Deliverable

D3.6 Publication on strategy for cohort development in omics studies	
Version | Status	V1 | final
Work package	WP3
Lead beneficiary	INSERM
Due date	31.12.2022 (M60)
Date of preparation	29.04.2024
Target Dissemination Level	Public
Author(s)	Kornelia Ellwanger (EKUT), Birte Zurek (EKUT)
Reviewed by	Holm Graessner (EKUT)
Approved by	Gisèle Bonne (INSERM)

The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257.

Explanation according to GA Annex I:

Publication on strategy for cohort development in omics studies.

Part of the data presented in this deliverable report has been published by Zurek et al. in Eur J Hum Genet. 2021 Sep;29(9):1325-1331. doi: 10.1038/s41431-021-00859-0.

Abstract:

Solve-RD aims to find a diagnosis for rare disease patients who did not get a molecular diagnosis yet. In addition to research re-analysis of existing exome and genome data, we used latest -omics technologies isolated or in combination in bespoke cohorts to solve the unsolved diseases and discover the underlying disease mechanisms.

To this end, bespoke rare disease cohorts provided by the European Reference Networks (ERNs) were analysed by novel (multi) omics tools that go beyond the exome/genome.

This report describes the applied Solve-RD strategy for cohort development in omics studies.

Introduction

Solve-RD has defined four types of cohorts to differentiate patients and diseases based on the data already available, the data generated in the project and the analysis approaches used to solve them (see Figure 1).

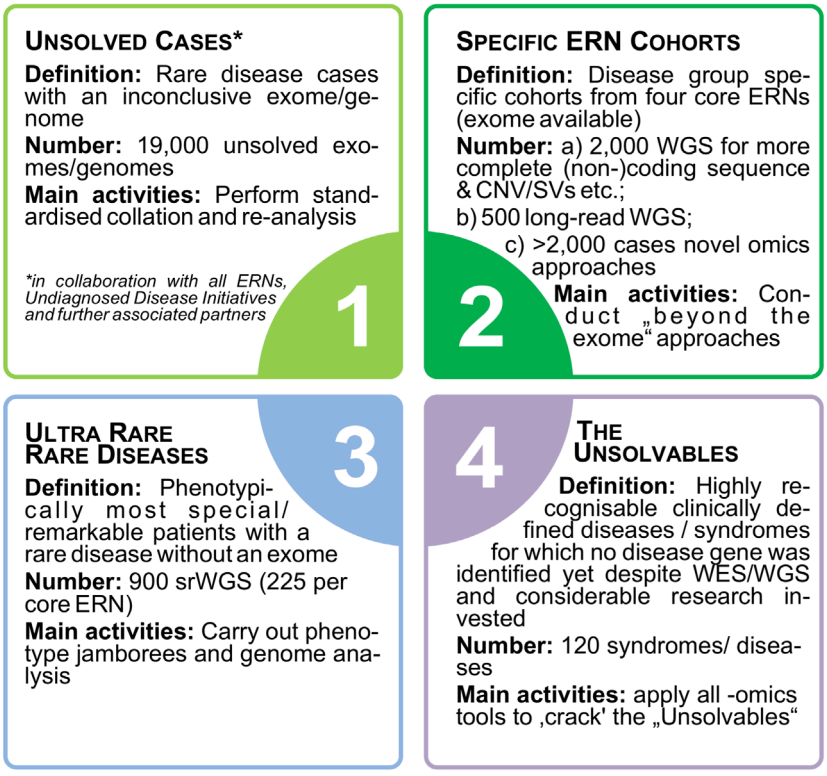


Figure 1: The four Solve-RD cohorts.

Cohort 1, "Unsolved Cases", comprises cases with an inconclusive sequenced exome or genome from any partnering or associated ERN center. These data underwent a comprehensive

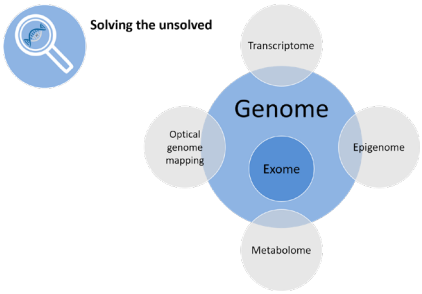
reanalysis. Cohort 2, “Specific ERN Cohorts”, represent disease group specific ERN cohorts that have been analysed by newly applied tailored -omics approaches. Cohort 3, “Ultra-Rare Rare Diseases”, includes (groups of) patients with unique phenotypes identified (and matched) by rare disease experts from all ERN participants. For the diseases included in Cohort 4, “The Unsolvables”, all relevant -omics methodologies have been 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.

Cohort formation

Rationale and inclusion of cases

Latest omics technologies isolated or in combination to solve the unsolved diseases (included in cohorts 2-4) and to discover the underlying disease mechanisms will be applied and **go beyond the exome/genome**. Innovative omics tools will be applied in order to ‚solve the unsolvable syndromes‘.

***From exomes to genomes*:**

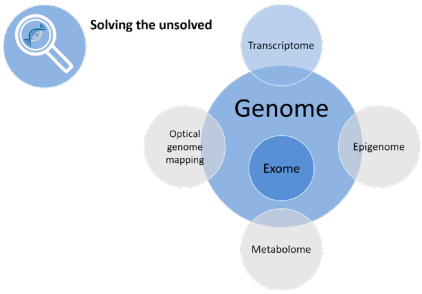


For 2,116 cases, Solve-RD has applied short-read WGS to achieve a more complete coding sequence for all cases, better CNV/SV detection and identification of non-coding mutations in functional elements. For hidden SVs, 519 cases were selected for long-read WGS.

For most disease cohorts, trio-based WGS was the method of choice. For unsolved diseases/cases with clear recessive inheritance two affected individuals per family have been analysed (if available).

730 samples have been collected and analysed by deep WES to identify disease-causing somatic mutations.

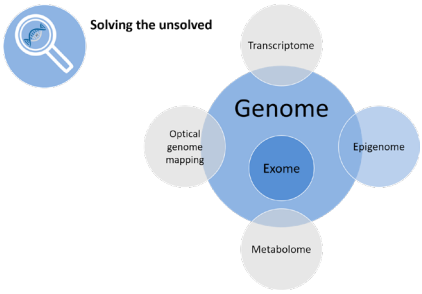
***Genomics and transcriptomics*:**



For 566 cases, short-read RNA sequencing has been applied. We have selected disorders for which we have affected tissues available and/or previous work has indicated RNA as a suitable read-out: Neuromuscular diseases (muscle and fibroblast – strong focus on myopathies/dystrophies and mitochondrial disorders), Ataxia (brain/cerebellum, fibroblasts; iPSC; PBMCs), (colorectal) cancer cases (fresh normal vs cancerous colon mucosa). The affected tissue work has been supplemented by transcriptome work from PBMC.

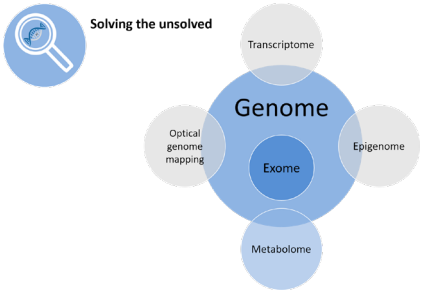
78 samples have been analysed by high coverage long-read RNA sequencing. Long-reads facilitate the identification of abnormal splicing effects and identification of full-isoforms available in a respective sample.

***Genomics and epigenomics*:**



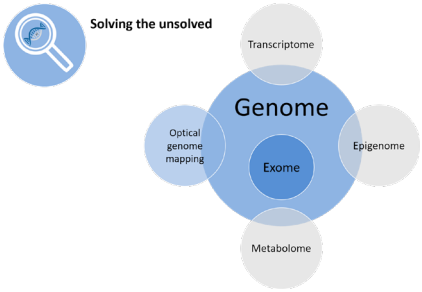
For specifically selected disorders from different ERNs (343 samples from 338 individuals) we have applied epigenome analysis by reduced representation bisulfite sequencing (RRBS) in two approaches. Initially, we identified differentially methylated CpGs and regions across various syndromes. Subsequently, we annotated significant CpGs and conduct functional enrichment analyses to unveil altered biological pathways. Conversely, the second approach can identify outlier methylation events unique to each patient, with the ultimate goal of pinpointing causal mutations.

***Genomics and metabolomics*:**



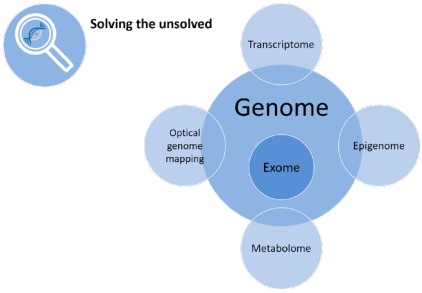
Metabolomic screens of blood-plasma of patients with undiagnosed ID and suspected metabolic disorders was performed (36 cases in total).

***Genomics and optical genome mapping (OGM)*:**



Optical genome mapping (OGM) was applied in selected patients that remained undiagnosed after standard-of-care analysis in the respective RD-expert centers and where evidence was provided that a hidden structural variant of >300 bp is the likely missed disease cause. A total of 209 samples from 90 families were enrolled.

***Multi-omics to solve the unsolvable syndromes*:**



The hypothesis was that unsolvable syndromes remained unsolved because they are caused by novel mechanisms and allow to unravel novel paradigms.

Therefore, we aimed to unravel these by integrated multi-omics approaches, combining all available omics tools. We have included the following diseases/disease groups: Pai syndrome, Oculoauriculofrontonasal syndrome (OAFNS), Gomez Lopez Hernandez syndrome, Aicardi syndrome and Hallerman-Streiff syndrome (ERN-ITH-ACA), axial myopathies, contractural phenotypes, OPDM, suspected titinopathies, suspected AD

calpainopathies and Inclusion-Body Myositis (IBM) (ERN-Euro NMD) and Hereditary Geniospasm (ERN-RND).

Identification of ultra-rare rare diseases: The rationale was to identify patients with unusual unique phenotypes which are likely to represent novel ultra-rare RD entities. This would then allow to perform a matchmaking with other patients with the same unusual phenotype and to perform a joint (re)analysis of the existing exome and/or a new genome. Each ERN used their own distinct strategy to identify those ultra-rare cases (also see deliverable report D2.8).

Exemplary subcohorts

The multi-omics approaches within the "Specific ERN Cohorts" (Cohort 2) and the ERN-specific "Unsolvables" (Cohort 4) are tailored to specifically address state of the omics-techniques towards the problems per disease utilizing the unique phenotypes identified by the clinical experts of each ERN. In addition, the ERNs serve as unique multi-tissue-sources for the respective analyses. Examples of the "Specific ERN Cohorts" (Cohort 2) and the ERN-specific "Unsolvables" are given in **Table 1**.

Table 1: Examples for the specific ERN cohorts and the unsolvables (cohort 2 and 4, respectively).

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

Results of sample collection and current state of analysis for bespoke Solve-RD cohorts

The following table (**Table 2**) provides an overview on the number of samples collected and analysed using the different omics technologies as well as the results. The number of solved cases is based on the status of cases in PhenoStore at the time of preparing this deliverable

and represents the results of multiple different efforts. For some omics, e.g. Epigenome analysis, solved cases (solved at the time of submission) were included in the analysis.

Table 2: Overview of novel omics samples.

Omics	Number of samples	Number of individuals	Number of index cases	Solved (index) cases
Short-read genomes	2,116	1,973	887	100
Long-read genomes	519	473	181	22
Deep-WES	730	412	403	8
Short-read RNAseq	566	470	319	41
Long-read RNAseq	78	25	20	4
Epigenomes	343	338	252	125[1]
Metabolomes	36	36	13	0
OGM	209	204	90	3

Conclusion:

The concept of matching clinically well-defined RD cohorts with state-of-the-art omics techniques or multi-omics approaches was successful in term of sample collection and implementation of analysis and interpretation. As analysis efforts are still ongoing, a conclusive statement on the contribution of tailored omics techniques or multi-omics approaches to diagnostic yield can, however, not done yet at present.

The identification of GAA repeat expansions in *FGF14* serving as a common cause of so far unsolved ataxia cases[2] or a very recent report demonstrating that a single variant in a non-coding RNA explains ~0.5% of all undiagnosed individuals with neurodevelopmental disorders (NDD)[3] provide, however, promising examples of success of this approach.

[1] Epigenome analysis included samples from families that were already solved at the time of submission.

[2] David Pellerin et al. Deep Intronic FGF14 GAA Repeat Expansion in Late-Onset Cerebellar Ataxia. N Engl J Med. 2023 Jan 12;388(2):128-141. doi: 10.1056/NEJMoa2207406. Epub 2022 Dec 14.

[3] Yuyang Chen et al. De novo variants in the non-coding spliceosomal snRNA gene RNU4-2 are a frequent cause of syndromic neurodevelopmental disorders. medRxiv 2024.04.07.24305438; doi: https://doi.org/10.1101/2024.04.07.24305438