Impact of mitochondrial epistatic interactions on evolution of different human subpopulations

Pramod Shinde[1,*], Harry J. Whitwell[2,3], Rahul Kumar Verma[1], Alexey Zaikin[3,4], Sarika Jalan[1,3,5,**]

[1] *Discipline of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Khandwa road, Simrol, Indore 453552, India.* [2] *Chemical Engineering, Imperical College London, London, UK.* [3] *Lobachevsky University, Gagarin avenue 23, Nizhny Novgorod, 603950, Russia.* [4] *Department of Mathematics, University College London, London, WC1E 6BT, UK.* [5] *Complex Systems Lab, Indian Institute of Technology Indore, Khandwa road, Simrol, Indore 453552, India.*
*Corresponding authors: *pramodshinde119@gmail.com, **sarikajalan9@gmail.com*

Abstract

Studies of human mitochondrial (mt) genome variation may be undertaken to investigate the human history and natural selection. By analyzing nucleotide co-occurrence over the entire human mt-genome, we have developed a network model to describe human evolutionary patterns. Using 7,424 unique polymorphic sites, we found evidence that mutation biases at second codon position and RNA genes were critical to producing continental-level heterogeneity among human subpopulations. Further, the analysis highlighted richer co-mutation regions of the mt-genome and thus provided evidence of epistasis. Specifically, a large portion of COX genes co-mutate in Asian and American populations whereas, in African, European and Oceanic populations, there was greater epistasis in hypervariable regions. Very interestingly, this study demonstrated hierarchical modularity as a crucial agent for a nucleotide co-occurrence network make-up. More profoundly, our ancestry-based nucleotide module analyses showed that nucleotide co-changes cluster preferentially in known mitochondrial haplogroups. It was also conceived that contemporary human mt-genome nucleotides most closely resembled the ancestral state and very few of them were ancestral-variants. Overall, these results demonstrated that subpopulation based factors such as intra-species evolution do exert selection on mitochondrial genes by favoring specific epistatic genetic variants.

Keywords: Human mitochondria, Genome evolution, Co-occurrence network, Epistasis, Hierarchical modularity

Background

Genetic polymorphism varies among species and within genomes and has important implications for the evolution and conservation of species. Polymorphism in the mitochondrial (mt) genome is routinely used to trace ancient human migration routes and to obtain absolute dates for genetic prehistory [1]. The human mt-genome is very small (16.6 kb), maternally inherited, evolves in both neutral and adaptive fashion, and shows a great deal of variation as a result of divergent evolution. An absence of recombination within mt-genome provides distinct polymorphic loci which have been used to define human genealogy by defining mt-genome haplogroups [1]. These haplogroups are formed as a result of the sequential accumulation of mutations through maternal lineages. Since mitochondria are essential to cellular metabolism, the mt-genome variation has been associated with multiple complex diseases including Alzheimer's disease in haplogroup U [2], idiopathic Parkinson disease within JT haplogroup [3] and age-related macular degeneration with the JTU haplogroup cluster [4]. Due to population migration, distinct lineages of mt-genome are associated with major global groups (African, American, European, Asian and Oceanic) raising the possibility that mt-genome variation could contribute to the differences in disease prevalence observed among both ethnic and racial groups [5].

Conventionally, analyses of mt-genome evolution have been focused on individual mutations, particularly in describing haplogroups, to understand and predict ancestral behavior. These approaches effectively iden-

tify single mutations. However, the evolutionary behavior of mt-genome often involves cooperative changes within and between genes which are difficult to detect using haplogroup analysis. For example, correlated mt-genome mutations were reported among different oxidative phosphorylation subunits which were found to affect population specific human longevity [6]. Besides, cooperative activities of both mitochondrial proteins and tRNA genes are critical for mt-genome evolution. The importance of co-mutational interactions has been well documented in the genomics field [7, 8]. Increasing evidence suggests that interactions among polymorphic sites may confer a cumulative association of multiple mutations with many diseases [8]. Interactions among polymorphic sites have also effectively been used to infer ancestry and functional convergence in the human populations using mt-genome co-mutations [9]. Commonly used methods include tree ensembles, functional nodal mutations and single nucleotide polymorphism (SNP) based enrichment [10]. Important information about mt-genome evolutionary behavior, which is contained in the correlated changes between nucleotide positions both within genes and between genes, is not captured by these techniques. Despite strong evidence that mt-genome variation plays a role in the development and progression of complex human diseases, the mitochondrial genetic variation has been largely ignored in the context of co-mutations and particularly the mechanisms by which these co-mutations occur [11, 12]. Investigation of joint polymorphism effects can, therefore, improve the explanatory ability of genetics twofold. Firstly, the interaction between two informative genomic positions to explain a part of the trait heritability. Secondly, finding significant statistical links between mutations could provide strong indications of molecular-level interactions that differ between different populations [13].

Complex network science revolves around the hypothesis that the behavior of complex systems can be elucidated in terms of structural and functional relationships between their constituents employing a graph representation [14, 15, 16, 17]. The basis of the current study is that genome positions can impact each other and co-occur within genomes [18, 19]. The interaction between two or more genetic loci is referred here as the co-occurrence of nucleotide positions. Herein, nucleotide positions are network nodes and if they co-occur together with a co-occurrence frequency, they form an edge. Co-occurrence among a pair of nucleotide positions is a ratio of frequency between a nucleotide pair co-occurring and nucleotides in a pair occurring separately in a set of genomes. There are previous studies which have used genomic co-occurrences as a basis of the evolution of human H3N2 and Ebola viruses [19, 20]. These viral genome models have identified the co-occurring nucleotide clusters, apparently underpinning the dynamics of virus evolution since these clusters were antigenic regions of the viral capsid proteins [19, 20]. In another study, Shinde et al. [18] demonstrated the impact of codon position bias while forming nucleotide co-occurrences using human mt-genome. These studies have considered perfect nucleotide co-occurrence as causing factor for co-mutations. However, the role of the co-occurrence frequency in these studies remains unclear. Here, we thoroughly examined a set of networks associated with a range of co-occurrence frequencies and chose a particular co-occurrence frequency for further network construction. Whilst, pair-wise nucleotide co-occurrences can be straightforwardly perceived, however, the identification of larger sized functional units is not straightforward. Here, we used community detection algorithms to enumerate lists of modules formed within networks and described the functional relationships among nucleotide positions forming these modules.

We set out to develop a comprehensive approach to understand mitochondrial diversity using mitochondrial co-mutations. To this end, we conducted a comparative analysis of 24,167 sequenced mitochondrial genomes. The paper is organized as follows. In the first two sections, we described the level of diversity observed among underlying subpopulations concerning polymorphic site variations in human mitochondrial genomes. In the third section, we provided a simple framework to investigate co-mutations which are critical in underlying complex mitochondrial evolution. To this end, we constructed nucleotide co-occurrence networks which were used to identify modules of co-mutations and also made their comparison with corresponding random networks. We found both similarities and crucial difference among networks under consideration. In the fourth section, we identified local topological phenomena which were crucial agents for genomic networks' make-up. We listed down modules comprised of co-mutations and demonstrated that the identified modules indeed correspond to ancestry based associations. Overall, revealing the importance of co-mutational biases among different human subpopulations, our analysis identified local preferences which were key agents in forming mt-genome epistatic interactions.

1. Methods and Material

1.1. Acquisition of genomic data

We prepared an extensive collection of mitochondrial genomes of geographically diverse *Homo sapiens* populations (Fig. 1) from the Human Mitochondrial Database (Hmtdb) [21]. All downloaded genome sequences were in FASTA format. In total, the dataset comprised of 24,167 mitochondrial genome sequences from the five world continents, including 3426 African (AF), 2650 American (AM), 8483 Asian (AS), 8060 European (EU) and 1548 Oceanic (OC) genomes. Antarctica was excluded from the present analysis since no data was available. These continents were termed as genome groups. It should be noted that these genome groups are the multiethnic cohorts representing the range of admixture populations across the continent. A brief description of all the genomes and their origin is provided in S1 File.

1.2. Construction of nucleotide co-occurrence networks

Nucleotide co-occurrence calculations were carried out on each genome group independently. For each of these five genome datasets, we constructed primary nucleotide co-occurrence networks in which nodes represent genome positions, and edges between nodes represent co-occurring genomic mutations. There were total M primary nucleotide co-occurrence networks constructed for each genome group. Subsequently, we constructed five final nucleotide co-occurrence networks for five genome groups using these primary nucleotide co-occurrence networks. The methodology for constructing primary and final nucleotide co-occurrence networks schematically represented in Fig 1B and also given as follows:

1.2.1. Primary nucleotide co-occurrence network

All the steps of primary nucleotide co-occurrence networks construction were explained as follows. (1) Since the current study was based on the analysis of specific nucleotide position in mt-genome, we considered genome sequence data that was already end to end aligned. (2) All non-variable genome positions within samples of a genome group were removed. We were thus left with only the polymorphic genome positions. The count of polymorphic sites (N_P) was given in Table 1. (3) Using only polymorphic nucleotide positions, we calculated the frequency of occurrence of all the nucleotide pairs $f(x_iy_j) = N(x_iy_j)/M$ where, $N(x_iy_j)$ denoted the number of co-occurrence pairs (x_iy_j) at position (i, j) and M denoted total number of samples in a genome group. Consequently, we calculated the frequency of individual occurrence of single nucleotides $f(x_i) = N(x_i)/M$ and $f(y_j) = N(y_j)/M$ where, $N(x_i)$ and $N(y_j)$ denoted the number of single nucleotides at their respective positions i and j [19]. (4) Co-occurrence of two nucleotides (C_F) at position (i, j) was denoted as,

$$C_{Fi,j} = \frac{f(x_iy_j)^2}{f(x_i)f(y_j)} \tag{1}$$

For a particular co-occurrence frequency threshold here we termed it as network efficiency score (described in the next subsection), we constructed primary nucleotide co-occurrence networks. A network can be represented mathematically by an adjacency matrix (A) with binary entries.

$$A_{\text{ij}} = \begin{cases} 1 & \text{if } C_{Fi,j} >= \alpha \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

As each genome sequence has its own information of co-occurring genome positions, a total M primary networks were generated for each genome group.

1.2.2. Final nucleotide co-occurrence network.

Each of M primary nucleotide co-occurrence networks possess information of statistically significance edges. This simplest form of information was used to construct final nucleotide co-occurrence networks. Specifically, we listed down unique edges from all M networks of a genome group which make a final nucleotide co-occurrence network. The total of five final nucleotide co-occurrence networks was further used for network analysis and community detection.

Table 1: **Data statistics and the properties of final nucleotide co-occurrence networks.** Here, N_P, N, N_C, $\langle k \rangle$, $\langle C \rangle$, r, Q represent the co-occurrence frequency, number of disconnected components, number of variable sites, number of nodes, number of connections, the average degree, the clustering coefficient and the assortativity coefficient. All five networks were sparsed, disassortative and modular in nature. Network statistics of the largest connected component and the disconnected components are given in S1 Table and S2 Table.

Network	α	N_P	N	N_C	$\langle k \rangle$	$\langle C \rangle$	r	Q
AF	0.99970	3716	2310	13721	12	0.33	-0.36	0.51
AM	0.99959	3581	2412	30283	25	0.31	-0.61	0.22
AS	0.999854	5405	2293	32705	29	0.22	-0.18	0.21
EU	0.999760	4557	2456	47952	39	0.18	-0.62	0.30
OC	0.998800	1565	1208	7304	12	0.54	-0.25	0.54

1.3. Selection of network efficiency score (α)

Network efficiency score (α) was used to filter a group of edges required for network construction. To select an α value for each network, should require scanning of C_F values among all pairs of polymorphic positions. To consider a network with C_F values of least 10^{-4} precision would require the construction of 24,167 $*$ 10^4 networks in total, which would be a very intensive computational process. Therefore, we performed statistical sampling on each genome group interdependently by selection analysis of m samples from each M population. Sample size was determined by Cochran's sample size formula [22] with critical value ($z = 1.96$). As the population was finite, the sample size was corrected by Cochran's adjustment [22].

1.4. Analysis of nucleotide co-occurrence networks

The degree of a node (k_i), which can be defined as a number of edges connected to the node ($k_i = \sum_{j=1}^{N} A_{ij}$). The clustering coefficient (C) is a measure of the extent to which nodes in a network tend to cluster together. An average clustering coefficient of a network can be written as $\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$. Another property of the network which turns out to be crucial in distinguishing the individual networks was the assortative coefficient (r), which measure the tendency of nodes with the similar numbers of edges to connect. The assortative coefficient, r, was defined as the Pearson correlation coefficient of degree between pairs of linked nodes [23]. The value of r being zero corresponds to a random network whereas the negative(positive) values correspond to dis(assortative) networks. For a stringent comparison, the precise information of the number of nodes and edges of real-world nucleotide co-occurrence networks were used to construct randomized networks. This comparison allowed to estimate the probability that a randomized network with certain constraints has of belonging to a particular architecture, and thus assessed the relative importance of different architectures. The randomized networks of size N and average degree $\langle K \rangle$ were constructed using the Erdös Rényi random network model [24] by connecting each pair of nodes with the probability (p) which was equal to $\frac{\langle K \rangle}{N}$.

1.5. Detection of module structures in the network

In recent years, numerous approaches were proposed to determine the modular organization of complex networks wherein Girvan and Newman simplified graph-partitioning problems by introducing the concept of modularity. Modularity is conceptualized as the most widespread quantity for measuring the quality of a network partition, P. In its original definition [25], an unweighted and undirected network that has been partitioned into communities has modularity (Q) as $\frac{1}{2N_C} \sum_{C \in P} \sum_{i,j \in C} \left[A_{i,j} - \frac{k_i k_j}{2N_C} \right]$. The indices i and j run over the N nodes of the network whereas C runs over the modules of the partition, P. Modularity calculates the number of edges between all combinations of nodes belonging to the same module and relates it to the expected number of such edges for an equivalent random graph. Therefore, modularity assesses how well a given partition incorporates the edges within the modules. We used the Louvain algorithm

for community detection for our networks [26]. The Louvain method was a simple, efficient and easy-to-implement method for identifying communities in large networks. The python package of Louvain algorithm was used to enumerate module structures.

2. Results

2.1. Characterization of polymorphic sites

The starting point of our analysis was 24,167 sequenced mitochondrial genomes. We screened common and independent polymorphic sites present among five genome groups. A total of 18,824 polymorphic sites were screened from all the genome groups resulting in 7424 unique polymorphic nucleotide positions out of 16,929bp genome size (43% as also reported by [18]). The number of polymorphic sites (N_P) for each genome group is summerised in Table 1 and Fig 2A. The number of independent polymorphic sites found in each population was 13.9% (AS), 7.3% (EU), 5.4% (AF), 4.1% (AM) and 1.4% (OC). Alignment of each genome against the Reconstructed Sapiens Reference Sequence (RSRS) showed the most and the least diverged sequences have 73 and 22 polymorphic sites respectively (Fig 2B). The mean number of polymorphic sites for each group was appromixely 50%, thus the extent of genomic diversity in each genome group was similar (Fig 2B). We also assessed the contribution of each genome in providing the count of polymorphic sites in an individual genome group by removing each time a single genome from a genome group and calculating the number of polymorphic sites prescribed by the rest of genomes (Fig 2C). The minimum and maximum contribution observed was one and ten unique polymorphic sites per genome respectively. More than 99% of polymorphic sites in each genome group were contributed by individual genomes having 1-4 unique polymorphic sites (Fig 2C), suggesting a genome group with a higher number of genomes would yield more polymorphic sites. However, the trend was not straight-forward (Table 1 and Fig 2C). When we calculated the ratio of number of genomes in a genome group to number of polymorphic sites yielded, it was observed that AM had the largest ratio (1: 1.35), followed by AF (1: 1.08), OC (1: 1.01), AS (1: 0.64) and EU (1: 0.57). Overall, molecular sequence-based diversions concerning polymorphic sites were very well conserved across genome groups as well as independently maintained within genome groups.

2.2. Classifying mt-genome Diversity

We evaluated the diversity in genetic regions of the mt-genome: 13 protein-coding genes, 22 tRNA and two rRNA genes, four loci in the non-coding region and a few other non-coding positions dispersed throughout protein-coding and non-coding region. To quantitatively assess this broad diversity, we examined the observed polymorphisms among genes against gene size and possible substitutions that can arise at each codon positions (Table 2).

2.2.1. Variations in Protein-Coding Genes

Considering only substitutions, 5,465 of 11,387 nucleotide positions (48%) located in protein-coding regions (double counting few polymorphic sites detected in overlapping regions of two genes) were variable. To understand the role of each codon position (CP), we broke the list of polymorphic sites down into the individual CPs and compared these to the total number of possible changes at each position. It was found that 60% of polymorphic sites was located at CP 3, 26% at CP 1 and 14% at CP 2. There was a good correlation between the observed and the maximum number of possible changes at CP 1 in each of the 13 protein-coding genes ($r^2 = 0.7063$; Fig 4). *COX1* and *ND4* had the lowest proportion of observed polymorphic sites compared to all possible ones, and *ATP6* had the largest proportion (Table 2). At CP 2 there was not only a smaller number of polymorphic sites but also a weak correlation ($r^2 = 0.3692$) between the number of observed and possible changes at this position. In which *COX1*, *ND4* and *ND5* had the smallest proportion of polymorphic sites and both ATP syntase genes, *ATP6* and *ATP8*, has the largest proportion of polymorphic sites (Table 2). Interestingly, there was a very strong correlation ($r^2 = 0.9986$) among the observed, and the maximum number of polymorphic sites at CP 3 and all genes unanimously showed high diversity at CP 3 (Table 2). Overall, both ATP synthase genes *ATP6* and *ATP8*

Table 2: **Gene- and codon-wise polymorphisms among 13 protein-coding genes.** The observed polymorphisms in each of 13 protein-coding genes show mutational biases at codon positions. ATP genes contained the most polymorphisms. CP 2 showed fewer polymorphisms as compare to CP 1 and CP 3.

Gene Names (OMIM Ids)	Gene (%)	CP 1(%)	CP 2(%)	CP 3(%)
ATP6 (516060)	69	69	51	88
ATP8 (516070)	66	67	54	78
COX1 (516030)	41	25	11	88
COX2 (516040)	47	34	20	87
COX3 (516050)	50	39	26	86
CYB (516020)	55	49	26	90
ND1 (516000)	47	34	21	87
ND2 (516001)	46	36	19	84
ND3 (516002)	41	33	18	73
ND4 (516003)	42	27	13	87
ND4L (516004)	39	27	8	82
ND5 (516005)	47	36	17	88
ND6 (516006)	47	32	21	88

have demonstrated high mutation ability at all three CPs (Table 2) which corresponds to a higher gene diversity among all protein-coding genes.

Out of 5465 polymorphic sites, 4670 (85%) have two alleles, 741 (14%) have three alleles, and 54 (1%) have four alleles (Fig 2D). Moreover, when we screened polymorphic sites with two alleles, it turned out that 95% mutations were transitional, giving a transition: transversion ratio of 1:17.2, indicating transversion mutations are more likely to occur than transitions. The dominance of transition substitutions in the evolution of animal mt-genome (not just human mt-genome) has long been appreciated [27].

2.2.2. Variations in the control region, RNA genes and noncoding sites inside the coding region

26.8% of polymorphic sites were observed in the non-coding region (including tRNA), dispersed throughout the 5539bp region. Out of 1502 polymorphic sites among non-coding DNA, 1161 positions have two alleles, 280 positions have three alleles, and 61 positions have four alleles (Fig 2D). The ratio of observed transitions and transversions was 1: 8.2 which was much smaller than protein-coding genes. When comparing observed and possible changes at non-coding positions in each of the seven non-coding regions, a good correlation ($r^2 = 0.8296$; Fig 4F) was obtained with HVS-I having the highest proportion of observed polymorphic sites. For each non-coding region, the number of polymorphic sites was 26% in 16srRNA; 27% in 12srRNA; 44% in D-loop; 47% in noncoding sites inside the coding region; 59% in HVS-II; 67% in HVS-III; and 68% in HVS-I. Furthermore, out of 490 polymorphic positions among all tRNAs, 434 positions have two alleles, 48 positions have three alleles, and only eight positions have four alleles (Fig 2D). Similarly to the non-coding region, the ratio of translations and transversions was 1: 8.64. Given that the sizes of the various tRNAs are quite similar, varying only from 61 to 84 bp, there were some interesting differences in variability between the tRNA genes. The genes for the tRNAs *Met*, *Tyr*, *Glu*, *Leu* and *Asn* showed the fewest polymorphisms (less than 15), whereas all the other tRNAs showed between 16 and 33 polymorphisms, except threonine *Thr*, which had 49 polymorphisms (Fig 4E). This observation renders the relationship between observed polymorphisms and size totally uneven ($r^2 = 0.05$).

2.3. Evolution of mitochondrial co-mutations

Beyond simply measuring co-occurrence of these polymorphic sites, we evaluated the degree of correlation among pairs of polymorphic sites within the context of genomic associations. The selection of a co-occurrence frequency threshold (α) was a critical task. As zero α value would give co-occurrence among each mutation with all others whereas α equal to one would give only those pairs of mutations which have co-occurred perfectly in a genome group. In other words, zero α value would result in the globally connected network

(Fig 3B) and $\alpha = 1$ would result networks with many globally connected small sub-graphs (Fig 3E). Even though the α value attended a value as high as 0.999, networks remained very densely connected (Fig 3C). Therefore, it was reasonable to propose a criterion to select an α value, otherwise generated networks would be saturated structures holding no information about nucleotide co-occurrences. In order to tackle this, we plotted $\langle k \rangle$ and N_{LCC} against all the α values. We observed a surprising network phenomena where at a particular α value, $\langle k \rangle$ is small whilst N_{LCC} was large. At this point, networks were sparser as compared to previous α values (Fig 3D). By a sparse network, we would mean that the majority of the elements of the adjacency matrix were zeroes. After attending this α value, the network breaks into several disconnected components.

Having this notion, we chose a particular α value for each genome group and constructed primary nucleotide co-occurrence networks. Although the α value applied to each genome group was very high (close to 1), this value was sufficient to capture more than 50% of the polymorphic sites in each genome group (except in AS; S4 Table and S2 Fig). We believed the α shortlisted evolutionarily important interactions among polymorphic sites as it preserved statistically important network edges. Further, it should be noted that α values for each genome groups were different (Table 1). This observation is intuitive since the number of polymorphic sites was different and therefore a preference for nucleotide co-occurrences should be different in each genome group.

2.3.1. Co-mutations displayed intra- and inter- genomic loci adaptation

Analysis of pairs of co-mutations provided an essential understanding of the relationship between two independent genome locations. Co-mutations can be formulated within a particular mitochondrial functional region (intra-loci) or between two functional regions (inter-loci). We compared co-occurrence configuration present among nine mt-genome functional regions. The number of polymorphic sites was normalized by the total count of co-occurring polymorphic sites in a genome group, and this information was stored in the co-occurrence configuration matrix and used to construct Circos plots (Fig 5). Nine mt-genome functional regions, comprising of four oxidative phosphorylation (OXPHOS) complexes, two RNA and three non-coding regions, displayed different preferences to co-mutate with other functional regions. In particular, OXPHOS complexes I, IV and HVS functional regions have a large contribution to the overall co-mutation configuration in each network. It was not surprising as these functional regions had large genomic lengths. Further, to know more on how each functional region has contributed in forming co-mutations, we plotted the count of co-mutations in each functional region against the corresponding functional region size for intra- and inter-loci (Fig 5). It was observed that co-mutations among functional regions were evenly distributed among both intra- and inter- loci in AM and AS. However, intra-loci were more evenly distributed as compared to inter-loci. Interestingly, we reported few functional regions found to be outside the 95% confidence intervals in both intra- and inter-loci (Fig 5). For intra-loci, rRNA was an outlier in all populations, HVS in AF and OC whereas COX in AM and EU. For inter-loci, HVS was an outlier in AF, EU, and OC whereas COX in AM and AS. Apart from that ATP and miscellaneous regions were outliers in AM, tRNA in AS and rRNA in OC. These statistical outlier regions should have an assertive evolutionary role in a population. To explore this further, we studied how these groups were separated from each other. We first calculated Frobenius distances among each pair of five co-occurrence configuration matrices and then performed hierarchical clustering of the calculated pairwise Frobenius distances. A dendrogram clearly showed the separation of five genome groups into two main branches *i.e.* {AM, AS} and {AF, EU, and OC} (Fig 6).

To investigate global level co-occurrence preferences between functional regions, we analyzed unique co-mutations from all the genome groups. Fewer co-mutation pairs were formulated among intra-loci than those of inter-loci. This relationship between co-mutations and the spatial proximity would occur to be natural in mt-genome since all 13 protein-coding genes formed heavily interacting OXPHOS complexes [28, 29]. However, co-mutation pairs formed among OXPHOS complex I or ND genes which make 38% of total mt-genome, participated in 31% of inter-loci co-mutations but only 13% of intra-loci co-mutations. Both D-Loop and all three hypervariable regions displayed a tendency to co-mutate with almost all other mt-genome loci (Fig 5). The rRNA genes make-up 15% of total mt-genome but they participated in only 9% of co-mutating sites. All 22 tRNA genes which make 9% of total mt-genome participated in 10% of co-mutating sites. Overall, co-mutations dispersed among mt-genome functional regions showed that formation

of co-mutations was driven mainly by local adaptive forces among each group. There are also preferences among functional regions to formulate inter- or intra-loci co-mutations.

2.3.2. Co-occurrence networks exhibited similar network properties

Pair-wise nucleotide co-occurrences were not sufficient to fully reveal the underlying structure of functionally related nucleotide positions. As described in Fig 1, a nucleotide co-occurrence network was constructed for each genome group where polymorphic sites forming co-mutations constituted the nodes, and the edges represented co-occurring nucleotide positions. The number of nodes and edges forming final nucleotide co-occurrence networks were found to be different for each genome group (Table 1). This observation was intuitive since the formation of co-mutations is totally determined by the number of polymorphic sites formed and the significance of their relationships within each group. To get an overview of the network-level organization of the genomic interactions, the topological properties of nucleotide co-occurrence networks were analyzed. It should be noted that selected networks were sparsed which means they have a very small average degree (Table 1). Furthermore, we investigated two essential network topological properties to understand local interactions in individual networks. All the networks exhibited high average clustering coefficient, $\langle C \rangle$ values (Table 1), suggesting that the nodes of these networks are densely connected. All five networks also displayed a highly negative value of the degree-degree coefficient (r) (Table 1). The negative r value suggested that networks were dis-assortative where high degree nodes, on average, tend to attach to low degree nodes [23]. Many paths between nodes in these networks were dependent on high degree node(s). Many biological and social networks displayed negative r value, suggesting that failure of a high degree node in a disassortative network have more impact on the connectedness of the network [23, 15, 30].

Overall, nucleotide co-occurrence networks have shown both the properties of high clustering and the presence of dissortative nature. This observation suggests the presence of dense subgraphs within the network and the presence of hierarchical structures. To explore more about the local interaction patterns in nucleotide co-occurrence networks, we investigated module structures within these networks.

2.4. High cohesiveness of nucleotide co-occurrence communities

This work is a first attempt to uncover the hierarchical organization of nucleotide co-occurrence networks. The major challenge for identifying modules in a hierarchical organization is to decide the depth to decompose network, as the Louvain algorithm can fragment networks and subsequently modules until it finds the greatest partition. In order to avoid large numbers of smaller modules (size 2), the size of the second largest connected component was used to decipher submodules among each hierarchy of parent modules. The size of the second largest connected component was 11, 8, 9, 6 and 12 for AF, AM, AS, EU, OC genome groups respectively. We calculated the modularity coefficient (Q) for five final nucleotide co-occurrence networks and also for corresponding random networks (Table 1). Q value was clearly reduced in the randomized networks, relative to the original data, indicating that our results on real nucleotide co-occurrence networks were not trivially reproduced in random networks. A high Q value will manifest if networks are modular in nature. There were 557, 571, 552, 622 and 227 modules obtained for AF, AM, AS, EU, and OC genome groups respectively. The full list of modules is provided in S2 File. In these networks, small sized modules (size less than 20) were predominant alongside one or two large sized modules *i.e.* AF (size of 119), AM (270), AS (217 and 216), AS (294) and OC (104) (S4 Fig). Interestingly, large sized modules were only comprised of polymorphic sites from non-coding regions (except in OC). Similarly to co-mutations, we also noted that polymorphic sites among each module could be from any of mt-genome loci. For example, in OC population module 59 had polymorphic sites only from *COX1* gene, whereas module 3 had all polymorphic sites from different genes (S2 File). We noted that protein-coding functional regions have a predominant role in the formation of modules (S5 Table and S6 Table). Particularly, ND and COX participated in >65% and >40% of modules in each of five networks, respectively. Additionally, we also observed a total of 391 modules out of a total of 2529 modules where all polymorphic sites in the module were from a single functional group. Such mono-functional region modules were also prevailed by ND and COX functional regions, 70% and 14% of total mono-functional region modules, respectively (S6 Table).

Table 3: **Statistics of modules and ancestral lineage polymorphism.** Count of modules among each genome group and percentage of nodes participating in those modules (brackets) is given. Mixed modules observed to the confined of the largest size modules.

	AF	AM	AS	EU	OC
		Modules			
Ancestral allele modules	501 (79%)	529 (76%)	488 (69%)	583 (75%)	205 (86%)
Ancestral-variant modules	26 (4%)	12 (2%)	11 (1%)	12 (1%)	12 (5%)
Mixed modules	30 (17%)	30 (22%)	53 (30%)	27 (24%)	10 (9%)
Total modules	557	571	552	622	227
		Ancestral lineage polymorphism			
ALPS[1]	163	192	136	177	122
Modules with atleast one ALPS	45	60	40	49	70
Modules with >1 ALPS	16	25	16	13	24
Modules with all ALPS	8	10	5	5	10

[1] Ancestral lineage polymorphic sites (ALPS)

2.5. Modules of co-occurring polymorphic sites indicated ancestral relationships

To investigate if the modules identified from the analysis of the network structure were evolutionarily related, we examined polymorphic sites in the individual modules with ancestral alleles from RSRS. If a non-RSRS allele was present in more than 1% of samples in a genome group, we termed it an ancestral-variant allele. Here, we used conventional definition of SNP to define ancestral-variant allele. Thus, we assigned ancestral-variant information to all of the network modules and noted three distinct types of modules (also schematically showed in S5 Fig). In the first and most common (more than 90% of total modules), all polymorphic sites were closely related to ancestral alleles (Table 3) and we termed them ancestral allele modules. All the polymorphic sites in these ancestral allele modules had ancestral alleles (or non-RSRS alleles present in $< 1\%$ of samples). Ancestral alleles were reported to be common throughout human mt-genome tree [31] and were also observed in large numbers in our genome group data (Table 3). In the second type of module, all the polymorphic sites were ancestral-variant alleles. We termed them as ancestral-variant modules and were of our particular interest. Ancestral-variant modules were observed the least out of three types of modules, both in terms of module count and the number of polymorphic sites present in these modules (Table 3). In the third type of module, polymorphic sites in a module were a mixture of ancestral and ancestral-variant alleles and we termed them mixed modules. The polymorphic sites among these modules were hypothesised to be recently diverged. Mixed modules comprised of the large-sized modules, therefore even though the module count was found to be lower, these mixed modules still possessed a higher number of nodes (Table 3). Full lists of these modules were mapped it to all known haplogroups and showed that each polymorphic site had contributed to one or many haplogroups. Although this observation was intuitive since every new polymorphism in the mt-genome have been successfully characterized in defining haplogroups [31], this mapping has given valuable information that entire module structure can be related to a single mt-genome haplogroup. For a complete list of modules and corresponding haplogroups for each polymorphic site in a module, see S2 File. Further, we showed the relationship among modules corresponding to ancestral haplogroup lineage markers (or top-level haplogroups). Thus, we characterized an entire list of modules for whether their polymorphic sites were related to ancestral lineage markers. Information of ancestral lineage markers was taken from the Mitomap database, and polymorphic sites among each module were mapped to ancestral lineage markers. Out of the total 350 ancestral lineage markers, most of them were present in the American population, followed by European, African, Asian and Oceanic populations (Table 3). These ancestral lineage markers were also observed to participate in the formation of entire module structures and there were a total of 38 such modules structures obtained (Table 3; File S1). Out of the

observed 38 modules, where all nodes were ancestral lineage polymorphic sites, 23 were ancestral-variant modules, 13 were ancestral modules, and two were mixed modules. Since all polymorphic sites among these 38 modules were the ancestral lineage markers, it would be reasonable to say that not only sub-level haplogroups but also top-level haplogroup markers have shown a tendency to be associated to each other.

3. Discussion

We used comparative genome analysis to investigate 24,167 mt-genomes and devised a network model comprising pairs of co-occurring nucleotides over the length of the human mt-genome. The method presented here can provide a new perspective on epistatic mutation as well as serving as a comparative tool for intra-species evolution. Our study showed the presence of heterogeneity in both epistatic mutations and functional modules across investigated genome groups.

3.1. Polymorphism among mt-genome loci

Mitochondrial DNA is one of the most preserved genomes which is highlighted by the observation that maximum divergence from RSRS was only 73 bp ($\sim$ 0.005% of mt-genome) and a large proportion of mt-genome polymorphic sites possessed as few as two alleles. However, our study also reported as high as 43% mt-genome positions were at least once mutated. These observations strongly suggest that although individual mt-genomes have very few divergences, a large number of genome positions have been utilized to provide intra-species separation. Furthermore, the comparison of observed polymorphisms with gene size clearly showed two essential features in providing maximum functional level diversity with the minimum level of genomic changes. First, genetic conservation at CP 2 but not at CP 3, was key to provide a structural diversity of mt-genome complexes. Second, the restriction of mutations in structurally important genes of tRNA and rRNA. These two observations of biases against mutations at CP 2 and RNA genes were earlier reported by Pereira et al. with 5140 human mt-genome [32]. The similar positive selection of CP 2 and tRNA genes was also reported among mitochondrial genomes of other primates including Macaca, Papio, Hylobates, Pongo, Gorilla, and Pan whereby the constraint of selection was determined in each lineage by the ancestral state of each codon position [33]. Among non-coding genes, all three HVS regions have displayed a higher level of polymorphisms whereas genes of rRNA and tRNA have shown lower levels of polymorphism. Our study, apart from providing the detailed enlisting of diversity present among five genome groups, reiterated that both codon level mutation bias and restriction of mutations among RNA genes were more evident at the subpopulation level which earlier reported to be at global level. Furthermore, given the ubiquitous variation in mt-genome, genetic flexibility may have evolved as a mechanism to maintain OXPHOS under a range of environments.

3.2. Evolution of co-mutations

Despite the clear and reasonable biases against polymorphisms at CP2 in the protein-coding genes, our analysis indicated similar other biases were also evident. Firstly, our results with intra- and inter- loci adaptations clearly suggested the dominance of polygenic mutations in human mt-genome. On a protein level, the richness of co-ordinated mutations between mitochondrial complexes will affect protein-protein interactions within individual OXPHOS complexes as well as supercomplex interactions between electron chain transport complexes I, III, and IV in the respirasome. The communication among OXPHOS complexes is driven by a highly constrained selection [34] and also by protein-protein interactions of Mito-interactome [35]. Second, our analysis highlights regions of the mt-genome rich in co-mutations and thus provides evidence of epistasis. In particular, a large portion of COX genes co-mutate in AS and AM populations whereas in AF, EU and OC populations, there was greater epistasis in functional regions of HVS. OXPHOS complexes consist of both nuclear and mitochondrial proteins, and nuclear-mitochondrial interactions are known to contribute epistatically in shaping mitochondrial evolution. However, epistasis has also been observed among mitochondrial genes, for example, the joint effect of genetic variants has been reported in Han Chinese family [36]. These two point mutations are known to act synergistically causing migration of gonadotropin-releasing hormone neurons [36]. Also, epistatic phenomena have been widely observed among

mitochondrial tRNA genes guided by homoplasy [37]. There were similar other mitochondrial co-mutations which have reported to have the role in mitochondrial diseases [38]. We extended information accompanied by the two-loci genome model broadly by constructing nucleotide co-occurrence networks.

Third, similar to polymorphic sites, co-mutations also showed biases at the subpopulation level. Consistent with the proposed importance of mt-genome variation in human adaptation, regional haplogroups are generally founded by one or more functionally significant polypeptide, tRNA, rRNA, and control region variants. These variants are retained in the descendant mt-genomes creating the haplogroups [39]. Therefore it was intuitive to observe patterns of variants at the subpopulation level. Genome group-wise comparison of co-mutations associated among mt-genome functional regions has helped in classifying these five human subpopulations into two prominent groups *i.e.* {AF, EU, OC} and {AS, AM}. This result was supported by a global mt-genome mutational phylogeny which implicated few peculiar routes of human migrations [31]. Asian haplogroup M and European haplogroup N arose from the African haplogroup L3 [40]. Haplogroup M gave rise to the haplogroups A, B, C, D, G, and F [40] in which Haplogroups A, B, C, and D populated East Asia and the Americas. In Europe, haplogroup N led to the European haplogroups H, J, T, U, and V [41] whereas Haplogroups S, P, and Q are found in Oceania [31]. Overall, variations probed by epistatic interactions have provided local preferences among different mt-genome loci. These local preferences might have helped in not only forming the closed-assembly of OXPHOS complexes but also classifying subpopulations.

3.3. Discontinuous transition in nucleotide co-occurrence network

In the nucleotide co-occurrence networks, we found that α was the main edge filter when intersecting networks and also produced the optimum network in relation to size and architecture. In our network models, the emergence of sparse networks was not a smooth, gradual process: the very dense largest connected component collapsed into a sparse largest connected component through a discontinuous transition (Fig 3). For all five genome groups, we encountered such a distinct phenomenon. A similar critical phenomenon was first observed by Erdös and Rényi through their random network model where the isolated nodes and tiny components observed for small $\langle K \rangle$ would collapse into one largest connected component [24]. Similarly, such phenomenon was also observed in computer traffic network and biomolecular interaction network in acute lung injury, etc. [43, 44]. However, in these networks with a largest connected component, several smaller disconnected components were observed similar to our networks. Interestingly, the nature of such transformations was earlier related to neutral genetic drift and hypothesized that biological processes have proceeded in discontinuous transitions [42].

We selected edges for inclusion in co-occurrence networks based on their best fit to a network sparseness. Sparseness is one of the essential property of real-world networks, particularly biological networks since biological networks are usually large, and there is evolutionarily cost involved to form more links and hence links are more difficult to create. It is well known that co-mutational events are very selective events which require a group of supporting mechanisms to perform co-activity [19]. Previously, the similar approximation for the inclusion of edges was used in microbial interaction networks using methods like WGCNA [45]. Using a network sparseness or similar data-driven approach avoids entirely arbitrary selections of network edges and provides a uniform rationale that can be implemented to generate co-occurrence network structures across different genome datasets. Therefore, it was reasonable to choose an α value where a network should have both the lowest value of $\langle k \rangle$ value and the largest component with a higher count of nodes.

3.4. Hierarical modularity

Following network construction, we utilized a modularity maximization algorithm to detect communities in the network. There was clear evidence for hierarchical modularity in our genome datasets, and the modular structure of the networks at all levels of the hierarchical patterns was reasonably similar across genome groups, suggesting that mt-genome functional modularity is likely to be a replicable phenomenon. This study provided a complete listing of the current knowledge of mt-genome variation in the human population, also with respect to higher level associations with hierarchical modules. Overall, we demonstrated that molecular changes, such as mutations, were not randomly distributed across the genome, but instead concentrated within modules. Modularity was one of the main features of nucleotide co-occurrence networks,

and evolutionary processes may favor the emergence of modularity by a combination of molecular interactions and natural selection [46]. Similar hierarchical modularity in brain network was related to functional regions in the brain and sub-set of brain functions have been reported to be associated among each hierarchy [47]. Therefore, it was reasonable to say that selection may favor modularity allowing both the specificity and autonomy of functionally distinct subsets of genomic positions. In this sense, the concentration of genomic positions within modules provided a way to understand module integration, favoring distinct functional roles developed by genomic positions in distinct modules. In the human mt-genome, modules were associated with mitochondrial subcomplexes that act in distinct steps of the electron transport assembly and function. Thus, the closed assembly of mitochondrial complexes may favor the emergence of highly integrated genomic subunits, in which effects of pairwise interactions may also activate indirect effects on non-interacting genomic positions associated with the same function [34]. Based on these results, we would expect that genome positions connecting modules were more conserved across evolution or, at least, less prone to failures that alter their function.

3.5. Co-occurrence patterns among network modules

It would be expected that module level associations reflected evolutionary relationships among underlying genomic positions as each module were constituted of ancestrally similar genomic polymorphisms. Our results of modules added that distinction between the ancestral and the derived mitochondrial polymorphisms was clear in very few cases where the entire module was made up of ancestral-variant polymorphic sites. However, a large number of modules (more than 90% of total modules) were made-up of ancestral polymorphic sites. In addition, the large number of nodes in mixed modules were of ancestral origin (S2 File). Using the list modules among all five networks, it would be reasonable to assert that contemporary human mt-genome nucleotide bases most closely resembled the ancestral state and very few of them were ancestral-variants. This observation was in agreement with previous studies which found co-occurrence among nucleotide positions to be higher between genetically similar taxa [48]. This fact was widely observed in our data as both sub-level, and top-level haplogroup markers were associated with each other in a closed group of network modules. Haplogroup level association of genetic markers, particularly Haplogroup T markers, recently shown to be involved in risk of colorectal cancer [49]. Therefore, these evolutionarily closed associations suggest that interactions among nucleotide positions might evolve within genetically related genomic polymorphisms (more likely of having similar functionality) responding to intra-species adaptation. These associations between mutations have been reported to be a major driver of co-occurrence patterns in intra-species evolution. For instance, modules of H3N2 viral genome co-mutations were found to be important agents in mediating protein binding sites [19].

Beyond identifying network modules, understanding their formation would require an extension of the described approaches to quantification of each module using the evolutionary information they possess. Here, we used simplistic information possessed by each genome position in terms of their underlying ancestral marker information. Previously, this ancestral marker information has been used in order to define taxa (precisely haplogroups) in mitochondrial phylotree which have provided exact mapping of mitochondrial signatures to infer the routes of human intra-species diversification events in the past [50, 51]. Similarly, our nucleotide co-occurrence modules have provided a detailed listing of mitochondrial co-mutations which were ancestrally associated together.

4. Conclusion

We constructed and investigated the human mt-genome nucleotide co-occurrence networks among geographical regions using a genomics and network theory framework. Our principal result was that mitochondria undergo substantial levels of co-mutational biases. Codon-level mutation bias, particularly at CP 2, and restriction of mutations in RNA genes was even evident at the continental level which was earlier reported to be among the global human population. The analysis highlighted regions of the mt-genome that were rich for co-mutations and thus provided evidence of epistasis. In particular, a large portion of COX and ND genes found to be co-mutated in AS and AM populations whereas in AF, EU and OC populations, there

was greater epistasis between regions of HVS and ND. Our networks identified differences in epistasis and codon-bias between human populations. From the co-occurrence network analysis of regional co-mutations, it was of great interest to investigate and verify the different co-occurrence patterns among mutations of various geographical regions. This analysis presented here can be extended further to study the complexity of the mt-genome evolution by forming various geographical groups as well as to understand alterations in personnel traits leading to complexity in mt-genome evolution. This understanding can provide the envelope of the network information which encodes the changes during the progression of the disease, where information of genomic alteration with time is available.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

PS acknowledges Inspire fellowship (IF150200) from the department of science and technology (DST), Government of India. SJ thanks the support by grant of the ministry of education and science of the Russian Federation (Agreement No. 074-02-2018-330) and grants of DST (EMR/2016/001921) and the council of scientific and industrial research (CSIR, 25(0293)/18/EMR-II), government of India. RKV acknowledges CSIR-NET fellowship (roll no.: 305089) from CSIR, Government of India.

Data Availability Statement

All data sources and related information is given in the manuscript and associated supporting information files.

Supporting information

S1 File. Information of genome samples considered in the study.
S2 File. Information of modules indetified in the study.
S1 Table. Network properties of Largest connected component.
S2 Table. Network properties of all disconnected components together except Lcc.
S3 Table. Level-wise community detection in five nucleotide co-occurrence networks.
S3 Table. Comparison of polymorphism α_{pre} and α_{post}.
S4 Table. Distribution of modules having atleast one polymorphic site among mt-genome functional groups. ND and COX functional groups were having maximum participation among modules.
S5 Table. Modules comprising all nodes as ancestral lineage polymorphic sites. The count was dominated by ND and COX.
S1 Fig. Degree distribution of five nucleotide co-occurrence networks.
S2 Fig. Gene-wise comparison of polymorphism before and after α.
S3 Fig. Transition and transversion.
S4 Fig. Distribution of module sizes.
S5 Fig. Identification and characterization of network modules.

Author Contributions

Conceptualization: PS, SJ
Data curation: PS
Investigation: PS, HW
Methodology: PS, SJ, RKV

Supervision: SJ
Writing - original draft: PS, SJ, HW, AZ, RKV

References

[1] Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. American journal of human genetics. 1995;57(1):133.

[2] Van Der Walt JM, Dementieva YA, Martin ER, Scott WK, Nicodemus KK, Kroner CC and et al. Analysis of European mitochondrial haplogroups with Alzheimer disease risk. Neuroscience letters. 2004;365(1):28-32.

[3] Hudson G, Nalls M, Evans JR, Breen DP, Winder-Rhodes S, Morrison KE and et al. Two-stage association study and meta-analysis of mitochondrial DNA variants in Parkinson disease. Neurology. 2013;80(22):2042-2048.

[4] Kenney MC, Hertzog D, Chak G, Atilano SR, Khatibi N, Soe K and et al. Mitochondrial DNA haplogroups confer differences in risk for age-related macular degeneration: a case control study. BMC medical genetics. 2013;14(1):4.

[5] Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S and et al. Natural selection shaped regional mtDNA variation in humans. Proceedings of the National Academy of Sciences. 2003;100(1):171-176.

[6] Raule N, Sevini F, Li S, Barbieri A, Tallaro F, Lomartire L and et al. The cooccurrence of mt DNA mutations on different oxidative phosphorylation subunits, not detected by haplogroup analysis, affects human longevity and is population specific. Aging cell. 2014;13(3):401-407.

[7] Lane HY, Tsai GE, Lin E. Assessing gene-gene interactions in pharmacogenomics. Molecular diagnosis and therapy. 2012;16(1):15-27.

[8] Chen JB, Chuang LY, Lin YD, Liou CW, Lin TK, Lee WC and et al. Preventive SNPSNP interactions in the mitochondrial displacement loop (D-loop) from chronic dialysis patients. Mitochondrion. 2013;13(6):698-704.

[9] Ioannidis JP, Ntzani EE, Trikalinos TA, ContopoulosIoannidis DG. Replication validity of genetic association studies. Nature Genetics. 2001;29:306-309.

[10] Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genetics. 2004; 5:32.

[11] Boles RG, Roe T, Senadheera D, Mahnovski V, Wong LJ. Mitochondrial DNA deletion with Kearns Sayre syndrome in a child with Addison disease. European journal of pediatrics. 1998;157(8):643-647.

[12] Goodman JE, Mechanic LE, Luke BT, Ambs S, Chanock S, Harris C. Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis. International journal of cancer. 2006; 118(7):1790-1797.

[13] Hartwig FP. SNP-SNP Interactions: focusing on variable coding for complex models of epistasis. Journal of Genetic Syndrome and Gene Therapy. 2013;4(189):10-4172.

[14] Albert R, Barabási AL. Statistical mechanics of complex networks. Review of Modern Physics. 2002;74:47.

[15] Shinde P, Yadav A, Rai A, Jalan S. Dissortativity and duplications in oral cancer. The European Physical Journal B. 2015 Aug 1;88(8):197.

[16] Shinde P, Jalan S. A multilayer protein-protein interaction network analysis of different life stages in Caenorhabditis elegans. EPL. 2015;112(5):58001.

[17] Whitwell HJ, Blyuss O, Menon U, Timms JF, Zaikin A. Parenclitic networks for predicting ovarian cancer. Oncotarget. 2018;9(32): 22717.

[18] Shinde P, Sarkar C, Jalan S. Codon based co-occurrence network motifs in human mitochondria. Scientific reports. 2018;8(1):3060.

[19] Du X, Wang Z, Wu A, Song L, Cao Y, Hang H, Jiang T. Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. Genome research. 2008;18(1):178-187.

[20] Deng L, Liu M, Hua S, Peng Y, Wu A, Qin FX and et al. Network of co-mutations in Ebola virus genome predicts the disease lethality. Cell research. 2015;25(6):753.

[21] Rubino F, Piredda R, Calabrese FM, Simone D, Lang M, Calabrese C and et al. HmtDB, a genomic resource for mitochondrion-based human variability studies. Nucleic acids research. 2011;40(D1):D1150-9.

[22] Cochran WG. Sampling techniques. John Wiley and Sons. 2007.

[23] Newman MEJ. Mixing patterns in networks Physical Review E. 2003;67:026126.

[24] Erdös P, Rényi A. On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences. 1960;5(1):17-60

[25] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E. 2004;69:026113.

[26] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;10:P10008.

[27] Lanave C, Tommasi S, Preparata G, Saccone C. Transition and transversion rate in the evolution of animal mitochondrial DNA. Biosystems. 1986; 19:273-283.

[28] Wong JT. A co-evolution theory of the genetic code. Proceedings of the National Academy of Sciences. 1975;72(5):1909.

[29] Thompson JN. The coevolutionary process. University of Chicago Press. 1994; ISBN: 0226-79759-7.

[30] Sarkar C, Yadav A, Jalan S. Multilayer network decoding versatility and trust. EPL. 2016;113(1):18007.

[31] Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D and et al. An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic acids research. 2006;35:D823-8.

[32] Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, Costa S and et al. The diversity present in 5140 human mitochondrial genomes. The American Journal of Human Genetics. 2009;84(5):628-640.
[33] Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K and et al. The role of selection in the evolution of human mitochondrial genomes. Genetics. 2006;172(1):373-387.
[34] da Fonseca RR, Johnson WE, O'Brien SJ, Ramos MJ, Antunes A. The adaptive evolution of the mammalian mitochondrial genome. BMC genomics. 2008;9(1):119.
[35] Schweppe DK, Chavez JD, Lee CF, Caudal A, Kruse SE, Stuppard R and et al. Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. Proceedings of the National Academy of Sciences. 2017;114(7):1732-1737.
[36] Wang F, Huang GD, Tian H, Zhong YB, Shi HJ, Li Z and et al. Point mutations in KAL1 and the mitochondrial gene MT-tRNA cys synergize to produce Kallmann syndrome phenotype. Scientific reports. 2015;5:13050.
[37] Moreno-Loshuertos R, Ferrn G, Acn-Prez R, Gallardo ME, Viscomi C, Prez-Martos A and et al. Evolution meets disease: penetrance and functional epistasis of mitochondrial tRNA mutations. PLoS genetics. 2011;7(4):e1001379.
[38] Morrow EH, Camus MF. Mitonuclear epistasis and mitochondrial disease. Mitochondrion. 2017;35:119-22.
[39] Wallace DC. Mitochondrial DNA variation in human radiation and disease. Cell. 2015;163(1):33-38.
[40] Wallace DC, Brown MD, Lott MT. Mitochondrial DNA variation in human evolution and disease. Gene. 1999;238(1):211-30.
[41] Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R and et al. Classification of European mtDNAs from an analysis of three European populations. Genetics. 1996;144(4):1835-50.
[42] Fontana W, Schuster P. Continuity in evolution: on the nature of transitions. Science. 1998;280(5368):1451-5.
[43] Liu R, Li M, Liu ZP, Wu J, Chen L, Aihara K. Identifying critical transitions and their leading biomolecular networks in complex diseases. Scientific reports. 2012;2:813.
[44] Ohira T, Sawatari R. Phase transition in a computer network traffic model. Physical Review E. 1998;58(1):193.
[45] Jackson MA, Bonder MJ, Kuncheva Z, Zierer J, Fu J, Kurilshikov A and et al. Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. PeerJ. 2018;6:e4303.
[46] Clune J, Mouret JB, Lipson H. The evolutionary origins of modularity. Proc. R. Soc. B. 2013;280(1755):20122863.
[47] Meunier D, Lambiotte R, Fornito A, Ersche K, Bullmore E. Hierarchical modularity in human brain functional networks. Frontiers in neuroinformatics. 2009;3:37.
[48] Chaffron S, Rehrauer H, Pernthaler J, von Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome research. 2010:104521.
[49] Li Y, Beckman KB, Caberto C, Kazma R, Lum-Jones A, Haiman CA, Le Marchand L, Stram DO, Saxena R, Cheng I. Association of genes, pathways, and haplogroups of the mitochondrial genome with the risk of colorectal cancer: the multiethnic cohort. PloS one. 2015;10(9):e0136796.
Steinberg B, Ostermeier M. Environmental changes bridge evolutionary valleys. Science advances. 2016;2(1):e1500921.
[50] Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N and et al. The promise of discovering population-specific disease-associated genes in South Asia. Nature genetics. 2017;49(9):1403.
[51] Derenko MV, Grzybowski T, Malyarchuk BA, Czarny J, Micicka-liwka D, Zakharov IA. The presence of mitochondrial haplogroup X in Altaians from South Siberia. The American Journal of Human Genetics. 2001;69(1):237-41.

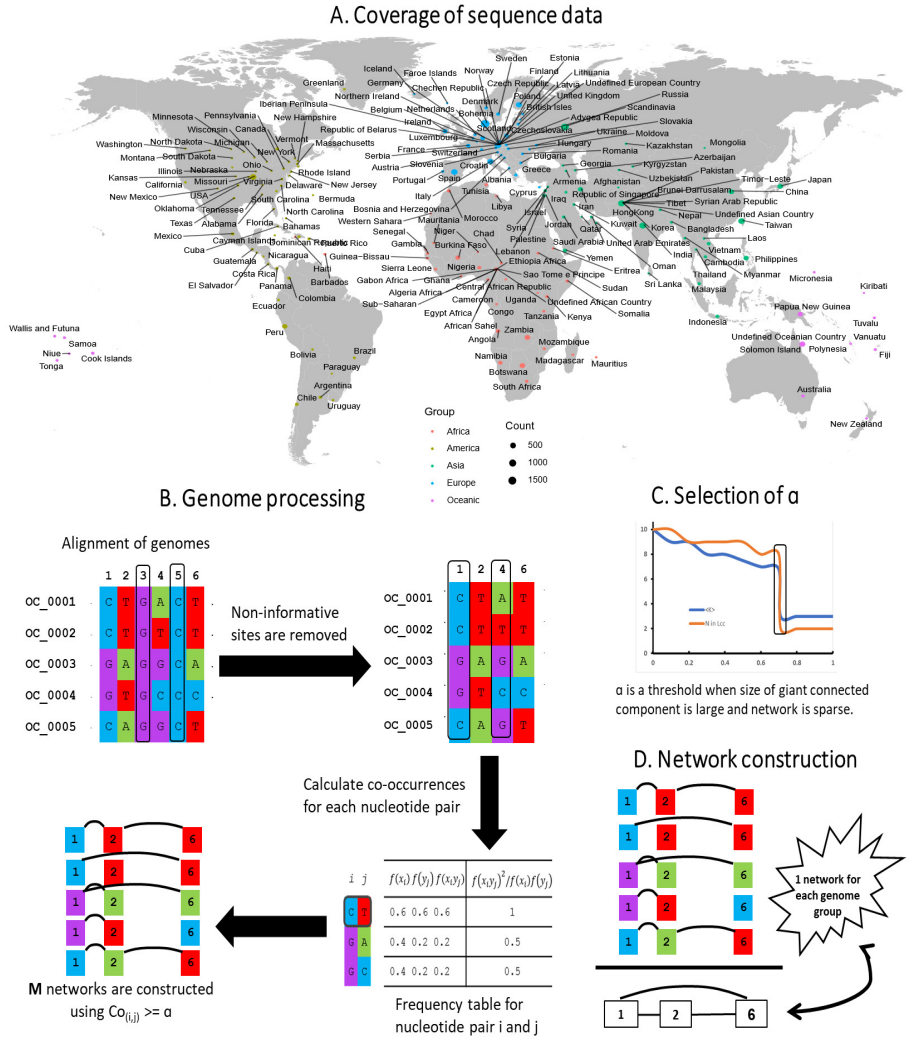

Figure 1: **Schematic representation of mtDNA nucleotide co-occurrence network construction and analysis.** (A) World map shows sequence data taken for the current study covered a good distribution across the entire globe. (B) A schematic diagram is drawn for a genome group with 5 sample sequences. Schematic diagram depicts (1) Alignment of genomes. All mitochondrial genomes in a genome group were end to end aligned, and therefore all aligned sequences had the same length. (2) Removal of non-informative sites. A genome position consist of a single nucleotide among all samples was removed from the analysis. (3) Calculation of co-occurrence frequency (C_F) for each nucleotide pair. (C) Selection of network efficiency score (α). α was a threshold when the average degree ($\langle k \rangle$) of a network is small, and the size of the largest connected component (N_{LCC}) is high. For each genome group, α was found to be different. (D) Each genome group has M genomes *i.e.* M networks. A unique list of edges was picked up from M networks from a genome group to construct a final weighted network for M networks in a genome group. Likewise, five networks were constructed for five genome groups.

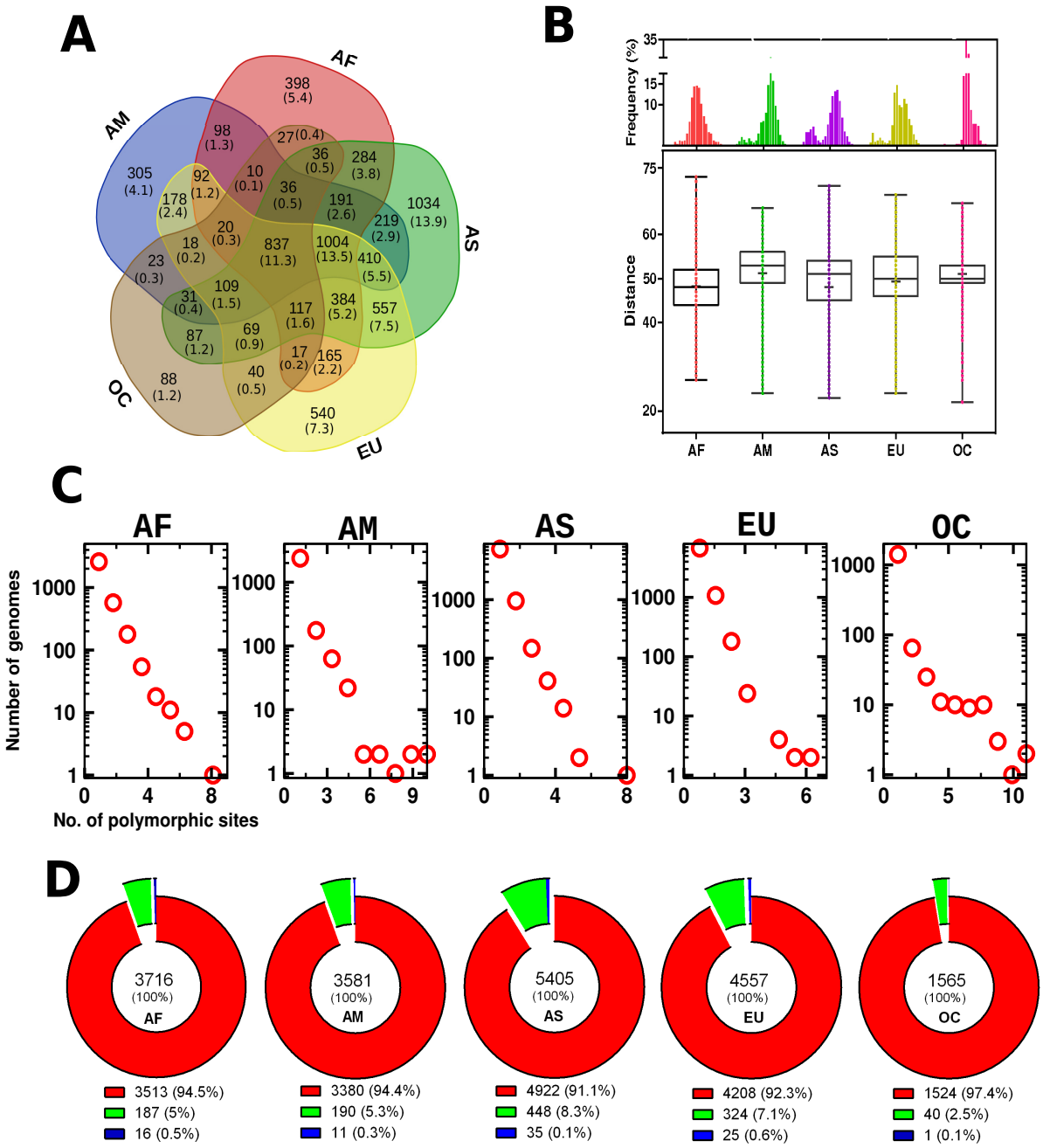

Figure 2: **Distribution of variable sites across genome groups and the distance between each genome from the reference sequence (RSRS).** (A) Variable sites occurred at similar genome positions in a different set of genome groups such as two, three, four and all five. Additionally, certain polymorphic sites were only found in a particular group. AS displayed the most number of unique polymorphic sites. The largest portion of variomes commonly occurred among {all five genome groups: 11.3%}, {AF, AM, AS and EU: 13.5%}, {AM,AF and AS : 5.5%}, {AF, AS and EU: 5.2%}, {AS and EU: 7.5%} sets of genome groups. These values indicated that human population habituating in AF, AM, AS and EU geographical regions shared more common polymorphism as compared to the OC geographical region. Also, AF and AS genome groups shared a large number polymorphisms with individual AM and EU populations. (B) The distance between each genome and RSRS sequence. All genome groups had a similar spread of distances. Multiple numbers of peaks were observed among AM, AS and EU populations. Each peak here would signify a subpopulation with a certain range of polymorphic sites. This would not be surprising as these genome groups span entire continents, there could be a huge amount of diversity within each continent. (C) Each genome posses new mutations as well as ancestral mutations. Set of all these mutations would make polymorphic sites in a genome group. Individual genome contribution in genome group follows behaviors where a large number of genomes gave only 1-4 unique polymorphic sites. (D) The majority of polymorphic sites possess two alleles (red) whereas polymorphic sites with three alleles (green) and four alleles (blue) were in less number.

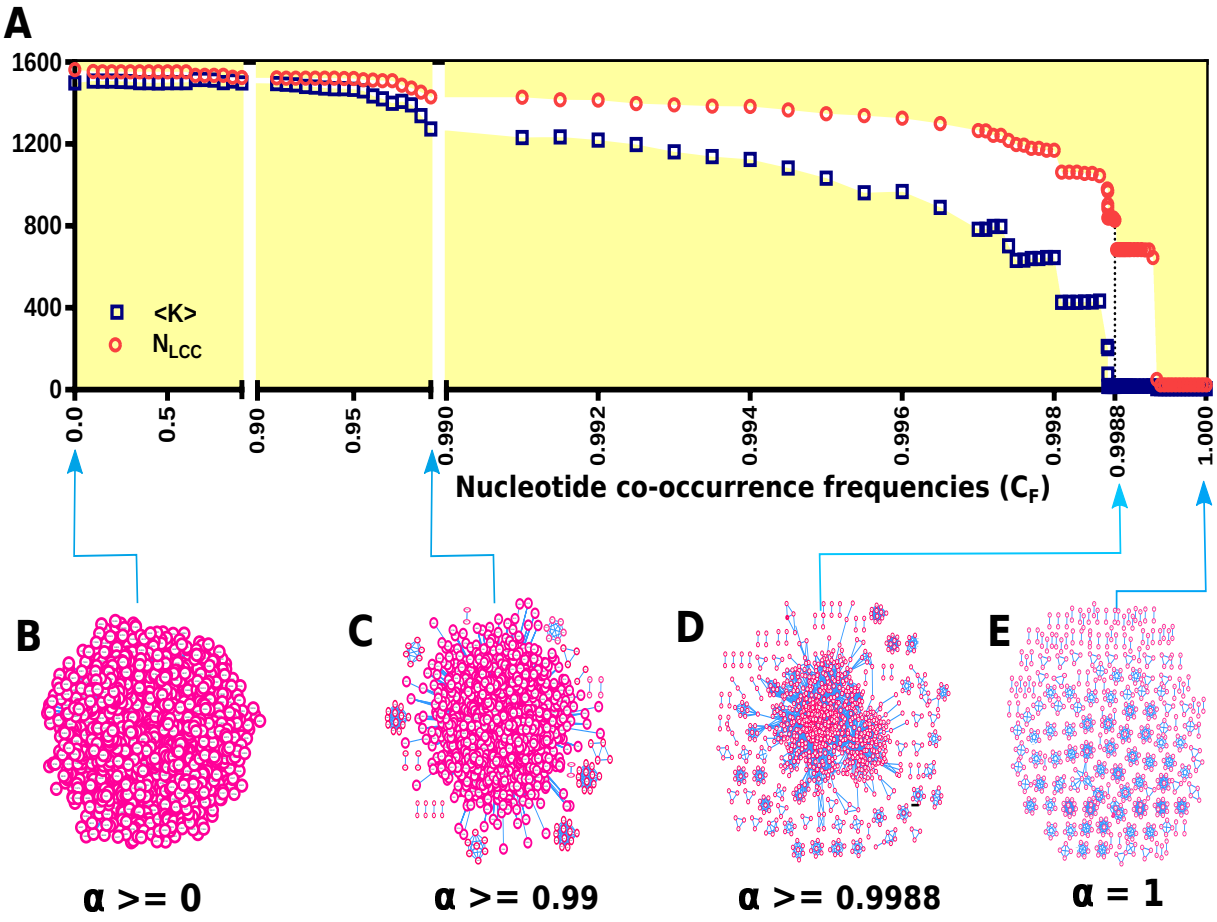

Figure 3: **Evolution of mitochondrial nucleotide co-occurrence network.** (A) The relative size of the largest component and the average degree of the largest component are plotted against nucleotide co-occurrence frequencies (C_F). The figure illustrates that at a particular α value (for OC, $\alpha = 0.9988$) nucleotide co-occurrence network has both a smaller value of the average degree and the number of nodes in the largest component are sufficiently in large number. We picked this C_F value for network construction. (B-E) A sample network at different C_F values show how it evolves from a globally connected network to the network with many disconnected components.

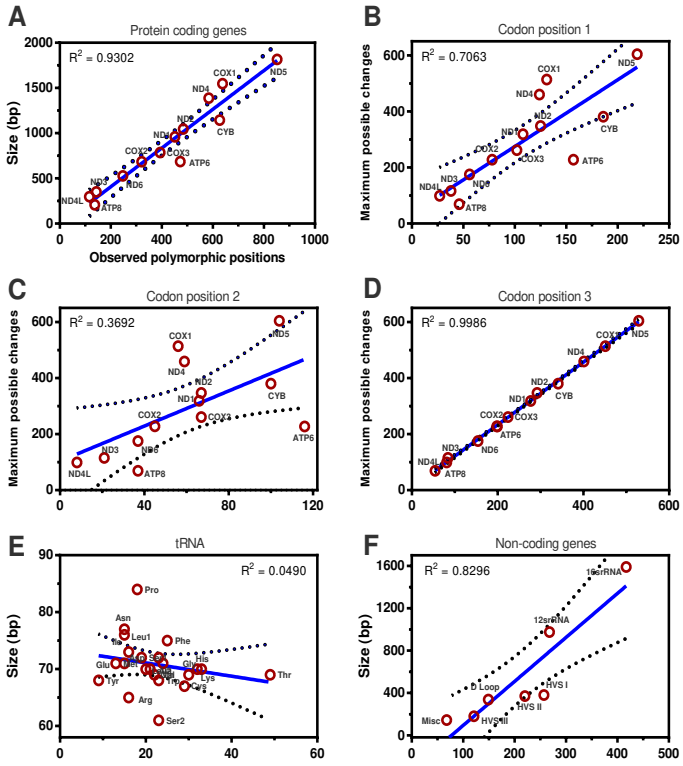

Figure 4: **Diversity among individual genome regions.** Correlation between the observed polymorphic positions and the gene size (bp) or maximum possible changes in (A) the 13 protein-coding genes, (B-D) the codon positions 1, 2 and 3 among the 13 protein-coding genes, (E) tRNA genes and (F) non-coding genes.

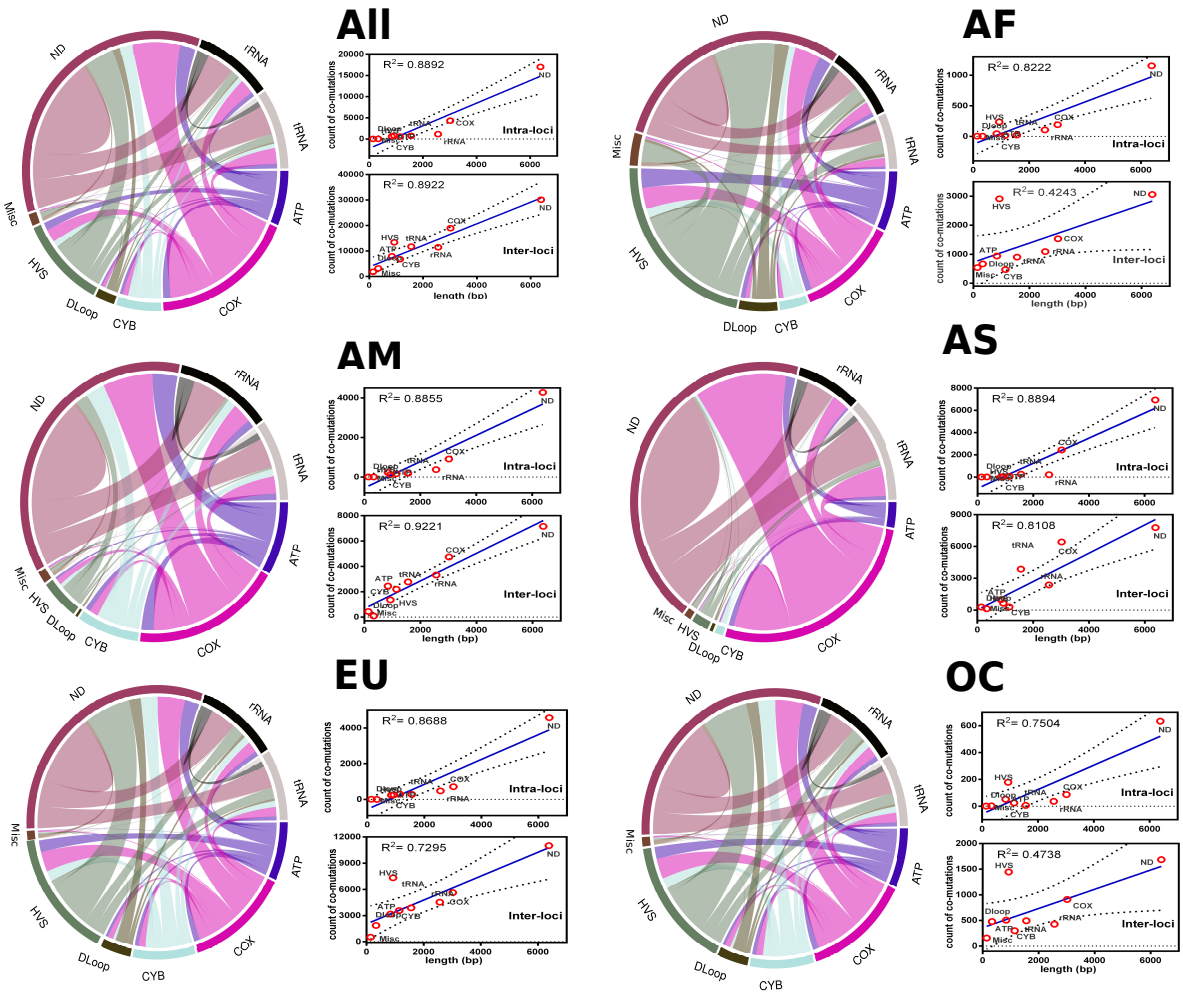

Figure 5: **Comparison of polymorphism among genomic loci.** A co-occurrence configuration in the human mtDNA nucleotide co-occurrence network consisting of nine functional regions. These nine regions were four mitochondrial complexes (ND, COX, ATP, and CYB), three non-coding regions (DLoop, HVS, Miscellaneous) and two RNA regions (rRNA and tRNA). Links or ribbons represent the frequency of CO pairs between two genomic loci. The four functional regions make the mitochondrial oxidative phosphorylation machinery. In large part, mtDNA-specified proteins are components of respiratory complexes: Complexes I (NADH dehydrogenase), Complex III (cytochrome c), Complex IV (cytochrome c oxidase) and Complex V (ATP synthase). The regression line is shown in blue (rigid) colour whereas 95% confidence interval is shown with black (dotted) colour. Circular maps were constructed using the rcirclize package in R.

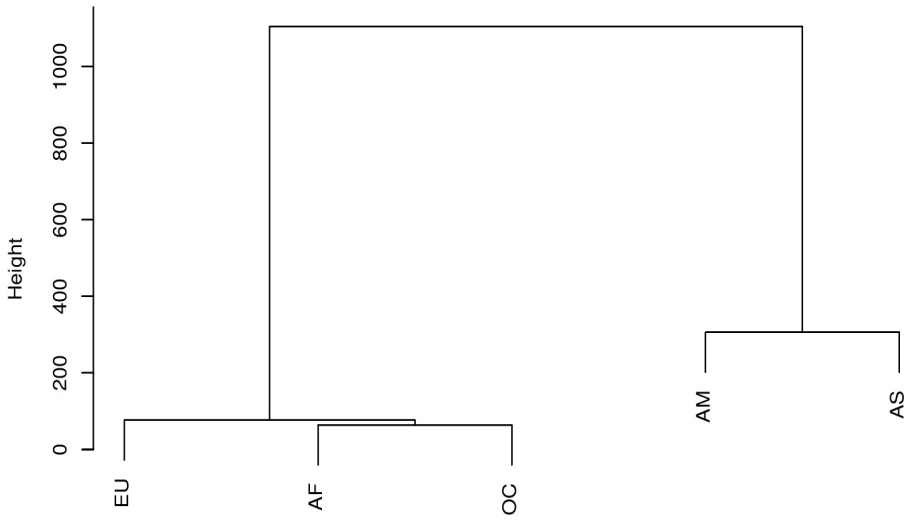

Figure 6: **Relationships among genome groups.** These relationships are predicted based on polymorphisms shown by their functional genomic loci. Five genome groups were classified as two main branches of the dendrogram , *i.e.* {AM, AS} and {AF, EU, and OC}.