

Cutting-edge NGS diagnostics: from interpretation to FAIRification & challenges of large-scale Rare Disease cohort analysis

Lennart F. Johansson & K. Joeri van der Velde
University Medical Center Groningen, Systems Genetics (i.e. Swertz group)
X-omics/BBMRI-NL workshop, January 20th 2021, 15:00

1 DNA error in 3 bil. can
cause “rare disease”



we know ~6,000 rare diseases

affecting 1:12 babies born

<https://www.mcrc.edu.au/content/rare-disease>

Context

genome diagnostics

find flaws in patient DNA that
explains their disease



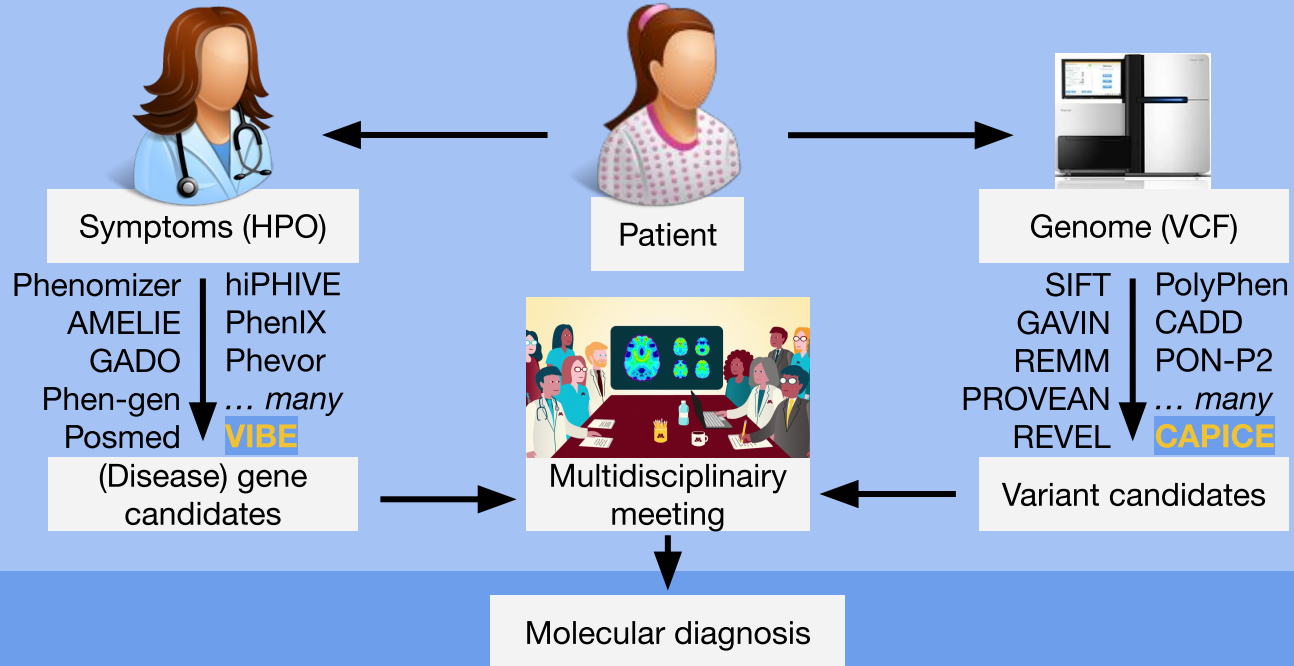
give a patient an **answer**,
a **prognosis**, and better
treatment options

The topics for today

FAIRification: sharing & reusing healthcare & research data (FAIR genomes)

Combine 19,000 patients: large-scale Rare Disease cohort analysis (Solve-RD)

Genome diagnostics: Variant Interpretation Pipeline (VIP)

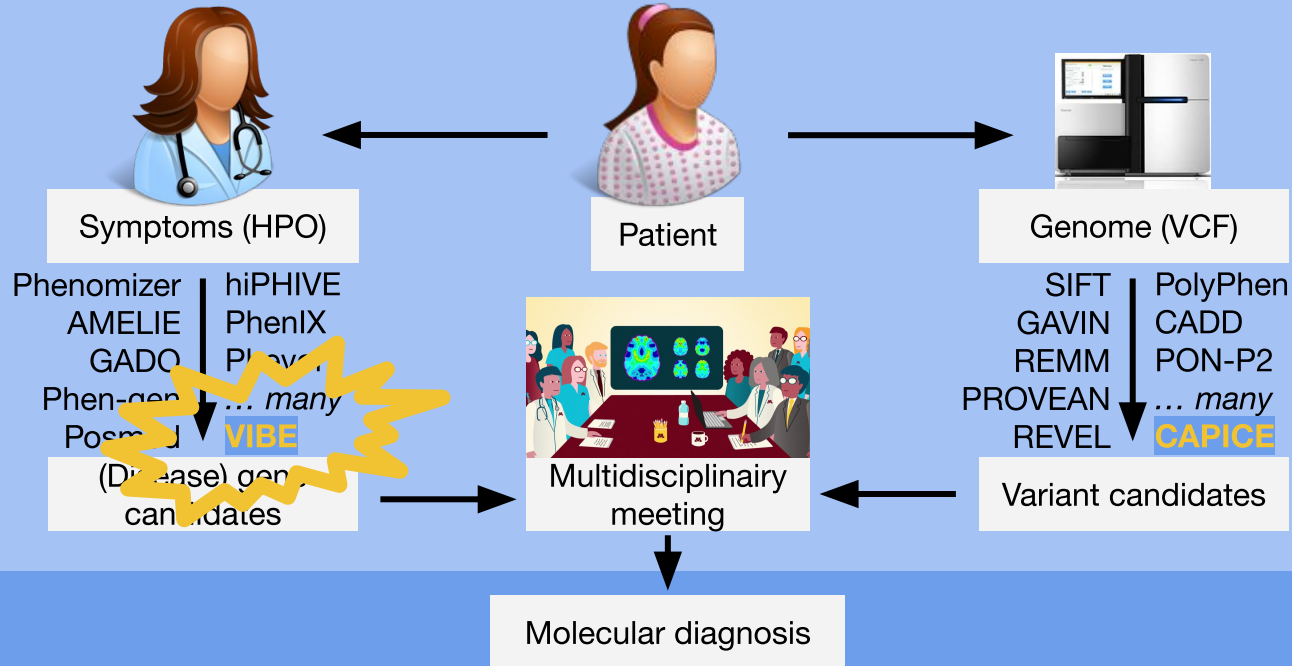


First up: VIBE

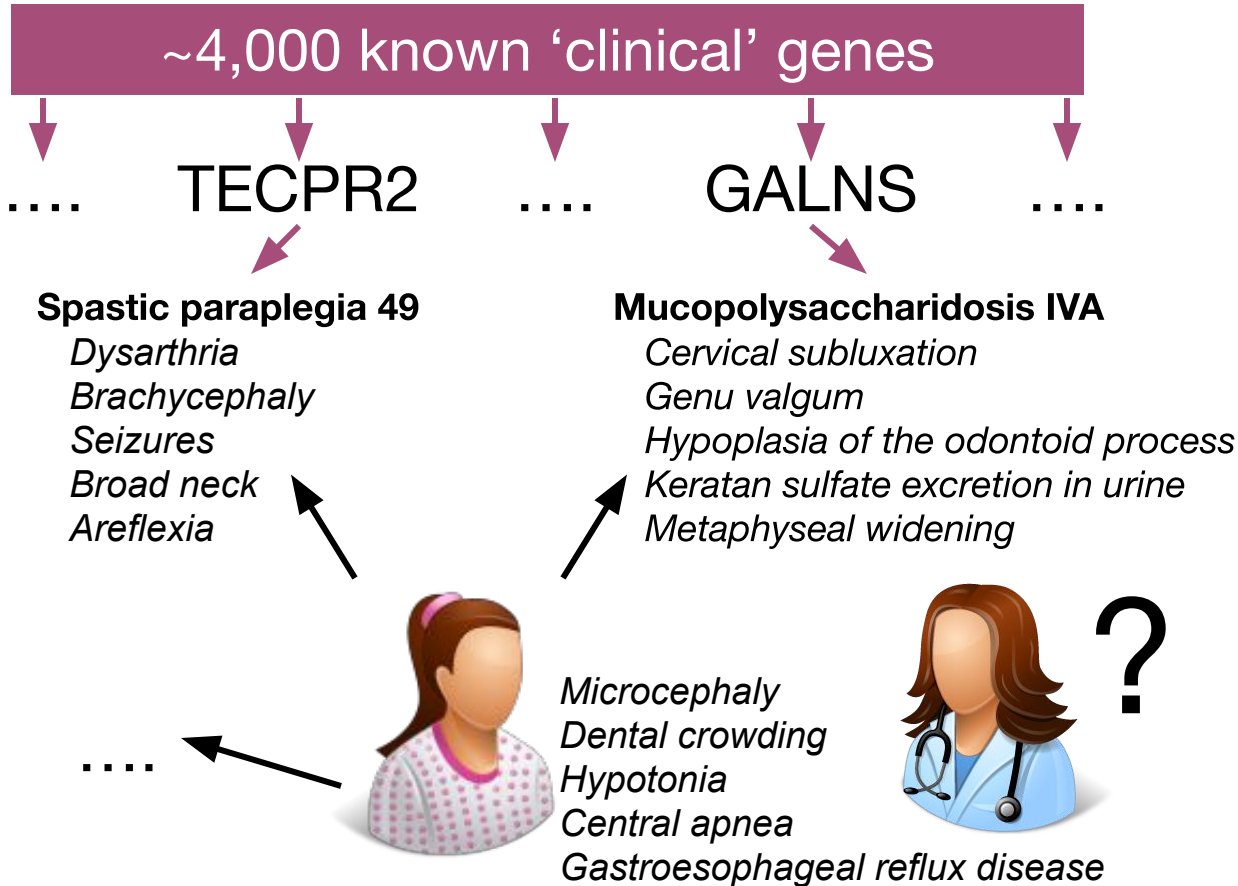
FAIRification: sharing & reusing healthcare & research data (FAIR genomes)

Combine 19,000 patients: large-scale Rare Disease cohort analysis (Solve-RD)

Genome diagnostics: Variant Interpretation Pipeline (VIP)



Symptoms to genes... ?



Few tools suitable for routine diag.

Tool	Suitable for routine diagnostics?
Phevor	No, online only
Phen-gen	No, embedded some variant prioritizer
OVA	No, online only
AMELIE	No, online only
PubCaseFinder	No, online only
SSAGA	No, closed source, unavailable
Phenomizer	No, closed source, online only
eXtasy	No, software has been abandoned
Posmed	No, software/site is gone without a trace
OMIM Explorer	No, software/site is gone without a trace
patient_sim	No, software/site is gone without a trace
Phenotips	Now commercial, cloud-based, if that is acceptable
PhenIX / hiPHIVE	Embedded in a variant prioritizer (<i>standalone possible</i>)
GADO	Yes, open source, offline cmdline executable available

FYI: new tool **LIRICAL**
did not exist yet when
VIBE was developed

VIBE: prioritize by real evidence

Abnormality of the
cerebral white matter
(HP:0002500)

+

Atrial septal defect
(HP:0001631)

+

Arachnodactyly
(HP:0001166)



VIBE

id	gene	pubmed	pubmed	pubmed
1	ADAMTS-1	16101611	16101611	16101611
2	ADAMTS-1	16101611	16101611	16101611
3	ADAMTS-1	16101611	16101611	16101611
4	ADAMTS-1	16101611	16101611	16101611
5	ADAMTS-1	16101611	16101611	16101611
6	ADAMTS-1	16101611	16101611	16101611
7	ADAMTS-1	16101611	16101611	16101611
8	ADAMTS-1	16101611	16101611	16101611
9	ADAMTS-1	16101611	16101611	16101611
10	ADAMTS-1	16101611	16101611	16101611
11	ADAMTS-1	16101611	16101611	16101611
12	ADAMTS-1	16101611	16101611	16101611
13	ADAMTS-1	16101611	16101611	16101611
14	ADAMTS-1	16101611	16101611	16101611
15	ADAMTS-1	16101611	16101611	16101611
16	ADAMTS-1	16101611	16101611	16101611
17	ADAMTS-1	16101611	16101611	16101611
18	ADAMTS-1	16101611	16101611	16101611
19	ADAMTS-1	16101611	16101611	16101611
20	ADAMTS-1	16101611	16101611	16101611
21	ADAMTS-1	16101611	16101611	16101611
22	ADAMTS-1	16101611	16101611	16101611
23	ADAMTS-1	16101611	16101611	16101611
24	ADAMTS-1	16101611	16101611	16101611
25	ADAMTS-1	16101611	16101611	16101611
26	ADAMTS-1	16101611	16101611	16101611
27	ADAMTS-1	16101611	16101611	16101611
28	ADAMTS-1	16101611	16101611	16101611
29	ADAMTS-1	16101611	16101611	16101611
30	ADAMTS-1	16101611	16101611	16101611
31	ADAMTS-1	16101611	16101611	16101611
32	ADAMTS-1	16101611	16101611	16101611
33	ADAMTS-1	16101611	16101611	16101611
34	ADAMTS-1	16101611	16101611	16101611
35	ADAMTS-1	16101611	16101611	16101611
36	ADAMTS-1	16101611	16101611	16101611
37	ADAMTS-1	16101611	16101611	16101611
38	ADAMTS-1	16101611	16101611	16101611
39	ADAMTS-1	16101611	16101611	16101611
40	ADAMTS-1	16101611	16101611	16101611
41	ADAMTS-1	16101611	16101611	16101611
42	ADAMTS-1	16101611	16101611	16101611
43	ADAMTS-1	16101611	16101611	16101611
44	ADAMTS-1	16101611	16101611	16101611
45	ADAMTS-1	16101611	16101611	16101611
46	ADAMTS-1	16101611	16101611	16101611
47	ADAMTS-1	16101611	16101611	16101611
48	ADAMTS-1	16101611	16101611	16101611
49	ADAMTS-1	16101611	16101611	16101611
50	ADAMTS-1	16101611	16101611	16101611
51	ADAMTS-1	16101611	16101611	16101611
52	ADAMTS-1	16101611	16101611	16101611
53	ADAMTS-1	16101611	16101611	16101611
54	ADAMTS-1	16101611	16101611	16101611
55	ADAMTS-1	16101611	16101611	16101611
56	ADAMTS-1	16101611	16101611	16101611
57	ADAMTS-1	16101611	16101611	16101611
58	ADAMTS-1	16101611	16101611	16101611
59	ADAMTS-1	16101611	16101611	16101611
60	ADAMTS-1	16101611	16101611	16101611
61	ADAMTS-1	16101611	16101611	16101611
62	ADAMTS-1	16101611	16101611	16101611
63	ADAMTS-1	16101611	16101611	16101611
64	ADAMTS-1	16101611	16101611	16101611
65	ADAMTS-1	16101611	16101611	16101611
66	ADAMTS-1	16101611	16101611	16101611
67	ADAMTS-1	16101611	16101611	16101611
68	ADAMTS-1	16101611	16101611	16101611
69	ADAMTS-1	16101611	16101611	16101611
70	ADAMTS-1	16101611	16101611	16101611
71	ADAMTS-1	16101611	16101611	16101611
72	ADAMTS-1	16101611	16101611	16101611
73	ADAMTS-1	16101611	16101611	16101611
74	ADAMTS-1	16101611	16101611	16101611
75	ADAMTS-1	16101611	16101611	16101611
76	ADAMTS-1	16101611	16101611	16101611
77	ADAMTS-1	16101611	16101611	16101611
78	ADAMTS-1	16101611	16101611	16101611
79	ADAMTS-1	16101611	16101611	16101611
80	ADAMTS-1	16101611	16101611	16101611
81	ADAMTS-1	16101611	16101611	16101611
82	ADAMTS-1	16101611	16101611	16101611
83	ADAMTS-1	16101611	16101611	16101611
84	ADAMTS-1	16101611	16101611	16101611
85	ADAMTS-1	16101611	16101611	16101611
86	ADAMTS-1	16101611	16101611	16101611
87	ADAMTS-1	16101611	16101611	16101611
88	ADAMTS-1	16101611	16101611	16101611
89	ADAMTS-1	16101611	16101611	16101611
90	ADAMTS-1	16101611	16101611	16101611
91	ADAMTS-1	16101611	16101611	16101611
92	ADAMTS-1	16101611	16101611	16101611
93	ADAMTS-1	16101611	16101611	16101611
94	ADAMTS-1	16101611	16101611	16101611
95	ADAMTS-1	16101611	16101611	16101611
96	ADAMTS-1	16101611	16101611	16101611
97	ADAMTS-1	16101611	16101611	16101611
98	ADAMTS-1	16101611	16101611	16101611
99	ADAMTS-1	16101611	16101611	16101611
100	ADAMTS-1	16101611	16101611	16101611

Find all matching disease genes from literature & databases

... and prioritize the genes, please

- Explain why: link to publications
 - Suitable for routine diagnostics
 - Stand-alone cmdline executable
 - Open Source & as web tool
- <https://molgenis.org/vibe>

Main data source: DisGeNET

- Curated
- Literature
- Animal models

Publication:
van der Velde, KJ, van den Hoek, S, van Dijk, F, et al. **A pipeline-friendly software tool for genome diagnostics to prioritize genes by matching patient symptoms to literature.** *Advanced Genetics*. 2020; 1:e10023. <https://doi.org/10.1002/ggn2.10023>



<https://github.com/molgenis/vibe>

Implementation: Sander van den Hoek

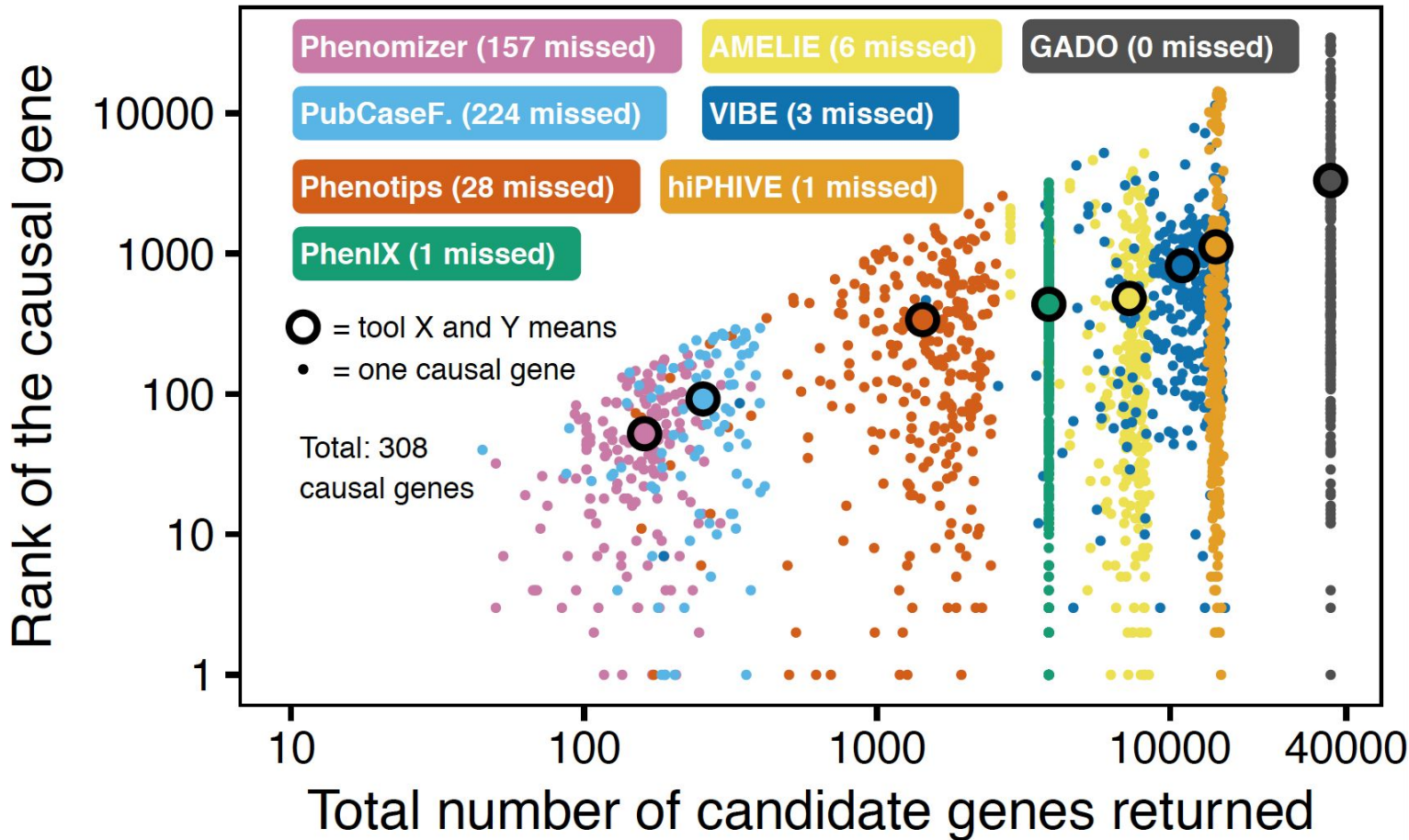
305 solved RD patient cases from Trujillano *et al.* (<http://dx.doi.org/10.1038/ejhg.2016.146>)

→ Per case: HPO terms & molecular diagnosis (i.e. causal variant in a disease gene)

Assess 8 different tools:

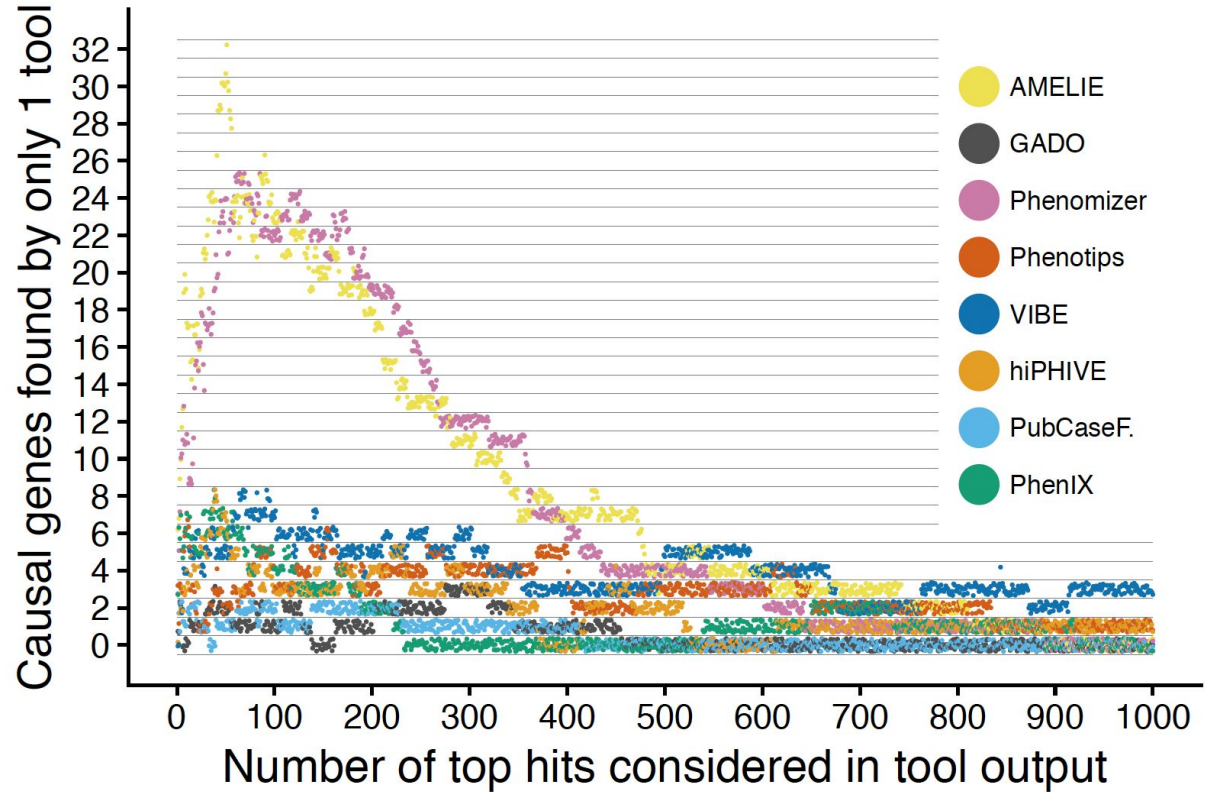
→ VIBE, GADO, AMELIE, hiPHIVE, PhenIX, PubCaseFinder, Phenotips, Phenomizer

Huge differences in output size



High amount of complementarity

- Consider **unique** hits among all tools
- Sliding rank cutoff 1-1000

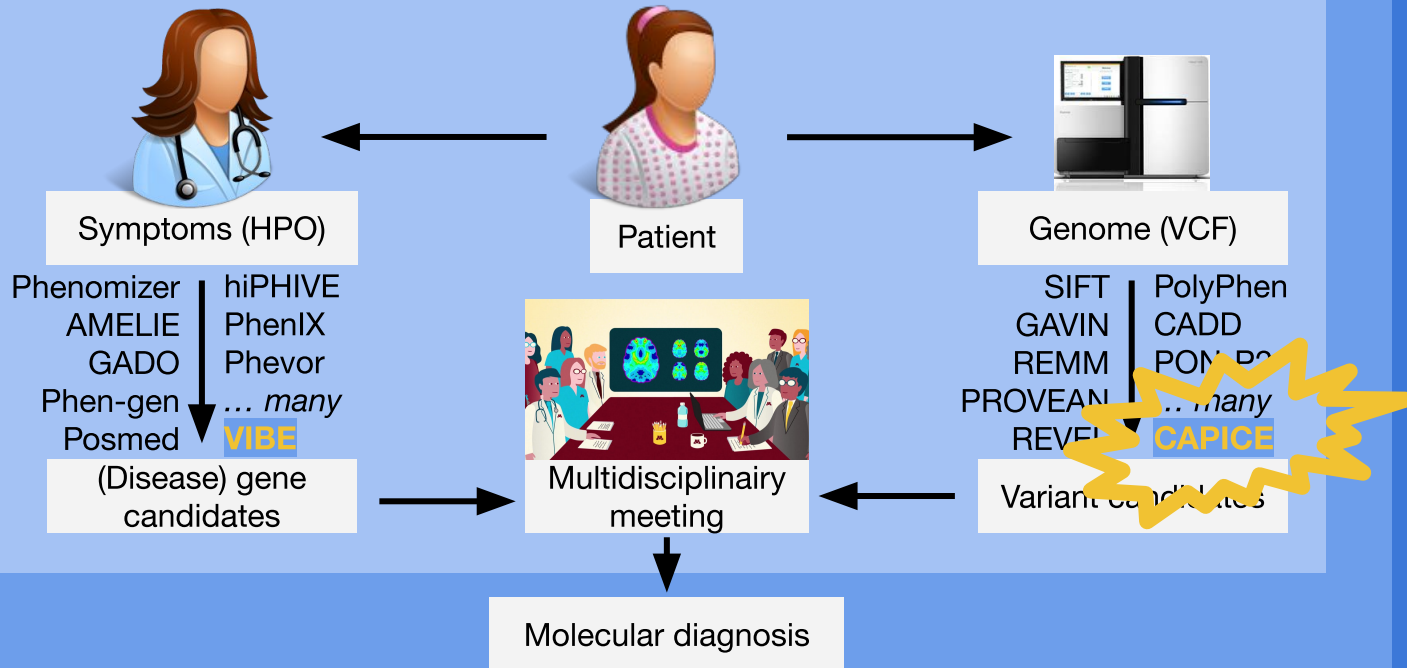


Second: CAPICE

FAIRification: sharing & reusing healthcare & research data (FAIR genomes)

Combine 19,000 patients: large-scale Rare Disease cohort analysis (Solve-RD)

Genome diagnostics: Variant Interpretation Pipeline (VIP)



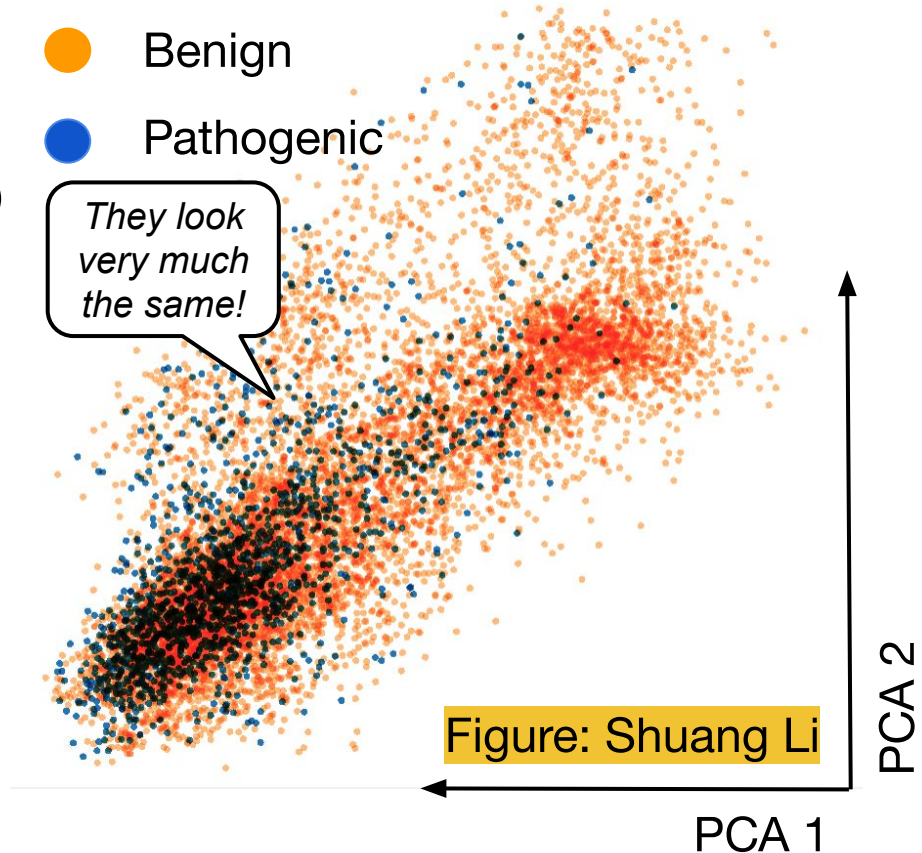
CAPICE: the challenge

*Dozens of 'variant pathogenicity prediction' tools, many limitations. Need a **fresh take**.*

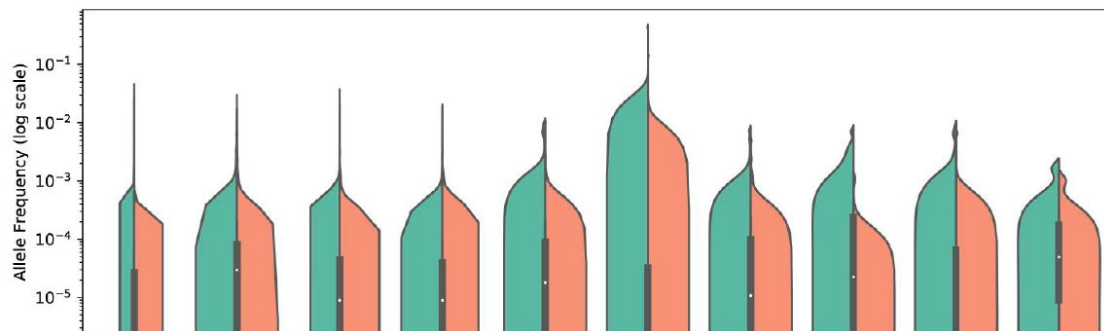
- 334,602 variants (gnomAD, ClinVar, VKGL)
 - 293,921 'neutral' DNA variants
 - 40,681 pathogenic DNA variants
- Added many genomic annotations
 - CADD 1.4 features ($n=92$)
 - Conservation scores (GERP)
 - Functional data (DNase hypersens.)
 - gnomAD AF (from exome & genome)
 - Transcript information (expression)
 - Protein-level scores (SIFT)
 - *etc*

- Benign
- Pathogenic

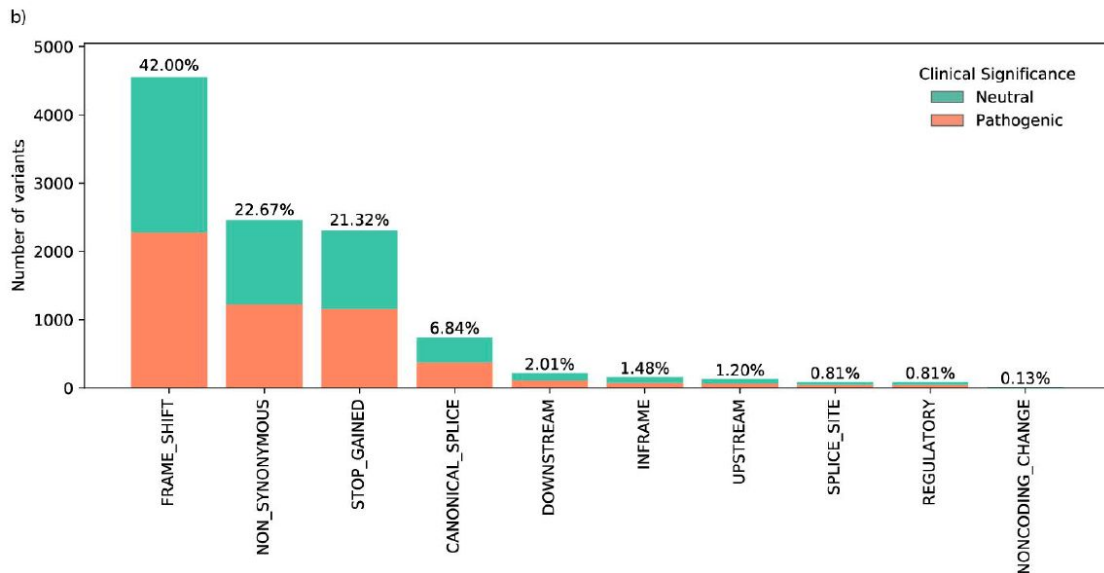
They look very much the same!



CAPICE trick: fully balanced data



← allele
frequency

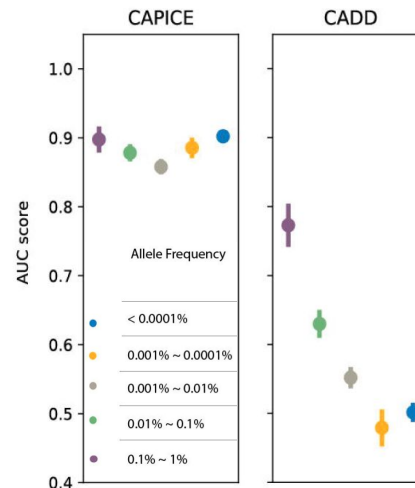
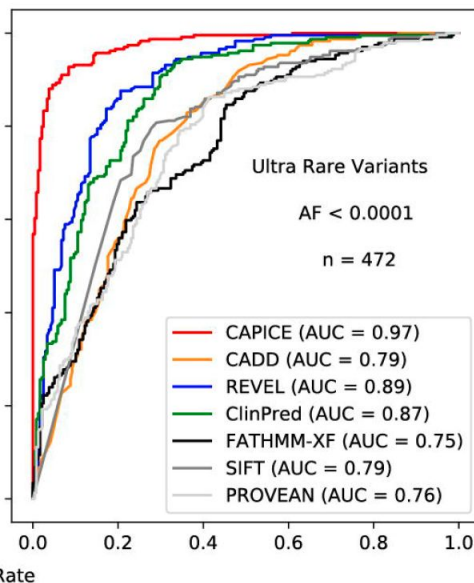
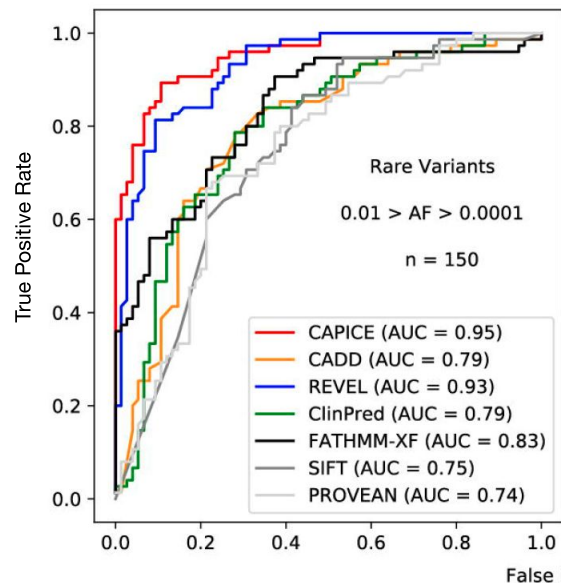


← molecular
effect

→ *into XGBoost
Machine Learning*

CAPICE: results

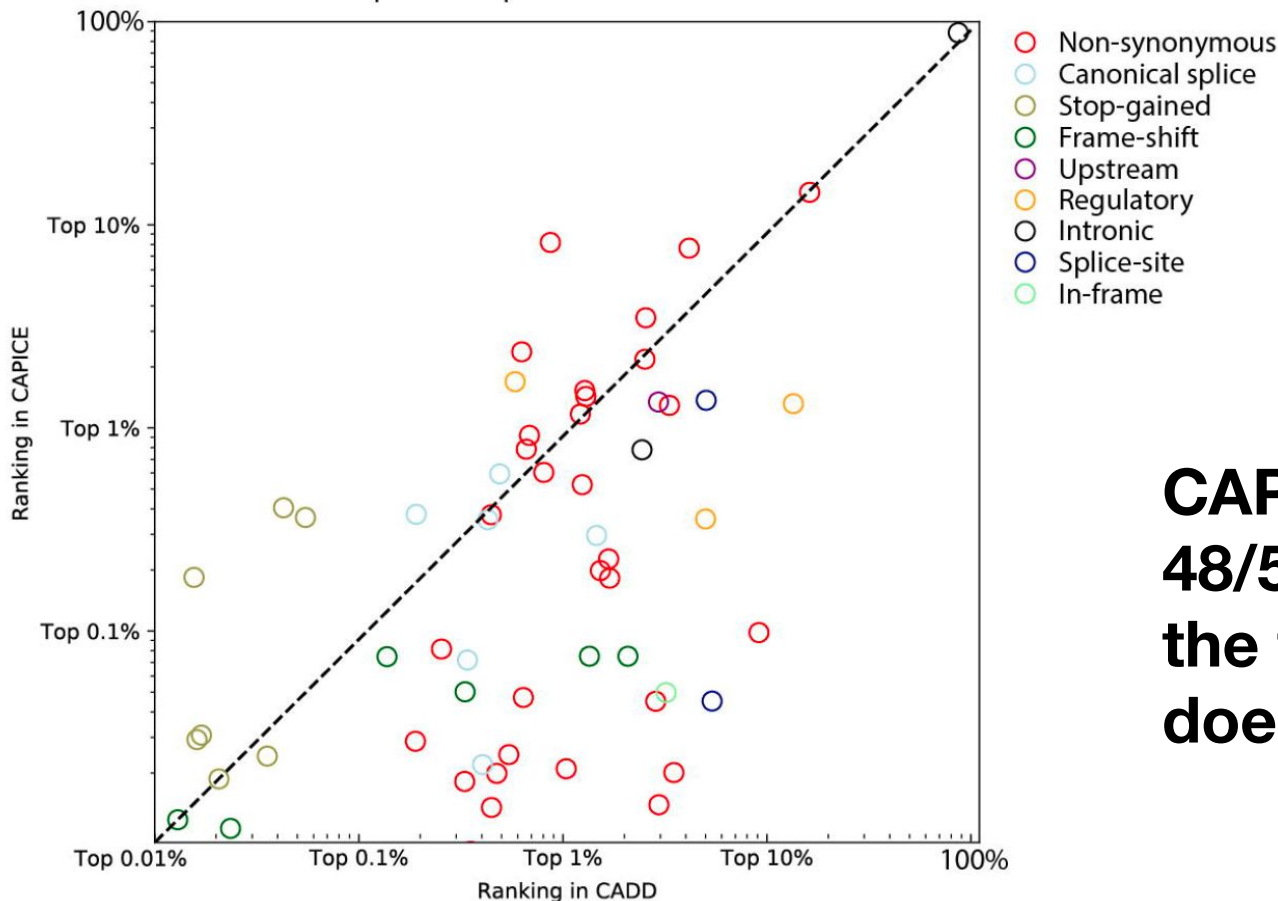
- Automated pathogenicity prediction for **any** DNA variant
- Retains high performance **especially** on ultra-rare variants



GitHub <https://github.com/molgenis/capice>

CAPICE: on solved patient cases

Comparison of performance in real cases



Publication:

Li, S., van der Velde, K.J., de Ridder, D. *et al.*
**CAPICE: a computational method for
Consequence-Agnostic Pathogenicity
Interpretation of Clinical Exome variations.**
Genome Med 12, 75 (2020).
<https://doi.org/10.1186/s13073-020-00775-w>

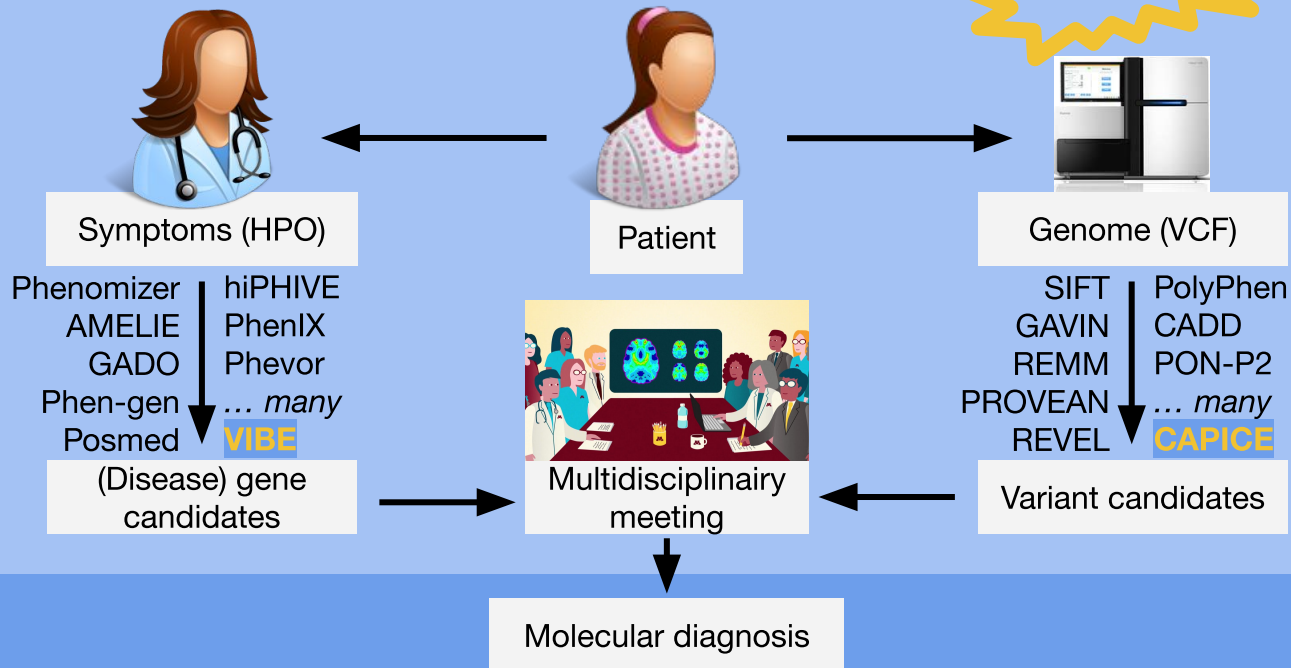
**CAPICE prioritizes
48/58 of variants in
the top 1%, CADD
does 35/58.**

Combine all tools: **VIP**

FAIRification: sharing & reusing healthcare & research data (FAIR genomes)

Combine 19,000 patients: large-scale Rare Disease cohort analysis (Poly-RD)

Genome diagnostics: Variant Interpretation Pipeline (VIP)



→ Collaborations



umcg
Genomdiagnostiek

VKGL

vereniging klinisch genetische
LABORATORIUMDIAGNOSTIEK



CINECA

Solve  RD
Solving the Unsolved Rare Diseases

→ Scope



→ Routine genome
diagnostics (GD)



→ Diagnostics in
development (GDIO)



→ Genetic (rare) disease
research

Variant Interpretation Pipeline

Goal: to reduce 100,000 DNA variants (VCF) to ~10 best candidates for multidisc. meeting



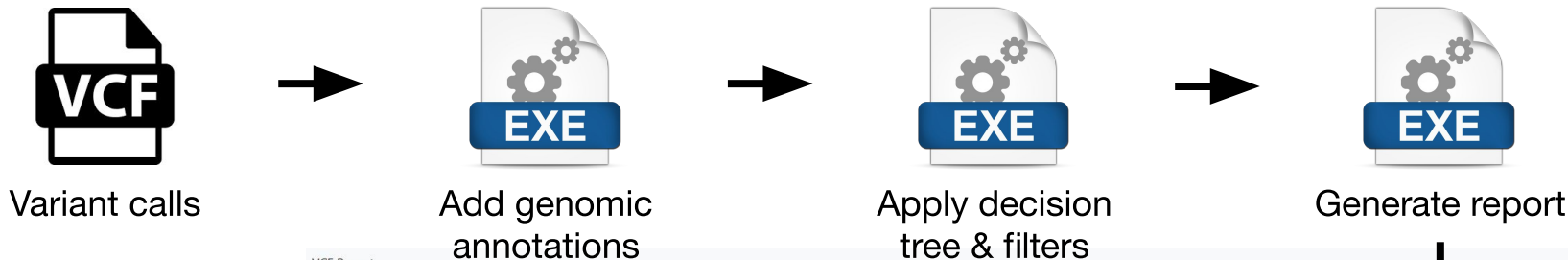
How: complex, time-consuming process - *automate where we can!*



Why?

1. **Reduce repetitive manual work** (more time to solve difficult cases)
2. **Freedom to design** exactly how we want it (not closed-source product)
3. **Open Source** for community use & development (e.g. via VKGL)
4. **Can prepare for WGS/*omics** (e.g. RNA-sequencing integration)
5. **Innovative new methods** can be quickly tested and adopted

Pipeline, a 'pipe dream'? No!



VCF Report

Phenotypes: HP:0004383 VIB

Position	Identifiers	Reference	Sample	Father	Mother	CAPICE	Effect	Symbol	HGVS C	HGVS P	gnomAD	VKGL	ClinVar	Literature
1:110,042,538		C	T C	C C	C C	0.0022	missense	NMNAT1	c.619C>T	p.Arg207Trp	0.00004	LP	P	PubMed
1:116,376,412		G	G A	A G	A G	0.7896						LP		
1:117,349,215		C	G C	C C	C C	0.9898	missense	SDHB	c.653G>C	p.Trp218Ser		LP	LP	
1:117,355,094		C	T C	C C	C C	0.9879						LP		
1:117,355,106		C	T T	T C	T C	0.9848	missense	SDHB	c.412G>A	p.Asp138Asn		LP		
1:117,355,175		G	A G	A G	A G	0.9819	stop_gained	SDHB	c.343C>T	p.Arg115Ter	0.00001	LP	P	PubMed
1:145,798,112		G	A A	A G	A G	0.9719	stop_gained	MUTYH	c.658C>T	p.Arg220Ter	0.00000	LP	P, LP	PubMed
1:145,798,130		G	A A	A G	A G	0.9877	missense	MUTYH	c.640C>T	p.Arg214Trp	0.00007	LP	P, P, LP	PubMed
1:156,106,982		G	A G	A G	A G	0.9735	missense	LMNA	c.1567G>A	p.Gly523Arg	0.00009	VUS	VUS, LP	PubMed
2:247,637,253		T C A	T T C A	T T C A	T T C A	0.9994	frameshift	MSH2	c.388_389del	p.Gln130ValfsTer2		LP	P	PubMed
4:106,320,294		G	A / G	A / G	A / G	0.7101	missense	PPA2	c.683C>T	p.Pro228Leu	0.00020	LP	P	PubMed
7:42,017,311		C	T / C	T / C	T / C	0.4905	missense	GLI3	c.1658G>A	p.Cys553Tyr		LP		
7:42,064,957		G A C T C	G / G A C T C	G / G A C T C	G / G A C T C	0.9912	frameshift	GLI3	c.1258_1261del	p.Glu420LeufsTer3		LP		
8:61,765,143		G	A / G	A / G	A / G	0.9759	stop_gained	CHD7	c.5981G>A	p.Trp1994Ter		LP		
8:145,140,500		C A G	C / C A G	C / C A G	C / C A G	0.9733	frameshift	GPA1	c.1477_1478del	p.Arg493GlyfsTer152	0.00013	LP		
9:107,546,633		A A A...T	A / A A A...T	A / A A A...T	A / A A A...T	0.3378	frameshift	ABCA1	c.6744_6748del	p.Phe2250ThrfsTer3		LP		
10:126,091,499		G	C / G	C / G	C / G	0.9334	stop_gained	OAT	c.897C>G	p.Tyr299Ter	0.00002	LP	P	PubMed
						0.7461	missense	JAM3	c.346G>A	p.Glu116Iys		LP	P	PubMed
						0.9622	missense	CLN5	c.578G>A	p.Cys193Tyr		LP	P	
						0.9203	frameshift	TTC8	c.963del	p.Met321IlefsTer15	0.00001	LP		



GitHub

<https://github.com/molgenis/vip>
<https://github.com/molgenis/vip-report>
<https://github.com/molgenis/vip-report-api>
<https://github.com/molgenis/vip-decision-tree>

Dennis Hendriksen
Bart Charbon, *et al.*

Pipeline has everything you expect



- Gene/transcript annotations (RefSeq, ENSEMBL)
- Transcript-specific effect predictions
- Allele frequencies (gnomAD, 1000G, GoNL, ..)
- MVLs (inhouse, VKGL, ClinVar, CLINVITAE, ..)
- Linkouts (OMIM, Orphanet, literature, ..)
- Inheritance matching (+CGD inheritance modes)
- Conservation scores (PhyloP, CADD, REMM, ..)
- Splice predictions (SpliceAI, MaxEntScan, ...)
- Gene panel filters (inclusive and exclusive)
- Quality filters, BAM file inspection, etc etc etc

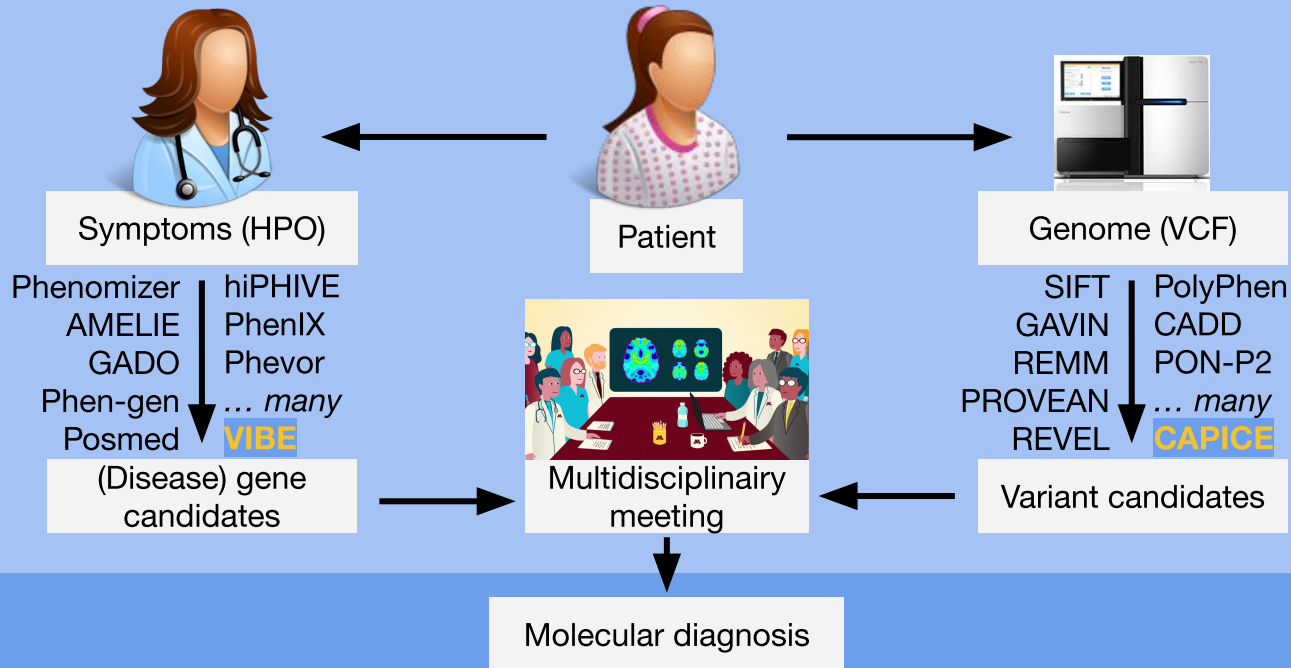
+
VIBE
CAPICE
others!

Zoom out a bit: FAIR Genomes

FAIRification: sharing & reusing healthcare & research data (FAIR genomes)

Combine 19,000 patients: large-scale Rare Disease cohort analysis (Solve-RD)

Genome diagnostics: Variant Interpretation Pipeline (VIP)



FAIR Genomes: re-use NGS in NL

- ➔ Previous NL sharing efforts (Fokkema *et al.* 2019): **better & faster variant classification**
- ➔ Consortium: **61** people from **14** Dutch institutes, working towards guideline & implementation

NGS analysis flow →



What was the phenotype of the patient?



What kind of sharing is allowed by the consent?



Which tissue was sampled?



Which sample prep kit was used?



What type of NGS machine was used?



What software was used to perform read mapping?



Which protocol was used to interpret the data?

The dream, when FAIR:
“I would like to please have all nationally available VCF files for PBMC samples from cardiomyopathy patients sequenced whole-exome on HiSeq machines processed with GATK 4.0, for which consent allows re-analysis.”

Join us at:



<https://github.com/fairgenomes>

in **1+MillionGenomes**

What is FAIR Genomes?

61 people from **14 institutes** (NL). F2F meetings, Zoom calls, focus workshops. **Interacting** with EJP-RD CDE, Solve-RD RD3, 1+MG, GA4GH, Phenopackets, X-omics, and others.

Currently **9 modules with 107 elements**:
Personal (13), Clinical (19), Material (14), Sample Preparation (9), Sequencing (12), Analysis (10), Informed Consent Form (9), Individual Consent (15), Study (6).

Together we define what meta data is needed to **find, share and reuse** NGS data in research and healthcare. Forming an evolving **semantic** model for properties and values.

Focus on being **harmonized** with EJP-RD Common Data Elements, RD3, PhenoPackets, MIABIS, etc. All models, coded lookups, applications **free & open source software**.

Join us at: <https://github.com/fairgenomes>

Together building semantic model

FAIR Genomes semantic metadata model

The FAIR Genomes semantic metadata model to power reuse of NGS data in research and healthcare. Version 0.0, 2020-12-19. This model consists of **9 modules** that contain **107 metadata elements** in total.

Module overview

Name	Description	Ontology	Nr. of elements
Study	A detailed examination, analysis, or critical inspection of a subject designed to discover facts about it.	NCIT:C63536	6
Personal	Data, facts or figures about an individual; the set of relevant items would depend on the use case.	NCIT:C90492	13
Informed consent form	A document explaining all the relevant information to assist an individual in understanding the expectations and risks in making a decision about a procedure. This document is presented to and signed by the individual or guardian.	NCIT:C16468	9
Individual consent	Consent by a patient to a surgical or medical procedure or participation in a clinical study after achieving an understanding of the relevant medical facts and the risks involved.	NCIT:C16735	15
Clinical	Data obtained through patient examination or treatment.	NCIT:C15783	19
Material	Natural substances derived from living organisms such as cells, tissues, proteins, and DNA.	NCIT:C43376	14
Sample preparation	A sample preparation for assay that preparation of nucleic acids for a sequencing assay.	OBI:0001902	9
Sequencing	The determination of complete (typically nucleotide) sequences, including those of genomes (full genome sequencing, de novo sequencing and resequencing), amplicons and transcriptomes.	EDAM:topic_3168	12
Analysis	Apply analytical methods to existing data of a specific type.	EDAM:operation_2945	10

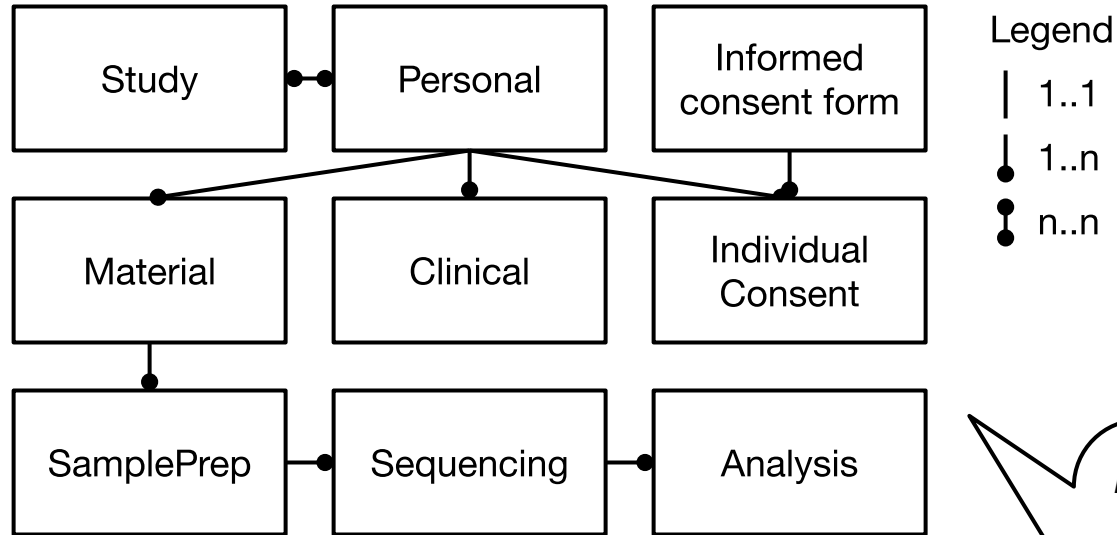
All modules, data elements and values are linked to **ontologies**.

You can use HL7 **NullFlavors** to explain why values are missing.

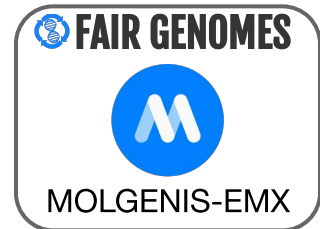
See:

<https://github.com/fairgenomes/fairgenomes-semantic-model>

Modules & links (“cardinality”)



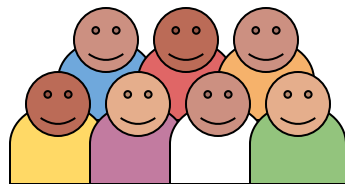
Derive relational database systems without effort



Core model: a YAML file

```
---
name: FAIR Genomes metadata model
description: The FAIR Genomes semantic metadata model to power reuse of NGS data in research and healthcare.
version: 0.0
date: 2020-12-19
lookupGlobalOptions: lookups/NullFlavors.txt
modules:
  - name: Personal
    description: Data, facts or figures about an individual; the set of relevant items would depend on the use case.
    ontology: NCIT:C90492 [http://purl.obolibrary.org/obo/NCIT_C90492]
    elements:
      - name: Personal identifier
        description: An alphanumeric identifier assigned to a specific patient.
        ontology: NCIT:C164337 [http://purl.obolibrary.org/obo/NCIT_C164337]
        values: UniqueID
      - name: Gender
        description: Biological sex is the quality of a biological organism based on reproductive function or organs.
        ontology: SIO:010029 [https://semanticscience.org/resource/SIO_010029.rdf]
        values: LookupOne [lookups/Gender.txt]
      - name: Genotypic sex
        description: A biological sex quality inhering in an individual based upon genotypic composition of sex chromosomes.
        ontology: PATO:0020000 [http://purl.obolibrary.org/obo/PATO_0020000]
        values: LookupOne [lookups/GenotypicSex.txt]
      - name: Country of residence
        description: Country of Residence at Enrollment.
        ontology: NCIT:C171105 [http://purl.obolibrary.org/obo/NCIT_C171105]
        values: LookupOne [lookups/Countries.txt]
      - name: Ethnicity
        description: The biological quality of membership in a social group based on a common heritage.
        ontology: SIO:001014 [http://semanticscience.org/resource/SIO_001014]
        values: LookupMany [lookups/Countries.txt]
```

Model is just the means to an end



Community development
How to FAIRify NGS data?



 **FAIR GENOMES**



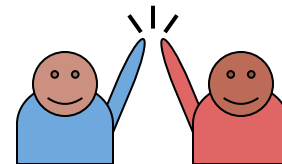
Semantic metadata model (YAML)



 **FAIR GENOMES**



Model transformer (Java app)



Community benefits:
Interoperable systems



 **FAIR GENOMES**



ART-DECOR

 **FAIR GENOMES**



Markdown

 **FAIR GENOMES**



MOLGENIS-EMX

 **FAIR GENOMES**



RDF-TTL

 **FAIR GENOMES**



OWL/XML

 **FAIR GENOMES**



... etc

Application #1: Nictiz

Gurnoor Singh, Jeroen Beliën, Sander de Ridder, K. Joeri van der Velde

are implementing a FAIR Genomes
ART-DECOR codebook for

Nictiz who develop & manage
information standards for exchange
of digital data in healthcare



<https://www.nictiz.nl/standaardisatie/art-decor/>

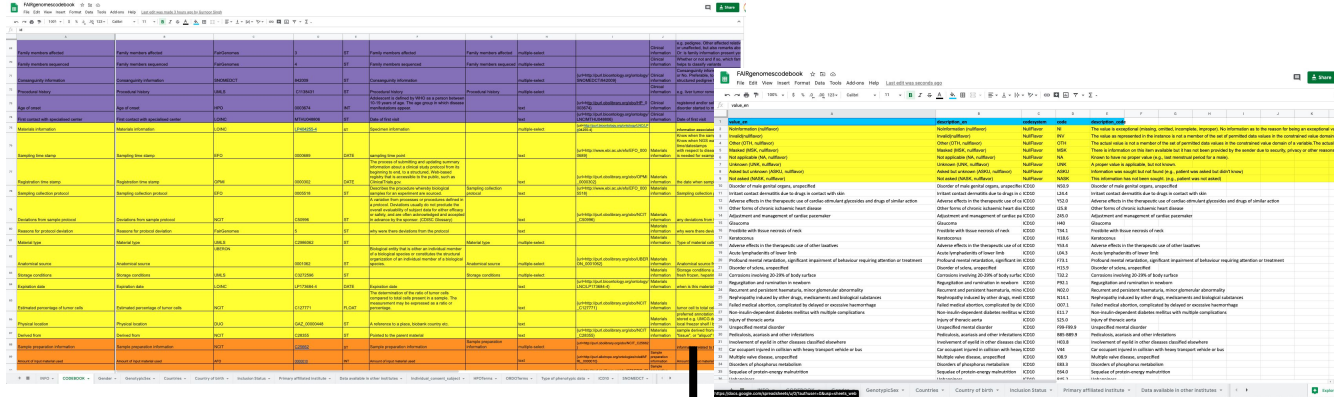


<https://art-decor.org/>

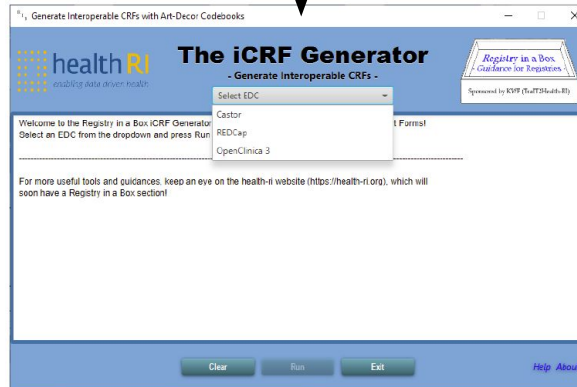
See:

https://github.com/fairgenomes/information/tree/master/fairgenomes_codebook_nictiz

Why an ART-DECOR codebook?



The image shows a complex spreadsheet titled 'F1000Research - ART-DECOR codebook'. It contains multiple columns of data, including medical terms, codes, and descriptions. A large black arrow points from the top of the spreadsheet down to the iCRF Generator application window.



iCRF Generator is a Java program that can generate the core of an interoperable electronic case report form (iCRF) for several of the major electronic data capture systems (EDCs)

de Ridder S and Beliën JAM. **The iCRF Generator: Generating interoperable electronic case report forms using online codebooks.** *F1000Research* 2020, 9:81 (<https://doi.org/10.12688/f1000research.21576.2>)

Application #2: MOLGENIS

K. Joeri van der Velde, Gurnoor Singh, Fernanda de Andrade, Dieuwke Roelofs-Prins, Lennart F. Johansson, Bart Charbon & MOLGENIS Team

MOLGENIS: scientific data platform with flexible model, tailor to your needs, 100+ instances running

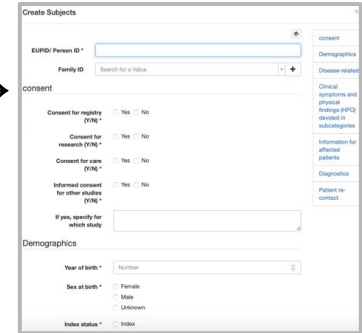


FAIR Genomes app

<https://github.com/fairgenomes/fairgenomes-semantic-model/tree/main/transformation-output/molgenis-emx>

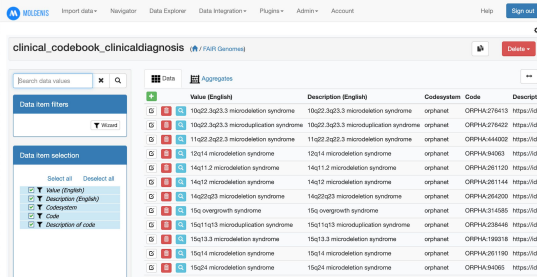
Powering ERNs

- GENTURIS →
- Ithaca
- Skin
- Cranio

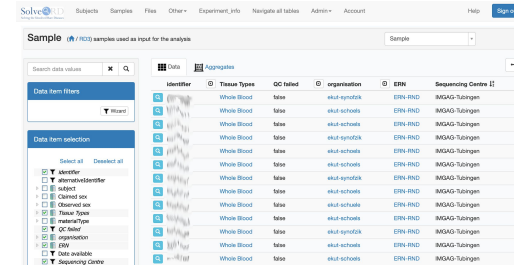
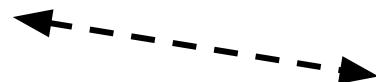


Solve-RD RD3 (https://github.com/molgenis/RD3_database)

Detailed NGS sample tracking



Value (English)	Description (English)	Codexsystem	Code	Describe
19q22.3q23.3 microdeletion syndrome	19q22.3q23.3 microdeletion syndrome	orphanet	ORPHA:27613	https://orpha.net/oc/27613
19q22.3q23.3 microduplication syndrome	19q22.3q23.3 microduplication syndrome	orphanet	ORPHA:27632	https://orpha.net/oc/27632
19q22.3q23.3 microdeletion syndrome	19q22.3q23.3 microdeletion syndrome	orphanet	ORPHA:44402	https://orpha.net/oc/44402
19q14 microdeletion syndrome	19q14 microdeletion syndrome	orphanet	ORPHA:94263	https://orpha.net/oc/94263
19q11.2 microdeletion syndrome	19q11.2 microdeletion syndrome	orphanet	ORPHA:261120	https://orpha.net/oc/261120
19q12 microdeletion syndrome	19q12 microdeletion syndrome	orphanet	ORPHA:261144	https://orpha.net/oc/261144
19q22q3 microdeletion syndrome	19q22q3 microdeletion syndrome	orphanet	ORPHA:264200	https://orpha.net/oc/264200
19q10.1q10.2 microdeletion syndrome	19q10.1q10.2 microdeletion syndrome	orphanet	ORPHA:314585	https://orpha.net/oc/314585
19q11.3 microduplication syndrome	19q11.3 microduplication syndrome	orphanet	ORPHA:238446	https://orpha.net/oc/238446
19q13.3 microdeletion syndrome	19q13.3 microdeletion syndrome	orphanet	ORPHA:199218	https://orpha.net/oc/199218
19q14 microdeletion syndrome	19q14 microdeletion syndrome	orphanet	ORPHA:261190	https://orpha.net/oc/261190
19q24 microdeletion syndrome	19q24 microdeletion syndrome	orphanet	ORPHA:94265	https://orpha.net/oc/94265



Identifier	Tissue Types	QC Includ	organisation	ERN	Sequencing Centre
skit-spr028	Whole Blood	Yes	skit-spr028	ERN_RND	IMGAG-Tuigen
skit-sch016	Whole Blood	Yes	skit-sch016	ERN_RND	IMGAG-Tuigen
skit-spr028	Whole Blood	Yes	skit-spr028	ERN_RND	IMGAG-Tuigen
skit-sch016	Whole Blood	Yes	skit-sch016	ERN_RND	IMGAG-Tuigen
skit-spr028	Whole Blood	Yes	skit-spr028	ERN_RND	IMGAG-Tuigen
skit-sch016	Whole Blood	Yes	skit-sch016	ERN_RND	IMGAG-Tuigen
skit-spr028	Whole Blood	Yes	skit-spr028	ERN_RND	IMGAG-Tuigen
skit-sch016	Whole Blood	Yes	skit-sch016	ERN_RND	IMGAG-Tuigen
skit-spr028	Whole Blood	Yes	skit-spr028	ERN_RND	IMGAG-Tuigen
skit-sch016	Whole Blood	Yes	skit-sch016	ERN_RND	IMGAG-Tuigen

Public demo online

FAIR GENOMES Import data ▾ Navigator Data Explorer Plugins ▾

FAIR GENOMES

VERSION 0.1 PUBLIC DEMO. THIS IS A PROTOTYPE FOR EVALUATION PURPOSES ONLY. DO NOT IMPORT ANY SENSITIVE DATA.

A national guideline to promote optimal (re)use of NGS data in research and healthcare.



Study



Personal



General Consent



Material



Clinical



Individual Consent



Sampleprep



Sequencing



Analysis



Codebooks



Information



Contribute

Please visit & give us feedback!

<https://fairgenomes-acc.gcc.rug.nl>

FAIR GENOMES Import data ▾ Navigator Data Explorer Plugins ▾ Help Sign in

CLINICAL

Clinical ID *

Visit 001

Unique label or human-interpretable identifier for this Clinical record

Phenotypic terms *

MESH:D010641

Unobserved phenotypes

HL7:C0442737

Type of phenotypic data *

DC:DCMIType

Clinical diagnosis

SNOMEDCT:39154008

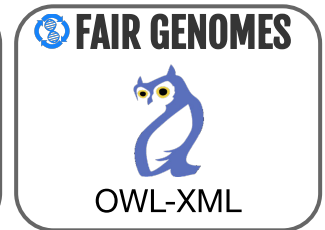
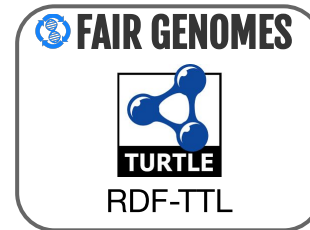
Genetic diagnosis

- Acanthosis nigricans-insulin resistance-muscle cramps-acral enlargement syndrome
- Achalasia-microcephaly syndrome
- Acral peeling skin syndrome
- Acrocallosal syndrome
- Acrocardiofacial syndrome
- Acrocephalosyndactyly**
- Acromegaloïd facial appearance syndrome

Also: expanded ontologies

We will create a 'FairGenomes' ontology including terms currently **not available** in ontologies:

- Genotypic sex (ie. karyotypes, ~10 items)
- NGS kits (~625 items, source: BioCompare)
- Sequencing instruments (adding ~10 items)
- Dutch hospitals (~110 items)
- *More definitions as needed*

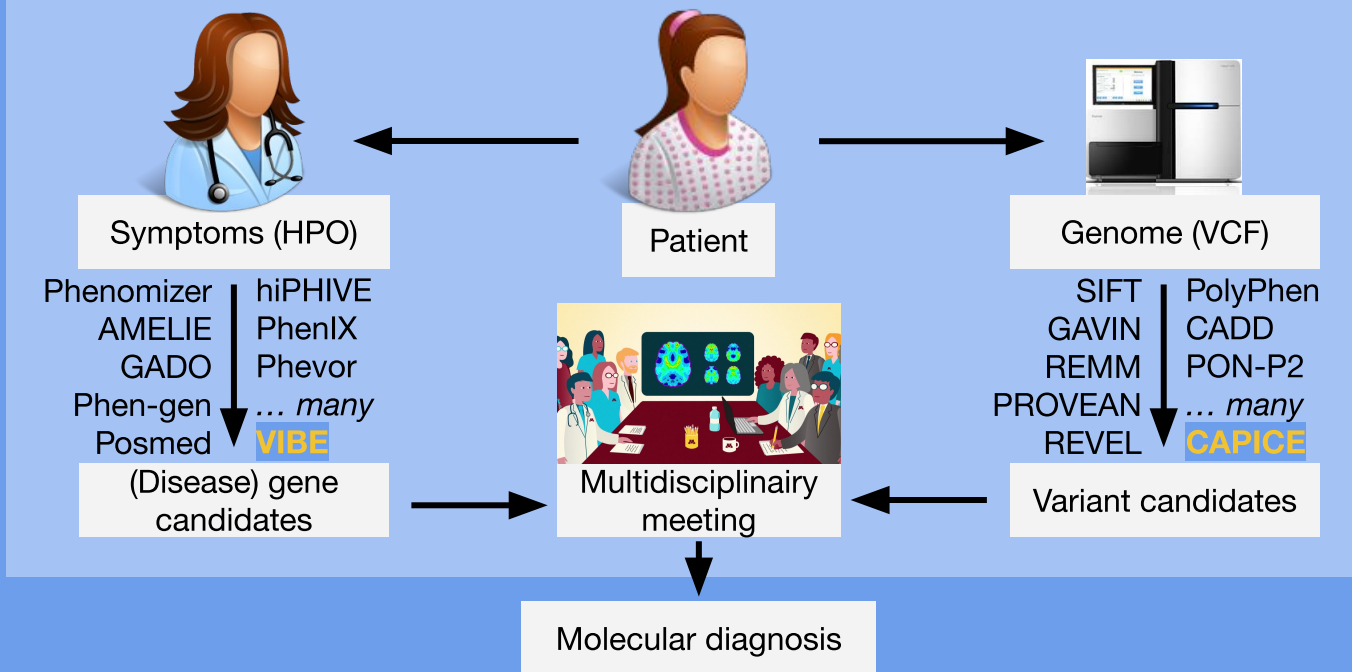


Large-scale analysis: Solve-RD

FAIRification: sharing & reusing healthcare & research data (FAIR genome)

Combine 19,000 patients: large-scale Rare Disease cohort analysis (Solve-RD)

Genome diagnostics: Variant Interpretation Pipeline (VIP)



- Omics analysis in health research is often performed using cohorts of patients with similar disease phenotypes
- For rare diseases often, these cohorts are small



- EU funded research project
- 1.1.2018 – 31.12.2022 (5 year project)
- 22 partners from 10 countries
- Coordinated by Olaf Riess & Holm Graessner (Tübingen)
- Co-coordinated by Han Brunner (Nijmegen) and Anthony Brookes (Leicester)

Two streams of data

- 19,000 Whole Exome Sequencing samples of rare disease patients for which no genetic diagnosis was made.

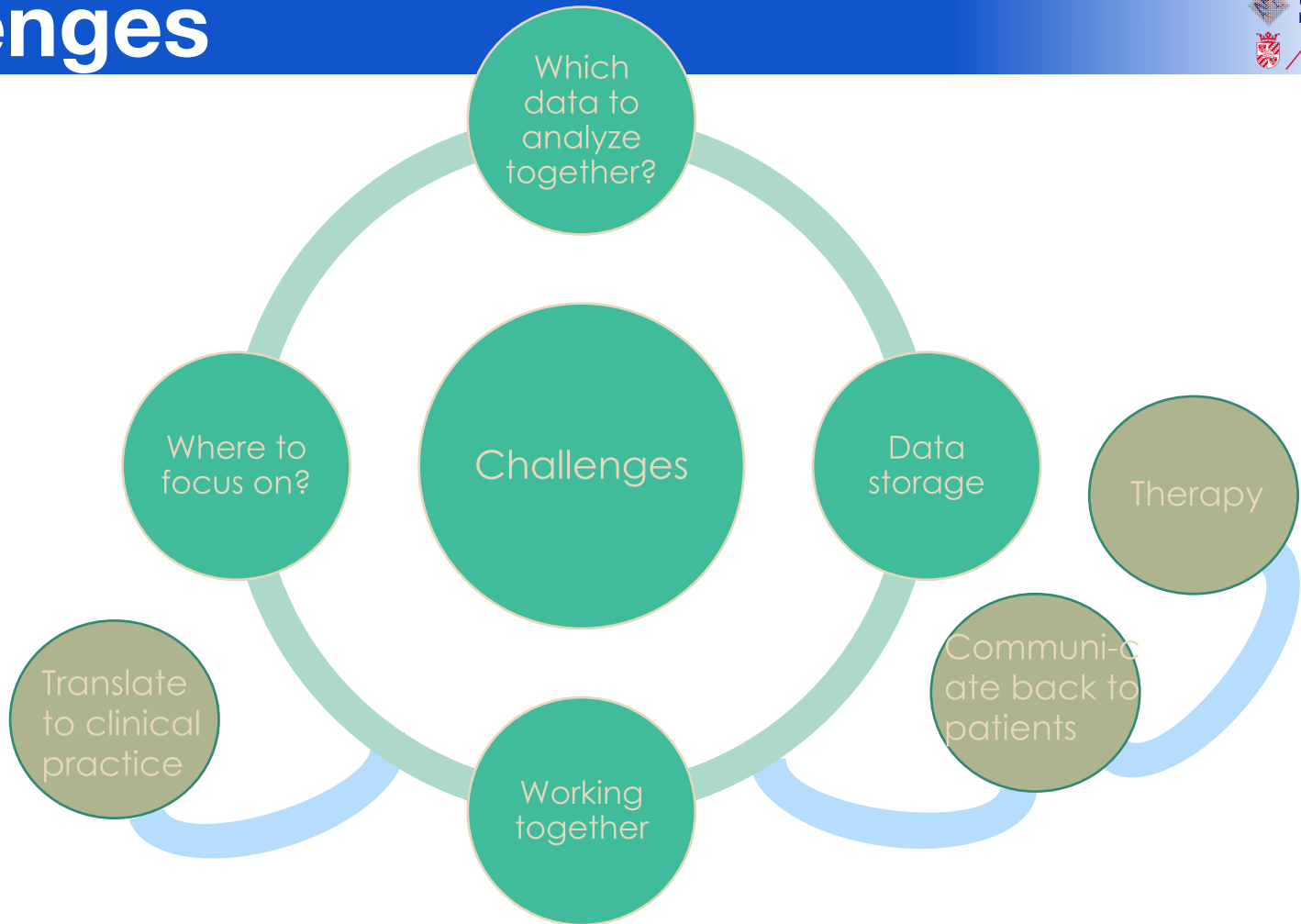
 - Newly generated novel omics data of those same patients
 - ◆ WGS (2,000)
 - ◆ Long-read sequencing (500)
 - ◆ Deep-WES
 - ◆ Transcriptomics
 - ◆ Proteomics
 - ◆ Methylation
- (2,000 + 120 multi-omics)

- **Core group of 4 European Reference Networks:** ERN-RND, ERN-EURO-NMD, ERN-ITHACA, ERN-GENTURIS
- **Associated networks:** 6 additional ERNs and 2 Undiagnosed Patient Programmes (Italy, Spain)
- **Existing RD infrastructures:** RD-Connect/ELIXIR, Orphanet, HPO, EuroGentest, Canadian Models and Mechanisms Network
- **Patient organisations:** EURORDIS, Genetic Alliance UK

Potentials

- Make new diagnoses
- Discover new geno-phenotype connections
- Discover new syndromes
- Discover target for treatment of a rare disease
- Develop new or improved analysis methods
- Bring together knowledge and know-how
- Ideal use-case to test functionality of infrastructure within challenging setting
- Ideal use-case to test functionality of (new) file format standards

Challenges



Which data to analyze together?

Challenge: Phenotypic similarity

→ Patients with similar phenotypes may have pathogenic variants in the same gene.



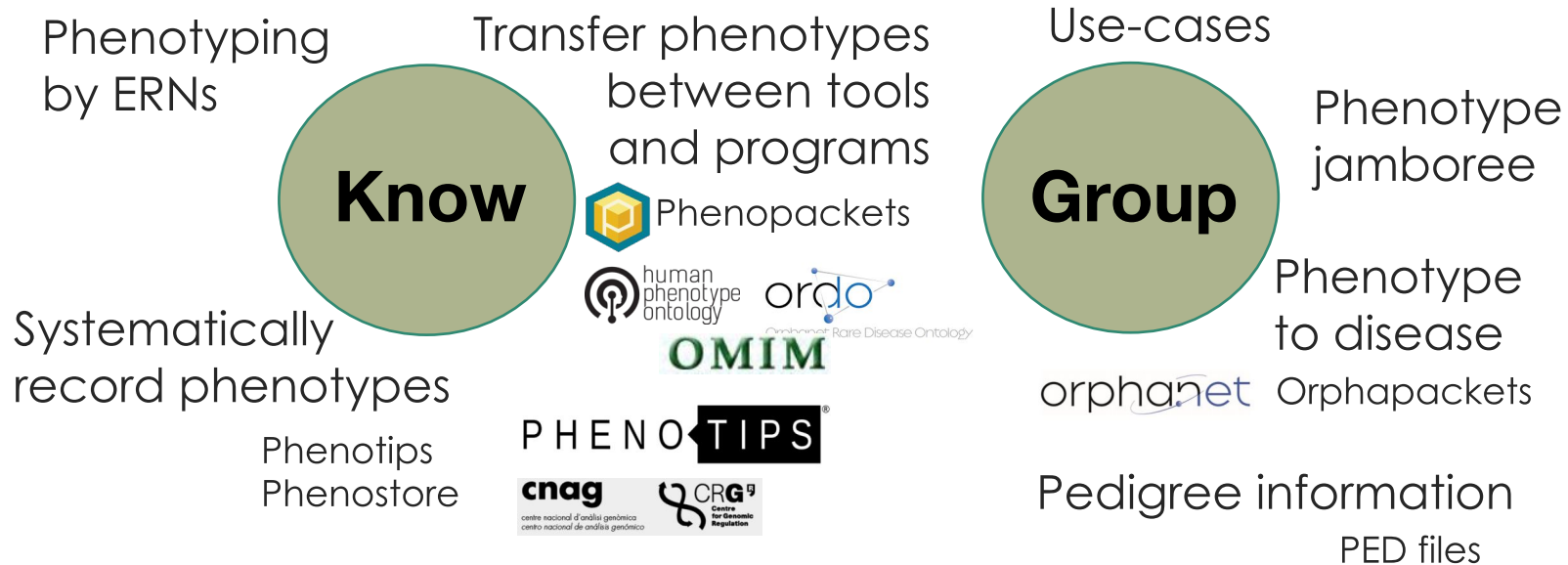
Know



Group

Which data to analyze together?

Challenge: Phenotypic similarity



Which data to analyze together?

Challenge: Technical similarity

→ WES data

- ◆ Produced at different centers
- ◆ Using different sequencers
- ◆ Using different enrichment kits

→ Different regions are covered

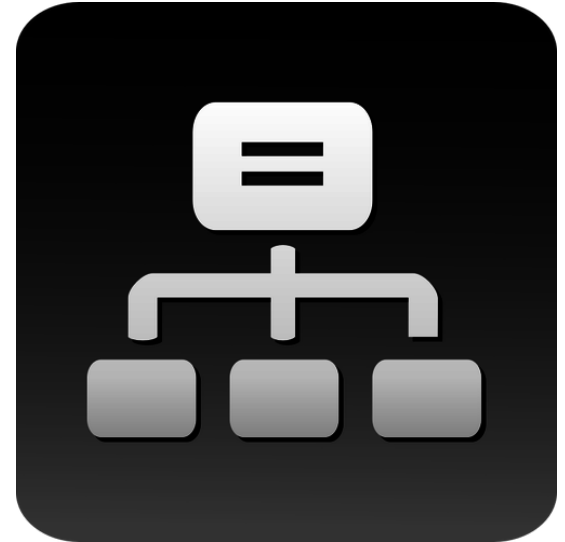
→ Different dynamic in coverage

Agilent_SureSelect_CRE_V1_54Mb
Agilent_SureSelect_CRE_V2_67Mb
Agilent_SureSelect_v1_39Mb
Agilent_SureSelect_v2_46Mb
Agilent_SureSelect_v3_51Mb
Agilent_SureSelect_v4_51Mb
Agilent_SureSelect_v5_50Mb
Agilent_SureSelect_v6_60Mb
Agilent_SureSelect_v7_36Mb
Baylor_Custom_v2.1_42Mb
Broad_AgilentCustom_v1.1_33Mb
Broad_Custom_v1_60Mb
IDT_xGen_ExomeResearchPanel_39MB
Illumina_NexteraExome_37Mb
Illumina_NexteraExpandedExome_62Mb
Illumina_Truseq_v1.2_45Mb
Illumina_TruseqExome_37Mb
Illumina_TruseqExpandedExome_62Mb
Illumina_TruseqOne_v1_12Mb
Nimblegen_SeqCapEZExome_v2_36Mb
Nimblegen_SeqCapEZExome_v2_47Mb
Nimblegen_SeqCapEZExome_v3_64Mb
Nimblegen_SeqCapEZMedExome_47Mb
Nimblegen_SeqCapEZMedExomePlusMT_47Mb
Prague_Custom_41Mb
Twist_Bioscience_Twist_Human_RefSeq_Panel
WGS

Which data to analyze together?

Solution: Technical similarity

- Realign all data using same pipeline
 - ◆ CNAG-CRG DNA analysis pipeline



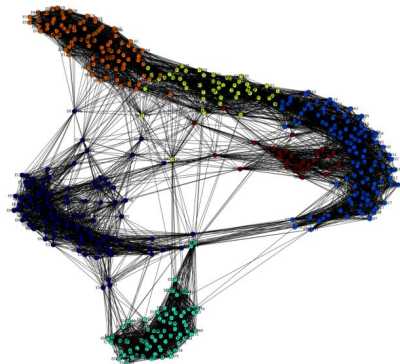
Which data to analyze together?

Solution: Technical similarity

→ Different dynamic in coverage

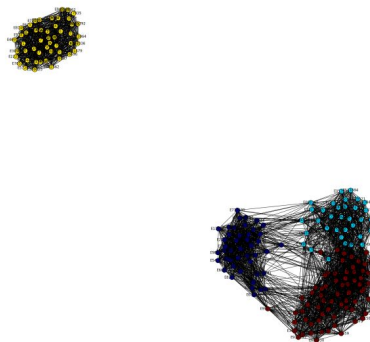
◆ CNV detection

Louvain clusters of samples with similar coverage patterns



Illumina Nextera Expanded Exome 62 Mb

Louvain clusters of samples with similar coverage patterns



Broad Custom 60 Mb

ClusterWES

Agilent_SureSelect_CRE_V1_54Mb
Agilent_SureSelect_CRE_V2_67Mb
Agilent_SureSelect_v1_39Mb
Agilent_SureSelect_v2_46Mb
Agilent_SureSelect_v3_51Mb
Agilent_SureSelect_v4_51Mb
Agilent_SureSelect_v5_50Mb
Agilent_SureSelect_v6_60Mb
Agilent_SureSelect_v7_36Mb
Baylor_Custom_v2.1_42Mb
Broad_AgilentCustom_v1.1_33Mb
Broad_Custom_v1_60Mb
IDT_xGen_ExomeResearchPanel_39MB
Illumina_NexteraExome_37Mb
Illumina_NexteraExpandedExome_62Mb
Illumina_Truseq_v1.2_45Mb
Illumina_TruseqExome_37Mb
Illumina_TruseqExpandedExome_62Mb
Illumina_TruseqOne_v1_12Mb
Nimblegen_SeqCapEZExome_v2_36Mb
Nimblegen_SeqCapEZExome_v2_47Mb
Nimblegen_SeqCapEZExome_v3_64Mb
Nimblegen_SeqCapEZMedExome_47Mb
Nimblegen_SeqCapEZMedExomePlusMT_47Mb
Prague_Custom_41Mb
Twist_Bioscience_Twist_Human_RefSeq_Panel
WGS

Where to focus on?

Opportunity

→ Through large cohort size possibly relatively large number of same condition.

Challenges

- Disease can be caused by
 - ◆ different variant types
 - ◆ variants located all over the genome
 - ◆ Somatic / mosaic variants
- Is the detected variant causal for the disease?

**Often not
detectable
by WES**

Where to focus on? | finding variants

- Group patients by phenotype
 - ◆ Determine consanguinity by RoH analysis
 - ◆ Gene burden analysis
- Specific use-cases
 - ◆ e.g. read expansion analysis in ataxias through long-read sequencing
 - ◆ Compound heterozygous variants of different types (e.g. SNV and CNV)

UNSOLVED CASES*

Definition: Rare disease cases with an inconclusive exome/genome

Number: 19,000 unsolved exomes/genomes

Main activities: Perform standardised collation and re-analysis

**in collaboration with all ERNs, Undiagnosed Disease Initiatives and further associated partners*

1

SPECIFIC ERN COHORTS

Definition: Disease group specific cohorts from four core ERNs (exome available)

Number: a) 2,000 WGS for more complete (non-)coding sequence & CNV/SVs etc.;

b) 500 long-read WGS;

c) >2,000 cases novel omics approaches

Main activities: Conduct „beyond the exome“ approaches

2

ULTRA RARE RARE DISEASES

Definition: Phenotypically most special/remarkable patients with a rare disease without an exome

Number: 1,200 exomes (300 per core ERN)

Main activities: Carry out phenotype jamborees and exome analysis

3

4

THE UNSOLVABLES

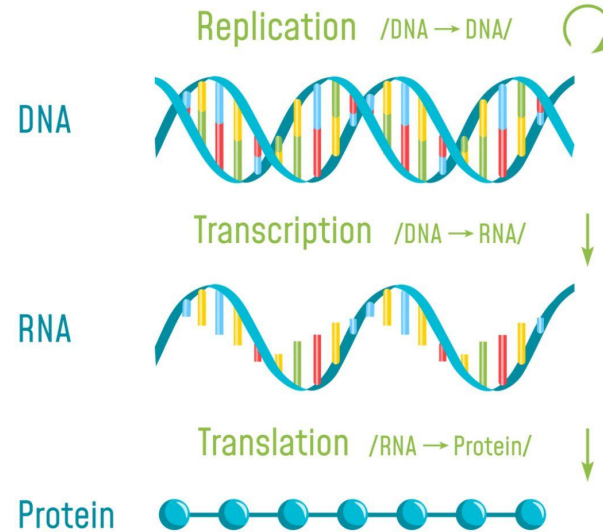
Definition: Highly recognisable clinically defined diseases / syndromes for which no disease gene was identified yet despite WES/WGS and considerable research invested

Number: 120 syndromes/ diseases

Main activities: apply all -omics tools to ‚crack‘ the „Unsolvable“

Where to focus on? | causality

- Genomics
 - WES/WGS/Long-read/Deep-WES
 - Methyloomics
- Transcriptomics
- Proteomics
- Seeding awards for confirmation through animal model

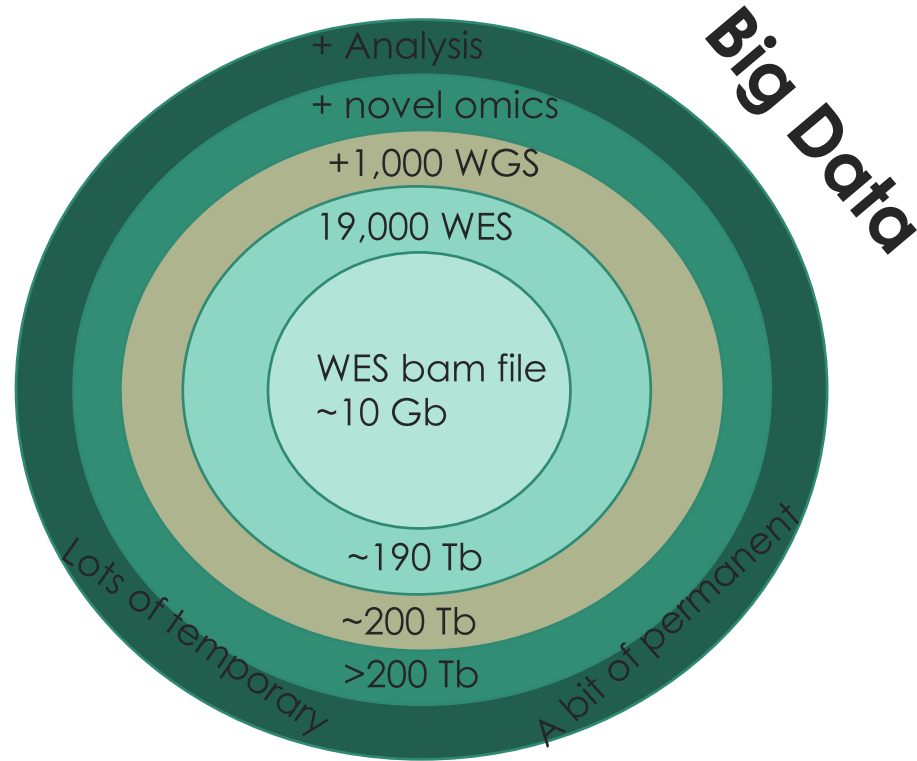


Where to focus on?

- Different working groups focus on different aspects
- Data Analysis Task Force (DATF)
 - ◆ Finding different types of variants in WES and novel omics data
 - E.g. SNV-inde, CNV, SV, methylation, splice-variants, etc.
 - ◆ Producing statistics on variants
 - E.g. Gene-burden and meta-analysis
- Data interpretation Task Force (DITF)
 - ◆ Determine use-cases
 - ◆ Determining which variants are pathogenic
 - ◆ Determining which variants are causal for the patients disease

Challenge

- Size
- Versioning
- Tracking files



Solutions

- Large storing capacity at the European Genome-Phenome Archive (EGA)
- Persistent identifiers | EGA Accession numbers
- Tracking data using RD3

Data Storage | RD3



MOLGENIS

RD3: Rare disease data about data

→ Metadata on subject, samples, files and experiments

The screenshot displays the RD3 web application interface, which is organized into several hierarchical views:

- Subject View:** Shows a search bar and a list of filters including Identifier, Claimed sex, FamilyID, MaternalID, PaternalID, Affected status, disease, phenotype, hasNotPhenotype, PhenopacketsID, and variant.
- Sample View:** Shows a search bar and filters for Identifier, alternativeIdentifier, subject, Claimed sex, Observed sex, Tissue Types, materialType, OC failed, and variant. It lists sample IDs like ED00003, DNC0003, P0001167, ED00005, DNC0005, and P0001165.
- Labinfo View:** Shows a search bar and filters for Identifier, sample, Enrichment kit, Library Source, flowcell, barcode, samplePosition, library, sequencer, and seqType. It lists Illumina_NexteraExpandedExome_G2Mb and Illumina_NexteraExpandedExome_G2Mb.
- Data Table:** A table with columns for EGA Accession Number, Filename, Checksum, typeFile, samples, and Created. It lists 19 rows of data, including accession numbers like EGAF00002756209 and filenames like 0544bb11c49404ec66460507dc3b31.

Challenges

- Data originates from many different locations
- Researchers are located at many different locations

Goal: Jointly analyze and interpret data

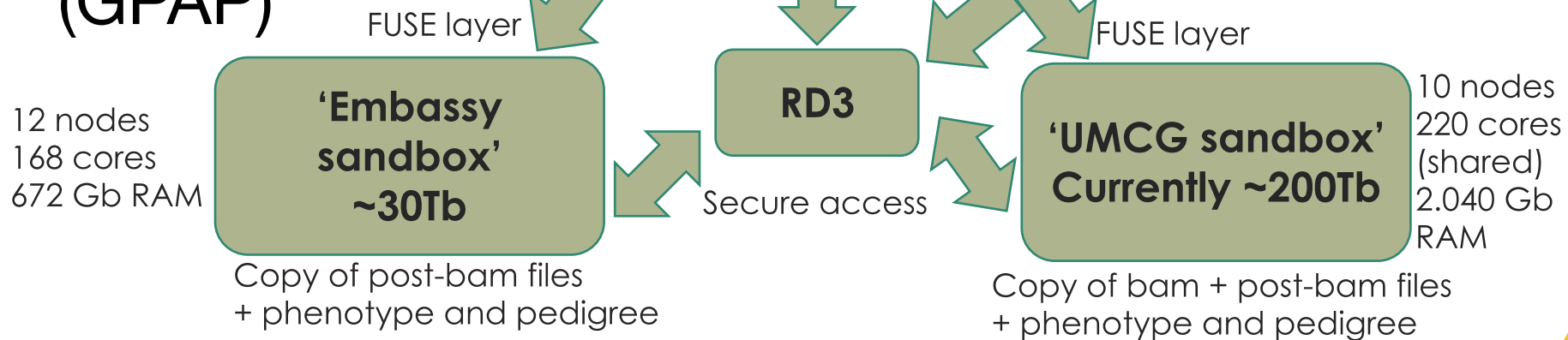
Solutions

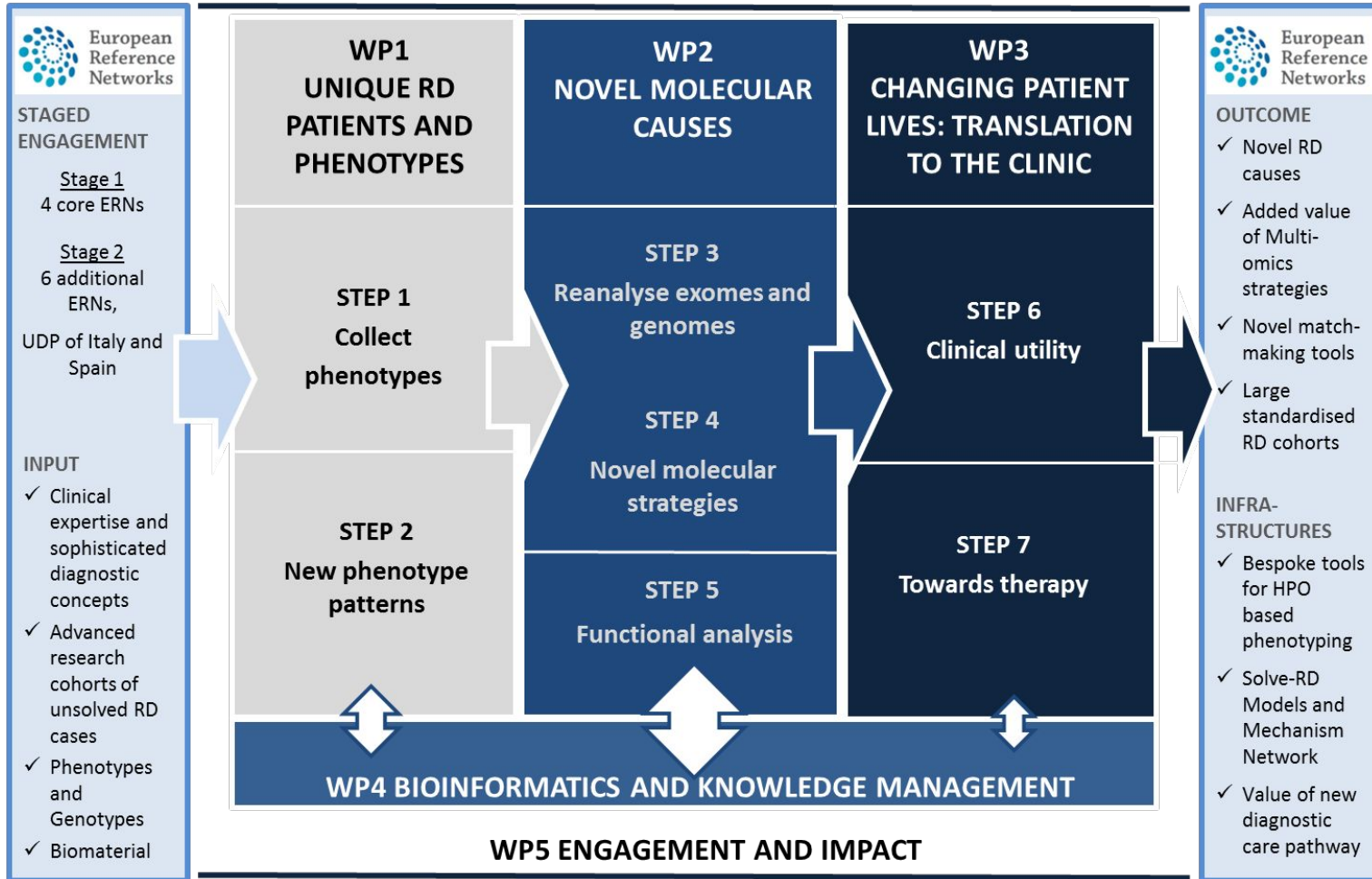
- Direct access to data at EGA
 - Filesystem in UserSpace (FUSE)
 - Still optimizing performance

Working together

Solutions

- Two 'sandboxes'
- Geno-Phenotype Analysis Platform (GPAP)





Take home messages

- Large cohorts of rare disease patients create the potential to diagnose patients and discover causal genes and variants
- Clear phenotypic information is essential
- Depending on the question different type of omics data are informative
- Large cohorts produce lots of data
-

Acknowl.: VIP/VIBE/CAPICE

Team VIP: Cleo van Diemen, Trijnie Dijkhuizen, Martine Meems-Veldhuis, Kristin Abbott, Inge Mulder, Birgit Raddatz, Marielle van Gijn, Dennis Hendriksen, Bart Charbon, Roan Kanninga, Gerben van der Vries, Lennart Johansson, Mariska Slofstra, Morris Swertz. **Team CAPICE:** Shuang Li, Robert Sietsma, Dick de Ridder, Aalt van Dijk, Dimitrios Soudis, Leslie Zwerwer, Patrick Deelen, Dennis Hendriksen, Bart Charbon, Marielle van Gijn, Kristin Abbott, Birgit Raddatz, Cleo van Diemen, Mieke Kerstjens-Frederikse, Richard Sinke, Morris Swertz. **Team VIBE:** Sander van den Hoek, Freerk van Dijk, Dennis Hendriksen, Cleo van Diemen, Lennart Johansson, Kristin Abbott, Patrick Deelen, Birgit Raddatz, Morris Swertz. **Plus many other wonderful (inter)national collaborators, colleagues, students. Thank you!**



university of
 groningen



umcg



BBMRI.nl
Biobanking and
BioMolecular resources
Research Infrastructure
The Netherlands



ZonMw



Solving the Unsolved Rare Diseases



Netherlands Organisation
for Scientific Research



WAGENINGEN
UNIVERSITY & RESEARCH



EUROPEAN JOINT PROGRAMME
RARE DISEASES



Disclaimer: picture taken in the times before COVID-19

Gurnoor Singh¹, K. Joeri van der Velde², Jeroen Beliën⁴, Jasmin Böhmer³, Daphne Stemkens⁵, Lisenka Vissers¹, Jeroen van Reeuwijk¹, Saskia Hiltemann⁷, Lennart F. Johansson², Nienke van der Stoep⁶, Daoud Sie⁴, Janneke Weiss⁴, Geert Frederix³, Marco Roos⁶, Erik van Iperen⁸, Terry Vrijenhoek³, Folkert W. Asselbergs³, Joris van Montfrans³, Rolf Sijmons², Hanneke van Deutekom³, Pieter Neerincx², Fernanda de Andrade², Anna Niehues¹, Hindrik H.D. Kerstens¹⁰, Mark Thompson⁶, Rajaram Kaliyaperumal⁶, Annika Jacobsen⁶, Katy Wolstencroft^{6,14}, Ies Nijman³, Marcel Nelen¹, Ariaan Siezen¹, Koen ten Hove¹, Nine Knoers², Christian Gilissen¹, Hans Scheffer¹, Stefan Willems³, Wendy van Zelst-Stams¹, Helger Ijntema¹, Kim Elsink³, Bart de Koning⁹, Bauke Ylstra⁴, Erik Sijm⁴, Patrick Kemmeren¹⁰, Henne Holstege⁴, Christine Staiger¹¹, Bastiaan Tops¹⁰, Susanne Rebers¹², David van Zessen⁷, Valesca Retèl¹², Edwin Cuppen¹³, Peter van Tintelen³, Esther van Enckevort², Lieneke Steeghs¹, Salome Scholtens², Jeroen Laros⁶, Leon Mei⁶, Cor Oosterwijk⁵, Andrew Stubbs⁷, Peter A.C. 't Hoen¹, Mariëlle van Gijn², Morris Swertz²

from:

¹Radboud University Medical Center, Nijmegen, The Netherlands, ²University Medical Center Groningen, The Netherlands, ³University Medical Center Utrecht, The Netherlands, ⁴Amsterdam University Medical Centers, location VUmc, NL, ⁵VSOP - Dutch Patient Alliance for Rare and Genetic Diseases, ⁶Leiden University Medical Center, The Netherlands, ⁷Erasmus Medical Center, Rotterdam, The Netherlands, ⁸Durrer Center for Cardiovascular Research, Utrecht, The Netherlands, ⁹Maastricht University Medical Center, The Netherlands, ¹⁰Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands, ¹¹Dutch Techcentre for Life Sciences, Utrecht, The Netherlands, ¹²Netherlands Cancer Institute, Amsterdam, The Netherlands, ¹³Hartwig Medical Foundation, Amsterdam, The Netherlands, ¹⁴Leiden Institute for Advanced Computer Science, Leiden University, Leiden, NL

THANKS!

The Solve-RD project team



*MOLGENIS & Genomics Coordination Center,
UMCG*

