Pangenomics

Sandra Smit

12 Oct 2023, BioSB Algorithms for Genomics

WAGENINGEN
UNIVERSITY & RESEARCH

Program today

- 9.00 Lecture
- 11.30 Break
- 11.00 Computer exercises (rosalind or project)
- 12.30 Lunch
- 13.30 Guest lecture by Erwin Datema from Keygene
- 14.15 Lecture about PanTools
- 15.00 Break
- 15.30 Computer exercises (rosalind or project or PanTools)
- 17.30 Closing

Learning outcomes

- Understand the need for pangenomes
- Know what a pangenome is
- Be aware of various applications
- Get a handle on this fast-developing field
    - Tools
    - Concepts
    - Data structures
    - Algorithms

Comparative genomics

Genotype x Environment

?

Phenotype

Comparative genomics

Genotype x Environment

?

Phenotype

Comparative genomics

Genotype x Environment

?

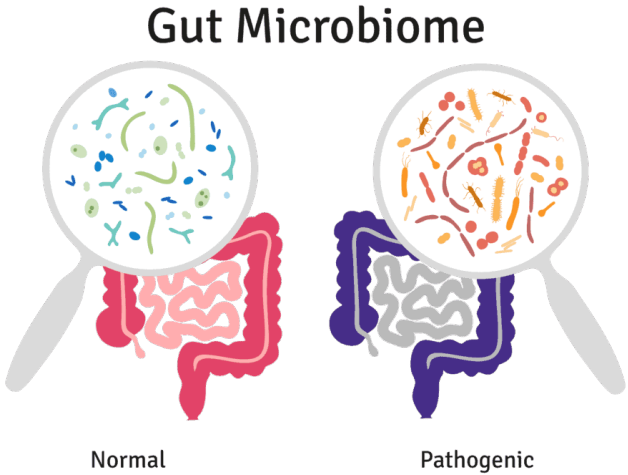
Phenotype

Comparative genomics

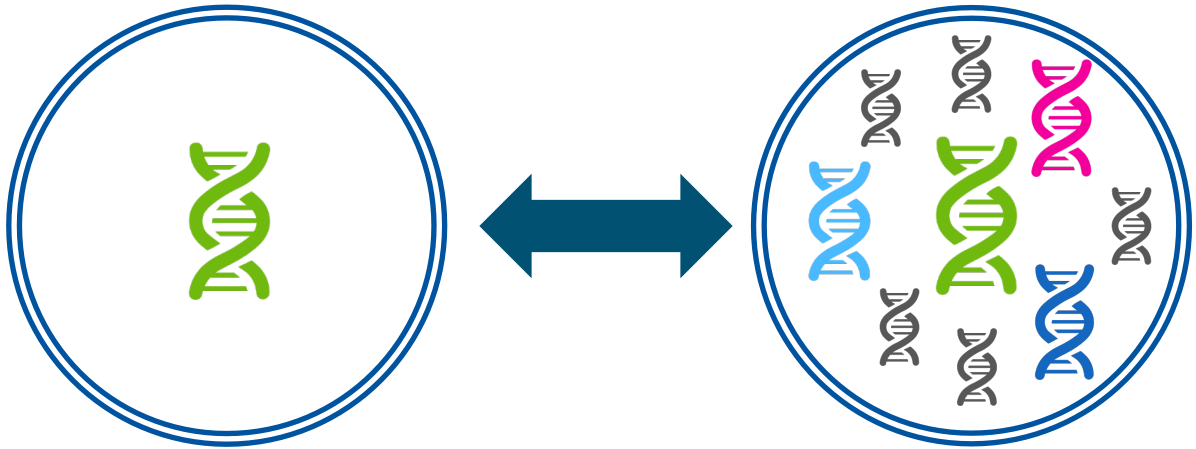
Genotype x Environment

?

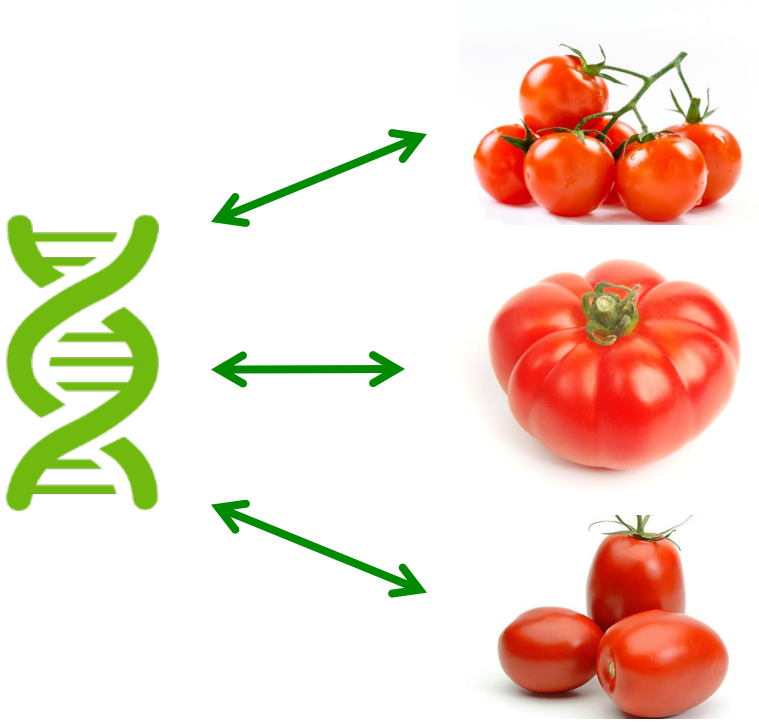
Phenotype



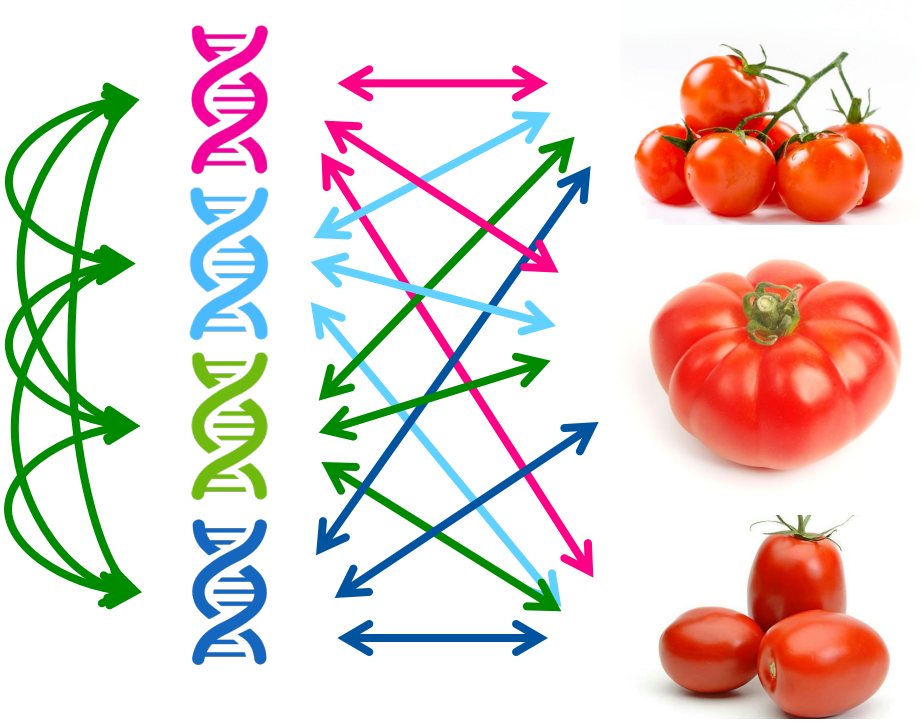
A paradigm shift…



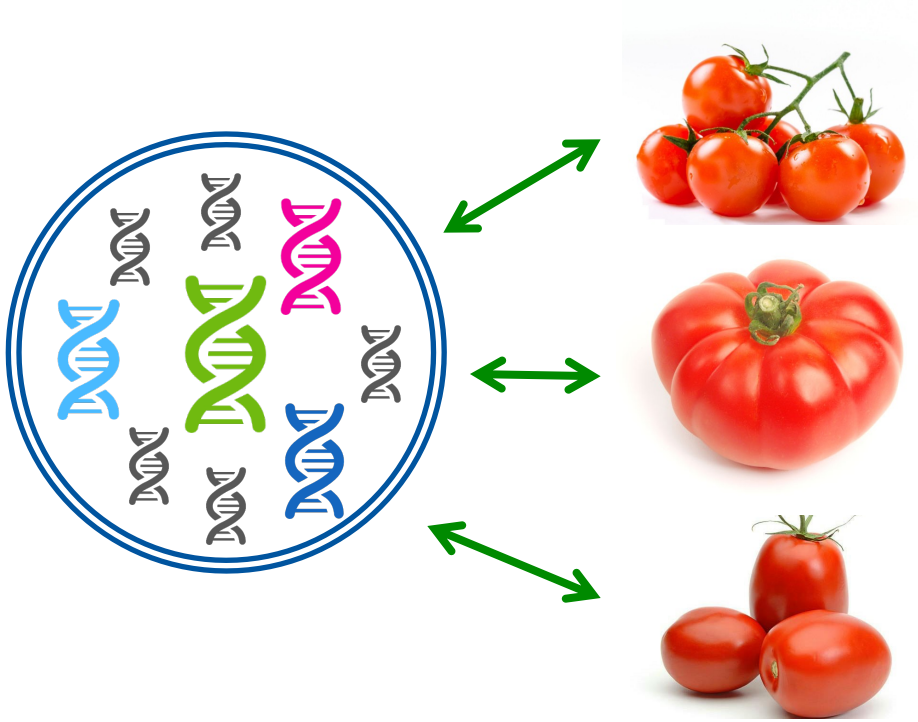
The reference genome



Variation and diversity



Pangenome

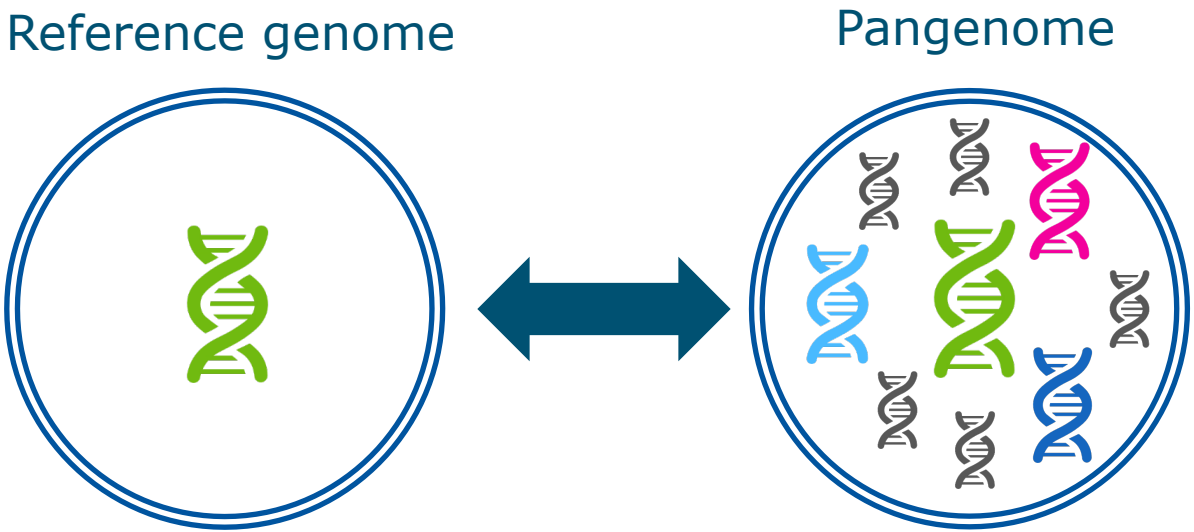


Pangenome

Single representation for many genomes

- Compress similarities
- Retain (all) variations

Why pangenomes?



Reference genome benefits

Reference genome

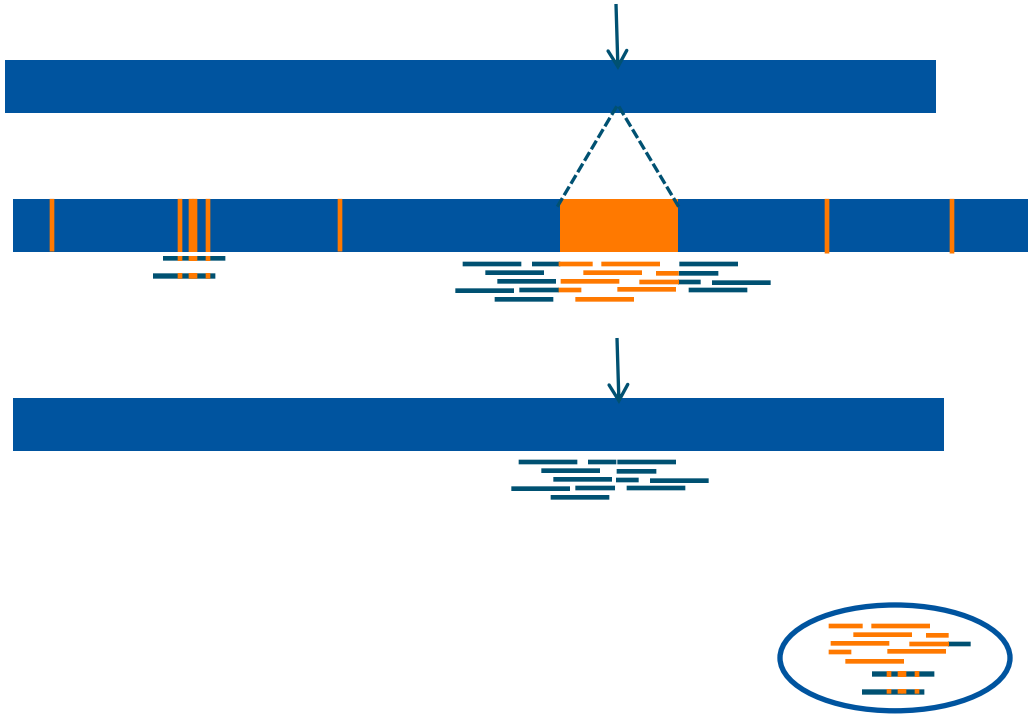
- Linear haploid representation
- Fast searching/mapping
- Indexing
- Sequence/string algorithms

Reference bias

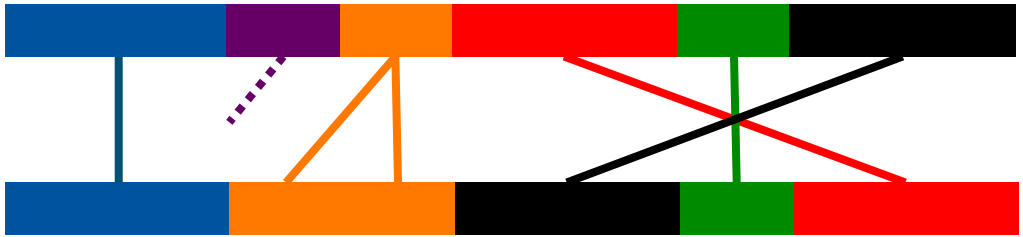
Reference genome

Sampled genome

Reference bias



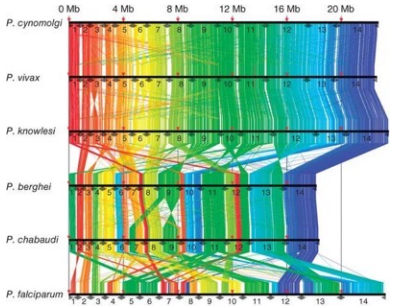
There is more than sequence variation



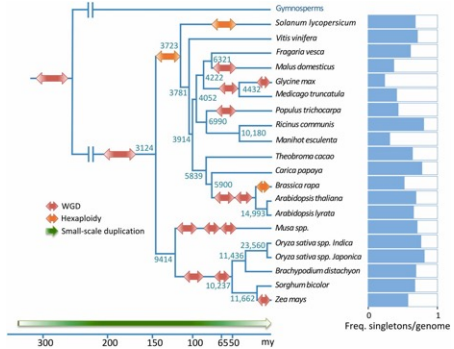
Reference genome

Sampled genome

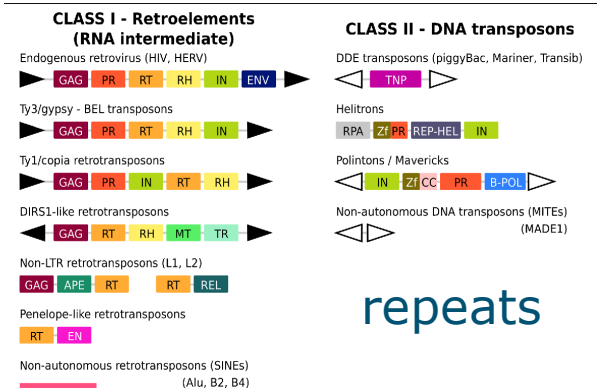
There is more than sequence variation



structural variations



ancient duplications



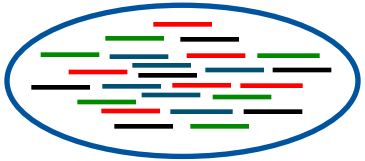
repeats

Polyploidy and heterozygosity

```
.......A.......T.......G.......   A
.......T.......C.......G.......   B
.......A.......T.......C.......   C
.......T.......T.......A.......   D
```

Polyploidy and heterozygosity

```
.......A.......T.......G.......   A
.......T.......C.......G.......   B
.......A.......T.......C.......   C
.......T.......T.......A.......   D
```



Tetraploid sample

Polyploidy and heterozygosity

```
.......A.......T.......G.......   A
.......T.......C.......G.......   B
.......A.......T.......C.......   C
.......T.......T.......A.......   D
```

```
Reference   .......T.......T.......G.......

SNPs              A       C       C
                                  A
```

Why pangenomes?

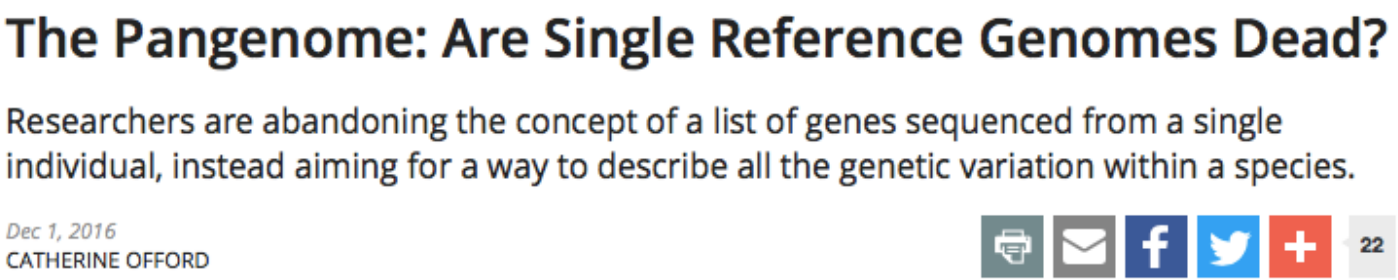
Why pangenomes?

Reference genome

Pangenome

Catch **intra**genomic and
intergenomic **variation**

Why pangenomes?



The Pangenome: Are Single Reference Genomes Dead?

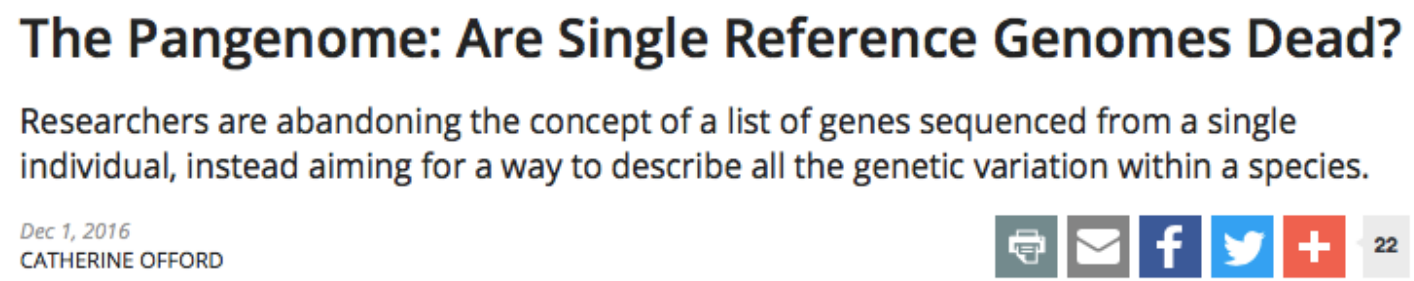
Researchers are abandoning the concept of a list of genes sequenced from a single individual, instead aiming for a way to describe all the genetic variation within a species.

Dec 1, 2016
CATHERINE OFFORD

22

https://www.the-scientist.com/features/the-pangenome-are-single-reference-genomes-dead-32458

Why pangenomes?



The Pangenome: Are Single Reference Genomes Dead?

Researchers are abandoning the concept of a list of genes sequenced from a single individual, instead aiming for a way to describe all the genetic variation within a species.

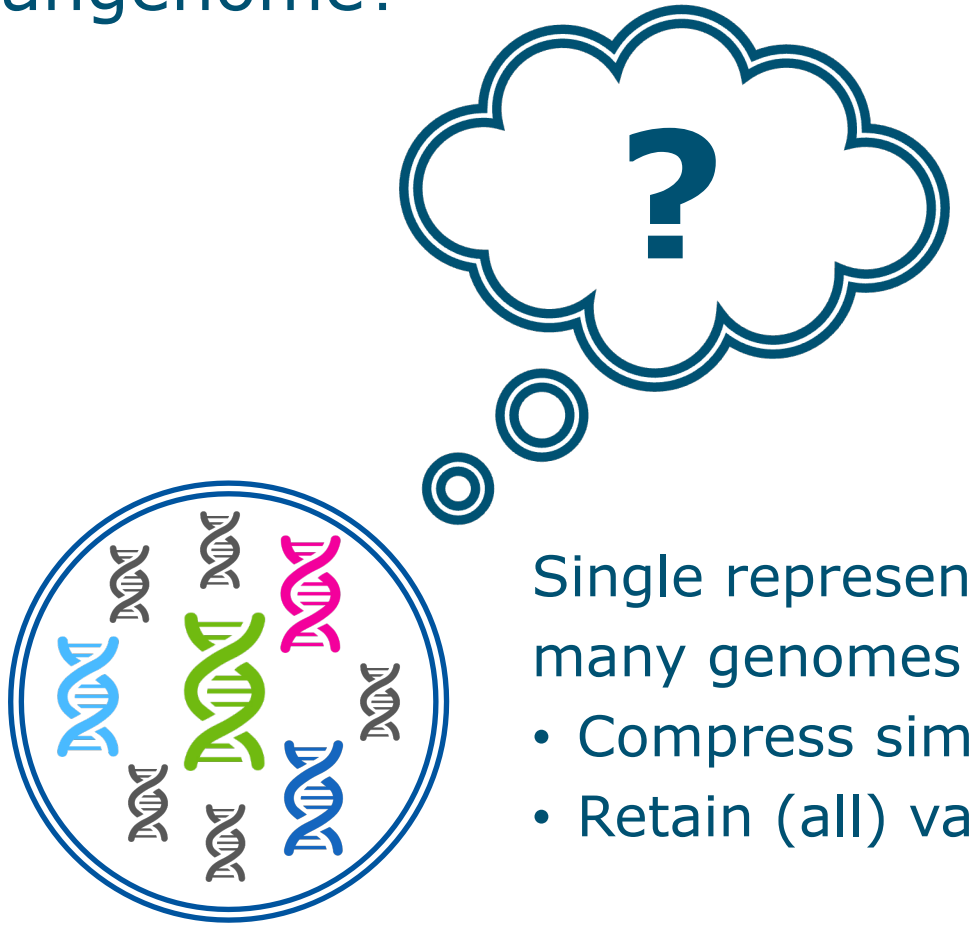
Dec 1, 2016
CATHERINE OFFORD

22

No!

- High-quality reference genomes are crucial input for pangenomes
- But only a single reference is not enough
- THE pangenome does not exist; it is a moving target
- Putting every genome available in a pangenome might not be useful

What is a pangenome?



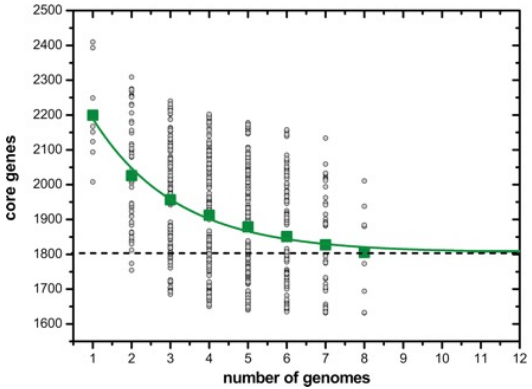
Single representation for many genomes

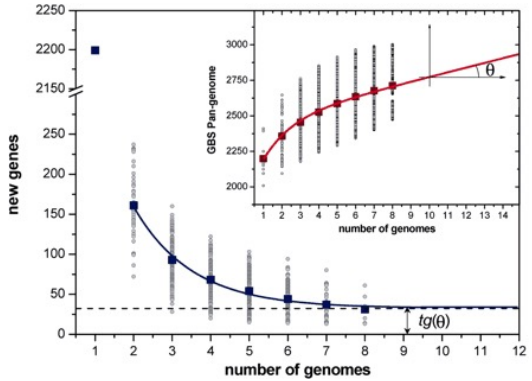
- Compress similarities
- Retain (all) variations

The first pangenome (Tettelin, 2005)

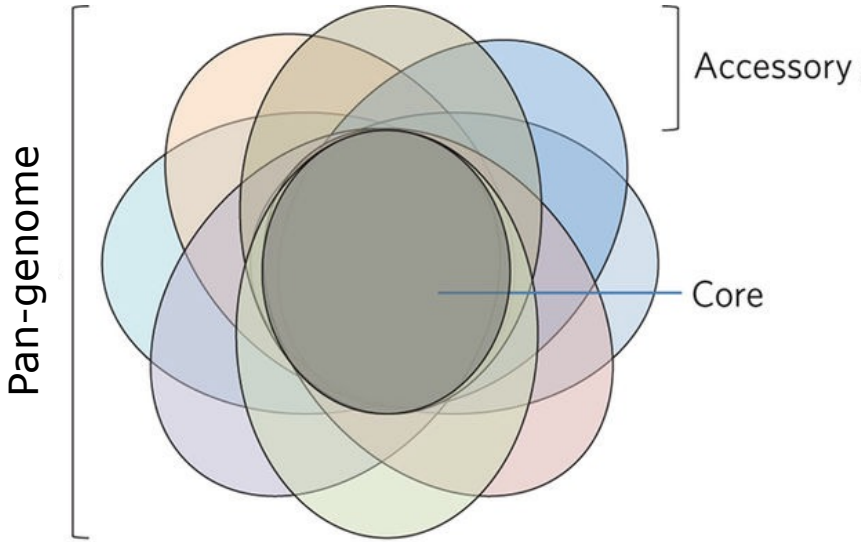
Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

How many genomes are necessary to fully describe a bacterial species?

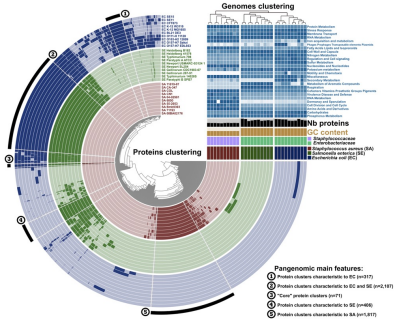




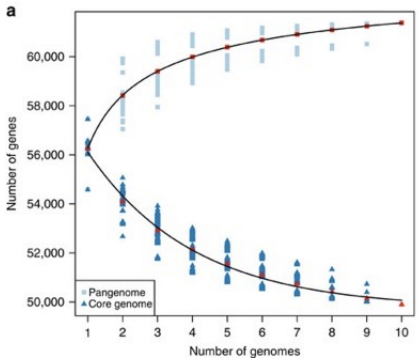
Gene-based pangenome



McInerney *et al.*, 2017

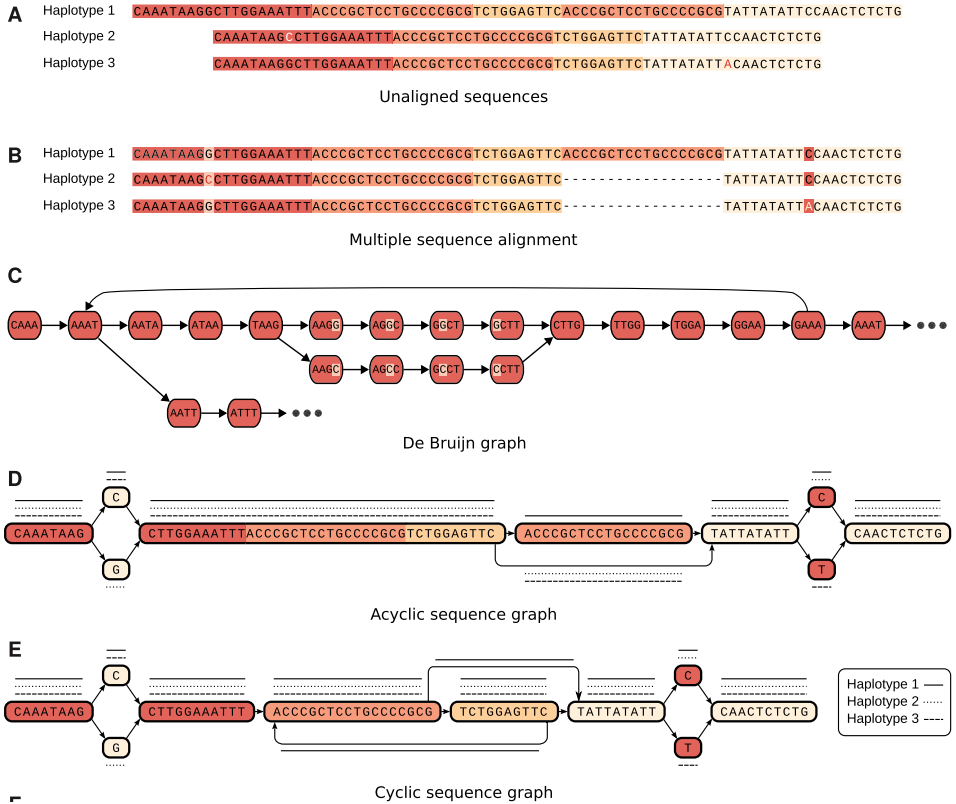


Eren *et al.*, 2015

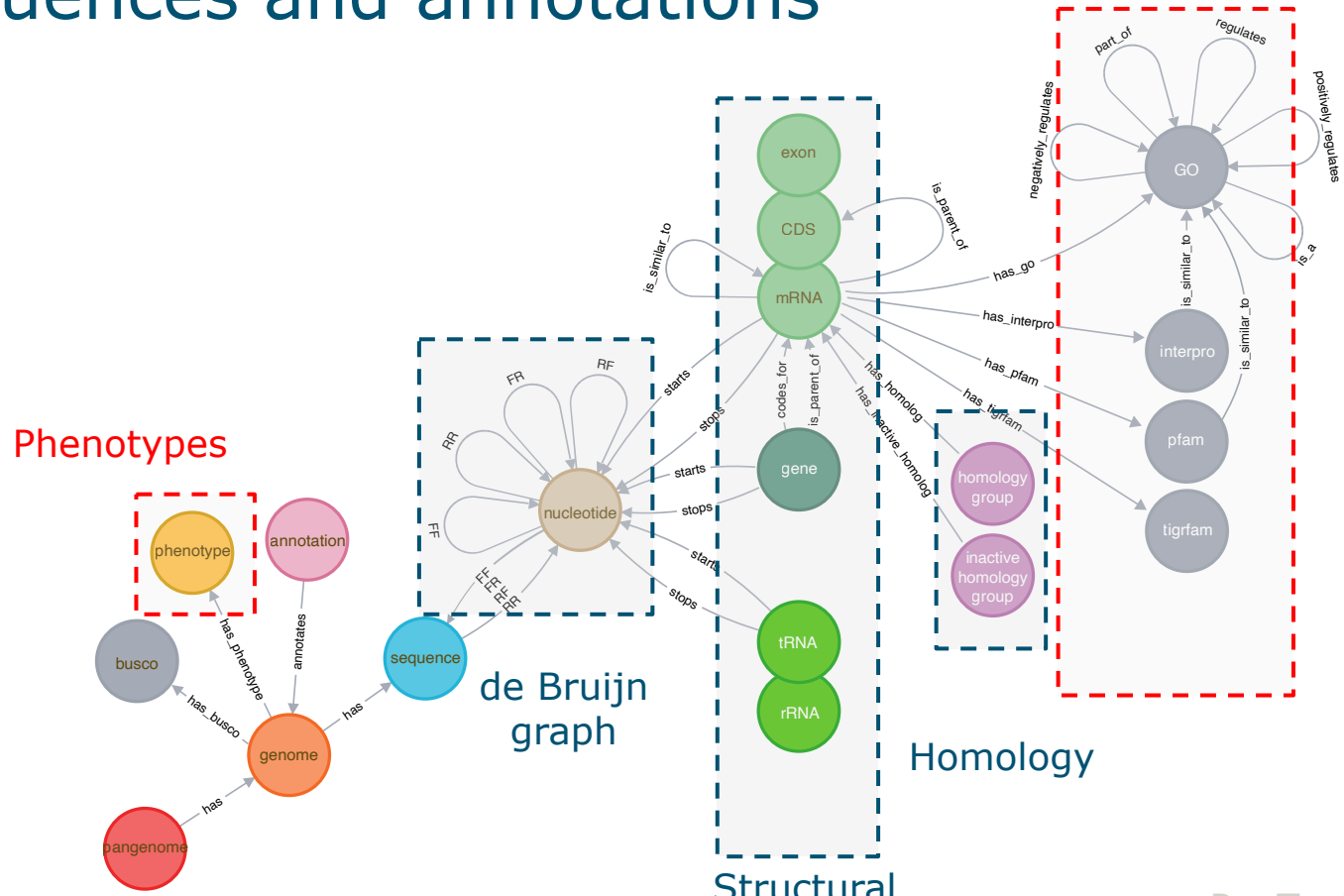


Golicz *et al.*, 2017

Sequence-level pangenome



Sequences and annotations

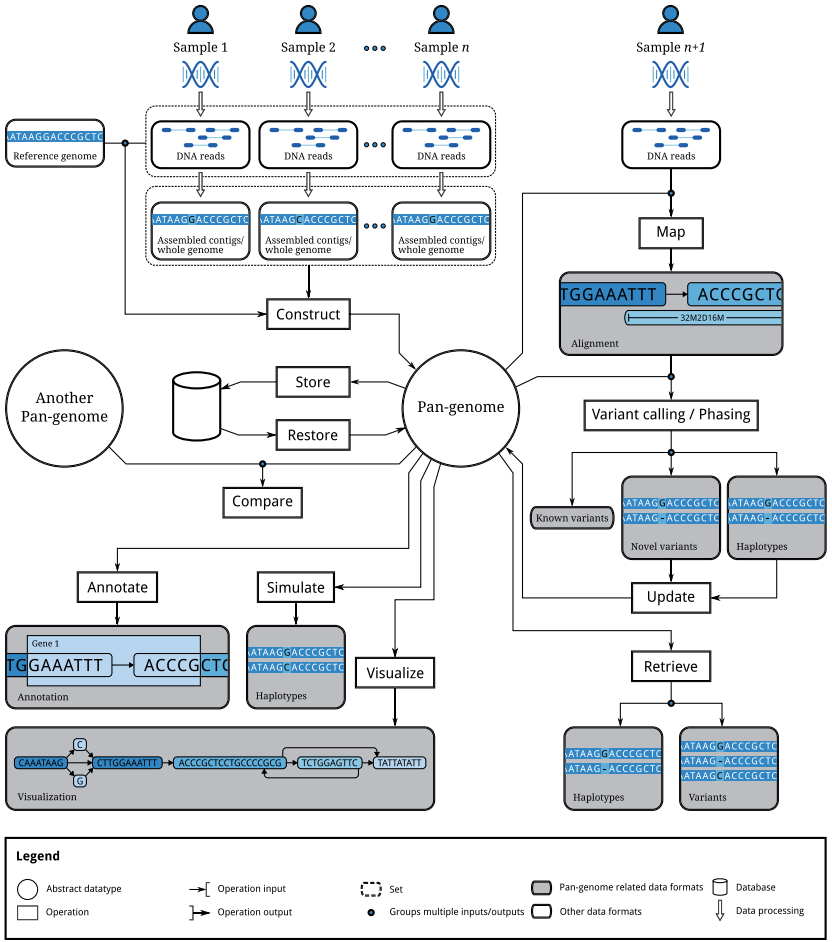


WAGENINGEN
UNIVERSITY & RESEARCH

PanTools v3
Jonkheer et al. (2023)

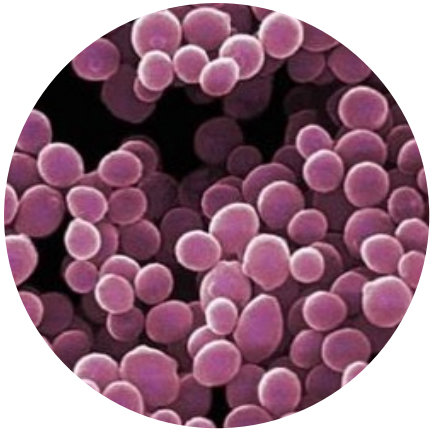
What is a pangenome?

- Variation-aware reference
  - Construct
  - Annotate
  - Map
  - Variant calling
  - Visualize
  - …



Marschall *et al.*, *2016*

Pangenome applications



Health

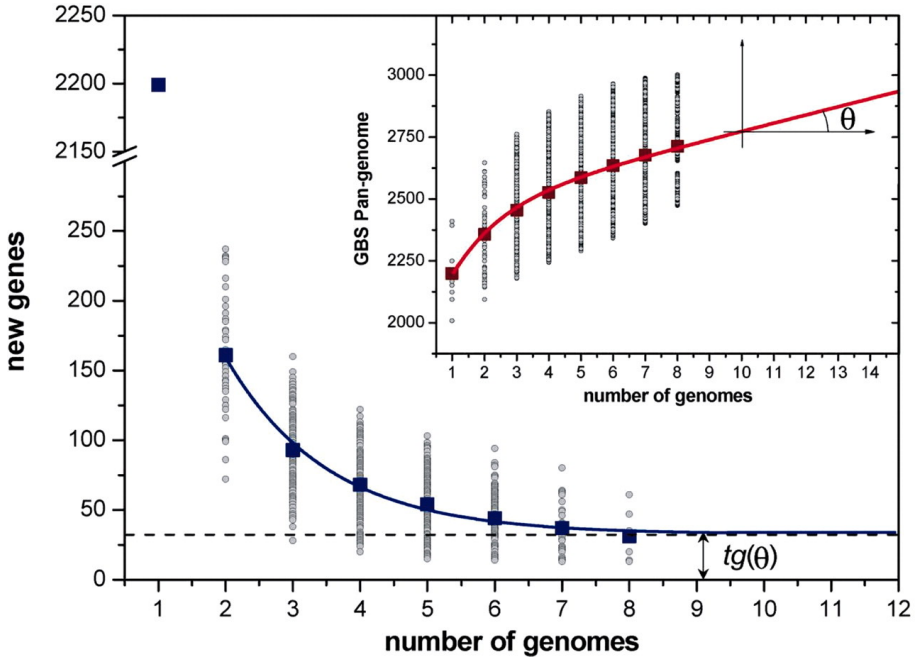
Agriculture

Biotechnology

The first microbial pangenome (2005)

- *Streptococcus agalactiae*
- Tettelin et al. 2005
- Gene-level pangenome

Unfortunately, the sequence of a single genome does not reflect how genetic variability drives pathogenesis within a bacterial species and also limits genome-wide screens for vaccine candidates or for antimicrobial targets.



Microbial pangenomes

PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome

Kanwal Naz, Anam Naz, Shifa Tariq Ashraf, Muhammad Rizwan, Jamil Ahmad, Jan Baumbach & Amjad Ali

BMC Bioinformatics **20**, Article number: 123 (2019)

panX: pan-genome analysis and exploration

Wei Ding, Franz Baumdicker, Richard A Neher

Nucleic Acids Research, Volume 46, Issue 1, 9 January 2018, Page e5, https://doi.org/10.1093/nar/gkx977

Published: 25 October 2017 **Article history**

Article | Published: 24 November 2021

Aspergillus fumigatus pan-genome analysis identifies genetic variants associated with human infection

Amelia E. Barber, Tongta Sae-Ong, Kang Kang, Bastian Seelbinder, Jun Li, Grit Walther, Gianni Panagiotou & Oliver Kurzai

Nature Microbiology **6**, 1526–1536 (2021) | Cite this article

Human applications

Variation graph toolkit improves read mapping by representing genetic variation in the reference

Erik Garrison[1], Jouni Sirén[1], Adam M Novak[2], Glenn Hickey[2], Jordan M Eizenga[2], Eric T Dawson[1,3,4], William Jones[1], Shilpa Garg[5], Charles Markello[2], Michael F Lin[6], Benedict Paten[2] & Richard Durbin[1,4]

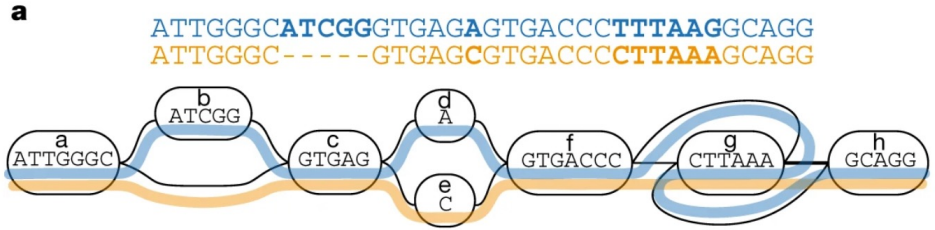
Assembly of a pan-genome from deep sequencing of 910 humans of African descent

Rachel M. Sherman, Juliet Forman, […] Steven L. Salzberg

Nature Genetics **51**, 30–35 (2019) | Download Citation

The first human pangenome (2023)

- 47 phased, diploid assemblies from a cohort of genetically diverse individuals



Article | Open Access | Published: 10 May 2023

A draft human pangenome reference

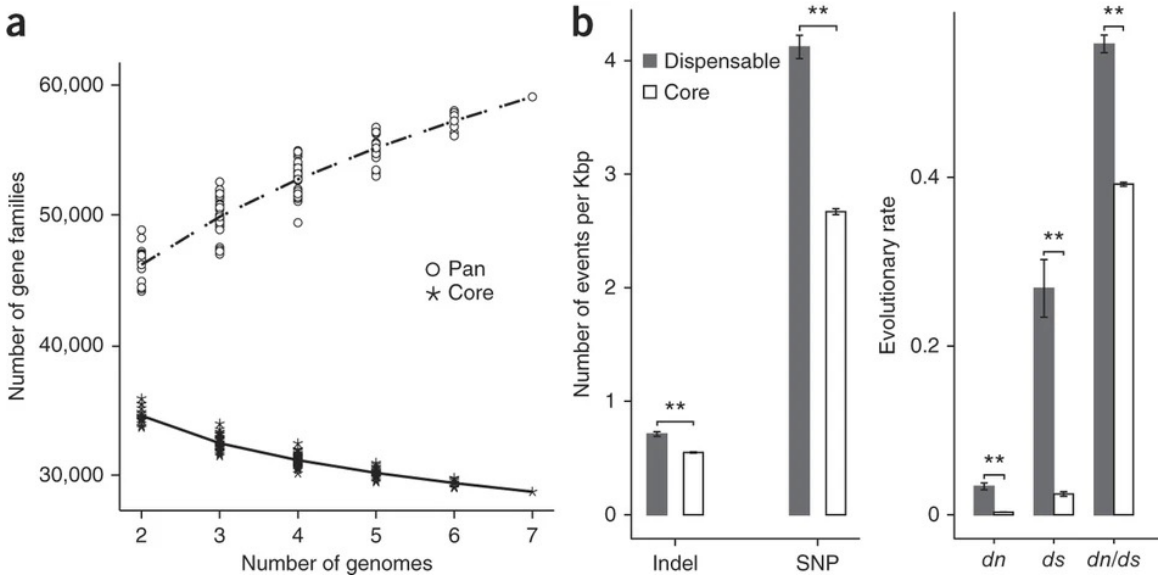
Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, Simon Heumos, ... Benedict Paten

+ Show authors

Nature **617**, 312–324 (2023) | Cite this article

The first plant pangenome (2014)

- Soybean
- Li et al. (2014)
- *Glycine soja*, the wild relative of cultivated soybean *Glycine max*
- 7 genomes



Since then…

Year published	Species	#accessions	Core genome
2014	*Brassica rapa*	3	87%
	Soybean	7	49%
	Rice	3	92%
	Maize	503	39%
2015	Rice	1483	?
2016	*Brassica oleracea*	10	81%
	Poplar	7	<90%
2017	*Brachypodium distachyon*	54	55%
	Medicago truncatula	15	33%
	Wheat	19	64%
2018	*Brassica napus*	53	62%
	Pepper	383	56%
	Rice	67	62%
	Rice	3010	54-62%
2019	Sesame	5	58%

WAGENINGEN
UNIVERSITY & RESEARCH

Adapted from
Bayer et al. (2020)

Pangenomics – plants

Article | Published: 13 May 2019

The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor

Lei Gao, Itay Gonda, Honghe Sun, Qiyue Ma, Kan Bao, Denise M. Tieman, Elizabeth A. Burzynski-Chang, Tara L. Fish, Kaitlin A. Stromberg, Gavin L. Sacks, Theodore W. Thannhauser, Majid R. Foolad, Maria Jose Diez, Jose Blanca, Joaquin Canizares, Yimin Xu, Esther van der Knaap, Sanwen Huang, Harry J. Klee, James J. Giovannoni ✉ & Zhangjun Fei ✉

Nature Genetics **51**, 1044–1051 (2019) | Cite this article

Pangenomics – plants

Article | Published: 13 May 2019

The tomato pan-gene
allele re

Article | Open Access | Published: 08 June 2022

Graph pangenome captures missing heritability and empowers tomato breeding

Yao Zhou, Zhiyang Zhang, Zhigui Bao, Hongbo Li, Yaqing Lyu, Yanjun Zan, Yaoyao Wu, Lin Cheng, Yuhan Fang, Kun Wu, Jinzhe Zhang, Hongjun Lyu, Tao Lin, Qiang Gao, Surya Saha, Lukas Mueller, Zhangjun Fei, Thomas Städler, Shizhong Xu, Zhiwu Zhang, Doug Speed & Sanwen Huang

Nature 606, 527–534 (2022) | Cite this article

28k 33k Accesses | 42 Citations | 163 Altmetric | Metrics

Pangenomics – plants

Article | Published: 13 May 2019

The tomato pan-gene[illegible]

Article | Open Access | Published: 08 June 2022

[illegible] Access | Published: 06 April 2023

[illegible]nome captures missing heritability and [illegible]breeding

[illegible]ging Lyu, Yanjun Zan, Yaoyao Wu, Lin Cheng, [illegible]g Gao, Surya Saha, Lukas Mueller, [illegible]8. Sanwen Huang

Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species

Ning Li, Qiang He, Juan Wang, Baike Wang, Jiantao Zhao, Shaoyong Huang, Tao Yang, Yaping Tang, Shengbao Yang, Patiguli Aisimutuola, Ruiqiang Xu, Jiahui Hu, Chunping Jia, Kai Ma, Zhiqiang Li, Fangling Jiang, Jie Gao, Haiyan Lan, Yongfeng Zhou, Xinyan Zhang, Sanwen Huang, Zhangjun Fei, Huan Wang, Hongbo Li & Qinghui Yu

Nature Genetics 55, 852–860 (2023) | Cite this article

12k Accesses | 27 Altmetric | Metrics

WAGENINGEN
UNIVERSITY & RESEARCH

Animal pangenomes

Science China Life Sciences
pp 1–14 | Cite as

Building a sequence map of the pig pan-genome from multiple *de novo* assemblies and Hi-C data

Authors

Authors and affiliations

ei Fu, Yan Li, Xihong Wang, Ming Li, Duo Du, Qianzi Tang, Yudong Cai, Yiming Long, Yue Zhao,

The Chicken Pan-Genome Reveals Gene Content Variation and a Promoter Region Deletion in *IGF2BP1* Affecting Body Size

Kejun Wang, Haifei Hu, Yadong Tian, Jingyi Li, Armin Sche... Chenyi Zhang, Yivi Li
Junfeng Wu, Lan Yang, Xuewei Fan ... Show more
Author Notes

Molecular Biology and Evolution, Volume 38, Issue 11, Nov
https://doi.org/10.1093/molbev/msab231
Published: 30 July 2021

Genome Res. 2022 Aug; 32(8): 1585–1601.
doi: 10.1101/gr.276550.122

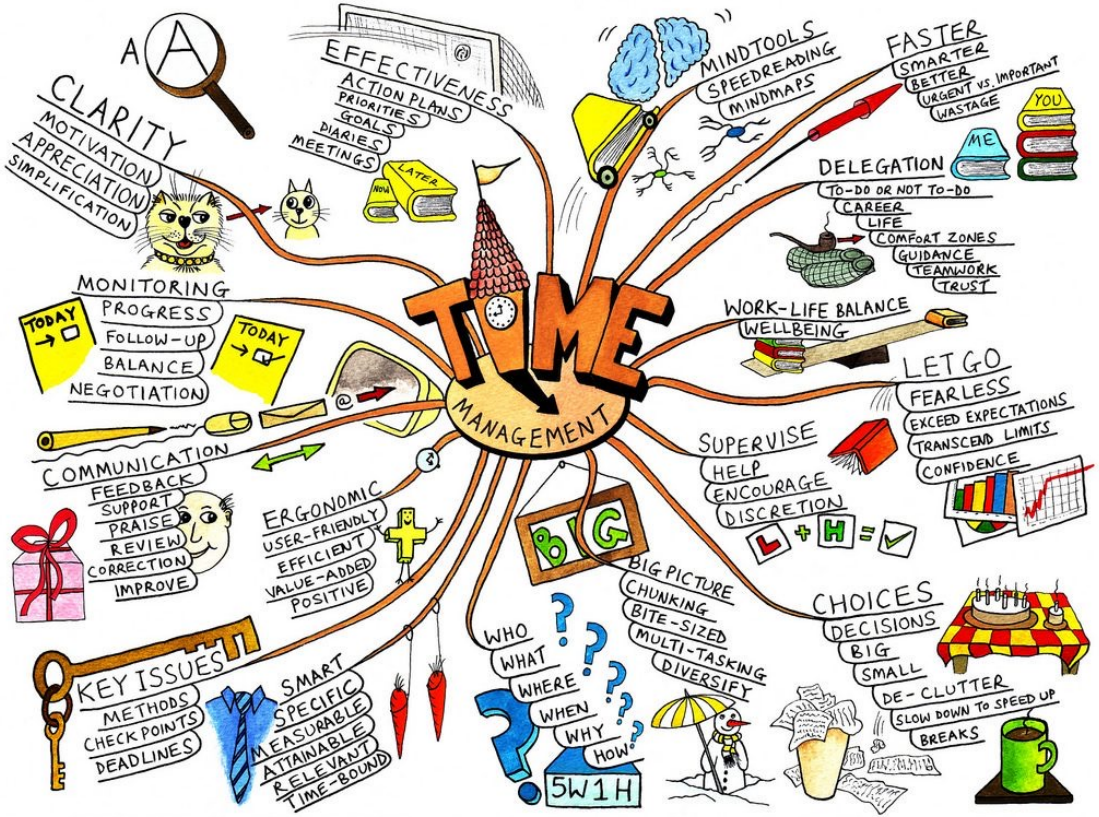
PMCID: PMC9435747
PMID: 35977842

Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history

Yang Zhou,[#1,6] Lv Yang,[#1,6] Xiaotao Han,[1] Jiazheng Han,[1] Yan Hu,[1] Fan Li,[1] Han Xia,[1] Lingwei Peng,[1] Clarissa Boschiero,[2] Benjamin D. Rosen,[2] Derek M. Bickhart,[3] Shujun Zhang,[1] Aizhen Guo,[4] Curtis P. Van Tassell,[2] Timothy P.L. Smith,[5] Liguo Yang,[1] and George E. Liu[2]

Author information · Article notes · Copyright and License information · PMC Disclaimer

Assignment: Mindmap around pangenomics



Pangenomics

A short history

- Concepts
- Algorithms
- Data structures
- Tools

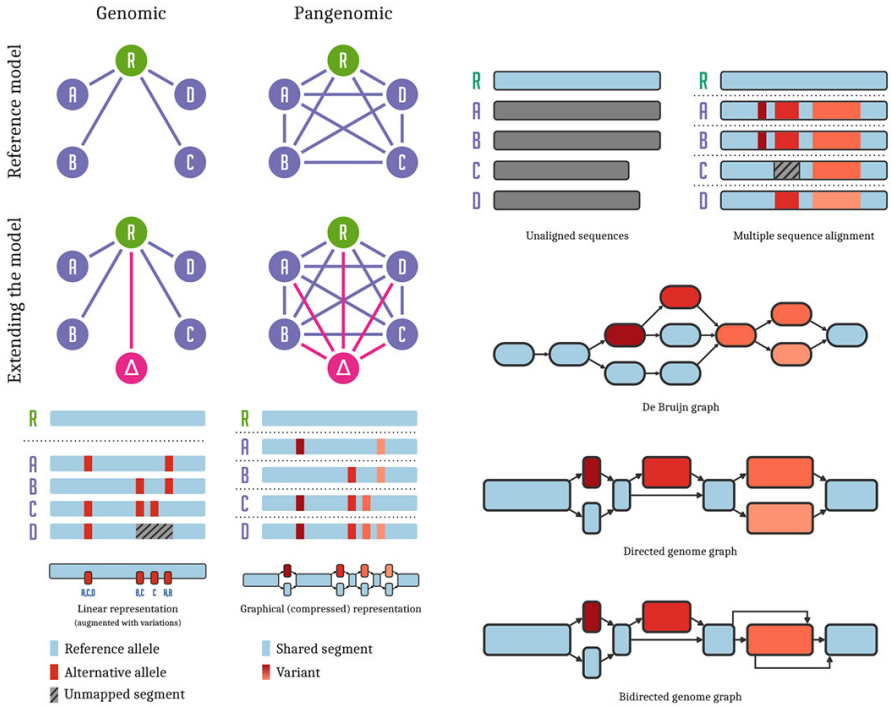
Highly recommended review

Pangenome Graphs.

Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M, Garg S, Paten B, Marschall T, Sirén J, Garrison E.

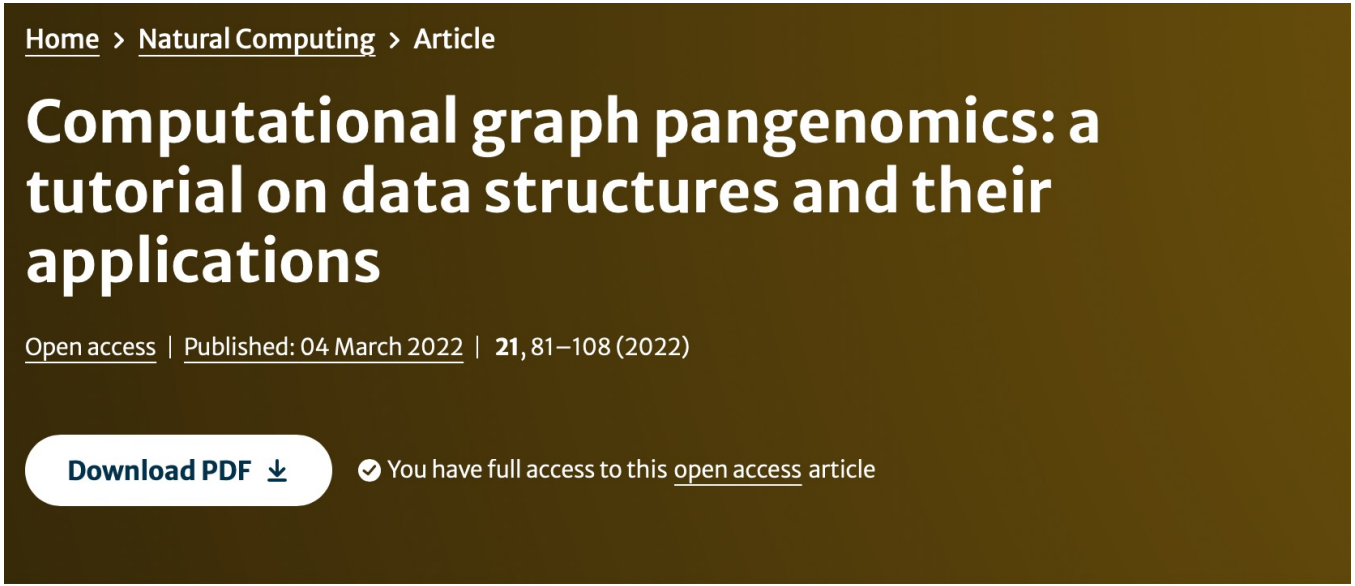
Annu Rev Genomics Hum Genet. 2020 Aug 31;21:139-162. doi: 10.1146/annurev-genom-120219-080406. Epub 2020 May 26.

PMID: 32453966 **Free PMC article.** Review.



Eizenga *et al.*, *2020*

Highly recommended review

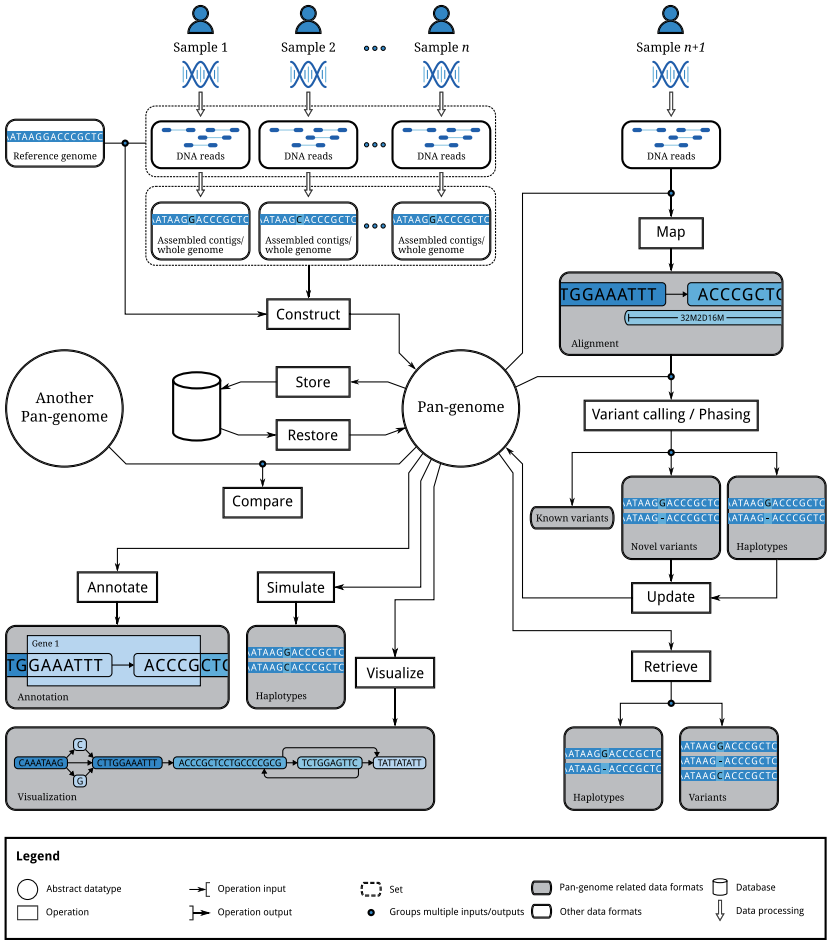


Jasmijn A. Baaijens, Paola Bonizzoni, Christina Boucher, Gianluca Della Vedova, Yuri Pirola, Raffaella Rizzi & Jouni Sirén

WAGENINGEN
UNIVERSITY & RESEARCH

What is a pangenome?

- Variation-aware reference
  - Construct
  - Annotate
  - Map
  - Variant calling
  - Visualize
  - …



Marschall *et al.*, *2016*

Pangenome construction

- Linear pangenome sequence
    - Iterative alignment
- Colored compacted De Bruijn Graph (cDBG)
    - Typically built from whole genomes
- Variation graph (Directed Acyclic Graph)
    - Typically built from reference + vcfs
    - Nowadays built from multiple assemblies

Iterative approach: concept

- Start with reference genome
- Map reads of another genome to reference
- Assemble the unmapped reads
- Extend the reference genome with the assembled sequences
- Continue with the next genome

Iterative approach: example

- Iterative mapping and assembly approach
- Nine morphologically diverse *B. oleracea* varieties and a wild relative—*Brassica macrocarpa*.

Open access | Published: 11 November 2016

The pangenome of an agronomically important crop plant *Brassica oleracea*

Agnieszka A. Golicz, Philipp E. Bayer, Guy C. Barker, Patrick P. Edger, HyeRan Kim, Paula A. Martinez, Chon Kit Kenneth Chan, Anita Severn-Ellis, W. Richard McCombie, Isobel A. P. Parkin, Andrew H. Paterson, J. Chris Pires, Andrew G. Sharpe, Haibao Tang, Graham R. Teakle, Christopher D. Town, Jacqueline Batley & David Edwards

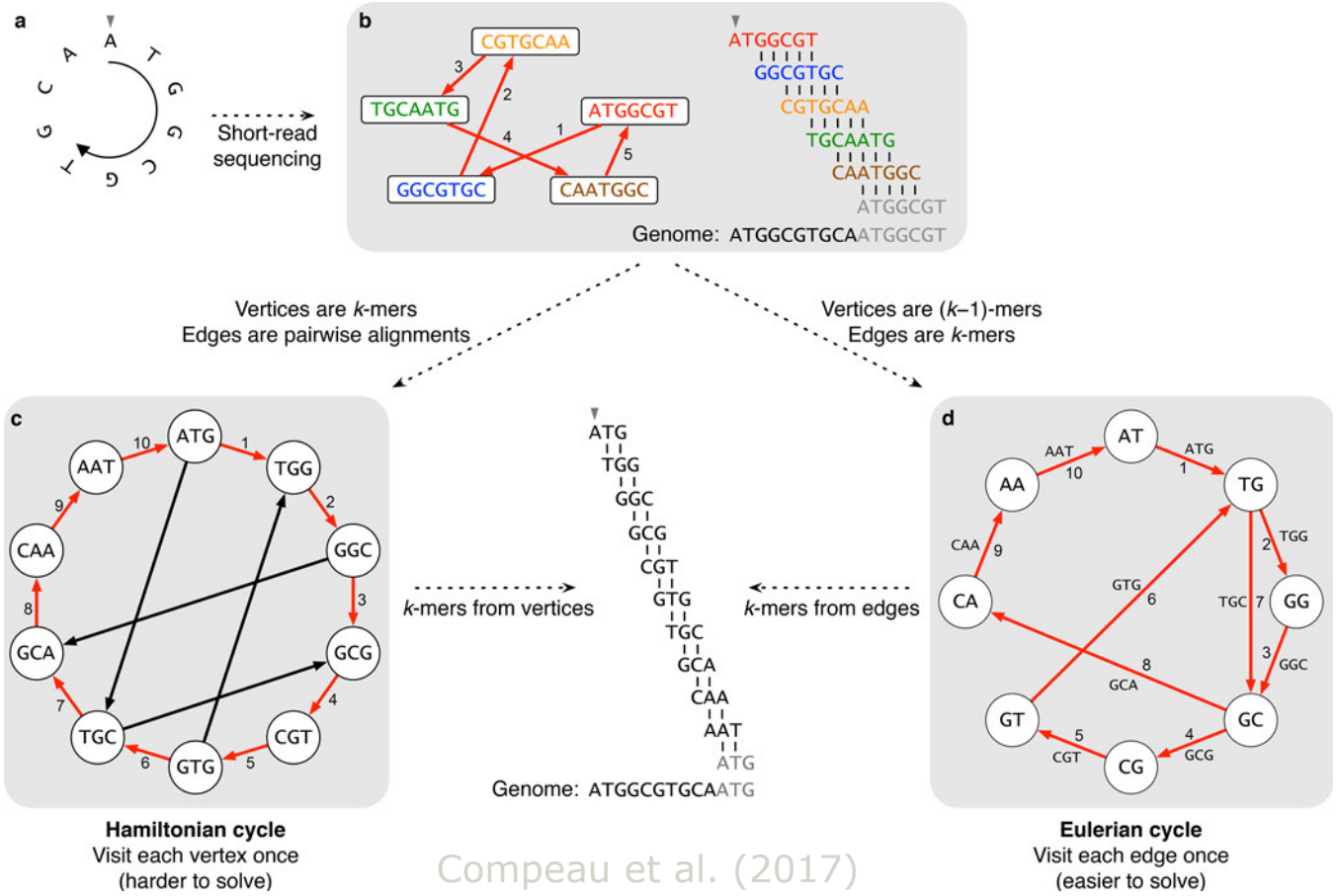
Nature Communications **7**, Article number: 13390 (2016) | Cite this article

Iterative approach: pros and cons

- Little storage required
- Can be processed with conventional reference-based tools
- Easy to annotate
- Great for gene content and PAV

- Reference bias
- Processing order matters
- Filtering of novel seqs is crucial
- Does not keep allelic variants
- Ignores intragenomic variation

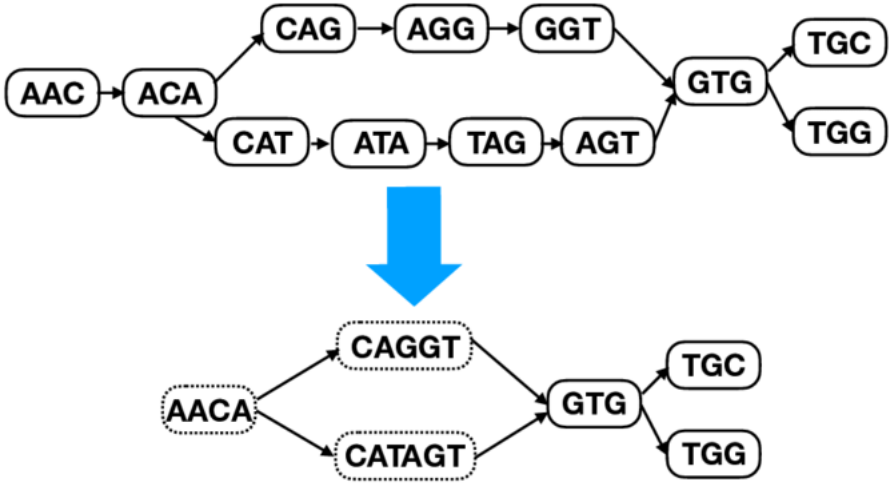
De Bruijn graph (known from assembly)



Compeau et al. (2017)

Compacted De Bruijn graph

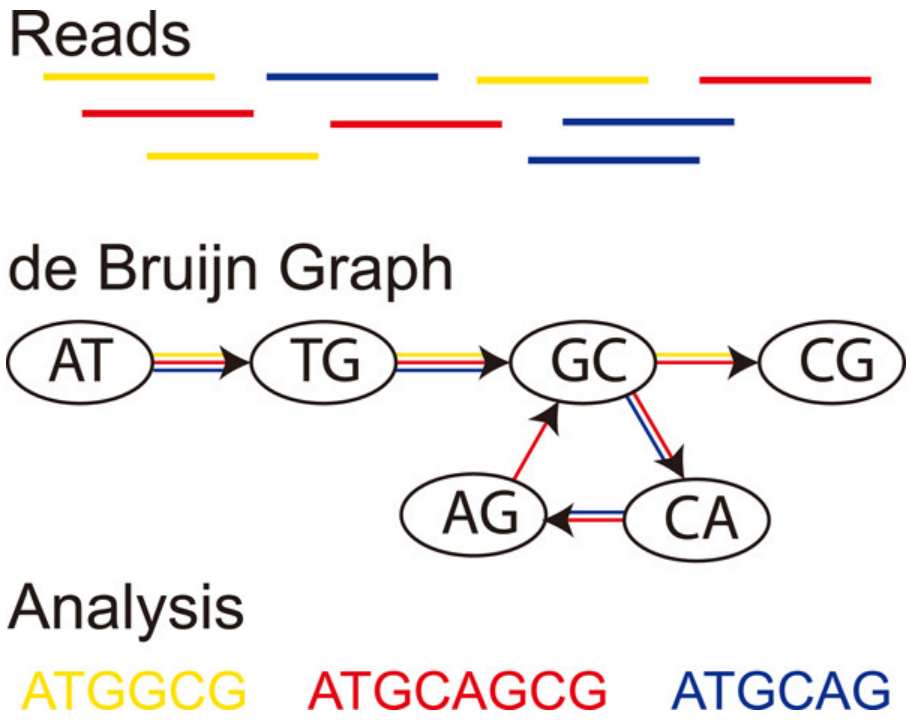
de Bruijn graph for the two sequences AACAGGTGC and AACATAGTGG



corresponding compacted de Bruijn graph

Menegaux & Vert (2020)

Colored (compacted) De Bruijn Graph



Dufault-Thompson & Jiang (2022)

De Bruijn Graph: tools

Bioinformatics. 2016 Jun 15;32(12):i201-i208. doi: 10.1093/bioinformatics/btw279.

Compacting de Bruijn graphs from sequencing data quickly and in low memory

bcalm 2

Rayan Chikhi [1], Antoine Limasset [2], Paul Medvedev [3]

Affiliations + expand

PMID: 27307618 PMCID: PMC4908363 DOI: 10.1093/bioinformatics/btw279

Bioinformatics. 2016 Sep 1;32(17):i487-i493. doi: 10.1093/bioinformatics/btw455.

PanTools: representation, storage and exploration of pan-genomic data

Siavash Sheikhizadeh [1], M Eric Schranz [2], Mehmet Akdel [1], Dick de Ridder [1], Sandra Smit [1]

Affiliations + expand

PMID: 27587666 DOI: 10.1093/bioinformatics/btw455

Software | Open access | Published: 17 September 2020

Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs

Guillaume Holley ✉ & Páll Melsted

Genome Biology **21**, Article number: 249 (2020) | Cite this article

Method | Open access | Published: 08 September 2022

Scalable, ultra-fast, and low-memory construction of compacted de Bruijn graphs with Cuttlefish 2

Jamshed Khan, Marek Kokot ✉, Sebastian Deorowicz & Rob Patro ✉

Genome Biology **23**, Article number: 190 (2022) | Cite this article

Extremely fast construction and querying of compacted and colored de Bruijn graphs with GGCAT

Andrea Cracco[1] **and Alexandru I. Tomescu**[2]

+ Author Affiliations

- Corresponding authors: alexandru.tomescu@helsinki.fi, andrea.cracco@univr.it

Published in Advance May 30, 2023, doi: 10.1101/gr.277615.122

Genome Res. 2023. 33: 1198-1207

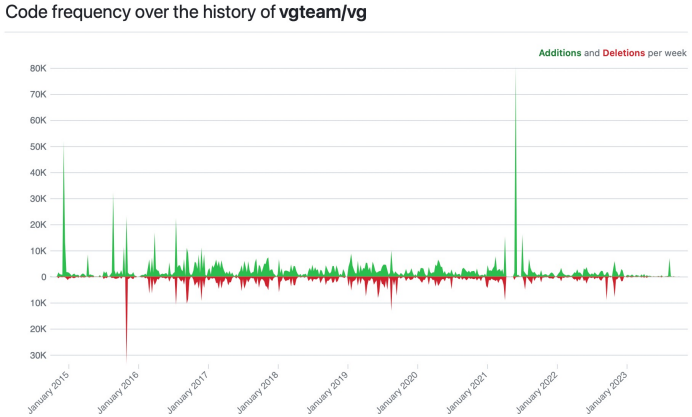
WAGENINGEN
UNIVERSITY & RESEARCH

De Bruijn Graph: pros and cons

- Strong compression
- Many applications
- Not alignment-based

- Lose sequence information

Variation graphs



Produces a directed acyclic graph (DAG), ordered along the reference genome, with bubbles at the sites of variation

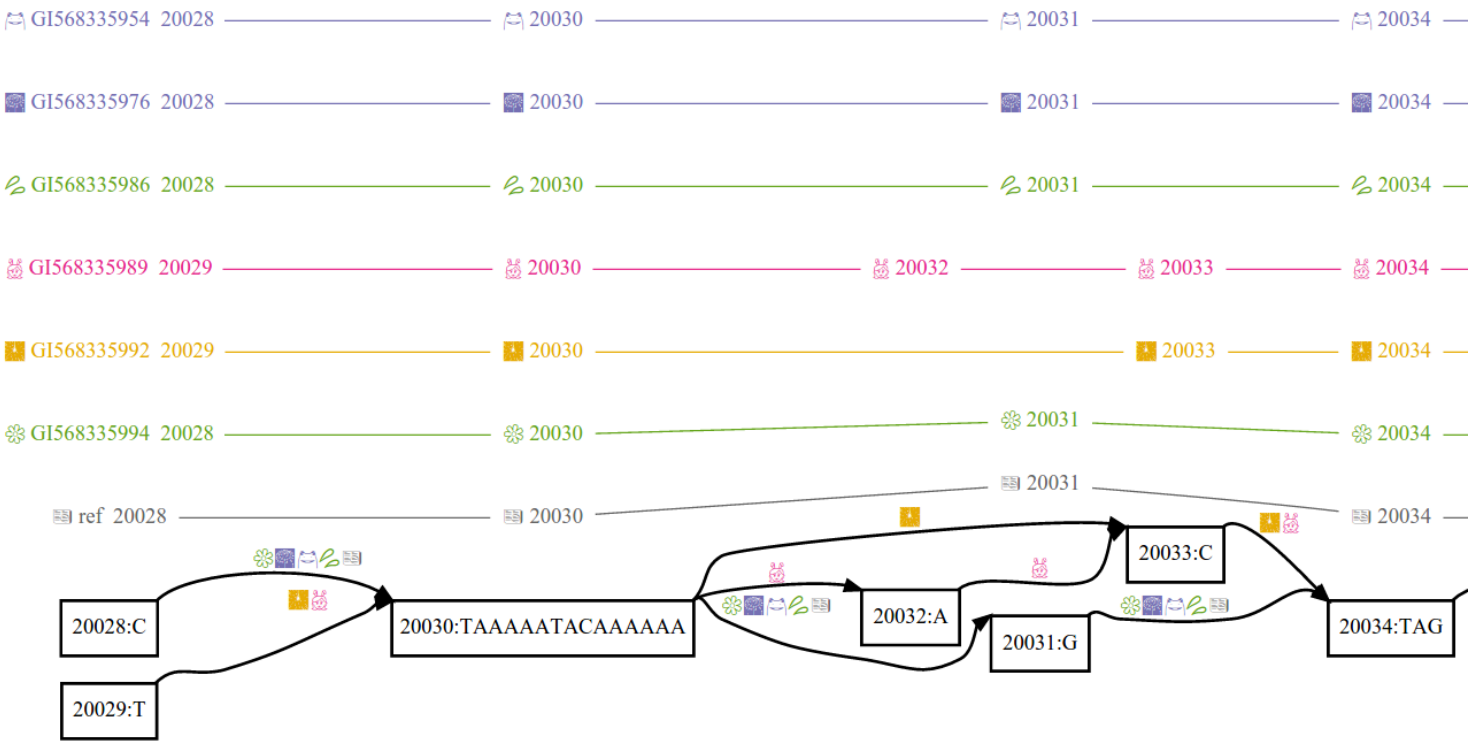
Variation graphs provide a succinct encoding of the sequences of many genomes. A variation graph (in particular as implemented in vg) is composed of:

- *nodes*, which are labeled by sequences and ids
- *edges*, which connect two nodes via either of their respective ends
- *paths*, describe genomes, sequence alignments, and annotations (such as gene models and transcripts) as walks through nodes connected by edges

This model is similar to sequence graphs that have been used in assembly and multiple sequence alignment.

https://github.com/vgteam/vg

Variation graphs



https://github.com/vgteam/vg

Variation graphs: algorithm description

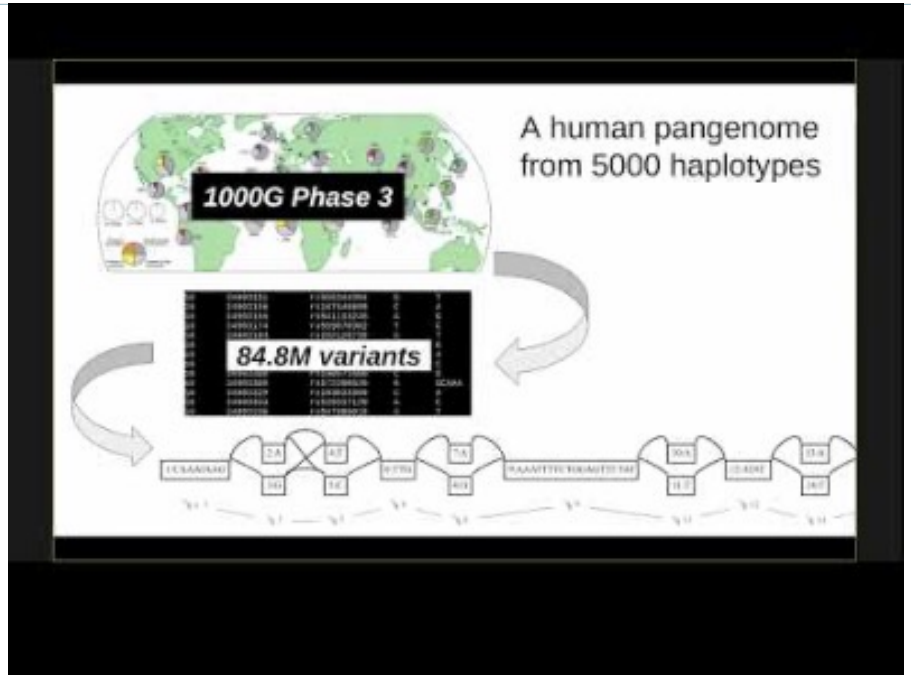
Published: 01 October 2018

Variation graph toolkit improves read mapping by representing genetic variation in the reference

Erik Garrison ✉, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten & Richard Durbin ✉

Nature Biotechnology **36**, 875–879 (2018) | Cite this article

Building and understanding pangenome variation graphs - Erik Garrison



https://www.youtube.com/watch?v=S8oSL4_Bqqw

Minigraph

- Motivation: there are no satisfactory solutions to the construction of reference pangenome graphs.

Method | Open access | Published: 16 October 2020

The design and construction of reference pangenome graphs with minigraph

Heng Li ✉, Xiaowen Feng & Chong Chu

Genome Biology **21**, Article number: 265 (2020) | Cite this article

- Our implementation, minigraph (https://github.com/lh3/minigraph), can construct a pangenome graph from twenty human assemblies in 3 h.

PanGenome Graph Builder (PGGB)

Building pangenome graphs

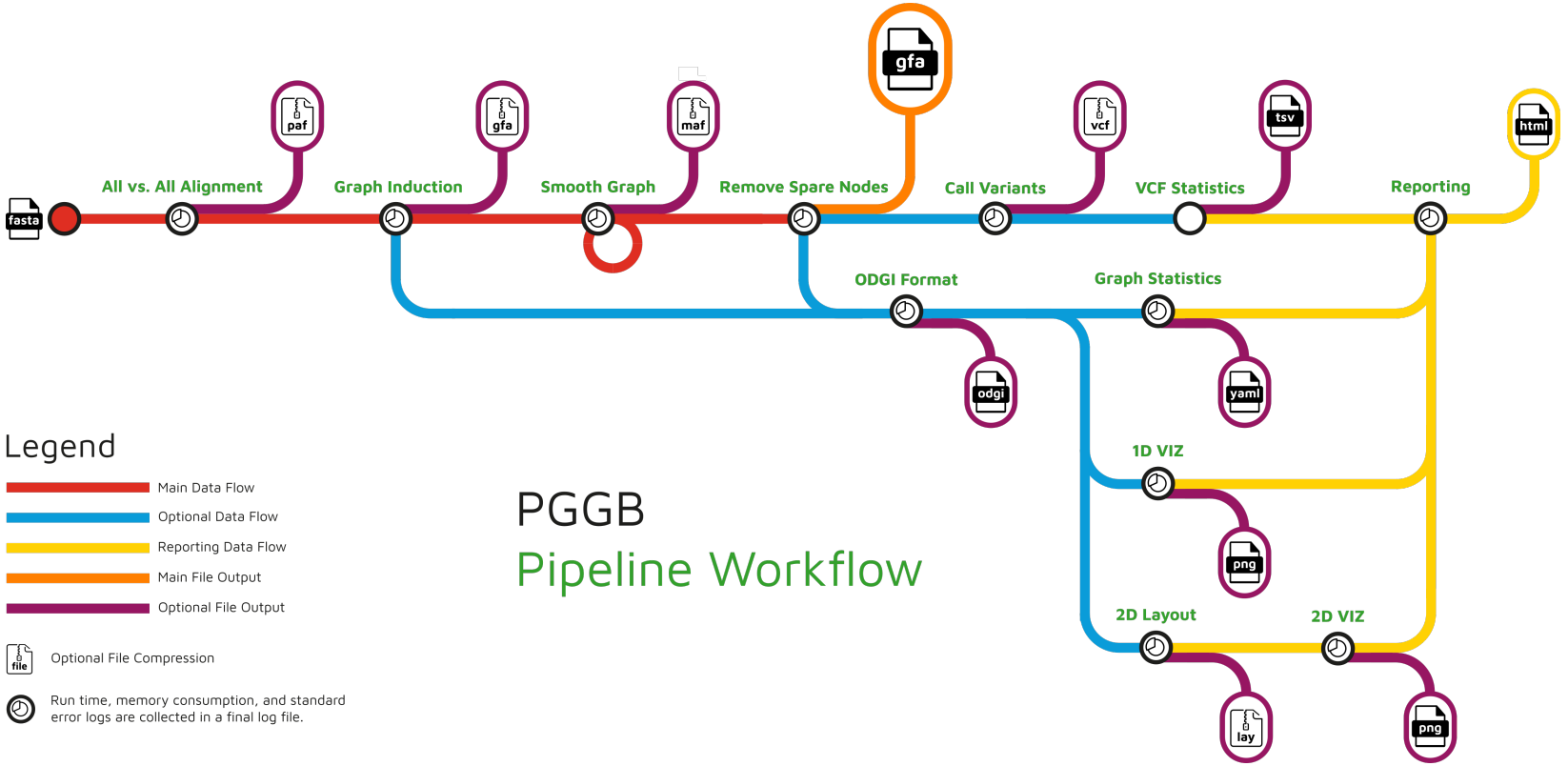
BioRxiv, April 2023

Erik Garrison, Andrea Guarracino, Simon Heumos, Flavia Villani, Zhigui Bao, Lorenzo Tattini, Jörg Hagmann, Sebastian Vorbrugg, Santiago Marco-Sola, Christian Kubica, David G. Ashbrook, Kaisa Thorell, Rachel L. Rusholme-Pilcher, Gianni Liti, Emilio Rudbeck, Sven Nahnsen, Zuyu Yang, Mwaniki N. Moses, Franklin L. Nobrega, Yi Wu, Hao Chen, Joep de Ligt, Peter H. Sudmant, Nicole Soranzo, Vincenza Colonna, Robert W. Williams, Pjotr Prins

doi: https://doi.org/10.1101/2023.04.05.535718

- Motivation: existing methods for constructing them are biased due to reference-guided approaches
- PGGB uses all-to-all whole-genome alignments and learned graph embeddings to build and iteratively refine a model in which we can identify variation, measure conservation, detect recombination events, and infer phylogenetic relationships

PanGenome Graph Builder (PGGB)



https://pggb.readthedocs.io

WAGENINGEN
UNIVERSITY & RESEARCH

Minigraph-cactus

Article | Published: 10 May 2023

Pangenome graph construction from genome alignments with Minigraph-Cactus

Glenn Hickey, Jean Monlong, Jana Ebler, Adam M. Novak, Jordan M. Eizenga, Yan Gao, Human Pangenome Reference Consortium, Tobias Marschall, Heng Li & Benedict Paten

Nature Biotechnology (2023) | Cite this article

Here we present the Minigraph-Cactus pangenome pipeline, which creates pangenomes directly from whole-genome alignments, and demonstrate its ability to scale to 90 human haplotypes from the Human Pangenome Reference Consortium.

WAGENINGEN
UNIVERSITY & RESEARCH

Research Briefing | Published: 22 May 2023

Combining reference genomes into a pangenome graph improves accuracy and reduces bias

Nature Biotechnology (2023) | Cite this article

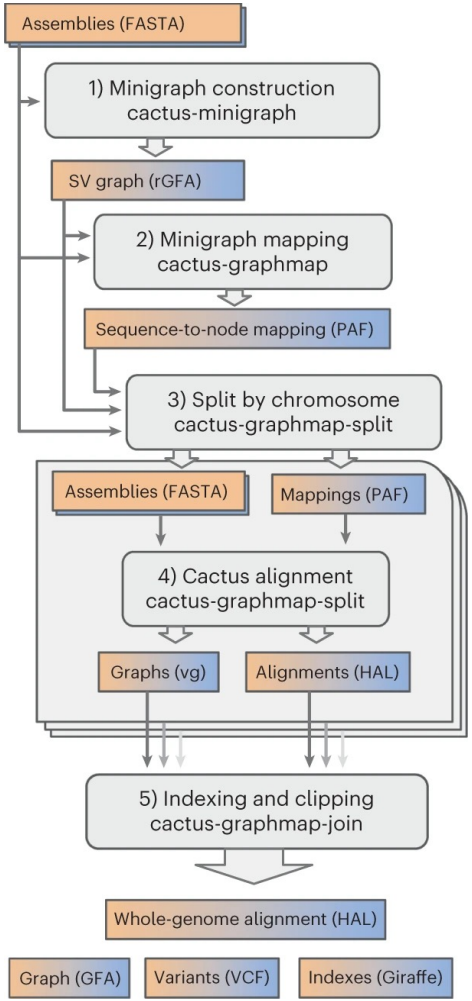
Minigraph-Cactus, a method to efficiently combine multiple reference genome assemblies into a pangenome reference graph, can be used to improve accuracy of read mapping and variant calling compared with a single reference in downstream applications.

Minigraph-cactus

This method proceeds by iteratively adding structural variants from each sample, then using these variants as anchors for computing a base-level realignment (Fig. 1).

The resulting graph consistently represents variation at different scales, from single-nucleotide polymorphisms to complex, nested intrachromosomal rearrangements.

Crucially, existing pangenomics software can be used with this method, as well as for read mapping and variant calling.

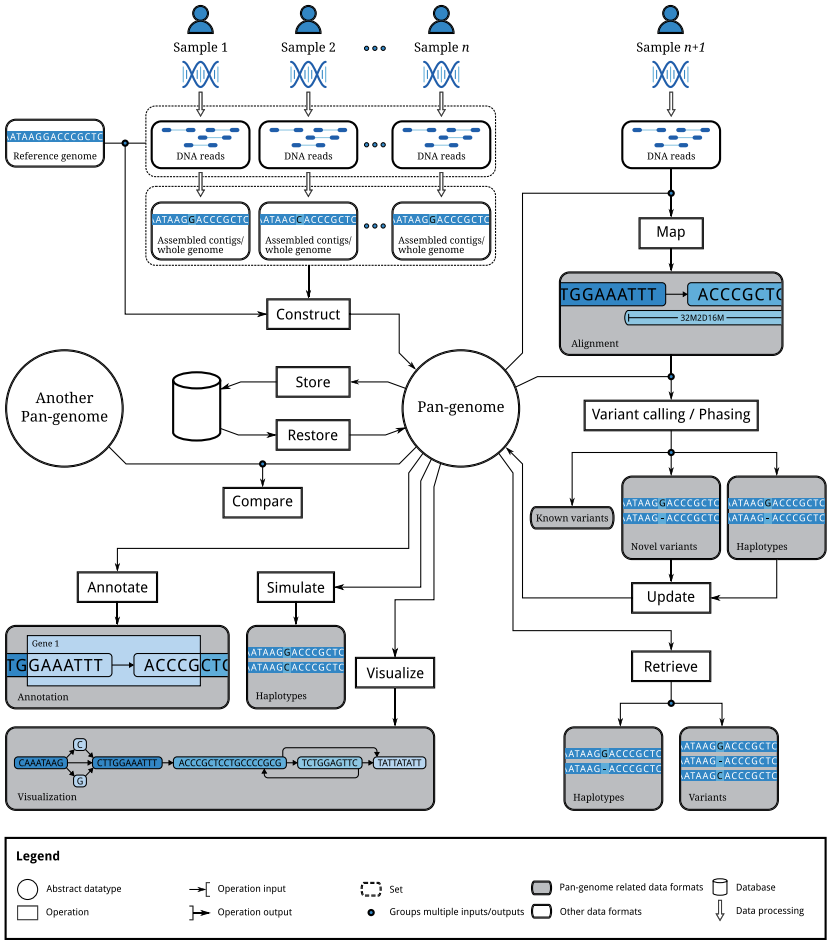


Not discussed

- Annotation
- Indexing

What is a pangenome?

- Variation-aware reference
  - Construct
  - Annotate
  - Map
  - Variant calling
  - Visualize
  - …



Marschall *et al.*, *2016*

Read mapping: tools

Table 1 Overview of some state-of-the-art sequence-to-graph aligners, categorized according to the type of graph they handle.

acyclic variation graph	general variation graph	assembly de Bruijn graph	de Bruijn graph as a reference
GenomeMapper	vg	BGREAT	deBGA
Graph Genome Aligner	GraphAligner	BrownieAligner	PuffAligner
HISAT2	Minigraph	GraphAligner	Beller & Ohlebusch
PaSGAL	Giraffe	SPAligner	**Nexus (this paper)**
V-MAP			
Vargas			

From: Depuydt et al. (2023).
https://assets.researchsquare.com/files/rs-2583159/v1/61d0b7c9b76925e42ee15cdd.pdf?c=1677080263

Read mapping: tools

Genome Biol. 2020; 21: 253.
Published online 2020 Sep 24. doi: 10.1186/s13059-020-02157-2

GraphAligner: rapid and versatile sequence-to-graph alignment

Mikko Rautiainen[1,2,3] and Tobias Marschall[4]

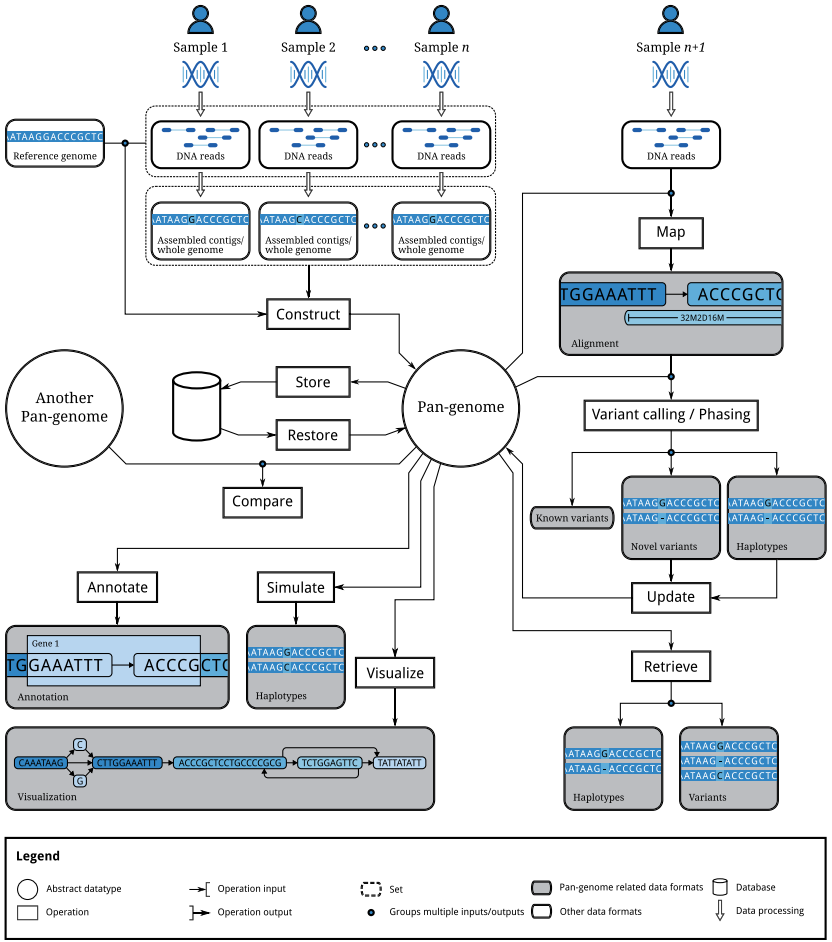
Mapping short reads with Giraffe

Jouni Siren edited this page on May 28 · 10 revisions

This tutorial will explain how to use `vg giraffe` to map short reads to a pangenome graph. See also Giraffe best practices.

What is a pangenome?

- Variation-aware reference
  - Construct
  - Annotate
  - Map
  - Variant calling
  - Visualize
  - …



Marschall *et al.*, *2016*

https://pangenome.github.io

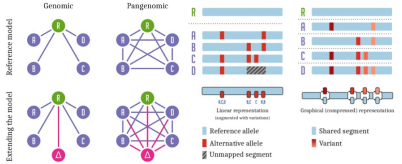
pangenome.github.io

Practical Graphical Pangenomics

tools and workflows based on genome variation graphs

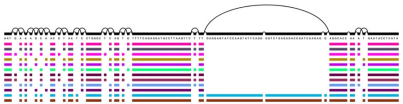
Pangenomic methods

Standard approaches to genome inference and analysis relate sequences to a single linear reference genome. This is efficient but has a fundamental problem: Differences from this reference are hard to observe and describe in a coherent way. Variation and sequence are separated.



Pangenomic methods allow us to relate all genomes or sequences in our analysis directly to each other. Sequence and variation are combined into a coherent data structure. This practice is still new, and research into ways to design, implement, and apply this model is ongoing. However, there is a growing consensus around best practices. Many methods work on an augmented sequence graph model and use a handful of common data formats for input and output.

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:



vg

The variation graph toolkit **vg** provides computational methods for creating and manipulating of genome variation graphs. It's pangenome representation of a set of genomes overcomes reference bias and improves read mapping. This is highlighted in the Nature Biotechnology publication. Users can receive support on **vg**'s Biostars page.

PanGenome Graph Evaluator (pgge)

This pangenome graph evaluation pipeline measures the reconstruction accuracy of a pangenome graph (in the variation graph model). Its goal is to give guidance in finding the best pangenome graph construction tool for a given input data and task.

xg

The succinct graph index **xg** presents a static index of nodes, edges and paths of a variation graph. **xg** can be used to annotate graph nodes with their reference path relative positions. It was a key component of early development in **vg**, and was use to scale short read mapping to large genomes. It implements the libhandlegraph API.

GWBT

PanGenome Graph Builder (pggb)

This pangenome graph construction pipeline renders a collection of sequences into a pangenome graph (in the variation graph model). Its goal is to build a graph that is locally directed and acyclic while preserving large-scale variation. Maintaining local linearity is important for the interpretation, visualization, and reuse of pangenome variation graphs. A Nextflow version of the pipeline is also available nf-core/pangenome.

Pangenome Graph Variation Format (PGVF)

PGVF is a hard fork of the GFAv1 format that allows the description of graph-to-graph alignments. It represents a collection of aligned graphs as a network of walks through an underlying merged sequence graph. While pangenome graphs let us represent differences between genomes, we have no mechanism to represent differences between pangenome graphs, or to combine multiple pangenome graphs into one structure without losing information. This motivates the development of a new biological data format.

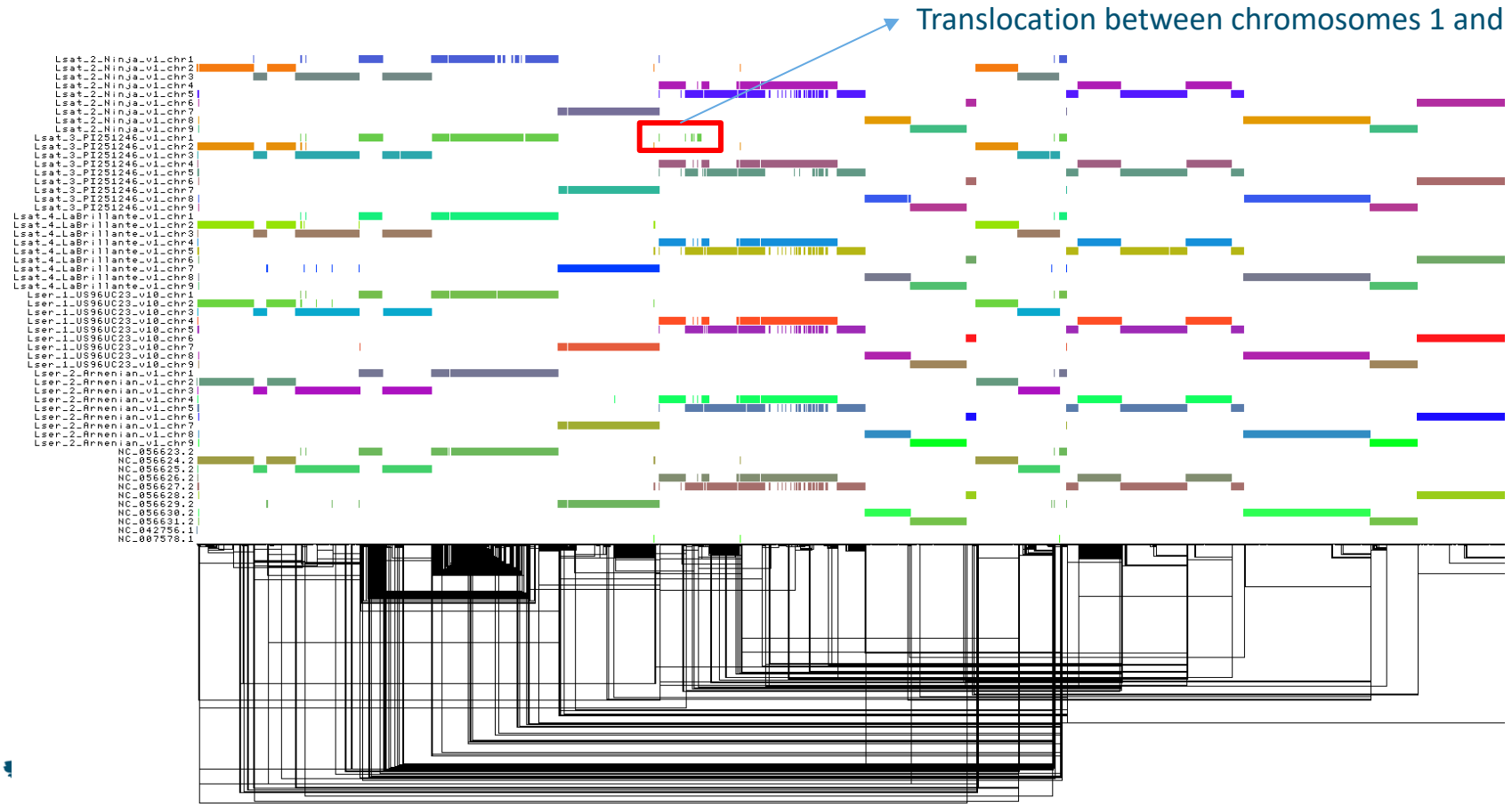
odgi

odgi, the Optimized Dynamic (genome) Graph Interface, links a thrifty dynamic in-memory variation graph data model to a set of algorithms designed for scalable sorting, pruning, transformation, and visualization of very large genome graphs. **odgi** includes python bindings that can be used to directly interface with its data model. The odgi manual provides detailed information about its features and subcommands, including examples.

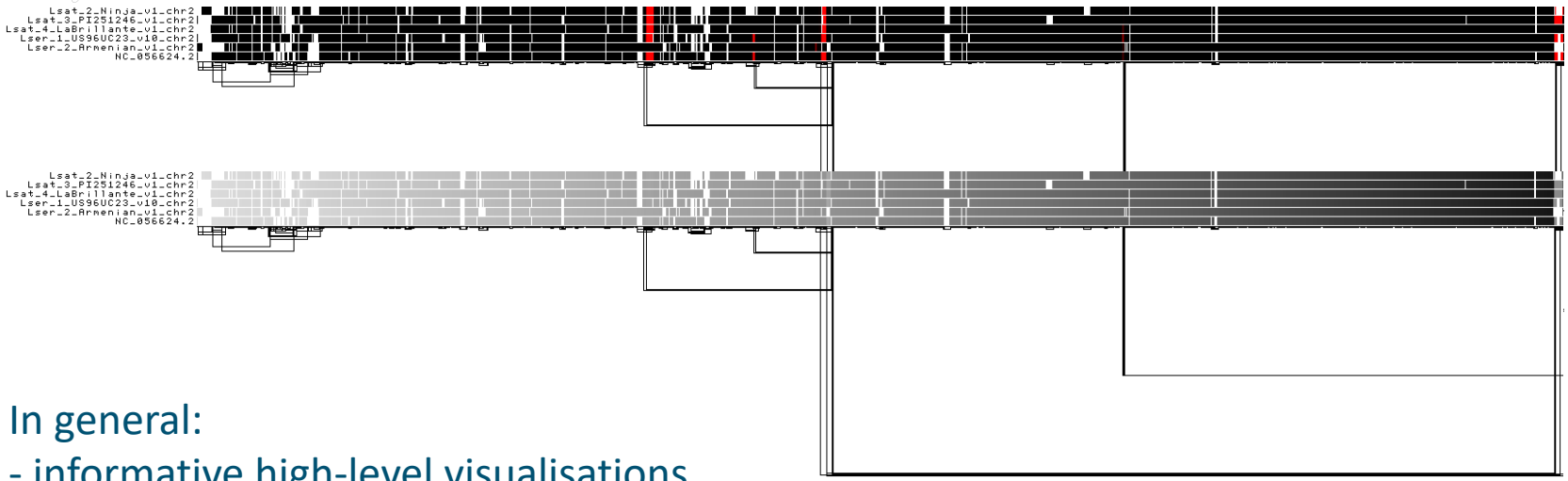
spodgi

WAGENINGEN
UNIVERSITY & RESEARCH

PGGB pangenome



PGGB pangenome chr2



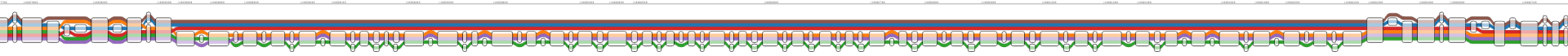
In general:

- informative high-level visualisations
- takes a long time to compute (±5 days for all chromosomes)
- alignment extremely dependent on parameters

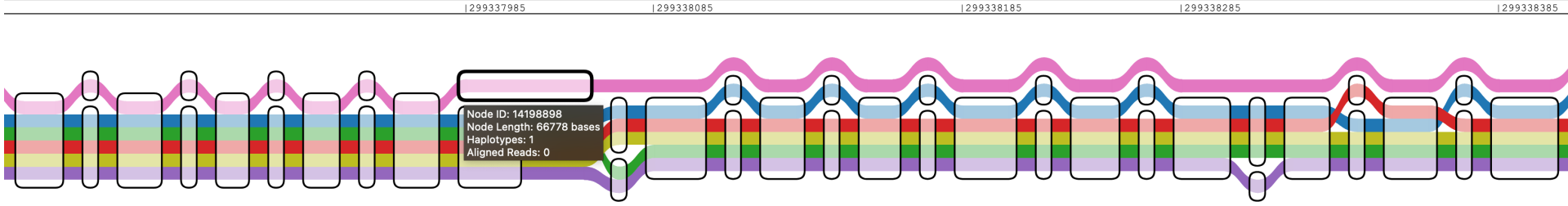
WAGENINGEN
UNIVERSITY & RESEARCH

MC pangenome

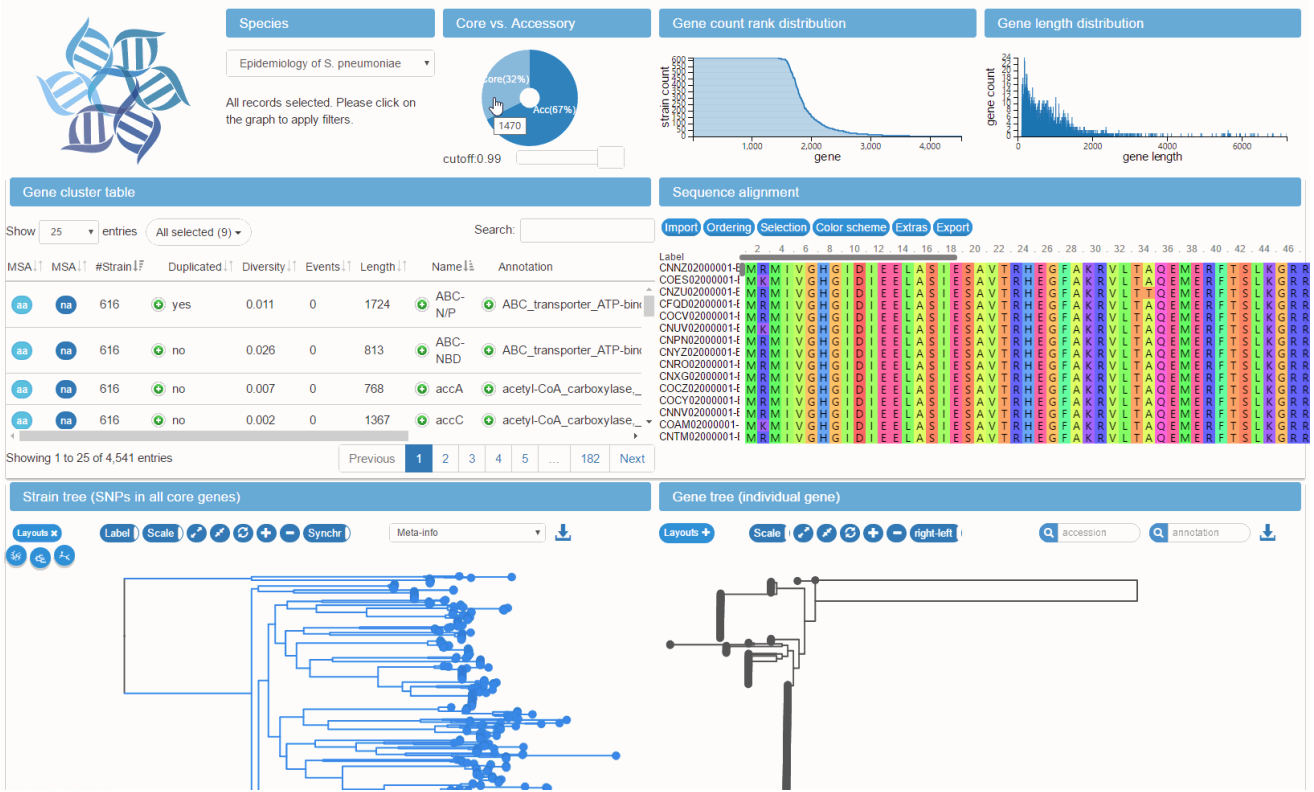
- "Minigraph-Cactus" pipeline (using "sequenceTubeMap" visualisation)
  - 4kb transposon insertion (*LsKN1*) clearly visible (insertion present in Salinas, Ninja, LaBrillante)



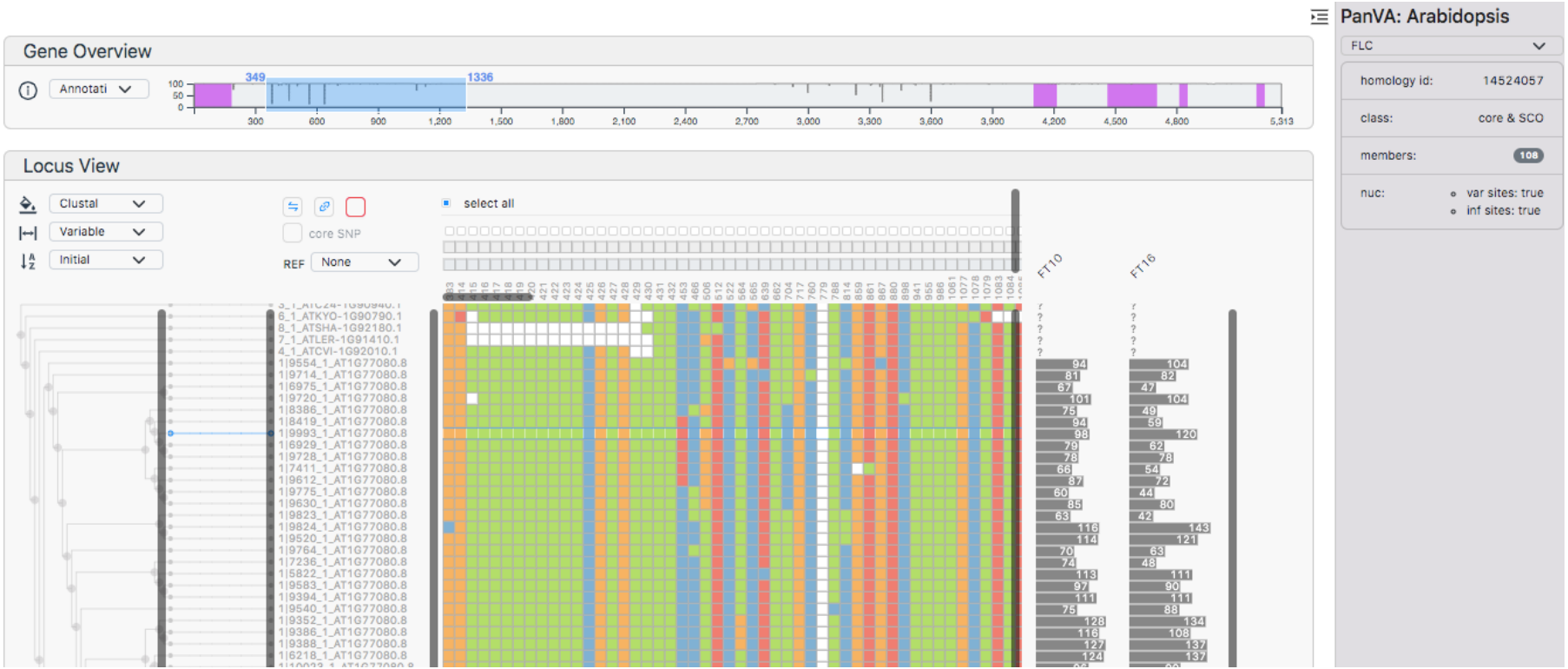
  - g2349.t1 is non-reference because of ±67kb part in Salinas which doesn't have it (transcript present in Ninja, PI251246, LaBrillante, US96UC23, Armenian 999)



Interactive visualization: PanX



Interactive visualization: PanVA



Conclusions on pangenomics

- Fast-developing field
    - There are still many challenges
- Graph representations are dominant
- Dominant tools: vg, pggb, minigraph-cactus
- Downstream applications are crucial to do comparative genomics
- Visualization is an important aspect of pangenome analysis

Program today

- 9.00 Lecture
- 11.30 Break
- 11.00 Computer exercises (rosalind or project)
- 12.30 Lunch
- 13.30 Guest lecture by Erwin Datema from Keygene
- 14.15 Lecture about PanTools
- 15.00 Break
- 15.30 Computer exercises (rosalind or project or PanTools)
- 17.30 Closing

Exercises

- Rosalind
    - Graphs (pangenomics): Complete questions 35 and higher
    - Supporting slides on graph algorithms available
- Project
    - Minimap & miniasm
- PanTools (afternoon)
    - Tutorial on https://pantools.readthedocs.io