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 Occasionally, unpublished data can be referred to, but only when essential and should never be used to substantiate any significant point.

07/16/2015 Word count: 3020

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}. 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, 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{} and ClinVar \cite{PMID:24234437} [[SK2ANS: Another resource,Humsavar]]. 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. 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 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 kind of change adds different kinds of functional constraints 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 molecular data \cite{PMID:22691493}[[SK2ANS: \cite{22691493}]]. [[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!]] 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}.[[ANS2MG: Tried to make this stronger!]]

[[dc2ANS]]Current overview / organization in introduction:

1) NGS is providing a lot of data, some of it gives deleterious variants in structure
2) Multiple potential aspects of constraint: destabilizing folds, interfering w/PPIs, etc
3) In the future, structural biologists will use NGS data to study proteins

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 \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%). Close to one-third of the rare coding variants are predicted to be deleterious (or harmful) and lower the fitness of the individual \cite{PMID:22604720}. A common mechanism to identify genes associated with a disease is to identify deleterious variants that affect genes within diseased individuals more often than in healthy populations. However, these variations might not be causative of a disease and may be in linkage disequilibrium (i.e., occur in a correlated fashion along with) [[SK2ANS: why throw in LD without describing.not sure structural biologist know about it]] with an unanalyzed causative mutation. 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 [[ANS2JC: If there is positive selection, isn’t some of the change leading to fitness in 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}. Hence, cataloging and characterizing the entire frequency spectrum of different variations is critical for understanding the fitness effects of different alleles.

The simple common disease-common variant hypothesis, which states that complex disease is largely attributable to a moderate number of common variants, dominated the field initially \cite{}. However, a number of studies show that common variants only explain a small minority of phenotypic difference in the human population implying that rare variants may be involved in complex disease etiology \cite{}. Hence, we need to continue sequencing a large number of individuals to characterize and catalog rare 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. This suggests that every individual has their own private set of variants and about 200,000-500,000 unobserved SNVs get discovered after each personal genome is sequenced \cite{1000 Genomes}. Because these variants are rare and novel, understanding their effect on function will be extremely challenging, and such an understanding is vital because the complex interactions between the deleterious and protective variants within a personal genome dictate the individual's health risk. As deleterious variants that are likely to increase disease susceptibility are under negative selection, the "clan genomics" concept posits that the combination of disease-causing alleles are likely to have arisen rather recently within extended familial lineages \cite{PMID:21962505}. According to this theory, the common variants within a population will have lesser influence on an individual's fitness than recently arisen rare variants and *de novo* mutations. It is imperative, however, to consider the whole collection of rare and common variants within an individual rather than to focus on specific variants. However, we need to annotate the effect of individual variants before we can predict the collective outcome of a large number of *de novo* variants.

Traditionally, structural biologists have utilized evolutionary conservation across species to identify functionally constrained regions within a protein family (Figure 2a)[[SK2ANS:we need lot of citations in this paragraph]] \cite{}. 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 human populations. While considering genomic variation within a species, regions under positive selection (alleles spreading [[JC2ANS150717: this is an odd word to use]] [[ANS2JC: I have seen it used often in literature]] within a species) could help identify a gain-of-function (such as a newly evolved protein-protein interaction) event within the human population \cite{} [[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...]] [[ANS2JC: I worked on it a bit]] Moreover, selection constraints, and thereby

conservation, are generally high within the protein-coding regions of the genome. [[ANS2JC:This is repeating some of the stuff I said earlier I think]] 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 (1000 Genomes). Furthermore, 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).

In addition to the number of common and rare variants on a gene, the ratio of nonsynonymous to synonymous variants (dN/dS) is commonly used to measure the selection pressure on the coding regions of the genome (Figure 2) \cite{PMID:19081788}. The dN/dS ratio is expected to exceed unity only if natural selection promotes changes in gene sequence \cite{PMID:865622}. 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. [[ANS2JC - can you please add a sentence next to figure ref saying something about ankyrin and how you are using it as an example to illustrate some of these points]]
[[specific examples -- ankyrin, and FGFR]]

[[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 (lastau) et al 2013, PLoS Comp Biol]][[ANS2JC/MG: I added the last paragraph and the clone genomics paragraph to highlight these points - some of the sequence based conservation part is in the next section]]
[[JC2ANS150715: do we want a few sentences about to protein-and-seq related technology like RNA-seq?]] [[ANS:I think there is no space]]

Deleterious Effects of Variations:

The effect of a deleterious variant can only be understood when all the functional constraints acting on a protein are known [[ANS2DC: Explain how information is incomplete and we cannot explain all disease-causing mutations in HGMD for FGF receptor here]]. The protein needs to

function within the cellular environment and during the course of its function, it needs to also interact with other biomolecular entities. 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 only focus on deleterious effects 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. 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. The interactome provides a good framework to measure the harmful effects of a variant. As shown in figure 4, deleterious mutations can either lead to the removal of a node (nonfunctioning protein) or the removal of an edge (a single PPI is lost).

Deleterious Effects of Variations on Nodes:

The protein sequence has several evolutionary constraints imposed upon it based on its biological function. Specifically, a sequence change should not hinder a protein from folding to its native state \cite{PMID:11295823}, bind to a specific ligand, and perform its function. If a

protein is unable to fold or function, it is equivalent to removing one node from the PPI network. While the number of structures resolved in the PDB database continues to grow, we have reached a stage where the discovery of new folds has begun to saturate (Figure 1). As a result, the stage is set to utilize this structural information to assess the effect of mutations on a protein's functional activity. Nonsynonymous amino acid substitutions that occur within the coding regions of healthy human populations is highly correlated with the frequency of amino acid occurrence in the human proteome \cite{}. Furthermore, the pattern of amino acid changes observed in inter-species sequence alignments, which is dominated by changes between chemically similar amino acids, is different from the pattern of mutations that occur within a species \cite{}. Both inter and intra-species sequence alignments of a gene or protein family are used to infer whether a naturally occurring variant would be benign or deleterious. 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 interactions) were developed to predict the deleterious effect of sequence variations on a protein's function \cite{}. 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}. 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}.

It is important to note, however, that 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. 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. Finally, mutations to the protein that occur distal to its active site can also affect its efficiency by affecting the dynamics or thermodynamic constant between its different states (Sarah Teichmann Science Article, 2014).

Deleterious Effects of Variations on Edges:

In addition to disease variants acting through disruptions in the nodes, 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}. 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{}. 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{} 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}.

The distinction between hub and non-hub proteins also extends to considerations regarding conformational heterogeneity. Hub proteins have been shown to generally exhibit greater degrees of conformational change than non-hubs. 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 signalling pathways, and this has not yet been examined very closely. As hub proteins undergo larger conformational changes on binding to their interaction partners, 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 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. 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{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 \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.