MEETING REPORT

Human Mutation

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

Clinical Interpretation of Variants from Next-Generation Sequencing: The 2016 Scientific Meeting of the Human Genome Variation Society

William S. Oetting,[1]* Anthony J. Brookes,[2] Christophe Béroud,[3] and Peter E. Taschner[4,5]

[1]Department of Experimental and Clinical Pharmacology, University of Minnesota, Minneapolis, Minnesota; [2]Department of Genetics, University of Leicester, Leicester, United Kingdom; [3]Aix Marseille Universite, Marseille, France; [4]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; [5]Generade Centre of Expertise Genomics and University of Applied Sciences, Leiden, The Netherlands

Communicated by Mark H. Paalman

Received 8 July 2016; accepted revised manuscript 23 July 2016.
Published online 5 August 2016 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.23059

Hum Mutat 37:1110–1113, 2016.

Introduction

The 2016 scientific meeting of the Human Genome Variation Society (HGVS; http://www.hgvs.org) was held on the 20th of May in Barcelona, Spain, with the theme of "Clinical Interpretation of Variants from Next-Generation Sequencing." The meeting was opened by William S. Oetting, of the University of Minnesota, United States. "Precision medicine" is the latest buzz words in healthcare, both in the literature and in government initiatives. Pharmacogenomics is one area where next-generation sequencing (NGS) will have an impact, but there are some issues that need to be addressed. Currently, the use of genomics in determining medication dosing has focused on high frequency variants with known functional consequences but these do not predict all of the genetic influences on subsequent drug levels, efficacy, or toxicity. The use of NGS presents clinicians with additional rare variants in candidate genes to consider but with unknown functionality. Whole genome sequencing (WGS) identifies over 3.5 million variants, many of which need to be selected and accurately interpreted for this information to be of use to the clinician for proper drug selection or dosing. This will be true for all clinical uses of NGS. This meeting explored multiple aspects of the clinical use of NGS, including whole exome and genome studies in inherited disorders, genomic analysis of somatic tissues, and issues of interpretation and standards of genetic variation for clinical use. The problems and possible improvements for the utilization of NGS in clinical care were presented in talks of this scientific meeting and will help clinicians use these results for the improvement of healthcare.

Whole Exome and WGS

The first session, Whole Exome and Whole Genome Sequencing, was chaired by Peter E. Taschner of the Leiden University Medical Center, The Netherlands. For the first talk, Tim Hubbard, of Genomics England, United Kingdom, spoke on "The 100,000 Genomes Project." The goal of the 100,000 genome project (http://www.genomicsengland.co.uk) is to provide help to individuals with unmet clinical needs by using WGS to determine the cause of their clinical symptom and identify possible treatments. The initial dataset for this project will include sequencing 100,000 individuals with undiagnosed diseases. The project will include over 80 hospitals as a source for samples for WGS and phenotypic information. The goal is to create a pipeline that results in a clinical report, based on WGS data, which is useable by the clinician for diagnosis of the primary condition. A few secondary uses will be allowed but must be consented for (i.e., cancer predisposition). Patients can be re-contacted if additional information or samples are required. It is expected that data models will need to be created for every genetic condition using disease specific phenotypic information. A standardized method of reporting information back to the hospitals and clinicians will be created. Data movement and analysis is done only within the system to ensure confidentiality for patients (data cannot move out of the datacenter), but robust research by investigators is encouraged. Researchers will have access to all health information of patients, but data will be held within the system and cannot be downloaded unless approved and in an anonymous form. Additionally, all datasets for publication will be reviewed. Funds for research projects are available.

The second speaker was Martin G. Elferink of University Medical Centre Utrecht (UMCU), The Netherlands, who spoke on "Comparing WGS to WES in a clinical setting." This study was part of the Genetics Clinic of the Future project within the UMCU (based on http://www.geneticsclinicofthefuture.eu). Currently most diagnostic sequencing at the UMCU is done using either a panel of selected genes or whole exome sequencing (WES). The use of WGS would result in a single test for all genomic analysis for single nucleotide variants (SNV), insertions and deletions (INDELs), copy number variation (CNV), and structural variation (SV), reducing the complexity inherent in multiple tests. To show the utility of WGS in a diagnostic setting, a comparison was made between the quality of SNV and INDELs in WES and WGS results. A genome in a bottle reference (GIAB) sample (NA12878) was used for comparison purposes. Precision and sensitivity were calculated using RTGtools (v3.6.1). For this analysis, only the GIAB high-confident regions of the coding exome were used for comparison. It was found that the sensitivity for SNVs was higher in WGS (99.5%) resulting in fewer false negatives compared with WES (99.1%) where most false negatives (~85%) in WGS were incorrectly called heterozygous instead of homozygous. Sensitivity for INDELs was also higher in WGS (97.0%) compared with WES (95.4%). Conversely, precision was greater in WES (99.2%) than WGS (98.4%) for non-filtered

*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

SNVs. Similar results were found for non-filtered INDELs: 91.4% for WES and 84.6% for WGS. For laboratories using diagnostic panels with WES, it was found that the quality of 40× WGS had a higher sensitivity then >>100× WES and that WGS may be the best approach for genetic analysis in a clinical setting. The gain of additive sequencing for WGS (increase to 80× or 120×) provided limited additional increase in the quality of the data.

The third talk in this session was by Lidia Feliubadaló of the Catalan Institute of Oncology (ICO-IDIBELL), Spain, who presented her talk entitled "Genetic testing for hereditary cancer: is exome sequencing ready or is there still room for ad hoc designed panels?" Until recently, Sanger sequencing of a panel of cancer predisposing genes was used to identify at risk patients and families. NGS allows the analysis of a panel of genes associated with cancer risk or even the entire exome (WES). In this study, a custom and a commercial hereditary gene panel were compared with WES for diagnostics purposes. In the control set, the three approaches found all pathogenic variants except for one in TruSight Cancer. Additional putatively pathogenic variants were found by all approaches in the discovery set. WES identified many more variants outside the genes covered by the ad hoc panels. The ad hoc designed panels had lower costs (about one-third the cost) and provided higher depth of coverage, reaching the minimum C30 coverage threshold for diagnostics in more than 99% of the regions of interest, whereas the exome covered just 94%. An additional consideration is that WES has greater ethical issues associated with the additional sequencing information created compared with focused candidate genes due to the identification of incidental findings.

Variant Calling and Nomenclature

The second session, Variant Calling and Nomenclature, was chaired by Anthony J. Brookes of the University of Leicester, United Kingdom. The first talk was by Ivo G. Gut, of the Centro Nacional de Análisis Genómico, Barcelona, Spain, with his provocative title "Are we done with variant calling?" When genetic information is used for important clinical decisions, the data (phenotypic and sequencing) and the analysis need to be of high quality. To ascertain the sequencing quality produced by diagnostic laboratories, a comparison was made between five large sequencing centers, using a set of samples including both tumor and associated normal genomic samples. It was found that the centers provided different qualities of data. Some of these differences resulted in problems with alignment and indel calling. To compare variant calling, a high quality set of sequencing data containing 1,200 somatic variants was provided to 20 different analytical teams. Comparing the results, all teams agreed on only 170 somatic variant calls. For somatic insertion/deletion variants everyone agreed on only one variant (16 submissions). Some teams reported many wrong calls and others provided fewer calls, but of higher quality. Reasons for poor calls included many false positives near centromeres (problems with repetitive areas). Some teams exhibited regions of poor calling with other regions having high quality calls where other teams that had poor quality calling throughout the sequence. It is difficult to compare calling algorithms to determine why wrong calls were being made because teams used "black boxes" for their analysis. To overcome these problems and improve the quality of data analysis, there needs to be a method of certification of sequencing laboratories, including datasets, for comparison. A framework for quality assessment of whole genome cancer sequences is being worked on by the PanCancer Analysis of Whole Genomes (PCAWG) project (dcc.icgc.org/pcawg). The sequencing of tumor samples is also highly variable, in part caused by the quality of the sample and amount of ploidy. It is not intended that researchers exclude the lower ranked cancer genomes, but to be wary of any conclusion based solely on poorer quality cancer samples. Some means of scoring the quality of tumor samples, such as a star rating system, is needed.

The next talk was by Jonathan K. Vis from the Leiden University Medical Center, The Netherlands, who spoke on "Towards formal specification of HGVS nomenclature enabling computational tool development." The HGVS nomenclature has provided valuable guidelines for the description of genetic variants, but there are still problems with the syntax and semantics of the HGVS language. Solving this is especially important in the structured environment of databases and for the design and implementation of computer algorithms for the analysis and comparison of sequence variants. At present, changes in the HGVS nomenclature occur as problems arise. One example is when variants created by two different events are next to each other. The HGVS nomenclature often requires replacement of the two original descriptions, although the two events may have occurred at different times. Separate descriptions of variants arising from these events may be important in understanding how each of these variants impacts the phenotype. The wish to have this reflected in the nomenclature specifications may introduce ambiguity. There is a need to make the HGVS nomenclature "future proof" by moving toward a set of formal specifications. Some steps include: The introduction of some concrete core principles, a strict set of leading rules (without discussion), the ability to distinguish between descriptions and annotations, acknowledgement and extension of the HGVS grammar in Extended Backus-Naur Form (EBNF), the addition of semantic rules and making sure that any new versions are consistent with the existing rules for backward compatibility.

The third talk was by Sohela Shah, from Qiagen, Redwood City, California, United States, who spoke on "An efficient and accurate end-to-end solution leveraging network analytics to infer patient syndrome and identify causal mutations in rare disease cases." A major reason for DNA sequencing is to identify the underlying genetic cause of disease but the identification of the causal variant can be challenging and time consuming. This is aggravated by the fact that 27% of variants cited in literature are either benign polymorphisms or are mis-annotated. A solution, termed the Hereditary Disease Solution, was presented that incorporates clinical and family history, a manually curated biological knowledge database of genes and functional disease causing variants, pathway analysis and tools that allows all of this to be brought together to identify the causal variant. The pipeline is a one-step variant calling to interpretation workflow that makes sure no variants are missed by checking for variants identified in one family member in the mapped sequencing reads of other family members. It incorporates extensive annotation from both public and private databases for candidate variants, literature annotation with supporting evidence and has a future plan for including functional studies of variants. In practice this pipeline has identified causal variants accurately 65% of the time and shortens the list of candidate variants to a minimum.

The final talk in this session was by Stephen E. Lincoln from Invitae, San Francisco, United States, who spoke on "What do public databases of clinical variants really tell us about classification concordance?" As previously stated, the literature is filled with both valid and invalid classifications of supposedly disease producing variants. Because of this variable quality, frequent disagreements are observed between papers and also between certain databases. Thus, clinical lab directors must critically evaluate all evidence of pathogenicity that they utilize in order to make sure that the variants they report are properly classified. This is a time-consuming

task requiring expertise in diverse scientific disciplines and is difficult, if not impossible to automate. Curated databases, such as many locus specific databases (LSDBs) on the other hand, are potentially more accurate and more consistent as a degree of expert review has already been performed. Working with collaborators at the University of California, they compared classifications of *BRCA1* and *BRCA2* variants submitted by major diagnostic laboratories to ClinVar, where most of these data were known to have received expert review consistent with the ACMG guidelines for variant classification. Concordance on a per-patient basis was high (99.8%) in terms of impact on clinical care. Far lower concordance was previously observed by other authors who had published a more naïve comparison of public databases in an apparent effort to discredit the public sharing of genetic data. It was concluded that curated public databases, such as ClinVar and certain LSDBs, when carefully used by experienced lab directors, can be an important asset. Moreover, they allow for inter-laboratory comparisons that improve test quality and patient care globally. Curating evidence for a large number of variants is a substantial amount of work, often poorly funded, and "crowd-sourced" initiatives such as ClinVar may represent a scalable solution.

Variant Annotation and Interpretation

The third session, Variant Annotation and Interpretation, was chaired by Christophe Béroud of the Aix Marseille Université, France. The first talk was by David Salgado, also from Aix Marseille Université, who spoke on "Variant annotation and filtration in NGS context." Modern high-throughput sequencing techniques facilitate the identification of new disease genes and disease causing variants but in the case of WES, the success rate in finding causative variants remains relatively low (between 23% and 26%). This low success rate is linked to various challenges, such as technical factors, the type of disease causing variant, the bioinformatics suite of tools and methods used to generate VCF files, or incorrect variant annotations and bad filtration practices. Variant annotations are used to distinguish "real" variants from sequencing artifacts and those that are potentially pathogenic from neutral and there are several systems that are suitable to gather such annotations at various granularity levels (variant, gene, and phenotypic levels) automatically from a VCF file. Automatic and manual filtration systems that are commonly used to highlight disease causing variants and the main advantages and drawbacks of these systems and processes were presented. Additionally, three systems were presented that are being developed to predict altered splicing sites; the UMD-Predictor system (http://umd-predictor.eu) to predict the pathogenicity of any human cDNA substitution; the Human Splicing Finder (http://umd.be/HSF3/), which is a reference system to identify variants that could impact splicing machinery; and the VarAFT system (http://varaft.eu) to facilitate annotation and filtration of variation from high-throughput sequencing technologies.

The second talk in this session was by Adam Frankish of the Wellcome Trust Sanger Institute, Cambridge, United Kingdom, who spoke on "Improving the annotation of clinically important genes to aid identification of missing causal variants." Identification of disease producing variants will require a complete set of high quality annotated genes. This includes the multiple alternatively spliced transcripts that are associated with most genes. This is a goal of the HAVANA team which produces the GENCODE reference gene set (http://www.gencodegenes.org). Early infantile epileptic encephalopathies (EIEE), associated with early onset seizures which occurs 3–5 per 10,000 live births, was used as a test case. In this study, 70 genes used on reference diagnostic panels for EIEE were re-annotated focusing on alternatively spliced transcripts including exon skipping and inclusion. A significant increase in exonic coverage was observed. The total number of novel transcripts found was 1,092 with 706 new exons, 1,132 new introns and 224 shifted splice junctions (SSJs) that extend or truncate existing exons. An important question is: Are these novel transcripts relevant to disease? This can be answered in part by determining if the transcript is expressed in the appropriate tissue and at what level, are proteins found associated with these novel transcripts and if there is conservation of the added sequence in other organisms. In this case, many of these transcripts were found to be expressed in the brain and approximately one-third of the novel CDS sequence conserved in other mammals. By expanding the number of novel transcripts for each candidate gene, additional regions for coding sequences can be analyzed for potential causative variants.

The third talk in this session was by Vijaya Ramachandran of South West Thames Regional Genetics Laboratory at St. George's University Hospital, London, United Kingdom, who spoke on "Sitting on the fence: variant interpretation in RASopathies." The RAS/MAPK pathway is essential for the proper regulation of the cell cycle and critical for normal development. Dysregulation of the RAS/MAPK pathway results in a number of clinically overlapping genetic disorders that include craniofacial and cardiovascular disorders such as Noonan syndrome and Costello Syndrome. Variants affecting this pathway have been identified in multiple genes including the novel genes *RIT1*, *RRAS*, *RASA2*, and *A2ML1*. To identify disease associated variants in patients that may be affected by alterations in any of the RAS/MAPK pathway, a 23 gene panel using Ion Torrent PGM Sequencer was established as a diagnostic service and tested in 243 patients. The minimum panel coverage was 97.43%. Variants were identified in 125 cases (51%). However, interpreting these variants in a clinical setting is challenging.

Ethics and Characterizing Disease with NGS

The fourth session, Ethics and Characterizing Disease with Next Generation Sequencing, was chaired by William Oetting. This session was opened by Heidi Carmen Howard, from Uppsala University, Sweden, who spoke on "Ethical issues of NGS in the clinic." To patients, genetics and genetic results can be a foreign language. Confusion or misunderstandings associated with genetic information can exacerbate ethical issues, and it can be conceptualized by ideas captured in "U3S" for different stakeholders; unexpected, unknown, and uncertain. Uncertainty exists for patients as they try to understand the probability or risk of developing a disease, the actions they should (not) take, or the meaning of results for them and their family. Uncertainties are also present for "experts" when trying to understand variants of unknown significance (VUS), incidental findings, incomplete penetrance, and phenotypic variability. These uncertainties can lead to misunderstanding of results for the patient and possibly incorrect treatments presented by the clinician. A general approach to help patients further understand complex information in genomics is encompassed through the notion of "SEED" (Stakeholder Engagement, Education, and Dialogue). Stakeholders include not just the patients but all of the health practitioners involved in medical care and experts in genetics or genomics. Sustained engagement between the patients and clinicians is needed, especially for genomic medicine and continual education is needed for all. For society, this may require more genomics in high school. Lastly, dialogue is needed to make sure that all individuals, especially the patients, understand the meaning of their results. Genetics

also produces novel ethical issues which have been amplified with the advent of NGS. Issues include the challenges to obtaining informed consent and supporting patient autonomy, as well as which or how to return (secondary or unsolicited) results when so many unknowns exist (VUS, actionability of results, and so on). These issues are made more complex in that the whole family can be impacted by genetic based results, understanding risks and benefits of this information, storage of phenotypic and sequencing data, incidental findings, opportunistic screening, and secondary use of data. All of this makes moving genomics into the clinic a challenge to existing ethical frameworks and norms that will only become more complex as clinical medicine moves from candidate gene analysis to WES to WGS. The time is right now to bridge "U3S" with "SEEDS."

The second talk of this session was by Celeste Bento of the Centro Hospitalar e Universtário de Coimbra, Portugal, who spoke on "Congenital hemolytic anemia study with a targeted NGS panel." Congenital hemolytic anemias (CHA) are a genetically heterogeneous disorder identified by an increased destruction of the red blood cells. Examples include sickle cell disease, G6PD deficiency, and hereditary spherocytosis. Many times these diseases can be diagnosed using a blood smear complemented with common screening tests for RBC disorders, but not always. In this report, NGS was used to identify the molecular causes in a group of 14 patients with unknown CHA. A panel of 26 genes including hemoglobins, membrane proteins, and enzymes were sequenced using NGS. Many variants were identified with each patient sample and for the 14 samples analyzed, eight had variants that were classified as definitively pathogenic and five samples had only benign variants. In some patients with pathogenic variants in multiple potential causative genes, their phenotype was more severe. This study showed that NGS of a panel of genes, along with in silico analysis, is a cost efficient means of identifying disease producing variants in individuals with CHA, but there are a significant number of affected individuals where causative variants are not identified. Use of WGS may be needed to identify variants in these individuals.

The final talk of this session and of the meeting was by Oscar Marin Sala of the Vall d'Hebron Institute of Research, Barcelona, Spain, entitled "Characterizing the intrinsic component of disease severity." Bioinformatics play an important role from the analysis of the genome to diagnosis. There is a need to improve the tools to characterize the effects of variants, especially missense variants. Machine learning techniques and the use of neural networks should help. Problems in variant calling include ambiguous annotation of variants in databases and the biological fact that different proteins have varying sensitivities to amino acid substitutions making universal predictions for a given substitution difficult. To study this, variants in coagulation factors VIII and IX were used to build a classification pipeline. Sequence structure, entropy, and hydrophobicity predictions were used to build the prediction model to classify variants as either mild or severe. The prediction pipeline worked well but it is still not good enough for clinical practice. Using this type of analysis may help to create a usable model. It was concluded that this is an intrinsic component in severity that allows its prediction using molecular and evolutionary properties.

Acknowledgments

The Scientific Program Committee would like to thank Rania Horaitis for her professional help in running this HGVS scientific meeting. This meeting of the HGVS was chaired by Peter Taschner, Anthony Brookes, Christophe Béroud, and William Oetting. The authors would like to thank the speakers for their help in the preparation of this report. This meeting was run in partnership with The Human Variome Project (http://www.humanvariomeproject.org).

Disclosure statement: The authors declare no conflict of interest.