GA LRA Workshop Supplementary Info: phasing and approaches to phasing

Linelle Abueg

04-06 July 2023

What is phasing? Why is phasing?

- **Phasing aims to partition contigs for an individual according to that sequence's haplotype-of-origin**
- Phasing prevents *switch errors*, when contigs switch from one haplotype to the other, creating a sequence that might not have been actually present in the original genome
- Phasing also helps prevent false duplications, which can arise when the two alleles for a particular locus look different enough from each other that the assembly algorithm thinks they're two different regions of the genome, resulting in the same region falsely being represented twice in the assembly

Types of assemblies: pseudohaplotype

- **Pseudohaplotype** assembly consists of long blocks phased by haplotype, separated by regions where haplotype cannot be distinguished (usually homozygous regions)
- **Primary assembly**: traditionally the more complete representation of an individual's genome – consists of homozygous regions and one set of loci for heterozygous regions
- **Alternate assembly:** consists of the alternate loci not represented in the *primary assembly* (that is, the other haplotype's allele for heterozygous loci). These sequences are often referred to as haplotigs.

Types of assemblies: trio-phased

- Trio-phasing requires sequencing the parents, in order to identify alleles that are present in the father AND the offspring but NOT the mother, and vice versa, in order to identify alleles that can properly segregate the offspring’s contigs.

Types of assemblies: Hi-C-phased

- A recent alternative to trio phasing uses Hi-C data from the same individual to try to phase contigs according to long-range Hi-C linkage information

Phasing approaches in order of preference

Rank	Approach	Details
1	**Trio**	Pros: ground truth. Gold standard for phasing. Cons: can be hard to acquire/identify parental sample, especially for non-human wild samples. Even if parents identifiable, is additional cost.
2	**Hi-C**	Pros: Hi-C data comes from same individual, so don't need to identify parents. Can get good phasing without hassle of trio logistics. Cons: need to do Hi-C prep, which requires whole, un-lysed cells. Can be tricky for certain sample types.
3	**None (pseudohaplotype)**	Pros: is your only option if you can only get HiFi data. Cons: not properly phased, contigs often have mixed hapmer content, and false duplications can remain (and purging is a messy process)

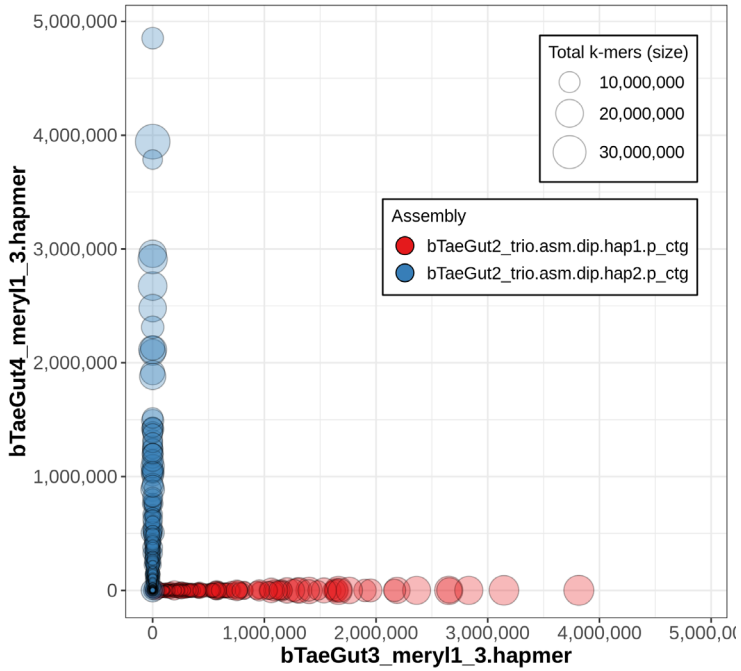
Hi-C phasing phases chromosomes

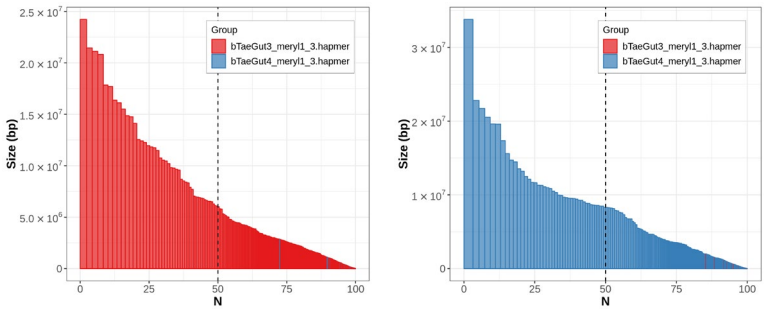
- Merqury blob plot shows that contigs for Hi-C phasing largely do not have mixed hapmer content
  - (cf. the pseudohaplotype assemblies' blob plot, where there are many contigs diagonally on the graph, meaning they contain both maternal and paternal hapmer content)
- This pattern remains even when looking at scaffolded Hi-C phased assembly
- The scaffolders are haplotype-unaware, so we can infer that chromosomes' constituent contigs were properly phased together into assemblies

Merqury blob and block plots for bTaeGut2 (*Taeniopygia guttata*)

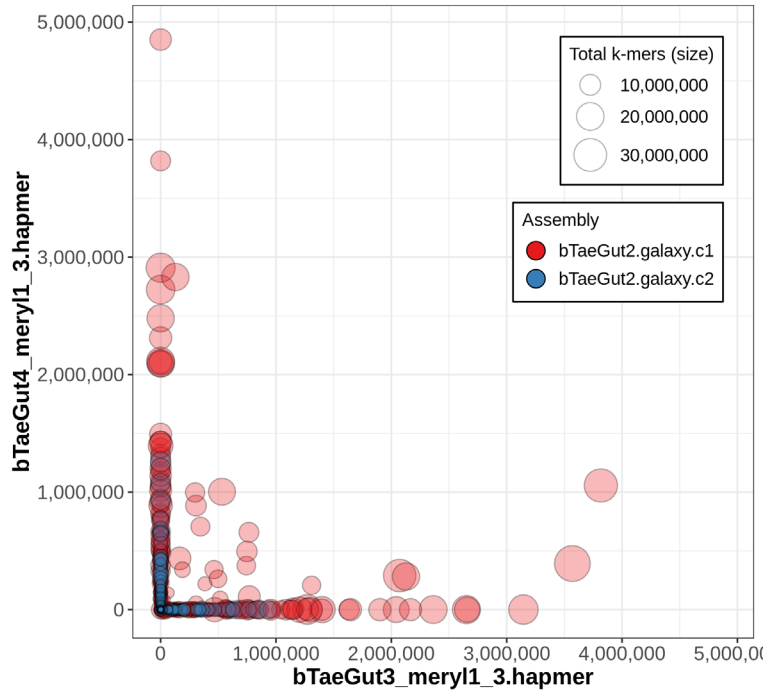
- blob plots: each blob is a contig, and its x,y position represents parental hapmer content, while color represents assembly-of-origin (e.g., pri, alt, hap1, hap2, mat, pat)
- block plots: each line is a contig, colored by majority hapmer content, each graph is an assembly (e.g., pri, alt, hap1, hap2, mat, pat)

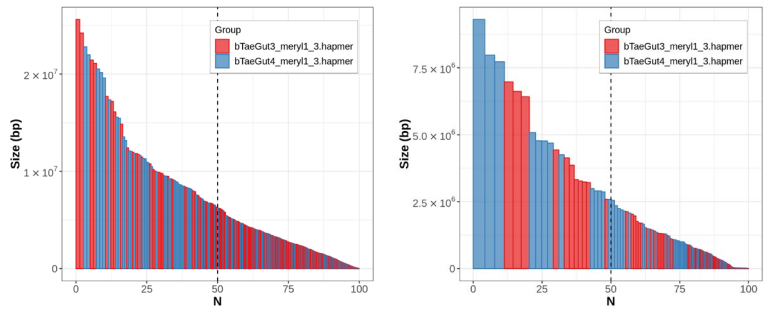
Trio



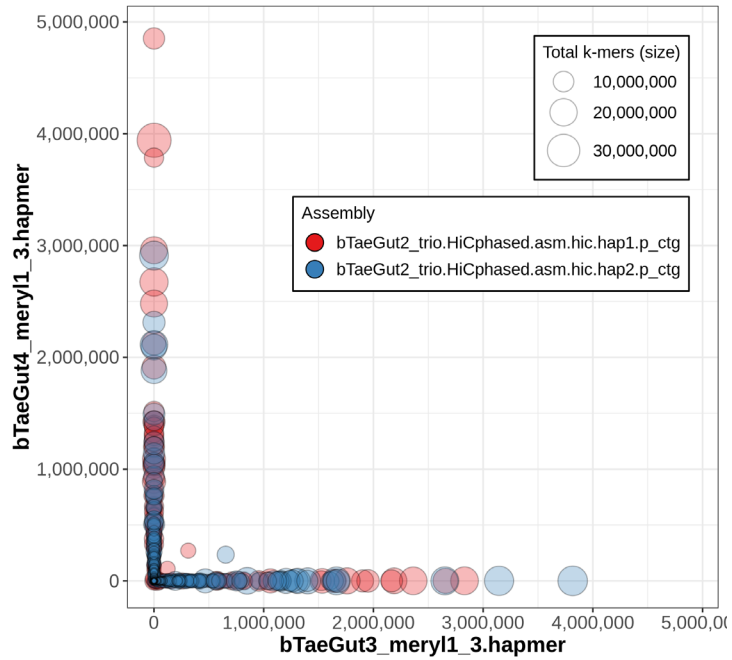


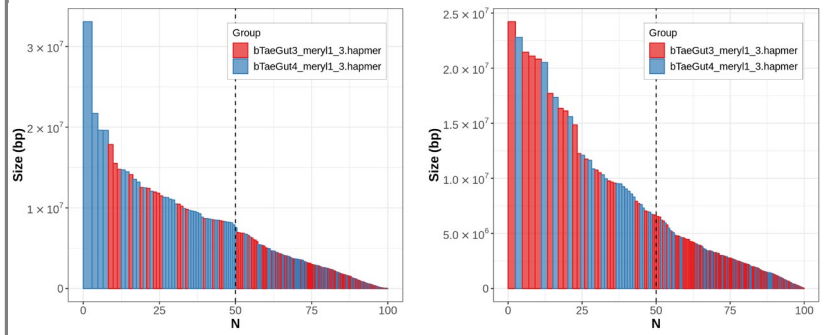
Unphased (pri/alt)





Hi-C

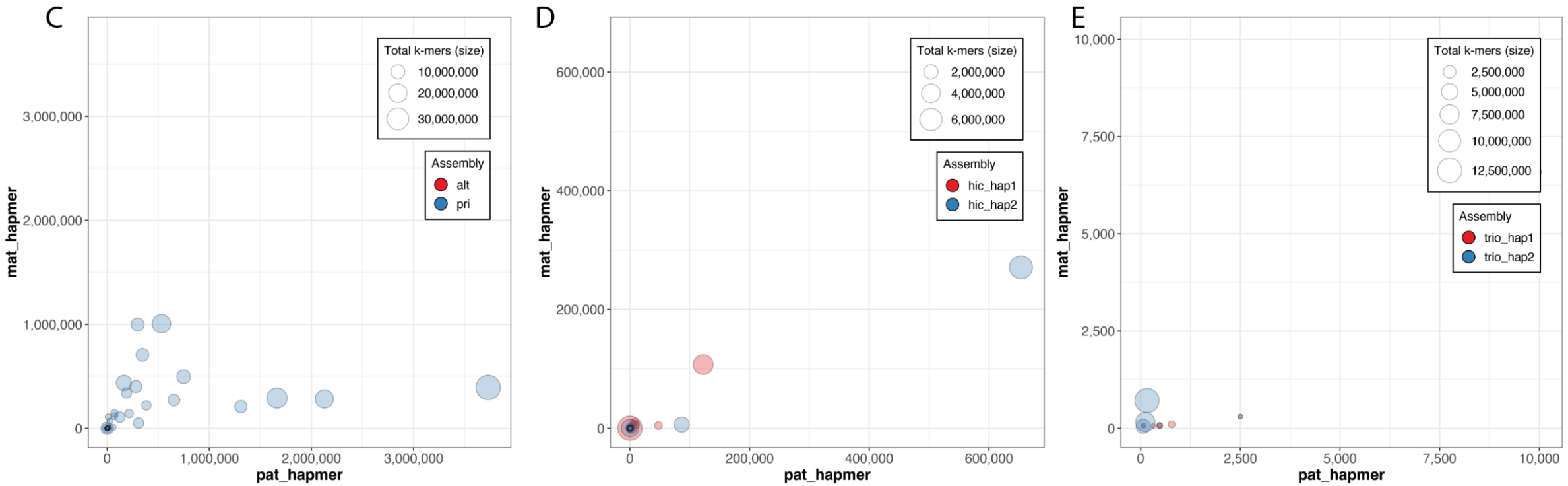




primary/alternate

Hi-C-phased

trio



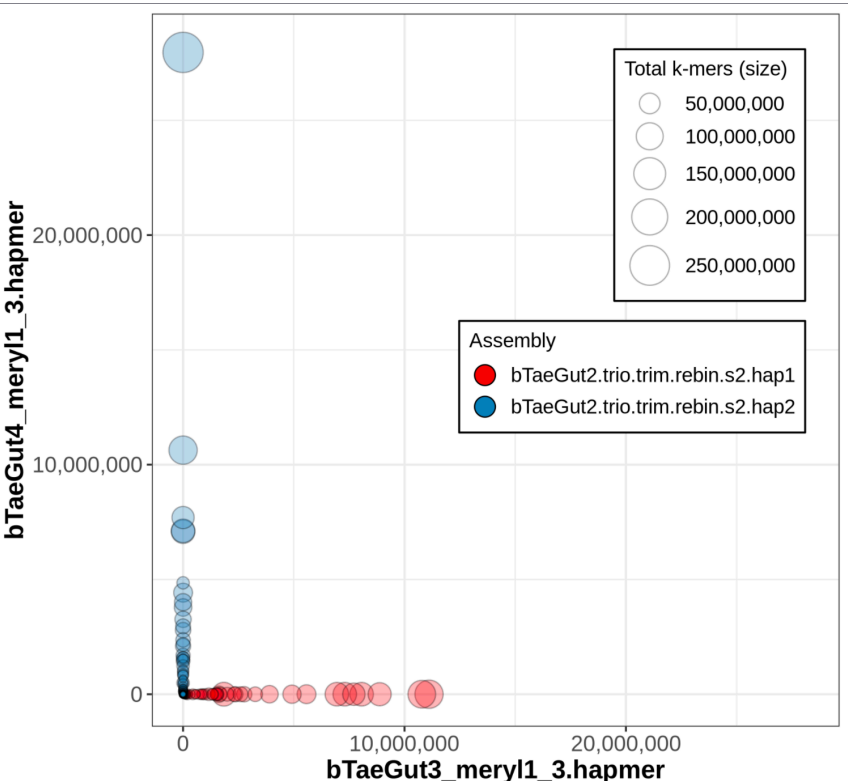
Contig-level blob plots with on-axes blobs (*i.e.*, ones with only one parent's hapmers) removed for increased visibility of mixed-hapmer-content contigs
(note that the axes scales are on a different order of magnitude)

Scaffold-level blob plots for the various assemblies

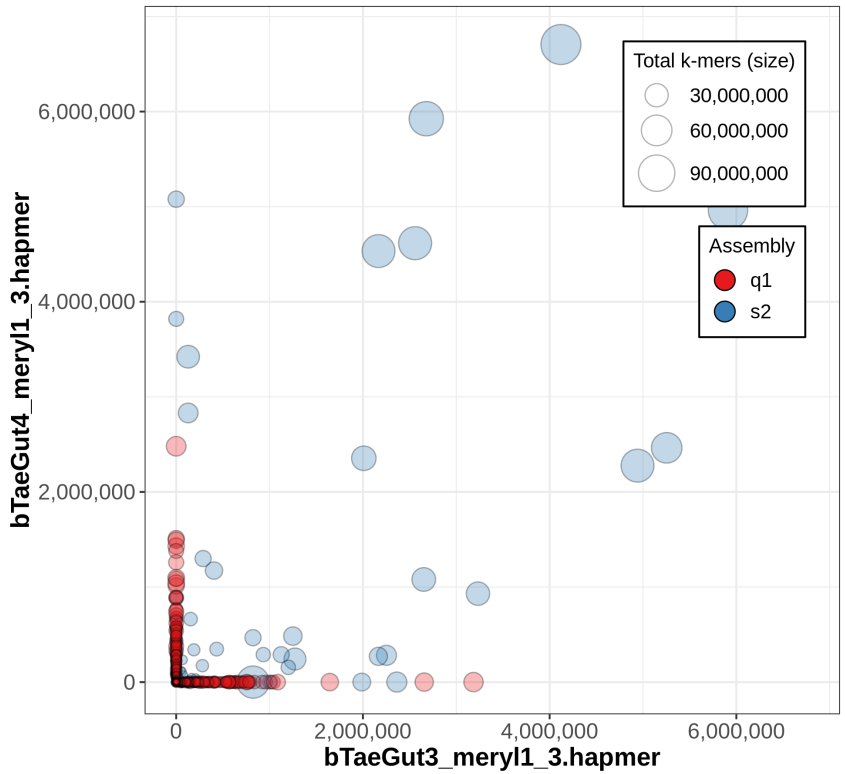
Hi-C scaffolds remain largely phased, and the scaffolders are haplotype-unaware, so we can conclude that chromosomes' constituent contigs were successfully phased together

(pri/alt has lots of alt contigs with only one hapmer because these are the alternate loci)

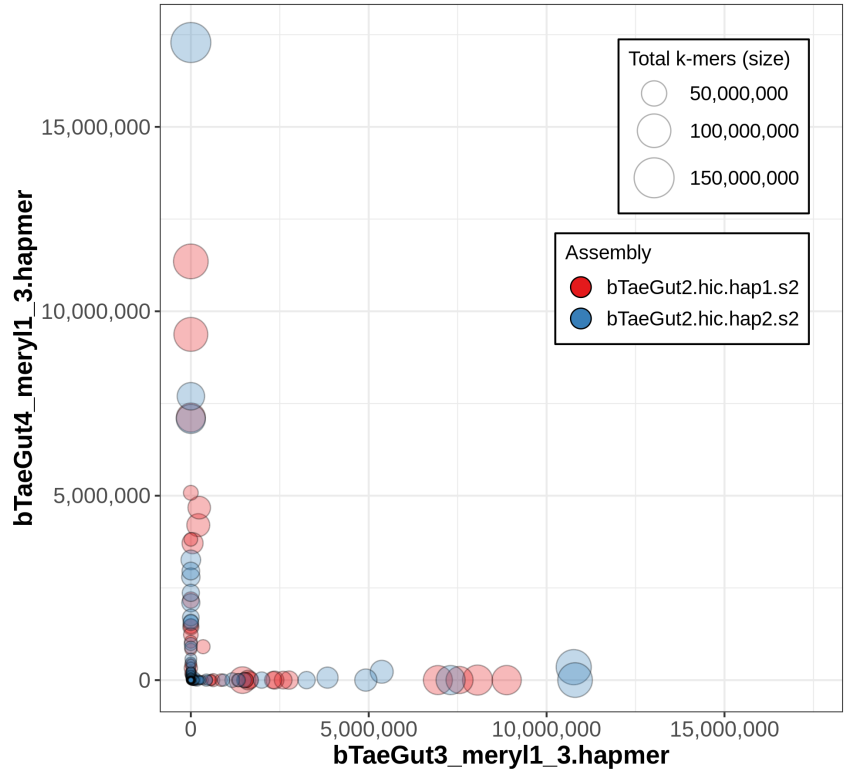
Trio



Pri/alt



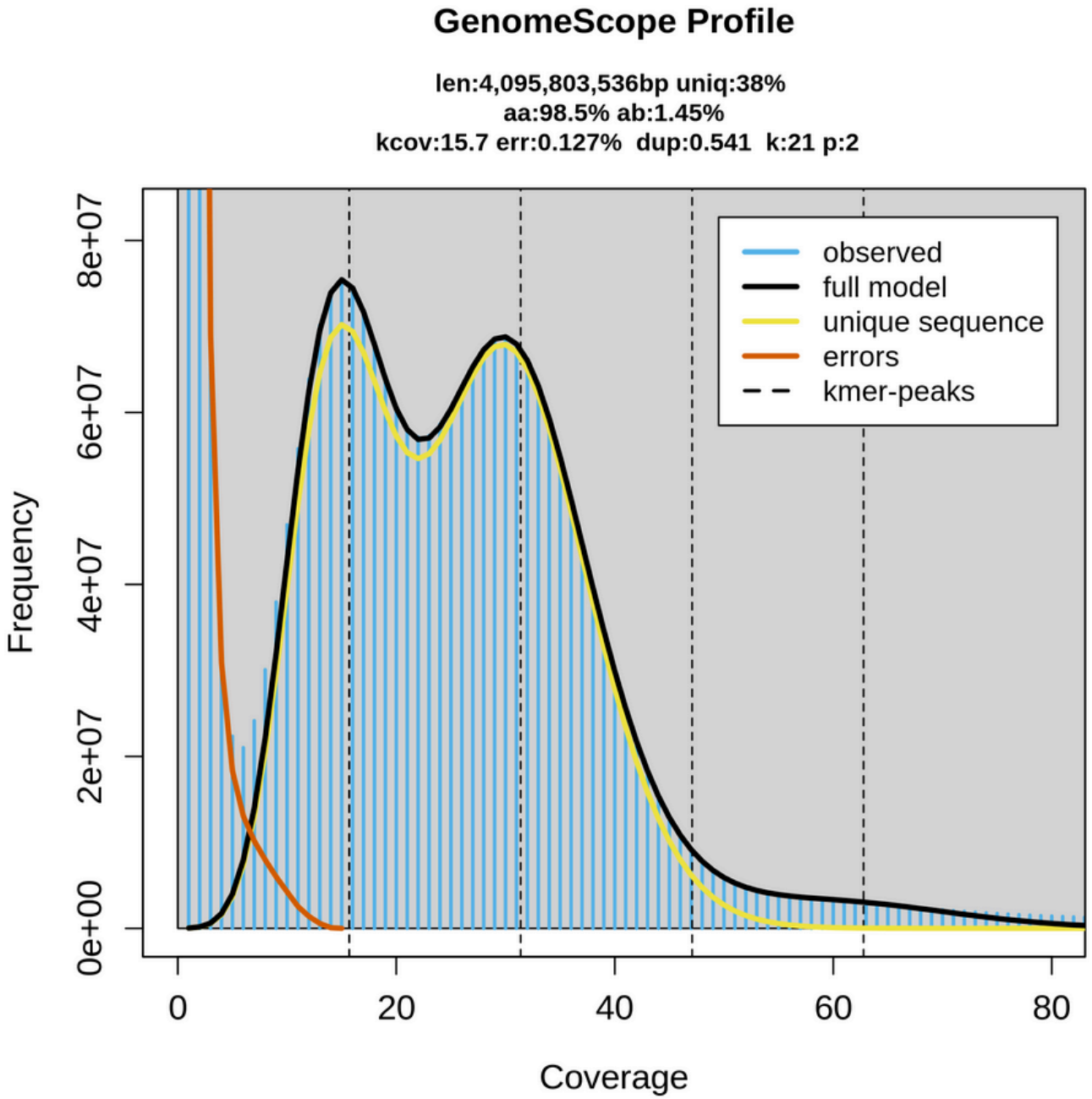
Hi-C



Hi-C phasing prevents false duplication

Case study: aGasCar1
Eastern narrow-mouthed toad
(*Gastrophryne carolinensis)*

pri/alt vs. HiC-phased assembly



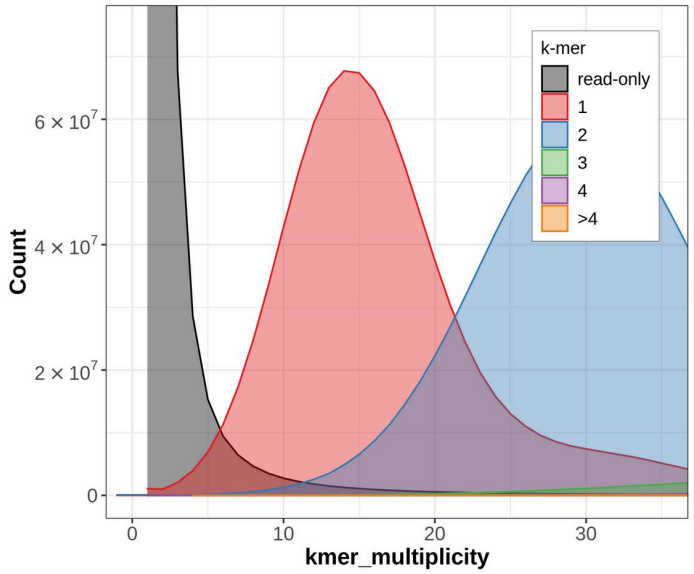
Hifiasm pri/alt assembly with purge_dups

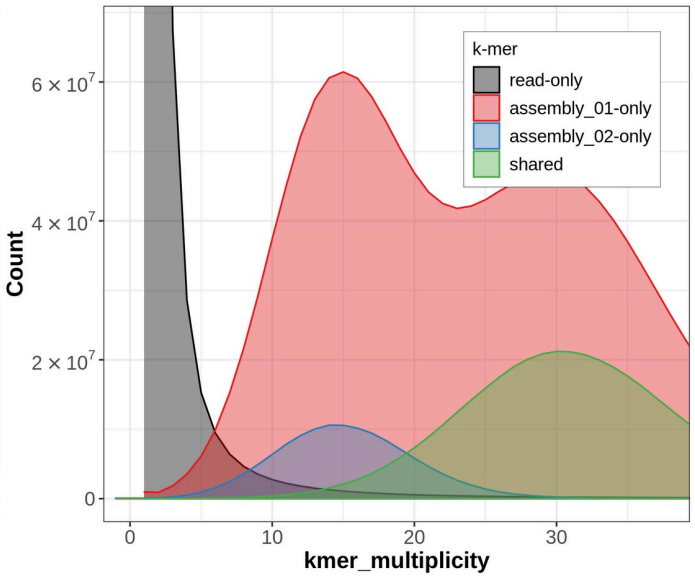
Eastern narrow-mouthed toad

spectra-cn

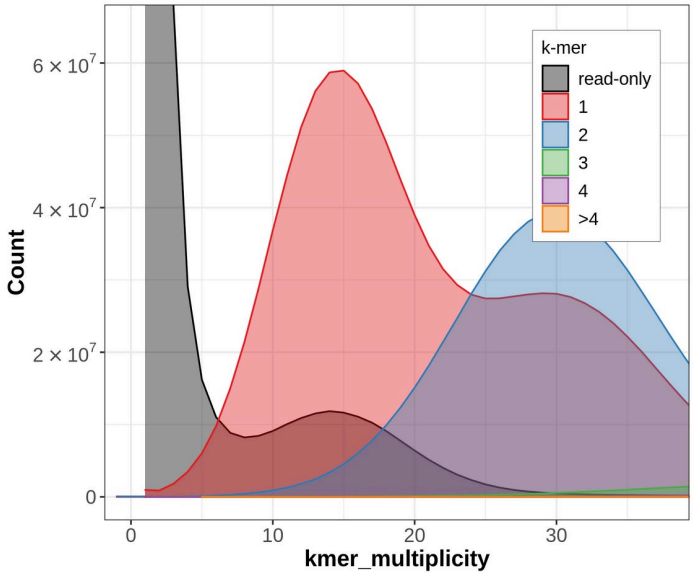
spectra-asm

both assemblies





primary only



results *before* purging

- primary has lots of duplicated BUSCO genes
- lots of 2-copy *k*-mers in the primary, at diploid coverage (diploid regions should be 1-copy, 1 for each haplotype)

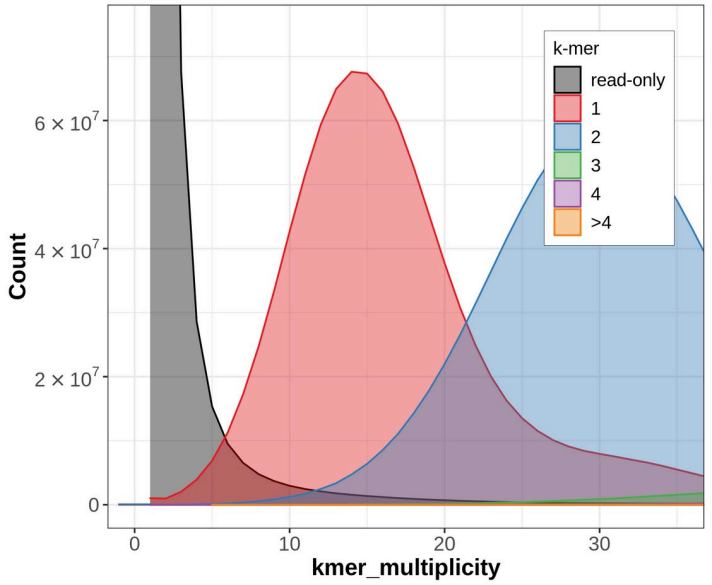
	Primary (unpurged)	Alternate (unpurged)
# of contigs	3,548	3,805
Total length	7,012,181,570	1,303,392,418
N50	5,385,022	1,773,313
L50	358	208

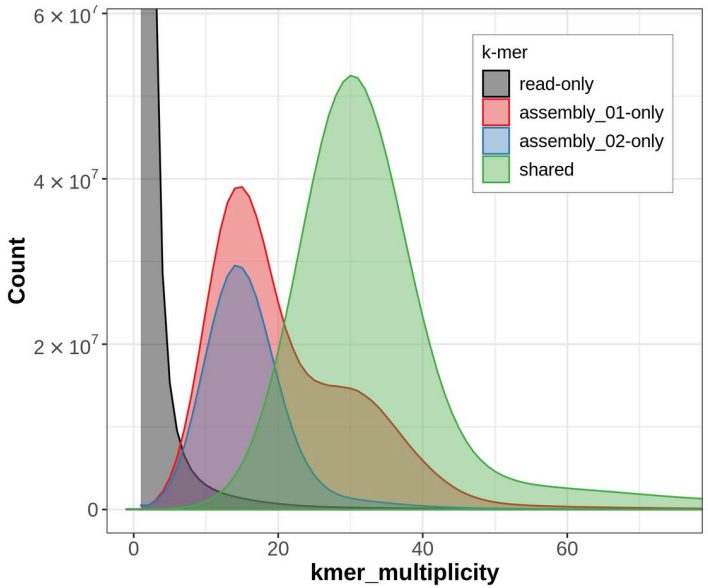
Hifiasm pri/alt assembly with purge_dups

spectra-cn

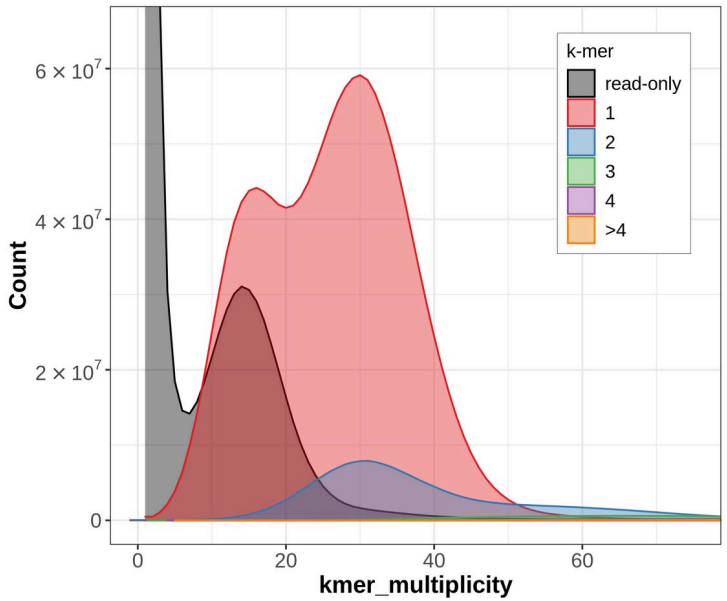
spectra-asm

both assemblies





primary only

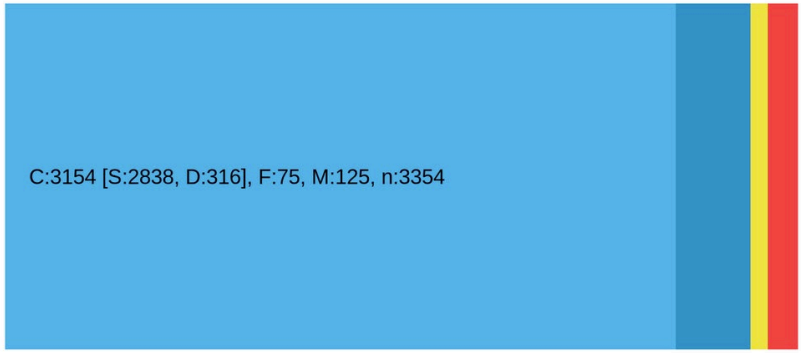


results *after* purging

- still some duplicated BUSCO genes
- primary assembly still has some 2-copy *k*-mers at diploid coverage
- trying to fine-tune purging could be messy

image: Alessandro Catenazzi

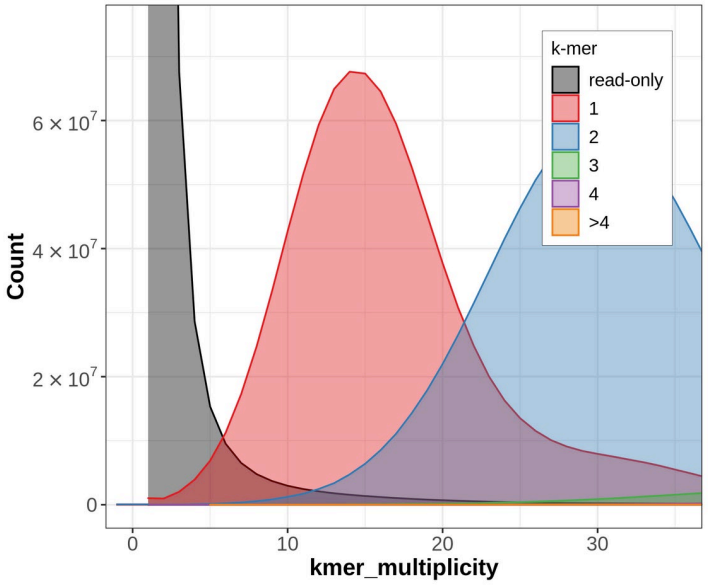
Eastern narrow-mouthed toad



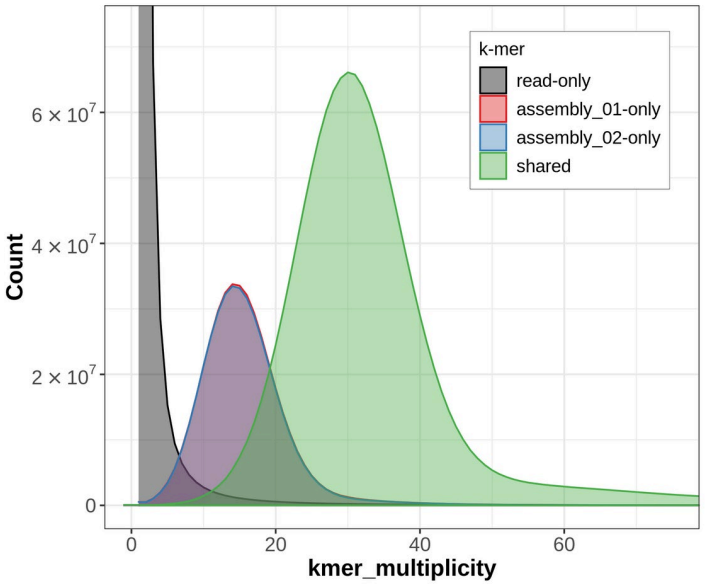
	Primary (post-purging)	Alternate (post-purging)
# of contigs	1,363	3,997
Total length	4,663,182,711	3,473,507,201
N50	7,405,859	2,218,743
L50	184	444

Hifiasm HiC-phased assembly (aGasCar1)

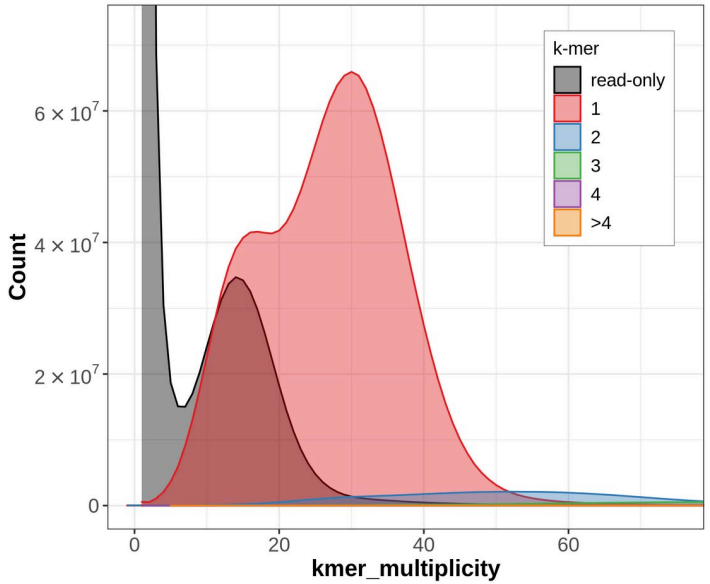
spectra-cn (both assemblies)



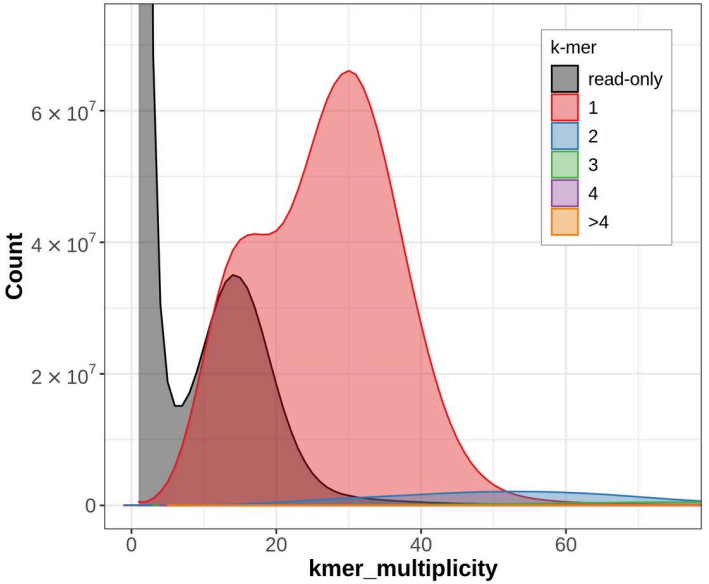
spectra-asm (both assemblies)



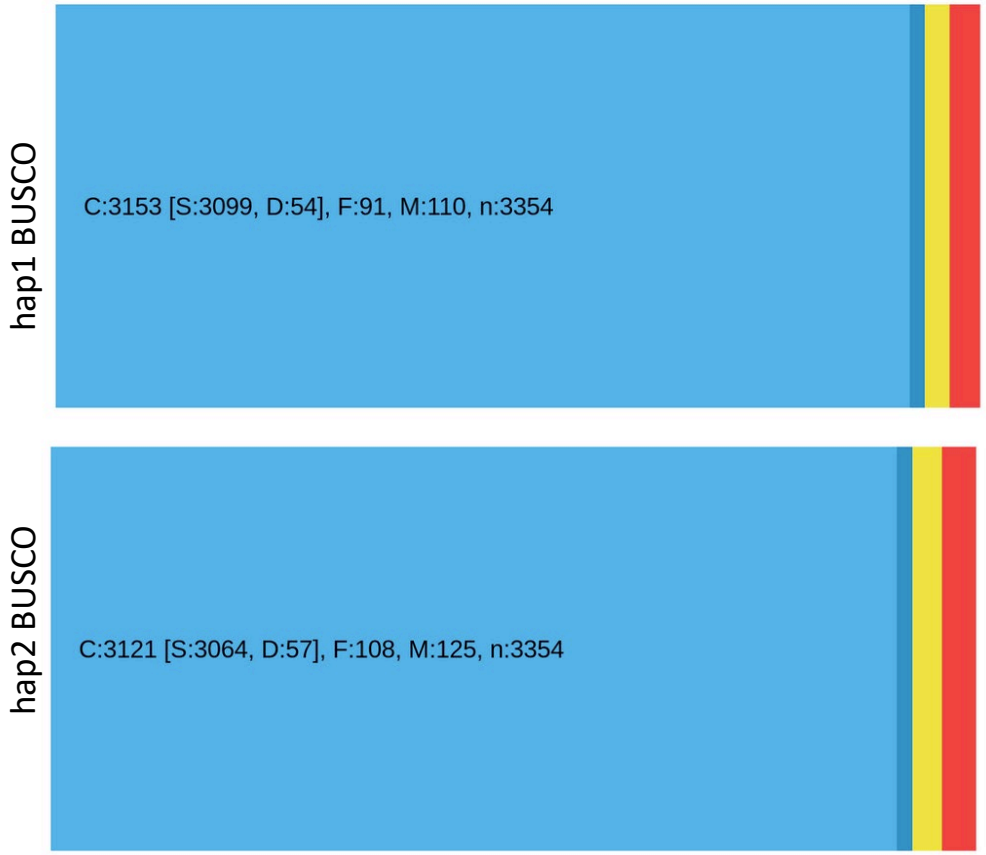
hap1 spectra-cn



hap2 spectra-cn



	Hap1	Hap2
# of contigs	2,511	2,172
Total length	4,339,321,113	4,302,114,182
N50	5,289,116	5,035,338
L50	232	240



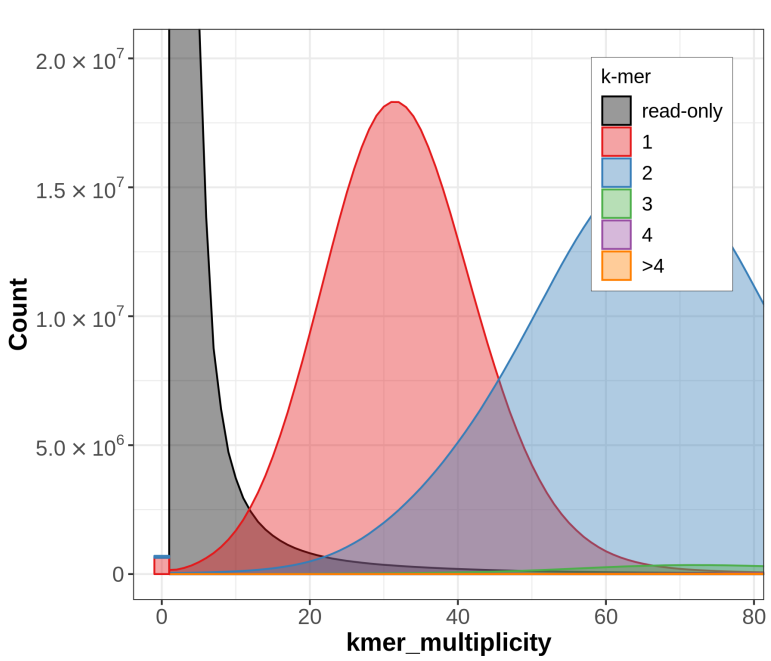
- Here’s the merqury plots for the previous zebra finch examples, in case anyone’s into that

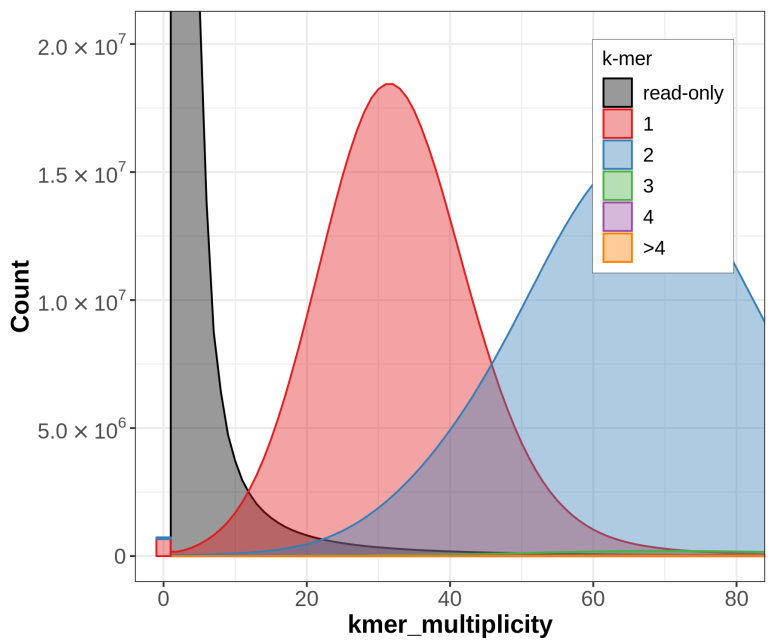
Trio

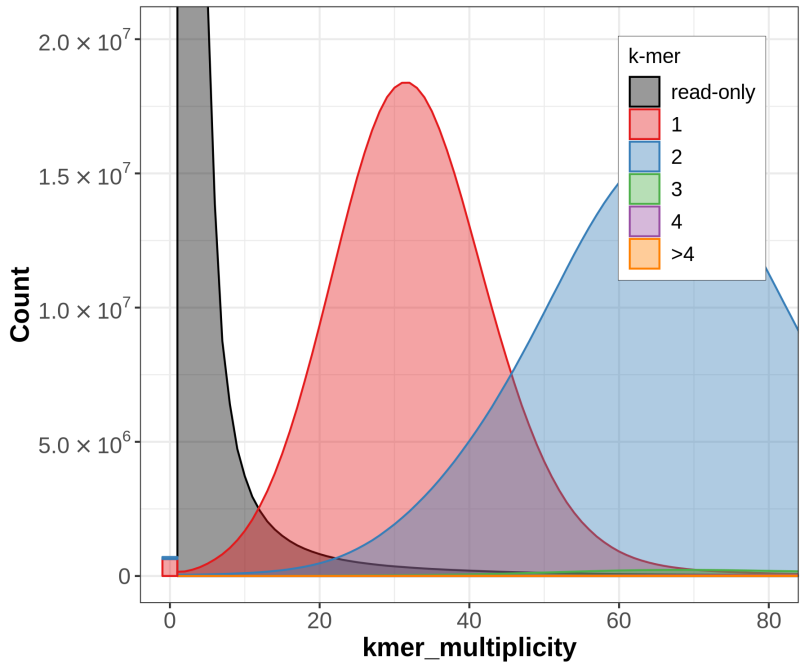
Unphased (pri/alt)

Hi-C

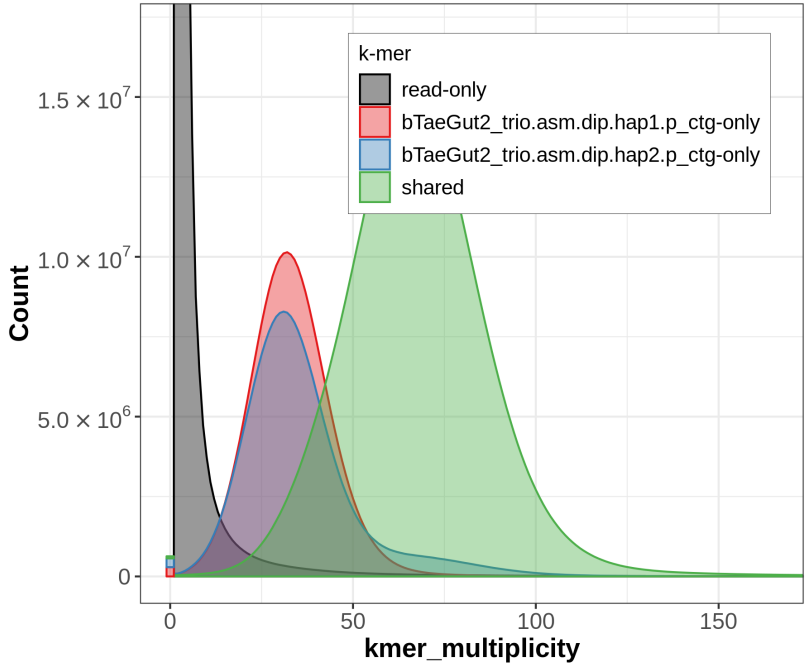
spectra-cn

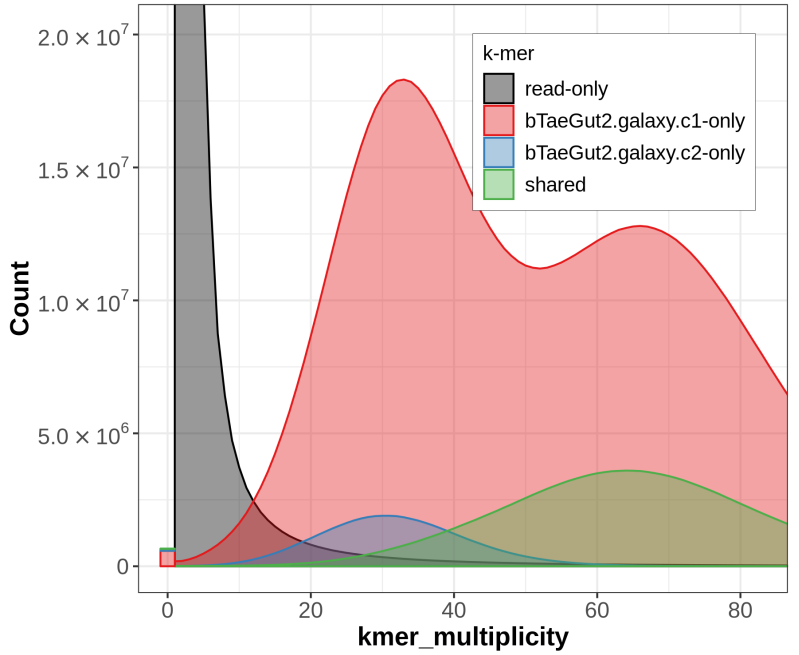


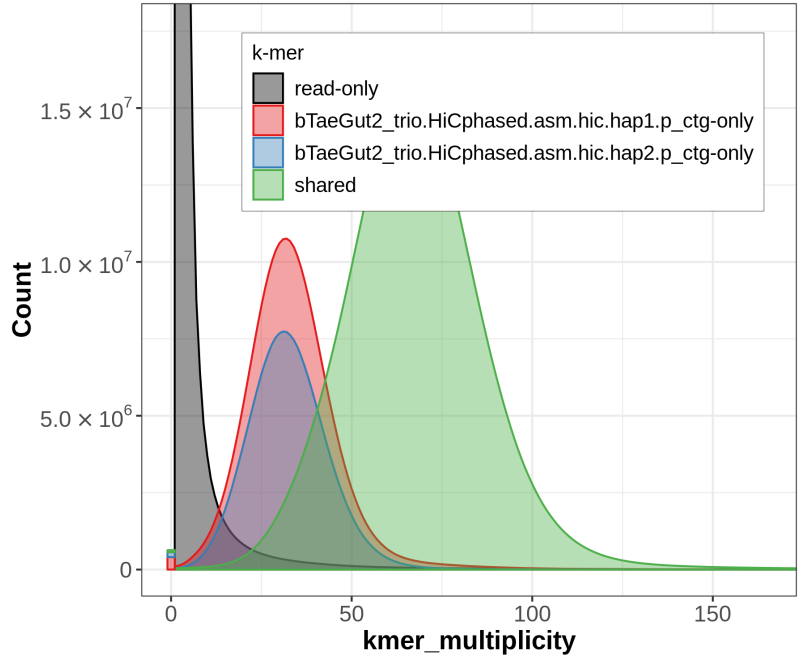


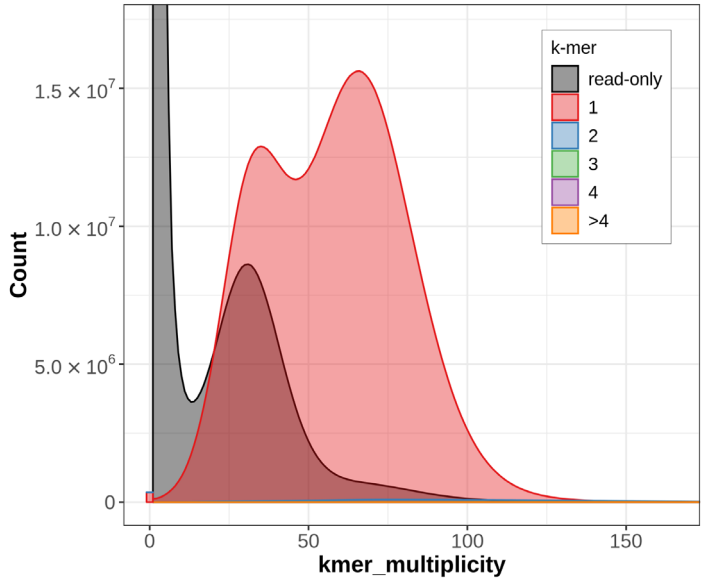


spectra-asm

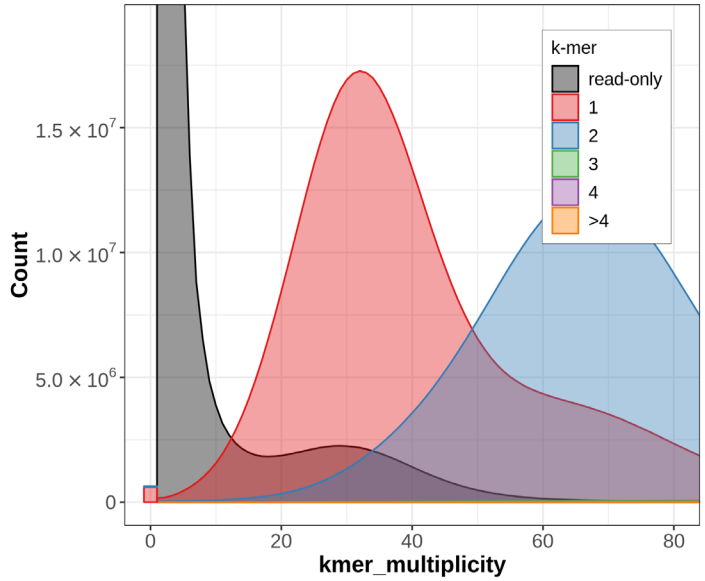




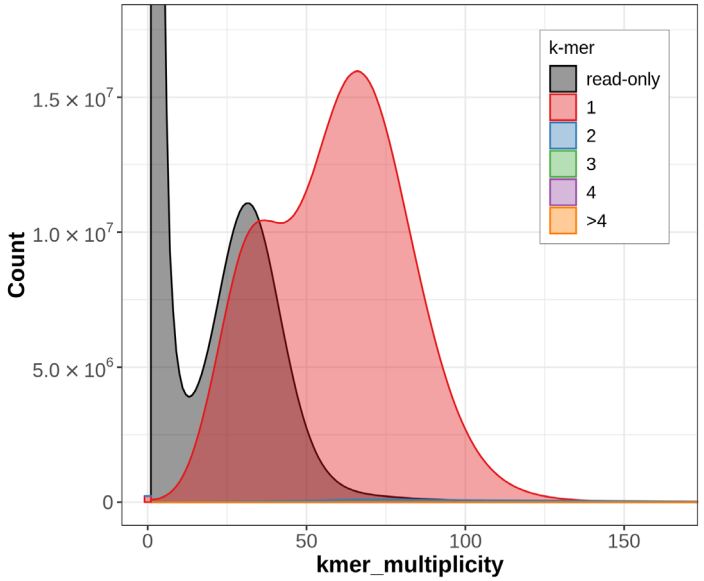




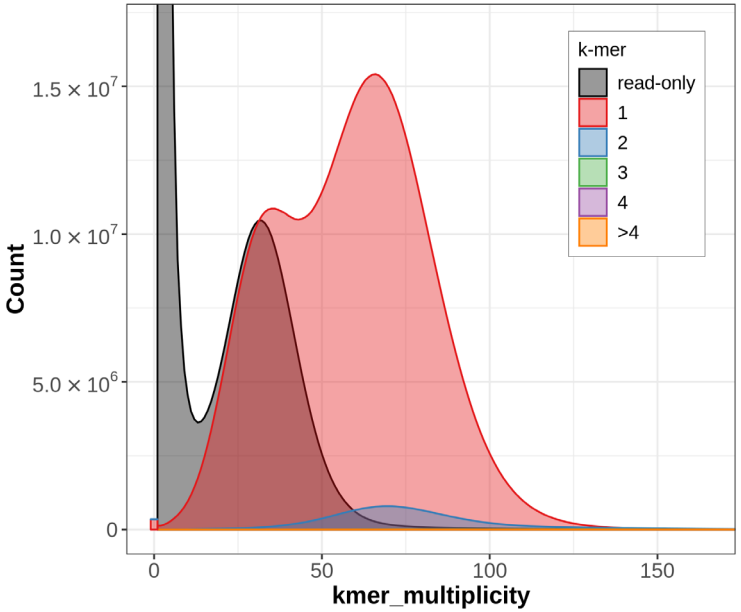
Hap1 (pat)



C1 (pri)

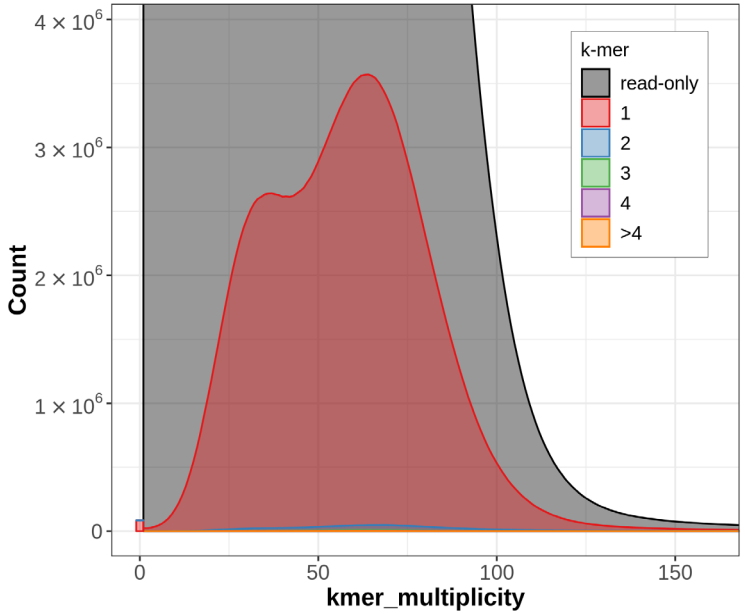


Hap1



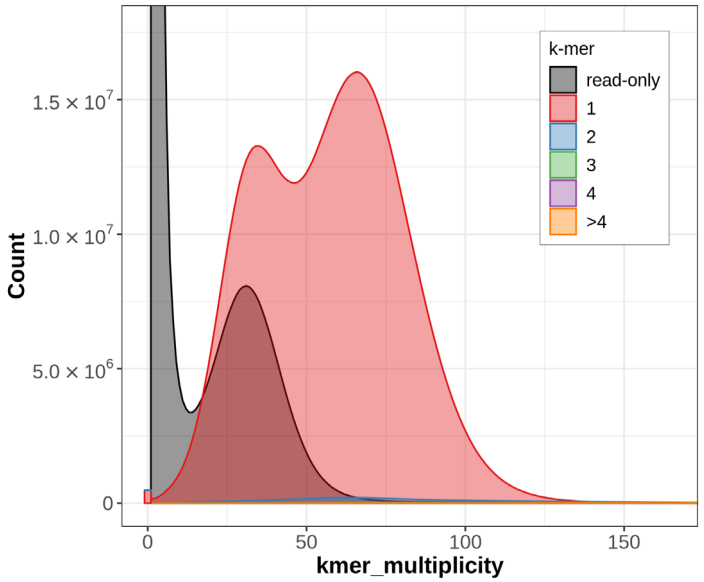
Hap2 (mat)

Trio asm



C2 (alt)

Unphased asm



Hap2

Hi-C asm