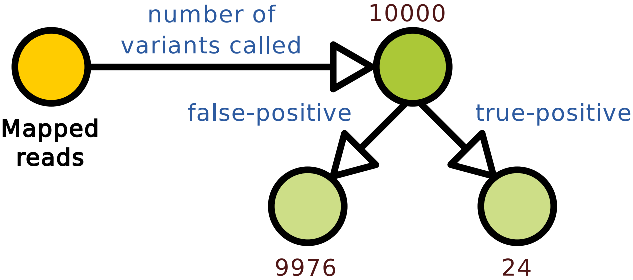
UMC Utrecht

Using SPARQL and RDF to analyze structural variants

Roel Janssen

September 4, 2017

About structural variant calling



Goals

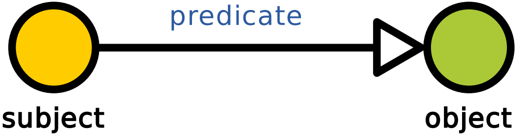
- Filter structural variant (SV) calls by position overlap*
- Filter or augment SV call information with regional information

* Idea by Mark van Roosmalen and Robert Ernst

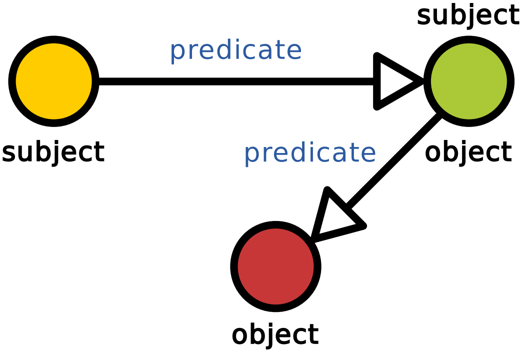
RDF and SPARQL

- Resource Description Framework (RDF)
  - is an information modeling method;
  - is a W3C recommendation since 1999;
  - EMBL-EBI made data accessible in RDF format.
- SPARQL Protocol and RDF Query Language (SPARQL)
  - is a language to query data in RDF format;
  - can be used in various programming languages (R, Python, Perl, ...).

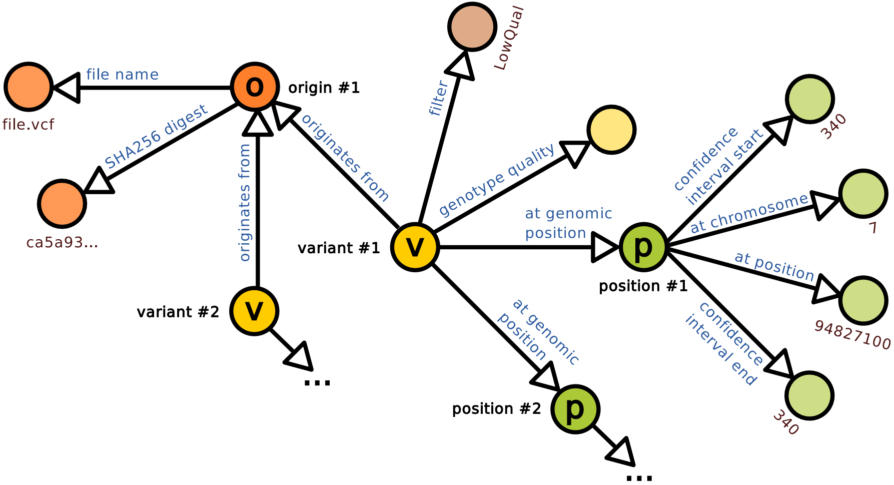
Describing information using RDF



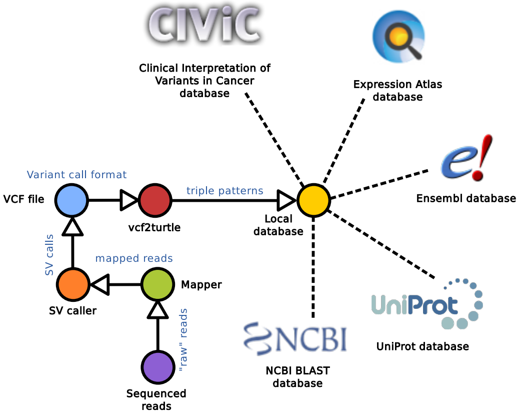
Describing information using RDF



Model: Extract triples from the Variant Call Format (VCF)



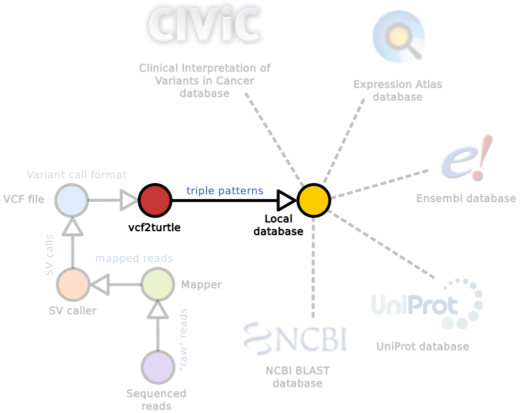
Tools: Extract triples from the Variant Call Format (VCF)



Source code for the `vcf2turtle`:

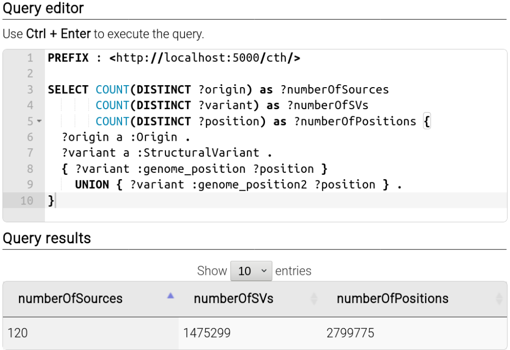
`https://github.com/UMCUgenetics/sparqling-svs`

Tools: Extract triples from the Variant Call Format (VCF)



Source code for the `vcf2turtle`:
`https://github.com/UMCUgenetics/sparqling-svs`

Tools: Query and ontology interface for quick exploration



Query editor

Use **Ctrl + Enter** to execute the query.

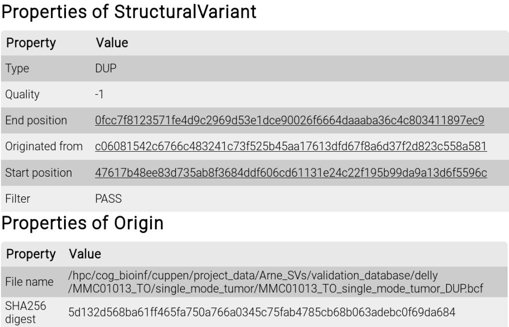
```
PREFIX : <http://localhost:5000/cth/>

SELECT COUNT(DISTINCT ?origin) as ?numberOfSources
       COUNT(DISTINCT ?variant) as ?numberOfSVs
       COUNT(DISTINCT ?position) as ?numberOfPositions {
  ?origin a :Origin .
  ?variant a :StructuralVariant .
  { ?variant :genome_position ?position }
    UNION { ?variant :genome_position2 ?position } .
}
```

Query results

Show 10 entries

numberOfSources	numberOfSVs	numberOfPositions
120	1475299	2799775



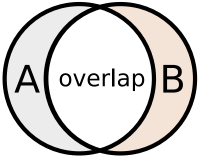
Properties of StructuralVariant

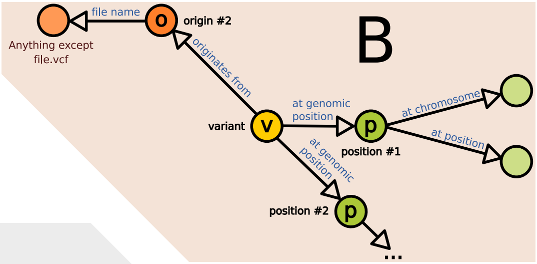
Property	Value
Type	DUP
Quality	-1
End position	0fcc7f8123571fe4d9c2969d53e1dce90026f6664daaaba36c4c803411897ec9
Originated from	c06081542c6766c483241c73f525b45aa17613dfd67f8a6d37f2d823c558a581
Start position	47617b48ee83d735ab8f3684ddf606cd61131e24c22f195b99da9a13d6f5596c
Filter	PASS

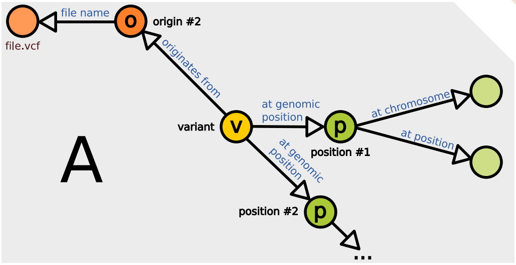
Properties of Origin

Property	Value
File name	/hpc/cog_bioinf/cuppen/project_data/Arne_SVs/validation_database/delly /MMC01013_TO/single_mode_tumor/MMC01013_TO_single_mode_tumor_DUP.bcf
SHA256 digest	5d132d568ba61ff465fa750a766a0345c75fab4785cb68b063adebc0f69da684

Filtering overlap

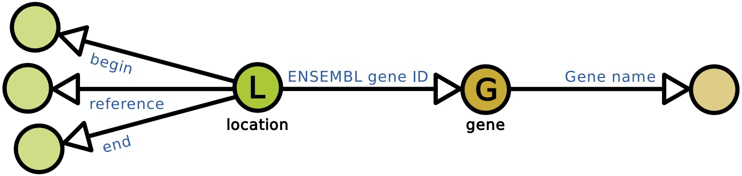




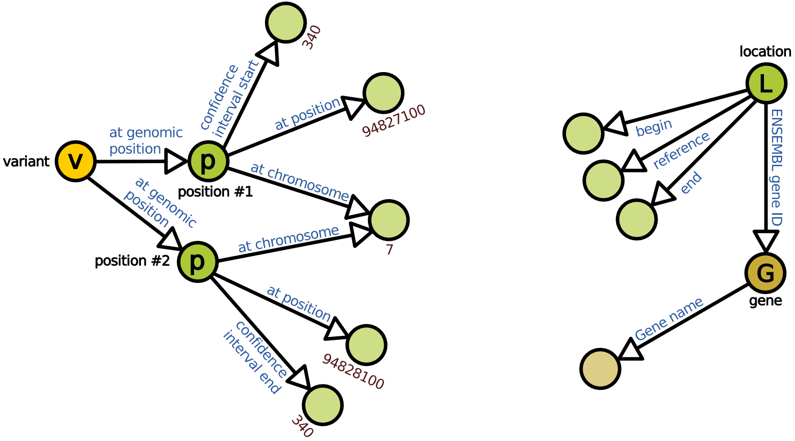


Ensembl gene regions

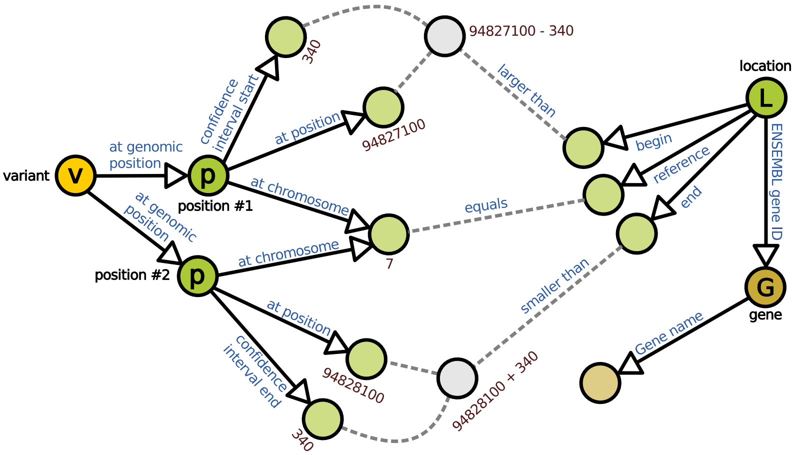
From `<http://rdf.ebi.ac.uk/resource/ensembl>`:



Linking Ensembl gene regions with our SVs



Linking Ensembl gene regions with our SVs



Wrapping up

- Triple stores scale by communicating with other triple stores;
- By describing your data using RDF, you can tap into other databases;
- ... and others could tap into yours (if you publish it);
- Linking with other databases needs to be driven by research questions;
  - Create a model to answer specific questions;
  - Don't over-engineer it.
- Slides will be available at: `https://github.com/UMCUGenetics/sparqling-svs`

Wrapping up

- Triple stores scale by communicating with other triple stores;
- By describing your data using RDF, you can tap into other databases;
- ... and others could tap into yours (if you publish it);
- Linking with other databases needs to be driven by research questions;
  - Create a model to answer specific questions;
  - Don't over-engineer it.
- Slides will be available at: `https://github.com/UMCUGenetics/sparqling-svs`

Wrapping up

- Triple stores scale by communicating with other triple stores;
- By describing your data using RDF, you can tap into other databases;
- ... and others could tap into yours (if you publish it);
- Linking with other databases needs to be driven by research questions;
  - Create a model to answer specific questions;
  - Don't over-engineer it.
- Slides will be available at: `https://github.com/UMCUGenetics/sparqling-svs`

Wrapping up

- Triple stores scale by communicating with other triple stores;
- By describing your data using RDF, you can tap into other databases;
- ... and others could tap into yours (if you publish it);
- Linking with other databases needs to be driven by research questions;
  - Create a model to answer specific questions;
  - Don't over-engineer it.
- Slides will be available at: `https://github.com/UMCUGenetics/sparqling-svs`

Wrapping up

- Triple stores scale by communicating with other triple stores;
- By describing your data using RDF, you can tap into other databases;
- ... and others could tap into yours (if you publish it);
- Linking with other databases needs to be driven by research questions;
  - Create a model to answer specific questions;
  - Don't over-engineer it.
- Slides will be available at: `https://github.com/UMCUGenetics/sparqling-svs`

Wrapping up

- Triple stores scale by communicating with other triple stores;
- By describing your data using RDF, you can tap into other databases;
- ... and others could tap into yours (if you publish it);
- Linking with other databases needs to be driven by research questions;
  - Create a model to answer specific questions;
  - Don't over-engineer it.
- Slides will be available at: `https://github.com/UMCUGenetics/sparqling-svs`

Acknowledgements

Arne Hoeck

Arnold Kuzniar

Joep de Ligt

Mark van Roosmalen

Robert Ernst

Edwin Cuppen