HAL
open science

Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases

Birte Zurek, Kornelia Ellwanger, Lisenka E L M Vissers, Rebecca Schüle, Matthis Synofzik, Ana Töpf, Richarda M de Voer, Steven Laurie, Leslie Matalonga, Christian Gilissen, et al.

To cite this version:

Birte Zurek, Kornelia Ellwanger, Lisenka E L M Vissers, Rebecca Schüle, Matthis Synofzik, et al.. Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. European Journal of Human Genetics, 2021, 10.1038/s41431-021-00859-0 . hal-03270971

HAL Id: hal-03270971

https://hal.sorbonne-universite.fr/hal-03270971v1

Submitted on 25 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

European Journal of Human Genetics
https://doi.org/10.1038/s41431-021-00859-0

VIEWPOINT

Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases

Birte Zurek[1] · Kornelia Ellwanger[1] · Lisenka E. L. M. Vissers[2,3] · Rebecca Schüle[4,5] · Matthis Synofzik[4,5] · Ana Töpf[6] · Richarda M. de Voer[2,7] · Steven Laurie[8] · Leslie Matalonga[8] · Christian Gilissen[2,7] · Stephan Ossowski[1] · Peter A. C. 't Hoen[7,9] · Antonio Vitobello[10] · Julia M. Schulze-Hentrich[1] · Olaf Riess[1,11] · Han G. Brunner[2,3,12] · Anthony J. Brookes[13] · Ana Rath[14] · Gisèle Bonne[15] · Gulcin Gumus[16] · Alain Verloes[17] · Nicoline Hoogerbrugge[2,7] · Teresinha Evangelista[15] · Tina Harmuth[1] · Morris Swertz[18] · Dylan Spalding[19] · Alexander Hoischen[2,7,20] · Sergi Beltran[8,21,22] · Holm Graessner[1,11] · Solve-RD consortium

Received: 14 October 2020 / Revised: 8 February 2021 / Accepted: 4 March 2021
© The Author(s) 2021. This article is published with open access

Abstract

For the first time in Europe hundreds of rare disease (RD) experts team up to actively share and jointly analyse existing patient's data. Solve-RD is a Horizon 2020-supported EU flagship project bringing together >300 clinicians, scientists, and patient representatives of 51 sites from 15 countries. Solve-RD is built upon a core group of four European Reference Networks (ERNs; ERN-ITHACA, ERN-RND, ERN-Euro NMD, ERN-GENTURIS) which annually see more than 270,000 RD patients with respective pathologies. The main ambition is to solve unsolved rare diseases for which a molecular cause is not yet known. This is achieved through an innovative clinical research environment that introduces novel ways to organise expertise and data. Two major approaches are being pursued (i) massive data re-analysis of >19,000 unsolved rare disease patients and (ii) novel combined -omics approaches. The minimum requirement to be eligible for the analysis activities is an inconclusive exome that can be shared with controlled access. The first preliminary data re-analysis has already diagnosed 255 cases form 8393 exomes/genome datasets. This unprecedented degree of collaboration focused on sharing of data and expertise shall identify many new disease genes and enable diagnosis of many so far undiagnosed patients from all over Europe.

Rare Diseases (RD) are individually rare but collectively a common health issue. Around 80% of RD are estimated to have a genetic cause [1]. The time to a genetic diagnosis however often takes several years and initial clinical diagnoses are incorrect in up to 40% of families [2]. Around 50% of patients with a RD remain undiagnosed even in advanced expert clinical settings where whole exome sequencing (WES) is applied routinely as a diagnostic approach. Depending on the exact diagnostic setting, the inclusion criteria and the type of RD, the diagnostic yield from WES ranges between 15 and 51% of cases [3, 4].

At least two scenarios allow boosting the current yield of WES. Firstly, there is a value in re-analysing WES data regularly [5] and on massive scale [6], but not every RD expert has access to tools enabling this systematically. Secondly, it is clear that moving beyond the exome can provide additional benefits [7, 8].

Solve-RD aims to solve a large number of unsolved RD, for which a molecular cause is not yet known, by implementing both strategies mentioned above. To this end, Solve-RD applies innovative ways to effectively organise expertise and data.

Cohorts

To structure its work Solve-RD has defined four types of *cohorts*. *Cohort 1*, "Unsolved Cases", comprises cases with an inconclusive WES or whole genome sequencing (WGS)

Members of the Solve-RD consortium are listed in Supplementary Information.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41431-021-00859-0.

✉ Holm Graessner
holm.graessner@med.uni-tuebingen.de

Extended author information available on the last page of the article.

Published online: 01 June 2021

Table 1 Examples for the specific ERN cohorts and the unsolvables.

Cohort	Rationale
Cohort 2: Long-read whole genome sequencing (LR-WGS)	
X-linked spinal and bulbar muscular atrophy (SBMA)	Suspected expansions of repeat disorder or other hidden structural variants (SV)
Hereditary ataxia	Suspected expansions of repeat disorder or other hidden SVs
Cohort 2: Genomics and Epigenomics	
Unexplained Intellectual Disability (ID): patient-parent trios	De novo mutation prioritisation very powerful filter for de novo methylation changes
Diffuse gastric cancer	Hypermethylation of cancer gene promoter known disease mechanism
Rare pheochromocytomas and paragangliomas	Hypermethylation of cancer gene promoter known disease mechanism
Cohort 4	
Unsolved syndromes available via ERN ITHACA	Aicardi syndrome, Gomez–Lopez Hernandes syndrome, Hallermann–Streiff syndrome are clinically well-defined entities and have been studied by WES and WGS globally and remain unsolved

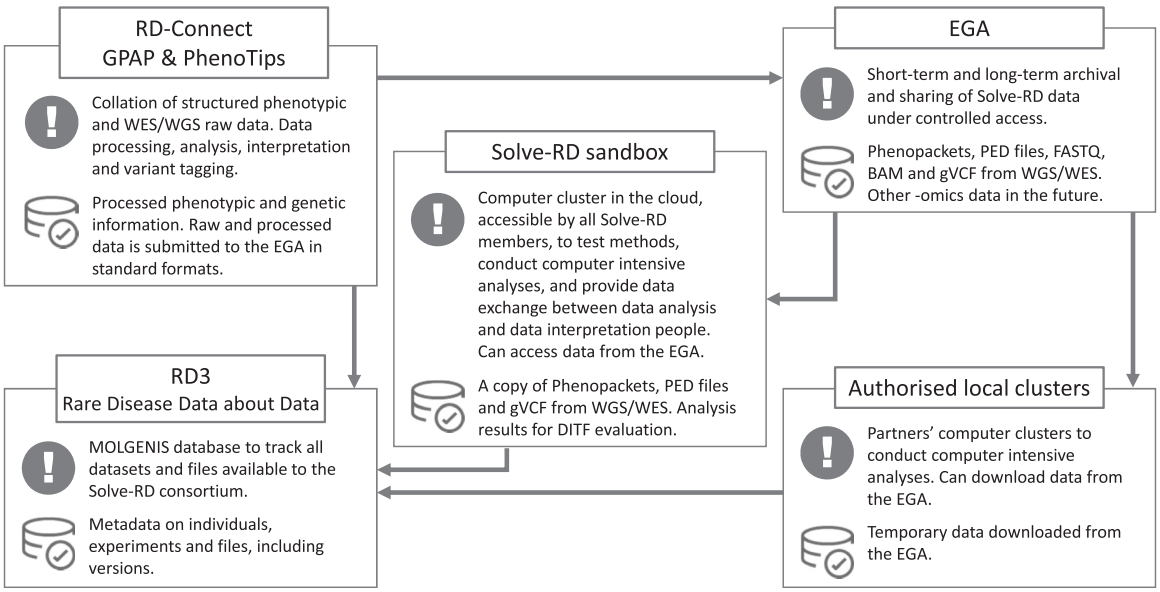

Fig. 1 Solve-RD data infrastructure. Key components of the Solve-RD infrastructure for multi-omics data analysis, illustrating main use and data available.

from any partnering or associated ERN center. These data undergo a comprehensive re-analysis effort. *Cohort 2*, "Specific ERN Cohorts", represent disease group specific ERN cohorts that are analysed by newly applied tailored -omics approaches. *Cohort 3*, "Ultra-Rare Rare Diseases", includes (groups of) patients with unique phenotypes identified (and matched) by RD experts from all ERN participants. For the diseases included in *Cohort 4*, "The Unsolvables", all relevant -omics methodologies will be used to solve highly recognisable, clinically well-defined disease entities for which the disease cause has not been found yet despite considerable previous research investigations including WES and WGS (Table 1).

In total, Solve-RD is targeting to re-analyse >19,000 datasets for cohort 1, sequence ~3500 short- and long-read WGS for cohorts 2, 3, and 4 and add >3500 additional -omics experiments including RNA sequencing, epigenomics, metabolomics, Deep-WES, and deep molecular phenotyping. Data collected and produced in Solve-RD shall be shared via the European Genome-Phenome Archive (EGA) and the RD-Connect Genome-Phenome Analysis Platform (GPAP) to allow controlled access by other RD initiatives and scientists.

Organisation of data

The Solve-RD strategy relies on the availability of large amounts of good quality, standardised genomic and phenotypic data and metadata from undiagnosed RD

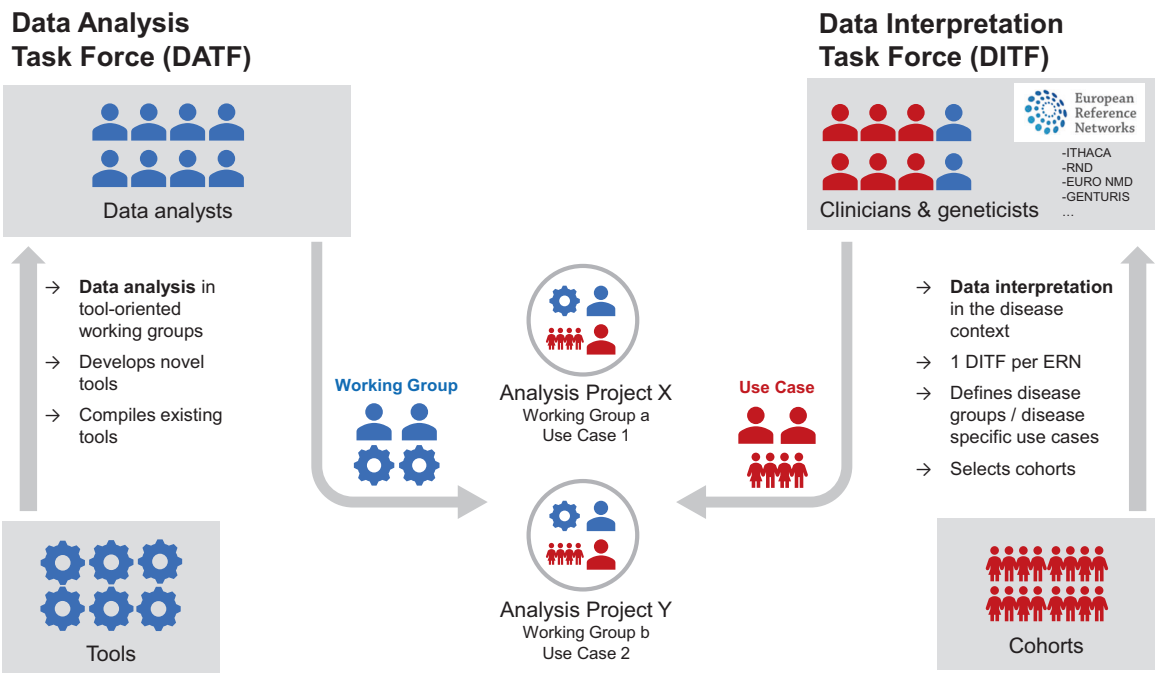

Fig. 2 The Solve-RD data analysis structure 'in action'. Consisting of the Data Analysis Task Force (DATF) and four Data Interpretation Task Forces (DITF)—one per core ERN involved. The DATF established working groups (WGs) for specific analyses. Working groups and DITFs jointly work on analysis projects based on use cases described by the DITF members.

patients and their relatives. Solve-RD follows a centralised approach, to enable all envisioned analyses. Data sharing in Solve-RD is regulated by policy documents, available on the project's website. To overcome the technical challenge of centralising large amounts of data, Solve-RD leverages existing infrastructures such as EGA, GPAP, and computing clusters from project partners (Fig. 1). In addition, Solve-RD is developing a cloud-based computing cluster for collaborative analysis and methods testing (the Solve-RD Sandbox) and a central database to control and view all the project's data and metadata (RD3; rare disease data about data) using the MOLGENIS open source data platform [9]. Clinical data and pedigree structure for all participating individuals is collated through standard terms and ontologies such as HPO, ORDO, and OMIM using GPAP-PhenoStore. To share data within the project and beyond, Solve-RD is an early adopter of the recently GA4GH-approved (Global Alliance for Genomics and Health, https://www.ga4gh.org) PhenoPackets standard to enable exchange of phenotypic and family information [10].

For each individual, WES and/or WGS data are submitted to GPAP in FASTQ, BAM, or CRAM format. The sequencing data are processed through a standard pipeline based on GATK (Genomic Analysis Toolkit variant calling software) best practices [11, 12]. After that, PhenoPackets, PED files (for pedigrees), raw data (FASTQ), alignments (BAM) and genetic variants (gVCF) are transferred to the EGA, where they are archived and made available to the consortium (and later on to the broader RD community) for further analysis. Furthermore, Solve-RD data are connected to MatchMaker Exchange via GPAP.

To reach the ambitious goal to collect 19,000 unsolved WES/WGS, Solve-RD has defined several deadlines to submit data to the project. After each deadline, all data are processed and released as a data freeze, which is amenable to corrections via patches. The first data freeze, released in early 2020, includes data from 8,393 individuals.

In parallel to the collection of existing data for cohort 1, new omics data are being generated for cohorts 2, 3, and 4. A common data workflow has been established for all these data types (Fig. 1). The data collated and generated by Solve-RD constitutes a unique collection that will be valuable beyond the project, and the consortium is committed to make it FAIR under controlled access, through the EGA and GPAP.

Organisation of expertise

Solve-RD works on the interphase of many disciplines relevant to solving the unsolved RD. Central to the RD field are clinical geneticists and clinical scientists organised in the respective ERNs. Solve-RD provides expertise in genomics and other -omics data analysis, through data scientists, molecular geneticists, and bioinformaticians.

To warrant the best exchange of expertise we have implemented two structures: (i) Data scientists and genomics experts are organised in a Data Analysis Task Force

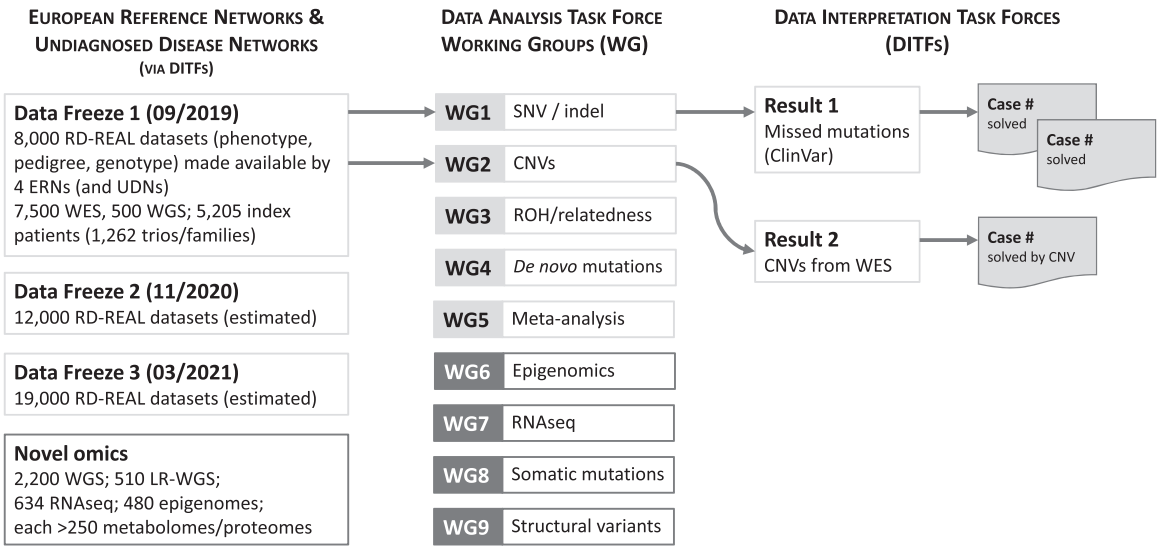

Fig. 3 Organisation of new result flow in Solve-RD. Working groups (WG) 1–5 will re-analyse existing sequencing data. Novel omics data will be analysed by all working groups (as appropriate). RD-REAL refers to Rare Disease - REAnalysis Logistics.

(DATF), (ii) Expert clinicians and geneticists from each ERN are organised in a Data Interpretation Task Force (DITF) (Fig. 2). The tasks for these structures are in brief: ►DITF: define needs of ERN for (a) data re-analysis and (b) novel -omics data; define use cases for re-analysis and novel analysis; discuss/test suitable data output formats for clinical scientists; coordinate collaborative data interpretation; discuss within respective ERN network and feedback to DATF. ►DATF: map expertise in Solve-RD and all (ERN-)partners; create *Analysis Projects* (Supplementary Table S1) based on ERNs needs; develop state-of-the-art analysis tools; analyse data: (a) data re-analysis and (b) novel omics data; optimise data sharing and output formats for DITF/ERNs.

The structure implemented for data re-analysis has proven efficient and versatile [13], and will therefore be applied for novel omics data analysis, with additional working groups for specific -omics technologies (Fig. 3).

To integrate expertise not available within the Solve-RD consortium, particularly with regards to molecular and functional validation of newly found genes, Solve-RD is implementing an innovative brokerage system (Rare Disease Models and Mechanisms Network—Europe (RDMM-Europe)) that has already been successfully used in Canada [14]. As of 4 December 2020, 14 "brokering" Seeding Grants have been awarded to external model investigators.

Achievements and challenges

The work of the first 3 years of Solve-RD resulted in a practical solution to share and jointly analyse 8393 datasets from all over Europe: Solve-RD organised RD expertise via a DITF and DATF with the respective working group structure described above. The first re-analysis approaches resulted in 255 newly diagnosed cases, mainly by leveraging latest ClinVar entries. As examples we refer to adjacent articles, published jointly in this issue [13, 15–18]. Many more candidate variants and new analysis results are under evaluation.

To achieve its current status Solve-RD has successfully addressed some critical challenges that are (a) European data sharing in accordance with GDPR, (b) heterogeneity in existing WES data (e.g. 26 WES kits so far; multiple sequencing platforms), (c) implementing a centralised analysis approach and (d) addressing the rarity of events.

It is the vision of Solve-RD that, by the end of the project, the Solve-RD dataset will be the largest well-annotated, standardised, multi-omics RD dataset on the diseases covered by the four core ERNs. In this sense, we hope that the Solve-RD dataset will be as useful to the RD community as the gnomAD consortium is for the genomics community [19], by making -omics data of RD populations available to the community.

Acknowledgements The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257. This research is supported (not financially) by four ERNs: (1) The ERN for Intellectual Disability, Telehealth and Congenital Anomalies (ERN-ITHACA)—Project ID No 869189; (2) The ERN on Rare Neurological Diseases (ERN-RND)—Project ID No 739510; (3) The ERN for Neuromuscular Diseases (ERN Euro-NMD)—Project ID No 870177; (4) The ERN on Genetic Tumour Risk Syndromes (ERN GENTURIS)—Project ID No 739547. The ERNs are co-funded by the European Union within the framework of the Third Health Programme.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Hartley T, Lemire G, Kernohan KD, Howley HE, Adams DR, Boycott KM. New diagnostic approaches for undiagnosed rare genetic diseases. Annu Rev Genomics Hum Genet. 2020;21:351–72.
2. EURORDIS AKFF. The Voice of 12,000 Patients. Experiences and Expectations of Rare Disease Patients on Diagnosis and Care in Europe. Eurordis; Paris, France; 2009.
3. Smith HS, Swint JM, Lalani SR, Yamal JM, de Oliveira Otto MC, Castellanos S, et al. Clinical application of genome and exome sequencing as a diagnostic tool for pediatric patients: a scoping review of the literature. Genet Med. 2019;21:3–16.
4. Wise AL, Manolio TA, Mensah GA, Peterson JF, Roden DM, Tamburro C, et al. Genomic medicine for undiagnosed diseases. Lancet. 2019;394:533–40.
5. Liu P, Meng L, Normand EA, Xia F, Song X, Ghazi A, et al. Reanalysis of clinical exome sequencing data. N. Engl J Med. 2019;380:2478–80.
6. Kaplanis J, Samocha KE, Wiel L, Zhang Z, Arvai KJ, Eberhardt RY, et al. Integrating healthcare and research genetic data empowers the discovery of 28 novel developmental disorders. bioRxiv. 2020. https://doi.org/10.1101/797787.
7. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature. 2018;555:611–6.
8. Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun. 2017;8:15824.
9. van der Velde KJ, Imhann F, Charbon B, Pang C, van Enckevort D, Slofstra M, et al. MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. Bioinformatics. 2019;35:1076–8.
10. Zhao M, Havrilla JM, Fang L, Chen Y, Peng J, Liu C, et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. NAR Genom Bioinform. 2020;2:lqaa032.
11. Laurie S, Fernandez-Callejo M, Marco-Sola S, Trotta JR, Camps J, Chacón A, et al. From wet-lab to variations: concordance and speed of bioinformatics pipelines for whole genome and whole exome sequencing. Hum Mutat. 2016;37:1263–71.
12. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
13. Matalonga L, Hernández-Ferrer C, Piscia D, Solve-RD SNV-indel working group, Vissers LELM, Schüle R, et al. Diagnosis of rare disease patients through programmatic reanalysis of genome-phenome data. Manuscript submitted to EJHG (703-20-EJHG).
14. Boycott KM, Campeau PM, Howley HE, Pavlidis P, Rogic S, Oriel C, et al. The Canadian Rare Diseases Models and Mechanisms (RDMM) Network: Connecting Understudied Genes to Model Organisms. Am J Hum Genet. 2020;106:143–52.
15. de Boer E, Ockeloen CW, Matalonga L, Horvath R, Solve-RD SNV-indel working group, Rodenburg RJ, et al. A pathogenic MT-TL1 variant identified by whole exome sequencing in an individual with unexplained intellectual disability, epilepsy and spastic tetraparesis. Manuscript submitted to EJHG (699-20-EJHG).
16. Schüle R, Timmann D, Erasmus CE, Reichbauer J, Wayand M, van de Warrenburg BPC, et al. Common pitfalls in genetic diagnosis of rare neurological diseases. Manuscript submitted to EJHG (705-20-EJHG).
17. Töpf A, Pyle A, Griffin H, Matalonga L, Schon K, Solve RD SNV indel working group, et al. Exome reanalysis and proteomic profiling identified TRIP4 as a novel cause of cerebellar hypoplasia and spinal muscular atrophy (PCH1). Manuscript submitted to EJHG (700-20-EJHG).
18. te Paske I, Garcia-Pelaez J, Sommer AK, Matalonga L, Starzynska T, Jakubowska A, et al. A Mosaic PIK3CA Variant in a Young Adult with Diffuse Gastric Cancer: Case Report. Manuscript submitted to EJHG (704-20-EJHG).
19. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.

Affiliations

Birte Zurek[1] · Kornelia Ellwanger[1] · Lisenka E. L. M. Vissers[2,3] · Rebecca Schüle[4,5] · Matthis Synofzik[4,5] · Ana Töpf[6] · Richarda M. de Voer[2,7] · Steven Laurie[8] · Leslie Matalonga[8] · Christian Gilissen[2,7] · Stephan Ossowski[1] · Peter A. C. 't Hoen[7,9] · Antonio Vitobello[10] · Julia M. Schulze-Hentrich[1] · Olaf Riess[1,11] · Han G. Brunner[2,3,12] · Anthony J. Brookes[13] · Ana Rath[14] · Gisèle Bonne[15] · Gulcin Gumus[16] · Alain Verloes[17] · Nicoline Hoogerbrugge[2,7] · Teresinha Evangelista[15] · Tina Harmuth[1] · Morris Swertz[18] · Dylan Spalding[19] · Alexander Hoischen[2,7,20] · Sergi Beltran[8,21,22] · Holm Graessner[1,11] · Solve-RD consortium

[1] Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany

[2] Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands

[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands

[4] Department of Neurodegeneration, Hertie Institute for Clinical Brain Research (HIH), University of Tübingen, Tübingen, Germany

[5] German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany

[6] John Walton Muscular Dystrophy Research Centre, Translational and Clinical Research Institute, Newcastle University and Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

[7] Radboud Institute for Molecular Life Sciences, Nijmegen, The Netherlands

[8] CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

[9] Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands

[10] Inserm—University of Burgundy-Franche Comté, Dijon, France

[11] Centre for Rare Diseases, University of Tübingen, Tübingen, Germany

[12] Department of Clinical Genetics, Maastricht University Medical Centre, Maastricht, The Netherlands

[13] Department of Genetics and Genome Biology, University of Leicester, Leicester, UK

[14] INSERM, US14—Orphanet, Plateforme Maladies Rares, Paris, France

[15] Sorbonne Université, INSERM UMRS 974, Center of Research in Myology, Paris, France

[16] EURORDIS-Rare Diseases Europe, Barcelona, Spain

[17] Genetics Department, APHP-Robert Debré University Hospital, Université de Paris, Paris, France

[18] Department of Genetics, Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

[19] European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridge, UK

[20] Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, The Netherlands

[21] Universitat Pompeu Fabra (UPF), Barcelona, Spain

[22] Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain