SOLVING THE UNSOLVED RARE DISEASE CASES

BY WORKING TOGETHER IN A SHARED ANALYSIS INFRASTRUCTURE WITH FLEXIBLE VARIANT CALLING AND INTERPRETATION TO ACCOMMODATE FOR SPECIFIC RESEARCH QUESTIONS

Johansson L, Been G, Hendriksen D, Charbon C, Ruvolo D, Neerincx P, van de Geijn G-J, van der Velde J , Swertz M.

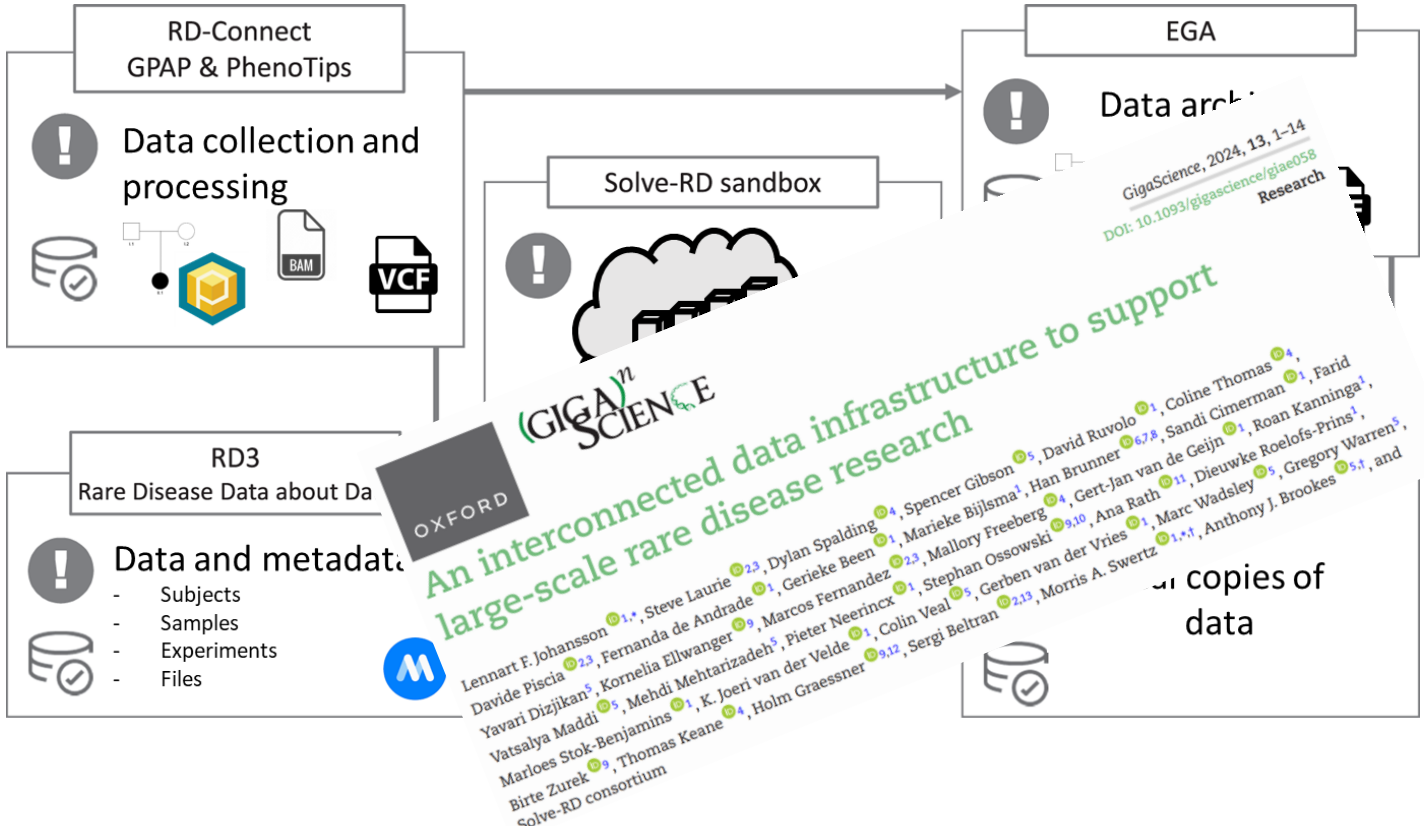
University Medical Center Groningen, The Netherlands

European Reference Networks

ERICA

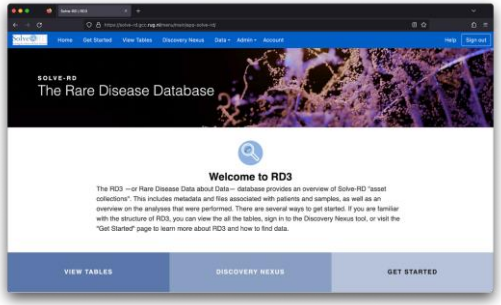
European Rare Disease Research Coordination and Support Action

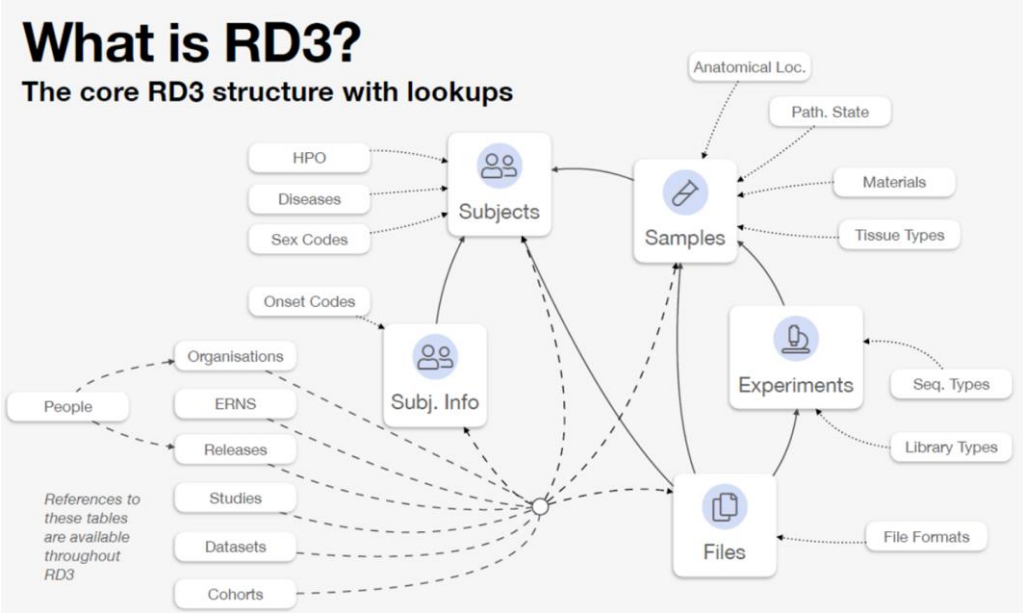
ERICA ERN Research conference 12 december 2024



Adapted from: Zurek, B., Elwanger K. Vissers, L.E.L.M.., Schüle, R. Synofzik M., et al. Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. *EJHG*. 2021:29:1325-1331. Figure 1.

RD3 database

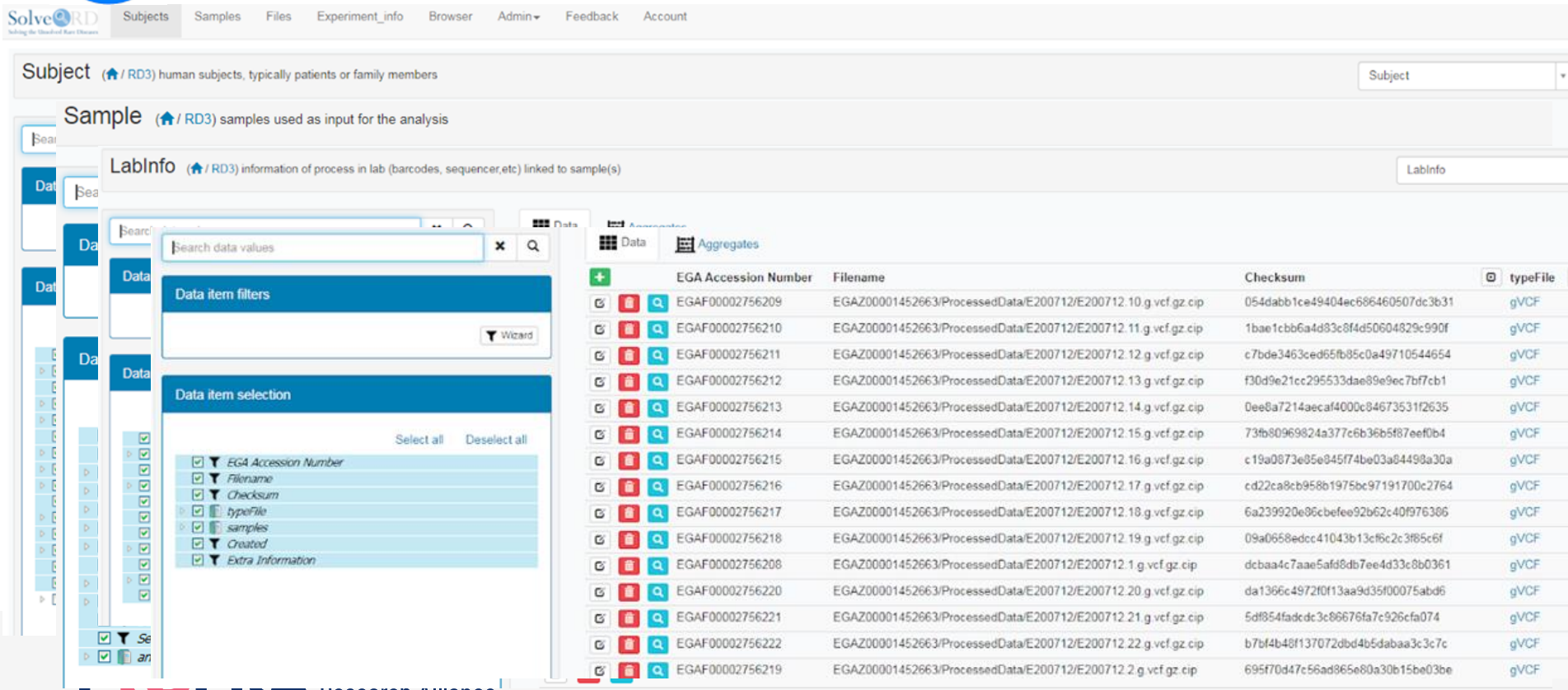




The Rare Disease Data about Data (RD3) database

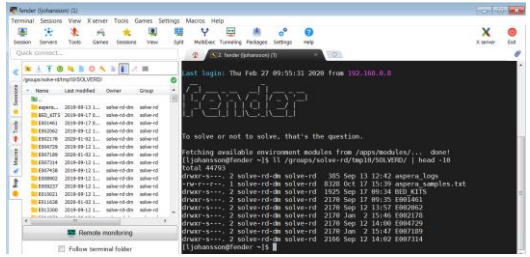
enables researchers and clinicians to find patients, samples, experiments and data files based on their characteristics.

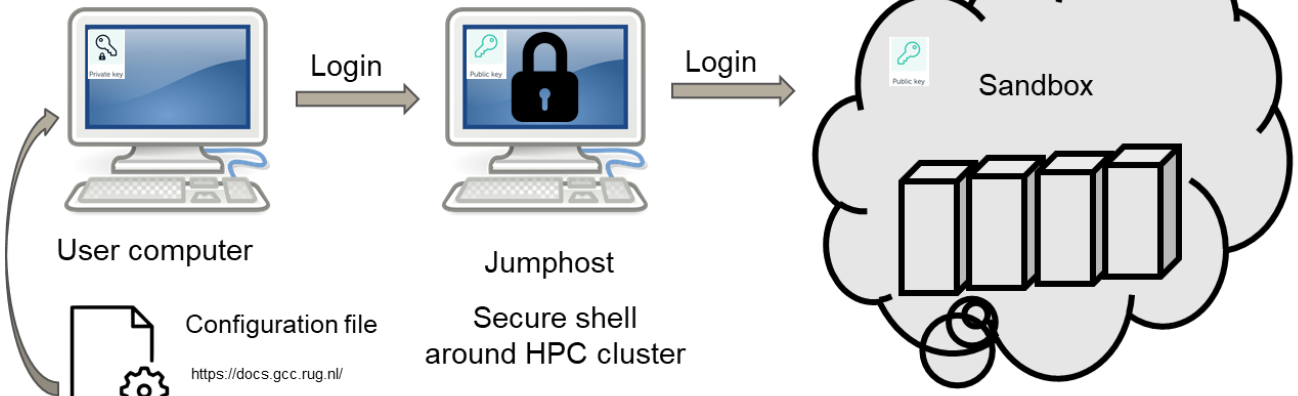
RD3 database

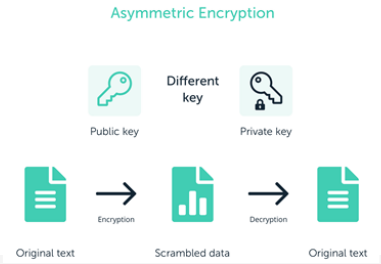


Research Alliance

Analytical Sandbox



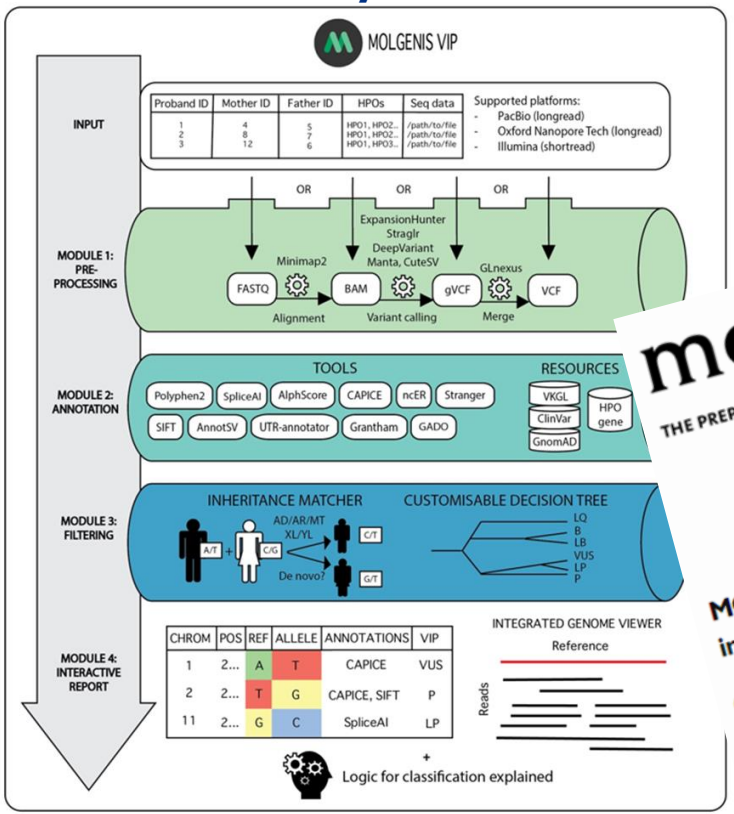




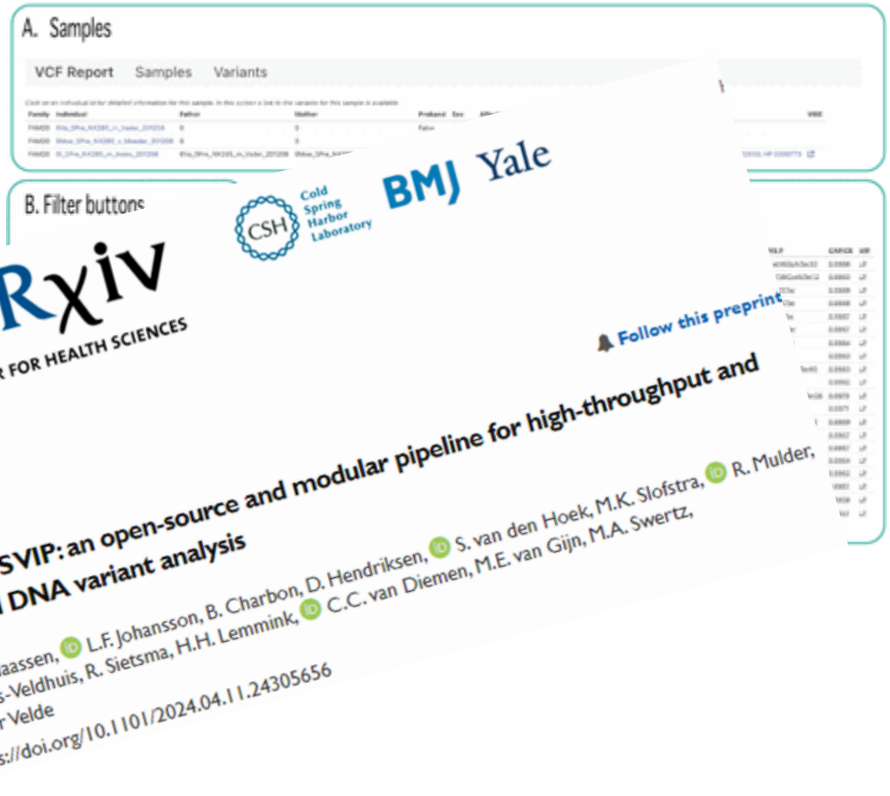
ERDERA European Rare Diseases Research Alliance

MOLGENIS Variant Interpretation Pipeline (VIP)

For DATF - **analysis** task forces



For DITF - **interpretation** task forces

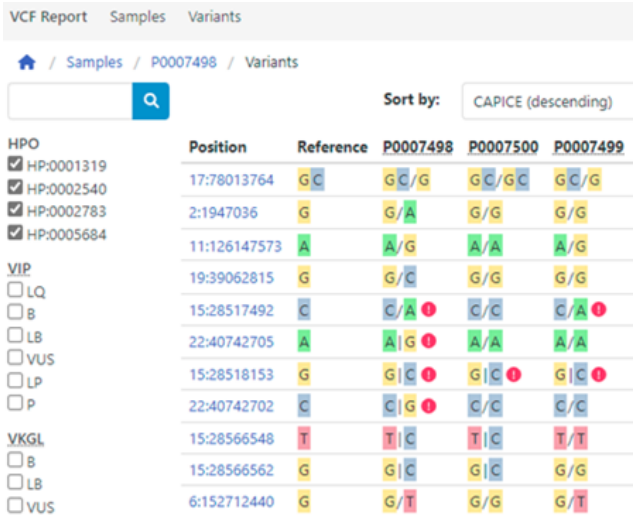


MOLGENIS Variant Interpretation Pipeline (VIP)

- Genetic variants (VCF, required)
- Pedigree information (PED)
- Patient phenotypes (Phenopackets/HPO)
- Aligned reads (BAM)

VIP

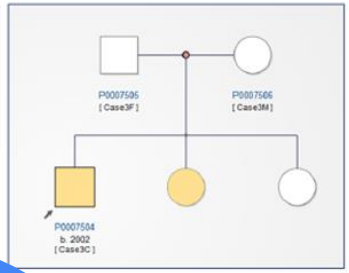
Part of the MOLGENIS data platform. https://molgenis.org



MOLGENIS Variant Interpretation Pipeline (VIP)

Case 3 description

Gender	Male
Age	16 years
Referral	Muscular dystrophy
Onset	Juvenile
Global pace of progression	Progressive
Main clinical features	• Muscle weakness • Dystrophic muscle biopsy • Quadriceps muscle atrophy • Myalgia



Information on the corresponding causative variants:

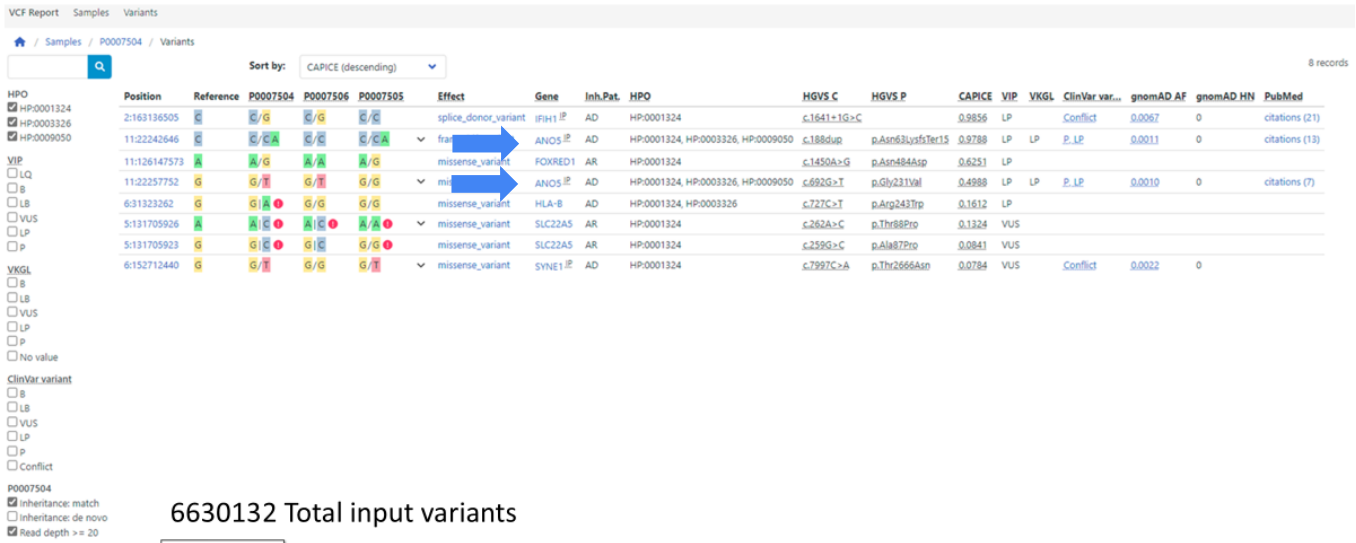
Dis[illegible] MUSCULAR DYSTROPHY, LIMB-GIRDLE, TYPE 2L; LGMD2L
Gene: ANO5
Chr Coordinates: 11:22257752G>T (mother allele) + 11:22242646C>CA (father allele)
Variants: c.692G>T p.Gly231Val (mother allele) + c.191dupA p.Asn64LysfsTer15 (father allele)
[illegible] this case in RD-Connect GPAP: https://playground.rd-connect.eu/genomics/[illegible]rs=59

Demo use-case

sample 3 from B1MG synthetic dataset
https://ega-archive.org/datasets/EGAD00002008392
Publicly available samples HApMap project
LP or P variant inserted with disease phenotype association.

FAM0001816	Index Case (male)	Father	Mother
Individual ID	Case3C	Case3F	Case3M
Phenopacket ID	P0007504	P0007505	P0007506
Genotype	**0/1; 0/1**	**0/0; 0/1**	**0/1; 0/0**
Clinical status	**Affected**	**Healthy**	**Healthy**

Variant Interpretation Pipeline (VIP)



6630132 Total input variants

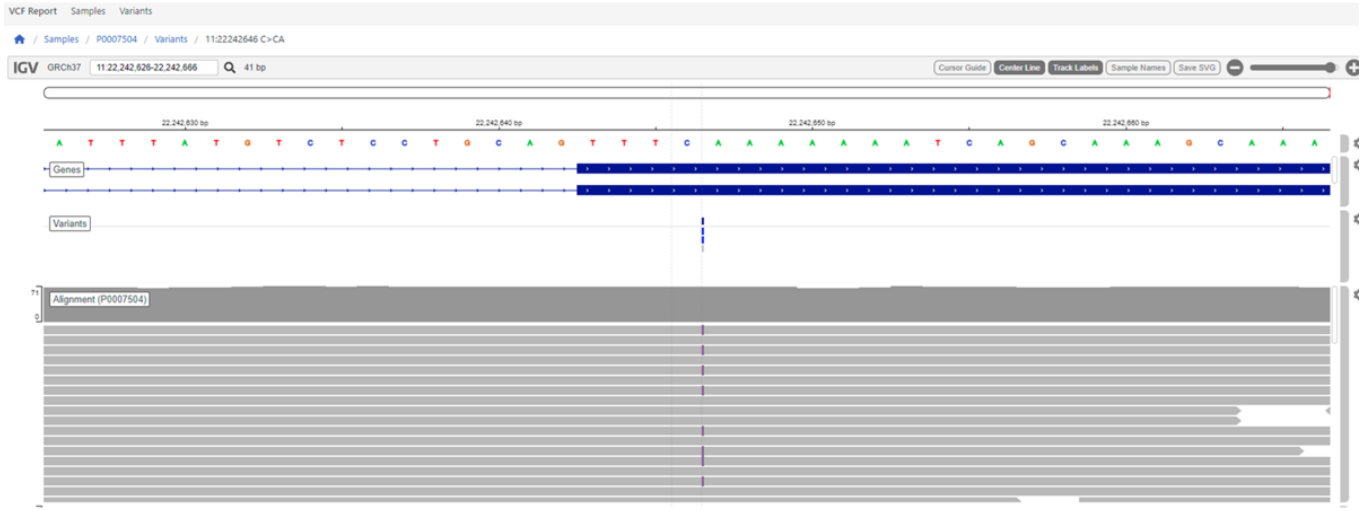
883 variants pass decision tree

42 variants in gene related to any of the patient phenotypes (ANO5 rank 5 and 15)

13 variants with inheritance match (ANO5 rank 2 and 4)

8 variants with coverage >=20 reads (ANO5 rank 2 and 4)

Variant Interpretation Pipeline (VIP)



Around each variant IGV view of bam file slice

Variant Interpretation Pipeline (VIP)

Record

Contig	11
Position	22242646
Reference allele	C
Alternate allele(s)	C A
Quality	758.12

Info

AC	2
AF	0.333
AN	6
BaseQRankSum	-0.535
DP	148
ExcessHet	3.9794
FS	0.791
MLEAC	2
MLEAF	0.333
MQRankSum	0
QD	7.74
ReadPosRankSum	0.178
SOR	0.666
VIPC_S	K

Samples

	GT	AD	DP	GQ	PL	VI	VIC	VID	VIG	VIM	VIPC_S	VIS
P0007504	C/C A	33, 27	60	99	643, 0, 828	AR, AD	11_22257752_G_T	0	203859	1	K, K	AD_IP, AR_C
P0007506	C/C	43, 0	43	99	0, 114, 1710							
P0007505	C/C A	30, 8	38	99	156, 0, 798							

CSQ

Consequence annotations from Ensembl VEP. Format:
Allele|Consequence|IMPACT|SYMBOL|Gene|Feature_type|Feature|BIOTYPE|EXON|INTRON|HGVSc|HGVSp|cDNA_position|CDS_position|Protein_position|Amino_acids|Codons|Existing_variation|ALLELE_NUM|DISTANCE|STRAND|FLAGS|PICK|SYMBOL_SOURCE|HGNC_ID|REFSEQ_MATCH|REFSEQ_OFFSET|SOURCE|SIFT|PolyPhen|HGVS_OFF

Allele	Effect	IMPACT	Gene	Gene	Feature type	Feature	BIOTYPE	EXON	HGVS.C	HGVS.P	cDNA position	CDS position	Protein_pos...	Amino acids	Codons	Existing_va...	ALLELE_NUM	STRAND	PICK	SYMBOL_SOURCE	HGVS OFFSET	ClinVar	PHENO	PubMed	SpliceAI_pr...	
A	frameshift_variant	HIGH	ANO5	IP	203859	Transcript	NM_001142649.2	protein_coding	5/22	c.188dup	p.Asn63LysfsTer15	596-597/6739	188-189/2739	63/912	N/KX	aat/aaAt	rs137854521	1	1		EntrezGene	7	likely_benign, likely_pathogenic, pathogenic	1	citations (13)	-48	-4
A	frameshift_variant	HIGH	ANO5	IP	203859	Transcript	NM_213599.3	protein_coding	5/22	c.191dup	p.Asn64LysfsTer15	599-600/6742	191-192/2742	64/913	N/KX	aat/aaAt	rs137854521	1	1	1	EntrezGene	7	likely_benign, likely_pathogenic, pathogenic	1	citations (13)	-48	-4

CSQ

Consequence annotations from Ensembl VEP. Format:
Allele|Consequence|IMPACT|SYMBOL|Gene|Feature_type|Feature|BIOTYPE|EXON|INTRON|HGVSc|HGVSp|cDNA_position|CDS_position|Protein_position|Amino_acids|Codons|Existing_variation|ALLELE_NUM|DISTANCE|STRAND|FLAGS|PICK|SYMBOL_SOURCE|HGNC_ID|REFSEQ_MATCH|REFSEQ_OFFSET|SOURCE|SIFT|PolyPhen|HGVS_OFF

ClinVar	PHENO	PubMed	SpliceAI_pr...	SpliceAI_pr...	SpliceAI_pr...	SpliceAI_pr...	SpliceAI_pr...	SpliceAI_pr...	SpliceAI_pr...	SpliceAI_pr...	CAPICE	CAPICE	HPO	Inc.Pen.	Inh.Pat.	VKGL	gnomAD	gnomAD AF	gnomAD HN	ClinVar ID	ClinVar var...	ClinVar var...	ClinVar sta...	VIP
benign, likely_pathogenic, pathogenic	1	citations (13)	-48	-4	11	43	0.01	0.00	0.00	0.00	ANO5	VUS	0.9788	HP:0001324, HP:0003326, HP:0009050	1	AD	LP	11:22242647-22242647	0.0011	0	2164	P_LP	424764:Likely_pathog	criteria_provided, _multiple_submitters, _no_conflicts	LP
benign, likely_pathogenic, pathogenic	1	citations (13)	-48	-4	11	43	0.01	0.00	0.00	0.00	ANO5	VUS	0.9781	HP:0001324, HP:0003326, HP:0009050	1	AD	LP	11:22242647-22242647	0.0011	0	2164	P_LP	424764:Likely_pathog	criteria_provided, _multiple_submitters, _no_conflicts	LP

For each variant IGV extended information per transcript

Highlighted features VIP

- alignment and variant calling for:
  - Short-read sequencing data: Illumina
  - Long-read sequencing data: Pacific BioSciences and Oxford Nanopore technologies
- Variant annotation
  - Coding sequences: e.g. CAPICE, PhyloP, SIFT, GnoMAD, Clinvar
  - Non-coding sequences: SpliceAI, UTR-Annotator, ReMM, ncER, FATHMM-MKL, constraint
- Inheritance matching
  - Based on gene and pedigree
- Phenotype matching

Can VIP help solve your cohort?

- We are looking for projects to help solve
- Which functionality would you like to be added to VIP?

ERDERA

EUROPEAN RARE DISEASES RESEARCH ALLIANCE

178 Organisations

40 funders

81 research performing organisations

9 patients' organisations

3 research infrastructures

22 private for-profit partners (industry & SME)

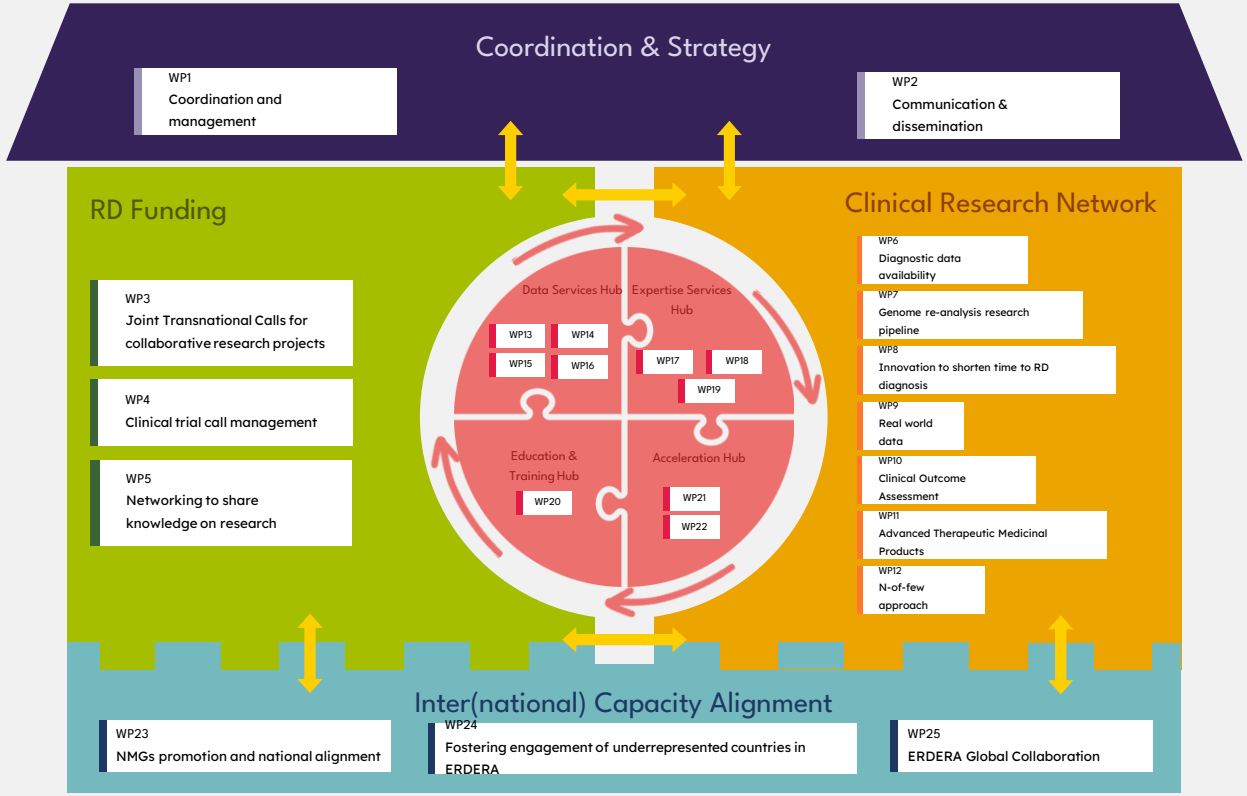
23 other (univ, hospital, non-profit, public administration)

37 Countries

25 EU member states

9 associated countries

3 non-EU



WP13
Rare Diseases-Virtual Platform (RD-VP): Finding and accessing the data ecosystem

WP14
Data readiness services

WP15
Data sharing and analysis services

WP16
Knowledge bases and ontologies for RD research

WP17
Mentoring and consultancy

WP18
Regulatory support service

WP19
Methodological Support

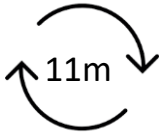
WP20
Education and training in rare diseases research

WP21
Technology accelerator

WP22
Public-Private Collaboration Accelerator

Co-funded by the European Union

WP15.1+2 operational infrastructure to CRN (y1)



DATA SUBMISSION

3 months

DATA ANALYSIS

3 months

DATA INTERPRETATION

3 months

Center notification

Submission slot opens

Data freeze

All data available for DATFs

Analysis freeze

Interpretation freeze

2m

2m

Center preparation

GPAP submissions

EGA 15d

Analysis Hubs 15d

DATFs

Interpretation by super DITFs

- Submission slots pre-granted
- Support to have all data ready for when submission slot opens
- Confirmation of experiments to be submitted

- Out of deadline – another slot will be attributed
- GPAP processing as data is received

Clinical information and experiment metadata automatically sent to RD3

- EGA submission can start before the freeze

- Hubs download data

EGA sFTP servers
EGA-ES do not envision transfer issues
Files location metadata automatically sent to RD3
Data is kept in EGA hot storage for 2 / 3 months

- WG analyses
- All results to be made available in X and link to the results/ platforms push to RD3

Data is kept in cold storage at EGA and removed at the end of the analysis freeze? Or the project?

- Interpretation to be done in X

DITFs access RD3 to link to all files and results
DITF interpretation is done in a single platform

Acknowledgements:
Colleagues at GCC, department of genetics, and our many (inter)national partners)

Feel free to contact me at:
l.johansson@umcg.nl

umcg

university of groningen

EUROPEAN JOINT PROGRAMME RARE DISEASES

eosc

CINECA

elixir Data for Life

BBMRI.nl
Biobanking and BioMolecular resources Research Infrastructure The Netherlands

SolveRD

ERDERA European Rare Diseases Research Alliance

FAIR GENOMES

European Genomic Data Infrastructure

health RI
enabling data driven health & life sciences