Altitude-wise analysis of co-occurrence networks of mitochondrial genome in Asian population

Rahul K Verma[1], Cristina Giuliani[2], Alena Kalyakuina[3,4], Ajay Deep Kachhvah[5], Mikhail Ivanchenko[3,4], and Sarika Jalan[1,4,5*]

[1]Discipline of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Khandwa Road, Simrol, Indore-453552, India
[2]Laboratory of Molecular Anthropology & Center for Genome Biology, Department of Biological, Geological and Environmental Sciences, University of Bologna, Italy
[3]Department of Applied Mathematics and Centre of Bioinformatics, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia
[4]Laboratory of Systems Medicine of Healthy Aging and Department of Applied Mathematics, Lobachevsky University, Nizhny Novgorod, Russia
[5]Complex Systems Lab, Discipline of Physics, Indian Institute of Technology Indore, Khandwa Road, Simrol, Indore-453552, India
[*]sarika@iiti.ac.in

ABSTRACT

Finding mechanisms behind high altitude adaptation in humans at the Tibetan plateau has been a subject of evolutionary research. Mitochondrial DNA (mt-DNA) variations have been established as one of the key players in understanding the biological mechanisms at the basis of adaptation to these extreme conditions. To explore cumulative effects and dynamics of the variations in mitochondrial genome at varying altitudes, we investigated human mt-DNA sequences from NCBI database at different altitudes by employing co-occurrence motifs framework. We constructed co-occurrence motifs by taking into account variable sites for each altitude group. Analysis of the co-occurrence motifs using similarity clustering revealed a clear distinction between a lower and a higher altitude region. In addition, the previously known high altitude markers 3394 and 7697 (which are definitive sites of haplogroup M9a1a1c1b) were found to co-occur within their own gene complexes indicating the impact of intra-genic constraint on co-evolution of nucleotides. Furthermore, an ancestral marker 10398 was found to co-occur only at higher altitudes supporting the fact that a separate root of colonization at these altitudes might have taken place. Overall, our analysis revealed the presence of co-occurrence motifs at a whole mitochondrial genome level. This study, combined with the classical haplogroups analysis is useful in understanding role of co-occurrence of mitochondrial variations in high altitude adaptation.

Introduction

The origin and inhabitation of humans in diverse geographical regions across the world has always been a topic of research for anthropologists and geneticists. Environmental diversity present in different geographical regions being one of the key factors in causing variability among human groups both at nuclear DNA and mtDNA level[1]. A wide range of environmental diversity exists in terms of temperature and altitude driven hypoxia all over the world. One of such environments exists in South-Central Asia at Tibetan plateau. The Tibetan plateau is known to be the highest altitude region ever inhabited by humans since the last Largest Glacial Maxima (LGM, 22−18 kya)[2]. The plateau has an average elevation of 4000m above sea level yielding extreme environments such as low oxygen concentration, high UV radiation and arid conditions[3]. The indigenous people of Tibet have acquired an ability to thrive in the hypoxic environment as a result of complex mechanisms of polygenic adaptations (both at nuclear and mtDNA level) [4]. Thus, biological study of the Tibetan plateau is of great interest due to its distinctive environment and migratory profile.

Mitochondria are the energy centers in eukaryotic cells and recent studies showed that the diversity of the mitochondrial genome may have a role in the adaptation to hypoxia in Tibetans[5]. Mitochondria have their own DNA of 16,569 bp encoding 13 proteins and 24 RNAs (2 ribosomal RNAs and 22 transfer RNAs) and are inherited solely through the maternal line (uniparental inheritance). Mitochondria play a regulatory role in oxygen metabolism through oxidative phosphorylation (OXPHOS). Following events may take place during hypoxic exposure; the ATP generation is down-regulated, the activities of respiratory chain complexes are reduced, and Reactive Oxygen Species (ROS) which are produced from the respiratory chain may cause cellular oxidative damage[6–8]. mtDNA mutations that affect OXPHOS could also affect metabolic rate modulation, oxygen utilization, and hypoxia adaptation[9]. Theoretically, it was accepted that migration and genetic drift play crucial role in controlling mtDNA haplotype frequencies and that mt-DNA variations in a species are selectively neutral[10]. However, recently it was reported hypothetically that mtDNA variations are result of natural selection[11,12]. The factors contributing to these variations are, (i) proteins from mtDNA interact with each other and with those imported from the cytoplasm, and consequently form four of the five complexes of the OXPHOS; and (ii) the presumption of total absence of crossing over in mtDNA, i.e., each genome has a set hierarchical history which is shared by all the genes. Even the highly mutating non-coding Control region cannot be assumed to have undergone neutral selection because of indirect genetic effects involving specific loci, affecting mitochondrial transcription and replication in significant ways[13]. Due to these reasons it was suggested that a site undergoing evolutionary pressure might have equally affected the genealogy of the whole genome[10].

Furthermore, networks provide an extremely powerful framework to understand and predict behavior of many large scale complex systems comprising of nodes and interactions (edges)[14]. For example, complex biochemical activities of a cell can be well understood by underlying protein-protein interaction networks. Network framework has been successful in revealing crucial proteins for breast cancer[15], to understand versatility of society[17], and to get insights into developmental changes in *C. elegans*[16]. Particularly motifs, which are complete sub-graphs of a network and repeat themselves, are considered to be building blocks of many complex systems[18]. Two-node motifs have been extensively studied as double negative feedback loops, double positive feedback loops, and auto-activation or repression loops[19,20].

We analysed two-node motifs of co-occurrence networks constructed for mt-DNA for varying altitudes, where nodes are variable sites and interactions are co-occurrence of these variations. Co-occurrence of variable sites have been investigated in variuos diseases, for example in understanding classification and prognosis of acute myeloid leukemia[21], for finding cause of female Duchenne muscular dystrophy[22], for finding co-occurrence of driver mutations in myeloproliferative neoplasms[23]; to understand evolution of influenza viruses[24]; in detection of pesticide resistance in *Aedes aegypti*[25] and recently in codon level analysis of human mt-DNA which revealed significance of codon-motifs in evolution of human sub-population[26].

The primary aim of this study was to investigate the evolutionary dynamics of co-occurrence motifs in human population at varying altitudes and the secondary being the characterization of these co-occurrence motifs at genetic levels. We analyzed 673 complete human mitochondrial genomes by categorizing them into different altitude ranges from the sea level to Tibetan plateau. We kept the interval of 500m between each altitude group taking into account the fact that oxygen percentage decreases by approximately 1% at every 500 meters above sea level[27]. Formal Random Forest classification based on variable sites as features demonstrates that the non-adjacent groups can be discriminated with at least 80% accuracy for the test set with 5-fold cross validation. The co-mutational cohorts were defined in terms of complete sub-graphs *aka* motifs. The motifs consist of variable sites in which minor alleles co-occurred equal to or above the set threshold of co-occurrence frequency (C_{th}). Through analysis of co-occurrence motifs, it was found that variable sites co-occur within their own gene or gene complex emphasising intra-genic constraint of genomic evolution. In addition, similarity analysis bifurcated all the altitude groups into two categories, lower and higher altitude regions. Characterization of co-occurrence motifs at genetic level revealed dominant role of cytochrome b in adaptation at Tibetan plateau.

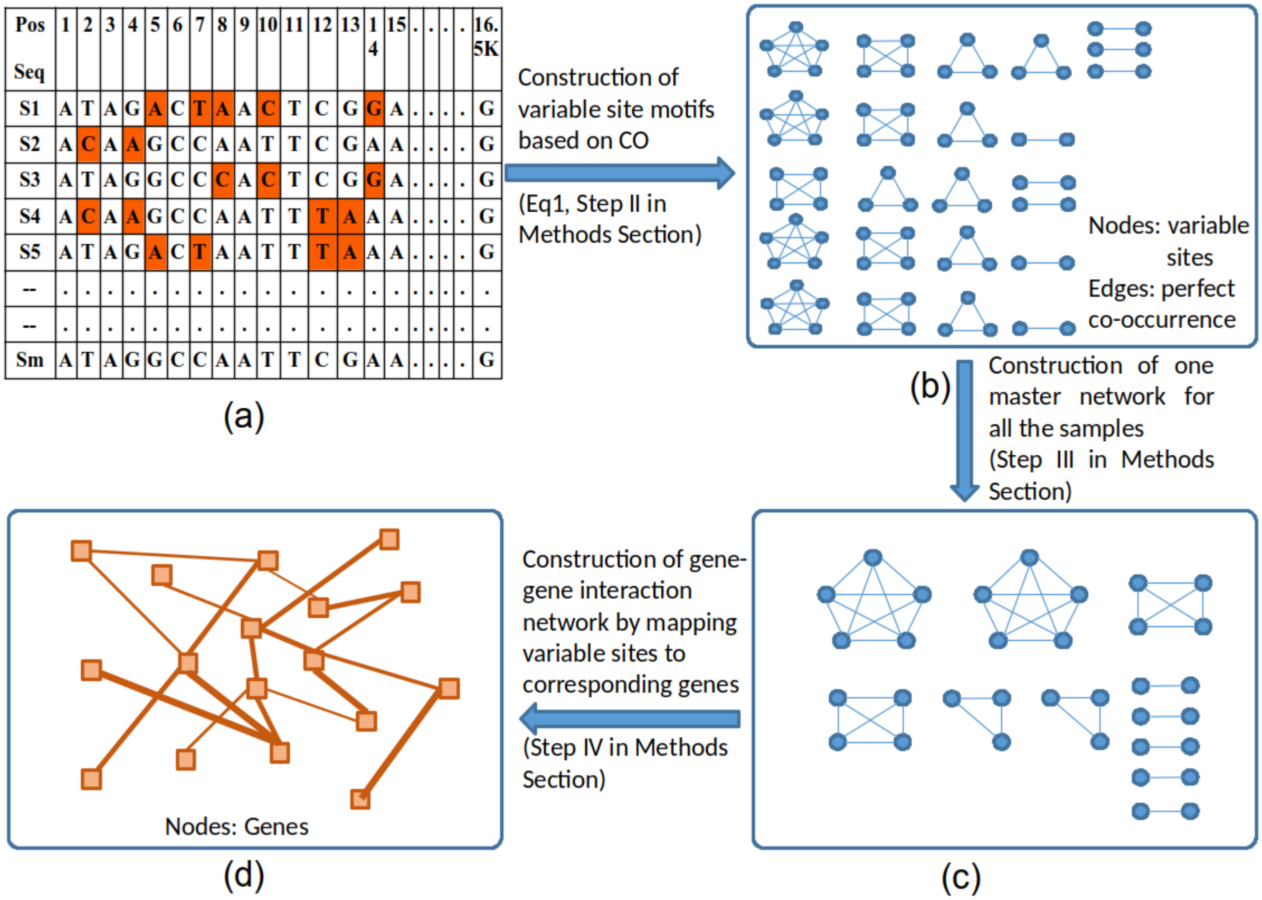
Results and Discussions

Characterization of Variable sites

A total of 3,829 variable sites exist for all the altitude groups out of which 3127 variable sites took part in motif formation. Among these sites ∼65% variable sites were located in protein-coding region (overlapped sites were double counted) with the rest lying in non-coding regions (Control region, t-RNAs and r-RNAs) (Table 1). This was expected since 11,395 (∼68%) sites out of total 16,569 sites belong to coding region. Most of the variable sites were bi-allelic, a few were having three alleles (tri-allelic) in all the groups and only group 3 had one site with four alleles (Table 1). These sites are well documented in Mitomap database for various genomic studies. Although much about the tri-allelic sites has not been understood, nonetheless their presence have been shown to be responsible for natural selection[28]. Approximately 90% of the variable sites were transitions, yielding a high transition to transversion ratio (Ts/Tv) (Table 1) which has already been reported[29] to be responsible for the conservation of structures at protein level among the individuals within a species[30]. In the context of varying altitudes, this ratio remains high which further emphasizes on the fact that even in changing environment the mt-DNA remains functionally stable. As we know that the functional stability comes from the structural integrity of proteins and the structural integrity arises due to specific interactions of amino acids[31]. These interactions of amino acids, in turn, are affected by mutations and their co-occurrence in the genome[32].

Discriminating Altitude Groups

According to the dataset, there is no unique correspondence between haplogroups and altitude groups. Therefore, we investigated the dergee of difference between altitude groups referring to the whole set of mt-DNA variable sites. This was achieved employing the formal machine learning Random Forest binary classification for each group pair, treating variable mt-DNA sites



Pos Seq	1	2	3	4	5	6	7	8	9	10	11	12	13	1 4	15	.	.	.	.	16. 5K
S1	A	T	A	G	A	C	T	A	A	C	T	C	G	G	A	.	.	.	.	G
S2	A	C	A	A	G	C	C	A	A	T	T	C	G	A	A	.	.	.	.	G
S3	A	T	A	G	G	C	C	C	A	C	T	C	G	G	A	.	.	.	.	G
S4	A	C	A	A	G	C	C	A	A	T	T	T	A	A	A	.	.	.	.	G
S5	A	T	A	G	A	C	T	A	A	T	T	T	A	A	A	.	.	.	.	G
--	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
--	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Sm	A	T	A	G	G	C	C	A	A	T	T	C	G	A	A	.	.	.	.	G

(a)
Construction of variable site motifs based on CO
(Eq1, Step II in Methods Section)
Nodes: variable sites
Edges: perfect co-occurrence
(b)
Construction of one master network for all the samples
(Step III in Methods Section)
(c)
Construction of gene-gene interaction network by mapping variable sites to corresponding genes
(Step IV in Methods Section)
Nodes: Genes
(d)

Figure 1. **(a)** Total variable sites are extracted from a particular group and co-occurrence threshold is applied (See Methods, Section 2.2). **(b)** For each sample, one set of motifs were constructed. **(c)** These motifs were then merged to construct one master network where nodes were variable sites. **(d)** This master network was then used to construct gene-gene interaction network by mapping the variable sites corresponding to each gene.

as features. The results demonstrated a good classification accuracy, over 80%, for the pairs of groups, which altitude difference was greater than $2000m$ (Table 2), thus confirming altitude association of the studied mt-DNA samples. However, the output lists of SNPs ranked by importance do not inform whether the definitive mutations are independent or co-occuring. This would be the general case for machine learning models, while the following co-occurence network analisys allows to go beyond.

Categorization of Variable sites

We categorized variable sites based on their occurrence in the network as follows; **(i)** Isolated nodes (variable sites which did not take part in network construction) and **(ii)** Connected nodes (variable sites which took part in network construction). Further, these two types of nodes were sub-categorized into **(a)** Global nodes (the nodes present in all the altitude groups), **(b)** Local Nodes (the nodes present exclusively in a particular group) and **(c)** Mixed nodes (the nodes present in more than one altitude groups but not in all). This categorization helped us to decipher the role of variable sites in terms of co-occurrence motifs. It is also observed that the percentage of local connected nodes is decreasing when going from low to high altitude while percentage of local isolated nodes is increasing when going from low to high altitude. Further, number of links (N_c) provide the information about the co-occurrence pairs formed by connected nodes having perfect co-occurrence frequency along with the average degree ($< k >$) (average number of connections of each node) of the network. Although, average degree was found to be nearly same in all the networks, it tells about the degree of sparsity of the network. The same are summarized in Table 3.

Table 1. Statistics of variable sites

Altitude group	Variable sites	Coding region sites	Non-coding region sites	Ts:Tv	No. of Tri-allelic sites
Group 1	474	286 (60.33%)	188 (39.6%)	16.5:1	5
Group 2	435	270 (62.06%)	165 (37.94%)	14.6:1	6
Group 3	644	423 (65.68%)	221 (34.32%)	14.2:1	7
Group 4	694	432 (62.24%)	262 (37.76%)	12.5:1	11
Group 5	429	274 (63.86%)	155 (36.14%)	17.8:1	4
Group 6	354	222 (62.71%)	132 (37.29%)	22.7:1	1
Group 7	357	212 (59.38%)	145 (40.62%)	17.0:1	1
Group 8	442	279 (63.12%)	163 (26.88%)	13.3:1	1

Table 2. Accuracy of binary classification between altitude groups, the cases of good discrimination, over 80%, are highlighted by green %

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Group 1		58.63	66.84	57.00	82.80	89.28	96.20	88.76
Group 2	58.63		77.23	66.27	82.29	92.60	97.60	93.13
Group 3	66.84	77.23		50.94	76.07	90.15	93.66	82.40
Group 4	57.00	66.27	50.94		79.99	89.22	94.11	89.14
Group 5	82.80	82.29	76.07	79.99		68.89	66.66	46.67
Group 6	89.28	92.60	90.15	89.22	68.89		59.78	73.58
Group 7	96.20	97.60	93.66	94.11	66.66	59.78		35.77
Group 8	88.76	93.13	82.40	89.14	46.67	73.58	35.77	

Isolated Nodes: A total of 702 isolated nodes were found in all the groups with an average of 88 nodes per altitude group ($\sim$25%) since we considered ($C_{th} = 1$). The isolated nodes represent molecular sites which undergo independent variation in the context of motif formation as these variable sites are not associated with any other variations because of the lack of perfect co-occurrence ($C_{th} < 1$). To gain further insights into these nodes, we mapped them with their corresponding genes and found that a significantly high number ($\sim$50%) of these nodes belonged to Control region (CR) in each altitude group (Fig 2). This significantly high contribution of CR in the isolated nodes suggests that the variants in this region are independent of the variants in other regions. It is already known that CR undergoes a higher rate of mutations as compared to the other mitochondrial genes[33]. These isolated nodes also contain the variable sites which show a high minor allele frequency (MAF) of nearly 40%. The presence of variable sites with such a high MAF in the isolated nodes and their correspondence to CR again suggest that this region is undergoing a high rate of mutation and the major allele is not fixed in the population. Further,

Table 3. Categorization of variable sites

Altitude	Connected Nodes			Isolated Nodes			N_c	<k>
	Local (%)	Mixed (%)	Global (%)	Local (%)	Mixed (%)	Global (%)		
Group 1	31.5	67.5	1.0	28.4	16.2	55.4	1855	9
Group 2	29.4	69.5	1.11	29.7	14.9	55.4	1324	7
Group 3	36.9	62.3	0.77	32.3	34.7	33.0	1658	6
Group 4	38.2	61.1	0.72	37.5	32.3	30.2	2219	8
Group 5	23.3	75.6	1.15	32.1	17.3	50.6	1443	8
Group 6	20.4	77.9	1.34	25.5	0.0	74.5	1182	8
Group 7	27.1	71.5	1.44	37.5	11.3	51.2	950	7
Group 8	25.9	73.1	1.10	35.9	11.5	52.6	1409	8

all the protein-coding genes and RNA genes contributed less than 10% to the isolated nodes at all the altitudes. Among the protein-coding genes, ND5 showed the highest percentage followed by ND2, CYB, ND4, CO1 and ATP6 while ATP8 showed the lowest contribution in the isolated nodes (Fig 2).

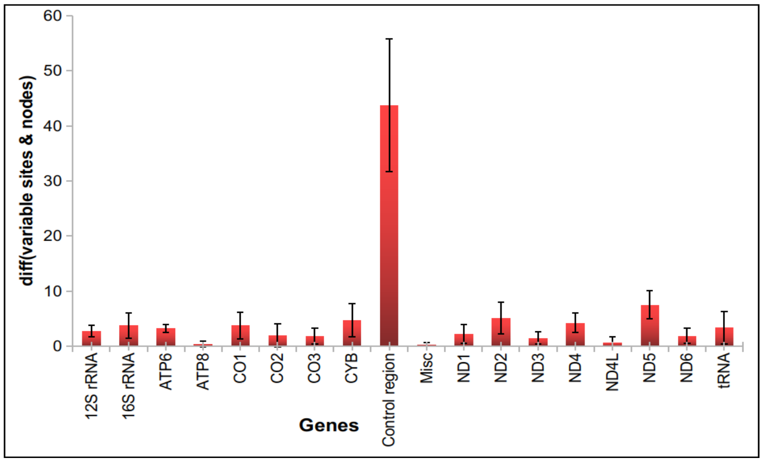


Figure 2. The mean of number of isolated nodes for corresponding gene across all the altitudes.

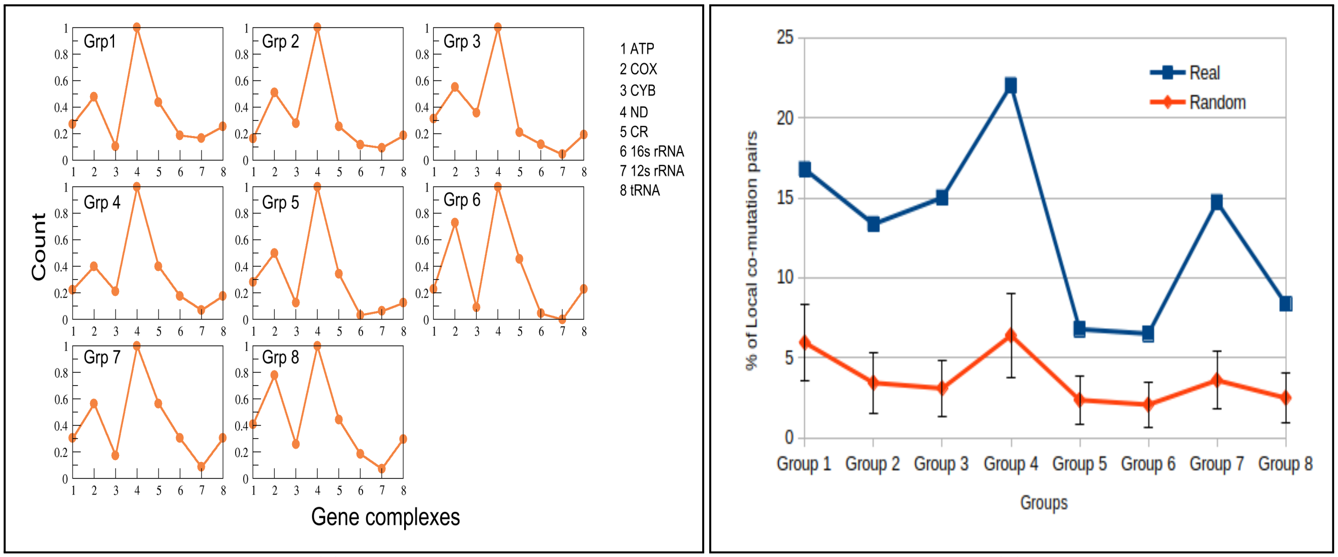
Global Nodes: 41 variable sites in the isolated nodes and only 4 variable sites in the connected nodes were found to be common in all the altitude groups, which were referred as 'global isolated nodes' and 'global connected nodes', respectively. It is quite apparent that the global connected nodes were found to be 10 fold less than the global isolated nodes. This suggests that the variations which are occurring globally are not necessarily co-occurring perfectly with the other variations and their co-occurrence may be population specific or more precisely in our case, altitude specific. The four global connected nodes are 9540, 10400, 10873 and 16327. Out of these four nodes, three, 9540, 10400 and 10873 were found to form a motif (perfectly co-occurred) in all the altitude groups except the group 3 where only 9540 and 10873 formed a motif whereas in group 4, despite having maximum number of nodes, surprisingly not a single node co-occur with these three global connected nodes. The 16327 node formed star like network with other nodes. Further, 9540, 10400, 10873 and 16327 belong to CO3, ND3, ND4 and CR genes, respectively. The genes corresponding to nodes which were co-occurring with global nodes of the coding region (9540, 10400 and 10873) are summarized in Table 4. From the table it is quite evident that these three global nodes prefer to co-occurred with the nodes belonging to the coding regions. The marker 10398 is an ancestral one belonging to L3 lineage, universally present in M haplogroup and found to co-occur with global nodes present only in higher altitude groups 6, 7 and 8.

Local Nodes: The variable sites which co-occurred exclusive in a particular group were considered as local nodes. In higher altitude groups, the percentage of local nodes was observed to be decreased significantly as compared to that of lower altitude groups (p-value < 0.05). To find the difference at the gene level between the local nodes of each pair of altitudes, we mapped local nodes to genes (Fig 3). Mapping with genes revealed a very peculiar property of mitochondrial genome. Although the local nodes are exclusively mutated sites for a particular altitude which are not found in any other altitude regions, however at the genetic level the variations remained similar for all the altitude groups. Among all the gene complexes, ND genes showed the highest count which is also expected since ND complex comprises seven sub-units. Apart from that we looked for

Table 4. Genes corresponding to variable sites co-mutating with global connected nodes of coding region

Altitude group	Genes
Group 1	ATP6, CYB, CYB
Group 2	ATP6, CYB, CYB, CYB
Group 3	ATP6, CYB, CYB
Group 4	-
Group 5	CR, ATP6, CYB, CYB, CYB
Group 6	ATP6, ND3, CYB, CYB, CYB
Group 7	ATP6, ND3, CYB, CYB, CYB
Group 8	CR, ATP6, ND3, CYB, CYB, CYB

co-occurrence pairs formed by these local nodes only and found that the lower altitude groups (group 1 to group 4) exhibited very high percent of such pairs as compared to the higher altitude groups (group 5 to group 8) and corresponding random networks as well.



. **(a)** **(b)**

Figure 3. (a)The local nodes were mapped to genes and gene counts were plotted. Each gene was showing similar count in each altitude. (b) The percentage of co-occurrence pairs consisting only local nodes.

High altitude markers: The variable sites 3010, 3394 and 7697 were reported as high altitude markers and have been associated with high altitude adaptation in Tibetan population. The markers 3010, 3394 and 7697 belong to 16s rRNA, ND1 and CO1 genes, respectively. Further, we wanted to investigate the motifs derived by these markers in order to determine their co-occurrence background. The marker 3010 was found only in groups 1, 6, 7 and 8. In groups 6, 7 and 8, it formed the motif with 14668 (ND6) and 8414 (ATP8) whereas in group 1 it formed a separate cohort with 295 (Control region), 462 (Control region), 8269 (CO2), 12612 (ND5) and 16069 (Control region). The marker 3394 was found in group 1 to form motifs with 11335 (ND4), in group 4 to form motif with 4832 (ND2) and 16258 (CR) and in groups 5 and 8 to form motif with 1041 (12s rRNA), 4491 (ND2) and 14308 (ND6) in both the groups. The marker 7697 was found only in higher altitude groups 5, 6, 7 and 8. It formed motifs with 711 (12s rRNA), 7142 (CO1), 9242 (CO3) in group 5, 453(CR), 711 (12s rRNA), 7142 (CO1), 9242 (CO3) and 14417 (ND6) in group 6, 711 (12s rRNA), 7142 (CO1), 9242 (CO3) and 14417 (ND6) in groups 7 & 8. Further, these high altitude markers tend to co-occur with their own gene complexes which emphasizes the intragenic influence on co-occurrence of nucleotides.

Grouping of altitudes based on co-occurrence motifs

Jaccard similarity coefficient was used to find out the similarity between each altitude group using mixed nodes (Fig 5). This similarity coefficient led to the formation of two major clads, one with groups 1, 2, 3 and 4 (lower altitude cohort) and the other with groups 5, 6, 7 and 8 (higher altitude cohort). The nodes of the dendrogram represent altitude groups. Two altitude groups were found to form doubletons within each clad. Moreover, it is deduced from the dendrogram that human population split up into two sub-populations giving rise to lower altitude cohort and higher altitude cohort. The lower altitude cohort furter segregated into two sub-groups forming one clad with Grp 2 and Grp 3 and another clad with Grp 1 and Grp 4. The higher altitude cohort furter segregated into two sub-groups forming one clad with Grp 5 and Grp 8 and another clad with Grp 6 and Grp 7. It is noteworthy that in lower altitude groups, Grp 1 and Grp 4 descended from common sub-group. Similarly, in higher altitude groups, Grp 5 and Grp 8 descended from common sub-group instead of having geographical distances between these altitude ranges. This segregated migration of human population suggests that, humans may have migrated through discrete path-ways for searching a better environment for establishment.

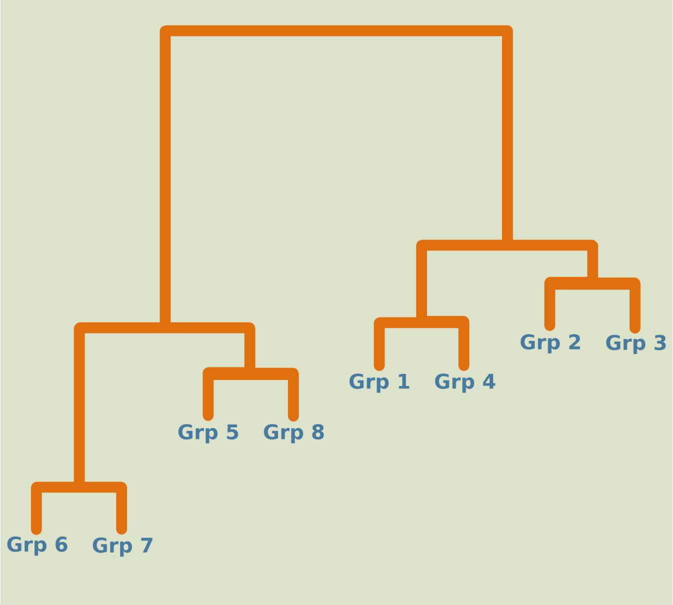


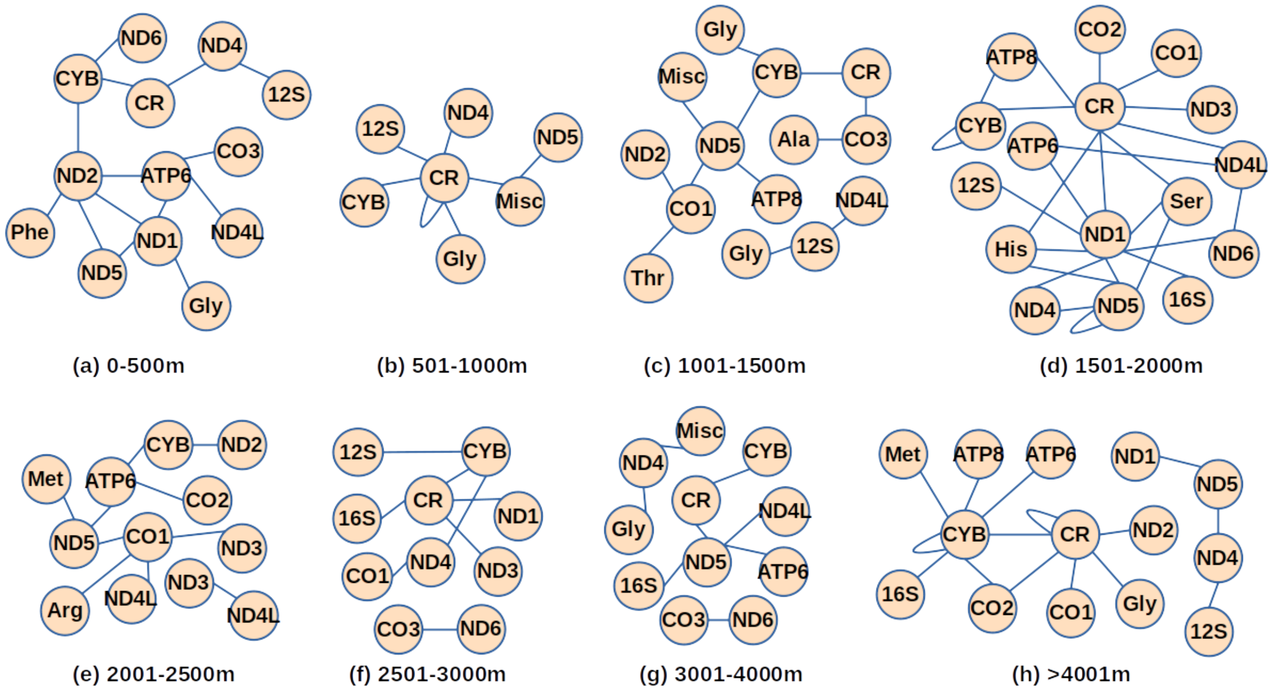
Figure 4. Cluster dendrogram was produced using common nodes between each altitude. It is clearly observed that two clusters are formed, one with groups 1 to 4 (lowest to middle) and other with groups 5 to 8 (middle to highest).

Gene-gene interactions

Gene-gene interaction networks were compared with the corresponding random networks. We found that only 72 (∼10%) of all possible gene-pairs were significantly deviated (falling out of the standard deviation range) from random networks (Fig 5). Moreover, out of total 73 gene-pairs, 46 were found to be present in any one of the altitude groups, 23 were found to be present in any two of the altitude groups, and 1 pair was found to be present in any 3, any 4 and any 7 altitude groups. There were certain pairs in which one of the gene was tRNA such as in group 1 tRNA-Gly. Gene-pair CYB-CR was found to be present in all the groups except in 5th group where these genes are present but interacting with other genes. Further, this pair has more weight than the corresponding random network in group 6 while less in other groups. Interestingly, there were only two genes which formed pairs with themselves, these genes are CYB and CR. The gene-pair CYB-CYB was found to be present in groups 4 and 8. In group 4 its weight was found to be less than the corresponding random network while in group 8 its weight was found to be more than the corresponding random network. The other gene-pair CR-CR was also found in group 2 and group 8. Moreover, in group 8, in seven out of fifteen gene-pairs, one of the gene is CYB. The gene-pair CO3-ND6 was found to be present in groups 6 and 7. Apart from CR gene and coding genes, despite having small length and less variable sites, certain tRNA genes were also found to form gene interaction pairs (Table 5). Particularly tRNA-Gly was present in groups 1, 2, 3 and 8. Interestingly, tRNA-Thr exhibited highest number of variable sites among tRNA genes, but it was found to be present only in group 3 co-occurring with CO1 gene. tRNA-Met was present only in groups 5 and 8. Overall, we found different gene-gene interactions at varying altitudes which can be further analyzed for adaptation or disease association.

Discussions and Conclusion

We investigated altitude driven co-occurrence of variations in Tibetan and lower altitude population using network motifs. Phylogenetic techniques are the traditional methods for studying DNA sequence evolution, particularly adaptive and migratory



(a) 0-500m

(b) 501-1000m

(c) 1001-1500m

(d) 1501-2000m

(e) 2001-2500m

(f) 2501-3000m

(g) 3001-4000m

(h) >4001m

Figure 5. Gene-gene interaction pairs which showed deviation from random network

Table 5. Statistics of gene-pairs showing deviation from random networks

Altitude group	No. of Gene-pairs	No.of genes	tRNA genes
Group 1	14	13	Gly, Phe
Group 2	7	7	Gly
Group 3	12	15	Ala, Asp, Gly, Thr
Group 4	30	16	His, Ser
Group 5	10	15	Arg, Met
Group 6	9	10	—
Group 7	9	13	Glu
Group 8	15	14	Gly, Met

evolution of populations in context. As such, machine learning approaches allow for building formal rules to discriminate different altitude groups with good accuracy, but fail to distinguish between independent and co-mutations. Here, we have used a network model to represent the whole mitochondrial genome as a complex genomic interaction network based on the co-occurring nucleotide pairs over the entire genome. Even though we took a perfect co-occurrence frequency, nearly 75% nodes (variable sites) took part in network construction which shows that mutual variations are abundant in human mitochondrial genome. This abundance of mutual variations suggest that these variations richly co-depend on each other. Further, apart from connected nodes, the isolated nodes showed high MAF which suggests that mutation with high frequency do not form co-occurrence pairs. Another category of nodes, the local nodes, constituted nearly 30% of total nodes for a particular group. The presence of local nodes provides evidence that human population has specific signature at nucleotide variation level for varying altitudes. This signature seems to disappear when we mapped these local nodes to their corresponding genes and counted the number of local nodes for each gene complex. This loss of signature reveals a peculiar property of mitochondrial genome that even with exclusive variations, the genomic functionality remains undisturbed, keeping fundamental molecular functions intact even in changing environments. This does not mean that we are suggesting that environment is not affecting the adaptive function of mitochondrial genome. Further, co-occurrence analysis of high-altitude markers revealed presence of intra-genic constrain in high altitude population. this suggests that the presence of particular variation is not sufficient for adaptation but that variation has to be assisted by other variations in the same gene or same gene complex. Variable sites of

12srRNA gene were found to co-occur with both the markers 3394 and 7697. 12srRNA gene encodes for a protein responsible for regulating insulin sensitivity and metabolic homeostasis.

Particularly, the co-occurrence of 7697 with 14417 is observed in the group 7 and 8. Since, the biological effects of high altitude are observed at >3000 m this co-occurrence pair seems a possible combination of variants that affects mitochondrial bioenergetics.

4491 of ND2 gene is found to be associated with High Altitude pulmonary edema (HAPE) susceptibles in low altitude population[34] whereas in our analysis, this 4491 was found to co-occur with 3394 and its presence only in higher altitude regions suggests its adaptative dependence. Variable sites forming global connected nodes 9540 and 10873 are the ancestral markers while 10400 and 16327 are markers of M sub-haplogroups[35]. The markers 9540 and 10873 are found to be present throughout the human population, 10400 is known to be specific to Asian population and 16327 is known to be the marker of C sub-haplogroup of M haplogroup which is specific to Siberian and American regions. Phylogeny based study has shown that 10398 resulted through selective sweep at colder geographical regions[36] which is supposed to lower the oxidative phosphorylation coupling leading to release of more heat. This possible tradeoff between ATP generation and thermogenesis seems to play key role in adaptation at colder higher altitudes. Substitution from A to G at 10398 corresponds to substitution of an alanine amino acid residue by a threonine at the carboxyl end of ND3 gene, a subunit of NADH-ubiquinone oxidoreductase (complex I). The presence of 10398 only at higher altitudes suggests that a separate ancestral population might have colonized these regions. Through gene-gene interaction network, we found that CYB gene was co-occurring significantly with other genes to co-evolve at Tibetan region shed light on its involvement in hypoxia and low temperature adaptation. Cytochrome b protein is an integral membrane protein subunit of the cytochrome bc1 complex encoded by mitochondrial CYB gene, this complex catalyzes the redox transfer of electrons from ubiquinone to cytochrome c in the mitochondrial electron transport chain. As the efficiency of the electron transport chain governs key aspects of aerobic energy metabolism[37], several investigators have suggested that functional modifications of redox proteins, such as cytochrome b, may be involved in physiological adaptation to different thermal environments[11, 12, 38]. The formation of local co-occurrence pairs and similarity clustering divides the altitude groups into higher and lower altitude regions. This division might be possible due two reasons, (i) due to migration and demographic dynamics or (ii) due to process of selection on mitochondrial variants that in combination optimize mitochondrial bioenergetics in extreme conditions, experienced by these populations that lived at high altitude. Thus, motifs identified at high altitude >3000m in group 7 and 8 can be suggested as candidate positions for a biological role in adaptation to these conditions. Overall co-occurrence network motifs provides detailed insight into finding the association of variable sites which are overlooked by haplogroup analysis alone.

Methods

Sequence Acquisition

We retrieved a total of 673 complete human mitochondrial sequences of healthy beings from GenBank (http://www.ncbi.nih.gov/), where metadata was available. These sequences were aligned using multiple sequence alignment tool, Clustal Omega[39] using default parameters. After alignment, all the sequences were mapped to master sequence, revised Cambridge Reference Sequence (rCRS). Then these sequences were grouped into eight altitude groups ranging from 0m to >4000m with an interval of 500m (Table 6).

Machine Learning Classification of Altitude Groups

We treated variable mt-DNA sites as seed features for the Random Forest algorithm to build and investigate binary classification models between each pair of altitude groups. The particular implementation was taken from Python package 'Scikit-learn' version 0.22; Python version 3.7.5[40]. To perform classification between each pair of altitude groups we built a Random Forest model with the number of decision trees equal 500. 5-fold cross-validation was applied to test the effectiveness of a built model, yielding the average accuracy over cross-validated models.

Construction of Mitochondrial Motif and Gene-Gene Interaction Networks

Step I: Any position within the samples having more than one allele is considered as a variable site. These variable sites were extracted from the aligned sequences for each group separately to construct co-occurrence motifs. For genomic equality, ambiguous nucleotides such as X, M, Y, etc were replaced with 'N' for all the sequences (Fig 1a).

Step II: For the construction of motifs, a node was represented by the position of a variable site and the edge was represented by co-occurrence frequency between any two nodes. We defined co-occurrence frequency for a pair of variable sites, $C_{(i,j)}$ as,

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

where $N(x_i y_j)$ is number of times x and y nucleotides co-occur at i^{th} and j^{th} positions, respectively, $N(x_i)$ is total number of times nucleotide x occurs at i^{th} position and $N(y_j)$ is total number of times nucleotide y occurs at j^{th} positions.
We consider variable sites as nodes and two nodes are connected if their co-occurrence frequency is equal to or greater than a given threshold value (C_{th}). Putting the $C_{th} = 0$, gives a globally connected network spanning all the nodes. Whereas, for $C_{th} = 1$, those pairs of nodes were connected which were perfectly co-occurring. In this way, a co-occurrence motifs network for each sample (Fig 1b).

Step III (Co-occurrence Network): The sugraphs constructed in Step II for each sample were merged to get a single master network for each altitude. This master motif network consists of all motifs of that particular group (Fig 1c). In this way we constructed eight master networks, where nodes were variable sites and edges were co-occurrence frequency (Fig 6).

Step IV (Gene-gene interaction Network): For each master network, nodes were mapped to their corresponding genes to achieve one gene-gene interaction network for each altitude (Fig 1d). Since, two or more motifs may belong to same gene or gene pair, each link is counted as many times it is found and this number was considered as weight of that gene-pair. For example, two-order motifs (3461-8715) and (4133-9157) yield two ND1-ATP6 gene-pairs, so this pair was counted twice and so on. Finally, we have got two types of networks, first were master networks where nodes are variable sites and second were gene-gene interaction networks where nodes were genes.

Table 6. Altitude and sample distribution.

Altitude group	Altitude range (m)	Sample size
Group 1	0-500	46
Group 2	501-1000	61
Group 3	1001-1500	95
Group 4	1501-2000	103
Group 5	2001-2500	89
Group 6	2501-3000	86
Group 7*	3001-4000	107
Group 8	4001-above	86

*Due to unavailability of data, Group 7 ranges from 3000m to 4000m.

Null models

Random networks were generated for each sample of each altitude group. We took same number of nodes as in real nework and a connection probability (p) based on number of connections present in real network. Here we defined p as,

$$p = \frac{N_c}{N(N-1)/2} \tag{2}$$

where, N_c is total number of connections and N is total number of nodes in real network. We compared the network properties of null networks with real networks.

Similarity Clustering

We wanted to look for similarity between altitude groups. For this, we calculated Jaccard similarity coefficient using common nodes between each altitude group.

$$J_{(i,j)} = \frac{|n_i \cap n_j|}{|n_i \cup n_j|} \tag{3}$$

where, $J_{(i,j)}$ is Jaccard similarity coefficient between two altitude groups i and j, and n_i and n_j are nodes of two altitude groups.

Acknowledgments

SJ acknowledges support from SERB grant (EMR/2016/001921), the council of scientific and industrial research grant (CSIR, 25(0293)/18/EMR-II) Govt of India. SJ, MI and AK acknowledges Ministry of education and science of the Russian Federation megagrant (075-15-2019-871). ADK acknowledges the council of scientific and industrial research grant (CSIR, 25(0293)/18/EMR-II) Govt of India for RA fellowship. RKV gratefully acknowledges University Grant Commission for SRF fellowship (305089).

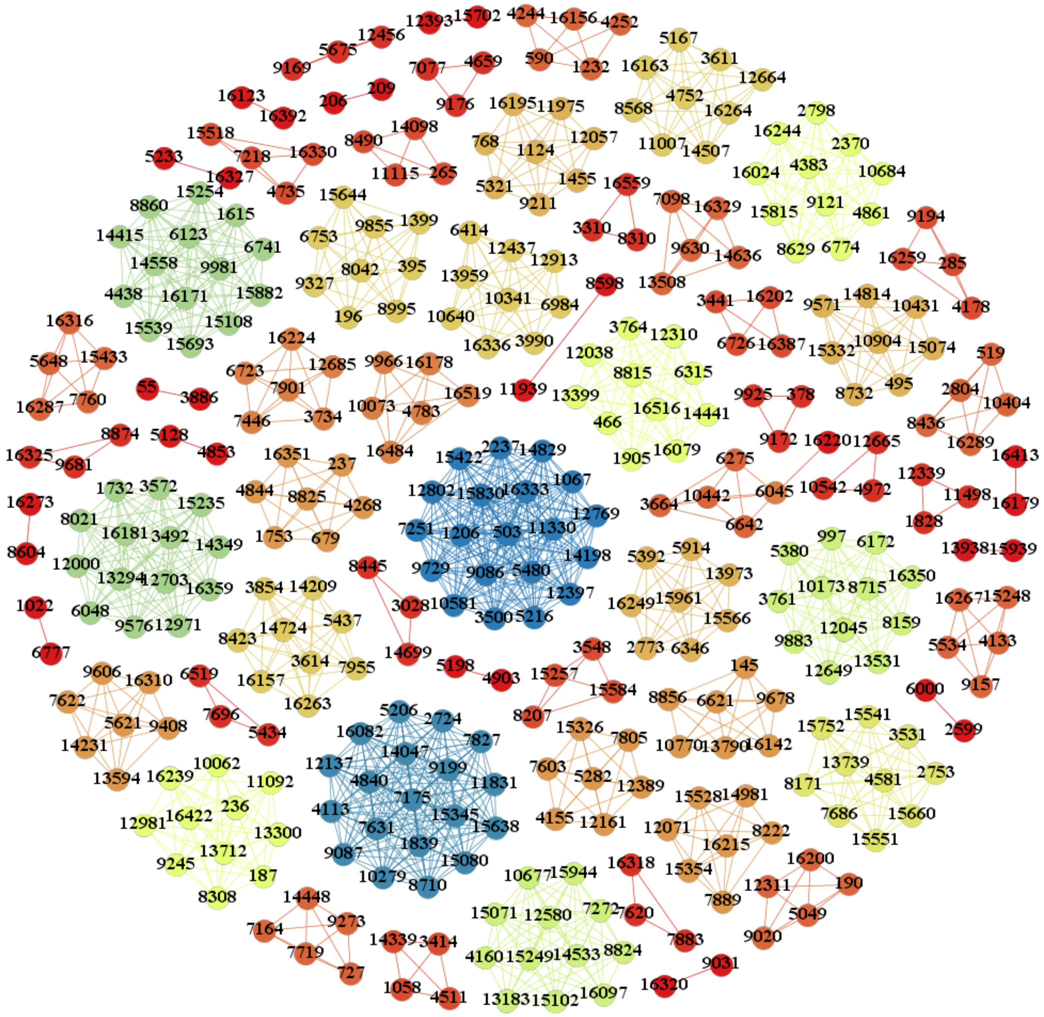


Figure 6. Co-mutation network for Tibetan population ($> 4001m$) with co-occurrence frequency equal to 1. The colour is representing the degree of nodes. Blue for higher degree while green for intermediate degree and red for lower degree nodes.

References

1. Witas H. W., Zawicki P. (2004). Mitochondrial DNA and human evolution: a review. Anthropol. Rev. 67 97–110

2. Zhao M, Kong Q-P, Wang H-W, et al. Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(50):21230-21235.

3. Dahlback A, Gelsor N, Stamnes JJ, Gjessing Y (2007) UV measurements in the 3000– 5000 m altitude region in Tibet. J Geophys Res Atmos 112:D09308.

4. Gnecchi-Ruscone GA, Abondio P, De Fanti S, et al. Evidence of Polygenic Adaptation to High Altitude from Tibetan and Sherpa Genomes. Genome Biol Evol. 2018;10(11):2919–2930. Published 2018 Nov 1.

5. Li, Q., Lin, K., Sun, H. et al. Mitochondrial haplogroup M9a1a1c1b is associated with hypoxic adaptation in the Tibetans. J Hum Genet 61, 1021–1026 (2016)

6. Magalhaes J, Ascensão A, Soares JMC, Ferreira R, Neuparth MJ, Marques F, Duarte JA. 2005. Acute and severe hypobaric hypoxia increases oxidative stress and impairs mitochondrial function in mouse skeletal muscle. J. Appl. Physiol. 99, 1247–125310.1152/japplphysiol.01324.2004

7. Fukuda R, Zhang HF, Kim JW, Shimoda L, Dang CV, Semenza GL. 2007. HIF-1 regulates cytochrome oxidase subunits to optimize efficiency of respiration in hypoxic cells. Cell 129, 111–12210.1016/j.cell.2007.01.047

8. Solaini G, Harris DA. 2005. Biochemical dysfunction in heart mitochondria exposed to ischaemia and reperfusion. Biochem. J. 390, 377–39410.1042/BJ20042006

9. Monge, C. and Leon-Velarde, F. Physiological adaptation to high altitude: oxygen transport in mammals and birds.

10. Ballard JWO, Rand DM. The population biology of mitochondrial DNA and its phylogenetic implications. Annual Review of Ecology Evolution and Systematics. 2005;36:621–642.

11. Mishmar, Dan et al. "Natural selection shaped regional mtDNA variation in humans." Proceedings of the National Academy of Sciences of the United States of America vol. 100,1 (2002): 171-6.

12. Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC, Effects of purifying and adaptive selection on regional variation in human mtDNA, Science. 2004 Jan 9; 303(5655):223-6.

13. Alzheimer's brains harbor somatic mtDNA control-region mutations that suppress mitochondrial transcription and replication. Coskun PE, Beal MF, Wallace DC Proc Natl Acad Sci U S A. 2004 Jul 20; 101(29):10726-31.

14. Sarika Jalan and Camellia Sarkar. Complex Networks: an emerging branch of science. Physics News 47, 3-4 (2017)

15. Rai, A., Menon, A. & Jalan, S. Randomness and preserved patterns in cancer network. Sci Rep 4, 6368 (2015)

16. Pramod Shinde and Sarika Jalan*. A multilayer protein-protein interaction network analysis of different life stages in Caenorhabditis elegans, EPL, 112, 58001 (2015)

17. Jalan S, Sarkar C, Madhusudanan A, Dwivedi SK (2014) Uncovering Randomness and Success in Society. PLOS ONE 9(2): e88249.

18. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U Network Motifs: Simple Building Blocks of Complex Networks Science. 2002 Oct 25; 298(5594):824-7.

19. Gardner TS, Cantor CR, Collins JJ Construction of a genetic toggle switch in Escherichia coli. Nature. 2000 Jan 20; 403(6767):339-42.

20. Kim JR, Yoon Y, Cho KH Coupled feedback loops form dynamic motifs of cellular networks. Biophys J. 2008 Jan 15; 94(2):359-65.

21. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. N Engl J Med. 2016;374(23):2209-2221.

22. Katayama, Y., Tran, V.K., Hoan, N.T. et al. Co-occurrence of mutations in both dystrophin- and androgen-receptor genes is a novel cause of female Duchenne muscular dystrophy. Hum Genet (2006) 119: 516.

23. Boddu, P., Chihara, D., Masarova, L. et al. The co-occurrence of driver mutations in chronic myeloproliferative neoplasms. Ann Hematol (2018) 97: 2071.

24. Du X, Wang Z, Wu A, et al. Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. Genome Res. 2008;18(1):178-87.

25. Kawada H, Oo SZ, Thaung S, et al. Co-occurrence of point mutations in the voltage-gated sodium channel of pyrethroid-resistant Aedes aegypti populations in Myanmar. PLoS Negl Trop Dis. 2014;8(7):e3032. Published 2014 Jul 31.

26. Shinde, P., Sarkar, C. & Jalan, S. Codon based co-occurrence network motifs in human mitochondria. Sci Rep 8, 3060 (2018)

27. Physiological Reviews. 1991, 71, 4, 1135-1172, doi: 10.1152/physrev.1991.71.4.1135 Peacock AJ. ABC of oxygen: oxygen at high altitude. BMJ. 1998;317(7165):1063–1066.

28. Cao, M. , Shi, J. , Wang, J. , Hong, J. , Cui, B. and Ning, G. (2015), Analysis of Human Triallelic SNPs by Next-Generation Sequencing. Annals of Human Genetics, 79: 275-281.

29. Pereira L, Freitas F, Fernandes V, et al. The diversity present in 5140 human mitochondrial genomes. Am J Hum Genet. 2009;84(5):628–640.

30. Guo C, McDowell IC, Nodzenski M, et al. Transversions have larger regulatory effects than transitions. BMC Genomics. 2017;18(1):394. Published 2017 May 19.

31. Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th edition. New York: W H Freeman; 2002. Chapter 3, Protein Structure and Function.

32. Harry C. Jubb, Arun P. Pandurangan, Meghan A. Turner, Bernardo Ochoa-Montaño, Tom L. Blundell, David B. Ascher, Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health, Progress in Biophysics and Molecular Biology, Volume 128, 2017, Pages 3-13.

33. Stoneking M, Hedgecock D, Higuchi RG, Vigilant L, Erlich HA. Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. American Journal of Human Genetics. 1991;48(2):370-382.

34. Sharma S, Singh S, Gupta RK, et al. Mitochondrial DNA sequencing reveals association of variants and haplogroup M33a2'3 with High altitude pulmonary edema susceptibility in Indian male lowlanders. Sci Rep. 2019;9(1):10975. Published 2019 Jul 29.

35. MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org, 2019

36. Balloux François, Handley Lori-Jayne Lawson, Jombart Thibaut, Liu Hua and Manica Andrea Climate shaped the worldwide distribution of human mitochondrial DNA sequence variation276Proc. R. Soc. B

37. Rolfe DF, Brown GC. Cellular energy utilization and molecular origin of standard metabolic rate in mammals. Physiol Rev. 1997;77:731–758.

38. Fontanillas P, Dépraz A, Giorgi MS, Perrin N, Nonshivering thermogenesis capacity associated to mitochondrial DNA haplotypes and gender in the greater white-toothed shrew, Crocidura russula. Mol Ecol. 2005 Feb; 14(2):661-70.

39. Szymon Chojnacki, Andrew Cowley, Joon Lee, Anna Foix, Rodrigo Lopez; Programmatic access to bioinformatics tools from EMBL-EBI update: 2017, Nucleic Acids Research, Volume 45, Issue W1, 3 July 2017, Pages W550–W553

40. Pedregosa et al.; Scikit-learn: Machine Learning in Python, JMLR 12, 2011, pp. 2825-2830.