K-mer analysis introduction

2023-05-22

Background – what can it be used for?

- Error levels
- Sequencing biases
- Completeness of sequencing coverage
- Contamination
- Metagenomics
- Identify problematic datasets
- Genomic complexity
  - Size
  - Karyotype
  - Levels of heterozygosity
  - Repeat content
- Guiding in choosing assembly parameter settings
- Transcriptomics

and more…

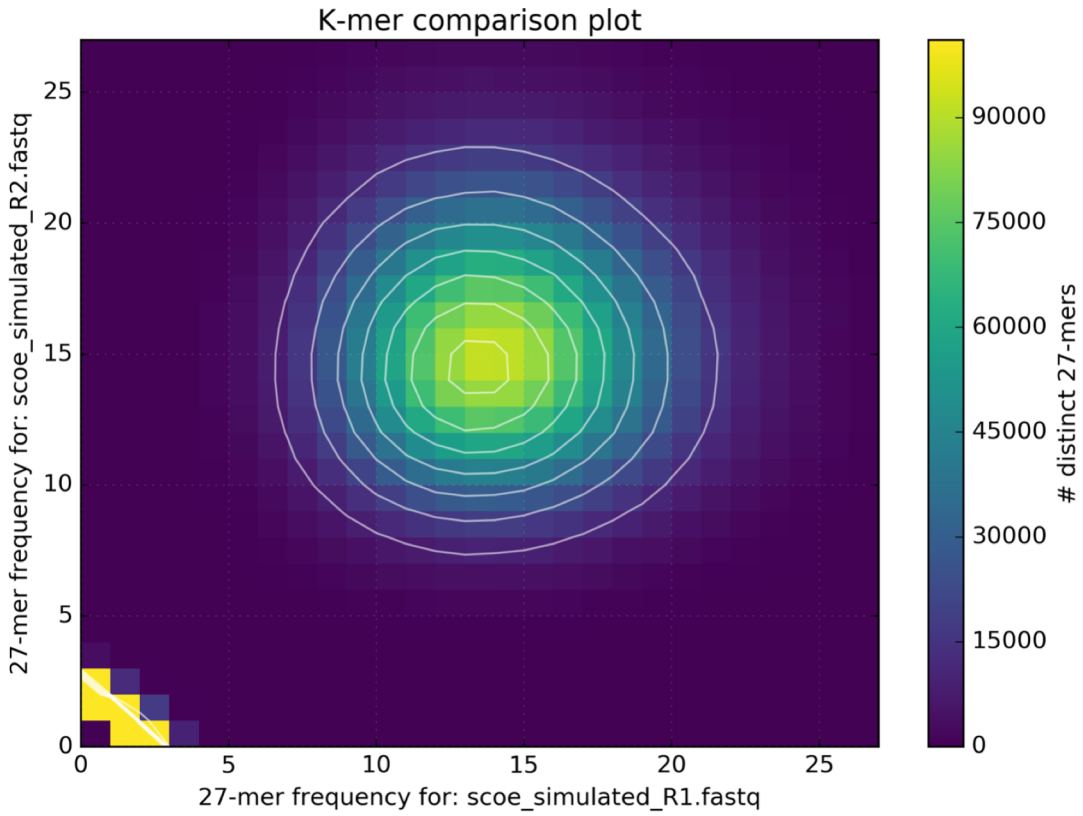
Tools - kmer counting

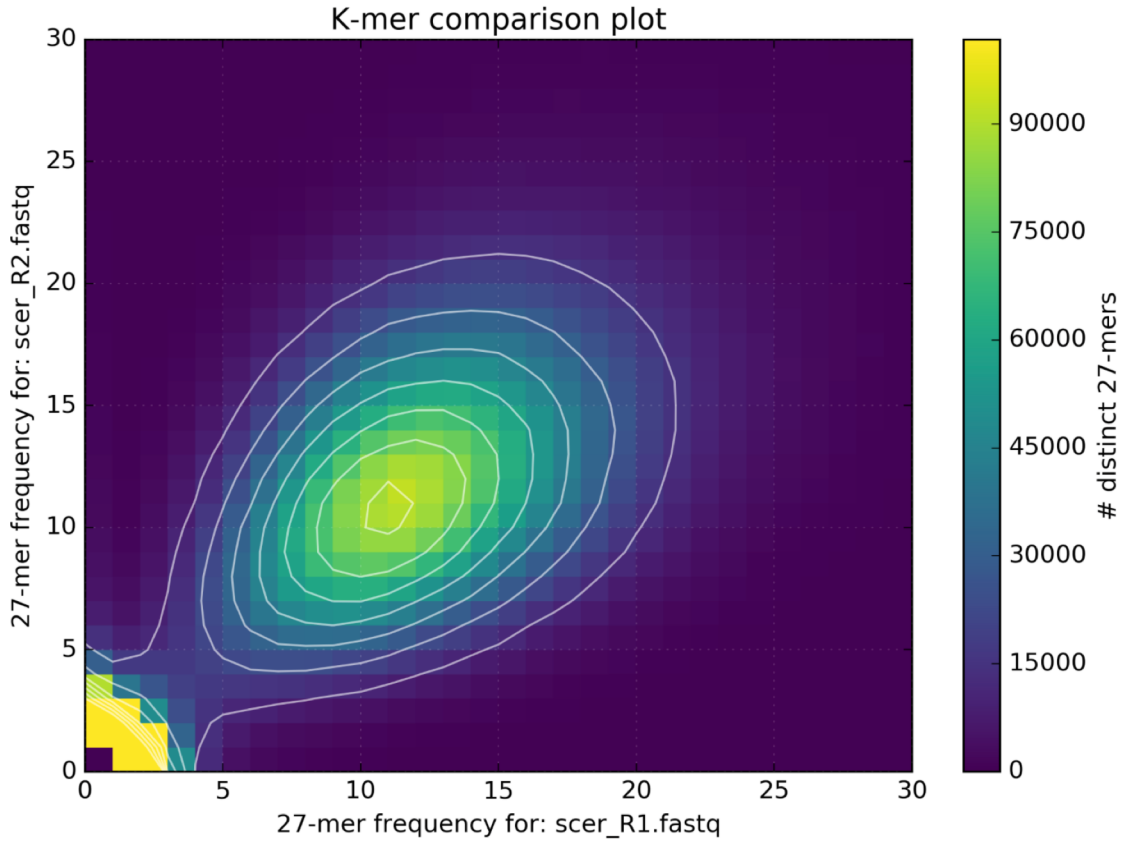
***k*-mer counting tool (year of publication) [citation]**	**(i) Data structure (ii) Algorithm (sorting/hashing, disk-based/in-memory)**	**Limit on *k* size**	**Supports online *k*-mer frequency retrieval?**	**Supports compressed file processing?**
KMC3 (2017) [37]	(i) Array, priority queue (ii) Sorting, disk-based approach, modified MSP (signature), radix sort, counting sort	Arbitrary large *k*-mer lengths	No	Yes
Gerbil 1.0 (2017) [31]	(i) Array (ii) Hashing, disk-based, modified MSP (signature), GPU implementation	< 521	No	Yes
GenomeTester4 (2015) [38]	(i) Custom hybrid data structure, array (ii) Sorting, disk-based	< 33	No	No
KCMBT 1.0 (2016) [43]	(i) Burst tries (ii) Sorting, in-memory, radix sort	< 33	Yes	No
MSPKmerCounter 0.1 (2015) [32]	(i) Hash table (ii) Hashing, disk-based, MSP (minimizer)	Arbitrary large *k*-mer lengths	Yes	No
Turtle 0.3 (2014) [42]	(i) Pattern block Bloom filter, array (ii) Sorting, in-memory, SAC	< 65	No	No
KAnalyze 2.0.0 (2014) [40]	(i) Array (ii) sorting, disk-based, dual-pivot quick sort (Java's Arrays.sort())	Arbitrary large *k*-mer lengths	No	Yes
DSK 2.2.0 (2013) [33]	(i) Hash table (ii) Sorting, disk-based approach, radix sort	Arbitrary large *k*-mer lengths	No	Yes
Jellyfish 2.2.6 (2011) [35]	(i) Hash table (ii) Hashing, in-memory, Jellyfish-BF(mode) (i) Bloom filter (ii) Disk-based	Arbitrary large *k*-mer lengths	Yes	No
BFCounter 1.0 (2011) [36]	(i) Bloom filter, hash table (ii) Hashing, in-memory	Arbitrary large *k*-mer lengths	No	Yes
Squeakr (2017) [34]	(i) CQF (ii) Hashing, in-memory	Arbitrary large *k*-mer lengths	Yes	Yes
Tallymer (2008) [22]	(i) Enhanced suffix arrays (ii) Counting using the lcp-interval tree, in-memory	Arbitrary large *k*-mer lengths	Yes	No

https://doi.org/10.1093/gigascience/giy125

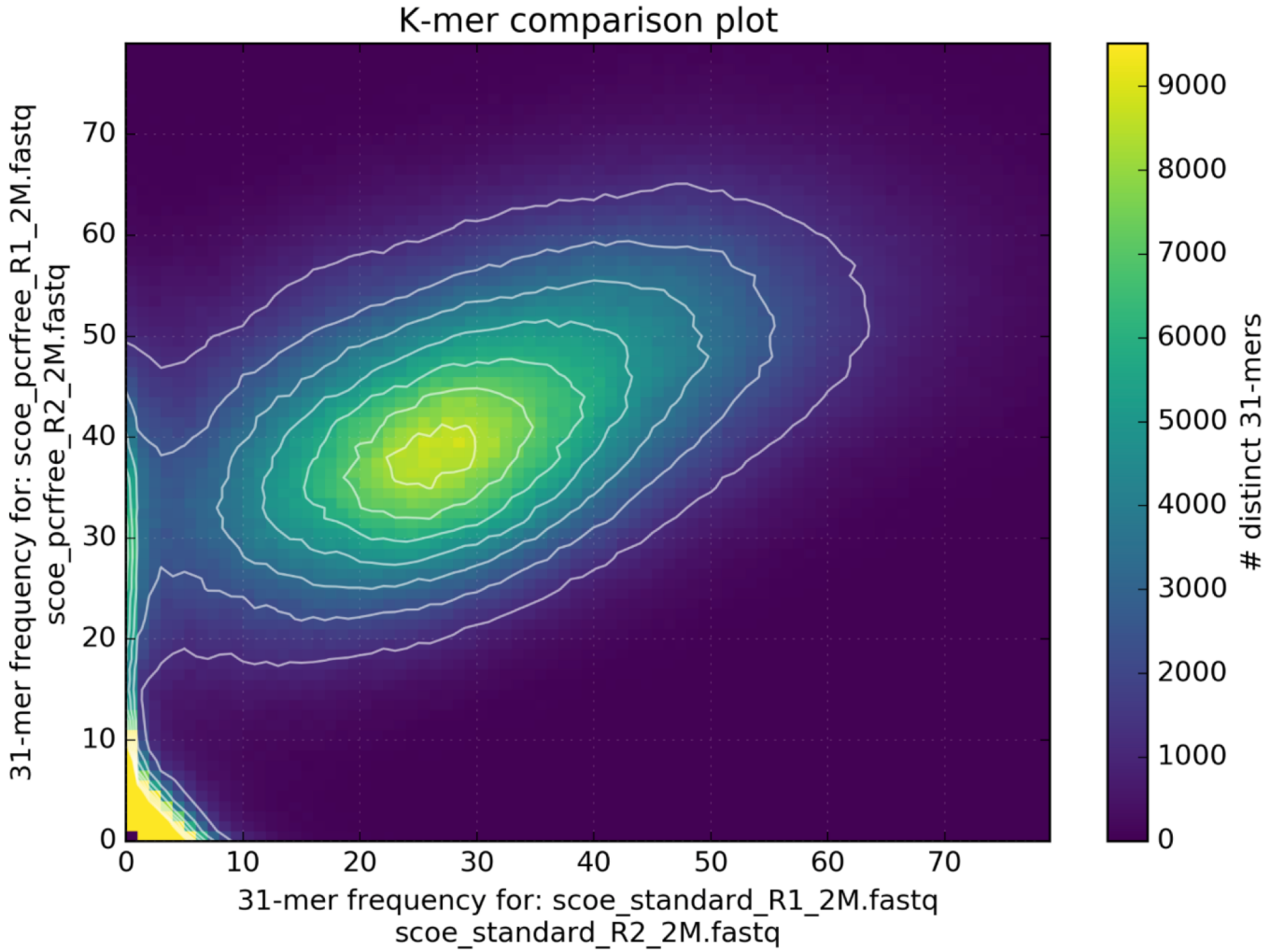
What does it look like?

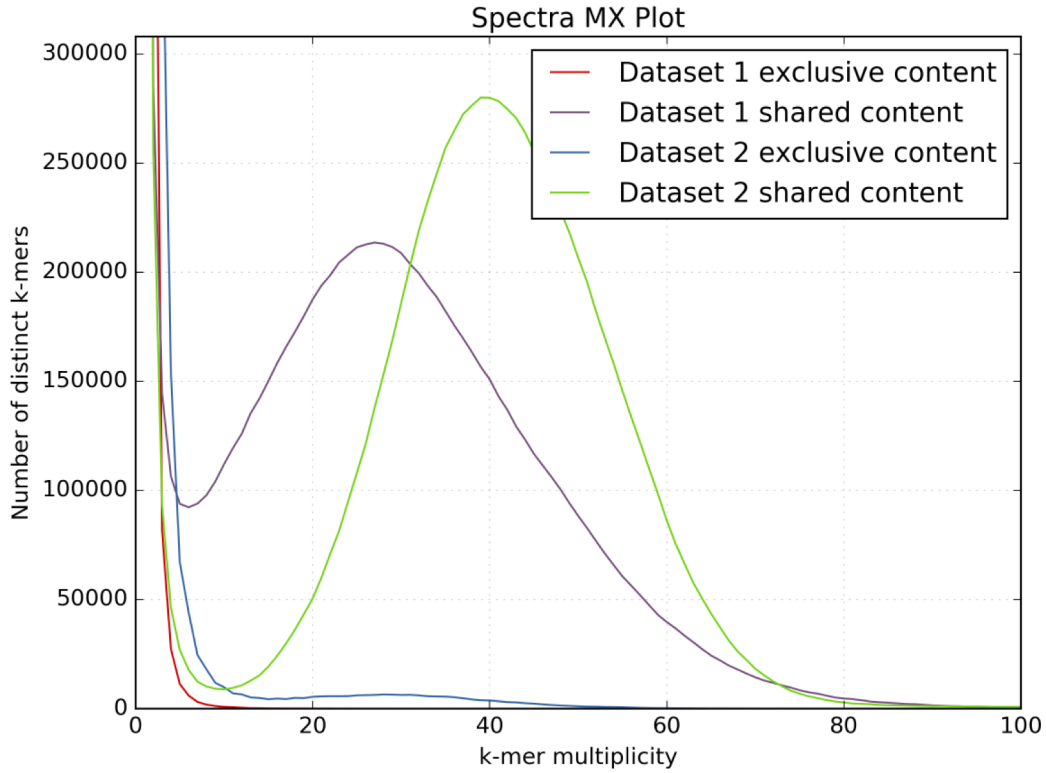
Basic QC





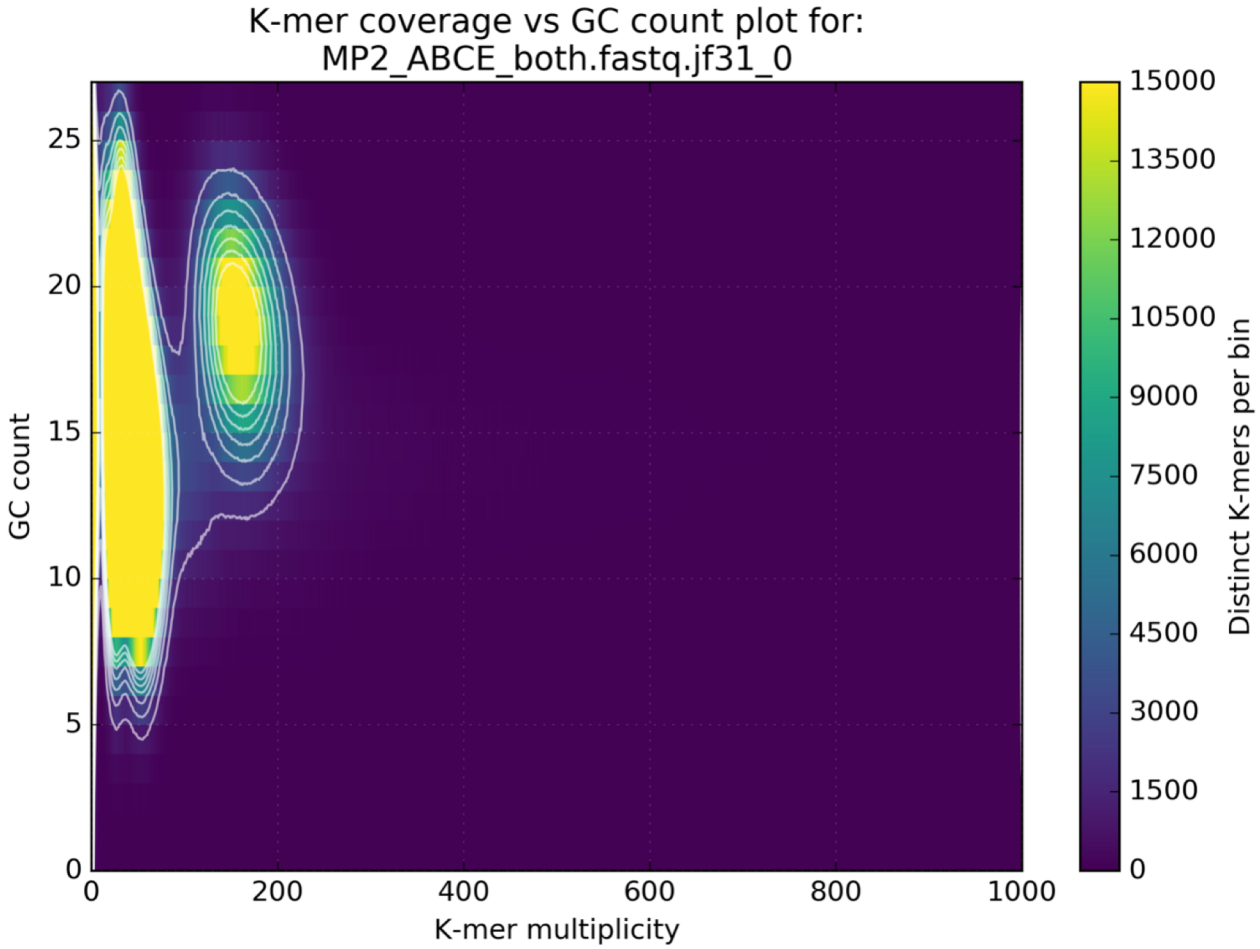
What does it look like?





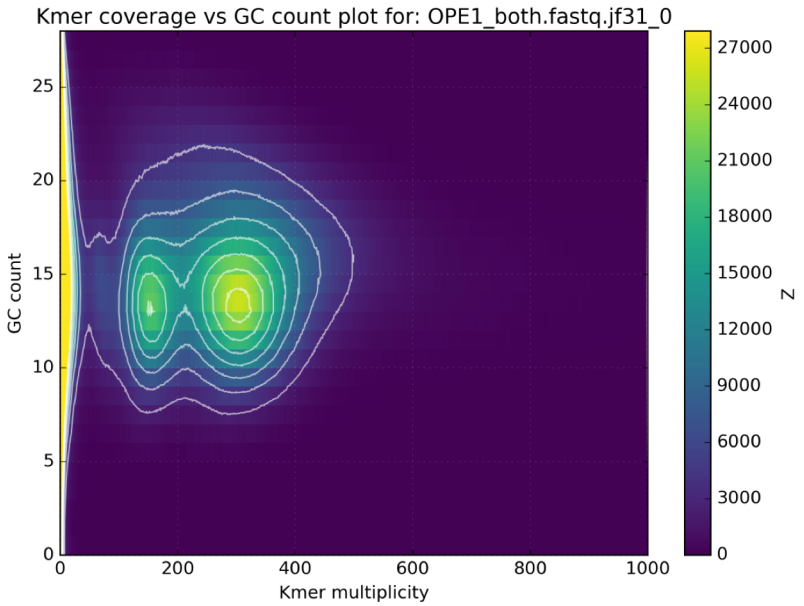
What does it look like?

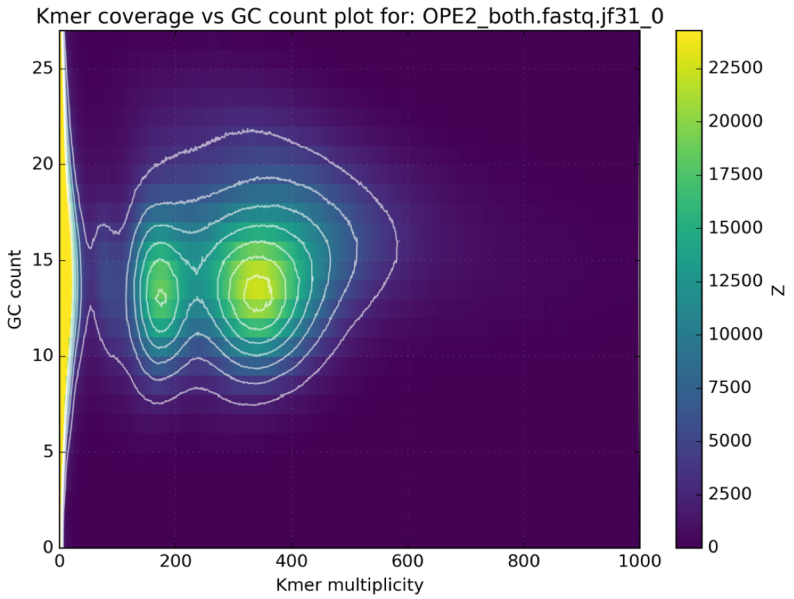
Contamination?

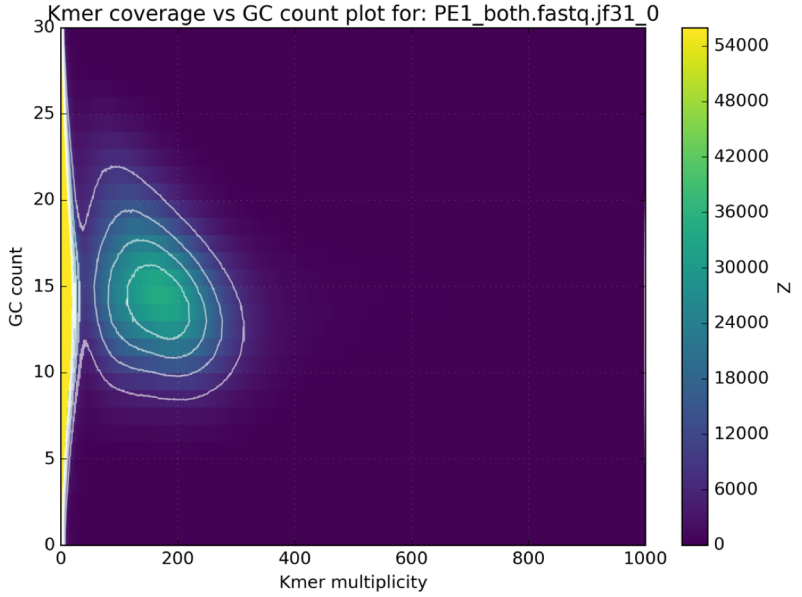


What does it look like?

Contamination? More libraries!

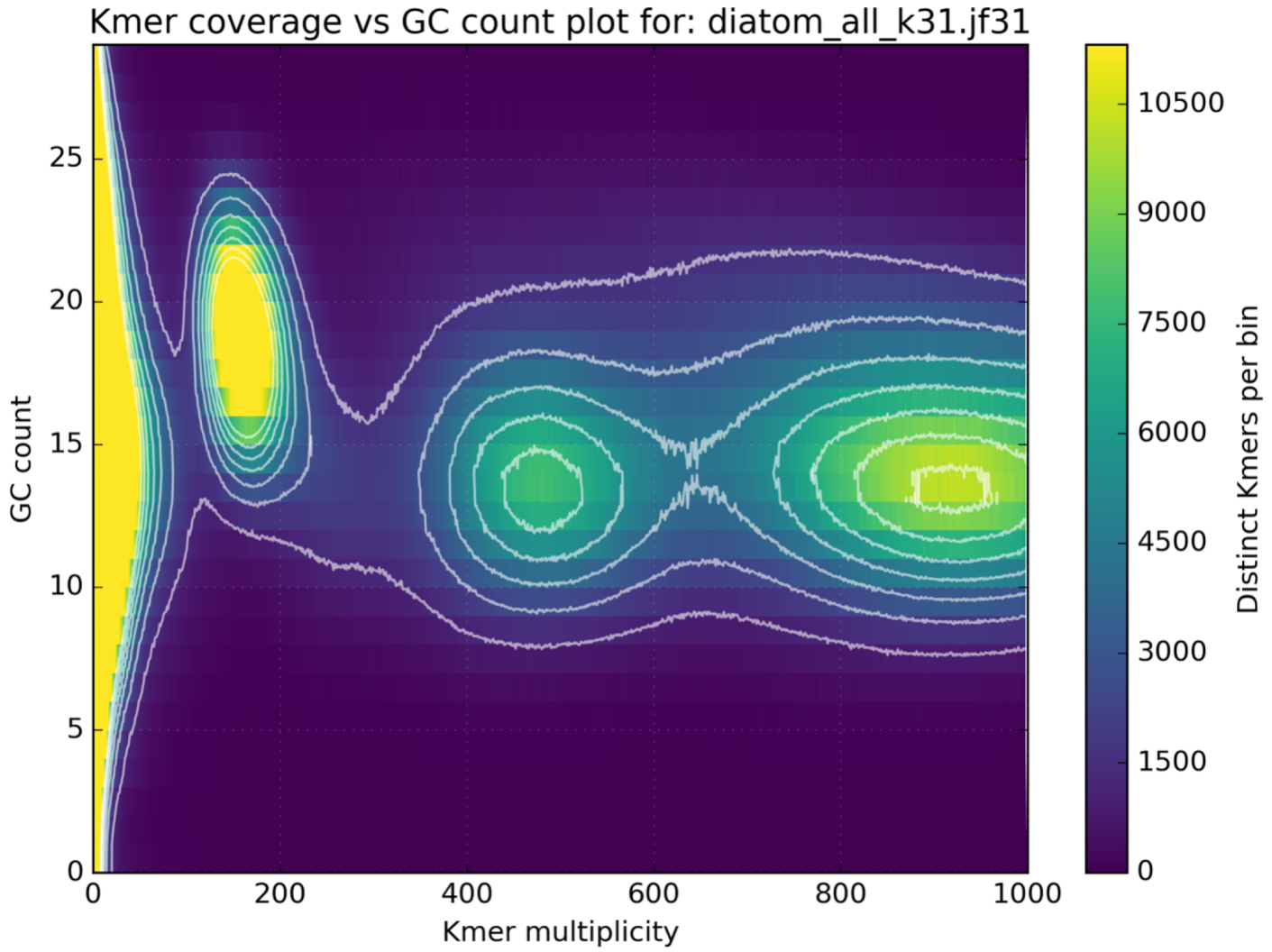






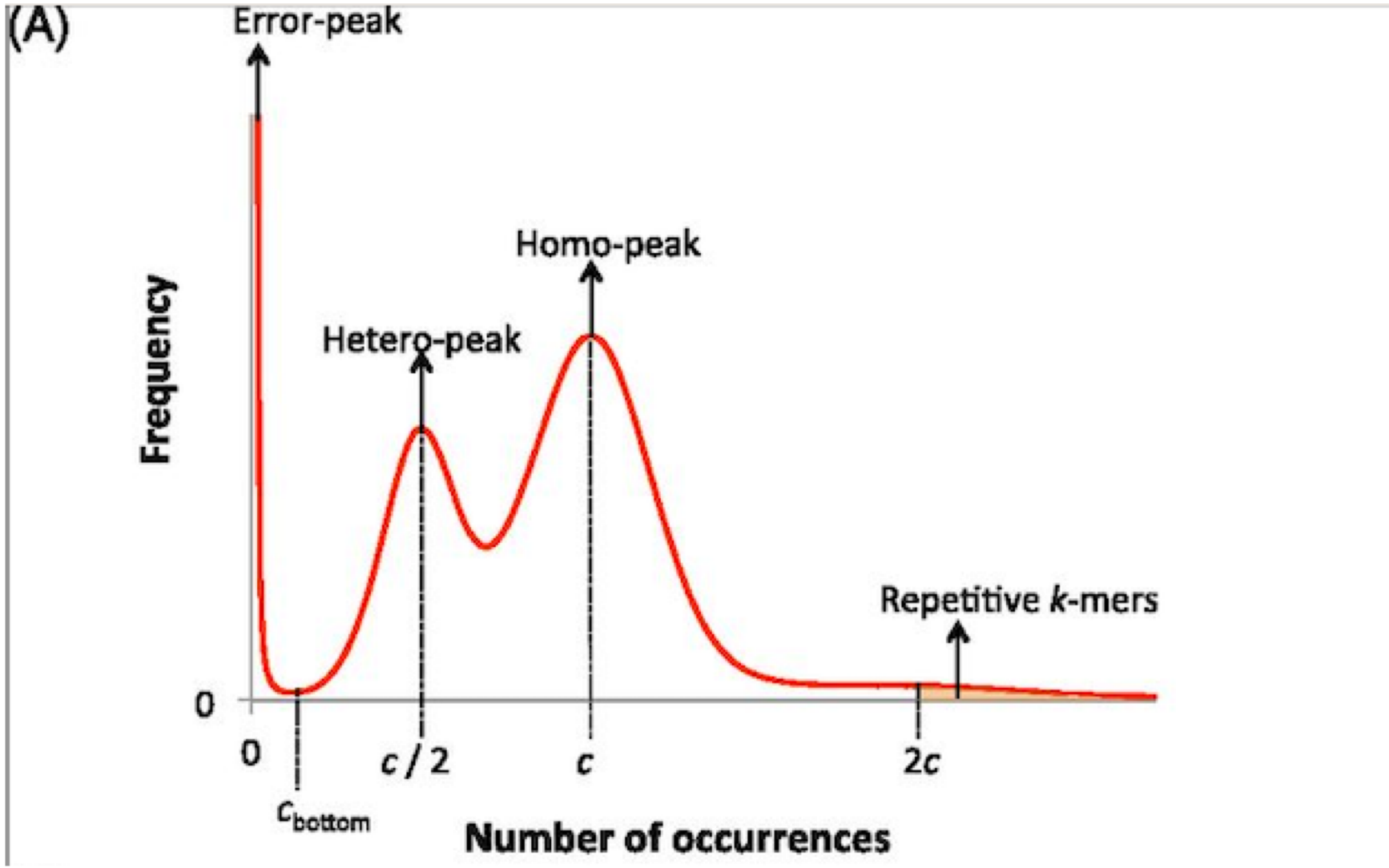
What does it look like?

Contamination? Merge of libraries!



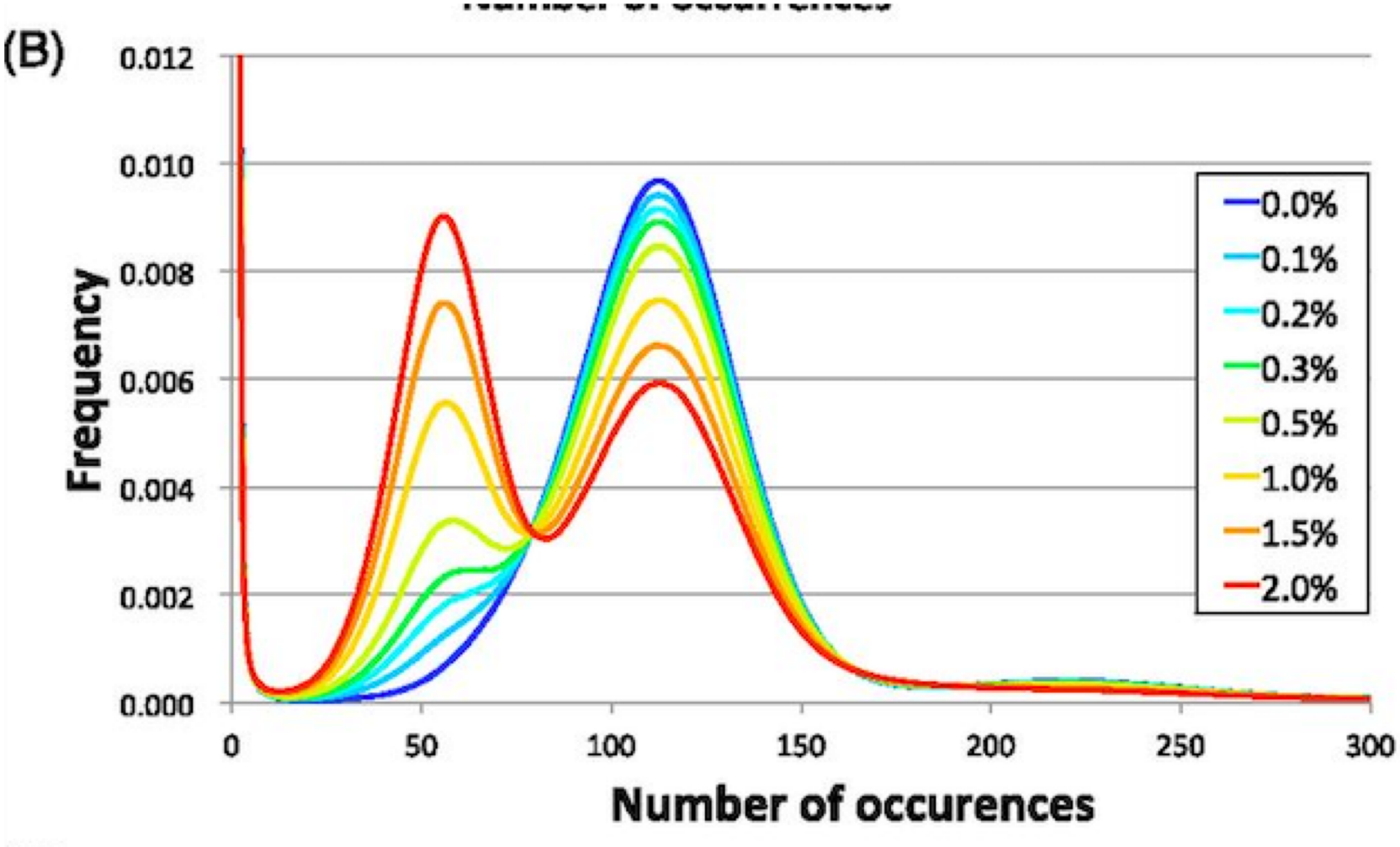
What does it look like?

heterozygosity



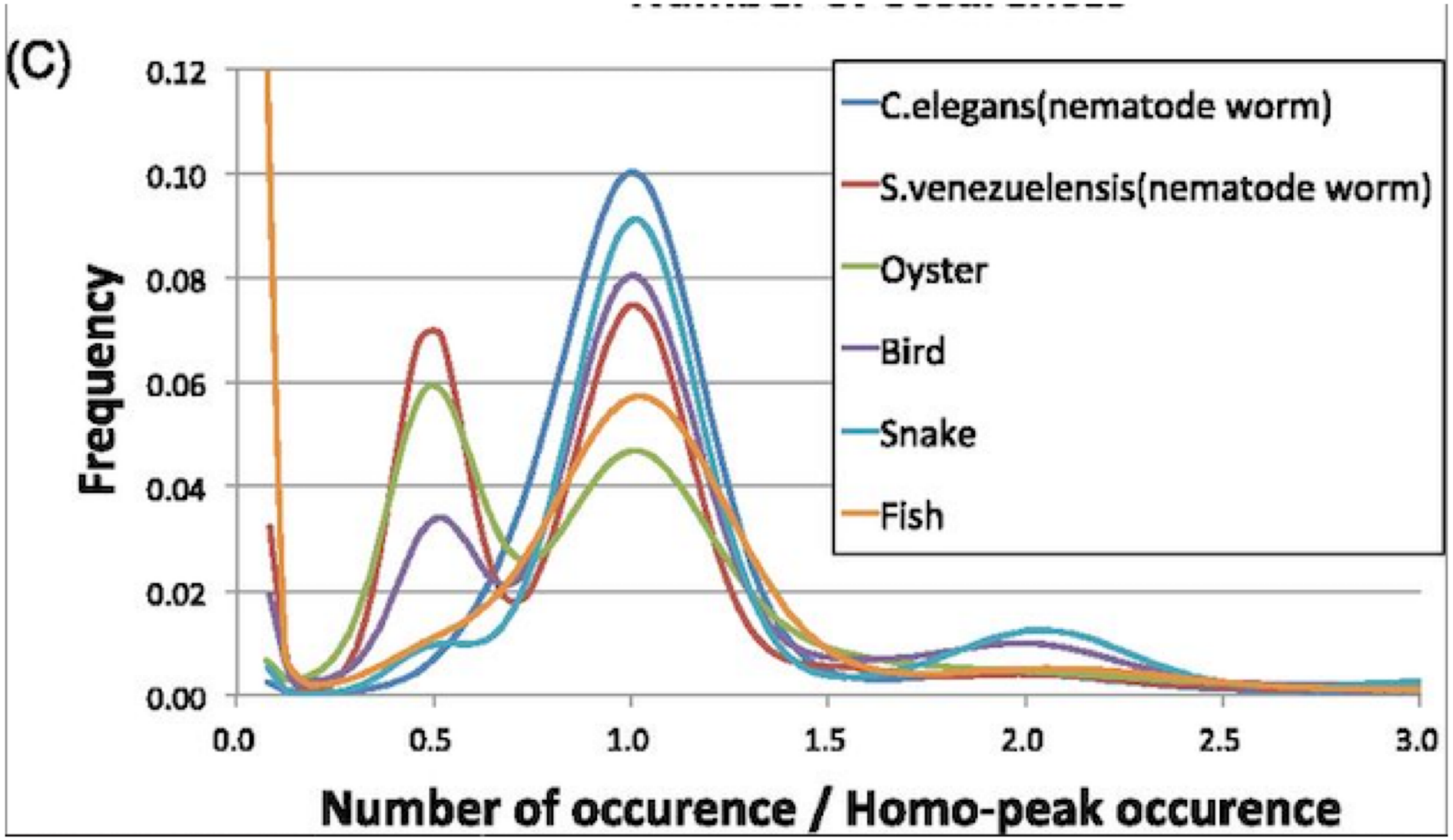
What does it look like?

heterozygosity



What does it look like?

heterozygosity



What does it look like?

- Genome size estimates