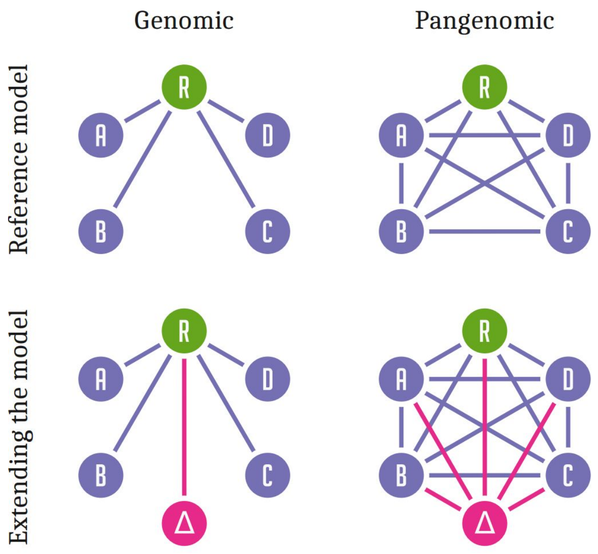
A pangenome for the expanded BXD family of mice

Flavia Villani[1], David G. Ashbrook[1], Andrea Guarracino[2], Simon Heumos[3], Christian Fischer[1], Hao Chen[1], ,Pjotr Prins[1], Erik Garrison[1], Robert W. Williams[1] and Vincenza Colonna[1,4]

[1]UTHSC, GGI, Memphis, TN, USA; [2]University of Tor Vergata, Biology, Rome, Italy; [3]Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany; [4]CNR, IGB, Naples, Italy

1. Genomic vs. pangenomic

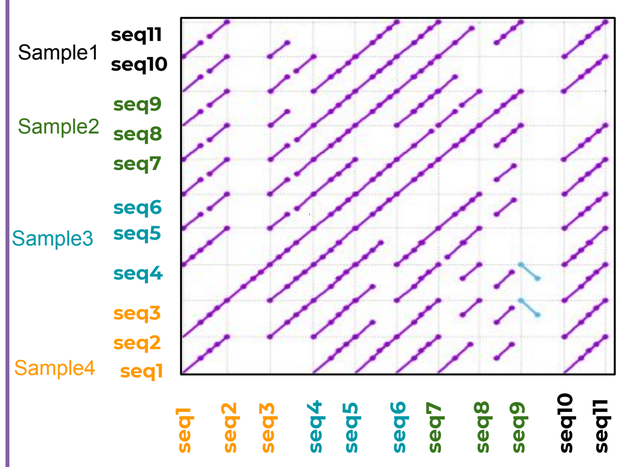


Eizenga[1] et al.(2020)

- In reference b(i)ased **genomic** analyses, all genomes are related to a reference genome.
- In a **pangenomic** setting[1], we model direct relationships between all the genomes in our analysis.

https://pangenome.github.io/

2. Build a pangenome with the PanGenome Graph Builder (PGGB)

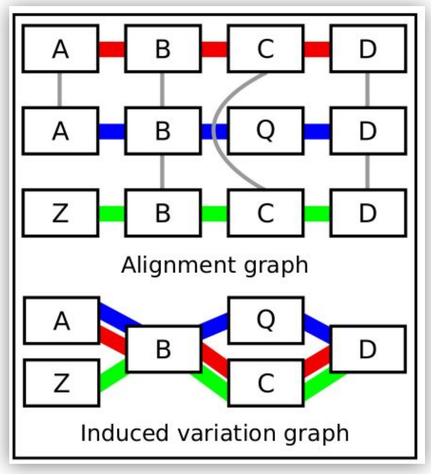


(A) ALL-VS-ALL ALIGNMENTS WITH WFMASH

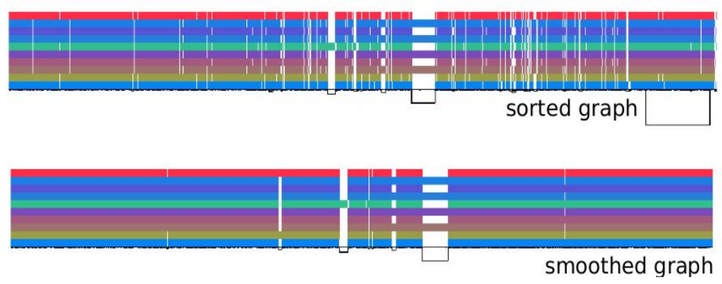
A hierarchical implementation of the wavefront algorithm[2] allows us to obtain base-level global alignments for all mappings.

(B) GRAPH INDUCTION WITH SEQWISH

To build a variation graph, we collapse the components of the graph connected by alignments and retrace the original paths through the graph.



Code available at https://github.com/pangenome/pggb



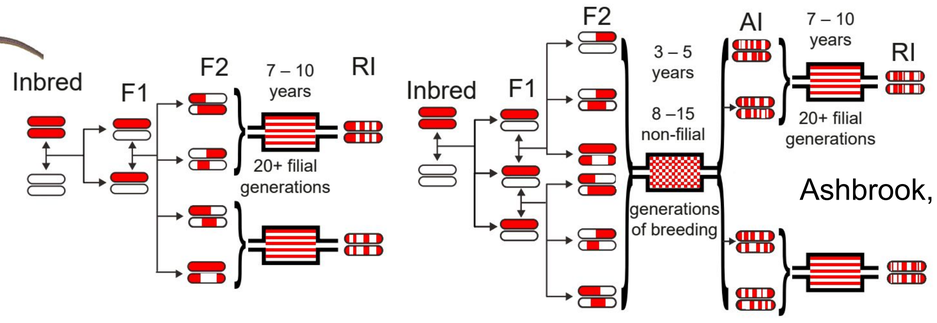
(C) GRAPH NORMALIZATION WITH SMOOTHXG

A partial order alignment is run for each part of the graph, resulting in a normalized graph with base-level resolution of all variant classes.

PGGB has three distinct phases (**A**) all-to-all alignment with wfmash, (**B**) graph induction with seqwish, and (**C**) normalization with smoothxg, which produces the resulting graph shown.

3. A pangenome for the BXD family

- 152 BXD strains sequenced (10x technology).
- 148 BXD strains assembled (30-60 thousand contigs).
- Contig length peak at 31kb.

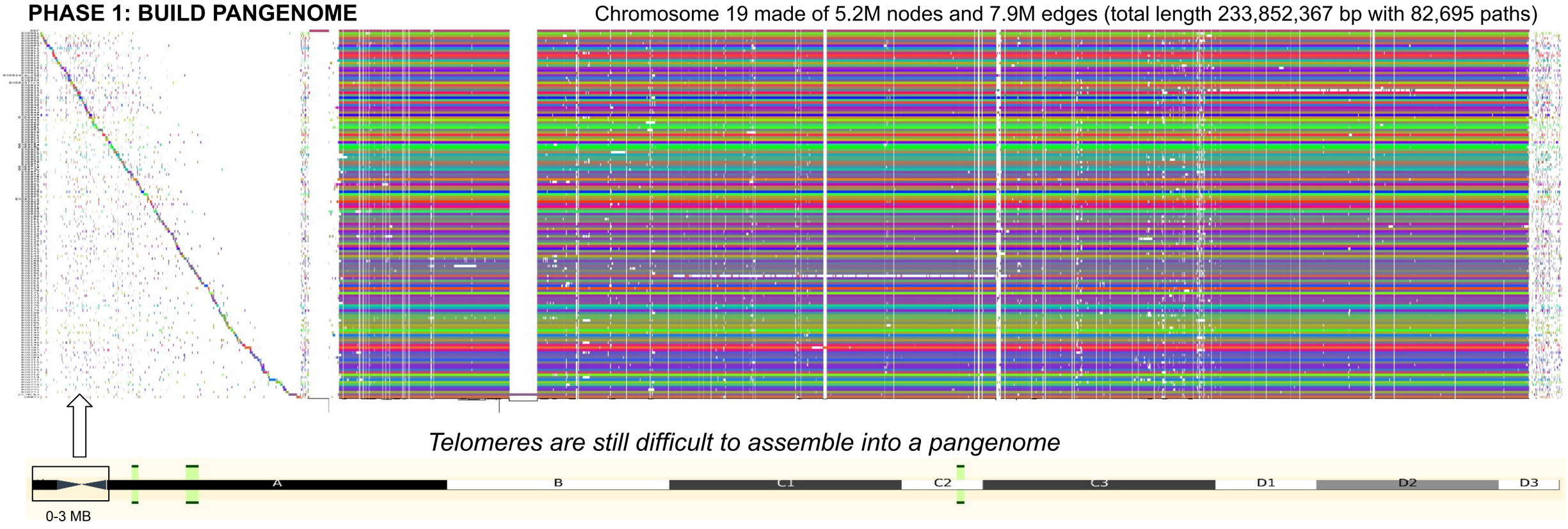


Ashbrook, D. G., et al[3]. Cell Systems (2021).

A graphical representation of the pangenome of 148 family members (all isogenic inbred)

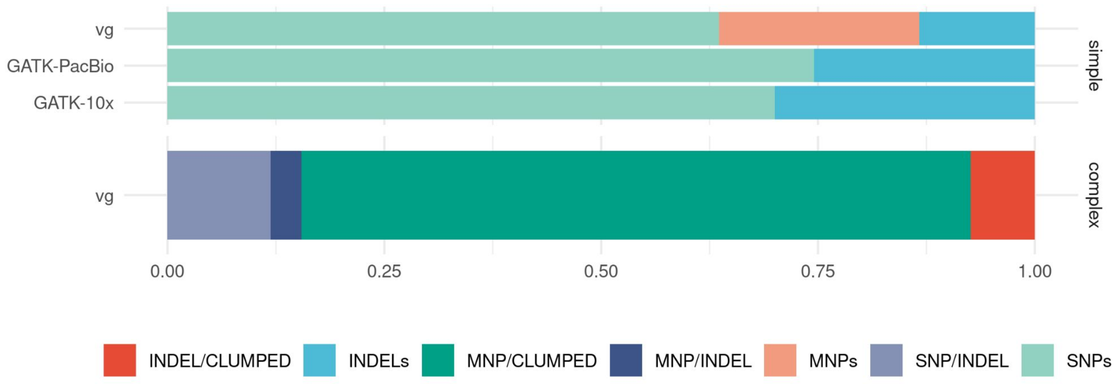
PHASE 1: BUILD PANGENOME

Chromosome 19 made of 5.2M nodes and 7.9M edges (total length 233,852,367 bp with 82,695 paths)

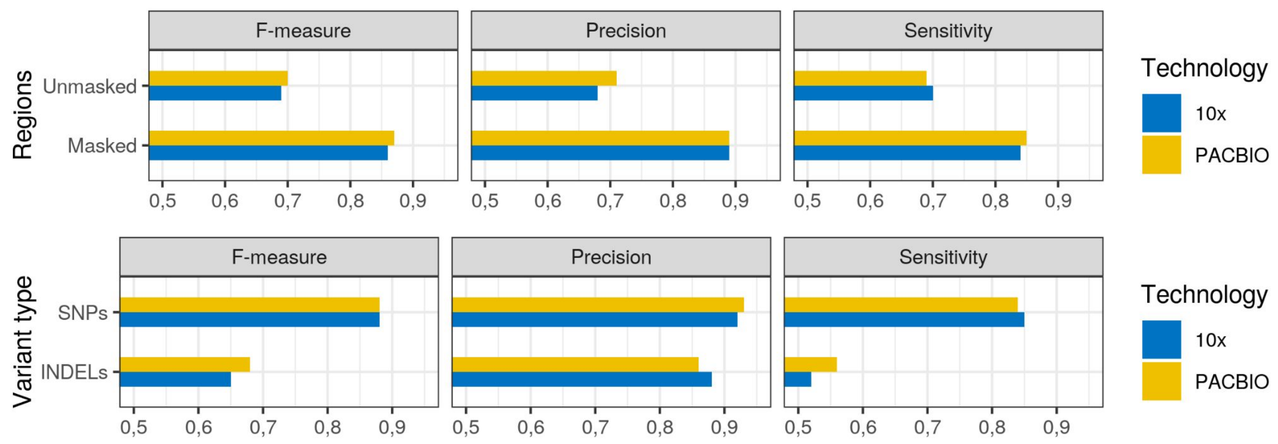


Telomeres are still difficult to assemble into a pangenome

PHASE 2: VARIANT CALLING



PHASE 3: VG *vs* TRUTH SETS



In DBA/2J we called from the pangenome with vg 218,139 simple microvariants (SNPs, MNPs, and INDELs), that is 45,208 and 39,937 more variants compared to GATK applied to 10x and PACBIO sequence data, respectively.

Future work

- Identification of complex variants.
- Use nanopore technology.

References

1. Eizenga et al. (2020). Pangenome Graphs. *Annual Reviews of Genomics and Human Genetics*, 21, 1.
2. Marco-Sola et al. (2020). Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics*.
3. Ashbrook, David G., et al. "A platform for experimental precision medicine: the extended BXD mouse family." *Cell Systems* 12.3 (2021): 235-247

Acknowledgments

IGB Consiglio Nazionale delle Ricerche Institute of Genetics and Biophysics Adriano Buzzati-Traverso

Consiglio Nazionale delle Ricerche

THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER