Title

The GA4GH Phenopacket schema defines a computable representation of clinical data

Journal

Nature Biotechnology, 40(6)

ISSN

1087-0156

Authors

Jacobsen, Julius OB

Baudis, Michael

Baynam, Gareth S

et al.

Publication Date

2022-06-01

DOI

10.1038/s41587-022-01357-4

Peer reviewed

The GA4GH Phenopacket schema defines a computable representation of clinical data

References

1. Dolgin, E. *Nature* **551**, 427–431 (2017).
2. Haynes, W. A., Tomczak, A. & Khatri, P. *Sci. Rep.* **8**, 1362 (2018).
3. Wood, V. et al. *Open Biol.* **9**, 180241 (2019).
4. Stoger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. *PloS Biol.* **16**, e2006643 (2018).
5. Edwards, A. M. et al. *Nature* **470**, 163–165 (2011).
6. Oprea, T. I. et al. *Nat. Rev. Drug Discov.* **17**, 317–332 (2018).
7. Dunham, I. *PLoS Biol.* **16**, e3000034 (2018).
8. Stoeger, T. & Nunes Amaral, L. A. *eLife* **9**, e61981 (2020).
9. Hutchison, C. A. III et al. *Science* **351**, aad6253 (2016).
10. Kustatscher, G. *Nat. Methods* https://doi.org/10.1038/s41592-022-01454-x (2022).
11. Sinha, S., Eisenhaber, B., Jensen, L. J., Kalbuaji, B. & Eisenhaber, F. *Proteomics* **18**, e1800093 (2018).
12. UniProt Consortium. *Nucleic Acids Res.* **47**, D506–D515 (2019).
13. Anton, B. P. et al. *PLoS Biol.* **11**, e1001638 (2013).
14. Gerlt, J. A. et al. *Biochemistry* **50**, 9950–9962 (2011).
15. Williamson, A. R. *Nat. Struct. Biol.* **7**, 953 (2000).
16. Koscielny, G. et al. *Nucleic Acids Res.* **45**, D985–D994 (2017).

Competing interests

T.G. is a shareholder of Westlake Omics Inc. T.I. is a cofounder of Data4Cure, is on the Scientific Advisory Board and has an equity interest. T.I. is on the Scientific Advisory Board of Ideaya BioSciences and has an equity interest. E.L. is advisor for Pixelgen technologies and Moleculent. E.M.M. is a cofounder, shareholder and scientific board member of Erisyon, Inc. G.K., T.C., A.-C.G., H.H., K.S.L., M.R. and J.R. declare no competing interests.

The GA4GH Phenopacket schema defines a computable representation of clinical data

To the Editor — Despite great strides made in the development and wide acceptance of standards for exchanging structured information about genomic variants, progress in standards for computational phenotype analysis for translational genomics has lagged behind. Phenotypic features (signs, symptoms, laboratory and imaging findings, results of physiological tests, etc.) are of high clinical importance, yet exchanging them in conjunction with genomic variation information is often overlooked or even neglected. In the clinical domain, substantial work has been dedicated to the development of computational phenotypes[1]. Traditionally, these approaches have largely relied on rule-based methods and large sources of clinical data to identify cohorts of patients with or without a specific disease[2–5]. However, they were not developed to enable deep phenotyping of abnormalities, to facilitate computational analysis of interpatient phenotypic similarity or to support computational decision support. To address this, the Global Alliance for Genomics and Health[6] (GA4GH) has developed the Phenopacket schema, which supports the exchange of computable longitudinal case-level phenotypic information for diagnosis of, and research on, all types of disease, including Mendelian and complex genetic diseases, cancers and infectious diseases. A Phenopacket characterizes an individual person or biosample, linking that individual to detailed phenotypic descriptions, genetic information, diagnoses and treatments (Fig. 1). The Phenopacket software is available at https://github.com/phenopackets/.

The 'PhenotypicFeature' is the central element of the Phenopacket schema. A 'PhenotypicFeature' can be used to describe any phenotypic characteristic, including signs and symptoms, laboratory findings, histopathology findings, and imaging and electrophysiological results, along with modifier and qualifier concepts. Each phenotypic feature is described using an ontology term. Although the Phenopacket schema does not mandate which ontology to use, it provides recommendations, such as the Human Phenotype Ontology[7] (HPO) for rare diseases and the National Cancer Institute Thesaurus (NCIT) for transmission of information about a cancer specimen (for example, pathological staging or more detailed information about histology or tumor markers)[8]. Within the schema, it is possible to indicate whether an abnormality was excluded during the diagnostic process (for example, whether a morphological cardiac defect was excluded by echocardiography) or to use other optional HPO terms to denote the severity, frequency (for example, number of occurrences of seizures per week), laterality (for example, unilateral) or other pattern of a phenotypic feature in the patient being described. Finally, the onset (and, if applicable, the resolution) of specific features can be indicated.

Other key elements of the schema are 'Measurement', which is used to capture quantitative (i.e., numerical), ordinal (for example, absent/present) or categorical measurements; 'Biosample', a description of biological material obtained from the individual represented in the Phenopacket and used for phenotypic, genotypic or other -omics analysis; and 'MedicalAction', which includes a hierarchical representation of medical actions, including medications, procedures and other actions taken for clinical management. The 'Treatment' element is a subelement of 'MedicalAction' and represents the administration of a pharmaceutical agent, broadly defined as prescription and over-the-counter medicines, vaccines and other therapeutic agents, such as monoclonal antibodies or chimeric antigen receptor (CAR)-T-cell therapy.

The 'Interpretation' element specifies interpretations of genomic findings. This element leverages complementary resources developed by the GA4GH Genomic Knowledge Standards Work Stream: the Variation Representation Specification (VRS) and VRS Added Tools for Interoperable Loquacious Exchange (VRSATILE)[6]. Further information on this and other elements is available in the online documentation (https://phenopacket-schema.readthedocs.io/).

The Phenopacket schema was designed to support several use cases. Phenotype-driven rare-disease genomic diagnostic software has previously used bespoke formats to represent phenotypic data (generally in the form of a list of HPO terms) and pedigree information. Phenopacket provides a standard input format for these tools that will simplify computational analysis pipelines, and the additional clinical information will enable analysis pipelines and algorithms to leverage other data, such as age of onset and excluded abnormalities. A number of databases have adopted the standard to represent the clinical data of individuals in the context of rare-disease genomics (European Genome-phenome Archive), registries (European Joint Programme on Rare Diseases and Western Australian Register of Developmental Anomalies), biosamples (EMBL-EBI BioSamples database) and biobanks (the Japanese Agency for Medical Research and Development Tohoku Medical Megabank project and National Center Biobank Network). In addition, Phenopackets can be used to store a computational representation of a case report, and we envision that authors could submit representations of patients as phenopackets to accompany published case reports

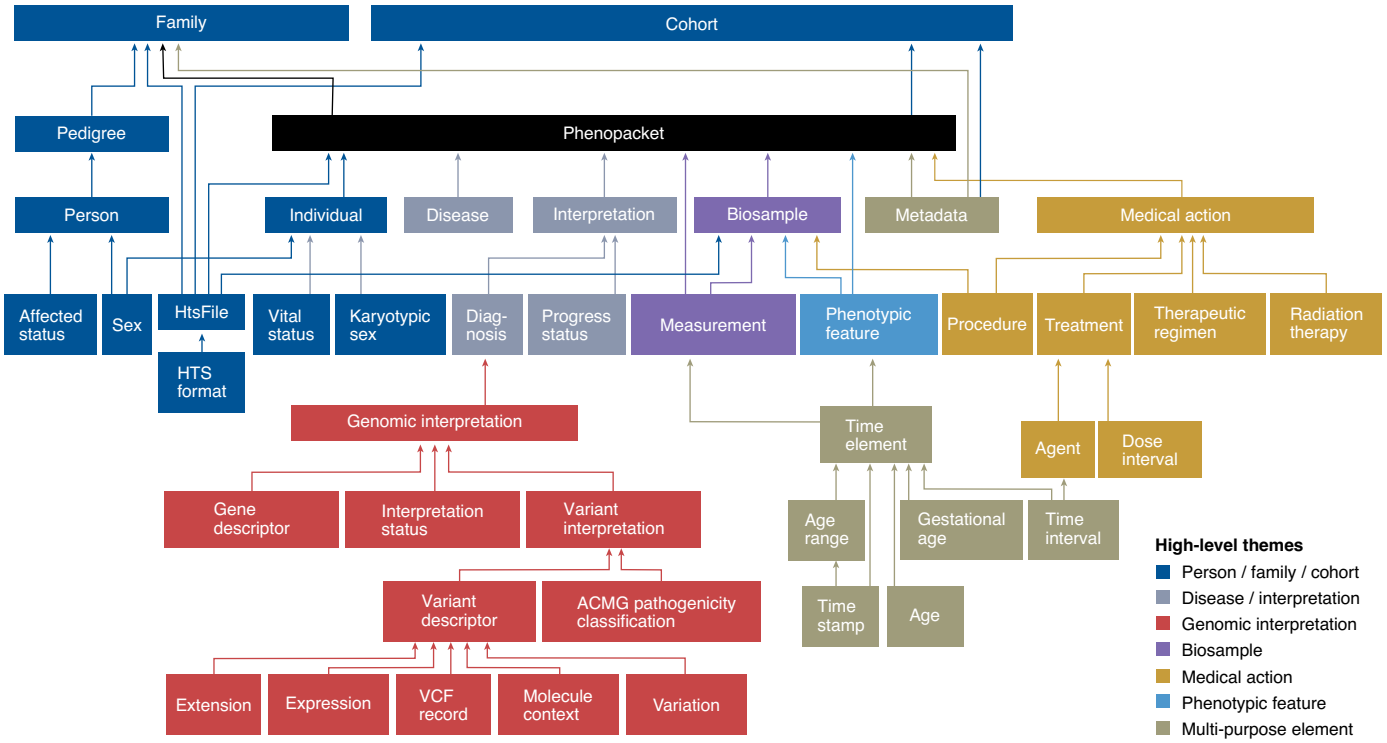

Fig. 1 | Phenopacket schema overview. The GA4GH Phenopacket schema consists of several optional elements, each containing information about a certain topic, such as phenotype, variant or pedigree. An element can contain other elements, which allows a hierarchical representation of data. For instance, Phenopacket contains elements of type Individual, PhenotypicFeature, Biosample and so on. Individual elements can therefore be regarded as building blocks that are combined to create larger structures. Colors represent the major themes of elements within the schema. ACMG, American College of Medical Genetics; HTS, high-throughput screening; VCF, variant call format.

and descriptions of genotype–phenotype correlations. In addition to these use cases, the Phenopacket schema is designed to interact with electronic health record (EHR) data. A longstanding challenge has been that computational phenotype analysis is poorly connected with EHRs and also that EHRs are not standardized across countries or even across institutions in a given country. To enable precision medicine, standards and tools are needed to improve machine-readable phenotypic characterization of patients beyond current standard EHR billing and clinical encounter data capture. To address this, we have created a Fast Healthcare Interoperability Resources (FHIR) implementation guide for representing a phenopacket within EHR systems (Supplementary Table 1).

Requirements and specifications for the standard were established through a community effort under the auspices of the GA4GH; Version 1.0 of the GA4GH standard was released in 2019 to elicit feedback from the community. Version 2.0, which is described here, was developed on the basis of this feedback and expanded the data model to include a better representation of temporality, medical actions and quantitative measures. The Phenopacket schema (version 2.0) was formally reviewed and approved as a GA4GH standard[6] in 2021. It is designed to be interoperable with other relevant standards, including the traditional PED (pedigree) file format as well as the GA4GH pedigree standard, the GA4GH Beacon[9] and the GA4GH Variation Representation Specification. The GA4GH has committed to coordinate its activities and future roadmaps with those of other standards development organizations, including the International Organization for Standardization (ISO) Technical Subcommittee for Genomics Informatics (ISO/TC215/SC1) and HL7 Clinical Genomics. Consequently, an FHIR implementation guide for Phenopacket interoperability has been developed, and the Phenopacket schema is at the approval stage of the ISO certification process (Supplementary Table 2).

The variant call format (VCF) standard for storing genotyping data allowed a wide range of research groups to write software for analyzing such data[10]. The GA4GH Phenopacket schema aspires to be similarly transformative in the landscape of genome analysis using phenotype data. The multiple providers of phenotypic data include patients and clinicians and convey data via a variety of mechanisms, including clinical notes and electronic health records, interfaces such as FHIR, app-based entry and mobile devices. The Phenopacket schema acts as a common model that can capture data from many sources with a unified software representation and, in its turn, can be used by multiple receivers of the phenotypic information, including journals, databases, registries and clinical laboratories. We anticipate that the Phenopacket schema will encourage the development of a collection of software for the analysis of genomic data in the context of clinical information that will accelerate innovation and discovery. Genomic data will become ever more important in translational research and clinical care in the coming years and decades. The Phenopacket schema represents a standard for capturing clinical data and integrating it with genomic data that will help to obtain the maximal utility of this data for understanding disease and developing precision medicine approaches to therapy. ❐

Julius O. B. Jacobsen[1]✉, Michael Baudis[2,3], Gareth S. Baynam[4,5,6],

Jacques S. Beckmann[7], Sergi Beltran[8,9,10], Orion J. Buske[11], Tiffany J. Callahan[12], Christopher G. Chute[13], Mélanie Courtot[14,15], Daniel Danis[16], Olivier Elemento[17], Andrea Essenwanger[18], Robert R. Freimuth[19], Michael A. Gargano[16], Tudor Groza[20], Ada Hamosh[21], Nomi L. Harris[22], Rajaram Kaliyaperumal[23], Kevin C. Kent Lloyd[24,25], Aly Khalifa[19], Peter M. Krawitz[26], Sebastian Köhler[27], Brian J. Laraway[12], Heikki Lehväslaiho[28], Leslie Matalonga[8], Julie A. McMurry[12], Alejandro Metke-Jimenez[29], Christopher J. Mungall[22], Monica C. Munoz-Torres[12], Soichi Ogishima[30], Anastasios Papakonstantinou[8], Davide Piscia[8], Nikolas Pontikos[31,32], Núria Queralt-Rosinach[23], Marco Roos[23], Julian Sass[18], Paul N. Schofield[33,34,35], Dominik Seelow[36,37], Anastasios Siapos[38], Damian Smedley[1], Lindsay D. Smith[15,39], Robin Steinhaus[36,37], Jagadish Chandrabose Sundaramurthi[16], Emilia M. Swietlik[40,41,42], Sylvia Thun[18], Nicole A. Vasilevsky[12], Alex H. Wagner[43,44], Jeremy L. Warner[45], Claus Weiland[46], The GAGH Phenopacket Modeling Consortium*, Melissa A. Haendel[12] ✉ and Peter N. Robinson[16,47] ✉

*[1]William Harvey Research Institute, Queen Mary University of London, London, UK. [2]Department of Molecular Life Sciences, University of Zurich, Zürich, Switzerland. [3]Computational Oncogenomics Group, Swiss Institute of Bioinformatics, Zürich, Switzerland. [4]Western Australian Register of Developmental Anomalies and Genetic Services of WA, King Edward Memorial Hospital, Perth, Western Australia, Australia. [5]Faculty of Health and Medical Sciences, Division of Paediatrics, University of Western Australia, Perth, Western Australia, Australia. [6]Genetic and Rare Diseases, Telethon Kids Institute, Perth, Western Australia, Australia. [7]Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland. [8]CNAG-CRG, Centre for Genomic Regulation (CRG), Bioinformatics Unit, The Barcelona Institute of Science and Technology, Barcelona, Spain. [9]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [10]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain. [11]PhenoTips, Toronto, Ontario, Canada. [12]Center for Health AI, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [13]Schools of Medicine, Public Health, and Nursing, Baltimore, Johns Hopkins University, Baltimore, MD, USA. [14]European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Hinxton, UK. [15]Genome Informatics, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [16]Genomic Medicine, The Jackson Laboratory, Farmington, CT, USA. [17]Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. [18]Core Facility Digital Medicine and Interoperability, Berlin Institute of Health at Charité–Universitätsmedizin Berlin, Berlin, Germany. [19]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA. [20]EMBL-EBI, Cambridge, UK. [21]Department of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA. [22]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [23]Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [24]Mouse Biology Program, University of California Davis, Davis, CA, USA. [25]Department of Surgery, University of California Davis School of Medicine, Sacramento, CA, USA. [26]University Hospital Bonn, Bonn, Germany, and Institute for Genomic Statistics and Bioinformatics, Bonn, Germany. [27]Ada Health GmbH, Berlin, Germany. [28]Sensitive Data Services, CSC–IT Center for Science, Espoo, Finland. [29]The Australian e-Health Research Centre, CSIRO, Herston, Queensland, Australia. [30]Tohoku University, INGEM, Sendai, Japan. [31]Institute of Ophthalmology, University College London, London, UK. [32]Genetics Service, Moorfields Eye Hospital, London, UK. [33]Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK. [34]Mammalian Genetics, The Jackson Laboratory, Bar Harbor, ME, USA. [35]The Alan Turing Institute, London, UK. [36]Bioinformatics and Translational Genetics, Berlin Institute of Health at Charité–Universitätsmedizin Berlin, Berlin, Germany. [37]Institute of Medical Genetics and Human Genetics, Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany. [38]Lifebit Biotech Ltd., London, UK. [39]Global Alliance for Genomics and Health, N/A, Toronto, Ontario, Canada. [40]40Medicine Department, University of Cambridge, Cambridge, UK. [41]Respiratory Medicine Department, Addenbrooke's Hospital, Cambridge, UK. [42]Cambridge Centre for Lung Infection, Royal Papworth Hospital, Cambridge, UK. [43]The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. [44]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, USA. [45]Departments of Medicine and Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. [46]Senckenberg–Leibniz Institution for Biodiversity and Earth System Research, Data and Modelling Centre, Frankfurt/Main, Germany. [47]Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: j.jacobsen@qmul.ac.uk; melissa@tislab.org; peter.robinson@jax.org*

Published online: 15 June 2022
https://doi.org/10.1038/s41587-022-01357-4

References

1. Richesson, R. L. et al. *J. Am. Med. Inform. Assoc.* **20**, e226–e231 (2013).
2. Hripcsak, G. & Albers, D. J. *J. Am. Med. Inform. Assoc.* **20**, 117–121 (2013).
3. Shivade, C. et al. *J. Am. Med. Inform. Assoc.* **21**, 221–230 (2014).
4. Wei, W.-Q. & Denny, J. C. *Genome Med.* 7, 41 (2015).
5. Richesson, R. L., Sun, J., Pathak, J., Kho, A. N. & Denny, J. C. *Artif. Intell. Med.* **71**, 57–61 (2016).
6. Rehm, H. L. et al. *Cell Genom.* **1**, 100029 (2021).
7. Köhler, S. et al. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
8. Sioutos, N. et al. *J. Biomed. Inform.* **40**, 30–43 (2007).
9. Fiume, M. et al. *Nat. Biotechnol.* **37**, 220–224 (2019).
10. Danecek, P. et al. *Bioinformatics* **27**, 2156–2158 (2011).

Acknowledgements

The authors gratefully acknowledge insight and feedback from Marian H. Adly, Pier Luigi Buttigieg, Janine Lewis, Manuel Posada de la Paz and Maria Taboada. This work was supported by 7RM1HG010860-02 (NHGRI). Additional funding was as follows. P.N.R. was supported by NLM contract 75N97019P00280, NIH NHGRI RM1HG010860, NIH OD R24OD011883 and NIH NICHD 1R01HD103805-01. H.H. was supported by NIH OD R24OD011883. G.I.S. was supported by ELIXIR, the research infrastructure for life-science data. C.G.C. was supported by NIH NCATS U24TR002306. K.C.L. was supported by NIH OD 5UM1OD023221. M.B. was supported by the BioMedIT Network project of the Swiss Institute of Bioinformatics (SIB) and Swiss Personalized Health Network (SPHN). A.H.W. was supported by NIH NHGRI K99HG010157 and NIH NHGRI R00HG010157. C.J.M., M.A.H., M.C.M.-T., J.A.M. and D.D. were supported by NIH NHGRI RM1HG010860 and NIH OD R24OD011883. A.M.-J. was supported by Australian Genomics. Australian Genomics is supported by the National Health and Medical Research Council (GNT1113531). D. Smedley and J.O.B.J. were supported by NIH NHGRI RM1HG010860, NIH OD R24OD011883 and NIH NICHD 1R01HD103805-01. M.D. was supported by NIH NHGRI U54HG004028, NIH NHGRI 5U01HG008473-03 and NIH NCATS OT2TR003434-01S1U54HG008033-01. G.S.B. was supported by the Roy Hill Community Foundation, Angela Wright Bennett Foundation, McCusker Charitable Foundation, Borlaug Foundation and Stan Perron Charitable Foundation. L.B. was supported by NIH NHGRI U41HG006834 (Clinical Genome Resource). M.C. was supported by EMBL-EBI Core Funds and Wellcome Trust GA4GH award number 201535/Z/16/Z. A.H. was supported by NIH NHGRI 1U41HG006627, NIH NHGRI 1U54HG006542 and NIH NHGRI 1RM1HG010860. P.N.S. was supported by The Alan Turing Institute. N.L.H. was supported by NIH NHGRI RM1HG010860, NIH OD R24OD011883 and US Department of Energy Contract DE-AC02-05CH11231. N.P. was supported by Moorfields Eye Charity. N.Q.-R. was supported by EU Horizon 2020 research and innovation programme grant agreement 825575 (EJP-RD). O.E. was supported by NIH grants UL1TR002384, R01CA194547 and P01CA214274, LLS SCOR grants 180078-01 and 7021-20 and Starr Cancer Consortium grant I11-0027. H. Lochmüller was supported by the CIHR Foundation Grant on Precision Health for Neuromuscular Diseases FDN-167281. R.T. was supported by CIHR postdoctoral fellowship award MFE-171275. L.D.S. was supported by Genome Canada and NIH NHGRI U24HG011025. S.O. was supported by AMED. D.P., L.M., A.P., S.B., M.R. and R.K. were supported by EU Horizon 2020 research and innovation programme grant agreements 779257 (Solve-RD) and 825575 (EJP-RD). R.R.F. was supported by NLM contract 75N97019P00280.

Competing interests

S.K. is an employee of Ada Health GmbH. N.P. is a director of Phenopolis Ltd. O.E. is supported by Janssen, Johnson & Johnson, Volastra Therapeutics, AstraZeneca and Eli Lilly research grants, and is a scientific advisor and equity holder in Freenome, Owkin, Volastra Therapeutics and One Three Biotech. A.R.M. is an employee of Philips Research North America. O.J.B. is an employee of PhenoTips. M.A. is an editor employed by Wiley. A.S. is an employee of Lifebit Biotech Ltd. The remaining authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01357-4.
Peer review information *Nature Biotechnology* thanks Kai Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

The GAGH Phenopacket Modeling Consortium

Myles Axton[48], Lawrence Babb[49], Cornelius F. Boerkoel[50], Bimal P. Chaudhari[43,44], Hui-Lin Chin[51,52], Michel Dumontier[53], Nour Gazzaz[52,54], David P. Hansen[29], Harry Hochheiser[55], Veronica A. Kinsler[56,57], Hanns Lochmüller[58,59,60], Alexander R. Mankovich[61], Gary I. Saunders[62], Panagiotis I. Sergouniotis[63], Rachel Thompson[58] and Andreas Zankl[64,65,66]

[48]John Wiley & Sons, Hoboken, NJ, USA. [49]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [50]Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [51]Khoo Teck Puat–National University Children's Medical Institute, National University Hospital, Department of Paediatrics, Singapore, Singapore. [52]Provincial Medical Genetics, Women's Hospital of British Columbia, Vancouver, British Columbia, Canada. [53]Institute of Data Science, Maastricht University, Maastricht, The Netherlands. [54]Department of Pediatrics, King Abdulaziz University Hospital, Jeddah, Saudi Arabia. [55]Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA. [56]Paediatric Dermatology, Great Ormond Street Hospital for Children, London, UK. [57]Mosaicism and Precision Medicine Laboratory, Francis Crick Institute, London, UK. [58]Molecular Biomedicine, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada. [59]Brain and Mind Research Institute, Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada. [60]Neuromuscular Centre, The Ottawa Hospital, Ottawa, Ontario, Canada. [61]Precision Diagnosis & Image-Guided Therapy, Philips Research North America, Cambridge, MA, USA. [62]ELIXIR Hub, ELIXIR, Cambridge, UK. [63]Division of Evolution, Infection and Genomics, University of Manchester, Manchester, UK. [64]Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia. [65]Department of Clinical Genetics, The Children's Hospital at Westmead, Westmead, New South Wales, Australia. [66]Kinghorn Centre for Clinical Genomics and Bone Division, Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia.