bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Impact of modular mitochondrial epistatic interactions on the evolution of human subpopulations

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

[1] *Discipline of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Khandwa road, Simrol, Indore 453552, India.* [2] *Chemical Engineering, Imperial College London, London, UK.* [3] *Department of Applied Mathematics and Centre of Bioinformatics, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia.* [4] *Department of Mathematics and Institute for Womens Health, University College London, London, WC1E 6BT, UK.* [5] *Complex Systems Lab, Discipline of Physics, Indian Institute of Technology Indore, Khandwa road, Simrol, Indore 453552, India.*
Corresponding authors: [a] *pramodshinde119@gmail.com* , [b] *sarika@iiti.ac.in*

Abstract

Investigation of human mitochondrial (mt) genome variation has been shown to provide insights to the human history and natural selection. By analyzing 24,167 human mt-genome samples, collected for five continents, we have developed a co-mutation network model to investigate characteristic human evolutionary patterns. The analysis highlighted richer co-mutating regions of the mt-genome, suggesting the presence of epistasis. Specifically, a large portion of COX genes was found to co-mutate in Asian and American populations, whereas, in African, European, and Oceanic populations, there was greater co-mutation bias in hypervariable regions. Interestingly, this study demonstrated hierarchical modularity as a crucial agent for these co-mutation networks. More profoundly, our ancestry-based co-mutation module analyses showed that mutations cluster preferentially in known mitochondrial haplogroups. Contemporary human mt-genome nucleotides most closely resembled the ancestral state, and very few of them were found to be ancestral-variants. Overall, these results demonstrated that subpopulation-based biases may

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

favor mitochondrial gene specific epistasis.

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

---

1. Introduction

Genetic polymorphism varies among a species as well as within genomes and carries 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 (Chen, at al., 1995). The human mt-genome is very small (16.6 kb), maternally inherited, evolves in both neutral and adaptive fashions, 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 referred to as mt-genome haplogroups (Chen, at al., 1995). These haplogroups are formed as a result of the sequential accumulation of mutations through maternal lineages. Since mitochondria are essential to cellular metabolism, mt-genome variation has been associated with multiple complex diseases including Alzheimer's disease in haplogroup U (Van, et al., 2004), idiopathic Parkinson disease within JT haplogroup (Hudson, et al., 2013) and age-related macular degeneration in the JTU haplogroup cluster (Kenney, et al., 2013). 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 (Mishmar, et al., 2003; Shriner, and Keita., 2016; Zanellati, et al., 2015).

Conventionally, analyses of mt-genome evolution have focused on individual mutations, particularly in describing haplogroups, and to understand and predict ancestral behavior. However, the evolutionary behavior of mt-genome often involves cooperative changes within and between genes which are diffi-

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cult 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 (Raule, et al., 2014; Fan, et al., 2016; Giuliani, et al., 2018; Conte, et al., 2018). 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 (Lane, et al., 2012; Chen, et al. , 2013; Haddad, et al., 2018). Increasing evidence suggests that interactions among polymorphic sites may confer a cumulative association of multiple mutations with many diseases (Chen, et al. , 2013). Interactions among polymorphic sites have also effectively been used to infer ancestry and functional convergence in the human populations (Ioannidis, et al., 2001). Commonly used methods include tree ensembles, functional nodal mutations, and single nucleotide polymorphism (SNP) based enrichment (Lunetta, at al., 2004). Important information about mt-genome evolutionary behavior, which is contained in the correlated changes between nucleotide positions both within 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, mitochondrial genetic variation has been largely ignored in the context of co-mutations and particularly the mechanisms by which these co-mutations occur (Boles, et al., 1998; Goodman, et al., 2006). Investigation of co-mutation 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 distant populations (Hartwig, , 2013).

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

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Albert, and Barabási., 2002; Shinde, et al., 2015; Shinde, and Jalan., 2015; Whitwell, et al., 2018; Rai, et al., 2018; Ho, et al., 2014). The basis of the current study is that genome positions can impact each other and co-mutate within genomes (Shinde, et al., 2018; Du, et al., 2008; Sun, et al., 2014). The interaction between two or more genetic loci is referred to here as the co-mutation of nucleotide positions. There are previous studies which have used genomic co-mutations as a basis of the evolution of human H3N2 and Ebola viruses (Du, et al., 2008; Deng, et al., 2015). These viral genome models have identified the co-mutating nucleotide clusters, apparently underpinning the dynamics of virus evolution since these clusters were antigenic regions of the viral capsid proteins (Du, et al., 2008; Deng, et al., 2015). In another study, Shinde et al. (Shinde, et al., 2018) demonstrated the impact of codon position bias while forming co-mutations using human mt-genomes. These studies have considered perfect co-mutation as causing factor for co-mutations. However, the role of the co-mutation frequency in these studies remains unclear. Here, we thoroughly examined a set of networks associated with a range of co-mutation frequencies and chose a particular co-mutation frequency for further network construction. Whilst pair-wise co-mutations can be straightforwardly perceived, 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 mt-genomes. The paper is organized as follows. In the first section, we briefly described the level of diversity observed among underlying subpopulations concerning polymorphic site variations in human mitochondrial genomes. In the second section, we described the framework to investigate co-mutations, which are critical in underlying complex mitochondrial evolution. For this, we constructed co-mutation net-

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

works which were used to identify modules of co-mutations and also compared these results with those of the corresponding random networks. In the third and fourth sections, we identified local topological phenomena, which were crucial agents for co-mutation 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.

2. Methods and Material

2.1. Acquisition of genomic data

Analysis of mt-genome variations in continental populations has revealed the most ancient of all human continent-specific haplogroups in Africa and their subsequent migration and settlement in other continents (Chen, at al., 1995). Therefore, continents are constituents to defining different global-ancestral lineages beyond being just landmasses. The global lineages among each continent have been shown to explain a variety of signatures including demographic history, climate and environmental changes, local-admixture patterns (Conte, et al., 2018; Fonseca, et al., 2008; Derenko, et al., 2001; Hudson, et al., 2013). Each continent has its own signatures as well as shared signatures, as human migrations are known to happen differently among different continents (Mishmar, et al., 2003).

Having this notion, we prepared an extensive collection of mitochondrial genomes of geographically diverse *Homo sapiens* populations (Fig 1) from the Human Mitochondrial Database (Hmtdb) (Rubino, et al., 2012). All downloaded genome sequences were in FASTA format. In total, the dataset comprised of 24,167 mitochondrial genome sequences from the five world continents (genome groups), including 3426 African (AF), 2650 American (AM), 8483 Asian (AS), 8060 European (EU) and 1548 Oceanic (OC) genomes.

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Antarctica was excluded from the present analysis since no data was available. It should be noted that these genome groups are multiethnic cohorts representing a range of populations across the continent (Fig 1A). A brief description of all the genomes and their origin is provided in S1 File.

2.2. Construction and preliminary analysis of co-mutation networks

Co-mutation calculations were carried out on each genome group distinctly. Co-mutation network construction is broadly divided into two parts, construction of primary networks for each genome sequence followed by the construction of final networks for each genome groups. Each primary co-mutation network represents an individual sequence, and thus for each genome group, M primary co-mutation networks were generated where M is the number of sequences in the genome group. In a co-mutation network (for both primary and final), nodes represent genome positions, and edges between nodes represent genomic co-mutations. We constructed five co-mutation networks for each genome group using their primary co-mutation networks. Final co-mutation networks comprise of qualitative information of interactions between genome positions. The methodology for constructing primary and final co-mutation networks is schematically represented in Fig 1 and described as follows:

2.2.1. Primary co-mutation network

(1) Genome sequence data was end to end aligned. (2) All non-variable genome positions within samples of a genome group were removed, leaving only polymorphic genome positions. The number of polymorphic sites (N_P) is given in Table 2. (3) Using only polymorphic nucleotide positions, we calculated the frequency of occurrence of all the nucleotide pairs $f(x_i y_j) = N(x_i y_j)/M$ where, $N(x_i y_j)$ denoted the number of co-mutation pairs $(x_i y_j)$ at position (i, j). We then calculated the frequency of 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 (Du,

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

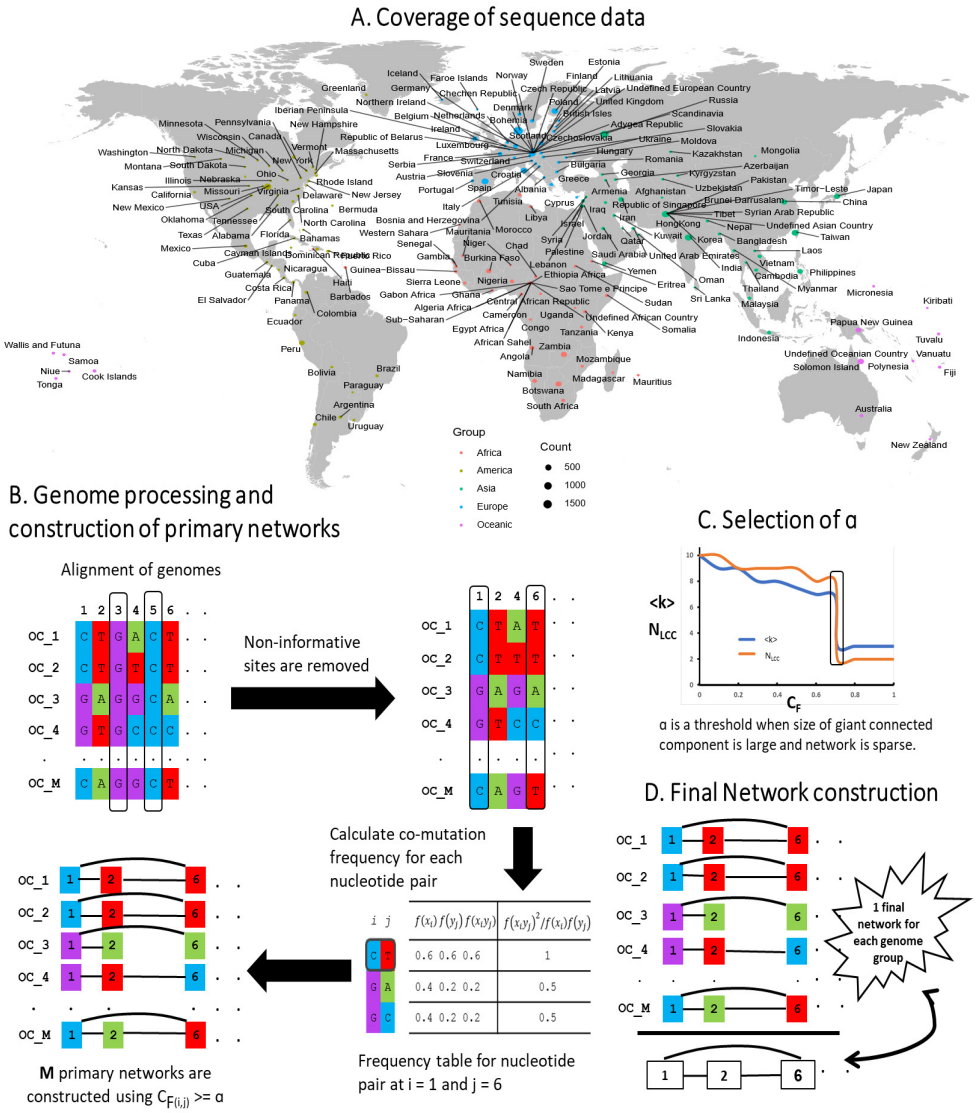

Figure 1: **Schematic representation of mtDNA co-mutation network construction and analysis. (A)** World map shows sequence data taken for the current study covered a good distribution across the entire globe. **(B)** A schematic diagram is drawn for a genome group with 5 sample sequences. The schematic 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-mutation 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.

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

et al., 2008). (4) Co-mutation of two nucleotides (C_F) at position (i, j) was denoted as,

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

For a particular co-mutation frequency threshold, here, termed as network efficiency score (see Section 2.3), we constructed primary co-mutation networks. A network can be represented mathematically by an adjacency matrix (A) with binary entries.

$$A_{\mathrm{ij}} = \{\, 1 \quad \text{if } C_{Fi,j} >= \alpha 0 \quad \text{otherwise} \tag{2}$$

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

2.2.2. Final co-mutation network

Unique edges from all primary co-mutation networks of a genome group were used to construct a final co-mutation network (Fig 1D). These five final co-mutation networks were used for network analysis and community detection.

We extracted hierarchical modules from final co-mutation networks and compared these networks with random networks, hypothesizing that hierarchical modularity is the underlying phenomena of co-mutation networks and is not a mere outcome of the random evolutionary process. Furthermore, we characterized the identified module structures using ancestral markers.

2.2.3. Preliminary analysis of co-mutation networks

The degree of a node (k_i) is defined as a number of edges connected to the node such as $k_i = \sum_{j=1}^{N} A_{ij}$ where N denoted the number of nodes in a network. The average degree connectivity $\langle k \rangle$ is the average nearest neighbor degree of nodes with degree k. The clustering coefficient (C) is a measure of the extent to which nodes in a network tend to cluster together. An

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 measures 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 (Newman, et al., 2003). The value of r being zero corresponds to a random network, whereas the negative (positive) values correspond to dis (assortative) networks.

2.3. Selection of network efficiency score (α)

Network efficiency score (α) was used to filter edges required for network construction. Selecting an α value for each network should require iterating through a range of C_F values. 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 computationally intensive process. Therefore, we performed statistical sampling on each genome group interdependently by selection analysis of m samples from each population. The sample size was determined by Cochran's sample size formula (Cochran,, 1997) with critical value ($z = 1.96$). As the population was finite, the sample size was corrected by Cochran's adjustment (Cochran,, 1997).

A zero α value would result in co-mutation between each mutation and all others, whereas α equal to one would give only those pairs of mutations which have co-mutated perfectly in a genome group. In other words, zero α value would result in the globally connected network (Fig 2B) and $\alpha = 1$ would result networks with many globally connected small sub-graphs (Fig 2E). Even when the α value was as high as 0.99, networks remained very densely connected (Fig 2C). Therefore, it was reasonable to propose a criterion to select an α value, otherwise generated networks would be saturated structures holding no information about co-mutations. In order to tackle this, we plotted $\langle k \rangle$ and the size of the largest connected component (N_{LCC}) against all the α values. We observed surprising network phenomena where at a particular

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

α value, $\langle k \rangle$ is small whilst N_{LCC} is large. At this point, networks are sparser as compared to previous α values (Fig 2D). By a sparse network, we would mean that the majority of elements of the adjacency matrix are zeroes. After exceeding this α value, the network breaks into several disconnected components.

With this criterion, we chose a particular α value for each genome group and constructed primary co-mutation 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 S3 Fig). A similar criterion of filtering network edges has been earlier used to construct gene co-expression network (Jackson, et al., 2018). It should be noted that α values for each genome groups were different (Table 2).

2.4. Detection of module structures in co-mutation networks

We used the Louvain algorithm, a modularity maximization algorithm, for community detection for our networks (Blondel, et al., 2008). 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 (Blondel, et al., 2008), and Gephi software was used for visualization (Bastian, et al. , 2009).

3. Results

Analysis of polymorphic sites is provided in supplementary materials and summarised in Table 1 and Fig 3. Codon position (CP) 2 showed fewer polymorphisms as compared to CP 1 and CP 3. Genes *ATP6* and *ATP8* demonstrated a higher level of polymorphisms at all three CPs. Similarly, all three HVS regions have displayed a higher level of polymorphisms, whereas genes of rRNA and tRNA have shown lower levels of polymorphisms.

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

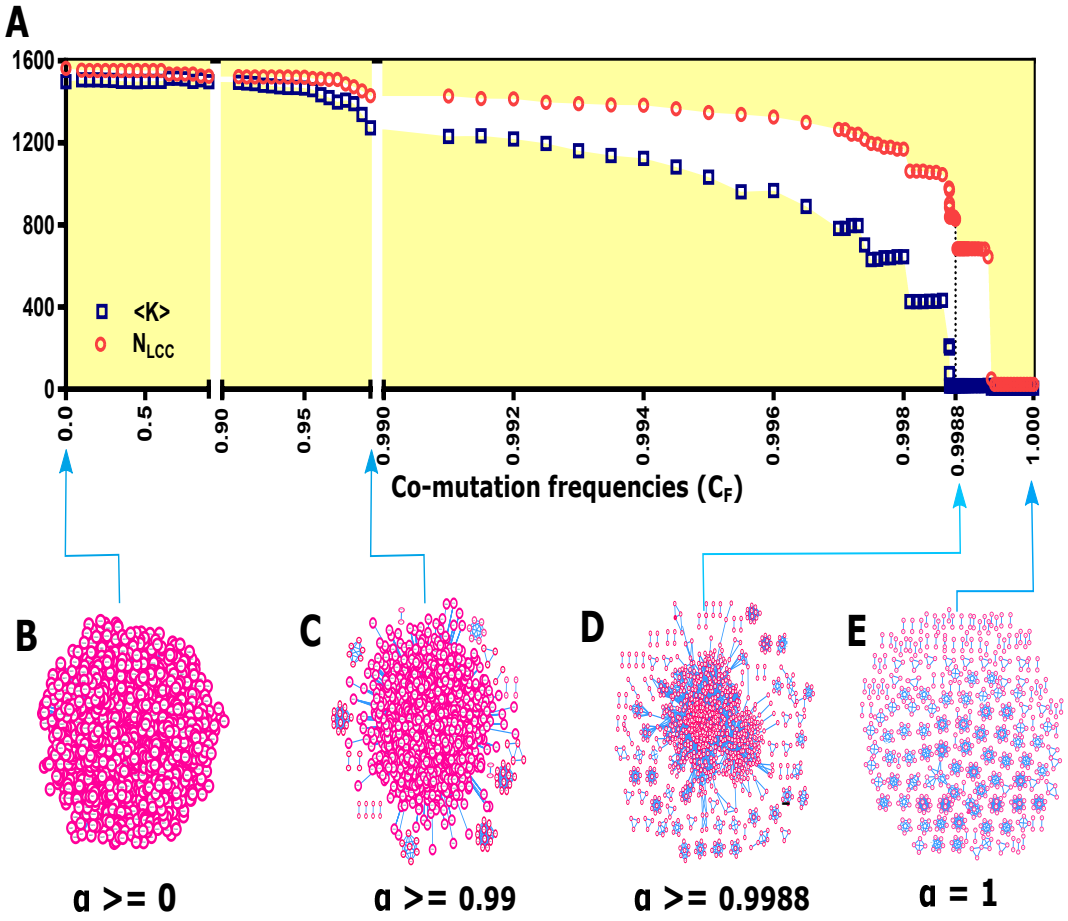

Figure 2: **Evolution of mitochondrial co-mutation network. (A)** The relative size of the largest component and the average degree of the largest component are plotted against co-mutation frequencies (C_F). The figure illustrates that at a particular α value (for OC, $\alpha = 0.9988$) co-mutation network 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.

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

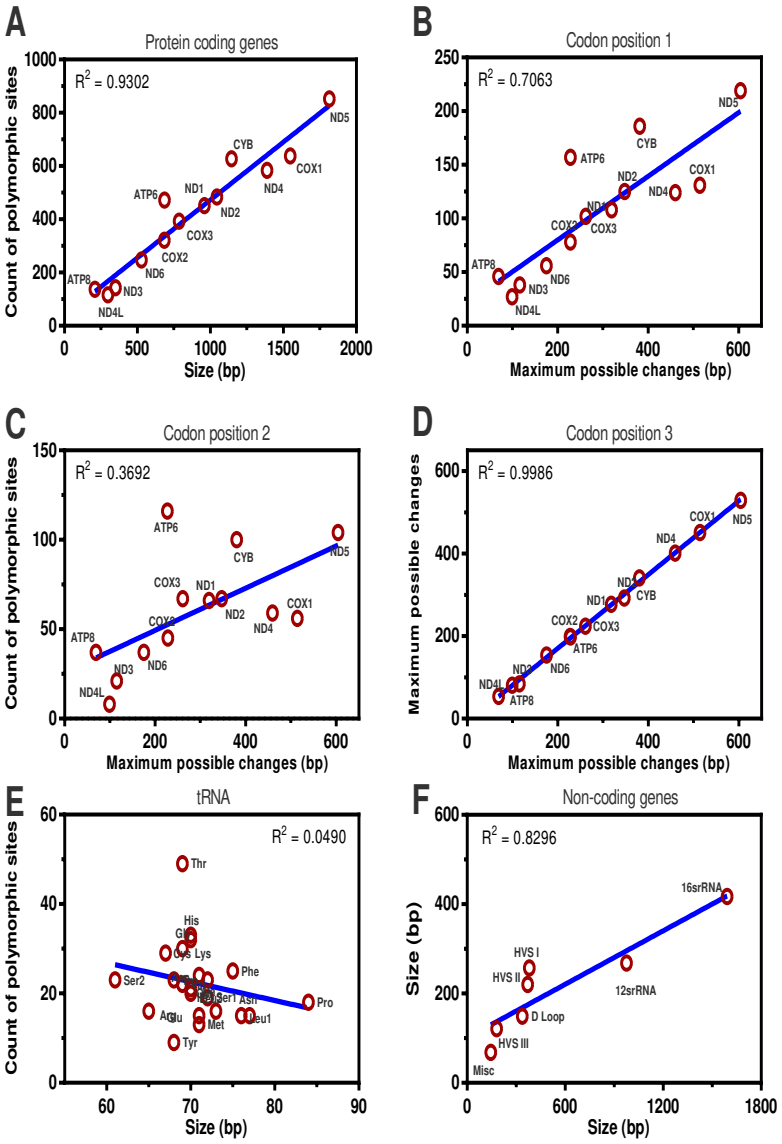

Figure 3: **Diversity among individual genome regions.** Correlation between the observed polymorphic positions and the gene size (bp) or maximum possible changes in **(A)** the 13 protein-coding genes, **(B-D)** the codon positions 1, 2 and 3 among the 13 protein-coding genes, **(E)** tRNA genes and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

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

Table 1: **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. *COX1* and *ND4* had the lowest proportion of observed polymorphic sites, and *ATP6* had the largest proportion.

3.1. Evolution of mitochondrial co-mutations

3.1.1. Co-mutations displaying intra- and inter- genomic loci biases

Analysis of pairs of co-mutations provides insight into the relationship between two distinct genome locations. Co-mutations can be formulated within a particular mitochondrial functional region (intra-loci) or between two functional regions (inter-loci). We enumerated co-mutations present among nine mt-genome functional regions. The number of polymorphic sites was normalized by the total number of co-mutating polymorphic sites in a genome group and used to construct Circos plots (Fig 4). 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

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

I, IV and HVS functional regions have a large contribution to the overall co-mutation configuration in each network. To know more on how each functional region has contributed in forming co-mutations, we plotted the number of co-mutations in each functional region against the corresponding functional region size for intra- and inter- loci (Fig 4). 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 4). 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. 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 calculated Frobenius distances between each pair of five co-mutation matrices and then performed hierarchical clustering. A dendrogram clearly showed the separation of five genome groups into two main branches *i.e.* {AM, AS} and {AF, EU, and OC} (Fig 5).

To investigate global level co-mutation preferences between functional regions, we analyzed unique co-mutations from all the genome groups. Fewer co-mutation pairs were formulated among intra-loci than inter-loci. This relationship between co-mutations and the spatial proximity is shown to be conserved in the mt-genome since all 13 protein-coding genes formed many interactions with OXPHOS complexes (Wong, et al., 1975; Thompson, et al., 1994). 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 4). The rRNA genes make-up 15% of total mt-genome

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

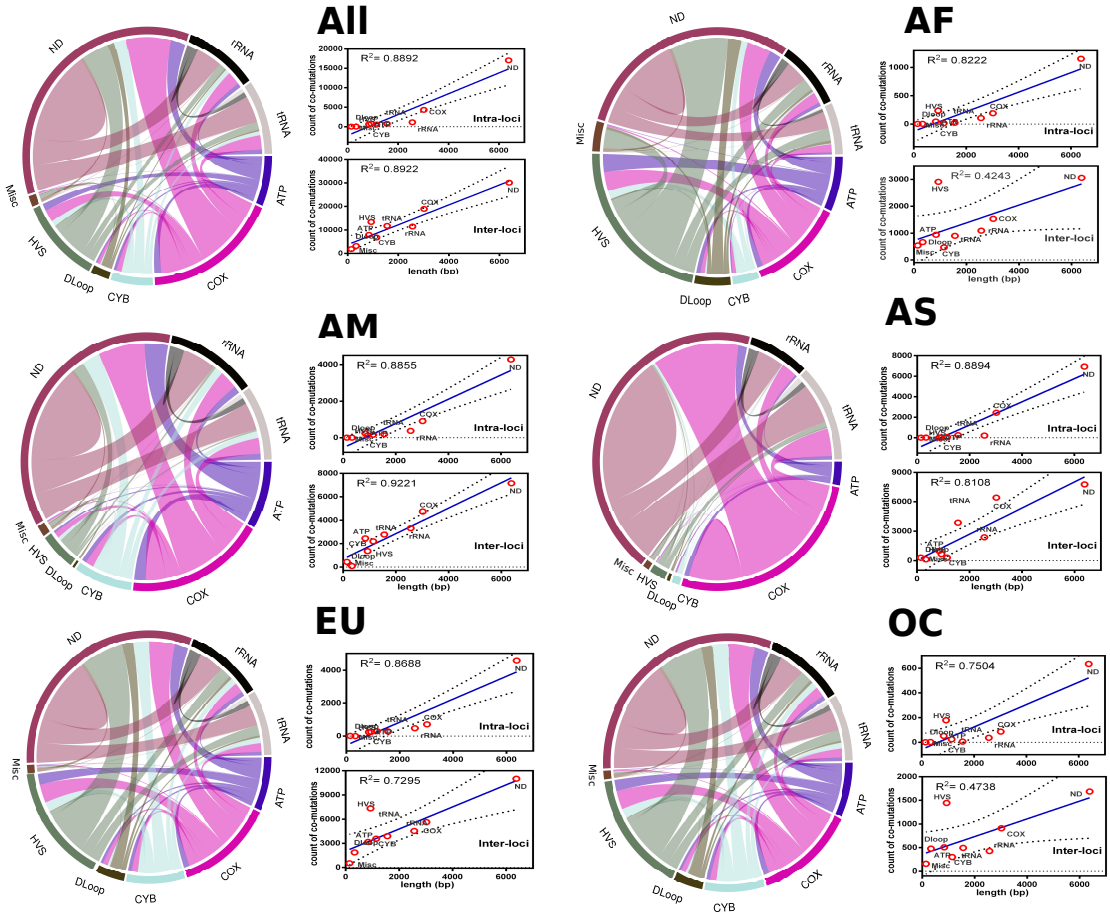

Figure 4: **Comparison of polymorphism among genomic loci.** A co-mutation configuration in the human mt-genome co-mutation network consisting of nine functional regions. These nine regions were four mitochondrial complexes (ND, COX, ATP, and CYB), three non-coding regions (DLoop, HVS, Miscellaneous) 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.

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Network	α	N_P	N	N_C	$\langle K \rangle$	$\langle C \rangle$	r	Q_{Real}	Q_{Rand}
AF	0.99970	3716	2310	13721	12	0.33	-0.36	0.51	0.20
AM	0.99959	3581	2412	30283	25	0.31	-0.61	0.22	0.10
AS	0.999854	5405	2293	32705	29	0.22	-0.18	0.21	0.09
EU	0.999760	4557	2456	47952	39	0.18	-0.62	0.30	0.07
OC	0.998800	1565	1208	7304	12	0.54	-0.25	0.54	0.21

Table 2: **Data statistics and the properties of final co-mutation networks.** Here, α, N_P, N, N_C, $\langle K \rangle$, $\langle C \rangle$, r, Q represent the co-mutation frequency, number of polymorphic sites, number of nodes, number of edges, the average degree, the clustering coefficient, the assortativity coefficient and modularity coefficient for both real-world co-mutation networks and random networks, respectively. All five networks were sparse, disassortative and modular in nature. Network statistics of the largest connected component and the disconnected components are given in S1 Table and S2 Table. The number of nodes and edges forming final co-mutation networks were found to be different for each genome group. 1000 degree sequence preserved random networks are constructed for comparison of modularity in each co-mutation network and standard deviation was found to be less than 0.002 in average Q values of corresponding random networks.

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 preferences within each group. Furthermore, to investigate whether the identified co-mutations can be mapped with previously known potential disease alleles, we extracted information about potential disease alleles from Mitomap (Ruiz-Pesini, et al., 2006), OMIM (Hamosh, et al., 2005), and COSMIC (Bamford, et al. , 2004) databases. There were 5, 74, 34, 8, and 10 co-mutations mapped for AF, AM, AS, EU, and OC genome groups respectively. The full list of co-mutations mapped with potential disease allele is given in File S3 and Table S6.

3.1.2. Co-mutation networks exhibited similar network properties

Pair-wise co-mutations were not sufficient to fully reveal the underlying structure of functionally related nucleotide positions. As described in Fig 1,

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

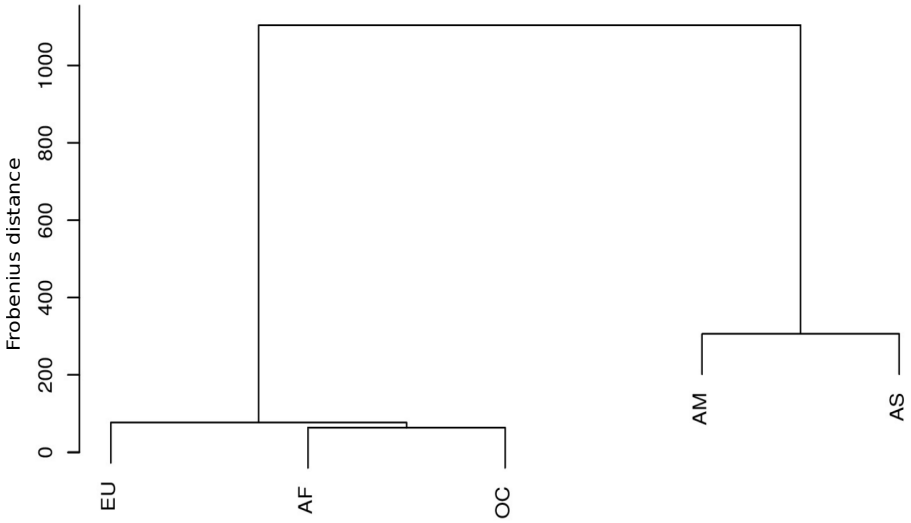

Figure 5: **Relationships among genome groups.** These relationships are predicted based on polymorphisms shown by their functional genomic loci. Five genome groups were classified as two main branches of the dendrogram , *i.e.* {AM, AS} and {AF, EU, and OC}. Here, Frobenius distance between co-mutation configuration matrices of different genome groups were used to define height of dendrogram. The branch separations shown in plot supports the routes of human migrations earlier discovered using global mt-genome mutational phylogeny. In particular, Asian haplogroup M and European haplogroup N arose from the African haplogroup L3 (Wallace, et al., 1999). Haplogroup M gave rise to the haplogroups A, B, C, D, G, and F (Wallace, et al., 1999) 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 (Torroni, et al., 1996) whereas Haplogroups S, P, and Q are found in Oceania (Ruiz-Pesini, et al., 2006).

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a co-mutation network was constructed for each genome group where polymorphic sites forming co-mutations constituted nodes, and edges represented co-mutating nucleotide positions. All five networks exhibited high average clustering coefficient, $\langle C \rangle$ values (Table 2), suggesting that the nodes of these networks are densely connected. Most real-world networks, particularly social networks, characterized by high $\langle C \rangle$ value, suggesting movie actors tend to create tightly knit groups by high compact ties (Sarkar, et al., 2016). In addition, all five networks displayed a highly negative degree-degree coefficient (r) (Table 2), suggesting that co-mutation networks were dis-assortative where high degree nodes, on average, prefer to link to low degree nodes (Newman, et al., 2003). Many biological and social networks have negative r values, suggesting that lack of a high degree node in a disassortative network has a large effect on the connectedness of the network (Newman, et al., 2003; Shinde, et al., 2015; Sarkar, et al., 2016). Overall, co-mutation networks have shown both the properties of high clustering and disassortative nature. This suggests the presence of dense subgraphs within the network and the presence of hierarchical structures. To explore more about the local interaction patterns in co-mutation networks, we investigated module structures within these networks.

3.2. High cohesiveness and hierarchical organization of co-mutation communities

Is real-world network organization driven by the non-random character, at least to some extent, by modules present in the network? If this is the case, it is expected that modules would be overrepresented in original co-mutation networks compared to their counterparts such as random networks of the same size (Prill, et al., 2005). To test this, we generated random networks, referred to as a configuration model, with the same degree sequence as the original co-mutation network (Csardi, and Nepusz., 2006). Random networks lack organizing principles, therefore, the presence of modules in a random network is determined by the density of edges (Itzkovitz, et al., 2003).

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

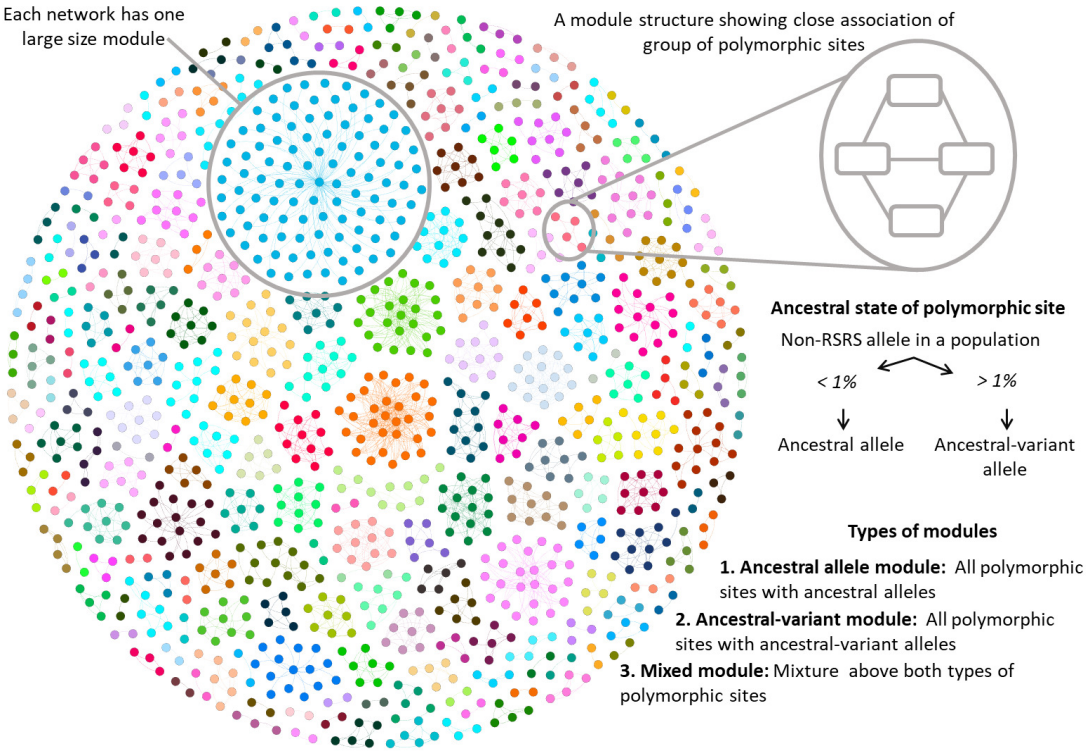

Figure 6: **Identification and characterization of network modules.** Network modules are identified using Louvain modularity algorithm. Here, the network view of modules is shown for OC network. Each network comprised of one large size module. Each polymorphic site in a module was classified as an ancestral or ancestral-variant polymorphic site. Further, modules are classified into three categories: ancestral allele module (ancestral polymorphic sites), ancestral-variant module (ancestral-variant polymorphic sites) and mixed module (both ancestral and ancestral-variant polymorphic sites).

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The major challenge for identifying modules in a hierarchical organization is to decide the depth to decompose the network, as the Louvain algorithm fragments networks and subsequently modules until it finds the greatest partition (Meunier, et al., 2009). 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 co-mutation networks and also for corresponding random networks (Table 2). Q value ranges between -1 and 1, where it takes positive values if there are more edges between same-group vertices than expected, and negative values if there are less (Blondel, et al., 2008). We tested the hypothesis that the average Q of random networks equals that of the co-mutation network. Q value was clearly reduced in the randomized networks (t-test, $p < 0.001$, for all co-mutation networks), relative to the original data, indicating that our results on real-world co-mutation 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) (S5 Fig). This signature of large-sized modules found in co-mutation networks was not displayed by corresponding random networks. 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 the 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

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

functional regions have a predominant role in the formation of modules (S7 Table and S8 Table). Particularly, ND and COX participated in >65% and >40% of modules in each of the 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 (S7 Table).

	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

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. [1] referes ancestral lineage polymorphic sites (ALPS)

3.3. Modules of co-mutating polymorphic sites indicate 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 for ancestral alleles from the Reconstructed Sapiens Reference Sequence (RSRS). If a non-RSRS allele was present in more than

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1% of samples in a genome group, we termed it an ancestral-variant allele. Thus, we assigned ancestral-variant information to all of the network modules and noted three distinct types of modules (Fig 6), which are explained in following.

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 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 (Ruiz-Pesini, et al., 2006) 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 because all polymorphic sites among these modules consist of the evolved character from RSRS. 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 hypothesized to have 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). We also confirmed that count of each module type in all five co-mutation networks is significantly different than those of corresponding random networks (t-test, $p < 0.0001$; Fig S8, S9, S10). In addition, ancestral-variant modules were difficult to produce in random networks.

Modules were mapped to all known haplogroups, which showed that each polymorphic site contributed to one or many haplogroups, and thus entire module structure can be related to a single mt-genome haplogroup (S2 File). Further, we investigated the relationship between modules corresponding to

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ancestral haplogroup lineage markers (or top-level haplogroups). Information of ancestral lineage markers was taken from the Mitomap database, and polymorphic sites among each module were mapped to ancestral lineage markers. These ancestral lineage markers were 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.

4. Discussion

We used comparative genome analysis to investigate 24,167 mt-genomes and devised a network model comprising pairs of co-mutating nucleotides over the human mt-genome. The method presented here provides a perspective on epistatic interactions using only sequence information as well as serves as a comparative tool to understand intra-species variations. Our study showed the presence of heterogeneity in both epistatic mutations and functional modules across investigated genome groups.

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 providing a protein diversity of mt-genome complexes. Second, the restriction of mutations in 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 sequences (Pereira, et al., 2009). The similar biases of CP 2 and tRNA genes were also reported among mt-genomes of other primates including Macaca, Papio,

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Hylobates, Pongo, Gorilla, and Pan whereby the strength of selection was determined in each lineage by the ancestral state of each codon position (Kivisild, et al., 2006). 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 despite infrequently reported to be at a 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.

As well as biases against polymorphisms at CP 2 in protein-coding genes, our analysis indicates other biases with co-mutations. First, our results with intra- and inter- loci preferences clearly suggested the dominance of polygenic mutations in the human mt-genome. These polygenic mutations are the outcome of a highly constrained organization of OXPHOS complexes (Fonseca, et al., 2008) and also due to protein-protein interactions of the mt-interactome (Schweppe, et al., 2017). Second, our analysis highlights regions of the mt-genome which are rich in co-mutations and thus suggests the presence 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 co-mutation bias in functional regions of HVS. Although mitochondrial genome epistasis is largely described in the context of mitochondrial-nuclear interaction due to the closed assembly of OXPHOS complexes (Fonseca, et al., 2008; Picard, et al., 2018; Connallon, et al., 2018; Schweppe, et al., 2017), there are many reports describing the presence of mitochondrial-mitochondrial epistasis, for example, shared family features in Han Chinese family (Wang, et al., 2015) and homoplasy guided by mt-tRNA genes (Moreno-Loshuertos, et al., 2011). Furthermore, few co-mutations were found to be mapped with previously known potential disease alleles, implicating association of disease

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

phenotypes within mt-genome sequences. Mitochondrial epistasis has been reported for its role in mitochondrial diseases (Morrow, and Camus., 2017; Smith, and Lusis., 2002; Pritchard, et al., 2010; Schrider, and Kern., 2017). We also note that these individual mapped co-mutations do not possess potential disease alleles in other genome groups. In view that mt-genome lineages are functionally different, it reflects that the same epistatic interaction can be advantageous in one environment and might be maladaptive in a different environment (Mishmar, et al., 2003). Hence, lineage-defining haplotypes could be contributing to bioenergetic disorders as the migration takes place (Mishmar, et al., 2003). Third, similar to polymorphic sites, co-mutations also showed biases at the subpopulation level. Genome group-wise comparison of co-mutations associated with 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 (Ruiz-Pesini, et al., 2006) showing the routes of human migrations (Fig 5). 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.

In our network model, 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 sudden change in the α curve (Fig 2). 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 (Erdos and Rnyi., 1960). Interestingly, the nature of discontinuous transformations was earlier reported in biological networks (Fontana, and Schuster., 1998; Liu, et al., 2012) and have been hypothesized that biological processes follow discontinuous transformation during their evolution (Fontana,

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and Schuster., 1998).

We selected edges for inclusion in co-mutation networks based on their best fit to a network sparseness. Sparseness is one of the essential properties of biological networks since links are more difficult to create due to the evolutionary cost involved in forming more links. It is well known that co-mutational events are very selective and require a group of cooperative supporting mechanisms (Du, et al., 2008). Previously, modules of highly correlated genes were identified using similar edge-filtering based methods like weighted correlation network analysis (Jackson, et al., 2018). Network sparseness or similar data-driven approach avoids arbitrary selections of network edges and provides a uniform rationale that can be implemented to generate co-mutation 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 $\langle K \rangle$ and the largest component with a higher count of nodes.

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. A combination of co-mutations at different mitochondrial regions, that are closely linked, tend to be inherited together. This study provides a complete listing of the current knowledge of mt-genome variation in the human population, also with respect to their higher level associations with hierarchical modules. Every set of co-mutations found to originate from and remain part of a preceding single group of co-mutations. This nested hierarchy suggests the conservation of ancestral as well as inherited co-mutations throughout human lineages. 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 (Meunier, et al., 2009). Modularity is one of the main features of co-mutation networks, and evolutionary processes

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

may favor the emergence of modularity by a combination of structural and functional preferences in forming molecular interactions (Clune, et al., 2013). Therefore, it was reasonable to say that evolutionary processes may favor modularity by allowing both the specificity and autonomy of functionally distinct subsets of genomic positions. Overall, we demonstrated that molecular changes, such as mutations, were not randomly distributed across the genome, but instead concentrated within modules. 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 might 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 (Fonseca, et al., 2008). In addition, the OXPHOS system is intrinsically incapable of evolving to a fixed and general optimum state; therefore, both functional and genetic heterogeneity plays a vital role in providing robustness to the evolving OXPHOS system (Enriquez., 2016). 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.

It was expected that module level associations would reflect evolutionary relationships between underlying genomic positions as each module consisted of ancestrally similar genomic polymorphisms. Our results added that the distinction between ancestral and ancestral-variant mitochondrial polymorphisms was clear when 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 of all five networks, it would be reasonable to assert

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

that contemporary 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-mutation among nucleotide positions to be higher between genetically similar taxa (Chaffron, et al., 2010). This fact was widely observed in our data as both sub-level, and top-level haplotype markers were associated with each other in a closed group of network modules. Furthermore, mitochondrial polymorphic positions adapt from ancestral state to ancestral-variant state (Keightley, and Jackson., 2018), is also demonstrated by a larger count of transient state mixed modules. Overall, these evolutionarily closed associations suggest that interactions between nucleotide positions might evolve within genetically related genomic polymorphic positions (more likely of having similar functionality) responding to intra-species biases (Du, et al., 2008).

Understanding the formation of the network would require an extension of the described approaches. Here, we used simplistic information possessed by each genome position in terms of their underlying ancestral markers. Previously, this ancestral marker information has been used in order to define taxa (precisely haplogroups) in mitochondrial phylotrees which have provided the exact mapping of mitochondrial signatures to infer the routes of human intra-species diversification events (Nakatsuk, et al., 2017; Derenko, et al., 2001). Similarly, our co-mutation modules have provided a detailed listing of mitochondrial co-mutations which were ancestrally associated together.

Consistent with the proposed importance of mt-genome variation in human adaptation (Wallace., 2015), regional haplotypes are generally founded by one or more functionally significant polypeptide, tRNA, rRNA, and control region variants. These variant traits, beyond being retained in the descendant population, manifest functional signatures about affecting phenotypic variations in the subpopulation. Particularly, a list of singular events involving both mt-genome variation and epistatic interactions have been evaluated in terms of affecting phenotypic variation in metabolism, fitness, and life-history traits

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Wallace., 2015; Shlush, et al., 2008; Li, et al., 2015). Hence, it is intuitive to observe epistatic patterns of both genotype and phenotype variants along with life-history traits at the subpopulation level. However, at the between-population level, the evidence in support of the relationship between human mt-genome variation and the metabolic rate is compelling.

5. Conclusion

We constructed and investigated human mt-genome co-mutation networks of continents using a combined framework of genomics and network theory. 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 also evident at the continental level, which was earlier only reported in the global human population. The analysis highlighted regions of mt-genome rich for co-mutations and thus suggested the presence 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 co-mutation bias between regions of HVS and ND, thus our networks identified differences in co-mutation bias between human populations. It was of great interest to investigate and verify different co-mutation patterns of various geographical regions. Importantly, we deduced hierarchical modular structures formed within co-mutation networks. Downstream analysis of these modules suggested that contemporary human population are dominated by ancestral states. In addition, ancestral-variant module structures are found to be in a lesser number, and such modules have found to be difficult to produce in corresponding random networks. The analysis presented here can be extended to study the complexity of mt-genome evolution by forming various geographical groups as well as to understand alterations in personal traits leading to complexity in mt-genome evolution.

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We are grateful to our the editor and reviewers for helping us to improve the manuscript. PS acknowledges Inspire fellowship [IF150200] from the Department of Science and Technology (DST), Government of India. SJ thanks the support by grant of DST [EMR/2016/001921], 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 supplementary material files.

Supplementary materials

S1 File. Information of genome samples considered in the study.
S2 File. Information of modules identified in the study.
S3 File. Information of co-mutations consist of potential disease alleles (PDA).
Table S1: Network properties of Largest connected component (Lcc).
Table S2: Network properties of all disconnected components together except Lcc.
Table S3: Level-wise community detection in five co-mutation networks.
Table S4: Comparison of polymorphism α_{pre} and α_{post}.
Table S5: List of hub nodes among co-mutation networks.
Table S6: List of co-mutations consist of potential disease alleles (PDA).
Table S7: Distribution of modules having at least one polymorphic site among mt-genome functional groups.

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table S8: Modules comprising all nodes as ancestral lineage polymorphic sites.

Table S9: Proportion of modules among co-mutation networks and corresponding random networks.

Table S10: Proportion of nodes in modules among co-mutation networks and corresponding random networks.

Figure S1: Distribution of variable sites across genome groups and the distance between each genome from the reference sequence (RSRS).

Figure S2: Degree distribution of five co-mutation networks.

Figure S3: Gene-wise comparison of polymorphism before and after α.

Figure S4: Transition and transversion.

Figure S5: Schematic description of assortativity.

Figure S6: Distribution of module sizes.

Figure S7: Identification and characterization of network modules.

Figure S8: Distribution of modules count in random networks.

Figure S9: Distribution of modularity coefficient among random networks.

Figure S10: Count of different type of modules among random networks

References

Albert, R., Barabási, A. L. (2002). Statistical mechanics of complex networks. *Rev Mod Phys*. 74:47. DOI: 10.1103/RevModPhys.74.47

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A. and et al. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer*, 91(2): 355. DOI: 10.1038/sj.bjc.6601894

Bastian, M., Heymann, S., Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Third international AAAI conference on weblogs and social media*.

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Boles, R. G., Roe, T., Senadheera, D., Mahnovski, V., Wong, L. J. (1998). Mitochondrial DNA deletion with Kearns Sayre syndrome in a child with Addison disease. *Eur J Pediatr*. 157(8):643-647. DOI: 10.1007/s004310050902

Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J Stat Mech Theory Exp*. 10. DOI: 10.1088/1742-5468/2008/10/P10008

Chaffron, S., Rehrauer, H., Pernthaler, J., von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res*. 104521. DOI: 10.1101/gr.104521.109

Chen, J. B., Chuang, L. Y., Lin, Y. D., Liou, C. W., Lin, T. K., Lee, W. C. and et al. (2013). Preventive SNP-SNP interactions in the mitochondrial displacement loop (D-loop) from chronic dialysis patients. *Mitochondrion*. 13(6):698-704. DOI: 10.1016/j.mito.2013.01.013

Chen, Y. S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S., Wallace, D. C. (1995). Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet*. 57(1):133.

Cochran, W. G. (1997). Sampling techniques. John Wiley and Sons. ISBN: 0-471-16240-X

Connallon, T., Camus, M. F., Morrow, E. H., Dowling, D. K. (2018). Coadaptation of mitochondrial and nuclear genes, and the cost of mother's curse. *Proc R Soc B*. 285(1871):20172257. DOI: 10.1098/rspb.2017.2257

Conte, M., Ostan, R., Fabbri, C., Santoro, A., Guidarelli, G., Vitale, G. and et al. (2018). Human aging and longevity are characterized by high levels of mitokines. *J Gerontol*. Gly:153.

Clune, J., Mouret, J. B., Lipson, H. The evolutionary origins of modularity. (2013). *Proc R Soc B*. 280(1755):20122863. DOI: 10.1098/rspb.2012.2863

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Csardi, G., Nepusz, T. The igraph software package for complex network research. (2006). *InterJournal Complex Systems* 1695(5):1-9.

Enrquez, J.A. Supramolecular organization of respiratory complexes. (2016) *Annu Rev Physiol*, 78: 533-561. DOI: 10.1146/annurev-physiol-021115-105031

da Fonseca, R. R, Johnson, W. E., O'Brien, S. J., Ramos, M. J., Antunes, A. (2008). The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics*. 9(1):119. DOI: 10.1186/1471-2164-9-119

Deng, L., Liu, M., Hua, S., Peng, Y., Wu, A., Qin, F. X. and et al. (2015). Network of co-mutations in Ebola virus genome predicts the disease lethality. *Cell Res*. 25(6):753. DOI: 10.1038/cr.2015.54

Derenko, M. V., Grzybowski, T., Malyarchuk, B. A., Czarny, J., Micicka-liwka, D., Zakharov, I. A. (2001). The presence of mitochondrial haplogroup X in Altaians from South Siberia. *Am J Hum Genet*. 69(1):237-241. DOI: 10.1086/321266

Du, X., Wang, Z., Wu, A., Song, L., Cao, Y., Hang, H., Jiang, T. (2008). Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. *Genome Res*. 18(1):178-187. DOI: 10.1101/gr.6969007

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

Fan, S., Hansen, M. E., Lo, Y., Tishkoff, S. A. (2016). Going global by adapting local: A review of recent human adaptation. *Science*. 354(6308):54-59. DOI: 10.1126/science.aaf5098

Fontana, W., Schuster, P. (1998). Continuity in evolution: on the nature of transitions. *Science*. 280(5368):1451-1455. DOI: 10.1126/science.280.5368.1451

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Giuliani, C., Sazzini, M., Pirazzini, C., Bacalini, M. G., Marasco, E., Ruscone, G. A. and et al. (2018). Impact of demography and population dynamics on the genetic architecture of human longevity. *Aging*. 10(8):1947. DOI: 10.18632/aging.101515

Goodman, J. E., Mechanic, L. E., Luke, B. T., Ambs, S., Chanock, S., Harris, C. (2006). Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis. *Int J Cancer*. 118(7):1790-1797. DOI: 10.1002/ijc.21523

Haddad, R., Meter, B., Ross, J.A. (2018). The genetic architecture of intra-species hybrid mito-nuclear epistasis. *Front Genet*, 9, 481. DOI: 10.3389/fgene.2018.00481

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids res*, 33: D514-D517. DOI: 10.1093/nar/gki033

Hartwig, F. P. (2013). SNP-SNP Interactions: focusing on variable coding for complex models of epistasis. *J Genet Syndr Gene Ther*. 4(189):10-4172. DOI: 10.4172/2157-7412.1000189

Ho, Y.Y., Cope, L.M. Parmigiani, G. (2014). Modular network construction using eQTL data: an analysis of computational costs and benefits. *Front Genet*, 5, 40. DOI: 10.3389/fgene.2014.00040

Hudson, G., Nalls, M., Evans, J. R., Breen, D. P., Winder-Rhodes, S., Morrison, K. E. and et al. (2013). Two-stage association study and meta-analysis of mitochondrial DNA variants in Parkinson disease. *Neurolog*. 80(22):2042-2048. DOI: 10.1212/WNL.0b013e318294b434

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., ContopoulosIoannidis DG.

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(2001). Replication validity of genetic association studies. *Nat Genet*. 29:306-309. DOI: 10.1038/ng749

Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G., Alon, U., 2003. Subgraphs in random networks. *Physical review E*. 68(2): 026127. DOI: 10.1103/PhysRevE.68.026127

Jackson, M. A., Bonder, M. J., Kuncheva, Z., Zierer, J., Fu, J., Kurilshikov, A. and et al. (2018). Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ*. 6:e4303. DOI: 10.7717/peerj.4303

Keightley, P.D., Jackson, B.C. (2018). Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics*. 209(3): 897-906. DOI: 10.1101/257246

Kenney, M. C., Hertzog, D., Chak, G., Atilano, S. R., Khatibi, N., Soe, K. and et al. (2013). Mitochondrial DNA haplogroups confer differences in risk for age-related macular degeneration: a case control study. *BMC med gen*. 14(1):4. DOI: 10.1186/1471-2350-14-4

Kivisild, T., Shen, P., Wall, D. P., Do, B., Sung, R., Davis, K. and et al. (2006). The role of selection in the evolution of human mitochondrial genomes. *Genetics*. 172(1):373-387. DOI: 10.1534/genetics.105.043901

Lanave, C., Tommasi, S., Preparata, G., Saccone, C. (1986). Transition and transversion rate in the evolution of animal mitochondrial DNA. *Biosystems*. 19:273-283. DOI: 10.1016/0303-2647(86)90004-3

Lane, H. Y., Tsai, G. E., Lin, E. (2012). Assessing gene-gene interactions in pharmacogenomics. *Molecular diagnosis and therapy*. 16(1):15-27. DOI: 10.1007/BF03256426

Li, Y., Beckman, K. B., Caberto, C., Kazma, R., Lum-Jones, A., Haiman, C. A. and et al. (2015). Association of genes, pathways, and haplogroups of the

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mitochondrial genome with the risk of colorectal cancer: the multiethnic cohort. *PloS One*. 10(9):e0136796. DOI: 10.1371/journal.pone.0136796

Liu, R., Li, M., Liu, Z. P., Wu, J., Chen, L., Aihara, K. (2012). Identifying critical transitions and their leading biomolecular networks in complex diseases. *Sci Rep*. 2:813. DOI: 10.1038/srep00813

Lunetta, K. L., Hayward, L. B., Segal, J., Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet*. 5:32. DOI: 10.1186/1471-2156-5-32

Meunier, D., Lambiotte, R., Fornito, A., Ersche, K., Bullmore, E. (2009). Hierarchical modularity in human brain functional networks. *Front Neuroinform*. 3:37. DOI: 10.3389/neuro.11.037.2009

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S. and et al. (2003). Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci*. 100(1):171-176. DOI: 10.1073/pnas.0136972100

Moreno-Loshuertos, R., Ferrn, G., Acn-Prez, R., Gallardo, M. E., Viscomi, C., Prez-Martos, A. and et al. (2011). Evolution meets disease: penetrance and functional epistasis of mitochondrial tRNA mutations. *PLoS Genet*. 7(4):e1001379. DOI: 10.1371/journal.pgen.1001379

Morrow, E. H., Camus, M. F. (2017). Mitonuclear epistasis and mitochondrial disease. *Mitochondrion*. 35:119-22. DOI: 10.1016/j.mito.2017.06.001

Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., Patterson, N. and et al. (2017). The promise of discovering population-specific disease-associated genes in South Asia. *Nat Genet*. 49(9):1403. DOI: 10.1038/ng.3917

Newman, M. J. (2003). Mixing patterns in networks. *Phys Rev E*. 67:026126. DOI: 10.1103/PhysRevE.67.026126

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Newman, M. J., Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys Rev E.* 69:026113. DOI: 10.1103/PhysRevE.69.026113
DOI: 10.1103/PhysRevE.58.193

Pereira, L., Freitas, F., Fernandes, V., Pereira, J. B., Costa, M. D., Costa, S. and et al. (2009). The diversity present in 5140 human mitochondrial genomes. *Am J Hum Genet.* 84(5):628-640. DOI: 10.1016/j.ajhg.2009.04.013

Picard, M., McEwen, B. S. (2018). Psychological stress and mitochondria: a conceptual framework. *Psychosom Med.* 80(2):126-40. DOI: 10.1097/PSY.0000000000000544

Prill, R. J., Iglesias, P. A., Levchenko, A. (2005). Dynamic properties of network motifs contribute to biological network organization. *PLoS biology*, 3(11), e343. DOI: 10.1371/journal.pbio.0030343

Pritchard, J. K., Pickrell, J. K., Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20(4):R208-15. DOI: 10.1016/j.cub.2009.11.055

Rai, A., Shinde, P., Jalan, S. (2018). Network spectra for drug-target identification in complex diseases: new guns against old foes. *Appl Netw Sci.* 3(1):51. DOI: 10.1007/s41109-018-0107-y

Raule, N., Sevini, F., Li, S., Barbieri, A., Tallaro, F., Lomartire, L. and et al. (2014). 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.* 13(3):401-407. DOI: 10.1111/acel.12186

Rubino, F., Piredda, R., Calabrese, F. M., Simone, D., Lang, M., Calabrese C. and et al. (2011). HmtDB, a genomic resource for mitochondrion-

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

based human variability studies. *Nucleic Acids Res.* 40(D1):D1150-9. DOI: 10.1093/nar/gkr1086

Ruiz-Pesini, E., Lott, M. T., Procaccio, V., Poole, J. C., Brandon, M. C., Mishmar, D. and et al. (2006). An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res.* 35:D823-8. DOI: 10.1093/nar/gkl927

Sarkar, C., Yadav, A., Jalan, S. (2016). Multilayer network decoding versatility and trust. *EPL.* 113(1):18007. DOI: 10.1209/0295-5075/113/18007

Schrider, D. R., Kern, A. D. (2017). Soft sweeps are the dominant mode of adaptation in the human genome.*Mol Biol Evol.* 34(8):1863-77. DOI: 10.1093/molbev/msx154

Schweppe, D. K., Chavez, J. D., Lee, C. F., Caudal, A., Kruse, S. E., Stuppard, R. and et al. (2017). Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry.*Proc Natl Acad Sci.* 114(7):1732-1737. DOI: 10.1073/pnas.1617220114

Shinde, P., Jalan, S. (2015). A multilayer protein-protein interaction network analysis of different life stages in Caenorhabditis elegans.*EPL.* 112(5):58001. DOI: 10.1209/0295-5075/112/58001

Shinde, P., Sarkar, C., Jalan, S. (2018). Codon based co-occurrence network motifs in human mitochondria.*Sci Rep.* 8(1):3060. DOI: 10.1038/s41598-018-21454-2

Shinde, P., Yadav, A., Rai, A., Jalan, S. (2015). Dissortativity and duplications in oral cancer.*Eur Phys J B.* 88(8):197. DOI: 10.1140/epjb/e2015-60426-5

Shlush, L.I., Atzmon, G., Weisshof, R., Behar, D., Yudkovsky, G., Barzilai, N., Skorecki, K. (2008). Ashkenazi Jewish centenarians do not demonstrate enrichment in mitochondrial haplogroup J. *PLoS One*, 3(10): 3425. DOI: 10.1371/journal.pone.0003425

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Shriner, D., Keita, S.O. (2016). Migration route out of Africa unresolved by 225 Egyptian and Ethiopian whole genome sequences. *Front Genet*, 7, 98. DOI: 10.3389/fgene.2016.00098

Smith, D. J., Lusis, A. J. (2002). The allelic structure of common disease.*Hum Mol Genet*. 11(20):2455-61. DOI: 10.1093/hmg/11.20.2455

Steinberg, B., Ostermeier, M. (2016). Environmental changes bridge evolutionary valleys. *Sci Adv*. 2(1):1500921. DOI: 10.1126/sciadv.1500921

Sun, X., Lu, Q., Mukherjee, S., Crane, P.K., Elston, R. and Ritchie, M.D. (2014). Analysis pipeline for the epistasis search-statistical versus biological filtering. *Front Genet*, 5, 106. DOI: 10.3389/fgene.2014.00106

Thompson, J. N. (1994). The coevolutionary process. *University of Chicago Press*. ISBN: 0226-79759-7. DOI: 10.7208/chicago/9780226797670.001.0001

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R. and et al. (1996). Classification of European mtDNAs from an analysis of three European populations. *Genetics*. 144(4):1835-50.

Van Der Walt, J. M., Dementieva, Y. A., Martin, E. R., Scott, W. K., Nicodemus, K. K., Kroner, C. C. and et al. (2004). Analysis of European mitochondrial haplogroups with Alzheimer disease risk. *Neurosci Lett*. 365(1):28-32. DOI: 10.1016/j.neulet.2004.04.051

Wallace, D. C. (2015). Mitochondrial DNA variation in human radiation and disease. *Cell*. 163(1):33-38. DOI: 10.1016/j.cell.2015.08.067

Wallace, D. C, Brown, M. D, Lott, M. T. (1999). Mitochondrial DNA variation in human evolution and disease. *Gene*. 238(1):211-230. DOI: 10.1016/S0378-1119(99)00295-4

Wang, F., Huang, G. D., Tian, H., Zhong, Y. B., Shi, H. J., Li Z. and et al. (2015). Point mutations in KAL1 and the mitochondrial gene MT-tRNA

bioRxiv preprint doi: https://doi.org/10.1101/505818; this version posted October 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cys synergize to produce Kallmann syndrome phenotype. *Sci Rep.* 5:13050. DOI: 10.1038/srep13050

Whitwell, H. J., Blyuss, O., Menon, U., Timms, J. F., Zaikin, A. (2018). Parenclitic networks for predicting ovarian cancer. *Oncotarget.* 9(32): 22717. DOI: 10.18632/oncotarget.25216

Wong, J. T. (1975). A co-evolution theory of the genetic code. *Proc Natl Acad Sci.* 72(5):1909. DOI: 10.1073/pnas.72.5.1909

Zanellati, M.C., Monti, V., Barzaghi, C., Reale, C., Nardocci, N., Albanese, A., and et al. (2015). Mitochondrial dysfunction in Parkinson disease: evidence in mutant PARK2 fibroblasts. *Front Genet*, 6, 78. DOI: 10.3389/fgene.2015.00078