CS4330: Combinatorial Methods in Bioinformatics

Hybrid genome assembly

Wong Limsoon

Based on the PhD thesis of Joshua Casey Darian

NUS
National University of Singapore
National University of Singapore

Bauhinia blakeana Aka Hong Kong orchid 洋紫荊

A hybrid between Bauhinia variegata and Bauhinia purpurea

Exercise

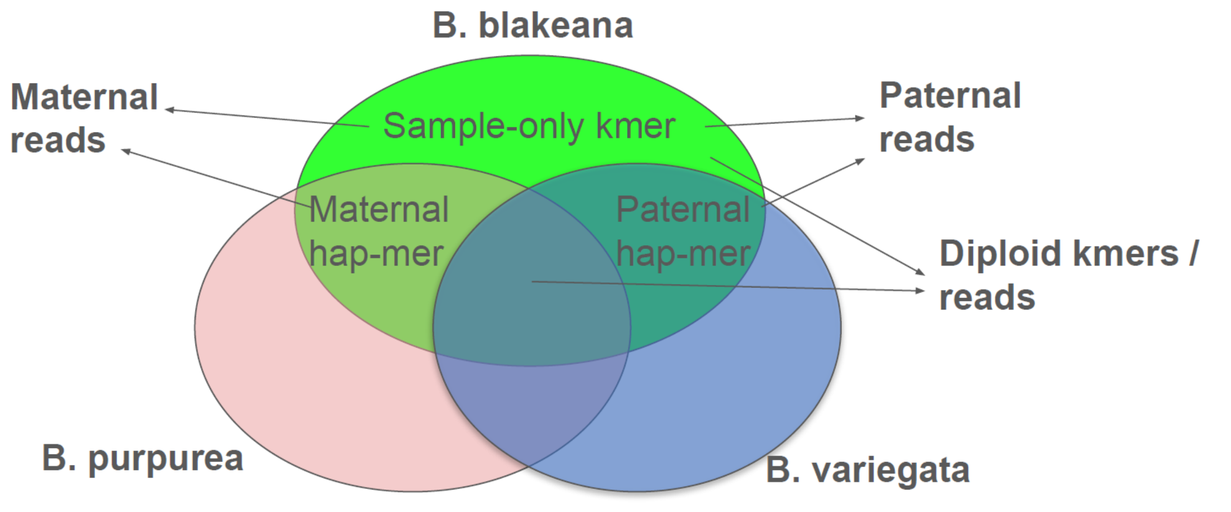
It is very difficult to perform genome assembly of a hybrid like the Hong Kong orchid

Why?

How to assemble hybrid genomes

Assemble all the reads together?

Split into paternal and maternal reads and assembly the two sets of reads separately?



Exercise

What will likely happen if we try to assemble all the reads together during hybrid genome assembly?

Hybrid genome assembly

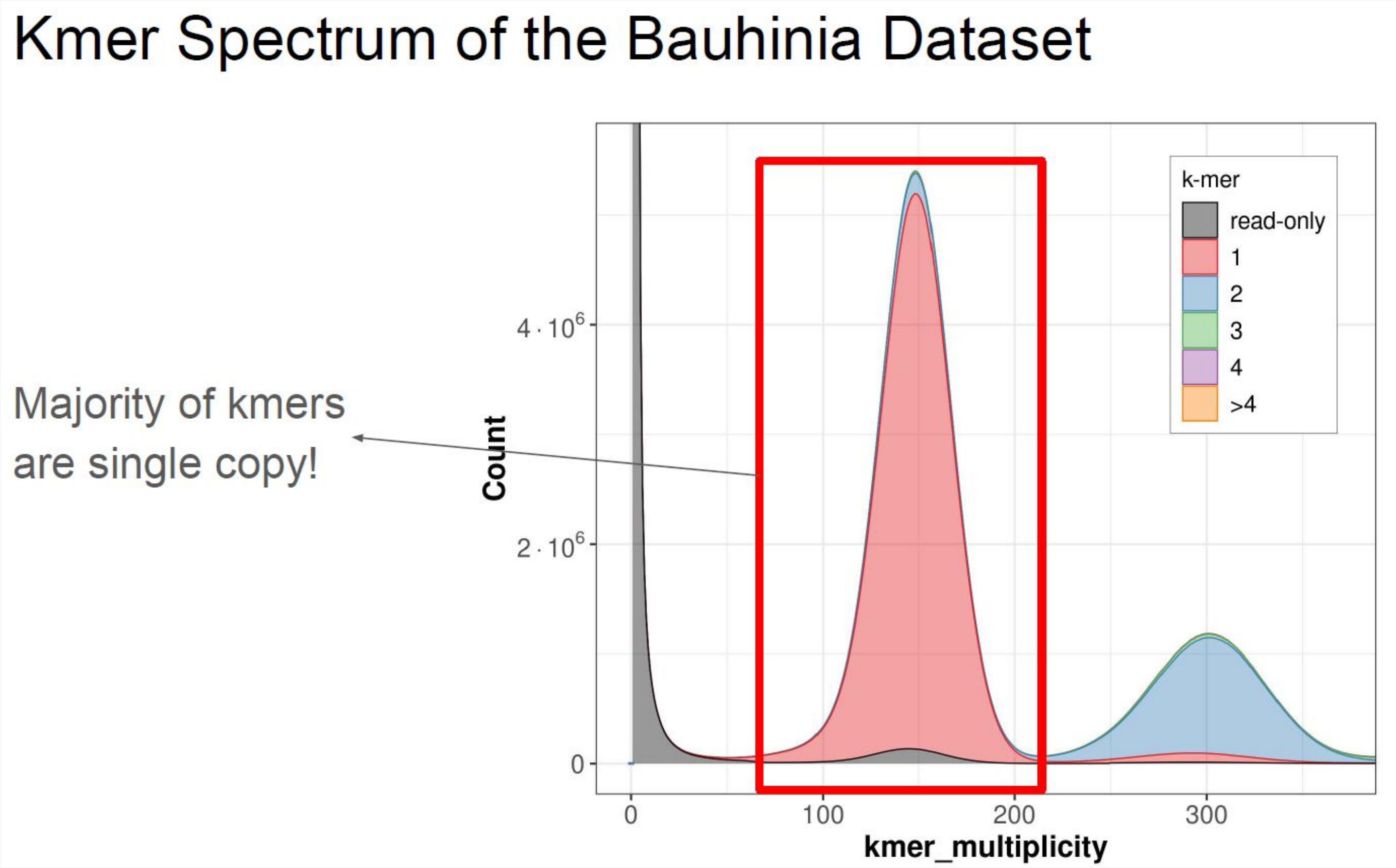
Sequence both parental species if their reference genomes are not already available

Determine solid K-mers of parental genomes

Classify a read from the hybrid genome
Paternal, if it is covered purely by paternal solid K-mers
Maternal, if it is covered purely by maternal solid K-mers
Ambiguous, otherwise

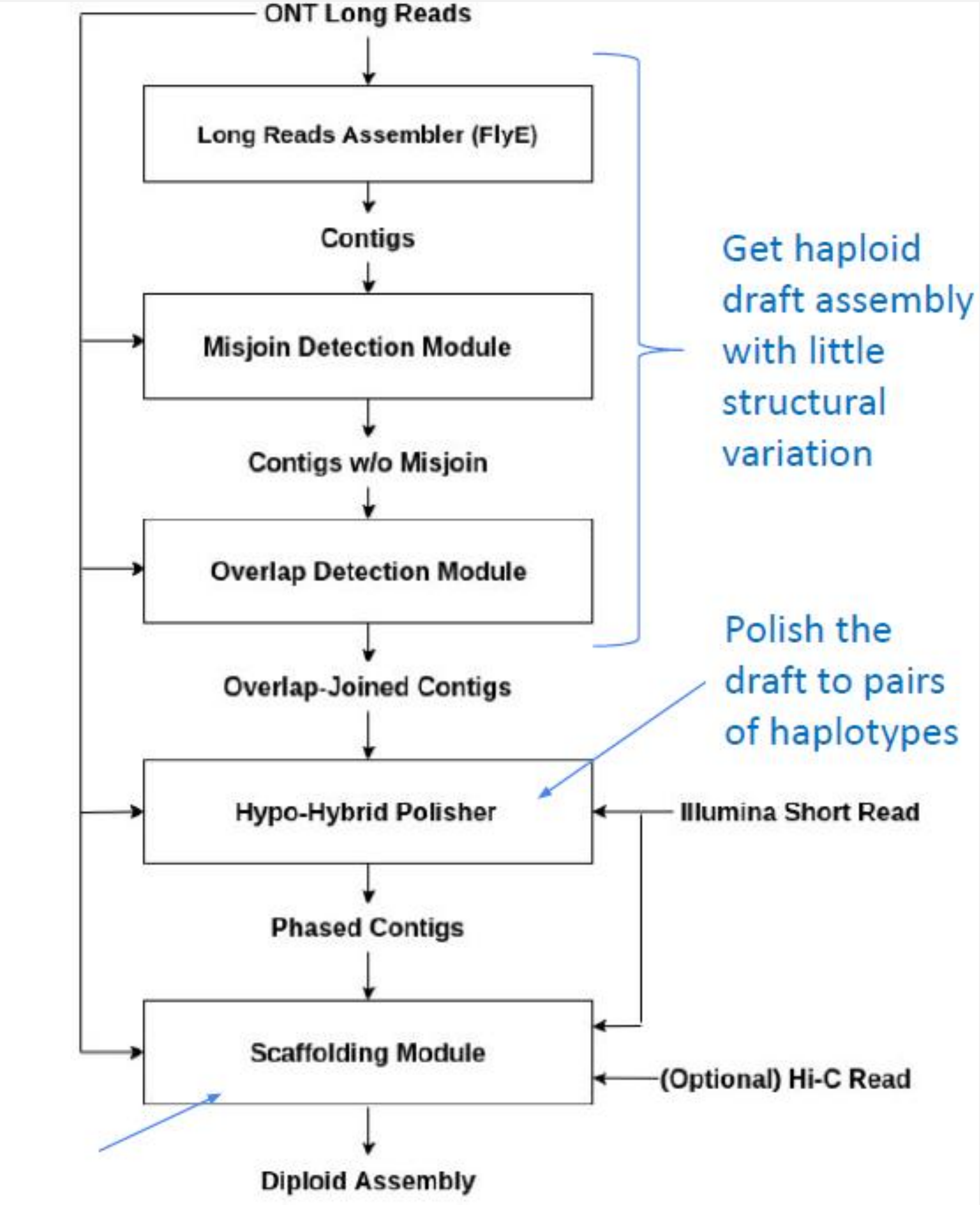
Separately assemble Paternal ∪ Ambiguous and Maternal ∪ Ambiguous

Solid K-mers: High-confidence K-mers that occur exactly once in the genome



Majority of kmers are single copy!

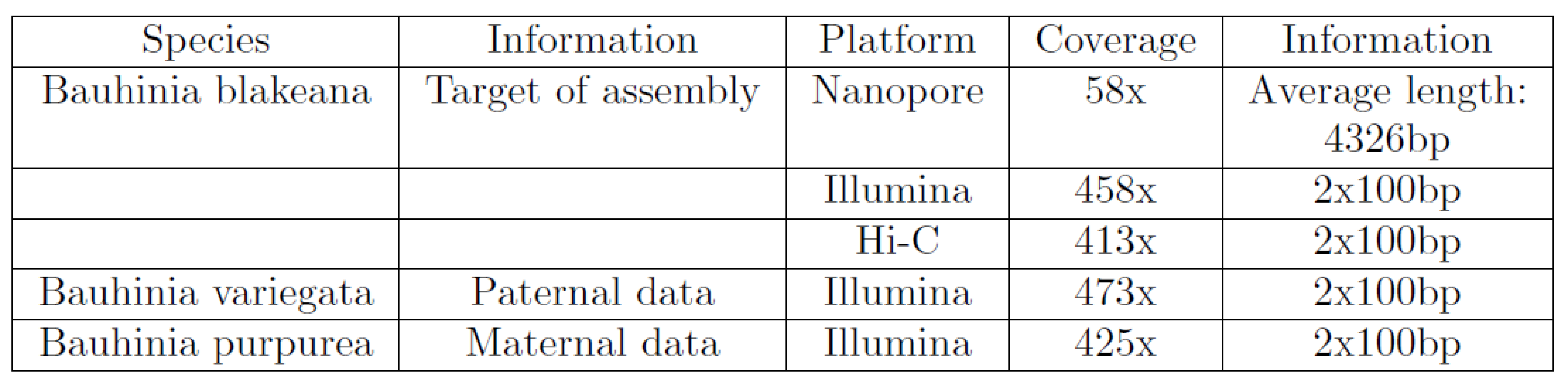
Genome assembly



Get haploid draft assembly with little structural variation

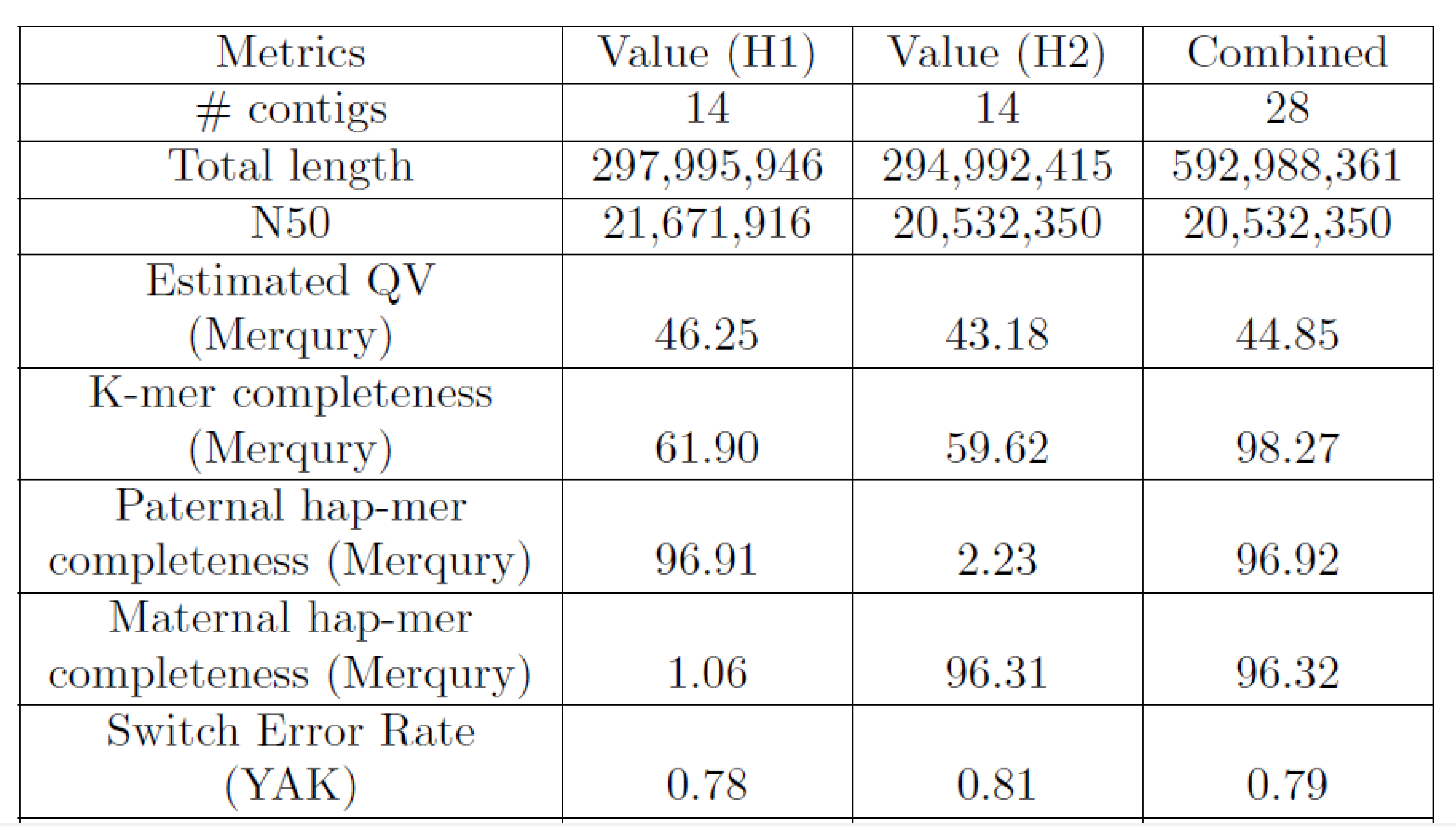
Polish the draft to pairs of haplotypes

Bauhinia dataset



Species	Information	Platform	Coverage	Information
Bauhinia blakeana	Target of assembly	Nanopore	58x	Average length: 4326bp
		Illumina	458x	2x100bp
		Hi-C	413x	2x100bp
Bauhinia variegata	Paternal data	Illumina	473x	2x100bp
Bauhinia purpurea	Maternal data	Illumina	425x	2x100bp

Bauhinia blakeana genome assembly

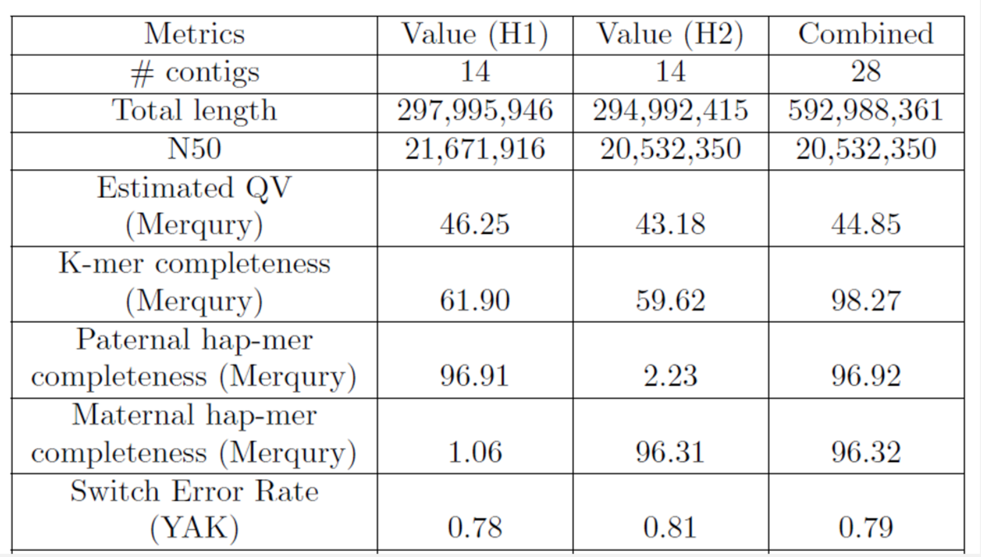


Metrics	Value (H1)	Value (H2)	Combined
# contigs	14	14	28
Total length	297,995,946	294,992,415	592,988,361
N50	21,671,916	20,532,350	20,532,350
Estimated QV (Merqury)	46.25	43.18	44.85
K-mer completeness (Merqury)	61.90	59.62	98.27
Paternal hap-mer completeness (Merqury)	96.91	2.23	96.92
Maternal hap-mer completeness (Merqury)	1.06	96.31	96.32
Switch Error Rate (YAK)	0.78	0.81	0.79

Exercise

How did Merqury estimate the completeness of genome assembly without a reference B. blakenana genome?

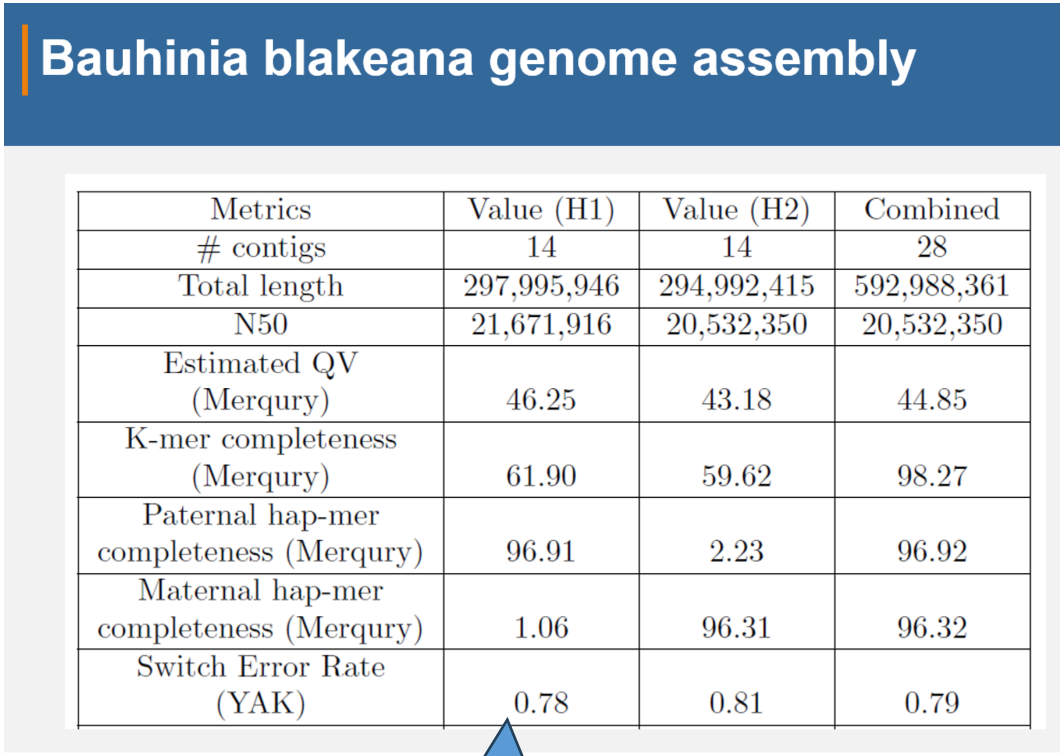
Bauhinia blakeana genome assembly



Metrics	Value (H1)	Value (H2)	Combined
# contigs	14	14	28
Total length	297,995,946	294,992,415	592,988,361
N50	21,671,916	20,532,350	20,532,350
Estimated QV (Merqury)	46.25	43.18	44.85
K-mer completeness (Merqury)	61.90	59.62	98.27
Paternal hap-mer completeness (Merqury)	96.91	2.23	96.92
Maternal hap-mer completeness (Merqury)	1.06	96.31	96.32
Switch Error Rate (YAK)	0.78	0.81	0.79

Exercise

What is switch error rate?



Bauhinia blakeana genome assembly

Metrics	Value (H1)	Value (H2)	Combined
# contigs	14	14	28
Total length	297,995,946	294,992,415	592,988,361
N50	21,671,916	20,532,350	20,532,350
Estimated QV (Merqury)	46.25	43.18	44.85
K-mer completeness (Merqury)	61.90	59.62	98.27
Paternal hap-mer completeness (Merqury)	96.91	2.23	96.92
Maternal hap-mer completeness (Merqury)	1.06	96.31	96.32
Switch Error Rate (YAK)	0.78	0.81	0.79

0.78 switches per kb

Acknowledgements

Most of the slides were adapted from Joshua Casey Darian’s PhD thesis and his viva slides