Pipeline modules - WGS

Phase 1 – Testing ACMGuru package

- Prepare analysis-ready datasets ✅
- Qualifying variant selection ✅
- Tailored filtering ❌
- Result interpretation ✅ (manual)

Phase 2 – Testing pipeline automation

- GATK pipeline optimise ✅
- Storage and costs ✅
- ACMGuru reports ❌ (template)

Phase 3 – Automated run ❌

Goals

WGS turnaround

- Single sample turn-around time
- Joint cohort analysis time / cost

Varian association tests

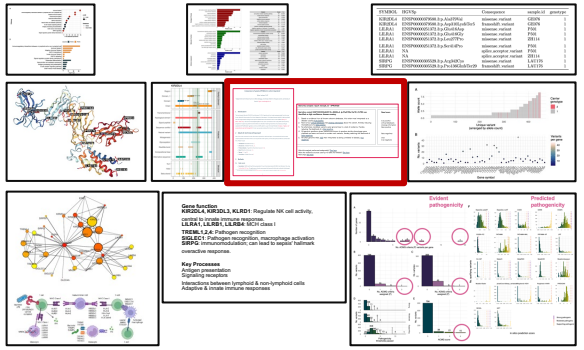
- Genetic association with outcome including multi-omic input

SPHN RDF Schema

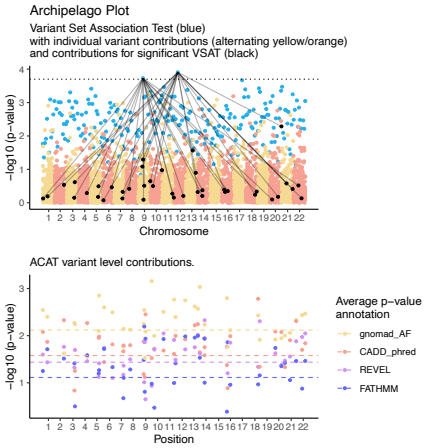
- Output directly

Formal test

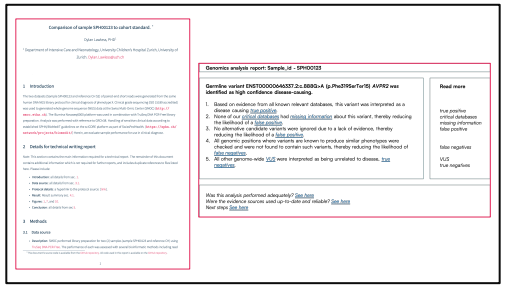
- genomebottle pangenome

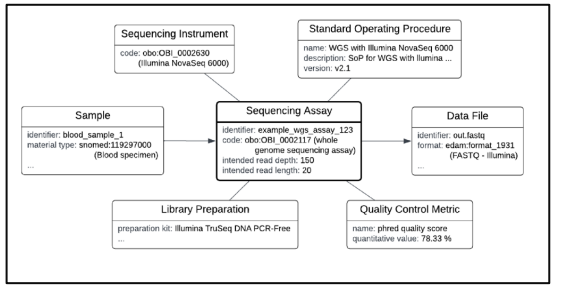


- **ACMGuru**
  - uniprotR
  - AutoDestructR
  - gene-illustrate
  - untangleR
  - getDiscussion
  - Interpretation
- **Analysis-ready dataset**
- **Clinical genetics report**
- **Technical quality report**
- **Association report**
- **Workflow code**
- **Logs**



Pre-print 2024





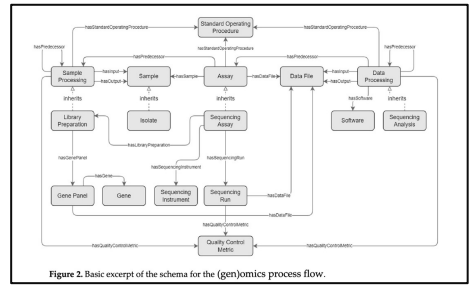
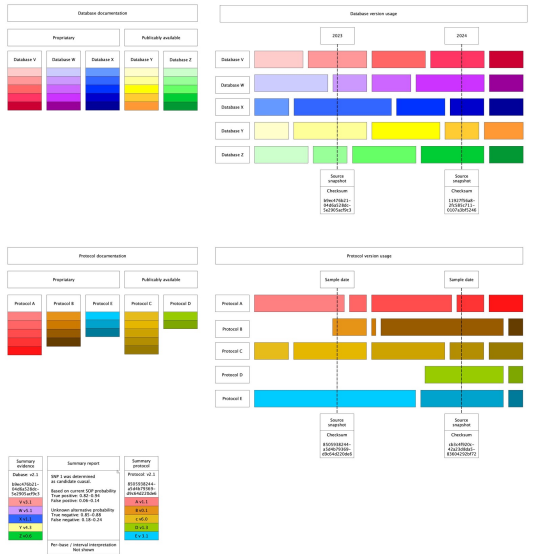


Figure 2. Basic excerpt of the schema for the (gen)omics process flow.

https://doi.org/10.20944/preprints202312.0373.v1

Future



Quantify integrity

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8
Database V	0.3	0.4	0.5	0.5	0.4	0.3	0.2	0.1
Database W	A	A	A	B	B	C	C	C
Database X	0	0	1	1	1	0	0	0
Database Y	0.3	0.4	0.5	NA	NA	0.3	0.2	0.1
Database Z	A	A	B	NA	NA	NA	C	C
Data metrics	0.9	0.9	0.9	0.9	0.9	0.9	0.9	NA
Evidence credibility	III	III	III	I	I	II	III	III

Per-base / interval interpretation

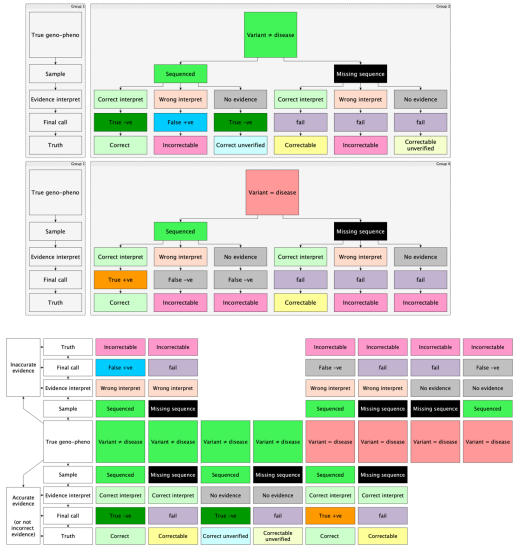
Pos 1–3:
Evidence confirmed informative for ALL databases.
Determined NOT CAUSAL candidate.

Pos4:
Critical evidence (database Y) not present.
Insuffient evidence.

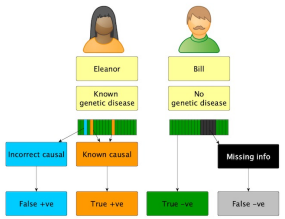
Pos 6:
Evidence confirmed informative for CRITICAL databases.
Determined NOT CAUSAL candidate.

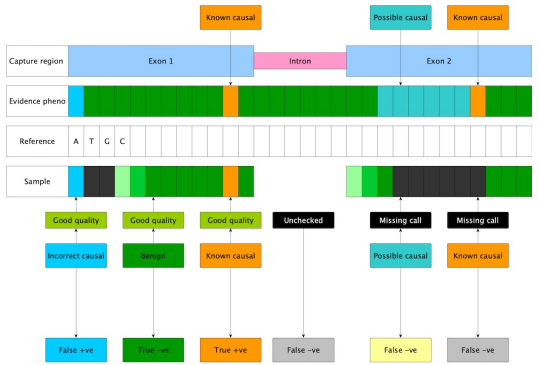
Pos 8:
Evidence sources all present.
Sample data insufficient.

Quantify evidence



Quantify a priori combination





Genomics analysis report

Germline variant *NM000492.3(CFTR):c.1521_1523del (p.Phe508del)* was identified as high confidence disease-causing.

1. Based on evidence from all known relevant databases, this variant was interpreted as a disease causing *true positive*.
2. None of our *critical databases* had *missing information* about this variant, thereby reducing the likelihood of a *false positive*.
3. No alternative candidate variants were ignored due to a lack of evidence, thereby reducing the likelihood of a *false positive*.
4. All genomic positions where variants are known to produce similar phenotypes were checked and were not found to contain such variants, thereby reducing the likelihood of *false negatives*.
5. All other genome-wide *VUS* were interpreted as being unrelated to disease, *true negatives*.

Read more

true positive
critical databases
missing information
false positive

false negatives

VUS
true negatives

Was this analysis performed adequately? See here
Were the evidence sources used up-to-date and reliable? See here
Next steps See here

Bayesian confidence diagnostic rate