Final Master Project

ANALYSIS OF THE MITOCHONDRIAL DNA OF SCHIZOPHRENIA PATIENTS BY NEXT GENERATION SEQUENCING

Marta, Rodríguez Balada

Msc in Omics Data Analysis

Manager: Lourdes, Martorell Bonet

Guarantor: Malu, Calle Rosingana

Vic, January 2014

INDEX

1. INTRODUCTION

1.1. SCHIZOPRENIA

Schizophrenia is a severe mental disorder characterized by a breakdown of thought processes and by a deficit of typical emotional responses. Common symptoms are delusions including paranoia and auditory hallucinations, disorganized thinking reflected in speech, and a lack of emotional intelligence. It is accompanied by significant social or vocational dysfunction. The onset of symptoms typically occurs in adolescents and early adulthood, and around 4 years earlier in males than females. The median incidence (25% and 75% quantiles) is 15.2 (10.2-22.0) per 100,000 persons, while the median lifetime prevalence (25%, 75% quantiles) is 4.0 (3.0-6.6) per 1,000 persons (McGrath et al. 2008). Aetiology and pathophysiology remain unknown; however, it is widely accepted that a combination of genetic, environment and psychosocial factors lead to the manifestation of the illness (van Os and Kapur 2009). Heritability in schizophrenia is around 80%, so genetic risk factors are important in the development of the disorder. The greatest risk factor for developing schizophrenia is having a first degree with the disease, with an odds ratio of nearly ten (Sullivan et al, 2005). A large amount of genome-wide linkage studies and association studies have been conducted to elucidate the genetic variations involved in schizophrenia but still have not been identified susceptibility genes

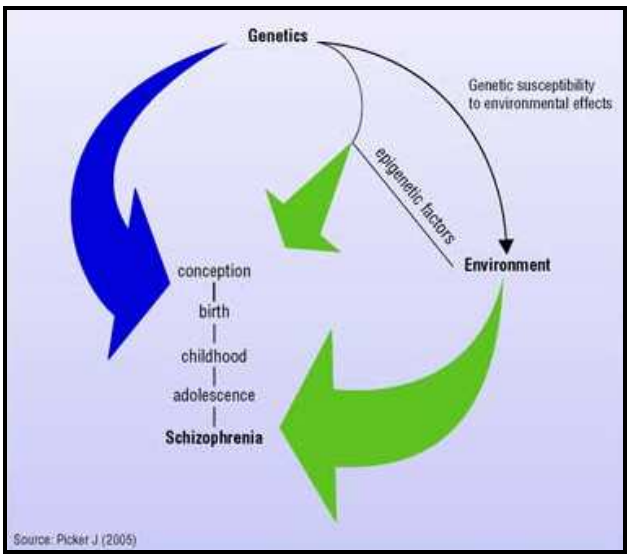

Figure 1. Schematic representation of the proposed etiopathogenic model for schizophrenia based on the interaction of genetic and environmental factors. Source: Picker J (2005)

1.2. MITOCHONDRIA

Mitochondria are subcellular organelles enriched in energetic tissues, such as muscle and brain, and located in the cytoplasm. Although there are over 1500 human mitochondrial genes (Wallace, 2005), only a small fraction of these are directly encoded by the mitochondrial DNA (mtDNA). The mitochondrial genome has several different characteristics from the nuclear genome. Human mtDNA, which is inherited in a matrilineal pattern (Giles et al., 1980), is a double stranded circular molecule of approximately 16.569 nucleotides (Wallace, 2005) containing 37 genes that encode two ribosomal RNAs, 22 transfer RNAs, and 13 polypeptides. Interestingly, it codifies for 13 of the 84 subunits of the mitochondrial respiratory chain that are crucial for its function. Also, it has a non-codifying region, the D-loop, involved in mtDNA replication and transcription regulation. One of the most important functions of mitochondria is to generate energy and for this reason it is important in tissues with high-energy requirements such as brain and muscle. The number of mitochondria in a somatic cell varies between 200 and 2000. Spermatozoa contain 16 mitochondria, whereas oocytes can hold up to 100,000.

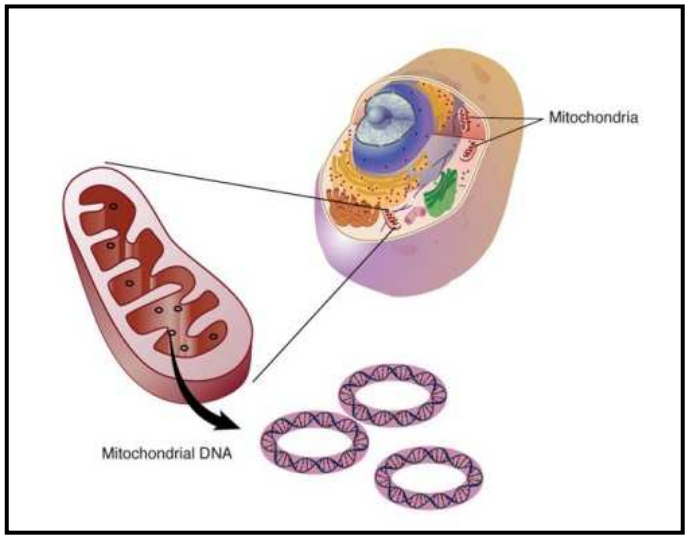

Figure 2. Mitochondrial DNA

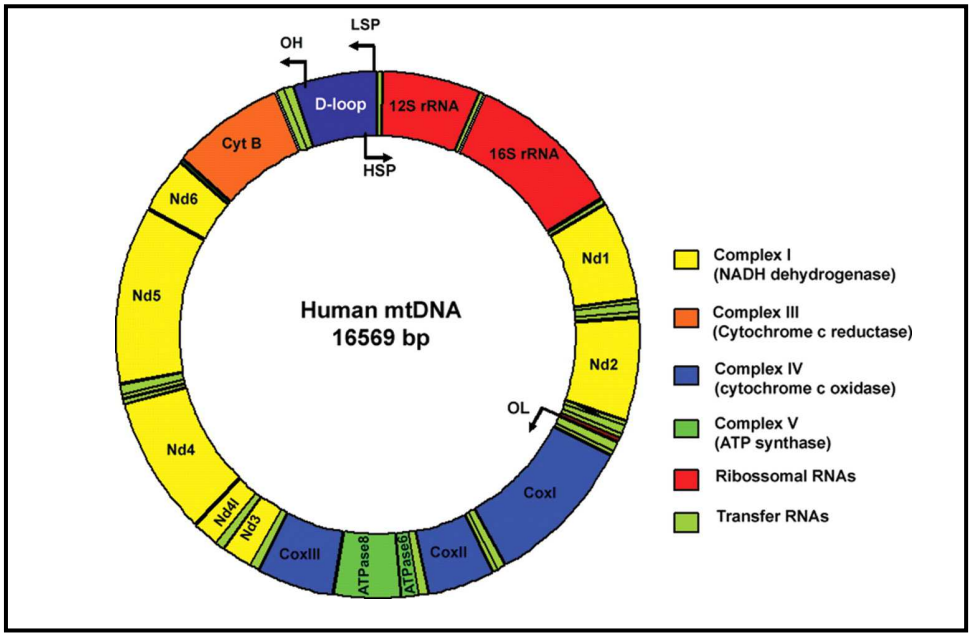

Figure 3. The human mitochondrial genome.
mtDNA is a 16,569 double-strand DNA molecule. The different regions of the mtDNA are shown in distinct colours. The 13 genes that encode for the essential components of the respiratory chain are shown in yellow (subunits of complex I), red (subunit of complex III), blue (subunits of complex IV) and green (subunits of complex V) .

Mitochondrial genetics involves specific characteristics that differ from nuclear genetics:

- **Uniparental inheritance**. MtDNA is only inherited from the mother (non-mendelian inheritance)
- Human mtDNA genetic code differs from the universal code

Table 1. Differences between the mtDNA and the nDNA

codon	mtDNA	nDNA
AUA	Initiation code (Met)	Arginine
AUG	Initiation code (Met)	
AGA, AGG	Termination codes	Arginine
UGA	Tryptophan	Termination
AUU	Isoleucine during elongation Methionine during initiation	

- **High substitution rate**. mtDNA is particularly susceptible to a higher rate of somatic deletions because mtDNA is not protected by histones and lacks of a complete set of DNA repair machinery as it is associated with nuclear DNA. Mutation rates are higher in mtDNA compared to nuclear DNA, and mutation rates are particularly high in non-coding regions of the mtDNA in human brain such as the hypervariable region of the mtDNA genome. The susceptibility of mtDNA to somatic mutations can lead to a gradual accumulation of germinal mutations along the transmission of the mtDNA through generations, giving rise to the differences among populations and among individuals. Also, it has been hypothesized that mtDNA may be also involved in complex traits including neurodegenerative disorders, ageing and cancer (Schon et al. 2012).

- **Poliploidy.** Each human cell contains different number of mitochondria, depending on its energy requirement, and also a lot of copies of mtDNA. All these copies of mtDNA are identical in healthy subjects but in some conditions, patients can present mutated molecules and wild type molecules. Homoplasmy refers to the condition in which all mtDNA molecules are identical and heteroplasmy refers to the existence of subpopulations of mtDNA genomes within an individual.

- **Highly polymorphic**. Specific patterns of polymorphisms are the basis of the mtDNA haplogroups. A mtDNA haplogroup is a certain group of polymorphisms that reflect the maternal ancestry of a particular individual that is characteristic and it is specific for different ethnic groups (although there are differences between individuals that are included in the same ethnic group). Each haplogroup contains individual haplotypes that are specific mitochondrial genotypes defined by a characteristic collection of mtDNA polymorphisms. When a sample is analysed, the first thing to do is to determine the haplogroup that it belongs to looking at the mitochondrial genotype (the mtDNA variants that define the haplotype) and the rest of the variants that are not defining the haplotype of the samples are those candidates to further studies. MtDNA variants have been reported to occur in almost every nucleotide position of the 16.569 bp mitochondrial genome. Although most reported common recurrent pathogenic mutations are heteroplasmic, the rare variants associated with common complex diseases such as diabetes or hypertension are often homoplasmic.

1.3. MITOCHONDRIAL DISORDERS

When we talk about mitochondrial disorders, we refer to disorders with mitochondrial dysfunction that can result from the nuclear DNA or mtDNA mutations. Mitochondrial disorders, resulting from point mutations or rearrangements of the mtDNA have been associated with a variety of clinical presentations. Point mutations are typically maternally inherited.

Mutations in the mtDNA cause mitochondrial disorders that can present either in infancy or adulthood. More than 150 genetic mitochondrial syndromes have been described affecting 1 in 5,000 live births (Skladal et al. 2003). Signature traits are lactic acidosis, skeletal myopathy, deafness, blindness, subacute neurodegeneration, intestinal dismotility and peripheral neuropathy. Though most organ systems can be affected it is also possible a highly tissue-specific dysfunction (Vafai and Mootha 2012)

1.4. THE mtDNA AND SCHIZOPHRENIA

The hypothesis of mtDNA involvement in the schizophrenia's genetic susceptibility is supported from several perspectives. Briefly, mitochondrial dysfunction, maternal inheritance and also comorbidity of schizophrenia and mitochondrial disorders related to mtDNA mutations have been reported (Verge et al. 2011; Anglin et al. 2012). Moreover, the mitochondrial dysfunction in schizophrenia has been hypothesised as an explanation for the heterogeneity of clinical and pathological manifestations and the mtDNA has been indicated as an underlying etiologic genetic factor. However, the mitochondrial dysfunction that has been proposed to be involved in schizophrenia may be related to an abnormal function of nuclear, mitochondrial or both types of genomes.

One possible indication of a mitochondrial disorder is a matrilineal inheritance (Wallace, 1994). For neuropsychiatric disorders such as schizophrenia, higher rates of disease are observed in offspring of maternal probands compared to offspring of paternal probands (Verge et al, 2011).

Mitochondria provide most of the energy for brain cells by the process of oxidative phosphorylation. Mitochondrial abnormalities and deficiencies in oxidative

phosphorylation have been reported in individuals with schizophrenia, bipolar disorder and major depressive disorder in transcriptomics, proteomics and metabolomics studies (Zaragoza MV et al, 2010. Cacabelos R, 2011).

Human mtDNA variants have been studied related to the human health and the illness. It is well known that specific mutations and some variants are involved in human disorders. It has also been described that some of these variants and mutations are related with neuropsychiatric disorders. And finally, severe mental conditions such as psychoses or depression have been described in adult patients of mitochondrial illness caused by a mtDNA mutation. The brain disorders that appear from mtDNA changes are divided in two groups: sporadic diseases mainly due to rearrangements in the mtDNA and maternally inherited diseases due to point mutations. One of the previous steps in the study of mtDNA in mental disorders was to study the mtDNA association in a small number of subjects, focused on rare mutations or haplogroup-defining SNPs and have no found compelling evidence for association (a case-control study). (Hudson G, 2013) There are few studies of mtDNA in schizophrenic patients and most of them do not support the association of particular mtDNA variant with an increased risk of schizophrenia (reviewed in Mosquera-Miguel et al, 2013).

Sequence variants in mtDNA can be present in all the mtDNA molecules of a participant (homoplasmy) or may occur only in a subpopulation of mtDNA copies (heteroplasmy). The percentage of mutation heteroplasmy can vary among different tissues, contributing to the high variability in clinical features and disease severity. The proportion of normal and mutant molecules in a specific tissue (heteroplasmy level) largely determines the clinical expression of a pathogenic mtDNA. Determining the heteroplasmy level is important because it is related to the degree of mitochondrial dysfunction and disease severity.

There is no single variant that should be considered an a priori candidate to be involved in schizophrenia. Thus, the study of the entire mtDNA is more appropriate. The two methods most widely used to identify mtDNA variants are direct Sanger sequencing and the Affymetrix GeneChip Human Mitochondrial Resequencing Array; however, these two methods are neither sensitive nor specific enough to detect heteroplasmy.

1.5. NEXT GENERATION SEQUENCING BY ION TORRENT (PGM)

In this project, the mitochondrial DNA resequencing was done by Next Generation Sequencing (NGS) using the ION TORRENT sequencer (PGM, Life Technologies). mtDNA from patients was first amplified in two overlapping fragments and purified prior to start with the sequencing protocol. DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. NGS allows large numbers of DNA molecules to be sequenced per run, and therefore, at low reagent costs.

In the benchtop group of sequencers, The Ion Torrent Personal Genome Machine (PGM) has been recently introduced. This sequencer is the first to use semiconductors and a non-optical based sequencing technology [Rothberg et al., 2011]. It is a *NGS* equipment that allows to sequence large amounts of nucleic acids chains. The detection system is based on changes of ionic potential. This PGM has an informatics server with software (Torrent Suite) where the data is filtered and also mapped with the reference genome that is required to be indicated and there is also a primary quality control. These semi-raw data have to be analysed using different tools as it is explained in the methodology section.

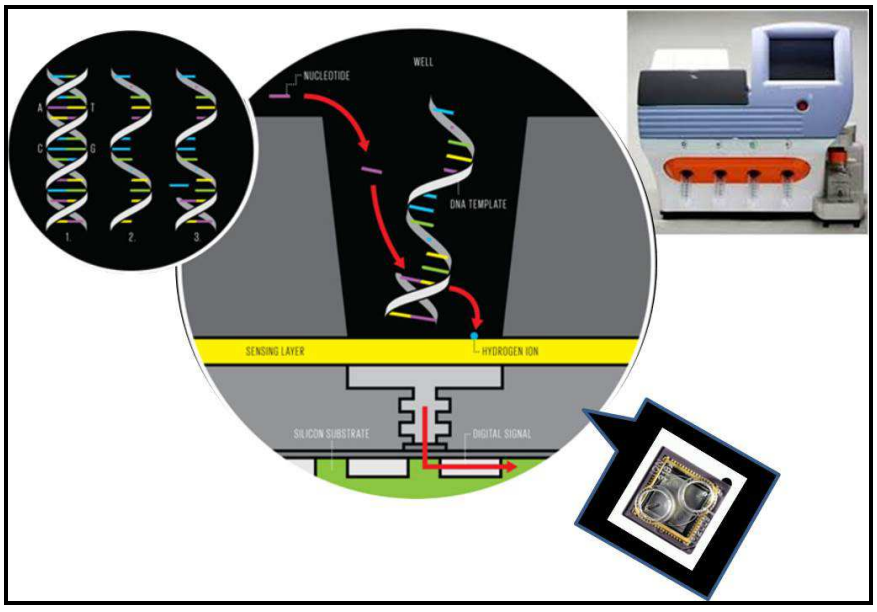

Figure 4. ION TORRENT PGM (Life Technologies) and a brief view of its chemistry.

2. OBJECTIVES

The aim of this final master project is to identify mtDNA variants or mutations in schizophrenia patients with an apparent matrilineal heredity of this disorder. This present study is involved in a project where the main goal is identify mtDNA variants or mutations in schizophrenia patients with an apparent matrilineal heredity of this disorder and to investigate whether the mtDNA mutations identified could contribute to the mitochondrial dysfunction described in schizophrenia.

This final master project has performed to accomplish three main objectives:

- To establish and standardize the bioinformatics analysis using free software of the NGS data obtained with the equipment Ion Torrent PGM.

- To identify new variants and apply bioinformatics prediction tools to assess the possible implications by the prediction of pathogenicity of non-synonymous variants and to determine if they can be related with schizophrenia.

- To identify whether some of the identified variants were more frequent in cases than in control subjects, and therefore, susceptible to account for schizophrenia risk.

3. SUBJECTS AND METHODS

3.1. SUBJECTS

This study was carried out with the approval of the Sant Joan University Hospital Ethics Committee from Reus and written informed consent was obtained from schizophrenia patients and control individuals after full explanation of the procedures. Blood samples were obtained from peripheral blood and the DNA isolated from leucocytes.
Patients and controls were from the same geographical region and had similar ethnic background. The control group was comprised by 38 individuals and they had no life history of psychiatric symptoms or first-degree relatives with psychiatric disorders. The group of patients was comprised by 55 schizophrenia patients with apparent matrilineal inheritance of the disease to identify variants and/or mutations.

3.2. mtDNA SEQUENCING BY NGS

This resequencing step was done by NGS with the Ion Torrent PGM (Life Technologies). This ionic sequencing technique allows sequencing high amount of nucleic acids in a short time. The detection system is based in ionic potential changes that happen when a nucleotide binds to the elongation sequence.

The mtDNA of each participant was amplified in two fragments of 8338 and 8647 bp using previously described primers (Gunnarsdottir et al., 2011). Long-range PCR was performed using 10 ng of DNA with Expand Long Range dNTPack (Roche, Barcelona, Spain), and the PCR products were purified using the QIAquick PCR Purification kit (Qiagen) following the manufacturer's instructions. Lastly, the two PCR fragments from each individual were mixed in equimolar ratios for the sequencing protocol using the Ion Torrent Personal Genome Machine (PGM) according to the manufacturer’s user guide. In short, 200-bp libraries were generated using the AB Library Builder System and the Ion Xpress Plus Fragment Library kit (Life Technologies). The samples were barcoded using the Ion Xpress Barcode Adapter 1-16 (Life Technologies), and the library concentration and amplicon size were assessed with the Agilent Bioanalyzer High Sensitivity DNA kit (Agilent, Madrid, Spain) using the Agilent Bioanalyzer. Equimolar pools of barcoded libraries were mixed to run the samples on the Ion 314 chips. The template of the library pool was prepared using the Ion OneTouch 200

Template kit v2 DL (Life Technologies), and sequencing was performed with the Ion PGM 200 Sequencing Kit (Life Technologies).
The PGM has an informatics server (Torrent server) were the sequencing data are dumped and by the Torrent Suite Software, a first filtering process and a quality step of the data are done. This data have been analysed by different bioinformatics tools established in a pipeline.

In the new NGS platforms, as Ion Torrent PGM, low percentages of heteroplasmy can be detected; however false positives variants are also identified, demonstrating low specificity (Tang S. et al, 2010). As we previously mentioned, determining the heteroplasmy level is important because it is related to the degree of mitochondrial dysfunction and disease severity. Therefore, it is important to indicate whether a change is homoplasmic or heteroplasmic and also it is of interest to know the technology capacity to detect it. Also, it is necessary to distinguish a true low level of heteroplasmy from an apparent heteroplasmy due to a technical artefact from poor DNA sequence quality. For this reason is needed to check the results obtained by Next Generation Sequencing by Sanger Sequencing. The same heteroplasmic change should be identified in both sequencing directions by Sanger sequencing and there should be no sequence background in the adjacent regions. If there are multiple heteroplasmic changes detected in one individual, further investigations are needed to verify the heteroplasmic finding.

3.3. SEQUENCING DATA ANALYSIS AND VARIANT IDENTIFICATION

The resulting sequences were aligned and compared to the revised Cambridge Reference Sequence, rCRS, NC_012920.1 (Andrews et al, 1999).
To perform a variant identification we established a basic pipeline using different bioinformatics tools.

a) Checking the quality of the reads. View the fastq files.
b) Sort the bam file.
c) Remove possible PCR duplicates.

d) Index the bam file. It generates a bai file that is needed for some display and snp calling programs.
e) Get some statistics using SAMtools. . (Li et al., 2009)
f) Perform variant calling with SAMtools.
g) Variant filtering. We need to filter variants from the raw vcf file. To do that we use samtools.
h) Visualize the SNPS on IGV (loading the bam file, sorted and without PCR duplicates).

```
#Sort the bam file

samtools sort P066_IonXpress_005_R_2013_07_30_05_00_15_user_PGM-
23-Xip_6_Helena_Auto_user_PGM-23-Xip_6_Helena_24.bam
p066_sorted.bam

# Remove possible pcr duplicates
samtools rmdup -s p066_sorted.bam.bam p066_sorted.noDup.bam.bam

# Index the bam file
samtools index p066_sorted.noDup.bam.bam

# GEt some stats using samtools
samtools flagstat p066_sorted.noDup.bam.bam

############################
### Perform variant calling with samtools

samtools mpileup -uf sequence.fasta p066_sorted.noDup.bam.bam |
bcftools view -vcg - >
p066_sorted.noDup.bam.bam.samtools.var.raw.vcf

.../samtools-0.1.19/bcftools/vcfutils.pl varFilter -Q 10 -d 15 -a
5 p066_sorted.noDup.bam.bam.samtools.var.raw.vcf >
p066_sorted.noDup.bam.bam.samtools.var.filtered.vcf


#less -s p066_sorted.noDup.bam.bam.samtools.var.filtered.vcf

# to obtain a txt file with the variant list
tail -n +2 p066_sorted.noDup.bam.bam.samtools.var.filtered.vcf |
cut -f2-10 > p066_sorted.noDup.bam.bam.samtools.var.filtered.txt

#ls -la *vcf


############################
# View aligned files using IGV
cd IGV_2.3.3/
./igv.sh
```

Figure 5. Pipeline to analyze the Ion Torrent data

3.4. BIOINFORMATIC TOOLS

To perform the bioinformatics analysis of the data obtained after sequencing all the samples, we need several bioinformatics tools to check the quality of the data, to filter and process it and also it is interesting to use a viewer to explore the data processed.

3.4.1. FASTQC

FastQC software allows making a post run quality check and making a read quality trimming. It aims to provide a tool to do some quality control checks on raw sequence data coming from the Next Generation Sequencer. It consists on a modular set of analyses which we can use to have a quick report to know if our data has any problems during the sequencing process of which we should be aware before doing any further analysis.

The main functions of FastQC are

- Import of data from BAM, SAM or FastQ files
- Providing a quick overview of the sequencing process
- Summary graphs and tables to quickly assess our data

To use this tool, FASTQ files are needed. FASTQ files extend FASTA files in that they provide both sequence and quality data. A FASTQ file thus typically consists of four lines of information.

1. The sequence identifier: a line starting with @ containing the sequence identifier
2. The actual sequence
3. A line starting with + after which the sequence identifier is optional
4. Quality data: a line with quality values which are encoded in ASCII space

Figure 5. Example of a FASTQ file

There are several quality parameters that we have to consider to explore the quality of our sequences:

3.4.1.1. Sequence Quality

The quality score introduced in the FASTQ files. It gives an idea about base call quality. Quality of reads often degrades over the course of a sequence. Looking to the bar plots that show the quantiles, we can have an idea about the spread of the quality.

3.4.1.2. Per Base Sequence Content

Base content at each position is an important quality indicator. When we want to know the per base sequence content, we assume that the contribution should be identical at each position. What we expect is to see straight lines; however the first few bases might indeed show some variability in base content. This is likely, due to non-completely random primers, but then we expect stability.

3.4.1.3. Introduction of errors and quality scores/encoding

As mentioned above the base caller assigns a quality score for each base. This score gives the estimated reliability for each individual. Ion Torrent has major issues with

homopolymers such as AAAAAA where the correct number of As can often not be determined exactly.

3.4.1.4. Preprocessing Steps

In order to avoid quality data problems, a common strategy is to remove low quality bases by **sequence quality trimming**. Typically one would remove lower quality bases from the 3' end using a sliding windows approach as per base quality gradually drops.

In addition to removing lower base quality data, one would also remove adapters, PCR primers and other artefacts. It can be done by Alternative clipping strategies (Adaptor clipping).In practice one would combine the adapter clipping with quality trimming approaches.

A typical quality control workflow consist of having a look to the quality plots from the raw data obtained by the sequencer (the raw data) and to evaluate them to know whether there have been problems during the library preparation process. Thereafter, make a trimming to remove errors. There are several tools that can be used, e.g. Trimmomatic, to remove bases presenting low quality values from the sequence ends and to remove the adapters from the reads. Finally, it is important to have a look to the post-trimming quality plots to judge the final quality of the sequences.

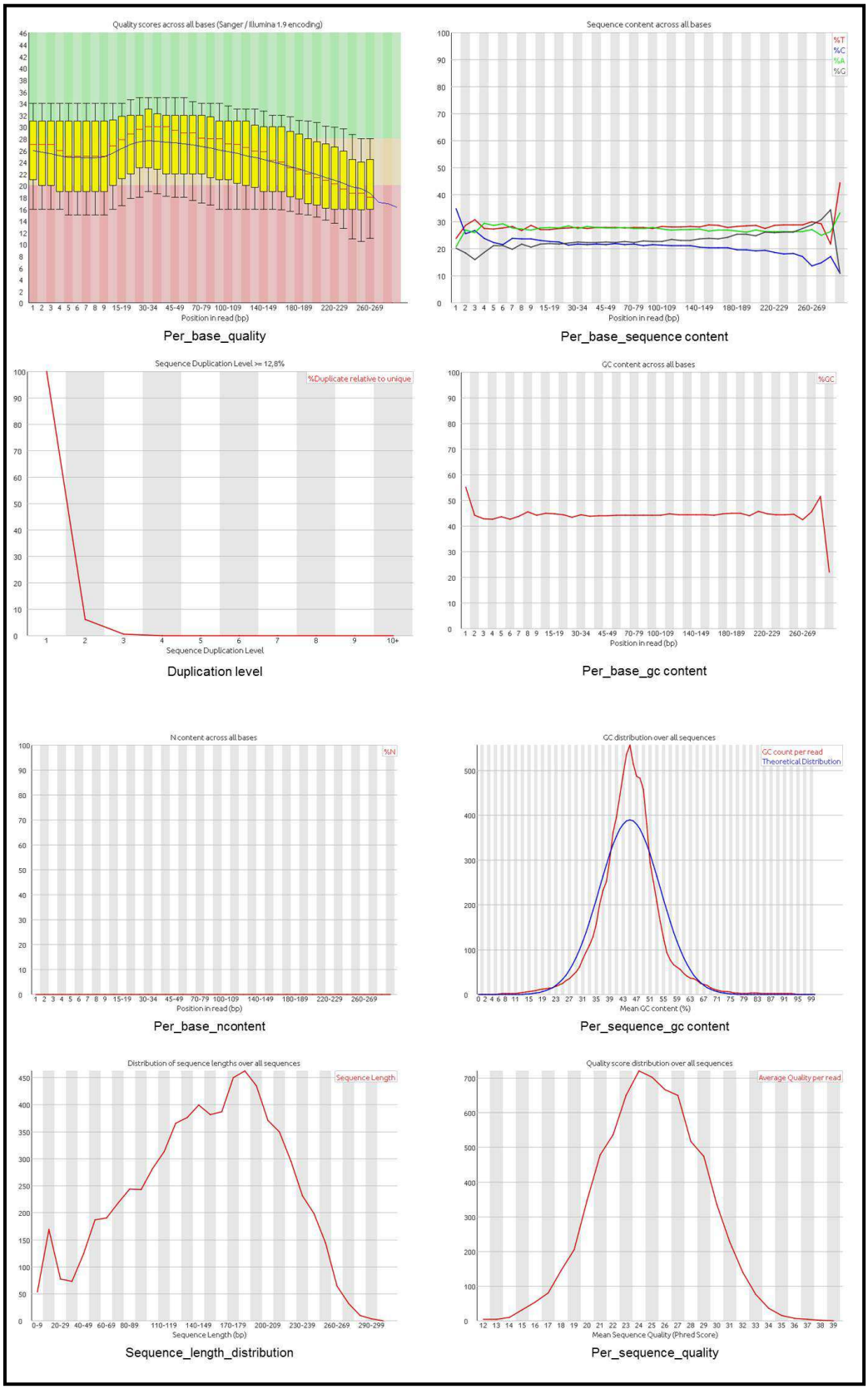

Figure 6. Report of an Ion Torrent chip generated by FASTQC

3.4.2. SAMTOOLS

SAM files consist of two types of lines: **headers and alignments**.

- Headers begin with @, and provide meta-data regarding the entire alignment file.
- Alignments begin with any character except @, and describe a single alignment of a sequence read against the reference genome.

Each read in a FASTQ file may align to multiple regions within a reference genome, and an individual read can therefore result in multiple alignments. In the SAM format, each of these alignments is reported on a separate line. Each alignment has 11 fields, followed by a variable number of optional fields. Each of the fields is described in the table 2.

Table 2. Mandatory fields and optional fields of a SAM file.

FIELD NAME	DESCRIPTION
QNAME	Unique identifier of the read; derived from the original FASTQ file
FLAG	A single integer value which encodes multiple elements of meta-data regarding a read and its alignment. Elements include: whether the read is one part of a paired-end read, whether the read aligns to the genome, and whether the read aligns to the forward or reverse strand of the genome
RNAME	Reference genome identifier.
POS	Left-most position within the reference genome where the alignment occurs.
MAPQ	Quality of the genome mapping. The MAPQ field uses a **Phred-scaled probability** value to indicate the probability that the mapping to the genome is incorrect. Higher values indicate increased confidence in the mapping
CIGAR	A compact string that (partially) summarizes the alignment of the raw sequence read to the reference genome. Three core abbreviations are used: M for alignment match; I for insertion; and D for Deletion
RNEXT	Reference genome identifier where the mate pair aligns. Only applicable when processing paired-end sequencing data.
PNEXT	Position with the reference genome, where the second mate pair aligns. As with RNEXT, this field is only applicable when processing paired-end sequencing data. A value of 0 indicates that information is not available.
TLEN	Template Length. Only applicable for paired-end sequencing data, TLEN is the size of the original DNA or RNA fragment, determined by examining both of the **paired-mates**, and counting bases from the left-most aligned base to the right-most aligned base. A value of 0 indicates that TLEN information is not available.
SEQ	The raw sequence, as originally defined in the FASTQ file.
QUAL	The **Phred quality score** for each base, as originally defined in the FASTQ file.

3.4.2.1. Converting SAM to BAM

In a NGS pipeline, the main objective is to identify the set of genomic variants that are present in the studied samples. The first step needed is the conversion of the SAM file into a BAM file, because all the following steps require a BAM file as the input file.

3.4.2.2. Removing possible duplicates

After mapping our sequences with the reference sequence, we should to eliminate those reads that are duplicates, to avoid problems in the SNP calling for over-representation of the reads. SAMtools allows removing those reads that can be PCR duplicates.

3.4.2.3. Sorting and Indexing

The next step is to sort and index the BAM file. Once we have sorted our BAM file, we can then index it creating a BAM index file which has the .bai extension.

3.4.2.4. Filtering

The called variants can be directly piped into the varFilter script that comes with SAMtools and filters for several parameters Examples of some of the parameters that can be applied are:

-Q INT	**minimum RMS mapping quality for SNPs**
-d INT	**minimum read depth**
-D INT	maximum read depth
-a INT	**minimum number of alternate bases**
-w INT	SNP within INT bp around a gap to be filtered
-W INT	window size for filtering adjacent gaps
-1 FLOAT	min P-value for strand bias (given PV4)
-2 FLOAT	min P-value for baseQ bias
-3 FLOAT	min P-value for mapQ bias
-4 FLOAT	min P-value for end distance bias
-e FLOAT	min P-value for HWE (plus F<0)
-p	print filtered variants

The filtered variants were not discarded immediately, however only variant that are marked as filtered in the VCF file will be taken into account for further analysis.

3.4.2.5. Identifying Genomic Variants

After all these steps, we can identify the genomic variants from our reads. It has to be done in two steps.

The first step uses the **SAMtools mpileup** command to calculate the genotype likelihoods supported by the aligned reads in our sample. The mpileup command automatically scans every position supported by an aligned read, computes all the possible genotypes supported by these reads, and then computes the probability that each of these genotypes is truly present in the studied sample. For each position assayed, SAMtools computes all the possible genotypes, and then outputs all the results in a binary call format (BCF)**.**

The second step is uses the SAMtools **bcftools** command, which is also packaged with SAMtools, but located in the bcftools directory. The bcftools view command uses the genotype likelihoods generated from the previous step to call SNPs and indels. The outputs display all identified variants in the **variant call format (VFC)**, the file format created for the 1000 Genomes Project and now widely used to represent genomic variants.

3.4.2.6. The VCF format

VCF files contain header lines, which begin with # or ## symbols, and data lines, which do not begin with a # symbol. The header contains important information, such as the name of the program that generated the file, the VCF format version number, the reference genome name, and information regarding individual columns within the file.

Each data line is required to have eight mandatory columns, including **CHROM** reference, **POS**ition within the chromosome, **REF**erence sequence, and **ALT** sequence. Directly after these eight columns there is a **FORMAT** column used to describe the format for all subsequent sample columns.

Within our VCF file, the FORMAT column is set to "**PL**". PL is defined as a list of Phred-scaled genotype likelihoods. DP indicates the position **read depth** and if in the FORMAT column there is PL, it indicates that all samples will include a list of genotypes likelihoods.

As an example, in the case that the alignments column presents the cryptic string "64,178,0,66,147,60", the translation to the likelihood p-value automatically calculated with SAMtools are the following:

Genotype	Phred Scaled Score	Likelihood p-value
TT	64	3.981-07
TG	178	2.512e-18
GG	0	1.0
TC	66	2.512e-07
GC	147	1.995e-15
CC	60	1.000e-06

This indicates that the GG genotype is most probably the true genotype in the sample.

3.4.3. IGV

The Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdottir et al., 2013) is a high-performance desktop tool for interactive visual exploration of diverse genomic data. Even for very large data sets, IGV supports real-time interaction at all scales of genome resolution, from whole genome to base pairs. IGV has a user-friendly and intuitive interface.

3.5. VARIANT IDENTIFICATION

To determine whether the sequence variants identified in the mitochondrial genomes of the schizophrenia patients had been previously described, predictive mtDNA haplogroups were determined using the Haplogrep software (http://haplogrep.uibk.ac.at/). Because of their unique population history, specific mtDNA haplogroups are identified in different geographic regions. Variant frequencies are different in different haplogroups. When identifying one or more rare variants from one individual, we have to check which haplogroup the individual belongs to. Then, we have to check the variants or mutations that are not linked to the haplogroup. There are several databases presenting mtDNA variants: the Human Mitochondrial Genome Database (www.mitomap.org), the Human Mitochondrial Genome Polymorphism Database (www.giib.or.jp/mtsnp/index_e.shtml) and the alignment of 928 European sequences (http://www.broad.mit.edu/mpg/tagger/mito.php) are widely used. It is important to notice that MITOMAP, the December 11, 2013, included information from 18363 GenBank sequences. When the genomic change was located in an encoding region, we used the MitoAnalyzer software (http://www.cstl.nist.gov/biotech/strbase/mitoanalyzer.html) and the mitowheel (http://mitowheel.org/mitowheel.html) to determine whether the variant triggered an amino acid change in the polypeptide sequence.

HaploGrep is a web-based java application for determining the mtDNA haplogroups of a specific sequence. HaploGrep imports .hsd files – these are text files, having a sample id, range, haplogroup if known (otherwise empty space) and the polymorphisms separated by tabulators. The ranges itself are separated by ";". The file contains the header information in the first row.

To import our file, we have to click on the Open button. Then, the search process starts by clicking on "Find haplogroups". After this, a results window is shown. A Haplogroup column is presented, that suggests the haplogroup status. The results can have three different colours: red, yellow and green, and the colour represent the quality of the haplogroup assignment.

- Red indicates that the haplogroup assignment is not very reliable.
- Yellow indicates that the haplogroup assignment is more accurate but still there's either sequence information lacking or many polymorphisms of the sample were not used for haplogroup assignment.
- Green indicates that the haplogroup assignment is quite reliable, as a large proportion of the sample polymorphisms are explained by its haplogroup affiliation.

By clicking on an entry on the results table, the lower part of the web-application will be filled with data, and mainly shows two areas: the information of the haplogroup in a tabular view and the corresponding graph showing the pathway from the rCRS.

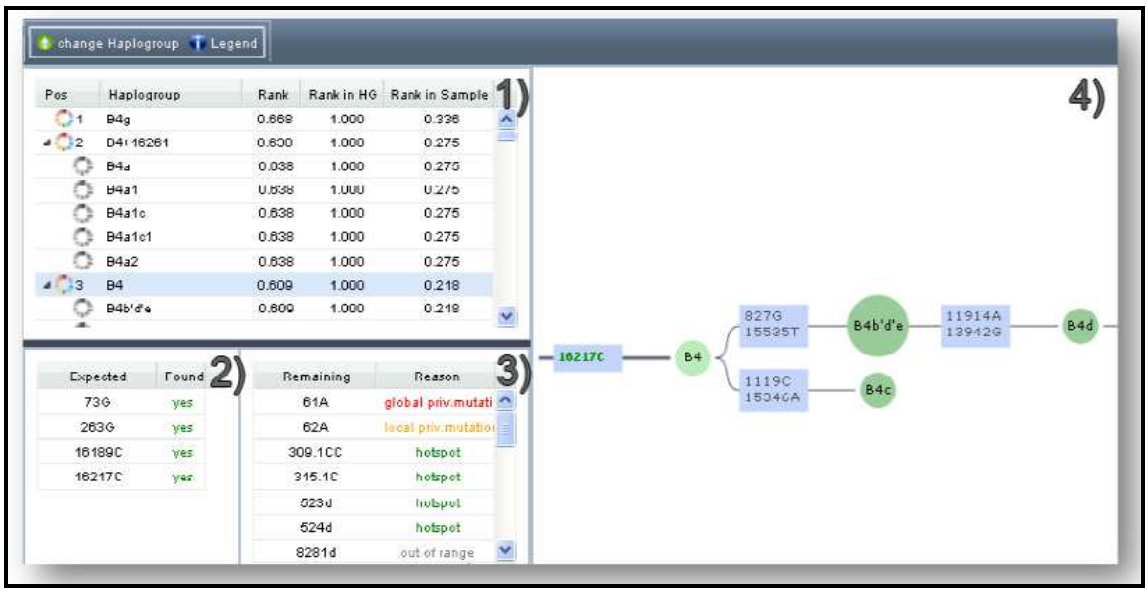

Figure 7. Information reported by Haploview

In this web tool, for every sample's haplogroup the following information is provided:

1) The haplogroups sorted by their final rank. The highest scoring haplogroup is per default used as the suggested haplogroup

2) The polymorphisms that are required by the current selected haplogroup. When a haplogroup-associated polymorphism is found in the sample, it is marked with yes, otherwise with no.

3) The remaining variants from the selected sample and the category which it belongs to. There are 4 categories:
    a. Hotspot: Mutation hotspots are 309.1C, 315.1C, 523-524d, 16182C, 16183C, 16193.1C and 16519
    b. Local private mutation: mutation observed in this sample, but not associated with the selected haplogroup.
    c. Global private mutation: mutation never observed in Phylotree, probably due to inconsistent alignments, phantom mutations or point heteroplasmies. Phylotree is a phylogenetic tree of global human mitochondrial DNA variation, based on both coding- and control-region mutations, and including haplogroup nomenclature. This mtDNA tree is meant as a framework for evolutionary anthropologists, medical geneticists, genealogists and forensic geneticists (van Oven and Kayser, 2009).

    d. Out of range: polymorphism lies outside the indicated range.

4) The graph showing the pathway from the rCRS to the currently selected Haplogroup. There are different colour-coding for the polymorphisms:

3.6. PATHOGENICITY PREDICTION

There are some criteria that suggest that a novel base change is pathogenic. After confirm that the variant is not a known polymorphism, it is important to check if the base change is affecting a site that is conserved during evolution.

To assess these possible implications of the variants detected, it is needed an in silico analysis by the utilization of computational algorithms, databases and online resources for the prediction of pathogenicity of missense variants. Some of these resources are tools as mtSNP (http://mtsnp.tmig.or.jp/cgi-bin/mtsnp/snpSearch/snpSearchEnglish.cgi) to determine the Grantham index of those variants that have been reported yet, SIFT and Polyphen that gave us a probability of pathogenicity of the aminoacidic change and

Align-GVGD that provides a classification of the aminoacid conservation (Grantham Score).

Protein sequence conservation analysis at a specific position is the first step to infer functional importance of an amino acid residue. The protein sequence containing the amino acid of interest should be used as a query for analysis using the NCBI protein website. We can Use ClustalW2—Multiple Sequence Alignment tool (http://www.ebi.ac.uk/Tools/msa/clustalw2/) for multiple sequence alignment.

The SIFT algorithm (sorting intolerant from tolerant) (http://sift.jcvi.org) is mainly based on sequence homology and the physical properties of amino acids to make predictions about protein function. It can be used to predict the likely effect of a non-synonymous substitution on protein function.

The PolyPhen (from polymorphism phenotyping) uses sequence-based features and structural information for characterizing the substitution to make predictions about the structure and function of the new protein. It is a structure-sequence-based amino acid substitution prediction method. The current version is PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2). It uses the data available in UniProtKB/UniRef100 and is based on conservation, protein folding, and crystal structure. This analysis classifies variants as likely benign, possibly damaging, or probably damaging when predicting pathogenicity.

With the mtSNP tool we can know the Grantham score of each variant that has been previously submitted, however, to know this score in the rest of the variants, we use ALIGN GVGD tool. Align-GVGD is a freely available, web-based program that combines the biophysical characteristics of amino acids and protein multiple sequence alignments (hand-made using Uniprot, selecting all the species that mtSNP tool uses to predict the Grantham score) to predict where missense substitutions in genes of interest fall in a spectrum of grades of clinical significance.
Align-GVGD is an extension of the original Grantham differences to multiple sequence alignments and true simultaneous multiple comparisons.
The biochemical variation at each alignment position is converted to a Grantham Variation score (GV). The difference between the properties of the wild-type amino acid and those of the amino acid variant being assessed are calculated and a Grantham

Difference score is generated (GD). These values are used as a measure of how likely the substitution is to be deleterious or neutral on a classification spectrum. The prediction classes form a spectrum (C0, C15, C25, C35, C45, C55, C65) with C65 most likely to interfere with function and C0 least likely (Tavigtian, SV et al, 2006).

None of these softwares have been clinically validated according to mitochondrial disorders because these applications are based on nuclear-encoded proteins. Proteins that are mtDNA-encoded have a different evolutionary conservation and there should be a source of errors when using these types of prediction tools. Bioinformatics prediction tools may be valuable as screening tools for identifying alleles of high pathogenicity for molecular and disease association studies. However, because the error rates in both nuclear and mitochondrial predictions are still high, current algorithms do not supplant the need for in vitro or in vivo studies.

However, we can take into account those variants that are in a region that is functionally important. This essentially means anywhere in the tRNA genes, and rRNA genes, and not only considering those that cause an amino acid change in the protein encoding genes.

3.7. STATISTICAL ANALYSIS

One of the objectives of this study is to identify whether schizophrenia patients present a higher number of variants in the mitochondrial DNA than control subjects. After the identification of all the variants that doesn't characterize the haplotype of a sample, we decided to infer whether there is an increased number of variants in cases than in controls, which could be related with the disease. With this goal, we applied an unpaired Wilcoxon test, which is a nonparametric alternative to the two-sample t-test when data do not follow a normal distribution and the sample size is small. .
Statistical analyses were performed using R language.

4. RESULTS AND CONCLUSIONS

4.1. RESULTS OF THE VARIANT IDENTIFICATION

In this study, the PGM equipment was used to sequence 93 mitochondrial genomes from 93 different samples. The use of barcoded adaptors allowed simultaneous sequencing of multiple samples with the PGM.

After applying all the steps of the bioinformatic analysis, and check each sample in Haplogrep to determine the haplogroup of each sample analyzed, we obtain a list of variants of each sample that doesn't define its haplogroup and also some variants that doesn't appear in Phylotree but they can be found in Mitomap or in the bibliography. We also obtain a list of variants that could not be found in any resource. According to these variants, they have not been submitted yet as a haplogroup variant, as an unclassified variant or as pathogenic variant. There are a lot of parameters that have to take into account to accept or discard a new variant. To determine the list of variants that present each sample and also to determine which of these variants can be candidates for additional studies, different analysis thresholds were established:

- A Phred-like quality score of the called genotype (Q call). The Q call established as a threshold was a value larger than 200 to detect all the homoplasmic variants. This threshold allows us to have a high reliability in the results. We detect also that all the variants with a low Q call value were heteroplasmic variants or artefacts of the technique. It is needed to use IGV viewer to determine whether those variants with a Qcall lower than 200 were definitively heteroplasmic variants. However, it is important to check them with Sanger Sequencing to definitively prove that it is a new variant. Moreover, some of the indels with a Qcall larger than 200 were checked with IGV viewer to determine the possibility of being a homopolimer artefact, that is, stretches of DNA containing the same base.

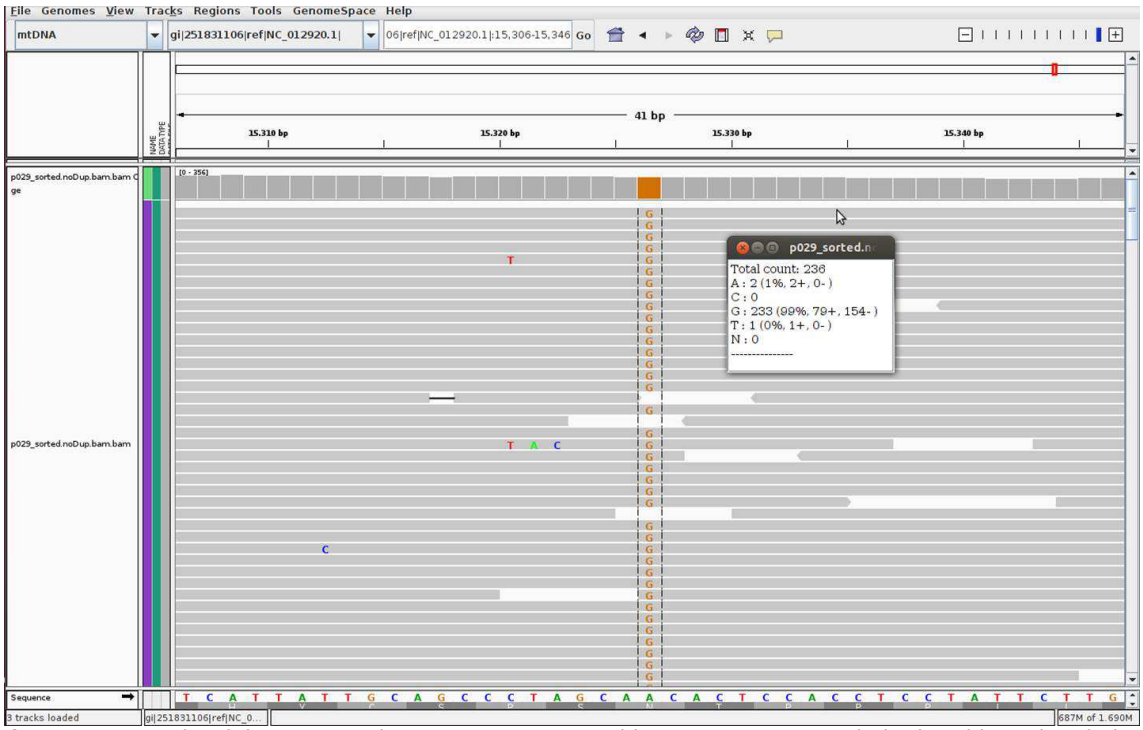

Figure 8. Example of the P029 variant m.15326A>G. In this case we can conclude that this variant is in homoplasmy.

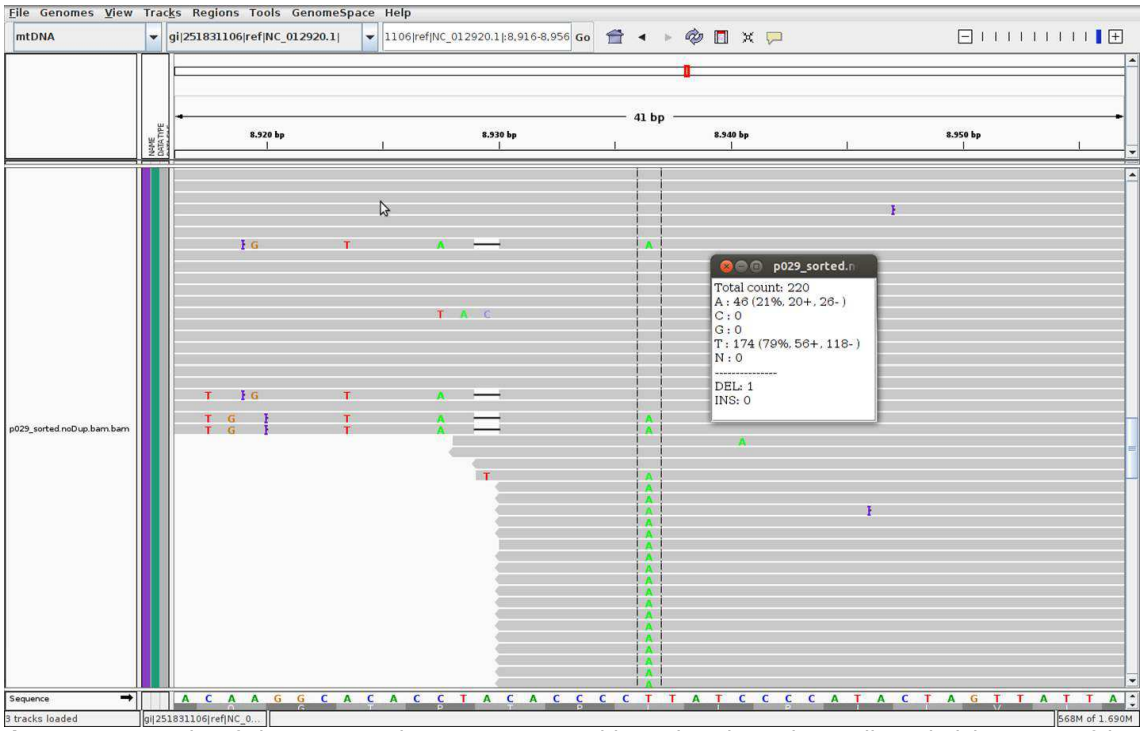

Figure 9. Example of the P029 variant c.8936T>A. This variant have been discarded because of its Qcall<200 and probably is a variant in heteroplasmy.

- Depth of coverage. We also establish a depth of coverage larger than 30X. This is an arbitrary value, however looking at the bibliography we can see that is a minimum value of coverage to have reliable results (Wang J, et al. 2012).

- The percentage of cases with the same variant. We proceed with the elimination of known polymorphisms and variants present in more than 1% of the cases (an arbitrary value found in the Bibliography as a correct value as the definition of rare mtDNA variants calculating a frequency of 93:1=1%) (Wang J, et al. 2012). Of those variants that have never been described, we only list those that are in less than 10% of the samples, assuming that those that are in a higher percentage can be artefacts of the technique, for example the deletion in the position 9546 that can be observed in the vast majority of the samples.

Table 3 (Supplementary files) presents each variant with the average of the Q call, the number of cases presenting the variants and whether they have been reported before. As we can see, some of the variants detected had never been submitted before. Those variants that weren't in the public databases were candidates for further analyse of its putative pathogenicity.

The known mutations and novel variants identified in the study are listed in table 4 and 5 (Supplementary files). The classification of each sample on its haplogroup, variants that don't appear in the haplogroup (i.e. variants that belong to another haplogroup, variants that have already been described and variants never described before), and the Genbank frequency are also presented.

We can observe that there are some variants that have not been previously reported. Some of the polymorphisms that have been described are related with a mitochondrial disease, like m.3310C>T that have been related with diabetes mellitus type II patients and with heart disease (Hattori Y, et al.2005).

Interestingly, there are twelve variants that have not been described before. Alignment of mitochondrial sequences that include different mitochondrial haplogroups, revealed synonymous as well as non-synonymous sequence variation that may affect mitochondrial protein function. In table 6 we can see the Polyphen prediction and also the Grantham Value of the missense variants. To know the possible implications of the mtDNA variant at the protein level, these variants have been submitted to *in silico*

analysis to obtain the Grantham score. Using Grantham score (Grantham, 1974), which categorize amino acid substitutions into classes of increasing chemical dissimilarity, 16 out of the 33 non-synonymous mutations were designated as conservative (score 0 to 50), 15 moderately conservative (score 51 to 100), and 2 moderately-radical (score 101 to 150) substitutions, according to the proposed classification (Li et al. 1984).

Several of the sequence variations were observed in some, but not all haplogroups, suggesting relatively early mutation events in mitochondrial evolution, whereas other variants segregated within haplogroups, suggesting relatively more recent mutation events, e.g. 1182C variation was found in samples C014, P007 and P025, all of them belonging to the U2d haplogroup.

Table 6. Variants not reported before

VARIANT	CHANGE EFFECT	LOCUS	ALIGN-GVGD PREDICTION	POLYPHEN PREDICTION
2861G		MT-RNR2 (rRNA)		
3106insC		MT-RNR2 (rRNA)		
3107T		MT-RNR2 (rRNA)		
5818T		MT-TC (tRNA)		
8209T	aa 209 SYN	MT-CO2		
8863A	p.V113M	MT-ATP6	48.95 (C0)	**PROBABLY DAMAGING**
9506T	aa 100 SYN	MT-CO3		
11812C	aa 351 SYN	MT-ND4		
11926T	aa 389 SYN	MT-ND4		
12568T	aa 78 SYN	MT-ND5		
13338A	p.F334L	MT-ND5	354 (C0)	**PROBABLY DAMAGING**
14593insC	aa 27 SYN	MT-ND6		

Only two non-reported variants were predicted as probably damaging, both detected in control samples. These were a valine to methionine substitution in the protein encoded by *MT-ATP6* and a phenylalanine to leucine in the protein encoded by *MTND5*, a core subunit of the mitochondrial respiratory chain NADH dehydrogenase (Complex I) that is

believed that belong to the minimal assembly required for catalysis. Complex I functions in the transfer of electrons from NADH to the respiratory chain.

Furthermore, we have also identified novel not missense variants. Because of there is no change in the amino acid sequence they were not considering for further study. However, it can be useful to develop a directed study to analyze the presence of these variants in a huge amount of patients and controls and also to develop functional studies of all these variants that are located in important regions like the control region where we cannot infer whether the change of this variant modify the functionality of this non-coding region.

4.2. RESULTS OF THE CASE-CONTROL STUDY

We apply the Wilcoxon Test to determine whether the schizophrenia patients present a distinct number of variants compared to the controls. This test assesses whether the distribution of values in two samples differ or not. Descriptive statistics and the results of the statistical test are shown in Table 7. In a first analysis we take into account only those variants that are not involved in defining the haplogroup for each participant. The differences in the number of variants present in cases (N=51; $\overline{X}$=0.9444, SD=1.07) and control subjects (N=33; $\overline{X}$ = 0.8684, SD=1.0698) are not significant (W=981, p=0.706).
We repeat this analysis but taking into account all the variants in each sample, in cases (N=214; $\overline{X}$=3.019, SD=2.4378) and controls (N=137; $\overline{X}$=2.737, SD=2.511425), that doesn't define the haplogroup where it belongs to, and the result is also not statistically significant (W= 925.5; p= 0.422).
In conclusion, we don't identify evidences of any difference between patients and controls in the number of variants present in their mtDNA molecules. However, this result should be taken carefully because probably the sample size of this study is insufficient to determine whether there is an association between the number of variants and the presence of the disease.

Table 7. Descriptive statistics and the Wilcoxon Test for comparing the number of variants in the two groups of individuals.

SAMPLES	QUANTILES	VARIANTS	
		NOT ASSOCIATED TO ANY HAPLOTYPE	ALL THE VARIANTS (INCLUDED THOSE THAT DEFINE OTHERS HAPLOTYPES)
CONTROLS	Min	0.0000	0.000
	1st Qu.	0.0000	1.000
N=38	Median	1.0000	2.000
	Mean	0.8684	2.737
	3rd Qu	1.0000	4.000
	Max	4.0000	10.000
	sd	1.0698	2.511425
	N	**33**	**137**
PATIENTS	Min	0.0000	0.000
N=55	1st Qu.	0.0000	1.000
	Median	1.0000	3.000
	Mean	0.9444	3.019
	3rd Qu	1.7500	4.000
	Max	4.0000	12.000
	sd	1.0714	2.4378
	N	**51**	**214**
	W	**981**	**925.5**
	p-value	**0.7058**	**0.4218**

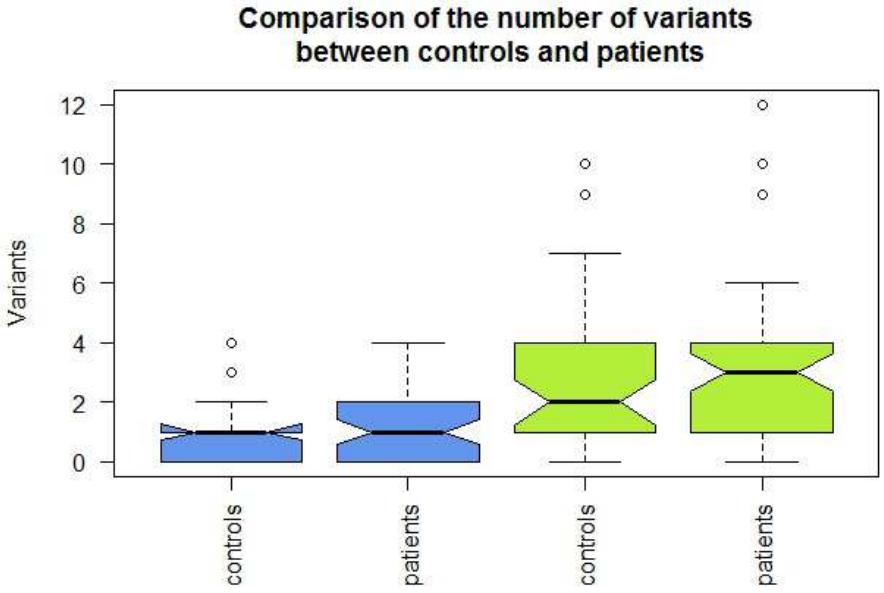

Figure 10. Boxplot of the comparison of the number of variants that present controls and samples. In blue we can see the boxes corresponding only to the number of variants excluding those that define another haplotype. In green we take into account all those variants that present the samples (excluding those that define their haplotype).

We also make this analysis taking into account the region of the mitochondrial DNA where the mutation is located. The mtDNA can be divided into two distinct important regions: the control region, also known as the D-loop and the coding region. We divide the variants detected into these two locations and perform the statistical analysis comparing the presence of variants between cases and control samples. The number of variants in the coding region in control subjects (N=68; X=1.789, SD=2.0816) and patients (N=98, X=1.782, SD=1.8628) is not statistically different. We also analyze the difference of the presence of variants in the control region between controls (N=32, X=0.8421, SD=0.8229) and patients (N=64, X=1.109, SD=0.8229) and neither is a statistically difference between these groups (W=876, p=0.1586). However, we observe that the presence of variants in the coding region in patients is slightly higher than in controls.

To determine if there are differences of the number of variants between the two mtDNA regions of the same group of individuals, we apply a paired Wilcoxon Test and in this case we observe that there are significant differences in both groups of samples, that is that the number of variants observed is significantly higher in de coding region in cases (V=499, p=0.02361) and in control subjects (V=320, p=0.006827).

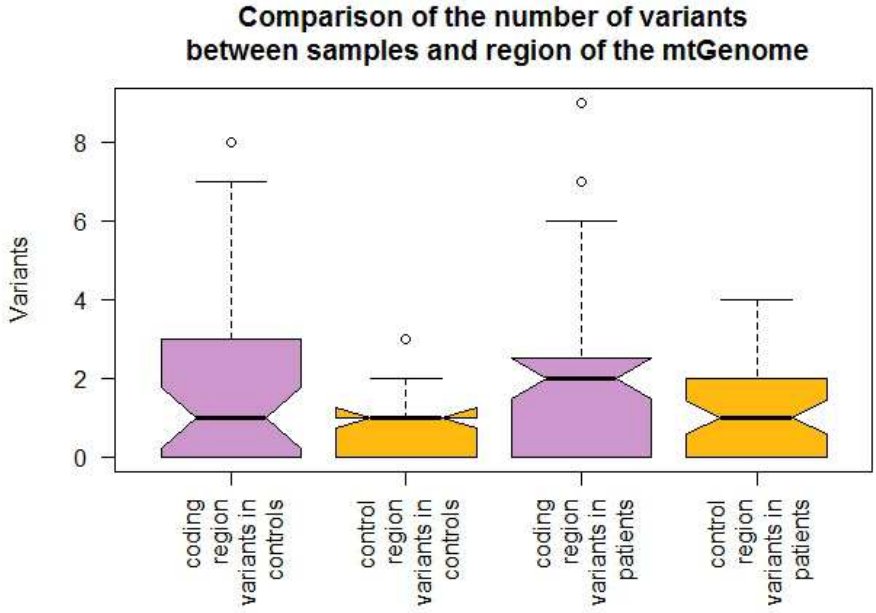

Figure 11. Boxplot of the comparison of the number of variants located in two main important regions of the mtDNA.

Table 8. Statistical results of the Wilcoxon Test comparing the number of variants of the diferent mtDNA regions that present the two groups of individuals and the quantile values of the boxplot.

SAMPLES	QUANTILES	MtDNA REGION		
		CODING REGION	CONTROL REGION	
CONTROLS	Min	0.000	0.0000	
	1st Qu.	0.000	0.0000	
N=38	Median	1.000	1.0000	
	Mean	1.789	0.8421	**V = 320**
	3rd Qu	3.000	1.0000	**p = 0.006827**
	Max	8.000	3.0000	
	sd	2.081552	0.8228597	
	N	**68**	**32**	
PATIENTS	Min	0.000	0.000	
	1st Qu.	0.000	0.000	
N=55	Median	2.000	1.000	
	Mean	1.782	1.109	**V = 499**
	3rd Qu	2.500	2.000	**p = 0.02361**
	Max	9.000	4.000	
	sd	1.862803	0.936358	
	N	**98**	**61**	
	W	**1002.5**	**876**	
	p-value	**0.7361**	**0.1586**	

4.3. CONCLUSIONS

We have studied the entire mtDNA of 55 schizophrenia subjects with a matrilineal transmission pattern of the illness to test the hypothesis that mutations present in their mitochondrial genomes contribute to the genetic basis of the illness and in 38 control subjects. We found considerable sequence diversity. When we compared the mtDNA of the patients with the revised Cambridge sequence, we found 147 homoplasmic variants that don't define the haplotype of the sample (they can be variants that define other haplogroups or other polymorphisms, rare variants or mutations). All variants were distributed along the mitochondrial genome and did not accumulate in a specific region. Of the 147 variants, twelve have not been previously reported and two of them were non-synonymous substitutions but both where only detected in controls. We also

found some variants that have been reported before and it can be hypothesized that they can be involved in the increased susceptibility to schizophrenia, e.g. m.4136A>G with a Grantham score of 194 and m.3368T>C with a Grantham score of 81. There are also some samples (e.g. patient P042) that have a larger number of missense variants; this is an important issue to take into account because, as we mentioned before, the susceptibility of mitochondrial disorders is not only due to the presence of one variant but to the accumulation of several variants.

The other variants found are located in the MT-RNR2 region (that codifies rRNA), MT-CO2, MT-CO3, MT-ND4, MT-ND5 and MT-ND6. We have observed that the number of synonymous variants is larger than the number of missense variants. This can be explained by the purifying selection that perhaps acts against amino acid changes in the protein-coding genes (Lawrie DS, et al. 2013). The vast majority of variants found during our study are synonymous variants or changes located in non-coding regions. It is important to consider that these variants can be also important for the functionality and replication of the mitochondria.

Nevertheless, after analyze the differences between cases and controls; we realize that there are no statistical differences in the presence of variants between the two groups or the different mitochondrial regions. This result can be due because of the small size of our control-case study; consequently a larger sample size study is needed to determine if there is a significant difference between the two populations.

4.4. LIMITATIONS OF THE STUDY AND FUTURE WORK

One of the most important limitations of this project was the lack of sequences obtained by Sanger sequencing. The resulting list of variants should be compared with the variants detected by Sanger Sequencing of the exact samples. As we mentioned, it is known that a 500X coverage in the NGS is needed to determine the heteroplasmic level. To further determine the correct threshold that we have to apply to detect heteroplasmic variants, an incoming step performing a Next Generation Sequencing with a 500X coverage and a Sanger Sequencing of several samples are needed to validate the different variants candidates to be heteroplasmic variants.

In a following step, some important things have to be done:

- An R package, homologous to ANNOVAR (that provides variant annotation), linked to haplogrep, to establish the haplogroup of each sample, and all this variants susceptible to be a pathogenic mutation, also looking at the SIFT and Polyphen tools. Since the end of 2013, in the ANNOVAR package there is also the possibility to annotate mitochondria variants. However there are several important caveats as the Reference Sequence used by ANNOVAR is not the correct sequence because the UCSC's hg19 assembly used the old version of the mitochondria genome. The problem is whether we align our data against the Reference sequence (NC_012920), we cannot really annotate our variants using UCSC's gene definition. Now, with ANNOVAR we have to look at those variants that align together with chrM of the UCSC

- To test matrilineal relatives. Targeted sequence analysis of the patient's mother and other matrilineal relatives is recommended to be done to determine if the variants detected are definitely matrilineal inherited. When a variant is homoplasmic in asymptomatic matrilineal adult relatives and is not cosegregating with the disease phenotype, then that variant, by itself, is unlikely to be the primary cause of the clinical symptoms. If a variant is absent or at low level of heteroplasmy in asymptomatic matrilineal relatives or is cosegregating with disease phenotype, then this variant may be pathogenic.

- Additional studies, including mitochondrial functional studies may be needed to further try to clarify the clinical significance of the variants found, including missense variants and also those non-codifying variants but nevertheless they can be important in any function like these variants located in the control region.

In conclusion, these results need to be validated by Sanger Sequencing and replicated in many more samples to further determine the validity of the obtained results. It would also be interesting to analyse in other case-control studies some of the novel variants found (after confirm their presence by Sanger).

5. BIBLIOGRAPHY

ANDREWS, R.M., KUBACKA, I., CHINNERY, P.F., LIGHTOWLERS, R.N., TURNBULL, D.M., HOWELL, N. "Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA". 1999; *Nat. Genet*. 23 (2), 147.

ANGLIN RE, TARNOPOLSKY MA, MAZUREK MF, ROSEBUSH PI. "The psychiatric presentation of mitochondrial disorders in adults". *The Journal of Neuropsychiatry and Clinical Neurosciences* 2012;24(4):394–409

ADZHUBEI IA, SCHMIDT S, PESHKIN L, RAMENSKY VE, GERASIMOVA A, BORK P, KONDRASHOV AS, SUNYAEV SR. "A method and server for predicting damaging missense mutations". *Nat Methods* **7**(4):248-249 (2010).

CACABELOS R, FERNÁNDEZ-NOVOA L, LOMBARDI V, CARRIL JC, CORZO L, CARRERA I, TELLADO I, MARTÍNEZ R, MCKAY A, TAKEDA M. "Genomics of Schizophrenia and Psychotic Disorders". *Science*, April 2011

GILES RE, BLANC H, CANN HM, WALLACE DC. "Maternal inheritance of human mitochondrial DNA". *Proc Natl Acad Sci U S A* 1980;77:6715–9.

GUNNARSDÓTTIR ED, LI M, BAUCHET M, FINSTERMEIER K, STONEKING M. "High-throughput sequencing of complete human mtDNA genomes from the Philippines". *Genome Res*. 2011 Jan;21(1):1-11

HATTORI Y, TAKEOKA M, NAKAJIMA K, EHARA T, KOYAMA MA**.** "Heteroplasmic mitochondrial DNA 3310 mutation in the ND1 gene in a patient with type 2 diabetes, hypertrophic cardiomyopathy, and mental retardation". *Exp Clin Endocrinol Diabetes*. 2005 Jun;113(6):318-23.

HUDSON G. "No evidence of an association between mitochondrial DNA variants and osteoarthritis in 7393 cases and 5122 controls". *Ann Rheum Dis* 2013; 72 :136 –139.

LAWRIE DS, MESSER PW, HERSHBERG R, PETROV DA. “Strong Purifying Selection at Synonymous Sites in *D. melanogaster*.” *PLoS Genet*. 2013. 9(5): e1003527. doi:10.1371/journal.pgen.1003527

MATHE E, OLIVIER M, KATO S, ISHIOKA C, HAINAUT P, TAVTIGIAN SV. “Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods”. *Nucleic Acids Res*. 2006 Mar 6;34(5):1317-25.

MCGRATH J, SAHA S, CHANT D, WELHAM J. “Schizophrenia: a concise overview of incidence, prevalence, and mortality”. *Epidemiologic Reviews* 2008;30:67–76

MOSQUERA-MIGUEL A, TORRELL H, ABASOLO N, ARROJO M, PAZ E, RAMOS-RÍOS R, ET AL. “No evidence that major mtDNA European haplogroups confer risk to schizophrenia”. *American Journal of Medical Genetics Part B, Neuropsychiatric Genetics* 2012;159B(4):414–21.

R CORE TEAM. “R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria”. 2013. URL http://www.R.project.org/.

ROBINSON JT, THORVALDSDOTTIR H, WINCKLER W,ET AL. “Integrative genomics viewer”. *Nat Biotechnol.* 2011; 29 :24–6

ROTHBERG JM, HINZ W, REARICK TM, ET AL. “An integrated semiconductor device enabling non-optical genome sequencing”. *Nature* 2011; 475:348–352

SCHON EA, DIMAURO S, HIRANO M. “Human mitochondrial DNA: roles of inherited and somatic mutations”. *Nature Reviews Genetics* 2012;13(12):878–90.

SKLADAL D, HALLIDAY J, THORBURN DR. “Minimum birth prevalence of mitochondrial respiratory chain disorders in children”. *Brain* 2003;126(Pt 8):1905–12.

SULLIVAN, P.F. “The genetics of schizophrenia”. *PLoS Med*. 2013. 2, e212.

TANG S, HUANG T. "Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system". *BioTechniques* 2010;48:287–96.

TAVTIGIAN SV, DEFFENBAUGH AM, YIN L, JUDKINS T, SCHOLL T, SAMOLLOW PB, DE SILVA D, ZHARKIKH A, THOMAS A. "Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral". *J Med Genet*. 2005 Jul 13

THORVALDSDÓTTIR H, ROBINSON JT, MESIROV JP. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". *Brief Bioinform*. 2013 Mar;14(2):178-92.

VAN OS J, KAPUR S. "Schizophrenia". *Lancet* 2009;374(9690):635–45.

VAN OVEN M, KAYSER M. "Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation". *Hum Mutat* 2009. 30(2):E386-E394. http://www.phylotree.org.

VERGE B, ALONSO Y, VALERO J, MIRALLES C, VILELLA E, MARTORELL L. "Mitochondrial DNA (mtDNA) and schizophrenia". *European Psychiatry* 2011;26(1):45–56.

WALLACE DC. "A Mitochondrial Paradigm of Metabolic and Degenerative Diseases, Aging, and Cancer: A Dawn for Evolutionary Medicine". *Annu Rev Genet*. 2005 ; 39: 359.

WANG J, SCHMIDT ES, LANDSVERK M L, ET AL. "An integrated approach for classifying mitochondrial DNA variants: one clinical diagnostic laboratory's experience". *Genetics in Medicine*. 2012; 14(6):620-626

ZARAGOZA MV, FASS J, DIEGOLI M, LIN D, ARBUSTINI E. "Mitochondrial DNA Variant Discovery and Evaluation in Human Cardiomyopathies through Next-Generation Sequencing". *Plos One.* 2010. 5(8):e12295.

6. SUPPLEMENTARY FILES

Table 3. List of variants detected with a Q call>200

VARIANT	QCALL AVERAGE	SAMPLES	REPORTED PREVIOUSLY
93G	222	P014	Yes
143A	222	P020	Yes
146C	222	P052	Yes
152C	222	C045,P008,P012,P015,P042	Yes
186A	222	P021	Yes
195C	222	P015,P036	YES (disease associated?)
204C	222	P027	Yes
207A	222	P027	Yes
234G	222	C038	Yes
239C	222	P059	Yes
295T	204	C039	Yes
309.1CT	208	P002,P005,P020,C001,C030,C048	Yes
524.1AC	214	P007,P025,P045,C029,C014,C032, C036	Yes
534T	222	P037	Yes
709A	222	C014,C015, P007, P025, P052	Yes
729C	222	P028,	Yes (disease associated?)
1027G	222	P002	Yes(disease associated?)
1189	222	C039	Yes
1192T	222	P035	Yes(disease associated?)
1694C	222	P056	Yes
2501T	222	P059	Yes
2558G	222	P039	Yes
2861G	222	P043	No
3010A	222	P009	Yes
3083C	222	P042	Yes
3084G	222	C045	Yes
3106.1C	214	P020	No
3107T	214	C032,C041	No
3310T	222	P026	Yes (disease associated?)
3368C	222	P010	Yes
3391A	222	P021	Yes
3540C	222	C039	Yes
3774G	225	P039	Yes
4029T	222	C001,C040	Yes
4136G	222	C036,P031	Yes
4619C	222	C048	Yes
5046A	222	P065	Yes
5585A	222	P023	Yes

VARIANT	QCALL AVERAGE	SAMPLES	REPORTED PREVIOUSLY
5818T	222	C021	No
5960T	222	P056	Yes
6095G	222	C036	Yes
6182A	209	P062	Yes
6212G	222	P022,	Yes
6249G	222	C048	Yes
6503G	222	P007,P025,C014	Yes
6590C	222	P050	Yes
6734A	222	P040	Yes
6908C	222	P062	Yes
7220C	222	P036	Yes
7280T	222	P022	Yes
7317G	222	C028	Yes
7468T	222	P022	Yes
7775A	222	C039	Yes
7830A	222	P066	Yes
8020A	222	C039	Yes
8209T	222	C047	No
8269A	222	C040,C001	Yes
8281-8289del	214	P002	Yes
8475T	222	P064	Yes
8513T	222	C021	Yes
8639C	222	P046	Yes
8701G	222	P042	Yes
8706G	222	C028	Yes
8838A	222	C028	Yes
8863A	222	C032	No
9182A	222	P038,P065	Yes
9490T	222	P054	Yes
9506T	222	P018	No
9899C	222	P015	Yes
9909C	222	P042	Yes
9966A	221	C043	Yes
10199T	222	P050	Yes
10248C	222	C025	Yes
10398G	222	P055	Yes
10544T	222	P056	Yes
11063T	222	P035	Yes
11404G	222	P032	Yes
10463C	222	P053	Yes
11653G	222	C013	Yes
11654G	222	P042	Yes

VARIANT	QCALL AVERAGE	SAMPLES	REPORTED PREVIOUSLY
11812C	222	P007, P025, C014	No
11836G	222	P040	Yes
11914A	222	C048, P043	Yes
11926T	222	C032	No
11971T	222	P035	Yes
12103A	222	C006	Yes
12373G	222	P042	Yes
12568T	225	P039	No
12582G	222	P004	Yes
12612G	222	P009	Yes
12684A	222	P029	Yes
12945C	222	P004	Yes
13105G	222	P042,P062,	Yes
13135A	222	C001, C040	Yes (disease asssociated?)
13161C	222	C033	Yes
13191C	222	P035	Yes
13230T	222	C045	Yes
13242G	222	C005	Yes
13338A	222	C006	No
13359A	222	P027	Yes
13464T	225	P026	Yes
13557G	222	C039	Yes
13575T	225	P047	Yes
13608C	222	C048	Yes
14002G	222	P043	Yes
14003T	222	C006	Yes
14016A	111	P028	Yes
14034C	222	P002	Yes
14178C	222	P066	Yes
14259A	222	P054	Yes
14544A	204	P010,P064	Yes
14572G	222	P046	Yes
14593.1C	222	P053	No
14755G	222	P066	Yes
14869A	222	P038	Yes
15257A	222	P026	Yes
15514C	222	C001, C040	Yes
14861A	222	P066	Yes
15916C	222	P035	Yes
15924G	222	C048,P018	Yes
15927A	222	P025	Yes(related with disease?)

VARIANT	QCALL AVERAGE	SAMPLES	REPORTED PREVIOUSLY
15977T	222	P055	Yes
15984C	222	P025	Yes
16042A	222	P050	Yes
16092C	222	P032	Yes
16093C	222	P014,P045,C038,C043	Yes
16111T	222	P032,P045	Yes
16129A	222	C031	Yes
16162G	222	C001,C040	Yes
16163G	222	P015	Yes
16171G	222	P058	Yes
16180G	218	P047	Yes
16183d	214	P040	Yes
16193T	222	C036	Yes
16189d	202	P015,P055	Yes
16241G	222	P040	Yes
16249C	203	P047	Yes
16261T	222	P052	Yes
16270T	222	P038,P065	Yes
16286A	191	P002	Yes
16290T	222	C041	Yes
16298C	222	P031	Yes
16311C	222	P004,P038,P065	Yes (disease associated?)
16319A	222	P011	Yes
16342C	222	P042	Yes
16362C	222	C026,P036,P001	Yes
16519C	222	C001, C010, P044	Yes

Table 4. Variants (that doesn't define the haplogroup of the sample) identified in the 38 controls, sequenced by NGS.

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
C001	H1e1a	309	CT insertion	MT-DLOOP				6	-
		4029	C > T	MT-ND1	241			2	2
		8269	G > A	MT-CO2	228			2	252
		13135	G > A	MT-ND5	267	A > T	58 **(185-C0)**	2	194
		15514	T > C	MT-CYB	256			2	110
		16162	A > G	MT-DLOOP				2	223
		16519	T > C	MT-DLOOP				3	11617
C005	T2c1D1a	200	A > G	MT-DLOOP				1	447
		13242	A > G	MT-ND5	302			1	1
C006	U5a1+!16192	4796	C > T	MT-ND2	109			1	7
		10550	A > G	MT-ND4L	27			1	915
		12103	C > A	MT-ND4	448			1	4
		13338	C > A	MT-ND5	334	F > L	**354 (C0)**	1	
		14003	C > T	MT-ND5	556	T > I	89 **(161-C0)**	1	6
		14527	A > G	MT-ND6	48			1	11
		14893	A > G	MT-CYB	49			1	32
		14971	T > C	MT-CYB	75			1	34
		16278	C > T	MT-DLOOP				1	1845
C007	U1a1b	523-524	delAC	MT-DLOOP				11	
C008	T2b6a								
C009	H2a5a1								
C010	V1a	13215	T > C	MT-ND5	293			1	82
		16216	A > G	MT-DLOOP				1	17
		16519	T > C	MT-DLOOP				3	11617

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
C012	H10b	9448	A > G	MT-CO3	81	Y > C	**194 (C65)**	2	5
C013	J1c3e1	11440	G > A	MT-ND4	227			1	127
		11653	A > G	MT-ND4	298			1	149
		15737	G > A	MY-CYB	331	D > N	**23.01 (C0)**	1	3
		16153	G > A	MT-DLOOP				1	175
C014	U2d	524	insCA	MT-DLOOP				7	9
		709	G > A	MT-RNR1				5	2404
		930	G > A	MT-RNR1				3	419
		4164	A > G	MT-ND1	286			3	130
		6503	A > G	MT-CO1	200			3	4
		10343	C > T	MT-ND3	95			3	12
		11812	A > C	MT-ND4	351			3	
		12127	G > A	MT-ND4	456			3	44
		16266	C > T	MT-DLOOP				3	204
		16525	A > G	MT-DLOOP				2	9
C015	U4a1b	709	G > A	MT-RNR1				5	2404
C016	R0a4								
C019	J1c3	12681	T > C	MT-ND5	115			1	29
		16213	G > A	MT-DLOOP				1	185
C020	H5a1	523-524	AC > del	MT-DLOOP				11	
		5567	T > C	MT-TW				1	17
		7570	A > G	MT-TD				1	4
C021	H1	5818	C > T	MT-TC		tRNA		1	
		8513	C > T	MT-ATP8	50	P > S	**214 (C0)**	1	2
		10253	T > C	MT-ND3	65			1	8

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
C023	H1j1								
C025	H1	7269	G > A	MT-CO1	456	V > M	**29 (C0)**	1	28
		10248	T > C	MT-ND3	64			1	20
		13188	C > T	MT-ND5	284			1	111
C026	J2b1a1	16362	T > C	MT-DLOOP				3	3089
C027	H1ba								
C028	HV0+195	7317	A > G	MT-CO1	472	I > V	**29 (C0)**	1	1
		8706	A > G	MT-ATP6	60			1	10
		8838	G > A	MT-ATP6	104			1	40
C029	K1a4a1e	524	insCA	MT-DLOOP				7	9
C030	I1a1b	309	insCT	MT-DLOOP				6	-
C031	K1a4a1	16129	G > A	MT-DLOOP				1	2159
C032	K1a+195	524	insCA	MT-DLOOP				7	9
		3107	N > T	MT-RNR2		rRNA		2	
		8863	G > A	MT-ATP6	113	V > M	**49 (C0)**	1	2
		11926	A > T	MT-ND4	389			1	
C033	U5b2b	9300	G > A	MT-CO3	32	A > T	58	2	84
		13161	T > C	MT-ND5	275			1	12
		13958	G > C	MT-ND5	541	G > A	60 **(209-C0)**	1	68
		16296	C > T	MT-DLOOP				1	388
C035	H4a1a2								
C036	H4a1a1a	524	insCA	MT-DLOOP				7	9
		4136	A > G	MT-ND1	277	Y > C	194	2	23
		6095	A > G	MT-CO1	64			1	2

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
		16193	C > T	MT-DLOOP					221
C037	H3q								
C038	J1c2e2	234	A > G	MT-DLOOP				1	80
		523-524	delAC	MT-DLOOP				11	
		16093	T > C	MT-DLOOP				4	984
C039	HV	295	C > T	MT-DLOOP				1	837
		1189	T > C	MT-RNR1		rRNA			749
		3540	T > C	MT-ND1	78			1	15
		7775	G > A	MT-CO2	64	V > I	**29 (C0)**	1	24
		8020	G > A	MT-CO2	145			1	213
		13557	A > G	MT-ND5	407			1	6
		15904	C > T	MT-TT		tRNA		1	286
C040	H1e1a	4029	C > T	MT-ND1	241			2	2
		8269	G > A	MT-CO2	228			2	252
		13135	G > A	MT-ND5	267	A > T	58	2	194
		15514	T > C	MT-CYB	256			2	110
		16162	A > G	MT-DLOOP				2	263
C041	U4a1	524	insCA	MT-DLOOP				7	9
		3107	N > T	MT-RNR2		rRNA		2	
		16290	C > T	MT-DLOOP				1	606
C042	H10b	9448	A > G	MT-CO3	81	Y > C	**194 (C65)**	2	5
C043	H1e1a1	9966	G > A	MT-CO3	254	V > I	29	1	117
		16093	T > C	MT-DLOOP				4	984

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
C045	H1bf	152	T > C	MT-DLOOP				6	3807
		3084	A > G	MT-RNR2		rRNA		1	4
		13230	C > T	MT-ND5	298			1	3
C047	H5a	8209	C > T	MT-CO2	209			1	
C048	N1b1a2	309	insCT	MT-DLOOP				6	-
		4619	T > C	MT-ND2	50			1	3
		6249	G > A	MT-CO1	116	A > T	58 **(104-C0)**	1	33
		7283	T > C	MT-CO1	460				110
		11914	G > A	MT-ND4	385			2	2063
		13608	T > C	MT-ND5	424			1	7
C049	U2e1a1	15924	A > G	MT-TT		tRNA		2	645

Table 5. Variants (that doesn't define the haplogroup of the sample) identified in 55 cases, sequenced by NGS

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
P001	U6a1a1	9300	G > A	MT-CO3	32	A > T	58	2	84
		13635	T > C	MT-ND5	433			1	45
		16293	A > G	MT-DLOOP				1	393
		16362	T > C	MT-DLOOP				3	3089
P002	U4a1	309	insCT	MT-DLOOP				6	-
		1027	A > G	MT-RNR1		rRNA		1	5
		8281-8289	del	MT-NC7				1	-
		9804	G > A	MT-CO3	200	A > T	58	1	56
		14034	T > C	MT-ND5	566			1	79
		16286	C > A	MT-DLOOP				1	10
P004	U5a1b1	12582	A > G	MT-ND5	82			1	5
		12945	T > C	MT-ND5	203			1	4
		16294	C > T	MT-DLOOP				1	1743
		16311	T > C	MT-DLOOP				3	3659
P005	T2b3b	309	insCT	MT-DLOOP				6	-
P007	U2d	524	insCA	MT-DLOOP				7	9
		709	G > A	MT-RNR1				5	2404
		930	G > A	MT-RNR1				3	419
		4164	A > G	MT-ND1	286			3	130
		6503	A > G	MT-CO1	200			3	4
		10343	C > T	MT-ND3	95			3	12
		11812	A > C	MT-ND4	351			3	
		12127	G > A	MT-ND4	456			3	44

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
		16266	C > T	MT-DLOOP				3	204
P008	K1a1b1	152	T > C	MT-DLOOP				6	3807
		3736	G > A	MT-ND1	144	V > I	29	1	42
P009	T1a1+!152	3010	G > A	MT-RNR2		rRNA		1	3108
		12612	A > G	MT-ND5	92			1	924
P010	H5a1	523-524	delAC	MT-DLOOP				11	
		3368	T > C	MT-ND1	21	M > T	81	1	6
		14544	G > A	MT-ND6	44			2	56
P011	U5b1f	16319	G > A	MT-DLOOP				1	1059
P012	H1j1	152	T > C	MT-DLOOP				6	3807
P013	H1bx								
P014	K1a+195	93	A > G	MT-DLOOP				1	427
		16093	T > C	MT-DLOOP				4	984
P015	T	152	T > C	MT-DLOOP				6	3807
		195	T > C	MT-DLOOP				2	3688
		9899	T > C	MT-CO3	231			1	217
		16163	A > G	MT-DLOOP				1	258
		16189	del	MT-DLOOP				2	971
P016	H6a1b2								
P018	H1j	9506	C > T	MT-CO3	100			1	
		15924	A > G	MT-TT		tRNA		2	645
P019	U5b3								

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
P020	W6a	143	G > A	MT-DLOOP				1	356
		309	insCT	MT-DLOOP				6	-
		3106	insC	MT-RNR2		rRNA		1	-
P021	H1i	186	C > A	MT-DLOOP				1	220
		3391	G > A	MT-ND1	29	G > S	56	1	26
P022	K1a+195	6212	A > G	MT-CO1	103			1	3
		7280	C > T	MT-CO1	459			1	7
		7468	C > T	MT-TS1		tRNA		1	5
P023	H1e	5585	G > A	MT-NC3		Non-coding		1	90
P024	H27								
P025	U2d	524	insCA	MT-DLOOP				7	9
		709	G > A	MT-RNR1				5	2404
		930	G > A	MT-RNR1				3	419
		4164	A > G	MT-ND1	286			3	130
		6503	A > G	MT-CO1	200			3	4
		10343	C > T	MT-ND3	95			3	12
		11812	A > C	MT-ND4	351			3	
		12127	G > A	MT-ND4	456			3	44
		15927	G > A	MT-ATT			tRNA	1	205
		15984	T > C	MT-ATT			tRNA	1	3
		16266	C > T	MT-DLOOP				3	204
		16525	A > G	MT-DLOOP				2	9
P026	K1a4a1	3310	C > T	MT-ND1	2	P > S	**158 (C0)**	1	8
		13464	C > T	MT-ND5	376			1	1

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
		15257	G > A	MT-CYB	171	D > N	23	1	242
P027	T2b	204	T > C	MT-DLOOP				1	991
		207	G > A	MT-DLOOP				1	814
		13359	G > A	MT-ND5	341			1	60
P028	H3	523-524	delAC	MT-DLOOP				11	
		729	T > C	MT-RNR1			rRNA	1	1
		14016	G > A	MT-ND5	560			1	70
P029	H1c	523-524	delAC	MT-DLOOP				11	
		12684	G > A	MT-ND5	116				
P031	V+!16298	4136	A > G	MT-ND1	277	Y > C	194	2	23
		16298	T > C	MT-DLOOP				1	960
P032	H1e5	11404	A > G	MT-ND4	215			1	38
		16092	T > C	MT-DLOOP				1	234
		16111	C > T	MT-DLOOP				2	486
P033	H1j1								
P034	U6a1a1								
P035	U6a1	1192	C > T	MT-RNR1				1	3
		11063	C > T	MT-ND4	102			1	3
		11971	C > T	MT-ND4	404			1	4
		13191	T > C	MT-ND5	285			1	21
		15916	T > C	MT-TT		tRNA		1	10
P036	H5a1c1a	195	T > C	MT-DLOOP				2	3688
		523-524	delAC	MT-DLOOP				11	

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
		7220	T > C	MT-CO1	439			1	16
		16362	T > C	MT-DLOOP				3	3089
P037	H1a3a	534	C > T	MT-DLOOP				1	36
P038	H56	9182	G > A	MT-ATP6	219	S > N	46	2	21
		14869	G > A	MT-CYB	41			1	50
		16270	C > T	MT-DLOOP				2	1003
		16311	T > C	MT-DLOOP				3	3659
P039	H13a1a	2558	A > G	MT-RNR2		rRNA		1	1
		3774	A > G	MT-ND1	156			1	-
		12568	C > T	MT-ND5	78			1	
P040	J1c8a	6734	G > A	MT-CO1	277			1	94
		11836	A > G	MT-ND4	359			1	2
		16183	Del	MT-DLOOP				1	23
		16241	A > G	MT-DLOOP				1	97
P041	J2b1a1								
P042	K1a4a1	152	T > C	MT-DLOOP				6	3807
		523-524	delAC	MT-DLOOP				11	
		3083	T > C	MT-RNR2		rRNA		1	13
		8701	A > G	MT-ATP6	59	T > A	58	1	6017
		9909	T > C	MT-CO3	235	F > L	22	1	7
		11654	A > G	MT-ND4	299	T > A	58	1	58
		12373	A > G	MT-ND5	13	T > A	58	1	13
		13105	A > G	MT-ND5	257	I > V	29	2	1503
		16342	T > C	MT-DLOOP				1	110

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
P043	H3b5	2861	A > G	MT-RNR2		rRNA		1	
		11914	G > A	MT-ND4	385			2	2063
		14002	A > G	MT-ND5	556	T > A	58	1	65
P044	U5b1g	16519	T > C	MT-DLOOP				3	11217
P045	K1a+195	524	insCA	MT-DLOOP				7	9
		8639	T > C	MT-ATP6	38	I > T	88 (C0)	1	7
		14572	A > G	MT-ND6	34			1	3
		16093	T > C	MT-DLOOP				4	984
		16111	C > T	MT-DLOOP				2	486
P047	K1a	13575	C > T	MT-ND5	413			1	3
		16180	A > G	MT-DLOOP				1	7
		16249	T > C	MT-DLOOP				1	363
P049	H1+16189								
P050	H3ap	6590	T > C	MT-CO1	229			1	2
		10199	C > T	MT-ND3	47			1	10
		16042	G > A	MT-DLOOP				1	14
P052	H13a1a1	146	T > C	MT-DLOOP				1	3613
		709	G > A	MT-RNR1				5	2404
		16261	C > T	MT-DLOOP				1	1339
P053	J1c2e	523-524	delAC	MT-DLOOP				11	
		10463	T > C	MT-TR		tRNA		1	902
		14593	insC	MT-ND6	27			1	

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
P054	J1c2e2	523-524	delAC	MT-DLOOP				11	
		9490	C > T	MT-CO3	95	A > V	64	1	9
		14259	G > A	MT-ND6	139	P > S	**264 (C0)**	1	14
P055	T1a1	10398	A > G	MT-ND3	114	T > A	58	1	7972
		15977	C > T	MT-ATT			tRNA	1	3
P056	H1e1a	1694	T > C	MT-RNR2			rRNA	1	11
		5960	C > T	MT-CO1	19			1	9
		10544	C > T	MT-ND4L	25			1	3
P058	H1	16171	A > G	MT-DLOOP				1	26
P059	H1at	239	T > C	MT-DLOOP				1	240
		2501	C > T	MT-RNR2			rRNA	1	1
P062	H3at1	6182	G > A	MT-CO1	93			1	64
		6908	T > C	MT-CO1	335			1	6
		13105	A > G	MT-ND5	257	I > V	29	2	1503
P064	T2c1d1	8475	C > T	MT-ATP8	37	P > L	**58 (C15)**	1	1
		11928	A >G	MT-ND4	390	N > S	46	1	57
		13056	C > T	MT-ND5	240			1	13
		14544	G > A	MT-ND6	44			2	56
P065	H56	5046	G > A	MT-ND2	193	V > I	29	1	401
		9182	G > A	MT-ATP6	219	S > N	46	2	21
		16270	C > T	MT-DLOOP				2	1003
		16311	T > C	MT-DLOOP				3	3659
P066	A2	523-524	delAC	MT-DLOOP				11	
		7830	G > A	MT-CO2	82	R > H	29	1	20

Individual	Haplo-group	Nucleotide position	Base change	Gene	Amino acid position	Amino acid change	Grantham value*	Number of samples that present the variant	Frequency in GenBank (18363 sequences)
		14178	T > C	MT-ND6	166	I > V	29	1	473
		14755	A > G	MT-CYB	3			1	51
		14861	G > A	MT-CYB	39	A > T	58	1	46

Variants that are not defining the haplotype of the sample are shaded and variants never reported are bold in red.

* The Grantham value was retrieved from mtSNP database and reflects the physicochemical differences between the original and altered amino acid residues (Grantham et al, 1974). Grantham values larger than 50 are described as radical amino acid replacements and those with Grantham values less than 50, are considered as conservative replacements. In bold blue there is shown the predicted Grantham value and the class assigned by Align-GVGD.