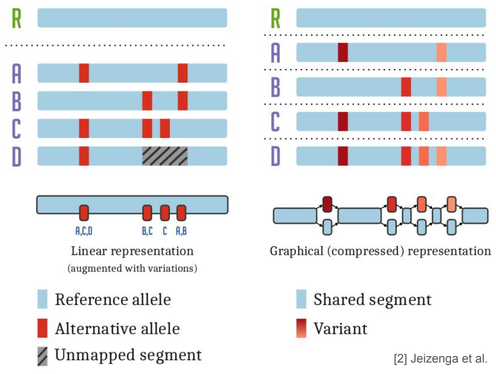
Semantic Variation Graphs - A Pangenome Ontology

Toshiyuki T. Yokoyama[1*], Simon Heumos[2*], Josiah Seaman[3], Dmytro Trybushnyi[4], Torsten Pook[5], Andrea Guarracino[6], Erik Garrison[7,8] and Jerven T. Bolleman[9]

[1]*The University of Tokyo, Chiba, Japan,* [2]*Quantitative Biology Center (QBiC) Tübingen, University of Tübingen, Tübingen, Germany,* [3]*Max Planck Institute for Developmental Biology, Tübingen, Germany,* [4]*Karlsruhe Institute of Technology, Karlsruhe, Germany,* [5]*Center of Integrated Breeding Research, University of Goettingen, Goettingen, Germany,* [6]*University of Rome Tor Vergata, Via della Ricerca Scientifica 1, Rome, Italy,* [7]*Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA,* [8]*Biomolecular Engineering and Bioinformatics, University of California Santa Cruz, Santa Cruz, CA, USA,* [9]*SIB Swiss Institute of Bioinformatics, Geneva, Switzerland, *Contributed equally.*

Variation graphs are a novel way to describe genomic variation across a population. Existing toolkits have limited capabilities in integrating biological annotation and providing FAIR[1] interfaces for large scale visualizations. Borderless technology such as the Semantic Web allows variation graph toolkits and pangenome tools to focus on their core competence while allowing bioinformaticians to integrate, analyze, and visualize the data. We show how the vg RDF and Pantograph RDF can represent data for the Semantic Web and how we can combine existing data from INDSC and UniProt without conversions or loss of information into a single Variation and Knowledge Graph.

Variation graphs encode pangenomes



A graphical pangenome[2] models the full set of genomic elements in a given species or clade. It can be encoded in the form of a variation graph, a particular type of sequence graphs which embed the linear sequences of the pangenome as paths in the graphs themselves.

https://bit.ly/PangenomeGraph

Pantograph Pangenome View

Each column of an MSA is a homologous base across all the individuals.

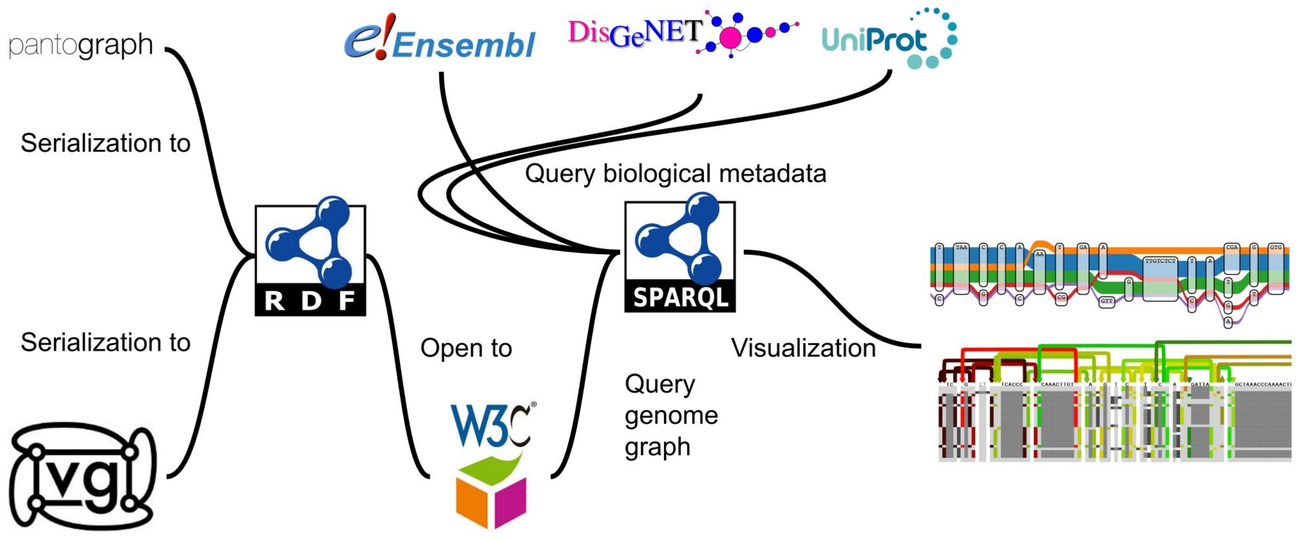
```
TCCAA---CTCTCTGIGGTTCCGGTTGCTAT*
TCCAAGGTCTCTCTGIGGTTCC----GCTAT*
TCCAA---CTCTCTGGGGTTCCGG-TGCTAT*
TCCAAGGTCTCTCTGAGGTTCC----GCTAT*
TCCAAGGTCTCTCTGIGGTTCCGG-TGCTAT*
```

In the Pangenome Sequence, bases are encoded in a binary matrix. Structural rearrangement are visualized with Links.

Pangenome Sequence is aggregated into Bins forming bin-sized ZoomLevels.

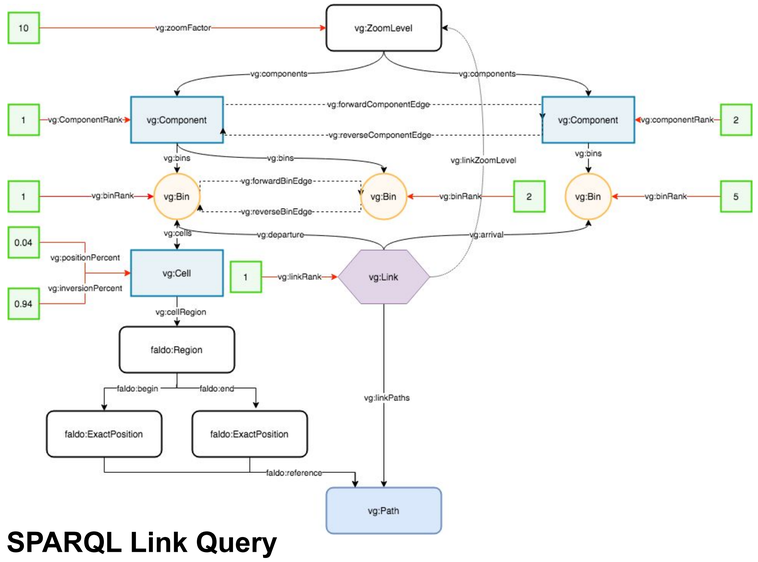
https://bit.ly/PantographBrowser

SPARQL Querying a Pangenome Graph



Variation Graph Ontology Snippet

Classes and Properties of Pantograph Ontology



SPARQL Bin Query

```
PREFIX vg: <http://biohackathon.org/resource/vg#>
PREFIX faldo: <http://biohackathon.org/resource/faldo#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?cell ?inversionpercent ?positionpercent ?beginpos ?endpos
WHERE {
        ?zoomlevel a vg:ZoomLevel;
                   vg:components ?components;
                   vg:zoomFactor ?zoomfactor .
        FILTER(?zoomfactor = 1)
        ?components vg:bins ?bin .
        ?bin vg:cells ?cell;
             vg:binRank ?binrank .
        FILTER(?binrank < 6 && ?binrank > 0)
        ?cell vg:cellRegions ?faldoregion;
              vg:inversionPercent ?inversionpercent;
              vg:positionPercent ?positionpercent .
        ?faldoregion faldo:begin ?begin;
                     faldo:end ?end .
        ?begin faldo:position ?beginpos .
        ?end faldo:position ?endpos .
        ?begin faldo:reference ?reference .
}
```

SPARQL Link Query

```
PREFIX vg: <http://biohackathon.org/resource/vg#>
PREFIX faldo: <http://biohackathon.org/resource/faldo#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?link ?path ?arrivalbinrank ?departurebinrank ?zoomfactor
WHERE {
        ?link a vg:Link;
              vg:arrival ?arrivalbin;
              vg:departure ?departurebin;
              vg:linkPaths ?path ;
              vg:linkZoomLevel ?zoomlevel .
        ?arrivalbin vg:binRank ?arrivalbinrank .
        ?departurebin vg:binRank ?departurebinrank .
        FILTER((?arrivalbinrank < 6 && ?arrivalbinrank > 0)
            || (?departurebinrank < 6 && ?departurebinrank > 0))
        ?zoomlevel vg:zoomFactor ?zoomfactor .
        FILTER(?zoomfactor = 1)
}
```

SPARQL Sequence Query

```
PREFIX vg:<http://biohackathon.org/resource/vg#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT (SUBSTR(group_concat(?sequence; separator=''), 1,5) as ?panSeq) {
    SELECT
        *
    WHERE {?s a vg:Node;
           rdf:value ?sequence .
    }
    ORDER BY ?s
}
```

Variation Graph of a SARS-CoV-2 Pangenome

- SpOdgi: Translate odgi graph into linked pangenome
- Part of the CWL Public Sequence Resource workflow
- Triple store of more than 1300 viral genomes integrated with information from SIB, INDSC, UniProt[3], Bgee, neXtProt, OMA, Rhea
- Ready to ask complex biological questions like "Are the active sites for PL1-PRO conserved?"

https://bit.ly/SpOdgi
https://bit.ly/PublicSequenceResource
https://bit.ly/COVID19StoreSIB

How to use it?

- SpOdgi creates Turtle for vg ontology
- Component Segmentation emits Turtle file for Pantograph ontology
- FALDO[4] for linking (pan)genome positional information across ontologies

https://bit.ly/PangenomeOntology
https://bit.ly/CompSegOntology

Future work

- Direct integration with FHIR and clinical data sources
- Improve scalability
- Integrate query optimizations
- We are one Javascript library away from running the Pantograph Browser on the Pangenome ontology!

References

1. Wilkinson et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3, 160018.
2. Eizenga et al. (2020). Pangenome graphs. *Annual Reviews of Genomics and Human Genetics*. 21.
3. The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res., 47, D506–D515.
4. Bolleman et al. (2016). FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. *Journal of Biomedical Semantics*, 7.