EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Quantitative Biology Center

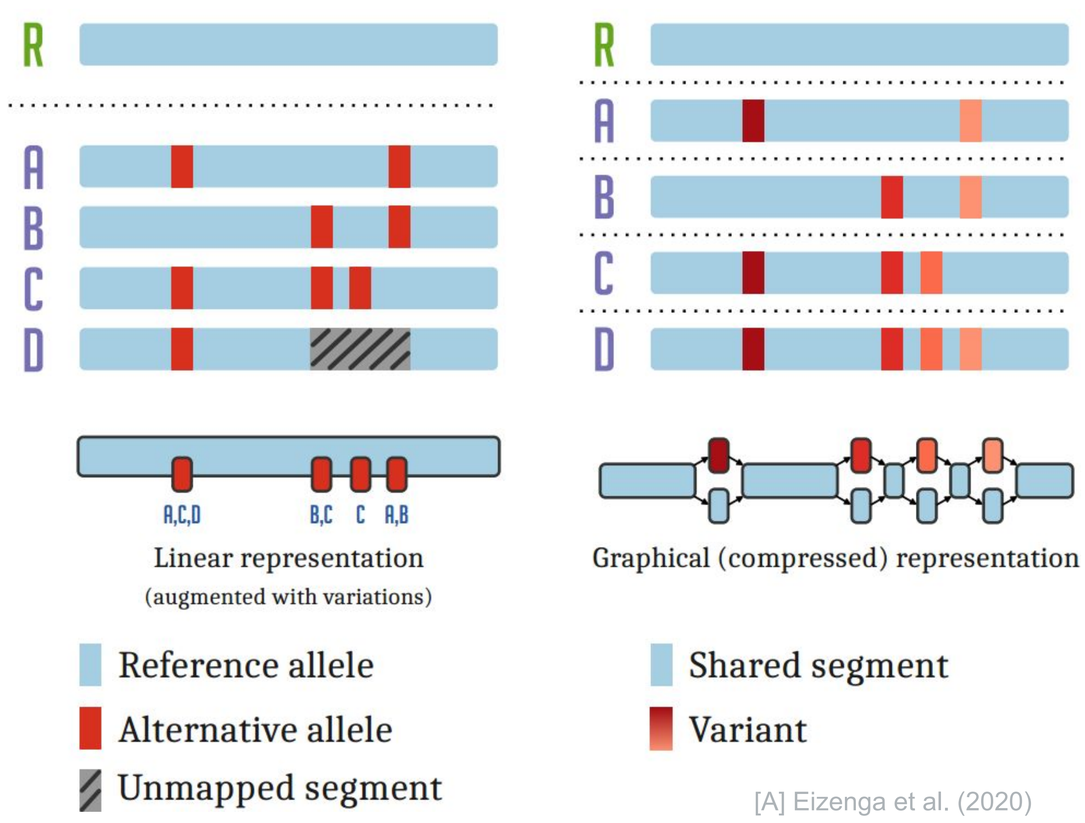
QBiC

Pangenome Graphs

Simon Heumos[1], Andrea Guarracino[2,3], Pjotr Prins[3], Erik Garrison[3], Sven Nahnsen[1]

Thanks to advances in sequencing technology, new telomere-to-telomere genome assemblies are produced at a high rate. Pangenome graphs can model the full set of genomic elements in a given species or clade, reducing the reference-bias. At QBiC, we are exploring algorithms to learn human-readable 1D or 2D projections of a graph. ODGI's 1D linearization algorithm is already the basis of the PanGenome Graph Builder (PGGB) pipeline and its Nextflow descendant nf-core/pangenome. We also develop and apply methods to understand the underlying biology of a pangenome graph.

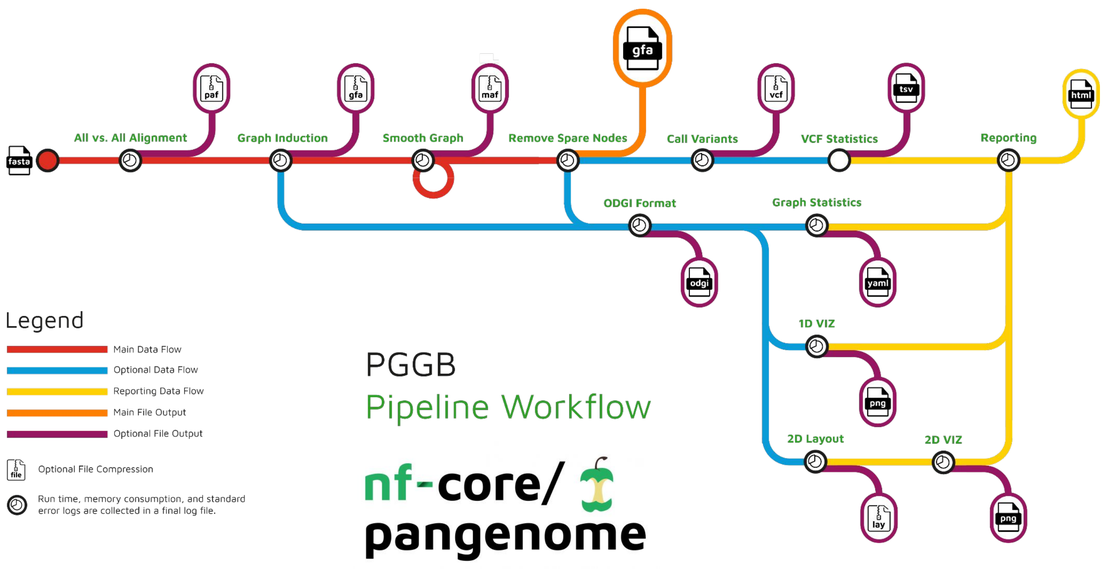
VARIATION GRAPHS ENCODE PANGENOME



[A] Eizenga et al. (2020)

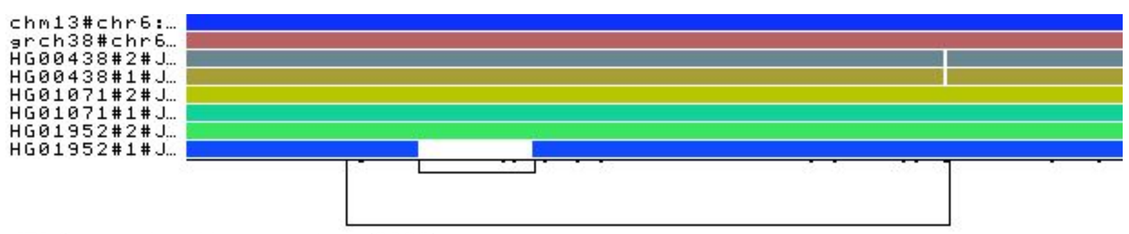
A pangenome[A] models the full set of genomic elements in a given species or clade. It can efficiently be encoded[B] in the form of a variation graph, which embeds the linear sequences of the pangenome as paths in the graphs themselves.

BUILDING PANGENOME GRAPHS



Our built pangenome graphs are precise enough to encode all kinds of structural variants, even complex ones like segmental duplication, centromeres, or satellites sequences. All the alignments represented are on a base-pair level.

UNDERSTANDING PANGENOME GRAPHS

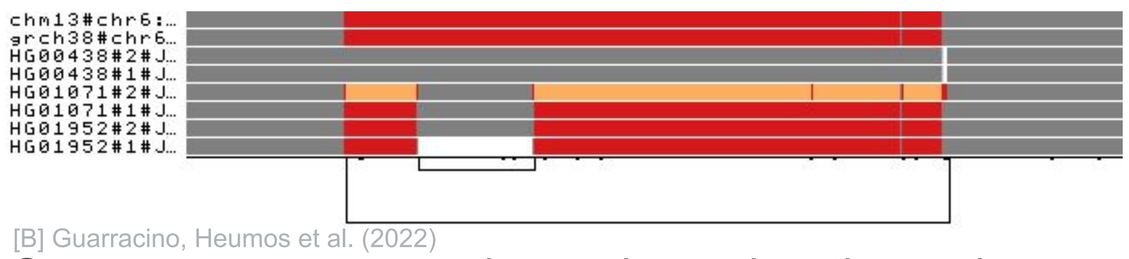


[B] Guarracino, Heumos et al. (2022)

Pangenome graph of the C4 locus of the human major histocompatibility complex. With CHM13 and GRCh38 on top, 8 human haplotypes are shown.

- The graph nodes are arranged from left to right, forming the pangenome sequence.
- The colored bars represent the paths versus the pangenome sequences in a binary matrix.
- The path names are placed on the left.
- The black lines under the paths are the links, which represent the graph topology.

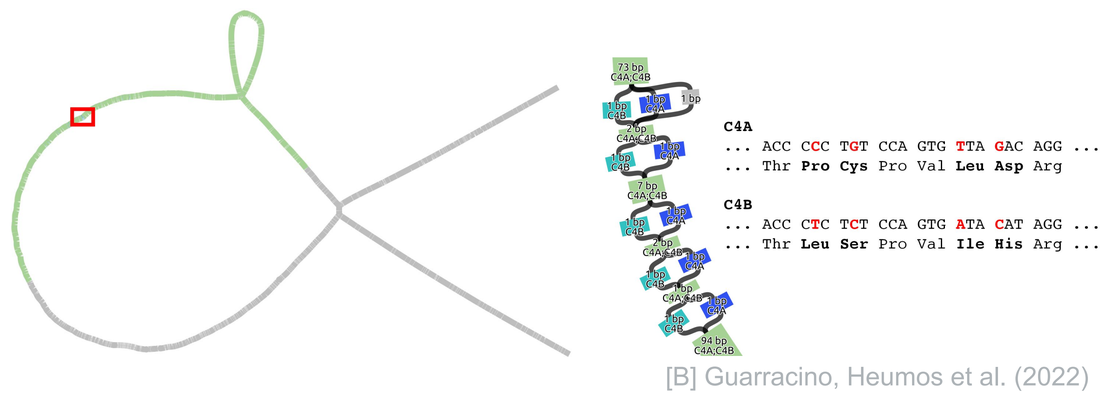
INSPECTING COMPLEX REGIONS



[B] Guarracino, Heumos et al. (2022)

Same pangenome graph as above, but the nodes are colored by the number of times, they are crossed by a path. Shown are four levels of depth, white indicates no depth, while gray, red and yellow indicate depth 1, 2 and greater than or equal to 3, respectively.

ANNOTATING PANGENOME GRAPHS



[B] Guarracino, Heumos et al. (2022)

Annotated 2D layout of the C4 locus where the C4A and C4B genes differ due to single nucleotide variants leading to changes in the encoded protein sequences. Bubbles in the figure indicate regions where paths diverge or repetitive loci.

1

Universität Tübingen
Quantitative Biology Center (QBiC)
Auf der Morgenstelle 10, D-72076 Tübingen
Phone +49 7071 29-76499 · http://qbic.life

2

HUMAN TECHNOPOLE

3

THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER

LITERATURE

A. Eizenga et al. (2020). Pangenome Graphs. *Annual Reviews of Genomics and Human Genetics*, 21, 1.
B. Guarracino, Heumos et al. (2022). ODGI - understanding pangenome graphs. *Bioinformatics*, btac308.