Biomedical provenance metadata to support FAIR genomic research data: *an intracranial aneurysms use case*

BC2 workshop T7
8th of September 2025

Alban Gaignard, Institut du Thorax, CNRS, Nantes, France

Biomedical context

ICAN cohort: 34 univ. hospitals / 3000 subjects

3.000 ToF MRIs
600 whole genomes (analysis in progress)

Predict the evolution ?

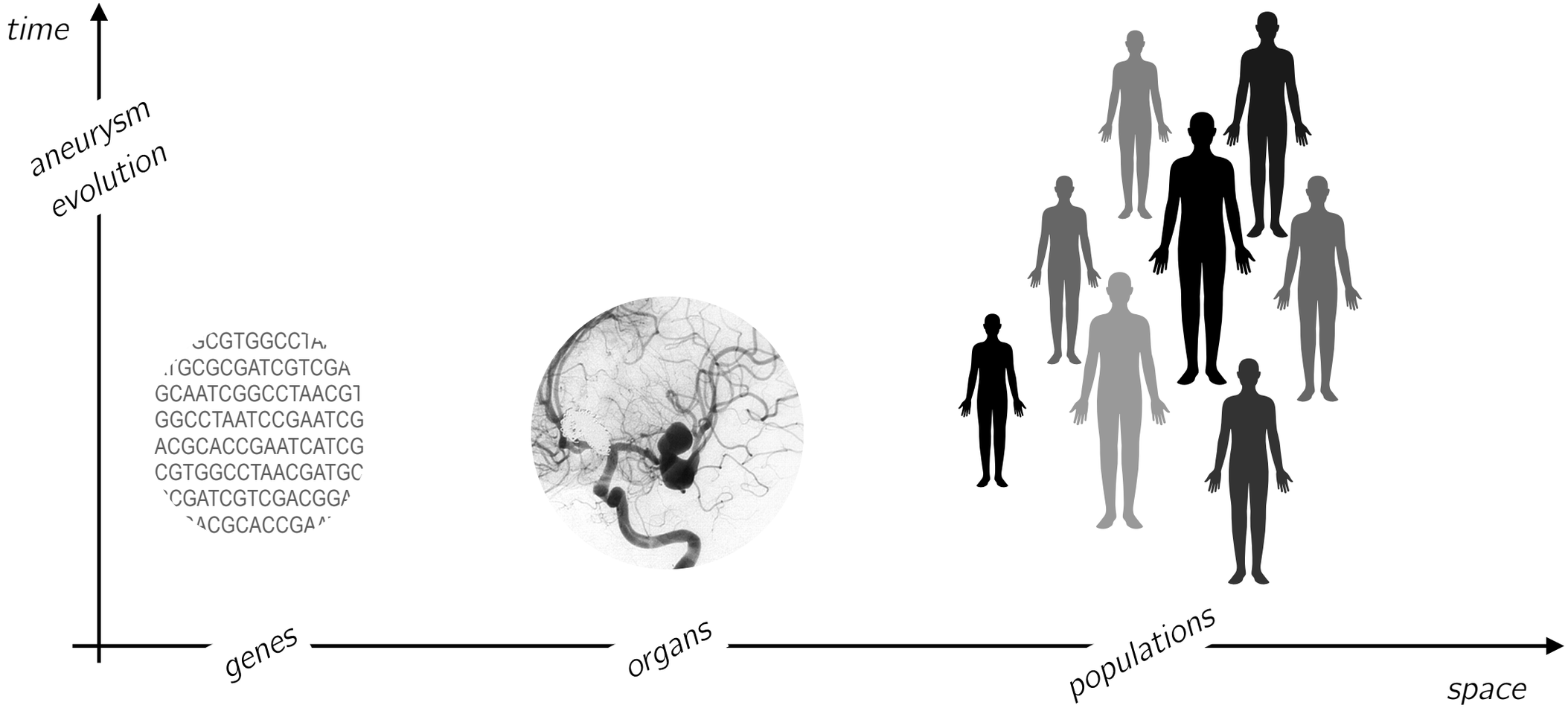
Understand the pathology ?

Prevention strategies ?

Bourcier R, Chatel S, Bourcereau E, Jouan S, Marec HL, Daumas-Duport B, Sevin-Allouet M, Guillon B, Roualdes V, Riem T, Isidor B, Lebranchu P, Connault J, Tourneau TL, Gaignard A, Loirand G, Redon R, Desal H; ICAN Investigators. Understanding the Pathophysiology of Intracranial Aneurysm: The ICAN Project. Neurosurgery. 2017 Apr 1;80(4):621-626. https://doi.org/10.1093/neuros/nyw135

Multi-factorial disease $\longrightarrow$ multi-scale data

- Intracranial aneurysms: a complex & multifactorial disease
- Inter-disciplicary efforts needed for a better understanding of the pathology
- Specific data produced at very specific scales



Multi-modal data analysis

Using deep learning for an automatic detection and classification of the vascular bifurcations along the Circle of Willis

Rafic Nader [a], Romain Bourcier [a], Florent Autrusseau [a,b,*]

[a] Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
[b] Nantes Université, Polytech'Nantes, LTeN, U-6607, Rue Ch. Pauc, 44306, Nantes, France

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Vascular bifurcations
Circle of Willis
Deep learning

ABSTRACT

Most of the intracranial aneurysms (ICA) occur on a specific portion of the cerebral vascular tree named the Circle of Willis (CoW). More particularly, they mainly arise onto fifteen of the major arterial bifurcations constituting this circular structure. Hence, for an efficient and timely diagnosis it is critical to develop some methods being able to accurately recognize each Bifurcation of Interest (BoI). Indeed, an automatic extraction of the bifurcations presenting the higher risk of developing an ICA would offer the neuroradiologists a quick glance at the most alarming areas. Due to the recent efforts on Artificial Intelligence, Deep Learning turned out to be the best performing technology for many pattern recognition tasks. Moreover, various methods have been particularly designed for medical image analysis purposes. This study intends to assist the neuroradiologists to promptly locate any bifurcation presenting a high risk of ICA occurrence. It can be seen as a Computer Aided Diagnosis scheme, where the Artificial Intelligence facilitates the access to the regions of interest within the MRI. In this work, we propose a method for a fully automatic detection and recognition of the bifurcations of interest forming the Circle of Willis. Several neural networks architectures have been tested, and we thoroughly evaluate the bifurcation recognition rate.

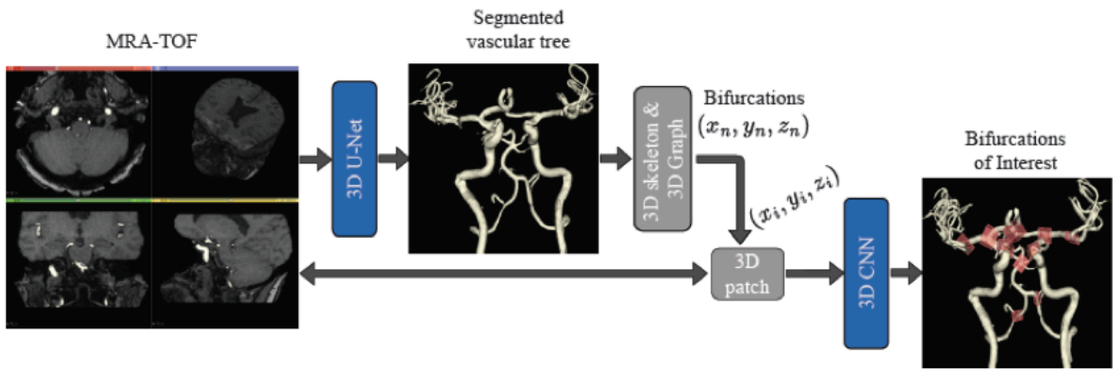


Fig. 1: General flowchart of the bifurcation recognition process.

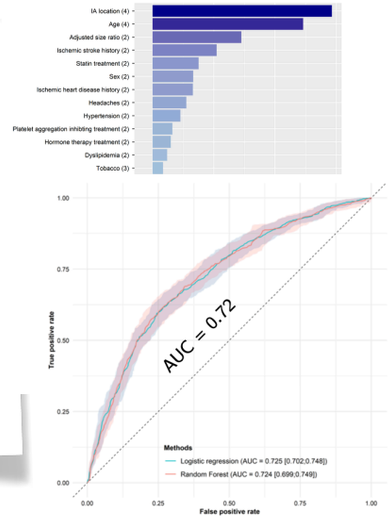
Observational Study > J Neurol Neurosurg Psychiatry. 2021 Feb;92(2):122-128.
doi: 10.1136/jnnp-2020-324371. Epub 2020 Oct 23.

Location of intracranial aneurysms is the main factor associated with rupture in the ICAN population

Olivia Rousseau [1], Matilde Karakachoff [1], Alban Gaignard [2], Lise Bellanger [3], Philippe Bijlenga [4], Pacôme Constant Dit Beaufils [1], Vincent L'Allinec [5], Olivier Levrier [6], Pierre Aguettaz [7], Jean-Philippe Desilles [8], Caterina Michelozzi [9], Gaultier Marnat [10], Anne-Clémence Vion [2], Gervaise Loirand [2], Hubert Desal [11], Richard Redon [2], Pierre-Antoine Gourraud [1], Romain Bourcier [12]; ICAN Investigators

Collaborators, Affiliations + expand

PMID: 33097563 DOI: 10.1136/jnnp-2020-324371

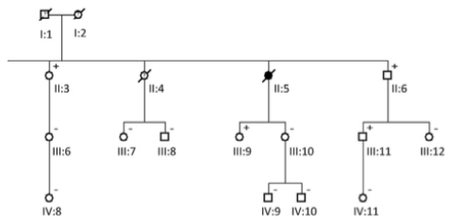


ARTICLE

Rare Coding Variants in *ANGPTL6* Are Associated with Familial Forms of Intracranial Aneurysm

Romain Bourcier,[1,2] Solena Le Scouarnec,[1] Stéphanie Bonnaud,[1,3] Matilde Karakachoff,[1,3] Emmanuelle Bourcereau,[3] Sandrine Heurtebise-Chrétien,[1] Céline Menguy,[1] Christian Dina,[1,3] Floriane Simonet,[1,3] Alexis Moles,[4] Cédric Lenoble,[2] Pierre Lindenbaum,[1] Stéphanie Chatel,[1,3] Bertrand Isidor,[5] Emmanuelle Génin,[6] Jean-François Deleuze,[7] Jean-Jacques Schott,[1,3] Hervé Le Marec,[1,3] ICAN Study Group, Gervaise Loirand,[1,3,8] Hubert Desal,[1,2,8,*] and Richard Redon[1,3,8,*]

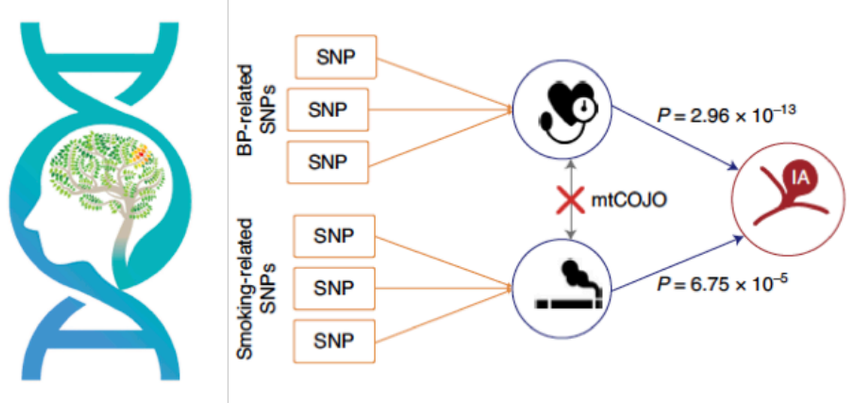
Intracranial aneurysms (IAs) are acquired cerebrovascular abnormalities characterized by localized dilation and wall thinning in intracranial arteries, possibly leading to subarachnoid hemorrhage and severe outcome in case of rupture. Here, we identified one rare nonsense variant (c.1378A>T) in the last exon of *ANGPTL6* (Angiopoietin-Like 6)—which encodes a circulating pro-angiogenic factor mainly secreted from the liver—shared by the four tested affected members of a large pedigree with multiple IA-affected case subjects. We showed a 50% reduction of ANGPTL6 serum concentration in individuals heterozygous for the c.1378A>T allele (p.Lys460Ter) compared to relatives homozygous for the normal allele, probably due to the non-secretion of the truncated protein produced by the c.1378A>T transcripts. Sequencing *ANGPTL6* in a series of 94 additional index case subjects with familial IA identified three other

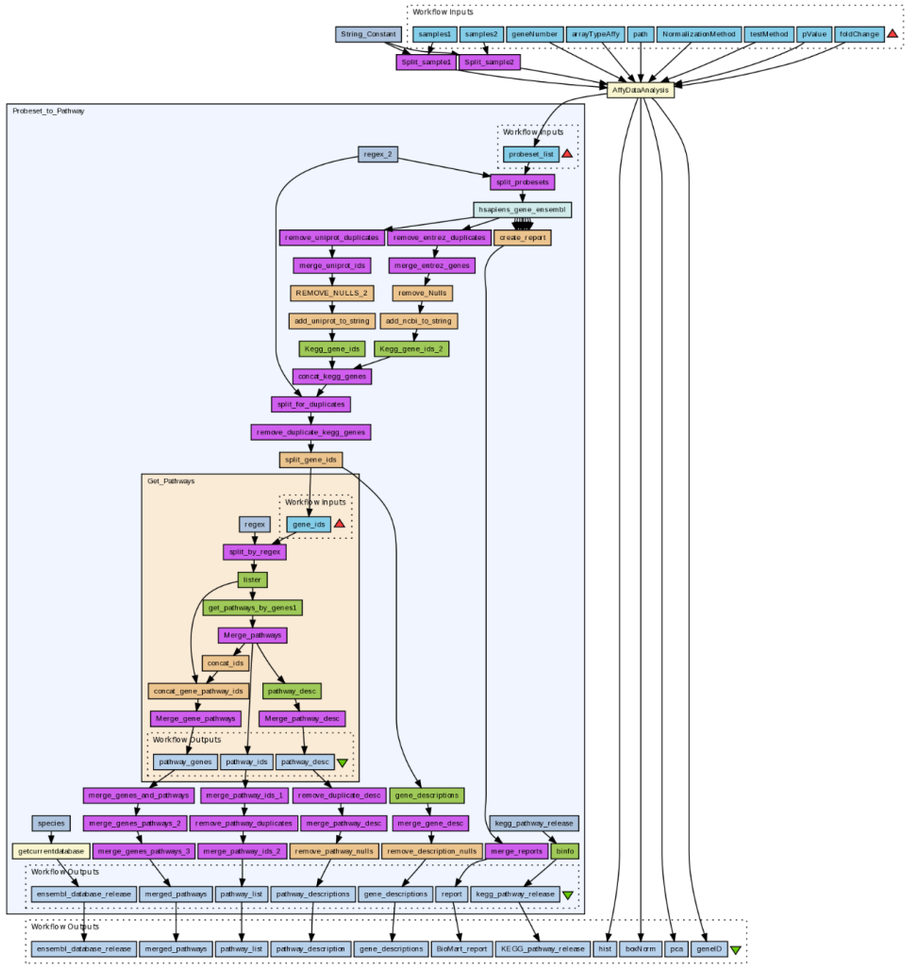


C

Filtering steps	Remaining variants	
Individuals :	*III-1*	*III-5*
Functional variants	25,674	23,456
MAF < 0,1% in ExAC (NFE)	456	436
Shared by III.1 & III.5	29	
In IBD haplotypes	10	
Shared by all affected	8	
Predicted LOF	1 (*ANGPTL6*)	

Genome-wide association study of intracranial aneurysms identifies 17 risk loci and genetic overlap with clinical risk factors





Workflows to enhance trust in scientific results :

→ automation (scalability)

→ abstraction (methods sharing)

→ traceability (provenance)

nextflow

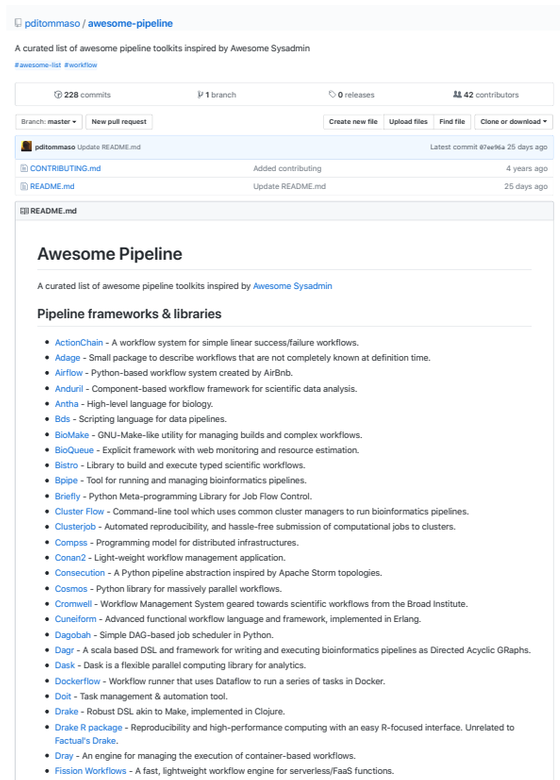
snakemake

Taverna

TensorFlow

COMMON WORKFLOW LANGUAGE

Galaxy COMMUNITY HUB



Galaxy platform

Galaxy COMMUNITY HUB

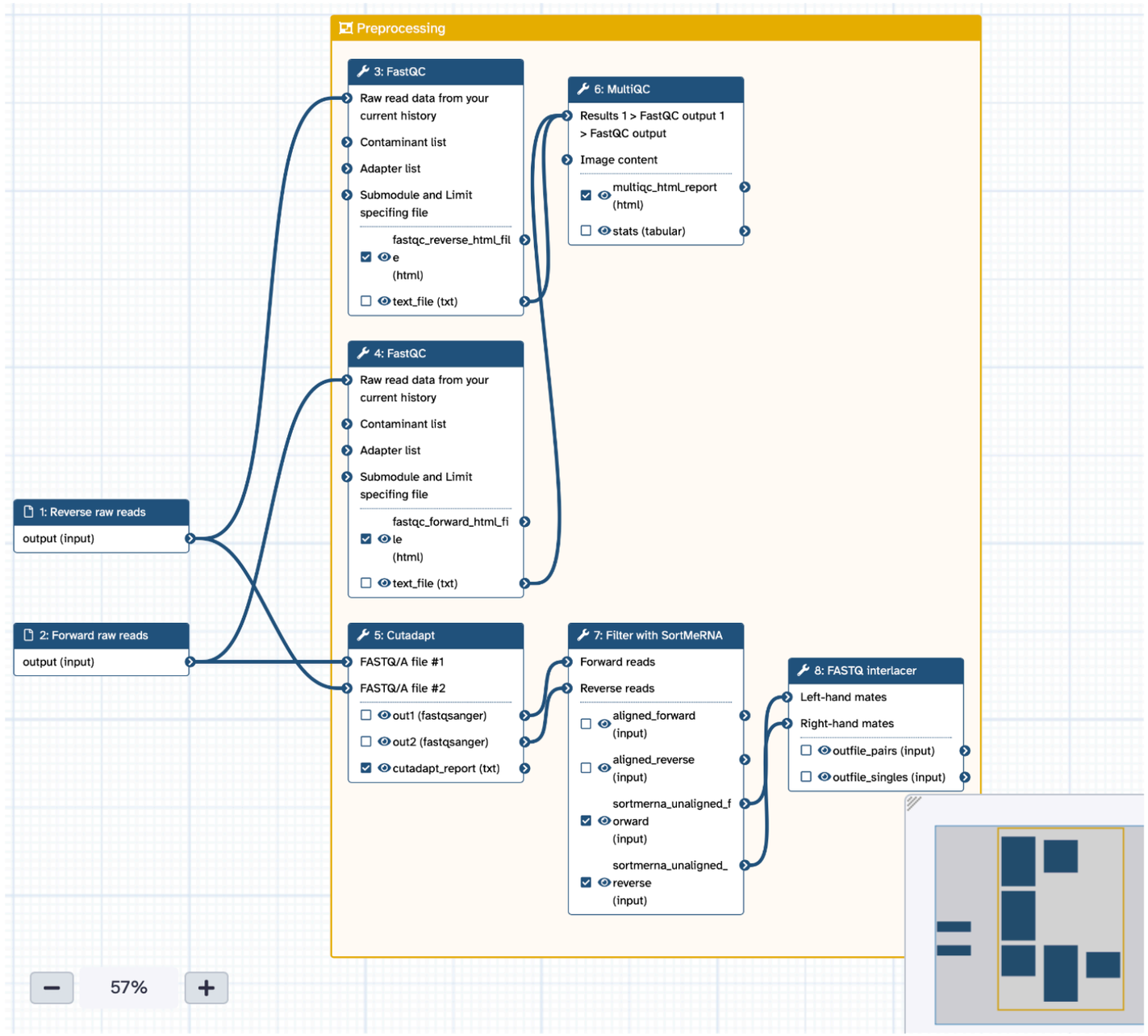
Galaxy

Upload
Tools
Workflows
Visualization
Histories
Pages

Login or Register

Workflow Preview

Download | Import | Run | ×

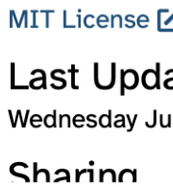


About This Workflow

Metatranscriptomics analysis using microbiome RNA-seq data - Workflow 1: Preprocessing - Version 2

Author

bebatut



All published Workflows by bebatut

Creators

- Bérénice Batut
- Pratik Jagtap
- Subina Mehta
- Ray Sajulga
- Emma Leith
- Praveen Kumar
- Saskia Hiltemann
- Paul Zierep
- Christine Oger

Description

Metatranscriptomics analysis using microbiome RNA-seq data (short)

Tags

microbiome

License

MIT License

Last Updated

Wednesday Jun 18th 8:56:01 2025 GMT+2

Sharing

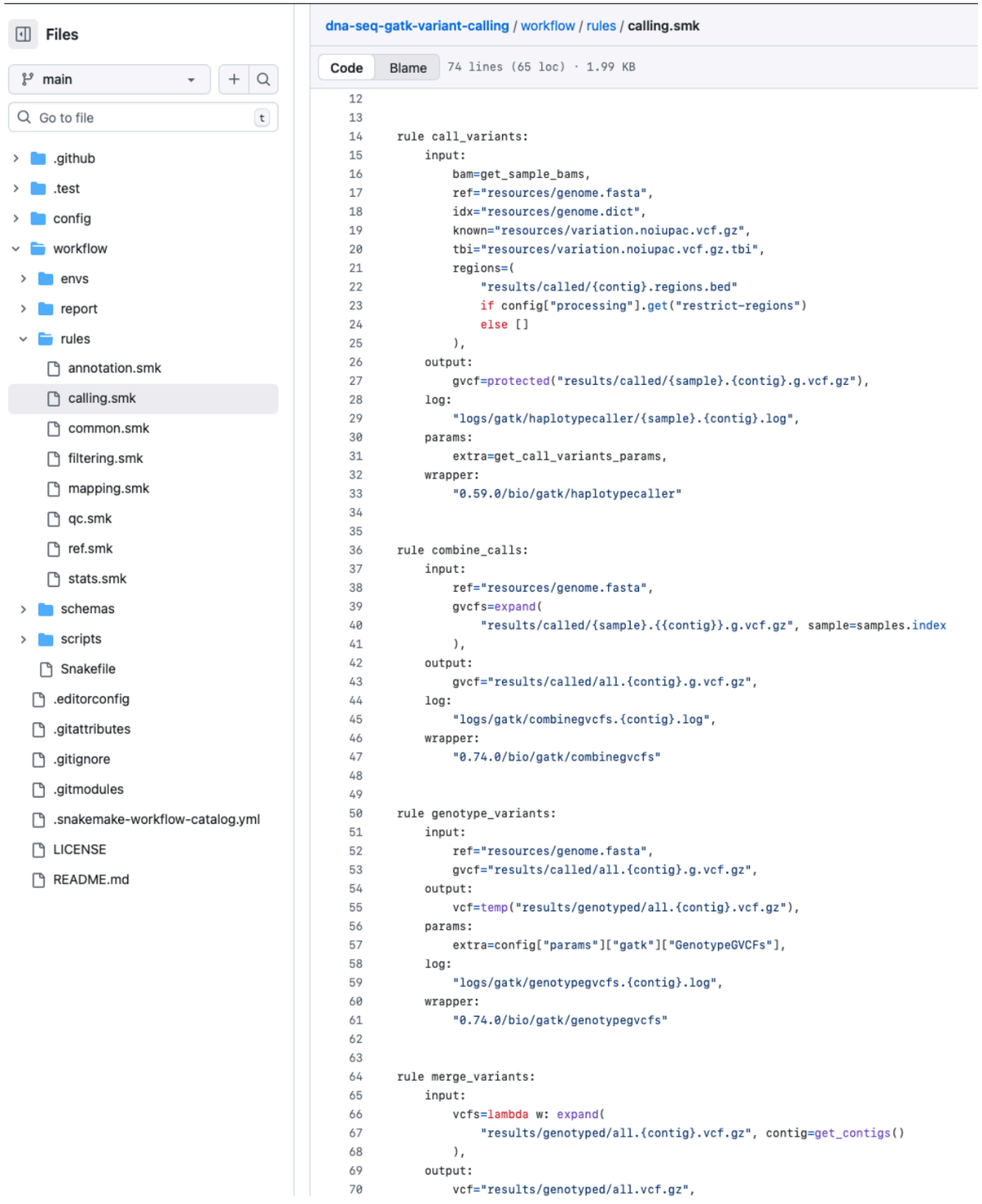
http://usegalaxy.fr http://usegalaxy.eu

Other workflow languages / engines

- When more control is needed on
  - data organization
  - storage
  - HPC clusters / GPUs ..

- Rule-based data tranformation

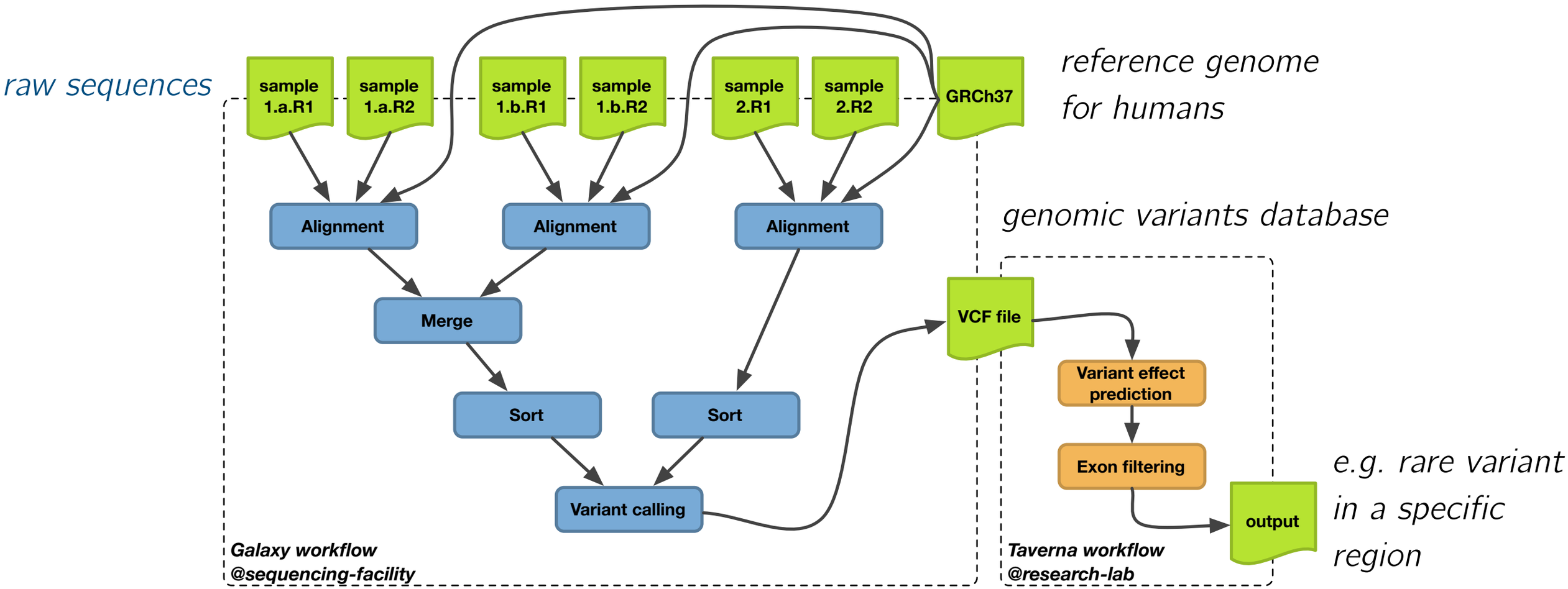
- Programming skills are needed



Massively produced data

Can we **reuse** data rather than **re-compute** ?

Reusing processed data ?



« Which alignment algorithm was used when predicting this pathogenic score ? »

« A new version of a reference genome is available, which genome was used when predicting these phenotypes ? »

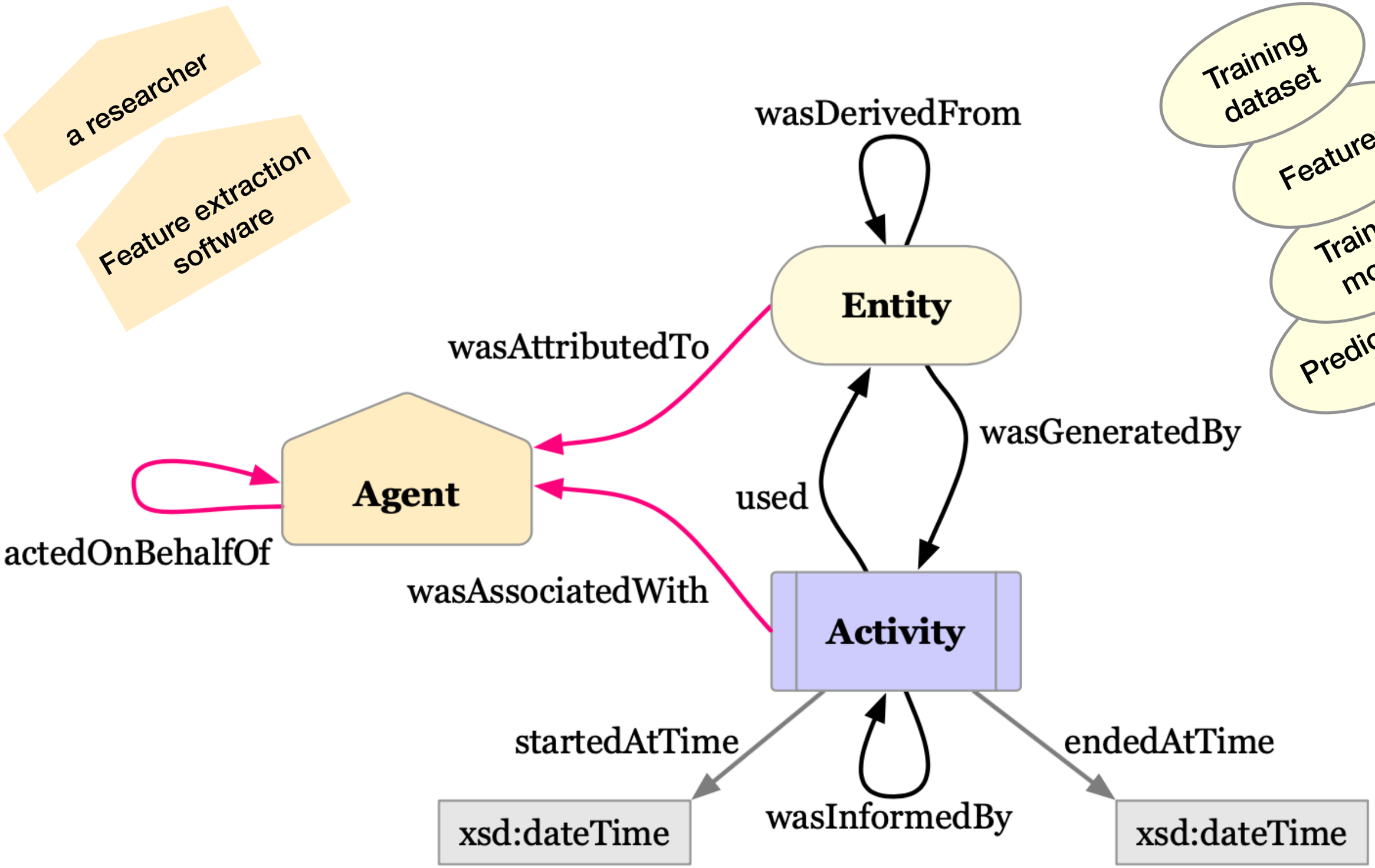
Need for an overall tracking of provenance over multiple workflows !

Hopefully, we have

W3C PROV

a researcher

Feature extraction software



Training dataset

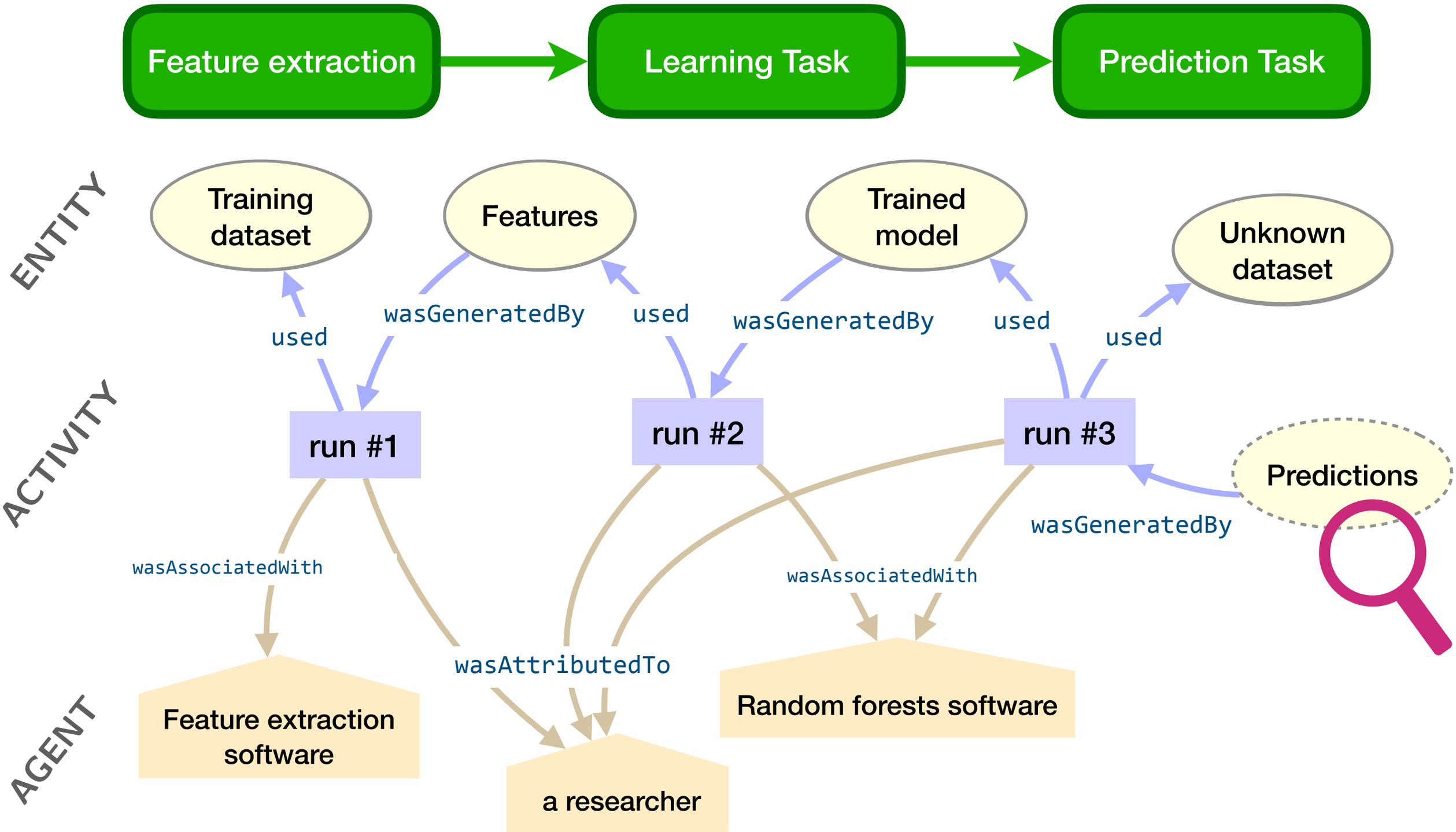
Features

Trained model

Predictions

Scikit-learn run #1

Linked data-algorithms-softwares-esearchers



PROV is domain-agnostic 🤩 🤔

Open Access | Published: 06 December 2016

Sharing brain mapping statistical results with the neuroimaging data model

Camille Maumet ✉, Tibor Auer, Alexander Bowring, Gang Chen, Samir Das, Guillaume Flandin, Satrajit Ghosh, Tristan Glatard, Krzysztof J. Gorgolewski, Karl G. Helmer, Mark Jenkinson, David B. Keator, B. Nolan Nichols, Jean-Baptiste Poline, Richard Reynolds, Vanessa Sochat, Jessica Turner & Thomas E. Nichols

Scientific Data **3**, Article number: 160102 (2016) | Cite this article

https://doi.org/10.1038/sdata.2016.102

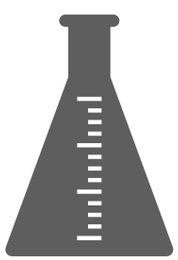
Research | Open Access | Published: 06 January 2022

End-to-End provenance representation for the understandability and reproducibility of scientific experiments using a semantic approach

Sheeba Samuel ✉ & Birgitta König-Ries

Journal of Biomedical Semantics **13**, Article number: 1 (2022) | Cite this article

https://doi.org/10.1186/s13326-021-00253-1



Research | Open Access | Published: 31 January 2022

Structure-based knowledge acquisition from electronic lab notebooks for research data provenance documentation

Max Schröder ✉, Susanne Staehlke, Paul Groth, J. Barbara Nebe, Sascha Spors & Frank Krüger

Journal of Biomedical Semantics **13**, Article number: 4 (2022) | Cite this article

https://doi.org/10.1186/s13326-021-00257-x

TaPP 2021 PAPERS A USENIX Publication ACCEPTED PAPERS

Astronomical Pipeline Provenance: A Use Case Evaluation

Authors:
Michael A. C. Johnson, *Institute of Data Science (DLR) and Max Planck Institute for Radio Astronomy;* Marcus Paradies and Marta Dembska, *Institute of Data Science (DLR);* Kristen Lackeos, Hans-Rainer Klöckner, and David J. Champion, *Max Planck Institute for Radio Astronomy;* Sirko Schindler, *Institute of Data Science (DLR)*

https://doi.org/10.48550/arXiv.2109.10759

Towards a Provenance Management System for Astronomical Observatories

Mathieu Servillat ✉, François Bonnarel, Catherine Boisson, Mireille Louys, Jose Enrique Ruiz & Michèle Sanguillon

Conference paper | First Online: 09 July 2021

Part of the Lecture Notes in Computer Science book series (LNISA,volume 12839)

https://doi.org/10.1007/978-3-030-80960-7_20

Many expectations ...

Comparability, transparency, confidence +

- **Citing** researchers and organisations
- Identifying **critical** data / software **resources** associated to scientific results
- Identifying possible **bias** when reusing / sharing datasets / pre-trained models

We need **guidance** for documenting Life-Science provenance

MIRAPIE

for genomic workflow ?

WF#1 Rare variants analysis

Raphaël Blanchet

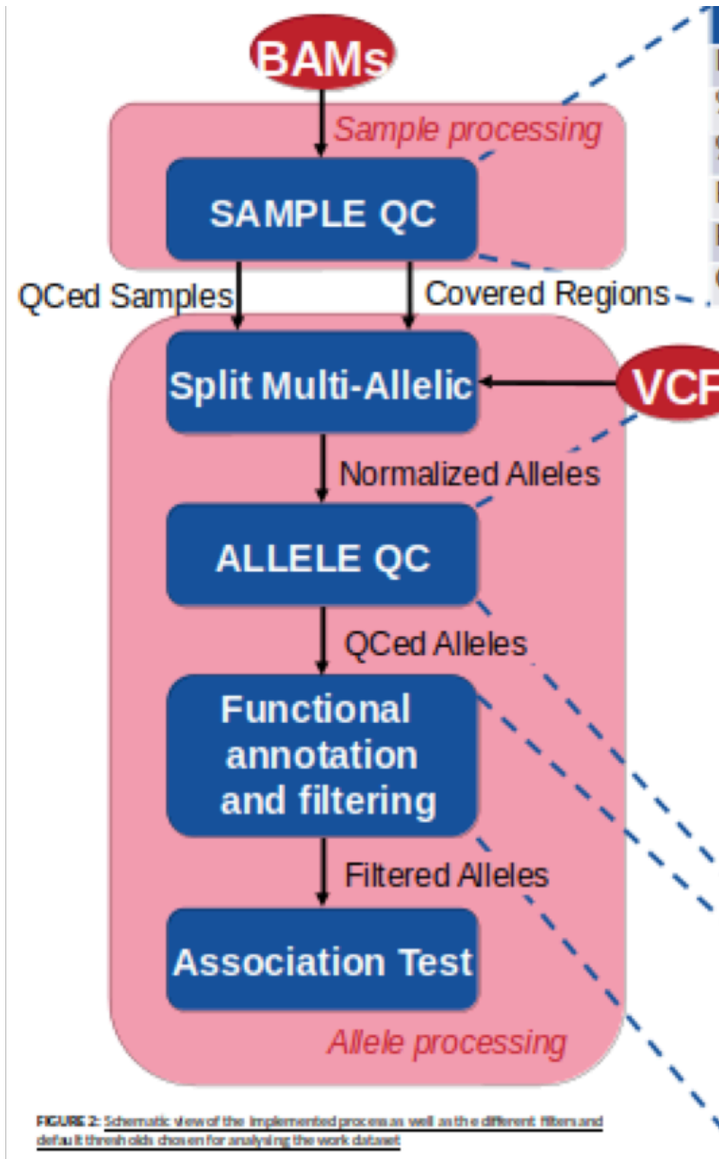


FIGURE 2: Schematic view of the implemented process as well as the different filters and default threshold chosen for analysing the work dataset

SAMPLE QC
Insert Size
% of unmapped/duplicated reads
% of IN-target reads
PCA
Relatedness
Coverage

First pipeline takes BAM files as input and returns a list of regions covered above a given threshold (20X here) in average, among both cases and controls. A report summarizes the distribution of the different metrics listed on the left, in order to detect potential outliers/ poor quality samples.
The 2nd pipeline returns a soft filtered VCF and a report containing association test results and filtering stats.

ALLELE QC	
FILTER	**THRESHOLD**
Spanning Deletions	
Average Depth (DP)	> 20X
Quality by Depth (QD)	> 2
Mapping Quality (MQ)	> 40
Mapping Quality Rank Sum	> -12.5
Fisher Strand (FS)	< 20
Strand Odd Ratio (SOR)	< 4
Mappability (MAP)	= 1.0
Genotyping Rate (GR)	> 90%
Genotyping Quality for singletons (GQ)	> 90
Low Complexity Regions (LCR)	

BIOLOGICAL FILTERS	
FILTER	**THRESHOLD**
Allele Frequency (Internal)	< 3%
Allele Frequency (gnomAD)	< 1%
Impact prediction	non synonymous

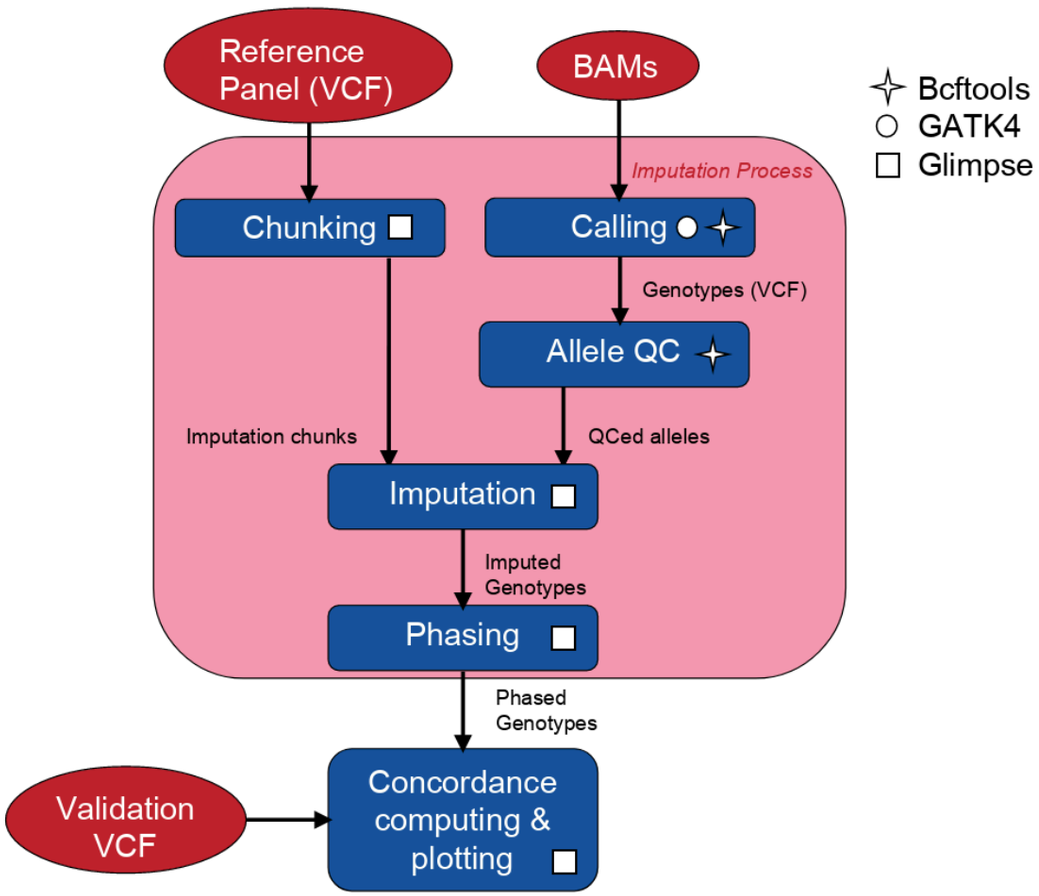
TOOLS & MODULES

- BCFTools
- SAMTools
- Picard
- Akt
- Mosdepth
- Somalier
- Deeptools
- Jvarkit
- Qqman (R)
- PyVCF
- Rvtests
- Snakemake

Provenance issue : tracking all these (hidden) domain-specific parameters ?

WF#2 Common variants analysis (LPS)

Raphaël Blanchet



- Low Pass Sequencing
- WF designed on 2600 whole genomes with very low coverage (1x, 2x, 4x)
- 1000 whole genomes for reference panel
- Statistical imputation

Provenance issues: which reference panel ? which imputation method ?

W_7 = Who - When - Where - Why - What - hoW - Which ?

Who ?

- "Raphael Blanchet", as a PhD Student, working at ITX, funded by Inserm
- ITX, as a research lab
- Inserm, as a research organization, funding institute

When ?

- Acquisition of input data
- Sequencing of biological samples
- Design of the data analysis workflow
- Run of the data analysis workflow

Where ?

- Location of the researcher
- Location of the biobank
- Sequenced biological samples
- Designed of the data analysis workflow
- Run of data analysis workflow

W_7 = Who - When - Where - Why - What - hoW - Which ?

What ?

WHAT is a sequence of events <e1, e2, ..., en> that affect a data object during its life time.

- workflow execution, fine-grained

Why ?

- **Are rare variant associated with the formation of intracranial aneurysms ?**
- **Are rare variant associated with the rupture of intracranial aneurysms ?**

Which (device) ?

- bcftools (http://bio.tools/bcftools)
- picard (http://bio.tools/picard)
- WF#1 rare variant analysis workflow
- **GLICID HPC infrastructure**

hoW ?

"How" documents actions that lead to the occurrence of an event."

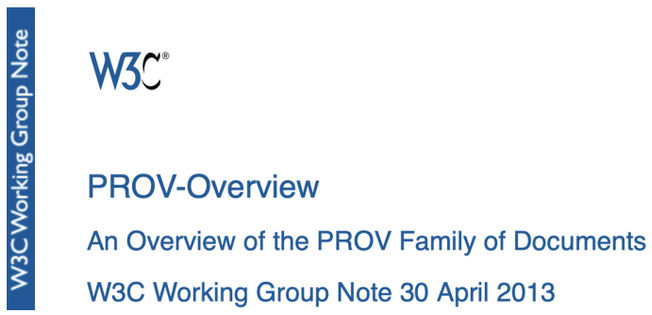
- WF#1 rare variant analysis workflow
- **reference guidelines, protocols, papers ?**

(partial) Implementation with PROV-O

PROV-O metadata captured at runtime with a slightly modified implementation of the Snakemake workflow engine.

```
<http://snakemake-provenance#activity-e5a4336a-3495-4821-9bd2-24d444b61637>
        a prov:Activity ;
        rdfs:comment """
                bwa mem -t 1 -M                -H '@CO\tProject:test Sample:Sample1 Date:2019-01-23 CWD:/Users/gaignard-a/D
        """ ;
        prov:wasAssociatedWith <https://bio.tools/bwa_cloudifb> ;
        prov:startedAtTime "2019-01-23T15:27:59.661594"^^xsd:dateTime;
        prov:endedAtTime "2019-01-23T15:27:59.661606"^^xsd:dateTime;
        prov:used <../testdata/human_g1k_v37.chr22.fasta> ;
        prov:used <../testdata/Sample1_ATGCCTAA_L001_R1_001.fastq.gz> ;
        prov:used <../testdata/Sample1_ATGCCTAA_L001_R2_001.fastq.gz> ;
        prov:used <../testdata/human_g1k_v37.chr22.fasta.bwt> ;
.

<Samples/Sample1/BAM/Sample1.p1.aligned.sam>
        a prov:Entity;
        prov:wasGeneratedBy <http://snakemake-provenance#activity-e5a4336a-3495-4821-9bd2-24d444b61637>;
        prov:wasAssociatedWith <https://bio.tools/bwa_cloudifb> ;
        rdfs:label "Samples/Sample1/BAM/Sample1.p1.aligned.sam";
        prov:wasDerivedFrom <../testdata/human_g1k_v37.chr22.fasta> ;
        prov:wasDerivedFrom <../testdata/Sample1_ATGCCTAA_L001_R1_001.fastq.gz> ;
        prov:wasDerivedFrom <../testdata/Sample1_ATGCCTAA_L001_R2_001.fastq.gz> ;
        prov:wasDerivedFrom <../testdata/human_g1k_v37.chr22.fasta.bwt> ;
.
```



https://github.com/albangaignard/fresh-toolbox/blob/master/provenance.ttl

https://github.com/albangaignard/galaxy-PROV/blob/master/notebooks/galaxy_prov.ttl

What is missing in PROV for biomedical use cases ?

... to increase findability, trust and reuse

- Link with the software containers / computing platform / **hardware**
- Fine-grained **biological parameters** (thresholds) ?
- Data & software **access conditions**
  - out of the scope of PROV (DUO ontology ?)
- **Reference methods / papers** associated with data analysis
  - PAV ? SIO ? MicroPublication ? CITO ?
- **Research context** ? questions ? outcomes ? data management plans ?

PEPR Santé numérique

- National project funded on Digital Health : ShareFAIR

- Focus on workflows for reproducibility / provenance / FAIR protocols & data
- 6 PhD projects in progress + 1 software developer

Any questions ?

Inserm cnrs Nantes Université CHU NANTES l'institut du thorax