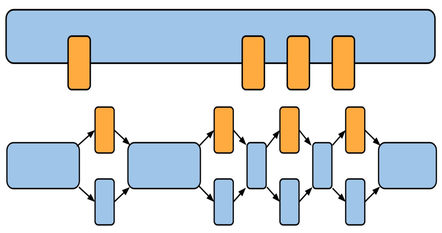
Quantitative Biology Center (QBiC) [1]
Swiss Institute of Bioinformatics (SIB) [2]
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Semantic Genome Graphs

Simon Heumos[1] [0000-0003-3326-817X], Jerven Bolleman[2] [0000-0002-7449-1266]

Current linear reference based methods of representing genomic variation are limiting our insights into the variation between genomes. Genome graphs are a set of techniques that can accurately represent large structural variation as well as single nucleotide polymorphisms. As any graph can be serialized as an RDF (Resource Description Framework) one, we show some advantages and disadvantages of making a Genome Graph available on the Semantic Web in a FAIR[1] (Findable Accessible Interoperable Reusable) way. We demonstrate how we can use SPARQL to drive visualizations and integrate with non genome graph knowledge.

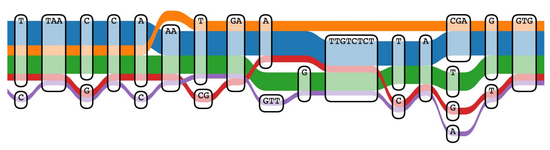
Reference bias[2] and genome graphs



Calling a di-nucleotide repeat in a high allele frequency region, for example the TATA box of gene UGT1A1, can be tricky due to the mapping bias and influences diagnostics. Genome graph implementations, like the vg toolkit[3], can help to deal with that problem.

http://bit.ly/tataUGT1A1

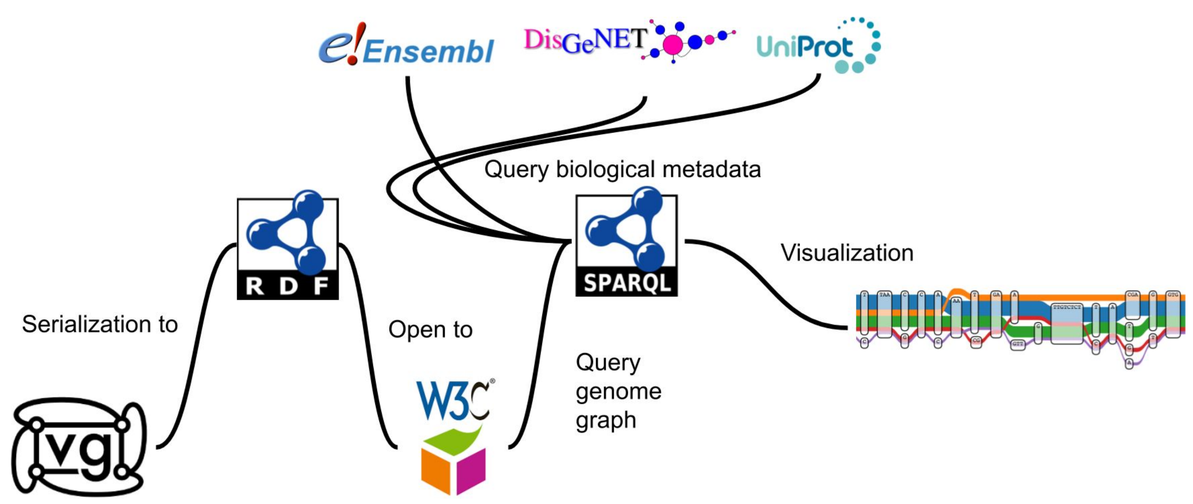
What does a genome graph look like?



Sequence Tube Map[4] presents sequence graphs in a tube map layout. Nodes represent shared or unique sequences of variable length, through which genome paths pass to reveal complete genomic sequences.

http://bit.ly/SequenceTubeMap

SPARQL querying a vg genome graph




```
BASE <http://rdf.ebi.ac.uk/resource/ensembl/97/saccharomyces_cerevisiae/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX faldo:<http://biohackathon.org/resource/faldo#>
PREFIX ensembltranscript:<http://rdf.ebi.ac.uk/resource/ensembl.transcript/>
PREFIX vg:<http://biohackathon.org/resource/vg#>

SELECT *

WHERE {
  # Our path matches an Ensembl linear genome
  BIND(<R64-1-1/VIII> AS ?ref) .
  #Limit the search to steps between nucleotide 200000 and 220000
  VALUES (?rangeBegin ?rangeEnd){(200000 220000)}
  ?target a vg:Step ;
    faldo:reference ?ref ;
    faldo:location ?stepLinearLocation .
  ?stepLinearLocation faldo:begin ?bp ;
    faldo:end ?ep .
  ?bp faldo:position ?stepBegin .
  ?ep faldo:position ?stepEnd  .
  FILTER ((?stepBegin >= ?rangeBegin && ?stepBegin <= ?rangeEnd)
          || (?stepEnd >= ?rangeBegin && ?stepEnd <= ?rangeEnd ))

  SERVICE <https://www.ebi.ac.uk/rdf/services/sparql/> {
    ?target faldo:location ?stepLinearLocation .
    ?stepLinearLocation faldo:begin ?bp ;
             faldo:end ?ep .
    ?bp faldo:position ?stepBegin ;
        faldo:reference ?ref .
    ?ep faldo:position ?stepEnd  ;
        faldo:reference ?ref .
    FILTER ((?stepBegin >= ?rangeBegin && ?stepBegin <= ?rangeEnd)
            || (?stepEnd >= ?rangeBegin && ?stepEnd <= ?rangeEnd ))
  }
}
```

This query part is on the Genome Graph, Using both VG schema ontology and FALDO coordinates space.

This part of the query is answered by the EBI Sparql end point. Finds features in Ensembl matching the VG path

How are genome graphs modeled in RDF?

A Node in the vg RDF is equivalent to a Node in the vg data model. A Path is a number of Steps that represent a sequence of Node visits revealing its linear biological sequence. Each Step connects a Node into a Path.

http://bit.ly/vgOntology

Zero extra cost SPARQLable genome graphs

Converting and loading a genome graph into an RDF datastore can incur significant costs in storage. Spodgi shows we can use python RDFLib to run SPARQL on native genome graph (Odgi and XG) data formats without extra storage related costs.

Future work

- Direct integration with FHIR and clinical data sources
- Improve scalability
- SPARQL update support for genome graph data formats
- Integrate query optimizations
- RDFLib federated query support

Are semantic genome graphs FAIR[1,2]?

http://bit.ly/fairPrinciples

FAIR principle	Implementation status
F1: Data are assigned a globally unique and eternally persistent identifier.	Graph, Step, not yet Node
A1: Data are retrievable by their identifier using a standardized communications protocol.	SPARQL and HTTP
I1: Data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	RDF using shared ontologies
I2: Data use vocabularies that follow FAIR principles.	FALDO[5], VG
I3: Data include qualified references to other (meta)data.	e.g. Ensembl via shared Path IRIs

Funding

Acknowledgments S.H. acknowledges Florian Battke of the CeGaT GmbH for pointing at the UGT1A1 example. S.H. acknowledges Lukas Heumos and Andreas Friedrich of the Quantitative Biology Center for valuable and constructive feedback.

References

1. Wilkinson et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3, 160018.
2. Ballouz et al. (2019). 35(24):5318-532035(24):5318-5320 *Genome Biology*, 20:159.
3. Garrison et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*. 36:9. Pages 875-879.
4. Beyer et al. (2019). Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics*, btz579.
5. Bolleman et al. (2016). FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. *Journal of Biomedical Semantics*, 7.