The First complete Zoroastrian-Parsi Mitochondria Reference Genome: Implications of mitochondrial signatures in an endogamous, non-smoking population

Authors Names and Affiliations:

Villoo Morawala Patell*[1,2,3], Naseer Pasha[1&2], Kashyap Krishnasamy[1&2], Bharti Mittal[1&2], Chellappa Gopalakrishnan[1&2], Raja Mugasimangalam[2&4], Naveen Sharma[1&2], Arati-Khanna Gupta[1], Perviz Bhote-Patell[1], Sudha Rao[2&4], Renuka Jain[1&2], The Avestagenome Project®

[1]Avesthagen Limited, Bangalore, India

[2]The Avestagenome Project® International Pvt Ltd, Bangalore, Karnataka, India-560005

[3]AGENOME LLC, USA

[4]Genotypic Technologies Private Limited, Bangalore 560094

**Corresponding Author:*

Address correspondence to Villoo Morawala Patell, THE dry lab, Avesthagen Limited Yolee Grande, 2nd Floor, Pottery Road, Richard's Town, Bangalore, 560005, Karnataka, India, Email: villoo@avesthagen.com

Abstract:

The present-day Zoroastrian-Parsis have roots in ancient pastoralist migrations from circumpolar regions[1] leading to their settlement on the Eurasian Steppes[2] and later, as Indo Iranians in the Fertile Crescent[3]. From then, the Achaemenids (550 - 331 BC), and later the Sassanids (224 BC - 642 AD) established the mighty Persian Empires[2]. The Arab invasion of Persia in 642 AD necessitated the migration of Zoroastrians from Pars to India where they settled as Parsis and practiced their faith, Zoroastrianism. Endogamy became a dogma, and the community has maintained the practice since their arrival in India. Fire is the medium of worship[4] as it is considered pure and sacrosanct; Social ostracism practiced against smokers resulted in a non-smoking community, thus forming a unique basis for our study.

In order to gain a clearer understanding of the historically recorded migration of the Zoroastrian-Parsis, decipher their phylogenetic relationships and understand disease association to their individual mitochondrial genomes, we generated the first complete *de novo* Zoroastrian-Parsi Mitochondrial Reference Genome, AGENOME-ZPMS-HV2a-1. Phylogenetic analysis of additional 100 Parsi mitochondrial genome sequences, showed their distribution into 7 major haplogroups and 25 sub-haplogroups and a largely Persian origin for the Parsi community. We have generated individual reference genomes for each major haplogroup and assembled the Zoroastrian Parsi Mitochondrial Consensus Genome (AGENOME-ZPMCG V1.0) for the first time in the world.

We report 420 variants, specifically 12 unique mitochondrial variants in the 100 mitochondrial genome sequences compared with the revised Cambridge Reference Sequence (rCRS) standard.

Disease association mapping showed 217 unique variants linked to longevity and 41 longevity associated disease phenotypes across most haplogroups. Our results indicate none of the variants are linked to lung cancer. Mutational signatures, C>A, G>T transitions[36], linked to tobacco carcinogens were found at extremely low frequencies in the Zoroastrian-Parsi cohort.

Our analysis of gene-coding, tRNA and the D-Loop regions revealed haplogroup specific disease associations for Parkinson's, Alzheimer's, Cancers, and Rare diseases.

These disease signatures investigated in the backdrop of generations of endogamy, in the rapidly declining, endangered Zoroastrian-Parsi community of India, provides exceptional universal opportunity to understand and mitigate disease.

Keywords: Mitochondria, Haplotypes, Phylogeny, Human migration, Endogamous, Non-smoking, Longevity, Cancers, Neurodegenerative disorders, Rare Diseases, t-RNA, D-loop variants, Population genetics, Unique mitochondrial variants, Zoroastrian Parsis, Persia, Iran, India, Precision healthcare.

Abbreviations: mt DNA-Mitochondrial DNA ; rCRS-revised Cambridge Reference Sequence ; NGS Next Generation Sequencing ; ZPMS- Zoroastrian Parsi Mitochondrial Sequence ; ZPMRG-Zoroastrian Parsi Mitochondrial Reference Genome ; ZPMCG- Zoroastrian Parsi Mitochondrial Consensus Genome ; AD- Alzheimers Disease ; PD- Parkinsons Disease

Introduction :

The Burden of History- Travelogue of the Zoroastrian Mitochondrion

Zoroastrian-Parsis of India are followers of the ancient prophet Zarathushtra, claimed by the Greek historian Herodotus to have been born circa 6,450 BC[1]. Zarathushtra, advocated the first known monotheistic concept of one supreme intelligence termed Ahura Mazda - 'Majestic Creator'[2].

The ancient homeland of the present-day Zoroastrian-Parsis finds mention in their sacred Avestan text *Vendidad*, and the location indicated is the North Polar Arctic region[3]. Sanskrit scholar B G Tilak's study *Arctic Home in the Vedas* is also corroborated by Bennet, suggesting that the Indo-European culture originated in the Hyperborean Regions of Northern Siberia and the islands of the circumpolar regions[4].

Around 12000 years ago, this region suffered a natural calamity and became ice-clad[1] necessitating southward migrations of these pastoralist inhabitants, and by 4,000 BC the Indo Europeans took over the Eurasian Steppe[7].

From the late second to early first millennium BC, the Indo Europeans, mostly on the basis of religious worship, split with the Indo Aryans who moved further south and crossed the Hindu Kush, while the Indo Iranians (Medes, Persians, and Parthians) began populating the western portion of the Iranian plateau, close to the Alborz and Zagros Mountains and northern Mesopotamia to Southeast Anatolia, in what is called the Fertile Crescent where significant innovation in agriculture occurred[8].

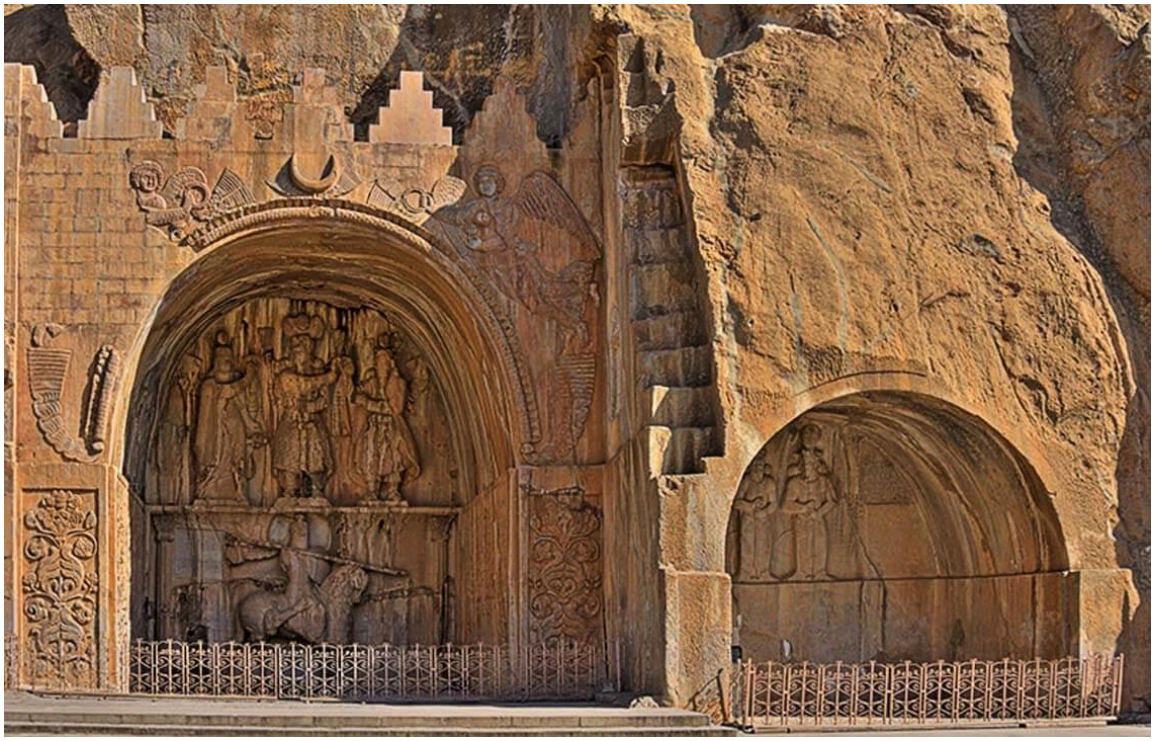
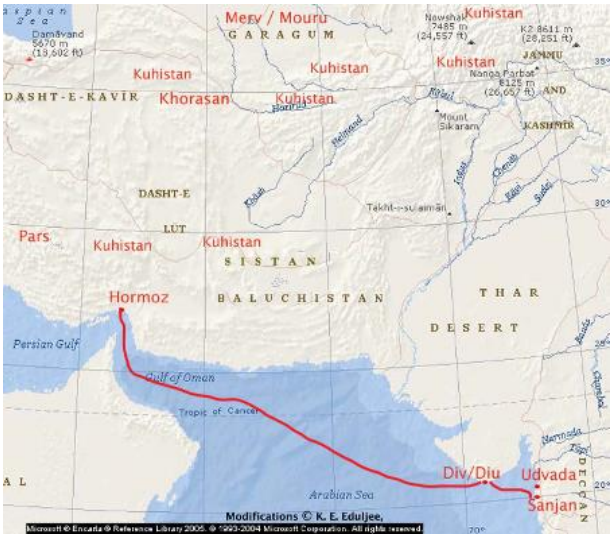


Image of ***Taq-e Bostan*** means "Arch of the Garden" or "Arch made by stone" , a site with a series of large rock reliefs from the era of the Sassanid Empire of Persia (Iran), carved around the 4th century CE. Image courtesy of Irandestination.com

In 550 BC, Cyrus the Great overthrew the leading Median rule and conquered the Kingdom of Lydia and the Babylonian Empire after which he established the Persian Zoroastrian Achaemenid Empire (the First Persian Empire), while his successors Dariush I (522-485 BC) Xerxes I, Artaxerxes and others extended its borders to encompass several countries across three continents, namely Europe, Africa and Asia. A second Zoroastrian dynasty of Sassanian Kings followed, who ruled Persia starting with Ardashir 1 (224 BC). It was the Golden age of the Persian Empire. During the time of Zoroastrian Achaemenid and Sassanid empires, Persia became a global hub of culture, religion, science, art, gender equality, and technology.

The Persians under Yezdezard III were defeated by the Arabs in two decisive battles – (Qadisiyah-636 AD and Nahavand – 642 AD) resulting in the fall of the Zoroastrian Persian Empire.

It was almost a hundred years later in the 8th century that a few boatloads of Zoroastrians left Paars and Khorasan from the port of Hormuz to sail south towards India. The boats first touched shore on Diu island on the west coast of India where the refugees stayed for around 19 years. The environment being non-conducive to progress, they once again set sail and arrived in Sanjan, Gujarat. Vijayaditya of the Chalukya dynasty (aka Jadi Rana) the ruler, hesitated to give refuge, but on being explained the principles of Zoroastrianism and observing the similarities with the Vedic religion, the Parsis were given refuge.



Map showing early migration of Parsis from Iran. Image courtesy Microsoft Encarta. Reference Library 2005

Endogamy became the norm to preserve their identity, and for the last 1300 years the community has maintained this practice[9,10]. Fire being the purest of all elements is considered sacred by Zoroastrian-Parsis. Strict measures are employed to maintain the purity of fire, hence the strict social ostracism practiced against smokers in the community.

Today, the Zoroastrian-Parsis, are a small community of <52000 in India (2011 Census, Govt of India). We present the genetic data of the conserved Zoarastrian-Parsi mitochondrion,

encapsulated in resilience of thousands of years of magnificent history: of struggles and overcoming them; of building something out of nothingness; of achievement gained with ethical standards; and philanthropy.

In recent decades, the analysis of the variability of maternally inherited mitochondrial DNA (mtDNA) has been commonly used to reconstruct the history of ethnic groups, especially with respect to maternal inheritance. The lack of genetic recombination in mtDNA, results in the accumulation of maternally inherited single nucleotide polymorphisms (variants). The accumulation of Variants along maternally inherited lineages results in phylogenetically traceable haplotypes[15] which can be used to follow maternal genealogies both historically and geographically. This approach has provided insightful findings into the origins and disease etiologies associated with another well documented endogamous European community: the Icelandic people (rev in [13]).

Human mtDNA (mitochondrial DNA) is a double stranded, circular (16,569 kb) genome of bacterial origin[16] primarily encoding vital subunits of the energy generating oxidative phosphorylation and electron transport chain (ETC) pathway that generates Adenosine Tri-Phosphate (ATP), the primary energy substrate of the eukaryotic cell. In addition, 22 tRNAs and 2 rRNAs are also encoded by the mtDNA[17].

In this study, our first aim was to gain a clear understanding and genetic impact of the historically recorded migration of the Zoroastrian-Parsis from Persia to India, and to link socio-cultural, ritualistic practices followed within the community over several millenia manifesting in genetic outcomes. To shed further light on the impact of migrations followed by integration into

communities, where ritual and social practices are strictly followed within communities and between communities resulting in specific traceable signatures.

Secondly, we have attempted to elucidate the genetic basis of commonly occurring diseases in this endogamous community. To address these issues, we generated the first complete *de novo* Zoroastrian-Parsi Mitochondrial Genome (AGENOME-ZPMS-HV2a-1) and used it to arrive at the mitochondrial haplotype specific Reference Genomes from a hundred Zoroastrian-Parsi individuals. Our study for the first time, has assembled the Zoroastrian Parsi Mitochondrial Consensus Genome (AGENOME-ZPMCG V 1.0) thereby creating the first Mitochondrial Consensus Genome for the Zoroastrian-Parsi community.

Our phylogenetic analysis confirmed that present day Zoroastrian-Parsis are closely related to Persians, and like most endogamous communities have comparatively lower genetic diversity and tend to be predisposed to several inherited genetic disorders[5,11]. They also possess longevity as a trait and are a long-living community[6] with lower incidences of lung cancer[12]. The study of the genealogic history of a close-knit community like the Parsis provides an unique opportunity, to understand the link between disease and social behaviour, thus providing the direction for population genetics as a basis for personalized healthcare.

Materials and Methods

Sample collection and ethics statement

One hundred healthy nonsmoking Parsi volunteers residing in the cities of Hyderabad-Secunderabad and Bangalore, India were invited to attend blood collection camps at the

Zoroastrian centers in their respective cities under the auspices of The Avestagenome Project™. Each adult participant (>18 years) underwent height and weight measurements and answered an extensive questionnaire designed to capture their medical, dietary and life history. All subjects provided written informed consent for the collection of samples and subsequent analysis. All health-related data collected from the cohort questionnaire were secured in The Avestagenome Project™ database to ensure data privacy. All procedures performed in this study involving human participants were in accordance with the ethical standards of the institution (Avesthagen Limited, Bangalore, India) and in line with the 1964 Helsinki declaration and its later amendments. This study has been approved by the Avesthagen Ethics Committee (BLAG-CSP-033).

Genomic DNA extraction

Genomic DNA from the buffy coat of peripheral blood was extracted using the Qiagen Whole Blood and Tissue Genomic DNA Extraction kit (cat. #69504). Extracted DNA samples were assessed for quality using the Agilent Tape Station and quantified using the Qubit™ dsDNA BR Assay kit (cat. #Q32850) with the Qubit 2.0® fluorometer (Life Technologies™). Purified DNA was subjected to both long-read (Nanopore GridION-X5 sequencer, Oxford Nanopore Technologies, Oxford, UK) and short-read (Illumina sequencer) for sequencing.

Library preparation for sequencing on the Nanopore platform

Libraries of long reads from genomic DNA were generated using standard protocols from Oxford Nanopore Technology (ONT) using the SQK-LSK109 ligation sequencing kit. Briefly, 1.5 µg of high-molecular-weight genomic DNA was subjected to end repair using the NEBNext Ultra II End Repair kit (NEB, cat. #E7445) and purified using 1x AmPure beads (Beckman Coulter Life

Sciences, cat. #A63880). Sequencing adaptors were ligated using NEB Quick T4 DNA ligase (cat. #M0202S) and purified using 0.6x AmPure beads. The final libraries were eluted in 15 µl of elution buffer. Sequencing was performed on a GridION X5 sequencer (Oxford Nanopore Technologies, Oxford, UK) using a SpotON R9.4 flow cell (FLO-MIN106) in a 48-hr sequencing protocol. Nanopore raw reads (fast5 format) were base called (fastq5 format) using Guppy v2.3.4 software. Samples were run on two flow cells and generated a dataset of ~14 GB.

Library preparation and sequencing on the Illumina platform

Genomic DNA samples were quantified using the Qubit fluorometer. For each sample, 100 ng of DNA was fragmented to an average size of 350 bp by ultrasonication (Covaris ME220 ultrasonicator). DNA sequencing libraries were prepared using dual-index adapters with the TruSeq Nano DNA Library Prep kit (Illumina) as per the manufacturer's protocol. The amplified libraries were checked on Tape Station (Agilent Technologies), quantified by real-time PCR using the KAPA Library Quantification kit (Roche) with the QuantStudio-7flex Real-Time PCR system (Thermo). Equimolar pools of sequencing libraries were sequenced using S4 flow cells in a Novaseq 6000 sequencer (Illumina) to generate 2 x 150-bp sequencing reads for 30x genome coverage per sample.

Generation of the d*e novo* Parsi mitochondrial genome (AGENOME-ZPMS-HV2a-1)

a) Retrieval of mitochondrial reads from whole-genome sequencing (WGS) data:

A total of 16 GB of raw data (.fasta) was generated from a GridION-X5 Nanopore sequencer for AGENOME-ZPMS-HV2a-1 from WGS. About 320 million paired-end raw reads were generated for AGENOME-ZPMS-HV2a-1 by Illumina sequencing.

Long Nanopore reads (. fastaq5) were generated from the GridION-X5 samples. The high-quality reads were filtered (PHRED score =>20) and trimmed for adapters using Porechop (v0.2.3). The high-quality reads were then aligned to the human mitochondrial reference (rCRS) NC_12920.1 using Minimap2 software. The aligned SAM file was then converted to a BAM file using SAMtools. The paired aligned reads from the BAM file were extracted using Picard tools (v1.102).

The short Illumina high-quality reads were filtered (PHRED score =>30). The adapters were trimmed using Trimgalore (v0.4.4) for both forward and reverse reads, respectively. The filtered reads were then aligned against a human mitochondrial reference (rCRS[21]) using the Bowtie2 (v2.2.5) aligner with default parameters. The mapped SAM file was converted to a BAM file using SAMtools, and the mapped paired reads were extracted using Picard tools (v1.102).

b) *De novo* mitochondrial genome assembly

Mapped reads were used for *de novo* hybrid assembly using the Maryland Super-Read Celera Assembler (MaSuRCA-3.2.8) tool. The configuration file from the MaSuRCA tool was edited by adding appropriate Illumina and Nanopore read files. The MaSuRCA tool uses a hybrid approach that has the computational efficiency of the de Bruijn graph methods and the flexibility of overlap-based assembly strategies. It significantly improves assemblies when the original data are augmented with long reads. AGENOME-ZPMS-HV2a-1 was generated by realigning the mapped mitochondrial reads from Illumina as well as Nanopore data with the initial assembly.

Confirmation of Variants in the *de novo* Parsi mitochondrial genome using Sanger sequencing

To validate the *de novo* Parsi mitochondrial sequence, AGENOME-ZPMS-HV2a-1, selected variants were identified and subjected to PCR amplification. Genomic DNA (20 ng) was PCR amplified using LongAmpTaq 2X master mix (NEB). The PCR amplicons of select regions were subjected to Sanger sequencing and BLAST analysis to confirm the presence of eight Variants using primers listed in Supplemental Table 1.

Generation of the Zoroastrian Parsi mitochondrial consensus genome (AGENOME-ZPMCG-V1.0) and Parsi haplogroup-specific consensus sequences

a) Retrieving mitochondrial reads from 100 Parsi whole-genome sequences

The whole-genome data from 100 Parsi samples were processed for quality assessment. The adapters were removed using the Trimgalore 0.4.4 tool for paired end reads (R1 and R2), and sites with PHRED scores less than 30 and reads shorter than 20 bp in length were removed. The processed Illumina reads were aligned against a human mitochondrial reference sequence (rCRS[18], NC_012920.1) using the Bowtie 2 (version 2.4.1) aligner with default parameters. Mapped reads were further used for the *de novo* assembly using SPAdes (version 3.11.1) and Velvet and IVA (version 1.0.8). Comparison of the assembly and statistics were obtained using Quast (version 5.0.2). The assembled scaffolds were subjected to BLASTn against the NCBI non-redundant nucleotide database for validation.

b) Variant calling and haplogroup classification

Sequencing reads were mapped to the human mitochondrial genome (rCRS[21]) assembly using the MEM algorithm of the Burrows–Wheeler aligner (version 0.7.17-r1188) with default parameters. Variants were called using SAMtools (version 1.3.1) to transpose the mapped data in a sorted BAM file and calculate the Bayesian prior probability. Next, Bcftools (version 1.10.2) was used

to calculate the prior probability distribution to obtain the actual genotype of the variants detected. The classification and haplogroup assignment were performed for each of the 100 Parsi mtDNAs after variant calling and after mapping reference and alternate alleles to the standard haplogroups obtained from MITOMAP (**Appendix 4**).

c) Haplogroup-based consensus sequence

Ninety-seven of 100 full-length Parsi mtDNA sequences were segregated based on haplogroups and separately aligned using the MUSCLE program to obtain the multiple sequence alignments. The Zoroastrian Parsi Mitochondrial Reference Genome (ZPMRG) and the Parsi haplogroup-specific consensus sequences were generated after calculation of the ATGC base frequency by comparison of the nucleotides in an alignment column to all other nucleotides in the same column called for other samples at the same position. The highest frequency (%) was taken to build seven Parsi haplogroup ZPMRGs and the seven Parsi haplogroup-specific consensus sequences.

Phylogeny build and analysis

Ninety-seven of 100 full-length Parsi mtDNA sequences generated as described above were compared with 100 randomly chosen Indian mtDNA sequences derived from NCBI Genbank under the accession codes FJ383174.1-FJ 383814.1[22], DQ246811.1-DQ246833.1[23], and KY824818.1-KY825084.1[24] and from previously published data on 352 complete Iranian mtDNA sequences[25]. All mtDNA sequences were aligned using MUSCLE software[26] using the “maxiters 2” and “diags 1” options, followed by manual verification using BioEdit (version 7.0.0). Following alignment, the neighbor-joining method, implemented in MEGAX[27], was employed to reconstruct the haplotype-based phylogeny. The neighbor-joining method was used because it is more efficient for large data sets[28].

Variant disease analysis

One hundred Parsi mitochondria sequences extracted from the WGS were uploaded into the VarDiG®-R search engine (https://vardigrviz.genomatics.life/vardig-r-viz/) on AmazonWeb Services. VarDiG®-R, developed by Genomatics Private Ltd, connects variants, disease, and genes in the human genome. Currently, the VarDiG®-R knowledgebase contains manually curated information on 330,000+ variants, >20 K genes covering >4500 phenotypes, including nuclear and mitochondrial regions for 150,000+ published articles from 388+ journals. Variants obtained from Parsi mitochondria were mapped against all the published variants in VarDiG®-R. Associations with putative diseases was ascertained for each variant through VarDIG®-R.

Seventeen tRNA SNP sites were identified in the 100 Parsi mitochondrial SNP data. The PON-mt-tRNA database[42] was downloaded to annotate the tRNA Variants for their impact and disease associations. This database employs a posterior probability-based method for classification of mitochondrial tRNA variations. PON-mt-tRNA integrates the machine learning-based probability of pathogenicity and the evidence-based likelihood of pathogenicity to predict the posterior probability of pathogenicity. In the absence of evidence, it classifies the variations based on the machine learning-based probability of pathogenicity.

For annotation of disease pathways associated with Variants, we employed MitImpact (https://mitimpact.css-mendel.it/) to predict the functional impact of the nonsynonymous Variants on their pathogenicity. This database is a collection of nonsynonymous mitochondrial Variants and their functional impact according to various databases, including SIFT, Polyphen, Clinvar, Mutationtester, dbSNP, APOGEE, and others. The disease associations, functional classifications,

and engagement in different pathways were determined using the DAVID and UNIPROT annotation tools.

Haplogroup and disease linkage

Principal component analysis (PCA) was performed to visualize the linkage of the haplogroup with disease. XLSTAT (Addinsoft 2020, New York, USA. https://www.xlstat.com) was used for statistical and data analysis, including PCA.

Data Accessibility:

The GenBank (http://www.ncbi.nlm.nih.gov/genbank) accession numbers for the 105 novel complete mtDNA sequences (97 ZPMS, 7 ZPMRG and 1 ZPMCG) reported in this paper are MT506242-MT506346. The raw reads for 97 ZPMS mitochondrial genome sequences have been deposited with BioProject ID: PRJNA636291. The SRA accession numbers for the 97 ZMPS: SRR11888826-SRR11888922.

Results

Assembly of the first complete Zoroastrian Parsi mitochondrial sequence, AGENOME-ZPMS-HV2a-1

The first complete *de novo* non-smoking Zoroastrian Parsi mitochondrial sequence, AGENOME-ZPMS-HV2a-1, was assembled from a healthy Parsi female sample by combining the sequence data generated from two next-generation sequencing (NGS) platforms using a protocol, as outlined in Materials and Methods. Our approach combines the sequencing depth and accuracy of short-read technology (Illumina) with the coverage of long-read technology (Nanopore). QC parameters

for mitochondrial reads, mitochondrial coverage, and X-coverage were found to be optimal, as seen in **Supplementary Figure 1**. The hybrid Parsi mitochondrial genome was assembled as a single contig of 16.6 kb (with 99.82% sequence identity), resulting in the consensus sequence for the *de novo* Parsi mitochondrial genome with 99.84% sequence identity to the revised Cambridge Reference Sequence (rCRS[21]).

Identification of 28 unique Variants in AGENOME-ZPMS-HV2a-1

The variants identified from both the Illumina and Nanopore data were considered to be significant for this *de novo* Zoroastrian Parsi mitochondrial genome, henceforth referred to as AGENOME-ZPMS-HV2a-1.

A total of 28 significant variants (i.e., variants) were identified by BLAST alignment between the Parsi mitochondrial hybrid assembly and the rCRS[21] (**Figure 1, Table 1**). To confirm the authenticity of the identified variants, we selected a total of 7 identified variants from the D-loop region and one SNP from the *COI* gene (m.C7028T, A375A) and subjected them to Sanger sequencing using primers. All 8 predicted variants were verified and confirmed for their presence in the consensus Parsi mitochondrial genome (**Figure 2**).

The majority (n=11) of the variants identified in the AGENOME-ZPMS-HV2a-1 were found in the hypervariable regions (HVRI and HVRII) of the D-loop. Of the remaining 17 variants, eight were found to represent synonymous variants, while four were in genes for 12S, 16S-rRNA (n=3) and tRNA (n=1) (**Figure 1**). The remaining 5 nonsynonymous variants were located in the genes for *ATPase6* (m8860G>A), *COIII* (m.9336 A>G), *ND4* (m.11016 G>A), and two in the *CytB* gene (m15326 A>G and m15792 T>C, (**Table 1**). Except for the *ATPase6* gene variant, which has been

found to be associated with hypertrophic cardiomyopathy in Iranian individuals[29], no associations were found in the published literature for these gene variants, and they need to be further investigated.

Given that the Zoroastrian Parsis are known to have originated in Persia and have practiced endogamy since their arrival on the Indian subcontinent, we wished to determine the mitochondrial haplogroup associated with the first complete Zoroastrian Parsi mitochondrial genome. We therefore compared the variants associated with ZPMS-HV2a-1 to standard haplogroups obtained from MITOMAP and determined the haplogroup to be HV2a (**Figure 1**). This haplogroup is known to have originated in Iran[25], suggesting Persian origins for this Parsi individual, based on maternal inheritance patterns.

Seven major haplogroups identified in 100 Zoroastrian Parsi individuals

Keeping in mind the endogamous nature of the Indian Parsis and to understand the extent of the diversity of the mitochondrial haplogroups in this population, we analyzed mitochondrial genomes from 100 consenting Parsi individuals. Our study had an equal representation of both genders, and 60% of the subjects were of age 30–59 (mean age 50±1.6) (**Figure 3**). Complete analysis of the variants in the 100 Parsi samples identified a total of 420 unique Variants (**Figure 4**, **Appendix 1**). QC analysis of the 100 mitochondrial genomes sequenced were found to be optimal: PHRED>30 (**Supplementary Figure 2**). Variant distribution in the coding region normalized to gene length showed the *ND6* gene has the highest number of variants (**Supplementary Figure 3**). The 100 Zoroastrian Parsi mitochondrial genomes were subjected to haplogroup analysis using haplogroup specific variant assignment matrix from MITOMAP (**Appendix 4**). The haplogroup

assignment based on variants classified the genomes into seven principal haplogroups (HV, U, T, M, A, F, and Z) and 25 sub-haplogroups were also identified within the principal haplogroups (**Figure 5**). The variant count across all sub-haplogroups ranged between 14-64 (**Figure 6A**). Analysis of the sub-haplogroups demonstrated that HV2a was the single largest representative sub-haplogroup within the Parsi population (n=14, n=9 females, n=5 males, (**Figure 6B**), that includes the AGENOME-ZPMS-HV2a-1.

The sub-haplogroup HV2a (n=14 subjects) contained 28 variants observed in the AGENOME-ZPMS-HV2a-1 are common across all 14 subjects. In total, the HV2a sub-haplogroup had 38 variants, with the highest number in the HVR II region (n=8). Coding region mutations constituted 20/38 variants, with equal distribution between synonymous (n=10) and non-synonymous substitutions observed for this sub-haplogroup (n=10). Among the coding regions, the largest number of Variants was found in the gene encoding *COI* (n=6, **Supplementary Figure 4A**). Four *COI* Variants distributed across all of the 14 subjects in the HV2a sub-haplogroup (m.6104 C>T, m.6179 G>A, m.7028 C>T, and m.7193 T>C) constitute synonymous mutations (amino acid change: F67F, M92M, A375A, and F430F, respectively). Two Variants (m.7080 T>C and m.7146 A>G), found to occur in one subject each in the sub-haplogroup HV2a, were nonsynonymous substitutions (F393L and T415A, respectively). Further analysis of rare Variants (occurring only in single subjects or n<8/14) showed their presence in the 16S-RNR2 gene (m.1883 G>A and m.1888 G>A), as well as the *COII*, *COIII* (m.8203 C>T and m.9540 T>C), and HVR I (m.16153 G>A and 16274 G>A) genes, which were synonymous substitutions in these coding genes, while we found nonsynonymous substitutions in the *COII* (m.7650 C>T; T22I), *ND5* (m.12358 C>T; T8A), and *CYTB* (m.14954 A>G; T70A) genes in our analysis. We found a variant in the gene

encoding for tRNA[R] at m.10410 T>C (n=14 subjects), but no mutations were observed in the D-loop region for the entire group under analysis.

The sub-haplogroup HV12b (n=1 subject) contained 17 Variants. HVR II harbors four Variants, while the coding genes together contain six Variants that encode three synonymous and three nonsynonymous substitutions. We observed Variants encoding nonsynonymous substitutions in this sub-haplogroup in *ATPase6* (m.8860 A>G; T112A), *ND5* (m.13889 G>A; C518Y), and *CYTB* (m.15326 A>G; T194A). Three Variants were found in 12S-RNR1, two Variants in 16S-RNR2. In the non-coding regions 5 variants were observed in HVRII, 1 in HVR I and 1 in the D-loop region (m.16519 T>C). No Variants were observed in the genes coding for tRNAs in the HV12b sub-haplogroup.

The 21 subjects analyzed that fell into the U haplogroup consisted of four sub-haplogroups U1a (n=1), U4b (n=11), U2e (n=3), and U7a (n=6). The U1a sub-haplogroup contained 44 Variants distributed across 19 positions in the mitochondrial genome. Twenty-one Variants were observed in the coding region (17 synonymous, 4 nonsynonymous). *ND5*, containing a coding region, contains six Variants, the most for any position within the U1a haplogroup. All *ND5* Variants coded for synonymous substitutions, while nonsynonymous substitutions were observed for *ND2* (m.4659 G>A; A64T), *ATPase6* (m.8860 A>G; T112A), and *CYTB* (m.14766 C>T; T7I and m.15326 A>G; A190T). 21/44 variants fell within coding genes, while the rest were distributed across HVR I (n=4 Variants), HVR II (n=3 Variants), HVR III (n=5 Variants), 12S-RNR1 (n=2 Variants), 16S-RNR2 (n=4 Variants), the D-loop region (n=1 SNP), and control regions (n=2 Variants). Two Variants were found in regions coding for tRNA[D] and tRNA[L:CUN].

The U4b sub-haplogroup is the most common sub-haplogroup among the U haplogroup in our analysis. In all, 64 Variants were observed for the U4b sub-haplogroup, with most of the variants (n=20) found in the gene encoding 16S-RNR2 (**Supplementary Figure 4B**). Twenty-one Variants were found in coding regions (14 synonymous and 7 nonsynonymous substitutions), with the highest number seen in the gene coding for *COI* (n=6 Variants). Five of six Variants coded for synonymous substitutions, while m.6366 G>A coded for a nonsynonymous substitution (V155I). Three Variants were found in the gene encoding *CYTB* and were distributed across all subjects (n=11) in the U4b sub-haplogroup. All three encoded nonsynonymous substitutions, m14766 C>T (T7I), m.15326 A>G (T194A), and m.15693 T>C (M316T), and need to be further investigated. Four tRNA mutations were observed in this sub-haplogroup and one mutation in the D-loop region.

A total of 52 variants were observed across all samples in the U7a subgroup (**Supplementary Figure 4B**). Twenty-seven Variants were found in noncoding regions, 12S-RNR1, 16S-RNR2, and the D-loop region. Twenty-five Variants were found in the coding region (17 synonymous and 8 nonsynonymous substitutions), with 17/25 distributed among the ND genes coding for *ND1–6*. *ND5* (n=6 Variants) encodes five synonymous mutations, with a nonsynonymous mutation observed at m.14110 T>C (F592L, in 4/6 subjects).

A total of 55 Variants was observed for U2e, with the majority (n=33 Variants) falling in the noncoding regions (HVRI-III and D-loop) and the 12S-RNR1, 16S-RNR2, and tRNA genes. Twenty-two Variants fell within the coding region (15 synonymous and 7 nonsynonymous substitutions), of which 8 fell in the ND gene complex (four *ND2*, four *ND5*) and four in the *CYTB* gene. While all the Variants in the *ND2* and *ND4* genes are synonymous substitutions, all the

Variants in the *CYTB* gene encoded nonsynonymous mutations (m.14766 C>T; T7I in 3/3 subjects, m.15326 A>G; T194A in 3/3 subjects; m.14831 G>A; A29T and m.15479 C>T; F245L, both in 1/3 subjects).

Five subjects in our analysis (n=100) fell within the T haplogroup. We found four sub-haplogroups within this haplogroup (T1a, 2 subjects: T2b, T2i, and T2g, with 1 subject each). Our analysis indicated a total of 39 Variants (**Supplementary Figure 4C**) for T1a, with 21/39 Variants found in noncoding regions, including 12S-rRNA, 16S-rRNA, tRNAs, and control regions, including the D-loop. Eighteen Variants were observed in the coding region, with the greatest number occurring in the *CYTB* gene (n=5 Variants). Three Variants within the *CYTB* gene coded for nonsynonymous mutations, including m.14776 C>T, m.14905 G>A, and m.15452 C>A, coding for T7I, T194A, and L236I substitutions, respectively.

The T2b, T2g, and T2i sub-haplogroups contained 35, 42, and 34 Variants, respectively, in total. We found that *CYTB* contained the majority of the Variants found in the coding regions in these sub-haplogroups, except for the T2i group in which the *CYTB* Variants (n=5) constituted the majority of the Variants found in coding and noncoding regions of the genome. Two Variants, m.14766 C>T and m.15326 A>G, seen in all three groups code for nonsynonymous substitutions, and m.15452 C>A was seen in T2g and T2i and codes for a nonsynonymous mutation. Single mutations were seen for m.15497 G>A and m.14798 T>C and code for nonsynonymous substitutions and need further investigation.

The A haplogroup in our study consists of the sub-haplogroup A2v (n=3 subjects). The subjects in the A2v sub-haplogroup had a total of 17 Variants (**Supplementary Figure 4D**) distributed across the mitochondrial genome. Twelve of seventeen Variants were found in the noncoding regions (HVR I, II) and in the 12S rRNA and 16S rRNA genes. Five Variants were distributed in the coding region across *ND2* (m.4769 A>G and m.6095 A>G), *ATPase6* (m.8860 A>G), *ND4* (m.11881 C>T), and *CYTB* (m.15326 A>G). Two nonsynonymous substitutions were observed in the *ATPase6* and *CYTB* genes that need further investigation.

F1g (n=1 subject) is a sub-haplogroup, along with Z1a (n=1 subject). A total of 33 and 32 Variants, respectively, were identified in these groups. Nine *CYTB* Variants were observed in total for both groups. Two encoded nonsynonymous substitutions, m.14766 C>T (T7I) and m.15326 A>G (T194A), while the seven other Variants resulted in synonymous mutations. Variants for *ND4L* are seen only across Z1a and F1g, with the m.10609 T>C SNP in F1g resulting in a nonsynonymous shift (M47T), while the Z1a SNP resulted in a synonymous substitution (**Supplementary Figure 4D)**.

The M haplogroup (n=52 subjects) consists of 12 sub-haplogroups, the most number for a haplogroup in our study (**Supplementary Figure 4E)**. M30d is the sub-haplogroups with the highest number of subjects in the M haplogroup (n=11 subjects). Fifty-one Variants were identified in this sub-haplogroup in total, of which 28 Variants were seen in the noncoding regions (HVR I, II, III), the D-loop region, and the 12S-RNR1 and 16S-RNR2 genes. The remaining 23 Variants were part of the coding region within *CYTB* (n=8 Variants) and *ND4* (n=5 Variants) and formed a majority. Nine of thirteen Variants in *CYTB* and *ND4* code for synonymous substitutions, while

four Variants in *CYTB* resulted in nonsynonymous substitutions (m.14766 C>T; T7I, m.15218 A>G; T158A, m.15326 G>A; T194A, and m.15420 G>A; A229T).

M39b (n=10 subjects) is one of the largest sub-haplogroups, and a total of 59 Variants were seen for this sub-haplolgroup. The noncoding regions, 12S, 16S, and control regions, together constitute 33/59 of the Variants. Of the remaining 26 Variants, the 5 Variants in the *CYTB* complex constitute the greatest number, while the ND gene complex accounts for 12 Variants (2 *ND1*, 1 *ND2*, 2 *ND3*, 2 *ND4*, 3 *ND5*, and 2 *ND6*). Of the nine remaining Variants, six are seen in the *COI*, *II*, and *III* genes (two each), while three Variants are found in the *ATPase6* gene.

The M2 sub-haplogroup consists of M2a (n=2 subjects) and M2b (n=1 subject). A total of 110 Variants was observed in total for M2a and M2b **(Supplementary Figure 4E).** In M2a, 23/53 Variants occurred in noncoding regions (HVR I, II, III), the 12S-RNR1 and 16S-RNR2 genes, the control region (OL), and the D-loop region. Thirty Variants occurred in the coding regions, making this one of the sub-haplogroups in which Variants in the coding region outnumber the Variants in the noncoding region. *CYTB* harbors seven Variants, followed by three Variants in *ND4* and three Variants in *ATPase8*, *ATPase6*, and *COI*. A total of 55 Variants was observed for M2b, in which 31/55 Variants occurred in the noncoding regions. Twenty-four Variants were observed in genes coding for COI, III; *ND1,2,3,4,5; ATPase6,8*; and *CYTB*. The six Variants in *CYTB* constitute the greatest number of Variants in the coding region. The M2a/b sub-haplogroup is also conspicuous by the presence of Variants in the *ATPase8* gene, which is not observed in any sub-haplogroup besides U4b. The complete distribution of the Variants across all the sub-haplogroups is presented in **Table 2**.

The M3a sub-haplogroup (n=8 subjects) consists of 38 variants, with 12/38 variants in the HVR I, II, III, D-loop regions **(Supplementary Figure 4E).** 19/38 variants were observed in the protein coding regions, with the most variants in this region occurring in *CYTB* (n=5). We found 15 coding for synonymous substitutions and 5 for non-synonymous variants (Supplementary Figure 4E)

M52b sub haplogroup (n=9 subjects) contained a total of 90 variants. 29/90 variants were observed in HVR I, II, III and the D-loop **(Supplementary Figure 4E).** 31 variants were observed for protein coding genes. *CYTB* (n=9 variants) contains the most variants for this region. 2 variants were found in t-RNA coding genes. 22 variants coded for synonymous substitutions while 9 variants coded for non-synonymous substitutions.

M24a subhaplogroup (n=8 subjects) contains a total of 48 variants, 12/48 are seen in HVR I, II, III and D-loop **(Supplementary Figure 4E).**. 22/48 are found in protein encoding genes with the most on *CYTB* (n=5 variants). 13 synonymous variants and 7 non-synonymous variants are seen in this sub-haplogroup. The rest of the variants are seen in 12S, 16S-rRNA. No variants for t-RNA genes were observed in this sub-haplogroup.

M27b (n=1 subject) has a total of 41 variants **(Supplementary Figure 4E).** 16/41 are seen in HVR I, II, III and the D-loop. 22/41 variants are seen in protein encoding genes with the highest variant count in *CYTB* (n=6 variants). 14 synonymous and 8 non-synonymous variants are observed for this sub-haplogroup and 1 variant for t-RNA coding gene.

M4a (n=1 subject) contains a total of 40 variants. 15/40 variants are seen in the non-coding regions of HVRI, II, III and D-loop **(Supplementary Figure 4E).** 21 variants are seen in the protein coding region with *CYTB* gene (n=5 variants) containing the highest variant count. Like M27b, M4a contains 14 synonymous and 7 non-synonymous variants and 1 variant on the t-RNA coding gene.

A total of 45 variants was seen in M5a sub-haplogroup (n=2 subjects) **(Supplementary Figure 4E).** 19/45 was seen in protein coding genes with *CYTB* (n=7 variants) representing the highest variants in the protein coding region. 13 variants code for synonymous substitutions while 6 code for non-synonymous variants. 1 variant is observed for a t-RNA coding gene.

M35b sub-haplogroup (1 subject) contains a total of 40 variants (**Supplementary Figure 4E)**. 15/40 variants are seen in HVR I, II, III and D-loop and 20/40 variants are found in protein encoding regions with the most variants observed in *CYTB* gene (n=5 variants). 14 code for synonymous substitution while 7 code for non-synonymous substitutions. 1 variant is observed for a t-RNA coding gene.

M33a sub-haplogroup (n=1 subject) contains 39 variants (**Supplementary Figure 4E)**. 15/39 variants are observed in HVR I, II, III and D-loop, 19/39 variants are seen in the protein coding region, with the highest count seen for *CYTB* (n=5 variants) for this region. 12 are synonymous and 7 are non-synonymous substitutions.1 variant for t-RNA coding gene is also observed in this sub-haplogroup. This haplogroup is unique amongst the 25 sub-haplogroups owing to the presence of a variant (m.8562 C>T) at *ATPase6/8* gene.

Phylogenetic analysis of the Parsi mitochondrial haplotypes with those of Iranians and Indians

To further investigate the substructure of the major haplogroups identified in the Parsi cohort, a comparative analysis of haplotypes from 452 complete mtDNA sequences, including 352 Iranians[25] and 100 Indian mitochondrial genome sequences, was undertaken. The rationale for selection of these two populations centered around the ancestral migration patterns of the Parsis of India[30]. This grouping also complements the model of the Parsi origin stemming from the ancient Iranian plateau[31].

Analysis of the haplogroups identified in the Parsis compared with the Iranians, of whom the Persians (n=180) and the Qashqais (n=112) were the most frequent representatives, demonstrated that a) all seven Parsi haplogroups were found within the Iranian haplogroup set and b) a marked lack of haplogroup diversity was observed in the Parsi datatset (n=7 principal haplogroups) compared with the Persians and Qashqais (n=14 principal haplogroups, **Figure 7A, B**). The reason for this lack of haplotype diversity likely lies in the practice of endogamy, which has been strictly adhered to in the Parsi community for centuries, following their arrival from the Iranian plateau. Contemporary populations of Iranians in the Iranian plateau represent diverse haplogroupings, possibly due to admixture following political upheavals in the region after the departure of Parsis from ancient Iran around 745 AD[31].

The presence of the predominantly Eurasian mtDNA haplotypes HV, T, and U in our study cohort was remarkable, given that Parsis have resided on the Indian subcontinent for over 1200 years.

While the majority of Parsis with M haplogroups can be linked to Persian descent, 2 sub-haplogroups (M2a, M2b) and 1 subject from M30d (n=4 subjects in total) were found to be related to relic tribes of Indian origin within the M haplogroups in our analyses.

A detailed phylogenetic clustering of the Parsis to establish more precise ethnic relationships was next undertaken. Our analysis revealed that the Parsis predominantly clustered with populations from Iran (Persians and people of Persian descent, **Figure 8A, 8E**), and the most common HV group showed that all Parsis in the HV2a tree (n=14) clustered with Persians and Qashqais (neighbour-joining tree weight > 0.72/72% (**Figure 8A and Table 3**), while the single Parsi in the HV12b (n=1) haplotype demonstrated a strong association with other Iranian ethnicities, including the Khorasani and Mazandaranis, in addition to the Qashqai and Persians **(Table 3).**

A total of 20 Parsi individuals in the U macro-haplogroup were found to fall into four subclades, U7a (n=6), U2e (n=3), U4b (n=10), and U1a (n=1), with the highest representation in U4b and U7a (**Figure 8B**). Phylogentic analysis demonstrated that the Parsis in the U haplogroup cluster with the Persians most frequently, while a few cluster with Kurds, Armenians, Mazandarani, Azeris, and Khorasanis, who all claim descent from Mesopotamia and the older Persian empire (https://journals.openedition.org/asiecentrale/480). Among the U haplogroup, U4b and U7a (the dominant branch of U7) haplotypes are distributed throughout the Near East and South Asia[24] with subclades specific to Central Asia in the Volga-Ural region[33], Mediterranean, and Southeast Europe, with lower frequencies in populations around the Baltic Sea, such as in Latvians and Tver Karelians[33]. Haplogroup U2 harbors frequency and diversity peaks in South Asia, whereas its U2d and U2e subclades are confined to the Near East and Europe[24].

The T haplogroup in the Parsi cohort was found to consist of T1a, T2g, T2i, and T2b, with an even distribution of samples across the subgroups (n=2, 1, 1, 1, respectively). Similar to the haplogroups HV and U, the Persians and Qashqais form the largest ethnic denomination associated with the Parsis with respect to the T haplogroups (>60%, **Figure 8C**). Five Parsi individuals of the haplogroups A2v (n=3), F1g (n=1), and Z1a (n=1) were observed to be phylogentically related to Persian, Kurd, Turkmen, and Iranian ethnicities, further attesting to their origin in the Iranian plateau (**Figure 8C**). The T haplogroup is also well distributed in Eastern and Northern Europe, as well as in the Indus Valley and the Arabian Peninsula. Younger T subclades are reported to have expanded into Europe and Central Asia during the Neolithic transition[34]

Unlike the HV, U, and T haplogroups, within which Parsi's cluster closely with Persians, Parsis harboring the M haplogroup appear to demonstrate more diversity in their mitochondrial genomes. This study showed the following breakdown: 8/12 M sub-haplogroups of the 29 Parsi M haplotypes (M24a [n= 8], M33a [n=1], M5a [n=2], M4a [n=1)], M3a [n=7], M52b [n=8], M27b [n=1], and M35b [n=1]) clustered with the Persians, Qashqais, Azeris of Iranian ethnicity, and others of Persian descent (**Figure 8D, Table 3**). Only two sub-haplogroups in our study (M2a and M2b [n=21], M30d [n=1], **(Figure 8D)** clustered more closely with relic tribes of Indian origin. Our phylogenetic analyses further showed that 19 Parsi individuals belonging to the M30d (n=10) and M39d (n=9) haplogroups did not cluster either with Indian or Iranian ethnic groups (**Figure 8D**) but remained clustered within their own subgroups.

Outgroup sampling is of primary importance in phylogenetic analyses, affecting ingroup relationships and, in placing the root, polarizing characters. Accordingly, we used AGENOME-

OUTGROUP-Y2b to root the phylogenetic tree. AGENOME-OUTGROUP-Y2b did not associate with the Zoroastrian-Parsis, Indians and Iranians attesting to the robustness of the method employed for phylogenetic analysis (**Figure 8E**, black line)

Assembly of the Zoroastrian Parsi mitochondrial consensus genome (AGENOME-ZPMRG-V1.0) and Parsi haplogroup-specific reference sequences

The Parsis of India are a nonsmoking, long-living community despite the prevalence of many genetic disease manifestations. This prompted us to generate a Parsi-specific mitochondrial consensus genome to better understand the nuances of disease and wellness in this unique community. Considering this goal, we classified the Parsi mitochondrial genome based on the seven identified major haplogroups, HV, M, U, T, A, F, and Z. The haplogroup-specific Parsi mitochondrial sequences were aligned, and a consensus call for each nucleotide was made based on the maximal frequency of a base called at each position in the mtDNA genome sequence (**Appendix 2**).

Using this approach, we derived the Zoroastrian Parsi mitochondrial reference sequences for each haplogroup: AGENOME-ZPMRG-HV-V1.0 (n=15 sequences), AGENOME-ZPMRG-U-V1.0 (n=20 sequences), AGENOME-ZPMRG-T-V1.0 (n=5 sequences), AGENOME-ZPMRG-M-V1.0 (n=52 sequences), AGENOME-ZPMRG-A2v-V1.0, AGENOME-ZPMRG-F1a-V1.0, and AGENOME-ZPMRG-Z-V1.0 (**Table 4**). Additionally, using all 100 Parsi mitochondrial genomes generated in this study (see Materials and Methods), we built the first standard Zoroastrian Parsi mitochondrial consensus genome (AGENOME-ZPMCG-V1.0). The consensus Parsi mtDNA sequence was found to have 31 unique Variants (**Table 5**), of which five Variants (A263G, A750G, A1438G, A4769G, and A15326G) were found to be common to the reference sequences of all

seven haplogroups considered (**Table 5**). While the number of Variants unique to each of the seven haplogroups ranged from 11 to 33, haplogroup M did not appear to have any unique Variants when compared with the overall consensus sequence, AGENOME-ZPMRG-V1.0. The utility of this newly generated reference standard could be found in the accurate mitochondrial-based analyses involving the global Zoroastrian Parsi population as well as for individuals of Western Asian, Indo-European and Indian origin.

Disease-specific associations of mtDNA variants predict the prevalence of commonly occurring diseases in the non-smoking Parsi cohort

As demonstrated in this paper (**Figure 7B**), the practice of intermarriage has likely restricted the genetic diversity of the Parsis, as measured by the paucity of haplogroups in our cohort compared with the Persian and Qashqai populations, possibly contributing to a number of autosomal recessive and other genetic diseases. In previous studies, Parsis were found to be disproportionately affected with certain diseases, such as prostate and breast cancers[5,11], Parkinsons disease (PD), and Alzheimers disease (AD). However, the Parsis are also considered to be a long-living community[6] with lower incidences of lung cancer[12].

In order to determine whether diseases known to be prevalent in the Parsi community could in fact be predicted by association using the collective mitochondrial variants discovered in this study, we first analyzed variants identified in tRNA genes in the mitochondrial genome that have previously been implicated in rare and degenerative diseases. We found a total of 17 tRNA-associated variants, with a pathogenic variant (amino acid change: G1644A) implicated significantly in LS/HCM/MELAS, a genetically inherited mitochondrial disease[47]. We also found

a total of six tRNA mutations associated with non-syndromic hearing loss, hypertension, breast/prostate cancer risk, and progressive encephalopathies in the analysis of our 100 Parsi individuals **(Table 6).**

Further analysis of the nucleotide transitions and transversions that constitute the 420 variants revealed that the mutational signatures (C>A and G>T) found in tobacco smoke-derived cancers[36] were found at an extremely low frequency (<6% compared to other mutational signatures) on both the H and L strands of the mitochondrial genomes of the Parsi population (**Figure 9**), who refrain from smoking due to their religious and social habits.

Variant analysis

Furthermore, we found that the 420 variants analysed were associated with 41 diseases. SNP disease-association analysis revealed that Parkinson's disease is highly associated with our variants (178 Variants, **Supplementary Figure 5**). Other neurodegenerative diseases, rare diseases of mitochondrial origin, and cardiovascular and metabolic diseases associated with the variants in our study were also predicted (**Supplementary Figure 5)**.

While a predisposition to 41 diseases were spread across 25 sub-haplogroups, many diseases were found to be recurring across haplogroups, totalling 188 diseases (**Figure 10A**). Haplogroup U4b harbored 15 diseases associations, while the majority of M and T groups had five diseases (Figure 6B). Some of the mitochondrial rare diseases, such as mitochondrial encephalomyopathies, MELAS syndrome and cytochrome c oxidase deficiency were found to be associated with M2a and U1a, U4b and M2b sub-haplogroups respectively (**Figure 10B**).

Haplogroup and disease linkage

Since the 420 variants identified fell into 25 sub-haplogroups contributing to 41 diseases and conditions, Principal component analysis (PCA) showed the grouping of variants and haplogroups (**Figure 11**). Alzheimers disease, breast cancer, cardiomyopathies, and Parkinsons disease were represented in all the 25 sub-haplogroups (**Appendix 3**), and longevity was represented in 23 sub-haplogroups, with the exception of HV12b and U1a groups. Our tRNA pathogenicity analysis showed that the variability in tRNA was highest in the U, T, and M haplogroups compared with other haplogroups (**Table 6**).

Analysis of Variants in tRNA genes and the D-loop region in the mitochondrial genome

While most of the variants in mtDNA genome sequences do not affect mitochondrial function, unlike synonymous/neutral variants, nonsynonymous/non-neutral variants may have functional consequences, and their effect on mitochondrial metabolism may be strongly deleterious, mildly deleterious, or even beneficial. We thus analysed, a SNP dataset obtained from 100 Parsi subjects for nonsynonymous mutations and identified 63 such Variants located within different mitochondrial genes (**Figure 12**). Twenty of sixty-three variants were found in genes encoding *CYTB* (n=13) and *ND2* (n=7), followed by *ND5* and *ND1*. Disease-association analysis showed that these genes were implicated in the onset of neurodegenerative conditions like AD, PD, cancers of colorectal and prostate origin, metabolic diseases such as type 2 diabetes, and rare diseases such as LHON (*CYTB* and *ND2*), (**Figure 13, Figure 14**). Variants implicated in longevity were observed in our study and distributed across the *ND2* gene **(Figure 10B**). As observed earlier, we

found no association of the nonsynonymous variants in our data set that linked to lung cancer or a risk of lung cancer.

To understand the mitochondrial pathways affected by the variants in our study, we annotated the pathways associated with Variants with DAVID and UNIPROT and found that the major genes *CYTB* and *ND2* were implicated in pathways that include the mitochondrial respiratory complex (*COI/COII/COIII/COIV*), OXPHOS, and metabolic pathways implicated in mitochondrial bioenergetics. Critical disease-related pathways in Parkinsons disease, Alzheimers disease, and cardiac muscle contraction were also associated with *CYTB*- and *ND2*-specific Variants, which possibly explains the high incidence of these disease in the Zoroastrian-Parsi population (**Figure 15**).

A total of 87 variants, including 6 unique variants, were observed in the D-loop region across all 25 sub-haplogroups (n=100 subjects, **Table 2**). 74/100 Parsis in our study, were found to have the polymorphism m.16519 T>C that is associated with chronic kidney disease[43], an increased risk for Huntingtons disease, migraine headache, and cyclic vomiting syndrome[44] and schizophrenia and bipolar disorder[48] a. While six subjects of the M52 sub-haplogroup were found to have m.16525 A>G. The rest of the variants were found at m.16390 G>A (n=4 subjects) and m.16399 A>G, m.16401 C>T, and m.16497 A>G (all with n=1 subject each). Taken together, these results warrant a deeper investigation into the D-loop variants in the Zoroastrian-Parsi community.

Identification of unique, unreported variants from the 100 Parsi-Zoroastrian mitogenome analysis

We performed a comparative analysis of the 420 variants in the Zoroastrian-Parsi community with MITOMASTER[45], a database that contains all known pathogenic mtDNA mutations and common haplogroup polymorphisms, to identify unique Variants in our population, that are not reported previously. Our analysis showed the presence of 12 unique Variants distributed across 27 subjects that were not observed in MITOMASTER and additionally in the VarDIG®-R disease association dataset (**Figure 16**). These unique variants were observed across different gene loci. 12S-rRNA (2 variants), 16S-rRNA (5 Variants), 1 each at ND1, COII, COIII, ND4 and ND6. The SNP haplogroup association showed that they fell into 4 major haplogroups and 13 sub haplogroups; HV2a=1, M24a=4, M2a=1, M30d=3, M35b=1, M39b=2, M3a=1, M4a=1, M52b=4, M5a=1, T2b=1, U4b=6, U7a=1. Of the 12 variants identified, no disease associations were observed for on analysis with MITOMASTER and VarDIG®-R and needs to be further investigated.

Discussion

The first *de novo* Parsi mitochondrial genome, AGENOME-ZPMS-HV2a-1 (Genbank accession: MT506314) from a healthy, non-smoking female of haplogroup HV2a when compared with the revised Cambridge Reference Standard (rCRS) showed 28 unique variants. Upon extending our mitochondrial genome analyses to an additional 99 Parsi individuals, we found that 94 individuals separate into four major mitochondrial haplogroups, HV, U, T, and M, while 5 individuals belong to the rarer haplogroups A, F, and Z. The largest sub-haplogroup was found to be HV2a (n=14).

Due to the strict endogamy practiced by the Zoroastrian Parsis, their maternally inherited mitochondrial lineage seems to have remained aligned with those of their ancestors in Old Persia, prior to 642 AD. On comparison of the major mitochondrial haplogroups in our Parsi cohort with 352 Iranian[25] and 100 Indian mitochondrial genomes, we observed that the Zoroastrian-Parsi genomes are phylogenetically related to the Persians and Qashqais[25] in HV, T, U, F, A and Z haplogroups, those associated with peopling of western Europe, Central Asia and the Iranian plateau.

The haplogroup HV2, dated at 36–42 kya, most likely arose in Persia between the time of the first settlement by modern humans and the last glacial melt, and the subclade HV2a has a demonstrated Persian ancestry. HV12b, a branch of the HV12 clade, is one of the oldest HV subclades and has been found in western Iran, India, and sporadically as far as Central and Southeast Asia. It has strong associations with the Qashqais, who are Turkic-speaking nomadic pastoralists of southern Iran and who previously resided in the Iranian region of the South Caucasus[32,35]. The presence of these predominantly Eurasian mtDNA haplotypes, HV, T, U, F, A and Z in our Parsi cohort attests to their practice of endogamy, given that Parsis have resided on the Indian subcontinent for over 1300 years.

Despite the large grouping of the M haplogroup (the largest haplogroup in the Indian subcontinent [34]) in our Parsi cohort, phylogenetic analysis showed that 47/51 Parsis belonging to the M haplogroups in our study, cluster with the Persians, suggesting Persian descent, with a small minority of Parsis found to be related to relic tribes of India. This observation suggests minimal

gene flow from indigenous Indian females into the Parsi gene pool, as had been previously proposed[30].

Phylogenetic analysis also revealed that two Parsi M sub-haplogroups, M30d and M39b formed a unique cluster that needs further resolution.

We further present the first complete Zoroastrian Parsi Mitochondrial Consensus Genome (AGENOME-ZPMCG V1.0), built from the mitochondrial genomes of 100 non-smoking, Parsi individuals representing seven mitochondrial haplogroups. The need for the generation of such an ethnic-specific consensus genome, specifically for the Parsis, is self-evident for studies involving comparative analyses, designed to precisely understand patterns of maternally inherited mitochondrial DNA and aid in reconstructing the history and prevalent disease associations in this unique community.

We found that *CYTB* gene contained the maximum number of variants (n≥5) in the coding region of haplogroup M, besides having maximal representation in F1g, T, and HV12b. Haplogroups U, A2v, and Z1a showed dominance for the ND complex genes *ND5* and *ND2*, while the *COI* genes were the most highly represented in HV2a and U4b. Variants in the *CYTB* gene are associated with Alzheimers disease, diabetes mellitus, cognitive ability, breast cancer, hearing loss, and asthenozoospermia and associated with changes in metabolic pathways, cardiac contraction and rare diseases such as Huntington's disease, whereas the *ND2* and *ND5* variants were associated with prostate, ovarian cancer, rare mitochondrial neuronal diseases, such as LHON, cardiomyopathy, Alzheimers disease and Parkinsons disease.

tRNA disease-association analysis in our study showed that these genes were implicated in the onset of neurodegenerative conditions, such as Alzheimers disease, Parkinsons disease, cancers of colorectal and prostate origin, metabolic diseases, such as type 2 diabetes, and rare diseases, such as LHON (*CYTB* and *ND2*). The D-loop SNP analysis showed the prevalence (74/100 subjects) of the m.16519 T>C polymorphism, which has been implicated in chronic kidney disease[43], an increased risk for Huntingtons disease, migraine headache, and cyclic vomiting syndrome[44]. Taken together, these results warrant a deeper investigation into the tRNA and the D-loop variants in the Zoroastrian-Parsi community.

Consanguineous marriages amongst the Parsis has given rise to longevity[11] and number of associated diseases, including colon, prostate, breast cancers[9,10], Parkinson's disease (PD), Alzheimer's disease (AD), and lower incidences of lung cancer[12] . Interrogation of the 420 variants across seven haplogroups in the Parsi cohort using the VarDIG®-R database revealed that Parkinson's disease, known to be prevalent in the Parsi community[36], was predicted to have the highest prevalence with 178 of the 420 variants represented. Not surprisingly, longevity, which often co-occurs with Parkinsons disease, was also predicted to be highly prevalent in the Parsi cohort, but with a notable absence in the U1 sub-haplogroup, an interesting observation that warrants further investigation.

Analysis of additional disease associations revealed that Alzheimer's disease (also related to ageing), breast cancer, and cardiomyopathies[38,39,40], were predicted to be associated with all the 25 Parsi sub-haplogroups. Additionally, the observed low birth rate among Parsi could be predicted from the presence of variants associated with asthenozoospermia[37]; a condition associated with

reduced sperm motility.

It is noteworthy that previously published studies demonstrating lower rates of lung cancer amongst the Parsis[41], appears to hold true, given that no haplogroup in the Parsi cohort demonstrated a predicted predisposition to lung cancer. The lack of mitochondrial signatures for lung cancer in all haplogroups examined in this non-smoking Parsi population coupled with the low frequency of mutational signatures for tobacco smoke-derived cancers was not surprising, particularly since the Zoroastrian-Parsi venerate fire, and smoking would lie in gross violation of that sacred tenet. Other diseases predicted to occur at high frequencies in our analyses await further investigation in the Parsi community.

The Parsi haplogroup specific variant-disease association analysis has shed predictive light on the association of mitochondrial variants linked to longevity, neurodegenerative diseases, cancers of the colon, breast and prostate and low birth rate, among others; diseases that have been well documented to occur in the Parsi community. The Parsis thus represent a small but unique, non-smoking community where genomic disease signatures, both mitochondrial and nuclear, can be investigated in the backdrop of generations of endogamy thus providing exceptional opportunities to understand and mitigate disease.

Conclusion

We have generated the first *de novo* Zoroastrian Parsi Mitochondrial Reference Sequence (AGENOME- ZPMS-HV2a-1) and the Zoroastrian Parsi Mitochondrial Consensus Genome (AGENOME-ZPMCG V1.0). This newly generated reference standards will contribute in the

analysis of mitochondrial genomes of not only the Zoroastrian Parsi population but also other populations. We have also provided evidence that the Zoroastrian Parsis of India, through centuries of endogamy, have retained their Persian genetic heritage, distinct traits of longevity and associated diseases. We have shown the rôle of social habits in genetic signatures exemplified by the lack of mitochondrial variants associated with lung cancer.

In sum, The Parsi haplogroup specific variant-disease association analysis has shed predictive light on the association of mitochondrial variants linked to longevity, neurodegenerative diseases, cancers of the colon, breast and prostate and low birth rate, among others; diseases that have been well documented to occur in the Parsi community. The Parsis thus represent a small but unique, non-smoking community where genomic disease signatures, both mitochondrial and nuclear, can be investigated in the backdrop of generations of endogamy thus providing exceptional opportunities to understand and mitigate disease.

References

1. Mistry, R. K. *Glimpses of Parsi history, Insights Into The Zarathustrian Religion, p.20.*
2. Nariman, R. F. *The Inner Fire – Faith, Choice, and Modern Day Living in Zoroastrianism,* p. 20-21
3. The Vendidad: The Zoroastrian Book Of The Law Paperback – September 10, 2010. I, 1-2 & II, 5. Charles. F. Horne. ISBN-10: 1162910089; ISBN-13: 978-1162910086. Kessinger Publishing, LLC (September 10, 2010)
4. Bennet, J. G. The Hyperborean Origin of the Indo-European Culture, Journal Systematics. *J Syst.* **1**, (1963).
5. Jussawalla, D. J., Yeole, B. B. & Natekar, M. V. Histological and epidemiological

features of breast cancer in different religious groups in greater bombay. *J. Surg. Oncol.* (1981) doi:10.1002/jso.2930180309.

6. Jussawalla, D. J. The persistance of differences in cancer incidence at various anatomical sites 1300 years after immigration. *Recent Results Cancer Res.* (1975) doi:10.1007/978-3-642-80880-7_22.

7. Anthony, DW, (2007), The Horse, The Wheel, And Language. How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World, Princeton University Press. p. 9.

8. Alizadeh, A. The Rise of the Highland Elamite State in Southwestern Iran. Current. *Curr Anthropol.* **51**, 353–383 (2010).

9. Shroff Z, C. M. The potential impact of intermarriage on the population decline of the Parsis of Mumbai, India. *Demogr Res.* **25**, 545–564 (2011).

10. Karkal, M. Marriage among Parsis. *Demogr. India* **4**, 128 (1975).

11. Barnabas-Sohi, N. *et al.* Breast carcinoma in a high-risk population: Structural alterations in neu, int-2, and p-53 genes. *Breast Dis.* (1993).

12. Jussawalla, D. J. & Jain, D. K. Lung cancer in Greater Bombay: Correlations with religion and smoking habits. *Br. J. Cancer* (1979) doi:10.1038/bjc.1979.199.

13. Helgason, A., Sigurðardóttir, S., Gulcher, J. R., Ward, R. & Stefánsson, K. mtDNA and the origin of the Icelanders: Deciphering signals of recent population history. *Am. J. Hum. Genet.* (2000) doi:10.1086/302816.

14. Wallace, D. C. Mitochondrial DNA Variation in Human Radiation and Disease. *Cell* (2015) doi:10.1016/j.cell.2015.08.067.

15. Wallace, D. C., Brown, M. D. & Lott, M. T. Mitochondrial DNA variation in human evolution and disease. *Gene* (1999) doi:10.1016/S0378-1119(99)00295-4.

16. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The Origin and Diversification of Mitochondria. *Current Biology* (2017) doi:10.1016/j.cub.2017.09.015.

17. Garcia, I., Jones, E., Ramos, M., Innis-Whitehouse, W. & Gilkerson, R. The little big genome: The organization of mitochondrial DNA. *Front. Biosci. - Landmark* (2017) doi:10.2741/4511.

18. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: Implications for human health and disease. *Nature Reviews Genetics* (2015) doi:10.1038/nrg3966.

19. Bussard, K. M. & Siracusa, L. D. Understanding mitochondrial polymorphisms in cancer.*Cancer Research* (2017) doi:10.1158/0008-5472.CAN-17-1939.

20. Alston, C. L., Rocha, M. C., Lax, N. Z., Turnbull, D. M. & Taylor, R. W. The genetics and pathology of mitochondrial disease. *J. Pathol.* **241**, 236–250 (2017).

21. Andrews, R. M. *et al.* Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA [5]. *Nature Genetics* (1999) doi:10.1038/13779.

22. Chandrasekar, A. *et al.* Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian corridor. *PLoS One* **4**, e7447–e7447 (2009).

23. Rajkumar, R., Banerjee, J., Gunturi, H. B., Trivedi, R. & Kashyap, V. K. Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages. *BMC Evol. Biol.* **5**, 26 (2005).

24. Sahakyan, H. *et al.* Origin and spread of human mitochondrial DNA haplogroup U7. *Sci. Rep.* **7**, 46044 (2017).

25. Derenko, M. *et al.* Complete mitochondrial DNA diversity in Iranians. *PLoS One* (2013)

doi:10.1371/journal.pone.0080673.

26. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

27. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

28. Tamura, K., Nei, M. & Kumar, S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11030–11035 (2004).

29. Houshmand, M. *et al.* Is 8860 variation a rare polymorphism or associated as a secondary effect in HCM disease? *Arch. Med. Sci.* (2011) doi:10.5114/aoms.2011.22074.

30. Chaubey, G. *et al.* 'Like sugar in milk': Reconstructing the genetic history of the Parsi population. *Genome Biol.* (2017) doi:10.1186/s13059-017-1244-9.

31. López, S. *et al.* The Genetic Legacy of Zoroastrianism in Iran and India: Insights into Population Structure, Gene Flow, and Selection. *Am. J. Hum. Genet.* (2017) doi:10.1016/j.ajhg.2017.07.013.

32. Quintana-Murci, L. *et al.* Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *Am. J. Hum. Genet.* **74**, 827–845 (2004).

33. Shamoon-Pour, M., Li, M. & Merriwether, D. A. Rare human mitochondrial HV lineages spread from the Near East and Caucasus during post-LGM and Neolithic expansions. *Sci. Rep.* **9**, 14751 (2019).

34. Farjadian, S. *et al.* Discordant Patterns of mtDNA and Ethno-Linguistic Variation in 14 Iranian Ethnic Groups. *Hum. Hered.* **72**, 73–84 (2011).

35. Thangaraj, K. *et al.* In situ origin of deep rooting lineages of mitochondrial

Macrohaplogroup ‘M’ in India. *BMC Genomics* **7**, 151 (2006).

36. Alexandrov LB, Ju YS, Haase K, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016;354(6312):618‑622.

37. E. Ruiz-Pesini, A.C. Lapeña, C. Díez, E. Alvarez, J.A. Enríquez, M.J. López-Pérez Seminal quality correlates with mitochondrial functionality. Clin. Chim. Acta., 300 (2000), p. 97 105.

38. Fang, H., Shen, L., Chen, T. et al. Cancer type-specific modulation of mitochondrial haplogroups in breast, colorectal and thyroid cancer. BMC Cancer 10, 421 (2010).

39. Van der Walt JM, Dementieva YA, Martin ER, Scott WK, Nicodemus KK, Kroner CC, Welsh-Bohmer KA, Saunders AM, Roses AD, Small GW, Schmechel DE, Murali Doraiswamy P, Gilbert JR, Haines JL, Vance JM, Pericak-Vance MA**.** Analysis of European mitochondrial haplogroups with Alzheimer disease risk. Neurosci Lett. 2004 Jul 15; 365(1):28-32.

40. van Oven M, Kayser M Hum Mutat. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. 2009 Feb; 30(2): E386-94.

41. Balkrishna Bhika Yeole, AP Kurkure, SH Advani, Sunny Lizzy; An Assessment of Cancer Incidence Patterns in Parsi and Non Parsi Populations, Greater Mumbai. Asian Pacific Journal of Cancer Prevention, Vol 2, 2001; 293-298

42. Niroula A, Vihinen M. PON-mt-tRNA: a multifactorial probability-based method for classification of mitochondrial tRNA variations. Nucleic Acids Res. 2016;44(5):2020-2027. doi:10.1093/nar/gkw046

43. Chen JB, Yang YH, Lee WC, et al. Sequence-based polymorphisms in the mitochondrial D-loop and potential SNP predictors for chronic dialysis. PLoS One. 2012;7(7):e41125.

doi:10.1371/journal.pone.0041125

44. Zaki EA, Freilinger T, Klopstock T, et al. Two common mitochondrial DNA polymorphisms are highly associated with migraine headache and cyclic vomiting syndrome. Cephalalgia. 2009;29(7):719-728. doi:10.1111/j.1468-2982.2008.01793.x

45. Brandon MC, Ruiz-Pesini E, Mishmar D, et al. MITOMASTER: a bioinformatics tool for the analysis of mitochondrial DNA sequences. Hum Mutat. 2009;30(1):1-6. doi:10.1002/humu.20801.

46. Narasimhan VM, Patterson N, Moorjani P, et al. The formation of human populations in South and Central Asia. *Science*. 2019;365(6457):eaat7487. doi:10.1126/science.aat7487.

47. Menotti F, Brega A, Diegoli M, Grasso M, Modena MG, Arbustini E. A novel mtDNA point mutation in tRNA(Val) is associated with hypertrophic cardiomyopathy and MELAS. *Ital Heart J*. 2004;5(6):460-465.

48. Schulmann A, Ryu E, Goncalves V, et al. Novel Complex Interactions between Mitochondrial and Nuclear DNA in Schizophrenia and Bipolar Disorder. Mol Neuropsychiatry. 2019;5(1):13‐27. doi:10.1159/000495658

Contributions

VMP conceptualized and guided the experiments; NP, CG analysed the sequences, bioinformatics analysis, interpreted the results; RM, SR, NS co-ordinated wet-lab work flows and data analysis; NP, BM, KK analysed data, plotted graphs and figures; VMP, AKG, PB, KK, RJ drafted the manuscript with inputs from RM, SR, NS, CG. All authors reviewed the manuscript

All authors researched data for the article, made substantial contribution to discussion of content, and wrote, reviewed, and edited the manuscript before submission.

Sources of funding

The project was funded by a grant (GG-0005), Cancer risk in smoking subjects assessed by next generation sequencing profile of circulating free DNA and RNA) by the Foundation for a Smoke-Free World, New York, USA.

Competing interests

The authors declare no competing interests

Acknowledgements:

We thank the Foundation for Smoke Free World who is advancing global progress in smoking cessation and harm reduction for funding this project and Dr.Derek Yach for his support to this important project.

We would like to express our thanks to Prof. Dr. Partha. P. Mazumdar, National Institute of Bio Medical Genetics (NIBMG), Kolkata and Prof. Dr. Kumarasamy Thangaraj, Center for Cellular and Molecular Biology (CCMB), Hyderabad; for their leadership, expertise and coordination of sequencing our samples.

We would like to express our deep gratitude to the Zoroastrian Parsi community of India for their patience and cooperation at multiple instances.

Parzor Foundation for its partnership in outreach among the Zoroastrian-Parsi community in India and across the globe.

Dr.Sami Gazder, Kouser Sonnekhan and the The Avestagenome Project™ project team of yore.

Main Figures

Figure 1 : Identification of 28 variants in the de novo Parsi mitochondrial genome, AGENOME-ZPMS-HV2a-1

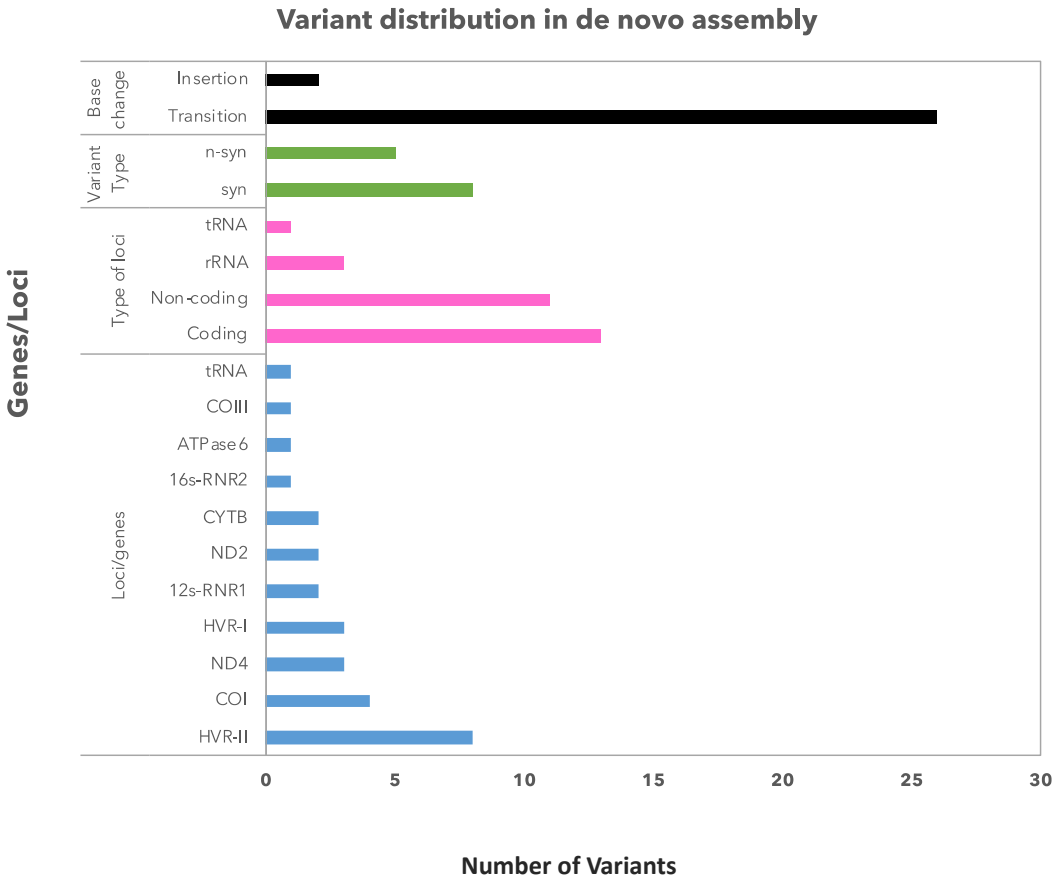
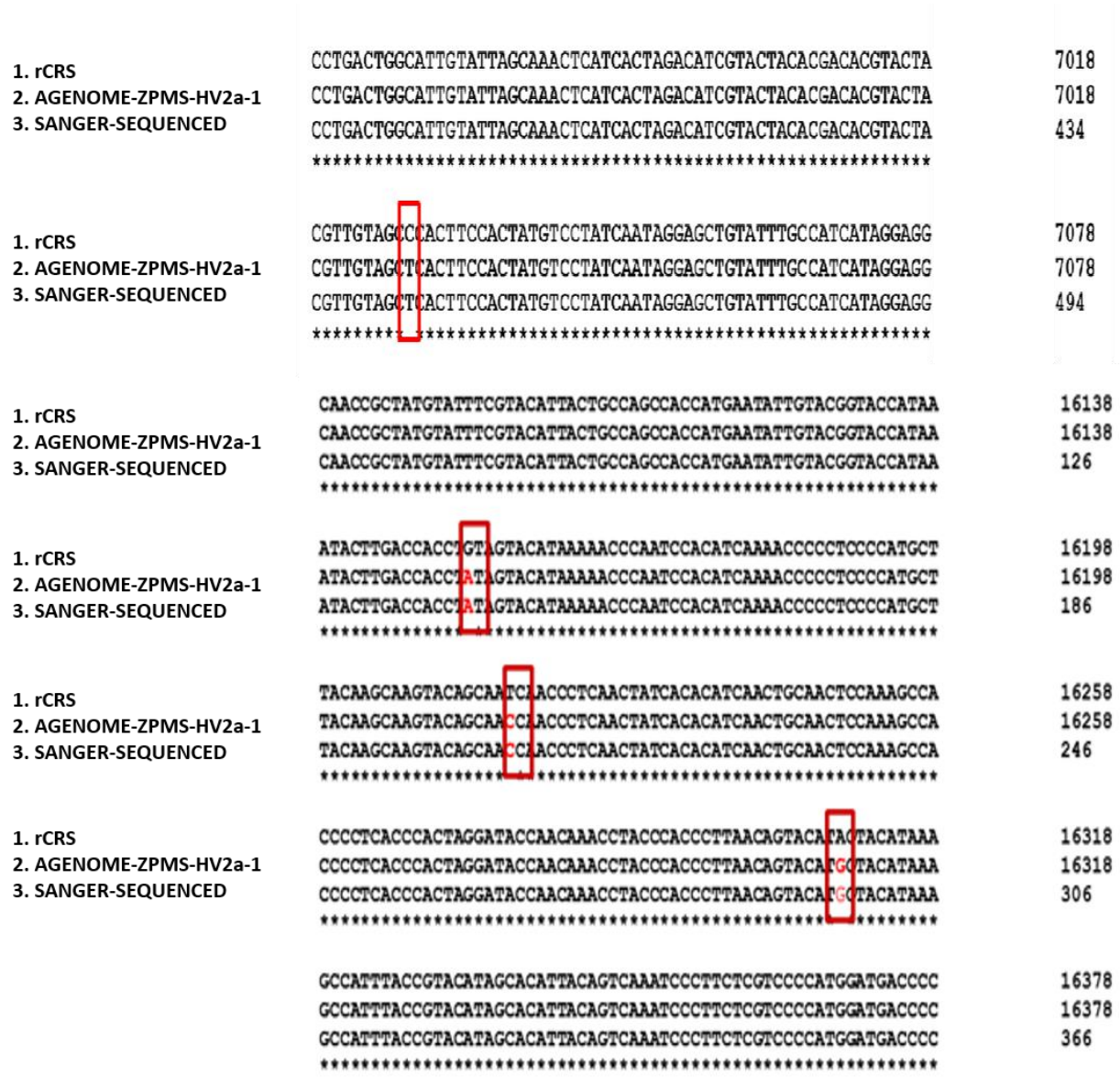


Figure 1: Distribution and classification of Variants in the AGENOME-ZPMS-HV2a-1. Representative histogram showing the base change, variant type, type of loci and distribution of variants across genes in the *de novo* mitochondrial genome AGENOME-ZPMS-HV2a-1

Figure 2 : Validation of variants in the AGENOME-ZPMS-HV2a-1 by Sanger sequencing



```
1. SANGER-SEQUENCED   CAGGACATCCCGATGGTGCAGCCGCTATTAAAGGTTCGTTTGTTCAACGATTAAAGTCCT   181
2. AGENOME-ZPMS-HV2a-1 CAGGACATCCCGATGGTGCAGCCGCTATTAAAGGTTCGTTTGTTCAACGATTAAAGTCCT   3060
3. rCRS               CAGGACATCCCGATGGTGCAGCCGCTATTAAAGGTTCGTTTGTTCAACGATTAAAGTCCT   3058
                      ************************************************************

1. SANGER-SEQUENCED   ACGTGATCTGAGTTCAGACCGGAGTAATCCAGGTCGGTTTCTATCTAC-TTCAAATTCCT   240
2. AGENOME-ZPMS-HV2a-1 ACGTGATCTGAGTTCAGACCGGAGTAATCCAGGTCGGTTTCTATCTAC-TTCAAATTCCT   3119
3. rCRS               ACGTGATCTGAGTTCAGACCGGAGTAATCCAGGTCGGTTTCTATCTACNTTCAAATTCCT   3118
                      ********************************************** ***********

1. SANGER-SEQUENCED   CCCTGTACGAAAGGACAAGAGAAATAAGGCCTACTTCACAAAGCGCCTTCCCCCGTAAAT   300
2. AGENOME-ZPMS-HV2a-1 CCCTGTACGAAAGGACAAGAGAAATAAGGCCTACTTCACAAAGCGCCTTCCCCCGTAAAT   3179
3. rCRS               CCCTGTACGAAAGGACAAGAGAAATAAGGCCTACTTCACAAAGCGCCTTCCCCCGTAAAT   3178
                      ************************************************************

1. SANGER-SEQUENCED   GATATCATCTCAACTTAGTATTATACCCACACCCACCCAAGAACAGGGTTTGTTAAGATG   360
2. AGENOME-ZPMS-HV2a-1 GATATCATCTCAACTTAGTATTATACCCACACCCACCCAAGAACAGGGTTTGTTAAGATG   3239
3. rCRS               GATATCATCTCAACTTAGTATTATACCCACACCCACCCAAGAACAGGGTTTGTTAAGATG   3238
                      ************************************************************

1. SANGER-SEQUENCED   -------------------------------------------------------CCCCA   5
2. AGENOME-ZPMS-HV2a-1 ACAATTGAATGTCTGCACAGCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCA   300
3. rCRS               ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA   300
                                                                          * ***

1. SANGER-SEQUENCED   AACCCCCCCCTCCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAA   65
2. AGENOME-ZPMS-HV2a-1 AACCCCCCCCTCCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAA   360
3. rCRS               AACCCCCC--CTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAA   358
                      ********    **********************************************

1. SANGER-SEQUENCED   AAACAAAGAACCCTAACACCAGCCTAACCAGATTCAAATTTATCTTTTGGCGGTAGCACT   125
2. AGENOME-ZPMS-HV2a-1 AAACAAAGAACCCTAACACCAGCCTAACCAGATTT-------------------------   395
3. rCRS               AAACAAAGAACCCTAACACCAGCCTAACCAGATTT-------------------------   393
                      *********************************
```

Figure 2 : Confirmation of variants identified with next-generation sequencing (NGS) data and confirmation by Sanger sequencing. Sequences obtained from desired regions were analyzed for presence of variants/Variants. Low quality bases were trimmed from both ends of the sequences and used for alignment with the reference Mitochondrial Genome (rCRS). A total of 13 variants/Variants from D-loop and internal region of mitochondrial genome were verified.

Figure 3 : Representation of Males and Females in the 100 Zoroastrian-Parsi whole mitogenome study

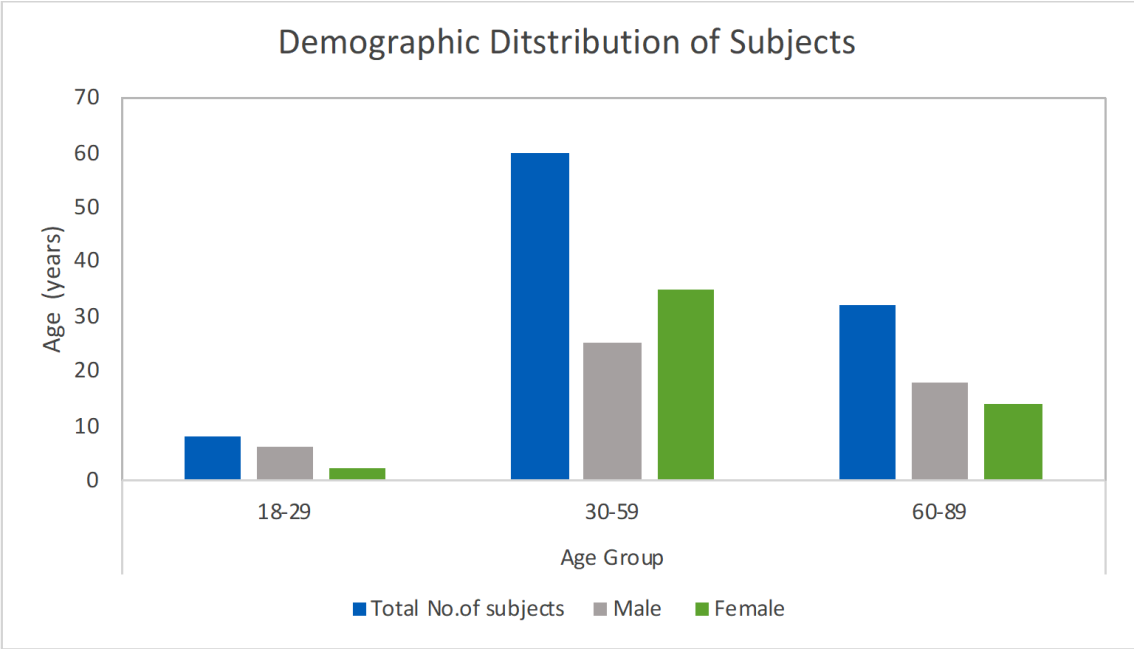
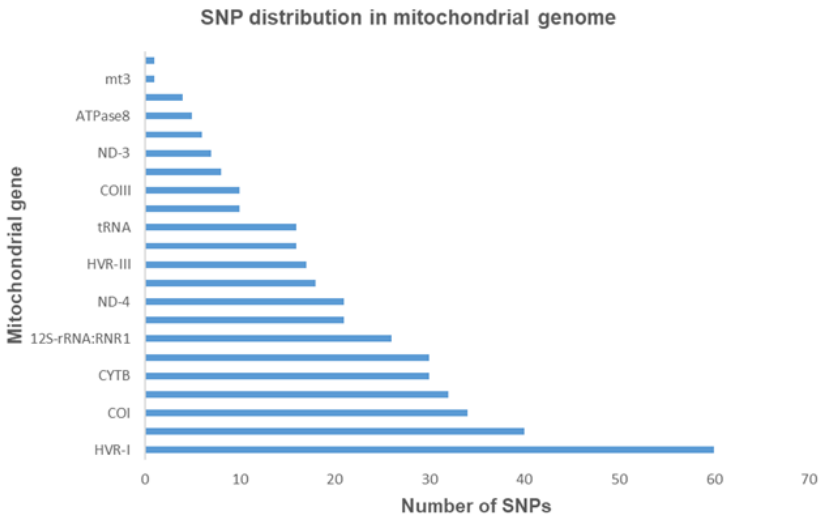


Figure 3: Distribution of 100 Parsi subjects. Distribution of the subjects classified based on gender and age. The bars on the histogram depict further segmentation of the total number of subjects, Male and Female numbers according to their age range.

Figure 4 : Distribution of 420 variants across gene loci in the 100 Zoroastrian-Parsi whole mitogenomes



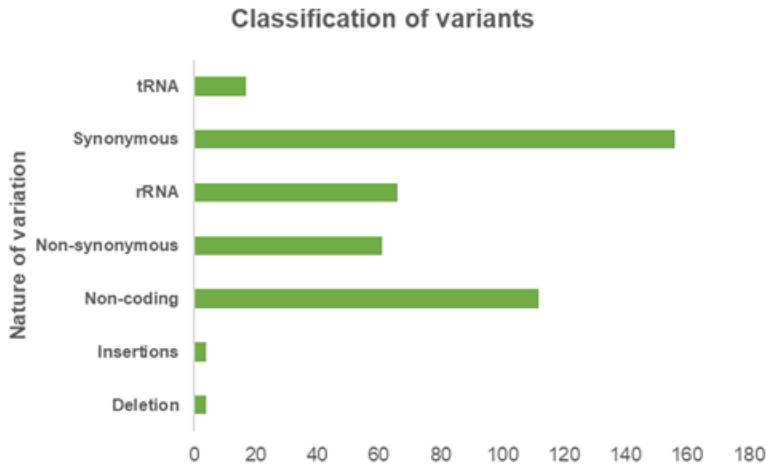


Figure 4 : Annotation and distribution of 420 variants across 100 Parsi complete mitogenomes

Figure 5 : Identification of 25 sub-haplogroups in the 100 Zoroastrian-Parsi study group

Table: Haplogroup and Haplotype count in Parsis

Major haplogroup	Sub-haplotypes	Number of Parsis
HV	HV2a	14
	HV12b	1
U	U7a	6
	U2e	3
	U4b	11
	U1a	1
T	T1a	2
	T2g	1
	T2i	1
	T2b	1
M	M5a	2
	M39b	9
	M33a	1
	M52b	9
	M24a	8
	M3a	8
	M30d	11
	M2a	2
	M4a	1
	M2b	1
	M35b	1
	M27b	1
A	A2v	3
F	F1g	1
Z	Z1a	1

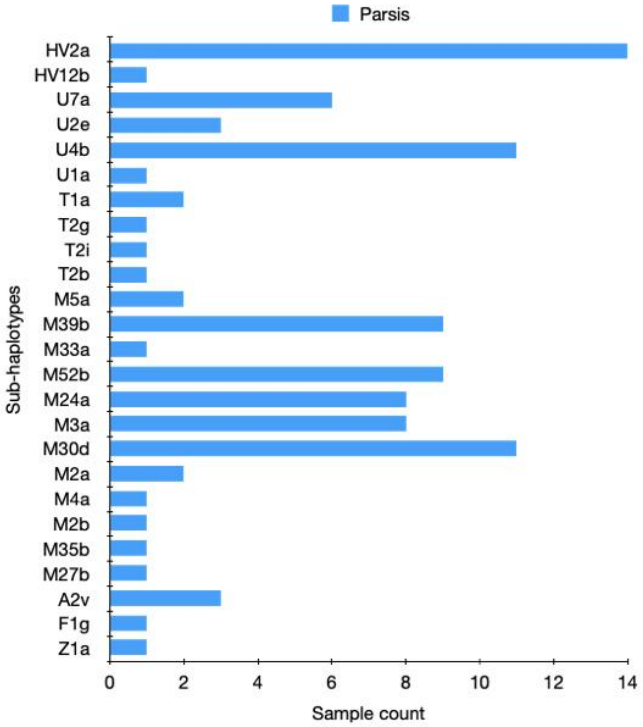
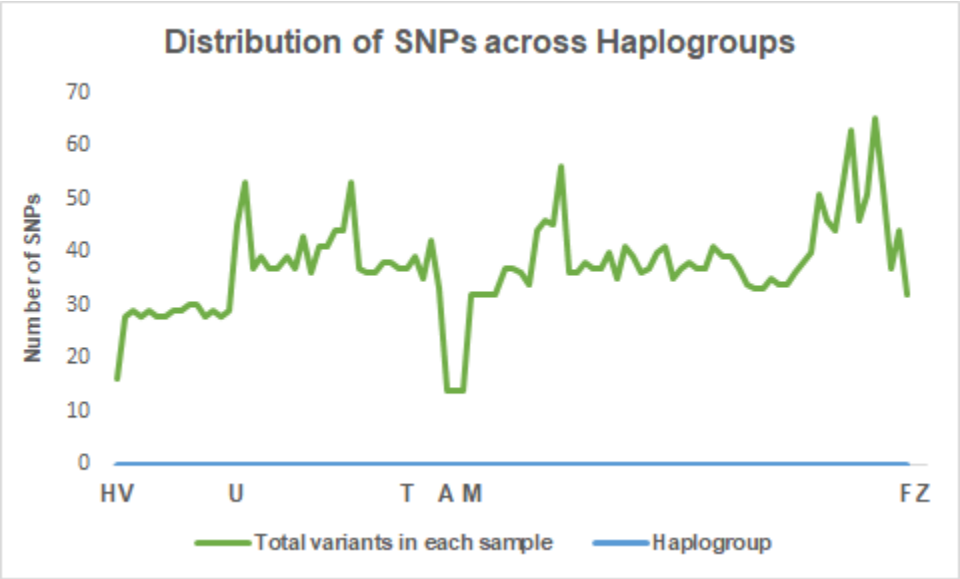


Figure 5 : Distribution of Parsis across major haplogroups and sub-haplogroups. The table and the histogram shows the distribution of 100 Parsi subjects across 7 major haplogroups and 25 sub-haplogroups

Figure 6 : Distribution of variants across haplogroups and demographic classification of the 100 Parsi study group

A



B

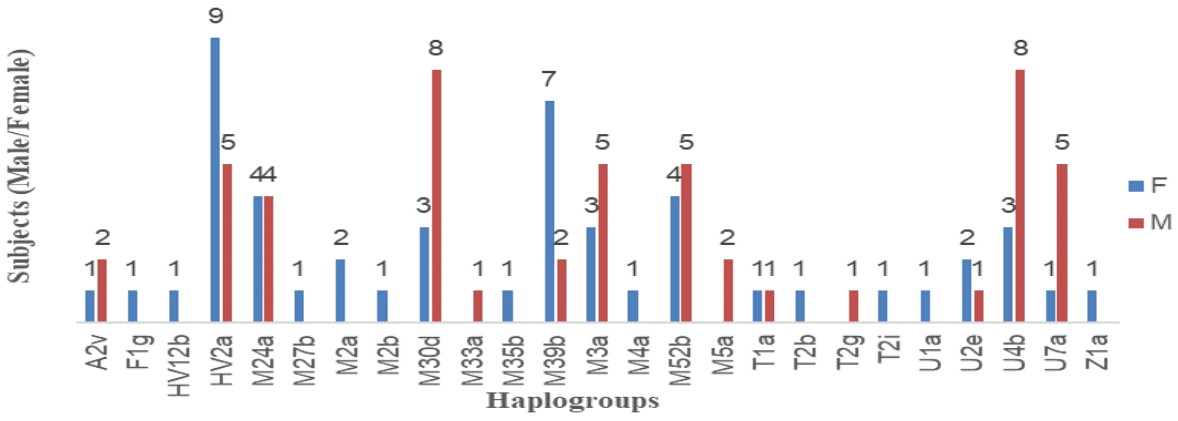


Figure 6: Distribution across the 100 Zoroastrian-Parsi subjects. (A) Representative graph depicting the distribution of SNP's count across the 7 major haplogroups (B) Graph depicts the distribution of the subjects classified based on gender across 25 sub-haplogroups

Figure 7 : Lack of haplogroup diversity in the Parsi cohort suggesting endogamy

A **B**

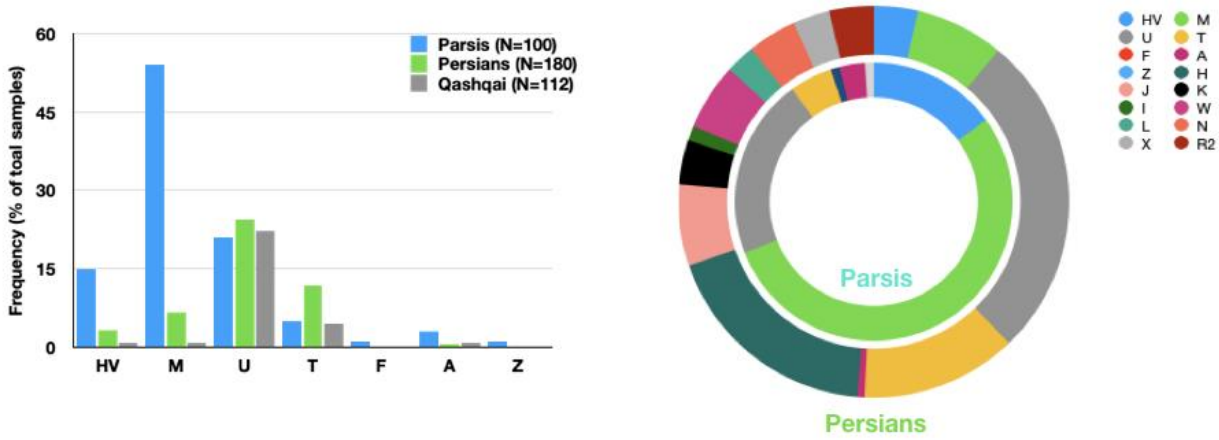


Figure 7: Comparative analysis of Major haplogroup distribution in the Parsis and populations of Iranian ethnicities (Persians, Qashqais) (A) Histogram depicting the analysis of The 7 Major haplogroups across Parsis (n=100), Persians (n=180) and Qashqais (n=112) (B) Representative figure showing the diversity of major haplogroups in the Parsis and the Persians

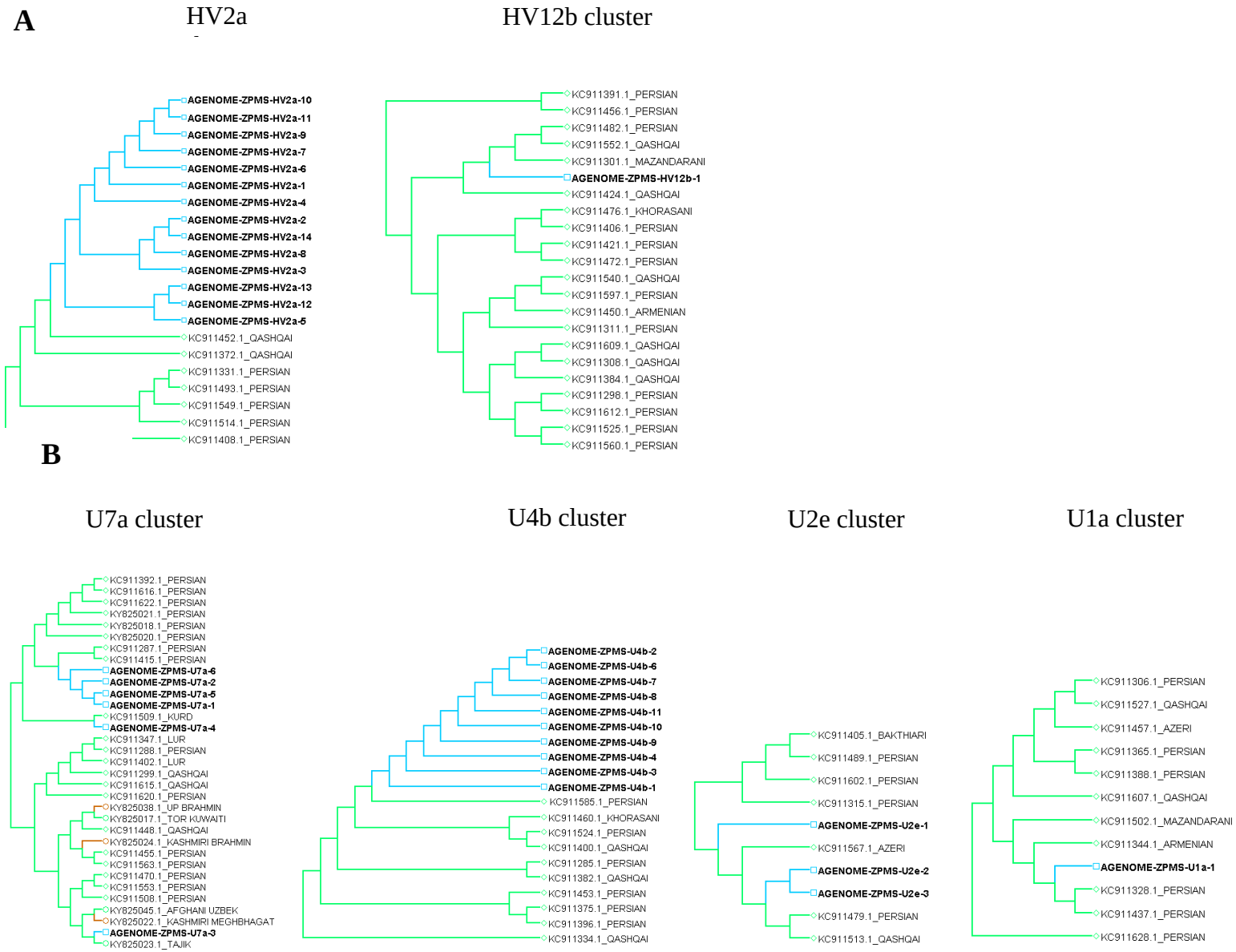


Figure 8: Phylogenetic analysis depicting individual sub-haplogroup clusters of 97 Parsis, 352 Iranian and 100 relic tribes of Indian origin ((A) Representative cladograms of the HV sub-haplogroup (B) Representative cladograms of the U sub-haplogroup

C

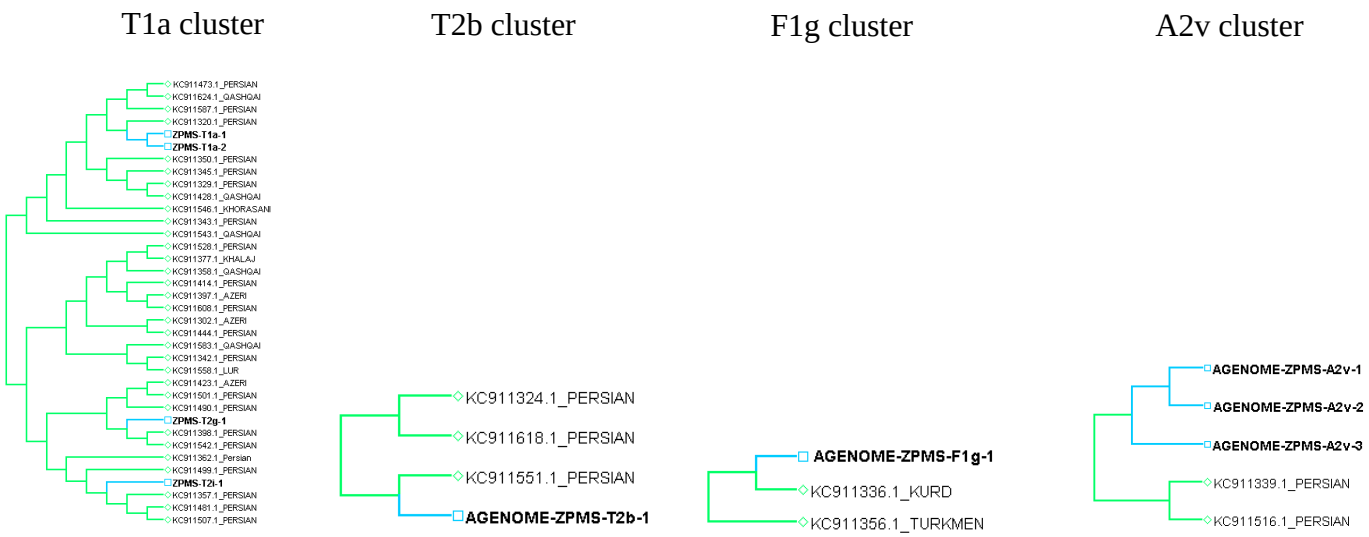


Figure 8: Phylogenetic analysis depicting individual sub-haplogroup clusters of 97 Parsis, 352 Iranian and 100 relic tribes of Indian origin (C) Representative cladograms of the T, F and A sub-haplogroup

D

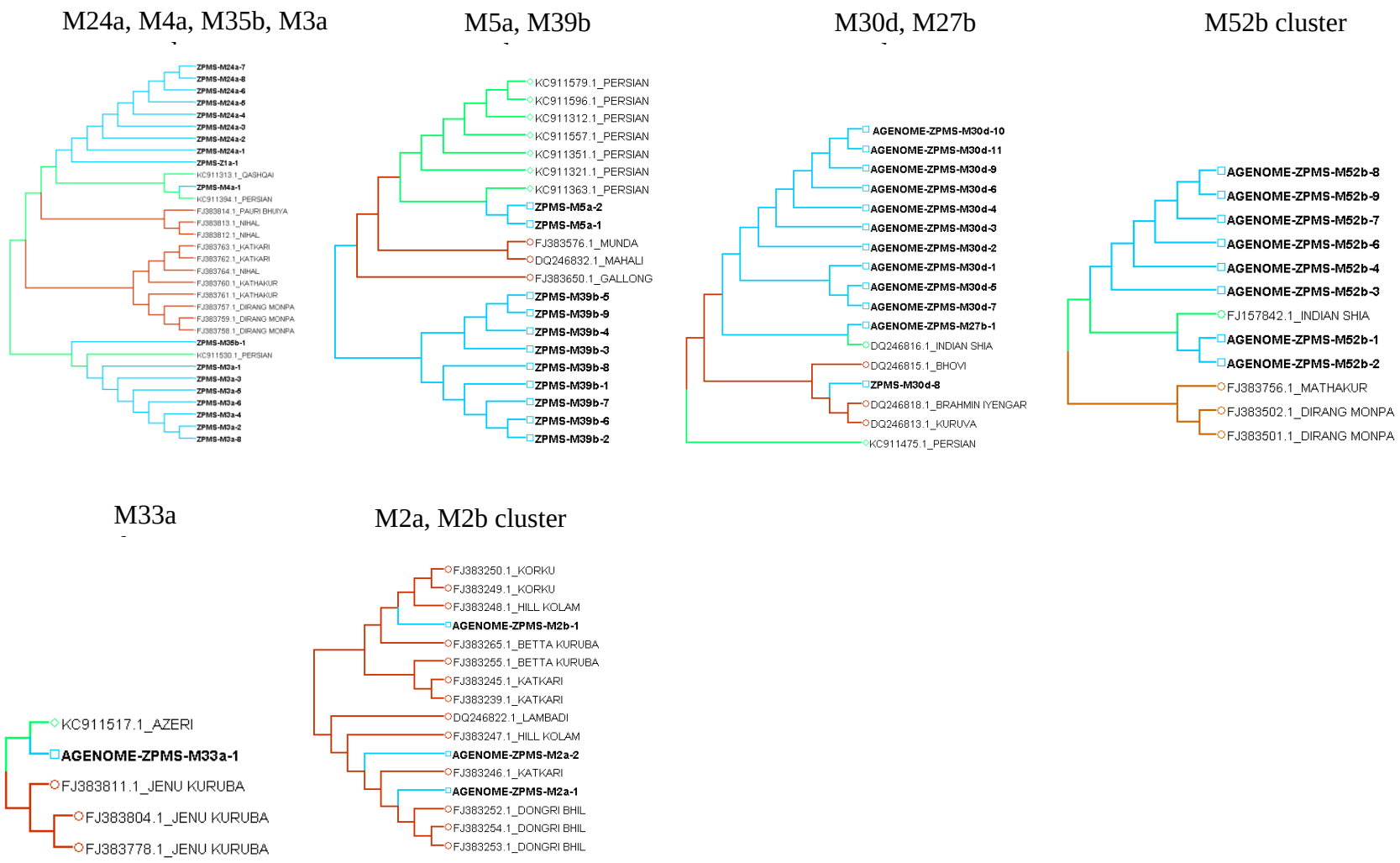
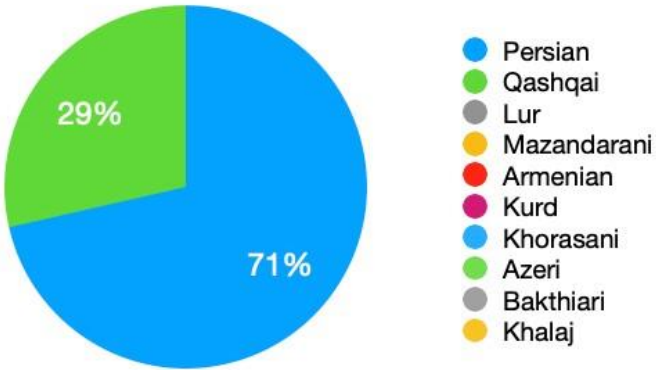


Figure 8: Phylogenetic analysis depicting individual sub-haplogroup clusters of 97 Parsis, 352 Iranian and 100 relic tribes of Indian origin (A-D) Representative cladograms of the each sub-haplogroup

E

Phylogeny analysis

Population	Parsi samples
Iranians and Iranian origin	74
Relic tribes	4
Distinct cluster	19



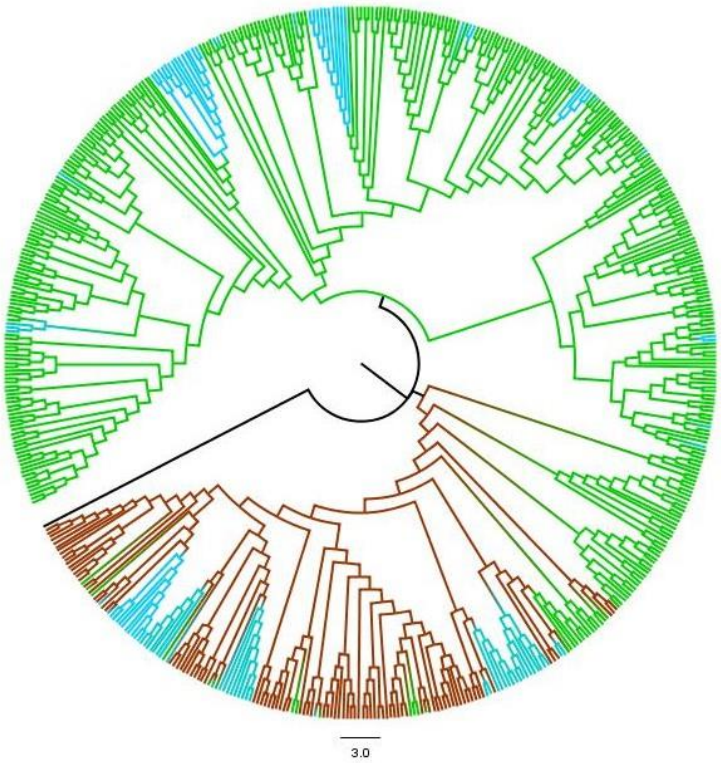


Figure 8E: Table indicates the number of Zoroastrian-Parsis who cluster with Persians or people of Persian origin, relic tribes of Indian origin. Pie chart indicates the percentage of clustering of the HV2a Zoroastrian-Parsis in the phylogenetic clustering analysis. Circular dendrogram of the complete Phylogenetic clustering analysis of Parsis (Blue clades) with Iranian mitogenomes (Green clades) and Indian mitogenomes (Brown clades). Outgroup is indicated by the black line.

Figure 9: Lack of smoking induced mutational signatures in the Parsi cohort

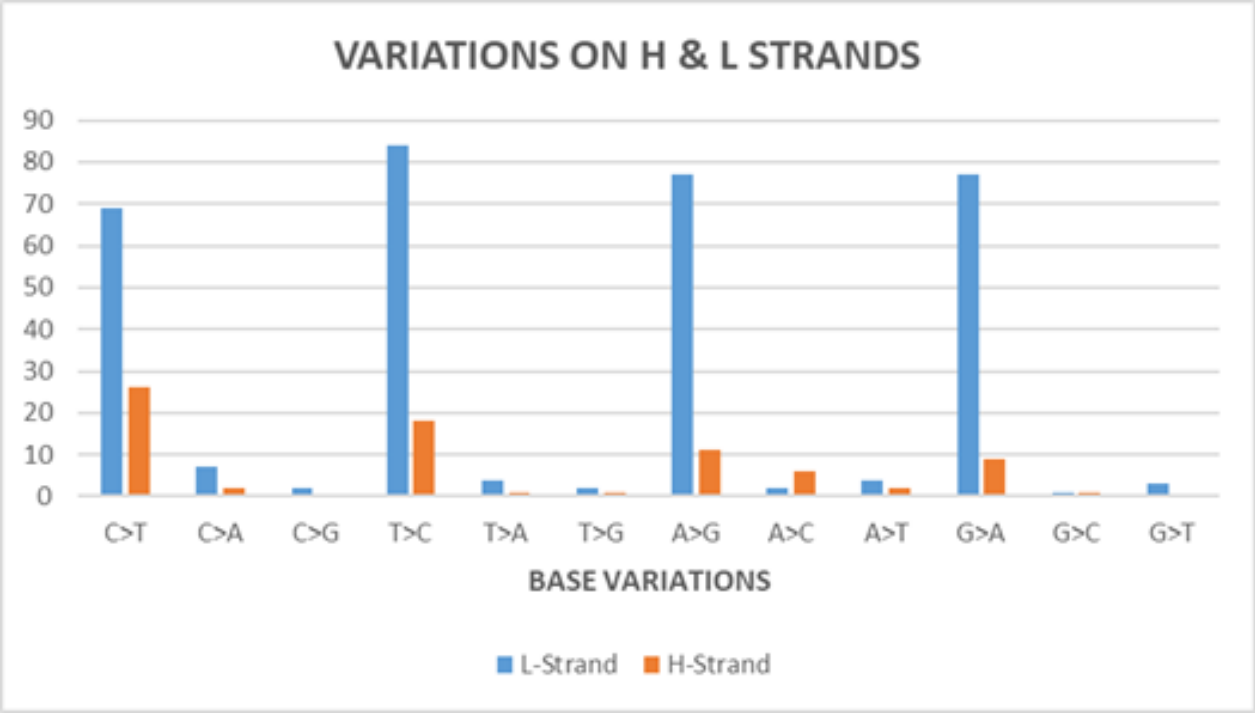
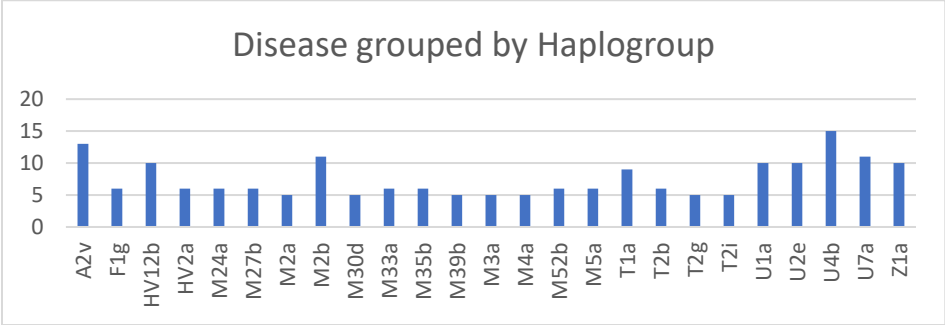


Figure 9: Mutational signatures observed in the 100 mitochondrial genomes of Parsis. Graph depicts the quantification of both transitions and transversions on both H&L strands of the 100 mitochondrial genomes of Parsis.

Figure 10: Observation of Longevity variants across all sub-haplogroups and predisposition of U and M haplogroups to diseases

A



B

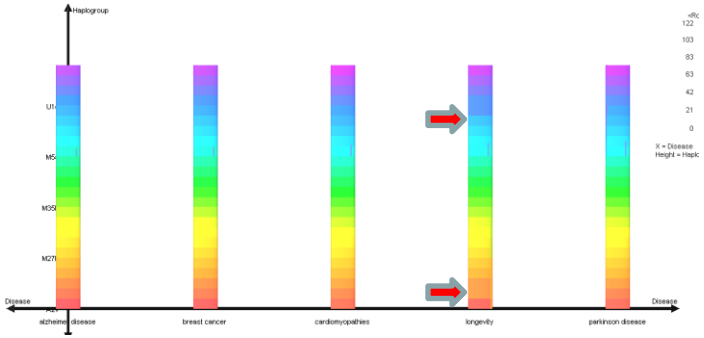
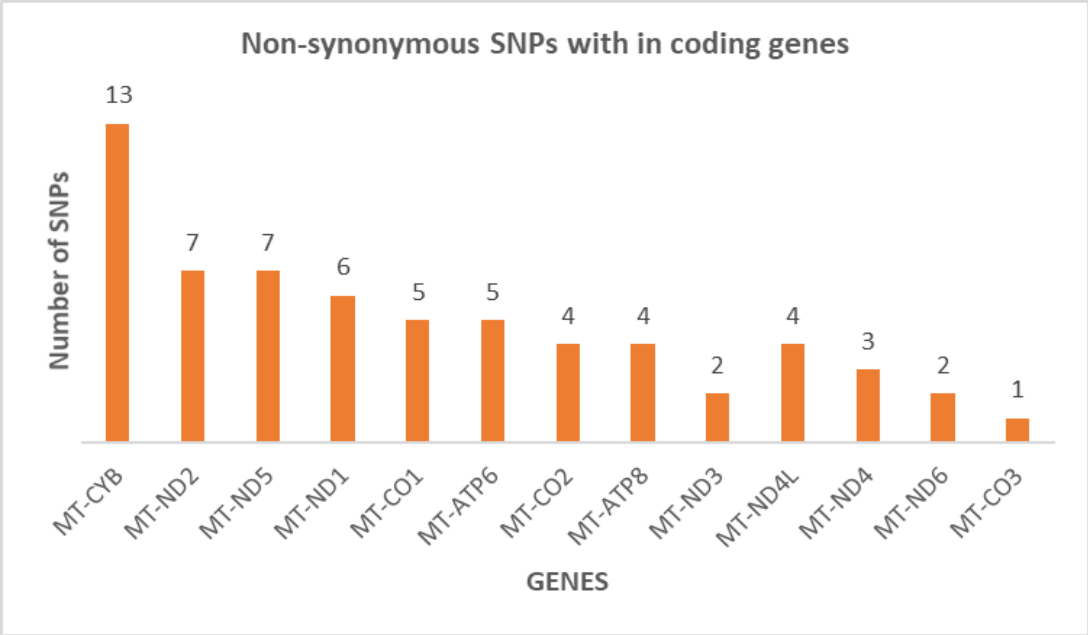


Figure 10: Haplogroup specific distribution of diseases. (A) Distribution of 188 diseases across 25 sub-haplogroups of the 100 Parsi subjects analyzed in this study (B) Histogram depicting longevity and disease prevalence across U1a, M52b, M35b, M27b

Figure 11: PCA analysis shows absence of Longevity variants in U1a and HV12b sub-haplogroups

Figure 11: Principal Component Analysis of disease associations with sub-haplogroups in the Parsi-Zoroastrian group under study

Figure 12: CYTB gene has the highest occurrence of non-synonymous variants in this study



Gene_symbol	Count
MT-CYB	13
MT-ND2	7
MT-ND5	7
MT-ND1	6
MT-CO1	5
MT-ATP6	5
MT-CO2	4
MT-ATP8	4
MT-ND3	2
MT-ND4L	4
MT-ND4	3
MT-ND6	2
MT-CO3	1

Figure 12: Analysis of the non-synonymous variants within 420 variants in the 100 Parsi mitochondrial genome sequences. The histogram and the table show the location of the non-synonymous variants in the coding gene loci in the mitochondrial genome analysed with MitImpact database

Figure 13: Non-synonymous variants among 420 variants and their disease associations

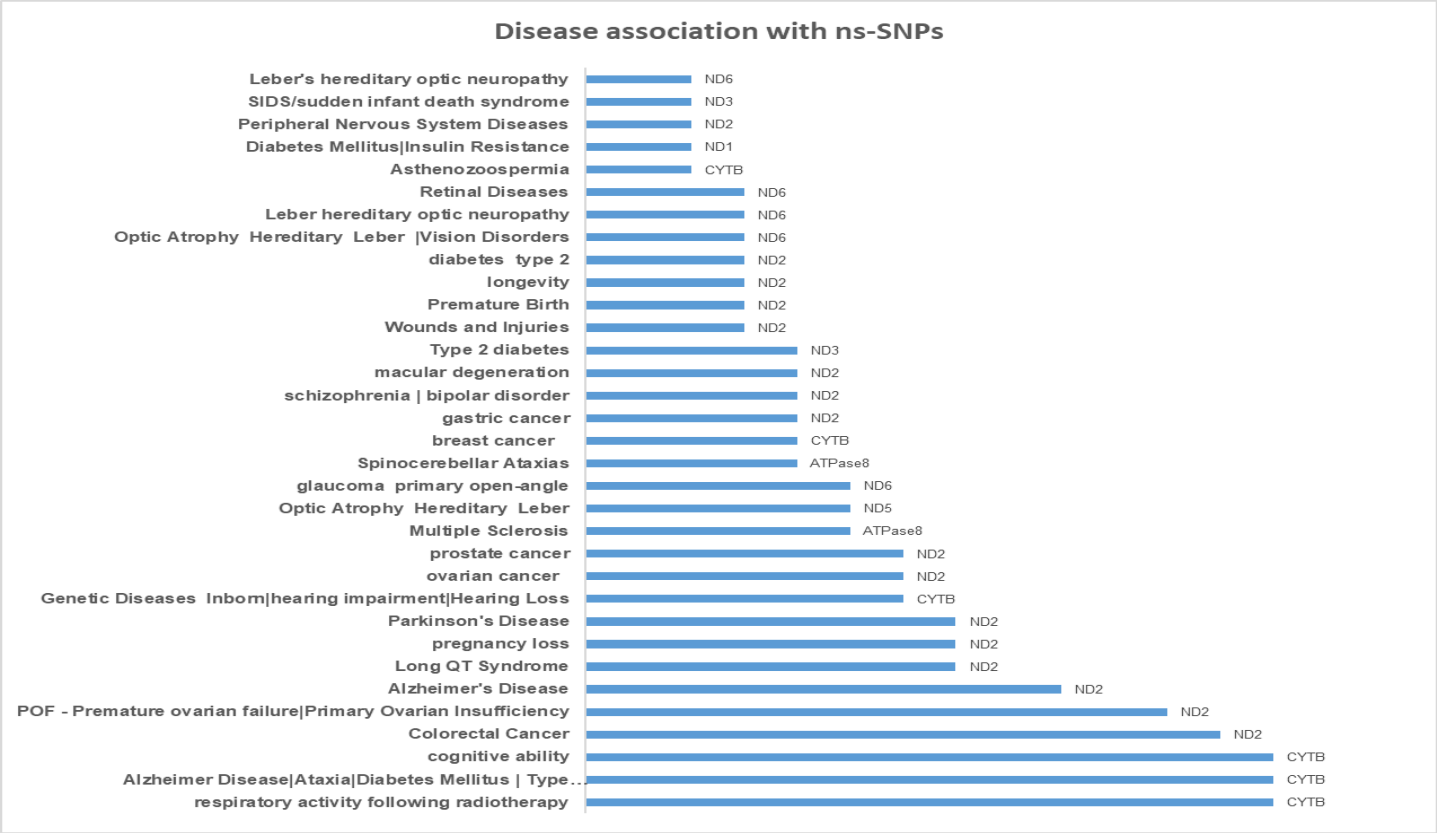


Figure 13: Analysis of the non-synonymous variants within 420 variants in the 100 Parsi mitochondrial genome sequences for and their disease associations.

Figure 14: Non-synonymous variants among 420 variants and their associations with mitochondrial function

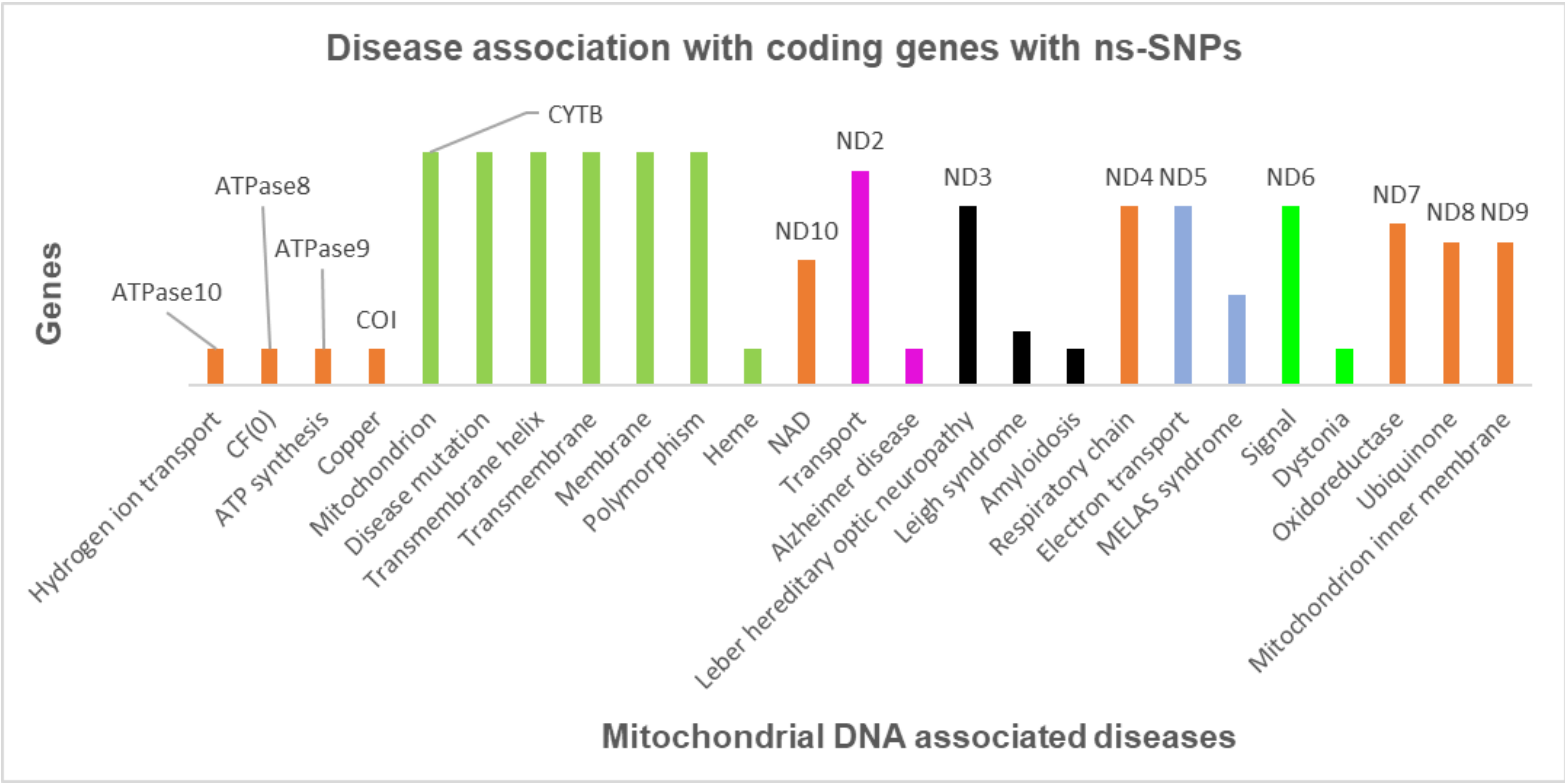


Figure 14 : Distribution of non-synonymous Variants across coding genes. Analysis was performed on the 420 Variants linked to the 100 Parsi mitochondrial genomes.

Figure 15: Gene ontology associated with non-synonymous variants among 420 variants

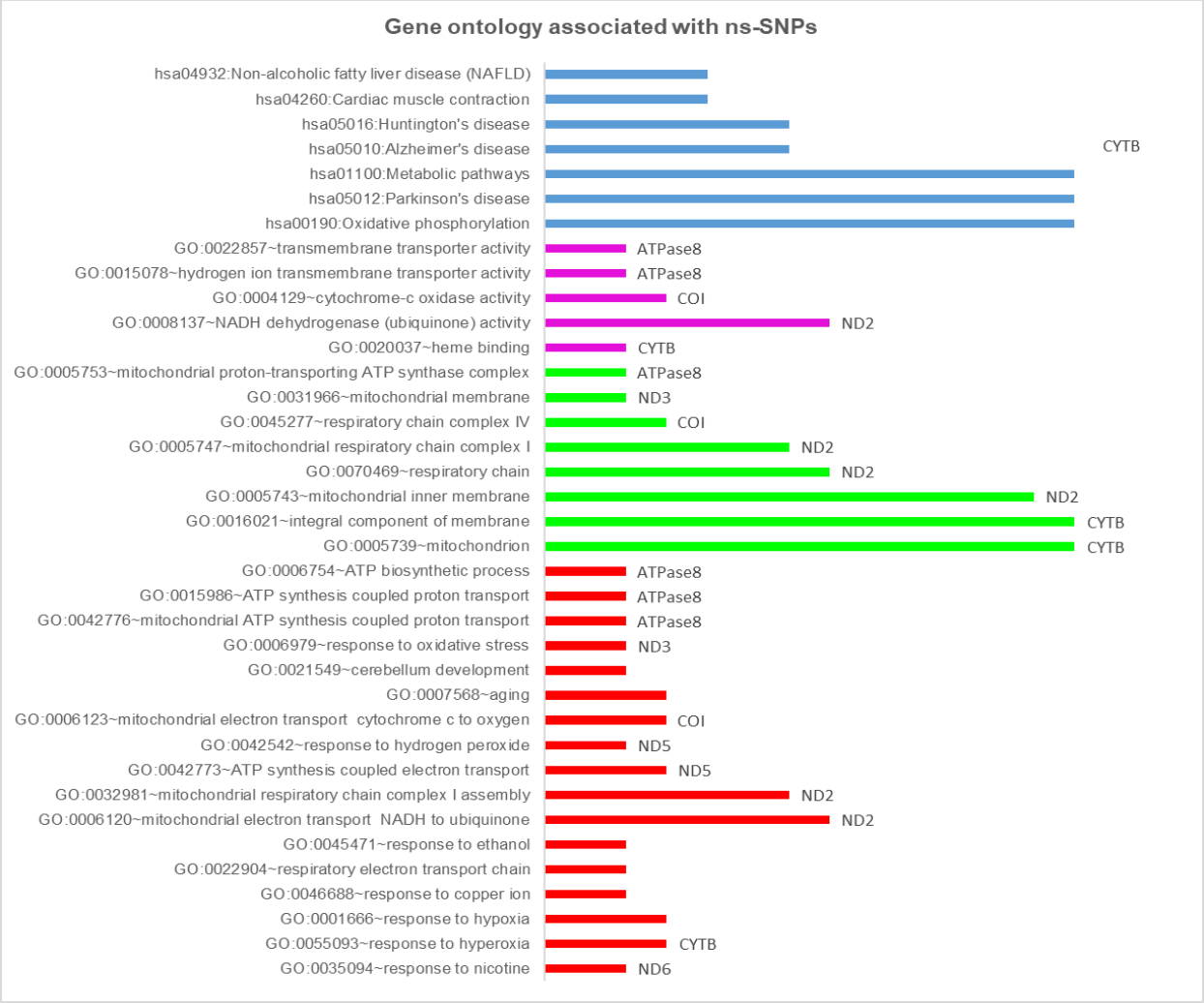


Figure 15: Analysis of non-synonymous mutations and their functional classification, engagement in different pathways respectively using DAVID and UNIPROT annotation tools.

Figure 16: 12 unique variants found in the current study

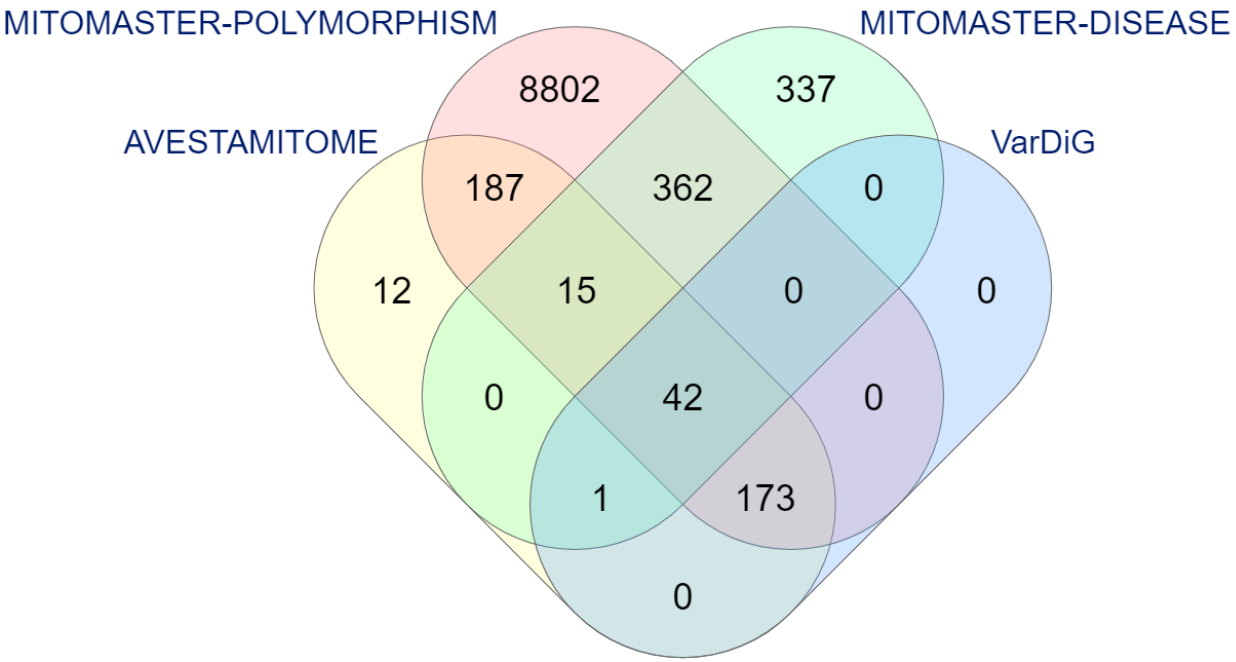


Figure 16: Comparative analysis of the 420 variants in the AVESTAMITOME™ Zoroastrian-Parsi community dataset with common and disease associated polymorphisms in MITOMASTER database and VarDiG®-R

Main Tables

Table 1: Annotation of 28 variants in the AGENOME-ZPMS-HV2a-1

Reference_position	72	73	152	195	263	309.1	309.2	310	750	1438	2706	4769	5075	6104	6179	7028	7193
Reference_base	T	A	T	T	A	.	.	T	A	A	A	A	T	C	G	C	T
AGENOME-ZPMS-HV2a-1	C	G	C	C	G	C	T	C	G	G	G	G	C	T	A	T	C
Mitochondrial genome loci	HVR-II								12S-rRNA:RNR1		16S-rRNA:RNR2	ND2		COI			
Amino Acid change	nc	nc	nc	nc	nc	nc	nc	nc	rRNA	rRNA	rRNA	M100M	I202I	F67F	M92M	A375A	F430F
Conservation index									98%	87%	84%	24%	44%	100%	100%	100%	100%
Protein Position												100	202	67	92	375	430
Variant Type												syn	syn	syn	syn	syn	syn
Type of base change	trans	trans	trans	trans	trans	ins	ins	trans	trans	trans	trans	trans	trans	trans	trans	trans	trans

Reference_position	8860	9336	10410	11016	11935	12061	15326	15792	16153	16217	16309	Haplogroup
Reference_base	A	A	T	G	T	C	A	T	G	T	A	
ZPMS-HV-1	G	G	C	A	C	T	G	C	A	C	G	HV2a
Mitochondrial genome loci	ATPase6	COIII	tRNA [R]	ND-4			CYTB		HVR-I			
Amino Acid change	T112A	M44V	tRNA	S86N	T392T	N434N	T194A	I349T	nc	nc	nc	
Conservation index	71%	16%	22%	7%	89%	69%	18%	58%				
Protein Position	112	44		86	392	434	194	349				
Variant Type	n-syn	n-syn		n-syn	syn	syn	n-syn	n-syn				
Type of base change	trans	trans	trans	trans	trans	trans	trans	trans	trans	trans	trans	

Table 1: Annotation of the de novo Parsi mitochondrial genome AGENOME-ZPMS-HV2a-1. B) The table indicates the Variants (n=28) found in the AGENOME-ZPMS-HV2a-1 in relation to the revised Cambridge Reference Sequence (rCRS, Reference bases

Table 2: Distribution of 420 variants for each sub-haplogroup for protein coding regions, D-loop of 100 Parsi mitogenomes

Association of coding region, D-loop with sub-haplogroup

Sub-haplogroup	Coding gene SNPs	Gene with max SNPs	D-loop
HV2a	20	6 COI	1
HV12b	6	2 CYTB	0
U4b	21	6 COI	4
U2e	22	4 CYTB, 4 ND2, 4 ND5	2
U7a	25	6 ND5	2
U1a	21	6 ND5	2
T1a	18	5 CYTB	1
T2b	17	3 CYTB	0
T2g	24	6 CYTB	1
T2i	17	5 CYTB	1
A2v	5	2 ND2	0
F1g	20	7 CYTB	0
Z1a	17	3 ND5	1
M30d	24	8 CYTB	1
M33a	19	5 CYTB	1
M35b	21	5 CYTB	1
M39b	22	5 CYTB	0
M3a	19	5 CYTB	1
M4a	20	5 CYTB	1
M52b	31	9 CYTB	2
M5a	21	6 CYTB	1
M2a	29	7 CYTB	2
M2b	25	6 CYTB	2
M27b	22	6 CYTB	1
M24a	20	5 CYTB	1

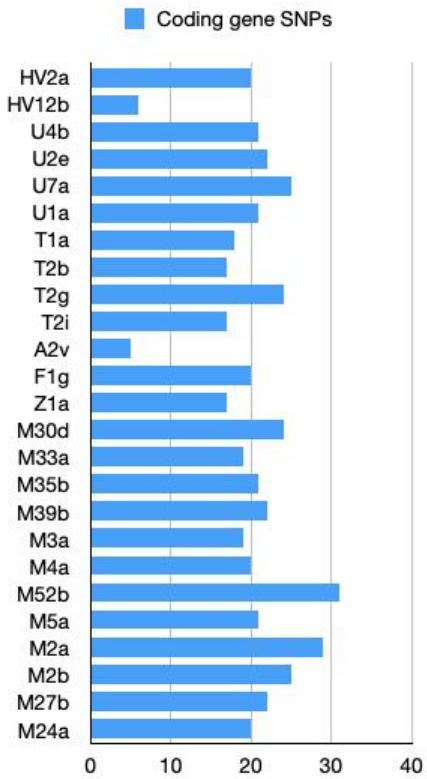


Table 2: Distribution of Variants across coding genes, D-loop across all the 25 sub-haplogroup

Table 3: Phylogenetic clustering of complete mitogenomes of Parsis with 352 Iranian and 100 relic tribes of Indian origin

Table: Clustering of Parsis with population of Persian and Indian descent

Major haplogroup	Sub-haplogroups	People of Persian origin (PO)	People of Indian & Relic tribal origin (IO)	Max BS value to nearest PO	Max BS value to nearest IO
HV	HV2a	Persian	N.A	0.7270	0
	HV12b	Persian, Qashqai, Mazandarani	N.A	0.6550	0
U	U7a	Persian, Kurd, Tajik	N.A	0.8980	0
	U2e	Persian, Qashqai, Azeri	N.A	1.000	0
	U4b	Persian, Khorasani, Qashqai	N.A	0.5100	0
	U1a	Persian, Armenian	N.A	0.6850	0
T	T1a	Persian	N.A	0.7320	0
	T2g	Persian	N.A	0.4880	0
	T2i	Persian	N.A	0.4480	0
	T2b	Persian	N.A	0.4320	0
M	M5a	Persian	Munda, Mahali	0.9860	0.6270
	M39b	Unique cluster			
	M33a	Azeri	Jenu Kuruba	0.2250	0.0960
	M52b	Indian Shia Muslim	Mathakur, Dirang Monpa	0.7950	0.1170
	M24a	Persian, Qashqai	Pauri Bhaiya, Nihal	0.8560	0.0200
	M3a	Persian	N.A	0.9380	0
	M30d	Unique cluster	1 M30d with Brahmin Iyengar, Bhovi	0	0.4020
	M2a	N.A	Lambadi, Hill Kolam, Katkari, Dongri Bhil	0	0.6110
	M4a	Persian	N.A	0.8560	0
	M2b	N.R	Korku, Hill Kolam	0	0.9400
	M35b	Persian	N.A	0.3860	0
	M27b	Indian Shia Muslim	N.A	0.4220	0
A	A2v	Persian	N.A	0.4690	0
F	F1g	Kurd, Turkmen	N.A	0.9970	0
Z	Z1a	Qashqai, Persian	N.A	0.2470	0

Table 3 : Results of the Phylogenetic clustering of the 100 Parsis mitochondrial genomes with 352 mitochondrial genomes of Iranian origin and 100 mitochondrial genomes of relic tribes of Indian origin through Neighbour Joining method. BS indicates Boot-Strap values between each sample. *N.A. indicates *No Association,* indicating a lack of representation of samples in the specific sub-haplogroup

Table 4: Variants associated with haplogroup specific Zoroastrian Parsi Mitochondrial Reference Genome (n=7) and Zoroastrian Parsi Mitochondrial Consensus Genome (n=1) mitochondrial genome sequences

Consensus Sequence	Number of Variants	Variants
AGENOME-ZPMCG-V1.0	31	T65TT, A73G, A263G, C309CCCT, T310C, T489C, G513GCA, A567ACCCCCC, A750G, A1438G, A2706G, A3158AT, A4769G, C7028T, A8701G, A8860G, T9540C, A10398G, C10400T, T10873C, G11719A, C12705T, C14766T, T14783C, G15043A, G15301A, A15326G, C16169CC, A16182AC, C16223T, T16519C
AGENOME-ZPMRG-**A2v**-V1.0	11	A263G, C309CCT, T310C, A750G, A1438G, A4769G, A8860G, C11881T, A15326G, C16168T, C16239T
AGENOME-ZPMRG-**HV**-V1.0	26	T72C, A73G, T152C, T195C, A263G, C309CCT, T310C, A750G, A1438G, A2706G, A4769G, T5075C, C6104T, G6179A, C7028T, T7193C, A8860G, A9336G, T10410C, G11016A, T11935C, C12061T, A15326G, T15792C, T16217C, A16309G
AGENOME-ZPMRG-**M**-V1.0	29	T65TT, A73G, A263G, C309CCCT, T310C, T489C, A567ACCCC, A750G, A1438G, A2706G, A4769G, C7028T, A8701G, A8860G, T9540C, A10398G, C10400T, T10873C, G11719A, C12705T, C14766T, T14783C, G15043A, G15301A, A15326G, C16169CC, A16182AC, C16223T, T16519C
AGENOME-ZPMRG-**U**-V1.0	25	A73G, A263G, C309CCCT, T310C, G499A, G513GCA, A567ACCCCCC, A750G, A1438G, A1811G, A2706G, A3158AT, A4769G, C7028T, A8860G, C11332T, A11467G, G11719A, A12308G, G12372A, C14620T, C14766T, A15326G, T16189TT, T16519C
AGENOME-ZPMRG-**T**-V1.0	28	A73G, A263G, C309CCT, T310C, G709A, A750G, A1438G, G1888A, A2706G, T4216C, A4769G, A4917G, C7028T, G8697A, A8860G, T10463C, A11251G, G11719A, G13368A, C14766T, G14905A, A15326G, C15452A, A15607G, G15928A, T16126C, C16294T, T16519C
AGENOME-ZPMRG-**F1g**-V1.0	32	A73G, A248d, A263G, C315CC, CA514d, A750G, A1438G, C2389T, A2706G, T3398C, C3970T, T3999C, A4769G, T6392C, G6962A, C7028T, A8589G, A8860G, G10310A, T10609C, G11719A, G12406A, C12882T, G13928C, C14766T, A15326G, T15916C, A16183C, T16189C, C16193CC, T16304C, T16519C
AGENOME-ZPMRG-**Z**-V1.0	33	A73G, C151T, T152C, A263G, C315CC, T489C, A750G, A1438G, A2072d, A2706G, A4769G, C7028T, A8701G, A8860G, T9540C, A10149T, A10398G, C10400T, C10556T, T10873C, G11719A, G12007A, C12705T, C14766T, T14783C, G15043A, G15301A, A15326G, G15346A, T15784C, C16223T, T16311C, T16519C

Table 4: List of unique variants associated with the Haplogroup specific Zoroastrian Parsi Mitochondrial Reference Genomes (ZPMRG) for A2v, HV, M, U, T, F1g, Z and overall unique variants in the Zoroastrian Parsi Mitochondrial Consensus Genome (ZPMCG)

Table 5: Variants associated with Zoroastrian Parsi Mitochondrial Reference Genome (ZPMRG) and unique variants of each ZPMRG compared to Zoroastrian Parsi Mitochondrial Consensus Genome (ZPMCG)

AGENOME-ZPMRG-A2v-V1.0	AGENOME-ZPMRG-HV-V1.0	AGENOME-ZPMRG-M-V1.0	AGENOME-ZPMRG-T-V1.0	AGENOME-ZPMRG-U-V1.0	AGENOME-ZPMRG-F-V1.0	AGENOME-ZPMRG-F-V1.0
C11881T	A16G		C6G	A21G	A248d	C151T
C16168T	T72C		G709A	G499A	CA514d	A2072d
C16239T	T195C		G1888A	A1811G	C2389T	C10556T
	T5075C		T4216C	C11332T	T3398C	G12007A
	C6104T		A4917G	A11467G	C3970T	G15346A
	G6179A		G8697A	A12308G	T3999C	T15784C
	T7193C		T10463C	G12372A	T6392C	T16311C
	A9336G		A11251G	C14620T	G6962A	
	T10410C		G13368A		A8589G	
	G11016A		G14905A		G10310A	
	T11935C		C15452A		T10609C	
	C12061T		A15607G		G12406A	
	T15792C		G15928A		C12882T	
	T16217C		T16126C		G13928C	
	A16309G		C16294T		T15916C	
					A16183C	
					C16193CC	
					T16304C	

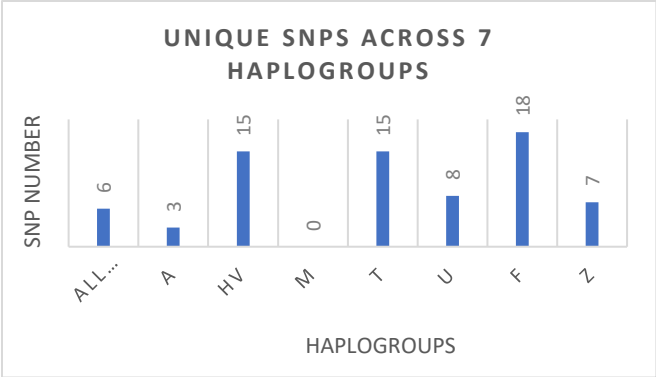


Table 5: (A) Unique Variants found in the haplogroup specific Reference Genomes (ZPMRG) compared to the Zoroastrian-Parsi Consensus Genome (AGENOME-ZPMCG-V1). The histogram (right) lists the exact number of variants in each ZPMRG compared to ZPMCG

Table 6: mt-t-RNA variants in our study and their disease association

mt-tRNA	Variation	Probability_of_ pathogenicity	Classification	Frequency %	Haplogroup	Disease association
Phe	T593C	0.16	Neutral	0.06	M52b	Non-syndromic hearing loss (Reported)
Val	G1644A	0.67	Pathogenic	0.01	U4b	LS/HCM/MELAS (Reported)
Val	T1654C	0.12	Neutral	0.01	M3a	
Met	T4454C	0.13	Neutral	0.02	M5a	Possible contributor to mito dysfuntion / Hypertension (Reported)
Asp	G7521A	0.46	Likely neutral	0.01	U4b	
Asp	T7561C	0.33	Neutral	0.01	U7a	
Asp	T7581C	0.42	Likely neutral	0.01	U1a	
Arg	T10410C	0.17	Neutral	0.14	Hv2a	
Arg	T10463C	0.31	Neutral	0.04	T1a,T2g,T2i	
His	A12172G	0.53	Likely pathogenic	0.01	U4b	
His	C12191G	0.11	Neutral	0.01	M27b	
Leu(CUN)	A12279G	0.37	Likely neutral	0.06	M52b	
Leu(CUN)	A12308G	0.41	Likely neutral	0.21	U4b,U7a	Stroke, CM, CPEO, Breast/Renal/Prostate cancer risk, Altered brain pH(Reported)
Glu	A14696G	0.26	Neutral	0.01	A2v	Progressive Encephalopathy (Reported)
Thr	A15907G	0.23	Neutral	0.03	U2e	
Thr	T15908C	0.5	Likely pathogenic	0.01	M33a	Deaf Helper mutation (Reported)
Thr	T15916C	0.33	Likely neutral	0.01	F1g	

Table 6: Analysis of the occurrence of the 420 variants in the tRNA and their disease associations annotated with the PON-mt-tRNA database. A frequency score ≥0.5 – pathogenic, =0.5 – likely pathogenic, <0.5 – neutral