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 Occasionally, unpublished data can be referred to, but only when essential and should never be used to substantiate any significant point.

Number of words: 2730

The amount of genomic information continues to grow 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}. While a large proportion of these mutations occur in noncoding regions of the genome, a few medically-relevant mutations and rare variations occur within proteins. Many disease-associated mutations are collected 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}, and ClinVar \cite{PMID:24234437}. It is essential to incorporate structural information for inferring the mechanistic basis of the evolutionary pressure preventing these variations and for developing drugs to combat the effects of disease-causing changes to the protein sequence. In addition, we envision a future in which structural biologists will utilize the genetic variation within human population to aid the interpretation of functional constraints within a protein family. [[MG: structural biology is going to change because we will have 100s of thousands of exome data and we can understand the structure a lot better in light of this information]] [[ANS2MG: Done!]] However, it remains challenging to annotate the physical effects of these mutations on proteins due to the multi-hierarchical [[JC2ANS170715: u mean in terms of SCOP/pfam classes eg folds?]][[ANS2JC:Is it clearer now?]] nature of the structural constraints on a protein's function and an incomplete knowledge of these constraints. A mutation in protein structure may cause local perturbations or large changes in structure [[SK2ANS:global transitions??]] or it could also have a massive impact on the protein-protein interaction (PPI) network, and each kind of change adds different kinds of functional constraints on the protein.

An Abundance of Sequence Variation:

[[MG: Existing headings are those that struct biologists often see and know -- we should also incl the presentation of variation -- ie, allele frequencies, selection in a population context, etc. Human polymorphism data is not the same thing as cross-species conservation (cross-species is a much longer-term and acting set of pressures). Struct biologists are not as acquainted w/the data and thinking assoc. w/next-gen sequencing as applied to human populations. Rare mutations have different types: de-novo mutation that is disease-causing, or just benign. A lot of stuff that struct biologists don't relate to. Include JC's figures -- to some extent, this is LIKE cross-species conservation, but it is not the exact same -- cross-humans conservation is not exactly what most would think in all cases. This can play out in the context of structures. Why (ie, what are these human-specific phenomena)? Maybe b/c there's a new interaction interface that's human-specific. Or it could be POSITIVE selection, etc. Expl. to struct biologists: pilot 1000G, phase I, ExAC, etc -- what does it mean when the numbers go up -- why get more and more sequences? Partially for better significance. # variants per exome = ? How many mutations would you expect in a given structure, etc? JC can maybe fill in the numbers of common and rare variants in a typical exome. If you aggregate many people, all the variants are rare. YZ can give summary of phase 3]] [[ANS2MG: Done!]]

There is a phenomenal growth in genomic data acquisition - both in the form of whole genome and exome sequencing. 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. On average, the genome of any individual contains 20,000-25,000 coding variants (Table 1), of which 9,000-11,000 are nonsynonymous changes (i.e., result in a change in amino acid) \cite{PMID:20981092,PMID:22604720,PMID:23128226,PMID:24092746}. About 25-50% of the rare non-synonymous variants within healthy individuals were estimated to be harmful or 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 [[JC2ANS150715: also because the rare variants are usually heterozygotes; most genes are haplosufficient]] [[ANS2JC:modified]][[JC2ANS150715: I added 'deleterious' since not all rare are necessarily hetero]] \cite{PMID:23128226,PMID:24092746}. The majority of genetic variation within coding regions are due to distinct single nucleotide variants (SNVs), each of which occur very rarely within the human population (minor allele frequency < 0.5%) [[JC2ANS150715: might be less controversial to use "variants" SNVs; SNP has a traditional connotation of >=1%]] [[ANS2JC:Done!]]. A *de novo* mutation introduced into a family for the first time either due to a mutation in one of the germ cells within the parents or even within the fertilized egg can be benign or harmful depending upon its genomic location [[JC2ANS150715: why are we talking about de novo?]]. Close to one-third of the rare coding variants are predicted to be deleterious and lower the fitness of the individual \cite{PMID:22604720}. As rare variants may be involved in complex disease etiology, we need to continue to sequence a large number of individuals to characterize and catalog rare variants 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. This suggests that every individual has their own private set of variants, which is shared between very few or no individuals. [[ANS2JC: can you add something about what it means when the numbers go up?]] Because these variants are rare and novel, understanding their effect on function will be extremely

challenging. Furthermore, different genes display different amounts of variation, with some proteins being enriched in SNVs when compared with other proteins, [[JC2ANS 150717: I am dubious about this ensuing part-sentence - you mean 'neutral' or 'adaptive' mutation? I dont think there are alot of adaptive mutations; most are neutral probably or just LD passengers since in same gene]] probably because they do not affect the individual's survival or because some of them may play a role in adaptation to a particular environment. For example, some signaling and immunological proteins that sense and react to the environment are highly enriched in nonsynonymous SNVs \cite{PMID:23128226,PMID:24092746}.

Traditionally, structural biologists have utilized evolutionary conservation across species to identify functionally constrained regions within a protein family (Figure 2a). Regions that vary among different species are used to denote functionally unimportant regions. There is an important distinction between interpreting inter-species conservation and conservation within the human populations. While considering genomic variation within a species, regions under positive selection (alleles spreading [[JC2ANS150717: this is an odd word to use]] within a human population) could help identify a new function (such as a newly evolved advantageous protein-protein interaction) for the protein within the human population [[JC2ANS/MG: this sentence seems a tad odd. But I wont change for now, maybe MG has ideas. I will rework later. For now, I will just add on...]] Moreover, selective constraints, and thereby conservation, are generally high within the protein-coding regions of the genome. As such, we can turn to intra-human comparisons to uncover more human- or domain-specific features (Figure 2). For instance, by comparing conservation of homologous sequences within the human population, human-specific features can be uncovered. In contrast to sequence comparisons across species, quantification can be accomplished by using an enrichment of rare variants as a proxy for conservation. Further, one can align homologous regions within a single human genome, such as protein repeat domains originating from the same structural domain family. This can especially elucidate domain-specific features (Figure 2b). . 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.

[[JC2ANS150715: I think adding some interplays between rare v common, ns v s variants etc within the context of protein structures/PPI/isoforms and related amino acids and molecules can be nice; also maybe functional impact (SIFT, polyphen etc) based on seq conservation, structure etc; physicochemical BLOSUM]] [[de Beer, Thornton (Iastau) et al 2013, PLoS Comp Biol]]
[[JC2ANS150715: do we want a few sentences about to protein-and-seq related technology like RNA-seq?]]

Effect of Mutations on Protein Folding:
The folding of a protein into its native conformation is typically essential for its function and mutations that affect the folding of protein into the native state [[JC2ANS150715: arguably, mutations can affect also intermediate states]] [[ANS2JC: I agree and my point is that this is neglected currently. Thats how I end the section. If this is not coming out, please let me know]] can have profound effects on its activity \cite{PMID:11295823}. In

addition, mutations that induce misfolding of proteins are also associated with neurodegenerative diseases such as Alzheimer's and Parkinson's disease \cite{PMID:15931380}. The guiding principle that a given structure dictates function motivated the concept of protein folds as a means of cataloguing proteins on the basis of common structural features \cite{PMID:7723011,PMID:10775657}. We have reached a stage where the discovery of new folds has begun to saturate (Figure 1B) and the stage is set to assess the effect of mutations on the stability of these structures.

There are several computational tools that predict the effect of a mutation on structural stability (citations). Disease-associated mutations are found to be highly enriched in the interior of proteins (22% of all mutations in HGMD and OMIM) and are predicted to destabilize the protein \cite{PMID:26027735}. However, mutations not only affect the native state of the protein but 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.[[SK2ANS: done]] Furthermore these models overlooks the role of heterogeneity in the native contact energetics, which is considered essential in determining functional characteristic of proteins. In addition, mechanistic insight into the mutation induced structural changes requires knowledge of the folding kinetics, which still remain elusive in these models.

Effect of Mutations on Protein Function:
Individual X-ray structures provide only static snapshots of macromolecular architecture, yet such models may at times suffice [[JC2ANS150715: arguable?]] [[ANS2JC: modified a bit]] to elucidate the essential features regarding ligand binding. The model of DNA as a double helix may come to mind, whereby the model [[JC2ANS150715: choice of word? hmm]] [[ANS2JC: modified a bit]] hinted at the mechanism for DNA replication (Watson et al, 1953). The Uniprot database annotates the ligand binding site and post-translational modification sites that are essential for the activity of a protein. As the ligand-bound structures of all proteins have not been crystallized, homology modeling of *holo* structures can extend the ligand-binding sites for proteins with no known structure or proteins that have only been crystallized in the *apo* state (citation). Incorporation of sequence variation with structural information indicates that, as expected, rare variants are highly enriched on active sites of a protein as these mutations have a profound effect on its functional activity \cite{PMID:20981092,PMID:22604720,PMID:23128226,PMID:24092746}. In addition, missense mutations occurring on active sites may explain about 11% of the pathologic variants in the HGMD and OMIM databases while a small number of disease-associated mutations also occur on post-translational modification sites \cite{PMID:26027735}.

Effect of Mutations on Protein Dynamics:
While mutations close to the active site of a protein are relatively easy to interpret in the presence of the appropriate structure, a few variants in distal sites might may also affect its functional efficiency. Proteins are dynamic entities that constantly fluctuate among many different configurational ensembles (or thermodynamic states) at room temperature, and these dynamical fluctuations are utilized to regulate the functional behavior of proteins (citations). The conceptual framework for the understanding of proteins as structurally heterogeneous yet functionally specific macromolecules was provided by energy landscape theory (Bryngelson et al, 1995). Mutations to the protein can also affect its efficiency by affecting the dynamics or thermodynamic constant between its different states (Sarah Teichmann Science Article, 2014). While various methods have been developed and applied to identify allosteric hotspots (binding site for

allosteric ligands) (citations) and/or mutations that could affect the intracommunication pathways (citations) within known allosteric proteins, these methods have not yet been utilized to study the effect of rare variants or disease causing mutations on the functional efficiency of the corresponding proteins.

[[dc writing transition text into networks -- and mention hierarchy + ref fig.]]

Effect of Mutations on the Interactome:

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{} and we focus on the PPI network here. Various experimental and computational approaches were applied to characterize the human PPI network \cite{} and these networks have been invaluable in interpreting the role of evolutionary constraints on a protein family. Mutations at the 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{}. 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}.

While structures of individual protein-protein complexes have provided an excellent resource to interpret the effect of disease-associated mutations on individual interactions, the system properties of the network have also aided in understanding the effect of these mutations. 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{maybe see Kim et al, 2007 paper in PNAS}. 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{}. 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{} and proteins associated with similar diseases tend to occur within the same module \cite{}. 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. [[JC2ANS150715: maybe a sentence about compensatory mutations and/or redundant pathways?]] [[ANS2JC: modified next sentence]] 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.

In an effort to bridge the information gained from individual structures with network properties, Kim, et al., \cite{} 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 \cite{PMID:23505346}. One understudied mechanism by which mutations could potentially affect protein complex formation is by hindering or causing a change in the motions required during biomolecular complex formation. As hub proteins undergo larger conformational changes on binding to their interaction partners \cite{PMID:21826754}, such mutations could also have large effect on the PPI network and 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 PPI 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 significant proportion of mutations may be associated with diseases because they disrupt the interaction network of the protein. Even though the interactome remains incompletely characterized \cite{}, the underlying basis of a large number of diseases can be inferred utilizing the network context of the disease-associated biomolecules \cite{PMID:25700523}.

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{}. However, these studies typically neglect the protein isoform even though the interactions a protein is involved in is highly dependent on its isoform \cite{Kim, Babu}. 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.

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{Dunker}. 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 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}. Cancer driver mutations are enriched in these alternatively-spliced disordered motifs showing that they are important for understanding the phenotypic effects of sequence variations in the human genome.

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 together sequences, structures, and networks in a multi-hierarchical fashion to understand the functional constraints on a protein family. Ultimately, we need to develop methods to predict the phenotype from a person's genotype and allow physicians to incorporate personalized medicine in their daily practice.