CAGI✱✱ Workshop

Accepted Abstracts

Northeastern University
East Village, 17th Floor
291 St. Botolph Street, Boston, MA 02115

September 29-October 1, 2023

The role of maternal-fetal genetic incompatibility in Gestational Diabetes mellitus

Maha Aamir* and Rafael F. Guerrero
North Carolina State University
maamir@ncsu.edu

Complications during pregnancy can lead to harmful outcomes for both the mother and the offspring. One suggested factor underlying these adverse pregnancy outcomes (APOs) is the incompatibility between maternal and fetal genotypes, which have been implicated in preeclampsia, spina bfida, schizophrenia, low birth weight, and gestational diabetes mellitus (GDM) among others. Here, we study maternal-fetal genetic interactions associated with GDM in nulliparous women, focusing on participants of the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be (nuMoM2b) cohort. We leverage genotype data from 3869 mother offspring dyads to infer previously unreported genetic interactions and to evaluate the enhancement of polygenic risk scores in GDM. Preliminary results indicate that incorporation of maternal fetal interactions in a logistic regression model for PRS of GDM improves risk prediction for individuals in the numom2b cohort. With calculations of the area under the receiver operating characteristic curve (AUC) being 0.6413 when incorporating information from both maternal and fetal sources as opposed to 0.6112 when relying solely on maternal genetic information. Our results suggest that a better understanding of genetic interactions in the maternal-fetal unit can aid in understanding the etiology of adverse pregnancy outcomes.

Inferring the phenotypic impact of variants of liquid–liquid phase separation proteins

Peiran Jiang[1], Jose Lugo-Martinez*

Computational Biology Department, School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
[1] peiranj@andrew.cmu.edu
* jlugomar@andrew.cmu.edu

Background Liquid–liquid phase separation (LLPS) compartmentalizes nucleic acids and proteins into functional micron-scale bodies. Such liquid-like, membraneless biomolecular condensates organize various biochemical reactions in cells. Mounting evidence has shown that genetics variants could alter the phase separation behavior of proteins and lead to aberrant LLPS-driven protein aggregation or fibrillation. However, computational predictors dedicated to inferring and monitoring pathogenic variants of LLPS are still lacking, which hinders the genotype-phenotype link discovery and potential diagnostic. With the help of current curated condensate and human mutation recourses, we are fortunate to model and infer the impact of variants with both biological principles and machine intelligence.

Goals Our goal is to develop a novel machine learning tool to identify pathogenic variants of LLPS proteins and provide interpretable mechanism underlying functional or structural pathogenic alterations. Prioritized mutations are missense mutation, frame-shift mutation, and non-frame-shit mutation respectively. All the genetic variants could become candidates for experimental assessment and preventative screening markers in LLPS-associated diseases.

Method

Data resource preparation We constructed a comprehensive data recourse of disease associated genetics variants in phase separation proteins such as genes, post-translational modifications (PTMs), single nucleotide polymorphisms (SNPs), etc. **Ontology construction**. We will construct and optimize an ontology based on three aspects of properties. 1. Multivalent driving forces of LLPS including pi-pi interaction, charge-charge interaction, cation-pi interaction, dipole-dipole interaction, and intrinsically disordered regions (IDRs). 2. Oligomerization. 3. Functional residues including macromolecule binding (DNA, RNA, and proteins) and PTM sites. **Algorithm development.** We are computationally inferring the phenotypic impact of a variant in LLPS proteins and explain the effect based on the ontology.

Results and Implications

Resulting in the rank of property scores to evaluate the concordance of reported mechanisms of diseases, we can identify the high confident pathogenic effect and putative mechanisms of LLPS related diseases immediately without experiments. As part of the CAGI challenges of variants understanding and interpretation, the tool could also be used as a highly specified genome interpreter to guide pre-diagnostic screening of LLPS-associated diseases clinically.

Joint analyses of phenotypically diverse Mendelian disease patients reveals significantly perturbed genes and pathways

Shilpa Nadimpalli Kobren[1,†], Mikhail Moldovan[1,†], Rebecca Reimers[2], Daniel Traviglia[1], Xinyun Li[3], Richard Sherwood[4], Joel Krier[5], Isaac S. Kohane[1], Undiagnosed Diseases Network Tool Building Coalition Working Group, Shamil R. Sunyaev[1,*]

1 Department of Biomedical Informatics, Harvard Medical School, Boston, MA
2 Scripps Research Translational Institute, La Jolla, CA
3 Computational Biology & Bioinformatics Program, Yale University, New Haven, CT
4 Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
5 Department of Genetics, Atrius Health, Boston, MA

†Lead authors: shilpa_kobren@hms.harvard.edu and mikhail_moldovan@hms.harvard.edu
*Corresponding author: ssunyaev@hms.harvard.edu

Rare genetic conditions are estimated to affect nearly 1 in 20 people worldwide. Despite their cumulative prevalence, patients are still typically evaluated case-by-case or in small silos of similarly-presenting individuals. Analyses of genomic data across cohorts of undiagnosed patients has the potential to reveal regions that harbor a higher mutational burden than expected; these recurrently perturbed genes and gene pathways would be strong diagnostic candidates even across diversely-presenting patients. Indeed, previous cohort-based analyses of neurodevelopmental cases, though limited to exome-only data, revealed that newly-associated disease genes result in more heterogeneous patient presentations than well-established disease genes. Here, we harmonize and jointly call whole genome variants from a cohort of 4,300 patients enrolled in the Undiagnosed Diseases Network. To increase our ability to detect perturbed genes with significance, we incorporate a highly accurate genome-wide mutational model with basepair resolution, Roulette, and newly consider non-exonic variants with predicted splicing impacts. We verify the functional impact of many of these deeply intronic variants across a broad range of SpliceAI scores through the design and implementation of a massively parallel splicing assay. We find five genes across fourteen patients with a significant excess of deleterious *de novo* variants, recapitulating six known diagnoses and highlighting three new likely diagnoses. We additionally introduce a rigorous approach to identify statistically surprising compound heterozygous variants in individual patients, uncovering about ten known and new strong candidate diagnoses. Finally, we identify recurrently perturbed gene pathways by selecting genes harboring compelling variants per patient and applying gene set enrichment analyses on patient subgroups. We find six enriched pathways covering 21 distinct genes across 20 patients, recapitulating six known diagnoses, lending further support to nine previously-identified candidate diagnostic genes, and revealing four new likely candidates in undiagnosed patients. Our approach includes an expert clinical evaluation of the phenotypic concordance between candidate diagnostic variants and affected patients. To streamline and scale this typically arduous and time-consuming process, we develop and calibrate a semi-quantitative, objective evaluation metric to reduce manual case review time. When applied to the candidate diagnostic variants highlighted by our cohort-based analyses, our clinical evaluations revealed a high correlation between top candidates and diagnostic feasibility. Taken together, these results demonstrate that comprehensive, joint analyses of diverse, suspected Mendelian disorders can provide valuable insights into new disease genes and pathways, ultimately paving the way for improved diagnostics and targeted therapeutics.

DITTO: Automated tertiary rare disease diagnosis pipeline using explainable Machine Learning

Tarun Karthik kumar Mamidi, Manavalan Gajapathy, Elizabeth A. Worthey*
Center for Computational Genomics and Data Science, The University of Alabama at Birmingham, 912 18th St S, Birmingham, AL, 35233, USA
Department of Genetics, Heersink School of Medicine, The University of Alabama at Birmingham, 720 20th Street S, Birmingham, AL, 35294, USA

* Correspondence: eaworthey@uabmc.edu

Introduction

The primary goal of rare disease diagnostics is identification of molecular variants responsible for the patient's clinical presentation. Manual interpretation of hundreds of variants even after filtering is time consuming. Methods and tools that can aid in accurate and efficient diagnosis and prognosis, even in presence of high degree of variability in phenotype presentation, are of critical importance. We hypothesized that application of machine learning based approaches based on omic and phenotypic data when combined with existing experimentally and clinically derived knowledge would allow us to accurately identify pathogenic variants and generate prognostically useful information.

Methods

We developed a neural network-based model (DITTO) to predict the likelihood of variant being deleterious using several publicly available databases. DITTO was trained and tested on 696,546 known pathogenic and benign variants reported in ClinVar, which had been annotated with various variant frequency, impact, damage predictions, and disease association using OpenCravat. We built a pipeline using nextflow that can annotate each CAGI6 RGP whole genome and makes deleterious predictions for all variants. We then used HPO terms to rank the predicted deleterious variants based on phenotype matching for accurate diagnosis.

Results

Our "DITTO" model demonstrated exceptional performance during training with 0.99, 0.99, and 0.99 for Precision, Recall, and Accuracy scores respectively, outperforming other classification methods. We then tested DITTO on 35 CAGI6 RGP probands and were able to accurately classify 91% (38/42) variants with probability score > 0.99. We then tested the pipeline on 30 test proband datasets and accurately classified 77% (20/26) variants with probability score > 0.99. Lowering the probability threshold and applying an existing rule based weighting supported classification of additional variants. A notable percentage of the variants that remained misclassified were found to be intronic; these need further investigation to evaluate their pathogenicity.

Conclusions

We developed a neural network-based method to predict variant deleteriousness with integration of existing functional knowledge to aid in functional impact prediction. To validate our method, we used DITTO to classify and prioritize variants in 65 CAGI6 RGP samples. With the highest stringency cutoff we accurately classifying 81% (58/68) variants as solves. Application of DITTO also supported identification of potentially misclassified variants. These advances are of critical importance in the rapid, sensitive, and accurate diagnosis of patients with rare diseases. We will discuss methods and findings.

Optimizing variant impact prediction in autoinflammatory disease

Brynja Matthiasardottir[1, 2], Daniel L. Kastner[2], and Stephen M. Mount [1]

1. Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, 20742, USA
2. Inflammatory Disease Section, National Human Genome Research Institute, Bethesda, MD, 20892, USA.

Autoinflammatory diseases are immunological disorders caused by dysregulation of the innate immune system, characterized by seemingly unprovoked attacks of inflammation without evidence of infection, high-titer autoantibodies, or antigen-specific T cells. The identification of genes underlying these diseases has revolutionized our understanding of innate immunity and provided the basis for life-saving therapies. We have sequenced the exomes of approximately 2000 patients and their unaffected family members. Exomes have been analyzed for protein coding variants by family-based and genotype-first approaches. The diagnostic yield of our cohort remains low, and many variants of uncertain significance (VUS) occur in known inflammatory genes.

We have found that validated pathogenic variants in genes associated with the immune system often score below standard thresholds for variant impact prediction. This indicates that we might be missing variants, or a diagnosis, due to a high false negative rate in the genes underlying these illnesses. Methods for variant impact prediction have relied on training models of known variants found in ClinVar. Variants associated with autoinflammatory diseases are underrepresented in ClinVar, so it is possible that existing mutation pathogenicity prediction tools are not well-suited to these genes. By statistically comparing experimentally validated variants found in ClinVar to validated variants found in the Infevers database, we've shown that the optimal accuracy thresholds for binary classification of variants associated with autoinflammation are lower than those for other genes. The optimal accuracy thresholds of predictors such as REVEL and CADD in autoinflammatory genes are 0.324 and 17.24 vs. 0.65 and 20.0, respectively. This same trend is seen for ten other predictors included in the analysis. The area under the curve is the same, indicating that we do not introduce more noise. We speculate that this difference is due to the nature of selection on these genes, with episodes of strong selection for increased inflammation due to pathogen exposure balanced by selection for reduced inflammation otherwise, which may change the relationship between conservation and functional importance. In support of the hypothesis that autoinflammatory genes are under strong but variable selection, we have found that many genes with extreme extended haplotype (iHS) scores are involved in innate immunity. We are exploring the relationship between measures of selection and optimal accuracy thresholds further.

Our data support the idea that optimal variant curation guidelines may depend on the history of selection for some genes, including those affecting autoinflammatory disease.

ReCIPE: Reconnecting Clusters In Protein Embeddings

Faith Ocitti, Charlotte Versavel, Lenore J Cowen*

Department of Computer Science, Tufts University, Medford, MA 02140
Corresponding authors: cowen@cs.tufts.edu

The Diffusion State Distance (DSD) measure introduced by Cao et al. [1] was paired with spectral clustering in the resulting similarity space to cluster proteins into shared pathways in a top-performing method for the 2016 DREAM disease module identification challenge [2]. The resulting set of non-overlapping clusters in addition to revealing groups of proteins that were involved in disease, was shown to create groups of proteins that were functionally enriched. However, a drawback of the method was that many clusters contained disjoint components, i.e. disconnected groups of proteins that do not have edges between them. This led to clusters that are functionally enriched but not easily biologically explainable. We theorized that the non-overlapping clusters often consisted of nodes that shared connections to the same set of minor hub proteins, and returning these proteins to the clusters would both improve their functional enrichment as well as make the clusters more biologically meaningful.

In order to test this, we introduced ReCIPE. ReCIPE (Reconnecting Clusters in Protein Embeddings) re-establishes the relationships between low-degree hub proteins and disjoint components within a cluster. We accomplish this by relaxing the condition that the clusters are non-overlapping but search for new cluster members in a parsimonious fashion; only including new proteins that best reconnect the existing clusters.

When given a list of disjoint clusters, ReCIPE generates overlapping clusters in three key steps. First, we generate a list of candidate proteins for each cluster. A “candidate protein” is any extra-cluster protein considered for addition. In order to be considered, the protein must have edges to a particular ratio of disjoint components. Second, the candidate proteins are ranked. This work explored three ranking measures: protein degree, number of connecting components, and finally, a combination of protein degree and connecting components. Lastly, we greedily add proteins to the cluster until a stop criterion is met.

We demonstrate that ReCIPE leads to better functional enrichment than the original clusters and better functional prediction accuracies across the DREAM 1, DREAM 2 and DREAM 3 data sets. Moreover, the overlapping clusters show a higher level of connectivity compared to the original clusters.

References

[1] Cao et al., PloS one 8, no. 10 (2013): e76339.

[2] Choobdar et al., Nature Methods 16, no. 9 (2019): 843-852.

Evaluating the clinical utility of missense variant effect predictors for CAGI6

Ruchir Rastogi[1]*, Ryan Chung[2], Sindy Li[3], Steven Brenner[2,3], Nilah Ioannidis[1,2,4*]

[1]Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA, USA.
[2]Center for Computational Biology, University of California Berkeley, Berkeley, CA, USA.
[3]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA, USA.
[4]Chan Zuckerberg Biohub – San Francisco, San Francisco, CA, USA.

Corresponding author emails: nilah@berkeley.edu, ruchir_rastogi@berkeley.edu

Regular, systematic characterization of the performance of computational tools used to predict the pathogenicity of missense variants is necessary to evaluate their clinical utility and suggest directions for future improvement. Here, we benchmark missense variant effect predictors on an evaluation dataset of rare missense variants with high-quality pathogenicity and benignity assertions made after the close of the CAGI6 challenge in October 2021. Our assessment includes models submitted to the CAGI6 annotate-all-missense challenge, other predictors commonly used by the clinical genetics community, and recent deep learning methods for variant effect prediction. We find that meta-predictors generally outperform individual predictors. Based on overall AUROC on our evaluation dataset, high performing meta-predictors include ClinPred, MetaRNN, and BayesDel_addAF and high performing individual predictors include Model 1 submitted by Team 1, MutPred, and VEST4. However, we find that model rankings differ when we focus on the high-sensitivity and high-specificity regions of the ROC curve separately, suggesting that different models may be best suited to different clinical applications. We also show that top-performing meta-predictors that incorporate allele frequency as a predictive feature often struggle to distinguish pathogenic variants from very rare benign variants, which is an important use case in clinical practice. Lastly, we calibrate model predictions to determine score thresholds that correspond to ACMG/AMP evidence strengths. Together, these results help illuminate the clinical utility of missense variant effect predictors and identify settings in which these models can be improved.

Prioritizing genomic variants through neuro-symbolic, knowledge-enhanced learning

Azza Althagafi[1,2], Fernando Zhapa-Camacho[1], Robert Hoehndorf[1]*
* Corresponding author
[1]Computer, Electrical and Mathematical Sciences Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia. [2]Computer Science Department, College of Computers and Information Technology, Taif University, Taif 26571, Saudi Arabia.
robert.hoehndorf , azza.althagafi , fernando.zhapacamacho@kaust.edu.sa

Whole-exome and genome sequencing have become common tools in diagnosing patients with rare diseases. Despite their success, this approach leaves many patients undiagnosed. A common argument is that more disease variants still await discovery, or the novelty of disease phenotypes results from a combination of variants in multiple disease-related genes. Interpreting the phenotypic consequences of genomic variants relies on information about gene functions, gene expression, physiology, and other genomic features. Phenotype-based methods to identify variants involved in genetic diseases combine molecular features with prior knowledge about the phenotypic consequences of altering gene functions. While phenotype-based methods have been successfully applied to prioritizing variants, such methods are based on known gene--disease or gene--phenotype associations as training data and are applicable to genes that have phenotypes associated, thereby limiting their scope. In addition, phenotypes are not assigned uniformly by different clinicians, and phenotype-based methods need to account for this variability. We developed an Embedding-based Phenotype Variant Predictor (EmbedPVP), a computational method to prioritize variants involved in genetic diseases by combining genomic information and clinical phenotypes. EmbedPVP leverages a large amount of background knowledge from human and model organisms about molecular mechanisms through which abnormal phenotypes may arise. Specifically, EmbedPVP incorporates phenotypes linked to genes, functions of gene products, and the anatomical site of gene expression, and systematically relates them to their phenotypic effects through neuro-symbolic, knowledge-enhanced machine learning. We demonstrate EmbedPVP's efficacy on a large set of synthetic genomes and genomes matched with clinical information.

Disease associations of proteins with multiple functions

Giulia Babbi*, Elisa Bertolini, Castrense Savojardo, Pier Luigi Martelli*, and Rita Casadio*

Biocomputing Group, FABIT- University of Bologna, Italy

*E-mail:
GB giulia.babbi3@unibo.it
PLM pierluigi.martelli@unibo.it
RC rita.casadio@unibo.it

The characterization of multifunctional proteins is an expanding research area aiming to elucidate the complexities of biological processes. We introduced the term “multifaceted”, to collect under a unique label multitasking proteins addressed in literature as pleiotropic, multidomain, promiscuous (concerning enzymes catalysing multiple substrates) and moonlighting (with two or more molecular functions).

We recently developed MultifacetedProtDB (https://multifacetedprotdb.biocomp.unibo.it), a curated database providing a comprehensive collection of 1103 multifunctional human proteins, of which 812 are enzymes. MultifacetedProtDB has been built by:

i) merging the available public dataset, MoonDB (http://moondb.hb.univ-amu.fr), MoonProt (http://www.moonlightingproteins.org), and MultitaskProtDB II (http://wallace.uab.es/multitaskII), that when restricted to human collect 47, 103, and 185 proteins, respectively;

ii) searching new multifunctional proteins reported in the recent literature;

iii) collecting enzymes endowed with multiple EC codes.

We collected associations among proteins in MultifacetedProtDB and diseases merging the information reported in UniProt, Humsavar, Monarch initiative, and ClinVar and reported the disease nomenclatures provided by MONDO ontology, OMIM and Orphanet catalogues, and the ICD10 classification scheme. Some 30% of proteins in our database (321 enzymes and 110 non-enzymes) are associated with 895 MONDO diseases classified into 213 ICD10 categories and in 17 out of the 19 ICD10 main chapters, after excluding chapters not describing diseases with a genetic component (namely, XX: “External causes of morbidity and mortality”, XXI: “Factors influencing health status and contact with health services”, and XXII: “Codes for special purposes”). The most represented chapter is XVII: “Congenital malformations, deformations and chromosomal abnormalities” accounting for 226 diseases associated with 135 multifaceted proteins, followed by IV: "Endocrine, nutritional and metabolic diseases", VI: "Diseases of the nervous system", II: "Neoplasm", with 134, 105, and 48 associated multifaceted proteins, respectively. Out of the 895 diseases, 323 are included in the Orphanet catalogue of rare diseases. Overall, when considering the phenotypic characterization of diseases, 430 proteins are linked to terms out of the Human Phenotype Ontology (HPO). Proteins with specific domains seem more involved than others in diseases (e.g.: PF00029 -Connexin, PF00069 -Protein kinase domain, and PF00067 - Cytochrome P450, associated with 49, 24 and 19 diseases respectively).

The database is useful for characterizing the involvement of a multitasking protein in the cell molecular complexity and associated diseases. A list of drugs is provided, when available. Some protein-coding genes of interest in past CAGI challenges are multifaceted proteins, included in Multifaceted DB, such as TP53, CDKN2A, BRCA1, LDLR, PTEN, and MAPK1.

The actual content of 1103 proteins, inclusive of 812 enzymes, is a possible underestimation of the total number of multifaceted proteins in human reference proteome, and possibly future research will enlarge the collection and improve their annotation, particularly in relation to different isoforms and their involvement in disease.

Accurate proteome-wide missense variant effect prediction with AlphaMissense

Jun Cheng[1]*, Guido Novati[1], Joshua Pan[1]†, Clare Bycroft[1]†, Akvilė Žemgulytė[1]†, Taylor Applebaum[1]†, Alexander Pritzel[1], Lai Hong Wong[1], Michal Zielinski[1], Tobias Sargeant[1], Rosalia G. Schneider[1], Andrew W. Senior[1], John Jumper[1], Demis Hassabis[1], Pushmeet Kohli[1]*, Žiga Avsec[1]*

[1]Google DeepMind, London, UK; †These authors contributed equally to this work
Corresponding authors: jucheng@google.com (J.C.); pushmeet@google.com (P.K.); avsec@google.com (Z.A.)

The vast majority of missense variants observed in the human genome are of unknown clinical significance, with only an estimated 2% (of ~4 million) clinically classified as pathogenic or benign. This gap remains an ongoing challenge in human genetics, limiting the diagnostic rate of rare diseases, as well as the development or application of clinical treatments that target the underlying genetic cause. While multiplexed assays of variant effect (MAVE) systematically measure protein variant effects and can accurately predict the clinical outcomes of variants, a proteome-wide survey of variant pathogenicity remains incomplete due to the cost and labor of such experiments. Machine learning approaches could close this variant interpretation gap by exploiting patterns in biological data to predict the pathogenicity of unannotated variants.

We present AlphaMissense, which combines advances of the highly-accurate structure prediction model, AlphaFold, and population variant data to predict missense variant pathogenicity. We demonstrate state-of-the-art predictions on clinical variant labels and experimental MAVE benchmarks, without explicitly training on such data. We calibrated our predictions against a gold standard set of clinically curated pathogenic and benign variants, such that AlphaMissense scores (ranging between 0 and 1) can be interpreted as the approximate probability of a variant being clinically pathogenic. We used these calibrated scores to classify variants into three discrete categories consistent with American College of Medical Genetics (ACMG) guidelines: likely pathogenic, likely benign, and ambiguous, using score thresholds chosen to have 90% expected precision estimated on ClinVar. Due to higher predictive performance, the fraction of ClinVar test variants that we can confidently classify with 90% precision has increased by 25.8 percentage points (from 67.1% to 92.9%) compared to the recent well-performing unsupervised model EVE.

We make AlphaMissense predictions for all possible single amino acid substitutions in the human proteome available for the research community. This includes confident classifications for 89% of all 71M possible missense variants. Combined with other data, these predictions could assist clinicians in prioritizing *de novo* variants for rare disease diagnostics. Among clinically-actionable genes (ACMG) we confidently classify the vast majority of all possible missense variants as either ‘likely pathogenic’ or ‘likely benign’, and outperform competitor models for the majority of these genes. Further, we find that gene-specific AlphaMissense scores are predictive of genes essential to cell survival, and this property holds amongst the 22% of smaller genes, which methods based only on population cohort data lack statistical power to detect reliably. AlphaMissense predictions could also boost power in studies of complex trait genetics that utilize annotations of rare, likely deleterious mutations. Variants classified as ‘likely pathogenic’ have similar rates of disease associations as predicted loss of function (pLoF) variants, thus representing a much larger pool of variation that can be exploited to discover disease-linked genes. Our predictions of all amino acid substitutions could also be used as a starting point for designing and interpreting MAVE experiments, thereby accelerating efforts to understand the molecular effects of variants on protein function.

Machine Learning Study of Allosteric Sites Prediction in the Structures of Gene Products

Estelleta Hackshaw and Mary Jo Ondrechen*

Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115 USA

Estelleta Hackshaw: hackshaw.e@northeastern.edu
Corresponding author: Mary Jo Ondrechen mjo@neu.edu

Allostery is a very important phenomenon to understand how biochemical processes are regulated. Allostery controls the activity and shape of an enzyme via a ligand that can be bound to a site other than the biochemically active site. Allosteric enzymes thus incorporate both an active site and allosteric site. These enzymes are essential for control of cellular processes as well as potential targets for allosteric modulators for drug discovery and properties for materials chemistry. Allosteric sites of allosteric enzymes will be predicted using the machine learning method called POOL (Partial Order Optimal Likelihood). We have identified twenty-three enzymes known to be allosterically regulated where the three-dimensional structures of both the active and inactive forms have been reported and with a noticeably small difference between the structures of the two states (< 1 Å RMSD). Results are presented for three-dimensional structures of the active as well as inactive states. It is shown that both catalytic and allosteric sites may be predicted by POOL for these structures, using computed electrostatic and chemical properties as input features. Supported by NSF MCB-2147498

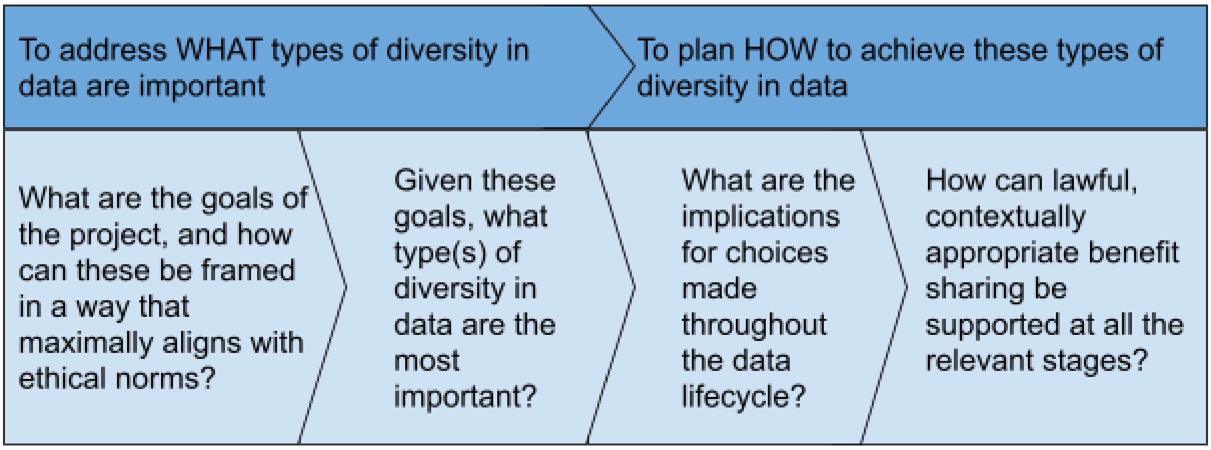
Considerations for how Genetic and Genomic Researchers Should Approach Thinking About Diversity in Data

Anna C F Lewis, Brigham and Women's Hospital and Harvard Medical School

Poster submission

In this poster I will share a newly developed policy framework by the Global Alliance for Genomics and Health (GA4GH) concerning diversity in data. At the time of the workshop this policy framework will just be pending final steering committee approval.

Calls to increase the diversity in datasets used for research are prominent within the genomics community. By increasing diversity in genomic data, the hope is that the benefits of advances in genetics and genomics can be applied and distributed more equitably and reliably. However, many of these calls do not specify what exactly is meant by diversity in datasets, and when this is specified, usually the language of ancestral diversity is used. However, diversity of genetic ancestry is only part of the need. Moreover, much care is needed in how human diversity is framed: the concepts and terms we use matter. This policy makes two contributions. First it proposes that researchers identify **what** types of diversity are important in data, via considering the outcome that this diversity is designed to deliver. This will often be linked to achieving equity. A consequence of this approach is that it demonstrates that diversity is not necessarily about representativeness. Second, this policy proposes a systematic way for **how** researchers can enable achieving the benefits that the identified diversity is designed to enable. Researchers must reckon with the complex reasons for the lack of diversity present in current data. We emphasize that data collection is only one stage within the data lifecycle, and consideration is needed across all stages. We also emphasize that attention to lawful, contextually appropriate benefit sharing is needed across each of these stages. We illustrate this proposal using two examples: Minimizing reporting false positive pathogenic variants, and the clinical impact of differential predictive performance of polygenic scores.



To address WHAT types of diversity in data are important

To plan HOW to achieve these types of diversity in data

What are the goals of the project, and how can these be framed in a way that maximally aligns with ethical norms?

Given these goals, what type(s) of diversity in data are the most important?

What are the implications for choices made throughout the data lifecycle?

How can lawful, contextually appropriate benefit sharing be supported at all the relevant stages?

Figure 1: Questions genetic and genomic researchers should answer when thinking about diversity in data. *We note that these steps can be iterated on.*

VIPdb 2.0, a genetic Variant Impact Predictor Database

Yu-Jen (Jennifer) Lin[1,†], Arul S. Menon[1,2,†], Zhiqiang Hu[3], Constantina Bakolitsa[3], Steven E. Brenner[1,*]

[1] Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA

[2] Division of Data Science, University of California, Berkeley, California 94720, USA

[3] Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA

[†] Yu-Jen Lin and Arul S. Menon are joint first authors.

[*] Steven E. Brenner is the corresponding author.

Steven E. Brenner: brenner@compbio.berkeley.edu

Yu-Jen (Jennifer) Lin: jenniferyjlin@berkeley.edu

Genome sequencing unveils a vast number of genetic variants. Predicting the molecular and clinical effects of these variants is one of the preeminent challenges in the realm of human genetics. Accurately predicting the impacts of genetic variants deepens our understanding of how genetic information is conveyed to molecular and cellular processes and stands as a critical milestone on the path toward precision medicine. We summarized over one hundred tools/resources developed specifically for this purpose in 2019 (Hu *et al.*, 2019). Building upon this foundation, our current work delves further, revealing the emergence of an additional three hundred variant impact prediction tools that have been developed after 2019. These tools, along with their distinctive characteristics, have been summarized within the genetic Variant Impact Predictor Database version 2.0 (VIPdb 2.0). By curating this database, we aim to provide researchers and clinicians with a resource that facilitates the exploration and selection of the most fitting tools for their specific inquiries. Moreover, VIPdb aspires to catalyze the evolution of methodologies by informing development strategies of current variant impact predictors.

Abstract:

In the latest SCOPe (Structural Classification of Proteins — extended) stable release, we identified and labeled heterogeneous domains that exhibit distinct structural characteristics within the same SCOPe family. Notably, some domains share similar sequences but display significant structural heterogeneity. This observation prompted us to conduct structure-based alignments, aiming to extract additional structural insights and potentially enhance the existing sequence-based automated classification pipeline. In this study, we evaluated three widely-used structural alignment algorithms: DaliLite, FATCAT, and Combinatorial Extension (CE), encompassing six distinct implementations. Our benchmarking efforts focused on a challenging SCOPe subset consisting of primarily manual curated domains.

Our analysis reveals that DaliLite outperforms its counterparts in terms of overall performance. However, it has limitations as it can only align proteins compatible with DSSP. For instance, it cannot align CA-only structures. While FATCAT-flexible offers enhanced alignment flexibility, its vulnerability to false positives often outweighs its advantages. In addition, all programs assessed in this study performed less effectively on all-alpha domains (class a in SCOPe). It appears that they tend to emphasize similarities in individual helical structures over the overall topology. A noteworthy finding is that all six implementations successfully identified some homologous structures that evaded detection by BLAST, primarily due to the absence of significant sequence similarity. These results underscore the importance of defining criteria when employing structural alignment tools and highlight the value of complementing sequence-based methods with structural approaches for comprehensive protein analysis.

Machine Learning for protein function prediction and for understanding gene product function

Lakindu Pathira Kankanamge, Atif Shafique, Suhasini M. Iyengar, Kelly K. Barnsley, Penny J. Beuning, and Mary Jo Ondrechen*

Corresponding author: Mary Jo Ondrechen mjo@neu.edu

Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115 USA

Our Machine Learning methodology, Partial Order Optimum Likelihood (POOL) is used to predict biochemically active amino acids in the structures of gene products (proteins). Computed electrostatic and chemical properties of individual amino acids serve as input features. Our most recent applications of POOL are described. From predicted local sites of biochemical activity, the biochemical functions of Structural Genomics proteins of unknown function are predicted by local structure matching of predicted spatial arrays of active amino acids with those of proteins of known function. Examples from the Haloacid Dehalogenase superfamily, with experimental testing of function by direct biochemical assay, are featured. POOL analysis also uncovers the types of interactions that enable protein structures to transform the weakly acidic or basic side chains of amino acids into strong acids, strong bases, or nucleophiles in the active site. Finally, the prediction and characterization of druggable sites is presented for SARS-CoV-2 targets. Supported by NSF MCB-2147498

Interface-guided phenotyping of coding variants in the transcription factor RUNX1 with SEUSS

Kivilcim Ozturk[1,2], Rebecca Panwala[3], Jeanna Sheen[4], Kyle Ford[3], Nathan Payne[4,5], Dong-Er Zhang[5], Stephan Hutter[6], Torsten Haferlach[6], Trey Ideker[1,2,5], Prashant Mali[3*], Hannah Carter[1,2,5*]

1. Division of Medical Genetics, Department of Medicine, University of California San Diego, La Jolla, CA, USA
2. Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, USA
3. Department of Bioengineering, University of California San Diego, La Jolla, CA, USA
4. School of Biological Sciences, University of California San Diego, La Jolla, CA, USA
5. Moores Cancer Center, University of California San Diego, La Jolla, CA, USA
6. MLL Munich Leukemia Laboratory, Max-Lebsche-Platz 31, 81377 Munich, Germany

* Corresponding authors

pmali@ucsd.edu
hkcarter@health.ucsd.edu

Abstract

Understanding the consequences of single amino acid substitutions in cancer driver genes remains an unmet need. Perturb-seq provides a tool to investigate the effects of individual mutations on cellular programs. Here we deploy SEUSS, a Perturb-seq style approach, to generate and assay mutations at physical interfaces of the transcription factor RUNX1, with the potential to perturb different interactions, and therefore produce transcriptional readouts implicating different aspects of the RUNX1 regulon. We measured the impact of 115 mutations on RNA profiles in single myelogenous leukemia cells and used the profiles to identify three functionally distinct groups of RUNX1 mutations: wild-type (WT)-like, loss-of-function (LOF)-like and hypomorphic, characterize their effects on cellular programs, and study the implications of cancer mutations. Notably, the largest concentration of functional mutations (non-WT-like) clustered at the DNA binding site and contained many of the more frequently observed mutations in human cancers. Hypomorphic variants shared characteristics with loss of function variants but had gene expression profiles indicative of response to neural growth factor and cytokine recruitment of neutrophils. Additionally, DNA accessibility changes upon perturbations were enriched for RUNX1 binding motifs, particularly near differentially expressed genes. Overall, our work demonstrates the potential of targeting protein interaction interfaces to better define the landscape of prospective phenotypes reachable by amino acid substitutions.

Bioinformatics and machine-learning approaches for clinical assessment of copy number variants in rare disease patients

Francisco Requena [1], Nigreisy Montalvo [1], David Salgado [2,3] , Valérie Malan [4,5]
, Damien Sanlaville [6,7] , Frédéric Bilan [8,9],Christophe Béroud [2,10] , Antonio Rausell [1,4,*]

(1) Université Paris Cité, INSERM UMR1163, Imagine Institute, Clinical Bioinformatics Laboratory, Paris, F-75006, France.
(2) INSERM, Marseille Medical Genetics, Aix Marseille University, Marseille, F- 13385, France.
(3) CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Evry, F- 91057, France.
(4) AP-HP, Necker Hospital for Sick Children, Fédération de Génétique et Médecine Génomique, Service de Médecine Génomique des Maladies Rares, Paris, F-75015, France.
(5) Université Paris Cité, INSERM UMR1163, Imagine Institute, Developmental Brain Disorders Laboratory, Paris, F-75015, France
(6) CHU de Lyon HCL-GH Est, Service de génétique, Bron, F-69677 France
(7) Université Lyon 1, CNRS, INSERM, Physiopathologie et Génétique du Neurone et du Muscle, UMR5261, U1315, Institut NeuroMyoGène, Lyon, F-69008, France.
(8) Service de Génétique, CHU de Poitiers, Poitiers, F-86000, France.
(9) Laboratoire de Neurosciences Expérimentales et Cliniques INSERM U1084, Poitiers, F-86073, France.
(10) APHM, Hôpital d'Enfants de la Timone, Département de Génétique Médicale, Marseille, F- 13385 France

(*) Correspondence to: antonio.rausell@institutimagine.org

Abstract: Copy number variants (CNVs) are a major cause of rare paediatric diseases. The adoption of whole-genome sequencing as a first-line genetic test has significantly enlarged the load of CNVs identified in single genomes. Together with such increased throughput, clinical interpretation is further challenged by small-size CNVs identified in non-coding genomic regions. In this talk I will present a panel of bioinformatics strategies for the assessment of CNVs in clinical settings recently developed in my laboratory. First, CNVxplorer, a web server suited for the functional interpretation of non-coding CNVs in a clinical diagnostic setting that mines a comprehensive set of phenotypic, genomic, and epigenomic features. Second, CNVscore, a supervised machine learning approach based on tree ensembles and trained on pathogenic and non-pathogenic CNVs from reference databases. Unlike previous approaches, CNVscore couples pathogenicity estimates with uncertainty scores, making it possible to evaluate the suitability of alternative models for the query CNVs. Finally, I will present recent developments on Federated learning (FL) settings that enable multiple institutions to collaboratively train machine-learning models without sharing their local datasets. Here we show that the FL strategies to classify pathogenic and benign variants reached competitive or superior performances as compared to the individual data owners or to the centralized-data model counterparts. Considered together, these developments will highlight the potential and current limitations of novel bioinformatics and ML approaches for clinical cytogenetics applications.

References:

Requena F, Salgado D, Malan V, Sanlaville D, Bilan F, Beroud C, Rausell A. CNVscore calculates pathogenicity scores for copy number variants together with uncertainty estimates accounting for learning biases in reference Mendelian disorder datasets. medRxiv 2022 https://doi.org/10.1101/2022.06.23.22276396 (preprint; under review)

Requena F, Abdallah HH, Garcia A, Nitschké P, Romana S, Malan Valérie, Rausell A*. CNVxplorer: a web tool to assist clinical interpretation of CNVs in rare disease patients. Nucleic Acids Research (2021) gkab347, https://doi.org/10.1093/nar/gkab347

How do catalytic lysines in enzymes gain their catalytic properties?

Atif Shafique, Michelle Mirabelli, Heidi Eren, Mary Jo Ondrechen*

Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115 USA

Corresponding author: Mary Jo Ondrechen   mjo@neu.edu

ABSTRACT

The lysine side chain is a weak base and is fully protonated at neutral pH in the free amino acid. How does it become a strong base or nucleophile in enzyme active sites? Computational analysis using Partial Order Optimum Likelihood (POOL) was performed on 75 enzymes that represent all six major enzyme classifications and a variety of different folds. All contain one or more lysine residues that have been found to be catalytically active in previously reported experiments. The catalytic lysine residues reported here have strong coupling interactions to other residues that obey one or both inequality expressions reported by Coulther, Ko and Ondrechen in 2021. Specifically, the catalytic lysines are strongly coupled to at least one other lysine residue with intrinsic pK_a difference within 1 pH unit, or else are strongly coupled with tyrosine or cysteine residues wherein these anion-forming residues have an intrinsic pK_a higher than that of the lysine. The interactions help us in identifying roles of these supporting residues and also identify arrays of interacting residues that are characteristic of specific biochemical functions, thus supplying functional information that can be used to annotate protein structures of unknown function. This analysis provides insight into how catalytic lysines achieve their catalytic power. Acknowledgement: NSF MCB-2147498 & USPakistan Knowledge Corridor; MM and HE supported by NSF DBI-2031778.

Understanding the mechanisms of incomplete penetrance: analysis of >100,000 human genomes

Mugdha Singh[1,2], Emily Groopman[1], Sarah Stenton[1,2], Sanna Gudmundsson[1,2,3#*], Anne O'Donnell-Luria[1,2#*]

1. Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA; 2. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 3. Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden.

#equal contribution

*sgudmund@broadinstitute.org
*odonnell@broadinstitute.org

Variant interpretation remains challenging and limits the diagnostic yield of sequencing studies. One factor complicating variant interpretation is incomplete penetrance, a phenomenon where only some individuals who carry a pathogenic variant will develop the disease or the full spectrum of the phenotype. We aim to identify these cases to perform deep molecular and phenotypic assessment to try to elucidate the mechanisms of incomplete penetrance. The final objective of the project is to increase the diagnostic yield by deciphering these diagnostic variants, as in cases of incomplete variance, the presence of these pathogenic variants in unaffected parents makes it difficult to conclude. We utilized gnomAD consisting of 76,156 genomes, wherein we determined 899 predicted loss of function (pLOF) but highly confident variants. After filtering out for the presence of secondary rescue variants, 131 variants in 39 genes identified in 216 individuals were shortlisted. Along with the initial cohort that focusses on gnomAD we have also identified a smaller subset of 16 individuals with gain of function (GoF) disorders, focusing on variants reported as pathogenic or likely pathogenic in ClinVar. We will expand the cohort by searching among individuals sequenced through the Broad Institute, Center for Mendelian Genomics (CMG) (n= 26,098 and growing) with pLoF in haploinsufficiency genes or pathogenic or likely pathogenic variants in ClinVar for dominant disease genes. To date, our analysis strategies have not considered incomplete penetrance, so these variants (and the diagnoses for these cases) have been overlooked. About 65% of the cohort remains undiagnosed, if ~2% have incomplete penetrance, we will identify >100 non-penetrant individuals with this approach. We will undertake one of the largest studies of non-penetrant individuals to date, to identify common molecular mechanisms of incomplete penetrance. We hypothesize that cis-regulatory variants mediate some cases of incomplete penetrance.

VARSTACK^2: AN UPDATED DATA RETRIEVAL TOOL FOR SOMATIC VARIANT INTERPRETATION IN CANCER

Nitin U Sreekumar[1], Shulan Tian[2], Alper Uzun[1,3,4], Ece D Gamsiz Uzun[1,3,4,*]
[1]Department of Pathology and Laboratory Medicine, Alpert Medical School, Brown University, Providence, RI, USA
[2] Biomedical Informatics, Mayo Clinic, Rochester, MN, USA
[3]Legorreta Cancer Center, Brown University, Providence, RI, USA
[4]Brown/Lifespan Center for Clinical Cancer Informatics and Data Science (CCIDS), Providence, RI, USA
*Corresponding author: Ece D Gamsiz Uzun, Email: dilber_gamsiz@brown.edu

Cancer, a complex and pervasive global health challenge, primarily stems from genetic abnormalities, particularly somatic variations. As cutting-edge sequencing technologies continue to evolve and the focus on amassing genomic data intensifies, a wealth of information becomes available[1]. The crucial task is harnessing this information effectively in clinical practice to advance precision medicine and improve health outcomes. Yet, organizing and deciphering genomic variant data scattered across multiple databases can be time-consuming. To address this challenge, we introduced Varstack, a web tool designed to streamline somatic variant interpretation[2]. Building upon Varstack's success, we present Varstack^2, an updated webtool for somatic variant interpretation. Varstack^2 retrieves data from several databases, including COSMIC, ClinVar, UCSC Genome Browser, cBioPortal, and ClinicalTrials.gov, significantly reducing the time required to access these datasets through a single search (Figure 1). Utilizing Application Programming Interfaces (APIs) ensures that data remains up-to-date, automatically updating whenever the source databases change. Searching for specific variants across all connected databases is made effortless through a user-friendly scroll-down interface. Furthermore, smart search feature of Varstack^2 allows users to supply gene information when amino acid or base changes are unknown, generating a list of variants from the databases for the specified gene. Varstack^2 represents a significant advancement, addressing VarStack's limitations and offering an improved web application for variant scientists, physicians and cancer researchers. The beta-version of Varstack^2 is accessible at https://nitthekit.github.io/varstack2/.

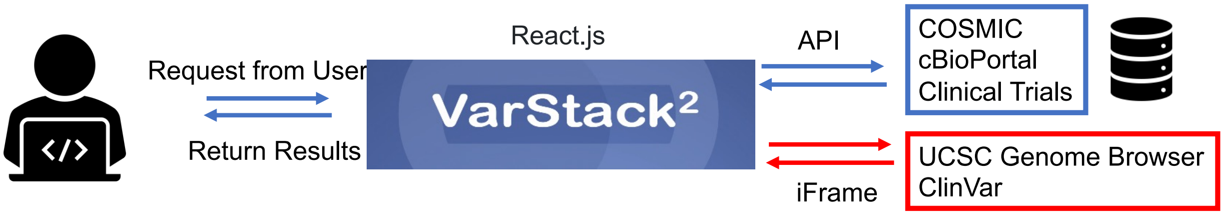


Figure 1: Varstack^2 workflow. Varstack^2 is built using the React.js framework. COSMIC, cBioPortal, and Clinical Trials are accessed using the site's Application Programming Interface (API), and the UCSC genome browser and ClinVar are viewed using an iframe.

References

1. Berger, M. F.; Mardis, E. R., The emerging clinical relevance of genomics in cancer medicine. *Nat Rev Clin Oncol* **2018,** *15* (6), 353-365.

2. Howard, M.; Kane, B.; Lepry, M.; Stey, P.; Ragavendran, A.; Gamsiz Uzun, E. D., VarStack: a web tool for data retrieval to interpret somatic variants in cancer. *Database (Oxford)* **2020,** *2020*.

From Variants to Vision: A Deep Dive into VCF with Variant Graph Craft.

Jennifer Li[1], Andy Yang[2], Benedito A Carneiro [3, 4], Ece Gamsiz Uzun [4, 5, 6], Lauren Massingham [7], Alper Uzun* [4, 5, 7]

[1]Department of Computer Science, Brown University, Providence, RI, USA;
[2] Department of Chemistry, Brown University, Providence, RI, USA;
[3]Lifespan Cancer Institute, Providence, RI, USA;
[4]Legorreta Cancer Center, Brown University, Providence, RI, USA;
[5]Department of Pathology and Laboratory Medicine, Alpert Medical School, Brown University, Providence, RI, USA;
[6]Brown/Lifespan Center for Clinical Cancer Informatics and Data Science (CCIDS), Providence, RI, USA;
[7]Department of Pediatrics, Alpert Medical School, Brown University, Providence, RI, USA;

* alper_uzun@brown.edu

The Variant Call Format (VCF) is a structured and comprehensive text file capturing critical genomic information, including variant positions, alleles, genotype calls, and quality scores. The intricate nature of VCF files makes their analysis and visualization a complex endeavor, demanding specialized resources and features. To address this challenge, we introduce "Variant Graph Craft (VGC)", a dedicated VCF visualization and analysis tool. VGC boasts an extensive suite of features tailored for genetic variation exploration. Not only can it extract and visualize variant data, but it also offers a graphical representation of samples, complete with genotype information. A standout feature is VGC's integration with external databases, allowing users to access gene function, pathway information from sources like Msig Database for GO terms, KEGG, Biocarta, Pathway Interaction Database, and Reactome. Additionally, VGC provides a dynamic connection to gnomAD for variant details and incorporates ClinVar data, highlighting pathogenic variants. Prioritizing user security, Variant Graph Craft operates locally, eliminating the need for VCF uploads to external clouds. It is compatible with the Human Genome Assembly Hg37, ensuring versatility with diverse datasets (Figure1). Furthermore, VGC is adaptive, accommodating user-specific requirements through optional phenotype input data. In summary, Variant Graph Craft presents a secure, intuitive, and comprehensive solution for genomic variation exploration. Its array of user-centric features empowers researchers and clinicians to navigate and decipher genomic variation data effectively, VGC is freely available at https://github.com/alperuzun/VGC.

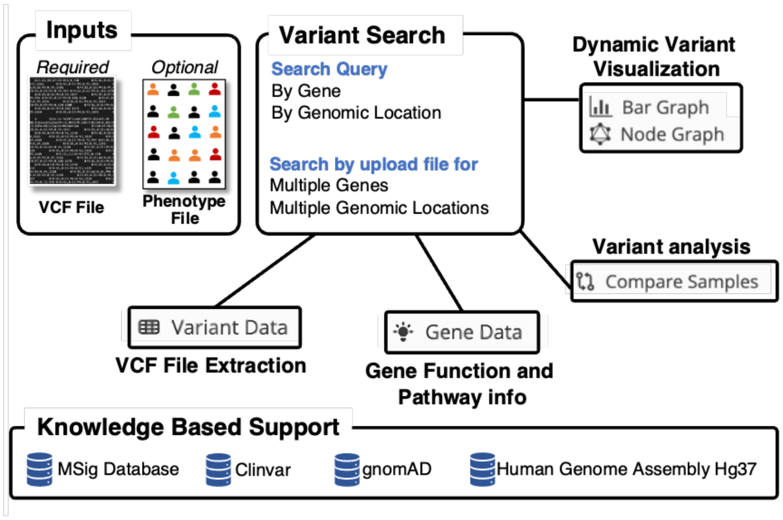


Figure 1. Features and integration of evidence-based information.

Modelling Abundance of Single-amino-acid Variants using Protein Language Models

Daniel Zeiberg[1], Shantanu Jain[1], Vikas Pejaver[2,3], Predrag Radivojac[1]

[1] Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA
{zeiberg.d, sh.jain, p.radivojac}@northeastern.edu
[2] Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
vikas.pejaver@mssm.edu
[3] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

As per the 2015 ACMG/AMP guidelines, measurements from functional assays serve as a strong line of evidence in determining the clinical pathogenicity of protein-coding variants. In this regard, stability and abundance are two measurable properties that may be altered by variants and are shared by all proteins. While the association between the former and variant pathogenicity is well established, a recently developed high-throughput assay called variant abundance by massively parallel sequencing (VAMP-seq) has shown promise in correlating the latter with variant pathogenicity. However, since protein stability and abundance are themselves correlated, it is unclear to what extent VAMP-seq abundance scores would improve the prediction of variant pathogenicity. Current predictive models of stability and pathogenicity are frequently evaluated in their correlation with abundance scores, but are neither trained nor calibrated to directly predict the abundance of single-amino-acid variants.

Here, we develop a model based on residue embeddings from the Prot-T5-XL-BFD protein language model to predict the abundance of variants and demonstrate that predictions from this model better correlate with true VAMP-seq scores, with an average correlation of 0.58, than do predictions from representative predictors of pathogenicity, stability and variant effect, including MutPred2, PremPS, VARITY-R, Envision, EVE, and Snap2. Additionally, we demonstrate how these protein language model variant representations can improve the predictive power of the MutPred2 pathogenicity predictor. We explore the extent to which these language model embeddings add signal to different subsets of the features used by MutPred2.

As more VAMP-seq data sets become available through IGVF and other Consortia, we anticipate our work to yield a high-quality predictor of variant abundance that can quickly predict the abundance of all possible single-amino-acid variants, serving as a means to obtain additional lines of evidence for pathogenicity, as well as more directly improving computational predictions of pathogenicity.

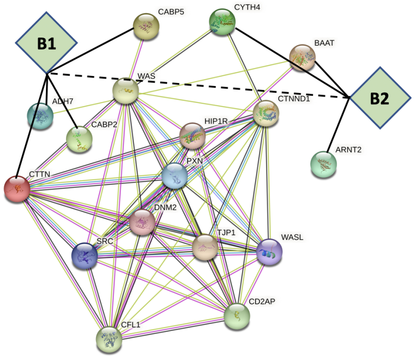
Melissa: Semi-Supervised Embedding for Protein Function Prediction Across Multiple Networks

Kaiyi Wu*[1], Di Zhou[2], Donna Slonim[2], Xiaozhe Hu*[1], Lenore Cowen*[1,2]

1.Department of Mathematics, Tufts University, Medford, MA 02140
2.Department of Computer Science, Tufts University, Medford, MA 02155
Co-corresponding authors: cowen@cs.tufts.edu, xiaozhe.hu@tufts.edu

Several popular methods exist to predict function from multiple protein-protein association networks. For example, both the Mashup algorithm, introduced by Cho, Peng and Berger [1], and deepNF, introduced by Gligorijevi, Barotand, and Bonneau [2], create embeddings by analyzing the diffusion in each network, to characterize the topological context of each node. Then, in Mashup, the high-dimensional topological patterns in individual networks are canonically represented using low-dimensional vectors, one per gene or protein, to yield the multi-network embedding. In deepNF, a multimodal autoencoder is trained to extract common network features across networks to yield the low-dimensional embedding. Neither embedding takes into account known functional labels; rather, these are then used by the machine learning methods applied after embedding.

MELISSA seeks to employ semi-supervised methods to incorporate known GO label information directly into the low-dimensional embedding to improve its quality and, thus, improve on the state-of-the-art function prediction methods. We augment each network with a sparse set of artificial new nodes, which are also embedded, encoding functional information with a set of "must-link" (ML) constraints and a set of "cannot-link" (CL) constraints. After augmentation, the resulting networks now contain positive and negative weights, and a signed version of the graph Laplacian and generalized diffusion state representations of each node are required to compute the embedding.



Melissa adds artificial nodes (B1; B2) representing biclusters of GO labels and genes, connected to the network with "must link" constraints (solid black lines) to pull together nodes with similar functional labels, and "cannot link" constraints (dashed lines) that separate the biclusters.

Once the embedding is formed, a variety of different classification methods can be applied to the embedded space. Because our focus is on improving the information content of the embedding, in this work we pair MELISSA with the simple kNN classifier, where comparing functional label prediction of competing embeddings gives a sense of the functional enrichment in local neighborhoods. In this setting, we show that MELISSA substantially improves the performance of the functional label prediction task, demonstrating its ability.

References

[1] H. Cho, *et al., Cell Systems*, 3(6):540–548, 2016.
[2] V. Gligorijević, *et al., Bioinformatics*, 34(22):3873–3881, 2018.

Exploring Population-Specific Genetic Variations and Their Impact on Protein-Molecule Interactions

Ziyang Gao[1], Mo Sun[1,2], Yanzhao Wang[3,4], Hongzhu Cui[1,6], Suhas Srinivasan[5,7,8], Dmitry Korkin[*1,5,9]

[1]Bioinformatics and Computational Biology Program, Worcester Polytechnic Institute, Worcester, MA
[2]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA
[3]Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA
[4]Regeneron, Basking Ridge, NJ
[5]Data Science Program, Worcester Polytechnic Institute, Worcester, MA
[6]Chromatography and Mass Spectrometry Division, Thermo Fisher Scientific, San Jose, CA
[7]Program in Epithelial Biology, Stanford School of Medicine, Stanford, CA
[8]Center for Personal Dynamic Regulomes, Stanford School of Medicine, Stanford, CA
[9]Computer Science Department, Worcester Polytechnic Institute, Worcester, MA

In this study, we aim to understand the population-specific genetic variation profile in molecular interactions mediated by proteins, including protein-protein interactions (PPIs) and protein-ligand interactions. We first developed a comprehensive catalog of population-specific edgetic effects, *i.e.*, events where a mutation disrupt a PPI, at the whole interactome level. Using the innovative 'edgotype' concept, we explore how mutations influence specific PPIs, shaping the human interactome and potentially contributing to population-specific traits. Next, we created a global population profile of genetic variation in protein-ligand interactions, with a focus on the variants within ligand-binding genes and their co-localization with the ligand-binding sites. We also investigated the diversity of population-specific ligand-binding mutations across various types of ligands.

To study the edgetic effects on PPIs, we carried out a systematic *in-silico* profiling of approximately 50,000 non-synonymous single nucleotide variants (nsSNVs) sourced from the diverse genetic backgrounds of the 1,000 Genomes Project. Employing the SNP-IN tool, a semi-supervised learning approach, we predicted the interaction-rewiring effects of these variants within a comprehensive set of over 10,000 structurally-characterized protein interaction complexes. By assessing the functional roles of these nsSNVs, we uncovered a remarkable capacity for the healthy populations to rewire PPIs. Surprisingly, we found that the interaction-disrupting mutations in these populations were linked to a wide array of biological functions and implicated in multiple diseases. Network analysis uncovered an intriguing phenomenon: the disease-associated modules exhibited a higher density of interaction-disrupting mutations from the healthy populations.

In our exploration of protein-ligand interactions, we employed a dataset comprising over 3,000 proteins with well-documented ligand interactions, sourced from BioLip2. These interactions were systematically mapped onto genetic variants extracted from the most recent population-specific dataset of variants, GnomAD3. Based on their biochemical properties and guided by their chemical structural similarities, the ligands were categorized into five distinct classes: ion, metal, metabolite, druglike, and regular. By calculating allele frequencies within ligand binding sites for various population subgroups, we unveiled unique population-specific patterns of the variant frequencies, which we further clustered using genetic distance metrics derived from the human phylogenetic tree, thereby generating distinct genetic signatures for each population. Importantly, our investigation highlighted specific variants within the drug binding sites that significantly impacted binding affinity, indicating potential clinical relevance. Moreover, we observed that populations characterized by a higher frequency of these variants exhibited reduced drug efficacy.

Integration of Gene-Level Annotation and Visualization in the Updated Functional Annotation of Variants Online Resource (FAVOR)

Hufeng Zhou[1], Eric Van Buren[1], Vineet Verma[1], Tom Li[1], Zhiping Weng[2], Shamil R Sunyaev[3,4], Xihong Lin[1,4,5*]

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA
[3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[4]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA
[5]Department of Statistics, Harvard University, Cambridge, MA, USA
[*]Correspondence may also be addressed to Xihong Lin. Email: xlin@hsph.harvard.edu

Abstract

We present Functional Annotation of Variants Online Resources (FAVOR), a comprehensive platform providing variant functional annotations for WGS analysis, result interpretation, and prioritizing disease-associated causal variants. FAVOR facilitates online queries and annotates genotype data from large-scale WGS studies, offering a multi-faceted variant functional annotation portal. It enables swift queries at variant, gene, and region levels, integrating functional information from various sources to uncover variant characteristics and prioritize causal variants influencing human phenotypes. To enhance user experience, we made significant enhancements to the front-end and back-end components, migrating the GUI from R shiny apps to the React Framework (next.js) and transitioning the backend database hosting to a local cloud platform based on open stack. These upgrades resulted in a more powerful database with faster response times for all core functions. Additionally, we introduced comprehensive gene-level functional annotation to provide diverse perspectives on gene function. Leveraging chat-gpt APIs, we developed FAVORGPT, an interactive data usage guide enabling users to engage with FAVOR naturally and gain comprehensive insights into each functional annotation channel. FAVOR now supports both HG19 and HG38 versions, broadening its compatibility and applicability. These advancements offer researchers and clinicians a valuable resource for comprehensive variant functional annotation, supporting the understanding of human disease genetics and targeted therapeutic strategies. The new version of FAVOR is accessible at https://favor.genohub.org.

Transcript-aware analysis of rare pLOF variants in the UK Biobank elucidate new isoform-trait associations

Rachel A. Hoffing[1]*, Aimee M. Deaton[1], Aaron M. Holleman[1], Lynne Krohn[1], Philip J. LoGerfo[1],
Mollie E. Plekan[1], Sebastian Akle Serrano[1], Paul Nioi[1], Lucas D. Ward[1]
Alnylam Pharmaceuticals[1]

Cambridge, MA 02142, USA

Email: rhoffing@alnylam.com

A single gene can produce multiple transcripts with distinct molecular functions. Rare-variant association tests often aggregate all coding variants across individual genes, without accounting for the variants' presence or consequence in resulting transcript isoforms. To evaluate the utility of transcriptaware variant sets, rare predicted loss-of-function (pLOF) variants were aggregated for 17,035 proteincoding genes using 55,558 distinct transcript-specific variant sets. These sets were tested for their association with 728 circulating proteins and 188 quantitative phenotypes across 406,921 individuals in the UK Biobank. The transcript-specific approach resulted in larger estimated effects of pLOF variants decreasing serum cis-protein levels compared to the gene-based approach (pbinom < $2x10^{-16}$). Additionally, 251 quantitative trait associations were identified as being significant using the transcriptspecific approach but not the gene-based approach, including PCSK5 transcript ENST00000376752 and standing height (transcript-specific statistic, P = $1.3x10^{-16}$, effect = 0.72 SD decrease; gene-based statistic, P = 0.02, effect = 0.05 SD decrease) and LDLR transcript ENST00000252444 and apolipoprotein B (transcript-specific statistic, P = $5.7x10^{-20}$, effect = 0.99 SD increase; gene-based statistic, P = $3.0x10^{-4}$, effect = 0.17 SD increase). This approach demonstrates the importance of considering the effect of pLOFs on specific transcript isoforms when performing rare-variant association studies.

The influence of genetic predisposition and physical activity on risk of Gestational Diabetes Mellitus in the nuMoM2b cohort

Kymberleigh A. Pagel[1,2], Hoyin Chu[3,4], Rashika Ramola[3], Rafael F. Guerrero[5], Judith H. Chung[6], Samuel Parry[7], Uma M. Reddy[8], Robert M. Silver[9], Jonathan G. Steller[6], Lynn M. Yee[10], Ronald J. Wapner[11], Matthew W. Hahn[1,12], Sriraam Natarajan[13], David M. Haas[14], *Predrag Radivojac[3]

[1]Department of Computer Science, Indiana University, Bloomington, IN
[2]Institute of Computational Medicine, Johns Hopkins University, Baltimore, MA
[3]Khoury College of Computer Sciences, Northeastern University, Boston, MA
[4]Dana-Farber Cancer Institute, Boston, MA
[5]Department of Biological Sciences, North Carolina State University, Raleigh, NC
[6]Department of Obstetrics and Gynecology, University of California, Irvine, CA
[7]Department of Obstetrics and Gynecology, University of Pennsylvania School of Medicine, Philadelphia, PA
[8]Department of Obstetrics, Gynecology, and Reproductive Sciences, Yale School of Medicine, Yale University, CT
[9]Department of Obstetrics and Gynecology, University of Utah School of Medicine, Salt Lake City, UT
[10]Department of Obstetrics and Gynecology, Northwestern University Feinberg School of Medicine, Chicago, IL
[11]College of Physicians and Surgeons, Columbia University, New York, NY
[12]Department of Biology, Indiana University, Bloomington, IN
[13]Department of Computer Science, The University of Texas at Dallas, TX
[14]Department of Obstetrics and Gynecology, Indiana University School of Medicine, Indianapolis, IN

Corresponding author(s): David M. Haas (dahaas@iu.edu), Predrag Radivojac (predrag@northeastern.edu)

Although polygenic risk scores (PRS) for Type II Diabetes Mellitus (T2DM) can improve risk prediction for Gestational Diabetes Mellitus (GDM), yet the strength of the relationship between genetic and lifestyle risk factors has not been quantified. This work assesses the effects of PRS and physical activity on existing GDM risk models and identifies patient subgroups who may receive the most benefits from receiving a PRS or activity intervention.

The study population included individuals who were enrolled in the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be (nuMoM2b) cohort in which nulliparous women were recruited from hospitals affiliated with 8 clinical sites in the US. The nuMoM2b study was established to study individuals without previous pregnancy lasting 20 weeks or more (nulliparous) and to elucidate factors associated with adverse pregnancy outcomes. A sub-cohort of 3,533 participants with European ancestry were used for risk assessment and performance evaluation. Self-reported total physical activity in early pregnancy was quantified as metabolic equivalent of tasks (METs) in hours/week. Polygenic risk scores were calculated for T2DM using contributions of 85 single nucleotide variants, weighted by their association in the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium data.

The findings show that in high-risk patient subgroups the addition of PRS resulted in increased risk of GDM diagnosis. The high-risk population subgroups (body mass index ≥ 25 or age ≥ 35), individuals with PRS in the top 25th percentile or METs below 450 have significantly increased odds of GDM diagnosis. Participants with both high PRS and low METs have three times higher odds of GDM diagnosis than the population. Conversely, increased physical activity is associated with decreased risk of GDM, particularly among individuals genetically predisposed to T2DM. The participants with high PRS and METs ≥ 450 do not exhibit increased odds of GDM diagnosis, and those with low METs and low PRS have reduced odds of GDM. The relationship between PRS and METs was found to be nonadditive. The findings suggest the benefits of targeted PRS ascertainment to encourage early intervention.

Pagel KA, Chu H, Ramola R, et al. Association of Genetic Predisposition and Physical Activity With Risk of Gestational Diabetes in Nulliparous Women. *JAMA Netw Open.* 2022;5(8):e2229158. doi:10.1001/jamanetworkopen.2022.29158

Comprehensive assessment of isoform detection methods for long-read sequencing

Yaqi Su[1,6], Zhejian Yu[1], Siqian Jin[1], Zhipeng Ai[2], Ruihong Yuan[1], Xinyi Chen[1], Ziwei Xue[1], Yixin Guo[1], Di Chen[1], Hongqing Liang[2], Zuozhu Liu[3], Wanlu Liu[1,4,5*]

1. Zhejiang University-University of Edinburgh Institute, Zhejiang University
2. Division of Human Reproduction and Developmental Genetics, Women's Hospital, and Institute of Genetics, Zhejiang University School of Medicine, Zhejiang University
3. Zhejiang University-University of Illinois at Urbana-Champaign Institute, Zhejiang University
4. Future Health Laboratory, Innovation Center of Yangtze River Delta, Zhejiang University
5. Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Zhejiang University
6. Department of Molecular & Cell Biology, University of California, Berkeley

* Corresponding author. Email: wanluliu@intl.zju.edu.cn

The development of long-read sequencing techniques has increased the length of sequencing reads to several kilobases, which facilitates the identification of alternative splicing (AS) events and isoform expressions. Numerous computational methods for isoform detection using long-read sequencing data have been developed recently. In this study, we systemically evaluate the performance of eleven methods implemented in eight computational tools capable of identifying isoform structures from long-read cDNA sequencing data. We assessed their performances using simulated datasets from an in-house simulator and experimental data encompassing various potential influencing factors, including diverse sequencing platforms. Our results indicate the guided mode of StringTie2 and Bambu achieved the best sensitivity and precision, respectively. This study informs future investigations of AS and the enhancement of tools for isoform detection using long-read data.