Introduction

Many well-known diseases can be linked to point mutations that affect important protein features such as enzyme active sites. A recent study from our group implicated a handful of variants in inflammatory bowel disease (IBD) via whole genome sequencing of five families with high incidence of IBD (1). One variant in particular was traced to a key structural location on the ubiquitin ligase protein TRIM11. However, with the tools available today, tracing each individual variant to its protein structural location is a complex manual task; mapping variants to structures on a large scale in order to explore patterns in their positions is currently infeasible.

Individual patients may carry point mutations that directly cause or make them susceptible to disease. An important goal of Precision Medicine is to first identify and interpret these individual mutations and then develop treatment options to mitigate their effects. The integration of genetic variation data and protein structures will help elucidate the molecular mechanisms involved in disease processes, and may lead to the discovery of drug targets for specific molecular defects. The power of this approach has been demonstrated in the treatment of cystic fibrosis (CF) patients who carry a specific gene mutation (2). By using the 3D protein structure to guide the selection of potential drug targets for specific mutations, new drugs have been developed and are now saving the lives of CF patients (http://www.kalydeco.com/how-kalydeco-works/treating-a-cftr-protein-defect).

To enable large scale and small scale interpretation of genetic variants at the protein level, we propose to interconnect four existing resources that currently do not all interoperate: the RCSB Protein Data Bank, the Kaviar database of human genetic variation, PeptideAtlas, and the Trans-Proteomic Pipeline. Individually, each of these resources supports thousands of researchers in their respective domains. We propose to develop an infrastructure to enable integration of the genomic, proteomic, and structural information in these resources, thereby enabling users to analyze, visualize and elucidate the functional and phenotypic impacts of genetic variation through the lens of protein sequence and structure.

- The **RCSB Protein Data Bank** (RCSB PDB, http://www.rcsb.org) provides access to 3D structures of biological macromolecules and is one of the leading resources in biology and biomedicine worldwide (3, 4). To enable a deeper analysis of the potential changes related to genetic variation (in the context of protein sequence and 3D structures) RCSB PDB has developed tools that facilitate mapping of any genetic location onto corresponding protein sequence isoforms and 3D protein structures.
- **Kaviar** is the largest public database of known human genetic variation (5), currently indexing 162 million single-nucleotide variants (SNVs) and 51 million short indels and substitutions, from over 77,000 individuals – over 13,000 of which have whole genome sequences. These variants are collected from 35 different data sources and normalized to maximize concordance across technologies. In addition to public data we also include private data, which is processed to maximize information to the user while minimizing privacy concerns.
- **PeptideAtlas** is a large and widely used compendium of uniformly-processed shotgun mass spectrometry proteomics data made publicly available to the community (6, 7). Users may explore the atlas for proteins or peptides of interest, explore consensus or individual spectra, download build results, or raw data. There are PeptideAtlas builds for human and 20 other species, and they include extensive information about detected variants, and post-translational modifications.
- The **Trans-Proteomic Pipeline (TPP)** (8, 9) is one of the most advanced, free and open source, complete mass spectrometry proteomics data processing suites available. It comprises tools to convert file formats, analyze spectra, perform statistical validation, and quantify analyte abundances (and more). It is actively used by thousands of researchers and provides the high-throughput infrastructure for processing all data for PeptideAtlas.

Goals and Activities

Our primary goal is to design a framework for the large-scale integration of genetic variation and 3D protein structure information, both for human research and medicine as well as for other species. This planning project will gather requirements and expertise from the community and the West Big Data Innovation Hub (WBDIH) on how a fully functional interconnected system could work, and develop a prototype infrastructure to demonstrate the feasibility of a later full implementation. We propose these activities over the one-year planning project as depicted in Figure 1:

- Create a requirements document that details various use cases, as well as the resources and infrastructure needed to support the use cases
- Develop several candidate designs for further presentation and discussion
- Arrange for the attendance to a workshop of approximately a dozen stakeholders and experts in various aspects of the proposed plan
- Host a workshop at the Institute for Systems Biology (ISB), with participation of all personnel on this proposal as well the stakeholders and experts
- Publish a workshop summary to capture the proposed ideas, reactions, discussion points, and recommendations for implementations
- Prototype software which implements parts of the framework to test its feasibility and potential value
- Prepare a manuscript, academic courses and conference presentations to disseminate our results to a community that is broader than the one included in the initial process.
- Prepare a full Big Data Spoke proposal with a broader list of participants (based on workshop participants and others identified as valuable) designed to bring our infrastructure from prototype to a community resource.

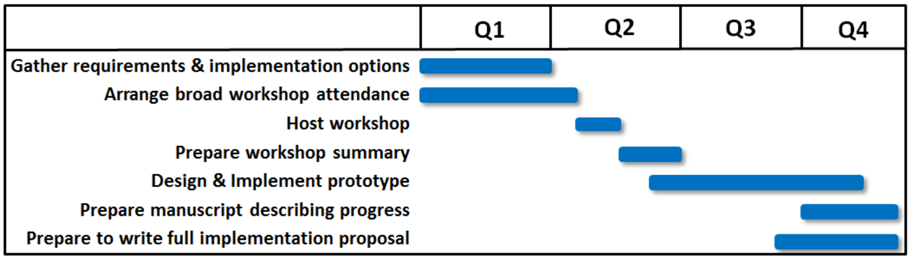

Figure 1. Proposed timeline for the activities of this one-year planning project, beginning with preparation for the workshop and culminating in the preparation of a full spoke proposal with an expanded group.

The workshop will be a major activity of this Planning Grant project, designed to gather specific feedback and additional ideas based on our initial plans from a wide range of stakeholders, domain experts, and other interested researchers. We will plan the workshop during the first quarter. This will entail gathering draft requirements and design options for the project, contacting potential participants, and arranging travel and logistics. At the workshop we will present the draft documents and ideas for broad discussion. After the workshop we will collate all ideas, critiques, and discussion into a workshop summary, which will then be the guide for prototype implementation and full spoke proposal. We will continue to coordinate with the participants after the workshop via periodic conference calls to discuss the prototype and prepare the proposal.

We have already begun contacting a few individuals with expertise in precision medicine, proteomics, genomics, protein structure, and big data technology. Letters of collaboration are attached for some of these individuals who will contribute to this project via their participation in the workshop and thereafter, including Gil Omenn (precision medicine), Emøke Bendixen (proteomics), and Hanna Carter (medical genetics).

In order to guide our design and development, we will include three main exemplar use cases.

- Kaviar data (including 5.8 million single-nucleotide variants that modify human proteins through amino acid changes) will be integrated with annotations of protein sequences and PDB structures as the model for large-scale interoperability between genomic variant and protein resources. This will enable the exploration of large-scale patterns in the locations of these variants. The information in these large-scale patterns may include variations as a function of their frequencies, affected amino acids, internal vs. external locations (active site and protein-protein interaction regions), and appearance within substructural components or domains.
- PeptideAtlas data will be integrated with PDB structures as the model for interoperability between mass spectrometry proteomics information and protein structure resources. PeptideAtlas contains information on the frequency with which individual variants have been observed in translated sequence, as well as high volumes of post-translational modifications. Similar broad patterns (described above) will emerge at the protein level, but likely with subtly different signatures.
- TPP tools will be integrated with Kaviar and PDB as the model for interoperability between user-focused data analysis tools and large-scale sequence and structure resources. This will enable researchers to view their own data in the wider context of these resources. The tools of the TPP enable researchers to analyze both RNA-seq as well as mass spectrometry data, with the intended co-analysis of the abundance and variation detected in transcripts and the abundance and variation detected at the protein level. Current tools can present detected variation and post translational modifications (PTMs) in a linear sequence context. However, a logical next step is missing: understanding the phenotypic implications of these sequence and PTM variations in the context of protein structure information.

Open source data sharing and community-driven data standards. The participating groups are leaders in defining open community standards for their respective domains. However, there are currently no good standards for mapping biological data across scientific domains and a lot of effort is required for establishing correct mappings of data. For genomic data, there are several standards for representing data in genomic coordinates, such as BAM, or bigWig files. However, there are limitations to what type of data can be represented using these file formats.

An example for a question that cannot easily be answered using current community standards is: if a site of genetic variation is projected to its position in 3D on a protein structure, what are all the amino acid residues that are in close proximity in 3D space? Do any of these sites have correlated mutations back on the level of the genome? To answer questions such as this, we not only need better standards, but also the appropriate infrastructure that allows the execution of such queries efficiently.

The activities of this grant will allow us to overcome these limitations by defining new standards based on Big Data technologies. For a larger Spoke proposal next year, we will use the initial network of key-stakeholders that we build up during this planning grant and expand into a more diverse set of participating groups. The prototype developed here will allow a demonstration of what kinds of questions can be answered by the system we are envisioning here. This will make it more convincing for a larger audience to participate in implementing and sharing a large framework for multi-scale scientific data-analysis in a larger Spoke proposal next year.

Broader impacts of the proposed work

Encouraging the genomics and MS proteomics communities to "think beyond linear". There is a strong perceived tendency for scientists in the genomic and MS proteomics communities to think about variants in linear terms, e.g. in terms of exons and introns, genomic coordinates, positions on linear protein sequences. However, the functional implications of variants and PTMs are strongly influenced by their 3D location on a protein structure. Due to the lack of readily available tools, this leap from a linear position to a 3D location is rarely made. Our infrastructure will encourage these communities to "think beyond linear".

*Fostering collaboration between several disjoint communities***.** The genomics, MS proteomics, and protein structure communities each represent vibrant and rapidly accelerating fields. Yet, the interaction between these communities is limited. By collaboratively developing an infrastructure to bring together data from diverse communities, there will be a natural effect of fostering further collaboration amongst these fields. We believe this will lead to new insights about how variation affects function.

Medicine. As precision medicine advances, it will become more common for patients to have their genome sequenced in order to assist in appropriate diagnosis and selection of treatment. The sequencing may often reveal several novel variants, the implications of which will be unclear. The mapping of these variants to proteins with known functions may begin to provide clues, but it will be far more useful to map the variants to structural locations, with tools that can predict if the variants are at sites that are likely or unlikely to disrupt the structure and function of a protein. The proposed framework will enable a deeper exploration of the relationship between genetic mutations and their consequences. The framework will allow the prioritization of SNVs for further research, for example for the analysis of somatic mutations in cancer patient datasets. SNV data can be filtered based on their protein level annotations, such as protein domains, active sites, or binding pockets. The framework will also allow novel ways to analyze genomic data not currently possible. For example, mutations that are (anti-) correlated based on their spatial proximity can be identified.

*Extending the "beyond linear" concept to research in many species***.** In addition to the benefits for medicine and research on the human system, these concepts will greatly benefit non-human research fields where the study of genetic variation is important. We include a collaboration with Emøke Bendixen of Aarhus University (letter of collaboration included) who is interested in the variants associated with several of the major breeds of cattle (*Bos taurus*). Dr. Bendixen has access to the extensive genomic sequencing of different cattle breeds as part of the 1000 Bull Genomes Project (http://www.1000bullgenomes.com). She is generating MS proteomics data on several of these breeds and will send them to PeptideAtlas to extend the current Cow PeptideAtlas. *Bos taurus* has the fifth highest number of structures of any species in PDB.

*Knowledge transfer: education***.** One of the impacts to the broader community will be through two existing courses at the Institute for Systems Biology. The curriculum for these courses evolves each year in response to the evolving science, community need, and anonymous feedback. Thus we have the flexibility to include our current work in developing the prototype, our best practices (i.e. disparate data integration and analyses, working virtually across platforms cohesively), our lessons learned (i.e. providing effective, low barrier-to-use interfaces), and outcomes (i.e. requirements gathering, manuscript) into these courses, to the benefit of the larger community. The Proteomics course, led by PI Deutsch, is taught several times per year, both at ISB and at various international locations (Vancouver, São Paulo, and Mumbai in 2015), focusing on the TPP and PeptideAtlas. We will introduce the students to variation-to-structure concepts and, eventually, teach them to use the tools to explore the phenotypic implications of the variations they discover in their datasets. The annual Systems Biology of Disease course includes a full day of lectures and hands-on "apply what you've learned" sessions to teach disparate data integration, including structural data, into analyses: "Boosting Signal in Biological Data".

Institutional roles and responsibilities

The ISB will act as the lead institution for this proposal. ISB has advanced mass spectrometry, genomics, and computational facilities, with staff who are leaders in these areas. ISB has been on the forefront of research in these areas for more than 15 years, and will bring the genomics and mass spectrometry expertise to our collaboration. Eric Deutsch will act as PI. Deutsch is a leader in the field of computational proteomics and currently serving as chair of Human Proteome Organization Proteomics Standards Initiative (http://psidev.info). Deutsch will be supported by team who are experts in genetics and whole genome analysis (Glusman, lead: Kaviar), software development (Farrah, developer: Kaviar, PeptideAtlas), and large-scale, multiple PI, multi-science project integration (Dougherty, senior program manager: a P50 National Center for Systems Biology). ISB will host the workshop.

The PDB is one of the most widely used open access, digital biological resources, serving over 300,000 users per month in over 160 countries. The role of PDB in this project will be to provide the expertise for how to integrate genomics information with protein sequence and 3D structural data in a scalable way (Prlic) and compressive structural bioinformatics and large scale distributed computing (Rose). In addition the PDB is located at the San Diego Supercomputer Center (SDSC), which is a member of WBDIH. The SDSC has world-class facilities such as cloud based storage, high performance networking, and a strong compute infrastructure. PDB will leverage these resources to scale our infrastructure.

Collaboration with West Big Data Innovation Hub (WBDIH)

Collaboration with the WBDIH will promote greater success in our goals and activities for this planning proposal. There are several advantages that the collaboration will confer. First, the WBDIH be invaluable in reaching out to an increased group of stakeholders and experts that can participate in the workshop and can advise and critique our design proposals. Since the WBDIH is a focus point for many collaborative research groups, it will be easier to link with individuals and groups interested in sequence variations and the consequences of their locations. Further, the hub itself has many experts in big data technologies, and adding such expertise to the workshop and subsequent prototyping will be valuable.

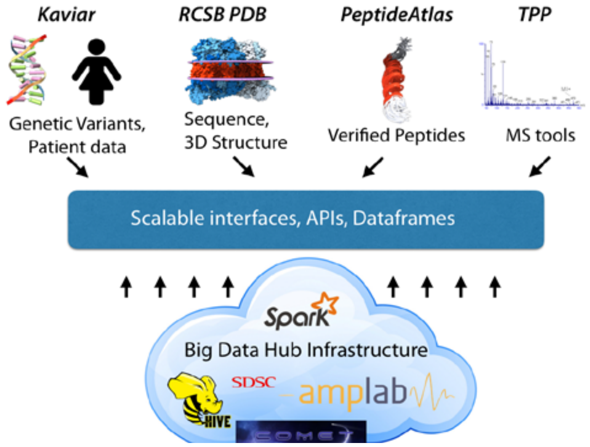

Figure 2. An outline how a big-data analytics prototype that integrates a diverse set of data resources could work. A prototype for a new technology layer will be developed, that will offer scalable interfaces, APIs, and DataFrames. This layer will provide the scalability needed to provide efficient data integration and analytics using the infrastructure provided by the West Big Data Innovation Hub.

The WBDIH will also be able to assist us in the evaluation of potential technology stacks and establish requirements for the development of a new layer on top of our existing database infrastructure that will allow us to scale queries across the diverse datasets (Figure 2). There are several key technologies that we will evaluate as part of our prototype: “Dataframes”, large “spreadsheets” that can be easily partitioned and distributed across technologies such as Apache Spark. Spark has been developed by the Amplab at UC Berkeley, which is part of the WBDIH. As such, the hub provides an excellent environment for the participating members of this spoke proposal, and the access to expertise within WBDIH on deploying these technologies will be invaluable. Other approaches that will be evaluated are high-performance APIs, that allow the retrieval of data from remote datasets, or the use of alternative data warehouse infrastructure, such as Apache Hive.

The WBDIH will also be able to assist us in the outreach and dissemination of our progress. The hub’s extensive network of traditional and social media outlets will enable us to reach many more researchers than we could on our own. And finally, the frequent workshops, conferences, and breakout sessions hosted by the hub will provide additional venues for disseminating our achievements and meeting up with other researchers interested in linking up with our team.

References

1. Stittrich, A. B., Ashworth, J., Shi, M., Robinson, M., Mauldin, D., Brunkow, M. E., Biswas, S., Kim, J.-M., Kwon, K.-S., Jung, J. U., Galas, D., Serikawa, K., Duerr, R. H., Guthery, S. L., Peschon, J., Hood, L., Roach, J. C., and Glusman, G. (2016) Genomic architecture of inflammatory bowel disease in five families with multiple affected individuals. *Human Genome Variation* 3, 15060.

2. Ren, H. Y., Grove, D. E., De La Rosa, O., Houck, S. A., Sopha, P., Van Goor, F., Hoffman, B. J., and Cyr, D. M. (2013) VX-809 corrects folding defects in cystic fibrosis transmembrane conductance regulator protein through action on membrane-spanning domain 1. *Molecular biology of the cell* 24, 3016-3024.

3. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res* 28, 235-242.

4. Rose, P. W., Prlic, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., Green, R. K., Goodsell, D. S., Westbrook, J. D., Woo, J., Young, J., Zardecki, C., Berman, H. M., Bourne, P. E., and Burley, S. K. (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* 43, D345-356.

5. Glusman, G., Caballero, J., Mauldin, D. E., Hood, L., and Roach, J. C. (2011) Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* 27, 3216-3217.

6. Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P., King, N. L., Eng, J. K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Hood, L., Katze, M. G., Kennedy, K. A., Kregenow, F., Lee, H., Lin, B., Martin, D., Ranish, J. A., Rawlings, D. J., Samelson, L. E., Shiio, Y., Watts, J. D., Wollscheid, B., Wright, M. E., Yan, W., Yang, L., Yi, E. C., Zhang, H., and Aebersold, R. (2004) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol* 6, R9.

7. Deutsch, E. W., Sun, Z., Campbell, D., Kusebauch, U., Chu, C. S., Mendoza, L., Shteynberg, D., Omenn, G. S., and Moritz, R. L. (2015) State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. *J Proteome Res* 14, 3461-3473.

8. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 1, 2005 0017.

9. Deutsch, E. W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z., and Moritz, R. L. (2015) Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics. Clinical applications* 9, 745-754.