MEETING REPORT

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Non-Coding Variation: The 2016 Annual Scientific Meeting of the Human Genome Variation Society

William S. Oetting,[1]* Christophe Béroud,[2] Steven E. Brenner,[3] Marc Greenblatt,[4] Rachel Karchin,[5] Sean D. Mooney,[6] and Shamir Sunyaev[7]

[1]*Department of Experimental and Clinical Pharmacology, University of Minnesota, Minneapolis, Minnesota;* [2]*Aix Marseille Universite, Marseille, France;* [3]*Department of Plant and Microbial Biology, University of California-Berkeley, Berkeley, California;* [4]*Department of Medicine, University of Vermont, Burlington, Vermont;* [5]*Departments of Biomedical Engineering and Oncology, Institute for Computational Medicine, Johns Hopkins University, Baltimore, Massachusetts;* [6]*Department of Biomedical Informatics, University of Washington, Seattle, Washington;* [7]*Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts*

Communicated by Mark H. Paalman

Received 5 December 2016; accepted revised manuscript 2 January 2017.
Published online 5 January 2017 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.23169

Hum Mutat 00:1–4, 2017.

KEY WORDS: HGVS; meeting report; non-coding; genetic variation; variant interpretation

Introduction

Significant advances have been made in recent years in the assessment and interpretation of the non-coding DNA that comprises 98% of our genome. Studying non-coding variants helps us better understand gene regulation and expression, non-coding RNA, and other non-coding genome functions. The 2016 annual scientific meeting of the Human Genome Variation Society (HGVS; http://www.hgvs.org) was held on the 18th of October in Vancouver, British Colombia, Canada, on the topic of "Non-Coding Variation." By using multiple data sets and wide-ranging methods including in vitro assays, high-throughput functional assays, statistical, machine learning, and other computational analyses, important sequence motifs, and the effect of variants in these regions, are being better understood. This year's meeting explored these methodologies and the effects of genetic variation in non-coding regions on human disease and other phenotypes.

Understanding Variation in Non-Coding Regulatory Elements

The meeting was opened and the first session chaired by Steven Brenner of the Department of Plant and Microbial Biology, University of California, Berkeley. Research has focused on coding variation and its effect on protein function, but we are now in a position where we can identify important non-coding regions that control gene expression. For a subset of genes, the effects of variation in non-coding sequences that alter gene expression levels have been identified. For our understanding of this type of variation to expand throughout the entire genome, we need to have better tools to study these regions and their variants.

*Correspondence to: William S. Oetting, Department of Experimental and Clinical Pharmacology, 7–115 Weaver-Densford Hall, 308 Harvard Street S.E., Minneapolis, MN 55455. E-mail: oetti001@umn.edu

To begin the discussion in this session, Nadav Ahituv of the Department of Bioengineering and Therapeutic Sciences and the Institute for Human Genetics at the University of California, San Francisco, spoke on "Functional characterization of gene regulatory elements." Enhancers can be defined as "the promoters of the promoters," controlling gene expression when and where promoters are turned on and at what levels. Unlike nucleotide substitutions in the coding regions of genes that are more easily understood, determining the functionality of nucleotide variants in enhancers is difficult. One well-studied enhancer, the sonic hedgehog (SHH) limb enhancer ZRS, is a case in point. Variation in ZRS in humans alters SHH expression and limb development. One variant in ZRS, a 13 bp insertion, resulted in the same phenotype as a single nucleotide change, though one would expect that the insertion would result in a much more severe phenotype. To understand why this is, we need a better understanding of the code and "grammar" of enhancers. One strategy would be to mutate every nucleotide in a previously characterized enhancer and measure functionality for each altered sequence. A method termed massively parallel reporter assays (MPRAs) was used for this approach. All nucleotides in three characterized liver enhancers were altered creating ~100,000 unique sequences. Each altered enhancer was placed in front of a minimal promoter and a reporter gene and the expression level of each construct was measured. A barcode was added to each construct to individually identify the expression products of each novel enhancer sequence. The entire library of constructs was put into a mouse using hydrodynamic tail vein injection, directing the expression vectors into the liver. After time, cells from the liver were processed for RNA and barcode expression levels were analyzed using RNA-seq. The presence of the barcode allowed each construct to be individually analyzed. Using this approach, specific transcription factor (TF) binding sites could be identified, based on their sequence changes having a large effect by either reducing or increasing this expression. For the telomerase reverse transcriptase (TERT) promoter, where some somatic variants in this element are found in several cancers and are associated with increased TERT expression in tumors (associated with increased telomere lengthening), it was found that variants at nucleotide –146 and –124 were associated with this increased expression. Additionally, MPRA was used to test regulatory "grammar" by testing how variation in the number of TF binding sequences, their combinations or their order can have an effect on gene expression. Using MPRAs to simultaneously test thousands of sequences for their regulatory activity will help us understand which nucleotides are important for normal gene expression and

how variation will alter expression levels. Experiments such as these will help investigators unlock the code for regulatory elements as well as understand the effect of variation on these elements.

The use of populations that have adapted to extreme environmental conditions can be helpful in identifying important regulatory elements that have allowed these populations to differentially respond to or tolerate these challenging environments compared with other populations. This was the topic presented by Rasmus Nielsen of the University of California, Berkeley, who spoke on "The genomics of human physiological adaptation in Inuits." Humans can adapt to a wide range of environments and genetic variation plays a major role in this adaptation. Variants can result in a positive selection, where it confers a fitness advantage, or negative selection, where sequences are conserved since variants are deleterious and removed from the population. In the case of the Inuits, they successfully live on a very rich fat diet, high in omega-3 poly-unsaturated fatty acids (PUFAs), due to consuming mainly seal and fish. Additionally they have had to adapt to a very cold environment. This successful adaptation is thought to be in part due to acquired genetic variants. These changes have not come without a cost. Individuals on the traditional diet of seal and fish do not have type 2 diabetes (T2D), but introduction of a high carbohydrate Western diet has greatly increased the incidence of this disease. To identify candidate genes, and their variants that provide greater fitness in a challenging environment or greater risk to T2D, comparisons were made between the Inuits and other populations. A variant in the TBC1 domain family, member 4 (TBC1D4) gene (p.Arg684Ter) was associated with increased risk of T2D. Individuals who were homozygous for this variant and ate a Western diet were at high risk for T2D. The T2D risk for individuals who were heterozygous for this variant was not much different than those homozygous for the original allele. TBC1D4 is associated with the regulation of glucose uptake. The traditional diet was low in glucose-yielding food but in the Western diet these are prevalent. Thus a change in lifestyle, along with the variants that alter glucose uptake, resulted in a profound risk for T2D. A second group of genes identified was the fatty acid desaturase enzymes FADS1 and FADS2. These genes encode the rate-limiting enzymes in the pathway responsible for endogenous synthesis of long-chained PUFAs. The genetic variants in Inuit downregulate the endogenous production of PUFAs to compensate for their increased dietary intake of these fatty acids. The existence of unique genetic variants controlling synthesis of PUFAs in Inuit implies that lessons from Inuit dietary intake of omega-3s, and fish oils generally, cannot be extrapolated to other populations. The fact that Inuit are extremely healthy on a diet rich in fatty acids from marine sources does not necessarily imply that other groups would be similar healthy on the same diet. Other population-specific variants that allow adaptation to extreme environments are being identified. In many cases, the adaptive variants are found in non-coding regions affecting the expression levels of the target gene. One example is TBX15, which has a number of functions including regulating the differentiation of Brown and brite adipocytes, which produce heat via limiting oxidation when stimulated by cold temperatures. It is hypothesized that this haplotype originated from the Denisovan population, which may also have been cold adapted. Overall, this presentation showed the strength of understudied populations to make discoveries, but also their limitations in extrapolating from one population to another.

The third talk was by Patrick Short of the Wellcome Trust Sanger Institute, Cambridge, United Kingdom, who spoke on "De novo mutations in fetal brain active conserved non-coding elements contribute to severe developmental disorders." De novo mutations are thought to play an important role in undiagnosed developmental disorders and most of the contributing genetic variants are thought to reside in non-coding regions. To study the role of these genetic variants, the Deciphering Developmental Disorders study has analyzed approximately 8,000 trios containing children with severe undiagnosed developmental disorders. In addition to the protein coding exons, an additional 4,000 highly conserved non-coding elements and 1,500 putative enhancers were sequenced in all trios. There was an excess of recurrently mutated regulatory elements but none were of genome wide significance. An excess of de novo mutations was identified in the conserved non-coding elements predicted to be active in fetal brain. In silico methods, including Combined Annotation Dependent Depletion (CADD), are strong predictors of selection strength in coding regions, but data from 7,000 unaffected parents showed that these tools were unable to identify strongly deleterious mutations in highly conserved non-coding elements. Over all, de novo mutations in regulatory elements can be identified but it is difficult to determine which are contributing to disease. Additional work needs to be done to understand the "grammar" of these regulatory elements so that we can determine if candidate variants are functional.

Continuing on this topic of interpreting variation in non-coding regulatory elements, Iuliana Ionita-Laza from the Department of Biostatistics, Columbia University, New York, spoke on "Integrative statistical approaches for functional prediction of genetic variation." An important goal is to create methods that predict the functional effects of genetic variation in a tissue/cell type-specific manner. Unsupervised statistical approaches toward this problem are appealing. Training sets consist of non-coding variants along with tissue-specific functional annotations (histone modifications and DNase hypersensitivity) from the ENCODE and Roadmap Epigenomics projects. Using a Bayesian approach, a posterior probability that a variant is functional in a specific tissue is calculated. The authors validate their predictions using results from the GTEx projects and summary statistics from GWAS studies. These tissue-specific functional predictions can result in a better understanding of genes and pathways and their impact on complex traits.

Variants Affecting Regulation and HGVS Nomenclature

The second session was chaired by Bruce Gottlieb of McGill University, Montreal, Quebec, Canada. Barbara Engelhardt of the Computer Science Department, Princeton University, New Jersey, and part of the GTEx Consortium, spoke on "Trans-eQTLs across 44 tissues." Expression quantitative trait loci (eQTLs) identify genetic variants that alter expression levels of a gene. eQTLs can be identified in studies that contain both genotypes and expression levels of multiple genes. Cis-eQTLs are variants that are typically near or in gene promoters that affect gene expression in an allele-specific way. Trans-eQTLs differ in that the variant is far removed from the targeted gene, or even on a different chromosome. Questions about trans-eQTLs include: Are they more or less tissue specific than cis-eQTL? What are the relative effect sizes of trans- versus cis-eQTLs? What are the mechanisms of distal regulation of gene expression levels? Can trans-eQTLs be reproduced across studies? Using GTEx data containing 7,051 transcriptomes from 44 tissues from 449 individuals, off-chromosome genotypes were tested against tissue-specific gene expression levels. Principal components analysis for population stratification, genotyping platform, and probabilistic estimation of expression residuals were used as co-variants in a linear model to test for association (matrix-eQTL), and the Benjamini–Hochberg procedure was used to quantify the false discovery rate. Many statistically significant hits were identified but most were artifacts, which

were filtered using a comprehensive post-processing pipeline. Overall, 81 genes were identified across 18 tissues with one or more significant trans-eQTLs. Thyroid and testis had the largest number of trans-eQTLs. A few inferences could be made about the identified trans-eQTLs. Trans-eQTLs appear to be substantially more tissue specific than cis-eQTLs. Additionally, the trans-eQTL effects could be reproduced using data from TwinsUK and in matched tissues. It appears that trans-eQTLs are master regulators effecting the expression of many genes downstream. Additionally, it is most likely trans-eQTLs will also have cis effects, while most cis-eQTLs will not have trans effects.

In the next talk, Xinshu (Grace) Xiao of the Department of Integrative Biology and Physiology, from the University of California, Los Angeles, spoke on "Genetic variations in the regulation of alternative splicing." Typically, variations in consensus splicing sites have been shown to affect splicing, but variations in intronic variants may also influence splicing. A method termed Intronic tag SNVs for Genetically Modulated Alternative Splicing (iGMAS) was created to identify variants, including intronic SNPs, which alter the splicing of exons. Using this technique to analyze nascent and mature RNA with changes in splicing, more than 600 GMAS SNVs were identified, among which >100 were in LD with significant GWAS SNPs, which may help identify the functional SNP. 71% of GMAS SNVs were found to be intronic. Possible mechanisms include altering protein binding sites involved in splicing such as the SRSF1 splicing factor binding. GMAS events demonstrate accelerated evolution with shared events occurring between different cell lines. A single GMAS event altering splicing in several tissues supports the involvement of GMAS events in complex diseases.

The final talk of this session was by Reece Hart of Invitae, San Francisco, California, who spoke on "HGVS Nomenclature update." Feedback from the 2015 HGVS annual meeting "Pathogenicity Interpretation in the Age of Precision Medicine" provided suggestions for improvements including updated documentation, additional and improved analytical tools, advocacy for increased data availability, and support for the community. Documentation for the HGVS nomenclature is now available at the Sequence Variant Nomenclature Website (http://varnomen.hgvs.org). In addition to existing tools such as Mutalyzer, SnpEff, and Variant Effect Predictor, several new and updated tools have appeared in the last year. The Variant Validator (http://variantvalidator.org) is a Web interface to the hgvs package. An imminent update to the hgvs package will support GRCh37 and GRCh38, including patch regions. The hgvs-eval project, which was originally proposed at the 2015 HGVS meeting as a way to assess features of HGVS-related tools, was initiated at a recent GA4GH hackathon. HGVS Nomenclature is frequently used both to present variants to human readers and to represent variants in computer systems. Unfortunately, human presentation forms, which are easy to interpret, can create challenges when used by computer algorithms. For example, some insertion/deletion variants can be written in multiple ways, resulting in a shifting of the nucleotide numbers. This can result in multiple names being assigned to the same identical variant. In matching this variant between databases, confusion may occur and important associated data would be lost or improper interpretation occurs. Slides for this presentation can be found at https://goo.gl/9wwVHl.

BRCA1/2 Variants—Sharing and Interpretation

The third session, focusing on the interpretation of variants in the breast cancer risk genes BRCA1 and BRCA2, was chaired by William Oetting of the Department of Experimental and Clinical Pharmacology, University of Minnesota, Minneapolis. The Session was opened by Heidi Rehm of the Harvard Medical School, Boston, Massachusetts, who spoke on "The landscape of BRCA data sharing" including the BRCA Challenge (http://www.genomicsandhealth.org/work-products-demonstration-projects/brca-challenge-0), ClinGen (https://www.clinicalgenome.org), and the Canadian Open Genetics Repository (COGR) (https://opengenetics.ca). The goal of the BRCA Challenge is to improve the care of patients at risk for breast and ovarian cancer by creating a decentralized global public repository (https://brcaexchange.org) bringing together all sources of BRCA variant data. These data assist in the analysis of BRCA1 and BRCA2 variants by collaborating groups including the ENIGMA consortium (https://enigmaconsortium.org). ClinGen, working in partnership with the ClinVar database, collects BRCA1/2 variant data from hundreds of sources and shares that data through ClinVar and the BRCA Exchange for free and unrestricted use by the community. Currently there are 11,460 interpreted BRCA1/2 variants in ClinVar and over 13,500 BRCA1/2 variants in the BRCA Exchange database. The COGR effort is developing consensus interpretations for 5,500 BRCA1/2 variants across 11 clinical laboratories in Canada and will be sharing those in ClinVar and the BRCA Exchange.

The BRCA interpretation session was generously sponsored by BRCA Share™. BRCA Share™ is a major effort to bring together BRCA1/2 variants from multiple sources (https://www.umd.be/BRCA1/ and https://www.umd.be/BRCA2/) and has support from LabCorp and Quest, as well as many partners. Christophe Béroud of the Aix Marseille Université, France, spoke on the "Origin of BRCA Share™". The goal of this new model of LSDB public/private partnership is to create an open user group and BRCA-Share™ databases with high-quality information from both academic institutions and private companies, including interpretation of variants of unknown significance (VUS) and associated evidences. These databases offer free access to academic researchers and licensed access for clinical diagnostic laboratories. The funds from licensed access pay for curation of the variants. The site was opened 1 year ago and currently has over 80,000 BRCA1/2 gene sequence tests, over 1,300 registered users from 71 countries and received more than 250,000 queries. They are currently developing a prioritization system for reclassification of VUS and will support functional studies, which could bring sufficient evidence for VUS reclassification.

Determining if a BRCA1/2 variant is functional is critical to both the clinician and the individual with the variant. A variant classified as a VUS makes for uncertainty when clinical decisions need to be made. Nicholas Woods of the Eppley Institute, University of Nebraska Medical Center, Omaha, spoke on this in his presentation "Functional Assays as a Tool for Clinical Annotation of BRCA1 VUS." Many variants have been identified in BRCA1, but not all are pathogenic. There are greater than 23 functional assays to assess BRCA1 VUS but many of these lack statistical validation and have poor sensitivity and specificity. A new method was created that is free from the limitations of genetic and epidemiology methods used to classify variants. This is a transcriptional test analyzing the BRCA1 transcriptional transactivation activity of the c-terminus (BRCT). The BRCT coding region, with variants inserted through site-directed mutagenesis, is fused to the DNA binding domain of GAL4. In the presence of a functional BRCT, the luciferase gene, with a GAL4 binding site in the promoter, is expressed. If the BRCA1 protein has a functional mutation, there is no activation of luciferase expression due to the lack of BRCT function. Over 260 mutations have been tested. Using the VarCall Probability model, functionality is predicted on a scale of 1—5, with 5 being pathogenic. VarCall worked better than other methods including SIFT,

PolyPhen-2, CADD, and Mutation Taster. This system will also work with multiple variants that are in cis. As an example, the mutations p.C1787S and p.G1788D are neutral individually but together in cis are pathogenic. This assay can be used with other proteins containing the BRCT domain such as microcephalin (MCPH1) and mediator of DNA damage checkpoint protein 1 (MDC1).

A second strategy for determining the functionality of BRCA1 mutations was presented by Lea Starita of the University of Washington, Seattle, in her presentation "Massively parallel functional analysis of missense mutations in BRCA1 for interpreting variants of uncertain significance." Increased testing of individuals at risk for breast cancer has resulted in many more VUSs in both BRCA1 and BRCA2 as well as in other genes associated with cancer. In an effort to predict the functionality of these variants, in silico methods have been used. Though computation prediction algorithms have high throughput, they do not have the accuracy needed for clinical decisions. Alternatively, testing each VUS one at time is very time consuming. A multiplex functional assay for measuring variant effect would be much more efficient. BRCA1 is required for double strand break repair. For this function, BRCA1 dimerizes with BARD1 at the RING domain resulting in E3 ligase activity. By assaying for these functions of the RING domain, the impact of variants can be ascertained. Using two assays, all mutation in the RING domain (approximately 100 amino acids) were created and the tested for E3 ligase activity and BARD1 binding in parallel. Results from these assays are more accurate at predicting damage to the full length protein than many commonly used variant-effect prediction algorithms.

Computer Models for the Prediction of Regulatory Regions and Functional Non-Coding Variants

The last session was chaired by Rachel Karchin of the Department of Biomedical Engineering and Oncology, Johns Hopkins University, Baltimore, Maryland. The first talk was by Ekta Khurana of the Weill Cornell Medical College, New York, New York, who spoke on "Computational method to identify non-coding cancer drivers." The switch from whole exome sequencing to whole genome sequencing is presenting a whole new class of variants that need to be interpreted. Most variants (98%–99%) identified are in non-coding regions. One mode of action for these non-coding variants is the alteration of TF binding sites. For example, alteration of the TERT promoter is common in many different cancer types and upregulation of TERT expression promotes tumorigenesis. The ability to identify such non-coding variants that drive cancer would greatly aid in identifying important functional somatic variants in tumors. CompositeDriver was presented as a method for classifying regions, such as promoters, enhancers and ncRNAs, which may contain non-coding variants as cancer drivers. This technique combines information about the functional impact of variants and their recurrence across multiple tumor samples. It accounts for mutational heterogeneity observed in somatic tissues, for example TF binding sites show increased mutation rates in melanoma and lung cancer due to impaired access of DNA repair complexes (PMID: 27075092, 27075101). The functional impact of variants is computed using the FunSeq scheme (http://funseq.gersteinlab.org). Functional studies validated some results of these two methods.

The second presentation of this session was by Ross Hardison of the Departments of Biochemistry and Molecular Biology and of Statistics, from Pennsylvania State University, University Park, who spoke on "Integrative analysis of epigenomes reveals roles of non-coding regions and variants in differentiation and diseases of blood cells." Hematopoiesis is a very active process, producing 2 million erythrocytes per second. The decisions for cellular differentiation in this process are determined by both TFs binding to cis-regulatory regions and epigenetic signals. Understanding the integration of epigenetic influences is important to understand this process. Assigning an epigenetic state to segments of DNA allows for the creation of an epigenome map to identify important non-coding regulatory regions. The method, Integrative and Discriminative Epigenome Annotation System (IDEAS), was presented as a way to identify these regions which have epigenetic signals (https://sites.stat.psu.edu/~yuzhang/IDEAS). An example for the use of this information is in sickle cell disease (SCD). The SCD variant in the beta-globin polypeptide produces the HbS form of the hemoglobin tetramer, which polymerizes under low oxygen conditions. The presence of gamma-globin, producing the HbF form of hemoglobin, ameliorates this polymerization. The presence of HbF in adults is well tolerated as exhibited by individuals with hereditary persistent fetal hemoglobin (HPFH) where the normally fetal produced HbF continues to be produced in adults. By knowing the location of critical regulatory regions, targeted changes in the non-coding regulatory regions can reactivate expression of the gamma-globin gene, resulting in an increase in HbF and a reduction in the polymerization of hemoglobin and pathogenicity of SCD. As shown in this example, identification of critical regulatory regions will help us better understand the function of non-coding variation.

The final talk of this session and of the meeting was by Olga Troyanskaya of the Lewis Sigler Institute of Integrative Genomics, Princeton University, New Jersey, who spoke on "Predicting the effects of non-coding variants with deep learning-based sequence model." As stated previously, most GWAS hits are in non-coding regions. Functional non-coding variants are most likely to be found in regions that regulate gene expression. These regions can be identified as TF binding sites, DNase hypersensitive sites, and regions of histone (epigenetic) modification. A model was presented that predicts functional non-coding variants using this information. This included flanking sequences (1,000 bps) exhibiting cooperativity between TF binding sites and epigenetic markers. An in silico method, termed DeepSEA (http://deepsea.princeton.edu), utilizes multi-task deep learning that leverages genomic context sequence information to predict important regulatory regions and the functionality of SNPs and their effects on chromatin. Accuracy of the models was also demonstrated by accurately predicting external experimental allele imbalance DNASeq data and histone mark QTLs. The model also far outperforms prior methods for predicting whether a SNP is disease-associated. A case study of alpha thalassemia was used to show how the method can infer disease associated variants and their mechanism for altering gene expression.

Acknowledgments

The Scientific Program Committee for the HGVS Annual Meeting consisted of Christophe Béroud, Marseille University, France; Steven Brenner, University of California-Berkeley; Marc Greenblatt, University of Vermont College of Medicine; Rachel Karchin, Johns Hopkins University; William Oetting, University of Minnesota; Sean Mooney, University of Washington; and Shamir Sunyaev, Harvard University. The Scientific Program Committee would like to thank Rania Horaitis for her professional help in running this HGVS annual scientific meeting. This meeting of the HGVS was chaired by Steven Brenner, Bruce Gottlieb, William Oetting, and Rachel Karchin. The authors would like to thank the speakers for their help in the preparation of this report. SEB's participation was supported by U41 HG007346.

Disclosure statement: The authors declare no conflict of interest.