Deliverable

D2.2 Guidelines for exome/genome re-analysis	
Version | Status	V2 | final
Work package	WP2
Lead beneficiary	CNAG-CRG (Sergi Beltran)
Due date	30.06.2019 (M18)
Date of preparation	15.05.2024
Target Dissemination Level	Public
Author(s)	Leslie Matalonga, Carles Garcia, Sergi Beltran (all CNAG-CRG), Christian Gilissen (RUMC)
Reviewed by	Christian Gilissen (RUMC), Stephan Ossowski (EKUT)
Approved by	Alexander Hoischen (RUMC)

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:

Guidelines for exome/genome re-analysis provided by the Data Analyses Task Force.

Abstract:

Over 5,000 genomic datasets have been collected by Solve-RD, processed using a standardized analysis pipeline (Laurie et al., 2016) and made available to all Solve-RD partners through the RD-Connect GPAP and the Solve-RD Sandbox.

To organise the re-analysis and interpretation of the data we have pulled together the consortium's bioinformatics and analysis expertise in a Data Analysis Task Force (DATF) and the clinical and biological expertise in 4 Data Interpretation Task Forces (DITFs), one per core ERN.

Solve-RD data analysis task force (DATF) has managed to setup a semi-automatized workflow for genomic data (re)analysis and interpretation involving partners and expertise from the different WPs and ERNs. Different working groups and use cases have been set up and largely documented to provide guidance to Solve-RD partners on how the (re)analysis of their data will be performed and which are the steps required for validation and feedback.

Introduction:

Over 5,000 datasets (mostly exomes, but also some genomes) integrated with their phenotypic information have been collected and processed by Solve-RD (deliverable 2.6). Genomic data is processed using a standardized analysis pipeline (Laurie et al., 2016) and made available in two ways to all Solve-RD partners:

1) In the RD-Connect Genome Phenome Analysis Platform (GPAP, https://platform.rd-connect.eu) through its API (Application Programming Interface) and its GUI (Graphical User Interface), with many functionalities and sources of information.

2) As files (FASTQ, BAM, gVCF, Phenopackets, etc), which are being transferred to the Solve-RD Sandbox and the EGA (workflow currently being tested).

To organise the re-analysis and interpretation of the data we have pulled together the consortium's bioinformatics and analysis expertise in a Data Analysis Task Force (DATF) and the clinical and biological expertise in 4 Data Interpretation Task Forces (DITFs), one per core ERN: NMD, ITHACA, RND and GENTURIS.

Each of the DITFs has proposed several Use Cases and approaches to try to reach a molecular diagnosis on their unsolved exomes and genomes, either at the individual level or by (re)analysing (sub-)cohorts based on phenotypic information encoded in HPO or ORDO terms. Together with the DATF, the Use Cases have been classified according to similarities in the technical approach to be followed and 5 Working Groups have been created: Relatedness and Runs of Homozigosity (lead: EKUT), SNV and InDel standard filtering (lead: CNAG-CRG), CNV analysis (lead: CNAG-CRG), De novo trio analysis (lead: RUMC) and Meta-analysis (lead: RUMC).

Report:

Over 5,000 genomic datasets have been collected by Solve-RD, processed using a standardized analysis pipeline (Laurie et al., 2016). We are now in the process of making all datasets available to all Solve-RD partners through the RD-Connect GPAP and the Solve-RD Sandbox.

The standard GPAP analysis pipeline described in Laurie et al., 2016 (PMID: 27604516) is summarized in Figure 1. Briefly, BAM or fastq sequencing files are processed using different

public software and applications to map all the reads (BWA-MEM), to remove their duplicates (PICARD), to realign Indels (GATK) and implement a base quality score recalibration (GATK). Finally, a variant calling program is used to obtain a gVCF file with all the genomic variants (GATK Haplotype Caller 3.6). Quality control is performed using FastQC.

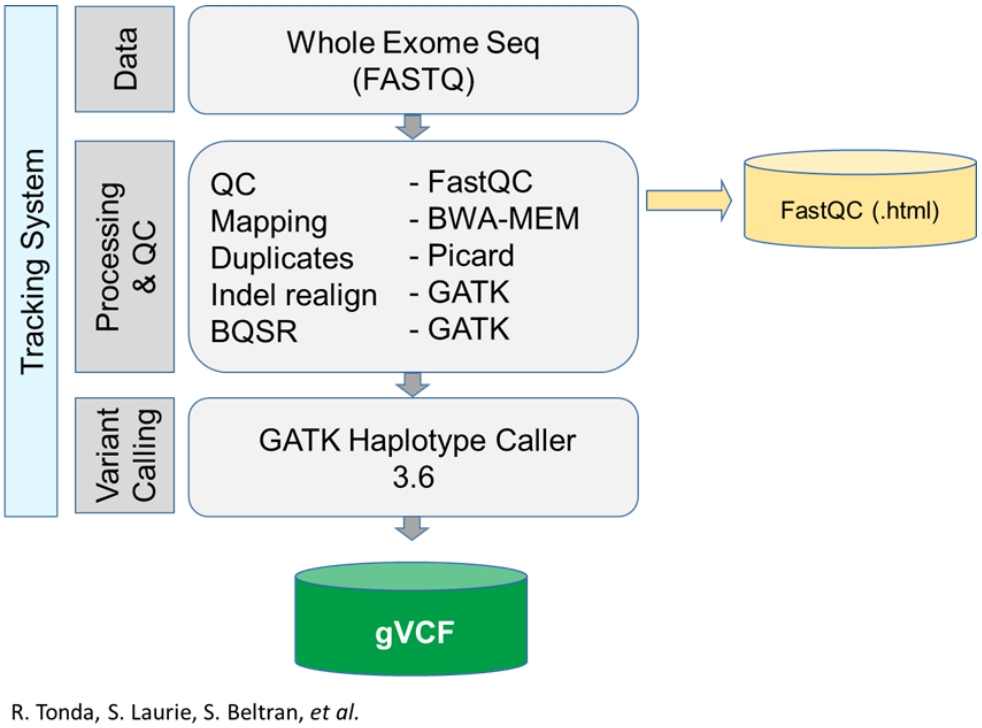

Figure 1: The standard GPAP analysis pipeline.

All gVCFs are then merged, joint genotyped and annotated as part of the uploading process to the RD-Connect GPAP. During this first part of the project we have updated the annotation system and have moved to Ensembl VEP, adding some additional information from gnomAD and ClinVar through dbSNFP. The final gVCF file is then ready to be analysed by RD-Connect GPAP users.

Guidelines for analysis and interpretation of variants with the RD-Connect GPAP are available as videos (https://rd-connect.eu/videos) and can be tested by the users on the Playground (https://platform.rd-connect.eu/#/playground), with many of the different examples provided.

In addition, we have conducted webinars (https://www.youtube.com/@solve-rd4075/videos) and a hands-on workshop at the Solve-RD Annual Meeting (February 8th, 2019, Nijmegen, the Netherlands).

The data is also available to the DATF Working Groups to proceed with the agreed down-stream analyses according to the use cases proposed by each of the Solve-RD core ERNs.

Based on these Use Cases, each of the Working Groups has drafted an analysis plan and roadmap. These plans have been approved by the DATF and DITFs.

Candidate variants identified by the DATF working groups will be transferred to the DITF (Data Interpretation Task Force) groups. DITFs will evaluate these candidate variants, and contact the original submitter for validation. This workflow will determine if the case can be considered as “solved”, or if it has to be re-evaluated by the DATF groups (Figure 2).

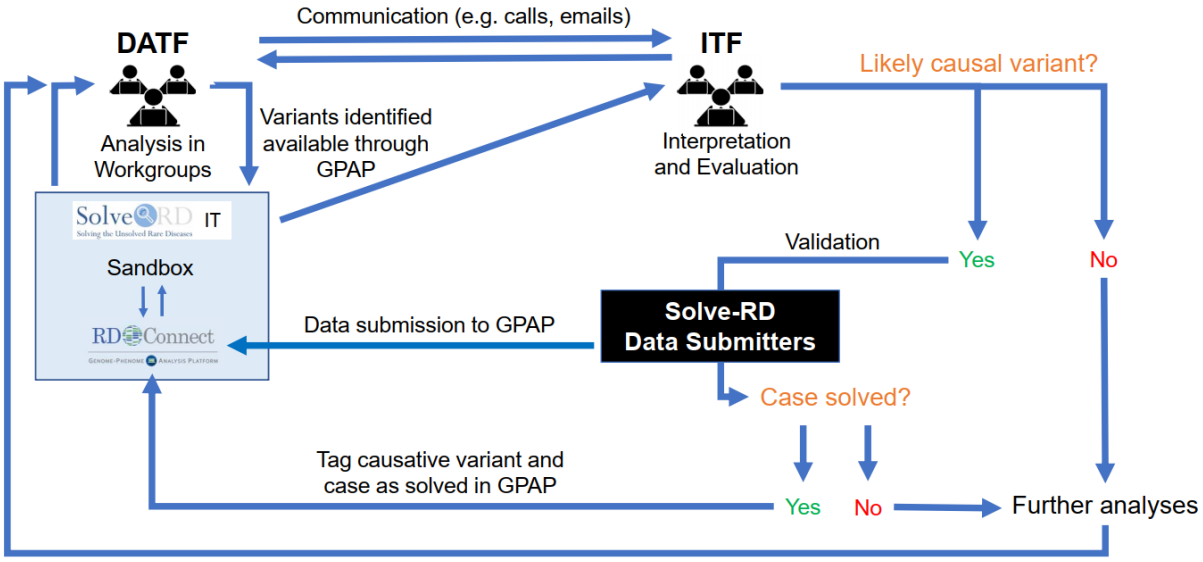

Figure 2: Workflow between Solve-RD data submitters, DATF and DITF.

At this stage, the following studies are being performed by DATFs working groups:

- ***Relatedness and Runs of Homozigosity*** (lead: EKUT). RoHs are being identified with PLINK as part of the standard pipeline, and a protocol to compute relatedness is being developed with the same software. In addition, new tools developed by EKUT (RoHhunter, SampleSimilarity) are being tested and benchmarked.
- ***SNV and InDel standard filtering*** (lead: CNAG-CRG). An automated tiered identification and tagging of likely pathogenic variants through the RD-Connect API is being developed. Very preliminary results indicate 2-6% of cases might be solved with this approach. There is the objective to also run Exomiser on all available datasets. In this sense, in a small group of patients suffering Congenital Myasthenic Syndromes (CMS) we have shown that increasing phenotypic annotation with HPO improves Exomiser's prioritization of the causative variant (Thompson et al. 2019, PMID:31231902).
- ***CNV analysis*** (lead: CNAG-CRG). The extensive expertise available in the consortium is allowing us to test several tools, taking advantage of the high number of datasets generated with the same exome capture kit (even if different disease types are included). These tools include ExomeDepth, CONIFER, ClinCNV and ClusterWES.
- ***De novo trio analysis*** (lead: RUMC): trio information is being collected, sample quality control is being performed, and a first overview of de novo tools for testing is being setup. This currently includes DeNovoGear and RUMC de novo tool.
- ***Meta-analysis*** (lead: RUMC): samples are being collected with the required meta-data, quality control and coverage analyses are being performed. Analysis of variant overlap will be performed first, de novo enrichment second, burden analysis third.

As an example, Figure 3 below describes the workflow of one of the approaches being followed by the SNV and InDel variant automatic filtration Working Group.

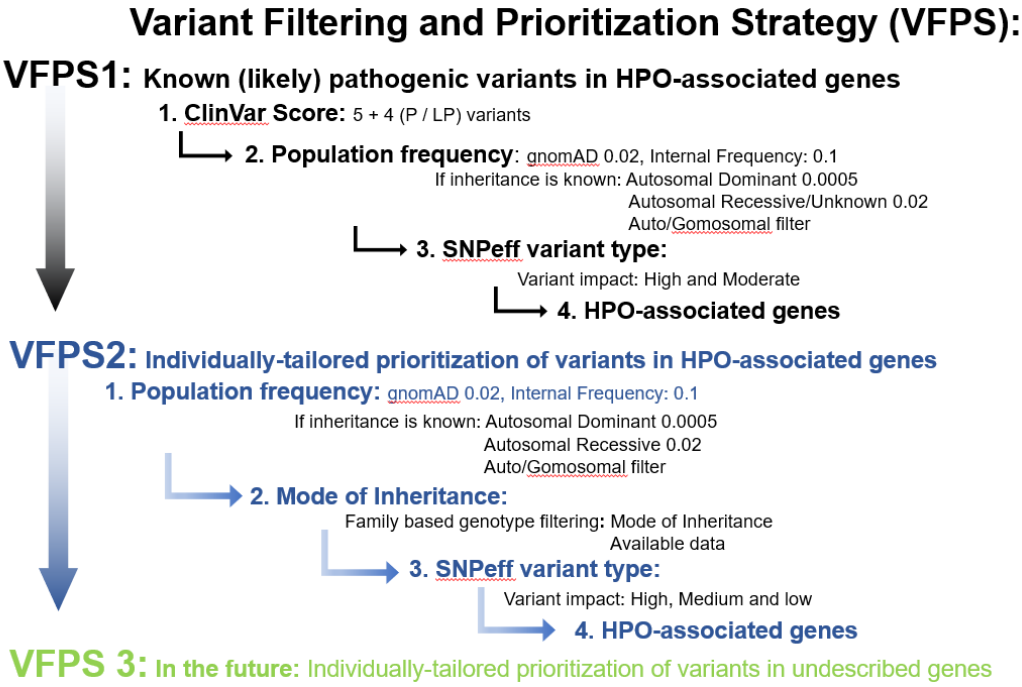

Figure 3: Workflow of one of the approaches followed by the SNV and InDel variant automatic filtration working group.

Monthly calls with all DATF members have been established. Each of the DITF has also established regular calls, and so have each of the working groups involving representatives from the DATF and each of the ERNs. Finally, overarching coordination is conducted in calls including DATF leads, DITF leads and Solve-RD coordination. Altogether, these calls are used to update members on the on-going work and to facilitate communication between partners involved in the project.

Conclusion:

Solve-RD data analysis task force (DATF) has managed to setup a semi-automatized workflow for genomic data (re)analysis and interpretation involving partners and expertise from the different WPs and ERNs. Different working groups and use cases have been set up and largely documented to provide guidance to Solve-RD partners on how the (re)analysis of their data will be performed and which are the steps required for validation and feedback. Training videos and example cases are available to guide users in the analysis and interpretation of data with the GPAP. In addition, we have conducted webinars and a hands-on workshop during the Solve-RD Annual Meeting.