Curating the Clinical Genome Meeting 2019

Abstracts Selected for Platform Presentation

Using semantic similarity analysis based on Human Phenotype Ontology terms to identify genetic etiologies for epilepsy and neurodevelopmental disorders

Ingo Helbig[1,2,3], Shiva Ganesan[1,2], Peter Galer[1,2], Colin A. Ellis[3], Katherine L. Helbig[1,2]

[1] Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA, 19104 USA
[2] Department of Biomedical and Health Informatics (DBHi), Children's Hospital of Philadelphia, Philadelphia, PA, 19104 USA
[3] Department of Neurology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, 19104 USA

More than 25% of children with intractable epilepsies have an identifiable genetic cause, and the goal of epilepsy precision medicine is to stratify patients based on genetic findings to improve diagnosis and treatment. Large-scale genetic studies in > 15,000 epilepsy patients have provided deep insights into causative genetic changes. However, the understanding of how these changes link to specific clinical features is lagging behind. This deficit is largely due to the lack of frameworks to analyze large-scale phenotypic data, which dramatically impedes the ability to translate genetic findings into improved treatment.

The Human Phenotype Ontology (HPO) has been developed as a curated terminology, with formal semantic relationships between more than 13,500 phenotypic terms, and has been adopted by many diagnostic laboratories, the UK's 100,000 Genomes Project, the NIH Undiagnosed Diseases Network, the GA4GH, and hundreds of global initiatives. The HPO also contains a rich vocabulary for epilepsy-related phenotypes, which we have developed in prior large research networks and expanded in current projects. However, analytical approaches leveraging this rich data resource have not been implemented so far. Given that reliable clinical data is pivotal for diagnosis and clinical care, there is a critical need to assess the ability of computable HPO-based epilepsy phenotypes to delineate subtypes of genetic epilepsies and define previously-uncharacterized disease entities that may reveal therapeutic strategies through novel biological mechanisms.

Within the current project, we used manually curated HPO data to assess the phenotypic range and specificity for known genetic epilepsies and neurodevelopmental disorders. We find that for most common genetic epilepsies, including genetic epilepsies due to variants in *SCN1A* and *STXBP1*, phenotypes based on HPO are significantly more similar than expected by chance, using algorithms using semantic similarity algorithms based on the information content (IC) of the most informative common ancestor (MICA). We also demonstrate that several genes for epilepsy and neurodevelopmental disorders are linked by phenotype rather than genotype and that semantic similarity can be used to provide additional statistical evidence for candidate gene-disease relationships that may be rare in the patient population and may have insufficient statistical evidence based on genetic findings alone.

Is Likely Pathogenic Really 90% Likely? A Look at the Data

Steven M. Harrison and Heidi L. Rehm
Broad Institute of MIT and Harvard; Laboratory for Molecular Medicine, Partners Healthcare, Cambridge, MA

The 2015 ACMG/AMP guideline for variant interpretation proposed that the term likely pathogenic (LP) be used to mean greater than 90% certainty of a variant being pathogenic. However, no study has analyzed the outcomes of LP variants to determine if the 90% certainty threshold is valid. Understanding LP classification confidence is necessary as many clinicians treat LP and P classifications equally, meaning reclassifications to VUS/LB/B are likely unanticipated. We sought to determine the true certainty threshold by calculating the reclassification rates of variants submitted to ClinVar, choosing variants assessed after Jan 2016 in hopes of restricting to variants classified with the 2015 ACMG/AMP guideline.

Between Jan 2016 and Jan 2019, there were 484,916 classifications submitted to ClinVar. By Jan 2019, 4,090 of these classifications had been reclassified, of which 94% moved to a classification category of more certainty and only 6% moved to a less certain (5.7%) or opposing (0.3%) category.

Of the 32,617 LP classifications in ClinVar, 569 were reclassified with 5 reclassified to LB/B, 99 reclassified to VUS, and 465 reclassified to P. If only including LP reclassifications to a more definitive category (LP to P or LP to LB/B), LP reclassification rates suggest a 99% (465/470) certainty of being P compared to B. However, 0.3% (99/32,617) of LPs dropped to VUS suggesting that some may eventually move to LB/B and the rate may be lower than 99%.

Variants were further interrogated to determine if certain variant types or genes were more likely to upgraded (LP to P) or downgraded (LP to VUS/LB/B). We found that 48% (222) of LP variants reclassified to P were LOFs and 52% (243) were missense or intronic. In contrast, 18% (18) of LP variants downgraded to VUS/LB/B were LOFs versus 82% (81) were missense or intronic. When comparing variants in ACMG59 cancer and cardio genes, 91% (138/151) of LP reclassified cancer variants were upgraded while only 76% (68/89) of LP reclassified cardio variants were upgraded.

In summary, current reclassification data from ClinVar shows that 99% of LP classifications move to P compared to LB/B, suggesting that application of the term is consistent with the intended confidence level. However, the vast majority of LPs still remain as LP within a three year window and a small subset (0.3%) dropped to VUS suggesting that more data and a longer time of analysis will be needed to more robustly evaluate the rate of LP reclassification.

Establishing an intralaboratory copy number variant conflict resolution process using ClinGen Dosage Sensitivity Map scores and internal classification discrepancies to improve retrospective data analyses and submissions to ClinVar

Authors: Zoe K. Lewis[1], Tim Tidwell[1], Adam Clayton[1], Julie L. Cox[1,2], Bo Hong[1,2], Allen N. Lamb[1,2], Denise I. Quigley[1,2], Roger Schultz[1,2], Reha Toydemir[1,2], Xinjie Xu[1,22], Erica F. Andersen[1,2]

[1]ARUP Laboratories, Salt Lake City, UT, USA
[2]University of Utah, Salt Lake City, UT, USA

Copy number variant (CNV) classification discrepancies are common in clinical laboratories and public databases such as ClinVar. Conflicts must be resolved to optimize patient care and maximize clinical utility of these databases. CNVs may be classified differently between and within laboratories for several reasons, including emerging evidence, evolving guidelines and processes for consistent classification of CNVs, as well as transcription errors. We present results from an internal process for identifying and resolving clinical CNV classification conflicts in our laboratory. A total of 3962 CNVs reported over a 6-year period were selected for review. Similar to a recent inter-laboratory CNV conflicts resolution process (Riggs et al. 2018, PMID: 30095202) we identified CNVs that overlap genes and regions with established haploinsufficiency/triplosensitivity based on curations from the ClinGen Dosage Sensitivity Map (DS Score=3, DS3), which had been classified as variant of uncertain significance (VUS) or likely benign (LB). We also identified multiply-encountered CNVs with ≥99% overlap and 99% size similarity (close-match) and discordant classifications. In total, 41 CNVs overlapping DS3 genes or regions previously classified as VUS/LB and 44 close-match CNVs (244 cases) were flagged for review. Of the 41 VUS/LB CNVs overlapping DS3 genes, 10 were flagged for reclassification to pathogenic (P) or likely pathogenic (LP) based on new evidence in the literature, 5 were flagged for reclassification from VUS to P due to X-linked carrier status in a female, and 14 were flagged due to transcriptional errors in classification selection not involving the clinical report. Of the 44 close-match CNVs with discordant classifications, 23 involved genes that confer a risk for autosomal recessive disease, underscoring the importance of standardized classification methods for these variants. Any LP/P reclassification resulted in amended clinical laboratory reports. Our findings illustrate the utility of accessing up-to-date information from the literature and ClinGen curations to guide patient care. We will discuss a few interesting cases involving these reclassified CNVs. This process will be used for prospective evaluation of clinical CNVs encountered in our laboratory. We aim to improve quality through consistency and standardization, as well as to ensure that valid, updated information enters the public space in ClinVar.

Constructing a Quantitative Metric for Evaluating the Clinical Significance of Recurrent Copy Number Variants

John Herriges[1*], Erica Andersen[2*], Bradley Coe[3], Laura Conlin[4], McKinsey Goodenberger[5], Benjamin Hilton[2], Vaidehi Jobanputra[6,7], Hutton Kearney[5], Brynn Levy[7], Prabakaran Paulraj[2], Erin Rooney Riggs[8], Ross Rowsey[5], Marsha Speevak[9], Romela Pasion[5], Cassie Runke[5], Justin Schleede[10], Kelly Toner[8], Erik Thorland[5] and Christa Lese Martin[8], on behalf of the Clinical Genome Resource (ClinGen)

1Children's Mercy Hospital, Kansas City, MO;
2ARUP Laboratories, University of Utah, Salt Lake City, UT;
3University of Washington, Seattle, WA;
4Children's Hospital of Philadelphia, Philadelphia, PA;
5Mayo Clinic, Rochester, MN;
6New York Genome Center, New York, NY;
7Department of Pathology & Cell Biology, Columbia University Medical Center
8Geisinger, Danville, PA;
9Trillium Health Partners, University of Toronto, Toronto, ON;
10Laboratory Corporation of America Holdings®, Research Triangle Park, NC;
(*Equal contribution)

Recurrent copy number variants (CNVs) represent one of the most common genomic changes detected by copy number analysis. The ClinGen Dosage Sensitivity Curation Working Group has historically used an established evidenced-based review process to evaluate the significance of recurrent CNVs, assigning a numerical score corresponding to strength of evidence associated with deletions and duplications of these regions (Riggs et al. 2012). While this process has been effective for evaluating recurrent CNVs associated with highly penetrant and specific/well-characterized clinical findings, such as the recurrent 7q11.23 deletion associated with Williams-Beuren syndrome, using this process to evaluate the significance of deletions and duplications for all recurrent CNVs has been challenging. For example, the 15q11.2 (BP1-BP2) and 16p12.2 regions exhibit incomplete penetrance and variable expressivity; and for many regions, segregation data, along with statistics for population-level frequency, are difficult to interpret clinically without well-defined syndromic features. In order to effectively and consistently assess evidence associated with recurrent CNVs, our group has developed revised criteria for dosage sensitivity scoring of these regions. Similar to other ClinGen curation efforts, we have developed a points-based scoring metric that incorporates phenotypic specificity and variability, segregation/inheritance data, and population-level statistics in order to objectively evaluate and appropriately weight information across different evidence categories. Using a test cohort of 12 previously scored recurrent CNV regions (24 total deletions and duplications), we evaluated the usability and accuracy of this scoring metric through review by two independent reviewers for each region. We observed concordance rates of 71% between the new point-based scoring metric and the historical scoring process, and 92% concordance rates between the independent reviewers using the new scoring metric. We will present results from this study as well as further testing and refinement of the metric. This work aims to assist in the clinical interpretation and reporting of recurrent CNVs, as well as provide better alignment with other quantitative ClinGen curation efforts.

References:

Riggs ER, et al. Towards an evidence-based process for the clinical interpretation of copy number variation. Clin Genet. 2012 May;81(5):403-12.

Genomics as a personalized medicine approach in disease risk prediction - P5.fi FinHealth

Heidi Marjonen[1],Minttu Marttila[1], Teemu Paajanen[1], Ari Haukkala[2], Helena Kääriäinen[1], Kati Kristiansson[1], Markus Perola[1,3]

[1]National Institute for Health and Welfare, Helsinki, Finland
[2]Faculty of Social Sciences, University of Helsinki, Finland
[3]Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Finland

In P5.fi study we utilize polygenic risk scores to provide personalized information on the individual disease risk related to three common diseases (coronary heart disease, type 2 diabetes and venous thromboembolism) for 3.400 volunteering participants. We study the value of returning genetic risk information and hypothesize that it would improve prevention, diagnosis and treatment.

We used polygenic risk scores containing up to 7 million genomic regions and validated them in whole genome genotyped population based FINRISK cohorts (N=20.000) using Cox regression models. Follow up data from national health care registers allowed us to model the impact of genetic and traditional risk factors such as smoking, cholesterol, blood pressure and BMI on a person's risk of disease within the next 10 years.

We observed that type 2 diabetes (T2DM) PRS significantly associates with the T2DM disease risk (HR:1.5 per 1 sd PRS, p-value:$<2*10^{-16}$). Also the top 8% of the FINRISK population who had inherited the highest PRS had fourfold increased risk for T2DM. Moreover, almost 30% of the people with >35 BMI and the highest PRS got T2DM in ten-year follow-up and T2DM incidences in >35 BMI group occurred at a younger age.

By combining the systemic genetic analyses with more traditional disease risk factors in the FINRISK cohort, we produced estimates on the impact of PRS and selected covariates on risk of T2DM. We use these estimates to assess the future risk of T2DM in P5.fi FinHealth participants who will receive this disease risk information including genetic risk via a web portal. We start returning personalized health reports to the P5 participants in May 2019. The impact of our intervention will be studied by following up the participants by questionnaires and through national health registers for five years. Our approach enables to identify the individuals within highest genetic risk and those with pre-disease symptoms.

ClinGen Linked Data Hub: scalable aggregation of diverse types of variant information to support pathogenicity assessment

Sai Lakshmi Subramanian[1], Andrew R. Jackson[1], Ronak Y. Patel[1], Arturo Lopez Pineda[3], Matthew Wright[3], Helio Costa[3], Bryan Wulf[3], Hannah Wand[3], Carlos D Bustamante[3], Sharon Plon[1,2], Aleksandar Milosavljevic[1], on behalf of the **Clinical Genome Resource**

1. Molecular and Human Genetics Department, Baylor College of Medicine, Houston, TX
2. Cancer Genetics Clinical and Research Programs, Texas Children's Hospital, Houston, TX
3. Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA

Abstract:

Variant pathogenicity assessment requires a diversity of functional, population-genetic, algorithmically-derived and other data types. ClinGen Linked Data Hub (LDH) aggregates large volumes of such data contributed from a growing number of diverse sources and makes it accessible for variant prioritization and curation. Using newly developed W3C Linked Data technologies, LDH enables contributors to link information about variants registered in the ClinGen Allele Registry and store a subset of relevant structured information within the LDH itself. LDH tracks provenance of the information and provides permanent reference for inclusion as supporting evidence in variant interpretation documents. W3C Web Annotation Model is employed as a mechanism for storing “quotes” to the effect that specific information about the variant was provided by a specific source at a specific time. The “quote content” is modeled using ClinGen/GA4GH/SEPIO community standards. LDH is implemented using the GenboreeKB platform, which also provides (as an optional service to contributors) cloud-hosted tools for modeling, validation, and hosting of detailed variant information which may then be linked through the LDH.

LDH is accessible via a web user interface linked through the ClinGen Allele Registry and also via Application Programming Interfaces (APIs) for use by ClinGen curation tools including the Variant Prioritization Tool, Variant Curation Interface, and the Pathogenicity Calculator as well as by the tools developed outside of ClinGen. These tools will have the option to subscribe to notifications about changes of variant information that may affect pathogenicity assessments, thus helping resolve discordances due to stale information and keep variant knowledge up to date.

We here present the first implementation of the LDH involving information about ~865,000 variants from 32 genes for use by the ClinGen Variant Prioritization Tool. We demonstrate tools for linking layers of variant information, tracking provenance and provision of permanent references. We demonstrate modeling and linking of information from GenboreeKB-hosted Functional Propositions repository that includes functional evidence from literature and

high-throughput experiments. Finally, we explore the future potential of the novel approaches for exchanging genomic information first implemented by LDH to empower clinical genomics and genetics and also to open new roads to discovery in genomics.

Abstract for CCG19

DECIPHER – Innovation in data-sharing in rare disease

HV Firth[1,2], AP Bevan[1], S Brent[1], J Foreman[1], B Hutton[1], D Perrett[1] K Samocha[1] and ME Hurles[1]

1 **Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, Hinxton CB10 1SA**
2 **Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Hills Road, Cambridge CB2 0QQ**

DECIPHER's mission is to map the clinically relevant elements of the genome and understand their contribution to human development and disease. Established in 2004, it is a major global platform for the visualization of genomic and phenotypic relationships. DECIPHER incorporates a Bayesian framework for ACMG variant classification enabling the user to see where on the continuum of risk from Benign (<0.001) to Pathogenic (>0.99) a given variant lies.

Three major recent innovations in DECIPHER are: (1) Visualising variants on the 3D protein structure (where a 3D structure exists) to aid assessment of the significance of missense variants, (2) Dynamic patient matching tunable by sex, inheritance, mutation consequence and variant similarity to show phenotypes seen in patients with identical or similar variants and (3) Aggregate data on the most frequent phenotypes seen in patients with pathogenic or likely pathogenic variants in a given gene.

DECIPHER (https://decipher.sanger.ac.uk) has a global network with >260 projects in 6 continents. DECIPHER is a pioneering partner in the GA4GH MatchMaker Exchange Application Programming Interface (API) enabling the federated discovery of similar variants in connected databases (eg. Phenome Central, MatchBox, MyGene2 and GeneMatcher).

DECIPHER promotes flexible data-sharing, enabling the extent of sharing to be varied proportionate to the clinical or scientific need to facilitate diagnosis or discovery. It hosts nearly 30,000 searchable, phenotype-rich, open-access records. Six consortia facilitate more limited sharing (including the NHS Consortium linking 20 UK Genetics services). More than 50,000 patient records are shared via consortia, including >120,000 phenotypes and their associated genomic variants.

Since its inception, DECIPHER has facilitated nearly 2000 publications in the peer-reviewed scientific literature; a testament to the importance of match-making in rare disease.

Title: Methodical curation of the human genome: challenges and progress in establishing an objective, consistent, and scalable process in a clinical laboratory.

Authors: Jackie Tahiliani, Kristin McDonald Gibson, Keith Nykamp, Rita Quintana, Brandon Bunker, Michelle Hogue, Jeanne Leisk, Amirah Khouzam, Invitae Clinical Genomics team, Robert L. Nussbaum, and Swaroop Aradhya.

Abstract: Clinical interpretation of molecular genetic variation in recent years has become more consistent due to improved guidelines for evaluating evidence and assigning pathogenicity classifications to variants. An important component of variant interpretation is an assessment of gene-disease relationships and the underlying biological mechanisms. The development of community-wide standards for evidence-based gene curation has begun recently, led by ClinGen and other groups. Modeled after these guidelines, we describe a gene curation workflow developed and implemented in a clinical laboratory setting at Invitae to support analysis of individual genes and multi-gene panels. In addition, using our curated database for clinical exome analysis has highlighted the need for standardized curation to ensure consistent reporting within and across laboratories.

We constructed a detailed and systematic approach for curating gene-level evidence and a custom database tool (Geneticus) for documentation. Observations such as affected probands, variant segregation, and functional characterization were evaluated for each gene and used to classify gene-disease relationships into three categories. To guide downstream variant classification, additional data were curated, including the mechanism of disease (e.g. haploinsufficiency), inheritance pattern, severity, and penetrance of the disease, age of onset, and detailed clinical features.

From evaluating ~14,000 genes, we have determined that ~2,900 genes have an established association with a disease, ~2,900 have a suggested relationship with a disease, and the remainder have little to no convincing evidence of a relationship with disease. Our methods for large-scale curation addressed several challenges, such as setting relative weights for each type of evidence assessed, minimizing areas of redundant evidence counting, and customizing rules for genes associated with rare versus common diseases.

To support clinical variant interpretation and exome analysis, we implemented a process to objectively assess gene-disease relationships and deployed a tool that enables large-scale curation. Our experience with >14,000 genes has revealed important nuances in capturing and evaluating specific types of evidence that impacts clinical interpretation. The methods we have applied provided insight that can inform the standardization of gene curation across laboratories and further foster consistency in clinical genetic testing.

Improving efficiency of gene curation to support reporting in a clinical laboratory setting.

Sarah A. Schmidt, Revathi Rajkumar, Vladimir G. Gainullin, Zinayida Schlachetzki, David R. Bentley, Ryan J. Taft, Denise L. Perry, Alison J. Coffey.

Illumina Clinical Services Laboratory, Illumina Inc., San Diego, CA, USA.

The Illumina Clinical Services Laboratory uses the ClinGen Gene-Disease Validity framework for gene curation to support reporting for clinical whole genome sequencing (cWGS) tests. This framework, whilst providing a robust method for gene curation, is a multi-step process involving extensive collection and evaluation of clinical and experimental evidence to allow the assignment of a gene-disease validity classification. The manual time investment required to curate an individual gene-disease association (GDA), ranges from two to over eight hours, and presents practical challenges to current reporting workflows and efforts to further reduce test turn-around-time.

Over 600 GDAs have been curated from November 2016 to February 2019 as part of our gene curation programme. Through our extensive experience, we have determined two time-intensive parts of the process that could benefit from automation: 1) automated gathering of information from external sources, such as frequency data, gene and disease names and other identifiers; and 2) automated identification of publications with relevant clinical and experimental data.

To address automated gathering of information, we use APIs from HGNC, NCBI, OMIM, Orphanet, GeneReviews, Genetics Home Reference, MONDO, GnomAD, PubMed, ClinVar and Uniprot to auto-populate sections of the curation template. We are developing further tools to aggregate remaining relevant information from these sources and other databases including FlyBase, MGI, ZFIN, MitoMap, and MSeqDR.

To address identification of relevant publications, we are developing automated literature searches based on the name(s) of the gene and disease under curation. Relevant literature can be sorted into different categories including the first report of a GDA, recent review articles, case reports and experimental evidence. We are also exploring natural language processing to auto-highlight key information in the publications to allow the curator to quickly focus on relevant clinical and experimental data.

Automated gathering of pertinent information for gene curation will allow the curation scientist to focus on evaluating evidence. This has the potential to decrease overall turn-around-time for case interpretation and reporting for our cWGS test, and overall hands-on-time for the highly skilled and trained curation scientists conducting this work.

The Gene Curation Coalition: A global effort to harmonize gene-level resources

Marina T. DiStefano[1], Joanna Amberger[2], Jonathan S. Berg[1], Ewan Birney[3], Ozge Birsoy[1], Carol Bocchini[2], Elspeth Bruford[4,5], Heather Collins[6,7], Fiona Cunningham[3], Louise C. Daugherty[8], Paul Denny[4], Helen Firth[9,10], David Fitzpatrick[9,10], Rebecca E. Foulger[8], Robert Fullem[11], Jenny Goldstein[1], Ada Hamosh[2], Matthew Hurles[9,10], Sarah E. Leigh[8], Ivone U.S. Leong[8], Antoine Marmignon[12], Christa L. Martin[1], Ellen M. McDonagh[8], Annie Olry[12], Sharon Plon[1], Nazneen Rahman[9], Erin M. Ramos[11], Erin Rooney Riggs[1], Charlotte Rodwell[12], Alan F. Scott[2], Ann Strydom[9], Emma Tudini[10], James S. Ware[13,14], Eleanor Williams[8], Nicola Whiffin[13,14], Caroline Wright[10], and Heidi L. Rehm[1]

Affiliations:
[1]Clinical Genome Resource (ClinGen), USA
[2]Online Mendelian Inheritance in Man (OMIM), Johns Hopkins University School of Medicine, USA
[3] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
[4] HUGO Gene Nomenclature Committee (HGNC), European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.
[5]Department of Haematology, University of Cambridge School of Clinical Medicine, Cambridge CB2 0XY, UK
[6]Genetics Home Reference, National Library of Medicine, Bethesda, MD,, USA
[7]ICF, 9300 Lee Highway, Fairfax, VA, 22031 USA
[8]Genomics England, Queen Mary University of London, Dawson Hall, Charterhouse Square London, EC1M 6BQ, UK
[9]Transforming Genetic Medicine Initiative, Wellcome Genome Trust, England
[10]Deciphering Developmental Disorders/Gene2Phenotype, England
[11]National Human Genome Research Institute, National Institutes of Health, USA
[12]Inserm-US14 - Orphanet, France
[13]National Heart & Lung Institute & MRC London Institute of Medical Sciences, Imperial College London, London UK
[14]Royal Brompton & Harefield Hospitals NHS Trust, London UK

The assessment of the evidence that a gene is linked to a particular disease is critical for variant curation. Unless a gene is convincingly linked to disease, the clinical impact of a variant cannot be interpreted. Thus, curation of gene-disease validity is a fundamental prerequisite for classifying variants. Several groups and resources provide information that pertains to the validity of gene-disease relationships; however, the standards and terminologies to define the evidence base for a gene's role in disease are still evolving. To tackle this issue, the Gene Curation Coalition (GenCC) was formed including members of the Clinical Genome Resource (ClinGen), Deciphering Developmental Disorders/Gene2Phenotype (DDD/G2P), Genetics Home Reference (GHR), Genomics England PanelApp (PanelApp), Online Mendelian Inheritance in Man (OMIM), Orphanet,

and the Transforming Genetic Medicine Initiative (TGMI). Together, this group is working to harmonize approaches to ensure gene-level resources are comparable and interoperable.

The GenCC drafted harmonized definitions for differing levels of gene-disease clinical validity. A Delphi survey was then performed to narrow the list of terms for clinical validity in the context of genes implicated in monogenic diseases. In the first round, members of the GenCC took the survey and chose from term sets already in use by current efforts or suggested new terms. In the second round, all previous and suggested terms were incorporated and the survey was sent out to the extended members of each of the GenCC groups. These responses were used to narrow down the term list. In January, the survey was sent to the larger international genetics community with an introductory video for context. The 220 community responses were used to further narrow the term list and finalize a harmonized term set. The GenCC recommended term set will be used to map all other terms used by each curation effort participating in the GenCC. The most common use case for terms generated by this survey is to determine which gene evidence levels should be used in diagnostic testing for monogenic disease as well as presymptomatic testing and other health screening. Terminology standardization will help to improve harmonization, facilitate the pooling of curation datasets across international curation efforts and in turn, improve consistency in genetic testing and variant interpretation.

Transcript curation for the clinical use case: LRG and MANE

J. Morales[1], J. Loveland[1], S Pujar[2], A. Astashyn[2], R. Bennett[1], C. Davidson[1], O. Ermolaeva[2], C. Farrell[2], L. Gil[1], J.M. Gonzalez[1], V. Joardar[2], M. Kay[1], V. Kodali[2], K. McGarvey[2], A. McMahon[1], S. Rangwala[2], G. Threadgold[1], F. Cunningham[1], A. Frankish[1], T. Murphy[2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK CB10 1SD.
[2]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, U.S.A. 20894

The accurate annotation of genes and transcripts is essential for clinical genomics. However, the large number of alternatively spliced transcripts and the lack of standardized reference sequences present challenges in the clinical context. To address this, the EBI and NCBI are working together on two related projects: the Locus Reference Genomic (LRG) and the Matched Annotation from the NCBI and EMBL-EBI (MANE). For a given locus, LRGs provide a stable genomic DNA sequence and a minimal set of transcripts to be used as reference standards. MANE aims to identify one representative transcript (MANE Select) to be used as default across genomics resources. In the clinical context, the goal of both is to converge on a high-confidence, genome-wide transcript set that is well-suited to the needs of the community. Our approach involves the use of automation in combination with manual review. For clinical genes, manual curation is essential. It first involves an update of annotation to incorporate newly released data, followed by the selection of transcripts based on a defined set of criteria. These include RNA expression (overall and tissue specific, if available), conservation, location of pathogenic variants, length and historical usage. We only select transcripts with evidence of functional potential and that contribute high-value features, such as well-supported alternatively spliced protein-coding exons. The minimal set of transcripts at each clinical locus is then included in an LRG record, with the MANE Select being one of these transcripts. The exception are LRGs for which the genomic backbone represents an alternate allele due to problems with the genome assembly at that locus whereas the MANE Select must always match GRCh38. Collaboration with locus-specific experts in the clinical community is a core component of our work. Currently, we are consulting with collaborators to prioritise genes with high gene-disease validity and to ensure that our method adequately fulfils demands. Over 1,260 records are available on the LRG website (http://www.lrg-sequence.org/) and on all genome browsers. MANE Select transcripts have been defined for 53% of protein-coding genes, with an increase expected in the Autumn of 2019. This work is supported in part by the intramural research program of the National Library of Medicine, National Institutes of Health, the Wellcome Trust (WT108749/Z/15/Z and WT200990/Z/16/Z) and the European Molecular Biology Laboratory.

Somatic variant curation standards enable improved identification of relevant clinical interpretations for tumor variants

AUTHORS

Alex H Wagner[1], Brian Walsh[2], Georgia Mayfield[2], David Tamborero[3], Dmitriy Sonkin[4], Kilannin Krysiak[1], Jordi Deu Pons[3], Ryan P Duren[5], Jianjiong Gao[6], Julie McMurry[2], Sara Patterson[7], Catherine Del Vecchio Fitz[8], Beth A Pitel[9], Ozman U Sezerman[10], Kyle Ellrott[2], Jeremy L Warner[11], Damian T Rieke[12], Tero Aittokallio[13], Ethan Cerami[8], Deborah Ritter[14], Lynn M Schriml[15], Robert R Freimuth[9], Melissa Haendel[16], Gordana Raca[17], Subha Madhavan[18], Michael Baudis[19], Jacques S Beckmann[20], Rodrigo Dienstmann[21], Debyani Chakravarty[6], Xuan Shirley Li[5], Susan Mockus[7], Olivier Elemento[22], Nikolaus Schultz[6], Nuria Lopez-Bigas[3], Mark Lawler[23], Jeremy Goecks[2], , Adam A Margolin[2], Malachi Griffith[1], Obi L Griffith[1]; Variant Interpretation for Cancer Consortium

AFFILIATIONS

[1]Washington University School of Medicine, St. Louis, MO, USA
[2]Oregon Health and Science University, Portland, OR, USA
[3]Institute for Research in Biomedicine, Barcelona, Spain
[4]National Cancer Institute, Rockville, MD, USA
[5]MolecularMatch, Houston, TX, USA
[6]Memorial Sloan Kettering Cancer Center, New York, NY, USA
[7]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
[8]Dana-Farber Cancer Institute, Boston, MA, USA
[9]Mayo Clinic, Rochester, MN, USA
[10]Acibadem University, Istanbul, Turkey
[11]Vanderbilt University, Nashville, TN, USA
[12]Charité University Medicine Berlin, Berlin, Germany
[13]University of Helsinki and FIMM, Helsinki, Finland
[14]Baylor College of Medicine, Houston, TX
[15]University of Maryland School of Medicine, Baltimore, MD, USA
[16]Linus Pauling Institute, Oregon State University, OR, USA
[17]University of Southern California, Los Angeles, CA, USA
[18]Georgetown University Medical Center, Washington D.C., USA
[19]University of Zurich, Zurich, Switzerland
[20]University of Lausanne, Lausanne, Switzerland
[21]Vall d'Hebron Institute of Oncology, Barcelona, Spain
[22]Weill Cornell Medicine, New York, NY, USA
[23]Queen's University Belfast, Belfast, United Kingdom

ABSTRACT

The promise of precision medicine–whereby a cancer patient's treatment is tailored to their specific cancer variants–requires concise, standardized, and searchable clinical interpretations (biomarker-disease associations that provide therapeutic or prognostic value). While many institutions curate interpretations for clinically relevant cancer variants, these efforts are siloed and non-interoperable. Through the Global Alliance for Genomics and Health (GA4GH), we formed a Variant Interpretation for Cancer Consortium (VICC; cancervariants.org). In this work, we evaluate six established sources of cancer variant interpretations hosted

by members of the VICC: the Clinical Interpretations of Variants in Cancer knowledgebase, OncoKB, the Jackson Laboratories Clinical Knowledgebase, the Precision Medicine Knowledgebase, MolecularMatch, and the Cancer Genome Interpreter. We describe a framework for aggregating and harmonizing variant interpretations, and use it to produce a meta-knowledgebase of 12,856 interpretations spanning 3,437 unique variants in 415 genes, 357 diseases, and 791 drugs. Our framework is informed by published guidelines for somatic variant curation, including the ClinGen Minimum Variant Level Data and AMP/ASCO/CAP standards, and we demonstrate how these guidelines fit to existing data models. In addition, we leverage existing resources to harmonize and link interpretation data elements, including the ClinGen Allele Registry for genomic variants and the Human Disease Ontology for diseases. We explore the utility of these resources and evaluate the strengths and challenges of using them in a somatic interpretation harmonization strategy.

Ultimately, we demonstrate large gains in overlap between resources across variants, diseases, and drugs as a result of our harmonization strategy. We subsequently evaluate 237,175 somatic mutations from 38,207 patient donors in the AACR Project GENIE cohort against our meta-knowledgebase, and find that our efforts identified potentially clinically significant genetic alterations in 57% of patients, compared to an average of 33% (range of 8% to 47%) identified using a single knowledgebase. Our analyses illuminate the need for open, interoperable, and standard-driven sharing of variant interpretation data. We also freely provide an available web interface (documented at docs.cancervariants.org) for exploring the harmonized meta-knowledgebase, and open-source code for all analyses, methods, and tools.

Implementation and cases studies of tumor whole-genome sequencing within Pediatrics Precision Medicine at MSKCC

Dominik Głodzik[1], Neerav Shukla*[2], Max Levine*[1], Gunes Gundem[1], Juan Medina[1], Yangyu Zhou[1], Juan E. Arango Ossa[1], Nancy Bouvier[2], Irene Rodriguez-Sanchez[2], Olivia Wilkins[2], Shanita Li[2], Geraldine Martinez[2], Agnès Viale[3], Michael F. Walsh[2], Elli Papaemmanuil[1], Andrew L. Kung[2]*

[1] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA;
[2] Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA;
[3] Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA;

** contributed equally*

Through cancer genome profiling we have a developed a near–complete understanding of the spectrum of genes that are recurrently mutated in cancer as well as the nature of the genetic events that drive malignant transformation. The cancer genomes are defined by patterns of acquired mutations, coding and non-coding, complex structural variations, and alterations in ploidy. Many of these are relevant to diagnosis, risk stratification and treatment decisions. Currently, a number of distinct assays are ordered to identify such genomic alterations in the clinic.

Whole genome sequencing (WGS) captures the spectrum of mutations in a cancer genome to include mutations in rarer cancer genes, non-coding regions, structural variants and ploidy. Additionally, information on mutational burden and mutational signatures can be extracted. Nevertheless, the utility of WGS analysis in clinical care, as well as the logistic tractability, remain to be proven. Here, we demonstrate a proof-of-concept WGS and RNA-seq workflow that is integrated into Pediatrics Precision Medicine at MSKCC.

To enable scalable processing of genomes, we developed the *Isabl* platform. Genome analysis is augmented by detailed annotations, comparisons to large datasets and intuitive visualizations. The workflow is optimized to deliver a comprehensive genome report within three weeks of tissue collection.

To date, we have analyzed 40 pediatric genomes. WGS data revealed at least 6 disease-defining gene fusions. Importantly, disease defining and prognostic mutations missed by gene panels due to their limited footprint were identified e.g. *ATRX* rearrangements in osteosarcoma and neuroblastoma. Other rearrangements could inform enrolment into clinical trials, e.g. *TP53* rearrangements precluding from *MDM2*-inhibitor trials. We additionally observed pediatric tumors with high mutational burden reaching 8 mutations per Mb, prompting assessment of inherited defects as well as candidacy to therapeutic interventions that exploit DNA repair deficiencies and immune checkpoint blockade. Among the highly mutated tumors, some show patterns of somatic mutations concordant with known DNA repair disorders. The mutational signatures can be used to phenotype putative oncogenic variants among the DNA repair genes.

The case studies highlight areas for improvement of analytical workflows, our understanding of genome biology and integration of genomes as biomarkers into clinical trials.