Methods and tools for assessing the impact of genetic variations: The 2017 annual scientific meeting of the Human Genome Variation Society

William Oetting, Christophe Béroud, Steven E Brenner, Marc S Greenblatt, Rachel Karchin, Sean D Mooney

Methods and Tools for Assessing the Impact of Genetic Variations: The 2017 Annual Scientific Meeting of the Human Genome Variation Society

William S. Oetting[1], Christophe Béroud[2], Steven E. Brenner[3], Marc S. Greenblatt[4], Rachel Karchin[5], Sean D. Mooney[6]

1. Department of Experimental and Clinical Pharmacology, University of Minnesota, Minneapolis, Minnesota, USA
2. Aix Marseille Universite, Marseille, France
3. Department of Plant and Microbial Biology, University of California, Berkeley, CA
4. Department of Medicine, University of Vermont, Burlington, VT
5. Departments of Biomedical Engineering/Oncology and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MA
6. Buck Institute for Research on Aging, Novato, CA

Contact information:

William S. Oetting, Ph.D.

Department of Experimental and Clinical Pharmacology

7-115 Weaver-Densford Hall

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.23393.

308 Harvard Street S.E.

Minneapolis, MN 55455

Email: oetti001@umn.edu

Grant Information: NIH U41 HG007346

Keywords: HGVS, meeting report, genetic variants, methodology

Introduction

The 2017 Annual Scientific Meeting of the Human Genome Variation Society (HGVS; www.hgvs.org) was held on the 17th of October in Orlando, Florida, USA with the theme of "Methods & Tools for Assessing the Impact of Genetic Variations". The meeting was opened by Marc Greenblatt of the University of Vermont Medical Center, Burlington, Vermont. This is an exciting time in genetic variation analysis. In the beginning of the Human Genome Variation Society (HGVS; http://www.hgvs.org), only a few individuals were interested in creating tools to predict the functional impact of genetic variants, but this has now become a mainstream portion of the genetics world. Both researchers and clinicians are awash in full exome and genome sequences, with the myriad of variants they harbor. 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 predict their phenotypic impact. In this meeting cutting-edge research and practical approaches involving methods for interpreting human genetic variants were presented.

Session 1

Session 1 was chaired by Marc Greenblatt. The first invited presentation was given by Mark Gerstein from the Program in Computational Biology and Bioinformatics at Yale University, Connecticut, who spoke on "Prioritizing somatic variants". Personal genomics will increasingly play an essential role in providing precision medicine to the general public through identification of functional genetic variants, both in the germline and somatic tissues. This is becoming apparent in individualizing cancer therapy. A tumor typically contains 2 to 8 driver mutations which confer a selective growth advantage to the tumor cell. Unfortunately the tumor also contains thousands of passenger mutations which have no obvious direct or an indirect effect on tumor progression. It is important to differentiate between these two classes of variants to better enable more precise diagnostics and targeted therapies. A number of approaches have been created to identifying key variants either by predicting the functional impact of the variant, or by considering the recurrence of a mutation in tumors. Two programs, Annotation of Loss-of-Function Transcripts (ALoFT; ALoFT.gersteinlab.org) and Frustration (github.com/gersteinlab/Frustration) provide a means to predict the functionality of a specific coding variant. A major problem which can confound this type of analysis is the existence of multiple transcripts from a single locus. The Variant Annotation Tool (VAT; http://VAT.gersteinlab.org) was created to take into account transcript isoforms when determining the functional consequences of a variant on protein function. Together with VAT, ALoFT or Frustration can characterize the functionality of variants and help identify the drivers of tumors. ALoFT creates an impact score, along with a confidence score, to determine if a variant is benign or deleterious. This method uses annotation of functional domains and conservation of sequence to derive the ALoFT score. Frustration attempts to determine the effect of a coding variant by predicting its effect on protein folding by evaluating localized frustration changes due to amino acid substitutions and their effect on changes in localized energies of protein folding. The

functional effects of non-coding variants can be predicted using FunSeq (http://FunSeq.gersteinlab.org) which integrates evidence from multiple datasets. The program uses annotation of transcriptional factor binding sites, open chromatin domains and sequence conservation. This is an entropy based method for weighing many genomic features which may impact transcription. Variation in somatic mutations is closely associated with chromatin structure (topologically associated domains; TADs) and replication timing which will alter the background mutation rate. MrTADFinder (http://github.com/gersteinlab/MrTADfinder) helps identify replication TAD boundaries as possible locations for functional non-coding variants. Mutation recurrence is another method to determine functionality of variants in tumors. Two methods, Mutations Overburdening Annotations Tool (MOAT; http://MOAT.gersteinlab.org) and Large-scale Analysis of Variants in noncoding Annotations (LARVA; http://LARVA.gersteinlab.org), were designed to analyze the probable functionality of a variant based on its reoccurrence in tumors. These methods require well curated databases. These different methods can help identify the driver mutations and better characterize (and treat) tumors.

The second talk of this session was by Ben Rodriguez of the Population Sciences Branch, NHLBI, Framingham, Massachusetts, who spoke on "GRASP v3: An updated GWAS catalog and contrast to similar catalogs". Mining GWAS data drives scientific discovery and research productivity. GRASP (Genome-wide Repository of Associations between SNPs and Phenotypes; http://grasp.nhlbi.nih.gov) is a leading repository of published single nucleotide polymorphism (SNP) associations with human traits, including methylation and expression quantitative trait loci (eQTL). GRASP began as an open access database in 2008 initially with results from 118 studies (56,000 data points) obtained from different cohorts including diabetes related, cardiovascular and neurological studies. GRASP v3.0 will contain results

from over 3,300 studies. The methodology used to obtain data for GRASP include: (1) controlled vocabulary searches, (2) review of tens of thousands of abstracts for inclusion criteria, (3) extraction of all associations with a $p<0.05$, and (4) associations with given SNP having a clearly distinct phenotype. Following QC, each record is associated with study-level information including phenotype, sample size, ancestry and publication information. It should be noted that eQTLs and associations in the MHC locus are over-represented among published significant results. GRASP contains more results for specific sub-phenotypes per SNP compared to the NHGRI-EBI catalog. This increased specificity may more accurately reflect GWAS study designs. The GRASP v3 update will expand clinically relevant results for chronic diseases that challenge public health.

The second invited speaker was Weiva Sieh, of Population Health Science and Policy and of Genetics and Genomic Sciences at the Icahn School of Medicine at Mount Sinai, New York, who spoke on "Predicting the pathogenicity of rare missense variants". Rare coding variants play an important role in disease but are often overlooked. Healthy individuals carry several hundred non-synonymous variants. Predicting which are functional is important to help identify disease-causing variants and tailor prevention and treatment according to each person's genetic risks. There are many *in silico* prediction tools for predicting the pathogenicity of variants, but they often disagree. Ensemble methods that combine scores from multiple tools often perform better. Of these methodologies, few focus on rare variants. This was the motivation for the creation of the Rare Exome Variant Ensemble Learner (REVEL; https://sites.google.com/site/revelgenomics/). The method was developed by combining 18 scores from 13 individual tools including MutPred, VEST, PolyPhen, SIFT and MutationTaster. The method was trained using recently reported HGMD disease mutations and rare neutral variants, and evaluated using two independent test sets. REVEL was more

accurate for differentiating between benign and disease producing variants than other methods when tested on 1,953 pathogenic and 2,406 benign variants from ClinVar, or 935 disease mutations from SwissVar and 141,051 neutral variants. REVEL was also more accurate for interpreting rare variants (MAF<0.5%). Pre-computed REVEL scores for >80 million theoretically possible human missense variants, along with the sensitivity and specificity for a wide range of cutpoints, can be found at sites.google.com/site/revelgenomics.

The last speaker of this session was Vikas Pejaver from the Department of Biomedical Informatics and Medical Education and the eScience Institute, University of Washington, Seattle, who presented his talk on "Probabilistic prediction of the different notions of missense variant impact". Sources of evidence for the functionality of a genetic variant incudes conservation of the amino acid between different species, co-segregation of the variant with disease, effect of the variant on protein structure and other functional properties and effects in model organism studies. Different computational tools are being used to predict functionality *in silico* based on the above criteria. More recently, machine learning approaches are being trained on either variants experimentally shown to alter protein function, or on variants shown to be associated with disease. But function altering variants are not always disease causing. In some cases reported functional effects are determined in other species and may not accurately predict their function in humans. Additionally, predicted changes, especially 'small' changes, in protein function may not result in a change in the phenotype. Despite this, previous studies have suggested that such predictors do predict disease-associated variants accurately. The converse question was presented, "Can predictors trained on binary labels of 'pathogenic' and 'benign' generalize to the prediction of real-valued functional effects of missense variants (*e.g.*, protein function, cellular phenotype)?" Using CAGI experiment (https://genomeinterpretation.org) based variants

which were associated with disease or on observed functionality the programs MutPred and MutPred2 (mutpred.mutdb.org) were tested and compared. MutPred2 uses conservation of sequence, structural and functional properties of protein and other factors within a neural network ensemble. Testing three tasks: 1) Identifying variants associated with disease (*MRN* and *CHEK2* and breast cancer), 2) Prediction of *in vitro* enzyme activity (N-acetyl-glucosaminidase glycolyase, *NAGLU*) and 3) Prediction of effects on CBS-dependent cellular growth using as determined by a yeast complementation assay, MutPred2 was consistently more accurate than MutPred. It was also found that predictors tend to perform better when biochemical and functional evidence is available.

Session 2

Session 2 was chaired by Steven Brenner of the University of California, Berkeley. The first invited speaker of this session was Christophe Béroud of the Genetics and Bioinformatics, Aix-Marseille University, France, who spoke on "Predicting the impact of mutations on splicing signals**".** Intronic variants which affect transcript splicing account for a significant number of mutations associated with disease. Additionally up to 30-50% of exonic variants might also impact transcript splicing. Because we will be collecting millions of variants through next-generation DNA sequencing (NGS), we need efficient methods to identify those that affect splicing. The role of splicing is to remove introns from pre-messenger RNA and is a complex process requiring the recognition of multiple RNA degenerated splicing motifs. Thus, the prediction of a variants' impact on those signals is difficult to assess and vary according to the splicing signal (splice sites, branch point or auxiliary sequences). We therefore need prediction tools able to predict the impact of variants on 5'splice site and 3'splice site motifs as well as branch point sites yet taking into account exonic splice site

motifs such as exonic splicing enhancers and silencers (ESE and ESS). To this end various approaches and tools have been created. These include Human Splicing Finder, MaxEntScan, NNSplice and SplicePort. Human Splicing Finder (HSF; http://www.umd.be/HSF3) uses a position weight matrices (PWM) approach. In testing pathogenic and non-pathogenic splice sites mutations from *BRCA1* and *BRCA2*, HSF had a positive predictive value (PPV) of 0.986 and a negative predictive value (NPV) of 1.0. This was more accurate for this set of mutations than MaxEntScan (PPV=0.714 and NPV=0.750). The HSF system also successfully predicts the impact of mutations on branch point sequences as all such mutations have been accurately predicted. Moreover, HSF now contains specific matrices for non-canonical splice sites and can be very helpful for VUS re-evaluation. Additional motifs which affect transcript splicing are ESEs and ESSs and intronic splicing enhancers (ISE) and silencers (ISS). Predicting the impact of mutations on ESE and ESS motifs is more difficult because of the paucity of experimental data. Nevertheless, HSF now provides an improved accuracy. Methods like HSF are now available for accurate interpretation of variants which affect splicing and will aid in helping select functional variants from the large number of variants identified in NGS data.

The topic of accurately identifying variants which affect splicing was continued by Gabe Rudy of Golden Helix, Bozeman, Montana, who presented his talk “Rethinking the 5 splice site algorithms used in clinical genomics”. There are several algorithms for splice site prediction. In this report, five methods were compared for their accuracy in predicting the consequences of variation at splicing motifs: Human Splice Finder-like, SpliceSiteFinder-like, MaxEntScan, NNSplice and GeneSplicer. The test set consisted of 20,000 known splice sites extracted from the human GRCh38 reference sequence using exon boundaries specified by NCBI RefSeq Genes. Additionally, 20,000 false splice sites from the HS3D splice site

dataset were added to the test set. For acceptor splice site results all methods did very well. For donor splice sites, GeneSplicer had the best performance for all metrics. MaxEntScan also performed well. SpliceSiteFinder-like and HumanSpliceFinder-like both had poor precision due to high false positive rates. This was due to not utilizing information in the dependencies between bases in which PWM based methods fail to capture because they that treat each base independently. It should be noted that the "-like" tools did not embody the full array of methods in the original authors' implementation. While these methods can provide useful *in silico* predictions to help prioritize and provide supplemental evidence when interpretation a variant in a clinical context, they cannot be used on their own to show that a variant disrupts splicing in a clinical context.

The second invited speaker in this session was Rachel Karchin of the Institute for Computational Medicine, Johns Hopkins Biomedical Engineering and Oncology Departments, Baltimore, Maryland, who spoke on "Evaluating the evaluation of cancer driver genes". Cancer is an evolutionary process driven by somatic alterations in driver genes and cancer cells are positively selected due to numerous alterations that result in an increased rate of clonal expansion compared to normal cells. The number of driver gene mutations in tumors is small compared to the number of benign passenger mutations. Identifying driver genes is important for precision cancer treatment but their identification is challenging. Informatics/statistical methods are critical for discriminating drivers and passengers. A number of different methods have been proposed to identify driver genes in tumors. A machine learning-based ratiometric approach was presented as an example (github.com/KarchinLab/2020plus). Ratiometric methods assess the composition of mutations in a gene normalized by total number of mutations in all genes. Several driver gene preditors identify significantly mutated genes by modeling the background somatic

mutation rate for all genes and then identifying genes exhibiting an elevated mutation rate. However, there can be difficulties in estimating the background mutation rate, because it is highly variable for many reasons that are not completely understood. Tumors with a high mutation rate can be particularly problematic for identification of "significantly mutated" genes, resulting in false-positive driver gene predictions. Dr. Karchin's team designed an evaluation protocol to compare different methodologies to identify driver genes in tumors, which includes a large somatic mutation dataset covering 30+ tumor types. The protocol is designed to work in the absence of a "gold standard" set of driver genes, and combines several criteria, including overlap of predicted drivers across multiple methods, a method's internal consistency, and proper calibration of p-values.

The final talk in this session was by Benjamin Moore of the European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom, who spoke on “The Ensembl Variant Effect Predictor (VEP)”. The Variant Effect Predictor (VEP; https://github.com/Ensembl/ensembl-vep) is a stand-alone tool with a web-interface that predicts the functional effects of genomic variants including SNPs, insertions/deletions (in/dels), copy number variants (CNVs) and other structural variants. Data input for VEP can be variant coordinates, a variant call file (VCF), HGVS nomenclature or variant IDs. The VEP returns detailed annotation for predicted effects of variants 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. Predicted consequences for regulatory regions can be limited to specific cell types using data from the Ensembl Regulatory Build. Pathogenicity predictions are accomplished using different algorithms such as SIFT,

PolyPhen, FATHMM and MutationTaster. Additionally, the Haplosaurus is a recently developed VEP-like tool that uses phased genotype data to predict the consequences of multiple variants. This can help identify cryptic stop gained codons or protein truncating variants (PTV) created by multiple variants as well as variants which rescue PTVs and frameshifts. This is particularly important for variants which frame rescue small in/dels or multiple missense variants in cis.

Session 3

Session 3 was chaired by Christophe Béroud. The first invited speaker was Sean Tavtigian of Oncological Sciences, Huntsman Cancer Institute, Salt Lake City, Utah, who spoke on "Validating and calibrating computational and functional approaches in BRCA and MMR genes". Identification of mutations driving tumors is critical to personalized treatment. Classification models attempt to score these variants on a continuum from 5 (pathogenic; posterior-probability >0.95) to 1 (neutral; posterior-probability of pathogenicity <0.001). Unfortunately, many somatic variants are classified as variants of unknown significance (UV/VUS; posterior-probability between 0.05 and 0.95). Additional information can update prior probability of their classification to (hopefully) a more accurate posterior classification using Bayes rule using likelihood ratios or odds ratios of causality. The updated *in silico* models from the Breast Cancer Information Core (BIC) to classify variants include co-segregation and co-occurrence with disease, summary family history and tumor histopathology to classify variants in breast cancer. This can be further enhanced by the knowledge that most functional missense mutations occur in specific protein domains including the ring domain and BRCTs of BRCA1 and the DNA binding domain and PALB2 interaction and RAD51 dissasembly domains of BRCA2. Calibration with these methods

improves the accuracy of variant classification. The creation of large datasets for validation and replication of the methodology are necessary for this. Test sets of variants from Myriad (68,000) and Ambry (100,000) were used to this end. In the case of mismatch repair (MMR) genes, variants were classified using a cell free *in vitro* assay for MMR activity. Analysis addressed the calibration of the assay, comparing the functional assay outputs with prior classifications that were independent of *in vitro* data. A final calibration curve successfully separated functional from non-functional variants. A 2-component classification, incorporating previously determined *in silico* scores of pathogenicity with functional assay results, could help classify MMR gene variants with only a small amount of additional clinical data. Using the American College of Medical Genetics (ACMG) combining rules provide better supporting evidence for pathogenicity scores. Improvement of these models should allow for more accurate computational scoring of variants.

The second invited speaker was Predrag Radivojac of the Department of Computer Science, Indiana University Bloomington, who presented his talk on "Predicting the molecular mechanisms of genetic disease for protein coding variants". There are several tools which predict whether a variant is pathogenic or not, but they do not predict the underlying mechanisms of disease. There is a need to include additional types of information to predict the functional consequences of variants. However, functional analysis of variants can be complex because of multiple functional sites in a single protein. For example, in the tumor suppressor protein p53, the p.Arg175His variant likely alters metal binding, the p.Val143Ala may alter stability, the p.Lys120Arg affects an acetylation site, and several others impact macromolecular binding. Additionally, a specific amino acid substitution can result in a loss of function or gain of function (e.g., a disease-associated variant p.Asp374Tyr in PCSK9 increases catalytic activity 10-fold). MutPred2 was created to predict these different types of

consequences of amino acid substitutions. The program was applied to identifying effects of *de novo* variants in a cohort of individuals with autism and to subsequently predict autism genes. Controls were unaffected sibs. The most significant signal was from variants found in proteins which altered macromolecular binding sites, with catalytic activity and phosphorylation also showing statistical significance. Experimental validation of variants with predicted effects on protein-protein binding was done using yeast 2 hybrid experiments. Preliminary work was also shown regarding differences of functional variants between populations. It was found that common sequence variants affect molecular function more often than rare variants. There is optimism that the effect of variants can be accurately determined, but there is still a long way to go in the quest of fully interpreting one's genome.

The final talk of the session and the meeting was by Steven Brenner who spoke on "Findings from CAGI, the Critical Assessment of Genome Interpretation," a community experiment to evaluate phenotype prediction. All individuals have many variants in their DNA sequence. Current computational methods to predict the impact of these variants are not accurate enough to evaluate all genomic variants for clinical use, with most producing many false positives. Each GAGI edition has about a dozen challenges to assess different prediction methods. In one challenge, variants within the α-N-acetylglucosaminidase (*NAGLU*) gene, associated with San Filippo disease (mucopolysacchraidosis type IIIB), were used as a test dataset. A total of 165 missense variants with known functional effects were presented to predictors. The 10 research groups participating in this challenge used 17 methods. Independent assessment of this challenge 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. Variants with a nearly complete loss of activity were predicted most accurately. Newer approaches employed in CAGI often enhance performance compared to more established methods. In a challenge using clinical data, predictors were able to identify causal variants that were overlooked in the clinical analysis; these variants were then clinically confirmed. 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 impact assignment. Some of the lessons learned from CAGI were: 1) In general predictions are statistically significant. 2) However, predictive accuracy for specific variants is low. 3) Bespoke approaches often enhance performance. 4) Missense methods tend to correlate better with each other than with experiments. 5) Predictors were able to identify causal variants overlooked by a clinical laboratory. 6) Interpretation of non-coding variants is not at the level of missense variants. 7) Use of multiple uncalibrated missense variants impact prediction methods and is not advised. To further the assessment of these predictive models, CAGI 5 has been launched (http://genomeinterpretation.org). In closing, Dr. Brenner also mentioned the comprehensive special issue on CAGI that was recently published (http://dx.doi.org/10.1002/humu.2017.38.issue-9).

Closing Remarks

The meeting concluded with an appreciation for how far the field has progressed in recent years, and optimism that the tools that were discussed will lead to the advances that will be required to interpret and manage the large volume of data that will be arriving as genome sequencing expands.

Acknowledgements

The Scientific Program Committee of Christophe Béroud, Steven E. Brenner, Anthony J. Brookes, Marc S. Greenblatt, Rachel Karchin and Sean D. Mooney would like to thank Rania Horaitis for her professional help in running this HGVS annual scientific meeting. Dr. Brenner's participation was supported by NIH U41 HG007346. This meeting of the HGVS was chaired by Marc Greenblatt, Steven Brenner, and Christophe Béroud. The authors would like to thank the speakers for their help in the preparation of this report.