GenomeScope 2.0 and Smudgeplots

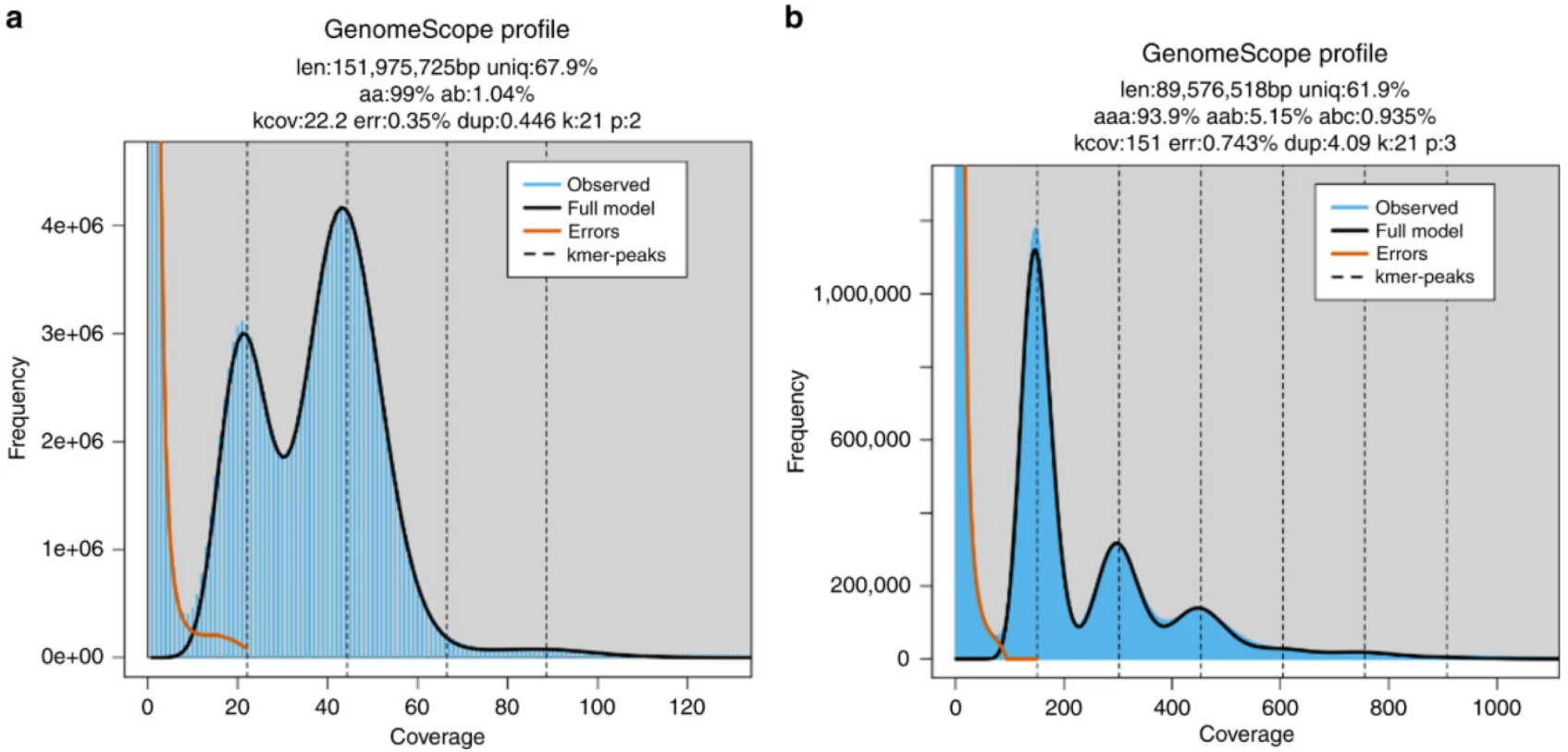
- Heterozygous and polyploid genomes.

- Polyploid-aware model
- Unassembled sequencing data*
- Visualize and estimate ploidy levels without prior knowledge.

- At least 15x coverage per homolog for GenomeScope and 25x coverage per homolog for Smudgeplot is required.

Expected plots:

From: [GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes]



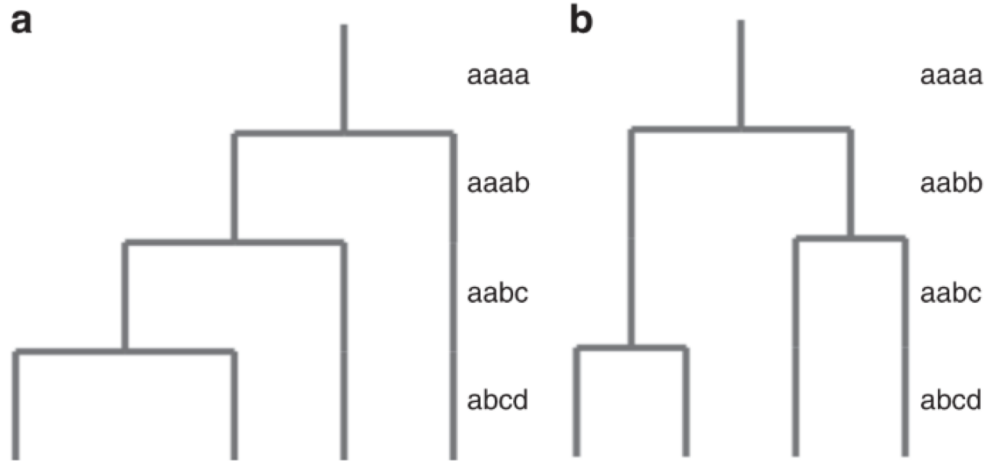
K-mer spectra and fitted models for (**a**) diploid *Arabidopsis thaliana* and (**b**) triploid *Meloidogyne enterolobii*. Note that the diploid plot has two major peaks, while the triploid plot has three major peaks. Both also have high frequency putative error k-mers with coverage near 1.

Diploid Triploid

In general: higher heterozygosity → higher first peak

Simulations vs. reality:

- Simulated data with several ploidy levels (3,4,5,6)
- Repetetiveness (0-20%)
- Heterozygosity (0, 0.5, 1, 1.5, and 2%)
- 15x coverage per homolog and 1% sequencing error (GenomeScope)
- 25x coverage per homolog and 1% error (Smudgeplots)



a The autotetraploid topology, notated as (, (, (,))); in Newick notation, corresponds to the following nucleotide heterozygosity forms: *aaaa*, *aaab*, *aabc*, *abcd*. **b** The allotetraploid topology, notated as ((,), (,)); in Newick notation, corresponds to the following nucleotide heterozygosity forms: *aaaa*, *aabb*, *aabc*, *abcd*.

Common name	Species name	Estimated genome size	Assembly size
Coastal redwood	*Sequoia sempervirens*	27.0 Gbp	26.5 Gbp
Cotton	*Gossypium barbadense*	2.293 Gbp	2.267 Gbp[27]
Cotton	*Gossypium hirsutum*	2.349 Gbp	2.347 Gbp[27]
Marbled crayfish	*Procambarus virginalis*	9.5 Gbp	3.3 Gbp[21]
Root-knot nematode	*Meloidogyne arenaria*	290.4 Mbp	163.7 Mbp[7]
Root-knot nematode	*Meloidogyne enterolobii*	268.7 Mbp	162.4 Mbp[7]
Root-knot nematode	*Meloidogyne floridensis*	201.7 Mbp	74.9 Mbp[7]
Root-knot nematode	*Meloidogyne incognita*	207.4 Mbp	122.0 Mbp[7]
Root-knot nematode	*Meloidogyne javanica*	280.2 Mbp	142.6 Mbp[7]
Potato	*Solanum tuberosum*	3.0 Gbp	778.7 Mbp[29]
Wheat	*Triticum aestivum*	14.1 Gbp	15.34 Gbp[23]

Smudgeplots

How are the heterozygous kmers defined?

1. Are exactly one SNP from each other.
2. Has to be unique pairs. Otherwise discarded.

This leads to ssignificant subsampling of the data in some cases →

For high heterozygosity, Smudgeplot will underestimate the ploidy level because real k-mers in a k-mer pair sometimes will have more than one nucleotide difference and will not be taken into account.

Age and mode of duplication is important.

Smudgeplots - repeats

Diploid: up to 39% repetitiveness,

Triploid: up to 38% repetitiveness,

Allotetraploid: data up to 43% repetitiveness,

Autotetraploid: data up to 38% repetitiveness,

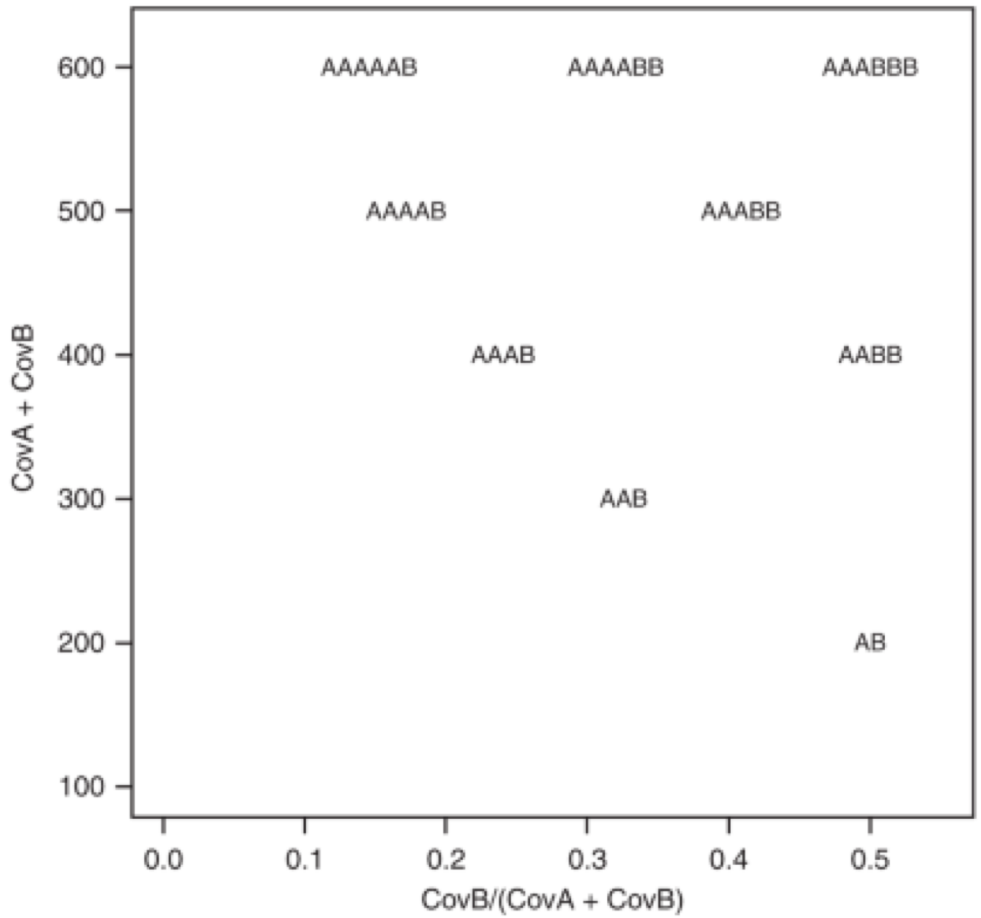
Higher repetitiveness → overestimation of the ploidy

because the signal from repetitive k-mers overshadows that from heterozygous k-mers.

Expectation for data with 100x monoploid coverage:

Fig. 7: Smudgeplot genomic structure locations.

a



b

Structure	CovA + CovB	$\frac{\text{CovB}}{\text{CovA + CovB}}$
AB	2λ	0.5
AAB	3λ	0.333
AAAB	4λ	0.25
AABB	4λ	0.5
AAAAB	5λ	0.2
AAABB	5λ	0.4
AAAAAB	6λ	0.166
AAAABB	6λ	0.333
AAABBB	6λ	0.5
...	...	...

Coordinates of individual genomic structures for a genome with monoploid coverage = 100 in **a** 2D space of coverage sums versus coverage ratios and in **b** a table of coordinates.

Beginning of the paper:
”Here, we present GenomeScope 2.0, which extends this approach with a polyploid-aware mixture model to computationally infer genome characteristics from unassembled sequencing data.”

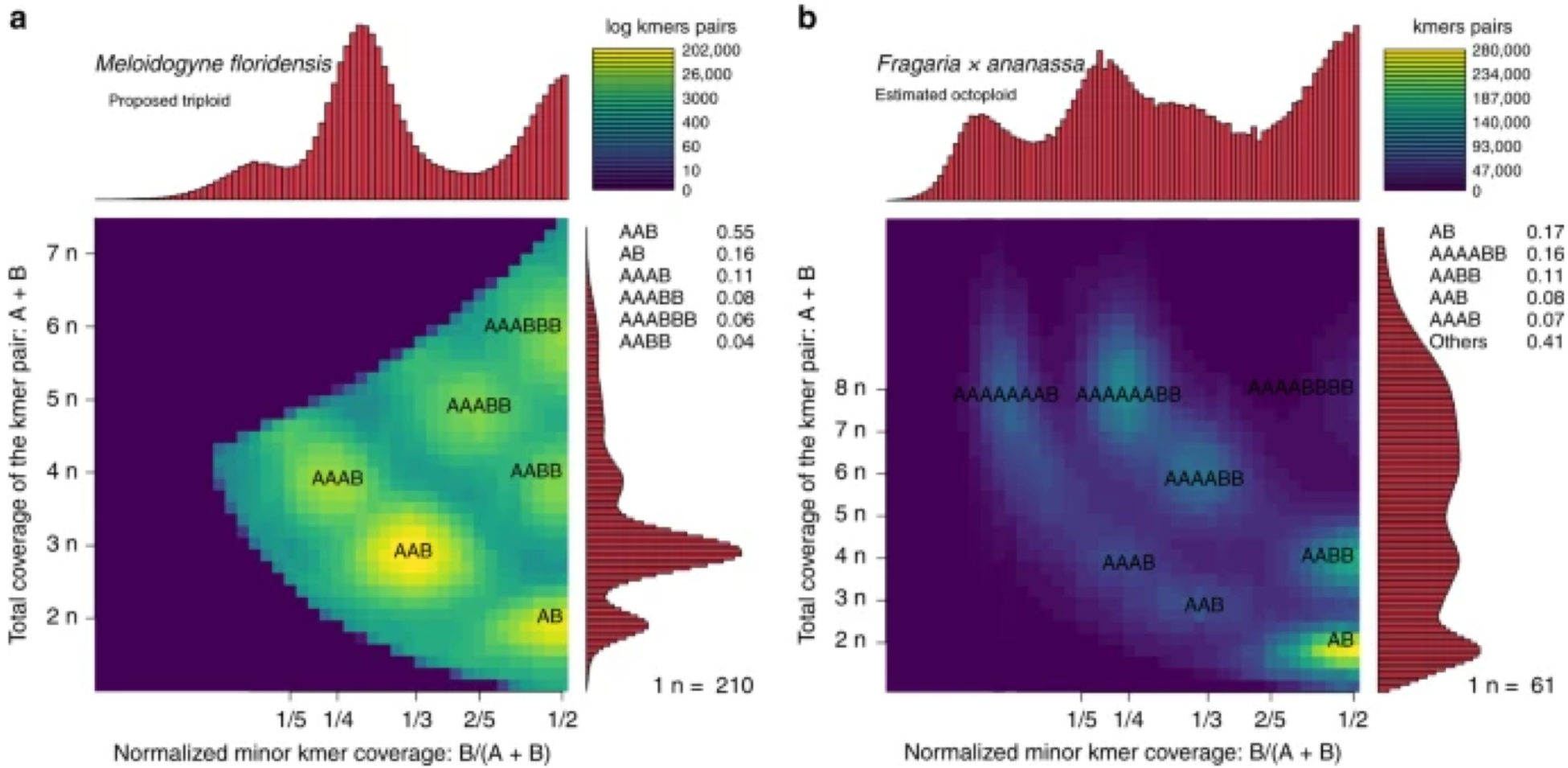
Marbled crayfish: ”It is clear that the assembly only spans one homolog of the triploid genome.”

Root-knot nematodes: ” GenomeScope estimates for genome size are 1.65–2.69 times larger than the current best assemblies[7], suggesting the assemblies have partially collapsed the homologous chromosomes”

Wheat: ”Bread wheat (*Triticum aestivum*) is an allohexaploid which consists of three subgenomes[22]. A Smudgeplot analysis inferred that the ploidy was diploid, because the individual subgenomes are highly divergent from each other.”

Discussion:
”Thus, one limitation of using a k-mer-based technique is that in these cases too few k-mers may actually be shared between the homologous copies. This can lead Smudgeplot to infer diploidy even for polyploid species.”
”It is important to run GenomeScope with the correct value for the ploidy parameter.”

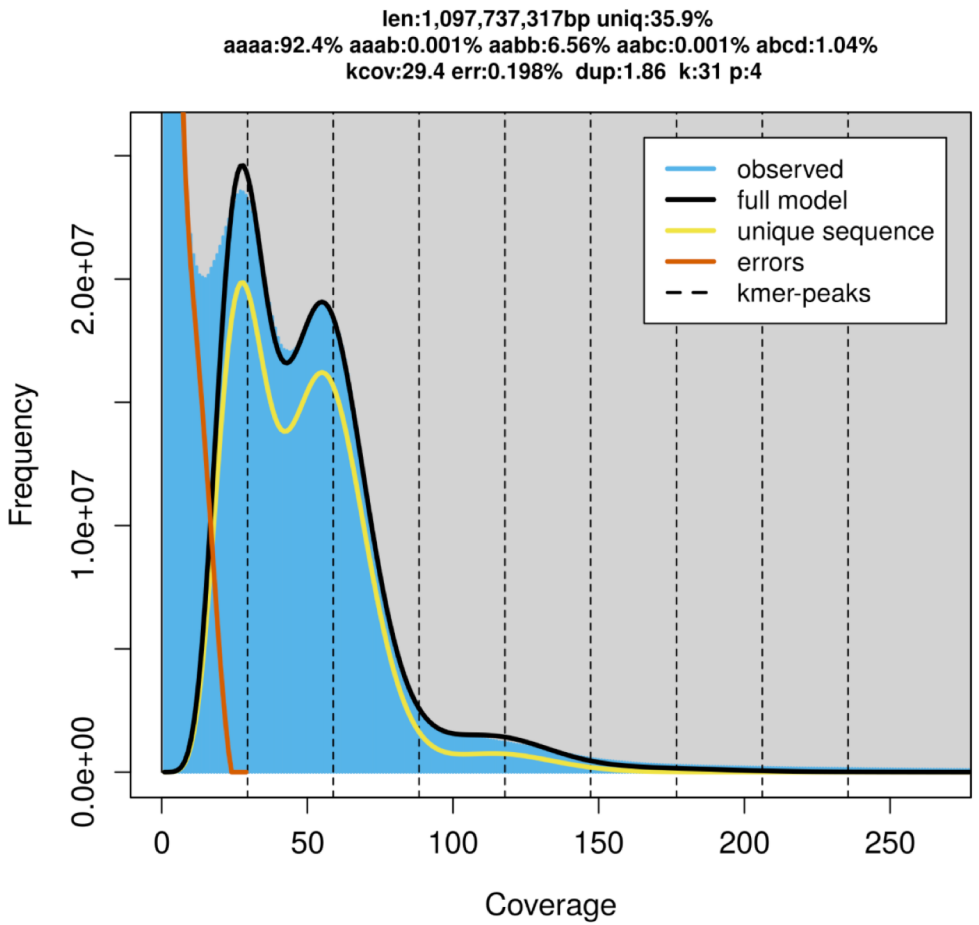
Fig. 8: Smudgeplots on real datasets.

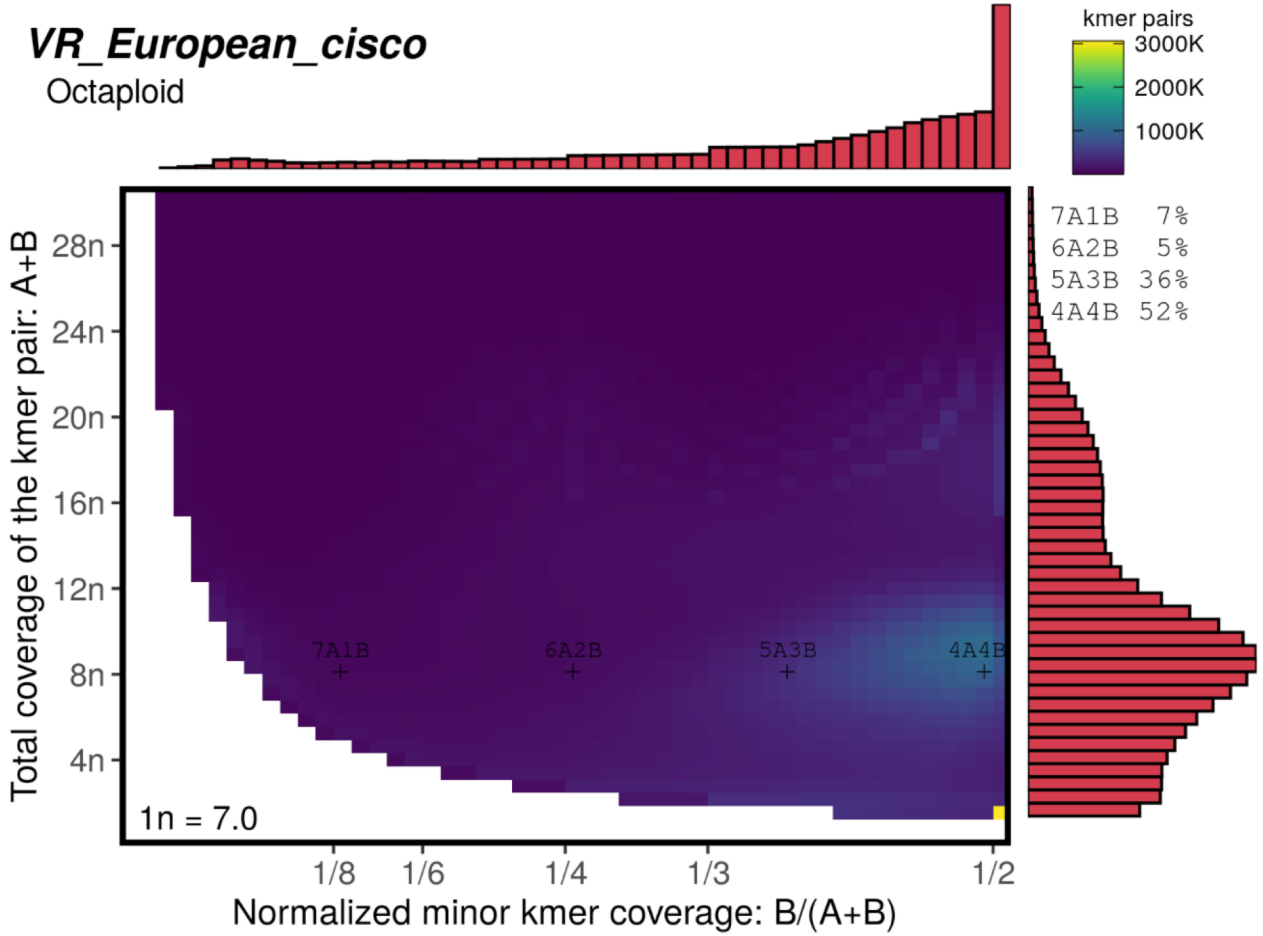


Smudgeplots for (**a**) the triploid root-knot nematode *Meloidogyne floridensis* and (**b**) the octaploid strawberry *Fragaria* × *ananassa*.

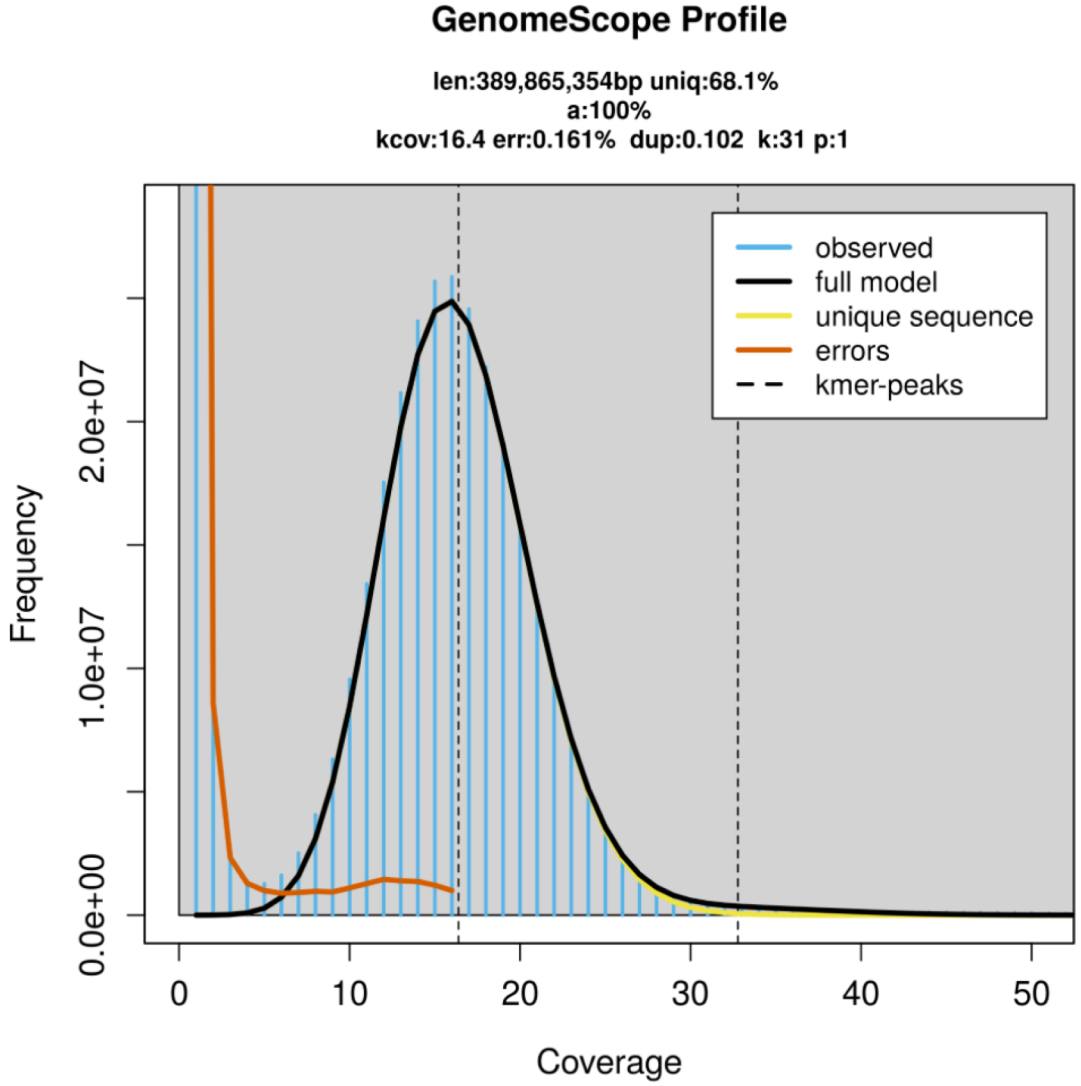
Our own data (polyploid):

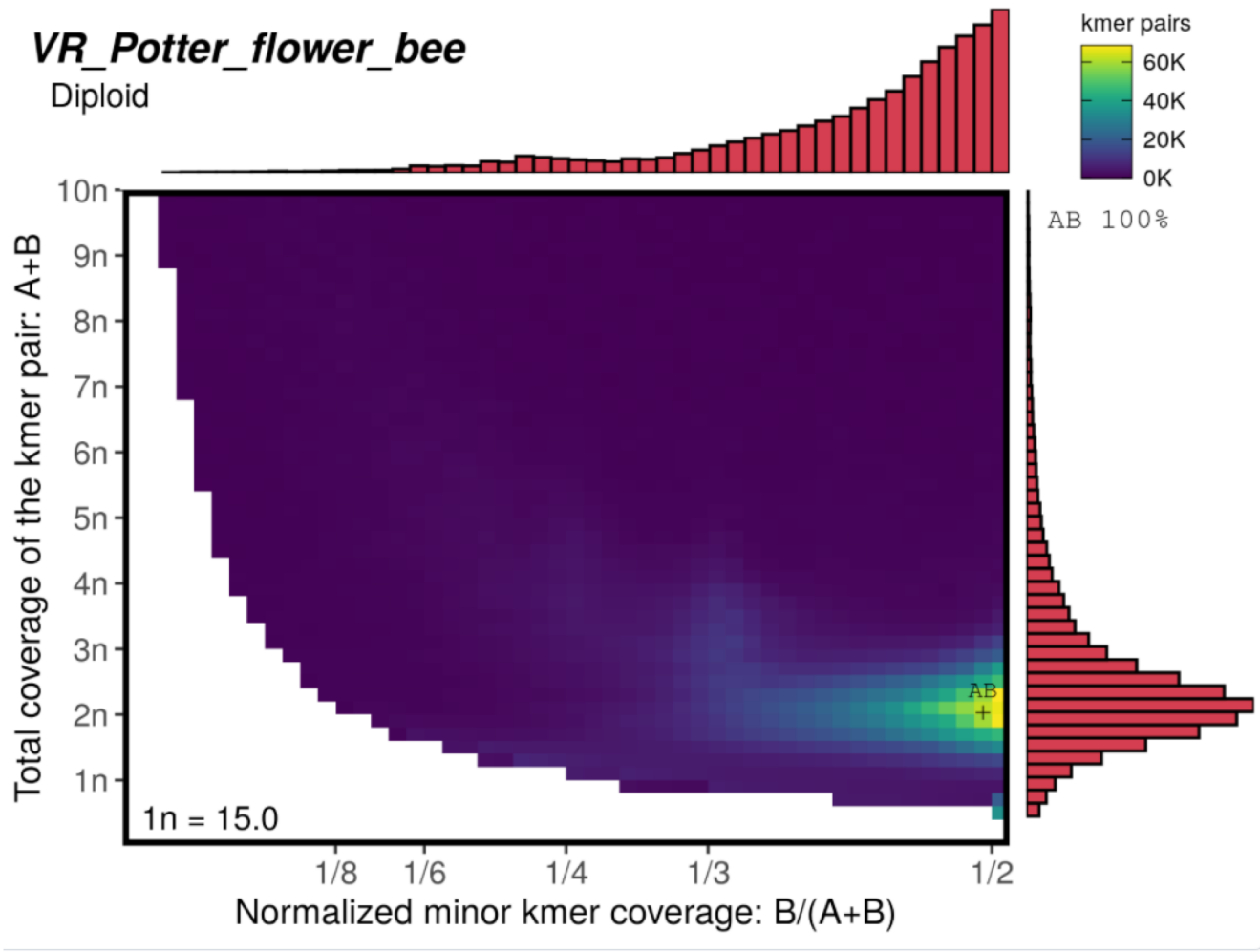
- Genome scope vs. smudgeplot





Our own data (haploid):





Take home messages:

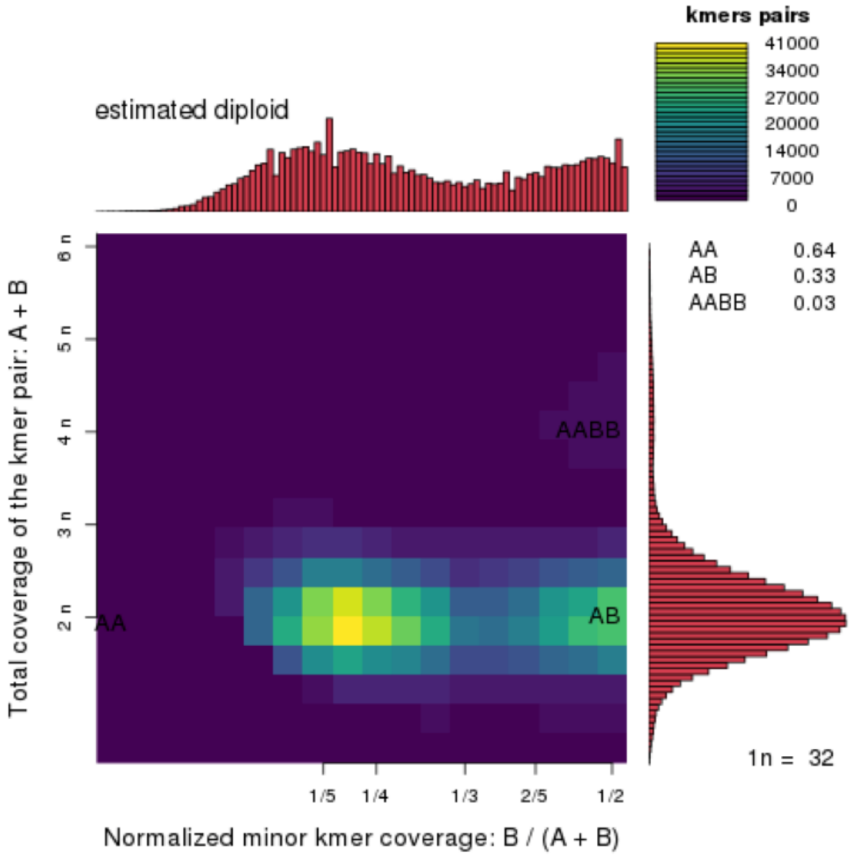
- Most likely needs to be iterated several times.
- A smudgeplot in isolation is not good enough to draw conclusions.
- Real live examples are quite far from the simulated examples:
  - Coverage of data most likely too low initially.
  - Knowledge of duplication history probaply not known.
  - Repeat content and heterozygosity is very often much higher.
  - etc. etc.

Yeast example from wiki:

https://github.com/KamilSJaron/smudgeplot/wiki/tutorial-saccharomyces

Setting thresholds are important:

Default:



After tweaking (a few rounds):