

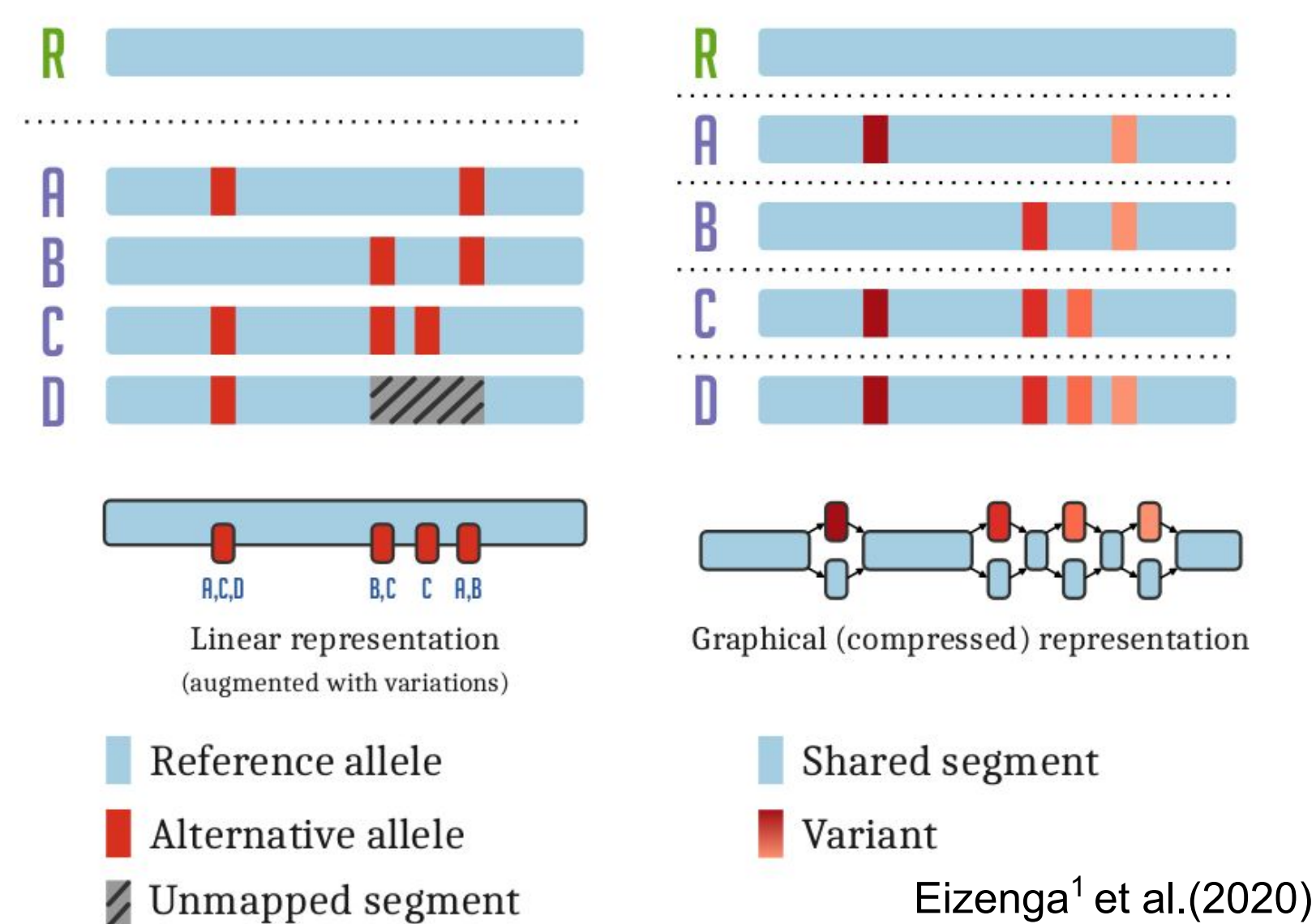
Scalable variant detection in pangenome models

Francesco Porto ^a, Flavia Villani ^b, Andrea Guarracino ^c, Christian Fischer ^d, Hao Chen ^e, Robert W. Williams ^d,
Vincenza Colonna ^b, Gianluca Della Vedova ^a, Erik Garrison ^f, and Pjotr Prins ^d

^a Department of Informatics, Systems, and Communication, University of Milano-Bicocca, Italy, ^b National Research Council, Institute of Genetics and Biophysics 'A. Buzzati-Traverso', Naples, Italy, ^c Centre for Molecular Bioinformatics, Department of Biology, University Of Rome Tor Vergata, Rome, Italy, ^d Department of Genetics, Genomics and Informatics, College of Medicine, UTHS, ^e Department of Pharmacology, Addiction Science, and Toxicology, The University of Tennessee Health Science Center, Memphis, TN, USA, ^f Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, United States.

We have implemented a two-step scalable approach to detect variants: first we construct a graph pangenome from a graphical fragment assembly (GFA) file that stores the fragments, where each fragment corresponds to a vertex of the graph, then we analyze the graph to detect all variants. We have tested our approach on a SARS-CoV-2 dataset with over 7800 fragments and on a dataset that contains all alternative sequences of the highly polymorphic human leukocyte antigen (HLA) complex.

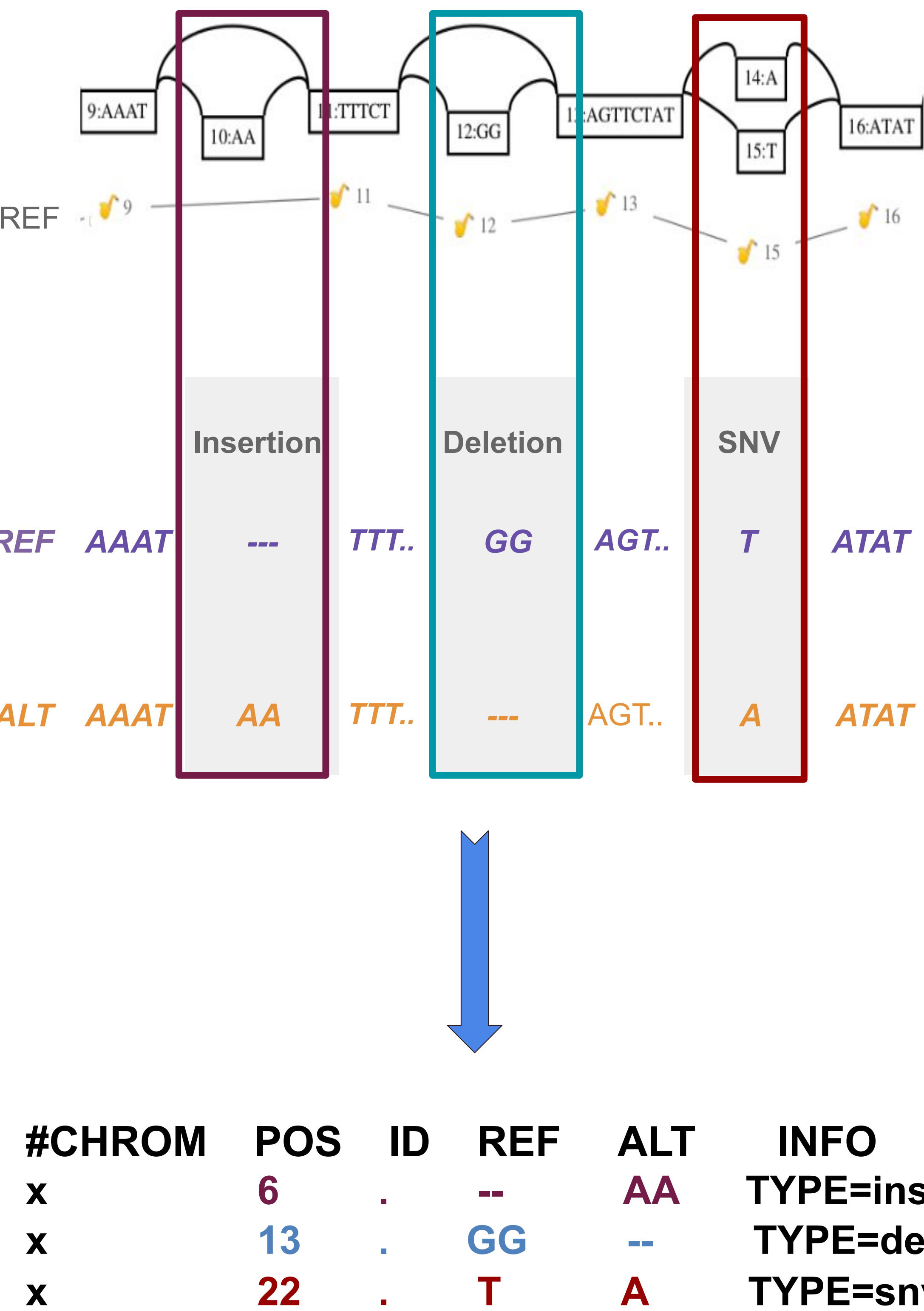
Variation Graphs encode pangenomes



A graphical [pangenome](#)¹ models the full set of genomic elements in a given species or clade.

The [variation graph](#) data model describes the all-to-all alignment of many sequences (genomes or genes for instance) as walks through a graph whose nodes are labeled with DNA sequences.

Variant detection in variation graphs



Code available at

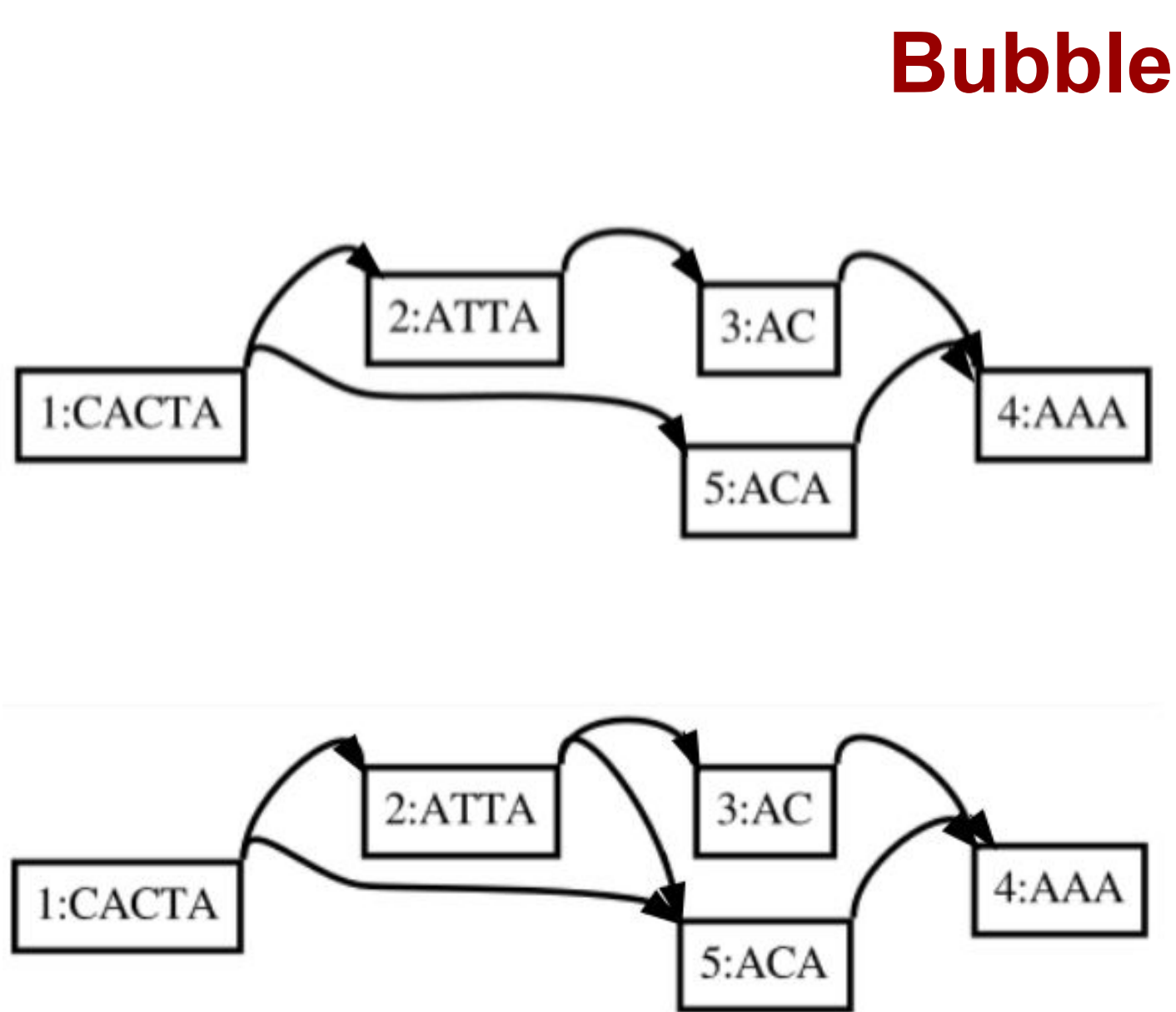
<https://github.com/HopedWall/rs-qfatovcf>

References

- Eizenga et al. (2020). Pangenome graphs. *Annual Reviews of Genomics and Human Genetics*. 21.
- Paten, Benedict, et al. "Superbubbles, ultrabubbles, and cacti." *Journal of Computational Biology* 25.7 (2018): 649-663.

Bubbles

In pangenome variation graphs, genetic variants appear as bubbles and ultrabubbles² (nested bubbles). These sites have a common starting context, a common exit point, and multiple possible paths that connect the two. Each path represents an allele.



Bubble

Ultrabubble

HandleGraph interface

A compact and efficient data structure to represent large genomic variation graphs. [ODGI](#) (**O**ptimized **D**ynamic **G**raph **I**mplementation) is a library implementing the HandleGraph interface with minimum memory overhead. This has required a careful encoding of the graph components

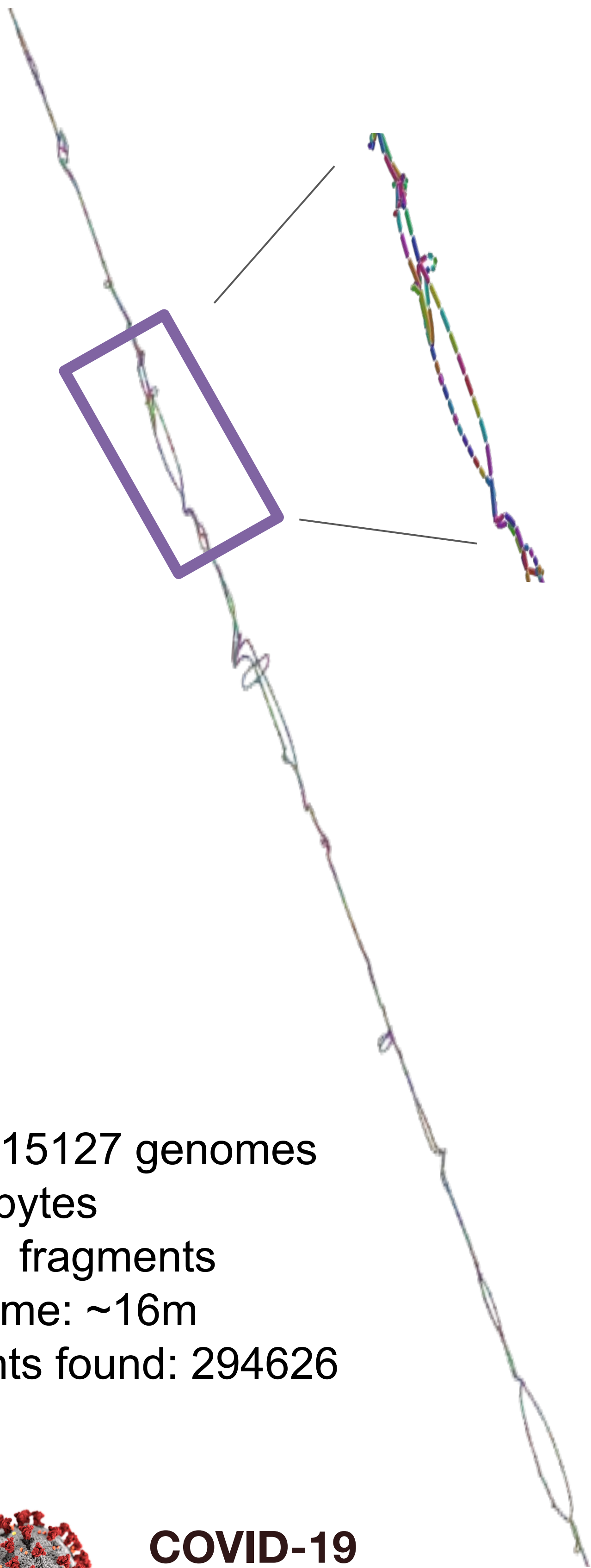
Why Rust?



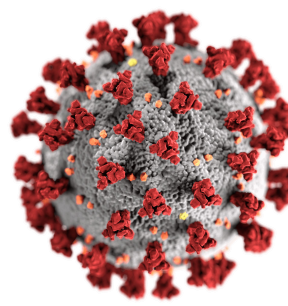
[Rust](#) is a programming language focused on performance and safety.

- ❖ Great **ecosystem** (Cargo, crates.io, docs.rs).
- ❖ Much **safer** than C++ while having a similar **speed**.
- ❖ Friendly and helpful **community**.
- ❖ Used in many open source projects, such as **Firefox**.

Dataset SARS-CoV-2 Pangenome



- ❖ From 15127 genomes
- ❖ 1.2 Gbytes
- ❖ 78571 fragments
- ❖ Run time: ~16m
- ❖ Variants found: 294626



COVID-19 PubSeq

Data available at

<http://covid19.genenetwork.org/>

Future work

- ❖ Parallel implementation to improve its speed.
- ❖ Identification of complex bubbles ([Superbubbles](#), [Ultrabubbles](#), and [Cacti](#)).

Dataset HLA-DRB1-3123 Pangenome

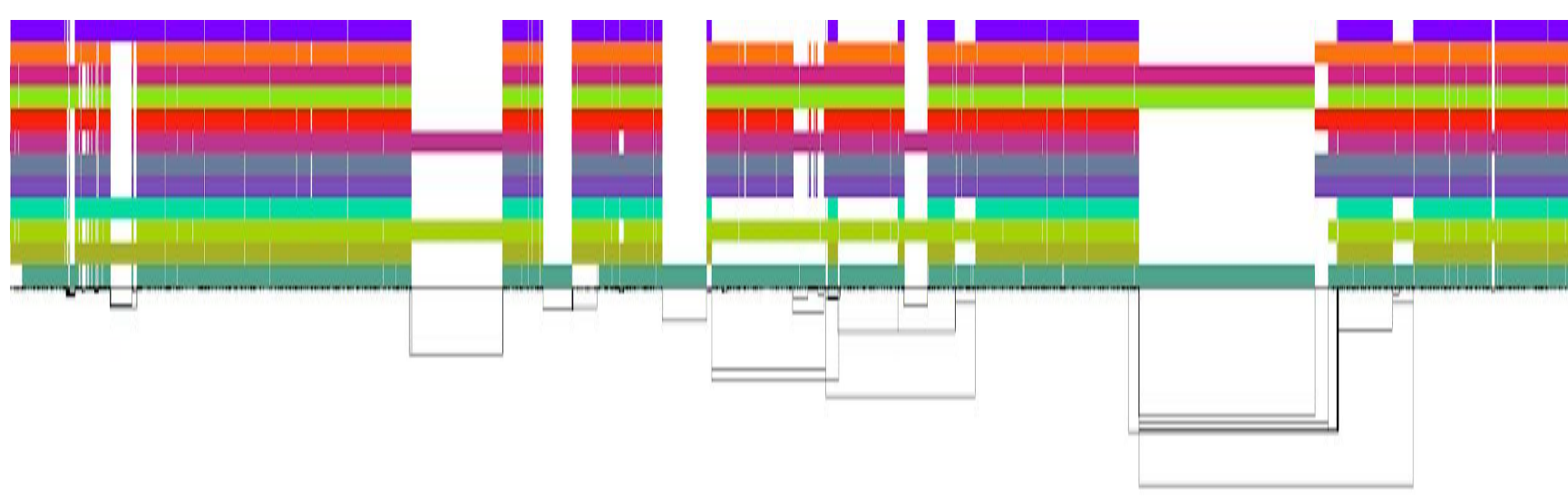


Image obtained via <https://github.com/vgteam/odgi>

- ❖ From 12 sequences
- ❖ Size: 163416 nucleotides
- ❖ Run time: ~0.1s
- ❖ Variants found: 7505



Data available at

<https://github.com/ekg/HLA-zoo>

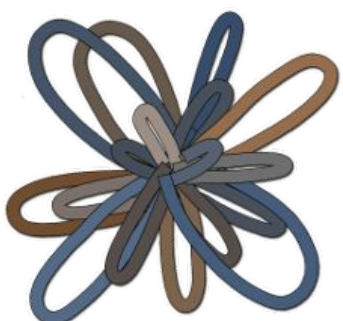


Image obtained via

<https://rwick.github.io/Bandage/>

Google Summer of Code

