of SN + N segments, the greatest accumulation was found in African populations. In the other macropopulations the number of segments of type SN + N is maintained with the lowest standard deviation of +/- 4. **(Table 1)**

TIPOS DE SEGMENTOS POLIMORFICOS	AFE	AFS	CAU	EUE	IND1	ASS3	LAT1	ASF2	BRZ
SN+N	79	98	51	39	44	39	40	44	40
SS+S	118	37	140	152	148	150	157	138	9
S+N	111	47	68	102	89	79	92	69	21
SS+SN	86	88	123	89	103	110	105	113	28
S+SS+SN+N	197	135	191	191	192	189	197	182	49

Table 1. Distribution of the type of polymorphic segment in the eight macropopulations and the BRZ population. The numbers represent the number of polymorphic segments found by each type (S, SS, SN, and N), considering a distribution by neighborhood ranges. In light blue are the highest numbers of polymorphic segment type combination.

Each types of polymorphic segment by macropopulation, it can be shared by all but with different CSI value. In other cases, a polymorphic segment is only shared by some or it is unique for a given macro population.

Likewise, in the case of LAT1, the populations that contribute to a greater number of polymorphic segments is PEL (Peru), which shows the largest number of S + N type, in LAT1. In the case of SS + S, the highest number was presented by PEL and CLM (Colombia - Medellín). Likewise, considering the types S + SS + SN + N, the populations of PEL and CLM have a similar level of diversity in the number of polymorphic segments. In the case of SN + N have a low standard deviation of +/- 4 for the number of segments of type SN + N and it is similar to those observed in the Macropopulations. **(Table 2)**

TIPOS DE SEGMENTOS POLIMORFICOS	PEL	MXL	MEX	COL	CLM
SN+N	31	34	40	42	32
SS+S	108	90	39	8	112
S+N	63	42	22	15	39
SS+SN	76	82	57	35	105
S+SS+SN+N	139	124	79	50	144

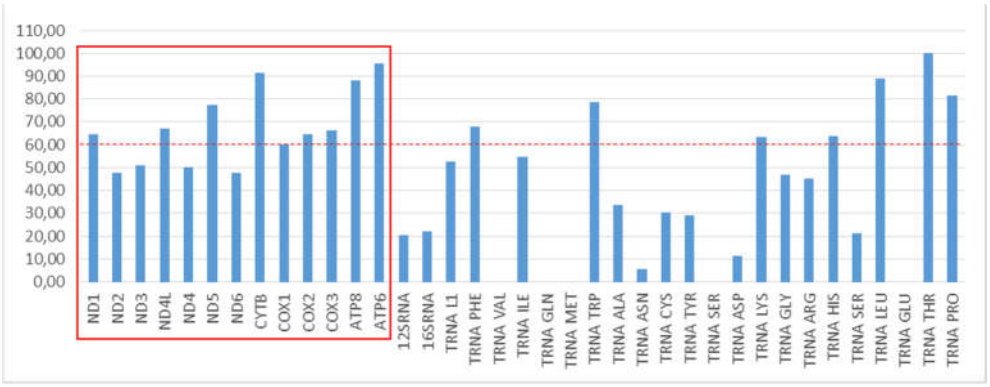
Tabla 2. Distribution of the type of polymorphic segment in the Latin American macropopulation. The numbers represent the number of polymorphic segments found by each type, considering a distribution by ranges by neighborhood. In light blue are the highest numbers of polymorphic segment type combination.

With the results, different population histories are evidenced by polymorphic segment and in each population are related to the action of neutrality and selection. For example, 144 polymorphic segments were found in CLM (Colombia - Medellín) and 79 polymorphic segments in MXL (Mexico) for the S + SS + SN + N. Although the CLM population has less haplotype diversity than the MXL population. Also in the case of the populations of PEL and CLM, where they have a similar level of SN + N type but in the analysis of haplotype diversity they are different and it being higher in PEL with respect to CLM.

e. <u>Distribution of polymorphic segments (S, SS, SN and N) according to the presence of mitochondrial genes:</u>

Greater accumulation of variation is observed in polymorphic segments associated with electron transporter complex genes (CPI-CPV) and some RNA transfer genes. This accumulation of type probably SN and N according to the ratio (SN + N / TOTAL x 100) were CPI in 58%, CPIV in 63%, CPIII in 91% and CPV in 91%. In the case tRNA, ten of them are close to polymorphic segments that on average reach 70% of content of type probably SN + N. However, it is very likely that intrinsically in tRNAs have an accumulation of around 25% of SN + N as specifically observed in tRNA contiguous with tRNA. In the case of polymorphic segments comprising rRNA, the accumulation of SN + N is close to 20% (Higher accumulation of SS + S). However, in the case of tRNA L1 (Coding for leucine 1) it presents an accumulation of polymorphic segments of 52% for SN + N type **(Figure 6a).** The coverage in base pairs of these polymorphic segments of type probably SN + N corresponds to 8439 bp of which type SN are 4400 bp and of type N are 4039 bp. Likewise, 80% of the 86 polymorphic segments of type SN and N segments have less than 150bp (Figure 6b). The SN and N content in the coding regions with respect to the hypervariable regions (I + II + III) is approximately 5 times more in number of base pairs, therefore they are less dense than the hypervariable regions.

A.



B.

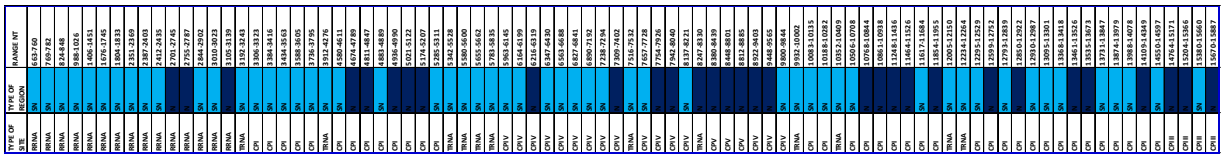


Figure 6. Distribution of polymorphic segments in the mtDNA and gene coverage (A), the red square indicates the greatest accumulation of these segments in the respiratory genes included. The dotted line indicates the accumulation greater than 60% of the polymorphic segments probably slightly neutral (SN) + Neutral (N). The difference (40%) in those same genes comprised the probably slightly selective (SS) + Selective (S) segments. The mtDNA distribution is shown in (B) corresponding to SN the light blue color and N in blue for the positions of covered genes.

f. <u>Multivariate analysis of CSIs in the mtDNA genome:</u>

With the normalization of CSI values (CSIn), the distribution of values for polymorphic segments was: 0.2≤ CSIn (S), 0.2 <CSIn <1 (SS), 1 <CSIn <4.5 (SN) and 4.5≥ CSIn (N). Thus, in the two-dimensional Heat Map analysis, we found a correlation of the distribution of the CSIn values of polymorphic segments with respect to the population trees obtained with Neighbor-Joining. However, the distribution of the probability of type S, SS, SN and N does not show significance (ANOSIM R = 0.08) when they are not grouped by neighborhood ranges of the same type in the genome. **(Figure 7)**

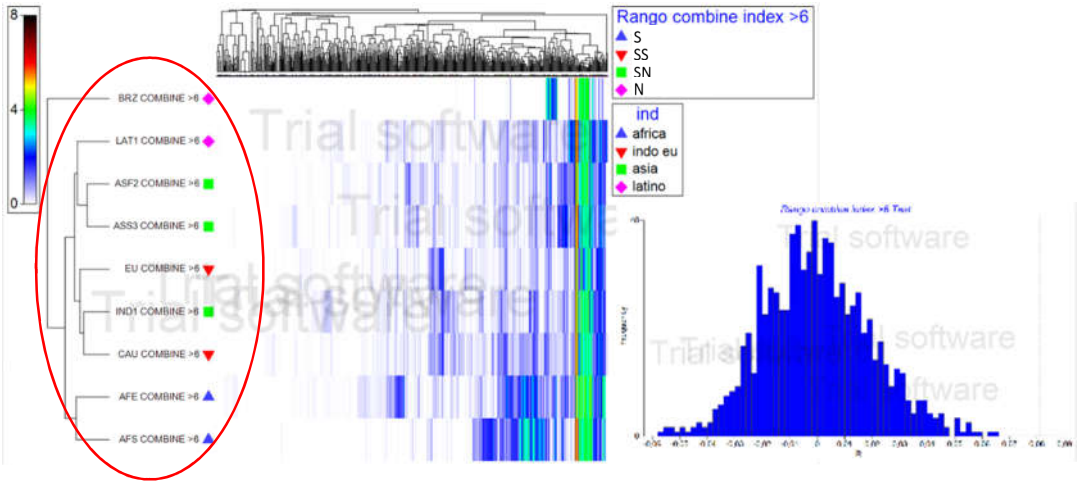


Figure 7. Two-dimensional multivariate analysis of the CSIn values of the polymorphic segments of the mtDNA by macropopulations (Tree on the Left) and the probability of type S, SS, SN and N (Top Tree) without grouping by type neighborhood range. In red circle, grouping of the distribution of macropopulations. The ANOSIM analysis histogram (Right) is displayed.

In the case of the distribution factor by type genes in the mtDNA for the CSIn values, it is observed when analyzing the averages for the areas that comprise ribosomal genes (Ribo), transfer RNA (Trans), Respiratory (Resp) and absence zones of genes (NG); the accumulation of SN and N types in respiratory gene zones. In addition, a higher value R=0.18 is obtained in ANOSIM for a multivariate structure. **(Figure 8)**

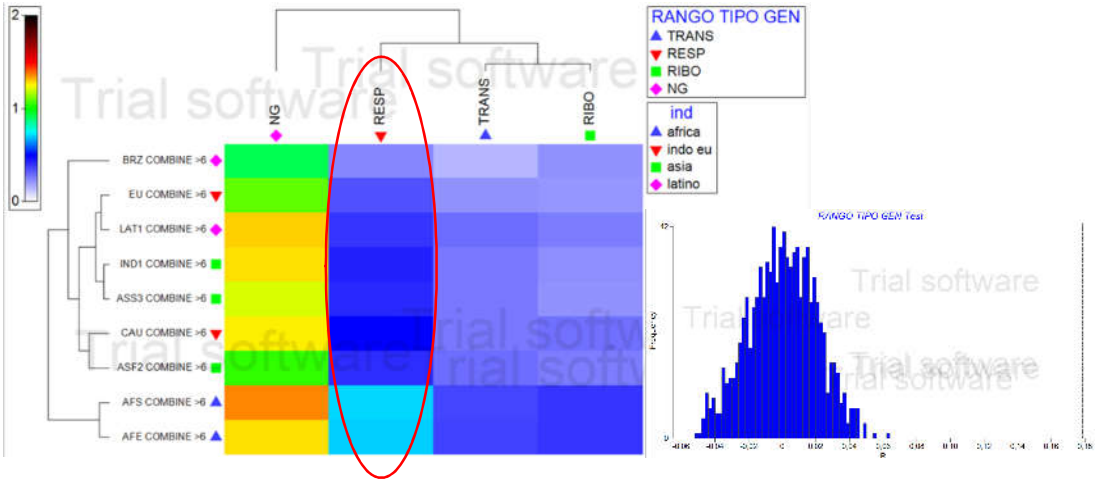


Figure 8. Two-dimensional multivariate analysis of the CSIn value of polymorphic segments in genes comprising ribosomal (Ribo), transfer RNA (Trans), Respiratory (Resp) and gene absence zones (NG). The ANOSIM analysis histogram is shown (right)

In this context, it is necessary to group the polymorphic segments by neighborhood ranges that include contiguous segments of similar type S, SS, SN and N. Each of these ranges is obtained through the average of the values of the polymorphic segments included and they are normalized as CSIn. In the case of CSIn values, a distribution of R = 042 with ANOSIM is observed, so we would face a multivariate structure for the types S, SS, SN and N. **(Figure 9)**

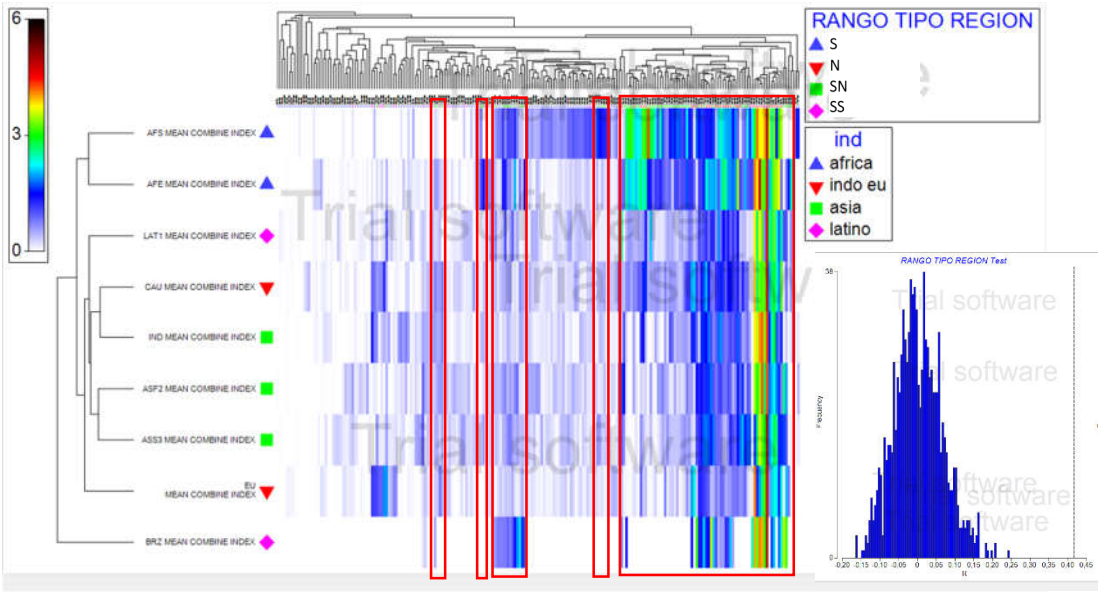


Figure 9. Two-dimensional Heat Map multivariate analysis of the CSIn values of polymorphic segments according to macropopulations and probability of type S, SS, SN and N. The red boxes indicate the areas that comprise polymorphic segments of type SN and N. The histogram of ANOSIM analysis (right) R=0.42.

In the case of Peruvian populations (PEL), it is located in LAT1 and S and SS polymorphic segments are characteristic of the Peruvian population and the same for SN and N in the two-dimensional Heat Map analysis. **(Figure 10)**. Likewise, PEL shares polymorphic segments with different CSIn values with CLM and MXL. PEL contribute with the most polymorphic segments in the LAT1 macropopulation, where theu are mostly S and SS.

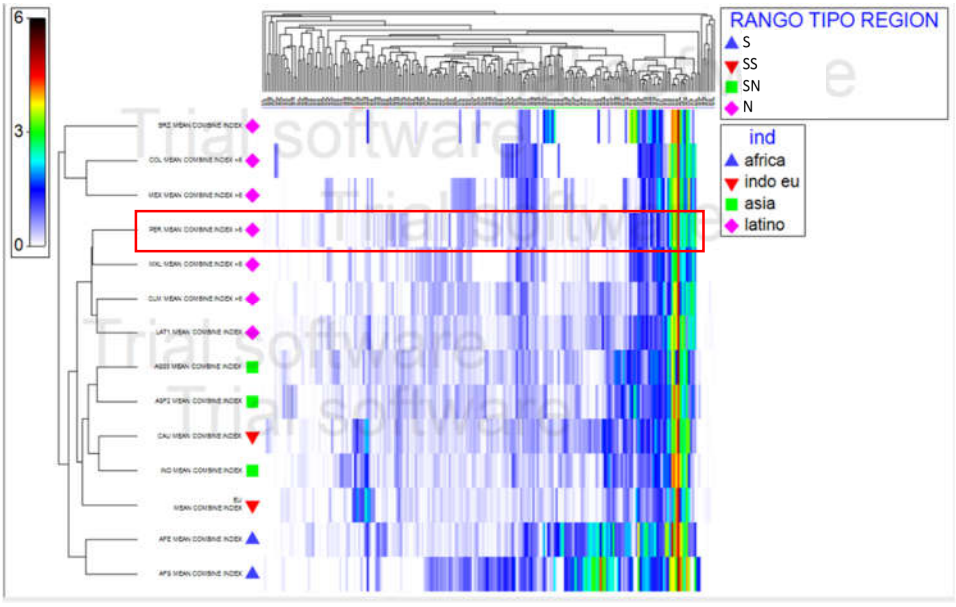


Figure 10. Two-dimensional Heat Map multivariate analysis of the average CSIn values of polymorphic segments according to macropopulations and probability of type S, SS, SN and N. The graph compares the individualized populations of the LAT macro population with respect to all macro populations found. In the red box Peruvian population PEL is indicated.

The distribution of the polymorphic segments of PEL (Peru) with respect to CLM and MXL is confirmed in MDS analysis (Stress = 0.07). In addition, there are three major groups LAT1-PEL-MXL-CLM, IND1-CAU-ASS3-ASF2 and AFE-AFS. In the case of the EUE group, it maintains a different diversity of polymorphic segments that it allows to remain as an isolated group. In the case of the COL, MEX and BZ populations, they also have a different diversity of the polymorphic segments, added to the fact that they are populations where their effective population numbers are smaller compared to the other populations, as we observed in relation to Haplotypic Diversity of the mtDNA genome without the hypervariable region. **(Figure 11)**

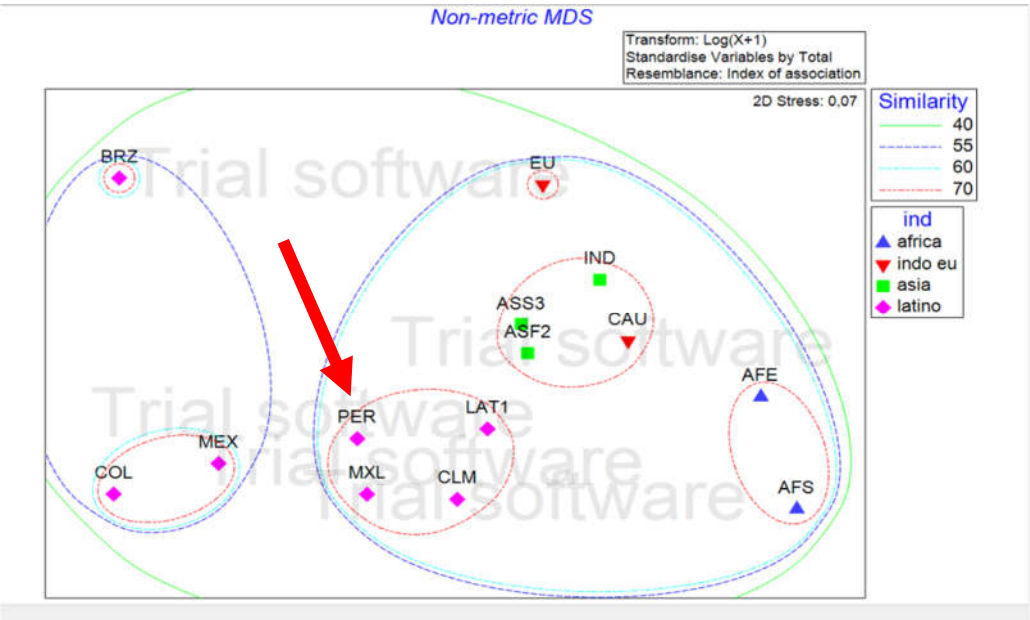


Figure 11. NMDS analysis of the CSIn values of the polymorphic segments with respect to the macropopulations. The red arrow indicates the place where the PEL (Peruvian population) is grouped

DISCUSION

Populations groupingt by genetic distances are essential for population genomics comparisons with the algorithm (GenoPopulation) developed. In that sense, eight macropopulations were found, with a distribution congruent to the migrations of human populations in other studies (Oppenheimer 2012). In the specific case of the Surui population is closer to ASF2 due to the significant genetic relationships with Australasia populations, (Skoglund et al 2015). This being further evidence of additional historical migrations in smaller magnitude (Via South America) to the majority made by the Bering Strait towards America (North-South direction).

On the other hand, the haplotypic diversity observed in the 47 populations (not including the mtDNA hivervariable regions) is 95% on average, this being comparable in magnitude to the diversity found in the hypervariable regions. Thus, the match probability of haplotypes are reduced within macro populations and even more between macro populations.

In addition, the existence of historical correlation in time of the genetic content in each macro population is verified for this type of sequences that it is related to the effective population number (Ne). In the case of Andean populations, a high haplotypic diversity has been observed due to high Ne (as is the case of urban Peruvian populations), unlike Amazonian areas where Ne is lower, (Lewis et al 2007, Cabana et al 2014). In that sense, the polymorphism observed in this study outside the hypervariable region of mtDNA is

complementary in increasing the improvement of haplotype diversity with respect to the hypervariable regions in a percentage that can reach up to 30% as found in others studies, (Just et al 2015) and it is dependent on the Ne in each population.

In that sense, the distribution of the polymorphic segments and the limits (segments of silence of variation) follow a pattern of accumulation more frequently in the sequences related to the respiratory genes and some genes of tRNA as is the case of tRNA L1. However, in some cases of tRNA with high polymorphism it may be influenced by being in proximity to the polymorphisms of neighboring respiratory genes. In the case of the most restrictive segments for the accumulation of polymorphism, there is a higher frequency of SS and S for rRNAs and most tRNAs.

Likewise, the distribution of polymorphic segments at the population level follows a pattern equivalent to that it shown in the neighbor-joinging tree obtained for the 47 populations studied. Thus, when we grouping the polymorphic segments by the criterion neighborhood ranges of selective or neutral probability type. The correlation is significant to classify the ranges of polymorphic segments according to the Combine Segment Index (CSI) and normalized (CSIn). It is according to whether they present lower polymorphism as selective (S) or slightly selective (SS) and if the segments have greater polymorphism they would correspond to the slightly neutral (SN) or neutral (N). The SN and N types accumulated most frequently in the respiratory genes corresponding to the length of greater base pairs but in relation to the total of polymorphic segments, the S and SS types are the most frequent and it are shorter in base pairs length on average than those of type SN and N.

In the case of the 8 macropopulations by the analysis of ranges by neighborhood of type S, SS, SN and N, the two-dimensional Heat Map cluster analysis shows correlation of the CSI value content with the origin of the population (time).Therefore, SN and N accumulate more neutral variants. In the case of selective S and SS, polymorphisms will be in smaller numbers due to the restriction of variation and it is characteristic of a given population because it was added in the time. Likewise, the effect of population admixture is a relevant aspect since it contributes to new variability in each polymorphic segments, whether they are of type S, SS, SN or N. As we observed in the case of the PEL, MXL and CLM populations, they have greater Ne than the populations of BRZ, COL and MEX within the Macropopulation LAT1. The degree of foreign introgression (approx. 75% foreign) is compatible with the reported in previous studies of genetics in Peruvian populations (Iannacone et al 2011 and Sandoval et al 2015)

In addition, it is interesting to note that the degree of standard deviation in the SN + N combination is the lowest of all combinations of the four types and therefore neutral nature is evidenced in the time of maximization of polymorphisms within a segment. However, we cannot expect that a high haplotype diversity correlate to a high distribution in number of these polymorphic segments. Therefore, polymorphism within segments is independent of population histories (migration). In the sense of the polymorphism, the accumulation in each segment is dependent if they are of type S, SS, SN or N.

CONCLUSION:

Based on the results in the populations studied, it supports the need to identify polymorphic segments in the control region and show the potential to improve the polymorphism and the possibility of obtaining more information under conditions of low mtDNA in the case to design specific primers for sequencing by NGS. These population genomics comparisons were achieved for the three indices developed in the algorithm that we have been tested through multivariate analysis.

In an applied sense, developing primers for the SN and N types will allow, on the one hand, to add variability to the hypervariable regions and in cases of low amount of DNA, the option to having more polymorphism information whicg it is increased than when we considering only the hypervariable regions sequencing in the analysis. On the other hand, the S and SS types, due to their polymorphism linked to certain populations, can be associated with phenotypes and used as an additional tool to strengthen identification.

Finally, the use of the designed algorithm (GenoPopulation) becomes a tool that helps to make population comparisons to find population genomic polymorphisms applicable to forensics. Also, it would be interesting automate the algorithm process in a programming language. For this raison we are loking for a parter to developed these goals.

ACKNOWLEDGEMEN´S

To Dr. Pablo Ramírez Roca, head of the Laboratory of Molecular Microbiology and Biotechnology of the Faculty of Biological Sciences of the Universidad Nacional Mayor de San Marcos for his supervision and support in this research.

To the Institute of Legal Medicine for the support provided to carry out the project for the improvement of bone remains workflow for identification purposes.

BIBLIOGRAPHY

Butler JM, The future of forensic DNA analysis. Philos Trans R Soc Lond B Biol Sci. 2015 Aug 5;370(1674)

Brinkmann C, Forster P, Schurenkamp M, Horst J, “Human Y-Cgromosomal STR haplotypes in a Kurdish population sample” Int J Legal Med (1999) 112:181-183

Cabana GS, Lewis CM Jr, Tito RY, Covey RA, Cáceres AM, Cruz AF, Durand D, Housman G, Hulsey BI, Iannacone GC, López PW, Martínez R, Medina Á, Dávila OO, Pinto KP, Santillán SI, Domínguez PR, Rubel M, Smith HF, Smith SE, Massa VR, Lizárraga B, Stone AC, Population genetic structure of traditional populations in the Peruvian Central Andes and implications for South American population history. Hum Biol. 2014 Summer;86(3):147-65.

Chaitanya L, Ralf A, van Oven M, Kupiec T, Chang J, Lagacé R, Kayser M Simultaneous Whole Mitochondrial Genome Sequencing with Short Overlapping Amplicons Suitable for Degraded DNA Using the Ion Torrent Personal Genome Machine. Hum Mutat. 2015 Dec;36(12):1236-47

Just RS, Scheible MK, Fast SA, Sturk-Andreaggi K, Higginbotham JL, Lyons EA, Bush JM, Peck MA, Ring JD, Diegoli TM, Röck AW, Huber GE, Nagl S, Strobl C, Zimmermann B, Parson W, Irwin JA.Development of forensic-quality full mtGenome haplotypes: success rates with low template specimens. Forensic Sci Int Genet. 2014 May;10:73-9.

Just RS, Scheible MK, Fast SA, Sturk-Andreaggi K, Röck AW, Bush JM, Higginbotham JL, Peck MA, Ring JD, Huber GE, Xavier C, Strobl C, Lyons EA, Diegoli TM, Bodner M, Fendt L, Kralj P, Nagl S, Niederwieser D, Zimmermann B, Parson W, Irwin JA. Full mtGenome reference data: development and characterization of 588 forensic-quality haplotypes representing three U.S. populations. Forensic Sci Int Genet. 2015 Jan;14:141-55

Iannacone GC, Parra R, Bermejo M, Rojas Y, Valencia , Portugues L, Medina M, Vallejo AR, Prochanow A, Peruvian genetic structure and their impact in the identification of Andean missing persons: A perspective from Ayacucho, Foren Sci Inter, 2011 Genetics Supplement Series Volume 3, Issue 1, Pages e291–e292

Iannacone GC, “Algorithm aligen: Simple method of encryption and matching with genetic profiles in a system of STR identification database”, Forensic Science International: Genetics Supplement Series 5 (2015) e159–e161

Lewis CM Jr1, Lizárraga B, Tito RY, López PW, Iannacone GC, Medina A, Martínez R, Polo SI, De La Cruz AF, Cáceres AM, Stone AC. Mitochondrial DNA and the peopling of South America. Hum Biol. 2007 Apr;79(2):159-78.

Oppenheimer S, Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map Philos Trans R Soc Lond B Biol Sci. 2012 Mar 19; 367(1590): 770–784

Sandoval JR, Salazar-Granara A, Acosta O, Castillo-Herrera W, Fujita R, Pena SD, Santos FR, Tracing the genomic ancestry of Peruvians reveals a major legacy of pre-Columbian ancestors. J Hum Genet. 2013 Sep;58(9):627-34.

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D. “Genetic evidence for two founding populations of the Americas”, Nature. 2015 Sep 3;525(7567):104-8