

SOLVING THE UNSOLVED RARE DISEASE CASES

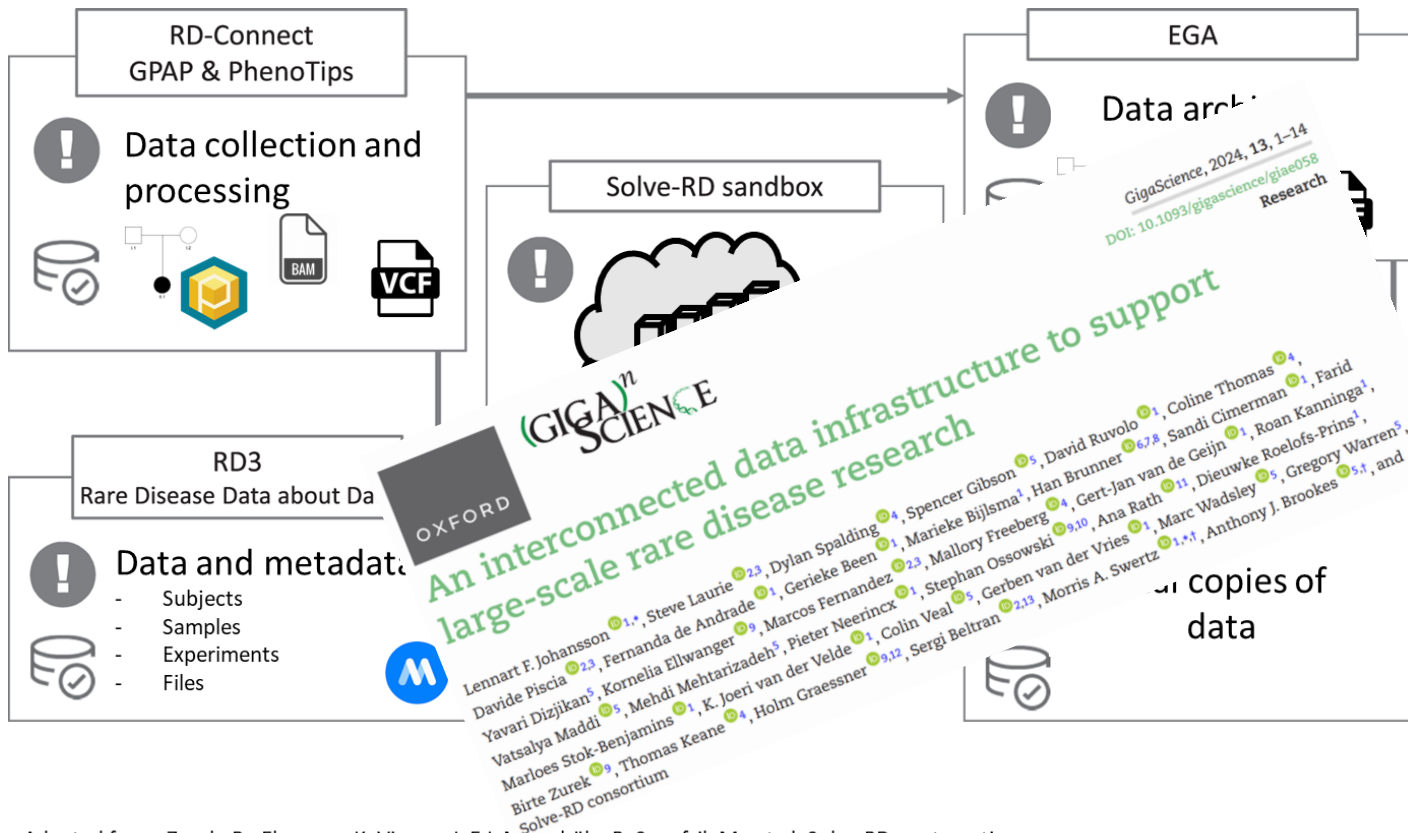
BY WORKING TOGETHER IN A SHARED ANALYSIS INFRASTRUCTURE WITH FLEXIBLE VARIANT CALLING AND INTERPRETATION TO ACCOMMODATE FOR SPECIFIC RESEARCH QUESTIONS

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

University Medical Center Groningen, The Netherlands



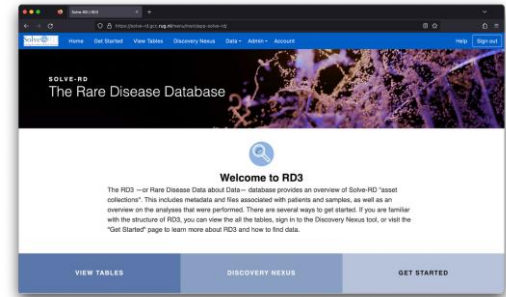
ERICA ERN Research conference 12 december 2024



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.

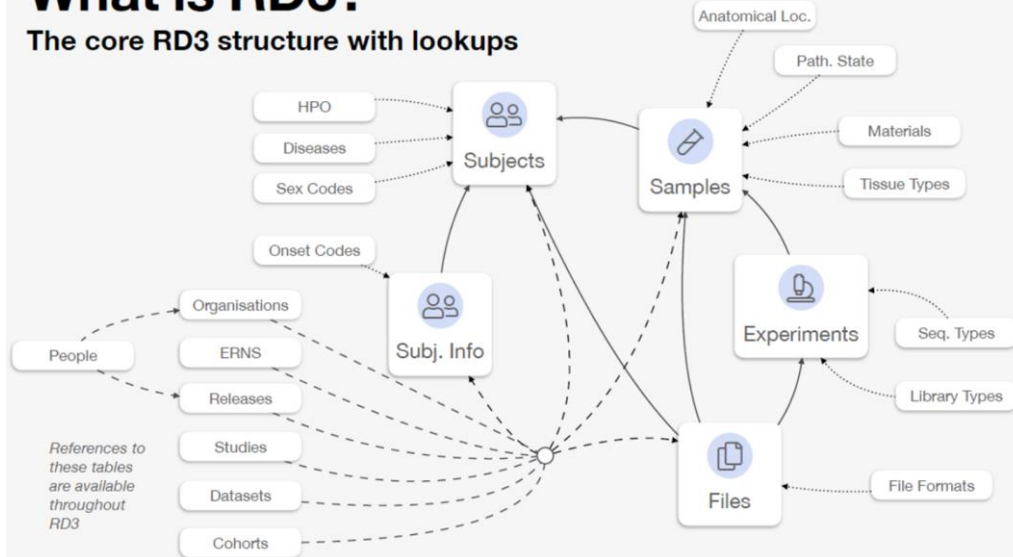


RD3 database



What is RD3?

The core RD3 structure with lookups



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

SolveRD
Solving the Unsolved Run Diseases

Subjects Samples Files Experiment_info Browser Admin Feedback Account

Subject (/ RD3) human subjects, typically patients or family members

Subject

Sample (/ RD3) samples used as input for the analysis

LabInfo (/ RD3) information of process in lab (barcodes, sequencer, etc) linked to sample(s)

LabInfo

Search data values

Data item filters

Wizard

Data item selection

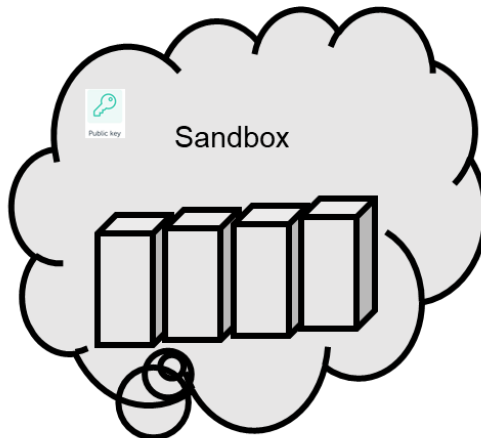
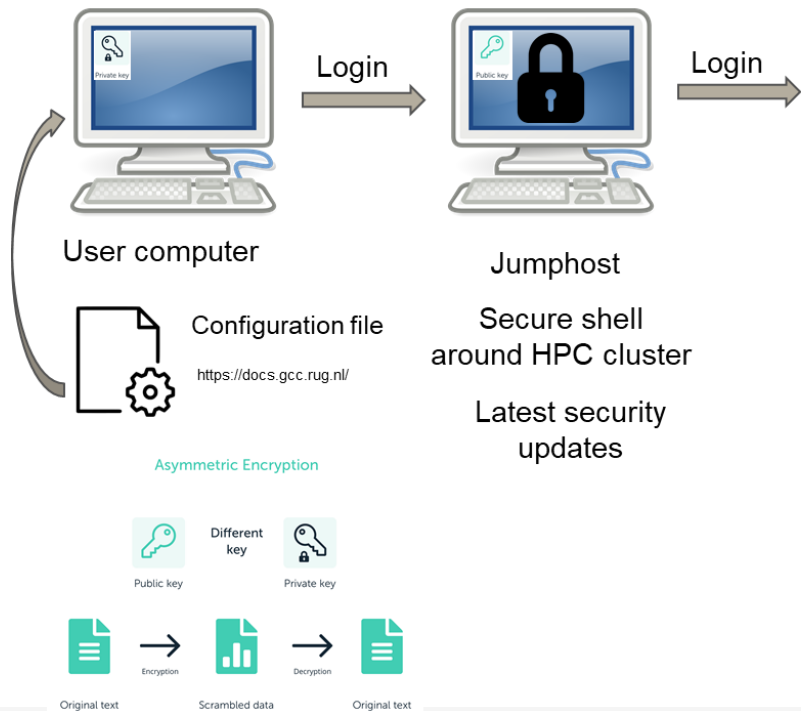
Select all Deselect all

- EGA Accession Number
- Filename
- Checksum
- typeFile
- samples
- Created
- Extra Information

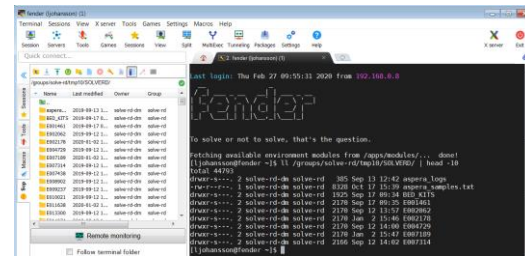
	EGA Accession Number	Filename	Checksum	typeFile
<input checked="" type="checkbox"/>	EGAF00002756209	EGAZ00001452663/ProcessedData/E200712/E200712.10.g.vcf.gz.cip	054dabb1ce49404ec686460507dc3b31	gVCF
<input checked="" type="checkbox"/>	EGAF00002756210	EGAZ00001452663/ProcessedData/E200712/E200712.11.g.vcf.gz.cip	1bae1cbb6a4d83c8f4d50604829c990f	gVCF
<input checked="" type="checkbox"/>	EGAF00002756211	EGAZ00001452663/ProcessedData/E200712/E200712.12.g.vcf.gz.cip	c7bde3463ced65fb85c0a49710544654	gVCF
<input checked="" type="checkbox"/>	EGAF00002756212	EGAZ00001452663/ProcessedData/E200712/E200712.13.g.vcf.gz.cip	f30d9e21cc295533dae89e9ec7bf7cb1	gVCF
<input checked="" type="checkbox"/>	EGAF00002756213	EGAZ00001452663/ProcessedData/E200712/E200712.14.g.vcf.gz.cip	Dee8a7214aecaf4000c84673531f2635	gVCF
<input checked="" type="checkbox"/>	EGAF00002756214	EGAZ00001452663/ProcessedData/E200712/E200712.15.g.vcf.gz.cip	73fb80969824a377c6b36b5f87eeff0b4	gVCF
<input checked="" type="checkbox"/>	EGAF00002756215	EGAZ00001452663/ProcessedData/E200712/E200712.16.g.vcf.gz.cip	c19a0873e85e845f74be03a84498a30a	gVCF
<input checked="" type="checkbox"/>	EGAF00002756216	EGAZ00001452663/ProcessedData/E200712/E200712.17.g.vcf.gz.cip	cd22ca8cb958b19f75bc97191700c2764	gVCF
<input checked="" type="checkbox"/>	EGAF00002756217	EGAZ00001452663/ProcessedData/E200712/E200712.18.g.vcf.gz.cip	6a239920e86cbe9e92b62c40f976386	gVCF
<input checked="" type="checkbox"/>	EGAF00002756218	EGAZ00001452663/ProcessedData/E200712/E200712.19.g.vcf.gz.cip	09a0658edcc41043b13cfc6c2c3f85c6f	gVCF
<input checked="" type="checkbox"/>	EGAF00002756208	EGAZ00001452663/ProcessedData/E200712/E200712.1.g.vcf.gz.cip	dcbaa4c7aae5af68db7ee4d33c8b0361	gVCF
<input checked="" type="checkbox"/>	EGAF00002756220	EGAZ00001452663/ProcessedData/E200712/E200712.20.g.vcf.gz.cip	da1366c4972f0f13aa9d35f00075abd6	gVCF
<input checked="" type="checkbox"/>	EGAF00002756221	EGAZ00001452663/ProcessedData/E200712/E200712.21.g.vcf.gz.cip	5df854fadcdc3c86676fa7c926cfa074	gVCF
<input checked="" type="checkbox"/>	EGAF00002756222	EGAZ00001452663/ProcessedData/E200712/E200712.22.g.vcf.gz.cip	b7bf4b48f137072dbd4b5dbaa3c3c7c	gVCF
<input checked="" type="checkbox"/>	EGAF00002756219	EGAZ00001452663/ProcessedData/E200712/E200712.2.g.vcf.gz.cip	695f70d47e56ad865e80a30b15be03be	gVCF



Analytical Sandbox



High Performance Cluster
functioning as a cloud for the user,
No direct connection to the outside world

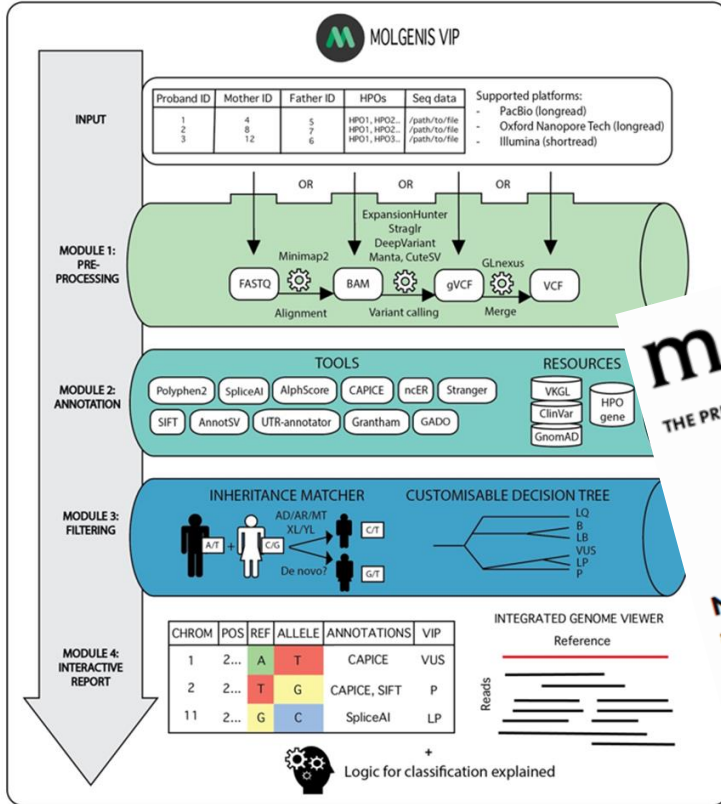




MOLGENIS Variant Interpretation Pipeline (VIP)

For DATF - **analysis** task forces

For DITF - **interpretation** task forces



A. Samples

VCF Report Samples Variants

Family	Individual	Gender	Proband	Sex	Age
PROBAND	PROBAND_HPO1_HPO2_1_1_1	0	0	Female	
MOTHER	MOTHER_HPO1_HPO2_1_1_1	0	0	Female	
FATHER	FATHER_HPO1_HPO2_1_1_1	0	0	Male	

B. Filter buttons



medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

Follow this preprint

MOLGENISVIP: an open-source and modular pipeline for high-throughput and integrated DNA variant analysis

W.T.K. Maassen, L.F. Johansson, B. Charbon, D. Hendriksen, S. van den Hoek, M.K. Slofstra, R. Mulder, M.T. Meems-Veldhuis, R. Sietsma, H.H. Lemmink, C.C. van Diemen, M.E. van Gijn, M.A. Swertz, K.J. van der Velde

doi: <https://doi.org/10.1101/2024.04.11.24305656>



MOLGENIS Variant Interpretation Pipeline (VIP)



Genetic variants (VCF, required)



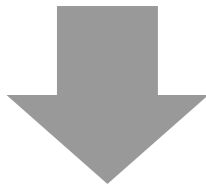
Pedigree information (PED)



Patient phenotypes (Phenopackets/HPO)



Aligned reads (BAM)



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

VCF Report Samples Variants

Home / Samples / P0007498 / Variants

Sort by: CAPICE (descending)

HPO	Position	Reference	P0007498	P0007500	P0007499
<input checked="" type="checkbox"/> HP:0001319	17:78013764	G C	G C / G	G C / G C	G C / G
<input checked="" type="checkbox"/> HP:0002540	2:1947036	G	G / A	G / G	G / G
<input checked="" type="checkbox"/> HP:0002783	11:126147573	A	A / G	A / A	A / G
<input checked="" type="checkbox"/> HP:0005684	19:39062815	G	G / C	G / G	G / G
<input type="checkbox"/> LQ	15:28517492	C	C / A	C / C	C / A
<input type="checkbox"/> B	22:40742705	A	A / G	A / A	A / A
<input type="checkbox"/> LB	15:28518153	G	G / C	G / C	G / C
<input type="checkbox"/> VUS	22:40742702	C	C / G	C / C	C / C
<input type="checkbox"/> LP	15:28566548	T	T / C	T / C	T / T
<input type="checkbox"/> P	15:28566562	G	G / C	G / C	G / G
<input type="checkbox"/> VKGL	6:152712440	G	G / T	G / G	G / T
<input type="checkbox"/> B					
<input type="checkbox"/> LB					
<input type="checkbox"/> VUS					



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	<ul style="list-style-type: none">• Muscle weakness• Dystrophic muscle biopsy• Quadriceps muscle atrophy• Myalgia

Pedigree chart showing Case 3C (P0007504) as the index case, with parents Case 3F (P0007505) and Case 3M (P0007506). Case 3C is highlighted in yellow.

Information on the corresponding causative variants:

Disease: 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)

View this case in RD-Connect GPAP: <https://playground.rd-connect.eu/genomics/2687859>

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)

VCF Report Samples Variants

Home / Samples / P0007504 / Variants

Sort by: CAPICE (descending) 8 records

HPO	Position	Reference	P0007504	P0007506	P0007505	Effect	Gene	Inh.Pat.	HPO	HGVSC	HGVSP	CAPICE	VIP	VKGL	ClinVar var...	gnomAD AF	gnomAD HN	PubMed
<input checked="" type="checkbox"/> HP:0001324	2:163136505	C	C/G	C/G	C/C	splice_donor_variant	IFIH1 ^{IP}	AD	HP:0001324	c.1641+1G>C		0.9856	LP		Conflict	0.0067	0	citations (21)
<input checked="" type="checkbox"/> HP:0003326	11:22242646	C	C/A	C/C	C/A	fr	ANOS ^{IP}	AD	HP:0001324, HP:0003326, HP:0009050	c.188dup	p.Asn63LysfsTer15	0.9788	LP	LP	P LP	0.0011	0	citations (13)
<input checked="" type="checkbox"/> HP:0009050	11:126147573	G	A/G	A/A	A/G	missense_variant	FOXRED1	AR	HP:0001324	c.1450A>G	p.Asn484Asp	0.6251	LP					
VIP	11:22257752	G	G/T	G/T	G/G	mi	ANOS ^{IP}	AD	HP:0001324, HP:0003326, HP:0009050	c.692G>T	p.Gly231Val	0.4988	LP	LP	P LP	0.0010	0	citations (7)
<input type="checkbox"/> LQ	6:31323262	G	G/A	G/G	G/G	missense_variant	HLA-B	AD	HP:0001324, HP:0003326	c.727C>T	p.Arg243Trp	0.1612	LP					
<input type="checkbox"/> LB	5:131705926	G	A/C	A/C	A/A	missense_variant	SLC22A5	AR	HP:0001324	c.262A>C	p.Thr98Pro	0.1324	VUS					
<input type="checkbox"/> VUS	5:131705923	G	G/C	G/C	G/G	missense_variant	SLC22A5	AR	HP:0001324	c.259G>C	p.Ala97Pro	0.0841	VUS					
<input type="checkbox"/> LP	6:152712440	G	G/T	G/G	G/T	missense_variant	SYNE1 ^{IP}	AD	HP:0001324	c.7997C>A	p.Thr2666Asn	0.0784	VUS		Conflict	0.0022	0	
<input type="checkbox"/> P																		
YKGL																		
<input type="checkbox"/> B																		
<input type="checkbox"/> LB																		
<input type="checkbox"/> VUS																		
<input type="checkbox"/> LP																		
<input type="checkbox"/> P																		
No value																		
ClinVar variant																		
<input type="checkbox"/> B																		
<input type="checkbox"/> LB																		
<input type="checkbox"/> VUS																		
<input type="checkbox"/> LP																		
<input type="checkbox"/> P																		
Conflict																		
P0007504																		
<input checked="" type="checkbox"/> Inheritance: match																		
<input type="checkbox"/> Inheritance: de novo																		
<input checked="" type="checkbox"/> Read depth >= 20																		

6630132 Total input variants



883 variants pass decision tree

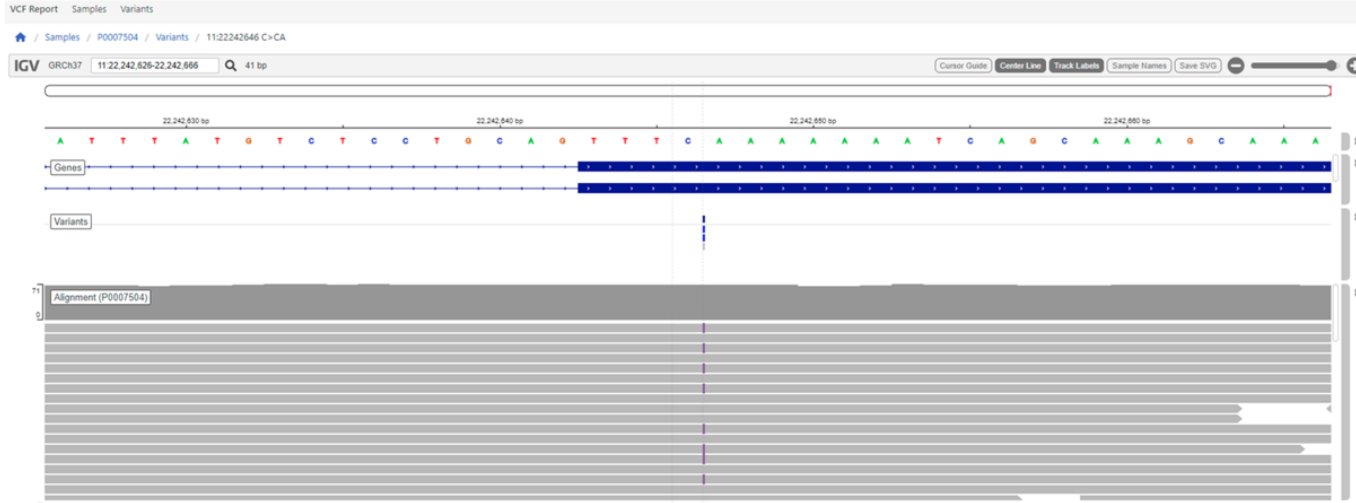
42 variants in gene related to any of the patient phenotypes (ANOS rank 5 and 15)

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

8 variants with coverage >=20 reads (ANOS 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 **K**
 Alternate allele(s) **C**
 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
 VPC_S K

Samples

	GT	AD	DP	GQ	PL	VI	VIC	VID	VIG	VIM	VIPC_S	VIS
P0007504	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/A	43,0	43	99	0,114,1710							
P0007505	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|HGVS_C|HGVS_P|DNA_position|CDS_position|Protein_position|Amino_acids|Codons|Existing_variation|ALLELE_NUM|DISTANCE|STRAND|FLAG|SPICK|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	DNA position	CDS position	Protein pos...	Amino acids	Codons	Existing va...	ALLELE NUM	STRAND	PICK	SYMBOL_SOURCE	HGVS OFFSET	ClinVar	PHENO	Published	SIFT	PolyPhen	HGVS_OFF
A	frameshift_variant	HIGH	ANOS1P	203859	Transcript	NM_001142649.2	protein_coding	5/22	c.188dup	p.Asn61ysfTer15	596-597/6739	188-189/2739	63/912	N/XX	aat/aaat	rs137854521	1	1	EntrezGene	7	likely_benign, likely_pathogenic, pathogenic	1	citations (13)	-48	-4		
A	frameshift_variant	HIGH	ANOS1P	203859	Transcript	NM_213599.3	protein_coding	5/22	c.191dup	p.Asn64ysfTer15	599-600/6742	191-192/2742	64/913	N/XX	aat/aaat	rs137854521	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|HGVS_C|HGVS_P|DNA_position|CDS_position|Protein_position|Amino_acids|Codons|Existing_variation|ALLELE_NUM|DISTANCE|STRAND|FLAG|SPICK|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	DNA position	CDS position	Protein pos...	Amino acids	Codons	Existing va...	ALLELE NUM	STRAND	PICK	SYMBOL_SOURCE	HGVS OFFSET	ClinVar	PHENO	Published	SIFT	PolyPhen	HGVS_OFF
benign, likely_pathogenic, pathogenic																											
benign, likely_pathogenic, pathogenic																											

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?

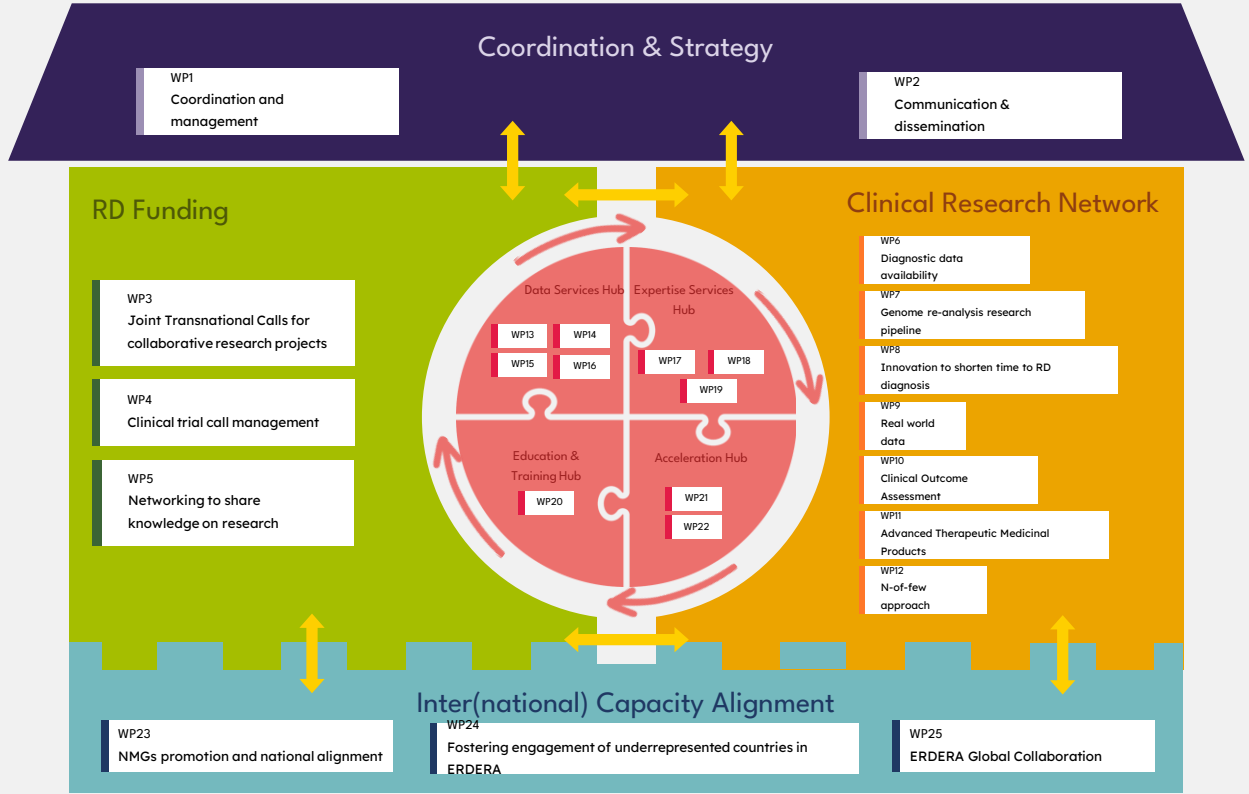
- We are looking for projects to help solve
- Which functionality would you like to be added to VIP?

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)

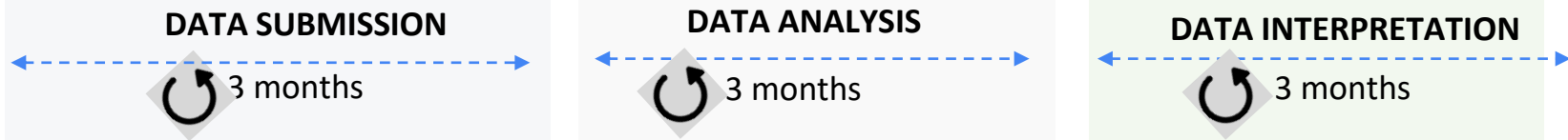
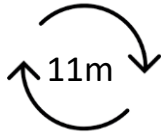
37 Countries

- 25 EU member states
- 9 associated countries
- 3 non-EU



<p>WP13 Rare Diseases-Virtual Platform (RD-VP): Finding and accessing the data ecosystem</p>	<p>WP16 Knowledge bases and ontologies for RD research</p>	<p>WP19 Methodological Support</p>
<p>WP14 Data readiness services</p>	<p>WP17 Mentoring and consultancy</p>	<p>WP20 Education and training in rare diseases research</p>
<p>WP15 Data sharing and analysis services</p>	<p>WP18 Regulatory support service</p>	<p>WP21 Technology accelerator</p>
		<p>WP22 Public-Private Collaboration Accelerator</p>

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



Center notification

Submission slot opens

Data freeze

All data available for DATFs

Analysis freeze

Interpretation freeze



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

- EGA submission can start before the freeze
- Hubs download data

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

- Interpretation to be done in X

Clinical information and experiment metadata automatically sent to RD3

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

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

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



eosc CINECA



E-DEERA European Rare Diseases Research Alliance



health RI
enabling data driven health & life sciences