
# Variation of *Homo* genus specific NUMTs: insight into the human evolution

**Lluís Zacarías Pons**
Máster Universitario en Bioinformática y Bioestadística (UOC-UB)
Área del trabajo final

**Xavier Jordana Comin**
**David Merino Arranz**

04/06/2019

## FICHA DEL TRABAJO FINAL

| | |
|---|---|
| **Título del trabajo:** | *Variation of* Homo *genus specific NUMTs: insight into human evolution* |
| **Nombre del autor:** | *Lluís Zacarías Pons* |
| **Nombre del consultor/a:** | *Xavier Jordana Comin* |
| **Nombre del PRA:** | *David Merino Arranz* |
| **Fecha de entrega (mm/aaaa):** | 06/2019 |
| **Titulación:** | *Máster Universitario en Bioinformática y Bioestadística (UOC-UB)* |
| **Área del Trabajo Final:** | *M0.128 TFM-Estadística y Bioinformática 29* |
| **Idioma del trabajo:** | *Inglés* |
| **Palabras clave** | *NUMTs; population genetics; paleogenetics* |

**Resumen del Trabajo (máximo 250 palabras):** Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.


En el género *Homo* se han identificado diversas inserciones nucleares de origen mitocondrial (NUMTs), algunas de las cuales se produjeron después de la divergencia de otros primates (1). En algunos estudios se ha estipulado que, dado su origen, estos polimorfismos son buenos candidatos a marcadores para estudiar las relaciones entre las especies del género *Homo* a través de la historia (2), pudiendo ofrecer una nueva perspectiva de la filogenia humana y de eventos ancestrales. Este trabajo propone estudiar la variación de 69 NUMTs exclusivos del género *Homo* (presentes en Neandertales y Denisovas y en poblaciones actuales), usando en este último caso 2.535 genomas (ya alineados) del consorcio *1000 Genomes Project* (3).

Sin embargo, la homología de dichas regiones con el ADN mitocondrial supone un obstáculo a la hora de estudiar su presencia, ya que material genético procedente de la mitocondria puede alinearse erróneamente con el nuclear. Para solventar dicho impedimento, en el presente estudio se ha desarrollado un protocolo de análisis que combina la detección de lecturas solapando las regiones de inserción del NUMT, las distancias entre lecturas apareadas y alineamientos locales con secuencias de referencia con y sin NUMT.

Dicho método ha permitido obtener un catálogo de variaciones de los NUMTs *Homo*-específicos, así como refinar algunas de las coordenadas conocidas. Dichos polimorfismos en poblaciones modernas, junto con algunas muestras de Neandertales y Denisovianas, pueden ofrecer una nueva perspectiva sobre la historia de poblaciones. Al final del trabajo, se ha visto cómo estas inserciones parecen ser buenos marcadores poblacionales.

Abstract (in English, 250 words or less):

*Insertions of mitochondrial DNA (mtDNA) sequences into nuclear genome (NUMTs) have been largely identified in* Homo *genus, some of them occurring after the divergence from other primates* (1)*. The low number of required insertions required to obtain a good level of population differentiation, altogether with the absence of homoplasy in NUMTs (given its mitochondrial origin) suggest that these insertions may be considered a useful marker to trace Homo populations through history* (2)*.*

*Therefore, NUMTs may offer a new perspective of human phylogeny and ancient events, such as the relation and inbreeding among different prehistorical populations. This study aims to study current populations using aligned genomes from 2,535 subjects of the Phase 3 of the 1000 Genomes Project* (3) *and ancestral Neanderthal and Denisovan samples through 69 different homo-specific NUMTs.*

*However, the homology with mtDNA (which may be mis-aligned to the nuclear genome) poses a serious problem to the NUMT presence polymorphism status determination. In addition, since these sequences are mostly located in non-coding regions, they are only available on Whole Genome Sequencing (WGS) samples thus resulting in a really low coverage. To handle these limitations and perform a large-scale analysis, we developed an approach based on overlapping insertion regions detection, paired-end distances and local alignments to the GRCh37 human reference sequence containing and lacking the studied NUMT.*

*This pipeline allowed to obtain a catalog of variation of Homo-specific NUMTs and to refine some given NUMTs position coordinates. Finally, NUMT insertions were concluded to perform a good population analysis.*

## Index

## List of figures



## List of tables

# 1. Introduction

## 1.1 Context and justification of the work

### *1.1.1. General description*

Insertions of mitochondrial DNA sequences into nuclear genome (NUMTs) have been identified in different eukaryotic species (4). In *Homo* genus, some of these events were found to occur after the divergence from other primates, since they are constantly occurring as an ongoing evolutionary process (5). Previous studies showed that a low number of insertions are required in contrast to other autosomal markers to obtain a comparable level of population differentiation. These facts, altogether with the absence of homoplasy in NUMTs (given its mitochondrial origin) suggest that these insertions may be considered a useful marker to trace *Homo* populations through history (2). Therefore, NUMTs may offer a new perspective of human phylogeny and ancient events, such as the relation and inbreeding among different prehistorical populations.

### *1.1.2. Justification of the project*

In a recent study, Riaño-Vivanco et al. (2017) (6) identified four *Homo*-specific NUMTs (absent in other primates) as polymorphic (absence/presence) in modern humans (across all the world) and present in Neanderthals and Denisovans. It was concluded that they must have occurred before the exit of Africa of the *Neanderthal* ancestors, given the low rate of NUMTs deletions (7). Therefore, it was discarded that they were produced with the inbreeding between anatomically modern humans and Neanderthals and Denisovans suggested by the analysis of their genome (Green et al., 2010 (8) and Reich et al., 2010 (9)).

Although these findings cannot refuse the inbreeding of these three *Homo* genus species, the analysis of the polymorphism of more *Homo* specific-NUMTs (whose coordinates were given by C. Santos, personal communication, Feb 27, 2019 based on the previous published database of the group Ramos et al, 2011 (10) and including additional NUMT coordinates discovered in Dayama et al, 2014 (5)) may elucidate their nature and give a new perspective and an approach to a new model of the human population history.

## 1.2 Objectives

### *1.2.1. General objectives*

1) To obtain a catalog of variation of *Homo* genus NUMTs in Neanderthal and Denisova, as well as in ancient and modern humans.

2) To use the NUMTs polymorphisms to offer a new perspective of human population phylogeny.

*1.2.2. Specific objectives*

1) To elaborate the NUMTs polymorphism catalog:

(a) Sample modern and ancient *Homo sapiens*, Neanderthal and Denisovan individuals for the analysis
(b) Elaborate a script to sample individuals across different databases
(c) Perform a script which unequivocally checks the presence/absence for known NUMTs controlling for mtDNA that might be erroneously aligned to the nuclear genome
(d) Develop an approach to differentiate three different categories for autosomic NUMTs for each individual (homozygous for deletion, heterozygous and homozygous for insertion)

2) To study human populations using NUMTs:

(a) Explore different approaches and bioinformatic tools (especially those within the R/Bioconductor environment) for phylogenetic analysis using presence/absence and sequence polymorphisms
(b) Perform a population analysis using R
(c) Compare the results to the existent bibliography about human population history

## 1.3 Selected approach

The *Homo sapiens* modern samples will be obtained from the Phase 3 of the 1000 Genomes Project (3), which uses the GRCh37 coordinates system (our known NUMTs coordinates (5,10) are given following this system). This database has been chosen for the accessibility of the NUMTs regions without downloading the whole BAM files (whose sizes management goes beyond the means of the author). Samples from Neanderthals and Denisovans containing the NUMTs regions will be also collected. In the latter cases, samples will be extracted from the public genomes data described and available at the web server of the Department of Evolutionary Genetics of the Max Planck Institute, accessible from the Genomes Project section of their website (11). In order to determine the presence/absence polymorphisms these BAM alignment files will be analyzed.

Two different approaches will be performed depending on whether the NUMT sequence is present or not in the GRCh37 reference genome.

In the first case, given the homology between NUMTs and mitochondrial DNA (that could be present in the samples when the sequencing and alignment were performed) some NUMTs regions may seem to have

nuclear aligned reads (thus suggesting a presence polymorphism) even when there is not any insertion. To assess this problem, an insertion will be confirmed only if there are some uncut reads overlapping the insertion coordinates. Additionally, two local alignments will be performed to assess the genotype status (one against a chain containing the full reference sequence, the other containing a deletion in the NUMT region) checking for deletion status.

Otherwise, to tackle the NUMT presence in the NUMTs that are not taken into account in the reference genome different parameters will be analyzed: the number of reads overlapping the insertion point (thus suggesting a deletion allele), the number of reads aligning to the mtDNA sequence and the number of reads around the insertion region whose mate is mapped in the mitochondrial chromosome.

After the classification, some NUMT positions will be manually checked (as a script quality control) in some individuals using the IGV viewer (12). Finally, the output variables from the latter scripts will be processed to determine the genotype for each individual. A phylogenetic analysis will be also performed within the R-Bioconductor environment.

## 1.4. Planification

The initial planification was designed as follows, even if some difficulties furtherly discussed originated delays and limited the planned goals to be achieved.

### *1.4.1. Tasks*

1) To elaborate the NUMTs polymorphism catalog:

A) To sample *Homo sapiens*, Neanderthal and Denisovan individuals for the analysis

- To explore databases and projects to find whole genome sequenced individuals from different genomic databases containing different current populations (04-17/03/2019)
- To find sequenced ancestral genomes from ancient *Homo sapiens* as well as from Neanderthals and Denisovans covering the human-specific NUMTs regions (04-17/03/2019)

B) To elaborate a script to sample individuals across different databases

- To find and explore R-Bioconductor tools for BAM files sampling and management (11-17/03/2019)
- To write the script (11-24/03/2019)
- To open some of the NUMT-region generated BAM files to check its quality (17-24/03/2019)

C) To perform a script which unequivocally checks the presence/absence for known NUMTs controlling for mtDNA that might be erroneously aligned to the nuclear genome

- To find and explore R-Bioconductor tools for BAM files analysis (11-17/03/2019)
- To write the script (18/03-07/04/2019)

D) To develop an approach to analyze the sequence polymorphisms in different individuals

• To think about a pipeline which allows to differentiate NUMTs and mitochondrial DNA sequences (11-24/03/2019)

- To write the script (01/04-21/04/2019)
- To check its results (22-24/04/2019)

2) To study the human phylogeny using NUMTs:

A) To explore different approaches and bioinformatic tools (especially those within the R/Bioconductor environment) for population analysis using presence/absence and sequence polymorphisms:

- Literature research on methods for population analysis (25/03-24/04/2019)
- To look for R-Bioconductor packages to perform these analysis (25/04-5/05/2019)

B) To perform the population analysis elaborating an R script (06-12/05/2019)

C) To compare the results to the existent bibliography about human populations history

- To look for literature about human populations history (25/04-12/05/2019)
- To evaluate how the generated population analysis differs from the general statements about human history (13/05-27/05/2019)

The Gantt Diagram of the Figure 1 illustrates the tasks list and its achievement.

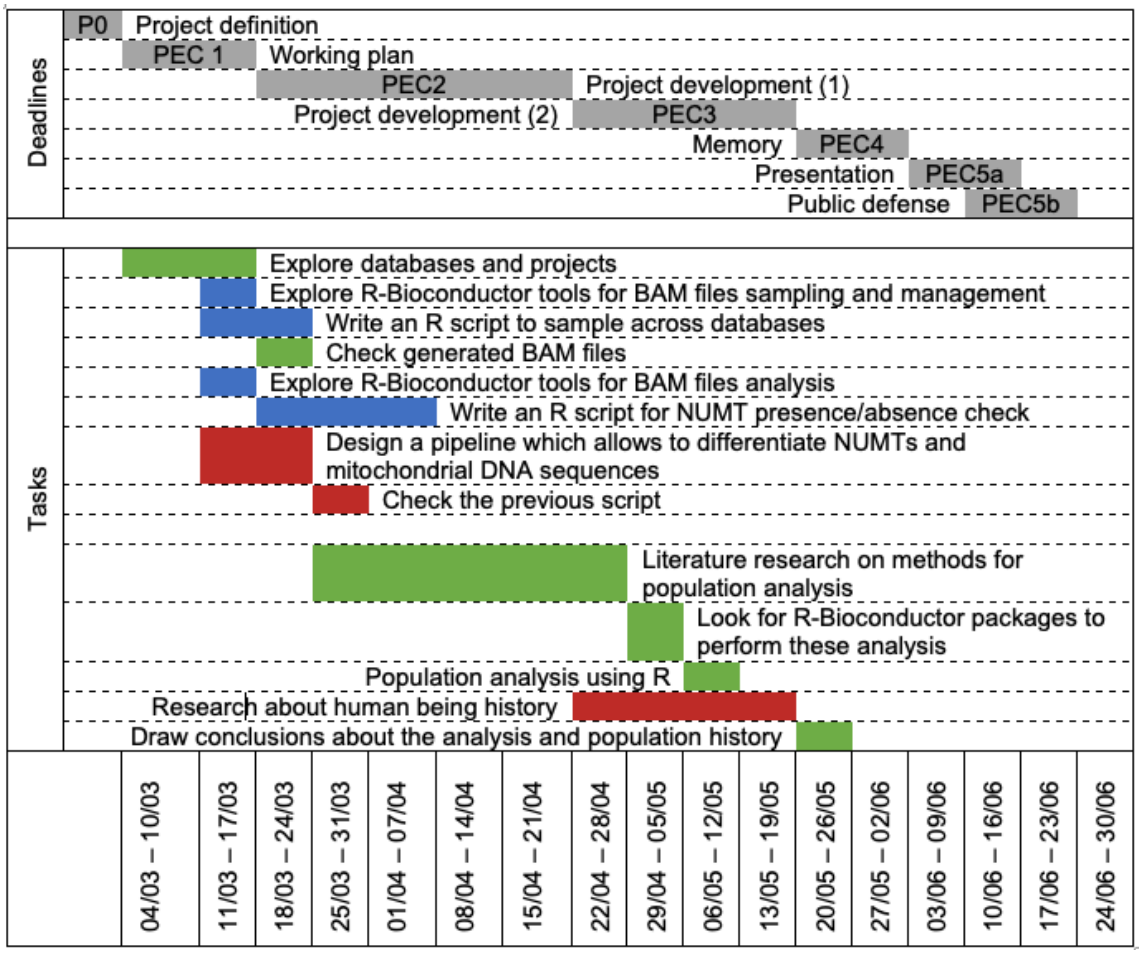



***Figure 1. Gantt Diagram*** *comparing the established deadlines and the programmed tasks at the beginning of the project. Performed tasks are colored in green, performed tasks in another programming language (Bash instead of R) are colored in blue and non-performed tasks are coded in red.*

### *1.4.2. Risk analysis*

1. Impossibility to study sequence polymorphisms due to the low coverage of Whole Genome Sequencing. In that case, other sources of information will be taken in account, such as Whole Exome Sequencing (also available from the 1000 Genomes Project). If none of the alternatives serves to tackle this risk, the phylogenetic analysis of presence/absence polymorphisms will be the only one performed.

2. Not enough human-specific polymorphic NUMTs to draw conclusions. In that case, this study will be limited to an elaboration of a polymorphic NUMTs catalog, although this is a really improbable scenario.

### 1.5. Brief summary of the obtained products

i. **Scripts**

   a. **1000genomesIndexScript.R** *Bash script used to prepare an index of links for BAM files*

   b. **1000genomesSampling.sh** *Bash script used to sample genomes from the 1000 genomes project*

   c. **referenceNUMTscript.sh** *Bash script used to analyze referenced NUMTs*

   d. **nonreferenceNUMTscript.sh** *Bash script used to analyze non-referenced NUMTs*

   e. **reference.sh** *Bash script used to generate NUMT and non-NUMT reference sequences*

   f. **Ranalysis.R** *Script used to analyze the output of Bash scripts*

ii. **IndexFiles** *Files used to iterate the sampling loop*

   a. **1000genomesindex.txt**

   b. **1000genomeslist.txt**

   c. **numtsbash.txt**

iii. **Intermediate files**

   a. **ModernIndividualsResults** (69 files) *Results of Bash scripts for modern samples to be processed using R*

   b. **AncestralResults** (69 files) *Results of Bash scripts for ancestral samples to be manually processed*

   c. **DenisovaHSAResults** (46 files) *Results of Bash scripts for the Denisovan sample on referenced NUMTs to be manually processed*

iv. **S1_Table.xlsx** *Table containing counts for different haplotypes of referenced NUMTs through modern populations*

v. **S2_Table.xlsx** *Table containing counts for different haplotypes of non-referenced NUMTs through modern populations*

### 1.6. Section description

2. **Genome sampling: Methods & Resources** *Description of the genomes sampling process and the source databases*

2.1. **Modern genomes: The 1000 Genomes Project** *Description of modern genomes sampling*

2.2. **Ancestral genomes: The Max Planck Institute Repository** *Description of ancestral genomes sampling*

3. **NUMT detection** *Workflow description of BAM files analysis using Bash*

    3.1. **NUMTs within the reference genome** *Pipelines for NUMTs included in the GRCh37 reference genome*

    3.2. **NUMTs not within the reference genome** *Pipeline for NUMTs not included in the GRCh37 reference genome*

4. **Genotype status and population analysis** *Phase of the analysis using R, in which genotype status for each individual is settled following three specific protocoles*

5. **Results**

    5.1. **NUMT insertion/deletion haplotypes and population frequencies** *Results of genotypes definitions and quality control plots for certain regions*

    5.2. **Hardy-Weinberg Equilibrium for each NUMT and population** *HWE to evaluate the analyzed NUMTs as possible population markers*

    5.3. **Principal Component Analysis** *Results for modern populations analysis*

    5.4. **NUMTs in ancestral individuals** *Catalog approach for for ancestral populations*

6. **Discussion** *Brief discussion about the obtained results*

7. **Conclusions** *Main conclusions of the project and future directions*

# 2. Genome sampling: Methods & Resources

Given its wide use and teaching during the Master's Program, the whole analysis in this project (including the genome sampling) was initially planned to be performed within the R-Bioconductor environment. In fact, there are packages allowing the extraction of a certain region from a BAM file located in a server, such as Rsamtools (13).

Nevertheless, some difficulties that appeared during the beginning of the project. Firstly, each region for each sample took a lot of space when stored into a variable. In addition, a really long time for sampling all individuals was estimated following this approach and considered unaffordable for the study. Finally, the process was not robust enough and crashed constantly during the iteration over all samples.

Therefore, all these problems led to conclude that R/Bioconductor tools for managing these files were more appropriate to manage small regions or a small sample size, but not the best for the addressed problem. Thus, leading to change the used language. The sampling was finally performed using SAMtools (14), written to be performed within the bash Unix shell command language.

## 2.1. Modern genomes: The 1000 Genomes Project

Between 2008 and 2015, the 1000 Genomes Project was conceived as a catalogue of the genetic human variation, sampling individuals from different populations across the whole world (Figure 1). At the beginning of this project, 2,535 Whole Genome Sequence (WGS) alignments from different individuals of 26 different populations were available at the International Genome Sample Resource (15), as a part of the Phase 3 of the project (3). They consist of paired-end reads aligned to the GRCh37 reference genome.


***Figure 2. Map of populations included in the 1000 Genomes Project.*** *Source: [http://www.international genome.org/home](http://www.international genome.org/home)*

In order to sample reads aligned to regions of all studied NUMTs (including their respective flanking sequences, 1000pb per each side), a bash script executing SAMtools (14) was iterated over every individual and region. It is available as *1000genomesSampling.sh* in the annex. Two additional files were provided as the output for the script: one for the address of each sample (*1000genomesindex.txt*) and another containing the desired regions (*numtsbash.txt*).

After approximately 3 days of computation using 5 parallel operations, all regions were sampled for every individual. All reads were stored into plain text files, with one file per individual and region. They were written in SAM format (the uncompressed version of BAM files) (16) to be processed in the next phase.

### 2.2. Ancestral genomes: The Max Planck Institute Repository

Besides the modern human genomes, Neanderthal and Denisovan alignments against the same reference genome (GRCh37) were sampled in order to evaluate the NUMT polymorphism status for ancestral individuals.

These samples were taken from the Genomes Projects repository of the Department of Evolutionary Genetics of the Max Planck Institute (11). There are four high-coverage samples (three from Neanderthal and one from Denisova) and different low coverage samples from 5 different Neanderthals (more recent than the 3 high-coverages samples) from 5 different archeological sites.

Three high-coverage samples came from the Denisova cave. Two of them were Neanderthals of ~50,000 and 120,000-130,000 years old (Vindija 33.19 (17) and Altai (18)) whose most reads were single end reads. The other one is a Denisovan individual (whose reads are completely paired-end) of approximately 74,000-82,000 years old (19). The last high-coverage sample is from and individual found in the Chagyrskaya cave (20), and the produced reads were mostly non-paired. Their sampling was performed in a similar fashion as previously done with the modern samples.

Finally, the five low-coverage samples come from the Goyet (Goyet Q56-1, 43,000-42,080 years before present) and the Spy caves (Spy94a, 39,150–37,880 cal. yr BP) in Belgium, Les Cottés cave (Les Cottés Z4-1514, 43,740–42,720 cal. yr BP) in France, the Vindija cave (Vindija 87, older than 44,000 uncalibrated years BP) in Croatia and from the Mazmaiskaya cave (Mezmaiskaya 2, 44,600–42,960 cal. yr BP) in Russia. Because of BAM index files (.bai) not being available at the server, the whole BAM files had to be download and then locally indexed using SAMtools (14). After being indexed, regions extraction was performed from these local BAM files.

| NUMTS | ID |
|---|---|
| HSA_NumtS_009_b1 | chr1:38077350-38077420 |
| HSA_NumtS_015_b1 | chr1:104163778-104163820 |
| HSA_NumtS_030_b1 | chr1:147332804-147332915 |
| Poly_NumtS_139 | chr1:170225627-170225631 |
| HSA_NumtS_038_b1 | chr1:215673139-215673177 |
| Poly_NumtS_1239 | chr2:33892476-33892481 |
| HSA_NumtS_050_b1 | chr2:33992539-33992587 |
| Poly_NumtS_1259 | chr2:53395587-53395589 |
| HSA_NumtS_058_b1 | chr2:81893601-81893852 |
| HSA_NumtS_092_b1 | chr2:149639295-149639426 |
| Poly_NumtS_1440 | chr2:231155927-231155930 |
| HSA_NumtS_133_b1 | chr3:25508995-25509033 |
| HSA_NumtS_143_b1 | chr3:96336032-96337354 |
| HSA_NumtS_171_b1 | chr4:12641918-12642262 |
| Poly_NumtS_1843 | chr4:29441246-29441248 |
| HSA_NumtS_179_b1 | chr4:47774289-47774381 |
| HSA_NumtS_182_b1 | chr4:56194327-56194457 |
| HSA_NumtS_188_b1 | chr4:78929684-78929923 |
| Poly_NumtS_1900 | chr4:83616162-83616166 |
| Poly_NumtS_1929 | chr4:120928121-120928123 |
| HSA_NumtS_200_b1 | chr4:160965748-160965826 |
| HSA_NumtS_201_b1 | chr4:163342526-163342693 |
| Poly_NumtS_2010 | chr5:32338582-32338584 |
| HSA_NumtS_214_b1 | chr5:73071708-73071757 |
| HSA_NumtS_215_b1 | chr5:79945841-79948187 |
| HSA_NumtS_228_b1 | chr5:134258999-134264217 |
| HSA_NumtS_232_b1 | chr5:165957424-165957466 |
| Poly_NumtS_2186 | chr6:36893177-36893179 |
| Poly_NumtS_2289 | chr6:138775077-138775079 |
| Poly_NumtS_2377 | chr7:57413680-57413682 |
| HSA_NumtS_276_b1 | chr7:68201515-68201620 |
| HSA_NumtS_289_b1 | chr7:145694423-145694525 |
| HSA_NumtS_304_b1 | chr8:36135128-36137214 |
| Poly_NumtS_2541 | chr8:6301423-63015431 |
| Poly_NumtS_2578 | chr8:97544159-97544382 |

| NUMTS | ID |
|---|---|
| HSA_NumtS_319_b1 | chr8:100508098-100508181 |
| Poly_NumtS_2611 | chr8:126230601-126230603 |
| Poly_NumtS_2653 | chr9:37621680-37621685 |
| Poly_NumtS_316 | chr10:92133616-92133618 |
| Poly_NumtS_430 | chr11:49883571-49883575 |
| Poly_NumtS_445 | chr11:69874823-69874826 |
| HSA_NumtS_398_b1 | chr11:73221706-73221868 |
| HSA_NumtS_410_b1 | chr11:122874314-122874385 |
| Poly_NumtS_531 | chr12:6475423-6475425 |
| HSA_NumtS_426_b1 | chr12:41757438-41757525 |
| HSA_NumtS_428_b1 | chr12:50211122-50211285 |
| HSA_NumtS_460_b1 | chr13:41342484-41342558 |
| HSA_NumtS_464_b1 | chr13:56545768-56545893 |
| Poly_NumtS_709 | chr13:64162211-64162213 |
| HSA_NumtS_472_b1 | chr13:110076472-110076727 |
| HSA_NumtS_474_b1 | chr14:32953304-32954324 |
| HSA_NumtS_512_b1 | chr17:42075084-42075151 |
| HSA_NumtS_513_b1 | chr17:51183094-51183746 |
| HSA_NumtS_518_b1 | chr17:78591376-78591422 |
| HSA_NumtS_519_b1 | chr18:2842198-2842352 |
| HSA_NumtS_522_b1 | chr18:45379615-45379808 |
| HSA_NumtS_543_b1 | chr20:9149571-9149612 |
| Poly_NumtS_1465 | chr20:9282578-9282580 |
| HSA_NumtS_544_b1 | chr20:13147959-13148001 |
| Poly_NumtS_1480 | chr20:21455100-21455451 |
| HSA_NumtS_546_b1 | chr20:55639110-55639179 |
| Poly_NumtS_1583 | chr22:27447873-27447875 |
| HSA_NumtS_560_b1 | chr22:36281719-36281765 |
| HSA_NumtS_569_b1 | chrX:125605687-125606435 |
| HSA_NumtS_570_b1 | chrX:125606450-125606718 |
| HSA_NumtS_573_b1 | chrX:142522269-142522941 |
| HSA_NumtS_574_b1 | chrY:4212822-4212892 |
| HSA_NumtS_578_b1 | chrY:8979505-8979570 |
| HSA_NumtS_585_b1 | chrY:21033988-21034133 |

***Table 1. Analyzed NUMTs*** *and their coordinates within the GRCh37 reference genome. NUMTs whose sequence appear in the reference genome are green-colored. Those insertions that are not within that sequence are colored in blue.*

# 3. NUMT detection

As previously mentioned, two kind of NUMTs (Table 1) were analyzed in this project: some of them were included in the GRCh37 consensus reference sequence, whereas some of them were not. Since this study's workflow focus on analyzing reads taken from BAM files (that is, they have been previously aligned against the GRCh37 reference sequence), different considerations are required to properly tackle both types.

Since analyzing one read at a time requires a lot of iterations, the NUMT performed was also performed using a Bash UNIX shell script, that might repeat the same loop faster than other languages (such as R).

NUMT sequence analysis was finally not performed, even if it is not discarded as a future working line.

## 3.1. NUMTs within the reference genome

The first approach, the one designed to handle referenced NUMTs, relies on the fact that reads are paired-end. One may recall that NUMTs are highly similar to mitochondrial sequences (given its origin). Therefore, some reads coming from the mitochondrial sequence may be misaligned to the NUMT reference sequence. To prevent mitochondrial reads to be considered as part of the individual's NUMT sequence, only reads whose mate is mapped in the NUMT flanking regions will be taken into account to compute the different parameters that will allow to state the genotype (deletion/deletion, deletion/insertion or insertion/insertion) for each individual.

Since the computation of the coverage within the NUMT region is a part of the presented pipeline (as a quality control rather than as a predictive parameter, as it will be further exposed), a special caution must be taken when it comes to large NUMTs. Because of read mates distance being limited and this approach considering only reads with a mate mapped in the flanking regions, within NUMT coverage might be underestimated for long NUMTs (where the whole pair of reads must be mapped within the studied region). Therefore, two slightly different "sub-approaches" were developed. One for NUMTs whose length was not longer than the mean insert size (i.e. reads length plus the distance separating them) minus 40, another for the rest.

### *3.1.1. Pipeline for "short" NUMTs*

The pipeline for "short" NUMTs (illustrated in the Figure 3) is basically based on three master lines:

i) Computation of the number of reads overlapping the NUMT insertion positions in an unbroken manner (i.e. without being "soft-clipped" in that position).

ii) Local alignments of reads overlapping the insertion position against the original reference sequence and against a “deletion” reference sequence (assembled from the original sequence using the *reference.sh* script). These alignments were performed using the MEM algorithm included in the BWA tool (21). A read is thus considered to come from a non-NUMT sequence when its alignment has a better CIGAR score when aligning the latter and vice versa.

iii) Computation of the within NUMT coverage and the flanking regions coverage.

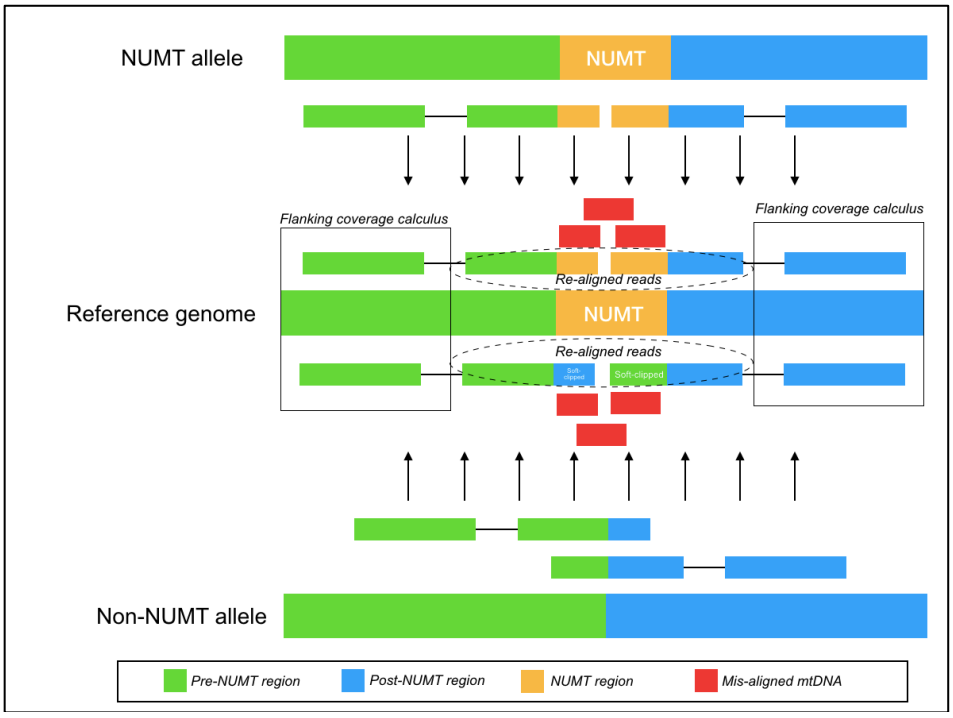



***Figure 3. Pipeline for “short” referenced NUMTs*** *applied to a deletion or an insertion allele. Black arrows represent the alignment to the GRCh37 reference genome that was originally performed to generate the sampled BAM file.*

### *3.1.2. Pipeline for “long” NUMTs*

This workflow is essentially equal to the previous one, although two slight differences:

i) Within NUMT coverage is not computed, since it might be underestimated

ii) When the NUMT is considerably long, it is possible to detect a NUMT deletion if a read pair has each pair in a different side of the NUMT whose insert size is considerably larger than the average insert size (as illustrated in the Figure 4). This fact was thus included into the pipeline to use it as an informative deletion marker.

Both methods were performed using the same script (*referenceNUMTscript.sh*), since the selected method must be selected for each individual even within the same NUMT (because of samples having different average insert size).

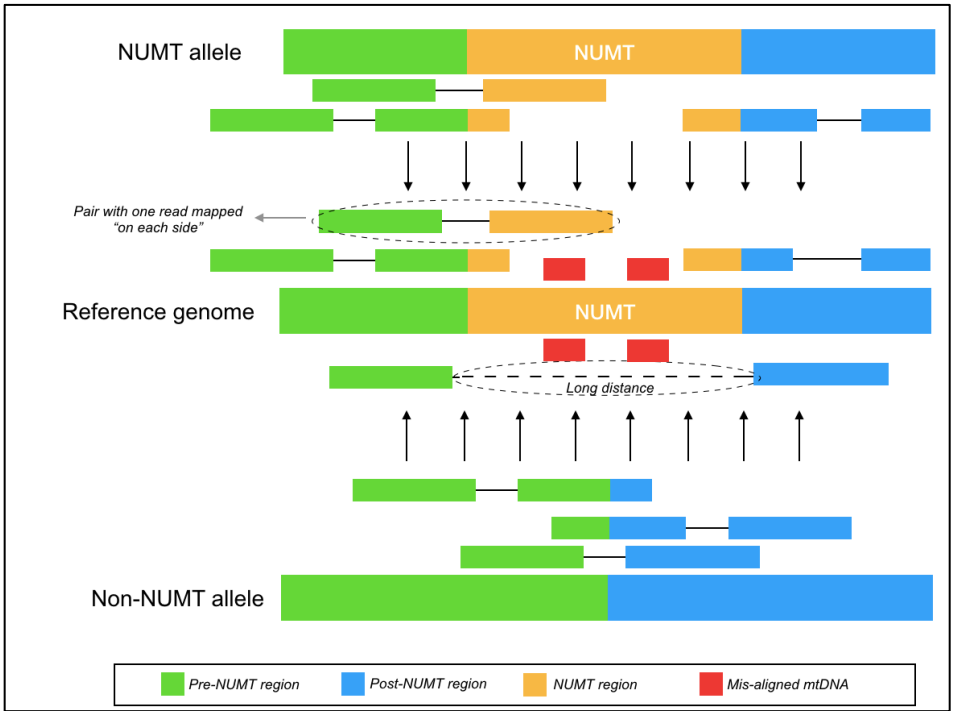



***Figure 4. Pipeline for "long" referenced NUMTs*** *applied to a deletion or an insertion allele. Black arrows represent the alignment to the GRCh37 reference genome that was originally performed to generate the sampled BAM file.*

### 3.2. NUMTs not within the reference genome

As previously explained, a different approach was implemented for NUMTs not present in the GRCh37 reference sequence. These NUMTs were discovered in Dayama et al. 2014 (5) and they have been processed in a similar fashion as proposed in the mentioned study, but using the Phase 3 of the 1000 Genomes Project and the sampled ancestral data.

Since these NUMTs are not in the reference sequence (used to map reads thus generating the BAM files), reads containing exclusively its sequences are not "placed" (mapped) within the studied regions (and therefore not appearing among the sampled reads). As a counterpart they will be mapped in the most similar region, that is, the mitochondrial chromosome.

In addition, those reads covering part of NUMT and part of flanking region will be probably mapped but soft-clipped after the "breakpoint" (position of the reference sequence where the NUMT is supposed to be). On the other hand, reads supporting the NUMT deletion will be fully aligned at that position.

Combining these facts, altogether with an alignment approach, a pipeline (visually described in Figure 5) was designed:

i) Computation of flanking regions coverage.
ii) Count of reads overlapping the breakpoint position in an unbroken manner.
iii) Alignment of reads mapped over the breakpoint against both nuclear and mitochondrial sequences, to check whether they also align to the mitochondrial chromosome.
iv) Count of reads whose mate is mapped in the mitochondrial chromosome (and should thus be mapped into a hypothetical proper reference sequence, where the NUMT is taken into consideration).

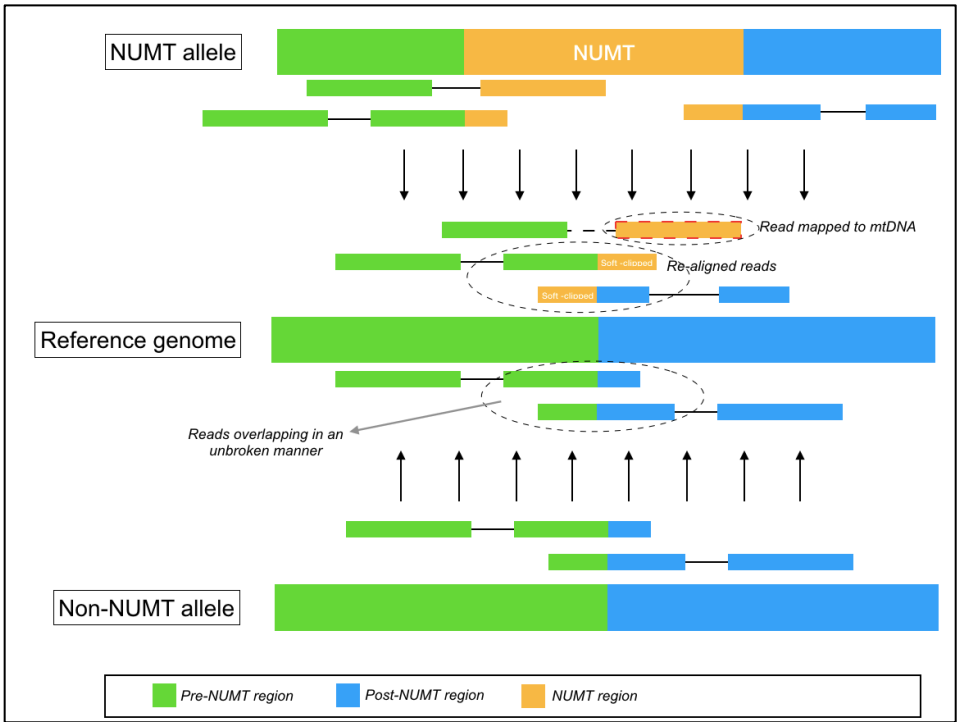



***Figure 5. Pipeline for non-referenced NUMTs*** *applied to a deletion or an insertion allele. Black arrows represent the alignment to the GRCh37 reference genome that was originally performed to generate the sampled BAM file.*

It is worth to mention that to properly perform the execution of the designed script (*nonreferenceNUMTscript.sh*), breakpoint positions for 4 NUMTs (Poly_NumtS_139, Poly_NumtS_1239, Poly_NumtS_2541, Poly_NumtS_1480) had to be refined. That review was achieved through a manual revision of reads in those samples in which some read was detected to have a mate mapped in the mitochondrial chromosome (using the IGV viewer).

# 4. Genotype status and population analysis

After executing the Bash UNIX shell script over all reads in all regions across every sampled individual, its outputs were stored in plain text files (one for each region for modern or ancestral individuals). They were then processed in the R environment, applying different conditions over each individual to determine his/her genotype status (absence/absence, presence/absence or presence/presence).

The workflow to define the genotype status is described in figures 6-8.

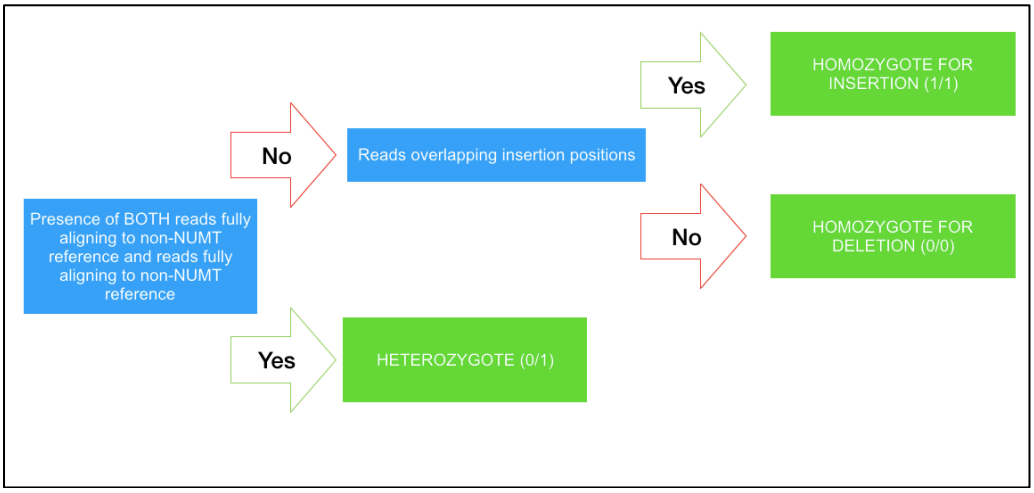



*Figure 6. Workflow to define the genotype from the output obtained from the Bash script described in section 3.1.1 (for "short" referenced NUMTs).*

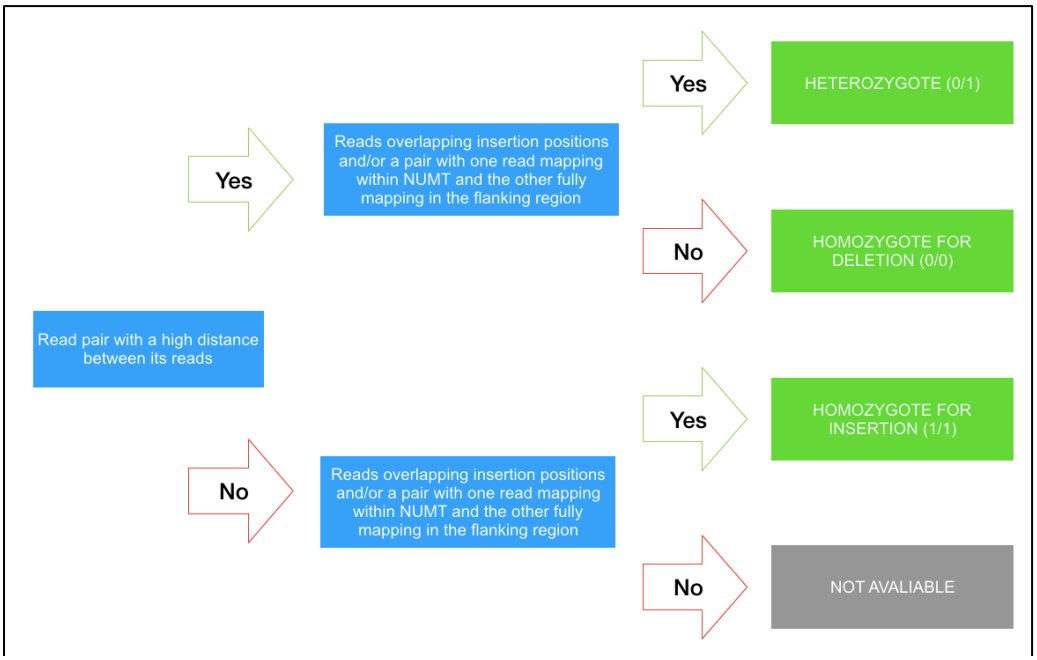



*Figure 7. Workflow to define the genotype from the output obtained from the Bash script described in section 3.1.2 (for "long" referenced NUMTs).*

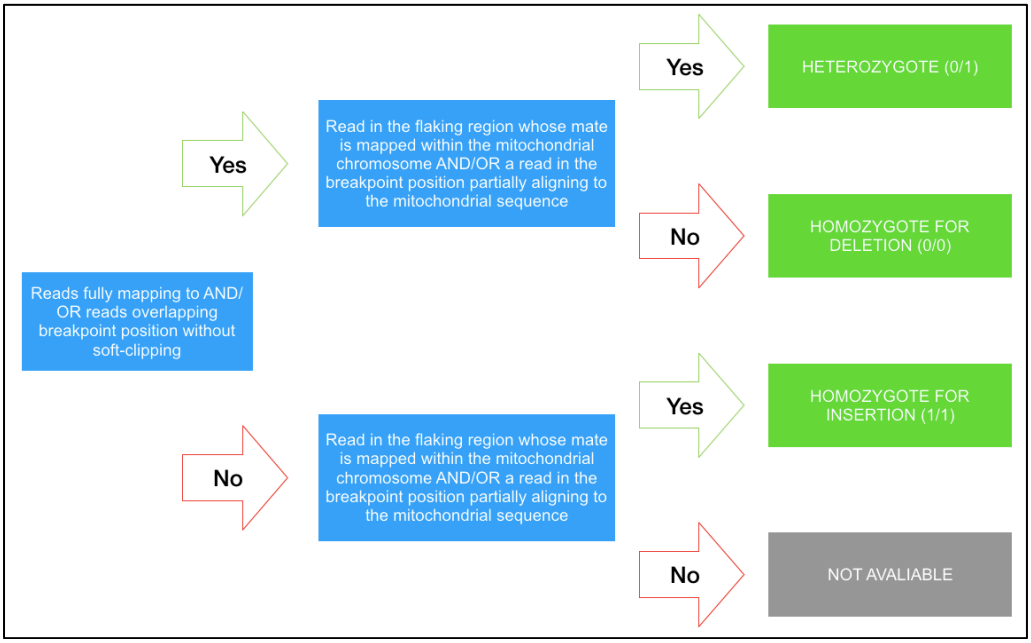



*Figure 8. Workflow to define the genotype from the output obtained from the Bash script described in section 3.2 (for non-referenced NUMTs).*

After defining a genotype for each individual, different plots were designed to validate the proposed approach (Figures 9-26) using the *ggplot2* package for R (22). For NUMTs whose sequence is reflected in the reference genome, coverage within NUMT and flanking coverage were compared. On the other hand, the count of reads overlapping the breakpoint position was compared to flanking coverage when it came to non-referenced NUMTs.

Additionally, Hardy-Weinberg equilibrium for polymorphic NUMTs was checked using the *HardyWeinberg* package for R (23). This test is widely used to consider proper population markers.

Finally, a NUMT-based population analysis was performed. It relies on the total population allele frequencies calculation resulting from aggregating biallelic data according to the proposed formula:

$$NUMT_{Freq} = \frac{(\,Number\ of\ heterozygotes\ \times\ 0.5\,)\ +\ Number\ of\ insertion\ homozygotes}{Number\ of\ individuals\ within\ population}$$

The resulting frequencies for each NUMT were then used to perform a Principal Component Analysis. It was computed separately for reference NUMTs and for non-reference NUMTs and then applied to the whole NUMT polymorphism information.

To draw conclusions about the ancestral individual outputs, bash UNIX shell script was manually checked to see whether NUMT and/or non-NUMT alleles were detected within each sample. The workflow was the same as used in modern humans in Denisovan sample, whereas a

different approach (not based on paired-end data) was used to evaluate the referenced NUMTs in other ancestral samples.

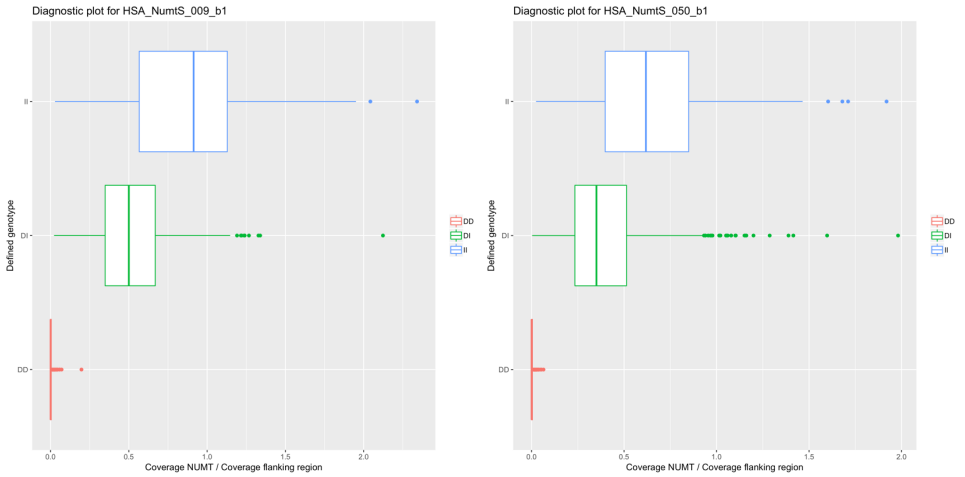
The script containing the R instructions for the analysis explained in this section is available as the *Ranalysis.R* script.

# 5. Results

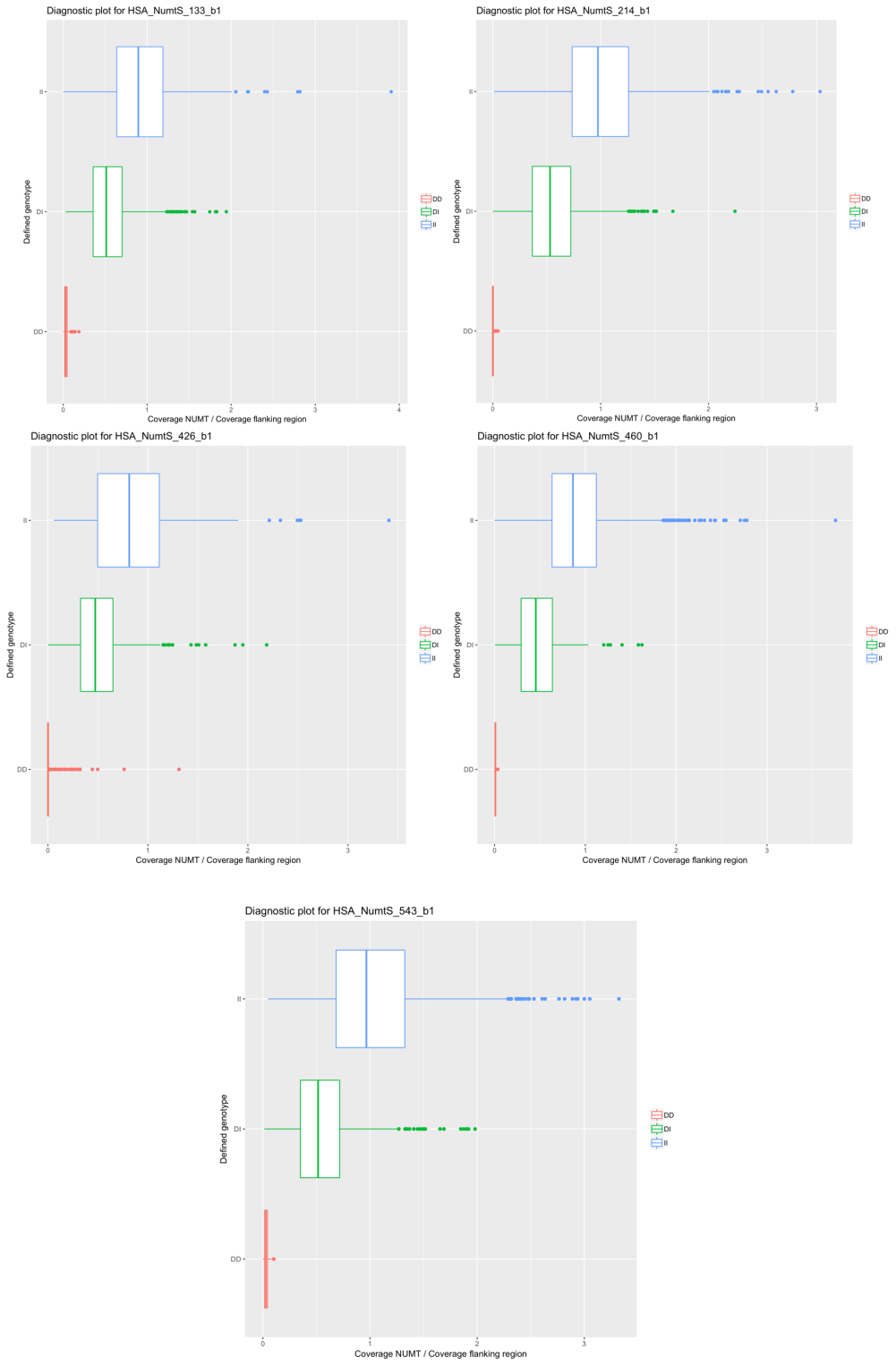
## 5.1. NUMT insertion/deletion haplotypes and population frequencies

For those NUMTs included in the human reference sequence, HSA_NumtS_009_b1, HSA_NumtS_050_b1, HSA_NumtS_133_b1, HSA_NumtS_214_b1, HSA_NumtS_426_b1, HSA_NumtS_460_b1, HSA_NumtS_543_b1, HSA_NumtS_474_b1 and HSA_NumtS_522_b1 turned to be polymorphic in modern populations. Their genotype counts for each population of the Phase 3 of the 1000 Genomes Project may be observed in Supplementary Table 1 (in Excel format). All other "referenced" NUMTs where considered to be present across every sampled individual. The HSA_NumtS_015_b1, was found to be present in low counts in all European populations (3 in CEU, 3 in GBR, 2 in IBS, 1 in FIN and 4 in TSI), in some American populations (5 in CLM, 3 in PUR and 2 in MXL) and in one South-Asian population (1 in BEB). However, it was not possible to obtain a proper genotype prediction for each individual. It was thus discarded from the rest of the analysis and must be furtherly addressed.

As a quality control, the proportion Within NUMT Coverage / Flanking coverage was computed for each individual and region and then plotted as seen in figures 9-15 (for those NUMTs checked using the protocol described in section 3.1.1). As one may observe, that proportion is mostly 0 for deletion/deletion haplotypes, 0.5 for deletion/insertion and 1 for insertion/insertion. These data suggest that the genotype determination was accurate when taking all individuals into account.



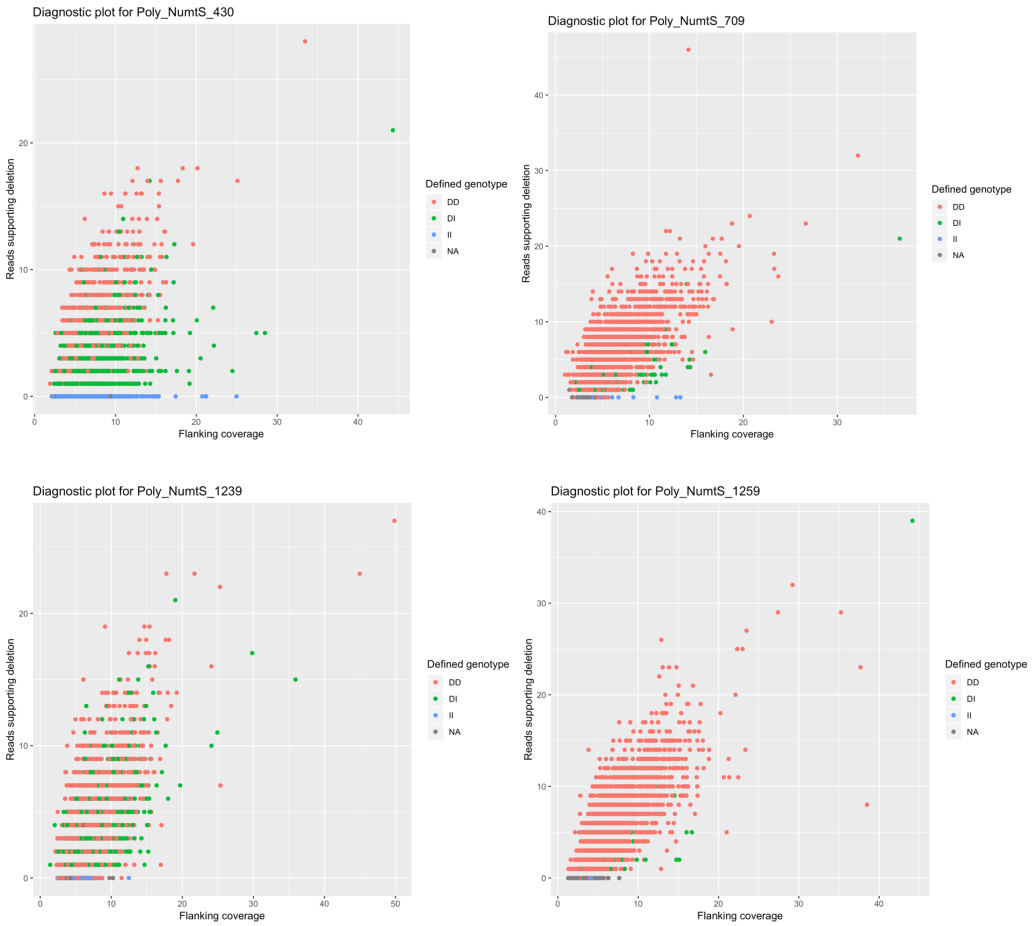

***Figures 9-10. Diagnostic plots*** *for HSA_NumtS_009_b1 and HSA_NumtS_050_b1. The NUMT coverage / Coverange flanking region proportions for each individual are plotted. Absence/absence haplotypes are plotted in red, heterozygotes are plotted in green and homozygotes for insertion are plotted in blue.*



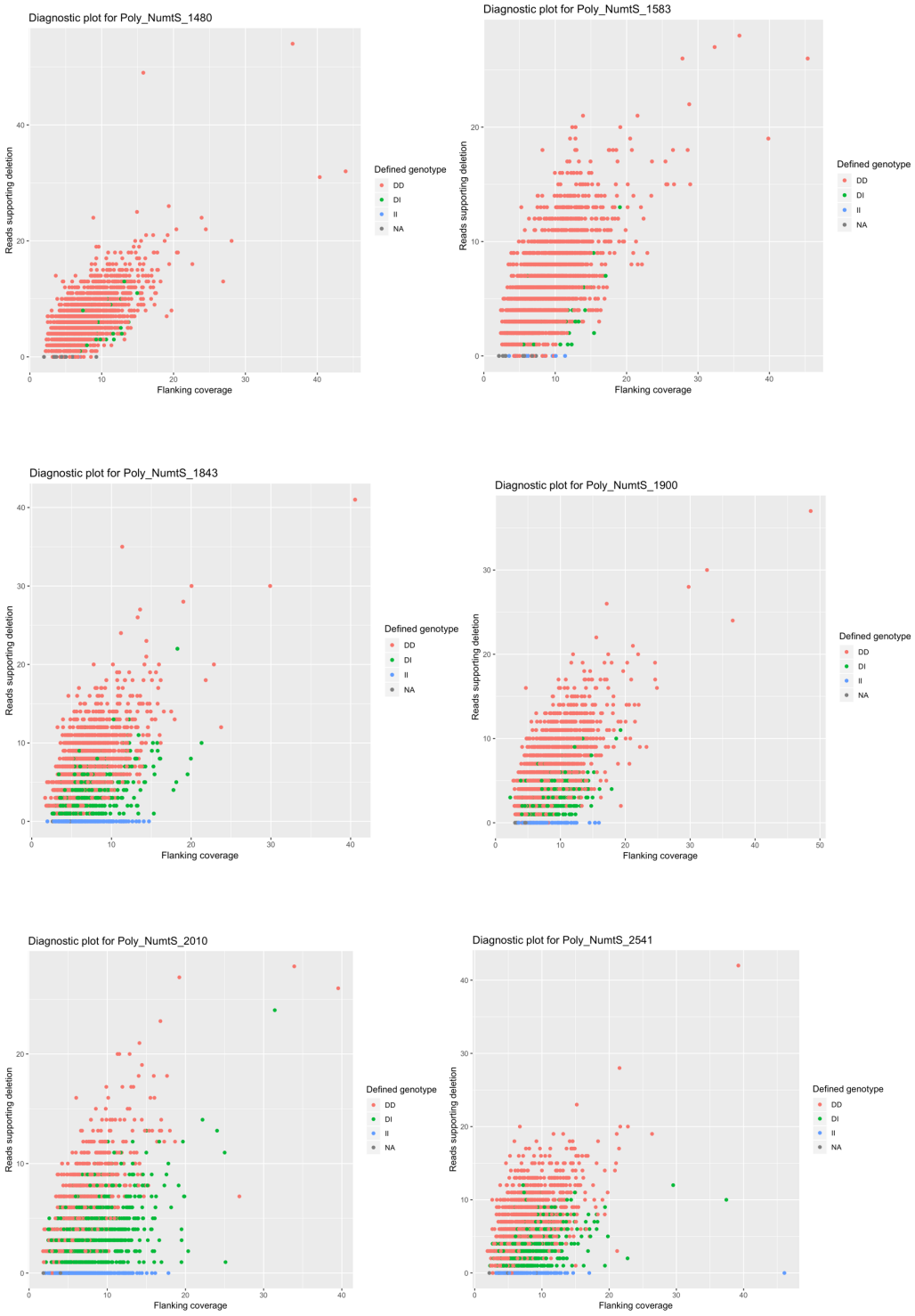

***Figures 11-15. Diagnostic plots*** *for HSA_NumtS_133_b1, HSA_NumtS_214_b1, HSA_NumtS_426_b1, HSA_NumtS_460_b1 and HSA_NumtS_543_b1. The NUMT coverage / Coverange flanking region proportions for each individual are plotted. Absence/absence haplotypes are plotted in red, heterozygotes are plotted in green and homozygotes for insertion are plotted in blue.*

On the other hand, those NUMTs not included in the reference genome where found to be widely present in different individuals were Poly_NumtS_1239, Poly_NumtS_1259, Poly_NumtS_139, Poly_NumtS_1465, Poly_NumtS_1480, Poly_NumtS_1583, Poly_NumtS_1843, Poly_NumtS_1900, Poly_NumtS_1929, Poly_NumtS_2010, Poly_NumtS_2289, Poly_NumtS_2377, Poly_NumtS_2541, Poly_NumtS_430 and Poly_NumtS_709. Poly_NumtS_2578 was not found in any individual and the other were found in minor frequencies. All frequencies are found in Supplementary Table 2 (in Excel format).

Additionally, the number of reads supporting an absence allele were plotted against the coverage within each region and individual. One may appreciate that deletion/deletion haplotypes have all 0 reads supporting an absence (given that the pipeline makes it a requisite to be included in that group). Deletion/insertion haplotypes tend to be in the right lower part of the plot whereas insertion/insertion points are usually plotted in the left upper part (except for Poly_NumtS_1239), thus suggesting that the genotype determination was also accurate in this case. Diagnostic plot for NUMTs present in a considerable number of individuals may be found in figures 16-25.




***Figures 16-19. Diagnostic plots*** *for Poly_NumtS_430, Poly_NumtS_709, Poly_NumtS_1239 and Poly_NumtS_1259. Red points represent homozygotes for deletion genotypes, heterozygotes are pointed in green and homozygotes for insertion in blue.*




***Figures 20-25. Diagnostic plots*** *for Poly_NumtS_1480, Poly_NumtS_1583, Poly_NumtS_1843, Poly_NumtS_1900, Poly_NumtS_2010 and Poly_NumtS_2541. Red points represent homozygotes for absence genotypes, heterozygotes are pointed in green and homozygotes for insertion in blue.*

## 5.2. Hardy-Weinberg Equilibrium for each NUMT and population

Hardy-Weinberg Equilibrium was also checked for each NUMT and population using the *HWExactMat()* (that performs an exact test to check whether the population is under the HW Equilibrium) function of the HardyWeinberg package for R. The p-values for each NUMT and population might be found in both Supplementary Tables 1 & 2. We may assume that a population is under the Hardy-Weinberg Equilibrium when the p-value is greater than 0.05, meaning that we cannot reject the null hypothesis stating that a specific population is under that equilibrium for a certain NUMT.

One may observe that most NUMTs can be assumed to be under the HWE in most populations (even if a considerable number of HWE violations may be considered in HSA_NumtS_522_b1 and HSA_NumtS_474_b1).

## 5.3. Principal Component Analysis

After computing the NUMT allele proportion in each population, a Principal Component Analysis was performed using i) all referenced NUMTs at a time, ii) all non-referenced NUMTs and iii) spread NUMTs in which HWE can be assumed.

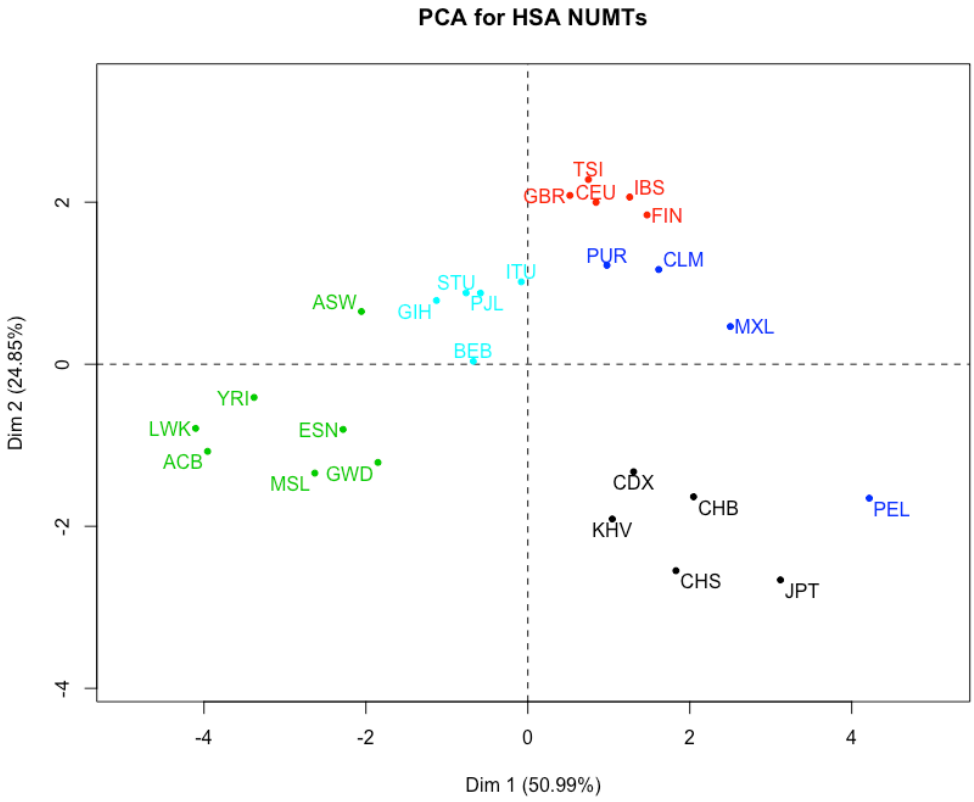



***Figure 26. Principal Component Analysis for all referenced NUMTs****. The two first components are plotted. 26 modern populations are plotted. The represented superpopulations are African (green), South Asian (black), East Asian (light blue), European (red) and American (blue). The specific code for each population is available [at http://www.internationalgenome.org/category/population/](http://www.internationalgenome.org/category/population/) .*

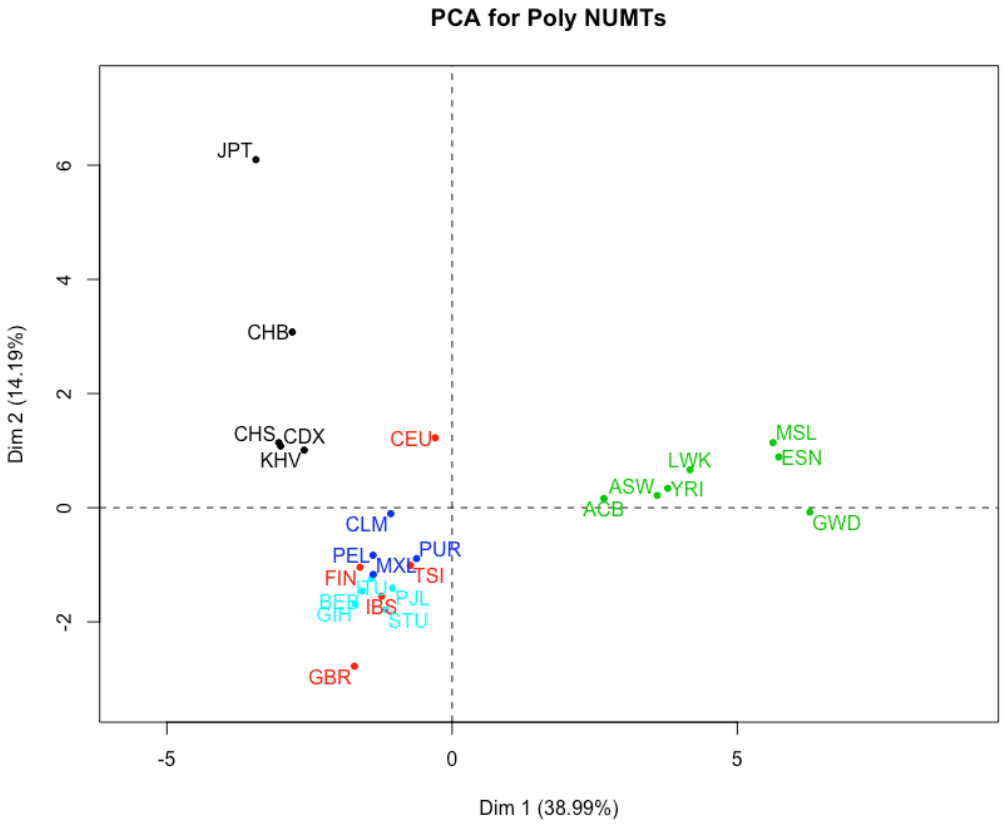



***Figure 27. Principal Component Analysis for all non-referenced NUMTs****. The two first components are plotted. 26 modern populations are plotted. The represented superpopulations are African (green), South Asian (black), East Asian (light blue), European (red) and American (blue). The specific code for each population is available [at http://www.internationalgenome.org/category/population/](http://www.internationalgenome.org/category/population/) .*

In both previous PCA analysis populations are grouped toward superpopulatons (European, South-Asian, East-Asian, African and American). However, in the second plot South-Asian, European and American populations do not appear to be clearly distinguished.

To perform the third PCA (Figure z), all referenced NUMTs except HSA_NumtS_522_b1 and HSA_NumtS_474_b1 were selected (those not under HWE in most populations). Among non-referenced NUMTs, Poly_NumtS_139, Poly_NumtS_430, Poly_NumtS_709, Poly_NumtS_1465, Poly_NumtS_1583, Poly_NumtS_1843, Poly_NumtS_1900, Poly_NumtS_1929, Poly_NumtS_2010, Poly_NumtS_2289, Poly_NumtS_2377 and Poly_NumtS_2541 were chosen (NUMTs having a low frequency and Poly_NumtS_1239 were discarded).

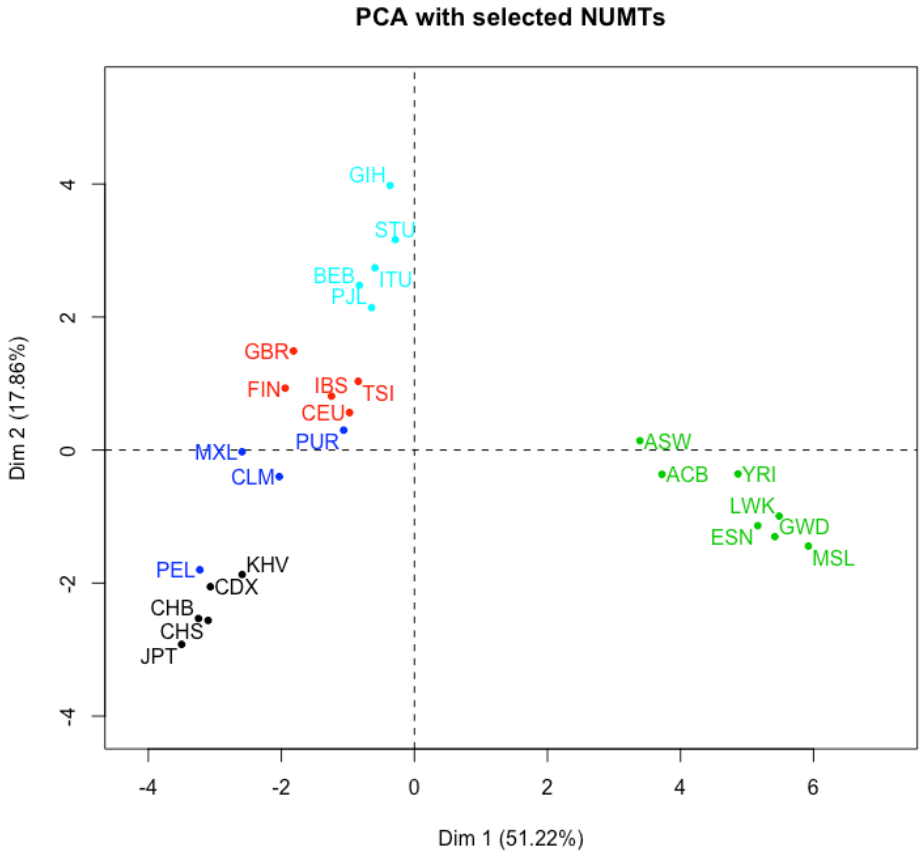



***Figure 28. Principal Component Analysis for selected NUMTs****. The two first components are plotted. 26 modern populations are plotted. The represented superpopulations are African (green), South Asian (black), East Asian (light blue), European (red) and American (blue). The specific code for each population is available [at http://www.internationalgenome.org/category/population/](http://www.internationalgenome.org/category/population/) .*

As previously observed, populations tend to be grouped in their respective superpopulations. The only exceptions (commented in the discussion) appear to be MXL, CLM and PUR.

## 5.4. NUMTs in ancestral individuals

After checking the output for the ancestral individuals, the table 2 was completed for each sample. In that table, one may find (for each NUMT an individual) whether there is (Y) or not (N) strong evidence of a insertion allele (In) and/or deletion allele (De). That is, an insertion allele will be considered if there is a read fully aligned against a NUMT reference sequence (NUMTs appearing in the genome reference sequence) or if there is a read whose mate is mapped to the mitochondrial chromosome or whose sequence partially aligns to the mtDNA. As a counterpart, a read fully mapping against a non-NUMT reference sequence (referenced NUMTs) or a read overlapping the breakpoint position in an unbroken manner (non-referenced NUMTs) will be interpreted as an evidence of a deletion allele. Therefore, if within a sample we find an insertion and a deletion evidence (Y/Y), that individual will probably be heterozygous (insertion/deletion) for that NUMT.

| NUMTS | DENIS | | ALTAI | | CHAG | | VINDIJ | | GOY | | COTT | | MEZ2 | | SPY94 | | VIN87 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | In | De | In | De | In | De | In | De | In | De | In | De | In | De | In | De | In | De |
| HSA_NumtS_009_b1 | N | Y | N | Y | N | Y | N | Y | N | N | N | N | Y | N | N | N | N | N |
| HSA_NumtS_015_b1 | N | Y | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | N |
| HSA_NumtS_030_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| Poly_NumtS_139 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | N |
| HSA_NumtS_038_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | N | N |
| Poly_NumtS_1239 | Y | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N |
| HSA_NumtS_050_b1 | Y | Y | N | Y | N | N | N | N | N | N | N | N | N | N | N | N | N | N |
| Poly_NumtS_1259 | Y | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N |
| HSA_NumtS_058_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | Y | N | N | N | Y | N |
| HSA_NumtS_092_b1 | Y | Y | Y | N | Y | N | Y | N | Y | N | N | N | Y | N | Y | N | Y | N |
| Poly_NumtS_1440 | Y | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | Y |
| HSA_NumtS_133_b1 | Y | Y | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | Y | N | N | N |
| HSA_NumtS_143_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | Y | N | Y | N | N | N |
| HSA_NumtS_171_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | Y | N |
| Poly_NumtS_1843 | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | Y | N | Y | N | N |
| HSA_NumtS_179_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N |
| HSA_NumtS_182_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N |
| HSA_NumtS_188_b1 | Y | N | Y | N | Y | N | Y | N | N | N | N | N | Y | N | Y | N | N | N |
| Poly_NumtS_1900 | Y | N | N | Y | N | Y | N | Y | N | N | N | N | Y | N | N | N | N | N |
| Poly_NumtS_1929 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y |
| HSA_NumtS_200_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_201_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N |
| Poly_NumtS_2010 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N |
| HSA_NumtS_214_b1 | Y | Y | N | Y | N | Y | N | Y | N | N | Y | N | N | N | N | N | N | N |
| HSA_NumtS_215_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_228_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_232_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N |
| Poly_NumtS_2186 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N |
| Poly_NumtS_2289 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N |
| Poly_NumtS_2377 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | N |
| HSA_NumtS_276_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_289_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N |
| HSA_NumtS_304_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | Y | N | N | N | N | N |
| Poly_NumtS_2541 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | Y | N | N |
| Poly_NumtS_2578 | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N |
| HSA_NumtS_319_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| Poly_NumtS_2611 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N |
| Poly_NumtS_2653 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | N |
| Poly_NumtS_316 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N |
| Poly_NumtS_430 | N | Y | N | Y | N | N | N | Y | N | N | N | N | N | Y | N | N | N | N |
| Poly_NumtS_445 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N |
| HSA_NumtS_398_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_410_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N |
| Poly_NumtS_531 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | N | N | N |
| HSA_NumtS_426_b1 | Y | Y | N | Y | N | Y | N | Y | N | N | N | N | N | N | N | N | N | N |
| HSA_NumtS_428_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_460_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_464_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| Poly_NumtS_709 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N |
| HSA_NumtS_472_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_474_b1 | Y | Y | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_512_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_513_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N |
| HSA_NumtS_518_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_519_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_522_b1 | Y | Y | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N |
| HSA_NumtS_543_b1 | Y | Y | Y | N | N | N | N | N | N | Y | N | Y | N | N | N | N | N | N |
| Poly_NumtS_1465 | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | N |
| HSA_NumtS_544_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| Poly_NumtS_1480 | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N | N | N | Y | N | N |
| HSA_NumtS_546_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| Poly_NumtS_1583 | N | Y | N | Y | N | Y | N | Y | N | N | N | Y | N | N | N | Y | N | N |
| HSA_NumtS_560_b1 | Y | N | N | Y | N | Y | N | N | N | N | N | N | N | N | N | N | N | N |
| HSA_NumtS_569_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_570_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_573_b1 | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | Y | N | N | N |
| HSA_NumtS_574_b1 | Y | N | N | N | N | N | N | N | N | N | N | N | Y | N | Y | N | N | N |
| HSA_NumtS_578_b1 | Y | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | N | N |
| HSA_NumtS_585_b1 | Y | N | N | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |

***Table 2. Results of NUMTs analysis for ancestral samples.***

# 6. Discussion

When observing the Principal Component Analysis results, one may conclude that NUMTs insertion/deletion polymorphisms appear to be good population markers. As one may observe, the different populations tend to be grouped according to its situation.

Regarding American populations situation, the Peruvian (PEL) proximity to East-Asian situations may be explained by the (hypothesized) colonization of the pre-colonial America coming from these populations. On the other side, European ancestor contribution to Mexican (MXL), Puerto Rican (PUR) and Colombian (CLM) described in Gravel et al. 2013 (24) may explain the proximity of the other three American populations to European individuals.

It is worth to mention that Poly_NumtS_1239 analysis might be carefully revised, since its diagnostic plot in which coverages are represented does not follow the desired pattern. That is, deletion/deletion and deletion/insertion haplotypes and reads supporting an absence appear to be unrelated when being plotted against flanking coverage. As commented in the results section, HSA_NumtS_015_b1 should be furtherly addressed (although it appears to be an European population marker, since it appears in all European populations and in CLM, MXL and PUR individuals).

There are several reasons that difficulted the ancestral samples analysis. Firstly, because of reads being aligned against a *Homo sapiens* reference sequence some information may be lost from the very beginning. In addition, only the Denisovan sample was mostly paired-end data, thus difficulting the approach for referenced NUMTs. Reads were neither soft-clipped in Neanderthal samples, thus losing a lot of information about the sequence that did not totally match with the reference genome. Finally, the low coverage in some samples also posed a problem to properly define the genotype for each individual.

However, the ancestral alleles table may serve, altogether with modern population haplotype frequencies, to hypothesize about cases of introgression between anatomically modern humans and those ancestral populations. Indeed, a further developed approach allowing to properly collect NUMT sequences from these samples could offer new information to refine those hypotheses.

# 7. Conclusions

## 7.1. Conclusions about the acquired competences

i. The open source data projects such as the 1000 Genomes Project enhance collaborative science. They also make small studies such as the presented here economically affordable.

ii. Even if R/Bioconductor is a powerful tool, it might not be appropriate when performing large-scale analysis of wide regions from BAM files.

iii. Using different programming languages in different steps of the same project allows to benefit from their different advantages.

iv. Even when performed at low coverage level, Next Generation Sequencing (NGS) technologies pose a powerful tool for large sample-size analysis

## 7.2. Specific conclusions from the project

i. NUMT insertions appear to be good population markers even when using just a few of them and at an insertion/absence level (without involving their sequence)

ii. Further improving in ancestral genomes alignments may help in analyzing Whole Genome Sequencing (WGS) data

iii. Sequence analysis of NUMT insertions may help to enhance their performance as population markers

## 7.3. Achieved objectives

Although with some inaccuracies that must be furtherly corrected, a modern human NUMT variation catalog was successfully performed. In addition, the potential of NUMTs as population markers was also proven.

However, even if a NUMT sequence analysis was thought to be performed at the very beginning of this project, it was discarded for time deadlines reasons. Nevertheless, an approach may be designed in the future.

Finally, a proper analysis and conclusions about NUMT frequencies and the deep hypothesis on how they may reflect the relation between populations through history (by carefully looking at the ancestral samples results) had to be left aside.

## 7.4. Working plan follow-up

Even if with some delay, the working plan was generally followed for the NUMT catalog elaboration and the evaluation of NUMTs as population markers.

The main change that had to be introduced was switching from the R/Bioconductor environment to the bash UNIX Shell language for reads sampling and its posterior analysis. As previously mentioned, R packages (*Rsamtools* (13)) were not robust and fast enough for sampling and managing such amount of information (one must recall that 2,535 samples over 70 NUMTs were analyzed in this project). Therefore, a new work pipeline had be designed and Bash language had to be in deep explored by the author. This handicap posed a serious delay on the project deadlines and conditioned its performance.

### 7.5. Future work

As previously mentioned, one of the future work lines to be explored is the NUMT sequence determination using the same samples. Since their coverage is considerably low, a population-based sequence analysis might be implemented to properly compare sequence variants between different populations.

Additionally, particular NUMT analysis still need to be properly reviewed (such as the Poly_NumtS_1239).

Finally, the obtained catalog opens an important line to deeply interpret analyzed NUMTs frequencies and distribution across modern and ancestral populations.

# 8. Glossary

| | |
|---|---|
| **BAI file** | Binary Alignment Index File |
| **BAM file** | Binary Alignment Map File |
| **HWE** | Hardy-Weinberg Equilibrium |
| **mtDNA** | Mitochondrial DNA |
| **NUMT** | Nuclear Mitochondrial DNA sequences |

# 9. References


1. Ricchetti M, Tekaia F, Dujon B. Continued colonization of the human genome by mitochondrial DNA. PLoS Biol. 2004;2(9).
2. Lang M, Sazzini M, Calabrese FM, Simone D, Boattini A, Romeo G, et al. Polymorphic NumtS trace human population relationships. Hum Genet. 2012;
3. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. The 1000 Genomes Project: Data Management and Community Access. 2012;9:459–62.
4. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 2010;6(2).
5. Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res. 2014;
6. Riaño Vivanco MA, Guardiola M, Dipierri JE, Ramos A, Santos C. Actas del XX Congreso de la Sociedad Española de Antropología Física: Human Specific Numts: Variation in Human Populations. In 2017. p. 311–25.
7. Bensasson D, Feldman MW, Petrov DA. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. J Mol Evol. 2003;57(3):343–54.
8. Lalueza-Fox C, Patterson N, Green RE, Malaspinas AS, Weihmann A, Verna C, et al. A Draft Sequence of the Neandertal Genome. Science (80-). 2010;328(5979):710–22.
9. Reich D, Shunkov M V., Stenzel U, Briggs AW, Good JM, Pääbo S, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468(7327):1053–60.
10. Mairal Q, Montiel R, Lima M, Aluja MP, Mateiu L, del Mar González M, et al. Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies. Mitochondrion. 2011;
11. Department of Evolutionary Genetics. Max Planck Institute. Genome Projects [Internet]. [cited 2019 May 15]. Available from: https://www.eva.mpg.de/genetics/genome-projects.html
12. Robinson JT, Thorvaldsdóttir, Helga Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
13. Morgan AM, Obenchain V. Package ‘ Rsamtools .’ 2019;
14. Ramirez-Gonzalez RH, Bonnal R, Caccamo M, MacLean D. Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments. Source Code Biol Med. 2012;7:1–6.
15. IGSR: The International Genome Sample Resource [Internet]. [cited 2019 Apr 1]. Available from: http://www.internationalgenome.org
16. The SAM/BAM Format Specification Working. SAM (Sequence Alignment / Map) Format Specification. 2019;(May):1–21. Available from: http://samtools.github.io/hts-specs/SAMv1.pdf
17. Prüfer K, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science (80- ). 2017;358(November):655–8.

18. Department of Evolutionary Genetics. Max Planck Institute. A high-quality Neandertal genome sequence [Internet]. [cited 2019 May 10]. Available from: https://www.eva.mpg.de/genetics/genome-projects/neandertal/index.html?Fsize=%27A%3D00
19. Meyer M, Kircher M, Gansauge M, Li H, Mallick S, Schraiber JG, et al. A High Coverage Genome Sequence From an Archaic. Science (80- ). 2013;338(6104):222–6.
20. Department of Evolutionary Genetics. Max Planck Institute. High coverage Chagyrskaya Neandertal [Internet]. Available from: https://www.eva.mpg.de/genetics/genome-projects/chagyrskaya-neandertal/home.html?Fsize=%27A%3D00
21. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;26(5):589–95.
22. Gómez-Rubio V. **ggplot2** - Elegant Graphics for Data Analysis (2nd Edition). J Stat Softw [Internet]. 2017;77(Book Review 2):3–5. Available from: http://www.jstatsoft.org/v77/b02/
23. Graffelman J. Exploring Diallelic Genetic Markers: The HardyWeinberg Package . J Stat Softw. 2015;64(3).
24. Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, et al. Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. PLoS Genet. 2013;9(12).