J Biosci (2022)47:63
DOI: 10.1007/s12038-022-00290-7

Review

Nucleotide-based genetic networks: Methods and applications

Rahul K Verma[1], Pramod Shinde[2] and Sarika Jalan[1,3]*

[1]*Department of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Khandwa Road, Simrol, Indore 453 552, India*

[2]*Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, CA 92037, USA*

[3]*Complex Systems Lab, Department of Physics, Indian Institute of Technology Indore, Khandwa Road, Simrol, Indore 453 552, India*

**Corresponding author (Email, sarika@iiti.ac.in)*

MS received 27 September 2021; accepted 25 April 2022

Genomic variations have been acclaimed as among the key players in understanding the biological mechanisms behind migration, evolution, and adaptation to extreme conditions. Due to stochastic evolutionary forces, the frequency of polymorphisms is affected by changes in the frequency of nearby polymorphisms in the same DNA sample, making them connected in terms of evolution. This article presents all the ingredients to understand the cumulative effects and complex behaviors of genetic variations in the human mitochondrial genome by analyzing co-occurrence networks of nucleotides, and shows key results obtained from such analyses. The article emphasizes recent investigations of these co-occurrence networks, describing the role of interactions between nucleotides in fundamental processes of human migration and viral evolution. The corresponding co-mutation-based genetic networks revealed genetic signatures of human adaptation in extreme environments. This article provides the methods of constructing such networks in detail, along with their graph-theoretical properties, and applications of the genomic networks in understanding the role of nucleotide co-evolution in evolution of the whole genome.

Keywords. Co-mutation; co-occurrence networks; complex networks; mitochondrial genome

1. Introduction

1.1 *Genomic variations*

Genomic variations (polymorphisms) play a significant role in defining phenotypic alteration in any given population. These variations could be categorized as neutral, beneficial, or deleterious based on their direct effects on the phenotype. The genomic variations could be further defined based on the presence of alleles, i.e., bi-allelic or tri-allelic, with two or three types of nucleotides at the same positions. Usually, bi-allelic sites are present in abundance compared with tri-allelic sites, and therefore are highly considered for evolutionary and phylogenetic analysis (Adelson *et al.* 2019). With the emergence of next-generation sequencing (NGS) data, genomics studies have been revolutionized. However, with larger genome lengths, even minute errors in sequencing could result in false variant calling. There could be possible errors while analyzing the collected data, such as contamination, degradation, missing private variations for priming, and machine failure (Robasky *et al.* 2014). One could apply Sanger sequencing (Mu *et al.* 2016), sample replicates (Kamps-Hughes *et al.* 2018), and reference sequences (Hardwick *et al.* 2017) to validate the variant calling. One of the possible ways to avoid sequencing error is the smaller size of genomes, such as viral, most bacterial, and extracellular plasmids or mitochondrial genomes. The smaller size allows the highest accuracy

This article is part of the Topical Collection Emergent dynamics of biological networks.

http://www.ias.ac.in/jbiosci
Published online: 12 October 2022

in sequencing the genome. In humans, the nuclear genome is vast and often disintegrates during isolation. In contrast, with its high copy number and small size, the mitochondrial genome became a favorite molecular tool for molecular, phylogenetic, and evolutionary scientists. The mitochondrial genome consists of 16569 nucleotides, 13 protein-coding genes, 2 rRNAs, and 22 tRNAs. The 13 protein-coding genes are part of the mitochondrial OXPHOS pathway (Wallace 2013). Humans originated in Africa and the first haplogroup established was the macrohaplogroup L (Wallace 2015). From this macrohaplogroup, L3 originated 65,000–70,000 years ago and gave rise to two major haplogroups, M and N, which later migrated out of Africa and populated the rest of the world (Wallace 2015). Various different haplogroups were founded by purifying selection of specific mitochondrial variations and enriched at regional levels (Mishmar *et al.* 2003). Apart from identifying the haplogroups and migration patterns, there are variations that also affect the bioenergetic system (Wallace 2013). It has been reported that about 10 to 20% of the tRNA variations, at least a few of the rRNA variations (Ruiz-Pesini and Wallace 2006), and 25% of the mtDNA protein sequence variations (Mishmar *et al.* 2003; Ruiz-Pesini *et al.* 2004) have played roles in altering mitochondrial coupling efficiency. The central role of mitochondria is to generate energy through the electron transport chain (ETC). The ETC oxidizes the reduced dietary components and, in the process, generates a proton gradient across the inner mitochondrial membrane. This proton gradient then leads to the formation of ATP with the help of complex V (ATP synthase) in the matrix. The efficiency with which the proton gradient is converted into energy production is known as coupling efficiency. The mtDNA variations that cause a reduction in this efficiency lead to the production of more heat than ATP, which is used in colder climates. One such variation is 3394C in the *ND1* gene, which is enriched in high-altitude Tibetans compared with low-altitude Asians (Ji *et al.* 2012). The mtDNA variations also affect the pH of the mitochondrial matrix and calcium dynamics in the mitochondria. Particularly, 8701A and 10398A lead to a decrease in the pH of the matrix and the uptake of calcium by mitochondria (Kazuno *et al.* 2006). Along with the phenotypic effect of mutations in coding genes, the variations in the control region also affect the overall dynamics of the mitochondrial genome. One such variant is 295T, which has been shown to enhance the binding of TFAM (mitochondrial transcription factor A) to the L-strand promoter of mtDNA, L-strand transcripts, and mtDNA copy number. The mtDNA mutations are also shown to be associated with a wide range of clinical phenotypes (Wei and Chinnery 2020). Large-scale deletions in mtDNA were, for the first time, shown to be responsible for optic neuropathy and myopathies (Holt *et al.* 1988; Wallace *et al.* 1988). Since mtDNA is present in a few to several copies per cell, depending upon the type of tissue, the mere presence of a mutation in a few molecules does not correspond to any clinical phenotypes. This condition where both mutated and wild-type mtDNA molecules are present is known as heteroplasmy. Mutated mtDNA molecules should be present in high frequency for any clinical implication of mutations to show up (Stewart and Chinnery 2021). Apart from the mtDNA-transcribed genes, there are ~1200 genes, which, if they were to get mutated, might lead to mitochondrial dysfunctions (Calvo *et al.* 2006). With the gain in understanding of the nature of mtDNA mutations through deep sequencing studies, new therapeutic methods are also emerging that target symptomatic interventions, pharmacological therapies, ATP and nitric oxide synthesis pathways, antioxidant defense, and improving mitochondrial quality and apoptosis (Bottani *et al.* 2020). However, in certain conditions, the mere presence of a variation is not enough to impart its effect, as in the case of the mitochondrial genome, since the number of mitochondria varies in human cells and each mitochondrion harbors hundreds to thousands of copies of their genome (D'Erchia *et al.* 2015), and the deleterious effect on the phenotype of individuals depends on the ratio of damaged mitochondrion or mutated genomes with respect to the healthy ones.

1.2 *Mutual effect of variations*

In the human mitochondrial genome, it is observed that a variation may have different effects depending upon the haplogroup/genome backgrounds (Ji *et al.* 2012). As we know, the 3394C variant confers high-altitude adaptation in Tibetans, and its haplogroup background plays a vital role in its phenotypic consequences. When 3394C is present on the M haplogroup on the M9 background in Tibetans and on the C4a4 background in the Indian Deccan plateau population, it has beneficial effects and does not affect the complex I activity. However, its presence in the N haplogroup reduces the complex I activity and associates with Leber hereditary optic neuropathy (LHON), suggesting the role of the haplogroup background in modulating 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 family from northern Italy (Caporali *et al*. 2018). It was reported that the variants, otherwise polymorphic, when present in a particular combination, led to reduced complex I activity and thereby the 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 (Charlesworth *et al*. 2000). Another phenomenon where the presence of mutual variations imparts their effect at the phenotypic level is observed in the form of epistasis, which usually deals with alterations in traits associated with those variations (Lehner 2011). Considering the fact that the relationship between phenotypic effects and the presence of variations is not direct, it becomes important 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 (Papp and Pál 2011). The heritability of complex diseases is minutely affected by the mere presence of single nucleotide polymorphisms (SNPs) (Jakobsdottir *et al*. 2009). However, the manifestation of such diseases depends on the interactions of SNPs (Cordell 2002; Marchini *et al*. 2005; Phillips 2008), and therefore it is important to study the collective effect of genomic variations. There are various ways to study the collective effect of variations, and the specific interactions between genes associated with specific traits (Gilbert-Diamond and Moore 2017). To select a particular cohort of the variations and their interactions responsible for the manifestation of complex phenotypes, various computational methods have been developed and implemented, among which principal component analysis, to infer groups of SNPs from linkage disequilibrium to evaluate multivariate SNP correlations for intragenic diversity coverage (Horne and Camp 2004), integrative scoring system based on their deleterious effects (Lee and Shatkay 2009), and the Pareto-optimal approach for identifying functionally and informatively significant SNPs (Lee *et al*. 2009), are the most popular. There exist other approaches based on pair-wise interactions such as two variations significantly interacting through logic regression (Schwender and Ickstadt 2007), predictive rule inference (Wan *et al*. 2009), 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 (Liu *et al*. 2020). These traditional methods of detecting genetic interactions based on SNPs rely on omitting interactions with minimal or no effect on the trait, thereby leading to the possibility of an increase in false-negative results. To overcome this, variable site pairs are considered based on their allele frequency in the population in question. This provides quantitative as well as qualitative (as major or minor allele) information of all the variable sites for that population. Such pairs of variable sites give rise to a network in which the nodes are defined as variable sites and edges are defined as the relative frequency of occurrence of their alleles together. In this way, we can start with most of the pairs of variable sites without worrying about false-negative errors since we are not omitting any variable site or its interactions based on phenotypic information. Next, with the appropriate threshold selection method we can focus on identifying significant variable sites and their interactions to further analyze the network properties. There are several structural properties such as degree, clustering coefficient, centrality measures, etc., which can be analyzed to extract specific information on important nodes as well as their interactions in the network (Jalan and Sarkar 2017). Various graph-theoretical measures and models are extensively reviewed to understand the biological significance and hidden properties of living systems (Pavlopoulos *et al*. 2011). With myriad biological components such as proteins (Pellegrini *et al*. 2004), transcription factors (Lee *et al*. 2002), metabolites, and metabolic reactions (Jeong *et al*. 2000), the systems were studied as a complex system of networks to advance the fundamental understanding to the origin of functioning of the system. The promising role of network science can be elucidated as a spectrum of structural and spectral properties that have been extensively applied to understand the complex behavior of multiple cancers (Rai *et al*. 2017) and to identify the crucial role of various proteins involved in each developmental stage of *C. elegans* (Shinde and Jalan 2015).

1.3 *Biological applications of co-occurrence/ co-mutation networks*

Role of compensatory mutations as co-mutations have been reported in influenza viral evolution; e.g., the presence of E375G is known to functionally compensate for deleterious effects of R384G while M239V enhances viral fitness in the NP gene (Berkhoff *et al*. 2005). With this information, co-occurrence networks were constructed for the human influenza virus (H3N2) to analyze the collective effect of all the variations on the evolution of the virus from viral genomes isolated between 1968 to 2006 (Du *et al*. 2008). These studies identified the role of connectivity maps between and within viral genes as a contributing factor of influenza virus evolution. However, rather than focusing on individual nucleotide pairs, these studies provided a wholesome picture of viral evolution based on evaluating the changes in co-occurrence network topology with respect to time. In a similar line of work, entropy (for genetic diversity) and information gain (for antigenic degree) were analyzed to identify antigenic critical amino acid positions on hemagglutinin (HA) protein in influenza virus to distinguish between avirulent and virulent strains (Huang *et al*. 2009). This distinction is critical for developing new vaccines for antigenic variants. The information gain from these studies was used to identify co-mutation pairs on different epitopes which could lead to antigenic drift. Association rule mining was performed to extract co-occurrence of mutations in the HA gene of human influenza A/H3N2, A/H1N1, and B viruses to predict their evolution and emphasized vaccine upgradation (Chen *et al*. 2016). Rule-based co-mutation networks identified the driver mutations during the H1N1 2009 pandemic by comparing the degree centrality of pandemic and post-pandemic networks. In an another study, co-occurring mutations in HA and neuraminidase (NA) genes of influenza A/(H1N1)pdm09 viruses were explored (Liu *et al*. 2020). The study showed that sore throat was associated with co-occurring mutations in hemagglutinin and neuraminidase genes. Moreover, apart from the influenza virus, a change in the degree centrality of co-mutation in the Ebola virus has been attributed to the accelerated viral evolution in recent outbreaks. Lethality of the disease with case fatality rate was also predicted by mapping the co-mutation networks (Deng *et al*. 2015). In recent studies, co-existing mutations have been used to classify Indian SARS-CoV-2 strains based on 22 groups (Sarkar *et al*. 2021), whereas co-mutation modules were explored to capture the evolution and transmission patterns of SARS-CoV-2 (Qin *et al*. 2021). The effect of co-occurring genetic alterations on non-small-cell lung cancer (NSCLC) progression and therapy resistance were studied for the first time on *CTNNB1* and *PIK3CA* genes. It was identified that co-occurring alterations in *CTNNB1* and *PIK3CA* genes cooperatively promote cancer progression (Blakely *et al*. 2017). A systems biology approach was presented to extract the impact of functional interactions between mutated genes in different cancers (Cui 2010). In the study, co-occurring (and anti-co-occurring) mutations were defined based on the presence (or absence) of mutated genes in a particular cancer. Analyzing such genetic interaction networks based on co-mutations, it was reported that mutated genes that co-occur in tumors shared signal transduction pathways and had functional similarities. In another study, candidate therapeutic pathways for personalized medicine were identified by utilizing the information of mutated genes in tumors of 14 different cancer types to construct co-mutation-based genetic interaction networks (Liu *et al*. 2020). The interaction of *WNT4* and *WNT5A* genes through rs2072920 and rs11918967 SNPs were shown to be associated with obesity in the Han Chinese people (Dong *et al*. 2017). Along with DNA sequences, co-evolution has been substantially studied in protein sequences (Morcos *et al*. 2011; Kamisetty *et al*. 2013). Various methods such as direct coupling analysis (Morcos *et al*. 2011), statistical coupling analysis (Russ *et al*. 2005), and evolutionary coupling analysis (Hopf *et al*. 2015) were developed and applied to establish the role of amino acid interactions in the structural stability and functionality of proteins. A relatively more developed method known as deep coupling scan was proposed, which takes care of patterns of evolutionary conservation in deep mutation scan data sets (Salinas and Ranganathan 2008). In this article, as a tutorial, we provide detailed step-by-step instructions to construct such co-occurrence/co-mutation networks along with a brief discussion of the results.

2. Network construction techniques

A network consists of a set of connected nodes or units, where connections are defined by an interaction type. There exist various network models incorporating various properties of real-world complex systems, among which three are most popular (box 1). Structural properties of real-world networks are readily compared with these corresponding model networks for their deviations and to detect system specific information. The prerequisite of constructing a network from a given set of genetic sequence data is the alignment of sequences usually in a FASTA format. For aligning the DNA sequences, the online tool Clustal Omega (*https://www.ebi.ac.uk/Tools/msa/clustalo/*) (Sievers *et al*. 2011), and for offline alignment, AliView software (Larsson 2014), are readily employed. Once the sequences are aligned, the next step is to pre-process them, by replacing ambiguous characters such as ‘M’, ‘Y’, ‘S’, ‘K’, etc., with one letter, say, ‘N’. This is done

Box 1. Popular network models and their definitions.

1. Random networks

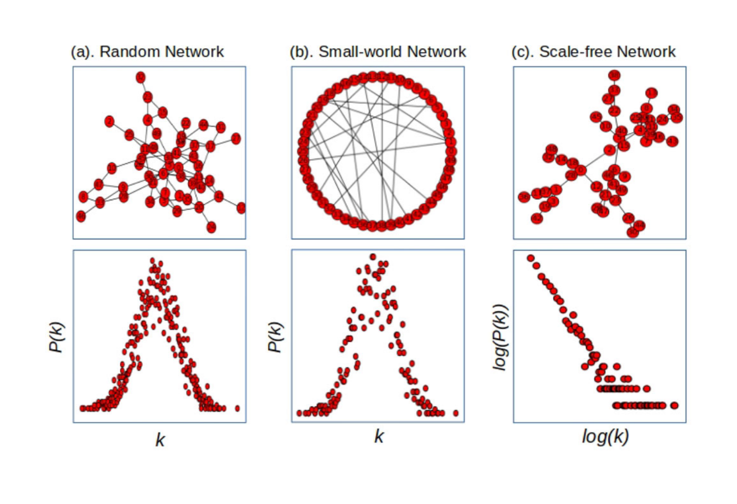
The Erdös–Rényi (ER) (Erdös and Rényi 1960) model is used for generating a random network. Starting with N nodes, it connects each pair of nodes with a probability p, which creates a graph with approximately $pN(N-1)/2$ randomly placed links. The node degrees follow a Poisson distribution (figure a), which indicates that most nodes have approximately the same number of links (close to the average degree $<k>$). This is the commonly used model to compare with real-world networks for their random behavior.

2. Small-world networks

Small-world networks (Watts and Strogatz 1998) 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, $l \sim \log N$. It follows degree distribution similar to ER network (figure b). Small-world behavior is depicted by bacterial metabolic (Wagner and Fell 2001) and brain networks (Bassett and Bullmore 2006) among many others.

3. Scale-free networks

Scale-free networks (Barabási and Pósfai 2016) are characterized by a power-law degree distribution with degree exponent $2 \le \gamma \le 3$. Such distributions are straight lines on a log–log plot (figure c). Yeast protein interaction maps (Uetz *et al.* 2000) and *E.coli* gene regulatory networks (Shen-Orr *et al.* 2002) are few examples which show scale-free topology. In a gene regulatory network, a scale-free topology suggests that a few general transcription factors regulate almost all the genes.



Visualization of model networks (upper panel), the degree distribution ($P(k)$) is plotted for each model network.

To visualize and perform statistical analysis of networks, Gephi (Bastian *et al.* 2009) and Cytoscape (Shannon *et al.* 2003) are well established, as are user-friendly open source softwares along with the network module (Hagberg *et al.* 2008) in Python.

to simplify the procedure of defining the variable sites. The first step to construct these nucleotide networks is identification of the variable sites. A variable site is defined as a position where more than one type of nucleotides is present. Such a site where only two types of nucleotides are present is considered a bi-allelic site, and the site with three nucleotides is considered a tri-allelic site. For our analysis, we only consider bi-allelic sites since the role of tri-allelic sites in evolution is still not clear. The allele frequencies substantially define the

Box 2. Basic terminology of genomics and their definitions.

Allele: An allele describes the different forms of a gene or a position of the gene. The presence of more than one nucleotide at a position in a gene for a given population renders that site as a variable site or polymorphic site.
Genome: Genome is the complete set of DNA present in an organism.
Genomics: It is the study of the whole genome of an organism including genome sequencing, structural organisation, and interactions.
Inheritance: Usually in higher organisms, offsprings get 50% of their genome from each of the parents. In humans, the mitochondrial DNA is solely inherited from the mother only. Hence, it is often used to trace genealogy in humans.
Major allele: The form of a gene which is present in a majority of the population is considered as a major allele.
Minor allele: The form of a gene which is present in a minority in a population is considered as a minor allele. Usually, the allele frequency is measured for the minor allele and known as minor allele frequency (MAF).
Whole genome sequencing: With advancement in sequencing techniques, it is now possible to sequence the whole genome of an organism rather than focusing on few genes or some section of DNA.

genetic structure of the population. However, the individual allele frequencies do not affect the construction of co-occurrence networks at all. Other than the bi-allelic or tri-allelic sites, we have sites with gaps and unknown nucleotides often represented by '–' or 'N' or '?'. The sites with gaps usually represent indels (insertions and deletions), and so there is no meaning in taking those as variable sites; 'N' could be any nucleotide out of the four standard nucleotides, and hence does not give any particular biological information about that polymorphic site. Therefore, we suggest such sites be ignored since they could give rise to artifacts while calculating the co-occurrence/co-mutation frequency between variable positions. In the co-occurrence and co-mutation networks, the nodes are variable sites and the connections are defined according to the type of network we are constructing (this will be discussed further).

2.1 *Co-occurrence networks*

Co-occurrence networks take into account the position of variable sites as nodes, and the connection between a pair of the nodes is defined based on the co-occurrence of alleles for the given population. For this reason, there exists one unique network for each available sequence of 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 defined our edges with respect to the alleles and their frequencies for pairs of variable sites in the population. The co-occurrence frequency between a pair of variable sites for the alleles in position is calculated as

$$Co_{ij} = \frac{\left(x_i y_j\right)^2}{(x_i)(y_i)}$$

where Co_{ij} is the co-occurrence frequency between the x and y alleles present at the i^{th} and j^{th} variable site for a particular sequence. The numerator $x_i y_j$ is the frequency of the presence of the x and y alleles together at the i^{th} and j^{th} sites, whereas x_i and y_j are the total frequencies of the allele x at the i^{th} position and the y allele at the 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 obtain 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 $\langle \mathrm{k} \rangle$ 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 a more or less 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 Nc_{LCC} and

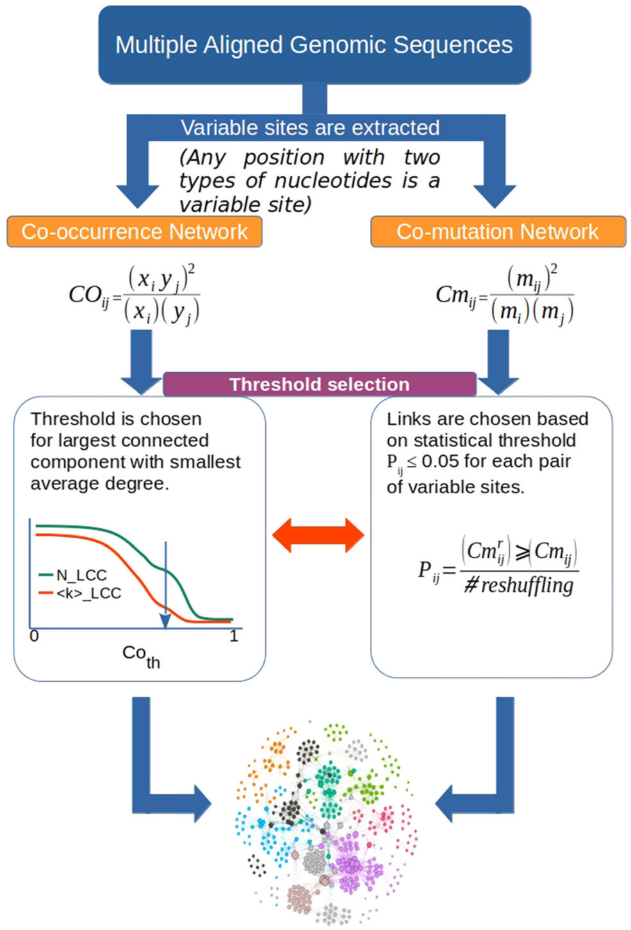

Multiple Aligned Genomic Sequences
Variable sites are extracted
(Any position with two types of nucleotides is a variable site)
Co-occurrence Network
Co-mutation Network
$CO_{ij} = \frac{(x_i y_j)^2}{(x_i)(y_j)}$
$Cm_{ij} = \frac{(m_{ij})^2}{(m_i)(m_j)}$
Threshold selection
Threshold is chosen for largest connected component with smallest average degree.
N_LCC
<k>_LCC
0
1
Co_th
Links are chosen based on statistical threshold $P_{ij} \leq 0.05$ for each pair of variable sites.
$P_{ij} = \frac{(Cm^r_{ij}) \geq (Cm_{ij})}{\#reshuffling}$

Figure 1. Schematic for construction of co-occurrence and co-mutation networks. Co_{ij} is co-occurrence frequency of two nucleotides at the variable positions i and j, $x_i y_j$ is the occurrence of x and y nucleotides at the i^{th} and j^{th} positions together, x_i and y_i are the occurrence of x and y nucleotides at the i^{th} and j^{th} position, respectively. Cm_{ij} and $Cm^r{}_{ij}$ represent the natural and random co-mutation frequencies respectively, m_{ij} represents the frequency of minor alleles present at both i^{th} and j^{th} positions together, m_i and m_j are the minor allele frequencies at the i^{th} and j^{th} positions respectively. Once Co_{ij} or Cm_{ij} is calculated, one can use either or both of the threshold selection methods.

$\langle k \rangle$. Following this procedure, we obtain 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 $Nc_{LCC} < N_{LCC}$, where Nc_{LCC} and N_{LCC} represent the number of connections and the number of nodes of the LCC, respectively (figure 1). We applied the above-mentioned method of threshold selection to all the networks generated for a given population, yielding as many sparse networks as the number of available sequences. In the next step, we constructed a master network by merging all the individual networks generated for each sequence. In this step, we obtained duplicate nodes and edges; however, we chose to perform a union operation on our networks (which can be performed by using codes in Python or Java or any other preferred programming language by the user). By performing a union operation on the edges, we obtained only one network in which all the nodes and edges of all the networks were taken into consideration only once, yielding a single undirected and unweighted co-occurrence network for all the samples. Then, various graph-theoretic structural properties were analyzed for this network, which are discussed in the results section.

2.2 *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. When we compare the minor allele with the reference sequence, we find that this minor allele is present as a major allele in the population. However, the site with such an allele would still be considered as a variable site for our analysis. 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

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

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

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

where $Cm^r{}_{ij}$ is a 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 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, and hence, the method is independent of number of samples.

2.3 *Network properties*

Degree (k): The degree of a node is defined as number of connections that a node has with other nodes (Barabási and Pósfai 2016). A node with a very high degree is referred to as the hub node and is known to play important functional roles in the corresponding system. In most networks, such hub nodes are very few, and hence the network becomes highly robust against random external attacks. On the other hand, a target attack to such nodes could drastically collapse the network. The degree of a node i is denoted as k_i and is calculated from a symmetric matrix as

$$k_i = \sum_{j=1}^{N} A_{ij}$$

where j is number of columns of the given adjacency matrix. The average degree ⟨k⟩ of a network is the average of the degrees of all the nodes in the network and is a measure of the sparseness (or denseness) of the underlying system.

Clustering coefficient: The tendency of the nodes in a system to form triangles is captured by the clustering coefficient (Barabási and Pósfai 2016). Most real-world networks show high clustering coefficients compared with the corresponding random network. The higher clustering also provides information on the existence of modularity in the network. The clustering coefficient of a node i is calculated as

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

where k_n is the number of connections between the neighbors of i.

Hierarchy: A decrease in the clustering coefficient with an increase in the degree of nodes suggests the presence of hierarchy in the network (Barabási and Pósfai 2016). This indicates that the nodes with small degrees belong to highly interconnected small modules. To quantify hierarchy, the local reaching centrality (C_R) is measured for a 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 the crucial role of cross-talking between any two or more modules. C_R is calculated as

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

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

$$h = \frac{\sum\left[C_R^{max} - C_R(i)\right]}{N-1}$$

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

Betweenness centrality: This is the measure of the importance of a node as a 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 is calculated as

$$\beta_i = \sum_{st} \frac{n^i_{st}}{g_{st}}$$

Modularity: In biological systems, the components form groups in order to perform relatively different functions at the molecular level (Barabási and Pósfai 2016). This property of a system is referred to as modularity. A high clustering coefficient is the signature of a network to have high modularity. This also gives 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 (Girvan and Newman 2002) or Louvian (Blondel *et al.* 2008).

3. Pseudo-codes/algorithms used to construct networks

In this section we have provided algorithms for constructing both types of networks. Algorithm 1 is common for both network construction methods.

Algorithm 1: Identify variable site

```
[] ← m
matrix ←sequences in fasta format
comment line with name of individual sequence starting with '>'
for each column in matrix do
        if two standard nucleotides are present then
                if '-' and 'N' are absent then
                        consider column index as variable site
                add column index in m
                end if
        end if
end for
```

Algorithm 2: Co-occurrence frequency calculation

```
for each sequence do
    for each (i, j) pair of variable sites do
        count occurrence of nucleotides, individually
        count occurrence of nucleotides, together
        Calculate co-occurrence frequency, (Coij )
        return i, j, Coij
    end for
end for
```

Algorithm 3: Co-mutation frequency calculation

```
for each (i, j) pair of variable sites do
    count occurrence of minor alleles, individually
    count occurrence of minor alleles, together
    Calculate co-mutation frequency, (Cmij )
    counter = 0
    for 10000 simulations do
        permute column i
        permute column j
        Calculate permuted co-mutation frequency, (Cmrij )
        if (Cmrij ) ≥ (Cmij ) then
            counter += 1
        end if
    end for
Calculate threshold (p) as,
                                        $p = \frac{counter}{10000}$
    if p ≤ 0.05 then
        return i, j, Cmij
    end if
end for
```

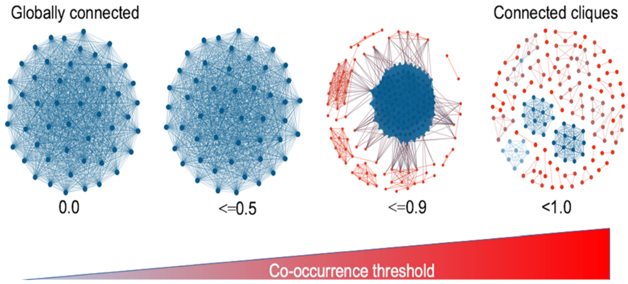


Figure 2. Evolution of co-occurrence network with change in co-occurrence threshold. In the first graph, when no threshold is taken, an all-to-all connected network is produced. In the second graph, when the threshold is increased to 0.5, only those edges are present which have co-occurrence frequency (C_o) 0.5. In the third graph, the threshold is further increased to 0.9, where we lose a lot of connections and get a sparse network. In the fourth graph, the threshold is set to 1.0, which means that only edges with C_o = 1.0 will be present in the network, giving rise to complete subgraphs.

4. Results

4.1 *Structural properties*

We have discussed two different methods of threshold selection to obtain a sparse network. One could apply either of these methods or even both of them together. Further, we considered a real-world example to understand these two methods of network construction. As a practical example, we started with ∼1500 human mtDNA from hmtDB from Oceania, giving rise to ∼1500 variable sites (dependent on the updates of the database), out of which only ∼450 take part in network construction with ∼470 connections. Figure 2a shows that the order of the largest connected component (N_{LCC}) and its average degree (k_{LCC}) decreases with an increase in the value of the co-occurrence frequency threshold. Initially, with no threshold, all the nodes (N) take part in network construction having degree $N-1$. In the inset of the graph in figure 3 (left), N_{LCC} and k_{LCC} from 0.9 to 1.0 are plotted to show the changes in the value of threshold to be considered to get a largest possible sparse network, which in this case is 0.9988. This method of threshold selection is perceived as the network efficiency score since it is employed to generate a sparse network. Similar to the network efficiency score, the effect of the p-value-based threshold selection method can be seen (figure 3, right). Here, we took the mtDNA of a Tibetan population with ∼85 samples, giving rise to ∼420 variable sites, of which ∼400 take part in network construction and ∼3400 significant connections. The number of nodes participating in the network construction is not affected as much as the change (more than 50%) in the number of connections (*Nc*) and the average degree ⟨k⟩, thus resulting in a sparse network. For this particular example, we considered a co-mutation network in which the minor allele frequency (*maf*) of a variable site plays a crucial role in determining the importance of the corresponding node in that network (figure 4). We plotted the degree of the node and its minor allele frequency to determine the role of the minor allele frequency in defining the co-mutation tendency of a variable site (figure 4a). The *maf* and degree show a positive correlation (0.4) for the given population, and there are a few nodes with smaller values of *maf* corresponding to a high degree in the network, and vice versa. In this context, the mutations which are usually the drivers of cancer evolution have been shown to have high *maf* (Spurr *et al.* 2018). It could be easily interpreted from the figure that there exist few nodes which have very high *maf* and most of the nodes have *maf* between 1 to 10 (figure 4a, x-axis). However, the two variable sites which co-mutate do not show any particular correlation in terms of *maf* (figure 4b).

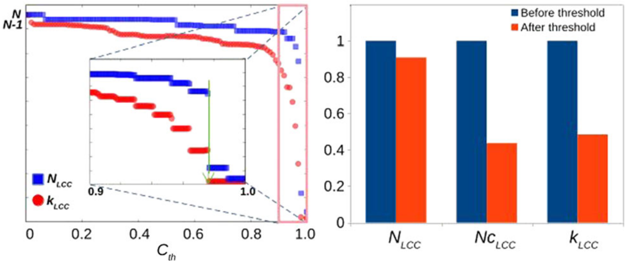


Figure 3. Left: Change in the order of largest connected component (blue squares) and its average degree (red dots) with respect to the co-occurrence frequency threshold. Right: The effect of p-value threshold on the order (N_{LCC}), size (Nc_{LCC}) and average degree (k_{LCC}) of largest connected component of a co-mutation network.

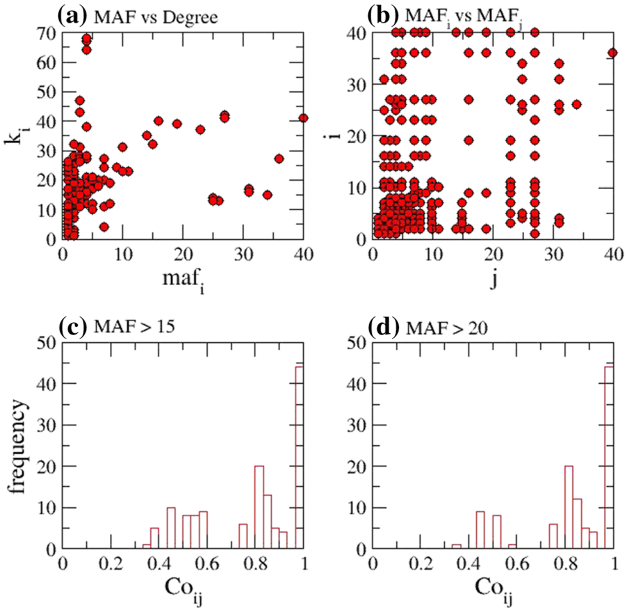


Figure 4. These networks show similar average shortest path length and higher clustering coefficient as compared with the corresponding random networks, further adding to the evidence on the small-world-like nature of these networks. (**a**) The *maf* and degree of each node showed a positive correlation. (**b**) The *maf* of nodes, *i* and *j* of each pair. (**c** and **d**) The effect of *maf* on the distribution of co-mutation frequency.

Variable sites with high *maf* can also co-mutate with nodes having relatively low *maf*. Moreover, *maf* could affect the distribution of edges with high co-mutation frequency (figure 4c and d). When the value of *maf* is increased from 15 to 20, the edges with $C_{oij} \geq 0.6$ are not affected, while edges with smaller C_{oij} are removed from the network. This observation provides a general idea that pairs of nodes in which both the variable sites have high *maf* would give a high value of co-mutation frequency (C_{oij}) in the network, since their association would be far less random (figure 4c and d).

As we know that the degree of a variable site provides the extent of co-evolution of that site with other sites, the degree is a prime property that indicates evolutionary change in the genome of an organism. The change in the average degree between viral genomes of two different seasons accounted for the sudden evolution of viral strains (Du *et al*. 2008). Similarly, the co-mutation of two amino acids of the HA protein in H3N2 virus was linked with antigenic drift or high co-evolution. A co-mutation score was calculated for each amino acid pair, and it was identified that the epitope regions (the antibody binding sites) showed a higher tendency of co-mutations as compared with non-epitope regions (Huang *et al*. 2009). The strong selection pressure on the mutated sites could also be identified by looking at the frequently co-mutating sites as in human influenza A/H3N2 and A/H1N1, and B viruses (Chen *et al*. 2016). In a large-scale study of mitochondrial genomic co-occurrence networks, it was found that the average degree is conserved across all the five continents (Shinde *et al*. 2018). This observation supports the consistent distribution of mtDNA variations across the whole world and the nature of its uniparental inheritance (Ruiz-Pesini *et al*. 2004; Wei and Chinnery 2020). The high-degree nodes usually considered as hub nodes were shown to be the mitochondrial haplogroup markers in three high-altitude populations (Verma *et al*. 2022).

Along with the degree of specific nodes, the degree distribution of all the nodes provides information about the nature of the network to compare with the model networks. The degree distribution (box 1) of these two types of the nucleotide networks follows a binomial distribution with a pronounced peak at $\langle k \rangle$ and decays exponentially for large degrees similar to small-world networks (Albert and Barabási 2002). However, the degree distribution could be affected by the choice of threshold for generating the final network. Additionally, these networks also show very high clustering coefficients, which suggests that the nodes in these networks tend to cluster together. The presence of a high clustering coefficient has been shown to be associated with the co-evolution of mitochondrial variations as a cohort between and among mitochondrial genes, in a dependable manner rather than individually, which is in line with the mechanism of haplogroup inheritance among subpopulations (Shinde *et al*. 2021; Verma *et al*. 2022). These networks show similar average shortest path length and higher clustering coefficient as compared with the corresponding random networks, further adding to the evidence about the small-world-like nature of these networks. In terms of sparseness, the network constructed using the p-value threshold method is more sparse as compared with that constructed using the efficiency score method. The high clustering also suggests the presence of high modularity in these networks. The modules formed in these networks are predominated by particular haplogroups and also specific genes. Thus, these modules harbor critical information, and these modules could be further analyzed at the genetic, haplogroup, or pathogenic levels.

The clustering co-efficient was reported to follow a negative power law with respect to degree (Ravasz *et al.* 2002), and hence provided evidence for the presence of hierarchy in these networks. The presence of hierarchy and modularity is consistent with the evolution of haplogroups over time and geographic space. As it is well established that humans migrated out of Africa and established settlements across Eurasia and other continents with time, various variations were also selected and enriched in the regional population while inheriting the previous root variants that gave rise to haplogroups. The hierarchy captures this characteristic behavior of mitochondrial evolution.

4.2 *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 $C_{mij} = 1.0$ and $P_{ij} \leq 0.05$, we get only those pairs of 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 2). The peculiar case of perfectly co-occurring sites could be interpreted in two ways: first, the nucleotides at the 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 (Bomba *et al.* 2017). To avoid this bias, one can either 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. Two- and three-order network motifs were analyzed for codon bias in the human population (Shinde *et al.* 2018). In this study, the codon positions of the mitochondrial genome were mapped to these network motifs given the position of the variable sites. The protein-coding gene codon positions were mapped as 1, 2, and 3, and the non-coding gene codon positions were mapped as 0. It was shown that synonymous positions (0 and 3) tend to co-occur more often than the non-synonymous positions. This suggests that protein-coding regions have played a selective role in human evolution. In an another study to compare the low- and high-altitude populations of Asia, these perfectly co-occurring network motifs were employed to identify the role of high-altitude marker sites in defining the co-evolution patterns at high altitudes (Verma *et al.* 2021)

4.3 *Gene mapping and genetic interaction networks*

The variable sites participating in the network construction after applying a threshold could be mapped to their corresponding genes. For this purpose, the reference sequence should be mapped to the multiple sequence alignment by introducing all the indels. In this way, one can generate a genetic interaction network based on the co-mutation/co-occurrence network. Since more than one variable site might belong to the same gene, or two separate sets of variable sites might belong to same gene pairs, the resultant network would be weighted but undirected. For example, there exists a link between the variable sites 45 and 89, and these two sites belong to same gene, yielding a self-loop for this gene, and in another case, variable sites 35 and 102, and 48 and 105, have a link. The variable sites 35 and 48 belong to gene 1, and 102 and 105 belong to gene 2. This will yield a link between gene 1 and gene 2 with a weight 2. Thus, this genetic interaction network contains the information of co-evolution of the gene qualitatively and quantitatively. Such networks were constructed to compare the three high-altitude populations: Tibet, Ethiopia, and Andes. Functional enrichment analysis of the identified gene sets provided information about their role in the evolution of the human population at these high-altitude regions (Verma *et al.* 2022). By mapping the variable sites to genes, the co-evolution of genes could be quantified. For example, co-operative changes were observed among and between influenza genes. It was observed that hemagglutinin genes underwent connectivity changes within themselves during a particular period and the neuraminidase genes underwent a similar evolutionary pattern as other genes (Du *et al*. 2008). In gene–gene networks, the flow of information was investigated in co-mutating genes using a model 'perturbed master equation' (pME) in order to identify the gene pairs responsible for network frailness in breast cancer (Bersanelli *et al*. 2020). In another study, enriched genes were identified in cancer pathways from co-mutation-based gene–gene networks of a large-scale study across 14 cancers with 2.5 million non-synonymous mutations and ~6700 tumor exomes (Liu *et al*. 2020). It was predicted based on this study that interactions between *BRCA2* and *TP53* were related to the sensitivity/resistance to anticancer drugs.

5. Conclusion and future prospects

This review described the methodology for constructing networks utilizing the information of variable sites of multiple DNA sequences. The variable sites could be defined as nodes given their allelic information. Based on this definition of allelic information of variable sites and their associations, 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 environments, specifically, fast-evolving viral genomes and mitochondrial genomes. The co-occurrence network motifs were studied in the mitochondrial genome in the human population of five continents to identify the role of codon positions (Shinde *et al*. 2018) and population-based biases in mitochondrial epistatic interactions of ancestral sites (Shinde *et al*. 2021) in shaping human evolution and migration patterns. Such motifs were also applied to analyze the human populations residing at different altitudes with respect to the Tibetan population (Verma *et al*. 2021), and co-mutation-based genetic networks identified the interplay of different gene sets in convergent evolution of highlanders globally (Verma *et al*. 2022). Network module identification and their analysis on these networks have shown to have potential applications in gaining information on human migration and geographic distribution of various haplogroups. The module-based study of amino acid substitutions in human influenza virus identified the change in the antigenic structure of viral proteins which could evade the recognition by antibodies (Du *et al*. 2008). Further, the evolution of viral genomes poses a greater threat to human health, such as the global pandemic of COVID-19, where the role of such networks becomes more important in predicting the evolution of potentially pathogenic strains of these viruses and similar pathogens (Qin *et al*. 2021; Sarkar *et al*. 2021). Genetic interaction networks have been successfully applied for mitochondrial (Verma *et al*. 2022) and viral genomes (Du *et al*. 2008) to identify the specific sets of genes responsible for evolution and adaptation. These networks could be further applied to identify unknown genes and their possible role in classifying viral or bacterial strains as virulent or hyper-virulent strains. These networks could also be applied to genomes of various organisms along with more sophisticated network science techniques, such as spectral techniques, as well as the established population genomics tools in advancing the understanding of genetic evolution.

Acknowledgements

SJ and RKV, respectively, acknowledge the support from DST POWER grant SERB/F/9035/2021-2022 and CSIR grant (CSIR, 25(0293)/18/EMR-II), University Grant Commission for SRF fellowship (305089). All the members of Complex Systems Lab are also acknowledged.

Availability of Codes

The codes to generate co-occurrence and co-mutation networks are freely available on our Github repository: *https://github.com/complex-systems-lab/Genomic-nucleotide-networks*.

References

Adelson RP, Renton AE, Li W, *et al*. 2019 Empirical design of a variant quality control pipeline for whole genome sequencing data using replicate discordance. *Sci. Rep.* **9** 16156

Albert R and Barabási AL 2002 Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74** 47–97

Barabási AL and Pósfai M 2016 *Network science* (Cambridge: Cambridge University Press)

Bassett DS and Bullmore E 2006 Small-world brain networks revisited. *Neuroscientist* **12** 512–523

Bastian M, Heymann S and Jacomy M 2009 Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Weblogs and Social Media* **3** 361–362

Berkhoff EGM, de Wit E, Geelhoed-Mieras MM, *et al.* 2005 Functional constraints of influenza A virus epitopes limit escape from cytotoxic T lymphocytes. *J. Virol.* **79** 11239–11246

Bersanelli M, Mosca E, Milanesi L, *et al.* 2020 Frailness and resilience of gene networks predicted by detection of co-occurring mutations via a stochastic perturbative approach. *Sci. Rep.* **10** 2643

Blakely C, Watkins T, Wu W, *et al.* 2017 Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. *Nat. Genet.* **49** 1693–1704

Blondel VD, Guillaume JL, Lambiotte R, *et al.* 2008 2008 Fast unfolding of communities in large networks. *J. Stat. Mech.* **P10** 008

Bomba L, Walter K and Soranzo N 2017 The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18** 77

Bottani E, Lamperti C, Prigione A, *et al.* 2020 Therapeutic approaches to treat mitochondrial diseases: "one-size-fits- all" and "precision medicine" strategies. *Pharmaceutics* **12** 1083

Calvo S, Jain M, Xie X, *et al.* 2006 Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet.* **38** 576–582

Caporali L, Iommarini L, La Morgia C, *et al.* 2018 Peculiar combinations of individually non-pathogenic missense mitochondrial DNA variants cause low penetrance Leber's hereditary optic neuropathy. *PLoS Genet.* **14** e1007210

Charlesworth B, Harvey PH and Barton NH 2000 Genetic hitchhiking. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **355** 1553–1562

Chen H, Zhou X, Zheng J, *et al.* 2016 Rules of co-occurring mutations characterize the antigenic evolution of human influenza A/H3N2, A/H1N1 and B viruses. *BMC Med. Genom.* **9** 69

Cordell HJ 2002 Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11** 2463–2468

Cui Q 2010 A network of cancer genes with co-occurring and anti-co-occurring mutations. *PLoS One* **5** e13180

Deng L, Liu M, Hua S, *et al.* 2015 Network of co-mutations in Ebola virus genome predicts the disease lethality. *Cell Res.* **25** 753–756

Dong S, Hu W, Yang T, *et al.* 2017 NP-SNP interactions between WNT4 and WNT5A were associated with obesity related traits in Han Chinese population. *Sci. Rep.* **7** 43939

Du X, Wang Z, Wu A, *et al.* 2008 Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. *Genome Res.* **18** 178–187

D'Erchia A, Atlante A, Gadaleta G, *et al.* 2015 Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity. *Mitochondrion* **20** 13–21

Erdös P and Rényi A 1960 On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5** 17–61

Gilbert-Diamond D and Moore JH 2017 Analysis of gene-gene interactions. *Curr. Protoc. Hum. Genet.* **95** 1.14.1-1.14.10

Girvan M and Newman MEJ 2002 Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826

Hagberg A, Swart P and Chult S 2008 Exploring network structure, dynamics, and function using NetworkX; in *Proceedings of the 7th Python in Science Conference (SciPy)* pp. 11–15

Hardwick S, Deveson I and Mercer T 2017 Reference standards for next-generation sequencing. *Nat. Rev. Genet.* **18** 473–478

Holt IJ, Harding AE and Morgan-Hughes JA 1988 Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. *Nature* **331** 717–719

Hopf TA, Morinaga S, Ihara S, *et al.* 2015 Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nat. Commun.* **6** 6077

Horne BD and Camp NJ 2004 Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genet. Epidemiol.* **26** 11–21

Huang JW, King CC and Yang JM 2009 Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. *BMC Bioinform.* **10** S41

Jakobsdottir J, Gorin MB, Conley YP, *et al.* 2009 Interpretation of genetic association studies: Markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet.* **5** e1000337

Jalan S and Sarkar C 2017 Complex networks: An emerging branch of science. *Phys. News* **47** 42–52

Jeong H, Tombor B, Albert R, *et al.* 2000 The large-scale organization of metabolic networks. *Nature* **407** 651–654

Ji F, Sharpley MS, Derbeneva O, *et al.* 2012 Mitochondrial DNA variant associated with Leber hereditary optic neuropathy and high-altitude Tibetans. *Proc. Natl. Acad. Sci. USA* **109** 7391–7396

Kamisetty H, Ovchinnikov S and Baker D 2013 Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. USA* **110** 15674–15679

Kamps-Hughes N, McUsic A, Kurihara L, *et al.* 2018 ERASE-Seq: Leveraging replicate measurements to enhance ultralow frequency variant detection in NGS data. *PLoS One* **13** e0195272

Kazuno Aa, Munakata K, Nagai T, *et al.* 2006 Identification of mitochondrial DNA polymorphisms that alter mitochondrial matrix pH and intracellular calcium dynamics. *PLoS Genet.* **2** e128

Larsson A 2014 AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30** 3276–3278

Lee PH and Shatkay H 2009 An integrative scoring system for ranking SNPs by their potential deleterious effects. *Bioinformatics* **25** 1048–1055

Lee PH, Jung JY and Shatkay H 2009 Functionally informative tag SNP selection using a pareto-optimal approach: playing the game of life. *BMC Bioinform.* **10** (Suppl 13) O5

Lee TI, Rinaldi NJ, Robert F, *et al.* 2002 Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298** 799–804

Lehner B 2011 Molecular mechanisms of epistasis within and between genes. *Trends Genet.* **27** 323–331

Liu C, Zhao J, Lu W, *et al.* 2020 Individualized genetic network analysis reveals new therapeutic vulnerabilities in 6,700 cancer genomes. *PLoS Comp. Biol.* **16** e1007701

Marchini J, Donnelly P and Cardon LR 2005 Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37** 413–417

Mishmar D, Ruiz-Pesini E, Golik P, *et al.* 2003 Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. USA* **100** 171–176

Morcos F, Pagnani A, Lunt B, *et al.* 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA* **108** E1293–E1301

Mu W, Lu HM, Chen J, *et al.* 2016 Sanger confirmation is required to achieve optimal sensitivity and specificity in next-generation sequencing panel testing. *J. Mol. Diagn.* **18** 923–932

Papp B and Pál C 2011 Systems biology of epistasis: Shedding light on genetic interaction network "hubs". *Cell Cycle* **10** 3623–3624

Pavlopoulos GA, Secrier M, Moschopoulos CN, *et al.* 2011 Using graph theory to analyze biological networks. *BioData Mining* **4** 10

Pellegrini M, Haynor D and Johnson JM 2004 Protein Interaction Networks. *Expert Rev. Proteomics* **1** 239–249

Phillips PC 2008 Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9** 855–867

Qin L, Ding X, Li Y, *et al.* 2021 Co-mutation modules capture the evolution and transmission patterns of SARS-CoV-2. *Brief. Bioinform.* **22** bbab222

Rai A, Pradhan P, Nagraj J, *et al.* 2017 Understanding cancer complexome using networks, spectral graph theory and multilayer framework. *Sci. Rep.* **7** 41676

Ravasz E, Somera AL, Mongru DA, *et al.* 2002 Hierarchical organization of modularity in metabolic networks. *Science* **297** 1551–1555

Robasky K, Lewis NE and Church GM 2014 The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* **15** 56–62

Ruiz-Pesini E and Wallace D 2006 Evidence for adaptive selection acting on the tRNA and rRNA genes of human mitochondrial DNA. *Hum. Mutat.* **27** 1072–1081

Ruiz-Pesini E, Mishmar D, Brandon M, *et al.* 2004 Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303** 223–226

Russ WP, Lowery DM, Mishra P, *et al.* 2005 Natural-like function in artificial WW domains. *Nature* **437** 579–583

Salinas V and Ranganathan R 2008 Coevolution-based inference of amino acid interactions underlying protein function. *eLife* **7** 34,300

Sarkar R, Mitra S, Chandra P, *et al.* 2021 Comprehensive analysis of genomic diversity of SARS-CoV-2 in different geographic regions of India: an endeavour to classify Indian SARS-CoV-2 strains on the basis of co-existing mutations. *Arch. Virol.* **166** 801–812

Schwender H and Ickstadt K 2007 Identification of SNP interactions using logic regression. *Biostatistics* **9** 187–198

Shannon P, Markiel A, Ozier O, *et al.* 2003 Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13** 2498–2504

Shen-Orr SS, Milo R, Mangan S, *et al.* 2002 Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31** 64–68

Shinde P and Jalan S 2015 A multilayer protein-protein interaction network analysis of different life stages in *Caenorhabditis elegans*. *Europhys. Lett.* **112** 58001

Shinde P, Sarkar C and Jalan S 2018 Codon based co-occurrence network motifs in human mitochondria. *Sci. Rep.* **8** 3060

Shinde P, Whitwell HJ, Verma RK, *et al.* 2021 Impact of modular mitochondrial epistatic interactions on the evolution of human subpopulations. *Mitochondrion* **58** 111–122

Sievers F, Wilm A, Dineen D, *et al.* 2011 Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7** 539

Spurr L, Li M, Alomran N, *et al.* 2018 Systematic pan-cancer analysis of somatic allele frequency. *Sci. Rep.* **8** 7735

Stewart JB and Chinnery PF 2021 Extreme heterogeneity of human mitochondrial DNA from organelles to populations. *Nat. Rev. Genet.* **22** 106–118

Uetz P, Giot L, Cagney G, *et al.* 2000 A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403** 623–627

Verma RK, Kalyakulina A, Giuliani C, *et al.* 2021 Analysis of human mitochondrial genome co-occurrence networks of Asian population at varying altitudes. *Sci. Rep.* **11** 133

Verma RK, Kalyakulina A, Mishra A, *et al.* 2022 Role of mitochondrial genetic interactions in determining adaptation to high altitude human population. *Sci. Rep.* **12** 2046

Wagner A and Fell DA 2001 The small world inside large metabolic networks. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **268** 1803–1810

Wallace DC 2013 Bioenergetics in human evolution and disease: implications for the origins of biological complexity and the missing genetic variation of common diseases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368** 20120267

Wallace DC 2015 Mitochondrial DNA variation in human radiation and disease. *Cell* **163** 33–38

Wallace DC, Singh G, Lott MT, *et al*. 1988 Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science* **242** 1427–1430

Wan X, Yang C, Yang Q, *et al*. 2009 Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics* **26** 30–37

Watts DJ and Strogatz SH 1998 Collective dynamics of 'small-world' networks. *Nature* **393** 440–442

Wei W and Chinnery P 2020 Inheritance of mitochondrial DNA in humans: implications for rare and common diseases. *J. Intern. Med.* **287** 634–644

Corresponding editor: Mohit Kumar Jolly