

# Placeholder title: **Deep Sequencing Meets Structure**

Theme of issue: PPI

Deadline to send for review: Mid August

Word Limit: The aim of the manuscript is to review recent articles, with particular emphasis on those articles published in the past two years. In addition to describing recent trends, you are encouraged to give your subjective opinion of the topics discussed, although you should not concentrate unduly on your own research. Your review should be approximately **2000 words** (not including references or reference notes), with approximately **50 references** and, as such, the review is intended to be a **concise view of the field as it is at the moment**, rather than a comprehensive overview. Our audience ranges from student to professor, so articles must be accessible to a wide readership. Please avoid jargon, but do not oversimplify: be accurate and precise throughout. Occasionally, unpublished data can be referred to, but only when essential and should never be used to substantiate any significant point.

-----  
08/08/2015 Word count: 3132 (main text)

## **Abstract:**

The increasing amounts of sequenced genomic information have added a new dimension to comparative genomics and structural biology. The advent of next generation sequencing technology has made it possible to not only analyze the variation of genetic information within healthy human populations but also to map out the genetic basis for a number of diseases. However, it remains a nontrivial task to understand the evolutionary constraints preventing a particular amino acid change on a protein. In this review, we emphasize that it is essential to integrate information from sequences, structures, and interaction networks to rationalize the phenotypic effects of these variations.

The amount of genomic information is growing at an astonishing pace due to vast improvements in next generation sequencing (NGS) technology (Figure 1A) \cite{PMID:26151137}. An essential goal of these efforts is to realize the objective of personalized medicine by analyzing genetic variation within healthy human populations as well as identifying pathological disease-associated variants \cite{PMID:21706342,PMID:21383744}. A large number medically-relevant

mutations occur within proteins, some of which appear in databases such as the Online Database of Mendelian Inheritance in Man (OMIM) \cite{PMID:15608251}, the Human Gene Mutation Database (HGMD) \cite{PMID:19348700}, Humsavar \cite{PMID:19843607}, and ClinVar \cite{PMID:24234437}. It is essential to utilize structural information to rationalize the evolutionary pressure preventing these variations and for developing drugs to combat the effects of disease-causing changes to the protein sequence. However, it remains challenging to annotate the physical effects of these mutations on proteins due to the assortment of functional constraints on a protein family and an incomplete knowledge of these constraints. In particular, a mutation in a protein structure may cause local perturbations or large changes in structure or it could also have a massive impact on the protein-protein interaction (PPI) network, and each category of change adds another layer of functional constraint on the protein. Conversely, as the amount of genomic data continues to grow, we envision a future in which biologists will utilize genetic variation within human population(s) to help interpret their structural data \cite{PMID:22691493}. Population genetic analysis of variation within human proteins has already been used to identify new species-specific functional constraints within a protein family \cite{PMID:16494531}. In addition, a number of fundamental insights about biological pathways can be garnered by analyzing new loci associated with a particular disease \cite{PMID:19812666}.

### **Classical Sequence Comparison:**

Structural biologists identify functionally constrained regions within a protein family by comparing homologous sequences from different species (Figure 2a) \cite{PMID:3709526, Book:Biological Sequence Analysis}. They focus on changes that take place on the longer timescales by comparing the dominant sequence within each species rather than focusing on changes that occur within each species. Nucleotides that do not change across different species are conserved over millions of years and are hence considered to be functionally important. Due to redundancy within the genetic code, some of the changes in the coding regions are silent as they occur without a corresponding change in the protein sequence (synonymous changes). While exceptions do exist, all synonymous changes and a majority of the nonsynonymous changes are expected to be neutral. A small fraction of the nonsynonymous changes can, however, either be harmful (deleterious) or beneficial to the fitness of the species. The ratio of nonsynonymous to synonymous variants (dN/dS) is commonly utilized to characterize the selection pressure on the coding regions of the genome (Figure 2) \cite{PMID:19081788}. If the dN/dS ratio for a coding region is less than 1, it indicates that a few of these mutations are harmful or deleterious and that these changes are under negative selection. On the other hand, a dN/dS ratio exceeding unity indicates that evolution is promoting a change in the protein

sequence and that this protein is under positive selection \cite{PMID:16494531}. Proteins undergoing positive selection may improve the fitness of an organism in different environments.

### **Introduction to Population Sequencing:**

The large amount of genomic and exome sequences available recently has provided us with the unique opportunity to characterize the genetic variation within the human population. The exome comprises the coding sequences of all protein-coding genes and is equivalent to approximately 1% of the total haploid genomic sequence (30 Mb) \cite{PMID:19684571}. Due to the reduced cost of exome sequencing and clinical relevance of variation within the coding regions of the genome, it is more widely used for genetic diagnosis. The variations within an individual's genome are either acquired hereditarily (germline mutations) or they occur during the person's lifetime (somatic mutations) due to errors during cell division. While germline mutations are typically present in every cell of the person, somatic mutations are only present in certain cells and are typically not passed on to the next generation. There are approximately 74 *de novo* (new) variants that occur during each cell division \cite{PMID:22805709} and these variants can either be germline (in egg or sperm cell or during fertilization) or somatic. As only germline mutations are passed on to the next generation, somatic mutations are not under evolutionary selection.

Various studies have found that there is extensive variation in the human genome \cite{PMID:20981092,PMID:22604720,PMID:23128226,PMID:24092746}. On average, any individual genome contains 20,000-25,000 coding variants (Table 1), of which 9,000-11,000 variants are nonsynonymous. As deleterious variants are expected to be under negative selection, the frequency with which a particular variation or allele occurs is used to characterize the evolutionary pressure on a variant. Although most of the variants within any particular individual are common variants (minor allele frequency > 5%), the majority of genetic variation within coding regions is due to distinct single nucleotide variants (SNVs), each of which occur very rarely within the human population (minor allele frequency < 0.5%). About 25-50% of the rare non-synonymous variants within healthy individuals were estimated to be deleterious indicating that the human proteome is highly robust to a large number of non-specific perturbations and because most rare deleterious variants are heterozygous with the cell also containing a functional copy of the gene \cite{PMID:23128226,PMID:24092746}. Even though the amount of genomic data is increasing, about 200,000-500,000 unobserved SNVs are still discovered after each personal genome is sequenced \cite{PMID:23128226,PMID:24092746}. As deleterious mutations are enriched within rare nonsynonymous variants, we need to continue

sequencing a large number of individuals to characterize and catalog these variants and their frequency within the human population. Indeed, the number of rare variants continues to grow even after the 1000 Genomes Consortium and Exome Aggregation Consortium data (60,706 individuals) data has become available.

As such, we can turn to intra-human comparisons to uncover more human- or domain-specific features (Figure 2). There is, however, an important distinction between interpreting inter- and intra-species conservation due to the reduced timescale of evolutionary changes within a species. In general, the higher selection constraints within coding regions of the genome imply that these regions are highly conserved. While performing such an analysis, one can also align homologous coding regions within a single human genome, such as proteins originating from the same structural domain family. This can especially elucidate domain-specific features (Figure 2b). Similar to the dN/dS ratio in cross-species comparisons, selective pressure on coding regions can be quantified using fraction of synonymous to nonsynonymous polymorphisms (pN/pS) at any site (Figure 2). In addition, evolutionary pressure can also be quantified during intra-species comparison using the ratio of rare to common variants at each site as rare variants are under negative selection and an enrichment of rare variants implies that the site is under higher selective pressure. Furthermore, genomic variations that are increasing in frequency within a human population (positive selection) could help identify a new gain-of-function event (such as a new protein-protein interaction) that increases the fitness of the species. Comparative genetics/genomics studies have already uncovered a growing list of genes that might have experienced positive selection during the evolution of human and/or primates [\cite{PMID:16494531}](#). These genes offer valuable inroads into understanding the biological processes specific to humans, and the evolutionary forces that gave rise to them.

There is one additional confounding factor to consider while identifying disease-associated variants. Genes associated with a disease are identified by detecting deleterious variants affecting genes within diseased individuals more often than in healthy populations. This might be misleading, however, because the variants associated with this gene might be correlated with other unanalyzed variants in the genome. Variants that are correlated to each other in the population are said to be under linkage disequilibrium. Hence, all variants (including the variants within a gene) statistically associated with a disease might not be causative and additional analysis may be required to identify the real disease-causing mutations. It is also important to note that while a single mutation may be associated with Mendelian diseases, the clan genomics hypothesis states that the combination of common, rare, and *de novo* variants that

arose recently within a clan increases a person's risk of getting a complex disease \cite{PMID:21962505}. We need to annotate the effect of individual variants, however, before we can predict the collective outcome of a large number of *de novo* variants.

### **Deleterious Effects of Variations on Protein Function:**

Each protein has several evolutionary constraints imposed upon it based on its biological function. The effect of a deleterious variant can only be understood when all these functional constraints acting on a protein are known and can be considered. [\[\[ANS2DC: Explain how information is incomplete and we cannot explain all disease-causing mutations in HGMD for FGF receptor here\]\]](#). A sequence change should not hinder a protein from folding to its native state, bind to a specific ligand, and perform its function \cite{PMID:11295823}. As shown in Figure 1, the number of folds in the PDB database has begun to saturate implying that we can model the structures of most proteins using homology modeling. We can utilize this structural information to assess the effect of mutations on a protein's stability as nonsynonymous changes that occur within the core of the protein or variants that disrupt the secondary structure of the protein could reduce its stability. Several computational tools based on sequence conservation (inter-species or intra-species) and/or several structural features (the physicochemical characteristics of the amino acid change, solvent accessibility, secondary structure, active site annotations, and protein-protein interfaces) were developed to predict the deleterious effect of sequence variations on a protein's function \cite{PMID:19561590, PMID:20354512, PMID:17526529, PMID:19734154}. Disease-associated mutations are highly enriched in the interior of proteins (22% of all mutations in HGMD and OMIM), and active sites of proteins \cite{PMID:20981092, PMID:22604720, PMID:23128226, PMID:24092746}.

It is important to note, however, that mutations may not only affect the native state of the protein but could also affect the stability of unfolded or misfolded intermediates within the folding pathway and this is typically ignored while assessing the effect of mutations on a protein's structure. In addition, mechanistic insight into the mutation induced structural changes requires knowledge of the folding kinetics, which still remains elusive in these models. Finally, while mutations that occur on the active site of the protein reduce ligand binding, some mutations may also affect protein activity even though they occur distant to the binding site \cite{PMID:25525255}. Such mutations that affect the thermodynamic stability of different allosteric states of a protein are typically ignored while predicting the deleteriousness of a putative variant.

### **The Network as a Framework to Understand Deleterious Variants:**

While structural and sequence information are invaluable in providing a rationale for the deleterious effects of certain disease-causing and rare variations, it is much harder to interpret the phenotypic effect of an individual variant unless one also considers the cellular environment within which these variants occur. As proteins are extensively involved in protein-DNA interactions (gene regulatory network), protein-RNA interactions (post-transcriptional regulation), and protein-protein interactions (PPI) within the cellular milieu, variants that disrupt these interactions could potentially affect the viability of the cell they are present in. As this review focuses on variation within the coding regions of the genome, we refer the reader to comprehensive essays on the phenotypic effect of noncoding variation \cite{PMID:23138309, PMID:25707927} and we only focus on deleterious effects on the protein-protein interaction (PPI) network here.

Various experimental and computational approaches were applied to characterize the human PPI network \cite{PMID:16189514, PMID:25416956} and these networks have been invaluable in interpreting the role of evolutionary constraints on a protein family. In the PPI network, a node represents a protein, while edge represents an interaction between the two proteins connected by the edge. Proteins that are highly interconnected in PPI networks (hubs) are under strong negative selection constraints while proteins at the periphery of the network are under positive selection in humans \cite{PMID:18077332}. Proteins that are more central in an integrated “multinet” formed by pooling biological networks from different context (PPI, metabolic, post-translational modification, GRN, etc.) are under negative selection within human populations \cite{PMID:23505346}. In agreement with this, perturbations to hub proteins are more likely to be associated with diseases than non-hub proteins \cite{PMID:17502601}. The PPI networks are organized in a modular fashion as proteins associated with the same function are more likely to interact with one another \cite{PMID:17353930} and proteins associated with similar diseases tend to occur within the same module \cite{PMID:17502601}. The system properties of the network have also been useful in interpreting how the human proteome is robust even in the presence of a large number of deleterious variants within healthy individuals. Most deleterious variants observed in healthy individuals occur on peripheral regions of the interactome, and have marginal effects on the interactome either due to compensatory mutations or due to the interactome’s redundant nature \cite{PMID:25261458}. Meanwhile, cancer-associated somatic deleterious variations occur in the internal regions of the interactome and tend to have larger structural consequences on the PPI network.

The interactome provides a convenient effect to measure the impact of a deleterious variant on the cell, as a deleterious variant would have a larger impact on the structure of the PPI network if it occurs on a hub. A deleterious variant can either remove a protein (and all its edges) from the PPI network by making a protein nonfunctional or it could lead to the loss of just one or more of its interactions (edgetic effects). Mutations at a PPI interface can have drastic effects on the biomolecular binding constant and several sequence and structure-based methods have been proposed to identify these interaction hotspots \cite{PMID:17630824, PMID:15855251}. While the discovery of structural folds has saturated, the discovery of new domain-domain interactions continues to grow (Figure 1). Even though we have incomplete information, it has been predicted that about 12% of all the HGMD and OMIM mutations occur at a PPI interaction \cite{PMID:26027735} while approximately 28% of experimentally-tested HGMD missense mutations affect one or more interactions emphasizing the importance of these interactions for annotating rare variants and disease-associated mutations \cite{PMID:25910212}.

In an effort to bridge the information gained from individual structures with network properties of the interactome, Kim, et al., \cite{PMID:17185604} combined the experimentally determined interactome with structural information from the iPfam database to form the structural interaction network (SIN) and were able to obtain a higher-resolution understanding of the selection constraints on the hubs. Using structural information, the hubs were classified into different groups based on the number of interfaces utilized for biomolecular complex formation and they showed that the hubs with two or more interfaces are more essential than hubs with one or two interfaces. Consistent with this interpretation, hub proteins in PPI network contain a higher fraction of disease-causing mutations on their solvent exposed surface, as compared to non-hub proteins indicating that a larger fraction of a hub's disease-associated mutations could affect its interactions [18].

As hub proteins interact with a large number of partners, they tend to be more flexible and conformationally heterogeneous than non-hub proteins \cite{PMID:21826754}. Furthermore, the number of distinct interfaces in hub proteins is correlated with degrees of conformational heterogeneity \cite{PMID:21826754}. To the extent that variants may enable or disable certain conformational states from being visited, such mutations could potentially affect protein complex formation and signaling pathways, and this has not yet been examined very closely. As deleterious mutations that affect hubs in networks tend to have a larger effect on the structures of could also have large effect on these networks, such variants could also affect the phenotype

of the cell. As proteins can utilize different interfaces for different (sets of) interactions, multiple mutations on the same protein can be associated with drastically different diseases based on the interface on which they occur. Such mutations would have different edgetic effects on the protein's interaction network - by breaking or weakening one of its interactions while the rest of its interactions remain intact - and a large proportion of HGMD and OMIM mutations are predicted to have edgetic effects on the PPI network \cite{PMID:22252508,PMID:25910212}.

As a mutation typically displays tissue-specific phenotypic effects, an understanding of functional constraints on a protein should also incorporate tissue information. While the gene regulatory network is being mapped out in a developmental time point and cell type-dependent fashion by several international consortia (cite ENCODE, REMC), the PPI network is largely treated in a static fashion. Recent work has tried to integrate proteome and gene expression profiles with PPI networks to create tissue-specific networks \cite{PMID:23028288}. However, these studies typically neglect the protein isoform even though the interactions a protein is involved in is highly dependent on its isoform \cite{PMID:22749401, PMID:22749400}. A structural study on the effect of sequence variations on isoform-dependent PPI complexes has not been performed and will improve the prediction of phenotypic effects due to missense mutations. However, it is likely that the high costs (both financial as well as in terms of experimental labor) associated with studying isoform-specific assays in various cell types have impeded these types of studies. We anticipate that isoform-specific protein-protein interaction network annotation will become easier and more accessible in the near future, which will present new opportunities to better annotate such networks.

### **Effect of Mutations on Disordered Regions:**

The discovery and prominent role (>30% of eukaryotic proteome) of intrinsically disordered regions within proteins that lack an ordered three-dimensional structure, has challenged the paradigm that structure determines the function of protein \cite{PMID:11381529}. The hubs in PPI networks tend to contain higher amount of disordered regions and these regions typically gain structure only after binding to a ligand or another biomolecule \cite{PMID:18364713,PMID24606139}. The assessment of a mutation on the activity of an intrinsically disordered protein is even more challenging because it would depend upon the effect of a mutation on either the unfolded ensemble and the structure gained in the presence of its interaction partner. Due to their flexibility, the unfolded ensembles of disordered proteins are difficult to characterize using either experimental or computational techniques



\cite{PMID:19162471,PMID:22947936}. However, the phenotypic effect of mutations on the functional viability of a disordered protein is important because a number of proteins also change their interaction partners in a tissue-specific manner based upon the dominant isoform of the protein in that tissue \cite{PMID:23633940}.

### **Conclusions:**

The exponential growth in genomic data has elucidated that a surprisingly large amount of genomic variation exists within the human population and it has also helped identify a vast number of rare variants and disease-associated variants. Though the motivation of developing methods to annotate the effects of variants that cause human disease are clear, it remains challenging to do so as it requires bridging disparate sources of information together to understand the functional constraints on a protein family. The network properties of the protein along with sequence and structural information regarding the nonsynonymous amino acid change need to all be considered in a single framework before predicting the phenotypic impact of an amino acid change.