HGVS
HUMAN GENOME
VARIATION SOCIETY

Methods & Tools for Assessing the Impact of Genetic Variations

Lake Nona Room
Hilton Orlando
Orlando, FL, USA

17th October 2017

Scientific Program Committee

Christophe Béroud, Marseille, France
Steven E. Brenner, CA, USA
Anthony J. Brookes, LT, UK
Marc S. Greenblatt - VT, USA
Rachel Karchin - MD, USA
Sean D. Mooney - WA, USA

Organising Secretariat

Rania Horaitis (Meeting Makers | www.meeting-makers.com)

Introduction

Both researchers and clinicians are awash in full exome and genome sequences, with the myriad of variants they harbour. For some purposes, well-studied variants provide research insight and clinical resolutions. But more often, the variants' roles are not conclusively known from previous studies, and therefore methods are necessary to help inform their phenotypic impact. The 2017 scientific meeting of the Human Genome Variation Society will present cutting-edge research and practical approaches involving methods for interpreting human genetic variants.

Invited Speakers

Christophe Béroud
Aix-Marseille University, Marseille, France

Mark Gerstein
Yale University, CT, USA

Rachel Karchin
Institute for Computational Medicine, Johns Hopkins Biomedical Engineering, MD, USA

Predrag Radivojac
Department of Computer Science, Indiana University, Bloomington, IN, USA

Weiva Sieh
Mt Sinai, NY, USA

Sean Tavtigian
Huntsman Cancer Institute, Salt Lake City, UT, USA

Methods & Tools for Assessing the Impact of Genetic Variations

Lake Nona Room, Lobby Level
Hilton Orlando
Orlando, FL, USA

17th October 2017

Program

8:00 - 8:55 Registration

8:55 - 9:00 Welcome

Session 1 **Moderator: Marc Greenblatt**

9:00 - 9:40 **KEYNOTE SPEAKER**

Prioritizing somatic variants

Mark Gerstein
Prof of Biomedical Informatics & Prof of Molecular Biophysics & Biochemistry, & of Computational Biology & Bioinformatics, Yale University, CT, USA

9:40 - 9:55 **Presentation from selected Abstract**

GRASP v3: an updated GWAS catalog and contrast to similar catalogs

Ben Rodriguez
Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, 73 Mt Wayte Ave, Framingham, MA, USA

9:55 - 10:35 **KEYNOTE SPEAKER**

Predicting the pathogenicity of rare missense variants

Weiva Sieh
Senior Faculty, Population Health Science & Policy & Genetics and Genomic Sciences, Mt Sinai, NY, USA

10:35 - 10:50 **Presentation from selected Abstract**

MutPred2 enables probabilistic interpretations of pathogenicity and impact on protein structure and function

Vikas Pejaver
Department of Biomedical Informatics and Medical Education and the eScience Institute, Univ. of Washington, Seattle, WA, USA

10:50 - 11:20 **Coffee Break & Poster Session**

Session 2 **Moderator: Steven Brenner**

11:20 - 12:00 **KEYNOTE SPEAKER**

Predicting the impact of mutations on splicing signals

Christophe Béroud
Genetics and Bioinformatics, Aix-Marseille University, Marseille, France

12:00 - 12:15 **Presentation from selected Abstract**

Rethinking the 5 splice site algorithms used in clinical genomics

Gabe Rudy
Golden Helix, 203 Enterprise Blvd, Suite 1, Bozeman, MT 59718, USA

12:15 - 12:55 **KEYNOTE SPEAKER**

Evaluating the evaluation of cancer driver genes

Rachel Karchin
Institute for Computational Medicine, Johns Hopkins Biomedical Engineering, Baltimore, MD, USA

12:55 - 13:10 **Presentation from selected Abstract**

The Ensembl Variant Effect Predictor (VEP)

Benjamin Moore
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK

13:10 - 14:15 **Lunch & HGVS Annual General Meeting**

Session 3	**Moderator: Christophe Béroud**
14:15- 14:55	**KEYNOTE SPEAKER** **Validating and calibrating computational and functional approaches in BRCA and MMR** **Sean Tavtigian** *Oncological Sciences, Huntsman Cancer Institute, Salt Lake City, UT, USA*
14:55 - 15:35	**KEYNOTE SPEAKER** **Predicting the molecular mechanisms of genetic disease for protein coding variants** **Predrag Radivojac** *Department of Computer Science and Informatics, Indiana University Bloomington, IN, USA*
15:35 - 15:50	**Presentation from selected Abstract** **Findings from CAGI, the Critical Assessment of Genome Interpretation, a community experiment to evaluate phenotype prediction** **Steven Brenner** *Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA*
15:50 - 16:00	Closing Remarks
16.00	**MEETING END (in time for ASHG Plenary @ 16.30)**

Poster Presentations

Assessing mutational signatures and impact of loss-of-function genetic variants
Predrag Radivojac

Unravelling the Genetic Mechanisms of Neonatal Diabetes Mellitus: An Egyptian Experience
Rasha Elkafass

Abstracts

SESSION 1

Prioritizing somatic variants

Mark Gerstein

Prof of Biomedical Informatics & Prof of Molecular Biophysics & Biochemistry, & of Computational Biology & Bioinformatics, Yale University, CT

My talk will focus on prioritizing genetic variants associated with cancer, to identify key variants driving cancer progression. First, I will look at the overall functional impact of the variants in cancer genomes, ranking them in terms of impact, for both coding and noncoding regions. For the coding analysis, we use the ALoFT and frustration tools, and for the noncoding analysis, we use FunSeq. Then, I will look at the recurrence of variants within cancer cohorts. Here we develop two approaches: one parametric (LARVA) and the other non-parametric (MOAT). Finally, I will put these methods together through application to kidney and prostate cancers.

GRASP v3: an updated GWAS catalog and contrast to similar catalogs

Ben A.T. Rodriguez[1], Caroline Zellmer[1,2], James Li[1,3], Ashwin Panda[1,4], Nicole Jensen[1], Ju-Ping Lien[1], John D. Eicher[1], Andrew D. Johnson[1*]

[1]*Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, 73 Mt Wayte Ave, Framingham, MA 01702*

[2]*University of Wisconsin-Madison, 702 West Johnson St, Madison, WI 53715*

[3]*The Ohio State University, 1800 Cannon Dr, Columbus, OH 43210*

[4]*Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213*

*johnsonad2@nhlbi.nih.gov

Access to the expanding number of GWAS studies is a vital tool to biomedical research, providing opportunity for novel scientific discovery, identification of targets for functional assessment and mechanistic insights into disease biology. Current GWAS catalogs each have unique design features as well as limitations. The goal of GRASP (https://grasp.nhlbi.nih.gov/) is to provide a simple, intuitive means for the scientific community to query a centralized repository of reported SNP associations ($P \leq 0.05$) with human traits, including studies of methylation and expression QTL. Current challenges to developing and maintaining GWAS databases include: (1) data sharing, (2) non-standardized formats, annotation practices across GWAS, (3) human curator study search, review and extraction strategies, and (4) criteria for inclusion. Here we describe efforts toward the release of GRASP v3.

The v3 update increases the total number of studies to >3,300. To facilitate results sharing summary statistics are now included in the GRASP web portal. To investigate differences among widely used catalogs, we contrast GRASP and the NHGRI-EBI. The latter catalog was downloaded May 5, 2017 and limited to results with clear SNPids. Random studies (n=50) common to both were selected for comparison.

At a genome-wide significant threshold (P<=5e-8) there was marked variation between the catalogs with respect to the studies' results. There were 656 SNP-association results unique to GRASP and 8 results unique to NHGRI-EBI. Focusing on 373 SNPs common to both we characterized consistency in associations for the studies. While most were consistent, differences in phenotype nomenclature were frequent. In general, GRASP contained more results for specific sub-phenotypes per SNP (e.g., sex-specific results, disease sub-classification). When comparing P-value associations for identical SNPs/publication/phenotype only a small number of associations (n=14/373) deviated by >1 log(P-value) unit, with others being highly correlated (r=0.99). Sampling of 50 more studies yielded similar results.

Study exclusion criteria and data extraction methods vary between catalogs and likely account for many differences. However, our results indicate marked heterogeneity across popular GWAS results catalogs and researchers may be best served by querying multiple resources instead of relying on a single one. Further characterization and analysis of GRASP v.3 will be presented.

Predicting the pathogenicity of rare missense variants

Weiva Sieh, MD, PhD, Associate Professor of Epidemiology and Genetics

Icahn School of Medicine at Mount Sinai, New York, NY

The vast majority of coding variants are rare, and have limited functional data available. Yet few existing tools have targeted the interpretation of rare variants. Better methods for predicting the pathogenicity of rare coding variants are needed to facilitate the discovery of disease variants from sequencing studies. We developed REVEL (rare exome variant ensemble learner), an ensemble method for predicting the pathogenicity of missense variants based on a combination of 18 scores from 13 individual tools: MutPred, FATHMM, VEST, Poly-Phen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, and phastCons. REVEL was trained on recently discovered pathogenic and rare neutral missense variants, excluding those previously used to train its constituent tools. We evaluated performance in two large independent test sets across a broad range of allele frequencies and found that REVEL performed well overall, and especially when applied to variants with allele frequencies less than 0.5%, compared to other methods. The REVEL score for an individual missense variant can range from 0 to 1, with higher scores reflecting greater likelihood that the variant is disease-causing. We provide pre-computed REVEL scores for all possible human missense variants to facilitate the identification of pathogenic variants in the sea of rare variants discovered as sequencing studies expand in scale.

MutPred2 enables probabilistic interpretations of pathogenicity and impact on protein structure and function

Vikas Pejaver,[1] *Predrag Radivojac,*[2*] *Sean D. Mooney*[1*]

1. *Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA 98109, U.S.A*
2. *Department of Computer Science and Informatics, Indiana University, Bloomington, IN 47405, U.S.A*

Email: predrag@indiana.edu; sdmooney@uw.edu

Several machine-learning approaches have been developed for the prediction of the impact of missense variants. These have operationalized impact in three different ways: (1) MutPred,[1] PolyPhen-2,[2] CADD,[3] and REVEL[4] are trained to predict disease-associated variants, (2) SNAP[5] and SNAP2[6] are trained to predict variants that affect *in vitro* protein function, and (3) other tools are trained to predict changes in specific protein properties such as stability[7] or macromolecular binding.[8] MutPred2[9] is a sequence-based tool that not only improves the prioritization of pathogenic missense variants but also serves as a first step towards unifying prediction across these three different *senses* of impact, leading to better interpretability.

For pathogenicity prediction, MutPred2's state-of-the-art performance results from a larger and heterogeneous training set, the inclusion of new features derived from over 50 in-built structural and functional property predictors and the use of a neural network ensemble model, which outputs prediction scores that approximate the posterior probability distribution. The latter leads to major implications for variant impact interpretation in two different *senses*. First, on genomes from healthy individuals, MutPred2 generates an exponentially decreasing score distribution that correlates with minor allele frequencies, resulting in fewer pathogenic predictions than PolyPhen-2. Second, on data sets from the Critical Assessment of Genome Interpretation (CAGI),[10] MutPred2 scores correlate with actual experimental measurements of different proteins' activities (tasks that it was not trained for).[11] Apart from these emergent properties, a distinguishing feature of MutPred2 is the probabilistic modeling of variant impact on specific aspects of protein structure and function that can serve to guide experimental studies and clinical interpretation. We demonstrate the utility of MutPred2 for this third *sense* of impact through the identification of the structural and functional mutational signatures relevant to Mendelian disorders and the prioritization of *de novo* variants associated with complex neurodevelopmental disorders.

Thus, MutPred2 builds upon variant impact prediction as a binary classification problem by generating continuous score distributions that capture some of the underlying biology, and providing additional context through the inference of molecular mechanisms of disease.

Acknowledgments
We thank Prof. Lilia Iakoucheva and Prof. Jonathan Sebat for their comments. We thank Dr. Matthew Mort and Prof. David N. Cooper for providing access to the Human Gene Mutation Database. This work was supported by the US National Institutes of Health grants R01 LM009722 and R01 MH105524, and the Indiana University Precision Health Initiative.

References

1. Li, B., et al., *Bioinformatics* **2009,** *25* (21), 2744-50.

2. Adzhubei, I. A., et al., *Nat Methods* **2010,** *7* (4), 248-9.

3. Kircher, M., et al., *Nat Genet* **2014,** *46* (3), 310-5.

4. Ioannidis, N. M., et al., *Am J Hum Genet* **2016,** *99* (4), 877-885.

5. Bromberg, Y.; Rost, B., *Nucleic Acids Res.* **2007,** *35* (11), 3823-35.

6. Hecht, M., et al., *BMC Genomics* **2015,** *16 Suppl 8*, S1.

7. Folkman, L., et al., *J Mol Biol* **2016,** *428* (6), 1394-1405.

8. Zhao, N., et al., *PLoS Comput Biol* **2014,** *10* (5), e1003592.

9. Pejaver, V., et al., *bioRxiv* **2017**, 134981.

10. Hoskins, R. A., et al., *Hum Mutat* **2017,** *38* (9), 1039-1041.

11. Pejaver, V., et al., *Hum Mutat* **2017,** *38* (9), 1092-1108.

SESSION 2

Predicting the impact of mutations on splicing signals

Christophe Béroud

Genetics and Bioinformatics, Aix-Marseille University, Marseille, France

The spliceosome is known to be one of the most complex macromolecule from the cell. It is a ribonucleoprotein (RNP) complex responsible for the excision of intronic regions from eukaryotic RNA polymerase II transcripts. Its ability to differentiate introns from exons is mediated through various sequence signals including the 5' splice site (5'ss) or donor splice site, the 3' splice site (3'ss) or acceptor signal, the branch point as well as auxiliary sequences known to either enhance or repress splicing: Exonic Splicing Enhancers (ESE) and Exonic Splicing Silencers (ESS). We have progressively evolved from a model of a simple interaction between the BP, the 5'ss and the 3'ss to a much more complex model including ESS and ESE. Depending on spatiotemporal expression of various binding factors, these signals act either antagonistically or synergistically to decide the exon/intron fate of the various transcripts. If we have not solved the exact impact of each element, we have collected enough knowledge about the major sequence signals in order to accurately predict the impact of mutations on these signals. Various software allow to identify splicing motifs in the human genome while only few are able to predict the impact of mutations.

This is an important challenge in the high throughput sequencing era, as millions of variations are now produced per experiment, most of them being localized in non-coding regions. Moreover, it has been reported that up to 50% of disease-causing mutations might affect splicing, resulting in needs for accurate predictions of mutations impact on all splicing signals: branch points, splice sites and auxiliary splicing sequences.

Here, we will review the recent developments of the Human Splicing Finder (HSF) reference system (http://www.umd.be/HSF3/), which now includes an expert system to automatically interpret information in a context-dependent manner. We will demonstrate that this expert system allows rapid and accurate prediction of the impact of mutations on 5’ and 3' splice site, branch points as well as ESE/ESS.

Rethinking the 5 Splice Site Algorithms Used in Clinical Genomics

Gabe Rudy, Nathan Fortier

Golden Helix, 203 Enterprise Blvd, Suite 1, Bozeman, MT 59718

To fully interpret variants in the context of clinical genomics, as outlined by the ACMG interpretation guidelines, variants near canonical splice boundaries must be evaluated for their potential to disrupt gene splicing and thus be classified as a gene damaging mutation. The five splicing algorithms SpliceSiteFinder-like, MaxEntScan, GeneSplicer, HumanSplicingFinder, NNSplice have been canonized for this purpose in the clinical testing market by being implemented and made easily accessible in the first-mover bioinformatics tool Alamut. Although these algorithms vary wildly in their performance characteristics such as sensitivity and specificity, they are treated as black-box oracles on equal footing when being used by variant curators to classify variants.

In this presentation, I will review these algorithms and their technical strengths and weakness from the perspective of re-implementing them to support modern variant curation in the clinical testing market. Their respective performance on both historical and updated splice site databases will be reviewed, as well as how their historical models can be retrained on up-to-date splice site annotations.

Finally, I will preset our approach to enable the variant interpretation process including splice site predictions in our clinical software platform VarSeq. I will demonstrate how we improve on the status quo by employing these algorithms both in the batch annotation and genomic visualization contexts. While treating these algorithms as black boxes may lead to equally weighting them in the interpretation process, providing context and educational materials on their relative performance and output ranges allows for an informed and more precise prediction of the effect of the variant and ultimate impact to the patient’s diagnosis.

Evaluating the Evaluation of Cancer Driver Genes

Rachael Karchin

Institute for Computational Medicine, Johns Hopkins Biomedical Engineering, Baltimore, MD, USA

Sequencing has identified millions of somatic mutations in human cancers, but distinguishing cancer driver genes remains a major challenge. Numerous methods have been developed to identify driver genes, but evaluation of the performance of these methods is hindered by the lack of a gold standard, that is, bona fide driver gene mutations. Here, we establish an evaluation framework that can be applied to driver gene prediction methods. We used this framework to compare the performance of eight such methods. One of these methods, described here, incorporated a machine-learning-based ratiometric approach. We show that the driver genes predicted by each of the eight methods vary widely. Moreover, the P values reported by several of the methods were inconsistent with the uniform values expected, thus calling into question the assumptions that were used to generate them. Finally, we evaluated the potential effects of unexplained variability in mutation rates on false-positive driver gene predictions. Our analysis points to the strengths and weaknesses of each of the currently available methods and offers guidance for improving them in the future.

The Ensembl Variant Effect Predictor (VEP)

William McLaren[1]*, Laurent Gil[1], Sarah E. Hunt[1], Benjamin Moore[1], Harpreet Singh Riat[1], Graham R.S. Ritchie[1], Anja Thormann[1], Paul Flicek[1] and Fiona Cunningham[1]

[1] *European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD*

* wm2@ebi.ac.uk

The Ensembl Variant Effect Predictor (VEP) [1] is an open source, free to use tool for the annotation of genomic variants [2]. It is available as an easy-to-use web interface, as a standalone perl script and can also be accessed through the Ensembl REST API.

The VEP supports the annotation of both sequence variants with specific and well-defined changes (including Single Nucleotide Variants (SNVs), insertions, deletions, multiple base pair substitutions, microsatellites, and tandem repeats); and larger structural variants, including those with changes in copy number or insertions and deletions of DNA. Annotation of variants can be performed for data submitted in a number of formats, including: HGVS notation, VCF and variant identifiers from databases including dbSNP and ClinVar. Therefore, the VEP is suitable for variant interpretation in a wide range of study designs, from the analysis of a single variant to the annotation of millions of variants identified in whole-genome or whole-exome variant calls.

For all input variants, the VEP returns detailed annotation for predicted effects on transcripts, proteins and regulatory regions, including functional consequences, pathogenicity predictions and HGVS notations relative to the transcript and protein sequences. For known or overlapping variants, allele frequencies, phenotype information and literature citations can also be retrieved from the Ensembl databases. The output consists of an HTML or text format summary file and a primary results file in tab-delimited, VCF, GVF, or JSON format and also provides filtering options allowing prioritisation of variants based on consequence, allele frequencies and known phenotypes.

Recently, we have also developed Haplosaurus [3]; a tool that annotates consequences taking multiple variants into account using phased genotypes from a VCF file. This approach offers an advantage over the VEP analysis, which treats each input variant independently. By considering the combined change contributed by all the variant alleles across a transcript, the compound effects the variants may have are correctly annotated, giving a personalised reference proteome.

[1] The Ensembl Variant Effect Predictor: http://www.ensembl.org/Tools/VEP

[2] McLaren, W. et al. “The Ensembl Variant Effect Predictor” Genome Biology 2016, 17:122 doi: 10.1186/s13059-016-0974-4

[3] Haplosaurus: https://github.com/Ensembl/ensembl-vep

Validating and calibrating computational and functional approaches in BRCA and MMR

Sean Tavtigian PhD, Professor of Oncological Sciences, Co-Leader, Cancer Center Population Sciences Program

Huntsman Cancer Institute, University of Utah School of Medicine

Growth in the scale of disease predisposition genetic testing has led to the realization that the human gene pool harbors enormous numbers of individually rare sequence variants of uncertain significance (VUS), in turn driving development of a new discipline within genetics, “variant classification science”.

Methods for evaluation and classification of BRCA1/2 VUS have been developing since the early 00’s, and methods for classification of mismatch repair (MMR) gene VUS since the late 00’s. Both rules-based qualitative methods and Bayesian quantitative methods for classification of VUS in these genes have been developed. To date, the qualitative methods have proven more efficient for VUS classification in practice because they are better tuned to the data actually gathered through the practice of clinical cancer genetics.

In principle, however, quantitative Bayesian methods could become extremely efficient if (and its a big if) four criteria can be met: (1) computational evaluations of VUS can be calibrated, (2) moderate- to high-throughput functional assays can be calibrated, (3) these calibrations can be validated, and (4) circular dependencies between the two calibrations are largely avoided.

Both the BRCA and MMR systems have calibrated computational evaluations of VUS in their genes. And teams from both gene systems are close to calibrating relevant functional assays. But there have been no published validations or existing calibrations, and logical circularities between computational and functional assay calibrations have not been much considered.

Here, we will discuss progress towards replication and validation of the computational analysis for BRCA gene missense VUS. We will also describe progress towards calibration and validation of a functional assay for MMR gene missense substitutions in a context that highlights the over-fitting that can ensue when logical circularities creep into variant classification science.

Predicting the molecular mechanisms of genetic disease for protein coding variants

Predrag Radivojac

Department of Computer Science and Informatics, Indiana, University Bloomington, IN, USA

Genome interpretation involves understanding the influence of genomic variants on molecular events such as the nature of disruption of protein function, but also includes systemic impact such as phenotypic alteration and disease. This talk will introduce our computational methodology for predicting the influence of coding variants on human disease based upon protein sequence, structure, and function data. We will then integrate these methods with systems approaches towards better understanding human phenotypes in a cohort of whole-exome sequenced individuals with neurodevelopmental disorders. Finally, we will show results that include experimental characterization of the computationally-identified significant variants.

Findings from CAGI, the Critical Assessment of Genome Interpretation, a community experiment to evaluate phenotype prediction

Gaia Andreoletti[1], Roger A Hoskins[1], John Moult[2,3,*], Steven E. Brenner[1,*], CAGI Participants

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

[2]Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850.

[3]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742.

**Corresponding authors:*

brenner@compbio.berkeley.edu Phone: (510) 643-9131

jmoult@umd.edu Phone: (240) 314-6241

Interpretation of genetic variants plays an essential role in cancer, in monogenic disease, and increasingly in complex trait disease. The needs for variant interpretation range from basic research to informing profound clinical decisions, however, currently the field lacks a clear consensus on what kind of methods provide useful tools to interpret the data. The Critical Assessment of Genome Interpretation (CAGI, \'kā-jē\) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. CAGI participants are provided genetic variants and make predictions of resulting phenotype. Independent assessors evaluate the predictions by comparing with experimental and clinical data.

CAGI challenges thus far have included prediction of the biochemical impact of non-synonymous variants and of the impact of non-coding regulatory variants on gene expression; prediction of the impact of mutations in cancer driver genes on cell growth; prediction of individuals' complex trait status based on exome data; matching personal genomes to phenotypic trait profiles; and matching variant data to clinical diagnoses. Results from the CAGI experiments are now described in more than 20 articles, shortly to appear in a special issue of *Human Mutation*.

There have been notable discoveries throughout the CAGI experiments, and general themes have emerged. Some examples: For a number of challenges, independent assessment has found that top missense prediction methods are highly statistically significant, but individual variant accuracy is limited. Missense methods also tend to correlate better with each other than with experiment (for reasons that may reflect biases in the predictive methods but also in the experimental assays). Although overall missense accuracy is limited, there is a subset of variants where methods may be sufficiently reliable to providing strong evidence for clinical use. Protein three-dimensional structure-based missense methods do well in a few cases, while sequence-based methods have more consistent performance. Bespoke approaches often enhance performance. Interpretation of non-coding variants shows promise but is not at the level of missense.

In challenges using clinical data predictors have been able to identify causal variants that were overlooked in the initial clinical pipeline analysis. The results have also highlighted possible diagnostic ambiguities. Additionally, the results suggest that running multiple uncalibrated methods and considering their consensus may result in undue confidence in a pathogenic assignment, so we advise against this procedure. For complex traits, CAGI results suggest that current methods do not yet fully use the genetic information, and so are spurring the development of more effective methods.

Detailed information about CAGI may be found at https://genomeinterpretation.org.

Assessing mutational signatures and impact of loss-of-function genetic variants

Kymberleigh Pagel[1], Vikas Pejaver[1], Guan Ning Lin[2], Hyun-jun Nam[2], Matthew Mort[3], David N Cooper[3], Jonathan Sebat[2], Lilia Iakoucheva[2], Sean D Mooney[4] and Predrag Radivojac[1]*

[1]Department of Computer Science and Informatics, Indiana University Bloomington, IN,USA [2]Department of Psychiatry School of Medicine University of California, San Diego, La Jolla, CA, USA. [3] Institute of Medical Genetics, Cardiff University, Cardiff, UK [4] Department of Biomedical Informatics, University of Washington, Seattle, WA, USA

predrag@indiana.edu

Loss-of-function genetic variants are frequently associated with severe phenotypes, yet are frequently present in the genomes of healthy individuals. Currently available methods to assess impact of these variants primarily rely upon evolutionary conservation and place little or no consideration on the structural and functional implications for the protein. Further, these methods do not output information regarding specific molecular alterations that can potentially lead to the disease. We investigate protein features that underlie loss-of-function variation and develop a machine learning method, MutPred-LOF, to discriminate between pathogenic and tolerated variants which can also generate hypotheses on specific molecular events disrupted by the variant. To this end, we investigate a large set of human variants from the Human Gene Mutation Database (http://www.hgmd.cf.ac.uk/ac/index.php), the ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) database and the Exome Aggregation Consortium (http://exac.broadinstitute.org/). Our prediction method shows an area under the Receiver Operating Characteristic curve of 0.85 for all loss-of-function variants and 0.75 for the subset of proteins in which both pathogenic and neutral variants have been observed. We then applied MutPred-LOF to a set of 1142 de novo variants derived from case/control studies of neurodevelopmental disorders and find enrichment of pathogenic variants in affected individuals. Overall, our results highlight the potential of computational tools to elucidate causal mechanisms underlying loss of protein function in loss-of-function variants.

Unravelling the Genetic Mechanisms of Neonatal Diabetes Mellitus: An Egyptian Experience

Rasha Elkaffas*[1,3], Hanan Madani[1] , Badawy Alkholy[1] ,Noha Musa[2], Yomna Shaalan[2], Rania M.H. Elkaffas[2] , Mona Hassan[2], Mona Hafez[2], Sarah E Flanagan[4], Elisa De Franco [4], Khalid Hussain[3,5]

1 Clinical and chemical pathology department, Cairo University, Egypt.

2 Pediatric department, Cairo University, Egypt.

3 UCL GOS Institute of Child Health, UCL, UK.

4 Institute of Biomedical and Clinical Science, University of Exeter Medical School, UK.

5 Sidra Medical & Research center, Qatar.

Corresponding author email: rasha.kaffas@kasralainy.edu.eg

Neonatal Diabetes Mellitus (NDM) isa rare monogenic form of Diabetes Mellitus presenting typically before the age of 6 months[1]. NDM can be transient (TNDM) with 70% of these cases caused by a methylation defect in 6q24 or permanent (PNDM) with 40% caused by mutations in potassium channel subunits coding genes (*KCNJ11, ABCC8*)[2,3]. It may present in an isolated form or as a part of syndrome[4]. Several studies have reported the genetic causes of NDM among different populations but to the best of our knowledge no previous studies had reported the genetic mechanisms of NDM among the Egyptian population.

Our aim was to understand the genetic causes of NDM in Egyptian neonates presenting before the age of 6 months referred to a tertiary diabetes centre between the period 2013-2016.

In a cohort of 22 patients, anti-GAD antibodies and C-peptide testing were done in some cases to exclude type 1 Diabetes Mellitus. Sanger sequencing for *KCNJ11, ABCC8, INS* and *EIF2AK*3 genes was done as a first tier genetic analysis for all cases. In absence of a causative mutation in these genes, we further undertook methylation analysis of 6q24 in cases presenting with TNDM, while in those presenting with PNDM targeted next generation sequencing (tNGS) for the coding regions and conserved splice sites of another 18 known genes to cause NDM were done. Targeted gene analysis was done in 4 syndromic cases with a distinctive phenotype suggesting a candidate gene.

Genetic causes were identified in 15/22 (68%) patients. These included 10 missense, 3 nonsense, 2 frameshift mutations and a complete loss of maternal methylation on chromosome 6q24. Mutations involved different genes; 4 in *ABCC8*, 3 in *GCK*,3 in *EIF2AK3*, 2 in *KCNJ11*, 1 in *NEUROD1*, 1 in *ZFP57* and 1 in *SLC19A2*. Most of these patients (11/15) had previously reported mutations. In 4 patients, no mutation was identified in the first tier analysis but insufficient DNA was available for tNGS. In 3 patients no genetic cause could be identified.

Acknowledgments:This study was financially supported by the Egyptian Ministry of Higher Education-Culture Affairs and Missions Sector and by the Wellcome Trust. SEF has a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number: 105636/Z/14/Z).

References:

1. Carmody D, Støy J, Greeley SA, Bell GI & Philipson LH (2016). A Clinical Guide to Monogenic Diabetes. In: Genetic Diagnosis of Endocrine Disorders, 2nd edition; Weiss RE, Refetoff S, Eds. Philadelphia, PA, Elsevier, pp. 19–30.
2. Ellard S et al. (2013). Improved genetic testing for monogenic diabetes using targeted next-generation sequencing. Diabetologia. 56: 1958–1963.
3. De Franco E et al. (2015). The effect of early, comprehensive genomic testing on clinical care in neonatal diabetes: an international cohort study. Lancet. 386: 957–963.
4. Murphy R, Ellard S, Hattersley AT (2008). Clinical implications of a molecular genetic classification of monogenic beta-cell diabetes. Nat Clin Pract Endocrinol Metab. 4: 200–213.