

OPEN

## Role of mitochondrial genetic interactions in determining adaptation to high altitude human population

Rahul K. Verma<sup>1</sup>, Alena Kalyakulina<sup>2</sup>, Ankit Mishra<sup>3</sup>, Mikhail Ivanchenko<sup>2,4</sup> & Sarika Jalan<sup>1,3</sup>✉

Physiological and haplogroup studies performed to understand high-altitude adaptation in humans are limited to individual genes and polymorphic sites. Due to stochastic evolutionary forces, the frequency of a polymorphism is affected by changes in the frequency of a near-by polymorphism on the same DNA sample making them connected in terms of evolution. Here, first, we provide a method to model these mitochondrial polymorphisms as “co-mutation networks” for three high-altitude populations, Tibetan, Ethiopian and Andean. Then, by transforming these co-mutation networks into weighted and undirected gene–gene interaction (GGI) networks, we were able to identify functionally enriched genetic interactions of *CYB* and *CO3* genes in Tibetan and Andean populations, while NADH dehydrogenase genes in the Ethiopian population playing a significant role in high altitude adaptation. These co-mutation based genetic networks provide insights into the role of different set of genes in high-altitude adaptation in human sub-populations.

Paramount success of network science banks heavily on the impact of pair-wise interactions among the constituents of complex systems on their evolution and performance. Various complex biological systems and phenomena have been studied through their underlying interaction networks<sup>1</sup>; for example, to infer the role of proteins<sup>2</sup>, to understand biological functions of neurons in various developmental stages of *C. elegans*<sup>3</sup>, to identify crucial structural patterns in diabetes mellitus II<sup>4</sup>, and to understand genes responsible for the evolution of specific characteristics of human sub-population<sup>5</sup>. Networks allow the modeling of real-world complex systems by a straightforward yet effective framework that consists of nodes and edges. Analyzing the structural and other features of the underlying network properties based on the information available of the edges (interactions) could reveal many system-level properties of corresponding complex systems<sup>6</sup>. The current work provides a model to capture mitochondrial genomic variations by considering them as co-mutation networks and analyze the role of polymorphic cohorts in high altitude (i) Andean Altiplano, (ii) Qinghai–Tibetan Plateau, and (iii) the Ethiopian human population. The mitochondrial genome plays a significant role in determining energy and heat production under specific conditions. It is now believed that about 25% of the mtDNA protein sequence variations<sup>7,8</sup> 10 to 20% of the tRNA variations, and at least a few of the rRNA variations have altered the mitochondrial coupling efficiency<sup>9</sup>. These alterations led to increased heat production at the expense of ATP production, eventually permitting humans to adapt to colder climates. However, the association and mutual role of polymorphisms in other regions of mtDNA as a complex system subject to investigation in establishing a conclusive role. These three high-altitude populations can be viewed as an outcome of independent replications of a natural experiment of convergent evolution. In such a case, an ancestral founding population moved from low to high altitude, and its descendants have been exposed for millennia to the opportunity for natural selection to improve their functions under high-altitude hypoxia<sup>10</sup>. The term ‘convergent evolution’ is defined as the development of the same or similar phenotypic adaptations under a similar external environmental condition as a consequence of natural selection. Although, it could not be denied that the recent multiple migrations from

<sup>1</sup>Department of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Khandwa Road, Simrol, Indore 453552, India. <sup>2</sup>Department of Applied Mathematics and Centre of Bioinformatics, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia. <sup>3</sup>Complex Systems Lab, Department of Physics, Indian Institute of Technology Indore, Khandwa Road, Simrol, Indore 453552, India. <sup>4</sup>Laboratory of Systems Medicine of Healthy Aging and Department of Applied Mathematics, Lobachevsky University, Nizhny Novgorod, Russia. ✉email: sarika@iiti.ac.in

corresponding lower-altitudes could affect the particular signature of high-altitude when compared with the respective low-altitude populations<sup>11</sup>.

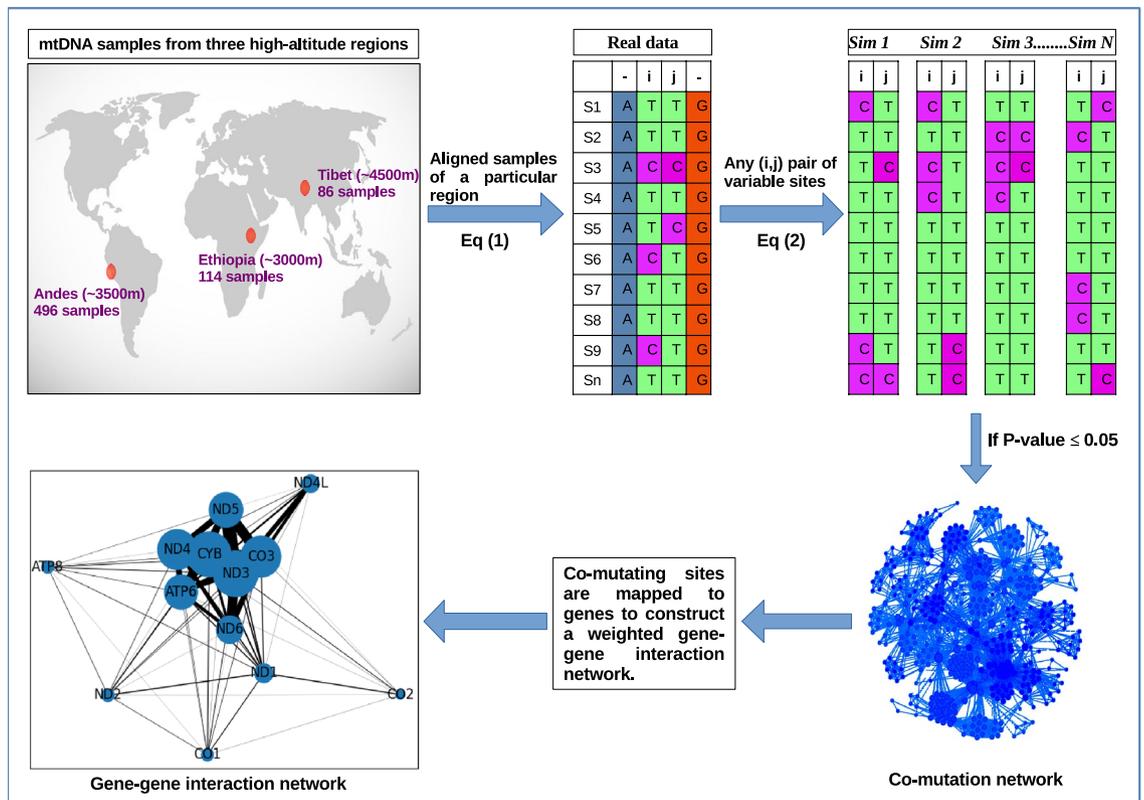
These three populations are believed to be evolved differently at the genetic<sup>12</sup> and physiological levels<sup>19</sup>. Various factors such as genetic pathways, molecular pathways, and phenotype levels have been attributed to the convergent evolution of humans and domestic animals<sup>10</sup>. Studies on the identification of nuclear genes for positive selection in highlander populations have provided evidence for natural selection in the genes responsible for hypoxia-related pathways<sup>13</sup>. In Tibetan highlanders two nuclear genes of the HIF (hypoxia inducible factor) pathway, HIF2A and PHD2 are known to be associated for positive selection<sup>14</sup>. The genetic variations in these genes were also found to be associated with hemoglobin concentration in Tibetans. Additionally, the presence of two non-coding SNPs, rs12097901 (C127S) and rs186996510 (D4E) were found to be as key variations in Tibetan highlanders<sup>15</sup>. Retrospectively, introgression from Denisovan or Denisovan-related individuals has been suggested to be affecting the pattern of high-altitude adaptation in Tibetans<sup>16</sup>. In Andean highlanders, out of 40 genes exhibiting positive selection, the  $\alpha$ -1 catalytic subunit of adenosine monophosphate-activated protein kinase (PRKAA1) gene has a significant role in high-altitude adaptation<sup>17</sup>. Further, among Ethiopian highlanders, the positive genetic signatures are known for aryl-hydrocarbon receptor nuclear translocator 2 (ARNT2), basic HLH family member e41 (BHLHE41), vav 3 guanine nucleotide exchange factor (VAV3), mitochondrial calcium uptake 1 (MICU), and thyroid hormone receptor (THRB) genes<sup>18</sup>. Among these three high-altitude populations, Andean and Tibetans represented similar set of genes for positive selection with specific attention to PHD2 gene than the Ethiopian population<sup>18</sup>.

The physiology is highly affected by less oxygen in the inhaled air at high altitudes, results in a lack of oxygen in the bloodstream flowing to the cells for oxygen-requiring energy-producing metabolic reactions in the mitochondria. Based on the factors contributing to arterial oxygen content like hemoglobin content, oxygen saturation, hemoglobin affinity, etc., there exist large pieces of evidences of the Andean-Tibetan difference for high-altitude adaptation. It has already been established that the three high-altitude populations possess significant differences at physiological and genomic levels from their respective low-altitude populations<sup>19,20</sup>. Andeans and Tibetans were reported to show increased hemoglobin concentration compared to corresponding low altitude individuals whereas, Ethiopian high-altitude dwellers reported no significant difference in their hemoglobin level with their low-altitude counterparts<sup>21</sup>. Among these three populations, Andeans showed the highest hemoglobin concentration in their blood. Another physiological trait associated with high altitude is oxygen saturation. Tibetan individuals showed the lowest oxygen saturation, followed by Andeans, and Ethiopians showed oxygen saturation values equivalent to sea level. These findings suggested that Andeans are less stressed by hypoxia than Tibetans, and Ethiopians can provide enough oxygen to their tissues even in a hypoxic environment. In summary, Andean characteristics are high hemoglobin concentration, higher arterial oxygen content, and low oxygen saturation than sea-level reference values. The Tibetans are characterized by sea-level hemoglobin concentration below 4000 m, moderate oxygen saturation, and lower arterial oxygen content than sea-level reference values. The Ethiopian patterns of hemoglobin concentration, oxygen saturation, and arterial oxygen content were found to be similar to those of low-altitude dwellers<sup>19</sup>. Apart from physiological differences, there exist specific polymorphisms belonging to mitochondrial genes *ND3* and *CYTb*, which are believed to be associated with high-altitude adaptation in the Tajiks population in Tibet native to China<sup>22</sup>. It has been observed that genetic polymorphisms rarely impart their effect at phenotypic levels individually but as a cohort of multiple interactions together<sup>23</sup>. Single nucleotide polymorphisms (SNPs) have been displayed to have a negligible impact on the heritability of few complex diseases<sup>24</sup>. Additionally, various studies have indicated that interactions of SNPs<sup>25–27</sup> are one of the critical factors in the manifestation of such complex diseases. Various computational methods have been developed and implemented to select a particular cohort of the variations and their interactions responsible for the manifestation of complex phenotypes. Such methods include principal component analysis to evaluate SNP correlations<sup>28</sup>, integrative scoring system based on their deleterious effects<sup>29</sup>, and Pareto-optimal approach<sup>30</sup>. There exist other approaches based on pair-wise interaction such that two variations significantly interacting through logic regression<sup>31</sup>, predictive rule inference<sup>32</sup>, and shrunken methodology<sup>33</sup>. Networks of variable sites were analyzed to identify genes of angiogenesis associated with breast cancer<sup>34</sup>, time-dependent weight dynamics in chickens<sup>35</sup>, feed efficiency in duroc and landrace pigs<sup>36</sup>, altitude-dependent interactions of mitochondrial genes in Asian population<sup>37</sup>.

In the current work, we analyze the possible role of mitochondrial co-mutations for these high-altitude populations in light of possible convergent evolution, using mtDNA genomes under the framework of networks. Foremost, we constructed the co-mutation network by selecting significantly interacting variations of the mitochondrial genome. These networks were found to follow the small-world behavior with high modularity. The weak ties, nodes with a low degree and high betweenness centrality, were found only in the Tibetan network and acted as haplogroup markers. Followed by that, a gene–gene interaction (GGI) network was constructed from the corresponding co-mutation networks for each population, and functional enrichment analysis was performed based on significantly interacting gene sets. Investigations of GGI networks pointed out essential role of *CYB* and *CO3* genes for high-altitude adaptation in Tibetan and Andean populations while *ND* genes for the Ethiopian population.

## Methods

**Sequence acquisition.** Complete human mitochondrial genome sequences were downloaded from the Human mitochondrial Database (HmtDB)<sup>38</sup> for the Ethiopia and Andes regions situated ~3000 m, and ~3500 m above sea level, respectively. For the Andes region, we have downloaded sequences from the Peru region since it inhabits the indigenous Andean (Aymara and Quechua) population. Tibetan sequences (~4000 m) were downloaded from the GenBank. Accession numbers and references are provided for all the sequences separately as



**Figure 1.** Construction of Co-mutation network and Gene–Gene Interaction (GGI) network for each high-altitude region as explained in “Methods” section.

supplementary information. All the sequences were aligned globally and mapped with a master sequence rCRS (revised Cambridge Reference Sequence)<sup>39</sup>.

**Construction of co-mutation network and gene–gene interaction (GGI) network.** We constructed two types of networks, first, the co-mutation networks where nodes were variable sites, and second, the weighted GGI networks where nodes were genes (Fig. 1) for each of the high-altitude population.

*Step 1 (co-mutation network)* Any position having more than one allele within the samples is considered a variable site. The variable sites were extracted from the aligned sequences for each region separately. For genomic equality, ambiguous nucleotides such as X, M, Y, etc. were replaced with ‘N’ for all the sequences and tri-allelic sites were not considered.

*Step 2* To construct a network for each high-altitude region, nodes were represented by the position of variable sites, and the edges were represented by co-mutation frequency between a pair of the nodes ( $C_{ij}$ ) defined, as

$$C_{ij} = \frac{(m_{ij})^2}{m_i m_j} \tag{1}$$

where ( $m_{ij}$ ) represents number of times the minor alleles occur together (defined as co-mutation pair) at *ith* and *jth* positions,  $m_i$  and  $m_j$  indicate total number of times the minor allele occurs at *ith* and *jth* position, respectively.

*Step 3 (p-value calculation)* To check the significance of any co-mutation pair, the threshold has been calculated as,

$$P_{ij} = \frac{[(C_{ij}^r) \geq (C_{ij})]}{10^4} \tag{2}$$

where ( $C_{ij}^r$ ) is the co-mutation frequency calculated after permuting the alleles at the *ith* and *jth* positions randomly. 10,000 random simulations were generated and  $P_{ij}$  was set to  $\leq 0.05$  (standard value) to consider a co-mutation pair significant.

*Step 4 (gene–gene interaction (GGI) network)* Nodes were mapped to their corresponding genes for each co-mutation network to achieve one weighted gene–gene interaction network for each high-altitude region. Since two or more co-mutation pairs may belong to the same gene or a gene pair, each link was counted as many times as it was found, and this number was considered as weight for that gene-pair. For example, the co-mutation pairs (3352–7623) and (4125–8054) would map to *ND1-CO2* gene pairs, so this edge was counted twice, and so on, similarly, the co-mutation pairs (3352–3489) would map to a self-loop for *ND1* gene. Since the variable

Region	Sample size	Nodes	Links
Tibet	86	398	3459
Ethiopia	114	838	13,770
Andes	496	1197	20,224

**Table 1.** Co-mutation networks: the number of nodes and connections in largest connected component.

sites contributed by each gene were in proportion to their lengths (except CR), the weight of a gene-pair was normalized by the sum of the total length of both the genes for that gene-pair. Additionally, we constructed GGI networks using the co-mutations of the (i) common nodes and (ii) exclusive nodes for all three regions. For (i), we took the nodes commonly present in all the networks and scanned for their co-mutating partners, and mapped these co-mutating nodes to respective genes. Similarly, exclusive nodes for each region were used to construct the corresponding GGI network.

**Null model and significant interactions.** Random networks were generated corresponding to each co-mutation network to compare with the real networks by taking the same number of the nodes as in the real network, and we connect them with a probability  $\alpha$  such that the total number of connections the random networks has the same as of the corresponding real-world;

$$\alpha = \frac{2N_c}{N(N-1)} \quad (3)$$

where,  $N_c$  is total number of connections and  $N$  is total number of nodes in the real network. Further, these random networks were used to get the random GGI networks as described in *Step 4*. We compared the real GGI networks with the corresponding random GGI networks and considered only those pairs significant whose weights fell two standard deviations away from the corresponding random one.

**Detection of community and role of nodes.** By calculating the modularity, we detected the communities using the algorithm given in<sup>40</sup> which is implemented in Python using *community module*. It is a modularity maximization algorithm. The role of each node in the communities is determined by its within-module degree,  $Z$  score, and the participation coefficient  $P$ . The within-module degree quantified the node's intra-modular connectivity and was calculated as the  $Z$ -score-transformed degree of centrality within the module. For a given node  $i$ ,  $Z_i$  is defined as,

$$Z_i = \frac{k_i - \bar{k}_i}{\sigma} \quad (4)$$

where  $k_i$  is degree of the  $i$ th node in its own community,  $\bar{k}_i$  is the average of  $k_i$  for all the nodes of that community, and  $\sigma$  is the standard deviation.  $Z_i$  takes a high value if degree of  $i$ th node is high within the cluster and vice versa.

Different roles of a node can also be deduced based on the number of connections the node makes with the nodes in the modules other than its own. For example, two nodes with the same  $Z$ -score will play different roles if one of them is connected to several nodes in other modules while the other is not. We define the participation coefficient  $P$  of node  $i$  as,

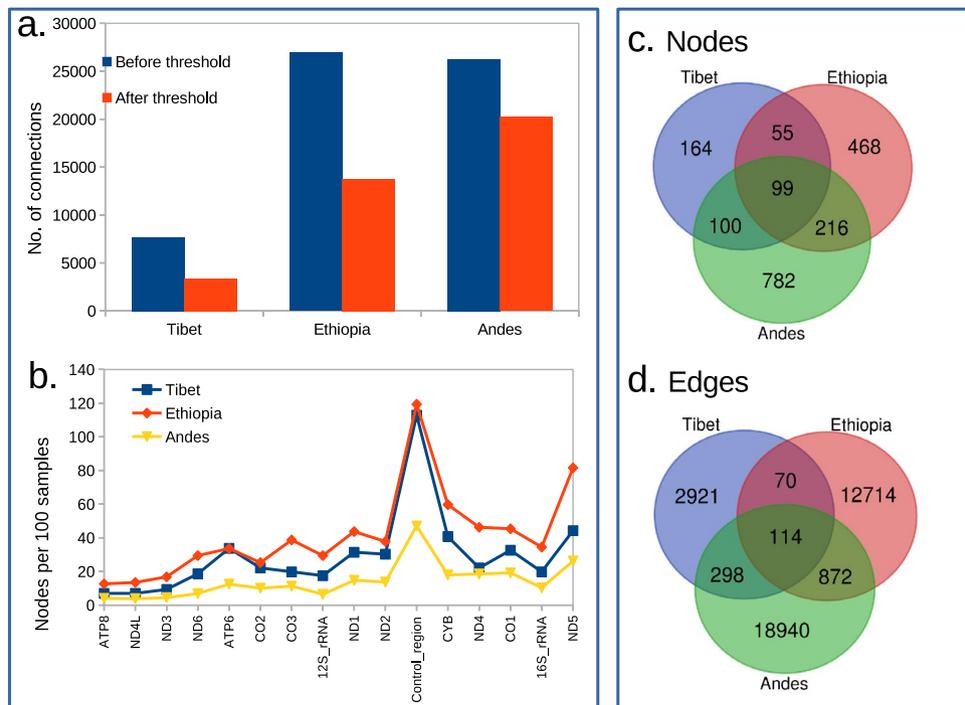
$$P_i = 1 - \sum_{S=1}^{Nm} \left( \frac{k_i}{K_i} \right)^2 \quad (5)$$

$K_i$  is the total degree of the node  $i$  in the whole network.  $S$  is the community and  $Nm$  is the total number of communities. The participation coefficient of a node is therefore close to 1 if its links are uniformly distributed among all the modules and 0 if all its links are within its own module.

**Ethics statement.** Ethical review and approval was not required for openly available human database. All methods were performed in accordance with relevant guidelines and regulations.

## Results

**Identification of significant interactions.** Among the three regions, the Andes region has the highest number of samples, nodes and connections (Table 1). More number of the samples rendered more number of interactions to be statistically significant, which is observed in the comparison of the number of connections ( $L$ ) at  $C_{ij} > 0$  between the Andes and Ethiopia. Both the regions were having an almost equal number of connections before applying the threshold. When significant pairs (with  $P_{ij} \leq 0.05$ ) were considered, the number of connections were decreased by  $\sim 23\%$  in Andes,  $\sim 49\%$  in Ethiopia and  $\sim 53\%$  in Tibet (Fig. 2a). It was further noted that above the threshold value, the number of co-mutations with low  $C_{ij}$  values ( $< 0.2$ ) were less, while co-mutations with high  $C_{ij}$  ( $> 0.2$ ) remained unaffected. This observation signifies the role of  $p$  value in determining the considerable interactions. Further, the distribution of  $C_{ij}$  for Andes population showed a heterogeneous distribution of the variations within the samples, i.e., there exist fewer co-mutations for  $C_{ij} > 0.8$ , indicating minor alleles in



**Figure 2.** (a) The change in number of connections with threshold, (b) Nodes which are participating in network construction were mapped to their respective genes and genes were counted and plotted on y-axis with their lengths on x-axis. Note that t-RNA genes are not shown. (c) Distribution of nodes and (d) co-mutation pairs across all the three regions.

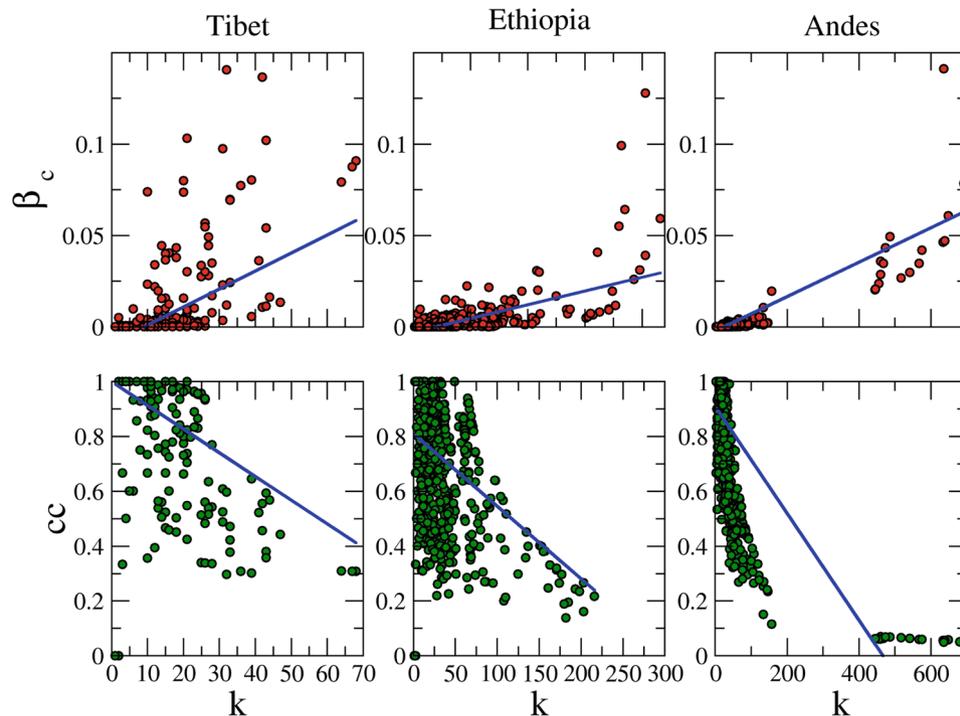
question are not always present in the same samples. The number of nodes and edges for all the three regions are also mentioned in Table 1. To get an overall idea for these three regions, we explored the common nodes and connections among these three co-mutation networks (Fig. 2c,d), to note down similarity and the differences among the underlying networks. It was found that in Tibet 41%, in Ethiopia 55%, and in Andes 65% nodes and ~90% connections among all the three regions were exclusively to each region. This suggests the co-evolution of mitochondrial variations pertaining to each geographic region.

**Structural properties of co-mutation networks.** As we have already established that in these co-mutation networks, a node was a nucleotide position and an edge was co-mutation frequency between any given pair of nucleotide positions. The degree of a node provided information about the frequency of co-mutation of any given variable site with the others. A node with a high degree (hub node) corresponds to a variable site undergoing high co-mutation with many variable sites. Such sites play a crucial role in shaping genome-wide co-evolution pattern of a population in view of multiple migrations and admixture events<sup>7</sup>. The hub node in Tibetan (12308, tRNA-Leu) and Andes (10398, ND3) co-mutation networks were commonly present in all three networks. In contrast, the hub node (4104, ND1) of Ethiopian co-mutation network was present in Andean and absent in the Tibetan network. It is noteworthy that these highest degree nodes were found to be haplogroup markers such that 12308 in Tibetan for K and U haplogroups of N lineage, 4104 in Ethiopian for L0, L1, L2, and L5 haplogroups of L lineage, and 10398 in Andean for multiple haplogroups of L (haplogroup frequency: 95%), M (99.5%) and N (17.1%) lineages. Since humans have migrated to the American continent much after the Eurasian migration, all the mtDNA haplogroups out of Africa descended from either M or N lineages. In the Andes, C and D haplogroups of M lineage contribute to 99% to haplogroup frequency. The revelation of these haplogroup markers as high degree nodes suggested the dominance of specific haplogroup backgrounds for each region's co-mutation of mtDNA variable sites. This also provides the biological relevance to network construction methodology along with the fundamental nature of co-evolution of haplogroup markers. Among the other high degree nodes, A15301G (CYB gene) node was found to be shared among all the three regions and was suggested to be a candidate site for functional analyses, and data association<sup>41</sup>.

Further, all the three networks were found to have small-world property characterized by high clustering coefficient<sup>42</sup>  $\langle C \rangle_{real} / \langle C \rangle_{rand} \gg 1$  and small diameter  $L_{real} / L_{rand} \sim 1$  as for many other real-world networks (Table 2)<sup>43,44</sup>. The small-world behavior shown by the brain networks suggests the swift flow of information in minimal steps from one region to another. Similarly, in co-mutation networks, the information of change in allele frequency of a certain nucleotide at one position sweeps to another nucleotide at another position in the same mtDNA sample. Although, for these co-mutation networks, it is a subject of further investigation that whether the two nodes connected through more than one step also share the information of change in allele frequency or not. In terms of co-mutation, this provides evidence for the fixation and inheritance of variations as a single cohort, and intragenic constraints<sup>37</sup> in the human mitochondrial genome. A high  $\langle C \rangle$  also implies

	Tibet	Ethiopia	Andes
Average degree	17	33	34
Clustering coefficient	0.8	0.7	0.8
Modularity	0.7	0.5	0.4
Average path length	4.3	2.8	2.3
Degree-degree correlation	0.4	0.2	-0.4

**Table 2.** Global properties of co-mutation networks.

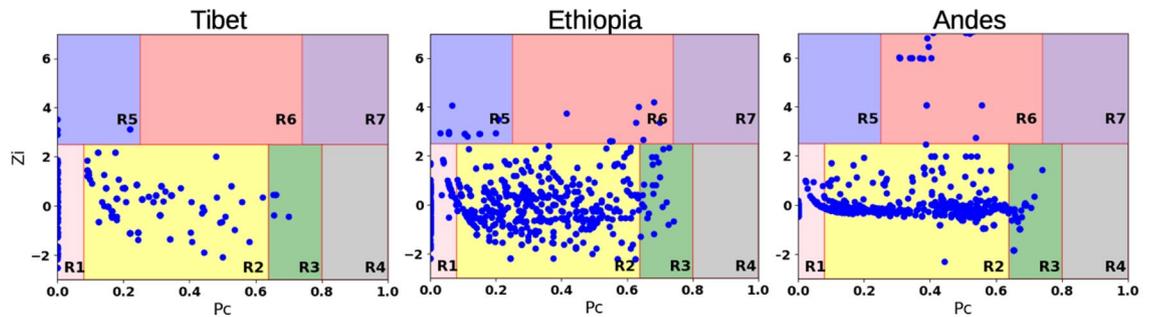


**Figure 3.** Betweenness centrality (upper panel) and clustering coefficient (lower panel) are plotted as function of degree for all the three regions.

that any given variable site prefers to co-mutate with all the other genes throughout the mitochondrial genome except for tRNA genes.

To capture hierarchy, another characteristic property of networks, the clustering coefficient of each node with its degree (Fig. 3, lower panel) was plotted. A decrease in the tendency of a variable site to form clusters with an increase in the degree of that node (Fig. 3) implies the presence of hierarchy<sup>45,46</sup> in these co-mutation networks. The hierarchical organization confers robustness and adaptability in complex biological networks<sup>47</sup>. mtDNA has acquired several variations depending on biotic and abiotic factors since humans have first migrated outside Africa throughout the world<sup>48</sup>. These enriched variations gave rise to multiple haplogroups. The presence of high clustering and hierarchy in these co-mutation networks might help capture this temporal and spatial co-evolution of variable sites of the mitochondrial genome in the form of haplogroups.

Resilience is an important property for a network, which is measured by the betweenness centrality ( $\beta_c$ )<sup>6</sup>. This centrality measure estimates the number of shortest paths between any given pair of nodes which increases if a node is removed. Usually, the nodes with a high degree tend to have high betweenness centrality. However, it was observed here that a few nodes, despite having a low degree, have high betweenness centrality and are considered as weak ties. Weak ties are the nodes that co-mutate with a few nodes but from different modules. Presence of weak ties suggests that mtDNA has evolved through co-evolution of few nodes (pertaining to low  $k$ ) of multiple discrete modules (pertaining to high  $\beta_c$ )<sup>49</sup>. Thus, these sites are significantly important since their removal can result in the breakdown of the network. In the Tibetan co-mutation network, we found four such variable sites (709, 15,927, 16,172, and 16,362). Interestingly, all these nodes were again found to be haplogroup markers (709: L6, G, T, and W; 15,927: G, B, and X; 16,172: L0 and F; 16,362: L4, D, G, and A). This suggests that haplogroup markers provide the necessary evolutionary background and play a key role in assisting the co-evolution of different clusters. Moreover, 15,927 node belongs to *tRNA-Thr* which is one of the highly mutated tRNAs among all the tRNAs in humans<sup>50</sup>, and similarly, variable sites 709 (*12S-rRNA*) in *rRNA*. tRNAs and rRNAs play a central role in protein synthesis. This signifies that tRNAs and rRNAs might play decisive roles in the co-evolution of



**Figure 4.** Roles of nodes in Z–P parameter space. Each node in a network can be characterized by its within-module degree and its participation coefficient. Nodes with  $Z > 2.5$  were classified as module hubs and nodes with  $Z < 2.5$  as non-hubs. Non-hub nodes can be naturally assigned into four different roles: (R1) ultra-peripheral nodes; (R2) peripheral nodes; (R3) non-hub connector nodes; and (R4) non-hub kinless nodes. Hub nodes can be naturally assigned into three different roles: (R5) provincial hubs; (R6) connector hubs; and (R7) kinless hubs.

different mutational cohorts in the Tibetan population. On the contrary, we did not find any such nodes in Ethiopian or Andean co-mutation networks. In these two networks, the nodes with high betweenness centrality also possessed a large degree. This suggests that in Ethiopian and Andean populations, the mtDNA has evolved through continuous co-evolution of many different nodes (of high  $k$ ) of multiple modules (of high  $\beta_c$ ) altogether.

Another characteristic property, the network diameter, defined as the longest of all shortest paths, was large in Tibet compared to Ethiopia and Andes. This large network diameter gives evidence about long-range co-mutation and high modularity in the Tibetan population. To detect modularity in these co-mutation networks we applied the Louvain community detection algorithm<sup>40</sup> implemented in the python *community module*. The high modularity indicates the formation of mutational cohorts of evolutionary constraints at the whole-genome level. We analyzed the genetic background of the communities formed in these co-mutation networks. On considering only the coding regions, it was observed that nodes of a few particular genes contribute more than other genes in each community in all the three co-mutation networks (Supplementary Table 1). Particularly, in Tibet, *ATP6*, *CYB*, *ND5* genes, in Ethiopia *ND5* gene, and in Andes *ND5*, and *CYB* genes showed considerable contribution among all the communities. *CO1* and *ND2* genes were also found to dominate at least in one community in each of the three populations. We also analyzed the highest degree nodes for the coding region in each community. These high degree nodes were considered "community cores". We found that none of these community cores are common among the three regions. Although the three regions have a certain number of common nodes (Fig. 2), the community structures are derived by independent nodes. This supports the fundamental nature of formation of various haplogroups due to migration patterns and events of natural selection which were derived by a few specific variations<sup>51</sup>. Upon individual inspection of the communities, we found that in each of the three regions, despite contributing few nodes *tRNA-Leu*, *tRNA-Lys*, and *tRNA-Gln* were found to be community cores in the Tibet, Ethiopia, and Andes regions, respectively. Apart from that, the *CR* was found to be evenly present in all the communities.

We investigated the localization properties of eigenvectors of these co-mutation networks. Localization of eigenvectors enjoy a wide range of applications in networks; in disease spreading<sup>52</sup>, perturbation of propagation in mutualistic networks<sup>53</sup>. Other applications of localization can also be found in<sup>54,55</sup>. To quantify localization, we used correlation dimension ( $D_2$ ) calculated by using the box-counting method for multifractal analysis of eigenvectors<sup>56</sup>. If  $D_2 \rightarrow 0$ , eigenvector is said to be localized while if  $D_2 \rightarrow 1$ , eigenvector is considered as delocalized. Thus,  $D_2$  provides insight into the degree of localization of eigenvectors. We focus on the eigenvectors of eigenvalues nearer to zero and  $D_2$  is averaged over all the eigenvectors inside width  $d\lambda = 0.5$ . Note that, slight increase or decrease in the width will not alter the results. Tibetan and Andean networks were more localized, with  $D_2$  being 0.43 compared to the Ethiopian network with  $D_2$  0.65. In these networks, the tendency of co-mutation is captured by  $D_2$  in terms of localization. A co-mutation occurs when minor alleles at any given two sites are present in considerable frequency in the population. The change in this co-mutation frequency is further affected by the introduction of new DNA samples harboring that particular minor allele. In other words the co-mutation is being localized around a few sites. Migration and natural selection are few of the events which might cause a change in allele frequency at certain positions, which further affects the tendency of that site to co-mutate. In Andean and Tibetan populations, the co-mutation has been localized compared to Ethiopia. This might be possible due to the recent population admixture experienced by the Ethiopian population<sup>11</sup> due to its comparatively lesser harsh environment than Tibet and Andes. Thus, population admixture might have played a role in observed localization behavior in these three populations.

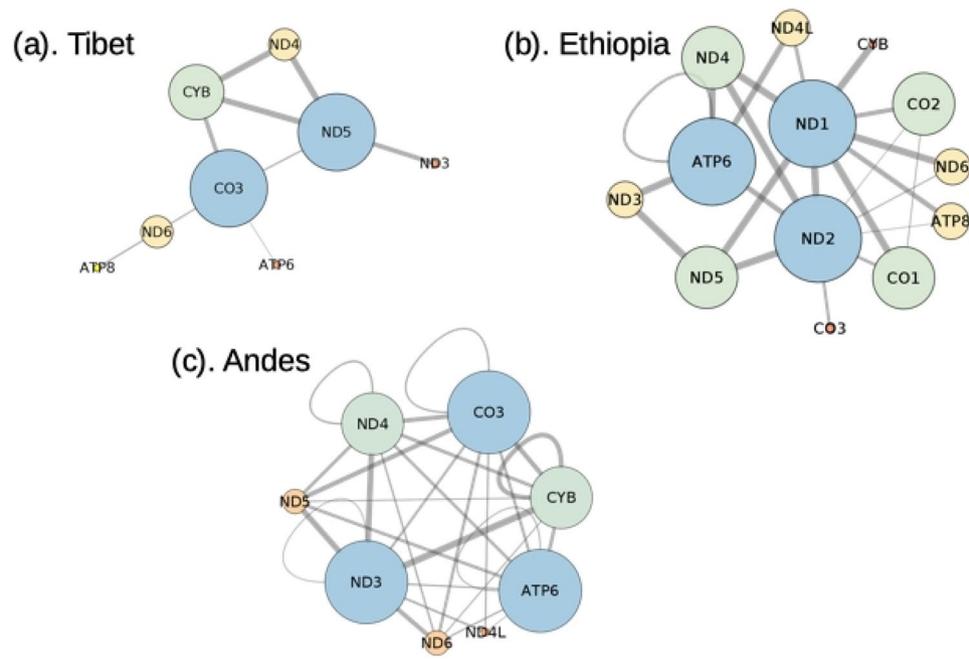
Further, for these networks, there exists no node with its degree distinctly very high than those of the other nodes, and hence the importance of a node cannot be assigned based on its degree only. Nevertheless, due to the presence of the high modularity, the importance of a node can be determined, to a great extent, by its within-module degree and participation coefficient, which defines how a node is positioned in its own module and with respect to other modules<sup>57,58</sup>. Based on the within-module degree and the participation coefficient, nodes were categorized as module hubs and non-hubs (Fig. 4). The highest degree nodes for the Tibetan co-mutation network were found in R3, non-hub connector category, while for the Ethiopian and Andean co-mutation networks, the highest degree nodes were found in the R6 connector hubs category. In the Tibetan co-mutation network, the

nodes in R5 category were 3010 (*16S\_rRNA*), 8414 (*ATP8*), 14,668 (*ND6*) and 12,361 (*ND5*). Variable site 3010 was shown to be a high-altitude marker in Tibetan population<sup>59</sup> and also reported from network motifs with variable sites 8414 and 14668<sup>37</sup> while, 12,361 was shown to be associated with nonalcoholic fatty liver disease<sup>60</sup>. It is noteworthy that in the Tibetan co-mutation network, there were no nodes in the R6 category, and in the Andean co-mutation network, there were no nodes in the R5 category, while in the Ethiopian co-mutation network nodes were present in both the R5 and R6 categories. As we know, provincial hubs (R5) tend to connect nodes within the same module while connector hubs (R6) tend to connect nodes from different modules. Based on the observation that the modules inherit the information of haplogroups, we can consider R5 nodes as intra-module hubs and R6 category as inter-module hubs. It can be inferred that in Tibetan population, intra-haplogroup co-evolution is prominent, while in the Andes, mtDNA inter-haplogroup co-evolution is prominent. On the other hand, Ethiopian mtDNA harbors both inter and intra-haplogroup co-evolution. This again provides evidence for recent admixture in Ethiopian region<sup>11</sup>. These co-mutation networks are then analyzed at the gene level through the gene–gene interaction networks, discussed further.

**Gene–gene interaction (GGI) networks.** The contribution of each gene was quantified by counting the number of variable sites from each gene (Fig. 2b). It was observed that the number of nodes in the network were proportional to the length of genes, hence we normalized the number of variable sites with the corresponding gene lengths (Fig. 2b). It was deduced that except for the Control region (CR), the occurrence of variable sites for each gene increased with an increase in the length of genes. CR is a mutational hot-spot in mtDNA, hence contributed the highest number of variable sites. Since CR does not code for any protein, we did not consider its interaction at the gene–gene network level to avoid bias due to the high mutation rate. It was evident that certain genes contributed more variable sites in the network construction than others in a particular region. Especially, *ATP6*, *CO2* and *ND2* genes are contributing equally in the Tibetan and Ethiopian networks while *12S-rRNA*, *16S-rRNA*, *CO3* and *ND4* were contributing equally in the Tibetan and Andean networks. Among the coding genes, the *ND5* gene showed the highest difference of contributing nodes with the minimum in Tibet and maximum in the Andes (Fig. 2b). The contribution of each gene per 100 samples for the network construction was highest in Ethiopia among all the regions. The nodes pertaining to CR displayed the largest participation in the network construction due to the fact that it is the highly variable part of mtDNA<sup>61</sup>. Contribution of the variable sites in each gene yields partial information about the interaction of the genes. To overcome this, we generated the gene–gene interaction networks by mapping the variable sites with the respective genes, as discussed in the *Methods* section. The gene–gene interaction networks provide a holistic and a reductionist approach to investigate interactions in the three high-altitude regions. We found 17 gene pairs in Tibetan, 23 gene pairs in Ethiopian, and 44 gene pairs in the Andes population after comparing with their corresponding random networks. Among these, the pair with the highest edge weight *ATP6-CYB* (Tibet), *ND4-ND5* (Ethiopia), and *CYB-ND4* (Andes). Four significant gene–gene pairs were found to be commonly present in all the three populations, which were *CO1-CO2*, *ND2-ND4*, *ND3-ND4* and *ND4-ND5*. All these genes are involved in the oxidative phosphorylation pathway (KEGG entry: 00190) and thermogenesis pathway (KEGG entry: 04714)<sup>62</sup>, along with that these genes are also found to be involved in cellular respiration (GO:0045333), and response to abiotic stimulus (GO:0009628)<sup>63</sup>. At higher-altitudes where low temperature and hypoxia are two main abiotic factors responsible for natural selection, the genes involved in thermogenesis and response to abiotic factors play an imperative role in determining the evolution, and adaptation<sup>20</sup>.

Since these three populations are believed to share a similar physical environment and to undergo the process of convergent evolution, for all the common nodes, we extracted their co-mutations and constructed corresponding GGI. The common variable sites categorized these three populations under the same haplogroups, while their co-mutations differ among these three populations. To capture this difference at the genetic level, significant genetic interaction (Fig. 5) of the common nodes were extracted based on the functional enrichment analysis for GO terms and KEGG pathways using DAVID<sup>64</sup>. It was found that in the Tibetan population *CO3*, *CYB* and *ND5* genes, in the Andean population *ATP6*, *CO3*, *CYB*, *ND3* and *ND4* genes, and in the Ethiopian population *ATP6*, *CO1*, *CO2*, *ND1*, *ND2*, *ND4* and *ND5* genes were significantly interacting with other genes. The functional enrichment of these gene sets is shown in Table 3. It is noteworthy that from the cytochrome oxidase complex, the *CO3* sub-unit was interacting in the Tibetan and Andean populations while *CO1* and *CO2* sub-units were interacting in the Ethiopian population. *CO3* sub-unit is the putative site for the entry of oxygen into the large cytochrome oxidase complex, thereby regulating its activity under hypoxic conditions<sup>65</sup>. Even though the Ethiopian gene set has not shown any feature related to the hypoxia adaptation in functional enrichment analysis, variations in *ND1* and *ND2* genes were shown to be associated with high-altitude hypoxia in Tibetan yak<sup>66</sup> and endemic Ethiopian rats<sup>67</sup>. Apart from the functional enrichment, the Tibetan and Andean gene-sets were also involved in non-alcoholic fatty liver disease (NAFLD) pathways. It has been shown that high altitude might improve the mitochondrion function and alleviate the NAFLD<sup>68</sup>.

Further to explore the role of such genetic interactions at the region-specific level, we extracted the co-mutation pairs pertaining to the exclusive nodes of each region and constructed GGI networks. In GGI networks with these exclusive nodes, we found specific interactions with significantly lower weights and others with significantly higher weights than the corresponding random ones (Fig. 6). It is readily observed from Fig. 6 that those genetic interactions that were significantly up in the Andean and Tibetan populations were significantly down in the Ethiopian population and vice versa. This suggests that both Tibetan and Andean populations have evolved at high altitudes through the interactions of *CYB* and *CO3* genes. In contrast, the Ethiopian population deviated in sharing the mitochondrial genetic interactions with the other two populations. This dissimilarity could be explained based on two facts; one is that Ethiopia is situated at the lowest altitude among all the three



**Figure 5.** Significant gene–gene interactions of common node co-mutations (the node size depicts the degree of the node and edge size represents the edge weight).

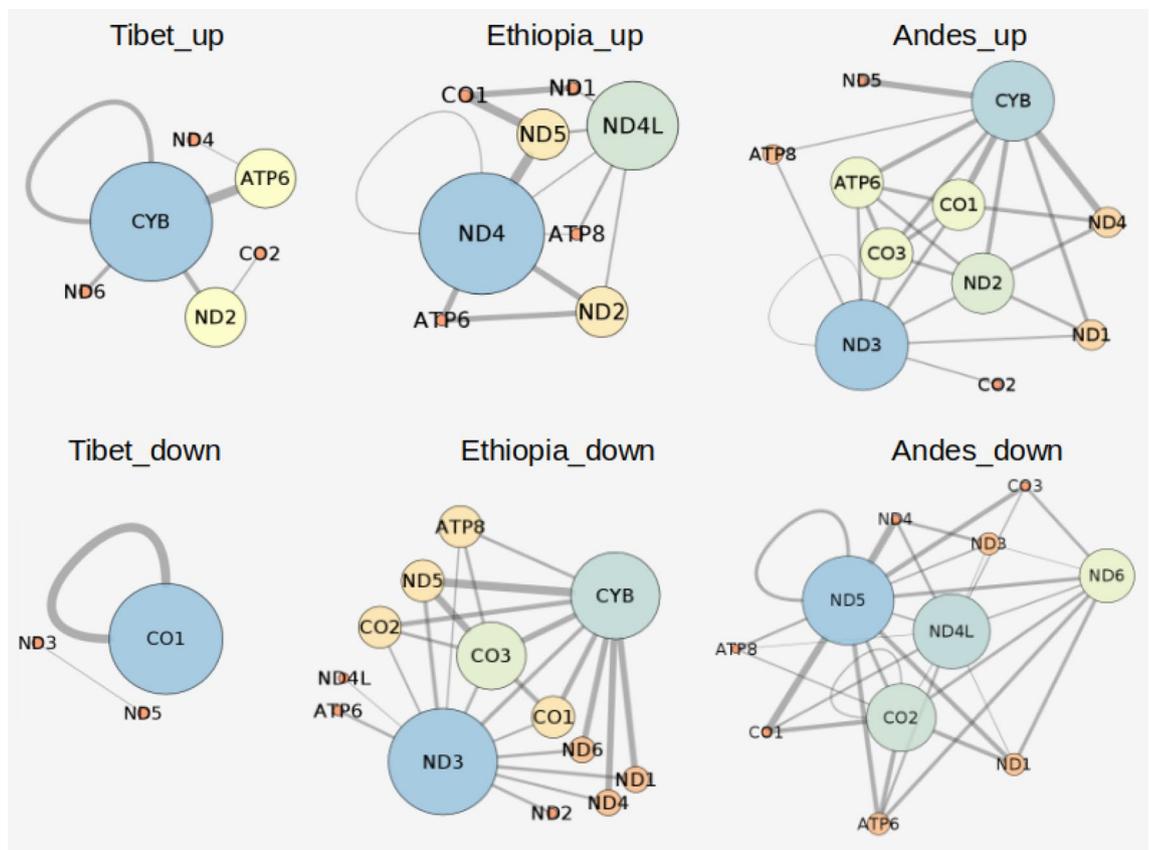
	Tibet (CO3, CYB, ND5)	Ethiopia (ATP6, CO1, CO2, ND1, ND2, ND4, ND5)	Andes (ATP6, CO3, CYB, ND3, ND4)
Response to hypoxia [GO: 0001666]	Yes	No	Yes
Response to hyperoxia [GO: 055093]	No	No	Yes
Respiratory electron transport [GO: 022904]	Yes	No	Yes
ATP synthesis coupled electron transport [GO: 042773]	No	Yes	No
Non-alcoholic fatty acid liver disease [KEGG: 04932]	Yes	No	Yes

**Table 3.** The GO terms and KEGG pathways for gene sets described in Fig. 5.

populations, and the second is that Ethiopia is believed to underwent frequent admixtures in its gene pool from a lower altitude populations<sup>69</sup>.

## Conclusion

Although haplogroups or specific mutations help us categorizing the human population geographically, the proposed co-mutation networks fortify the specific genetic interactions even in similar environmental backgrounds. Our analysis showed that mtDNA has evolved with similar biological mechanisms in Andean and Tibetan populations than the Ethiopian population. We found a heterogeneous set of genes in Ethiopian population than the Tibetan and Andean populations compared to corresponding random networks. Notably, CYB and CO3 genes are commonly present in Tibetan and Andean population, and are interacting with ND5 gene in Tibetan and ATP6, ND3 and ND4 genes in Andean populations. Whereas, in the Ethiopian population four NADH dehydrogenase genes (ND1, ND2, ND4, ND5) showed interactions with ATP6, CO1 and CO2 genes. It was noticeable that in exclusive GGI networks, Ethiopian population showed the contrasting behavior compared to the Tibetan and Andean population. Further, the  $D_2$  analysis also showed that Tibetan and Andean populations are similar in their localization behavior compared to Ethiopian population. The tendency of variable sites to co-mutate could be affected by introduction of either new samples or new variations in the existing samples, the  $D_2$  analysis would be employed to capture such admixture or multiple migration events or genetic drift in a particular population. To conclude, co-mutation based genetic interaction networks identified the gene sets which could have played critical role in establishing the human lineage and acclimatization to higher altitudes around the globe. These gene sets and pertaining variable sites provide a ground for further investigation of patterns of human migration and settlements across these three regions.



**Figure 6.** Significant gene–gene interactions of exclusive nodes co-mutations. The Region\_up and Region\_down represents those interactions which were showing higher and lower weight than the corresponding random networks for each region, respectively (the node size depicts the degree of the node and edge size represents the edge weight).

Received: 26 June 2021; Accepted: 17 December 2021  
Published online: 07 February 2022

## References

- Pavlopoulos, G. A. *et al.* Using graph theory to analyze biological networks. *BioData Min.* **4**, 10 (2011).
- Vazquez, A. *et al.* Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* **21**, 697–700 (2003).
- Shinde, P. & Jalan, S. A multilayer protein-protein interaction network analysis of different life stages in *Caenorhabditis elegans*. *EPL* **112**, 58001 (2015).
- Rai, A., Pawar, A. K. & Jalan, S. Prognostic interaction patterns in diabetes mellitus II: A random-matrix-theory relation. *Phys. Rev. E* **92**, 022806 (2015).
- Shinde, P., Sarkar, C. & Jalan, S. Codon based co-occurrence network motifs in human mitochondria. *Sci. Rep.* **8**, 3060 (2018).
- Jalan, S. & Sarkar, C. Complex networks: An emerging branch of science. *Phys. News* **47**, 3–4 (2017).
- Mishmar, D. *et al.* Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. U.S.A.* **100**(1), 171–176 (2003).
- Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V. & Wallace, D. C. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303**(5655), 223–226 (2004).
- Ruiz-Pesini, E. & Wallace, D. C. Evidence for adaptive selection acting on the tRNA and rRNA genes of human mitochondrial DNA. *Hum. Mutat.* **27**(11), 1072–1081 (2006).
- Witt, K. E. & Huerta-Sanchez, E. Convergent evolution in human and domesticated adaptation to high-altitude environments. *Philos. Trans. R. Soc. Lond. B* **374**(1777), 20180235 (2019).
- Simonson, T. S. Altitude adaptation: A glimpse through various lenses. *High Alt. Med. Biol.* **16**(2), 125–137 (2015).
- Bigham, A. W. & Lee, F. S. Human high-altitude adaptation: Forward genetics meets the HIF pathway. *Genes Dev.* **28**(20), 2189–2204 (2014).
- Beall, C. M. *et al.* Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proc. Natl. Acad. Sci. U.S.A.* **107**(25), 11459–11464 (2010).
- Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**(5987), 72–75 (2010).
- Lorenzo, F. R. *et al.* A genetic mechanism for Tibetan high-altitude adaptation. *Nat. Genet.* **46**(9), 951–956 (2014).
- Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**(7513), 194–197 (2014).
- Bigham, A. W. *et al.* Maternal PRKAA1 and EDNRA genotypes are associated with birth weight, and PRKAA1 with uterine artery diameter and metabolic homeostasis at high altitude. *Physiol. Genom.* **46**(18), 687–697 (2014).
- Huerta-Sánchez, E. *et al.* Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. *Mol. Biol. Evol.* **30**(8), 1877–1888 (2013).

19. Beall, C. M. Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. *Integr. Comp. Biol.* **46**(1), 18–24 (2006).
20. Storz, J. F. High-altitude adaptation: Mechanistic insights from integrated genomics and physiology. *Mol. Biol. Evol.* **38**(7), 2677–2691 (2021).
21. Gassmann, M. *et al.* The increase in hemoglobin concentration with altitude varies among human populations. *Ann. N. Y. Acad. Sci.* **1450**(1), 204–220 (2019).
22. Chen, Y. *et al.* Mitochondrial DNA genomes revealed different patterns of high-altitude adaptation in high-altitude Tajiks compared with Tibetans and Sherpas. *Sci. Rep.* **10**, 10592 (2020).
23. Papp, B. & Pál, C. Systems biology of epistasis: Shedding light on genetic interaction Network “hubs”. *Cell Cycle* **10**(21), 3623–3624 (2011).
24. Jakobsdottir, J. *et al.* Interpretation of genetic association studies: Markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet.* **5**(2), e1000337 (2009).
25. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37**(4), 413–417 (2005).
26. Cordell, H. J. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**(20), 2463–2468 (2002).
27. Moore, J. H. & Williams, S. M. Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* **85**(3), 309–320 (2009).
28. Horne, B. D. & Camp, N. J. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genet. Epidemiol.* **26**(1), 11–21 (2004).
29. Lee, P. H. & Shatkay, H. An integrative scoring system for ranking SNPs by their potential deleterious effects. *Bioinformatics* **25**(8), 1048–1055 (2009).
30. Lee, P. H., Jung, J. Y. & Shatkay, H. Functionally informative tag SNP selection using a pareto-optimal approach: Playing the game of life. *BMC Bioinform.* **10**(Suppl 13), O5 (2009).
31. Schwender, H. & Ickstadt, K. Identification of SNP interactions using logic regression. *Biostatistics* **9**(1), 187–198 (2008).
32. Wan, X. *et al.* Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics* **26**(1), 30–37 (2010).
33. Liu, Y. & Ng, M. Shrunk methodology to genome-wide SNPs selection and construction of SNPs networks. *BMC Syst. Biol.* **4**(Suppl 2), S5 (2010).
34. Lin, H. Y. *et al.* SNP-SNP Interaction network in angiogenesis genes associated with prostate cancer aggressiveness. *PLoS ONE* **8**(4), e59688 (2013).
35. Li, F. G. & Li, H. A time-dependent genome-wide SNP-SNP interaction analysis of chicken body weight. *BMC Genom.* **20**, 771 (2019).
36. Banerjee, P., Carmelo, V. & Kadarmideen, H. N. Genome-wide epistatic interaction networks affecting feed efficiency in duroc and landrace pigs. *Front. Genet.* **11**, 121 (2020).
37. Verma, R. K. *et al.* Analysis of human mitochondrial genome co-occurrence networks of Asian population at varying altitudes. *Sci. Rep.* **11**(1), 133 (2021).
38. Clima, R. *et al.* HmtDB 2016: Data update, a better performing query system and human mitochondrial DNA haplogroup predictor. *Nucleic Acids Res.* **45**(D1), D698–D706 (2017).
39. Andrews, R. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147 (1999).
40. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**(10), 1008 (2008).
41. Tranah, G. J. *et al.* Mitochondrial DNA variation in human metabolic rate and energy expenditure. *Mitochondrion* **11**(6), 855–861 (2011).
42. Brown, K. S. *et al.* The statistical mechanics of complex signaling networks: Nerve growth factor signaling. *Phys. Biol.* **1**(3–4), 184–195 (2004).
43. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**(6804), 651–654 (2000).
44. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. Biol. Sci.* **268**(1478), 1803–1810 (2001).
45. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**(6684), 440–442 (1998).
46. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999).
47. Mengistu, H., Huizinga, J., Mouret, J. B. & Clune, J. The evolutionary origins of hierarchy. *PLOS Comput. Biol.* **12**(6), e1004829 (2016).
48. Soares, P. *et al.* The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* **29**(3), 915–927 (2012).
49. Rai, A. *et al.* Understanding cancer complexome using networks, spectral graph theory and multilayer framework. *Sci. Rep.* **7**, 41676 (2017).
50. Zifa, E. *et al.* Mitochondrial tRNA mutations: Clinical and functional perturbations. *RNA Biol.* **4**(1), 38–66 (2007).
51. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–E394 (2009).
52. Goltsev, A. V., Dorogovtsev, S. N., Oliveira, J. G. & Mendes, J. F. Localization and spreading of diseases in complex networks. *Phys. Rev. Lett.* **109**, 128702 (2012).
53. Suweis, S., Grilli, J., Banavar, J. R., Allesina, S. & Maritan, A. Effect of localization on the stability of mutualistic ecological networks. *Nat. Commun.* **6**, 1–7 (2015).
54. Pastor-Satorras, R. & Castellano, C. Distinct types of eigenvector localization in networks. *Sci. Rep.* **6**, 18847 (2016).
55. Pradhan, P., Yadav, A., Dwivedi, S. K. & Jalan, S. Optimized evolution of networks for principal eigenvector localization. *Phys. Rev. E* **96**, 022312 (2017).
56. Mishra, A., Bandyopadhyay, J. N. & Jalan, S. Multifractal analysis of eigenvectors of smallworld networks. *Chaos Solitons Fractals* **144**, 110745 (2021).
57. Rives, A. W. & Galitski, T. Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U.S.A.* **100**(3), 1128–1133 (2003).
58. Han, J. D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**(6995), 88–93 (2004).
59. Luo, Y., Gao, W., Liu, F. & Gao, Y. Mitochondrial nt3010G-nt3970C haplotype is implicated in high-altitude adaptation of Tibetans. *Mitochondrial DNA* **22**(5–6), 181–190 (2011).
60. Lu, M. Y. *et al.* Mitochondrial polymorphism 12361A>G is associated with nonalcoholic fatty liver disease. *Transl. Res.* **159**(1), 58–59 (2012).
61. Stoneking, M., Hedgecock, D., Higuchi, R. G., Vigilant, L. & Erlich, H. A. Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Am. J. Hum. Genet.* **48**(2), 370–382 (1991).
62. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**(1), 27–30 (2000).
63. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**(D1), D330–D338 (2019).

64. Huang, D., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**(1), 44–57 (2009).
65. Lau, G. Y., Mandic, M. & Richards, J. G. Evolution of cytochrome c oxidase in hypoxia tolerant sculpins (Cottidae, Actinopterygii). *Mol. Biol. Evol.* **34**(9), 2153–2162 (2017).
66. Shi, Y. *et al.* Genetic diversities of MT-ND1 and MT-ND2 genes are associated with high-altitude adaptation in yak. *Mitochondrial DNA Part A* **29**(3), 485–494 (2018).
67. Bartáková, V., Bryjová, A., Nicolas, V., Lavrenchenko, L. A. & Bryja, J. Mitogenomics of the endemic Ethiopian rats: Looking for footprints of adaptive evolution in sky islands. *Mitochondrion* **57**, 182–191 (2021).
68. Song, K., Zhang, Y., Ga, Q., Bai, Z. & Ge, R. L. High-altitude chronic hypoxia ameliorates obesity-induced non-alcoholic fatty liver disease in mice by regulating mitochondrial and AMPK signaling. *Life Sci.* **252**, 117633 (2020).
69. Hodgson, J. A., Mulligan, C. J., Al-Meerri, A. & Raaum, R. L. Early back-to-Africa migration into the Horn of Africa. *PLoS Genet.* **10**(6), e1004393 (2014).

## Acknowledgements

SJ, RKV and MI, respectively, acknowledges support from SERB grant (EMR/2016/001921) and CSIR grant (CSIR,25(0293)/18/EMR-II) Govt of India, University Grant Commission for SRF fellowship (305089), and the Ministry of education and science of the Russian Federation mega grant (075-15-2021-639). We are thankful to Pramod Shinde and Aparna Rai for continuous discussions throughout and Cristina Giuliani for useful suggestions at the beginning of the project.

## Author contributions

S.J. and R.K.V. conceived and designed the project. R.K.V. and A.M generated the data and performed the network analysis. R.K.V., S.J., M.I. made the interpretations. R.K.V. and A.K. worked on the result and discussion sections. All the authors wrote and approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-05719-5>.

**Correspondence** and requests for materials should be addressed to S.J.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022