# The PanGenome Graph Builder

Andrea Guarracino[1]*, Simon Heumos[2]*, Flavia Villani[3], Emilio Rudbeck[4], Kaisa Thorell[4], Lorenzo Tattini[5], Christian Kubica[6], Sebastian Vorbrugg[6], Christian Fischer[7], Njagi Mwaniki[8], Sven Nahnsen[2], David Ashbrook[7], Robert Williams[7], Hao Chen[7], Vincenza Colonna[3], Pjotr Prins[7], Erik Garrison[7]

[1]University of Tor Vergata, Biology, Rome, Italy
[2]University of Tübingen, QBiC, Tübingen, Germany
[3]CNR, IGB, Naples, Italy
[4]University of Gothenburg, Infectious Diseases, Gothenburg, Sweden
[5]Université Côte d'Azur, CNRS, Nice, France
[6]MPI, Developmental Biology, Tübingen, Germany
[7]UTHSC, Genomics, Memphis, TN, USA
[8]KEMRI-Wellcome Trust, Training, Kilifi, Kenya
*Contributed equally

A pangenome contains the full genomic information of a species. Pangenome graphs provide a compact representation of the mutual alignment of collections of genomes. In these graphs, nodes represent sequences in the pangenome, and paths describe genomes as walks through the graph. We implement a method to construct pangenome graphs, the PanGenome Graph Builder (PGGB), that scales efficiently to large collections of eukaryotic genomes.

PGGB is a modular process consisting of three phases. First, we apply an approximate mapping step based on locality sensitive hashing to rapidly determine syntenic regions in the input. We then use a novel high-order implementation of the wavefront algorithm to obtain the base-level alignment of chromosome-scale mappings. Afterwards, the sequences and alignments between them are transformed into a variation graph by applying the seqwish graph induction algorithm. Finally, we normalize this graph by ordering it with an unsupervised machine learning method and applying partial order alignment to blocks in the sorted order. The output includes not only a normalized graph, but an equivalent multiple sequence alignment, and a set of consensus graphs describing the structural variation in the pangenome.

To evaluate PGGB, we apply it to pangenomes from diverse biological orders, building graphs from *S. cerevisiae*, *H. pylori*, *A. thaliana*, *M. musculus*, and *H. sapiens* datasets. For smaller pangenomes, PGGB runs quickly enough to provide interactive feedback on computing systems with minimal resources. But, it comfortably scales to multi-gigabase mammalian genomes, requiring ~12 hours on nine 128GB/48vCPU compute nodes to build a human pangenome from 92 haplotypes. While our analysis of pangenome representation accuracy shows that existing methods for pangenome graph construction are strongly reference biased, we find that PGGB allows unbiased evaluation of sequence variation of all types, including SNPs, small INDELs, as well as large structural variation. Large INDELs, VNTRs, and translocations are naturally represented in compact motifs in the graph. We anticipate a multitude of applications of the

resulting graphs, as they provide excellent targets for the mapping of short and long reads, and are a basis for comparative genomic applications.