nf-core/pangenome
Pangenome growth

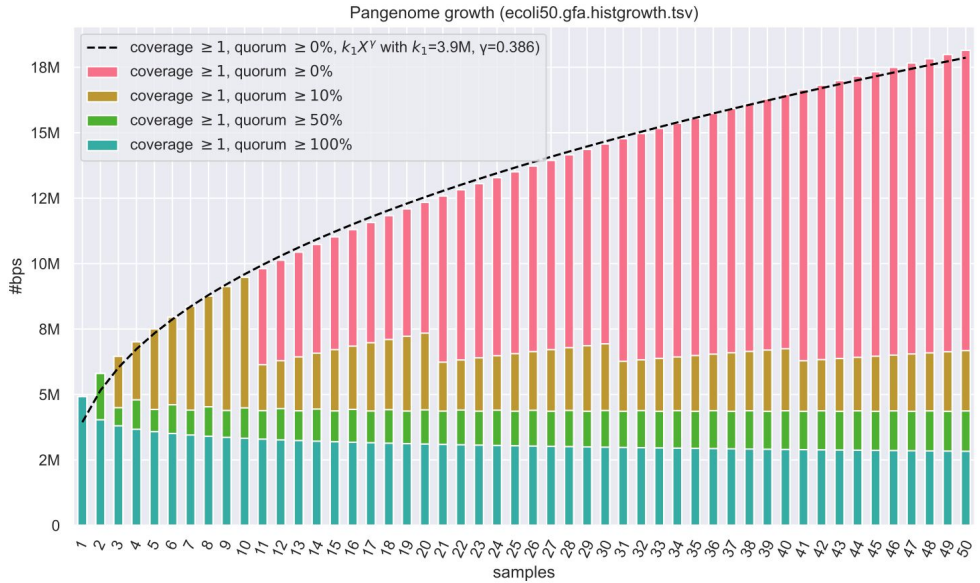
#MEMPANG24

Simon Heumos
PhD student at Quantitative Biology Center (QBiC), University of Tübingen, Germany

@SimonHeumos

University of Tennessee Health Science Center
930 Madison Ave, Freeman Auditorium at the Hamilton Eye Institute
Day 2b - 2024/05/19

nf-core/pangenome




nf-core/
pangenome

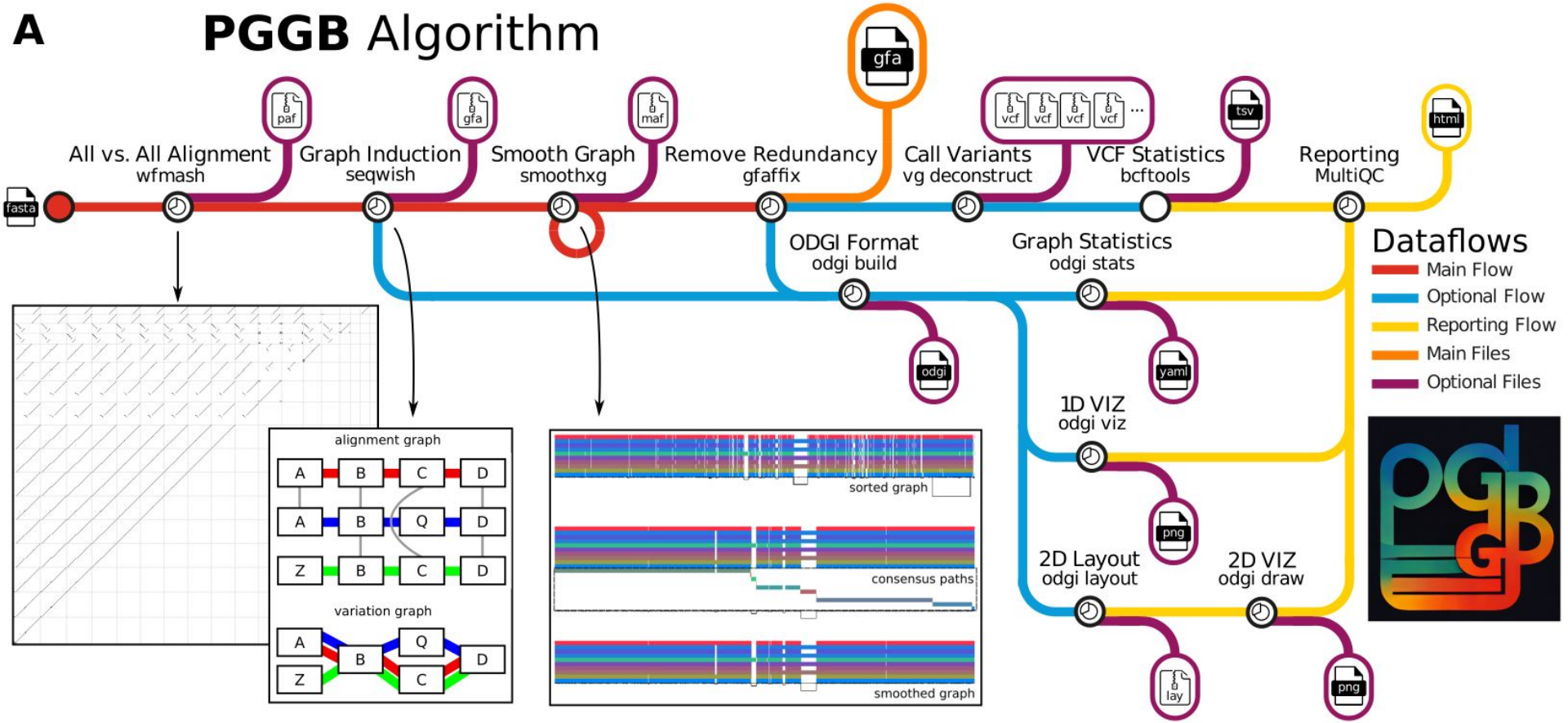
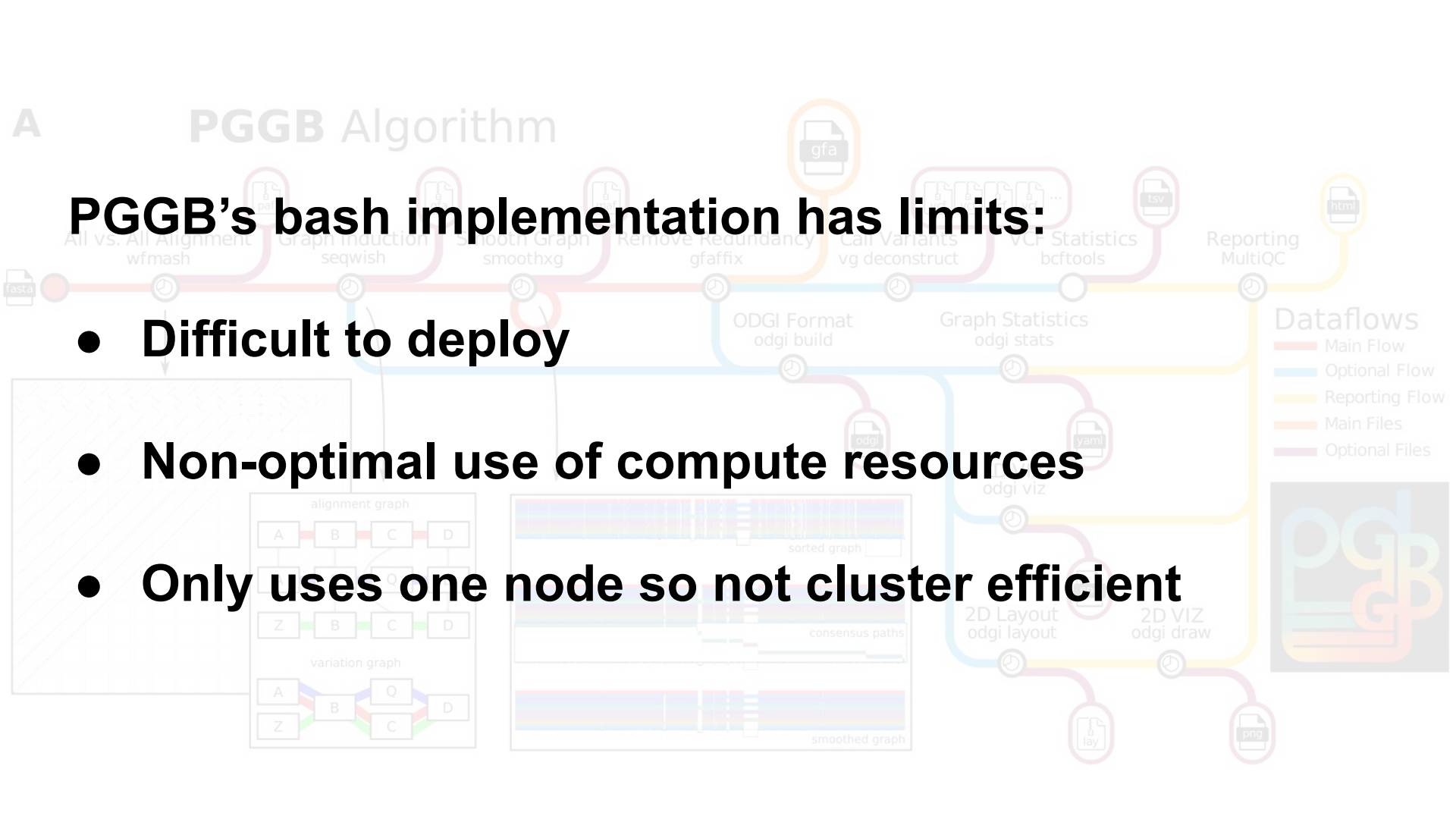


Figure from: [Garrison, Guarracino et al., 2023]



PGGB's bash implementation has limits:

- Difficult to deploy
- Non-optimal use of compute resources
- Only uses one node so not cluster efficient

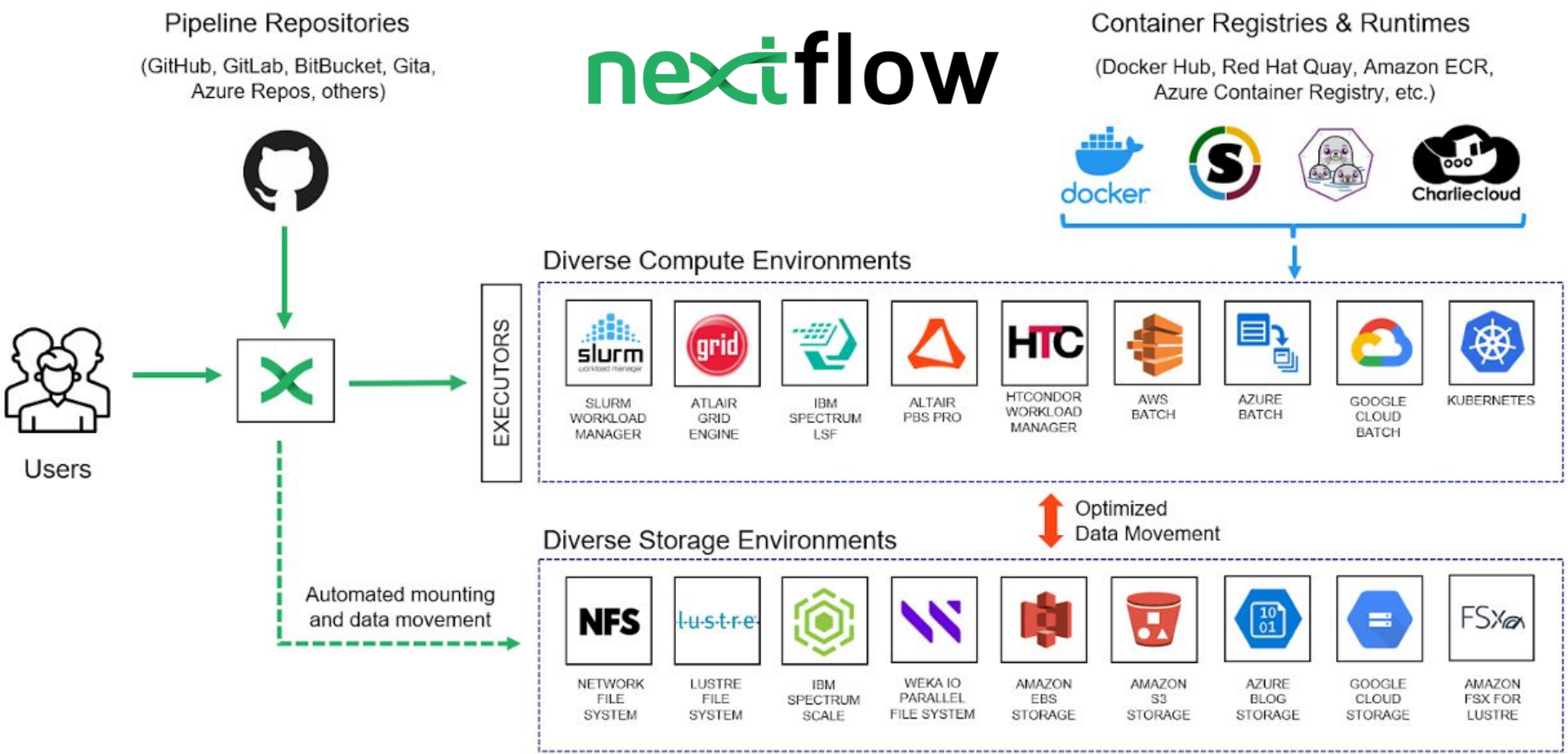


Figure from Di Tomasso P. et al., 2017.

nf-core

A community effort to collect a curated set of Nextflow pipelines

Stable releases

Packaged software

Documentation

Portable

Continuous integration

Cloud ready

Slide modified with permission by Gisela Gabernet

Figure from Ewels et al., 2020.

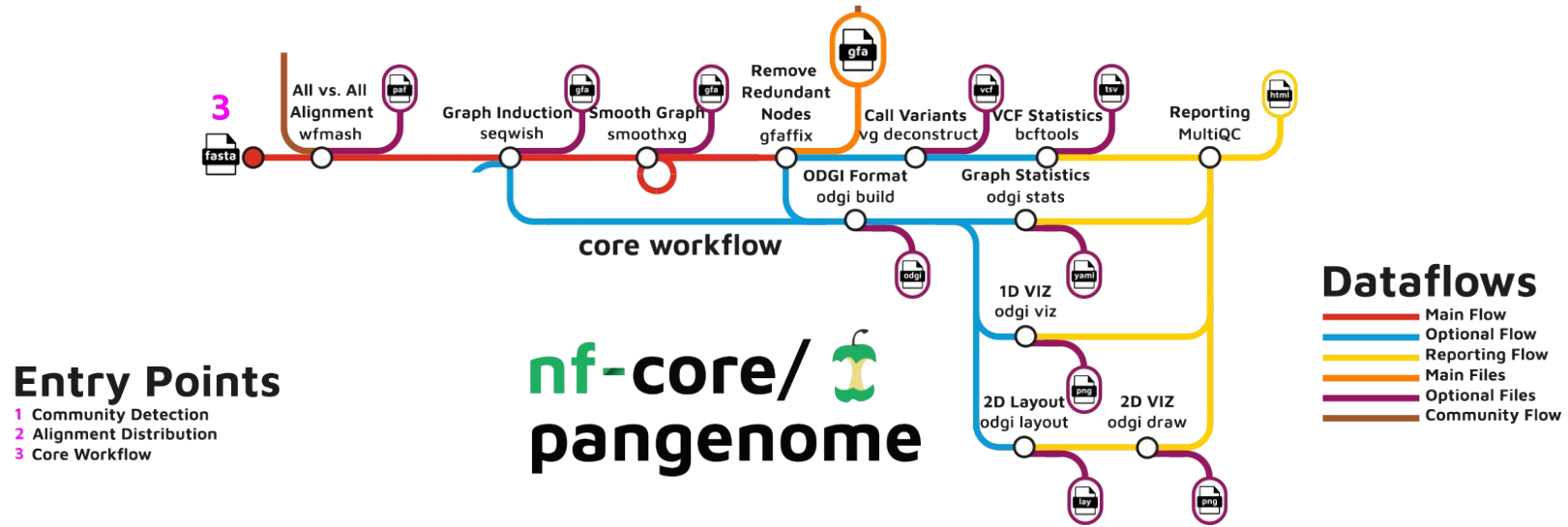


Figure from: [Heumos et al., 2024](#).

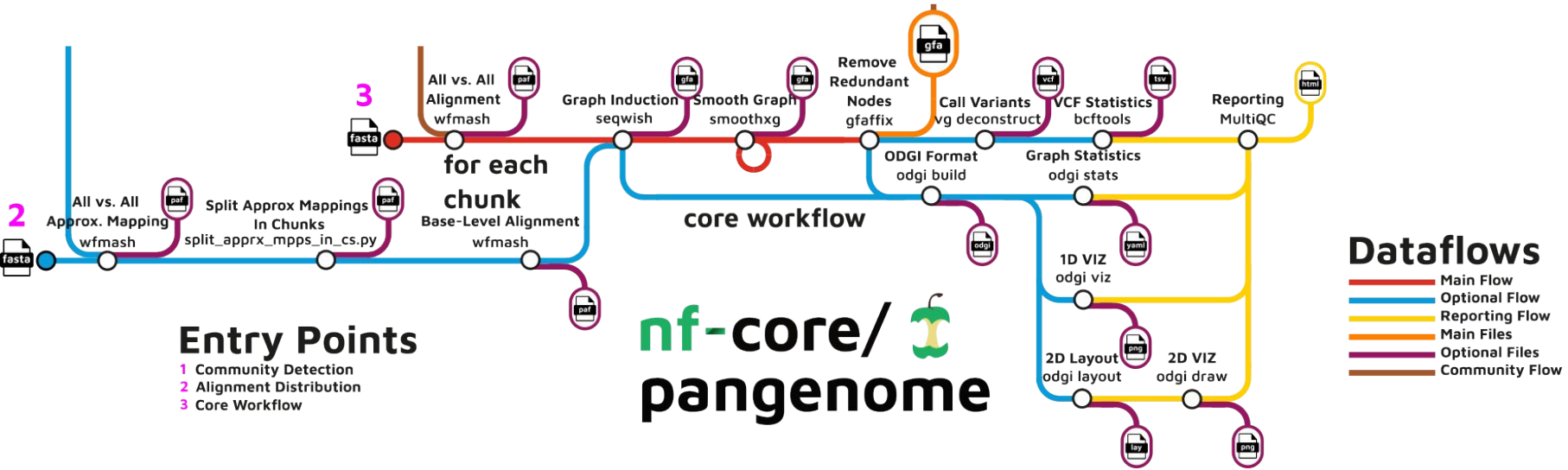
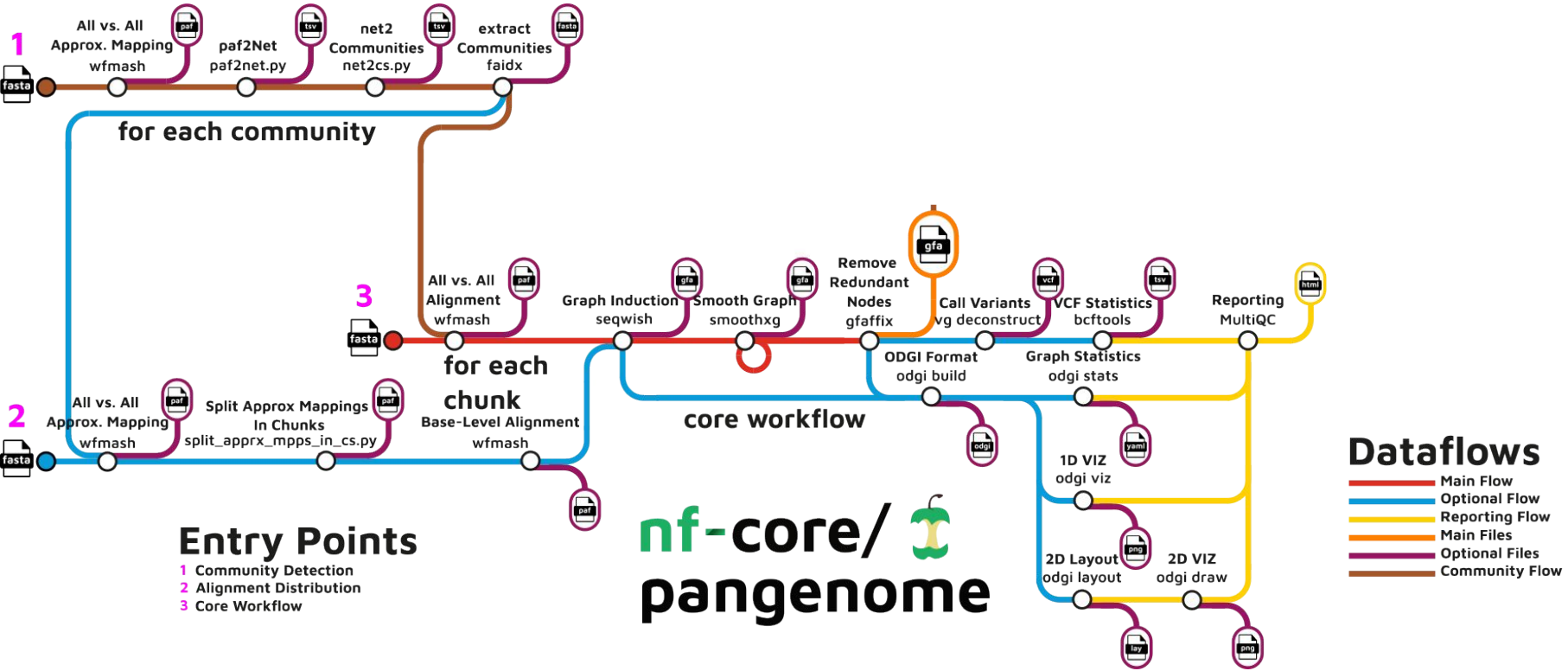


Figure from: [Heumos et al., 2024](#).



Clustering with the Leiden algorithm: Edge weight is `mapped_length * mapped_identity`

Leiden was applied in Guarracino et al., 2023.

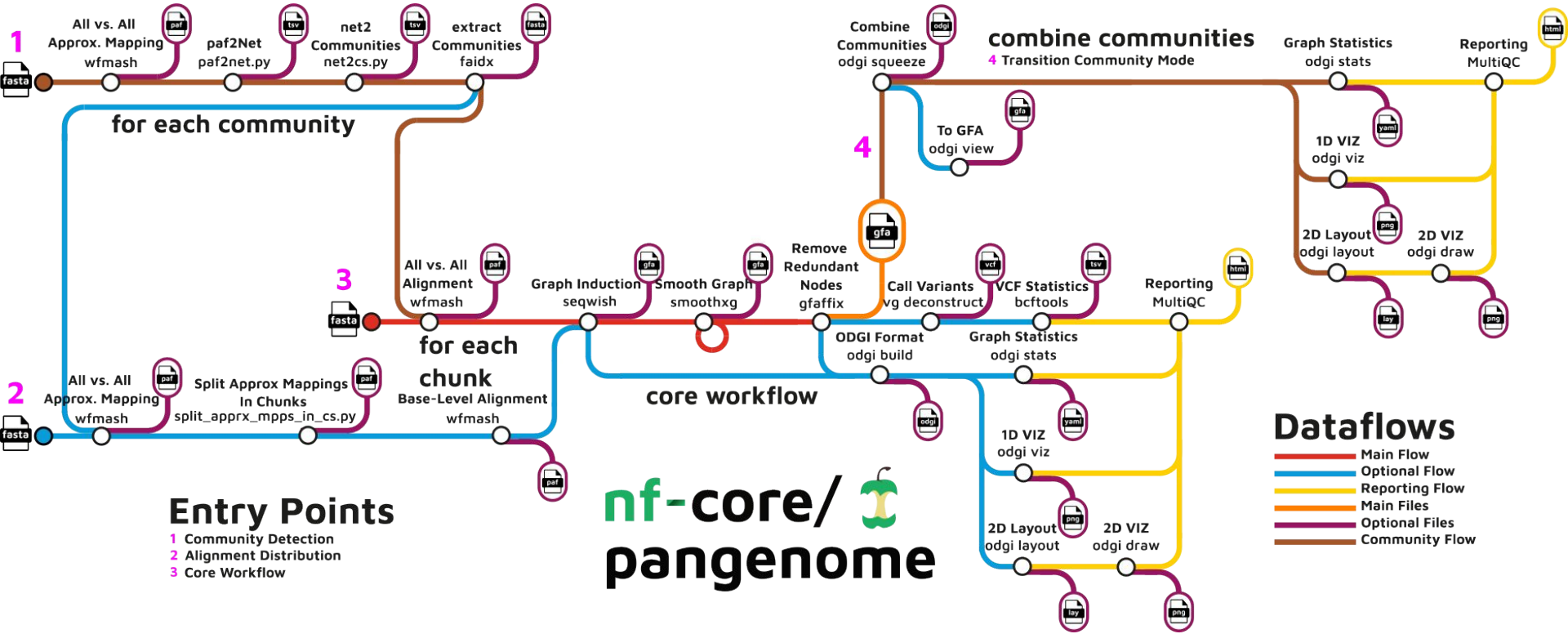


Figure from: [Heumos et al., 2024](#).

Building a cluster scalable version of PGGB

- Approximate mappings are split into chunks
- Processing of chunks is parallelized

⇒ Alignment can be distributed to cluster nodes

⇒ Time spent on base-pair level alignments is reduced linearly with increase in chunks

Time scales linearly with number of chunks

Base-pair level alignment of 1024 E.coli genomes

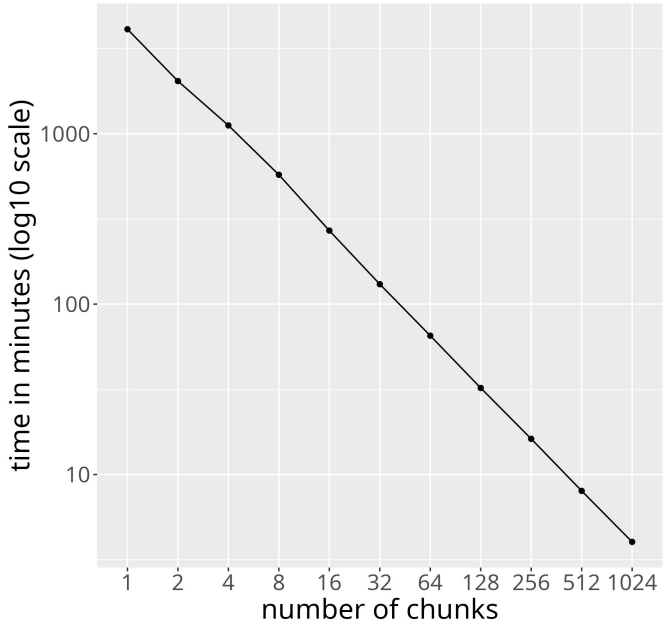


Figure from: Heumos et al., 2024.

Building a cluster scalable version of PGGB

- Approximate mappings are split into chunks
- Processing of chunks is parallelized

⇒ CO2 equivalent (CO2e) emissions do not rise with increasing number of chunks

CO2e exhaust is stable with number of chunks

Base-pair level alignment of 1024 E.coli genomes

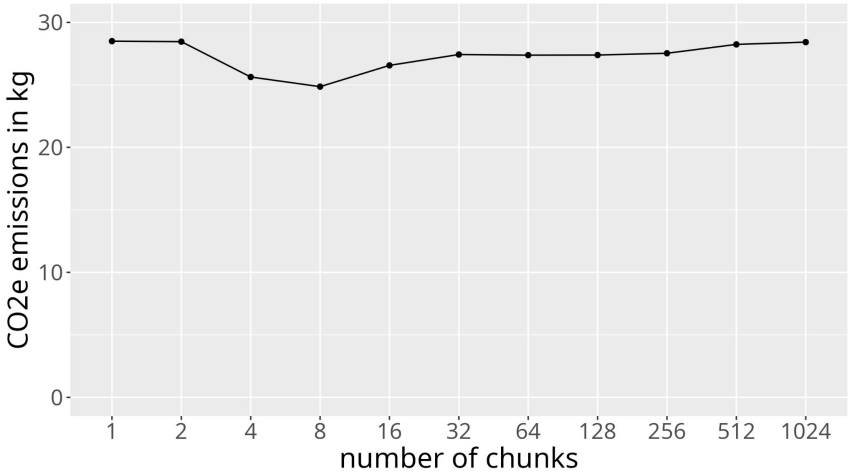


Figure from: Heumos et al., 2024.

Building a 1000 haplotypes 1KGP chr19 graph

nf-core/pangenome: 3 days

PGGB: 7 days

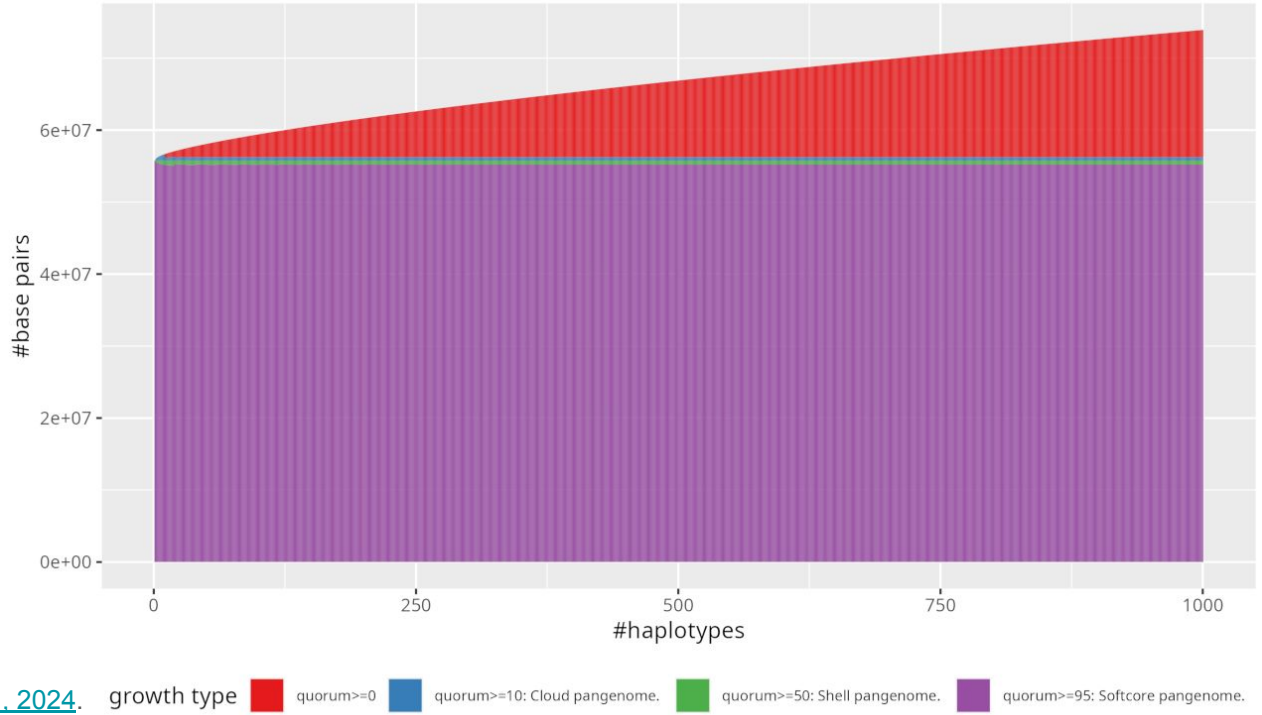


Figure from: [Heumos et al., 2024](#).

Building a 2146 sequences E. coli graph

nf-core/pangenome: 10 days

PGGB: didn't finish after 30 days

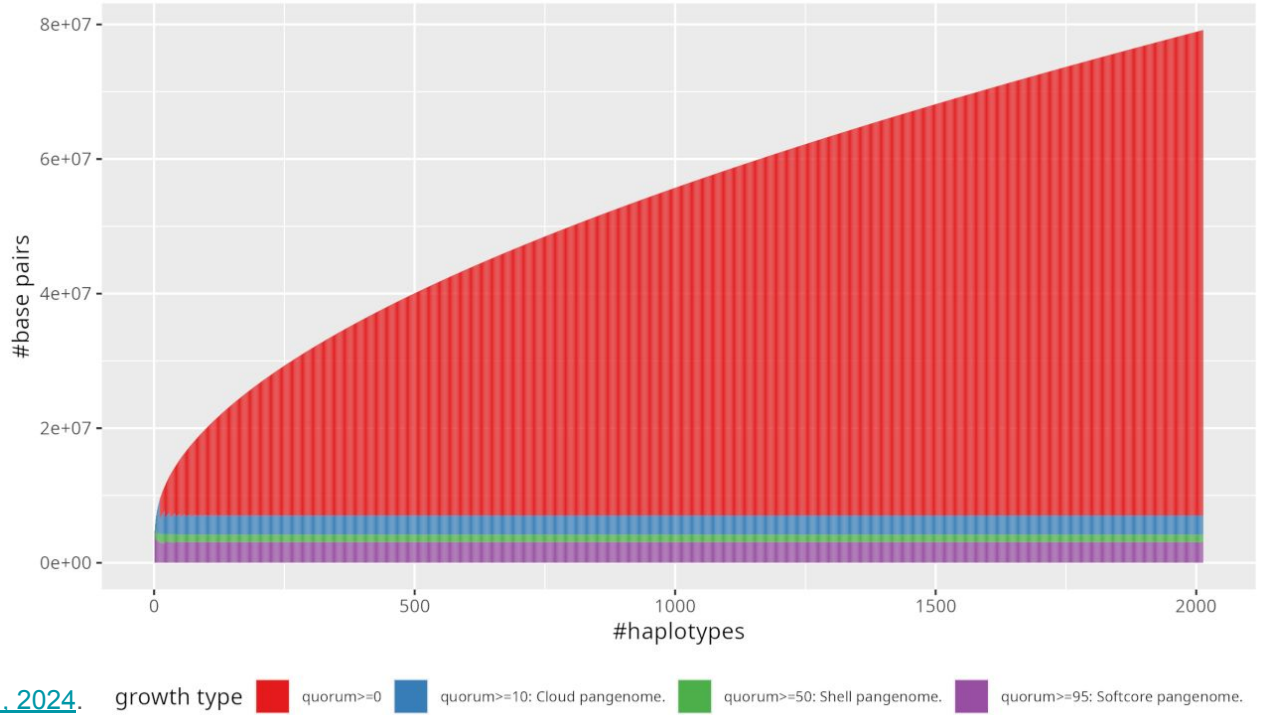


Figure from: Heumos et al., 2024.

nf-core/pangenome example run

```
nextflow run nf-core/pangenome -r a_brave_new_world -profile docker -config git/pangenome/conf/hla.config --input git/pggb/data/HLA/
DRB1-3123.fa.gz --wfmash_map_pct_id 70 --wfmash_segment_length 2000 --outdir DRB1-3123_pan --n_haplotypes 12
[ba/4d19e2] process > NFCORE_PANGENOME:PANGENOME:PGGB:WFMASH_MAP_ALIGN (DRB1-3123.fa.gz)                                [100%] 1 of 1 ✔
[f9/98e332] process > NFCORE_PANGENOME:PANGENOME:PGGB:SEQWISH (DRB1-3123.fa.gz)                                         [100%] 1 of 1 ✔
[34/01dc90] process > NFCORE_PANGENOME:PANGENOME:PGGB:SMOOTHXG (DRB1-3123.fa.gz)                                        [100%] 1 of 1 ✔
[81/ba4e25] process > NFCORE_PANGENOME:PANGENOME:PGGB:GFAFFIX (DRB1-3123.fa.gz)                                         [100%] 1 of 1 ✔
[88/081d82] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_BUILD (DRB1-3123.fa.gz.gfaffix)                              [100%] 2 of 2 ✔
[bd/b4a715] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_UNCHOP (DRB1-3123.fa.gz.gfaffix)                             [100%] 1 of 1 ✔
[6d/358413] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_SORT (DRB1-3123.fa.gz.gfaffix)                               [100%] 1 of 1 ✔
[ae/e3b97a] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_VIEW (DRB1-3123.fa.gz.gfaffix)                               [100%] 1 of 1 ✔
[9f/c04a92] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_STATS (DRB1-3123.fa.gz.gfaffix)                      [100%] 2 of 2 ✔
[3f/55484a] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_VIZ_COLOR (DRB1-3123.fa.gz.gfaffix.viz)              [100%] 1 of 1 ✔
[90/1aded7] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_VIZ_POS (DRB1-3123.fa.gz.gfaffix.viz_pos)           [100%] 1 of 1 ✔
[bf/1e22c8] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_VIZ_DEPTH (DRB1-3123.fa.gz.gfaffix.viz_depth)       [100%] 1 of 1 ✔
[2e/387271] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_VIZ_INV (DRB1-3123.fa.gz.gfaffix.viz_inv)          [100%] 1 of 1 ✔
[98/f89280] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_VIZ_COMPR (DRB1-3123.fa.gz.gfaffix.viz_O)          [100%] 1 of 1 ✔
[4e/9f113c] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_VIZ_UNCALLED (DRB1-3123.fa.gz.gfaffix.viz_uncalled) [100%] 1 of 1 ✔
[72/d583b1] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_LAYOUT (DRB1-3123.fa.gz.gfaffix)                    [100%] 1 of 1 ✔
[4b/ff48d1] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_DRAW_MULTIQC (DRB1-3123.fa.gz.gfaffix)               [100%] 1 of 1 ✔
[c2/469298] process > NFCORE_PANGENOME:PANGENOME:PGGB:ODGI_QC:ODGI_DRAW_HEIGHT (DRB1-3123.fa.gz.gfaffix)               [100%] 1 of 1 ✔
[06/0c9f72] process > NFCORE_PANGENOME:PANGENOME:CUSTOM_DUMPSOFTWAREVERSIONS (1)                                        [100%] 1 of 1 ✔
[50/a8350a] process > NFCORE_PANGENOME:PANGENOME:MULTIQC                                                                [100%] 1 of 1 ✔
-[nf-core/pangenome] Pipeline completed successfully-
```

Set pipeline resources

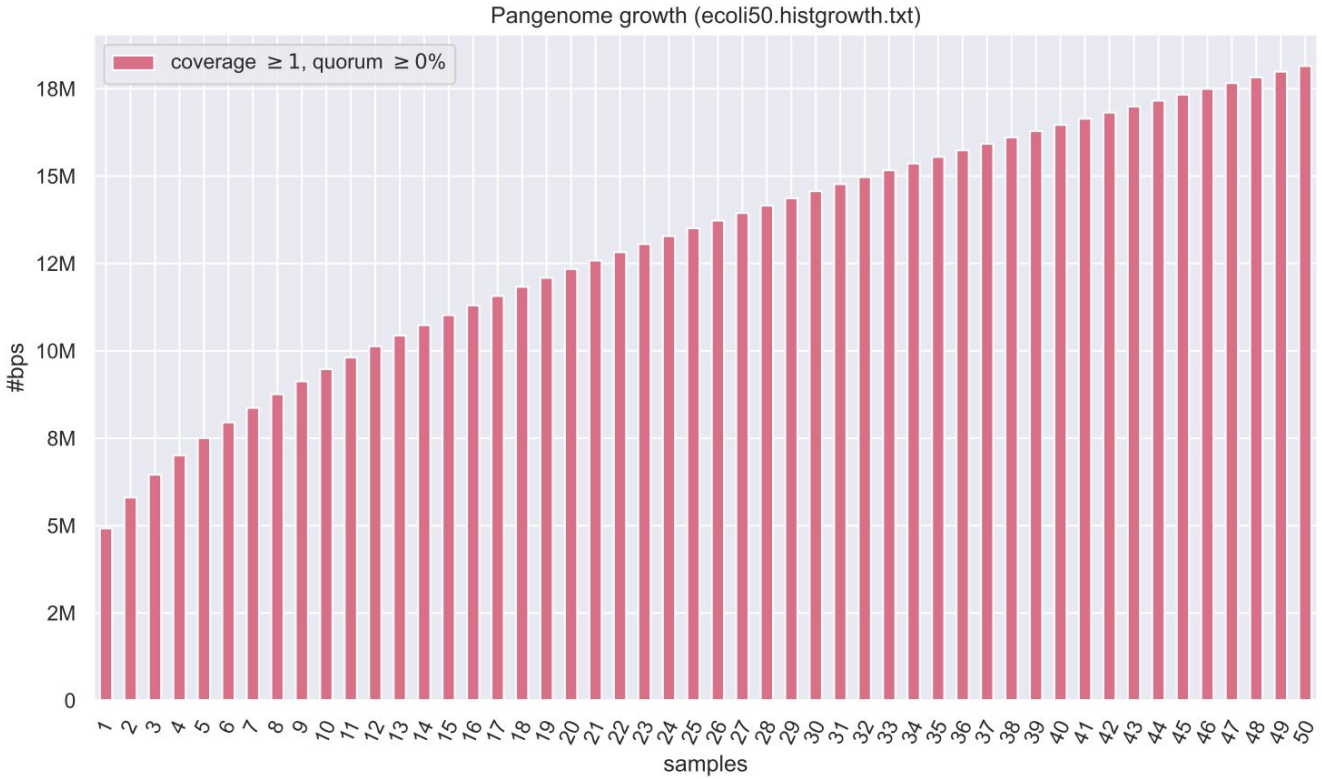
```
process {
    withName:'MULTIQC|MULTIQC_COMMUNITY|SAMTOOLS_FAIDX|CUSTOM_DUMPSOFTWAREVERSIONS' {
        cpus = 1
        memory = 1.GB
    }

    withName:'TABIX_BGZIP|ODGI_STATS|WFMASH_ALIGN|VG_DECONSTRUCT' {
        cpus = 4
        memory = 1.GB
    }

    withName:'WFMASH_MAP_ALIGN|WFMASH_MAP|SEQWISH|ODGI_BUILD|ODGI_UNCHOP|ODGI_SORT|ODGI_LAYOUT|WFMASH_MAP_COMMUNITY|ODGI_SQUEEZE' {
        cpus = 4
        memory = 4.GB
    }

    withName:'SMOOTHXG' {
        cpus = 4
        memory = 8.GB
    }

    withName:'GFAFFIX|ODGI_VIEW|ODGI_VIZ*|ODGI_DRAW|SPLIT_APPROX_MAPPINGS_IN_CHUNKS|PAF2NET|NET2COMMUNITIES|EXTRACT_COMMUNITIES' {
        cpus = 1
        memory = 4.GB
    }
}
```

What do we get?

- No new biological results or algorithms compared to PGGB
- Same CPU hours than PGGB
- Highly reduced wall clock time when running on a cluster
- The load of a cluster is more balanced
- Easily run the pipeline in all major clouds saving resources and money

Computing pangenome growth

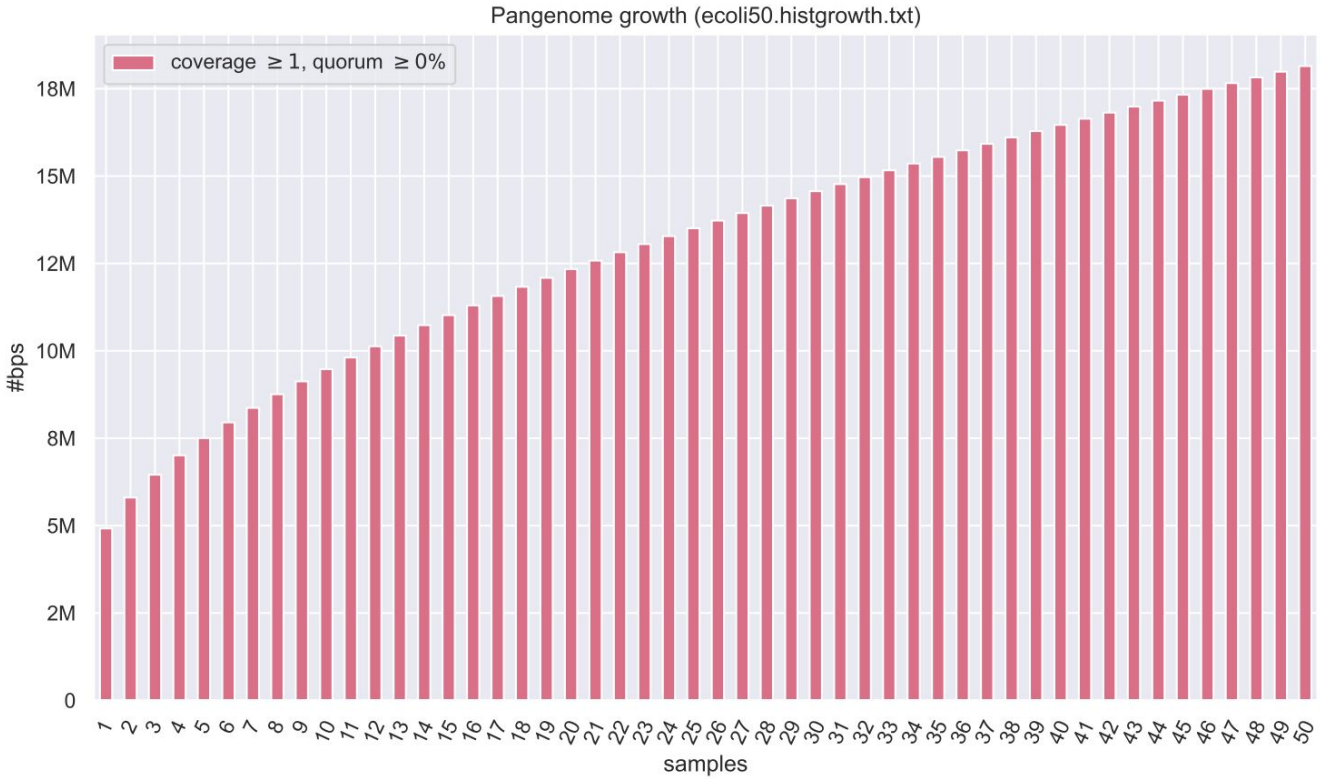
Slide modified from
Daniel Doerr et al.



- Genomic variation within a species

Computing pangenome growth

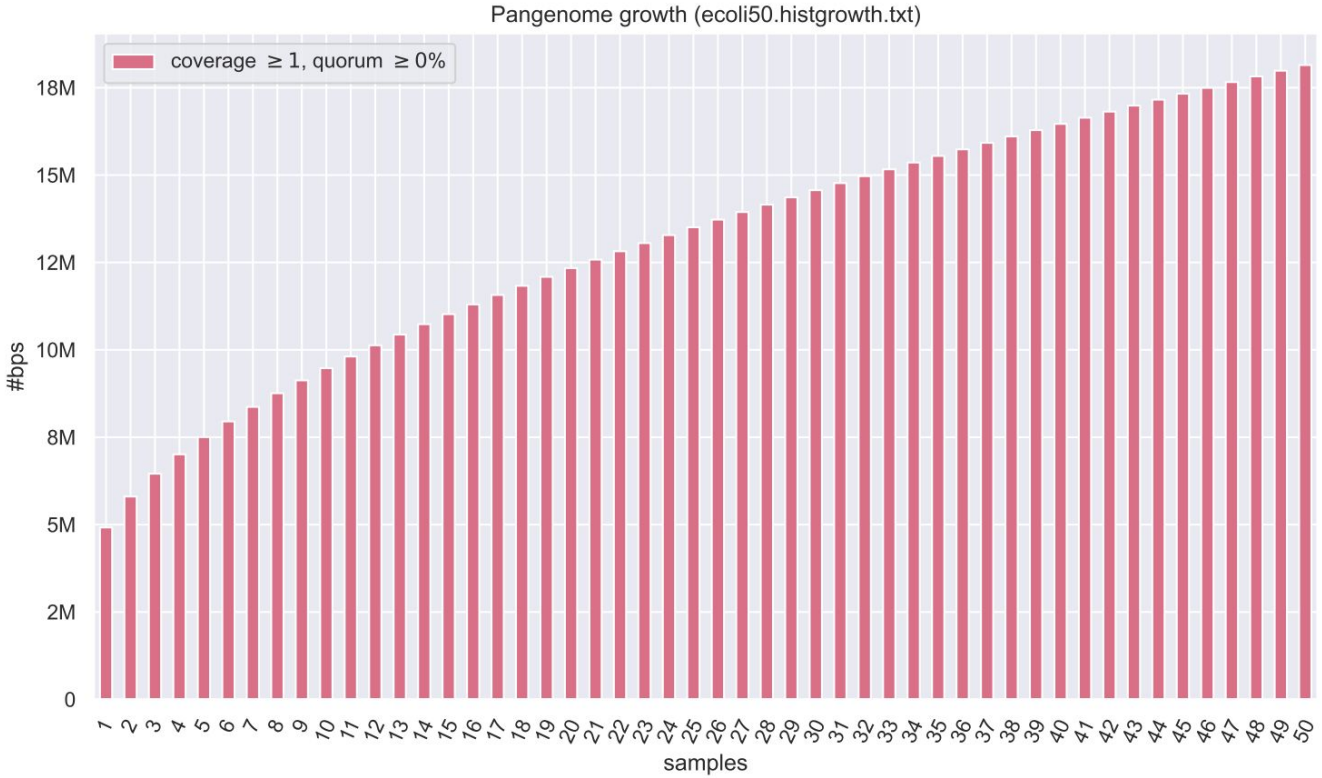
Slide modified from
Daniel Doerr et al.



- Genomic variation within a species
- Population structure

Computing pangenome growth

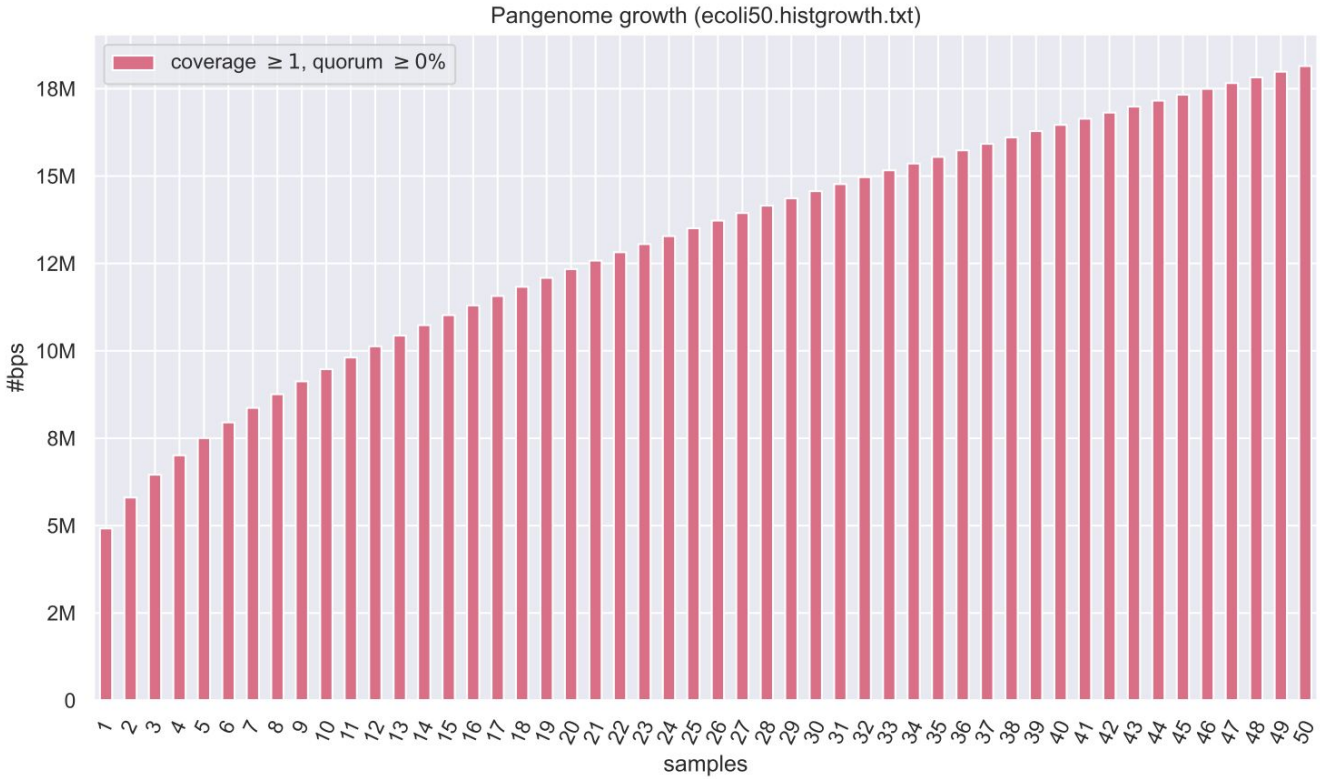
Slide modified from
Daniel Doerr et al.



- Genomic variation within a species
- Population structure
- Genetic evolution: Pathogen research

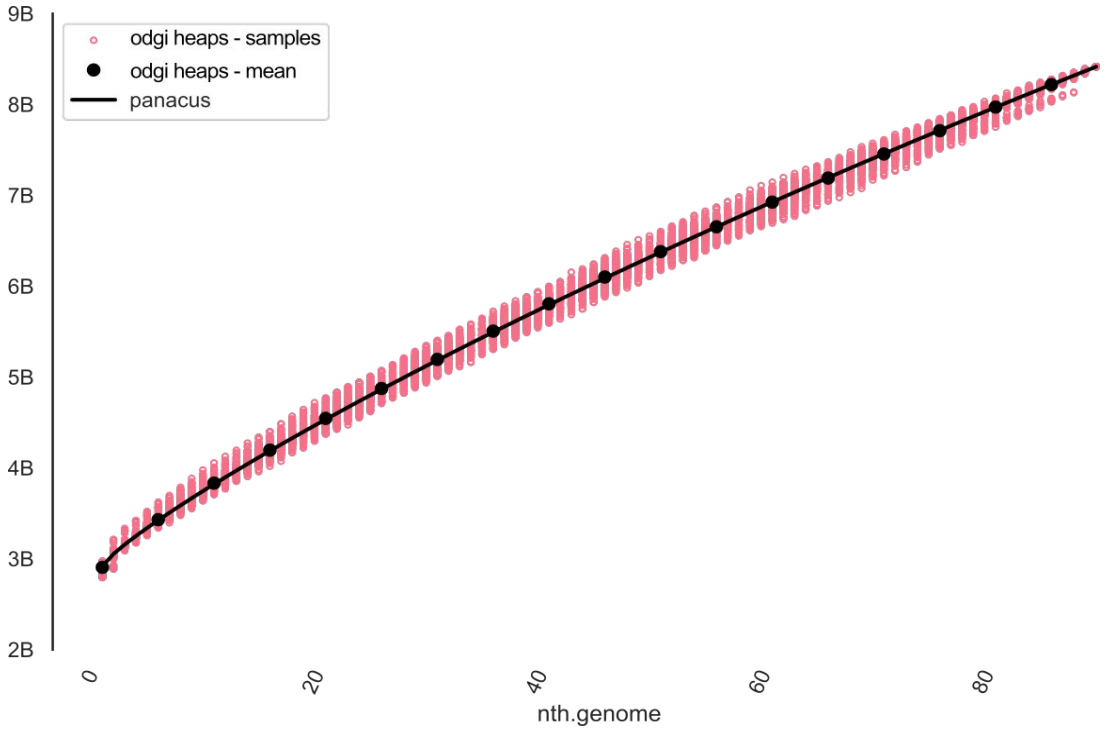
Computing pangenome growth

Slide modified from
Daniel Doerr et al.



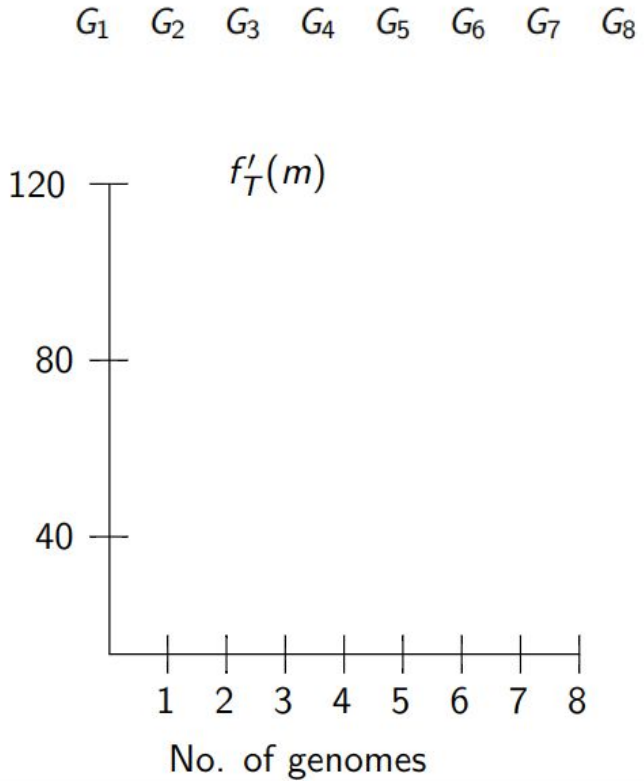
- Genomic variation within a species
- Population structure
- Genetic evolution: Pathogen research
- Biodiversity conservation

Computing pangenome growth statistics



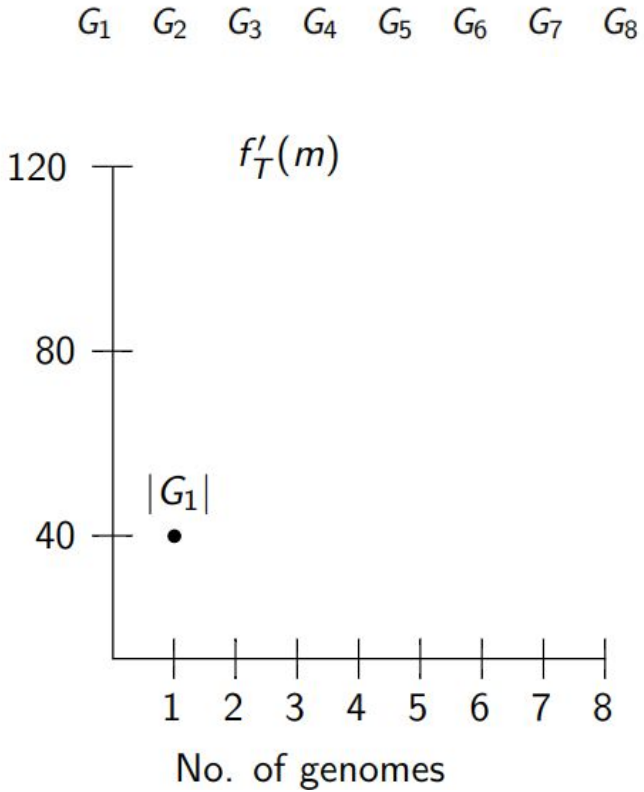
Slide taken from
[Daniel Doerr et al.]

Pangenome growth computation



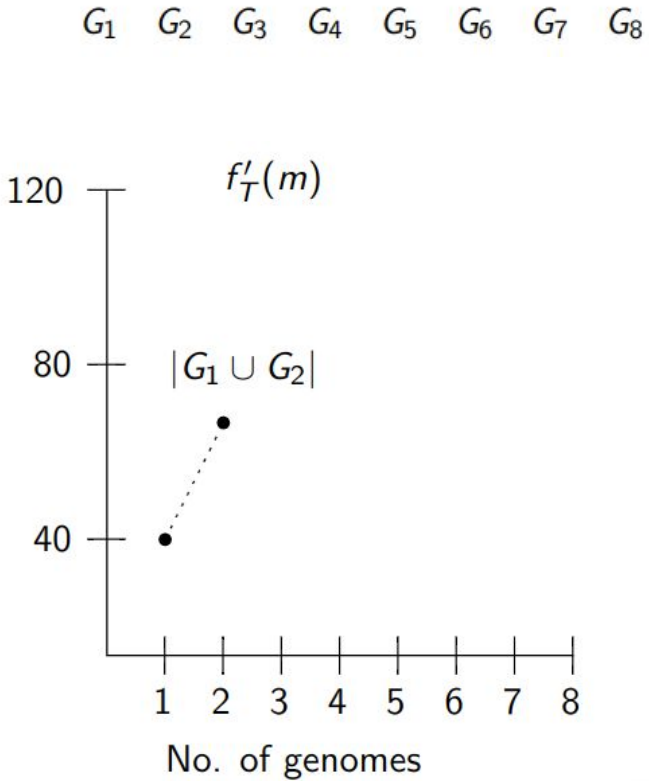
Slide taken from
Daniel Doerr et al.

Pangenome growth computation



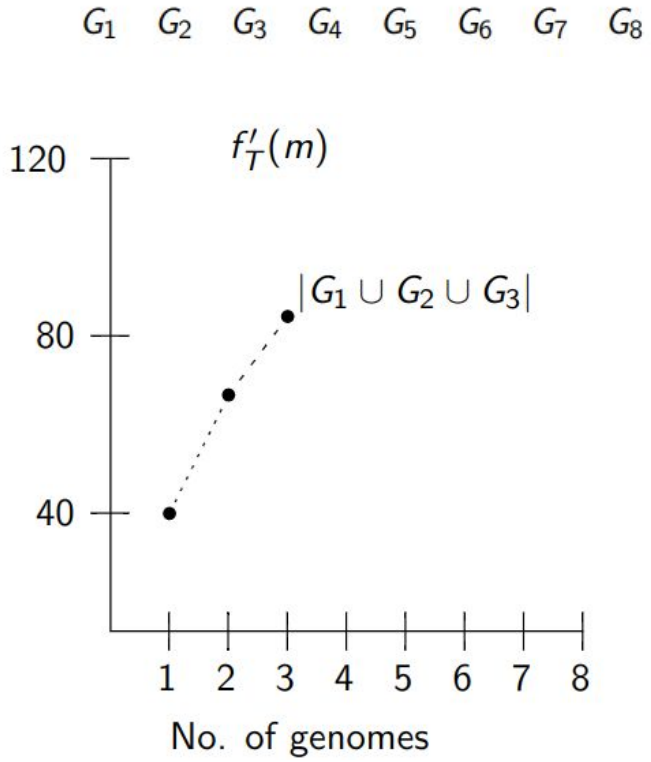
Slide taken from
Daniel Doerr et al.

Pangenome growth computation



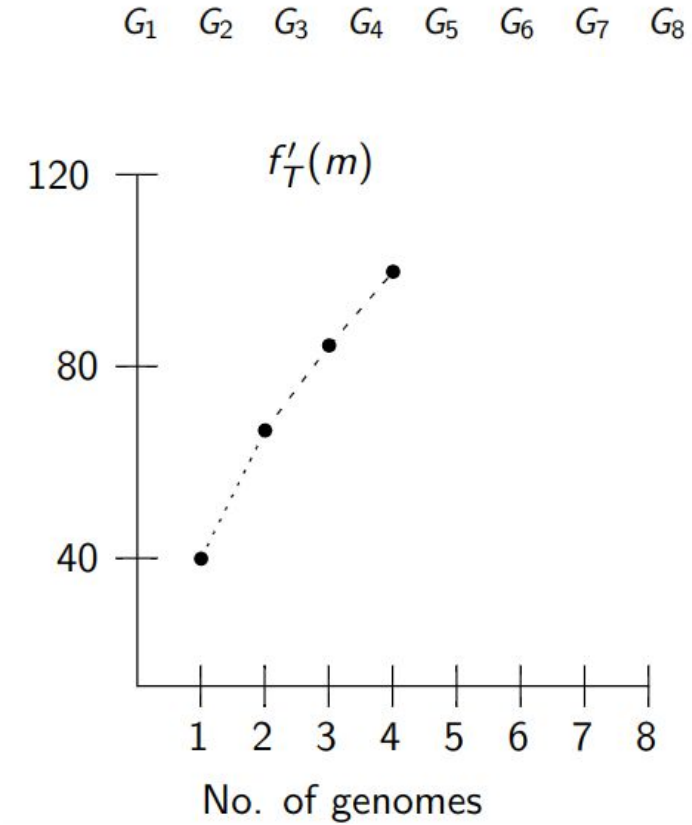
Slide taken from
Daniel Doerr et al.

Pangenome growth computation



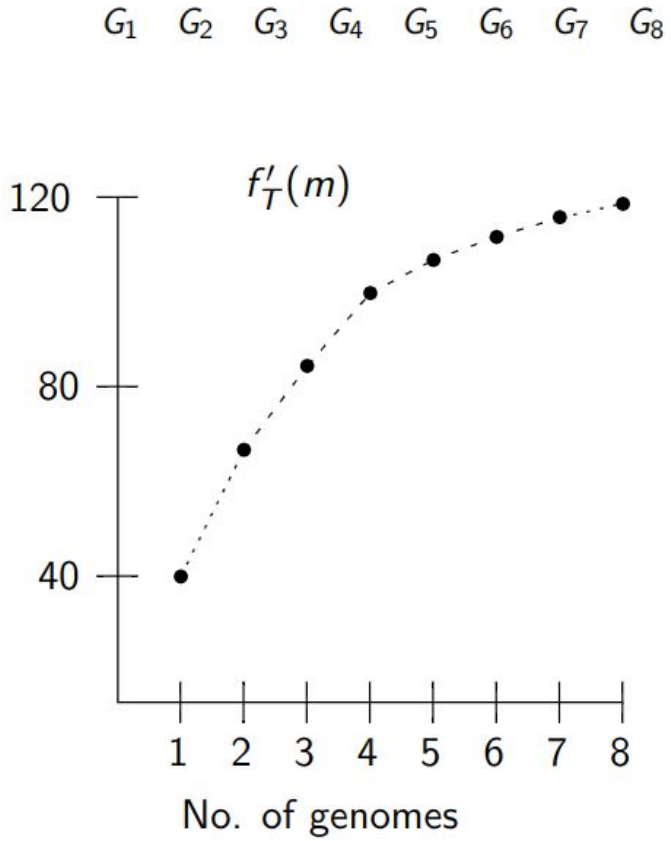
Slide taken from
[Daniel Doerr et al.]

Pangenome growth computation



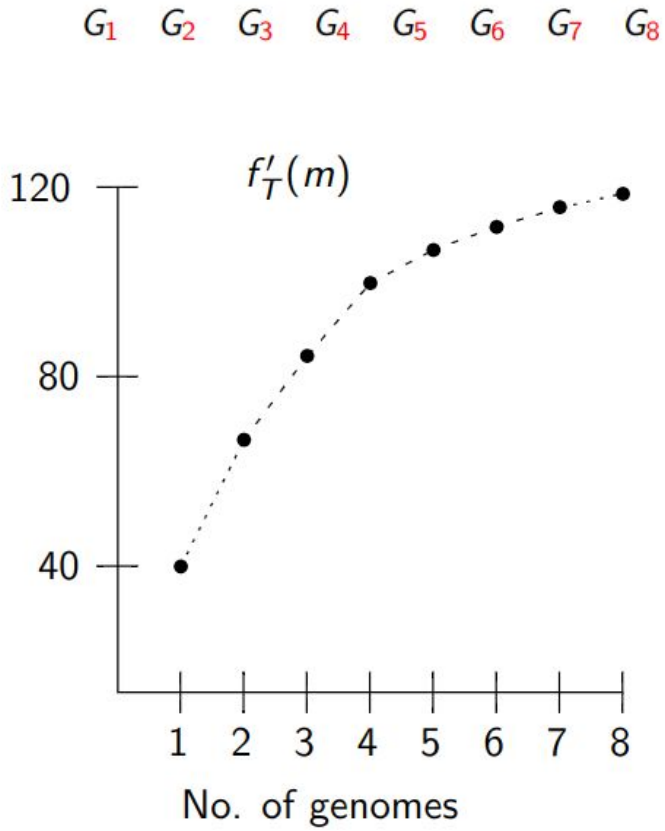
Slide taken from
Daniel Doerr et al.

Pangenome growth computation



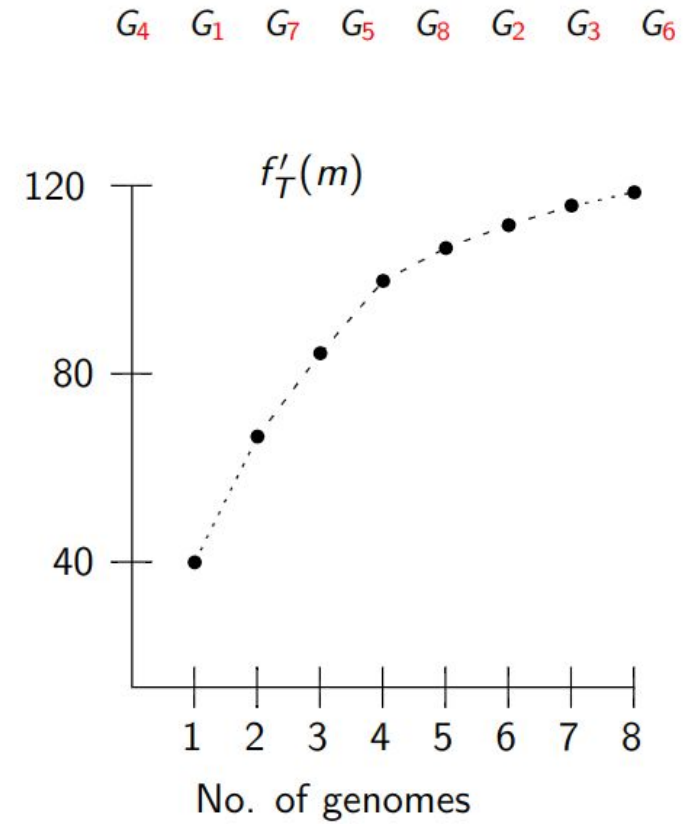
Slide taken from
Daniel Doerr et al.

Pangenome growth computation



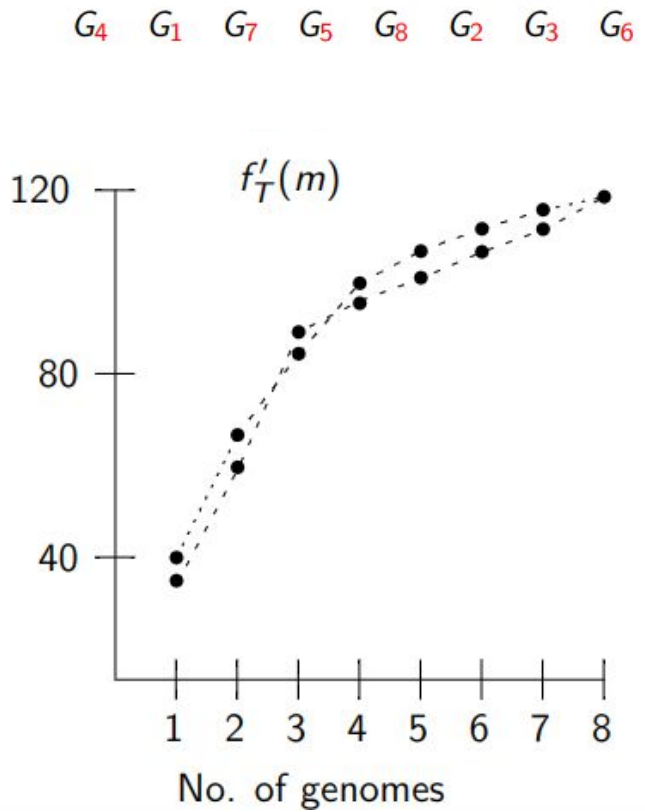
Slide taken from
Daniel Doerr et al.

Pangenome growth computation



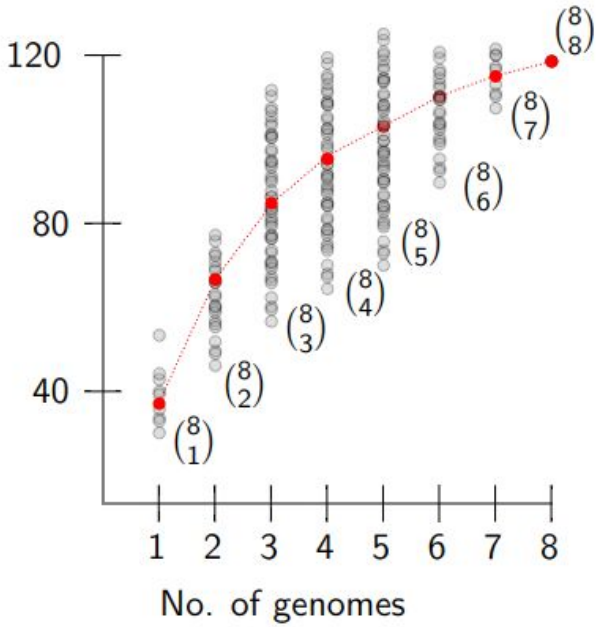
Slide taken from
Daniel Doerr et al.

Pangenome growth computation



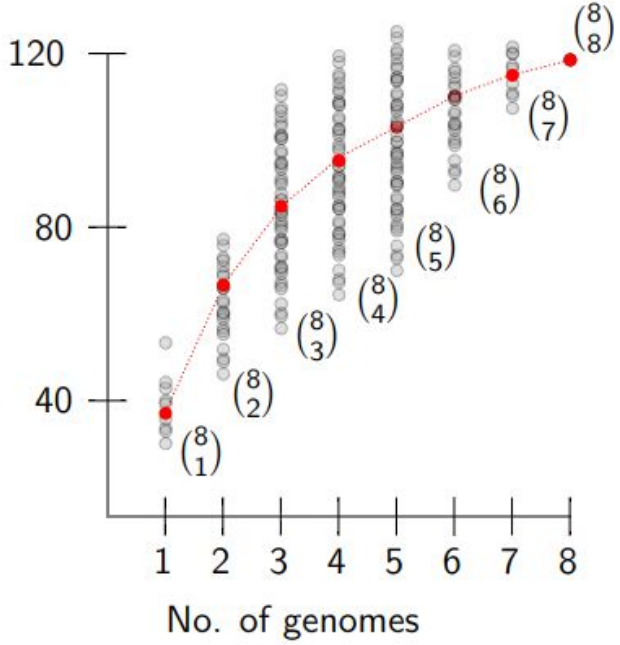
Slide taken from
Daniel Doerr et al.

Pangenome growth computation



Slide taken from
Daniel Doerr et al.

Pangenome growth computation - *odgi heaps*



$$f_{\text{tot}}(m) = \frac{1}{\binom{n}{m}} \sum_{\mathcal{G}' \in \mathcal{G}_m} \left| \bigcup_{G \in \mathcal{G}'} G \right|$$

m: total number of genomes

n: number of genomes in current growth computation

Slide modified from
Daniel Doerr et al.

Pangenome growth computation - *Panacus*

Slide modified from
Daniel Doerr et al.

$$f_{\text{tot}}(m) = \frac{1}{\binom{n}{m}} \sum_{\mathcal{G}' \in \mathcal{G}_m} \left| \bigcup_{G \in \mathcal{G}'} G \right| =$$

$h(i) =$ number of items occuring in **exactly** i genomes

- Time complexity $O(n^2)$ << Number of countables (nodes, edges, bps)
- 1971, Stuart H. Hurlbert[a]

[a]Hurlbert, S. H. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. Ecology, 52(4), 577–586
[b]A. Chacoma and D. H. Zanette. Heaps' law and Heaps functions in tagged texts:evidences of their linguistic relevance. Royal Society Open Science, 7(3),2020

Pangenome growth computation - *Panacus*

Slide modified from
Daniel Doerr et al.

$$f_{\text{tot}}(m) = \frac{1}{\binom{n}{m}} \sum_{\mathcal{G}' \in \mathcal{G}_m} \left| \bigcup_{G \in \mathcal{G}'} G \right| = \sum_{i=1}^{n} h(i) \left(1 - \frac{(n-i)^{\underline{m}}}{n^{\underline{m}}}\right)$$

$h(i)$ = number of items occuring in **exactly** i genomes

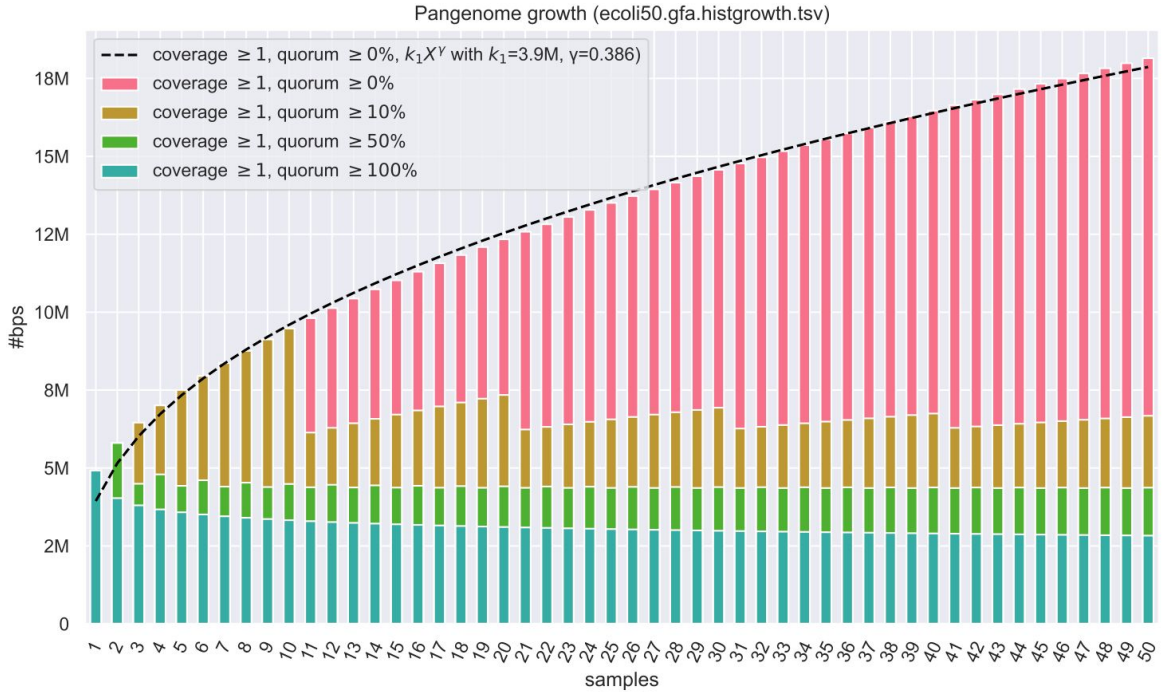
- Time complexity $O(n^2)$ << Number of countables (nodes, edges, bps)
- 1971, Stuart H. Hurlbert[a]
- 2020, A. Chacoma and D. H. Zanette[b]

> [...] “it seems to have passed unnoticed that its exact analytical expression has been available in the literature since at least four decades ago.”

[a]Hurlbert, S. H. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. Ecology, 52(4), 577–586
[b]A. Chacoma and D. H. Zanette. Heaps’ law and Heaps functions in tagged texts:evidences of their linguistic relevance. Royal Society Open Science, 7(3),2020

$$n^{\underline{m}} = \overbrace{n(n-1)\dots(n-m+1)}^{m \text{ factors}}$$

Pangenome growth curve with Panacus



quorum >= 0%: all seqs of all haps

quorum >= 10%: seqs traversed by at least 10% of haps - ***cloud pangenome***

quorum >= 50%: seqs traversed by at least 50% of haps - ***shell pangenome***

quorum >= 100%: seqs traversed by 100% of haps - ***core pangenome***

Pangenome growth curve HPRC

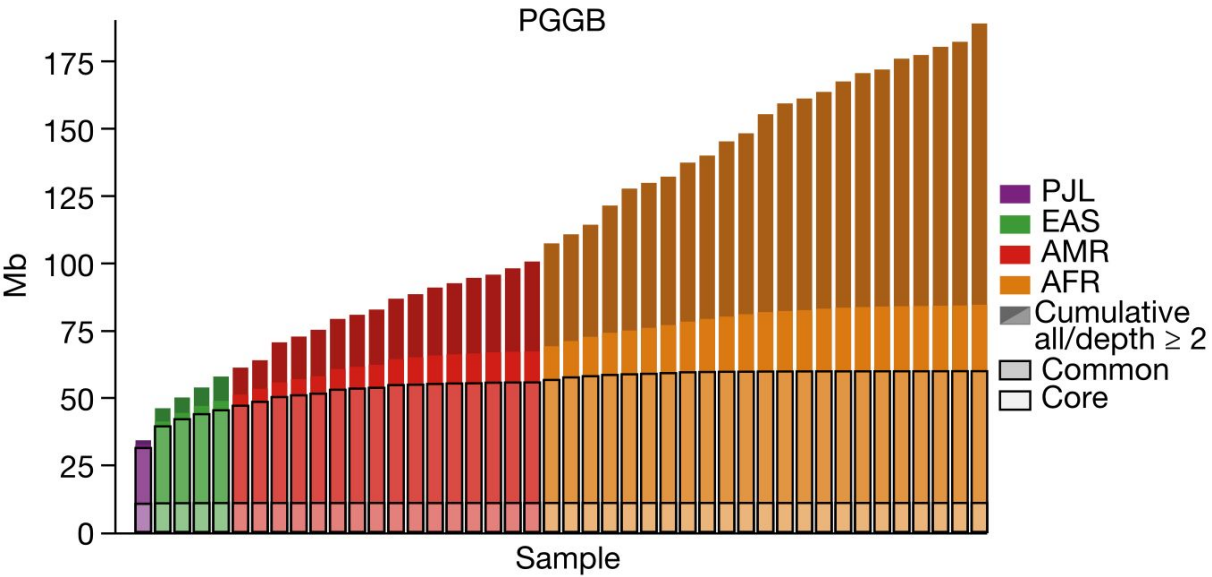


Figure from: Liao et al., 2023.

PJL: Punjabi in Lahore

EAS: East Asia

AMR: American

AFR: African

Common: Sequences traversed by at least 5% of haplotypes

Core: Sequences traversed by at least 95% of haplotypes

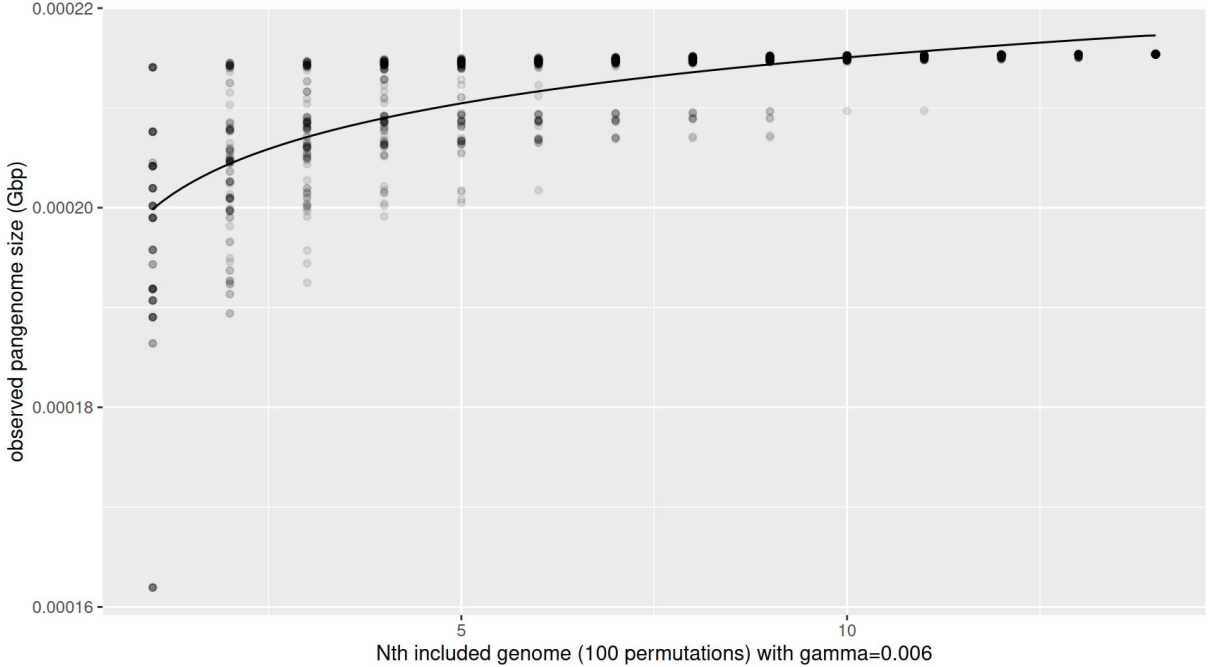
Activities

https://hackmd.io/snceiMesSECepJGzBdDIkw?view=

In this tutorial you will learn to

- run a nf-core Nextflow pipeline,
- configure the resources according to what is available,
- deal with alternative parameter names,
- understand the nf-core/pangenome pipeline output,
- and evaluate and interpret the growth of a pangenome.

Heap’s law for measuring pangenome openness



Pangenome growth curve of the LPA pangenome graph

$$n = \kappa N^{\gamma}$$

n is pangenome size

N is the number of genomes

κ and **γ** are fitting parameters

γ<0: closed pangenome

γ>0: open pangenome

Tettelin H. et al. (2008). Current Opinion in Microbiology, 11(5), 472-477.

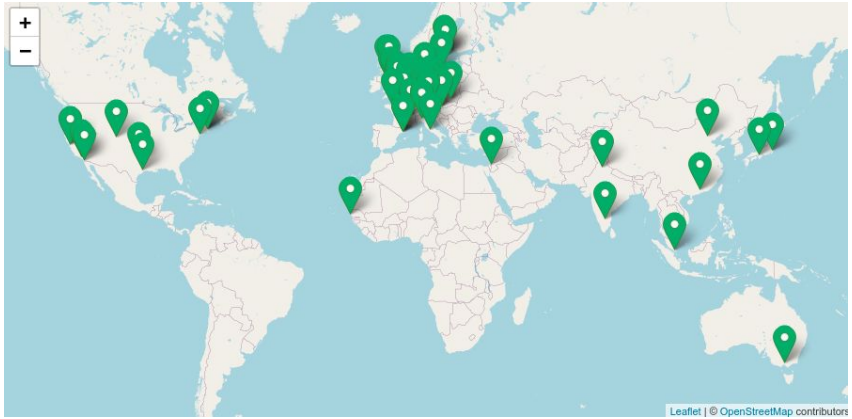
nf-core

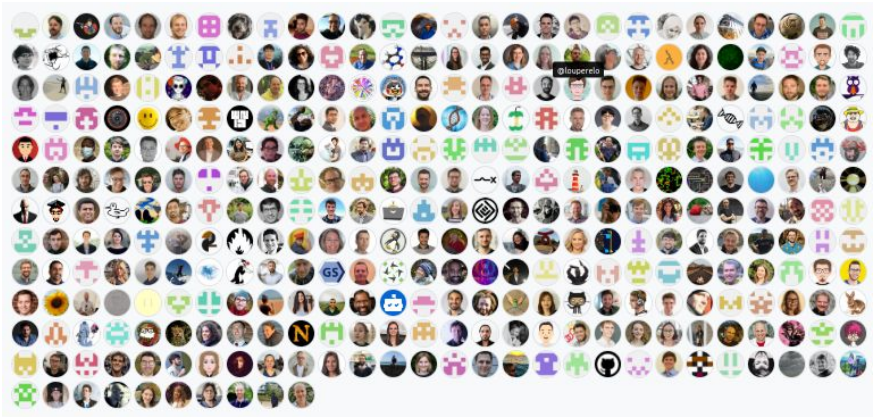
>6000

Slack Members

>700

GitHub Organisation Members





Slide with permission by James Fellows Yates

nf-core

49

✅ Released pipelines

19

🔧 Under development

Genomics	Metagenomics
Proteomics	Metabolomics
Epigenetics	(Bio)imaging
Transcriptomics	Analysis/Utility
Immunology	Economics*

* and astronomy and earth sciences coming soon ™

& many more…

https://nf-co.re

Slide modified with permission by James Fellows Yates

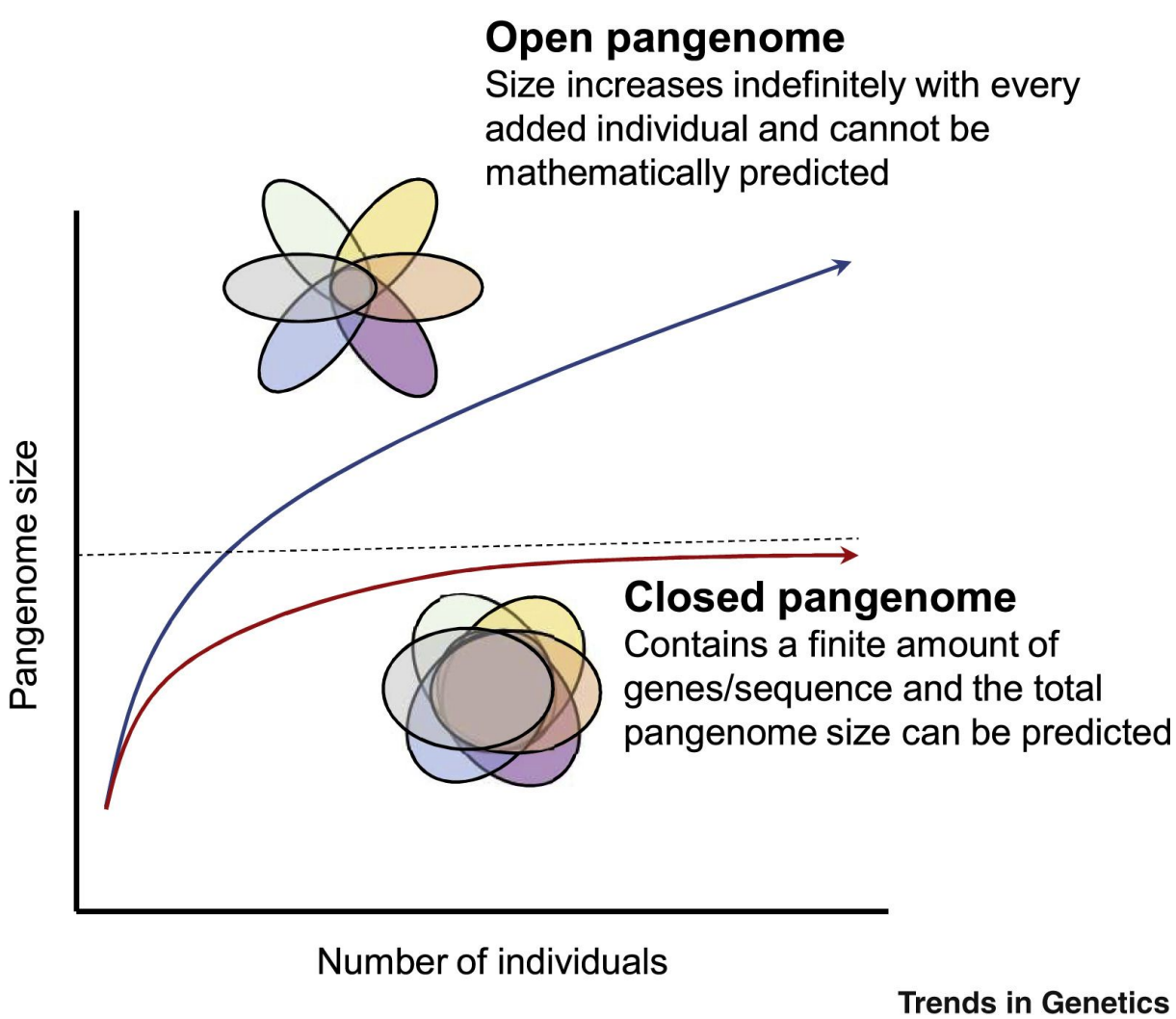
Perspective | Published: 23 September 2021

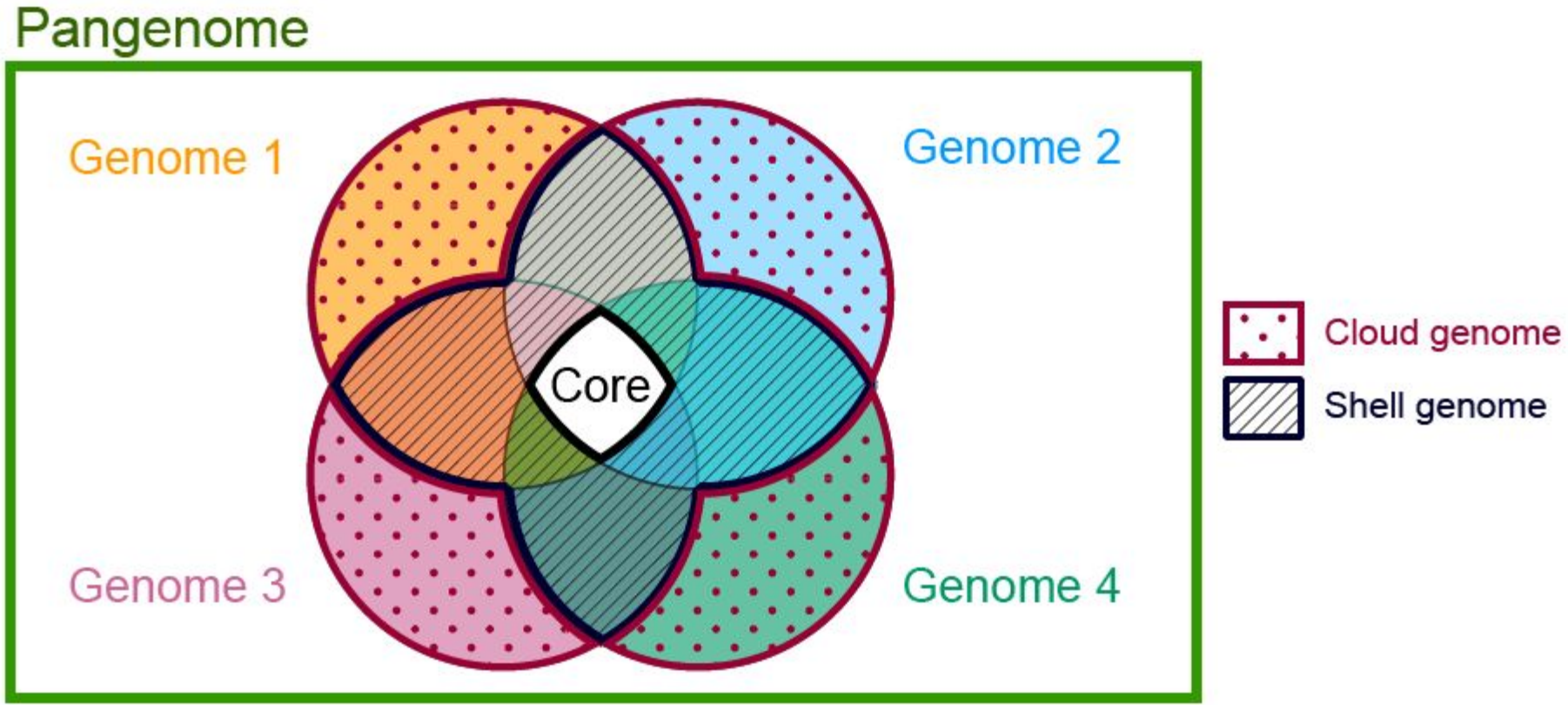
Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers

Laura Wratten, Andreas Wilm & Jonathan Göke ✉

Nature Methods **18**, 1161–1168 (2021) | Cite this article

Tool	Class	Ease of use[a]	Expressiveness[b]	Portability[c]	Scalability[d]	Learning resources[e]	Pipeline initiatives[f]
Galaxy	Graphical	●●●	●○○	●●●	●●●	●●●	●●○
KNIME	Graphical	●●●	●○○	○○○	●●◐	●●●	●●○
Nextflow	DSL	●●○	●●●	●●●	●●●	●●●	●●●
Snakemake	DSL	●●○	●●●	●●◐	●●●	●●○	●●●
GenPipes	DSL	●●○	●●●	●●○	●●○	●●○	●●○
bPipe	DSL	●●○	●●●	●●○	●●◐	●●○	●○○
Pachyderm	DSL	●●○	●●●	●○○	●●○	●●●	○○○
SciPipe	Library	●●○	●●●	○○○	○○○	●●○	○○○
Luigi	Library	●●○	●●●	●○○	●●◐	●●○	○○○
Cromwell + WDL	Execution + workflow specification	●○○	●●○	●●●	●●◐	●●○	●●○
cwltool + CWL	Execution + workflow specification	●○○	●●○	●●◐	○○○	●●●	●●○
Toil + CWL/WDL/Python	Execution + workflow specification	●○○	●●●	●◐○	●●●	●●○	●●○


Open pangenome
Size increases indefinitely with every added individual and cannot be mathematically predicted
Pangenome size
Closed pangenome
Contains a finite amount of genes/sequence and the total pangenome size can be predicted
Number of individuals
Trends in Genetics



https://upload.wikimedia.org/wikipedia/commons/9/9b/Parts_of_the_pangenome.png

Core: 100% of genomes, minimum fraction → quorum: 1.0
Shell: 50% of genomes, minimum fraction → quorum: 0.5
Cloud: 10% of genomes, minimum fraction → quorum: 0.1