Network analysis of mitochondrial genome

Ph.D. Thesis

By

Rahul Kumar Verma

Department of Biosciences and Biomedical Engineering
INDIAN INSTITUTE OF TECHNOLOGY, INDORE, INDIA
July 2022

Network analysis of mitochondrial genome

A THESIS

Submitted in partial fulfilment of the requirements
for the award of the degree
of

DOCTOR OF PHILOSOPHY

by

Rahul Kumar Verma

Department of Biosciences and Biomedical Engineering
INDIAN INSTITUTE OF TECHNOLOGY, INDORE, INDIA
July 2022

INDIAN INSTITUTE OF TECHNOLOGY INDORE

I hereby certify that the work which is being presented in the thesis entitled "**Network analysis of mitochondrial genome**" in the partial fulfillment of the requirements for the award of the degree of DOCTOR OF PHILOSOPHY and submitted in the DEPARTMENT OF BIOSCIENCES AND BIOMEDICAL ENGINEERING, Indian Institute of Technology Indore, is an authentic record of my own work carried out during the time period from June 2017 to July 2022 under the supervision of Dr. Sarika Jalan, Professor, IIT Indore.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

23-01-2023
Signature of the student with date

(**Rahul Kumar Verma**)

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

23-01-2023
Signature of Thesis Supervisor with date

(**Prof. Sarika Jalan**)

Rahul Kumar Verma has successfully given his/her Ph.D. Oral Examination held on 23-01-2023.

23-01-2023
Signature of Thesis Supervisor with date

(**Prof. Sarika Jalan**)

Acknowledgements

With the grace of Lord Shiva and many thanks are due to the numerous individuals and organizations that contributed their time, knowledge, and support in completing this dissertation work. It is my immense gratitude to my supervisor Prof. Sarika Jalan, for guiding and supporting me through this wonderful journey. You have strengthen my scientific understanding of the subject by putting me with the academic challenges throughout my Ph.D. at Indian Institute of Technology (IIT) Indore. Your enthusiasm, rigorous discussions, and fighter attitude kept me sailing through the tough phases of this voyage. You have helped me become more confident as a student and as a professional, more than you know.

My sincere regards to the PSPC committee members, Dr. Debasis Nayak (former), for several vital suggestions on genomic data, Dr. Sharad Gupta for his helpful insights and Dr. Kapil Ahuja for monitoring and providing constructive suggestion to improve my methodology. I am also thankful to the Director, IIT Indore, DPGC convener, Head (Biosciences and Biomedical Engineering), all the faculty members and staff members of Discipline of Physics, IIT Indore for their timely help and support throughout. I humbly acknowledge the sincere efforts of all the staff members of Academic Office (Mr. Tapesh Parihar and Mr. Neeraj Kumar), BSBE office and Physics office members (Mr. Amit Mishra (former), Mr. Murphy Ganveer, Mr. Arif Patel, Mr. Gaurav Singh, Mr. Sunny Namdev, Mr. Nitin Upadhyay, Mr. Prashant Gupta and Mr. Vedprakash Thakur) Accounts Section, R&D Section (Mr. Rahul Geed), IT Section and Central Library, Medical Team, Hostel Team, IIT Indore whose constant support ascertained smooth functioning of my research work.

The Ph.D. was a long and bumpy road filled with countless nights stay awake, and I would have never found my way out without the beautiful community that resides over the globe. Several personal and email communications which make this dissertation more useful and accurate. First and foremost I would like to thank Dr M. V. Ivanchenko (Lobachevsky State University of Nizhny), Dr. Cristina Giuliani (University of Bologna). I am thankful to all the journal reviewers to critically re-

view the manuscript and provide constructive suggestions to make the content of the thesis more correct.

I am feeling very much blessed and from the bottom of my heart I would like to say thanks to all my present and past research colleague at Complex Systems Lab Dr. Camellia Sarkar, Dr. Aparna Rai, Dr. Ajaydeep, Dr. Priodyuti Pradhan Dr. Alok, Dr. Pramod, Dr. Saptarshi, Dr. Anil and Ankit, current lab members Vasundhara, Tanu, Sanket, Ayushi, Jerry for their love, care and support throughout, making my time in IIT Indore so memorable. My best wishes are always with them. Also many thanks and best wishes to inter members Naveen, Pramodini, Neeraj, Ayesha, Bibhabasu, Rishabh, Sagar, Angeliya, Siddhant, Saket, Drashti, Shraddha, Namrata, Neeraj, Nikhil and Prashant for being a joyful companion of this beautiful journey. Along with my lab members I am greatly thankful to my batch-mates with whom I spent a joyful time at IIT Indore.

I apologize if I forgot to mention any of the names.

Words fall short for expressing my gratitude towards the most important people of my life, my family! Finally, I thank Lord Shiva for sending all these wonderful people to my life. Thank you!!

Rahul Kumar Verma

This thesis is

Dedicated

to

All my Teachers

Synopsis

Introduction

The mitochondrion plays a vital role in the life of unicellular to multicellular organisms by providing energy and heat as it is a site of various metabolic reactions. Mitochondria also have a unique feature which is the presence of its own genome. This unique feature is further amplified by their maternal inheritance only in higher organisms. These features make mitochondria a super-cell organelle. It is the main site of energy production, it has its own circular DNA molecules, and it is inherited in the next generation through females only (uniparental inheritance). The mitochondrial genome is a circular DNA molecule with 16,569 base pairs and harbors thirteen protein-coding genes, two rRNA genes, 22 tRNA genes, and one D-loop or Control region. The control region regulates the replication and transcription of the genome. The thirteen protein-coding genes express the proteins which are involved in oxidative phosphorylation, the energy-producing set of reactions, along with the other proteins which are produced by nuclear genes. Due to its small size, high mutation rate, and lack of recombination, the mitochondrial genome is a useful resource in studying genetic evolution and population genetics. Mitochondrial variations are employed to define various haplogroups, which help in the categorization of the human population according to their migratory and evolutionary path. In general, the individual mutations could be lethal for an individual and force natural selection leading to a selective sweep in the population. However, most of the mutations do not impose any change at the phenotypic level. Hence, remain in the population as a polymorphism (aka genetic variation). These genetic variations seem to have no impact on the fitness of a population individually, but their mutual presence does affect the phenotypic fitness of the population. Such a phenomenon is expressed as epistasis, where the presence of one variation might be deleterious along with the other variation, or it could be beneficial. Thus, an epistatic inter-

action could be referred to as a competitive or complementary interaction. In the human mitochondrial genome, it is observed that a variation may have different effects depending upon haplogroup/genome backgrounds. This could be explained by the hypothesis of genetic hitchhiking, which states that a selective sweep at one position in the genome could alter the allele frequency at a nearby position. Considering the fact that the relationship between phenotypic effects and the presence of variations is not direct, it becomes significant to assess the collective role of variations in structuring mitochondrial genetics in the human population. Genetic variations were observed to impart their effect at the phenotypic level as a cohort of multiple interactions and rarely individually. The heritability of complex diseases is minutely affected by the mere presence of single nucleotide polymorphisms (SNPs). However, the manifestation of such diseases depends upon the interactions of SNPs. Therefore it is important to study the collective effect of genomic variations. There are various ways to study the collective effect of the variations and the specific interactions between genes associated with specific traits. The genomic variations and their mutual presence could be readily analyzed for phenotypic adaptations under the framework of networks as these variations, and their presence represents a complex system. Network science has shown accountable success as an effective tool to study the behavior of complex systems through readily modeling them as networks in the past few decades. Alongside the emergence of Big data for biological systems such as proteomics, transcriptomics, genomics, or in general, the omics data provide us with a greater opportunity to understand the underlying mechanisms of these biological systems. However, the dimension of such data and its interacting entities poses limitations at statistical levels, which leads to the employment of an obvious network framework as an elementary yet mathematically robust tool to analyze and model such data. The research work performed in this thesis is to develop a network framework to model the occurrence of mitochondrial genomic variations in the human population (Chapter 1). The results obtained through analysis of network properties of these networks obtained, along with the biological relevance, contemplate to supplement the understanding of the impact of the mutual presence

of variations in mitochondrial genomes in a given population for its evolutionary adaptation and ancestral seeding (Chapter 2 and 3). However, the epistatic interactions occur pair-wise; their impact is collective. We analyzed the higher-order interactions among mitochondrial genes by constructing 2-order simplices of the co-mutating variable sites (Chapter 4).

Objectives

- To devise a network framework for mutual occurrence of variations considering the minor allele frequency in the human mitochondrial genome.
- To investigate the role of perfectly interacting two-order motifs of genomic variations in high-altitude adaptation in Asian human population.
- To investigate the role of co-mutating variations in convergent evolution of high-altitude populations of the three continents.
- To analyse the higher-order genetic interactions of co-mutation networks.

Background

Summary of the work done

Network Models: Construction and Background

This chapter is dedicated to the description of networks in general and the network construction methodologies used in the other chapters of the thesis. The variable sites could be defined as nodes given their allelic information. Based on this definition of allelic information of variable sites and their association, we categorized these networks into (i) Co-occurrence and (ii) Co-mutation networks. Such networks provide insights into the evolutionary patterns of given species under the

spectrum of external environmentsspecifically, fast-evolving viral genomes and mitochondrial genomes. In co-mutation networks, the variable sites are isolated based on minor allele frequency. Considering the minor allele may have one additional aspect, the comparison with the reference sequence since sometimes minor allele in the reference sequence might be the one present as major allele in the population in question. We discussed popular model networks, their structural properties, and their relevance with the network frameworks presented in this thesis.

Co-occurring motifs analysis in high to low altitude populations

This chapter discusses altitude-driven co-occurrence of variations in Tibetan and lower altitude populations using the two nodes network motifs. 673 complete mitochondrial genomes were grouped into eight groups based on their geographical origin according to altitude. The network motifs analysis of human mitochondrial genomic variations has furnished a new point of view on the role of genetic interactions based on mutual variations. Although the network motifs were formed by all the thirteen protein-coding genes, tRNA genes, rRNA genes, and the non-coding region of mtDNA, there are selective interactions that have played a critical role in distinguishing high-altitude populations from low-altitude dwellers and high-altitude adaptation. The nodes of these network motifs were categorized into three categories, (a) Local nodes, (b) Global nodes, and (c) Mixed nodes, based on their occurrence in different altitude groups. The tendency of co-occurrence is higher in the high-altitude population compared to low-altitude populations, suggesting the emergence of network motifs as a functional aspect of mitochondrial genomes in evolution at higher altitudes. The co-occurrence analysis of the high-altitude markers revealed the presence of intra-genic constraint in the high-altitude population. This suggested that the presence of a particular variation was not sufficient for adaptation, but that variation had to be assisted by other variations in the same gene or same gene complex. The formation of local co-occurrence pairs and similarity clustering divided the altitude groups into the higher and the lower altitude regions. This division might be possible due to two reasons, (i) migration and demographic

dynamics or (ii) the process of selection on mitochondrial variants that, in combination, optimize mitochondrial bioenergetics in extreme conditions experienced by these populations that lived at high altitudes.

Co-mutation network analysis of three high-altitude populations

In this chapter, we analyze the possible role of mitochondrial co-mutations for three high-altitude populations, Tibetan in Asia, Ethiopian in Africa, and Andean in America, in light of possible convergent evolution, using mtDNA genomes through networks. We constructed the co-mutation network by selecting significantly interacting variations of the mitochondrial genome for each sub-population. We performed the analysis of various structural properties of these co-mutation networks. 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 the essential role of CYB and CO3 genes for high-altitude adaptation in Tibetan and Andean populations while ND genes for the Ethiopian population.

Higher-order interactions among mitochondrial genes

In this chapter, we sought to investigate the extent to which the presence of higher-order interactions among the mitochondrial genes in three different altitude populations, low-altitude (0-500m), middle-altitude (2001-2500m), and high-altitude (>4000m) affect the overall behavior of mitochondrial genome according to demography and environmental constraint. The studies of two-order interactions, as epistatic interactions, provided ample evidence about the role of genetic background in the manifestation of a variation into phenotypic effect. However, these studies

also raised the concern of investigating higher-order epistatic interactions and their role in the phenotypic landscape. Here, we investigate the higher-order interactions in terms of 2-order simplices in co-mutation-based weighted genetic interaction networks of mitochondrial genes for three different altitude populations. We slightly modified the formula for calculating the co-mutation frequency for three variable sites by introducing the allele information of the third allele. In the mitochondrial genome, $\sim$10% higher-order interactions were found to be true simplices. In terms of vertex degree, these 2-order simplices captured distinct haplogroups in different altitude populations. In low-altitude, it captures M-haplogroup, in the middle altitude, C-haplogroup, and in high-altitude, K haplogroup. The genetic analysis of true weighted gene simplices revealed that in the lower altitude group, *ATP6 and ND* genes, in the middle altitude group, *CO1 and ND5* genes, and in the high altitude group, *CYB and ND5* genes are predominantly forming higher-order simplices.

Conclusion and future scope

Evolution is majorly regulated by selective inheritance of sudden and random changes in the genome with a direct impact on the fitness of individuals in a population. These random changes, considered as polymorphisms or genetic variations, get incorporated into the genome and are vastly inherited from one generation depending on their role in selective advantage in adaptation to given environmental conditions. Therefore, there are certain variations that are beneficial, some are deleterious or lethal, and some are just neutral in their phenotypic effect. Identification and characterization of these variations are core to molecular evolutionary biology as well as molecular pathological studies. However, under certain conditions, these variations give rise to phenotypic effects when mutually present, the phenomena explained through epistasis and genetic hitchhiking. The work done in this thesis discusses in detail developing a network framework to address the mutual role of genetic variations in mitochondrial genomes. We applied the developed network framework to two sets of populations, and by analyzing the network properties at the mutation and genetic levels, we were able to identify the characteristic differences in the different

populations at the molecular level.

The network framework of variable sites provides an effectively simple tool to analyze genomes for the possible role of epistatic interactions at an evolutionary and functional level. This work can be further extended in two ways, one, at the level of technique enhancement with more biologically relevant parameters, and second, at the level of the data being used. For the technique level, at a primary stage, the co-mutation network construction considers just the presence of variations in the population which can be modified at an advanced stage by considering (i) different minor allele frequencies for particular variations depending on their penetrance in the population, (ii) rare and very rare variations and (iii) different weights for pathological or deleterious mutations. Since it is already established fact that the mitochondria and nucleus work together at the molecular and genetic level, this study can be further extended to incorporate both the genomes together to extract the mito-nuclear interactions of variations and quantify their role in evolution, disease progression, and aging.

Keywords : Network Science, complex networks, co-mutation, mitochondrial genome, human evolution.

LIST OF PUBLICATIONS

Publications from thesis

1. **Verma, R.K.**, Kalyakulina, A., Giuliani, C. et al. Analysis of human mitochondrial genome co-occurrence networks of Asian population at varying altitudes. Sci Rep 11, 133 (2021). DOI: 10.1038/s41598-020-80271-8

2. **Verma, R.K.**, Kalyakulina, A., Mishra, A. et al. Role of mitochondrial genetic interactions in determining adaptation to high altitude human population. Sci Rep 12, 2046 (2022). DOI: 10.1038/s41598-022-05719-5

3. **Verma, R. K.**, Shinde, P., Jalan, S. (2022). Nucleotide-based genetic networks: Methods and applications. Journal of biosciences, 47(4), 63. DOI: 10.1007/s12038-022-00290-7

Other publications

4. Shinde, P., Whitwell, H. J., **Verma, R. K.**, Ivanchenko, M., Zaikin, A., Jalan, S. (2021). Impact of modular mitochondrial epistatic interactions on the evolution of human subpopulations. Mitochondrion, 58, 111122. DOI: 10.1016/j.mito.2021.02.004

5. Sarkar, C., Gupta, S., **Verma, R. K.**, Sinha, H., Jalan, S. (2018). Longitudinal network theory approaches identify crucial factors affecting sporulation efficiency in yeast. bioRxiv, 068270. DOI: 10.1101/068270

Table of Contents

List of Figures

List of Tables

Introduction

Chapter 1

Introduction

1.1 Overview

The identity of an individual entity lies in the wholeness of the system in which it is present. We observe numerous complex phenomena happening around us, and to study them, we define them as systems with particular entities leading to the commencement of those phenomena. Modeling these complex systems gives rise to the formation of complex networks. These networks represent the meaningful connections between the entities of the complex system. "I think the next century (21^{st}) will be the century of complexity", once said Stephen Hawking in light of the omnipresence of complex systems around us. The past two decades observed the immense potential of network science due to its holistic approach, flexibility, and applications to vast fields of scientific research. Network science has provided various models and algorithms under the umbrella of statistical physics to analyze natural and social sciences, including complex biological systems [1]. Like any other physical system, it is also required to identify and characterize the individual building blocks in complex biological systems and obtain and establish insights

into the interactions. The biological complex systems can be defined by multiple types of entities such as biomolecules (proteins and genes), pathways (metabolic, anabolic, and disease), cells (neurons), tissues (brain regions), and organs (human complexome) along with their defined interactions. In biological systems, interactions among cellular entities are not always straightforward as in social and physical networks. Hence, their interpretation becomes much more complicated, aided by the immense size, temporal dynamics, and non-linearity behavior. However, the vast diversity of biological systems allows us to define them at various levels into network models. At the subcellular level, protein-protein interaction networks [2], gene regulatory networks [3] and metabolic networks [4] have been characterized and discussed thoroughly. The analysis of these exemplary networks identified significant molecules and interactions and facilitated the prediction of possible interactions critical to cellular function. From a topological perspective, these networks exhibit non-random structures with scale-free or small-world properties and the presence of communities [5]. Along with the global behavior of these networks, there are several structural properties such as degree, clustering coefficient, degree-correlation, and several centrality measures, which can be analyzed to extract specific information about important nodes as well as their interactions in the network [6]. The analysis of these properties not only helped reveal the sub-cellular mechanisms but also improved our perception of disease prognosis [7]. In the following section, we briefly describe the mitochondrial genome, variations, and haplogroups, followed by the discussions regarding the complex behavior of genomic variations in the context of population and haplogroups.

1.2 Background

1.2.1 Mitochondria and mitochondrial genome

In this section, we discuss the mitochondria, and mitochondrial genome since the research work included in this thesis revolves around the mitochondrial genome. These cell organelles, previously known as α-proteobacterium, ended up in a precursor eukaryotic cell as a process of endosymbiosis [8]. During evolution, the mi-

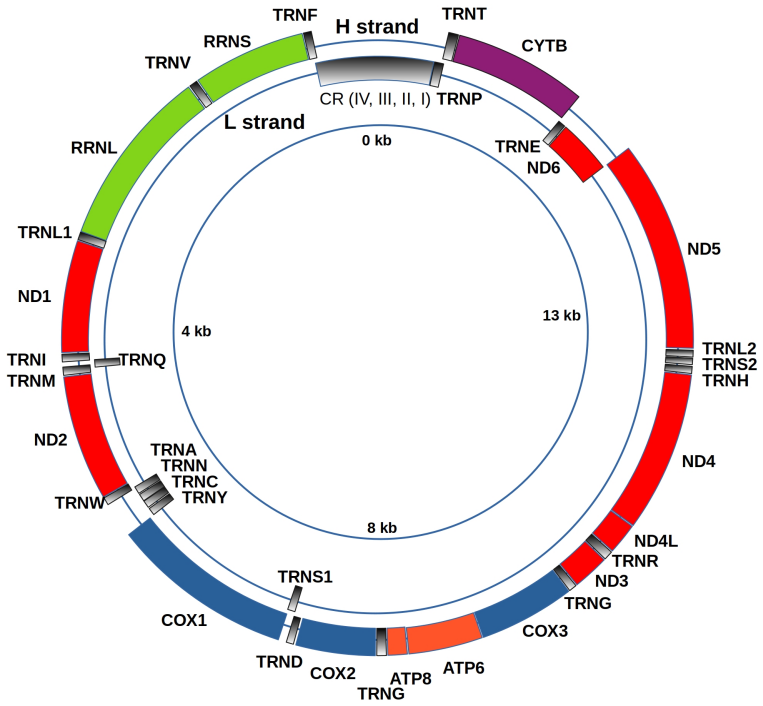

Figure 1.1: Map of human mitochondrial genome (reproduced from Mitomap).

tochondria retained bi-membrane structure while drastically reducing the genome size by losing or transferring genes to the nuclear genome [9]. The remaining genome is 16 kbp, circular, double-stranded with the presence of multiple copies in each mitochondrion, inherited maternally and in a semi-autonomous fashion. Among mammals, the structure and organization of the mitochondrial genome are highly conserved [10]. Due to the difference in the distribution of Guanine/Thiamine base composition in the two strands resulted in different buoyancy densities, which led to denoting them as L (light) strand and H (heavy) strand (Figure 1.1) [11]. The H strand encodes for 12 proteins, two rRNAs, and 14 tRNAs, while the L strand encodes for eight tRNAs and a single polypeptide. The introns are absent, and one regulatory region is present, along with a few overlapping regions and some intragenic bases. Each strand has its own origin of replication and three initiation sites, one for the L strand and two for the H strand in the regulatory region, also known as the D-loop region [12]. Apart from the mtDNA genes, the mitochondrion har-

bors $\sim$ 1000 proteins originating from the nucleus, whose composition might vary depending upon the spatiotemporal demands of the specific tissues [13]. Mitochondria are the energy centers in eukaryotic cells that produce adenosine triphosphate (ATP), the energy currency, by different oxidizing biomolecules through respiration. The 13 proteins encoded by mtDNA are core constituents of electron transport chain complexes I, III, IV, and V. The respiratory chain complexes create a proton gradient across the inner membrane, which is then fuelled to pump complex V, ATP synthase, to generate energy. The efficiency with which the proton gradient is converted into energy production is known as coupling efficiency. A tightly coupled mitochondria produce more ATP and less heat. When electrons are maximally reduced, they can be transferred to O2 to generate increased free radicals and oxidative stress. Alternatively, restricted calories and regular exercise keep the electrochemical gradient hyperpolarized, leading to limited oxidative stress. This can be achieved by reducing the coupling efficiency by alterations of complexes I, III, or IV by disconnecting electron transport from proton pumping or by expressing a proton channel which helps in leaking electrons. The structure and organisation of mitochondrial genome is highly controlled. mtDNA is highly condensed within mitochondrial matrix and termed as nucleoids. The primary protein involved in this compaction was identified as TFAM, *m*itochondrial *t*ranscription *f*actor *A* [14, 15]. TFAM also plays major role in actively controlling the transcription of mtDNA genes, mtDNA copy number and its maintenance [16, 17].

1.2.2 Mitochondrial variations and haplogroups

As we know that the mtDNA is inherited through maternal lineage only and present as haploid. Thus it is defined as a haplotype. The related group of haplotypes constitutes a haplogroup that is defined by shared mutations and with regional specificity. Human origin is rooted in Africa, based on mtDNA about 130,000 to 170,000 years before present (YBP). The distribution of mtDNA haplogroups worldwide reveals four striking regional mtDNA discontinuities. For the first $\sim$ 100,000 years, short-distance migrations gave rise to a plethora of Africa-specific haplogroups, L0-L3,

which established the first macrohaplogroup L [18]. Among these, L0 was the most ancient haplogroup found in the Koi San peoples, and L1 and L2 are found in Pygmy peoples. From this macrohaplogroup, L3 (youngest) originated 65,000-70,000 years ago in Sub-Saharan Africa. It gave rise to two major haplogroups, M and N, which later successfully migrated out of Africa and populated the rest of the world [18]. The N haplogroups radiated into Eurasian indigenous populations, giving rise to haplogroups H, T, U, V, W, X, I, J, and K. About 40,000-50,000 YBP, the Europeans separated from Africans. In Asia, lineage M gave rise to A, B, F, and lineage N gave rise to C, D, G, and others. Around 20,000 YBP, haplogroups A, C, and D became enriched in northeastern Siberia, which led foundation population for Native Americans by migrating across the Bering Land Bridge. Different haplogroups were founded by purifying selection of specific mitochondrial variations and enriched at the regional level [19]. These discontinuities posit the role of environmental constraint and natural selection on mtDNA diversity. The evidence for natural selection was provided by the analysis of nonsynonymous (Ka) to synonymous (Ks) mutation ratios (Ka/Ks) from the 13 mtDNA transcripts [20, 21]. It was revealed that the amino acid sequence of the *ATP6* gene was highly variable in the arctic but was strongly conserved in the tropics and temperate zone; *CYB* gene was hypervariable in temperate Europe but conserved in the tropics and arctic, and *CO1* was variable in tropical Africa but invariant in the temperate and arctic regions. Regional variation was also observed in multiple ND subunits [19]. Such regional gene-specific variation would not be expected if all mtDNA mutations were random and neutral.

1.2.3 Mutual presence of nucleotide variations

In the human mitochondrial genome, it is observed that a variation may have different effects depending upon haplogroup/genome backgrounds [22]. As we know that the 3394C variant confers high-altitude adaptation in high-altitude Tibetans, its haplogroup background plays a vital role in its phenotypic consequences. When 3394C is present on M haplogroup on M9 background in Tibetans and on C4a4

background in Indian Deccan plateau, it has beneficial effects and does not affect the complex I activity. Its presence in the N haplogroup reduces the complex I activity and associates with Leber hereditary optic neuropathy (LHON), suggesting the role of haplogroup background in modulating the bioenergetics. Similarly, the non-pathogenic missense variants cause low penetrance LHON, as shown in a study of complete mtDNA sequences of three families from southern Italy and one from Northern Italy [23]. It was reported that the variants, otherwise polymorphic, when present in a peculiar combination, lead to reduced complex I activity and thereby onset of LHON. There could be multiple reasons behind such complex observations, one being the hypothesis of genetic hitchhiking, which states that a selective sweep at one position in the genome could alter the allele frequency at a nearby position [24]. Another phenomenon where mutual variations impart their effect at the phenotypic level is observed in the form of epistasis, which usually deals with alterations in traits associated with those variations [25]. Because the relationship between phenotypic effects and the presence of variations is not direct, it becomes significant to assess the collective role of variations in understanding mitochondrial genetics in the human population. Genetic variations were observed to impart their effect at the phenotypic level as a cohort of multiple interactions and rarely individually [26]. The heritability of complex diseases is minutely affected by the mere presence of single nucleotide polymorphisms (SNPs) [27]. However, the manifestation of such diseases depends upon the interactions of SNPs [28–30] therefore it is essential to study the collective effect of genomic variations. There are various ways to study the collective effect of the variations and the specific interactions between genes associated with specific traits [31]. 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. Among which principal component analysis to infer groups of SNPs from linkage disequilibrium to evaluate multivariate SNP correlations for intragenic diversity coverage [32], integrative scoring system based on their deleterious effects [33], and Pareto-optimal approach for identifying functionally and informatively

significant SNPs [34], are most popular ones. There exist other approaches based on pair-wise interactions, such as two variations significantly interacting through logic regression [35], predictive rule inference [36], and shrunken dissimilarity measure, in which a gene-gene similarity value is calculated and pairs are selected if the similarity value crosses a set threshold value [37].

1.3 General Network Properties

A network consists of a set of nodes (N) or units that are connected, where an interaction type defines connections (Nc). A network can be represented by an adjacency matrix or an adjacency list. An adjacency matrix is a square matrix with size *N x N*, which could have binary entries (unweighted) or non-binary entries (weighted) depending upon the way the network is constructed. In a binary matrix 1 represents connection between given nodes i and j and 0 otherwise.

- **Adjacency matrix:** An adjacency matrix (**A**), is a square matrix with rows and columns equal to the total number of nodes present in the network. The adjacency matrix is defined by Eq. 1.1

$$A_{ij} = \begin{cases} 1 & \text{if nodes } v_i \text{ and } v_j \text{ are connected} \\ 0 & \text{Otherwise} \end{cases} \tag{1.1}$$

- **Degree:** The degree is a local property of a given node which represents number of connections formed by that node. For a node v_i, it is denoted as $k_{v_i} = \sum_{j=1}^{n} a_{ij}$ or simple k_i.

- **Degree sequence:** The degree sequence ($\{k_i\}_{i=1}^{n}$) of a graph $\mathcal{G}$ is the sequence obtained by listing the vertex degrees of $\mathcal{G}$ in increasing order, with repeats as necessary.

- **Average degree:** The average degree of the network is denoted by $\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i$.

- **Degree distribution:** Degree distribution of a network is represented as $p(k)$ which says fraction of vertices having degree k. We can calculate $p(k) = \frac{\Gamma_k}{n}$,

where Γ_k is the number of nodes having degree k [1].

- **Clustering coefficient:** The tendency of nodes in a system to form triangles is captured by the clustering coefficient [1]. Most real-world networks shows high clustering coefficient than corresponding random network. The higher clustering also straight forward provides information of existence of modularity in the network. The clustering coefficient of a node i is calculated as,

$$C_i = \frac{2k_n}{k_i(k_i - 1)} \tag{1.2}$$

where k_n is the number of connections between the neighbours of i. The average clustering coefficient is represented as,

$$\langle CC \rangle = \frac{1}{n} \sum_{i=1}^{n} C_i$$

and $\langle CC \rangle$ is the probability that two neighbors of a randomly selected node link to each other.

- **Degree-degree correlation:** The correlation between degree of nodes is measured by the Pearson correlation coefficient. it tells about the (dis)assortative nature of network and denoted as,

$$r_{deg-deg} = \frac{[m^{-1} \sum_{i=1}^{m} j_i k_i] - [m^{-1} \sum_{i=1}^{m} \frac{1}{2}(j_i + k_i)]^2}{[m^{-1} \sum_{i=1}^{m} \frac{1}{2}(j_i^2 + k_i^2)] - [m^{-1} \sum_{i=1}^{m} \frac{1}{2}(j_i + k_i)]^2} \tag{1.3}$$

where m is the total number of edges in the network and j_i, k_i are the degrees of nodes with i^{th} edge and $r_{deg-deg}$ value varies in between -1 to 1. When high degree nodes connected to other high degree nodes in a network, then $r_{deg-deg}$ values become positive and referred to as an assortative network. In case of high degree nodes connected to lower degree nodes, then $r_{deg-deg}$ becomes negative, and the network is said to be dis-assortative.

- **Hierarchy:** A decrease in the clustering coefficient with an increase in the degree of nodes suggests presence of hierarchy in the network [5]. This indicates that the nodes with small degree belong to highly interconnected small modules. To quantify the hierarchy, local reaching centrality (C_R) is measured for node i as the proportion of all the nodes that can be reached from node i. Hierarchy arises due to the fact that the modules are not completely

independent in a network where a few nodes play crucial role to cross talk between any two or more modules. It is calculated as,

$$C_R(i) = \frac{1}{N-1} \sum_{j:0<d_{i,j}<\infty} \frac{1}{d_{i,j}} \tag{1.4}$$

where $d_{i,j}$ is the shortest path length and N is the total number of nodes. Based on C_R, the hierarchy is defined as,

$$h = \frac{\sum_{i \in V} [C_R^{max} - C_R(i)]}{N-1} \tag{1.5}$$

where C_R^{max} is the highest reaching centrality in the network.

Betweenness Centrality: It is the measure of importance of a node as connector or bridge between two modules or communities independent of their degree. It is defined as the fraction of shortest paths between all the pair of nodes that pass through the node i, and calculated as,

$$\beta_i = \sum_{st} \frac{n_{st}^i}{g_{st}} \tag{1.6}$$

Modularity: In biological systems, the components form groups to perform relatively different functions at molecular level [38]. This property of a system is referred as modularity. High clustering coefficient is signature of a network to have high modularity. This also gives the information about the motifs which are highly connected subgraphs. Motifs are present in almost all the real-world networks that have been examined so far. Modularity can be calculated using various algorithms such as Newmann-Girvan [39] or Louvian [40].

1.4 Network Models

In this section we discuss about the popular model networks.

- **Erdös-Rényi random network:** The ErdösRényi (ER) model is used for generating a random network. Starting with N nodes, connects each pair of nodes with a probability p, which creates a graph with approximately $pN(N1)/2$ randomly placed links with most nodes have approximately the same number of links (close to the average degree $\langle k \rangle$). The node degrees follow binomial distribution [41]. This is commonly used model to compare

with real-world networks for their random behavior. To generate the ER random networks, we provide two parameters, one is the network size (n) and another is the probability (p).

- **Scale-free networks:** A scale-free (SF) networks are characterized by power-law ($p(k) \sim k^{-\gamma}$) degree distribution with degree exponent $2 \geq \gamma \geq 3$ [41]. Barabasi-Albert preferential attachment algorithm is used to construct this model network. Many real-world networks follow power-law degree distribution and one of the main reason for the popularity of this model network.

- **Small-world networks:** Small-world networks are characterized by high clustering coefficient and small average path length. To construct such networks one can start with a regular network and then rewire the edges for a given probability, p_r. For p_r = 0, a regular network is sustained while for p_r = 1, a completely random network is obtained. The mean path length is proportional to the logarithm of the network size, $l \sim logN$. It follows degree distribution similar to ER network [41].

- **Configuration model:** In the random or scale-free networks the degree sequence can not be controlled as real-world networks. The configuration model was proposed to generate a network randomly while keeping the degree sequence fixed. It is known that the generated networks through the configuration model having a fixed degree sequence, however, allow for self-loops and multiple edges [42].

1.4.1 Subnetwork patterns and functions

Network Motifs: Certain patterns of interconnections (subgraphs) were found to be significantly higher than those in randomized networks. These patterns were considered as building blocks of complex networks and defined as motifs [43]. There could be various contexts where the presence of motifs are observed in the networks in various forms, such as feed-forward loop motifs usually found in gene-transcriptional regulatory networks involved in regulating the gene expression [44],

triangular motifs observed in social and disease spreading networks [45, 46], and ecological networks [47]. Motifs detection have been applied in cancer diagnosis [48], studying neural networks [49], and brain functions [50].

Communities: Communities or modules are defined as relatively densely connected set of nodes to each other compared to other densely connected groups of nodes in the network. The identification of such modules is of particular importance. For example, in food webs, identification of communities based on habitat rather than the taxa [51], in metabolic networks, the nodes of a module correspond to a specific cellular metabolic activity [52], and in gene regulatory networks, the modules formed represent a group of genes with similar functions [53].

Dynamics in networks: In the past two decades, with the advancement of network science, our understanding of biological networks has significantly improved as we know that networks are not always static and have inherent dynamics embedded in them. The network dynamics handle the behavior of interactions depending upon their Spatio-temporal variations [54]. Since the large-scale *in vitro* methods detecting the biomolecular interactions do not provide the spatial-tempo-contextual information of these interactions, the dynamics underlying the biological networks have been either overlooked or limited to a few proteins or genes along with their interactions [55]. Apart from the dynamics associated with protein-protein interaction networks, and gene regulatory networks, there are dynamics associated with interactions of genetic variations and corresponding phenotypes. Genetic variations that correspond to changes in phenotype are known as eQTLS (expression quantitative trait loci). The analysis of these eQTLS revealed that gene expression variations are often linked to interactions between these loci or genetic variations (epistasis) [54]. Recently, a model has been developed to show the long time effects of physical dynamics on epistatic consequences [56].

1.5 Nucleotide networks construction

1.5.1 Co-occurrence networks

Co-occurrence networks take into account the position of variable sites as nodes, and the connection between a pair of nodes is defined based on the co-occurrence of alleles for the given population. In this way we get one network for each sequence in the population. Since we have already mentioned that a co-occurrence network considers the allele (major or minor) present at a variable site for a particular sequence, we define our edges with respect to the alleles and their frequencies for pair of variable sites in the population. The co-occurrence frequency between a pair of variable sites for the alleles in position is calculated as,

$$Co_{ij} = \frac{(x_i y_j)^2}{(x_i)(y_j)} \tag{1.7}$$

where, Co_{ij} is the co-occurrence frequency between x and y alleles present at i^{th} and j^{th} variable site for a particular sequence. The numerator $x_i y_j$ is the frequency of presence of x and y allele together at i^{th} and j^{th} sites whereas the x_i and y_j are total frequency of allele x at i^{th} position and y allele at j^{th} position, individually. The value of Co_{ij} gives information about the co-occurrence of two nucleotide positions in a given sample with respect to their presence in the whole population. The co-occurrence frequency ranges from [0 to 1], and we need to define a threshold value to filter out the possible noise to get a structurally meaningful sparse network. To define a threshold, we look for two structural properties of the network, one is the order of the largest connected component (LCC), and the other one is the average degree ($\langle k \rangle$) of LCC. These two properties help in constructing a sparse network that is structurally more meaningful by filtering out some connections. For calculating the threshold, we start to construct a network by considering all the pairs with $Co_{ij} > 0$, which gives rise to more or less a globally connected network. To get a meaningful sparse network from such a network, we gradually remove the links with $Co_{ij} \leq Co_{th}$ (co-occurrence threshold) and keep only those connections above a particular Co_{th} and simultaneously calculate the Nc_{LCC} and $\langle k \rangle$. Following this procedure, we get a threshold where the Nc_{LCC} consists of almost all the nodes but

as few connections as possible, yielding a very low $\langle k \rangle$ or $Nc_{LCC} \sim N_{LCC}$, where Nc_{LCC} and N_{LCC} represent number of connections and number of nodes of the LCC, respectively (Figure (1.2). We apply the above-mentioned method of threshold selection for all the networks generated for a given population yielding as many sparse networks as the number of available sequences. In the next step, we construct a master network by merging all the individual networks generated for each sequence. In this step, we get duplicate nodes and edges; however, we choose to perform a union operation on our networks. By performing union operation on the edges, we get only one network in which all the nodes and edges of all the networks are taken into consideration only once, yielding a single undirected and unweighted co-occurrence network for all the samples.

1.5.2 Co-mutation networks

For co-mutation networks, we again start with the multiple aligned DNA sequences. The variable sites are isolated based on minor allele frequency. Considering the minor allele may have one additional aspect: comparison with the reference sequence, since sometimes minor allele in the reference sequence, might be the one that is present as major allele in the population in question. Even though the site with such an allele would still be considered a variable site. As we calculated the co-occurrence frequency in the previous section, here, we will be defining and calculating the co-mutation frequency between two variable sites based on their minor allele frequencies as,

$$Cm_{ij} = \frac{(m_{ij})^2}{(m_i)(m_j)} \tag{1.8}$$

where, Cm_{ij} is co-mutation frequency, m_{ij} is number of times minor alleles at i^{th} and j^{th} position occurs together, m_i and m_j are minor allele frequency at i^{th} and j^{th} positions individually. Calculating the Cm_{ij} is the first step for constructing co-mutation networks whose range is again between [0, 1]. Further, we calculate a statistical correlation (P_{ij}) (popularly known as p-value test) to filter out interactions lying below a threshold value to get a meaningful network.

$$P_{i,j} = \frac{\#[(Cm^r_{ij}) \geq (Cm_{ij})]}{\#reshuffling} \tag{1.9}$$

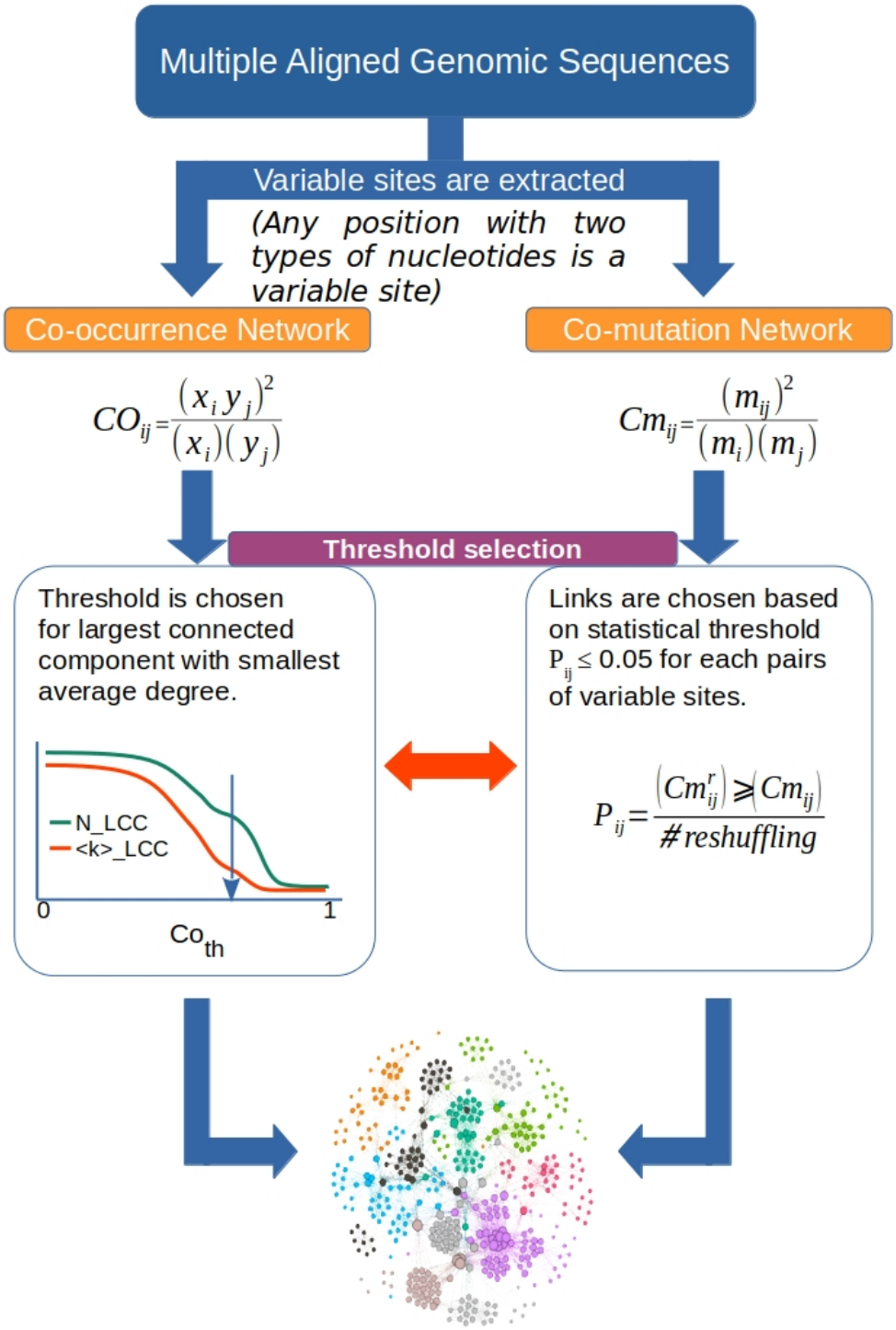

Figure 1.2: The schematic for nucleotide based co-occurrence and co-mutation networks.

where, Cm_{ij}^{r} is random co-mutation frequency, calculated after permuting the alleles at the i^{th} and j^{th} positions for a large number of times. As per the standard practice, we keep the threshold value at standard ≤ 0.05 to filter the interactions. This method yields only one network combining all the sequences, hence, the method is independent of number of samples.

1.5.3 Perfectly co-occurring sites

As discussed, we can apply a threshold value for generating a sparse network. In co-occurrence networks, if we keep the threshold of efficiency score at 1.0, and in co-mutation networks, if we consider only those pairs with $Cm_{ij} = 1.0$ and $P_{ij} \leq 0.05$, we get only those pairs of the variable sites that are perfectly co-occurring and co-mutating in the population, respectively. Such sites form complete subgraphs (all-to-all connected) and yield disconnected components (Figure 1.3). The peculiar case of perfectly co-occurring sites could be interpreted in two ways: first, the nucleotides at i^{th} and j^{th} positions are co-occurring in just one sample and not present in any other sample in that population; second, the nucleotides are co-occurring in many samples and are not present individually. In both these cases, we would get a perfect co-occurrence of the involved sites. However, in the first case, the significance of co-occurrence would be negligible for common variants but could have considerable importance for rare variants [57]. To avoid this bias, one can define the rare variants beforehand or consider only those sites with a higher minor allele frequency. The perfectly co-occurring sites give rise to disconnected complete subgraphs or motifs of order two or more.

1.6 Thesis Overview

In the upcoming chapters, the co-occurrence and co-mutation networks have been applied to mitochondrial genomes of different populations. In Chapter 2, the low and high-altitude populations are analyzed through a network framework, and fundamental aspects of evolution and adaptation are discussed. In Chapter 3, three different high-altitude populations are studied to analyze convergent evolution's ef-

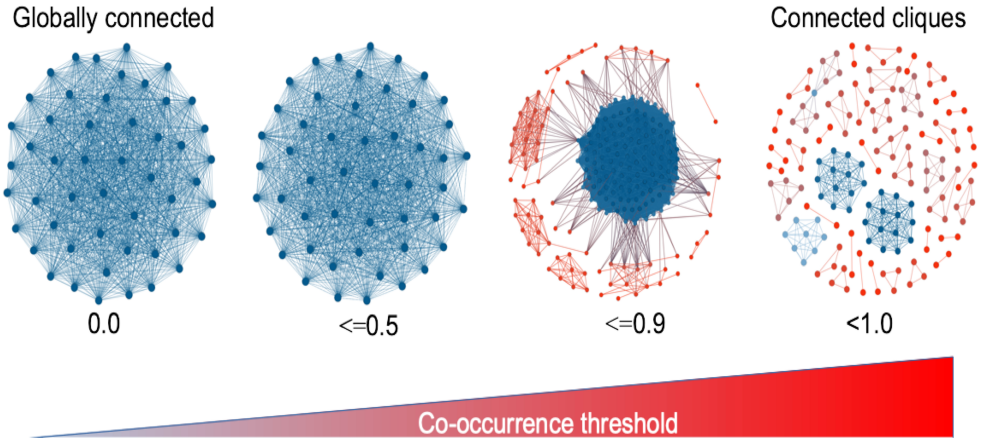

Figure 1.3: Evolution of co-occurrence network with change in co-occurrence threshold.

fect and identify the critical genes possibly involved in such evolution. In Chapter 4, we move one step further to identify the presence of higher-order interactions in three populations of different altitudes (lowest to highest). Chapter 5 provides the conclusion of the studies and discusses the future aspects of these studies.

Altitude based co-occurrence pattern in Asian population

Chapter 2

Altitude based co-occurrence pattern in Asian population

2.1 Introduction

To understand and predict behavior of many large scale complex systems, networks provide an extremely powerful framework consisting of nodes and interactions (edges) [58]. For example, complex biochemical activities of a cell can be well understood by underlying proteinprotein interaction (PPI) networks. Network framework has been successful in revealing crucial proteins for breast cancer [59], to understand versatility of society [60], to get insights into developmental changes in *C. elegans* [61]. Motifs, which are complete subgraphs of a network, considered to be building blocks of many complex systems [43]. These motifs have been shown to occur significantly in several biological networks such as gene regulatory networks, ecological networks and neural networks. Two-node motifs have been extensively studied as double negative feedback loops, double positive feedback loops, and auto-activation or repression loops [62, 63].

Motifs are complete subgraphs, and the two node motifs may primarily consist

of feed-forward or backward loop. Here we analyzed simple two-node undirected motifs of co-occurring nodes for mtDNA at varying altitudes. The nodes are variable sites and interactions are the co-occurrence of these variations. Co-occurrence of variable sites have been investigated for various diseases, for example in understanding classification and prognosis of acute myeloid leukemia [64], for finding cause of female Duchenne muscular dystrophy [65], for finding co-occurrence of driver mutations in myeloproliferative neoplasms [66]; to understand evolution of influenza viruses [67]; in detection of pesticide resistance in *Aedes aegypti* [68] and recently in codon level analysis of human mt-DNA which revealed significance of codon-motifs in evolution of human sub-population [69]. The origin and inhabitation of humans in diverse geographical regions across the world has always been a topic of research for anthropologists and geneticists. These studies have pointed out that the presence of environmental diversity in different geographical regions was one of the key factors in causing variability among human groups both at nuclear DNA and mtDNA level [70]. A wide range of environmental diversity existed in terms of temperature and altitude driven hypoxia all over the world. One such environment existed 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, 2218 kya) [71]. The plateau has an average elevation of 4000 m above sea level yielding extreme environments such as low oxygen concentration, high UV radiation and arid conditions [72]. 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) [73]. 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 [74]. 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 [75–77]. mtDNA mutations that affect OXPHOS could also affect metabolic rate modulation, oxygen utilization, and hypoxia adaptation [78]. Theoretically, it was accepted that migration and genetic drift play crucial roles in controlling mtDNA haplotype frequencies and that mt-DNA variations in a species are selectively neutral [79]. However, recently it was reported that mtDNA variations are the result of natural selection [80, 81]. 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. Due to these reasons, it was suggested that a site undergoing evolutionary pressure might have equally affected the genealogy of the whole mitochondrial genome [79]. In this chapter, we discuss the analysis of 673 complete human mitochondrial genomes by categorizing them into eight altitude groups from the sea level up 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 500m above sea level [82].

2.2 Source of data and network construction

We retrieved a total of 673 complete human mitochondrial DNA sequences of the healthy individuals from GenBank (http://www.ncbi.nih.gov/), where metadata was available. These sequences were aligned using multiple sequence alignment tools, Clustal Omega [83] using default parameters. After alignment, all the sequences were mapped to master sequence, revised Cambridge Reference Sequence (rCRS). Since we were interested in altitude-wise stratification of these sequences based on available geographic information, we divided these sequences into eight altitude groups ranging from 0m to>4000m with an interval of 500m based on specific coordinates provided for each sequence in the published dataset. In this way we have eight altitude groups with different numbers of mtDNA sequences for construction

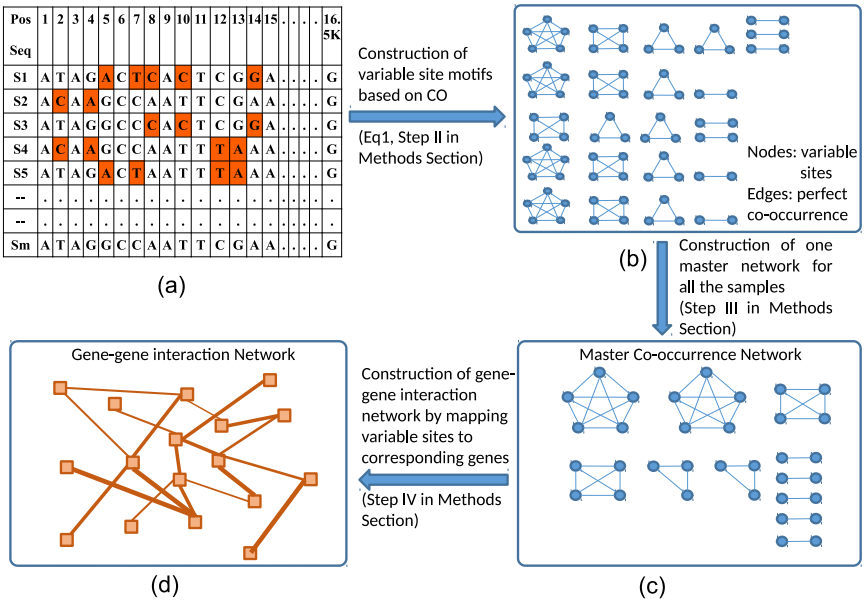

Figure 2.1: (a) Total variable sites are extracted from a particular group and co-occurrence threshold is applied. (b) For each sample, one set of motifs were constructed. (c) These motifs were then merged to construct one master network where nodes were variable sites. (d) This master network was then used to construct a genegene interaction network by mapping the variable sites corresponding to each gene.

of co-occurrence networks corresponding to each altitude group (Fig. 2.1).

2.3 Results

2.3.1 Characterization of variable sites (nodes)

A total of 3829 variable sites exist for all the altitude groups. Out of which 3127 variable sites took part in mutation cohort formation of motifs of order two or higher. However, here we have considered the simplest motifs of order two only for the further analysis. Among these variable sites, ∼65% sites were found to be located in protein-coding regions (overlapped sites were double counted) with the rest lying in non-coding regions (control region, t-RNAs and r-RNAs) (Table 2.1). This was not surprising as 11,395 (∼68%) sites out of the total 16,569 sites belong to the 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 2.1). These sites are well documented in the Mitomap database for

various genomic studies. Although much about the tri-allelic sites have not been understood, nonetheless their presence has been shown to be responsible for natural selection [84]. Approximately 90% of the variable sites were transitions, yielding a high transition to transversion ratio (Ts/Tv) (Table 2.1) which has already been reported [85] to be responsible for the conservation of structures at protein level among the individuals within a species [86]. In the context of varying altitudes, this ratio remained high which further emphasized the importance of survival and for mitochondrial functionality. 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 [87]. These interactions of amino acids, in turn, were shown to be affected by mutations and their co-occurrence in the genome [88].

Table 2.1: Statistics of variable sites

Altitude group	Variable sites	Coding sites	Non-coding sites	Ts:Tv	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

2.3.2 Categorization of variable sites

In this section we discuss about the categorization of 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 subcategorized into (a) global nodes (the nodes present in all the altitude groups), (b) local nodes (the nodes presented exclusively in a particular group) and (c) mixed nodes (the nodes presented in more than one altitude groups but not in all). This categorization helped us to decipher the role of variable sites in terms of two nodes co-occurrence motifs. It was also observed that the percentage of local connected

Table 2.2: 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	14.5	60.0	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

nodes was decreasing with increasing altitude while the percentage of local isolated nodes was increasing with increasing altitude. Further, the number of links provided the information about the co-occurrence pairs formed by the connected nodes having perfect co-occurrence frequency. The average degree of the network was found to be nearly similar in all the networks and also confirmed sparseness of the networks of all the groups. The real-world complex networks are found to be mostly sparse [89]. Here, the sparsity of co-occurrence networks meant that the variable sites are having very few perfectly co-occurring pairs.

2.3.3 Altitude classification through co-occurrence motifs

In the previous section we categorized the variable sites based on their presence in altitude groups. In this section we utilize that information for similarity analysis. Jaccard similarity coefficient was used to find out the similarity between each altitude group using the mixed nodes (Figure 2.2). This similarity coefficient led to the distinction of two major clusters, 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 in Figure 2.2 represented altitude groups. Two altitude groups were found to form doubletons within each cluster. Moreover, it was deduced from the dendrogram that the human population splits up into two subpopulations giving rise to lower altitude cohort and higher altitude cohort. The lower altitude cohort further segregated into two sub-groups forming one clade with Grp 2 and Grp 3 and

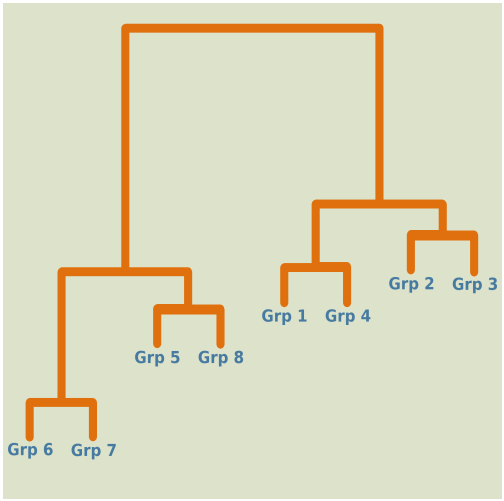

Figure 2.2: Cluster dendrogram was produced using common nodes between each altitude. It is clearly observed that two clusters are formed, one with groups 1 to 4 (lowest to middle) and other with groups 5 to 8 (middle to highest).

another clade with the Grp 1 and Grp 4. The higher altitude cohort further segregated into two sub-groups forming one clade with the Grp 5 and Grp 8 and another clade with the Grp 6 and Grp 7. It is noteworthy that in the lower altitude groups, Grp 1 and Grp 4 descended from a common sub-group. Similarly, in the higher altitude groups, Grp 5 and Grp 8 descended from a common sub-group despite having geographical distances between these altitude ranges. Many previous studies have pointed out that early humans have migrated towards higher altitudes in summer for hunting, whereas moved towards lower altitudes during winter season to avoid extreme harsh environments [90]. This seasonal migration is common even today in the plateau [91]. The segregation of the human population in lower and higher altitude groups observed in our analysis suggests that humans may have migrated through discrete pathways searching for a better environment for establishment.

2.3.4 Imact of nodes; CADD scores

To explore the predicted functional impact of selected variants, we extracted the various prediction scores from HmtVar database [92] and Combined Annotated Depletion Dependent (CADD) database [93]. The prediction scores for various variants which were found to be significant based on the co-occurrence networks. The

C-scores and PhyloP scores conveyed about the deleteriousness and conservation score, respectively. The positive PhyloP score predicted conserved sites while its negative value predicted fast-evolving sites. For most of the variants HmtVar scores were unavailable while we discussed the relative significance of those which were available. The high-altitude markers were shown to have only polymorphic nature with some degree of pathogenicity. The variant T3394C was predicted to have high pathogenicity along with disease-causing impact by MutPred, PhD-SNP, and SNPs-GO databases. This variant has been associated with Lebers hereditary optic neuropathy (LHON), diabetes mellitus, osteoarthritis, cardiomyopathy in non-Asian populations while in Asian population it has been associated with high-altitude adaptation. The co-mutating partners of T3394C were all fast-evolving sites except G4491A. There are four variants which co-occurred with G7697A in which two variants were highly evolving while two variants were found to be conserved. For most of the variants the predictions were not available in HmtVar while PhyloP score was available for all the variants. The PhyloP scores of pairs of Global connected nodes suggested that the Global connected nodes tend to pair with conserved sites across all the altitude groups. The C-scores for all the nodes for each altitude-group are shown in 2.3. The C-scores are ranging between2 and 4. Although the absolute C-scores had no meaning, they did have a relative meaning. The negative value predicted a site to be proxy neutral while a positive value predicted it to be a proxy deleterious site. It was observed that C-scores of the variants from the control region were close to 1 which suggested that the variants in the control region were neither deleterious nor beneficial across varying elevations. The C-scores for selected variants were checked where the highest deleterious score was 2.5 for C3310T variable site, however, this variant was exhibited to have low pathogenicity and likely to be neutral. By projecting the C-scores for all the nodes for each master co-occurrence network, we were able to look at the likely role played by each variant in contributing proxy-deleterious or proxy-neutral variants for co-occurrence network construction. We found that more than 60% nodes were falling in between 0 and 1 C-scores,$\sim$30% were found to show$>$1 C-score and $\sim$10% were found to

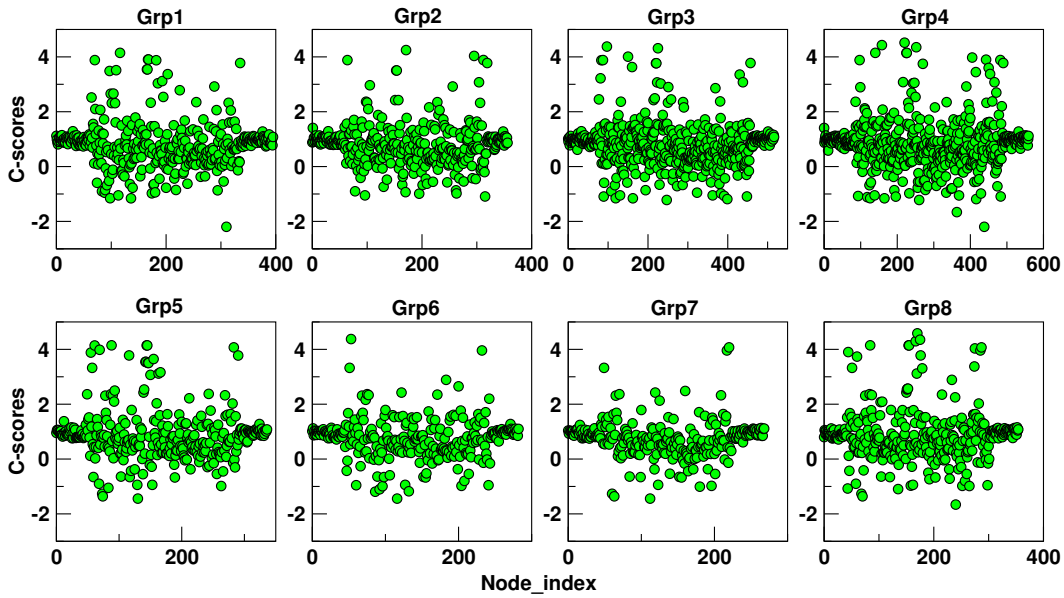

Figure 2.3: The CADD scores are plotted for all the nodes of each master co-occurrence network. The negative values show proxy neutral predictions while the positive values show proxy deleterious predictions for each variable site.

show <0 C-scores. This suggested that a few individual variants might have deleterious effects on population, however, mutational cohorts might be able to subsidise these predicted deleterious effects.

2.3.5 Altitude dependent Gene-Gene interactions

Gene gene interaction networks were compared with the corresponding random networks. We found that only 72 ($\sim$ 10%) of all possible gene-pairs were significantly deviated (falling out of the standard deviation range) from random networks (Figure 2.4). Moreover, out of total 72 gene-pairs, 46 were found to be exclusive to any one of the altitude groups, 23 were found to be exclusive to any two of the altitude groups, and only 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 genes was tRNA such as in group 1, *tRNA-Gly*. The gene-pair *CYB-CR* was found to be present in all the groups except in the 5th group where these genes were present but interacted with other genes. Further, this pair had more weight than that of the corresponding random network in group 6 while less weight than that of the corresponding random networks in the other groups. Interestingly, there existed only three genes which

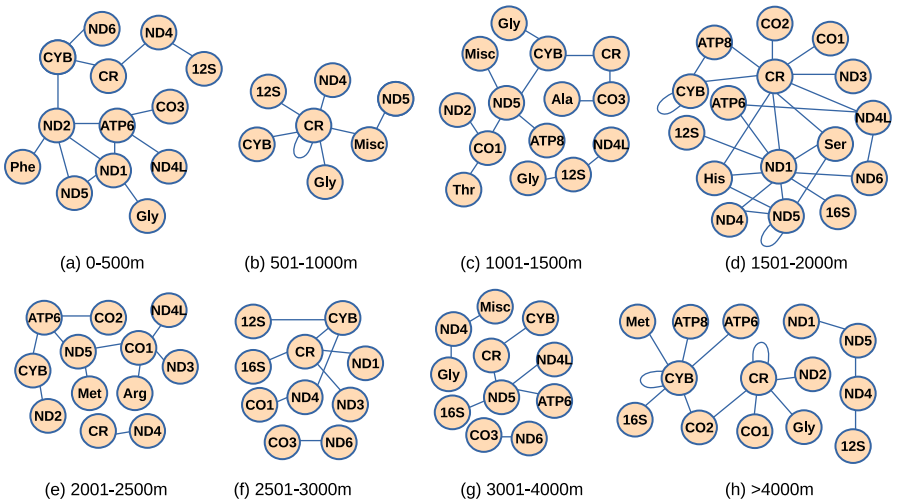

Figure 2.4: Gene gene interaction pairs which showed deviation from random networks.

formed pairs with themselves, these genes are *CYB, ND5* and *CR*. The gene-pair *CYB-CYB* was found to be present in the groups 4 and 8. In group 4, its weight was found to be less than that of 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 genes was *CYB*. Another gene-pair *CO3-ND6* was found to be present in groups 6 and 7. Apart from the CR and the coding genes, despite having small length and less variable sites, certain tRNA genes were also found to form gene interaction pairs. Particularly, *tRNA-Gly* was present in groups 1, 2, 3 and 8. Interestingly, *tRNA-Thr* exhibited the 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 genegene interactions at varying altitudes which can be further analyzed for possible adaptation or disease association.

2.4 Conclusion

In this chapter, we investigated altitude driven co-occurrence of variations in Tibetan and lower altitude population using the two nodes network motifs. Here,

we used a network model to represent the mitochondrial genome as a complex genetic 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 the network construction suggesting a widespread presence of mutual variations in the human mitochondrial genome. This widespread of mutual variations further suggested that these variations richly co-occur with each other. The rest of the 25% nodes, which we categorized as the Isolated nodes, were mostly contributed by the Control region which was a mutational hotspot in human mtDNA. These isolated nodes were also found to be more in the population belonging to the lower altitude groups as compared to those at the higher altitude groups. An absence of any selective pressure in lower altitude groups might be a reason for lower co-occurrence of the variable sites yielding high number of isolated nodes (more independent signals). Whereas, for the high-altitude groups which were accompanied with peculiar conditions (oxygen, temperature, UV, etc.) leading to more selective pressures which might be causing co-occurrence of the variants for their positive advantages. Further studies are needed to correlate these observations with peculiar phenotypes. Another category of the nodes, the local nodes, constituted nearly 30% of total nodes for a particular group. The presence of local nodes provided evidence that the human population had a specific signature at nucleotide variation level for varying altitudes, which might be arising by different genetic history of the population that colonized a particular area especially in the low/medium altitudes where the intensity of natural selection on human mt genome was likely to be low. This signature seemed 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 the signature revealed a peculiar property of the mitochondrial genome that even with exclusive variations, the genomic functionality remained undisturbed, keeping fundamental molecular functions intact. This did not imply that the environment did not affect the adaptive function of the mitochondrial genome but that the neutral variation was a common and well described phenomenon for the mitochondrial genome. Further, the co-occurrence analysis of

the high-altitude markers revealed the presence of intra-genic constraint in the high-altitude population. This suggested that the presence of particular variation was not sufficient for adaptation, but that variation had to be assisted by other variations in the same gene or same gene complex. Here we had analyzed only mt-DNA but it was likely that co-occurrence occurred between mt-DNA and the nuclear genomes as the role of nuclear variants in adaptation to high altitudes was well described [94, 111]. Further studies are necessary to identify these interactions between the two genomes. Variable site 711 of 12S rRNA gene was found to co-occur with both the markers 3394 and 7697. The variants 711C and 14417G defined the sub-haplogroups M9a1a1c1b1a, M9a1a1c1b1a1 and M9a1a1c1b1a2 which were widely distributed in East Asian and Southeast Asian populations, and prevalent in Tibetan population. Moreover, MT-RNR1 gene encoded for a protein responsible for regulating insulin sensitivity and metabolic homeostasis [96]. Particularly, the co-occurrence of 7697 with 14417 was observed in the group 7 and 8. Since, the biological effects of high altitude were observed at$>$3000 m existence of these co-occurrence pairs seemed a possible combination of variants that affects mitochondrial bioenergetics. Variant 4491 of *ND2* gene was found to be associated with high altitude pulmonary edema (HAPE) susceptible in low altitude population [97] whereas in our analysis, this variant was found to co-occur with 3394 and its exclusive presence in the higher altitude regions suggested its adaptive dependence. The variable sites forming the global connected nodes 9540 and 10873 were the RSRS50 ancestral variants [98] while 10400 and 16327 were markers of M sub-haplogroups [99]. The markers 9540 and 10873 were found to be present throughout the human population, 10400 was known to be specific to Asian population and 16327 was known to be the marker of C sub-haplogroup of M haplogroup which was specific to Siberian and American regions. The presence of 16327 in Asian population was not surprising since human migration in American continent took place through the Beringia bridge from Siberia [100, 101]. Phylogeny based study had shown that 10398 resulted through selective sweep at colder geographical regions [102] which was supposed to lower the oxidative phosphorylation coupling leading to release

of more heat. This possible tradeoff between ATP generation and thermogenesis seemed to play a key role in adaptation at colder higher altitudes. Substitution from A to G at 10398 corresponded 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 exclusively at higher altitudes suggested that a separate ancestral population might have colonized these regions. We found the existence of three variable sites belonging to the *CYB* gene which were co-occurring with global connected nodes; these variable sites are 14783, 15043 and 15301. Although we had collected the mt-DNA sequences for healthy individuals, these variants were predicted to be likely pathogenic and had been shown to be associated with Familial cancer of breast in ClinVar database. Through genegene interaction network, we found that *CYB* gene was co-occurring significantly with other genes at Tibetan region. This had shed light on its possible involvement in hypoxia and low temperature adaptation. Cytochrome b protein is an integral membrane protein subunit of the cytochrome bc1 complex encoded by *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, several investigators have suggested that functional modifications of redox proteins, such as cytochrome b, may be involved in physiological adaptation to different thermal environments [80, 81, 103]. It is interesting to note that variants located in *CYB* gene (such as C14766T) are known to be significantly higher in the high-altitude pulmonary edema a disorder arising due to acute exposure to high altitude above 3000 m and it has been hypothesized to play a role in high altitude sickness [97].

The formation of local co-occurrence pairs and similarity clustering divided the altitude groups into the higher and the lower altitude regions. This division might be possible due to the two reasons, (i) migration and demographic dynamics or (ii) 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, the two node motifs identified at high altitude$>$3000

m in the groups 7 and 8 can be suggested as candidate positions for a biological role in adaptation to these conditions. Overall, the co-occurrence network motifs provided detailed insight into finding the association of variable sites which are overlooked by haplogroup analysis alone and showed that selective pressures, such as high altitude, may generate constraints on the mitochondrial genomes, forcing the co-occurrence of certain variants on the mitochondrial genome.

Three high-altitude populations across the three continents

Chapter 3

Three high-altitude populations across the three continents

3.1 Introduction

In Chapter 2, we have presented the interplay of co-occurring variable sites in the form of two-order motifs and their impact on evolution at different altitude populations. The co-occurrence signature was prominently observed at genetic level and gene-gene interaction networks were shown to provide the genetic difference among all the altitudes. In this chapter, we focus on three different high-altitude populations (i) Andean Altiplano in South America, (ii) Qinghai-Tibetan Plateau in Asia, and (iii) the Ethiopian in Africa.

These three high-altitude populations can be viewed as an outcome of independent replications of a natural experiment of convergent evolution. In such cases, descendants of an ancestral founding population moved to high altitudes from relatively lower altitudes got the opportunity for natural selection due to exposure to high-altitude hypoxia for a sufficiently long time. Thus, it helps them to improve their physiological and genetic functions under hypoxic conditions [104]. The term

'convergent evolution' is defined as the development of the same or similar phenotypic adaptations under a common external environmental condition as a consequence of natural selection. Although, it could not be denied that the recent multiple migrations from corresponding lower-altitudes could affect the particular signature of high-altitude when compared with the respective low-altitude populations [105]. These three populations are believed to be evolved differently at the genetic [106] and physiological levels [107]. Various factors such as genetic pathways, molecular pathways, and phenotype levels have been attributed to the convergent evolution of humans and domestic animals [104]. 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 [108]. In Tibetan highlanders two nuclear genes of the HIF (hypoxia inducible factor) pathway, *HIF2A* and *PHD2* are known to be associated for positive selection [109]. 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 [110]. Retrospectively, introgression from Denisovan or Denisovan-related individuals has been suggested to be affecting the pattern of high-altitude adaptation in Tibetans [111]. In Andean highlanders, out of 40 genes exhibiting positive selection, the α-1 catalytic subunit of adenosine monophosphate-activated protein kinase (*PRKAA1*) gene has a significant role in high-altitude adaptation [112]. 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 [113]. 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 [113].

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 posses significant differences at physiological and genomic levels from their respective low-altitude populations [107, 114]. 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 [115]. Among these three populations, Andeans were found to have the highest hemoglobin concentration in their blood. Another physiological trait associated with high altitude is oxygen saturation. Tibetan individuals reported 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 4000m, moderate oxygen saturation, and lower arterial oxygen content than sea-level references values. The Ethiopian patterns of hemoglobin concentration, oxygen saturation, and arterial oxygen content were reported to be similar to those of low-altitude dwellers [107]. Apart from physiological differences, specific polymorphisms belong 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 [116]. In this chapter, we focus on analysing the possible role of mitochondrial co-mutations for these high-altitude populations in light of possible convergent evolution, using whole mitochondrial genomes under the networks framework. Foremost, we constructed the co-mutation networks by selecting significantly interacting variations of the mitochondrial genome. These networks were found to follow the small-world behavior

with the high modularity. The weak ties, nodes with a low degree and high betweenness centrality, were found only in the Tibetan network and were found to be haplogroup markers. Followed by that, a single 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.

3.2 Source of data and network construction

Complete human mitochondrial genome sequences were downloaded from the Human mitochondrial Database (HmtDB) [117] for the Ethiopia and Andes regions situated ∼3000m, and ∼3500m 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 (∼4000m) were downloaded from the GenBank. All the sequences were aligned globally and mapped with a master sequence rCRS (revised Cambridge Reference Sequence)[118].

3.2.1 Detection of Community and Role of nodes

By calculating the modularity, we detected the communities using the algorithm given in [40] which is implemented in Python using the *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 nodes 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,

$$Z_i = \frac{k_i - \bar{k_i}}{\sigma} \tag{3.1}$$

where k_i is degree of the i^{th} node in its own community, $\bar{k}$ is the average of k_i for all the nodes of that community, and σ 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,

$$P_i = 1 - \sum_{S=1}^{N_m} \left(\frac{k_i}{K_i} \right)^2 \tag{3.2}$$

K_i is the total degree of the node i in the whole network. S is the community and N_m 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.

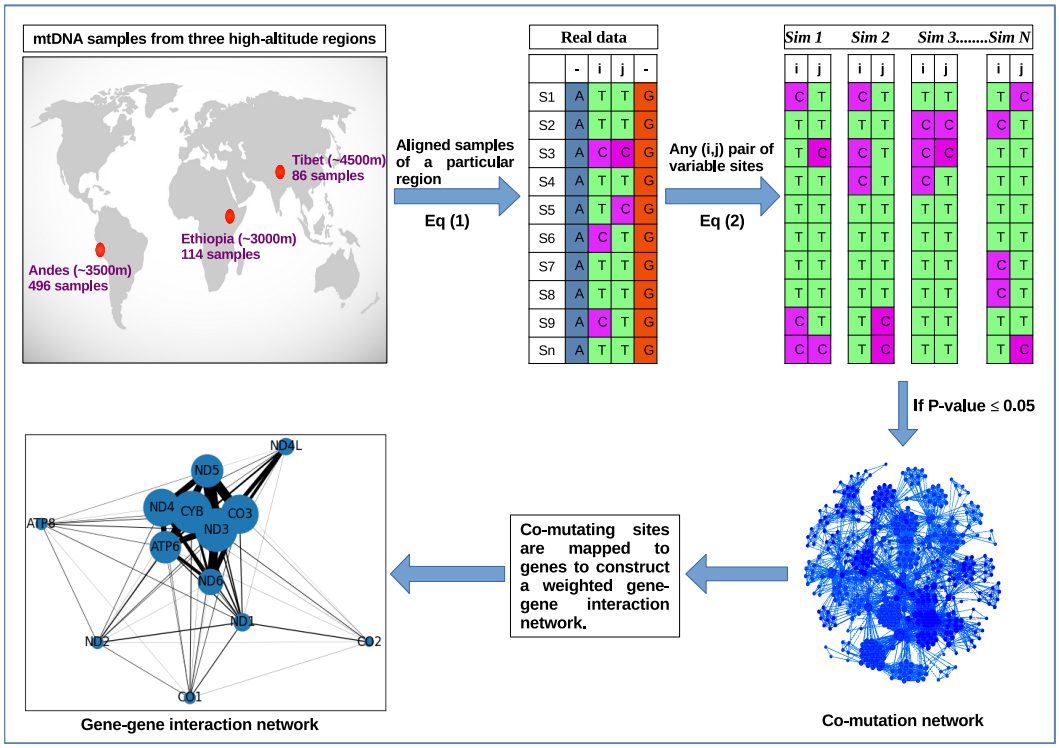

Figure 3.1: Construction of Co-mutation network and Gene-Gene Interaction (GGI) network for each high-altitude region.

3.3 Results and Discussion

3.3.1 Identification of significant interactions

Among the three regions, the Andes population had the highest number of samples, nodes and connections, (Table 3.1). It was observed that more samples rendered more interactions to be statistically significant when the number of connections (L) at $C_{ij} > 0$ between the Andes and Ethiopia was compared. Both these regions were having an almost equal number of connections before applying the threshold. When significant pairs (with $P_{ij} \leqslant 0.05$) were considered, the number of connections was decreased by $\sim$23% in Andes, $\sim$49% in Ethiopia and $\sim$53% in Tibet (Figure. 3.2a). It was further noted that above the threshold value, the number of co-mutations with lower C_{ij} values ($\leqslant 0.2$) were less, while co-mutations with high C_{ij} (> 0.2) values remained unaffected. This observation signifies the role of *p*-value in determining the considerable interactions. Further, the distribution of C_{ij} for Andes population portrayed a heterogeneous distribution of the variations within the samples, i.e., there exist fewer co-mutations for $C_{ij} > 0.8$, indicating that the minor alleles were not always present in the same sample(s). To get an overall idea for these three regions, we explored the common nodes and connections among these three co-mutation networks (Figure. 3.2c,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 $\sim$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.

Table 3.1: Co-mutation networks

Regions	Sample size	Nodes	Links
Tibet	86	398	3459
Ethiopia	119	838	13770
Andes	496	1197	20224

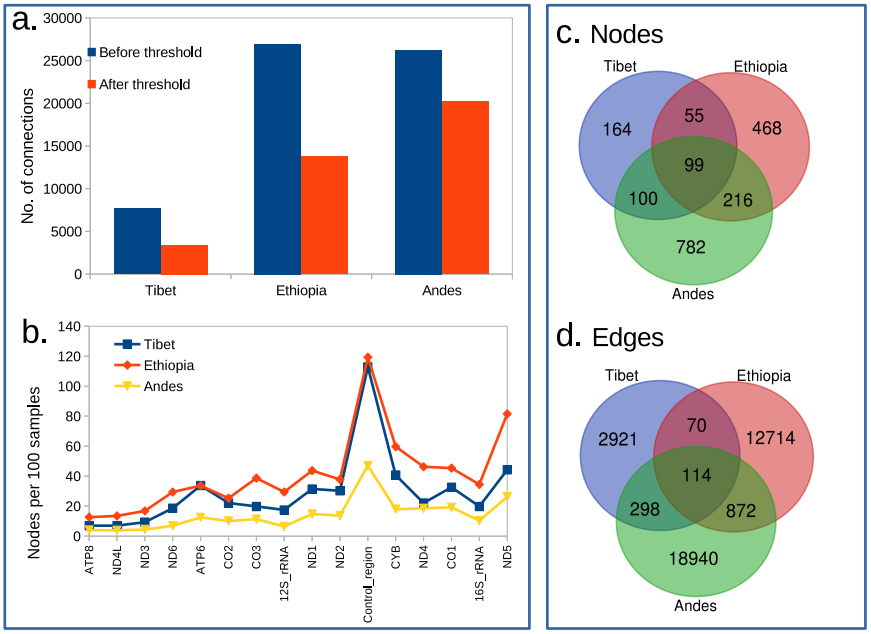

Figure 3.2: **(a)** The change in a number of connections with threshold **(b)** Nodes participating in network construction were mapped to their respective genes and genes were counted and plotted on the y-axis with their lengths on the x-axis. Note that t-RNA genes are not shown. **(c)** Distribution of nodes **(d)** and co-mutation pairs across all three regions.

3.3.2 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 that of 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[19]. 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 hub 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% of 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 commonly found in all three regions. This particular site was suggested to be a candidate signature for functional analyses, and data association [119].

Further, all the three networks were found to have small-world properties characterized by high clustering coefficient [120] $< C >_{real} / < C >_{rand} >> 1$ and small diameter $L_{real}/L_{rand} \sim 1$ as for many other real-world networks (Table 3.2) [121, 122]. 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. This provides evidence for the fixation and inheritance of variations as a single cohort, and intragenic constraints [123] in the mitochondrial genome in terms of co-mutation. A high $< C >$ also implies 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 (Figure 3.3, lower panel) was plotted. A decrease in the tendency of a variable site to form clusters with an increase in its degree implied the presence of hierarchy [124, 125] in these co-mutation networks.

Table 3.2: Global properties of co-mutation networks

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

The hierarchical organization confers robustness and adaptability in complex biological networks [126]. mtDNA has acquired several variations depending on biotic and abiotic factors since humans have first migrated outside Africa throughout the world [127]. 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.

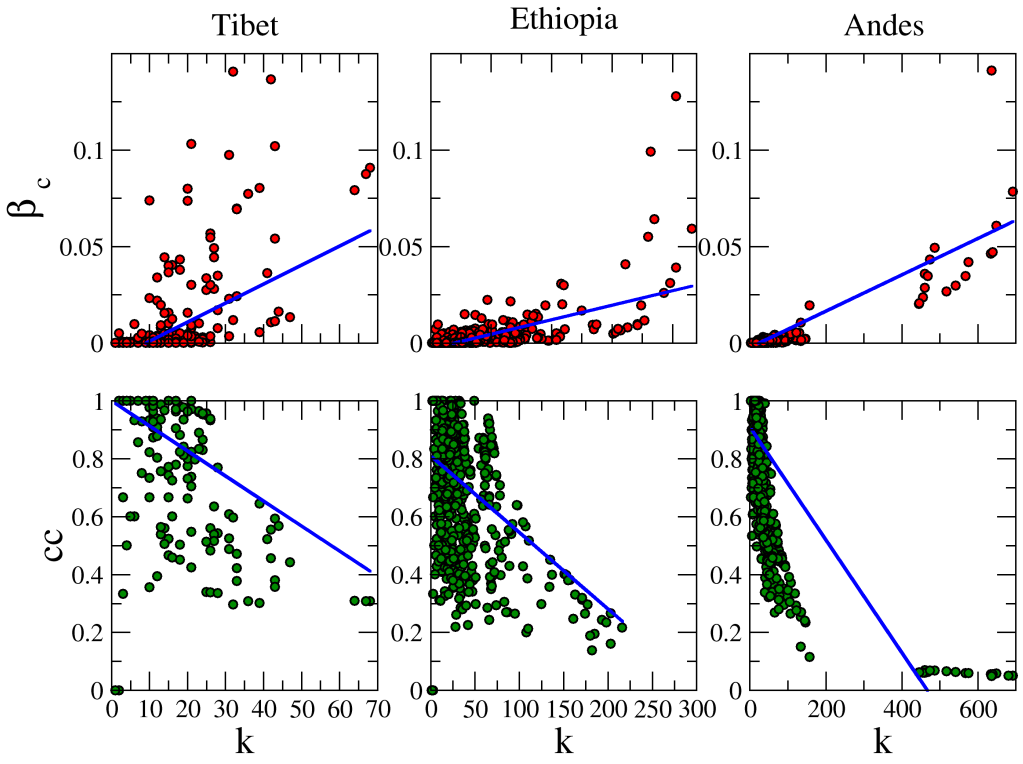

Figure 3.3: Betweenness centrality (upper panel) and clustering coefficient (lower panel) are plotted as a function of degree for all three regions.

Resilience is an important property for a network, which is measured by the

betweenness centrality (β_c)[58]. 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 is observed here that a few nodes, despite having a low degree, have high betweenness centrality and are considered weak ties. Weak ties are the nodes that co-mutate with a few nodes but from different modules. The presence of weak ties suggests that mtDNA has evolved through co-evolution of a few nodes (pertaining to low k) of multiple discrete modules (pertaining to high β_c)[128]. 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, 15927, 16172, and 16362) (Figure 3.3, upper panel). Interestingly, all these variable sites were also found to be haplogroup markers (709: L6, G, T, and W; 15927: G, B, and X: 16172: L0 and F; 16362: 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, 15927 node belongs to *tRNA-Thr* which is one of the highly mutated tRNAs among all the tRNAs in humans[129], 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 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 β_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 highest modularity in the Tibetan population than the other populations. 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 contributed more than other genes in each community in all the three co-mutation networks. 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 were common among the three regions. Although the three regions had a certain number of common nodes (Figure 3.2c), the community structures were 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[130]. 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 network applications; in disease spreading [131], perturbation of propagation in mutualistic networks [132]. Other applications of localization can also be found in [133, 134]. To quantify localization, we used correlation dimension (D_2) calculated using the box-counting method for multifractal analysis of eigenvectors [135]. If $D_2 \longrightarrow 0$, an eigenvector is localized while $D_2 \longrightarrow 1$, the eigenvector is considered delocalized. Thus, D_2 provides insight into the degree of localization of eigenvectors. We focused on the eigenvectors of eigenvalues nearer to zero and mean D_2 was 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 $D_2 \sim 0.43$ compared to the Ethiopian network with $D_2 \sim 0.65$. In these networks, D_2 captured the tendency of co-mutation in terms of localization.

A co-mutation occurs when minor alleles at any given two sites 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 [105] 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 [136, 137]. Based on the within-module degree and the participation coefficient, nodes were categorized as module hubs and non-hubs (Figure 3.4). The nodes with highest degree in the Tibetan co-mutation network were found in R3, non-hub connector category, while in the Ethiopian and Andean co-mutation networks, the nodes with highest degree 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*), 14668 (*ND6*) and 12361 (*ND5*). The variable site 3010 was shown to be a high-altitude marker in the Tibetan population [138] and also reported to form network motifs with variable sites 8414 and 14668[123] while, 12361 was shown to be associated with nonalcoholic fatty liver disease [139]. 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 the 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 showed both inter and intra-haplogroup co-evolution. This again provides evidence for recent admixture in Ethiopian region [105].

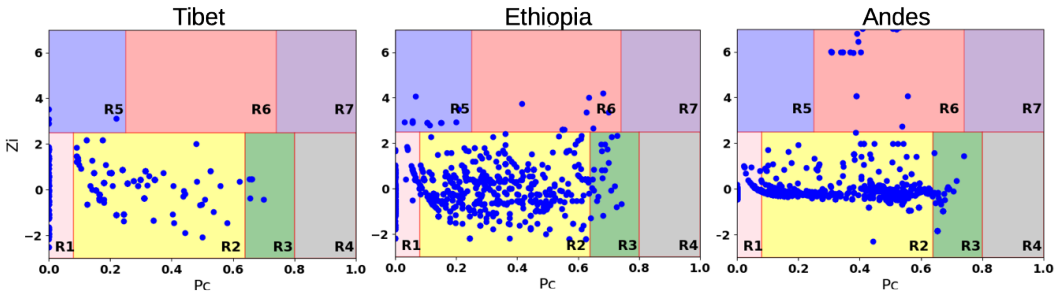

Figure 3.4: Roles of nodes in ZP parameter space. Each node in a network can be characterized by its within-module degree and its participation coefficient. Nodes with Z 2.5 were classified as module hubs and nodes with Z < 2.5 as non-hubs. Non-hub nodes can be naturally assigned into four different 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.

3.4 Gene-gene interactions

These co-mutation networks are then analyzed at the gene level through the construction of the gene-gene interaction networks, discussed further. The contribution of each gene was quantified by counting the number of variable sites from each gene (Figure 3.2b). 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 (Figure 3.2b). It is 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 contributing 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 any bias due to the high mutation rate. It was evident that specific genes contributed more variable sites in the network construction than others in a particular region. Especially, *ATP6, CO2* and *ND2* genes were contributing equally in the Tibet and Ethiopia networks while *12S-rRNA, 16S-rRNA, CO3* and *ND4* are contributing equally in the Tibet and the Andes 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 (Figure 3.2b). 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 most extensive participation in the network construction because it is the highly variable part of mtDNA [140]. 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 reductionist approach to investigating interactions in the three high-altitude regions. After comparing with the corresponding random networks, we identify 17 gene-pairs in Tibetan, 23 gene-pairs in Ethiopian, and 44 gene-pairs in the Andes population. Among these, the pair with the highest edeg weight *ATP6-CYB* (Tibet), *ND4-ND5* (Ethiopia), and *CYB-ND4* (Andes). Four significant gene-gene pairs were 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) [141], along with that these genes are also found to be involved in cellular respiration (GO:0045333), and response to abiotic stimulus (GO:0009628) [142]. 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 [114]. Since these three populations are believed to share a similar physical environment and to undergo the process of convergent evolution, for all the com-

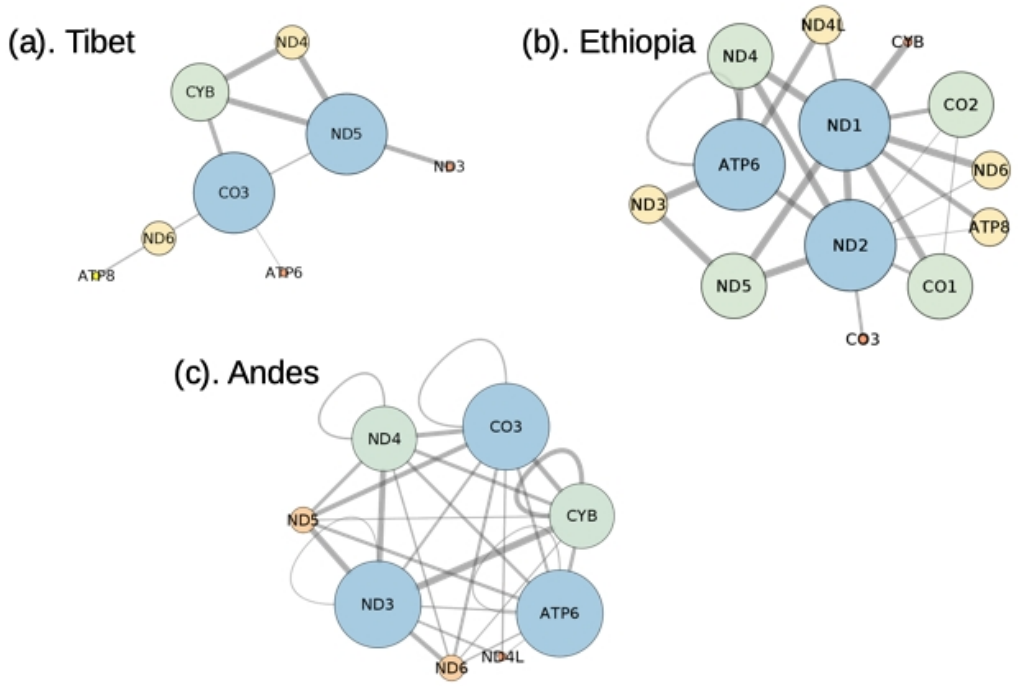

Figure 3.5: Significant gene-gene interactions of common node co-mutations and corresponding GO terms and KEGG pathway. (The node size depicts the degree of the node and edge size represents the edge weight.)

mon 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 interactions (Figure 3.5) of the common nodes were extracted based on the functional enrichment analysis for GO terms and KEGG pathways using DAVID [143]. 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 were shown in table 3.3. It was 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 reported as the putative site for the entry of oxygen into the large cytochrome oxidase complex, thereby regulating its activity

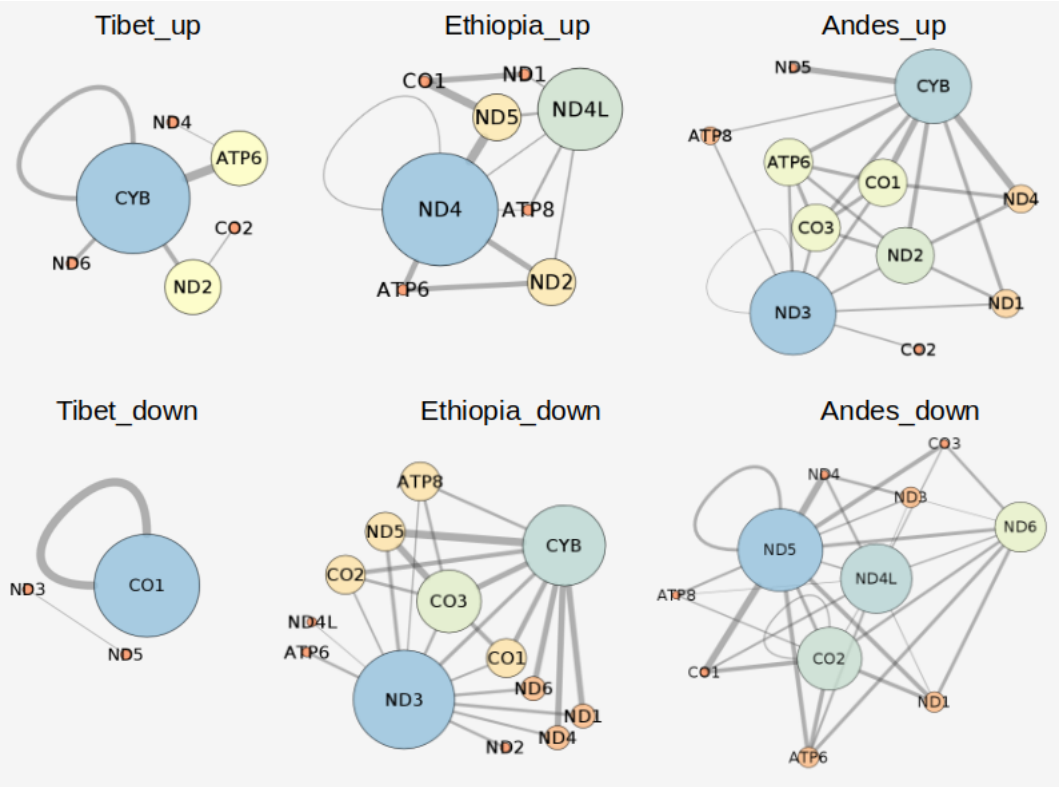

Figure 3.6: Significant gene-gene interactions of exclusive node co-mutations. (The node size depicts the degree of the node and edge size represents the edge weight.)

under hypoxic conditions [144]. 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 associated with high-altitude hypoxia in Tibetan yak [145] and endemic Ethiopian rats [146]. 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 [147]. Further to explore the population-specific role of such genetic interactions, 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 (Figure 3.6). It is readily observed from Figure 3.6 that those genetic interactions that were significantly up in the Andean and Tibetan populations were significantly down in the Ethiopian population and vice versa.

Table 3.3: Functional enrichment of gene sets for three regions

Region	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

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; Ethiopia is situated at the lowest altitude among all the three populations, and the second is that Ethiopia is believed to have undergone frequent admixtures in its gene pool from a lower altitude populations[148].

3.5 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. Analysis performed in this chapter showed that mtDNA has evolved with similar biological mechanisms in Andean and Tibetan populations than the Ethiopian population. It was found that there exists 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 pop-

ulations. 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 gene sets which might 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.

Higher-order interactions among mitochondrial genes

Chapter 4

Higher-order interactions among mitochondrial genes

4.1 Introduction

The understanding of a complex system comes with the study of patterns of interactions among its components. Networks provide us with a statistical approach to model complex systems [1]. Generally, the networks are studied in a statistical landscape through pair-wise interactions among their entities. In biological systems, such as cellular activities, biomolecules interact to perform many functions which includes protein-protein interactions [2], gene-transcription factor interactions [3], metabolic interactions [4], and so on. This drives biologists to explore further the structure and dynamics of these complex sub-cellular interactions facilitating the structural and evolutionary functionality of a cell [5]. The analysis of complex biological systems through the networks has assisted biologists in many fundamental ways in deducing biomarkers and designing new experiments. Such a network analysis under the lens of network theory has assisted in deducing biomarkers and designing new experiments [157, 158]. With an advent in network theory concepts,

many other details of real-world systems, such as edge weights emphasising on relative importance of interactions, directed edges [159] indicating flow of signal through an edge [160], multi-layer structure [161, 162], and evolution of network properties with time aka temporal networks [163, 164] got incorporated while constructing the corresponding model network. All these studies described complex systems through pairs of interacting units (pair-wise interactions) and demonstrated that properties and evolution of real-world complex systems indeed can be modeled as a net of interacting units. Nevertheless, quest for more accurate representations of properties of real-world interactions is continuing revealing many exciting facets, particularly, as in real-world systems, the interactions occur in much more detailed fashion. Particularly, many factors affect the very nature of interactions themselves. One of the most exciting revelation of these pursuits is discovery of higher-order interactions, which emphasize existence and importance of beyond pair-wise interactions in real-world complex systems. There are pieces of evidence that show the significance of higher-order interactions in many real-world complex systems such as ecological [165], biological [166], neuronal [167], and social systems [168]. In ecological networks, more than two species interact together and affect the ecological dynamics of the ecosystem. In collaboration networks, more than two authors appear in the same article [169]. In epistatic networks, more than two variable sites interact/co-occur for phenotypic benefits of the species, and reactions involve more than two biochemicals to perform a cellular function. The prediction of critical genes or nodes based on graphs assume that two adjacent nodes have similar functional contribution where we miss the information of other nodes contributing in that similar module (functional module). The representation of networks as hypergraphs is a spontaneous way of overcoming this assumption.

Thus, a $k-$simplex set out simultaneous interactions between $k+1$ nodes forming a set $I = \{v_1, v_2, v_3, \ldots, v_{k+1}\}$. Hence, $1-$simplex corresponds to $I = \{v_1, v_2\}$ depicting pair-wise interactions, $2-$simplex represent triangles with $I = \{v_1, v_2, v_3\}$ and so on.

Hypergraphs are powerful tools to extend the studies on ordinary graphs to

higher-order structures [170]. The hypergraphs differ from simplicial networks in terms of the presence of lower-order interactions. For example, a hypergraph may not contain a dyadic interaction, whereas a simplicial network contains all sub-simplexes. The order of a simplicial network is defined based on the presence of the highest order in it. Hypergraphs [171, 172] and simplicial complexes [173, 174] are two most often used frameworks to examine higher-order interactions. A hypergraph connects d nodes simultaneously with a hyperedge where $d \geq 2$. Not all the hyperedges need to have the same size. However, a hypergraph is referred to as a d-uniform hypergraph if all hyperedge connects the same number of nodes d. Here, we consider a 3-uniform hypergraph where each hyperedge connects three nodes simultaneously.

First ever analysis, to our knowledge, of the higher-order interactions, i.e. hypergraphs of protein complexes demonstrated an existence of scale-free degree distribution of both the nodes and the hyperedges [175, 176]. Additionally, signaling pathways [177] and protein interactions [178, 179] have been studied under the umbrella of higher-order interactions.

Furthermore, interactions between genetic variations (epistasis) and the genetic background are known to alter the phenotypic landscape of the respective population leading to various studies of mutations between pairs of the genes [180, 181]. All these investigations are comprised of only pair-wise interactions, and hence limit our understanding of genetic evolution [166]. The presence of a third mutation has been displayed to alter the fitness and also modify the way pair-wise interactions occur [182, 183]. The evidence of possible higher-order interactions was found in the crossing of two yeast strains for 46 different phenotypes while searching of missing heritability [184]. This study, identified few phenotypes that were not explained by two-loci interactions. In a large-scale study involving 500 genotypes of yeast tRNA and over 45000 interactions, it was revealed abundance of higher-order interactions which were responsible for dynamic across different genotypes [185]. In another study, higher order interaction of five different genes was found to affect the complex phenotype in *Saccharomyces cerevisiae* [186]. This and few other

studies involve the investigation of the presence and absence of mutations/variations individually or together and their effect on the fitness or morphology of the organism. However, Ref. [187] has quantified the digenic and transgenic interactions with respect to the fitness of the population by introducing a transgenic score, τ-SGA score. This τ-SGA score combines double and triple mutant fitness extracted from the experimental results. Further, the word co-occurrence hypergraphs have been studied to explore the conceptual landscape of mathematics [188].

All these investigations provide sufficient evidences of existence of higher-order interactions in complex systems and moreover highlight importance of such interactions and corresponding hypergraphs in evolution and functionality of underlying complex systems. In this chapter we investigated the co-mutation hypergraphs in the human mitochondrial genomes for three different altitude populations, low-altitude (0-500m), middle-altitude (2001-2500m), and high-altitude (>4000m). We consider 3-uniform hypergraph based on co-mutation frequency defined for three variable sites altogether for all the variable sites. Thereafter, we use a two-step threshold selection method to construct the unweighted d-regular hypergraph.

Note that in the following text, triangles and hyperedges are used interchangeably as we are considering 3-uniform hypergraphs only, where each hyperedge contains exactly the three nodes. By going one step forward form traditional pair-wise interactions to 2-order interactions, we specifically calculated the co-mutation frequency of three variable sites. A two-step threshold selection was used to filter significant hyperedges, one using a statistical score (P_{ijk}), and second using few specific network properties. Thereafter, to identify the significant genes, we calculated the weight of the each triangle, followed by the mapping of genes, and analyzed the genes involved in higher-order interactions for each altitude group. In the present chapter, we sought to investigate the extent to which the presence of higher-order interactions among the mitochondrial genes in three different altitude populations, low-altitude (0-500m), middle-altitude (2001-2500m), and high-altitude (>4000m) affect the overall behavior of mitochondrial genome according to demography and environmental constraint. We go one step forward from traditional pair-wise inter-

actions to 2-order interactions by explicitly calculating the co-mutation frequency of three variable sites. We define the higher-order interactions in terms of 2-order simplices for all possible three variable sites combinations where the order is irrelevant. A two-step threshold selection was used to filter significant simplices, one using a statistical score (P_{ijk}), and the second using network properties. Along with threshold selections to filter insignificant simplices, we also incorporated the information of genes in each simplex to define the higher-order interactions. In comparison with pair-wise interactions, we identified that there are certain higher-order interactions that exist in these genomes. In order to identify the significant genes, followed by mapping of genes, we calculated the weight of each two-order simplex. It was found that there are relevant genes involved in higher-order interactions for each altitude group.

Materials and Methods

mtDNA sequences were collected from [123] for three altitude groups, lowest (0-500m), middle (2001-2500m) and highest (4001m). The ambiguous nucleotides were replaced by the default letter 'N' for thee computational calculations. All the sequences were aligned together using Clustal Omega and then mapped to rCRS for gene annotation.

Construction of higher-order Gene-Gene Interaction (GGI) networks:

Two types of networks constructed, the co-mutation networks where nodes were variable sites and, the weighted GGI networks where nodes were genes (Figure 4.1) for each of the high-altitude population.

Step 1 (Co-mutation Network): Any position having more than one allele in 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 altitude group, nodes were represented by the position of variable sites, and the edges were represented by co-mutation frequency between three pairs of the nodes (C_{ijk}) defined as,

$$C_{ijk} = \frac{(m_{ijk})^3}{m_i m_j m_k} \tag{4.1}$$

where (m_{ijk}) represents number of times the minor alleles occur together at i^{th}, j^{th} and k^{th} positions, m_i, m_j and m_k indicate total number of times the minor allele occurs at i^{th}, j^{th} and k^{th} positions, respectively.

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

$$P_{ijk} = \frac{\#[(C^r_{ijk}) \geqslant (C_{ijk})]}{10^4} \tag{4.2}$$

where, (C^r_{ijk}) is the co-mutation frequency calculated after permuting the alleles at the i^{th}, j^{th} and k^{th} positions randomly. 10,000 random simulations were generated and P_{ijk} was set to $\leqslant 0.05$ (standard p-value) to consider a co-mutation triangle significant.

Step 5 (Higher-order and pair-wise interactions selection): To define a triangle as filled or unfilled, we map the variable sites in each triangle (i, j and k) to genes and if all the genes are different, then we consider it as filled triangle (true triangle).

Step 6 (Co-mutation threshold selection): To select a co-mutation threshold for filled triangles, we plotted the size of largest connected component (lcc) with the C_{ijk}. A threshold was selected based on the change in size of the lcc.

Relative weight of genes:

A relative weight of each gene was calculated for all the triangles as,

$$R_g = \frac{\sum_{s=1}^{N_s} W_g}{N_s} \tag{4.3}$$

where, R_g is relative weight of gene g, s represents gene triangle, W_g is the weight of triangle in which gene g is participating, and N_s is the total number of gene triangles having gene g.

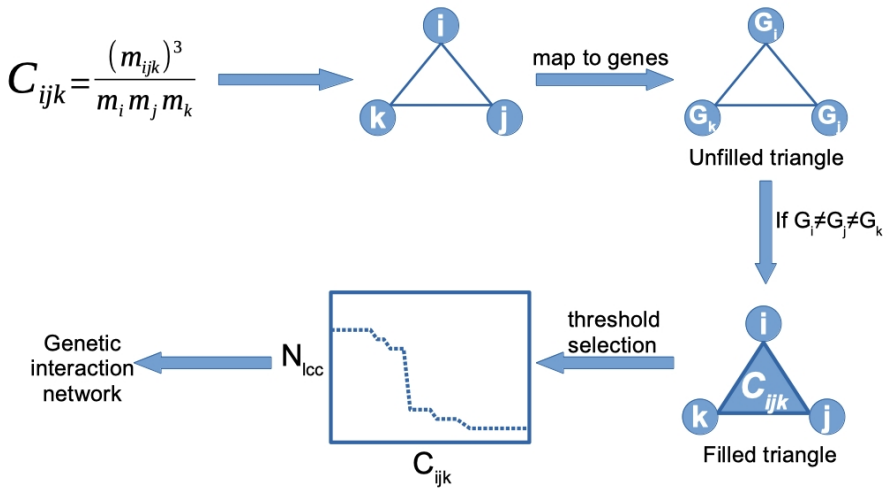

Figure 4.1: The schematic representation for constructing and considering higher-order interactions from co-mutations for genetic interaction networks. Note that the, C_{ijk} is used to determine the threshold for higher-order interactions and the weight of genetic interaction network is determined only by the number of interactions between genes.

Table 4.1: Number of triangles (i, j, k) before and after applying the threshold (shown in bracket). The number of gene simplices are shown without the Control region.

	Low altitude (C_{th} =0.334)	Mid altitude (C_{th} =0.500)	High altitude (C_{th} =0.251)
Total triangles (i, j, k)	81656	62390	74398
True triangles (i, j, k)	8758	6141	6671
True gene triangles (G_i, G_j, G_k)	808	660	746

4.2 Results

4.2.1 Characteristics of hyperedges

Initially, after applying the first step filtration using P_{ijk}, the co-mutation hypergraphs were quite dense, therefore a co-mutation threshold (C_{th}) was applied to make the network sparse. To identify C_{th}, we plotted the size of LCC against C_{ijk} and a C_{th} was selected such that the size of LCC was very small for a higher C_{ijk} (Figure 4.3). This threshold selection method has been defined as network efficiency score [154]. We started with ∼450 nodes in each altitude group with a potential of 0.1 million combinations, and we ended up with a very small number of total tri-

angles and true triangles (Table 4.1). For a given hyperedge, the C_{ijk} is subjected to the minimal co-mutation frequency value for all three triangle pairs. Hence, C_{ijk} is not a linear combination of pair-wise co-mutation frequencies for all the pairs of i, j, k variable sites. After getting the sparse network, we defined the true triangles such as no variable sites in any hyperedge belong to the same gene for all the hyperedges. As we know that at the genetic level, such a pair of variable sites where both the nodes belong to the same gene would give rise to a self-loop and also transform the triangle into a weighted edge. Hence such connections were also removed. In low-altitude, mid-altitude, and high-altitude $\sim$10% of all the triangles were found to be true triangles (Table 4.1).

4.2.2 Variable sites and haplogroups

From this small number of simplices, we looked for the variable sites (nodes) contributing to forming simplices in terms of 2-order degree (Table 4.2). In the low-altitude group, node 1598 represented M-haplogroup (M17, M28, M29, M30, M42, M52, B5 and N1). In the mid-altitude group C-haplogroup represented by 14318 (C4) and 15204 (C1, C4, C5 and C7), and in the high-altitude group, K-Haplogroup is represented by 9055 (Z3, K1, K2 and U8), 9698 (K1, K2 and U8), 10550 (K1 and K2) and 11299 (L0, K1, K2 and Y2). It is noted that in low-altitude, only one high-degree node was represented as a haplogroup marker. In contrast, most high-degree nodes in mid and high altitudes represented a few haplogroups. It is to note that despite having very low variable sites in low-altitude, tRNA genes, Trp, Tyr, Gly and Ser are contributing to the highest node degree. When the difference of pair-wise and triangle degrees was plotted, we found that there were certain variable sites with large differences compared to most of the nodes. This suggests that a few variable sites show a higher tendency to form higher-order interactions (Fig. 4.2 upper panel)

4.2.3 Overlapping edges and associated genes

Next, we checked for the edge degree of each hypergraph. As these are 3-uniform hypergraphs, the edge degree was defined as the number of triangles a given edge

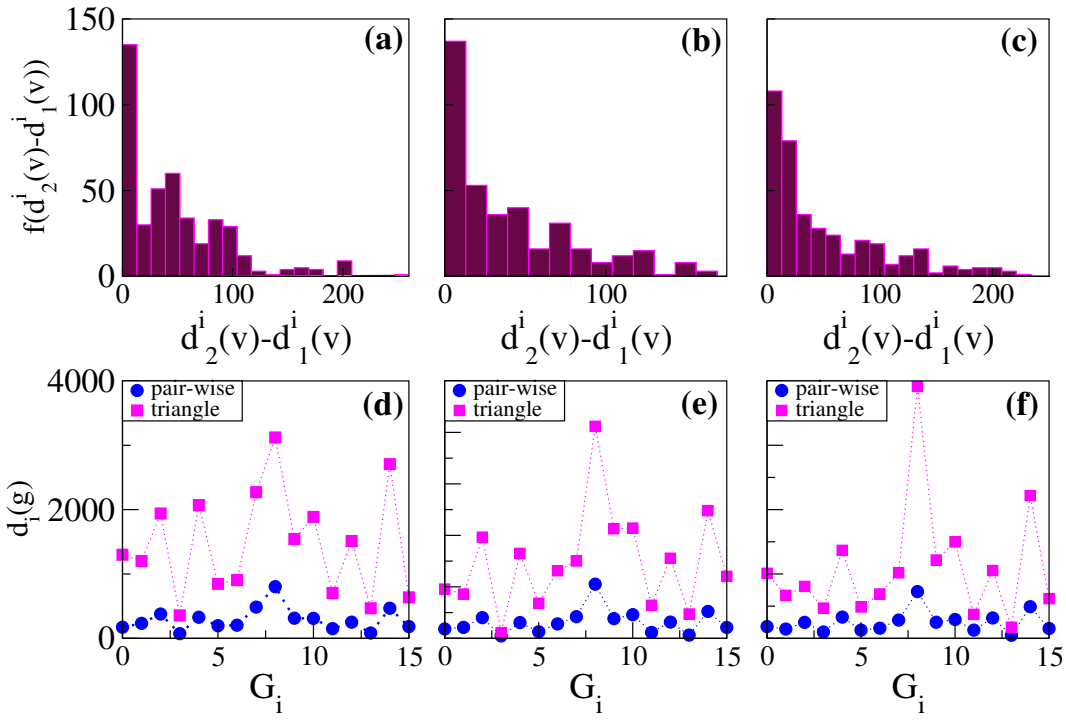

Figure 4.2: The distribution of difference of degrees between 2-uniform ($d_1^i(v)$) and 3-uniform ($d_2^i(v)$) variable sites for (a) low-altitude, (b) mid-altitude and (c) high-altitude (upper panel). The 1-simplex and 2-simplex degrees ($k_i(g)$) of each i^{th} gene (G_i) are plotted for (d) low-altitude, (e) mid-altitude and (f) high-altitude. The 3-uniform degrees were found to be comparatively higher than the 2-uniform degrees at all the altitudes (lower panel).

is part of. After calculating the edge degree, we mapped the edges (both nodes) to genes, giving us a gene-pair for each edge. As more than one edge can belong to the same gene-pair, here we get the gene-pair weight by adding the edge degree for all edges of a given gene-pair. For instance, edges (1,2) and (3,4) are part of two triangles each. So, the edge degree of both these edges will equal 2. Now, mapping these to genes, both edges belong to the same gene-pair, $G1-G2$. Hence, the gene-pair weight will equal the sum of both the edge degrees, i.e., four. This provided us the information of overlapping edges between overlapping triangles. In all the three altitude groups, *Control region* was commonly found in hyperedges with the highest weights. Particularly, in low and mid-altitude, it formed the highest weight hyperedge with *ND5* gene and in high-altitude with *CYB* gene. Apart from *Control*

region, in low-altitude *ND5* formed highest weight hyperedge each with *ND1, ND2*, in mid-altitude *ND5* formed highest weight hyperedge each with *CO1, ND1, ND4*, and in high-altitude *CYB* formed highest weight hyperedge each with *ATP6, ND2, ND5*. In summary, in low-altitude ND, genes yield overlapping hyperedges. In high-altitude CYB, ATP6 and ND genes yield overlapping hyperedges. The CYB and ATP6 genes have already been established to play a significant role in high-altitude adaptation [189].

4.2.4 Common and exclusive triangles

Since there are few variable sites common and exclusive among all and between altitudes, we investigated how these variable sites involve forming triangles by identifying the presence of shared and exclusive true triangles. Note that a variable site common to two altitudes could be part of two different triangles. Hence common variable sites are not necessarily part of common triangles. We found 181 common triangles among all the altitude groups, 123 common triangles between low and mid altitudes, 169 between low and high altitudes, and 612 between mid and high altitudes. A large proportion of triangles were found to be exclusive to each altitude group. We were interested in identifying the genes involved in forming triangles among these common and exclusive triangles for each altitude. In order to attain this objective, we mapped these triangles to corresponding genes to get weighted gene triangles. The triangles with the highest weight had {*CYB* and *ND3*} genes present ubiquitously among common triangles of all the altitudes. Among common triangles of low and middle altitudes, {*ND2* and *ND4*} genes, and among common triangles of middle and high altitudes, {*CO1* and *ND5*} genes were present ubiquitously in gene triangles with highest weights. To extend this in terms of genomics, the interaction between two particular genes is significant in forming triangles with a third different gene. This provides evidence and strengthens the presence of particular pair-wise interactions among mitochondrial genes.

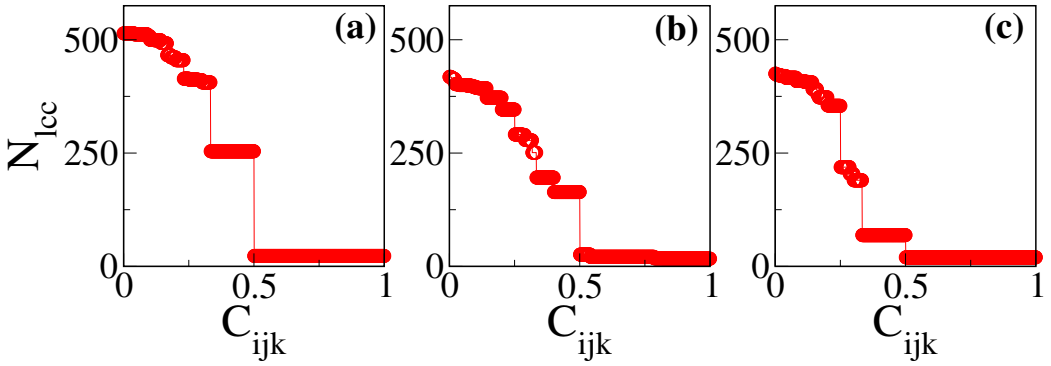

Figure 4.3: The change in size of largest connected component is shown with respect to C_{ijk} for (a) low-ltitude, (b) mid-altitude, and (c) high-altitude.

Table 4.2: The nodes with highest contribution in each altitude group.

Altitude	Highest 2-order degree nodes (with corresponding genes in bracket)
Low altitude	1598 (12S_rRNA), 4734 (ND2), 5557 (Trp), 5836 (Tyr), 7268 (CO1), 7490 (Ser), 10007 (Gly), 10631 (ND4L), 15307 (CYB)
Mid altitude	10304 (ND3), 14318 (ND6), 15204 (CYB)
High altitude	3480 (ND1), 7229 (CO1), 9055 (ATP6), 9698 (CO3), 10550 (ND4L), 11299 (ND4)

4.2.5 Codon positions in hyperedges

A variable site also possesses information about codon position (CP) in coding genes. The relation between co-evolving variable sites and codon positions has been described previously in the human population using two, and three-order motifs [69]. For a coding gene, the codon positions were set based on nucleotide position in codons 1, 2, and 3. For non-coding genes, all the nucleotide positions were set to 0. It was detected that CP 1 and 2 are highly conserved in forming the codon triangles, whereas CP 3 is versatile in forming these hyperedges (Figure 4.4). The CP 3 predominates the formation of hyperedges with other coding and non-coding positions. The third codon position is considered to be weakly responsible for amino acid selection during protein synthesis, hence supports the Wooble hypothesis and is not subjected to evolutionary constraint [190]. The codon-conservation has been observed in hypervariable regions of conopeptides revealing their accelerated evolution [191]. However, the hyperedges with all sites being CP 3 are also relatively less

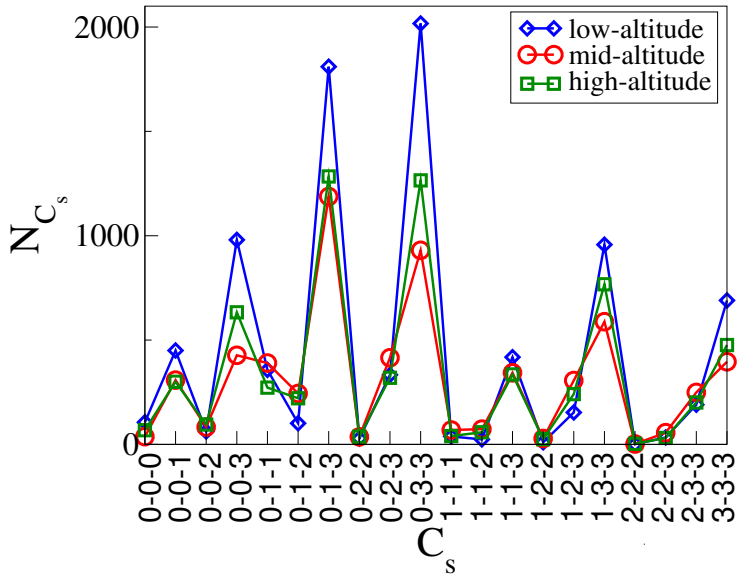

Figure 4.4: Distribution of codon simplex in different altitude groups.

abundant than other hyperedges with one or two CPs occupied by coding and non-coding positions. This suggests that the third CP is also subject to conformational stability provided by other CPs. Among all, CP 2 is the most conserved position, which supports the functional stability of amino acid selection during protein synthesis. Similarly, the hyperedges with all CP 0 were also found to be relatively less. Since these do not code for any proteins, the hyperedges represent the intra-Control region and intra-RNA genes. This suggests that non-coding regions least favor the formation of higher-order hyperedges.

4.2.6 Gene triangles

After identifying the true triangles, we calculated the actual number of true gene triangles by summing over their occurrence in the network, thereby calculating their weights. If we exclude the *Control region*, out of all the possible true gene triangles, only $\sim$10% were identified as true gene triangles (Table 4.1) in all the altitudes. The *Control region* is known to be the highly mutating region of the mitochondrial genome. Due to this, we excluded *Control region* to look at the triangles of coding and non-coding genes explicitly. We plotted the distribution of weights of all the

gene triangles and found that it follows power-law distribution (Figure 4.5). This demonstrates that the mitochondrial genes prefer forming particular triangles over others due to the conserved co-mutation patterns formed by variable sites. After

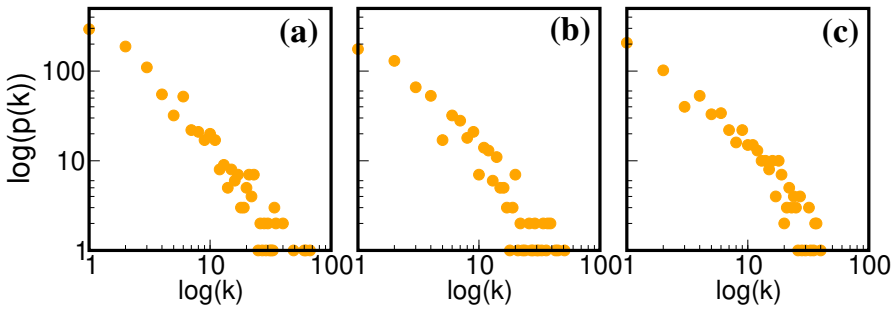

Figure 4.5: The distribution of weights of gene simplices (a) low-altitude, (b) mid-altitude, and (c) high-altitude.

selecting a specific C_{ijk}, all the variables sites of each triangle are mapped to corresponding genes, and the contribution of each gene is calculated as a relative weight for each altitude group (Figure 4.6). The relative weight of each gene is defined based on the number and weight of the hyperedges for each gene. Note that for this calculation *Control region* variables sites are not considered. *Control region* is a highly mutating region of mtDNA and hence, occurs in all the groups equally. In the high-altitude group, *CYB* gene is showing distinctly high relative weight among coding genes and *tRNA-Gln* and *tRNA-Met* among non-coding genes (Fig. 4.6). In low and mid altitudes, none of the coding genes showed distinctly high relative weight; however, in the mid-altitude group, *tRNA-Ile*, *tRNA-Arg* and *tRNA-Leu*, and in the low-altitude group, *tRNA-Phe*, *tRNA-Ser* and *tRNA-Tyr* showed distinct relative weights among non-coding genes (Fig. 4.6). It is to note that in low altitude, *tRNA-Phe* showed distinctly high weight among non-coding genes; however, it was not found to be a high degree node.

4.2.7 Gene hyperedges categories

Furthermore, we classified gene hyperedges into four categories based on the information of genes. These categories are, *all-coding* (with no non-coding gene),

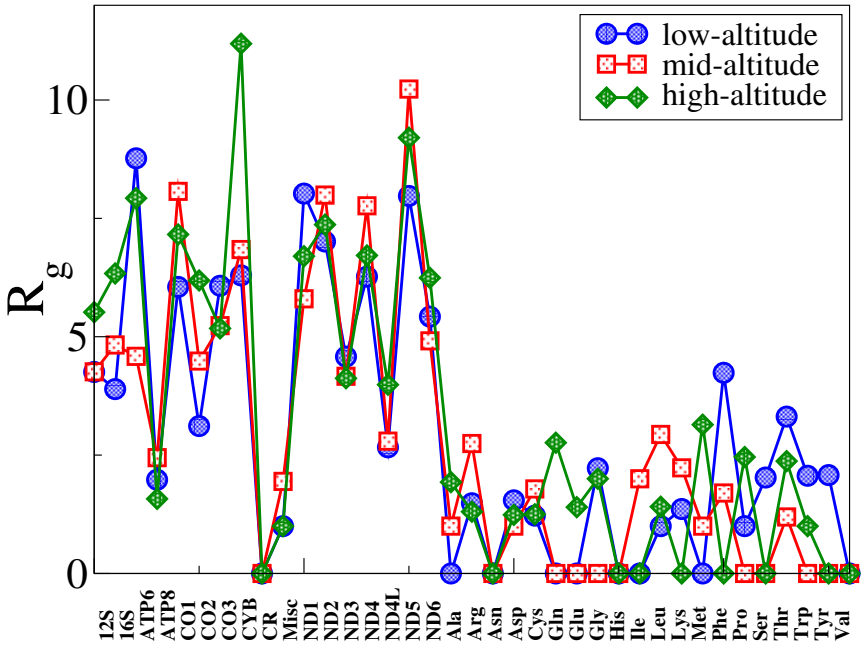

Figure 4.6: Relative weight of each gene for all three altitude groups.

2-coding (with two coding genes), *1-coding* (with one coding gene) and *0-coding* (with no coding genes). We found that ∼50% of gene triangles were from *2-coding* category and ∼1% from *all-coding* category in all altitude groups. *all-coding and 1-coding* were present equally around 25%. The relatively high percentage of 2-coding type in all the altitudes suggests that the presence of one non-coding gene favors the formation of higher-order interactions in the mitochondrial genome in general, independent of environmental conditions. However, the gene triangles with high weight are different for different altitudes. The non-coding gene in the 2-coding category with high weight was found to be *12SrRNA* in low altitude and *16SrRNA* in both middle and high altitude.

Primarily, we were interested in specific gene triangles having considerable importance in each altitude. To achieve this, we looked for gene triangles having high weights (Table 4.3). From the table, it is clear that low altitude population *ATP6 and ND* genes, in middle altitude population *CO1 and ND* genes, and in high altitude population *ATP6, CO1, CYB, and ND* genes contribute in forming significant

Table 4.3: Gene triangles with distinct weights

Altitude	Gene triangles with weight
Low altitude (0-500m)	ND1-ND2-ND5: 67 ATP6-ND2-ND5: 62 ATP6-ND1-ND5: 60
Mid altitude (2001-2500m)	CO1-ND4-ND5: 51 CO1-ND1-ND5: 45 ND2-ND4-ND5: 43
High altitude (>4001m)	CO1-ND1-ND5: 40 ATP6-CYB-ND2: 37 CYB-ND1-ND5: 37 ATP6-CO1-CYB: 36 ATP6-CYB-ND5: 36

higher-order interactions. In our previous study, where perfectly co-occurring pairwise interactions were studied, we found a similar set of genes with some additional genes for these different altitude groups. However, the contributing variable sites forming gene triangles differ from this study. This suggests that the formation of higher-order simplices is facilitated by a different set of variable sites but might be capturing the genetic interactions according to their role in environmental adaptability. An important observation is that in each gene triangle, one of the three 1-order simplices shows the considerable weight (discussed previously as edge weight), which is again observed when we looked for the variable sites contributing to high-weight gene triangles. In the low-altitude group, *ND1, ND5* genes or *ATP6, ND5* genes gave rise to variable sites with the highest contribution; in the mid-altitude, *CO1, ND5* or *ND2, ND5* genes, and in the high-altitude, *CO1, ND1* or *ATP6, CYB* or *ND5, CYB* or *ATP6, CO1* or *ATP6, CYB* genes in one or the other gene triangles gave rise to variable sites with the highest contribution.

4.3 Conclusion

In this chapter, we studied the co-mutation of variable sites through a framework of higher-order interactions. Here, we proposed a new method for constructing the hyperedges of variable sites by defining the co-mutation frequency between three

variable sites. These hyperedges were first filtered through the statistical score and second through gene-based information. We superimposed the variable site based hyperedges onto genes to realize the gene triangles finally. In the mitochondrial genome, $\sim$10% higher-order interactions were found to be true triangles. In terms of vertex degree, these hyperedges captured distinct haplogroups in different altitude populations. In low-altitude, it captures M-haplogroup, in mid-altitude, C-haplogroup, and in high-altitude, K haplogroup. Moreover, codon-based simplices provided evidence about the codon bias and conservation of codon usage throughout the mitochondrial genome for all the altitude groups. Based on gene triangles, we found that in the low-altitude group, *ATP6 and ND* genes, in the mid-altitude group, *CO1 and ND5* genes, and in the high-altitude group, *CYB and ND5* genes are predominantly forming higher-order simplices. The analysis presented here can be extended further to generalize all the higher orders rather than just 2-orders to get deep insights into higher-order interactions and identify the simplices-based communities in the mitochondrial genome.

Conclusion and future scope

Chapter 5

Conclusion and future scope

5.1 Conclusion

In this dissertation, we focus on understanding the mutual role of mitochondrial variable sites through co-occurrence and co-mutation networks for different human populations. Using the variables sites of mitochondrial genomic sequences under the framework of networks, we were able to capture phenomena of genomic evolution and pattern of polymorphisms in the human mtDNA genome. For our analysis, we primarily focused on bi-allelic variable sites for network construction. In the following, we provide chapter-wise conclusions of this thesis work.

The presence of one polymorphism and its dynamics affect the frequency and evolution of other polymorphisms in the genome. This is a well-known phenomenon in terms of epistasis and genetic hitchhiking. Motifs are building blocks of networks that are complete subgraphs, and the two-order motifs are the most abundant and easy to analyze motifs in any given network. The two-node motifs feed-forward or feed-backward loops of gene regulatory mechanisms. Mitochondria play an important role in oxygen metabolism, and hence, in chapter 2, we studied the mitochon-

drial genomes for different altitudes. Through the perfectly co-occurring two-order motifs analysis, we found that the polymorphic changes are also interdependent with the change in the physical environment, such as altitude. Categorizing variable sites revealed that all the variable sites do not take part in network construction for the given threshold as perfectly co-occurring two-order motifs. However, this is a method-based observation, and it was interesting to look for the biological aspect of this observation. We found that the approx. 50% nodes that did not take part in network construction were belonging to *Control region*, which is the most highly mutated part of the mtDNA. These sites were also found to show very high minor allele frequency. We also calculated the structural properties of these networks and found that these networks exhibit similarity at the structural level. The structural similarity showed that the overall pattern of two-order motifs does not alter much for different altitude groups. However, the individual variable sites taking part in the motif formation exhibit exclusivity. To look at this through the co-occurrence networks, we found only four variable sites globally present in all the networks, and approx. 40 variable sites as isolated global nodes. This showed that, even if the networks are structurally similar, they exhibit quite a difference at the level of nodes. These globally connected nodes tend to form the motifs with either themselves or mostly with coding genes. Similarly, the local connect nodes or nodes exclusively present in any given altitude group when mapped to genes showed similarity in patterns of motifs formation. This is an exciting observation since the local nodes and their co-occurrence sites were completely exclusive; the similarity at the genetic level shows a peculiar property of the genome, which is an evolution with genetic stability. Along with the global and isolated nodes, there are nodes that are present in more than one altitude group and not in all, and we refer to them as mixed nodes. We used these mixed nodes to calculate the Jaccard similarity index for finding the similarity in altitude groups. Applying the Jaccard similarity index to construct a dendrogram revealed that the eight altitude groups are clustered into two major clades with four lower and four higher altitude groups. The observation of this distinction goes in line with the previous studies on human migration

and settlements on and around the Tibetan plateau. There are reported high-altitude marker sites for the Tibetan plateau. These high-altitude markers were apparently found in higher-altitude groups, and when their co-occurring sites were mapped to genes, it was found that these sites tend to co-occur within their own gene region or OXPHOS complex. This suggests that the high-altitude marker sites either drive the intra-genic variations or are established by these intra-genic variations. When the two-order motifs are mapped to genes, we found different genetic interactions, and thereby genes for these altitude groups, particularly CYB, CO3, and ND6 genes in higher altitude groups interact more often than random networks.

In the previous chapter, we focused our investigation on different altitudes in the Asian population with respect to the Tibetan plateau. In chapter 3, we studied three high-altitude populations, (i) Andean Altiplano in North America, (ii) QinghaiTibetan Plateau in Asia, and (iii) the Ethiopian in Africa. For these populations, we modified the co-occurrence network construction methodology by including the minor allele frequency in our calculation. Unlike the two-order motif analysis performed in the previous chapter, here in this chapter, we performed structural analysis of networks as a whole along with community detection and, up to some extent, haplogroup-based analysis of these communities and other important nodes. These networks showed a very high clustering coefficient and similar path length to corresponding random networks, which shows that these networks exhibit small-world behavior. Hence, there were no distinctly observable hub nodes, and the top degree nodes were found to be haplogroup markers. A high number of nodes and connections were exclusive to each region which suggests the independent co-evolution of variable sites pertaining to each region. The presence of a hierarchy in these networks captures the temporal and spatial evolution of mtDNA genomes. The hierarchy and high clustering coefficient are evident in the presence of communities in these networks. It was found that none of the community cores in each region were commonly present. The community cores were high-degree nodes in each community and can be considered a driver of that community. Exclusivity of these community cores showed that the community formation in these networks is independent

of each other, which suggests the convergent evolution of these populations. On the evaluation of the genetic background of these communities, particular genes were found to be dominating in all the communities in each region. Since there are no distinguishable hub nodes in these networks, we calculated within module degree and participation coefficient for each node in these networks. With these two parameters, it was found that in the 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. With the help of the correlation dimension (D2), we calculated the localization properties of eigenvectors for these networks, and it was found that Tibetan and Andean networks were more localized as compared to Ethiopian networks. This localization parameter, in context to co-mutation, describes the events of admixture in the Ethiopian region. With the help of the gene-gene interaction networks, we identified certain genes in all three regions which were highly deviating from random networks. Upon functional enrichment analysis of these genes, it became clear that Tibetan and Andean populations have evolved in a similar fashion genetically than the Ethiopian population. In the last chapter, we introduced higher-order interactions onto the co-mutation of variable sites. Here, we constructed the 2-order simplices of variable sites by defining the co-mutation frequency between three variable sites. These 2-order simplices were first filtered through the statistical score and second through gene-based information. We superimposed the variable site based 2-order simplices onto genes to finally realize the gene simplices. In the mitochondrial genome, $\sim$10% higher-order interactions were found to be true simplices. In terms of vertex degree, these 2-order simplices captured distinct haplogroups in different altitude populations. In low-altitude, it captures M-haplogroup, in middle altitude, C-haplogroup, and in high-altitude, K haplogroup. Moreover, codon-based simplices provided evidence about the codon bias and conservation of codon usage throughout the mitochondrial genome for all the altitude groups. Based on gene simplices, we found that in the lower altitude group, *ATP6 and ND* genes, in the middle altitude group, *CO1 and ND5* genes, and in the high

altitude group, *CYB and ND5* genes are predominantly forming higher-order simplices. The analysis presented in here can be extended further to generalize all the higher orders rather than just 2-orders to get further insights into higher-order interactions and to identify the simplices-based communities in the mitochondrial genome.

5.2 Future scope

The present work mainly focuses on mitochondrial genomes to construct the connected, unweighted, and undirected networks. Complex systems, like the one we described here in the form of mitochondrial genomic variable sites, are studied under the framework of networks, and their applications have been narrated by many experimental-based biological studies. With the advancement of biological data generation with more preciseness and accuracy, many aspects of metadata are presented with the opportunity to analyze and model them in context to networks. The multidimensional depth of network science has been employed to bring together multifaceted 'omics' data and experimental observations to provide a wholesome picture of the biological aspects. However, with the growing inclusiveness of interdisciplinary fields, there is a broad horizon to analyze the data by developing new methodologies and models. The present work can be further extended by including much more sophisticated and curated metadata in process and downstream analysis.

- For the network construction methods, we have mainly focused on the presence of the variable sites and their minor allele frequency. It would be interesting to include genetically more relevant parameters such as penetrance and expressivity. By doing so, the genotype and phenotype relation would become accessible for statistical analysis through networks. Recently, simplified tools were presented to calculate the penetrance of SNPs. The expressivity is, however, more of a qualitative phenomenon, it can be quantified on a gradient scale. The co-mutations presented in this thesis were filtered through a frequency threshold, and hence, the resulting networks can have different sizes and structural properties. Since the frequency threshold is a

physical parameter, it results in statistically significant interactions. However, including the information on penetrance and expressivity at the genetic level of networks would be an organic extension of the whole analysis.

- Through the whole genome-based association studies, it has become possible to identify SNPs associated with any altered phenotypes. However, many variants susceptible to complex diseases remain unaccounted for. The genetic variants that are left out in most statistically powered association studies are believed to be major contributors to the missing heritability of many characters. Some low-frequency variants and rare variants might substantially affect the manifestation of phenotype individually or in combination. Naturally, considering the rare and very rare variants for the network construction could lead to a higher degree of noise in the network, which can be subjected to selective threshold filtration. Along with the information on the frequency of these variants, linkage disequilibrium, genotype certainty, and annotation of clinical and phenotypic consequences should also be considered in the network analysis.

- The gene regulatory networks and co-expression networks are well-studied networks in the context of the regulation and expression of genes for one or multiple developmental stages of an organism/cell. These networks are heavily dependent on the information curated from the experimental studies. Similarly, the gene-gene interaction networks analyzed in this thesis provide the opportunity to represent the interaction of genes based on variant information. The merit of these networks is that they indirectly depend on primary data supplied along with sequences. In biological systems, activities are specified spatially and temporally. The changes observed with time in cellular activities bring the dynamics to the scenario. It would be interesting to study the co-mutation-based weighted gene-gene interaction networks through a dynamical model with the understanding of the dynamic expression of given genes. Dynamical models present opportunities to subject the bio-

logical information through rigorous numerical simulations to assist molecular biologists in designing further experiments.

Bibliography

[1] Barabási, A. L. Albert, R. (2002), Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 4797. (DOI: 10.1103/RevModPhys.74.47)

[2] Maslov, S., Sneppen, K. (2002). Specificity and stability in topology of protein networks. Science (New York, N.Y.), 296(5569), 910913. (DOI: 10.1126/science.1065103)

[3] Maniatis, T., Reed, R. (2002), An extensive network of coupling among gene expression machines. Nature 416, 499506. (DOI: 10.1038/416499a)

[4] Stelling, J., Klamt, S., Bettenbrock, K. et al. (2002), Metabolic network structure determines key aspects of functionality and regulation. Nature 420, 190193. (DOI: 10.1038/nature01166)

[5] Barabási, AL., Oltvai, Z. (2004), Network biology: understanding the cell's functional organization. Nat Rev Genet. 5, 101113. (DOI: 10.1038/nrg1272)

[6] Jalan, S., Sarkar, C. (2017). Complex networks: An emerging branch of science. Phys. News, 47, 3-4.

[7] Barabási, A. L., Gulbahce, N., Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature reviews. Genetics, 12(1), 5668. (DOI: 10.1038/nrg2918)

[8] Lane, N., Martin, W. (2010). The energetics of genome complexity. Nature, 467(7318), 929934. (DOI: 10.1038/nature09486)

[9] Gabaldón, T., Huynen, M. A. (2004). Shaping the mitochondrial proteome. Biochimica et biophysica acta, 1659(2-3), 212220. (DOI: 10.1016/j.bbabio.2004.07.011)

[10] Wolstenholme D. R. (1992). Animal mitochondrial DNA: structure and evolution. International review of cytology, 141, 173216. (DOI: 10.1016/s0074-7696(08)62066-5)

[11] Kasamatsu, H., Vinograd, J. (1974). Replication of circular DNA in eukaryotic cells. Annual review of biochemistry, 43(0), 695719. (DOI: 10.1146/annurev.bi.43.070174.003403)

[12] Ojala, D., Montoya, J., Attardi, G. (1981). tRNA punctuation model of RNA processing in human mitochondria. Nature, 290(5806), 470474. (DOI: 10.1038/290470a0)

[13] Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S. E., Walford, G. A., Sugiana, C., Boneh, A., Chen, W. K., Hill, D. E., Vidal, M., Evans, J. G., Thorburn, D. R., Carr, S. A., Mootha, V. K. (2008). A mitochondrial protein compendium elucidates complex I disease biology. Cell, 134(1), 112123. (DOI: 10.1016/j.cell.2008.06.016)

[14] Ngo, H. B., Kaiser, J. T., Chan, D. C. (2011). The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. Nature structural molecular biology, 18(11), 12901296. (DOI: 10.1038/nsmb.2159)

[15] Rubio-Cosials, A., Sidow, J. F., Jiménez-Menéndez, N., Fernández-Millán, P., Montoya, J., Jacobs, H. T., Coll, M., Bernadó, P., Solà, M. (2011). Human mitochondrial transcription factor A induces a U-turn structure in the light strand promoter. Nature structural molecular biology, 18(11), 12811289. (DOI: 10.1038/nsmb.2160)

[16] Ekstrand, M. I., Falkenberg, M., Rantanen, A., Park, C. B., Gaspari, M., Hultenby, K., Rustin, P., Gustafsson, C. M., Larsson, N. G. (2004). Mitochondrial transcription factor A regulates mtDNA copy number in mammals. Human molecular genetics, 13(9), 935944. (DOI: 10.1093/hmg/ddh109)

[17] Shi, Y., Dierckx, A., Wanrooij, P. H., Wanrooij, S., Larsson, N. G., Wilhelmsson, L. M., Falkenberg, M., Gustafsson, C. M. (2012). Mammalian transcription factor A is a core component of the mitochondrial transcription machinery. Proceedings of the National Academy of Sciences of the United States of America, 109(41), 1651016515. (DOI: 10.1073/pnas.1119738109)

[18] Wallace D. C. (2015). Mitochondrial DNA variation in human radiation and disease. Cell, 163(1), 3338. (DOI: 10.1016/j.cell.2015.08.067)

[19] Wei, W., Tuna, S., Keogh, M. J., Smith, K. R., Aitman, T. J., Beales, P. L., Bennett, D. L., Gale, D. P., Bitner-Glindzicz, M., Black, G. C., Brennan, P., Elliott, P., Flinter, F. A., Floto, R. A., Houlden, H., Irving, M., Koziell, A., Maher, E. R., Markus, H. S., Morrell, N. W., Chinnery, P. F. (2019). Germline selection shapes human mitochondrial DNA diversity. Science (New York, N.Y.), 364(6442), eaau6520. (DOI: 10.1126/science.aau6520)

[20] Ingman, M., Kaessmann, H., Pääbo, S., Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. Nature, 408(6813), 708713. (DOI: 10.1038/35047064)

[21] Wallace, D. C., Brown, M. D., Lott, M. T. (1999). Mitochondrial DNA variation in human evolution and disease. Gene, 238(1), 211230. (DOI: 10.1016/s0378-1119(99)00295-4)

[22] Ji, F., Sharpley, M. S., Derbeneva, O., Alves, L. S., Qian, P., Wang, Y., Chalkia, D., Lvova, M., Xu, J., Yao, W., Simon, M., Platt, J., Xu, S., Angelin, A., Davila, A., Huang, T., Wang, P. H., Chuang, L. M., Moore, L. G., Qian, G., Wallace, D. C. (2012). Mitochondrial DNA variant associated with Leber hereditary optic neuropathy and high-altitude Tibetans. Proceedings of the National Academy of Sciences of the United States of America, 109(19), 73917396. (DOI: 10.1073/pnas.1202484109)

[23] Caporali, L., Iommarini, L., La Morgia, C., Olivieri, A., Achilli, A., Maresca, A., Valentino, M. L., Capristo, M., Tagliavini, F., Del Dotto, V., Zanna, C., Liguori, R., Barboni, P., Carbonelli, M., Cocetta, V., Montopoli, M., Martinuzzi, A., Cenacchi, G., De Michele, G., Testa, F., Carelli, V. (2018). Peculiar combinations of individually non-pathogenic missense mitochondrial DNA variants cause low penetrance Leber's hereditary optic neuropathy. PLoS genetics, 14(2), e1007210. (DOI: 10.1371/journal.pgen.1007210)

[24] Barton N. H. (2000). Genetic hitchhiking. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 355(1403), 15531562. (DOI: 10.1098/rstb.2000.0716)

[25] Lehner B. (2011). Molecular mechanisms of epistasis within and between genes. Trends in genetics : TIG, 27(8), 323331. (DOI: 10.1016/j.tig.2011.05.007)

[26] Papp, B., Pál, C. (2011). Systems biology of epistasis: shedding light on

genetic interaction network "hubs". Cell cycle (Georgetown, Tex.), 10(21), 36233624. (DOI: 10.4161/cc.10.21.17853)

[27] Jakobsdottir, J., Gorin, M. B., Conley, Y. P., Ferrell, R. E., Weeks, D. E. (2009). Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS genetics, 5(2), e1000337. (DOI: 10.1371/journal.pgen.1000337)

[28] Marchini, J., Donnelly, P., Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics, 37(4), 413417. (DOI: 10.1038/ng1537)

[29] Cordell H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human molecular genetics, 11(20), 24632468. (DOI: 10.1093/hmg/11.20.2463)

[30] Phillips P. C. (2008). Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. Nature reviews. Genetics, 9(11), 855867. (DOI: 10.1038/nrg2452)

[31] Cole, B. S., Hall, M. A., Urbanowicz, R. J., Gilbert-Diamond, D., Moore, J. H. (2017). Analysis of Gene-Gene Interactions. Current protocols in human genetics, 95, 1.14.11.14.10. (DOI: 10.1002/cphg.45)

[32] Horne, B. D., Camp, N. J. (2004). Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. Genetic epidemiology, 26(1), 1121. (DOI: 10.1002/gepi.10292)

[33] Lee, P. H., Shatkay, H. (2009). An integrative scoring system for ranking SNPs by their potential deleterious effects. Bioinformatics (Oxford, England), 25(8), 10481055. (DOI: 10.1093/bioinformatics/btp103)

[34] Lee, P. H., Jung, J. Y., Shatkay, H. (2010). Functionally informative tag SNP selection using a Pareto-optimal approach. Advances in experimental medicine and biology, 680, 173180. (DOI: 10.1007/978-1-4419-5913-3_20)

[35] Schwender, H., Ickstadt, K. (2008). Identification of SNP interactions using logic regression. Biostatistics (Oxford, England), 9(1), 187198. (DOI: 10.1093/biostatistics/kxm024)

[36] Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L., Yu, W. (2010). Predictive rule inference for epistatic interaction detection in genome-wide association

studies. Bioinformatics (Oxford, England), 26(1), 3037. (DOI: 10.1093/bioinformatics/btp622)

[37] Liu, C., Zhao, J., Lu, W., Dai, Y., Hockings, J., Zhou, Y., Nussinov, R., Eng, C., Cheng, F. (2020). Individualized genetic network analysis reveals new therapeutic vulnerabilities in 6,700 cancer genomes. PLoS computational biology, 16(2), e1007701. (DOI: 10.1371/journal.pcbi.1007701)

[38] Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-8582. (DOI: 10.1073/pnas.0601602103)

[39] Newman, M. E., Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), 026113. (DOI: 10.1103/PhysRevE.69.026113)

[40] Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008. (DOI: 10.1088/1742-5468/2008/10/P10008)

[41] Brede, M. (2012). NetworksAn Introduction. Mark E. J. Newman. (2010, Oxford University Press.). ISBN-978-0-19-920665-0. Artificial Life, 18, 241-242. (DOI:10.1093/acprof:oso/9780199206650.001.0001)

[42] Del Genio, C. I., Kim, H., Toroczkai, Z., Bassler, K. E. (2010). Efficient and exact sampling of simple graphs with given arbitrary degree sequence. PloS one, 5(4), e10012. (DOI: 10.1371/journal.pone.0010012)

[43] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science (New York, N.Y.), 298(5594), 824827. (DOI: 10.1126/science.298.5594.824)

[44] Shen-Orr, S., Milo, R., Mangan, S. et al. (2002), Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31, 6468. (DOI: 10.1038/ng881)

[45] Holland, P. W., Leinhardt, S. (1976). Local Structure in Social Networks. Sociological Methodology, 7, 145. (DOI: 10.2307/270703)

[46] Molina C, Stone L. (2012) Modelling the spread of diseases in clustered networks. J Theor Biol. Elsevier; 315: 110118. pmid:22982137. (DOI: 10.1016/j.jtbi.2012.08.036)

[47] Stone, L., Roberts, A. Competitive exclusion, or species aggregation?. Oecologia 91, 419424 (1992). (DOI: 10.1007/BF00317632)

[48] Chen, L., Qu, X., Cao, M. et al. Identification of breast cancer patients based on human signaling network motifs. Sci Rep 3, 3368 (2013). (DOI: 10.1038/srep03368)

[49] Messé, A., Hütt, M.T., Hilgetag, C.C. (2018) Toward a theory of coactivation patterns in excitable neural networks. PLOS Computational Biology 14(4): e1006084. (DOI: 10.1371/journal.pcbi.1006084)

[50] Sporns, O., Kötter, R. (2004) Motifs in Brain Networks. PLOS Biology 2(11): e369. (DOI: 10.1371/journal.pbio.0020369)

[51] Girvan, M., Newman, M.E. (2002) Community structure in social and biological networks. Proc Natl Acad Sci U S A. 99(12):7821-6. (DOI: 10.1073/pnas.122653799)

[52] Guimerà, R., Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. Nature, 433(7028), 895900. (DOI: 10.1038/nature03288)

[53] Segal, E., Shapira, M., Regev, A. et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34, 166176. (DOI: 10.1038/ng1165)

[54] Przytycka, T. M., Singh, M., Slonim, D. K. (2010). Toward the dynamic interactome: it's about time. Briefings in bioinformatics, 11(1), 1529. (DOI: 10.1093/bib/bbp057)

[55] Ni, L., Bruce, C., Hart, C., Leigh-Bell, J., Gelperin, D., Umansky, L., Gerstein, M. B., Snyder, M. (2009). Dynamic and complex transcription factor binding during an inducible response in yeast. Genes development, 23(11), 13511363. (DOI: 10.1101/gad.1781909)

[56] Husain, K., Murugan, A. (2020). Physical Constraints on Epistasis. Molecular biology and evolution, 37(10), 28652874. (DOI: 10.1093/molbev/msaa124)

[57] Bomba, L., Walter, K., Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. Genome biology, 18(1), 77. (DOI: 10.1186/s13059-017-1212-4)

[58] Jalan, S., Sarkar, C. (2017). Complex networks: An emerging branch of science. Phys. News, 47, 3-4.

[59] Rai, A., Menon, A. V., Jalan, S. (2014). Randomness and preserved patterns in cancer network. Scientific reports, 4, 6368. (DOI: 10.1038/srep06368)

[60] Jalan, S., Sarkar, C., Madhusudanan, A., Dwivedi, S. K. (2014). Uncovering randomness and success in society. PloS one, 9(2), e88249. (DOI: 10.1371/journal.pone.0088249)

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

[62] Gardner, T. S., Cantor, C. R., Collins, J. J. (2000). Construction of a genetic toggle switch in Escherichia coli. Nature, 403(6767), 339342. (DOI: 10.1038/35002131)

[63] Kim, J. R., Yoon, Y., Cho, K. H. (2008). Coupled feedback loops form dynamic motifs of cellular networks. Biophysical journal, 94(2), 359365. (DOI: 10.1529/biophysj.107.105106)

[64] Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., Potter, N. E., Heuser, M., Thol, F., Bolli, N., Gundem, G., Van Loo, P., Martincorena, I., Ganly, P., Mudie, L., McLaren, S., O'Meara, S., Raine, K., Jones, D. R., Teague, J. W., Campbell, P. J. (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. The New England journal of medicine, 374(23), 22092221. (DOI: 10.1056/NEJMoa1516192)

[65] Katayama, Y., Tran, V. K., Hoan, N. T., Zhang, Z., Goji, K., Yagi, M., Takeshima, Y., Saiki, K., Nhan, N. T., Matsuo, M. (2006). Co-occurrence of mutations in both dystrophin- and androgen-receptor genes is a novel cause of female Duchenne muscular dystrophy. Human genetics, 119(5), 516519. (DOI: 10.1007/s00439-006-0159-4)

[66] Boddu, P., Chihara, D., Masarova, L., Pemmaraju, N., Patel, K. P., Verstovsek, S. (2018). The co-occurrence of driver mutations in chronic myeloproliferative neoplasms. Annals of hematology, 97(11), 20712080. (DOI: 10.1007/s00277-018-3402-x)

[67] Du, X., Wang, Z., Wu, A., Song, L., Cao, Y., Hang, H., Jiang, T. (2008). Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. Genome research, 18(1), 178187. (DOI: 10.1101/gr.6969007)

[68] Kawada, H., Oo, S. Z., Thaung, S., Kawashima, E., Maung, Y. N., Thu, H. M., Thant, K. Z., Minakawa, N. (2014). Co-occurrence of point mutations in the voltage-gated sodium channel of pyrethroid-resistant Aedes aegypti populations in Myanmar. PLoS neglected tropical diseases, 8(7), e3032. (DOI: 10.1371/journal.pntd.0003032)

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

[70] Pakendorf, B., Stoneking, M. (2005). Mitochondrial DNA and human evolution. Annual review of genomics and human genetics, 6, 165183. (DOI: 10.1146/annurev.genom.6.080604.162249)

[71] Zhao, M., Kong, Q. P., Wang, H. W., Peng, M. S., Xie, X. D., Wang, W. Z., Jiayang, Duan, J. G., Cai, M. C., Zhao, S. N., Cidanpingcuo, Tu, Y. Q., Wu, S. F., Yao, Y. G., Bandelt, H. J., Zhang, Y. P. (2009). 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, 106(50), 2123021235. (DOI: 10.1073/pnas.0907844106)

[72] Dahlback A, Gelsor N, Stamnes JJ, Gjessing Y (2007) UV measurements in the 30005000 m altitude region in Tibet. J Geophys Res Atmos 112:D09308.

[73] Gnecchi-Ruscone, G. A., Abondio, P., De Fanti, S., Sarno, S., Sherpa, M. G., Sherpa, P. T., Marinelli, G., Natali, L., Di Marcello, M., Peluzzi, D., Luiselli, D., Pettener, D., Sazzini, M. (2018). Evidence of Polygenic Adaptation to High Altitude from Tibetan and Sherpa Genomes. Genome biology and evolution, 10(11), 29192930. (DOI: 10.1093/gbe/evy233)

[74] Li, Q., Lin, K., Sun, H., Liu, S., Huang, K., Huang, X., Chu, J., Yang, Z. (2016). Mitochondrial haplogroup M9a1a1c1b is associated with hypoxic adaptation in the Tibetans. Journal of human genetics, 61(12), 10211026. (DOI: 10.1038/jhg.2016.95)

[75] Magalhães, J., Ascensão, A., Soares, J. M., Ferreira, R., Neuparth, M. J., Marques, F., Duarte, J. A. (2005). Acute and severe hypobaric hypoxia increases oxidative stress and impairs mitochondrial function in mouse skeletal muscle. Journal of applied physiology (Bethesda, Md. : 1985), 99(4), 12471253. (DOI: 10.1152/japplphysiol.01324.2004)

[76] Fukuda, R., Zhang, H., Kim, J. W., Shimoda, L., Dang, C. V., Semenza, G. L. (2007). HIF-1 regulates cytochrome oxidase subunits to optimize efficiency of respiration in hypoxic cells. Cell, 129(1), 111122. (DOI: 10.1016/j.cell.2007.01.047)

[77] Solaini, G., Harris, D. A. (2005). Biochemical dysfunction in heart mitochondria exposed to ischaemia and reperfusion. The Biochemical journal, 390(Pt 2), 377394. (DOI: 10.1042/BJ20042006)

[78] Monge, C., León-Velarde, F. (1991). Physiological adaptation to high altitude: oxygen transport in mammals and birds. Physiological reviews, 71(4), 11351172. (DOI: 10.1152/physrev.1991.71.4.1135)

[79] Ballard, J.W.O., Rand, D.M. (2005) The population biology of mitochondrial DNA and its phylogenetic implications. Annual Review of Ecology Evolution and Systematics, 36, 621642. (DOI: 10.1146/annurev.ecolsys.36.091704.175513)

[80] Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M. D., Sukernik, R. I., Olckers, A., Wallace, D. C. (2003). Natural selection shaped regional mtDNA variation in humans. Proceedings of the National Academy of Sciences of the United States of America, 100(1), 171176. (DOI: 10.1073/pnas.0136972100)

[81] Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V., Wallace, D. C. (2004). Effects of purifying and adaptive selection on regional variation in human mtDNA. Science (New York, N.Y.), 303(5655), 223226. (DOI: 10.1126/science.1088434)

[82] Peacock A. J. (1998). ABC of oxygen: oxygen at high altitude. BMJ (Clinical research ed.), 317(7165), 10631066. (DOI: 10.1136/bmj.317.7165.1063)

[83] Chojnacki, S., Cowley, A., Lee, J., Foix, A., Lopez, R. (2017). Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. Nucleic acids research, 45(W1), W550W553. (DOI: 10.1093/nar/gkx273)

[84] Cao, M., Shi, J., Wang, J., Hong, J., Cui, B., Ning, G. (2015). Analysis of human triallelic SNPs by next-generation sequencing. Annals of human genetics, 79(4), 275281. (DOI: 10.1111/ahg.12114)

[85] Pereira, L., Freitas, F., Fernandes, V., Pereira, J. B., Costa, M. D., Costa, S., Máximo, V., Macaulay, V., Rocha, R., Samuels, D. C. (2009). The diversity

present in 5140 human mitochondrial genomes. American journal of human genetics, 84(5), 628640. (DOI: 10.1016/j.ajhg.2009.04.013)

[86] Guo, C., McDowell, I. C., Nodzenski, M., Scholtens, D. M., Allen, A. S., Lowe, W. L., Reddy, T. E. (2017). Transversions have larger regulatory effects than transitions. BMC genomics, 18(1), 394. (DOI: 10.1186/s12864-017-3785-4)

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

[88] Jubb, H. C., Pandurangan, A. P., Turner, M. A., Ochoa-Montaño, B., Blundell, T. L., Ascher, D. B. (2017). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. Progress in biophysics and molecular biology, 128, 313. (DOI: 10.1016/j.pbiomolbio.2016.10.002)

[89] Clauset, A., Newman, M. E., Moore, C. (2004). Finding community structure in very large networks. Physical review. E, Statistical, nonlinear, and soft matter physics, 70(6 Pt 2), 066111. (DOI: 10.1103/PhysRevE.70.066111)

[90] Zhang, D., Dong, G., Wang, H., Ren, X., Ha, P. U., Qiang, M., Chen, F. (2016). History and possible mechanisms of prehistoric human migration to the Tibetan Plateau. Science China Earth Sciences, 59(9), 1765-1778. (DOI: 10.1007/s11430-015-5482-x)

[91] Brantingham, P. J., Xing, G. (2006). Peopling of the northern Tibetan Plateau. World Archaeology, 38(3), 387-414. (DOI: 10.1080/00438240600813301)

[92] Preste, R., Vitale, O., Clima, R., Gasparre, G., Attimonelli, M. (2019). HmtVar: a new resource for human mitochondrial variations and pathogenicity data. Nucleic acids research, 47(D1), D1202D1210. (DOI: 10.1093/nar/gky1024)

[93] Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research, 47(D1), D886-D894. (DOI: 10.1093/nar/gky1016)

[94] Gnecchi-Ruscone, G. A., Abondio, P., De Fanti, S., Sarno, S., Sherpa, M. G., Sherpa, P. T., Marinelli, G., Natali, L., Di Marcello, M., Peluzzi, D., Luiselli, D., Pettener, D., Sazzini, M. (2018). Evidence of Polygenic Adaptation to

High Altitude from Tibetan and Sherpa Genomes. Genome biology and evolution, 10(11), 29192930. (DOI: 10.1093/gbe/evy233)

[95] Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X., Li, K., Gao, G., Yin, Y., Wang, W., Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature, 512(7513), 194197. (DOI: 10.1038/nature13408)

[96] Lee, C., Zeng, J., Drew, B. G., Sallam, T., Martin-Montalvo, A., Wan, J., Kim, S. J., Mehta, H., Hevener, A. L., de Cabo, R., Cohen, P. (2015). The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. Cell metabolism, 21(3), 443454. (DOI: 10.1016/j.cmet.2015.02.009)

[97] Sharma, S., Singh, S., Gupta, R. K., Ganju, L., Singh, S. B., Kumar, B., Singh, Y. (2019). Mitochondrial DNA sequencing reveals association of variants and haplogroup M33a2'3 with High altitude pulmonary edema susceptibility in Indian male lowlanders. Scientific reports, 9(1), 10975. (DOI: 10.1038/s41598-019-47500-1)

[98] MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org.

[99] Starikovskaya, E. B., Sukernik, R. I., Derbeneva, O. A., Volodko, N. V., Ruiz-Pesini, E., Torroni, A., Brown, M. D., Lott, M. T., Hosseini, S. H., Huoponen, K., Wallace, D. C. (2005). Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. Annals of human genetics, 69(Pt 1), 6789. (DOI: 10.1046/j.1529-8817.2003.00127.x)

[100] Tamm, E., Kivisild, T., Reidla, M., Metspalu, M., Smith, D. G., Mulligan, C. J., Bravi, C. M., Rickards, O., Martinez-Labarga, C., Khusnutdinova, E. K., Fedorova, S. A., Golubenko, M. V., Stepanov, V. A., Gubina, M. A., Zhadanov, S. I., Ossipova, L. P., Damba, L., Voevoda, M. I., Dipierri, J. E., Villems, R., Malhi, R. S. (2007). Beringian standstill and spread of Native American founders. PloS one, 2(9), e829. (DOI: 10.1371/journal.pone.0000829)

[101] Llamas, B., Fehren-Schmitz, L., Valverde, G., Soubrier, J., Mallick, S., Rohland, N., Nordenfelt, S., Valdiosera, C., Richards, S. M., Rohrlach, A., Romero, M. I., Espinoza, I. F., Cagigao, E. T., Jiménez, L. W., Makowski, K.,

Reyna, I. S., Lory, J. M., Torrez, J. A., Rivera, M. A., Burger, R. L., Haak, W. (2016). Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. Science advances, 2(4), e1501385. (DOI: 10.1126/sciadv.1501385)

[102] Rolfe, D. F., Brown, G. C. (1997). Cellular energy utilization and molecular origin of standard metabolic rate in mammals. Physiological reviews, 77(3), 731758. (DOI: 10.1152/physrev.1997.77.3.731)

[103] Fontanillas, P., Dépraz, A., Giorgi, M. S., Perrin, N. (2005). Nonshivering thermogenesis capacity associated to mitochondrial DNA haplotypes and gender in the greater white-toothed shrew, Crocidura russula. Molecular ecology, 14(2), 661670. (DOI: 10.1111/j.1365-294X.2004.02414.x)

[104] Witt, K. E., Huerta-Sánchez, E. (2019). Convergent evolution in human and domesticate adaptation to high-altitude environments. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 374(1777), 20180235. (DOI: 10.1098/rstb.2018.0235)

[105] Simonson T. S. (2015). Altitude Adaptation: A Glimpse Through Various Lenses. High altitude medicine biology, 16(2), 125137. (DOI: 10.1089/ham.2015.0033)

[106] Bigham, A. W., Lee, F. S. (2014). Human high-altitude adaptation: forward genetics meets the HIF pathway. Genes development, 28(20), 21892204. (DOI: 10.1101/gad.250167.114)

[107] Beall C. M. (2006). Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. Integrative and comparative biology, 46(1), 1824. (DOI: 10.1093/icb/icj004)

[108] Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., Li, C., Li, J. C., Liang, Y., McCormack, M., Montgomery, H. E., Pan, H., Robbins, P. A., Shianna, K. V., Tam, S. C., Tsering, N., Veeramah, K. R., Wang, W., Wangdui, P., Weale, M. E., Zheng, Y. T. (2010). Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. Proceedings of the National Academy of Sciences of the United States of America, 107(25), 1145911464. (DOI: 10.1073/pnas.1002443107)

[109] Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T., Ge, R. (2010).

Genetic evidence for high-altitude adaptation in Tibet. Science (New York, N.Y.), 329(5987), 7275. (DOI: 10.1126/science.1189406)

[110] Lorenzo, F. R., Huff, C., Myllymäki, M., Olenchock, B., Swierczek, S., Tashi, T., Gordeuk, V., Wuren, T., Ri-Li, G., McClain, D. A., Khan, T. M., Koul, P. A., Guchhait, P., Salama, M. E., Xing, J., Semenza, G. L., Liberzon, E., Wilson, A., Simonson, T. S., Jorde, L. B., Prchal, J. T. (2014). A genetic mechanism for Tibetan high-altitude adaptation. Nature genetics, 46(9), 951956. (DOI: 10.1038/ng.3067)

[111] Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X., Li, K., Gao, G., Yin, Y., Wang, W., Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature, 512(7513), 194197. (DOI: 10.1038/nature13408)

[112] Bigham, A. W., Julian, C. G., Wilson, M. J., Vargas, E., Browne, V. A., Shriver, M. D., Moore, L. G. (2014). Maternal PRKAA1 and EDNRA genotypes are associated with birth weight, and PRKAA1 with uterine artery diameter and metabolic homeostasis at high altitude. Physiological genomics, 46(18), 687697. (DOI: 10.1152/physiolgenomics.00063.2014)

[113] Huerta-Sánchez, E., Degiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., Cardona, A., Montgomery, H. E., Cavalleri, G. L., Robbins, P. A., Weale, M. E., Bradman, N., Bekele, E., Kivisild, T., Tyler-Smith, C., Nielsen, R. (2013). Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. Molecular biology and evolution, 30(8), 18771888. (DOI: 10.1093/molbev/mst089)

[114] Storz J. F. (2021). High-Altitude Adaptation: Mechanistic Insights from Integrated Genomics and Physiology. Molecular biology and evolution, 38(7), 26772691. (DOI: 10.1093/molbev/msab064)

[115] Gassmann, M., Mairbäurl, H., Livshits, L., Seide, S., Hackbusch, M., Malczyk, M., Kraut, S., Gassmann, N. N., Weissmann, N., Muckenthaler, M. U. (2019). The increase in hemoglobin concentration with altitude varies among human populations. Annals of the New York Academy of Sciences, 1450(1), 204220. (DOI: 10.1111/nyas.14136)

[116] Chen, Y., Gong, L., Liu, X., Chen, X., Yang, S., Luo, Y. (2020). Mitochondrial DNA genomes revealed different patterns of high-altitude adaptation in

high-altitude Tajiks compared with Tibetans and Sherpas. Scientific reports, 10(1), 10592. (DOI: 10.1038/s41598-020-67519-z)

[117] Clima, R., Preste, R., Calabrese, C., Diroma, M. A., Santorsola, M., Scioscia, G., Simone, D., Shen, L., Gasparre, G., Attimonelli, M. (2017). HmtDB 2016: data update, a better performing query system and human mitochondrial DNA haplogroup predictor. Nucleic acids research, 45(D1), D698D706. (DOI: 10.1093/nar/gkw1066)

[118] Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature genetics, 23(2), 147. (DOI: 10.1038/13779)

[119] Tranah, G. J., Manini, T. M., Lohman, K. K., Nalls, M. A., Kritchevsky, S., Newman, A. B., Harris, T. B., Miljkovic, I., Biffi, A., Cummings, S. R., Liu, Y. (2011). Mitochondrial DNA variation in human metabolic rate and energy expenditure. Mitochondrion, 11(6), 855861. (DOI: 10.1016/j.mito.2011.04.005)

[120] Brown, K. S., Hill, C. C., Calero, G. A., Myers, C. R., Lee, K. H., Sethna, J. P., Cerione, R. A. (2004). The statistical mechanics of complex signaling networks: nerve growth factor signaling. Physical biology, 1(3-4), 184195. (DOI: 10.1088/1478-3967/1/3/006)

[121] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabási, A. L. (2000). The large-scale organization of metabolic networks. Nature, 407(6804), 651654. (DOI: 10.1038/35036627)

[122] Wagner, A., Fell, D. A. (2001). The small world inside large metabolic networks. Proceedings. Biological sciences, 268(1478), 18031810. (DOI: 10.1098/rspb.2001.1711)

[123] Verma, R. K., Kalyakulina, A., Giuliani, C., Shinde, P., Kachhvah, A. D., Ivanchenko, M., Jalan, S. (2021). Analysis of human mitochondrial genome co-occurrence networks of Asian population at varying altitudes. Scientific reports, 11(1), 133. (DOI: 10.1038/s41598-020-80271-8)

[124] Watts, D. J., Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684), 440442. (DOI: 10.1038/30918)

[125] Barabasi, A. L., Albert, R. (1999). Emergence of scaling in random networks. Science (New York, N.Y.), 286(5439), 509512. (DOI: 10.1126/science.286.5439.509)

[126] Mengistu, H., Huizinga, J., Mouret, J. B., Clune, J. (2016). The Evolutionary Origins of Hierarchy. PLoS computational biology, 12(6), e1004829. (DOI: 10.1371/journal.pcbi.1004829)

[127] Soares, P., Alshamali, F., Pereira, J. B., Fernandes, V., Silva, N. M., Afonso, C., Costa, M. D., Musilová, E., Macaulay, V., Richards, M. B., Cerny, V., Pereira, L. (2012). The Expansion of mtDNA Haplogroup L3 within and out of Africa. Molecular biology and evolution, 29(3), 915927. (DOI: 10.1093/molbev/msr245)

[128] Rai, A., Pradhan, P., Nagraj, J., Lohitesh, K., Chowdhury, R., Jalan, S. (2017). Understanding cancer complexome using networks, spectral graph theory and multilayer framework. Scientific reports, 7, 41676. (DOI: 10.1038/srep41676)

[129] Zifa, E., Giannouli, S., Theotokis, P., Stamatis, C., Mamuris, Z., Stathopoulos, C. (2007). Mitochondrial tRNA mutations: clinical and functional perturbations. RNA biology, 4(1), 3866. (DOI: 10.4161/rna.4.1.4548)

[130] van Oven, M. and Kayser, M. (2009), Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat., 30: E386-E394. (DOI: 10.1002/humu.20921)

[131] Goltsev, A. V., Dorogovtsev, S. N., Oliveira, J. G., Mendes, J. F. (2012). Localization and spreading of diseases in complex networks. Physical review letters, 109(12), 128702. (DOI: 10.1103/PhysRevLett.109.128702)

[132] Suweis, S., Grilli, J., Banavar, J. R., Allesina, S., Maritan, A. (2015). Effect of localization on the stability of mutualistic ecological networks. Nature communications, 6, 10179. (DOI: 10.1038/ncomms10179)

[133] Pradhan, P., Yadav, A., Dwivedi, S. K., Jalan, S. (2017). Optimized evolution of networks for principal eigenvector localization. Physical review. E, 96(2-1), 022312. (DOI: 10.1103/PhysRevE.96.022312)

[134] Pastor-Satorras, R., Castellano, C. (2016). Distinct types of eigenvector localization in networks. Scientific reports, 6, 18847. (DOI: 10.1038/srep18847)

[135] Mishra, A., Bandyopadhyay, J. N., Jalan, S. (2021). Multifractal analysis of eigenvectors of small-world networks. Chaos, Solitons Fractals, 144, 110745. (DOI: 10.1016/j.chaos.2021.110745)

[136] Rives, A. W., Galitski, T. (2003). Modular organization of cellular networks. Proceedings of the National Academy of Sciences of the United States of America, 100(3), 11281133. (DOI: 10.1073/pnas.0237338100)

[137] Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P., Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 430(6995), 8893. (DOI: 10.1038/nature02555)

[138] Luo, Y., Gao, W., Liu, F., Gao, Y. (2011). Mitochondrial nt3010G-nt3970C haplotype is implicated in high-altitude adaptation of Tibetans. Mitochondrial DNA, 22(5-6), 181190. (DOI: 10.3109/19401736.2011.632771)

[139] Lu, M. Y., Huang, J. F., Liao, Y. C., Bai, R. K., Trieu, R. B., Chuang, W. L., Yu, M. L., Juo, S. H., Wong, L. J. (2012). Mitochondrial polymorphism 12361A>G is associated with nonalcoholic fatty liver disease. Translational research : the journal of laboratory and clinical medicine, 159(1), 5859. (DOI: 10.1016/j.trsl.2011.10.011)

[140] Stoneking, M., Hedgecock, D., Higuchi, R. G., Vigilant, L., Erlich, H. A. (1991). Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. American journal of human genetics, 48(2), 370382.

[141] Kanehisa, M., Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 2730. (DOI: 10.1093/nar/28.1.27)

[142] The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic acids research, 47(D1), D330D338. (DOI: 10.1093/nar/gky1055)

[143] Huang, d., Sherman, B. T., Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), 4457. (DOI: 10.1038/nprot.2008.211)

[144] Lau, G. Y., Mandic, M., Richards, J. G. (2017). Evolution of Cytochrome c Oxidase in Hypoxia Tolerant Sculpins (Cottidae, Actinopterygii). Molecular biology and evolution, 34(9), 21532162. (DOI: 10.1093/molbev/msx179)

[145] Shi, Y., Hu, Y., Wang, J., Elzo, M. A., Yang, X., Lai, S. (2018). Genetic diversities of MT-ND1 and MT-ND2 genes are associated with high-altitude adaptation in yak. Mitochondrial DNA. Part A, DNA mapping, sequencing, and analysis, 29(3), 485494. (DOI: 10.1080/24701394.2017.1307976)

[146] Bartáková, V., Bryjová, A., Nicolas, V., Lavrenchenko, L. A., Bryja, J. (2021). Mitogenomics of the endemic Ethiopian rats: looking for footprints of adaptive evolution in sky islands. Mitochondrion, 57, 182191. (DOI: 10.1016/j.mito.2020.12.015)

[147] Song, K., Zhang, Y., Ga, Q., Bai, Z., Ge, R. L. (2020). High-altitude chronic hypoxia ameliorates obesity-induced non-alcoholic fatty liver disease in mice by regulating mitochondrial and AMPK signaling. Life sciences, 252, 117633. (DOI: 10.1016/j.lfs.2020.117633)

[148] Hodgson, J. A., Mulligan, C. J., Al-Meeri, A., Raaum, R. L. (2014). Early back-to-Africa migration into the Horn of Africa. PLoS genetics, 10(6), e1004393. (DOI: 10.1371/journal.pgen.1004393)

[149] Albert R. (2005). Scale-free networks in cell biology. Journal of cell science, 118(Pt 21), 49474957. (DOI: 10.1242/jcs.02714)

[150] Jeong, H., Mason, S. P., Barabási, A. L., Oltvai, Z. N. (2001). Lethality and centrality in protein networks. Nature, 411(6833), 41-42. (DOI: 10.1038/35075138)

[151] Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., ... Davis, R. W. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. science, 285(5429), 901-906. (DOI: 10.1126/science.285.5429.901)

[152] Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. Science, 296(5568), 750-752. (DOI: 10.1126/science.1068696)

[153] Rai, A., Pawar, A. K., Jalan, S. (2015). Prognostic interaction patterns in diabetes mellitus II: A random-matrix-theory relation. Physical Review E, 92(2), 022806. (DOI: 10.1103/PhysRevE.92.022806)

[154] Shinde, P., Whitwell, H. J., Verma, R. K., Ivanchenko, M., Zaikin, A., Jalan, S. (2021). Impact of modular mitochondrial epistatic interactions on

the evolution of human subpopulations. Mitochondrion, 58, 111-122. (DOI: 10.1016/j.mito.2021.02.004)

[155] Verma, R. K., Kalyakulina, A., Mishra, A., Ivanchenko, M., Jalan, S. (2022). Role of mitochondrial genetic interactions in determining adaptation to high altitude human population. Scientific reports, 12(1), 1-12. (DOI: 10.1038/s41598-022-05719-5)

[156] Bromberg, K. D., Ma'ayan, A., Neves, S. R., Iyengar, R. (2008). Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. Science, 320(5878), 903-909. (DOI: 10.1126/science.1152662)

[157] Furlong, L. I. (2013). Human diseases through the lens of network biology. Trends in genetics, 29(3), 150-159. (DOI: 10.1016/j.tig.2012.11.004)

[158] Shinde, P., Marrec, L., Rai, A., Yadav, A., Kumar, R., Ivanchenko, M. Jalan, S. (2019). Symmetry in cancer networks identified: Proposal for multicancer biomarkers. Network Science, 7(4), 541-555. (DOI: 10.1017/nws.2019.55)

[159] Jalan, S., Bandyopadhyay, J. N. (2007). Random matrix analysis of complex networks. Physical Review E, 76(4), 046107.

[160] Butts, C. T. (2009). Revisiting the foundations of network analysis. science, 325(5939), 414-416. (DOI: 10.1126/science.1171022)

[161] Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardenes, J., Romance, M., ... Zanin, M. (2014). The structure and dynamics of multilayer networks. Physics reports, 544(1), 1-122. (DOI: 10.1016/j.physrep.2014.07.001)

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

[163] Holme, P., Saramäki, J. (2012). Temporal networks. Physics reports, 519(3), 97-125. (DOI: 10.1016/j.physrep.2012.03.001)

[164] Kachhvah, A. D., Jalan, S. (2022). First-order route to antiphase clustering in adaptive simplicial complexes. Physical Review E, 105(6), L062203.

[165] Grilli, J., Barabás, G., Michalska-Smith, M. J., Allesina, S. (2017). Higher-order interactions stabilize dynamics in competitive network models. Nature, 548(7666), 210-213. (DOI: 10.1038/nature23273)

[166] Sanchez-Gorostiaga, A., Baji, D., Osborne, M. L., Poyatos, J. F., Sanchez, A. (2019). High-order interactions distort the functional landscape of microbial consortia. PLoS Biology, 17(12), e3000550. (DOI: 10.1371/journal.pbio.3000550)

[167] Sizemore, A. E., Giusti, C., Kahn, A., Vettel, J. M., Betzel, R. F., Bassett, D. S. (2018). Cliques and cavities in the human connectome. Journal of computational neuroscience, 44(1), 115-145. (DOI: 10.1007/s10827-017-0672-6)

[168] Benson, A. R., Gleich, D. F., Leskovec, J. (2016). Higher-order organization of complex networks. Science, 353(6295), 163-166. (DOI: 10.1126/science.aad9029)

[169] Newman, M. E. (2001). Scientific collaboration networks. I. Network construction and fundamental results. Physical review E, 64(1), 016131. (DOI: 10.1103/PhysRevE.64.016131)

[170] Shi, D., Chen, Z., Sun, X., Chen, Q., Ma, C., Lou, Y., Chen, G. (2021). Computing cliques and cavities in networks. Communications Physics, 4(1), 1-7. (DOI: 10.1038/s42005-021-00748-4)

[171] Ghoshal, G., Zlati, V., Caldarelli, G., Newman, M. E. (2009). Random hypergraphs and their applications. Physical Review E, 79(6), 066118. (DOI: 10.1103/PhysRevE.79.066118)

[172] Estrada, E., Rodriguez-Velazquez, J. A. (2005). Complex networks as hypergraphs. arXiv preprint physics/0505137. (DOI: 10.48550/arXiv.physics/0505137)

[173] Courtney, O. T., Bianconi, G. (2016). Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. Physical Review E, 93(6), 062311. (DOI: 10.1103/PhysRevE.93.062311)

[174] Krishnagopal, S., Bianconi, G. (2021). Spectral detection of simplicial communities via Hodge Laplacians. Physical Review E, 104(6), 064303. (DOI: 10.1103/PhysRevE.104.064303)

[175] Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., ... Mewes, H. W. (2010). CORUM: the comprehensive resource of mammalian protein complexes2009. Nucleic acids research, 38(suppl_1), D497-D501. (DOI: 10.1093/nar/gkp914)

[176] Wong, P., Althammer, S., Hildebrand, A., Kirschner, A., Pagel, P., Geissler, B., ... Frishman, D. (2008). An evolutionary and structural characterization of mammalian protein complex organization. BMC genomics, 9(1), 1-16. (DOI: 10.1186/1471-2164-9-629)

[177] Ritz, A., Tegge, A. N., Kim, H., Poirel, C. L., Murali, T. M. (2014). Signaling hypergraphs. Trends in biotechnology, 32(7), 356-362. (DOI: 10.1016/j.tibtech.2014.04.007)

[178] Gaudelet, T., Malod-Dognin, N., Prulj, N. (2018). Higher-order molecular organization as a source of biological function. Bioinformatics, 34(17), i944-i953. (DOI: 10.1093/bioinformatics/bty570)

[179] Franzese, N., Groce, A., Murali, T. M., Ritz, A. (2019). Hypergraph-based connectivity measures for signaling pathway topologies. PLoS computational biology, 15(10), e1007384. (DOI: 10.1371/journal.pcbi.1007384)

[180] Gong, L. I., Bloom, J. D. (2014). Epistatically interacting substitutions are enriched during adaptive protein evolution. PLoS genetics, 10(5), e1004328. (DOI: 10.1371/journal.pgen.1004328)

[181] Puchta, O., Cseke, B., Czaja, H., Tollervey, D., Sanguinetti, G., Kudla, G. (2016). Network of epistatic interactions within a yeast snoRNA. Science, 352(6287), 840-844. (DOI: 10.1126/science.aaf0965)

[182] Taylor, M. B., Ehrenreich, I. M. (2015). Higher-order genetic interactions and their contribution to complex traits. Trends in genetics, 31(1), 34-40. (DOI: 10.1016/j.tig.2014.09.001)

[183] Mullis, M. N., Matsui, T., Schell, R., Foree, R., Ehrenreich, I. M. (2018). The complex underpinnings of genetic background effects. Nature communications, 9(1), 1-10. (DOI: 10.1038/s41467-018-06023-5)

[184] Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L. V., Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. Nature, 494(7436), 234-237. (DOI: 10.1038/nature11867)

[185] Domingo, J., Diss, G., Lehner, B. (2018). Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature, 558(7708), 117-121. (DOI: 10.1038/s41586-018-0170-7)

[186] Taylor, M. B., Ehrenreich, I. M. (2014). Genetic interactions involving five or more genes contribute to a complex trait in yeast. PLoS genetics, 10(5), e1004324. (DOI: 10.1371/journal.pgen.1004324)

[187] Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., ... Myers, C. L. (2018). Systematic analysis of complex genetic interactions. Science, 360(6386), eaao1729. (DOI: 10.1126/science.aao1729)

[188] Salnikov, V., Cassese, D., Lambiotte, R., Jones, N. S. (2018). Co-occurrence simplicial complexes in mathematics: identifying the holes of knowledge. Applied Network Science, 3(1), 1-23. (DOI: 10.1007/s41109-018-0074-3)

[189] Kang, L., Zheng, H. X., Chen, F., Yan, S., Liu, K., Qin, Z., ... Jin, L. (2013). mtDNA lineage expansions in Sherpa population suggest adaptive evolution in Tibetan highlands. Molecular biology and evolution, 30(12), 2579-2587. (DOI: 10.1093/molbev/mst147)

[190] Bofkin, L., Goldman, N. (2007). Variation in evolutionary processes at different codon positions. Molecular Biology and Evolution, 24(2), 513-521. (DOI: 10.1093/molbev/msl178)

[191] Conticello, S. G., Pilpel, Y., Glusman, G., Fainzilber, M. (2000). Position-specific codon conservation in hypervariable gene families. Trends in Genetics, 16(2), 57-59. (DOI: 10.1016/S0168-9525(99)01956-3)