Floria: metagenome strain haplotyping with short/long reads

Jim Shaw[1*], Jean-Sébastien Gounot[2*],
Hanrong Chen[2], Niranjan Nagarajan[2#], Yun William Yu[1,3#]

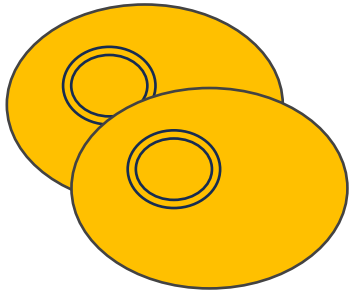
* Equal contribution

[1]University of Toronto
[2]Genome Institute of Singapore
[3]Carnegie Mellon University

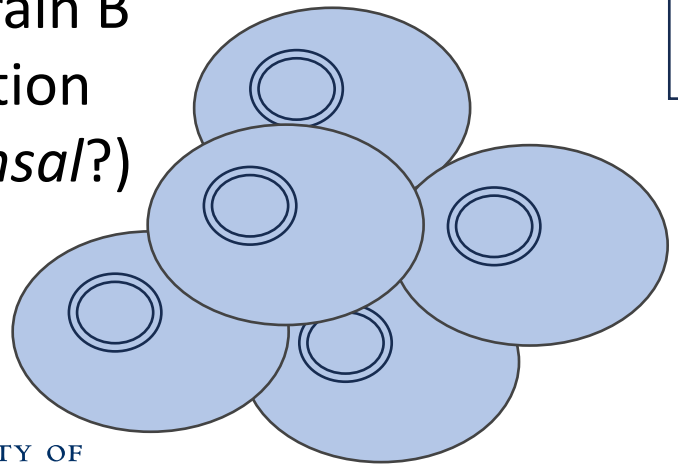
UNIVERSITY OF TORONTO

Strain-level heterogeneity in metagenomes is important!

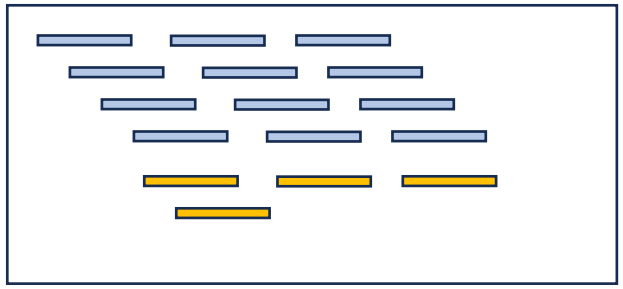
E. coli strain A population (*pathogenic*?)



E. coli strain B population (*commensal*?)



Strain-mixed metagenomic reads

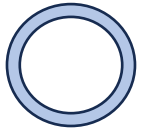


Assembly

What is assembled?

A) only

B) only



C) both

UNIVERSITY OF TORONTO

Answer: depends on technology + algorithms

Technology	Resulting assembly
Short reads (algorithm: SPAdes)	**One** strain (lost information)
Low-fidelity long-reads (algorithm: metaFlye)	**One** strain (lost information)
High-fidelity long-reads (algorithm: hifiasm or metaMDBG)	**Both** strains

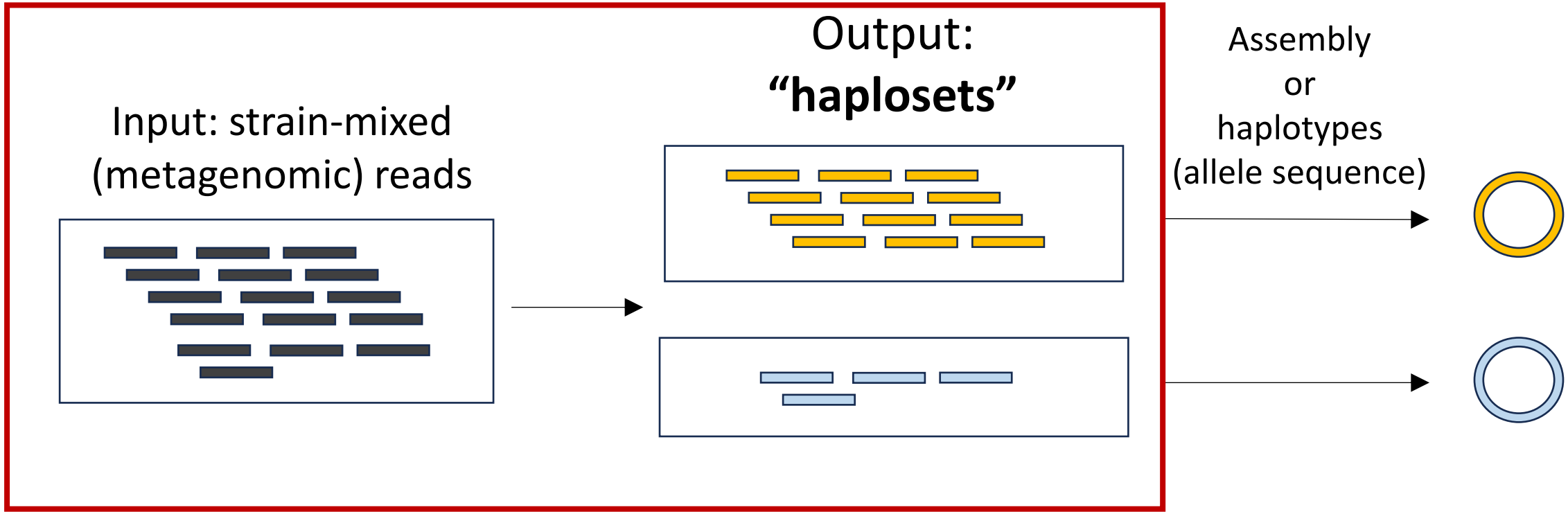
UNIVERSITY OF TORONTO

Presentation focus

Introduction:
Haplotyping and phasing

UNIVERSITY OF TORONTO

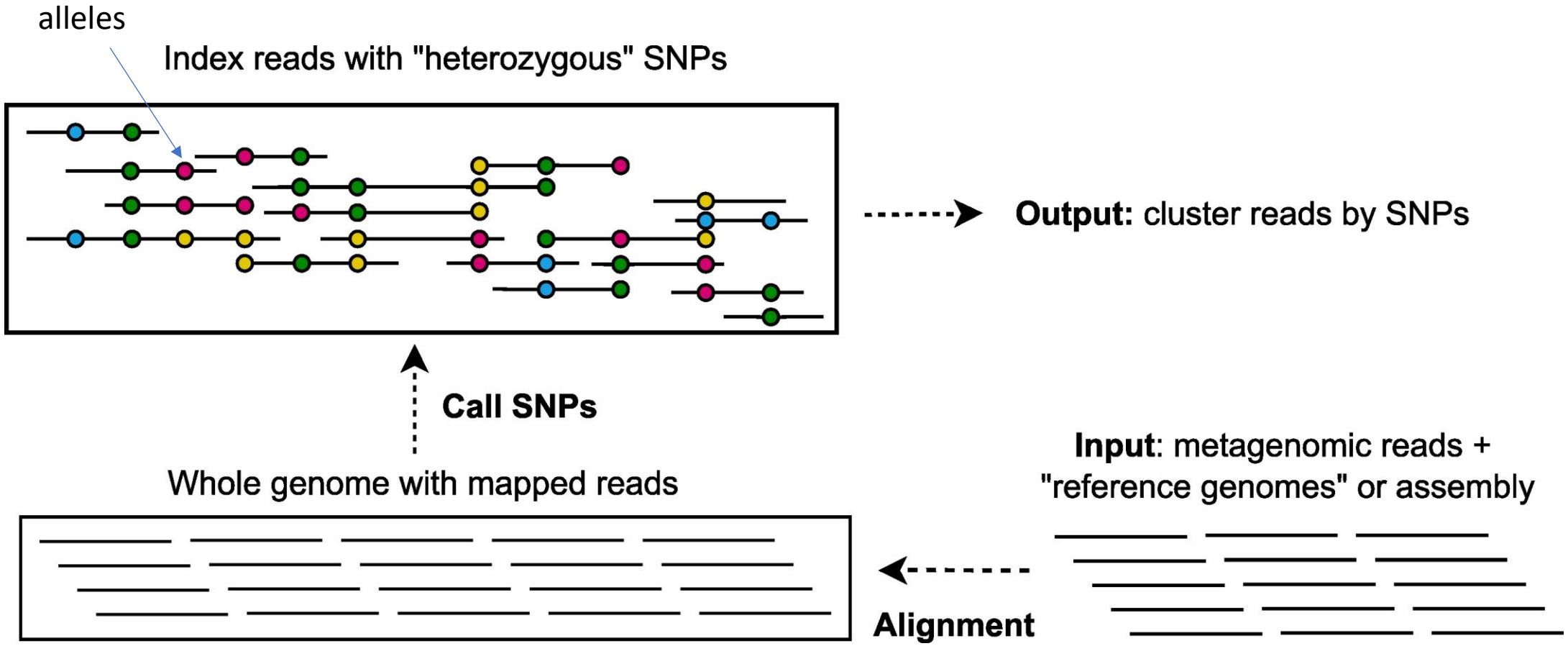
Computational goal:
reads → **strain-level** "haplosets"



UNIVERSITY OF TORONTO

Presentation focus

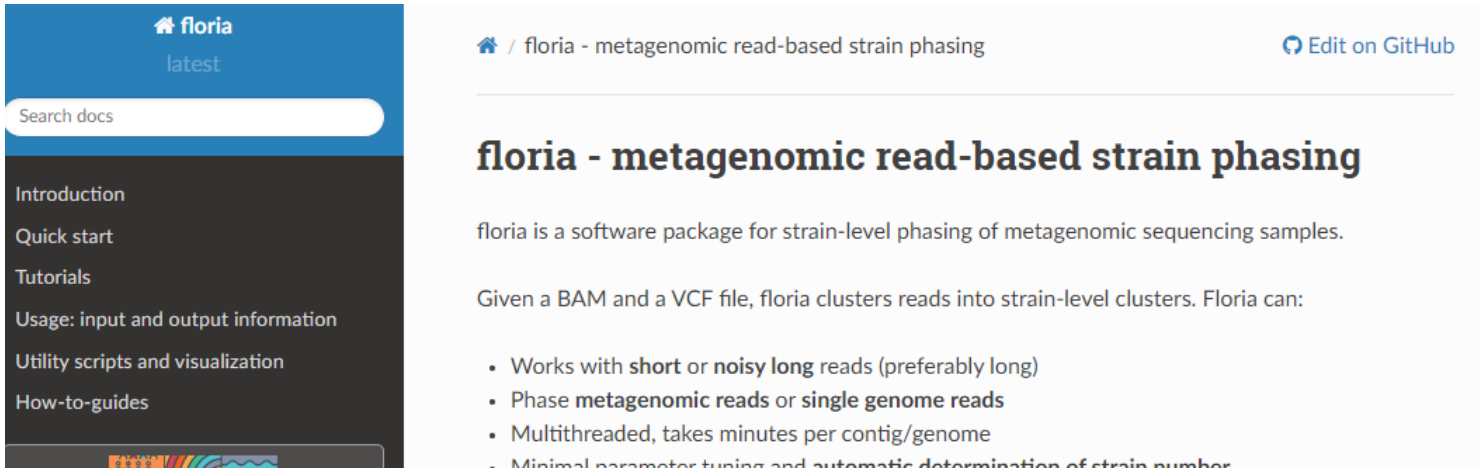
Main idea: computational phasing (haplotyping) using SNPs + read-linkage



UNIVERSITY OF TORONTO

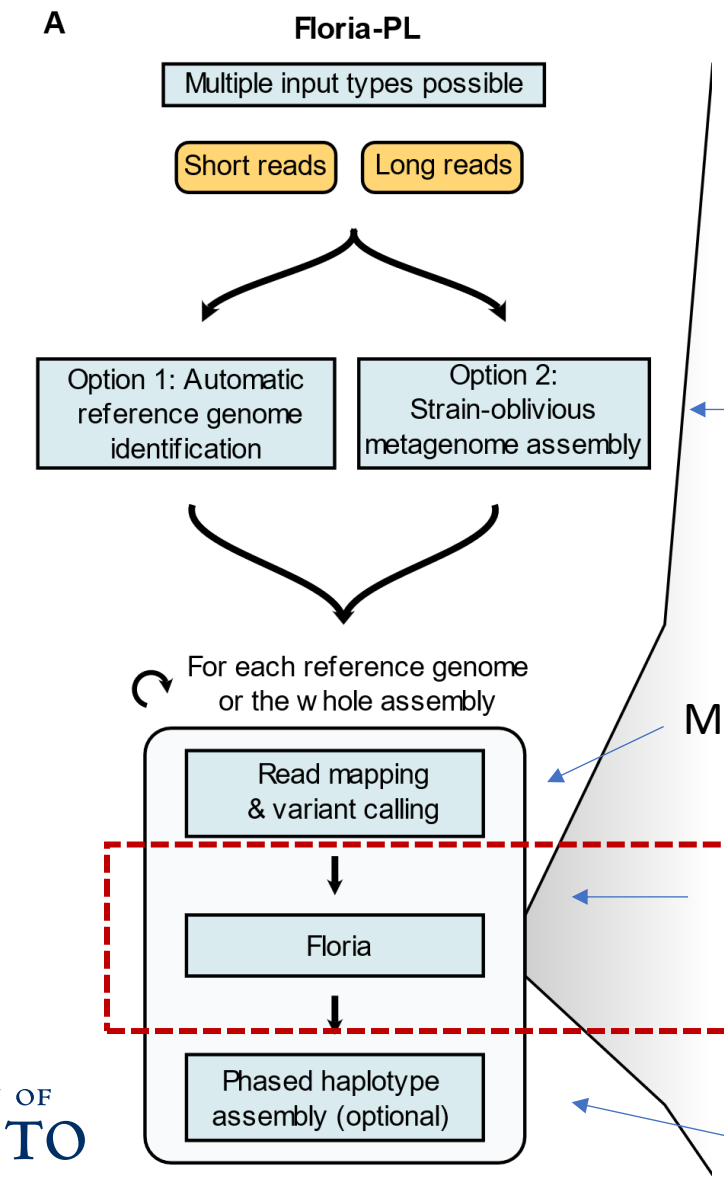
Contributions (Shaw and Gounot et al. 2024)

- **(1) Floria**: read clustering (phasing) algorithm from **alignments + SNPs**
  - Written in Rust; documentation + conda install
- **(2) Floria-PL**: end-to-end pipeline (fastq -> assemblies)
  - Written in Snakemake - **integrating** floria



UNIVERSITY OF TORONTO

Floria-PL: pipeline integrating floria



Generate "reference" genomes (profiling or assembly)

Map reads to genomes + call variants

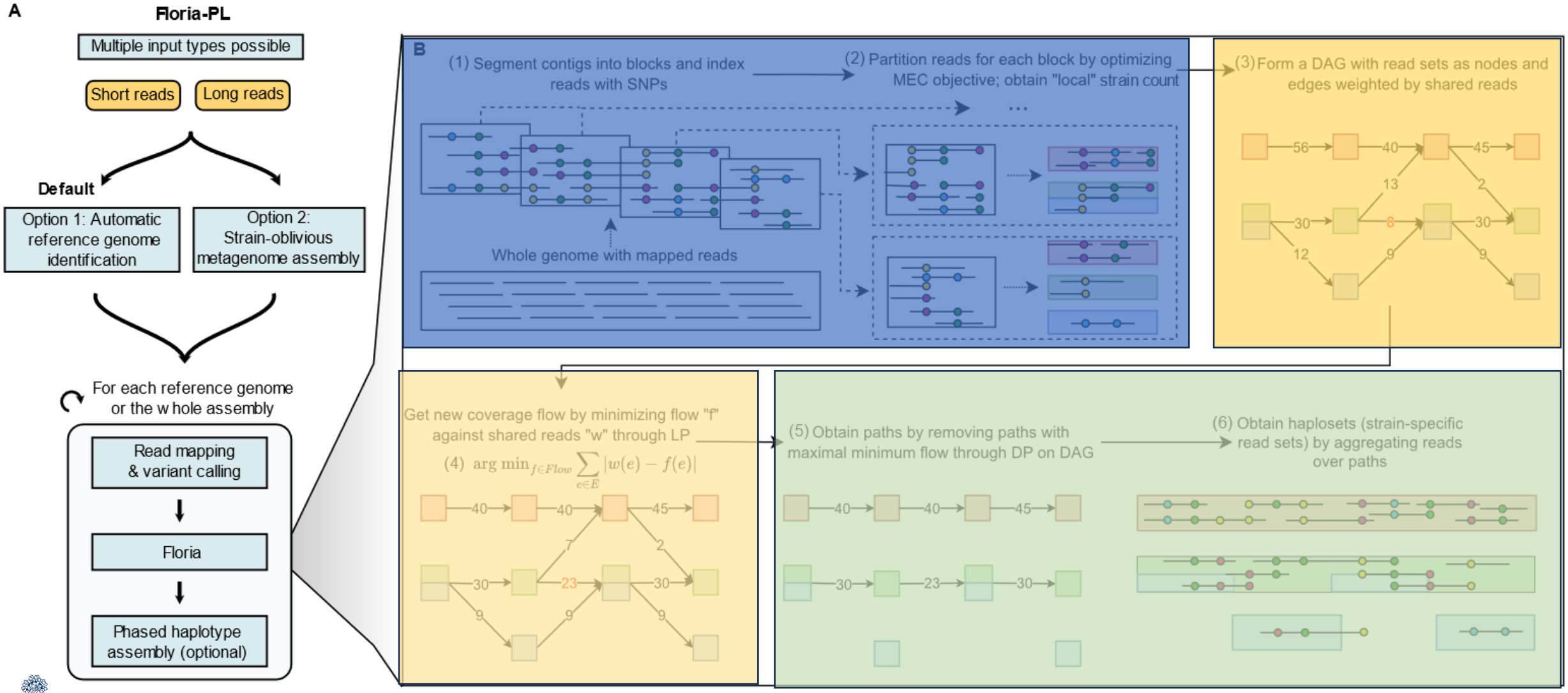
New algorithm: floria

Generate **haplosets from BAM + VCF**

Assemble haplosets - **OPTIONAL**

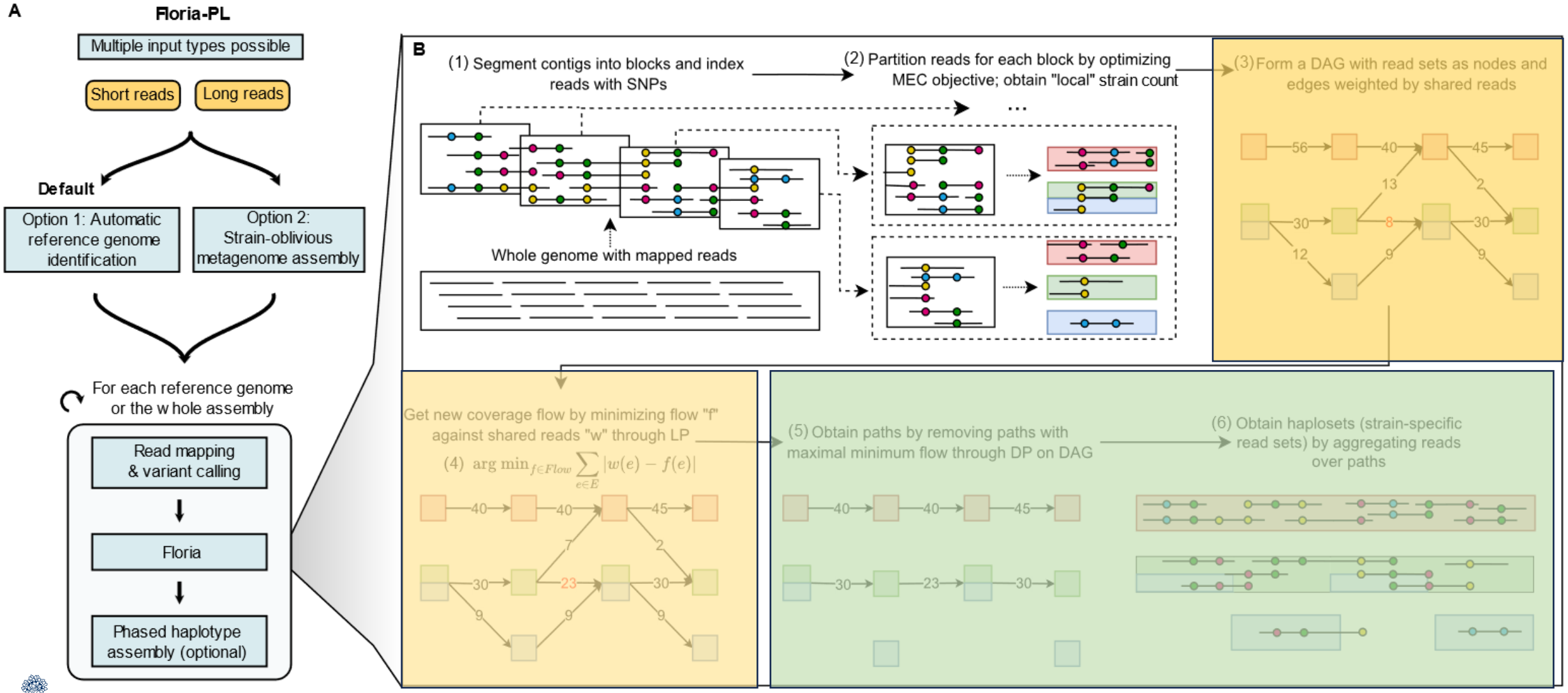
UNIVERSITY OF TORONTO

Floria: read clustering by optimization + network flows

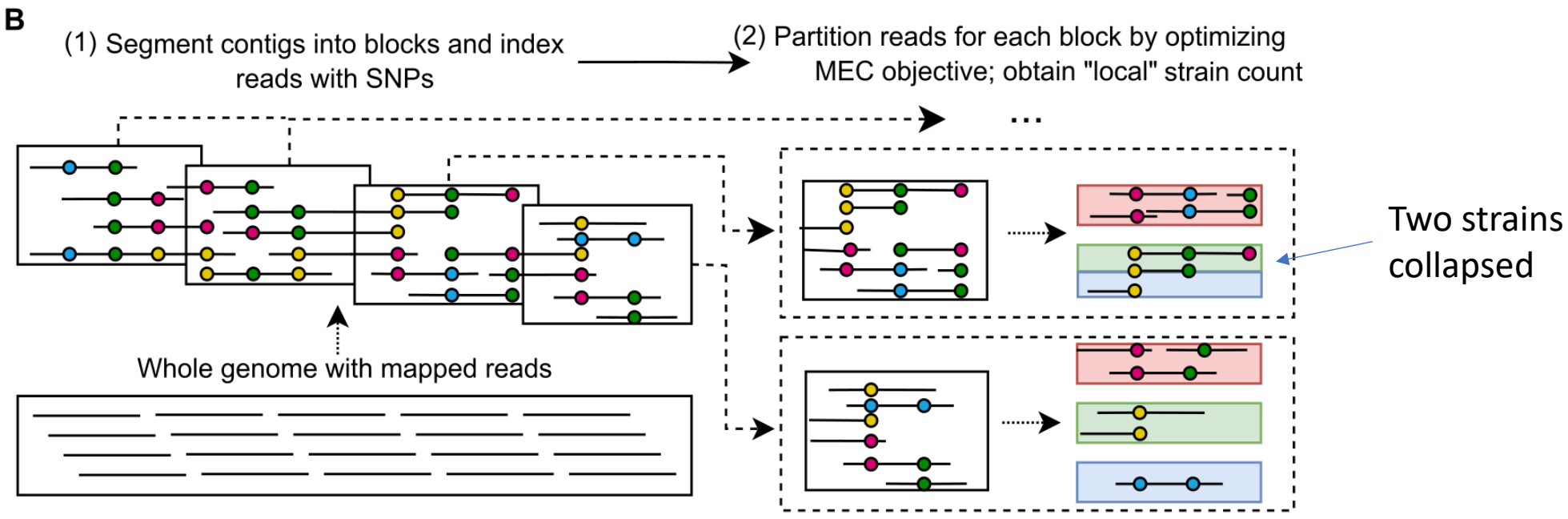


UNIVERSITY OF TORONTO

Step 1: local clustering



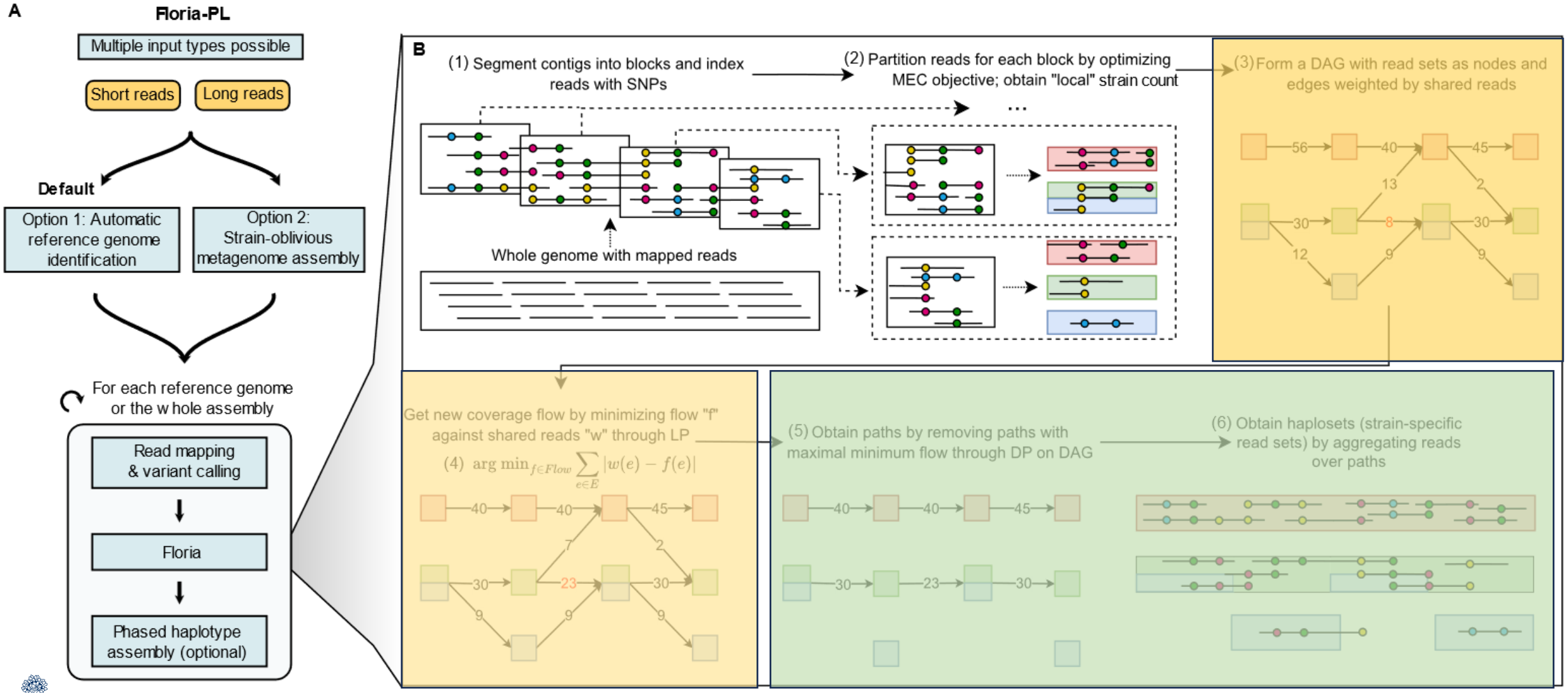
UNIVERSITY OF TORONTO



1. Clustering objective: **minimum error correction** (**MEC**) score
   - NP-Hard (Lancia et al., 2001)
   - **Floria:** beam search heuristic
     - often used in Natural Language Processing (NLP)
2. **# of strains?** → iteratively cluster until MEC score plateaus

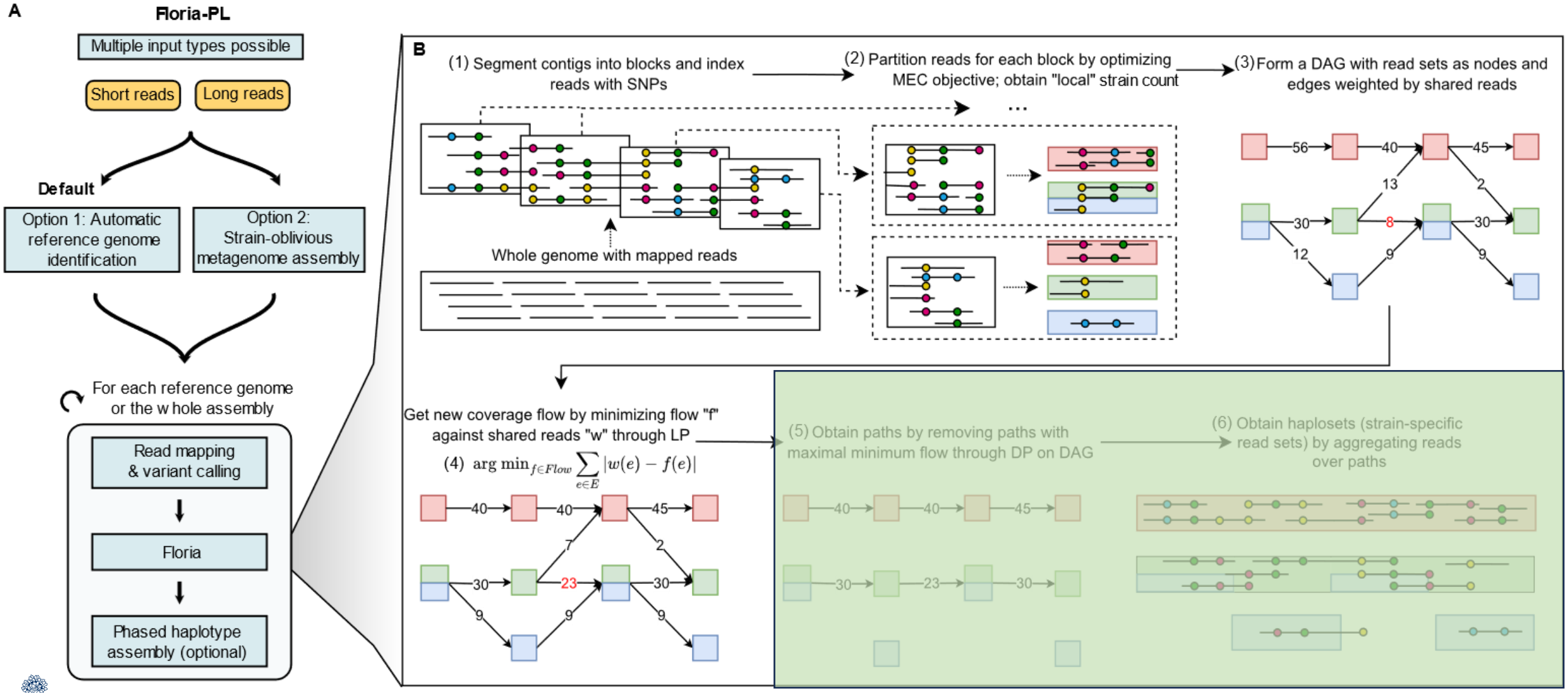
UNIVERSITY OF TORONTO

Step 1: local clustering



UNIVERSITY OF TORONTO

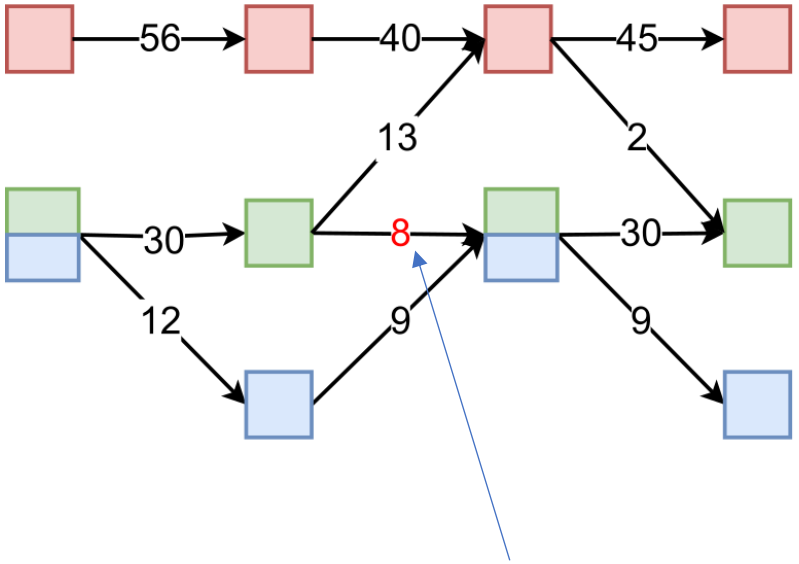
Step 2: **network flows**



UNIVERSITY OF TORONTO

(DAG = Directed acyclic graph)

(3) Form a DAG with read sets as nodes and edges weighted by shared reads



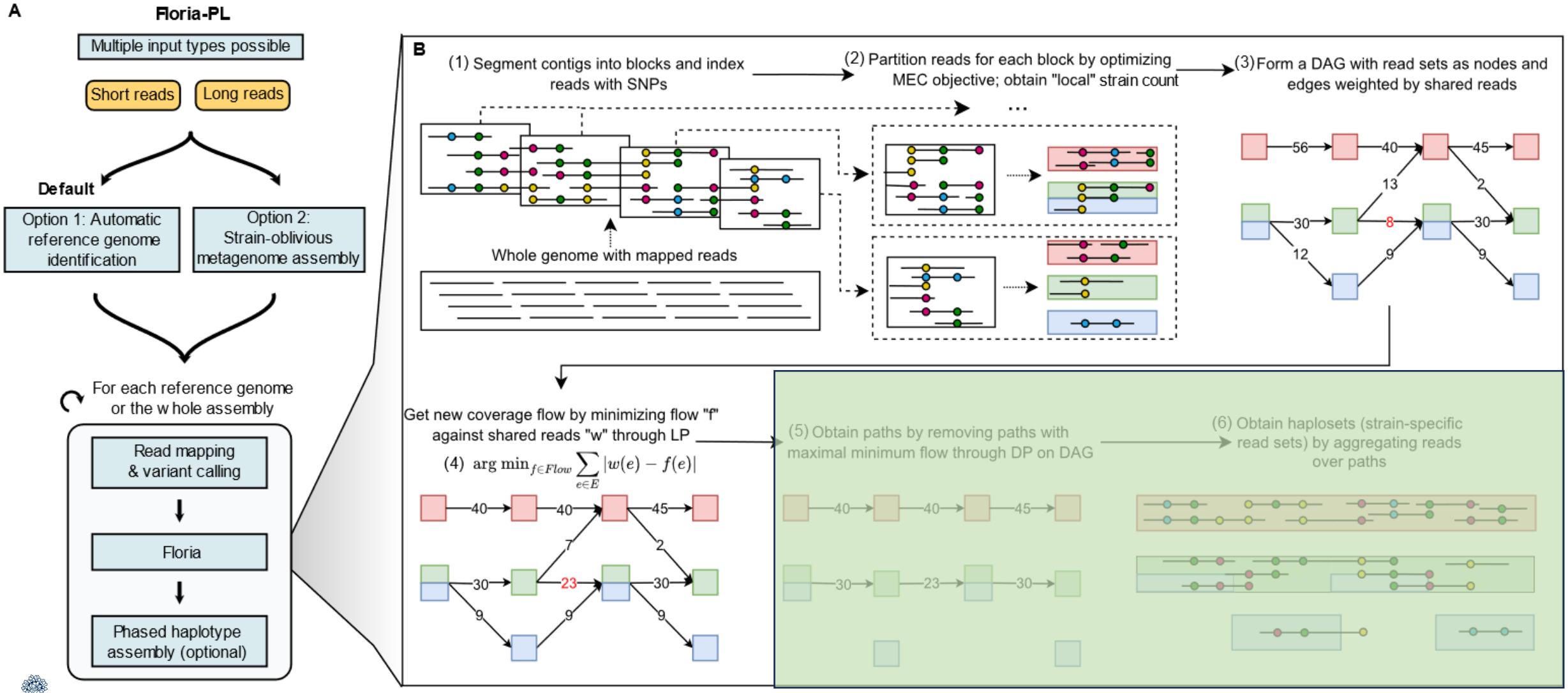
Error!
bad edge

By *linear programming:*

$$\arg\min_{f \in Flow} \sum_{e \in E} |w(e) - f(e)|$$

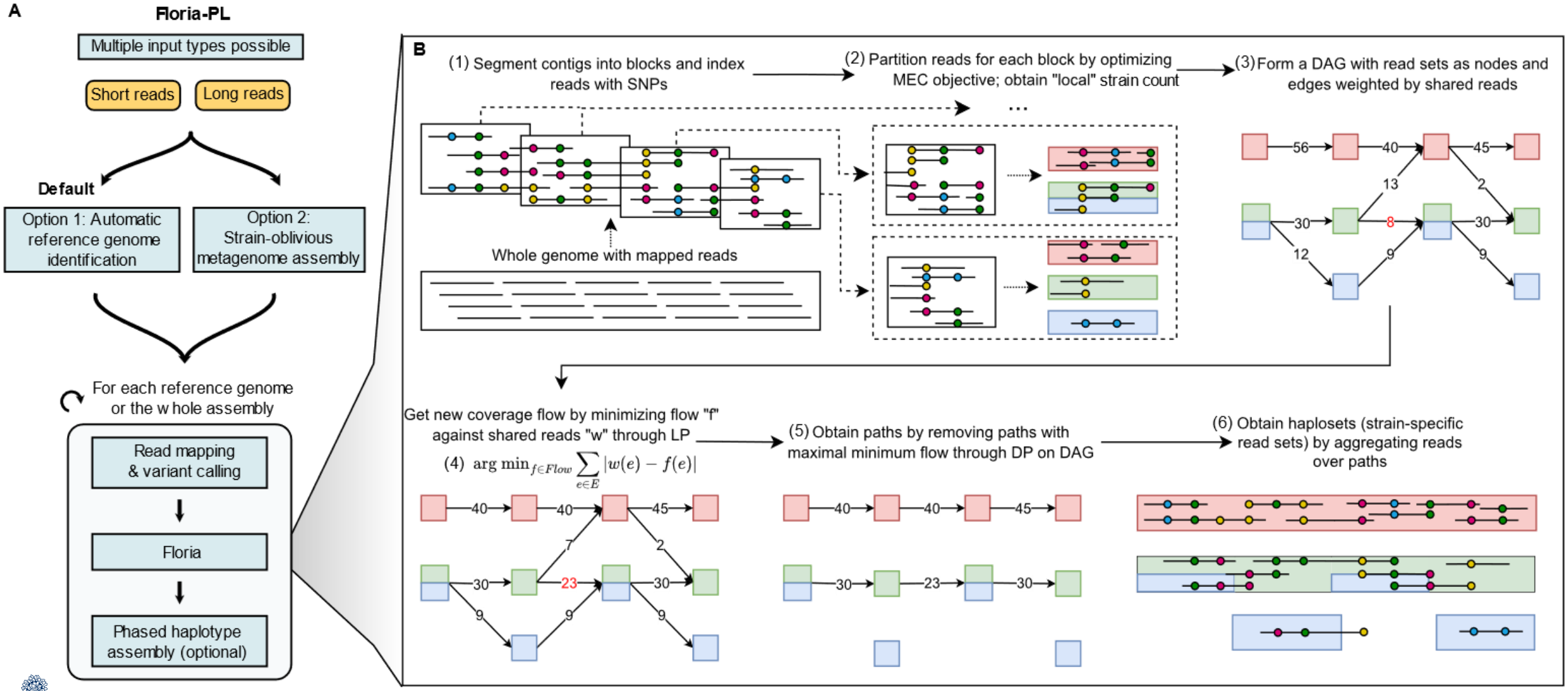
f – network flow
w – original weights

Step 2: **network flows**



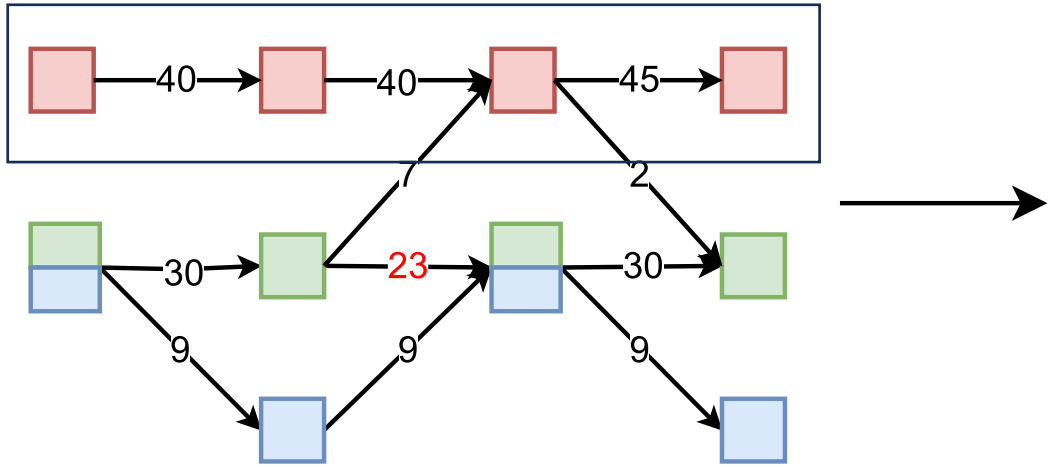
UNIVERSITY OF TORONTO

Step 3: Obtaining haploset paths

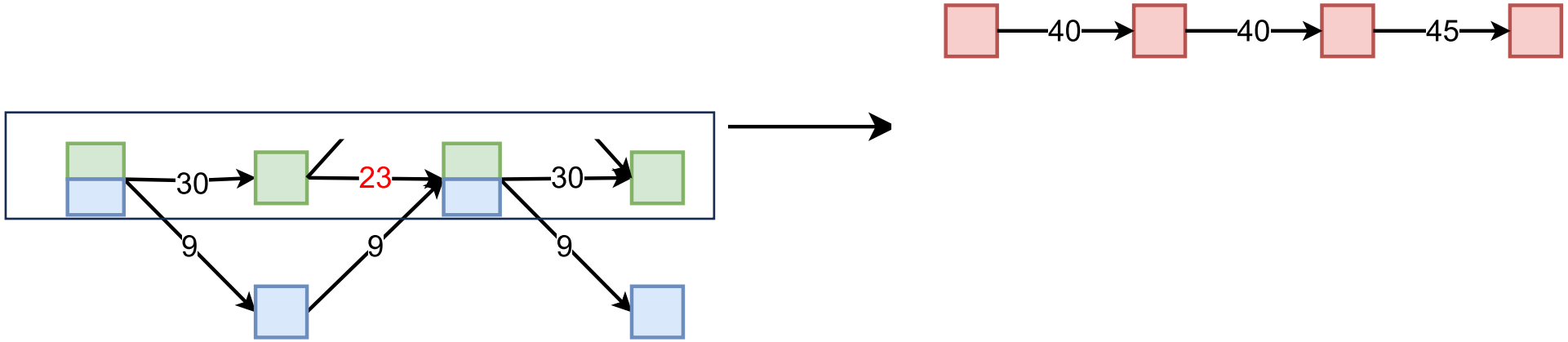


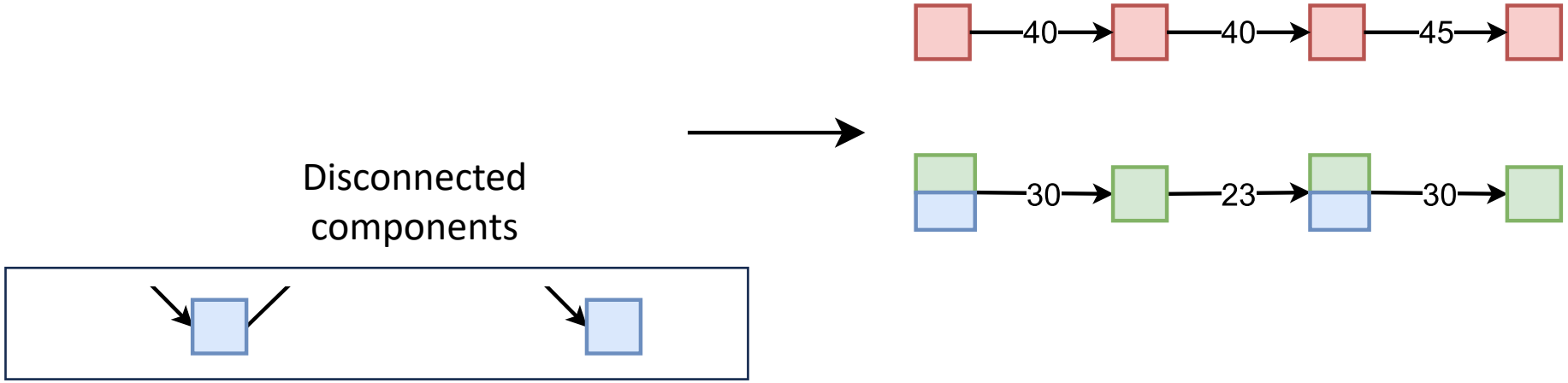
UNIVERSITY OF TORONTO

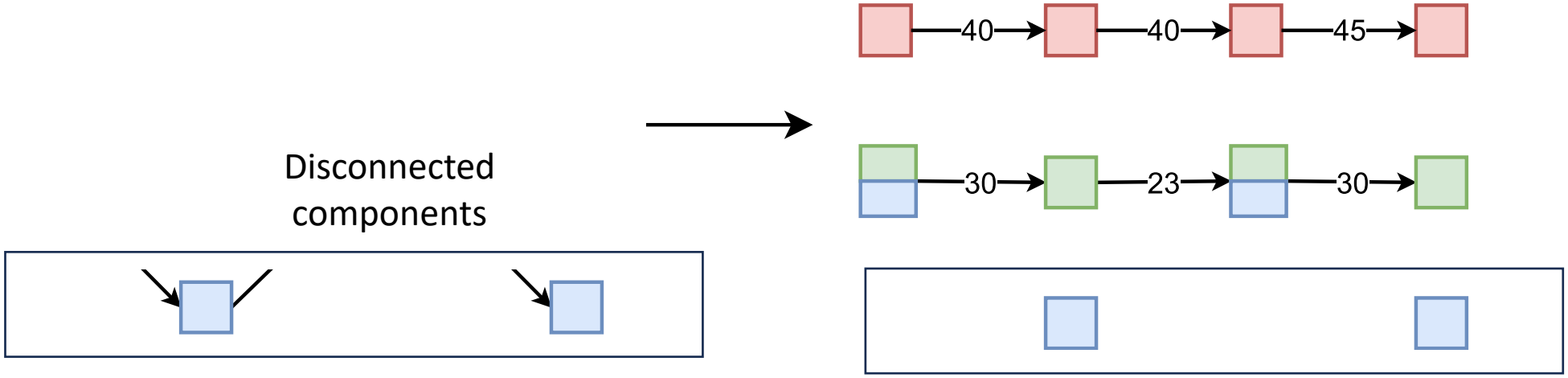
Largest *minimum* flow path
(via dynamic programming - DP)



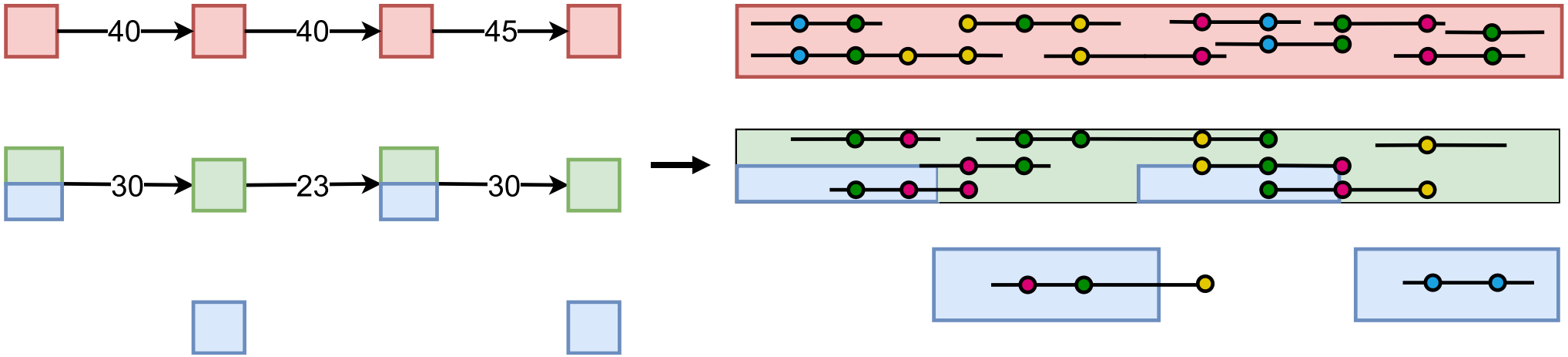
Largest *minimum* flow path
(via dynamic programming - DP)



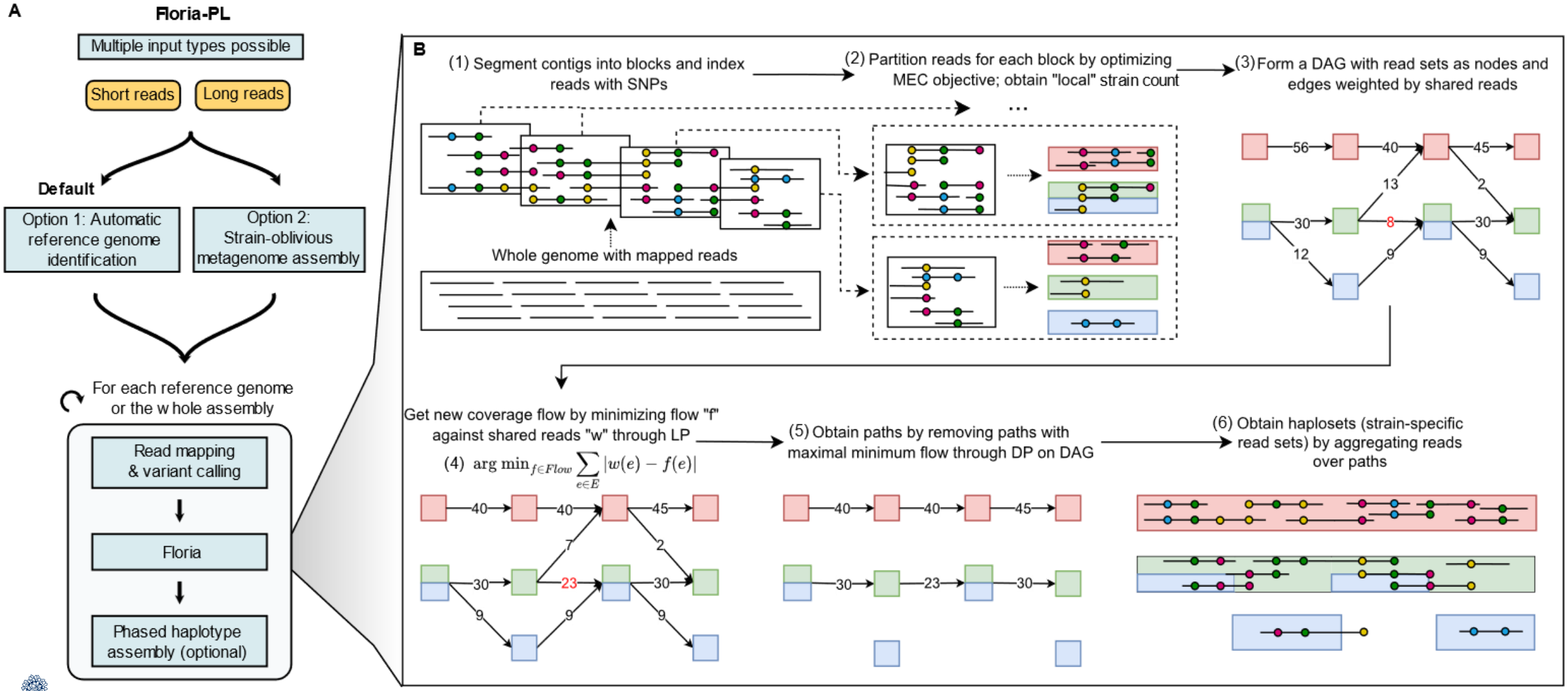

Disconnected
components
40
40
45
30
23
30


Disconnected
components
40
40
45
30
23
30

(6) Obtain haplosets (strain-specific read sets)
by aggregating reads over paths



Step 3: trim flow graph to obtain haplosets



UNIVERSITY OF TORONTO

Benchmarking results

Benchmarking: simulated metagenome

Synthetic metagenomes:

- 40 common gut species
- 1-5 strain genomes per species
- Synthetic noisy nanopore reads **(88% identity)**
- Random strain coverages between [5,25]

Comparison against:

1. metaFlye – metagenome assembler (Kolmogorov et al. 2020)
2. Strainberry – strain assembler (Vicedomini et al. 2021)

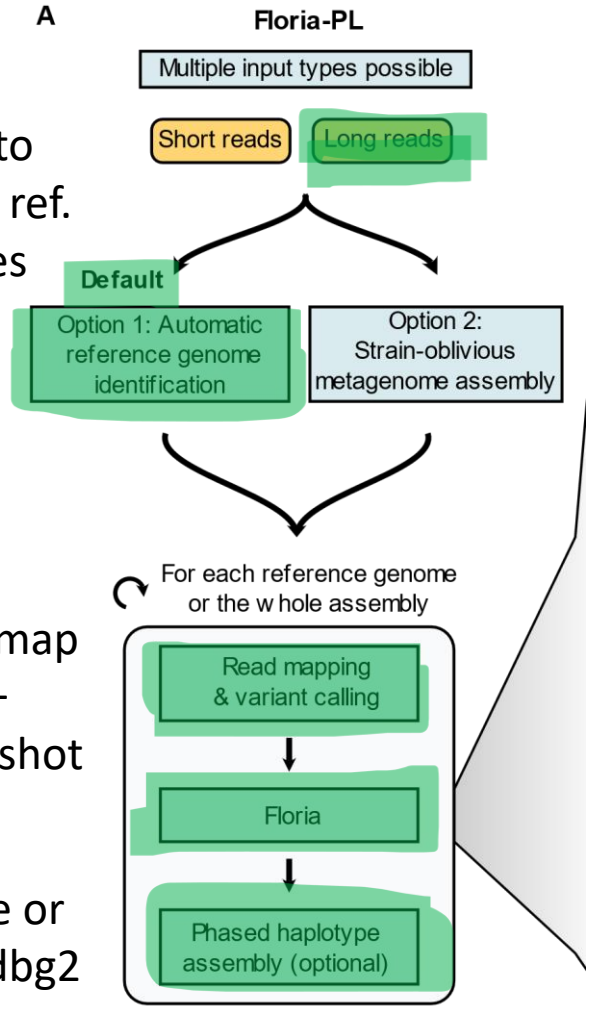
UNIVERSITY OF TORONTO

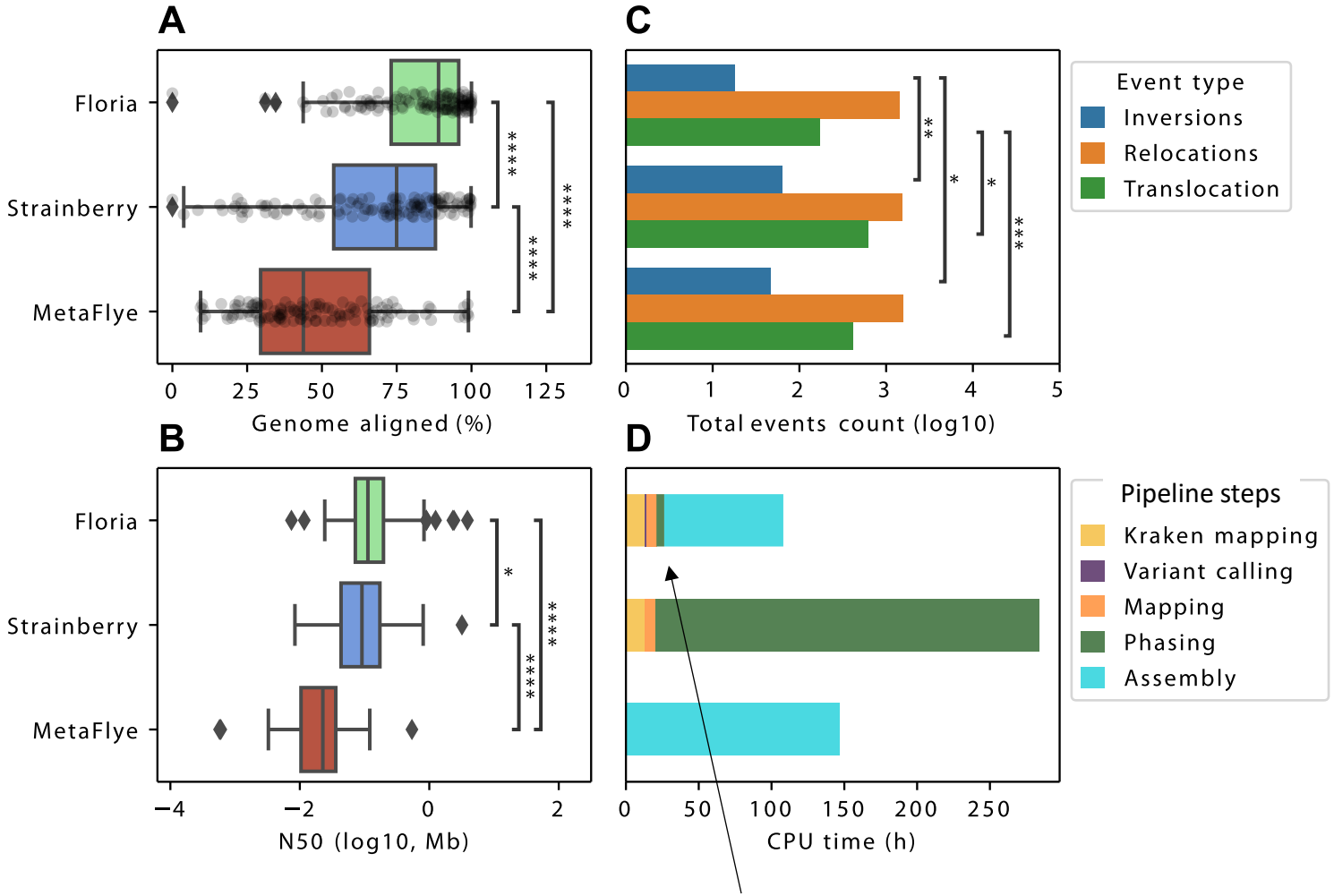
Assembly benchmarking

Kraken to identify ref. genomes

Minimap + Longshot

Flye or wtdbg2





Floria **phasing** time << **assembly** time!

UNIVERSITY OF TORONTO

Results: **real** metagenomes

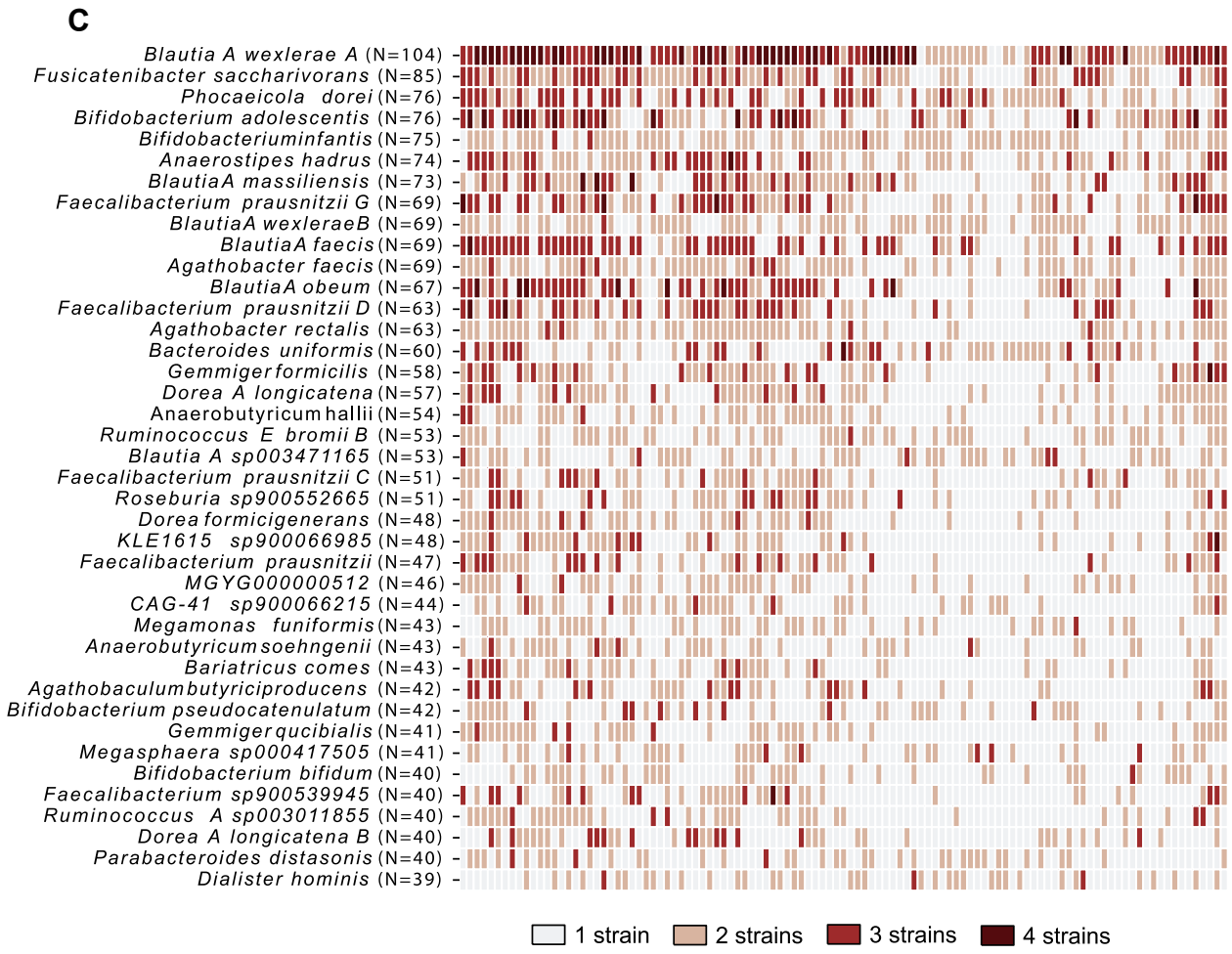
Long read **AND** short read

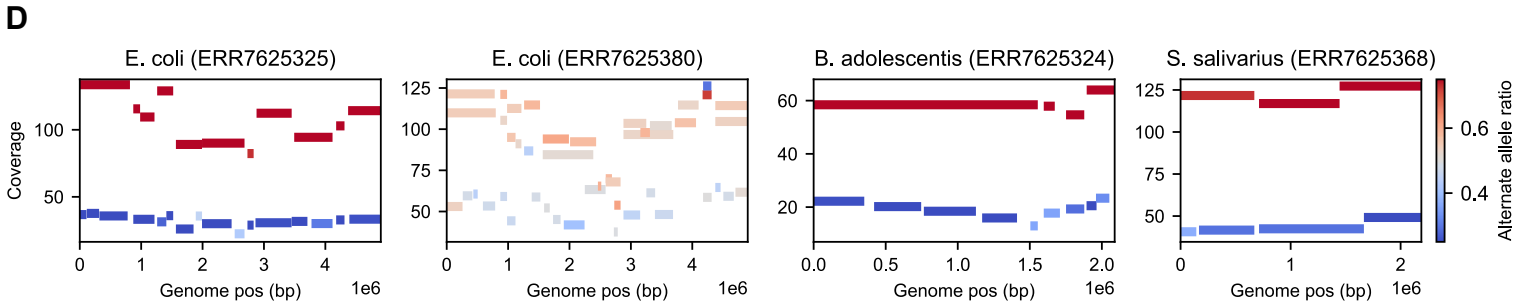
UNIVERSITY OF TORONTO

Floria on 109 gut nanopore samples!

(dataset: Gounot et al. 2022, Nat. Comms.)

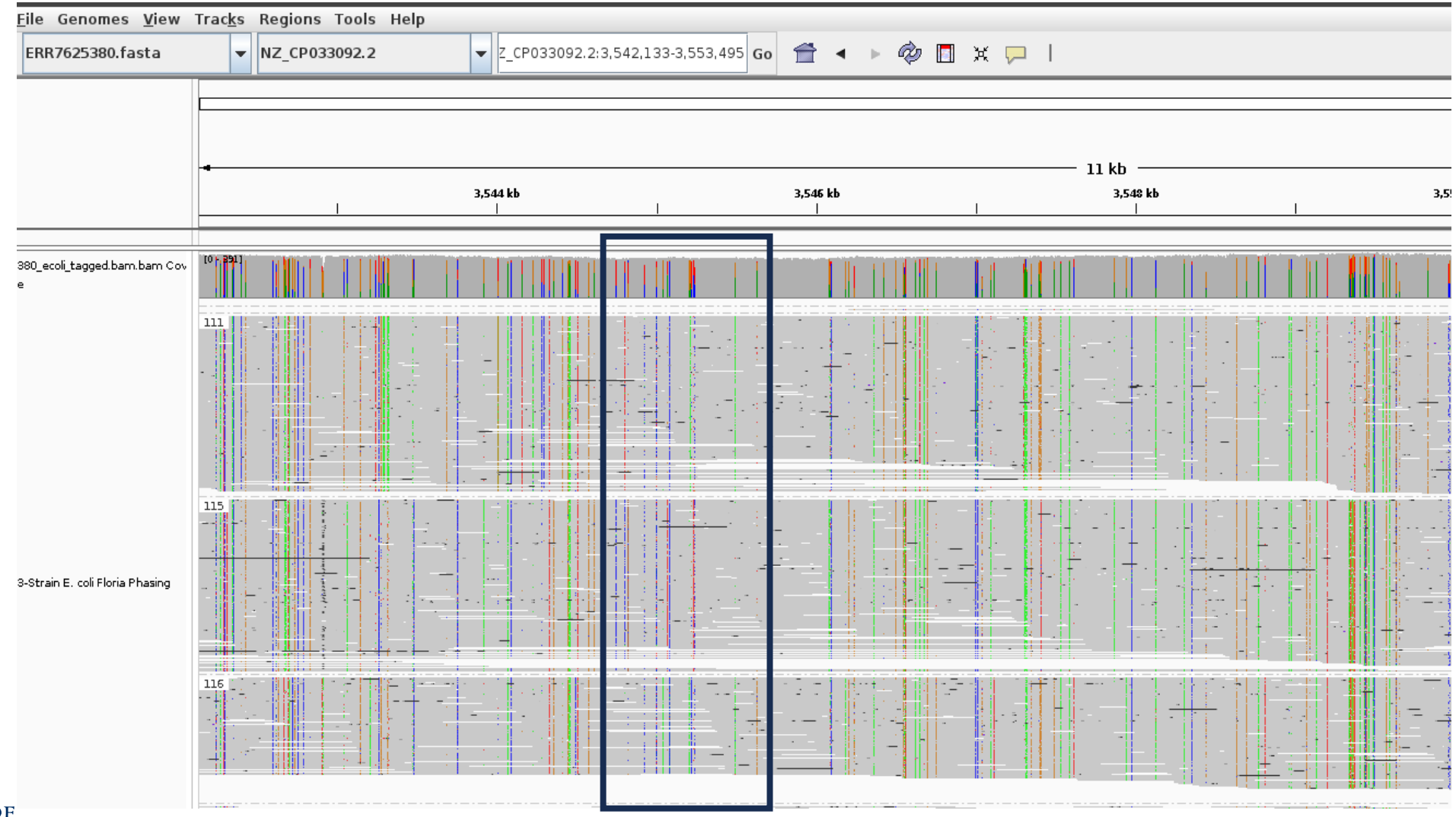
< 15 mins per sample for phasing!





UNIVERSITY OF TORONTO

Floria allows for **visualization** – 3-strain *E. coli*



Consistently different alleles

Longitudinal strain tracking with floria

- 24 longitudinal **SHORT-READ gut samples (636 days)**
  - Dataset from *"**Metabolic independence drives gut microbial colonization and resilience in health and disease**"* by Watson et al. (Genome Biology 2023)
- Run floria → obtain haplosets → track across time

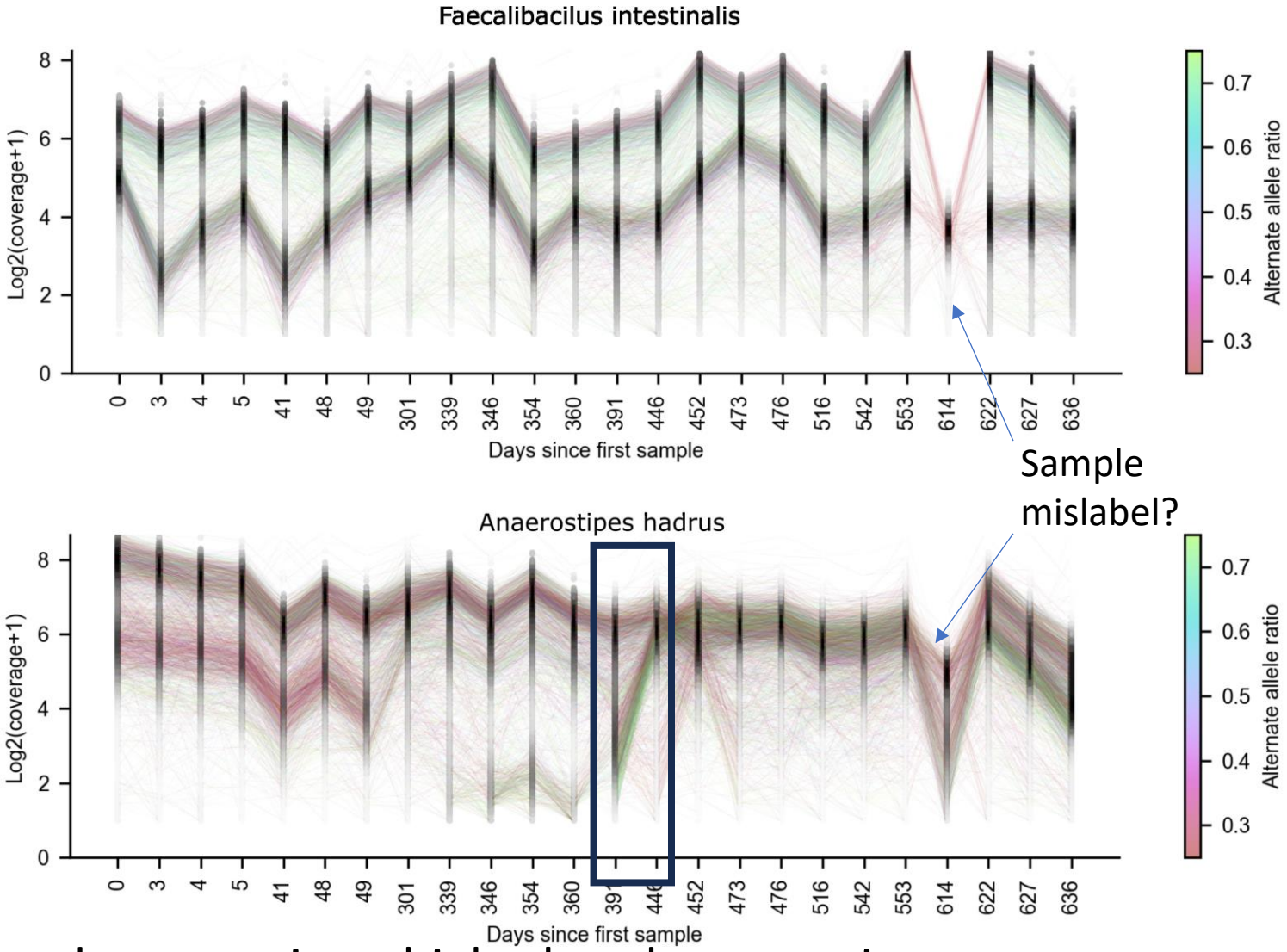
UNIVERSITY OF TORONTO

Line: haploset match across sample

Two species with interesting patterns:

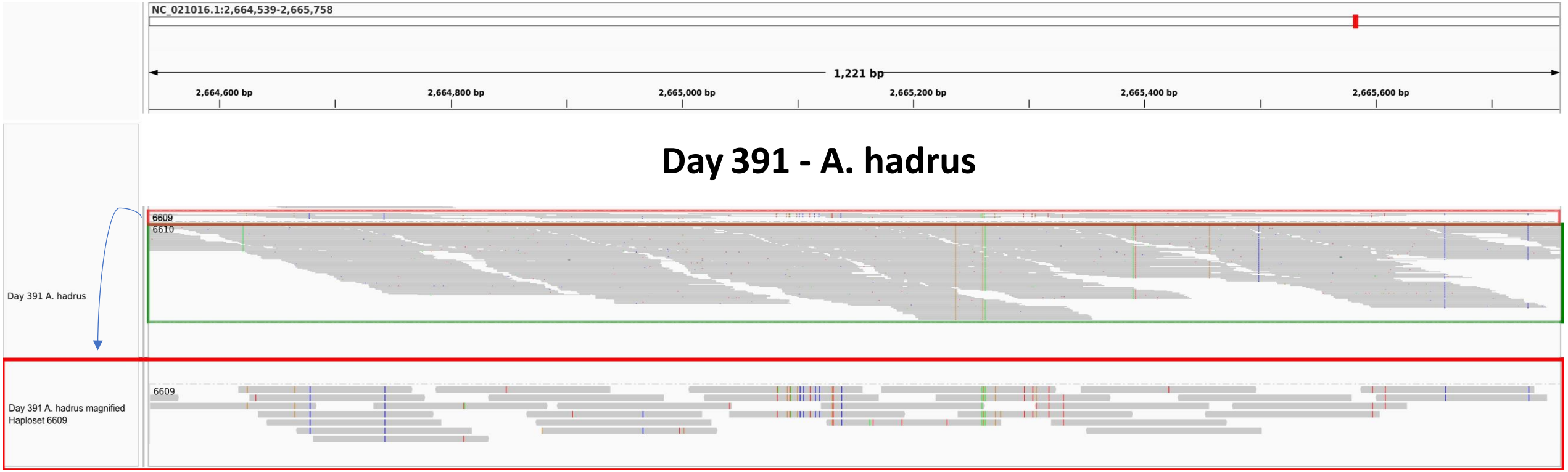
F. intestinalis - ***stable***

A. hadrus – ***transient***



Low-abundance strain to high-abundance strain emergence? **Visualize!**

UNIVERSITY OF TORONTO

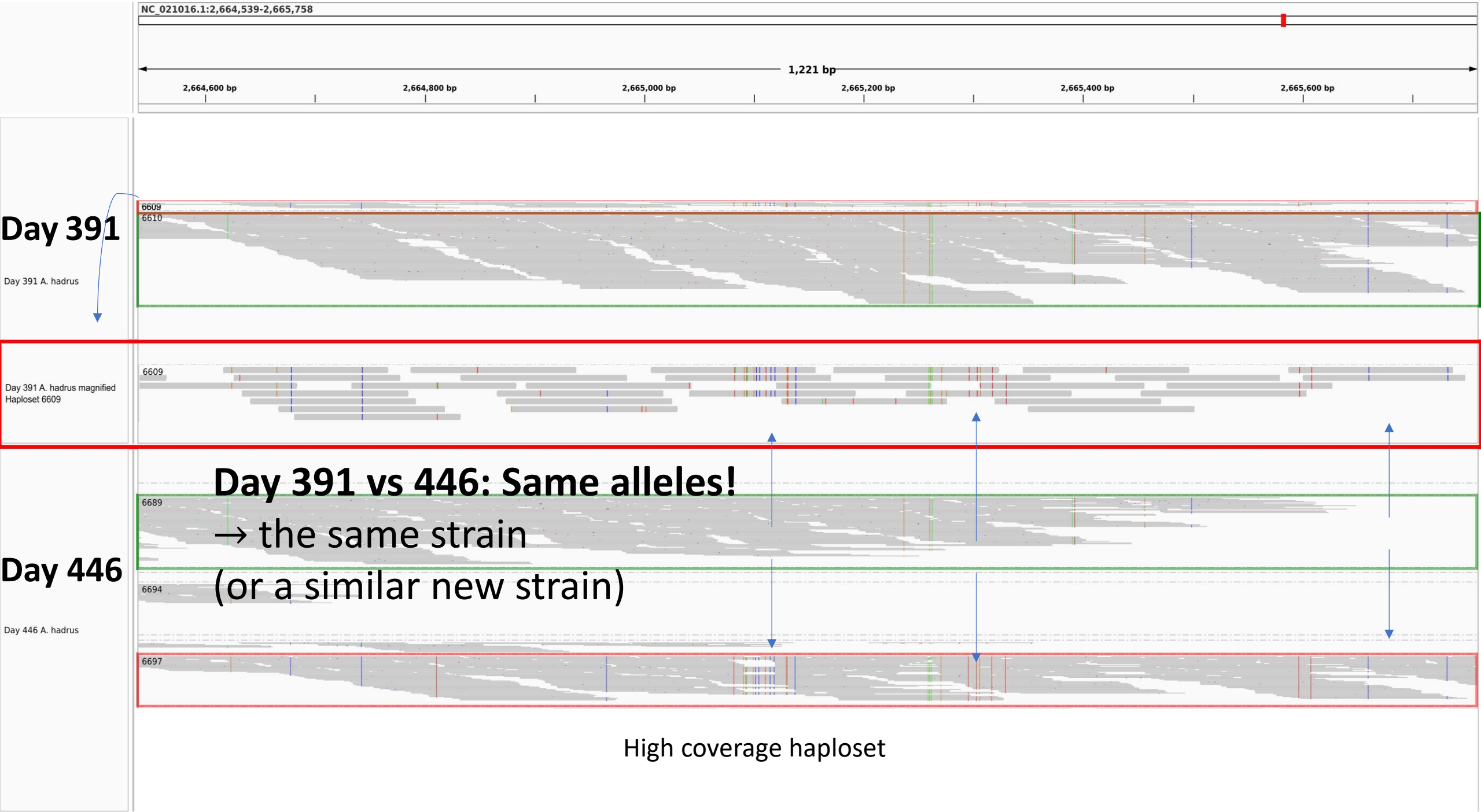


Day 391 - A. hadrus

Zoomed in on red haploset ("6609"):

LOW coverage

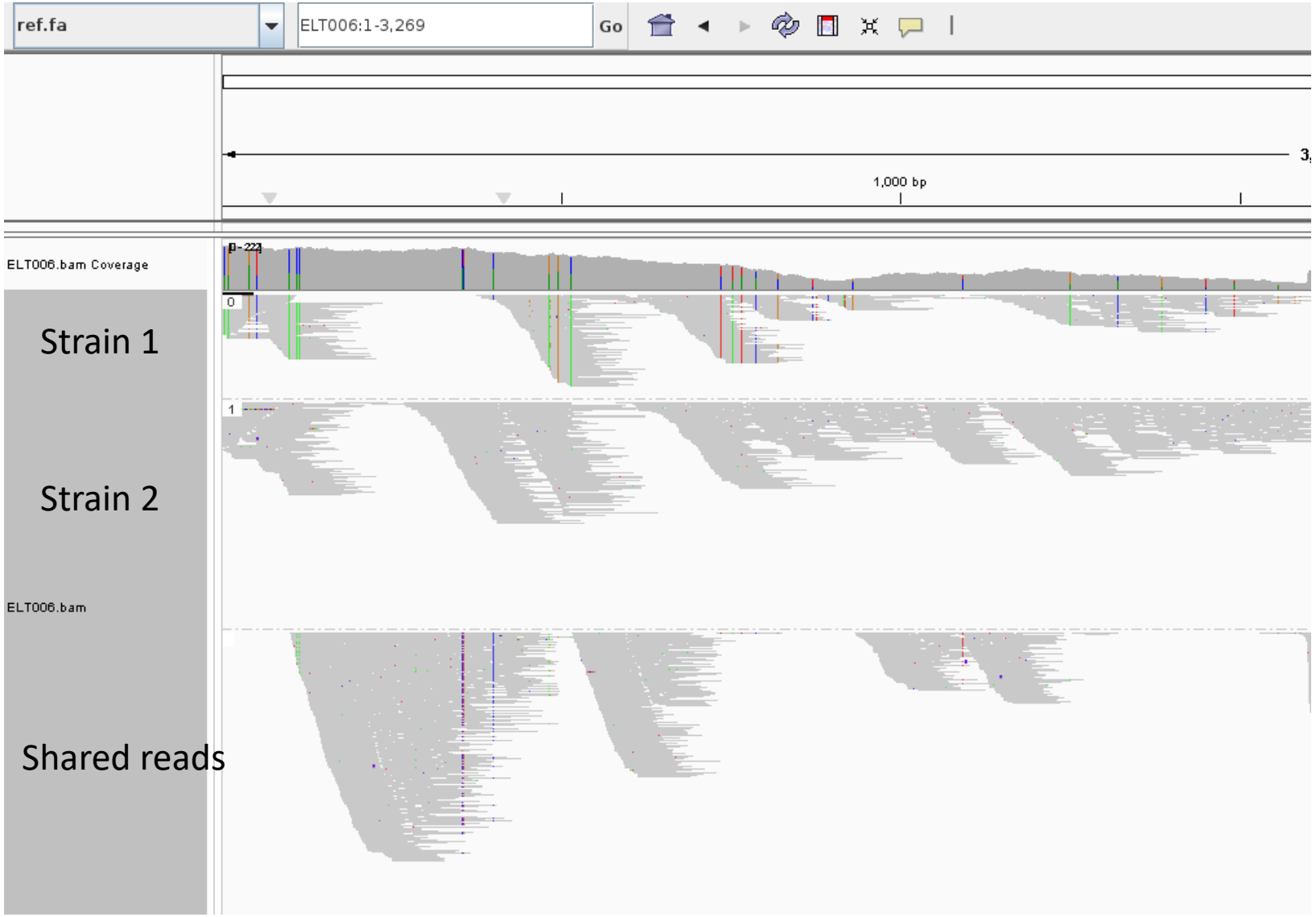
UNIVERSITY OF TORONTO


NC_021016.1:2,664,539-2,665,758
1,221 bp
2,664,600 bp
2,664,800 bp
2,665,000 bp
2,665,200 bp
2,665,400 bp
2,665,600 bp
Day 391
Day 391 A. hadrus
6609
6610
Day 391 A. hadrus magnified
Haploset 6609
6609
Day 391 vs 446: Same alleles!
→ the same strain
(or a similar new strain)
Day 446
Day 446 A. hadrus
6689
6694
6697
High coverage haploset

Floria works on **ancient viral metagenomes!?**

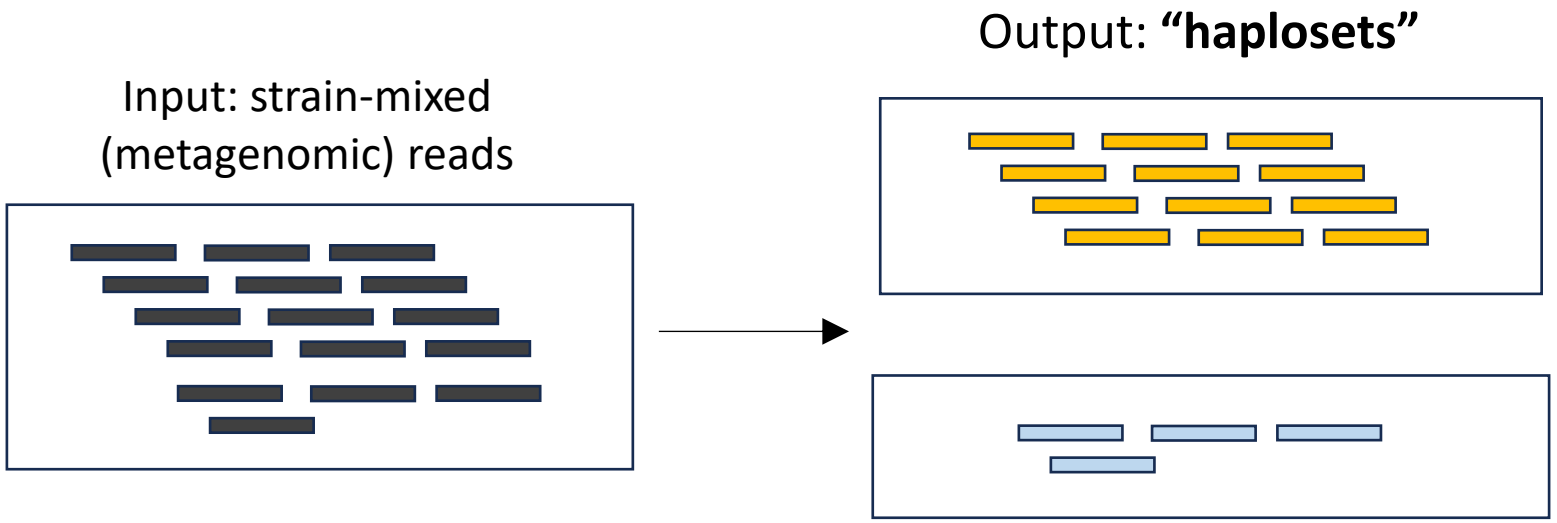
From Maxime Borry (Postdoc at Max Planck for Evo Anthro):

Ancient mixed infection of hepatitis B (Kocher et al., 2021, Science)



UNIVERSITY OF TORONTO

Conclusion



- Developed **floria**: a strain-level read clustering (phasing) algorithm
  - Short OR long reads
  - < 20 min per metagenome
  - assembly optional
- Fast, informative, and versatile
  - can do strain-level metagenomic assembly…
  - but even more useful for **hypothesis generation and data sleuthing**

UNIVERSITY OF TORONTO

Acknowledgements + authors

- Jim Shaw (**presenter**)
  - PhD at University of Toronto (2024)
  - incoming postdoc at Harvard Medical School / DFCI
- Jean-Sébastien Gounot
  - Research fellow at Genome Institute of Singapore
  - **co-lead author**

- Hanrong Chen
  - Postdoc at Genome Institute of Singapore

- Niranjan Nagarajan
  - Genome Institute of Singapore**, co-lead PI**

- Yun William Yu
  - Carnegie Mellon University, **co-lead PI**

UNIVERSITY OF TORONTO

Floria: fast and accurate strain haplotyping in metagenomes

Jim Shaw, Jean-Sebastien Gounot, Hanrong Chen, Niranjan Nagarajan ✉, Yun William Yu ✉ Author Notes

Bioinformatics, Volume 40, Issue Supplement_1, July 2024, Pages i30–i38, https://doi.org/10.1093/bioinformatics/btae252

Published: 28 June 2024

Agency for Science, Technology and Research SINGAPORE

Natural Sciences and Engineering Research Council of Canada | Conseil de recherches en sciences naturelles et en génie du Canada

Canada

Supported by NSERC CGS-D

Thanks to Maxime Borry for ancient DNA results