# Copy number estimation for de Bruijn graph *de novo* whole-genome shotgun assembly contigs

Yee Fay Lim
May 27, 2021

# Abstract

- ❏ High-throughput short reads widely used for reference-free whole genome assembly

- ❏ Multiplicity information for partly assembled sequences improves assembly quality

- ❏ However, a principled, general solution for multiplicity estimation had been unavailable

- ❏ Introduce novel, versatile copy number estimator for haploid & diploid genomes

- ❏ Reliable accuracy at resolving multiplicities up to a low maximum
  - ❏ Over a range of genome, sequencing, and assembly conditions

- ❏ Outperforms & far more versatile than closest alternatives

# Background and vocabulary

- ❏ Sequencing
  - ❏ **(high-throughput) short-read:** vs. long-read sequencing
  - ❏ **shotgun**: duplication and random fragmentation of genomic segments to generate overlapping reads
- ❏ Assembly
  - ❏ *de novo*: i.e. reference-free, vs. reference-based
  - ❏ **de Bruijn graph (DBG)**
- ❏ **contig**: partially assembled sequence
- ❏ **k-mer**: sequence of k characters (bases)
- ❏ **heterozygosity**: difference between homologous chromosomes at the same locus

# De Bruijn graph (DBG) assembly

❏ Dominant paradigm for *de novo* whole-genome high-throughput short read assembly

❏ 3 stages

  ❏ **k-mer de Bruijn graph construction**
  ❏ **Contig building**
  ❏ Scaffolding

# k-mer DBG construction

❏ Extract all distinct k-mers in read dataset

❏ Represent each distinct k-mer with a node

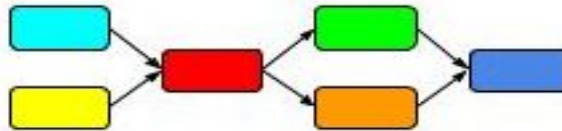❏ Create a directed edge between each pair of nodes with a k-1 suffix-prefix overlap
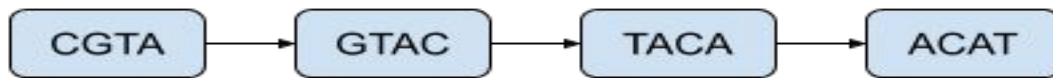


k-mer graph of the sequence "CGTACAT", with k = 4

# Contig building

❏ Nodes are merged along "linear" (non-forking) paths in the DBG
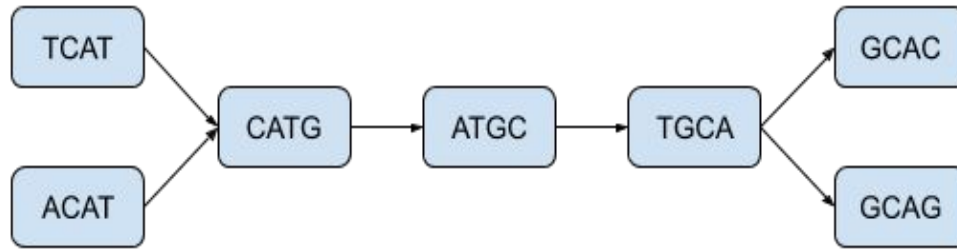
❏ **Contig**: the resulting string for each path
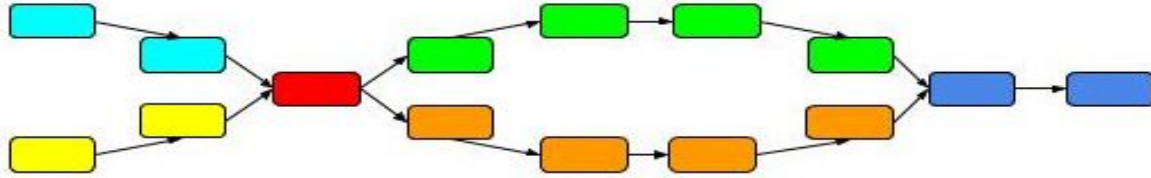
k-mer graph (k-mers omitted): Linear paths, each corresponding to a contig, colour coded

Contig graph: Linear path nodes merged into contigs; each node represents a contig.

# Repeat resolution

❏ The core difficulty in de novo whole-genome assembly

❏ Which of the 4 possible paths here are correct???

# Motivation

- ❏ Copy number with graph topology and long-range information can resolve repeats
  - ❏ i.e. resolve ambiguity in correct paths through any node with in- and out-degrees > 1

- ❏ Resulting in improvements to:
  - ❏ Genome coverage and size estimate accuracy
  - ❏ Structural and base-level correctness
  - ❏ Contiguity

# Repeat resolution: copy number & graph topology

❏ ABC has sufficient long-range support, but DBE doesn't
❏ But B has "remaining" copy number 1
❏ Gives unambiguous support to DBE based on graph topology

A: 1x    B: 2x    C: 1x

D: 1x             E: 1x
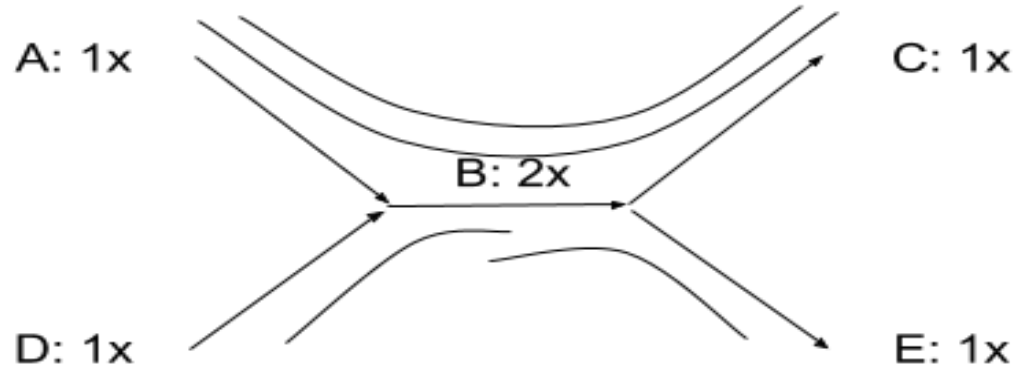
# Repeat resolution affects genome coverage
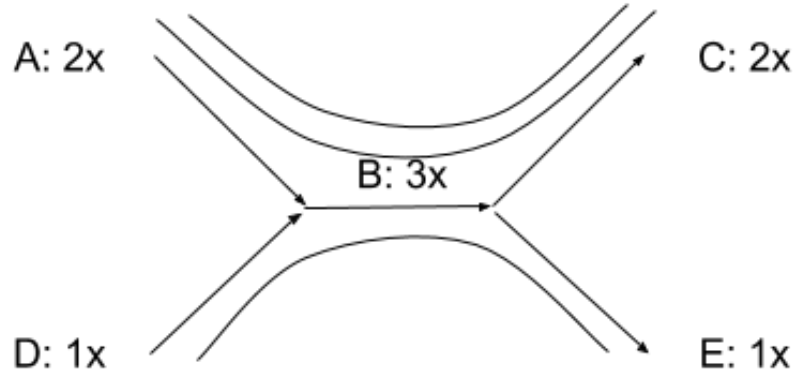


**Resolved**
Paths: ABC x2, DBE x1
Total path length:
$2L_A + 3L_B + 2L_C + L_D + L_E$

**Unresolved**
Paths: AB, DB, C, E
Total path length:
$L_A + 2L_B + L_C + L_D + L_E$
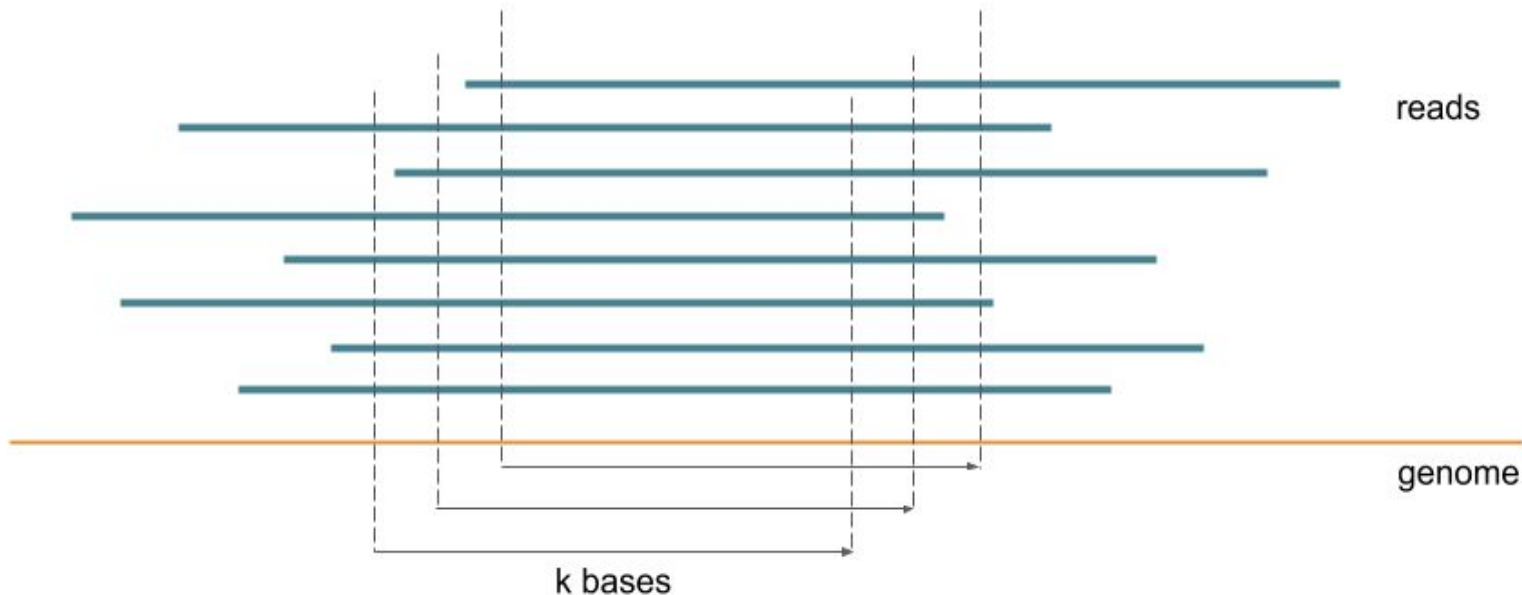Missing path length:
$L_A + L_B + L_C$

A: 2x

B: 3x

C: 2x

D: 1x

E: 1x

# Read coverage / depth

❏ The number of reads covering a given base or bases, e.g. a k-mer
❏ **Contig mean k-mer read depth**: mean # reads covering k-mers in a contig

# Copy number accounting

- Any given sequence can occur in a genome 0, 1, 2, 3, 4, ... times
- Copy number in haploid genome: equals # occurrences
- Copy number in diploid genome: -> 0, 0.5, 1, 1.5, 2, ...
  - Only consider 0, 0.5, 1, 2, ...

# Literature gap

- Accurate, statistically informed copy number estimation in *de novo* WGS assembly

- *Unicycler*: naive usage of read depth in graph-based copy number propagation

- *GenomeScope*: mixture model of k-mer frequency as function of read depth & copy number ($\leq 2$)

- *Magnolya*: Poisson mixture model of read counts for long OLC contigs

- Shortcomings
  - Dependence on a narrow range of sequencing coverage (*Unicycler*)
  - Inflexible: do not accommodate arbitrary contig lengths (*GenomeScope*, *Magnolya*)

# Methods

- Input: Contig sequences, lengths, mean k-mer read depths (coverage)
- **Repeat prevalence decreases rapidly as length increases**
- Partition contigs by length
  - Depth distributions differ across partitions: fit separate model on each
  - Flexible to differing length and depth joint distributions

# Contig mean k-mer depth across length strata
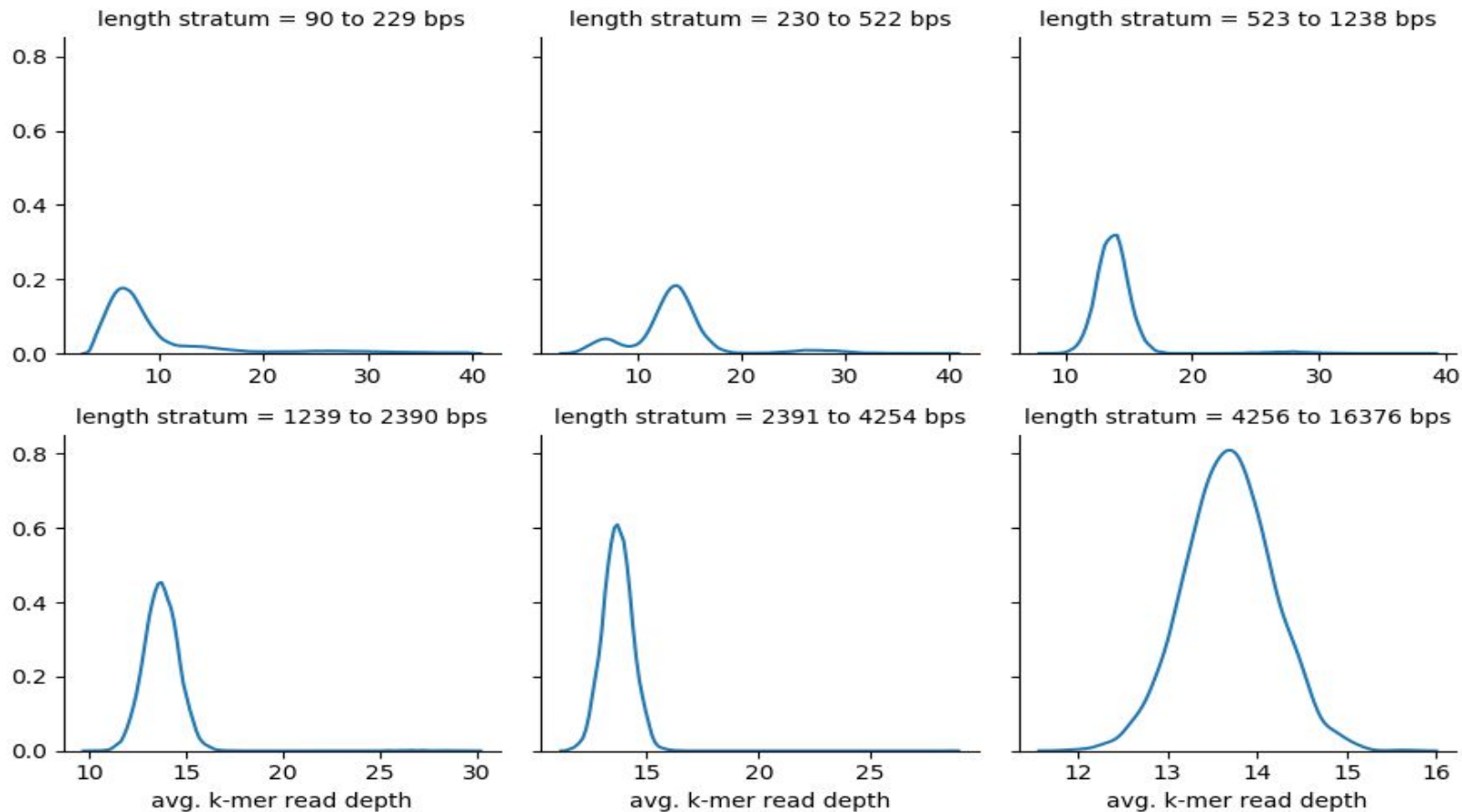
# Model formulation

- ❏ Mean k-mer read depth of sufficiently long contigs is approximately Gaussian

- ❏ Use for all contig lengths due to consistency, simplicity, and precedent

- ❏ $\mu$: mean k-mer read depth of copy # 1 contigs across entire dataset

- ❏ Copy # 1 depth variance, $\sigma^2$, specific to each length stratum

- ❏ Within stratum, copy number i mean and variance are $i\mu$ and $i\sigma^2$

- ❏ Where applicable:
  - ❏ High copy numbers modelled in aggregate using gamma distribution
  - ❏ Copy number 0 (spurious) contigs modelled using exponential distribution

# Model statement

❏ Within-stratum contig frequency, given mean k-mer depth:

$$f(x; \Theta) = G \cdot \{I_0 w_0 f_{exp}(x; r) + I_{0.5} w_{0.5} f_g(x; 0.5\mu, 0.5\sigma^2) + \sum_{i=1}^{c} w_i f_g(x; i\mu, i\sigma^2) + I_{c+1} w_{c+1} f_\gamma(x; \lambda, s, \theta)\}$$

Where

$\Theta \equiv \{r, \mu, \sigma^2, \lambda, s, \theta, I_0, I_{0.5}, I_\gamma, G, w\}$ is the parameter set

$f_{exp}, f_g, f_\gamma$ are the density functions of the exponential, Gaussian, and gamma distributions

$r$ is the rate parameter of the exponential distribution

$c$ is the highest copy number represented by a Gaussian in the stratum

$\lambda, s, \theta$ are gamma distribution location, shape, and scale parameters

$I_0, I_{0.5}$ are indicator variables for copy numbers 0 and 0.5 components

$I_{c+1}$ is an indicator variable for the component aggregating all copy numbers > c

$W$ is the set of component weights

# Model implementation

❏ Longest contigs (almost) exclusively single-copy (copy # 0.5 or 1)

❏ Empirical depth density with peak very near or at copy # 0.5 or 1 mean
  ❏ 0.5 for sufficiently heterozygous diploid data; 1 otherwise
  ❏ Correct copy # specified by user or chosen using model fit with lower AIC

❏ First fit model on longest contig stratum for precise estimate of $\mu$

❏ Iterate over strata in decreasing order of length, constraining $\mu$

❏ Each contig assigned copy number with highest estimated density given its depth value

# Copy number assignment boundaries
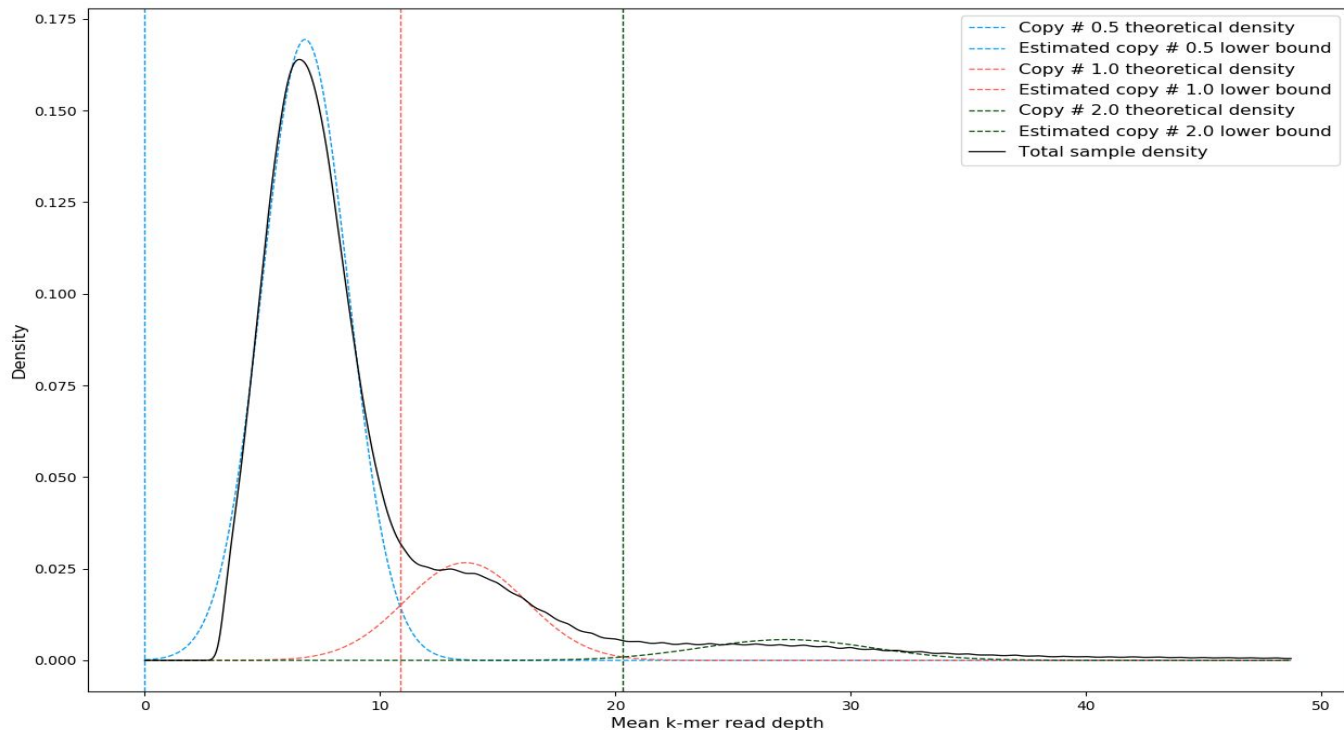
# Workflow

For contigs in each length stratum:

**Preprocess**: Compute exact model (copy numbers & probability distributions) to fit

**Fit**: contig frequency $f(x; \Theta)$ , where $x \equiv$ mean contig k-mer read depth, $\Theta \equiv$ model parameters

**Post-process**: To each contig, assign most probable copy number implied by $\Theta$

# Experiments: summary

| Data type | Species | Heterozygosity rate, approx. (%) | Mean per-base read coverage | *k* values |
|---|---|---|---|---|
| Simulated | *E. coli* | N/A | 30 | 60, 65, …, 115, 120 |
| | | | 50 | 65, 70, …, 120, 125 |
| | *C. elegans* | 0.0609 | 30 | 60, 65, …, 115, 120 |
| | | | 50 | 65, 70, …, 120, 125 |
| | | 0.2430 | 30 | 60, 65, …, 115, 120 |
| | | | 50 | 65, 70, …, 120, 125 |
| | | 1.1726 | 50 | 65, 70, …, 120, 125 |
| | *P. trichocarpa* | 0.0580 | 50 | 65, 70, …, 120, 125 |
| | | 0.2338 | | |
| | | 0.5852 | | |
| | *H. sapiens* | 0.0563 | 50 | 65, 70, …, 120, 125 |
| | | 0.1148 | | |
| Real | *E. coli* | N/A | 989 | 40, 45, …, 85, 90 |

# Experiments: details and comparisons

❏   2 types of classification performance evaluated:
   ❏   Binary (1-many), multi-class (copy numbers [0, 0.5,] 1, 2+)

❏   Performance evaluated against:
   ❏   GenomeScope 2.0 for diploid genomes
   ❏   Unicycler for *E. coli* at odd *k* values

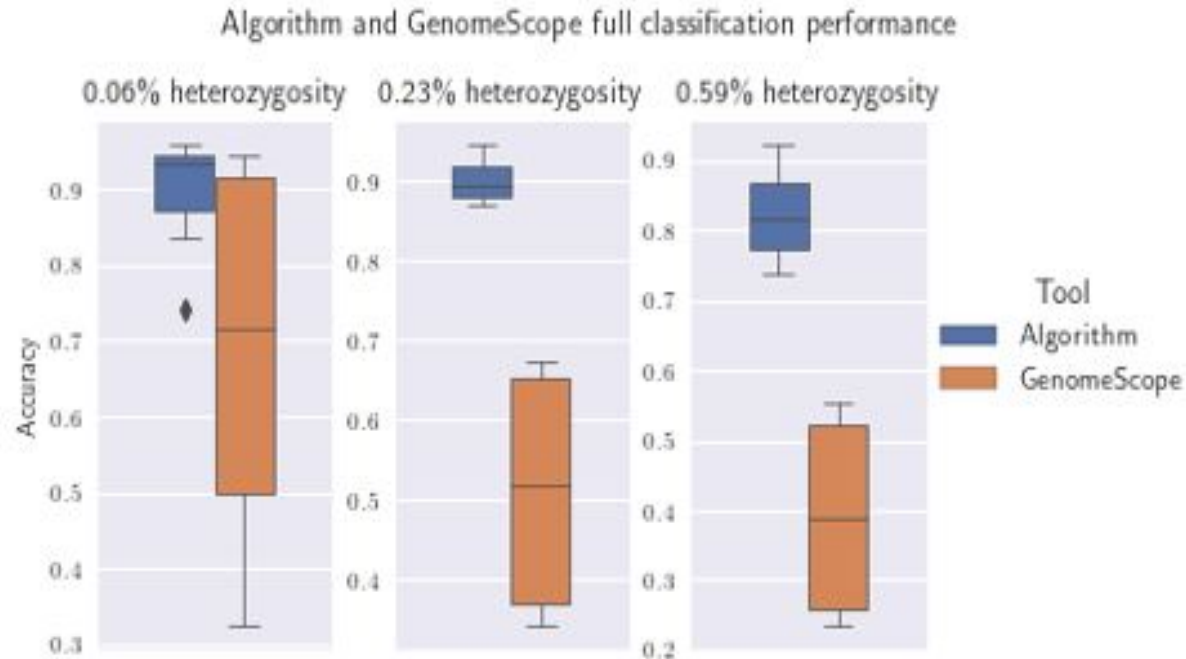# Results example 1: 1-many classification

❏   *P. trichocarpa*, simulated reads, across *k* values

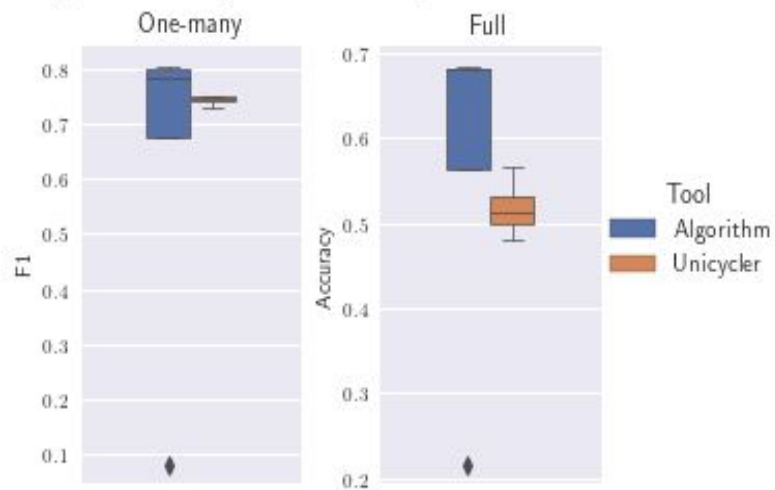❏   Note: GenomeScope sometimes doesn't converge, resulting in fewer data points

Algorithm and GenomeScope one-many classification performance
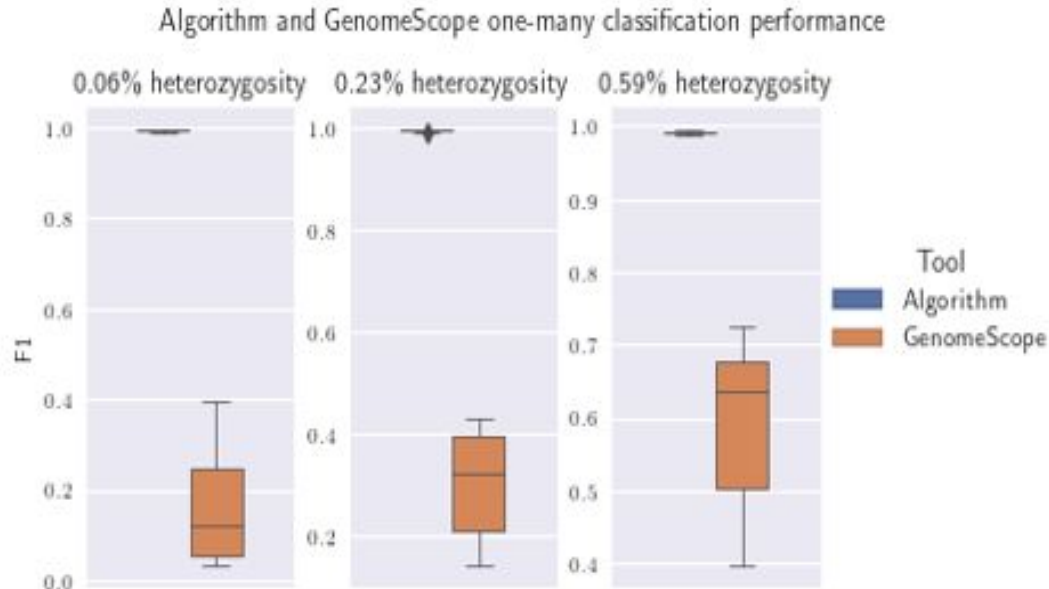
# Results example 2: multiclass classification

❏ *P. trichocarpa*, simulated reads, across *k* values



Algorithm and GenomeScope full classification performance

# Results example 3: real *E. coli* data
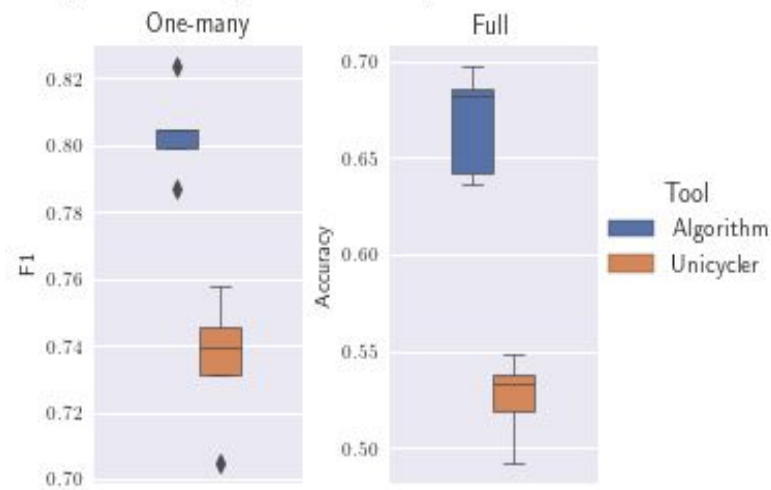
# Discussion

❏ Performs reliably across a range of conditions
- ❏ Genome characteristics: heterozygosity level, repeat content
- ❏ Sequencing and assembly characteristics: read coverage, $k$ value

❏ Outperforms closest existing alternatives, often by wide margin

❏ Performs better at 1-many than multiclass classification

❏ Shows that versatile, reliable DBG contig copy # estimation feasible

# Limitations and further work

- ❏ Performs poorly on copy number 0: depth & length data inadequate
  - ❏ Spurious contigs arise from several error reads having high overlap with correct reads
- ❏ Simulated data experiments may overstate performance on real data
  - ❏ Real data depth distributions depart from theoretical assumptions underlying algorithm
- ❏ Suggests fundamental limitations to using only length and depth information
- ❏ Different data, maybe also model, needed for substantial improvement
  - ❏ E.g. one or more measures of GC content