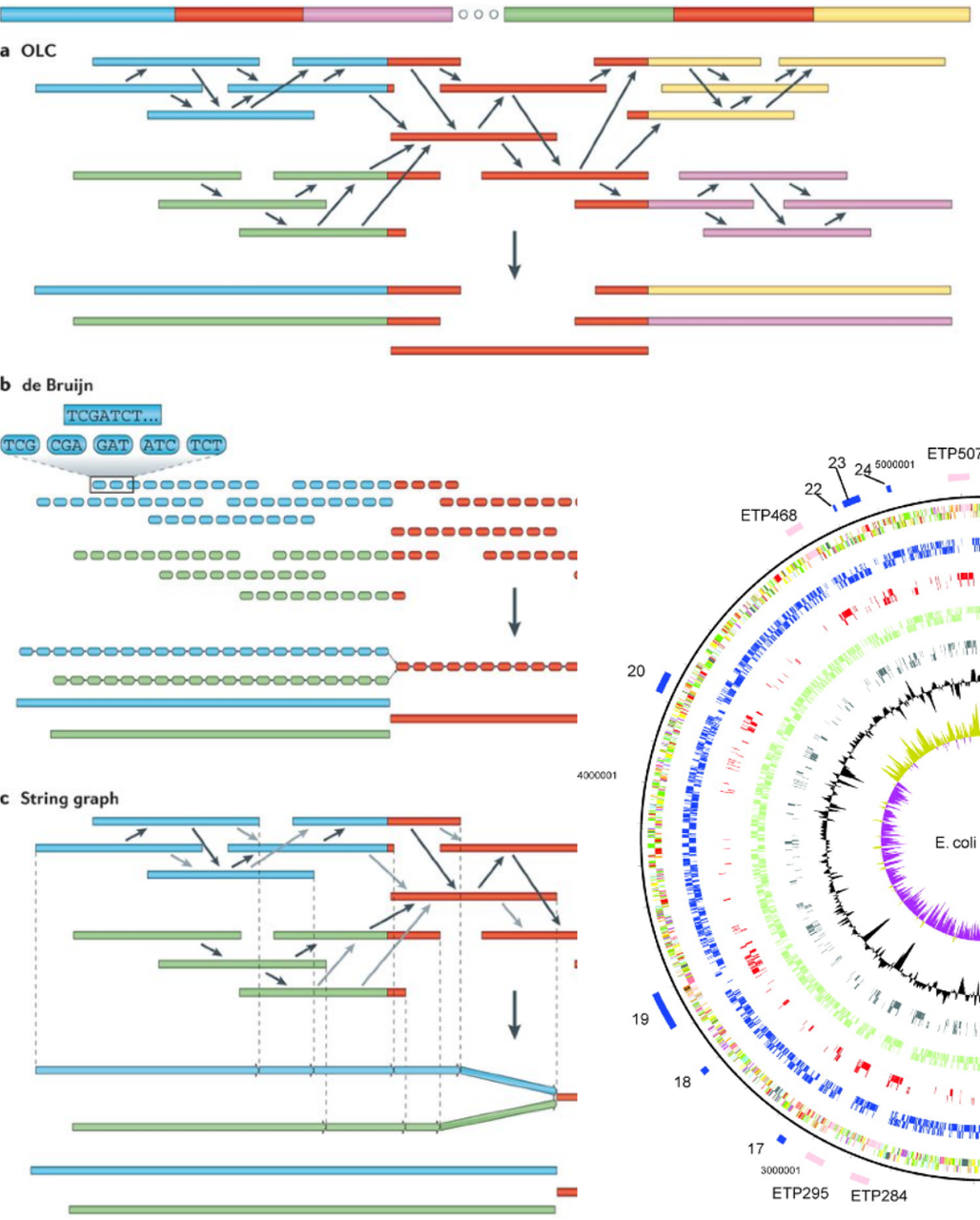
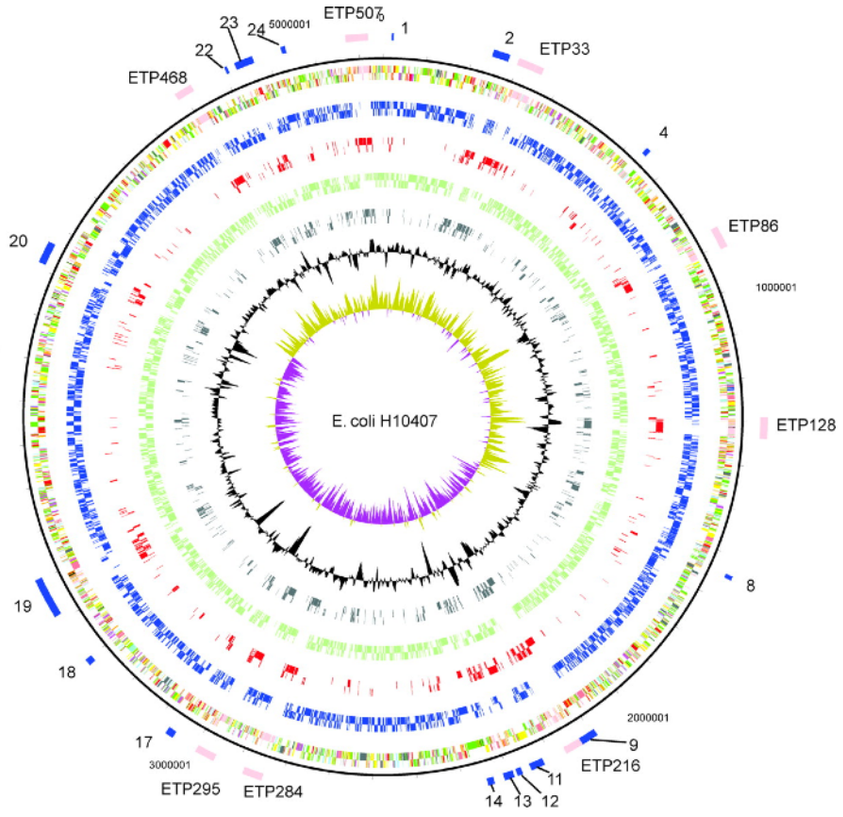
Assembly Quality Control





HTS Workshop Genomics & Transcriptomics KAUST 2019

Robert Lehmann
Octavio Salazar

Expected Assembly Size

"*k*-mer" is a substring of length *k*

S: GGCGATTCATCG

All 3-mers of *S*:

GGC
GCG
CGA
GAT
ATT
TTC
TCA
CAT
ATC
TCG

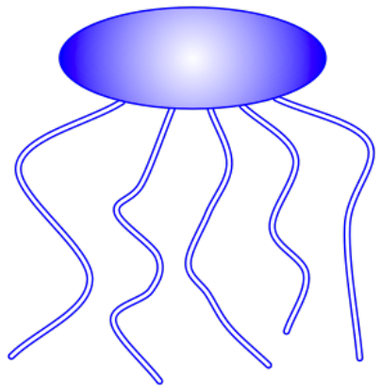
Genome Length L = 12

How many k-mers can it contain?

$n = (L - k) + 1$

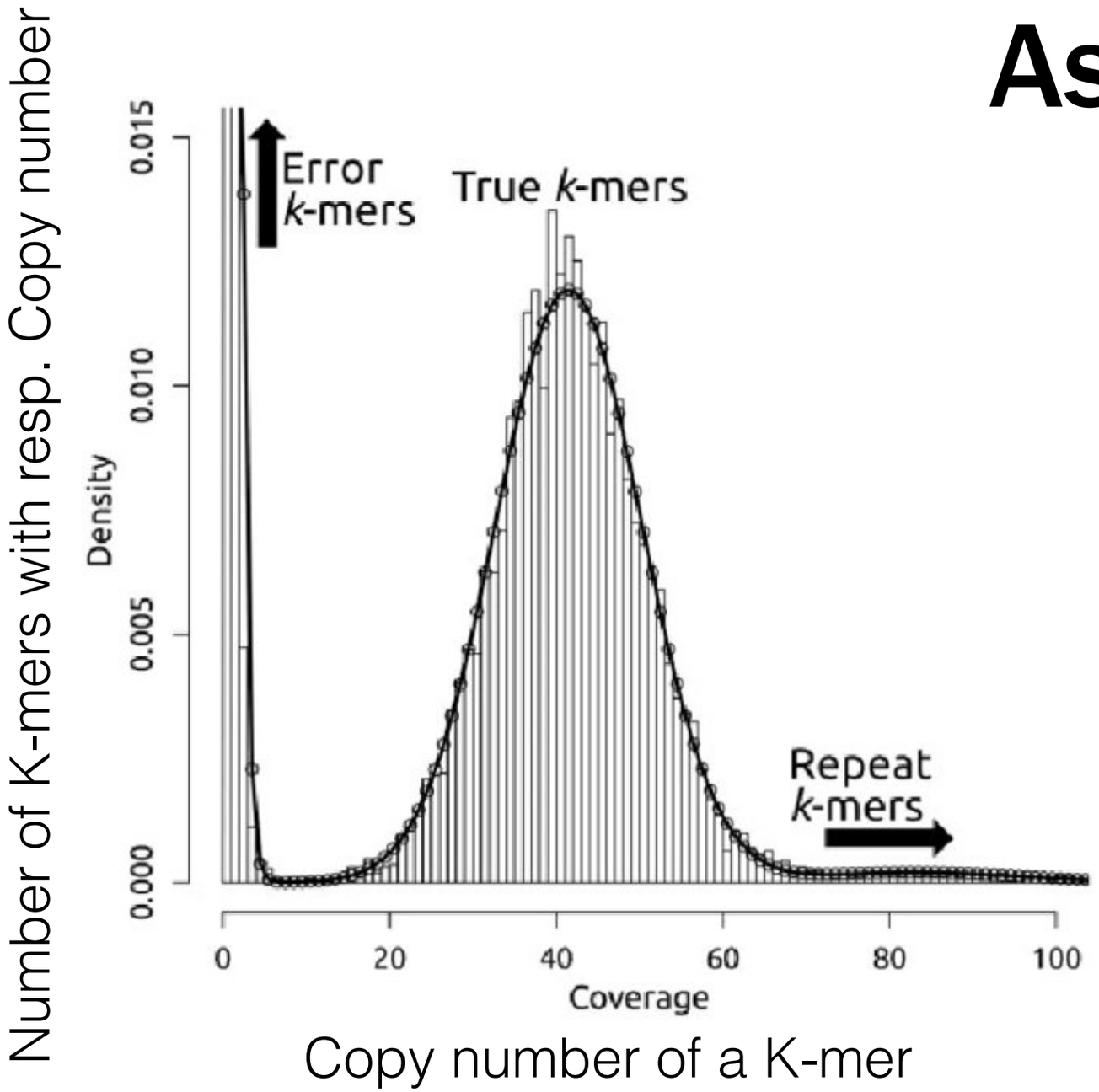
-> number of k-mers provides estimate of genome size with small error (0.0017% for 1Mb and k = 18)

https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/



https://www.cbcb.umd.edu/software/jellyfish/

Expected Assembly Size



Copy number of a K-mer

Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Laehnemann D, Borkhardt A, McHardy AC. Brief Bioinform. 2016 Jan;17(1):154-79

- underlying genome sequenced > 1x coverage

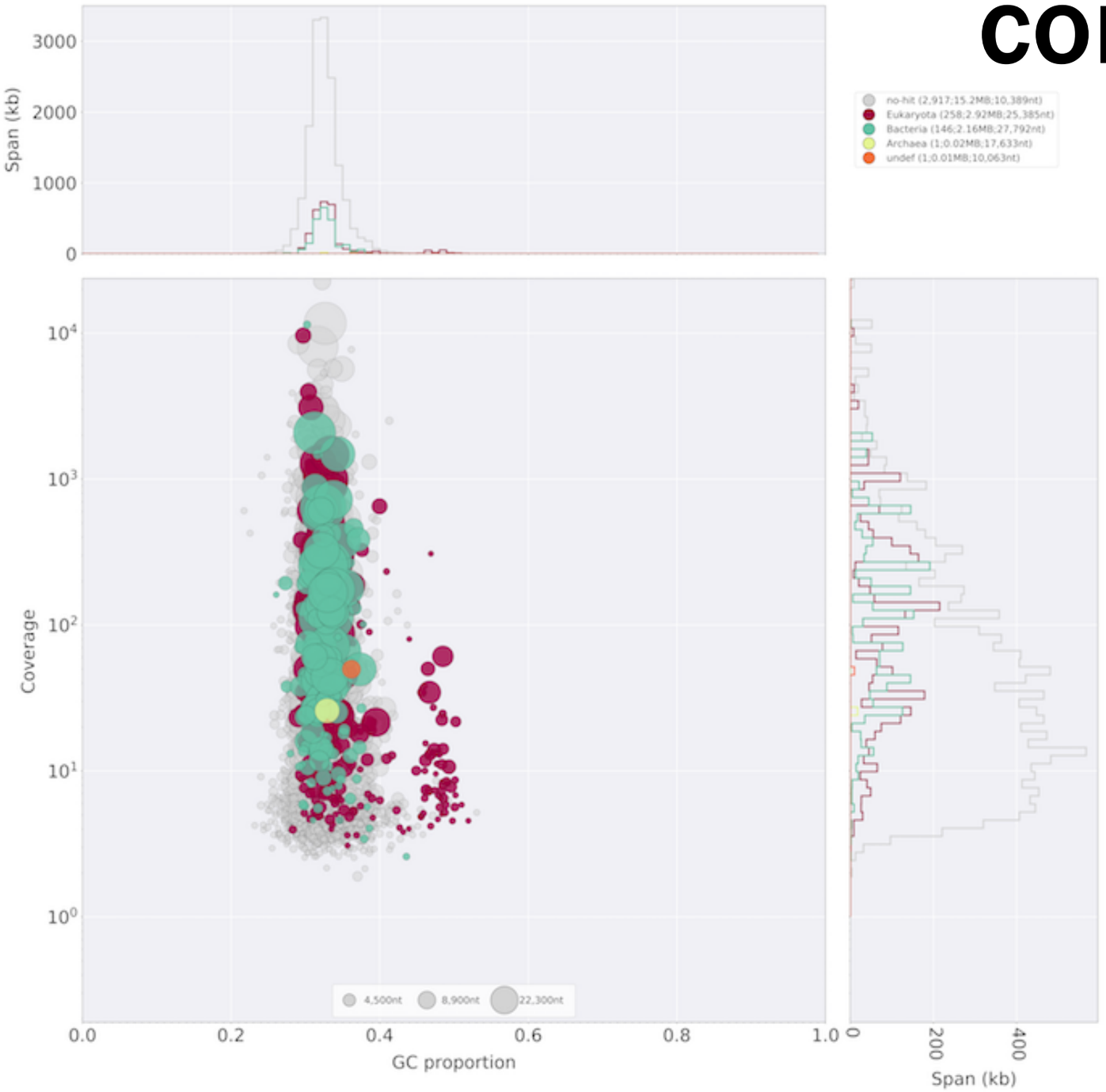
$$n = ((L - k) + 1) * C$$

$$C = \text{coverage}$$

- sequencing errors and repetitive sequence introduce rare and highly frequent K-mers
- does estimate fit to C-value estimate?

https://blobtools.readme.io/docs

Detecting contamination



Contamination:

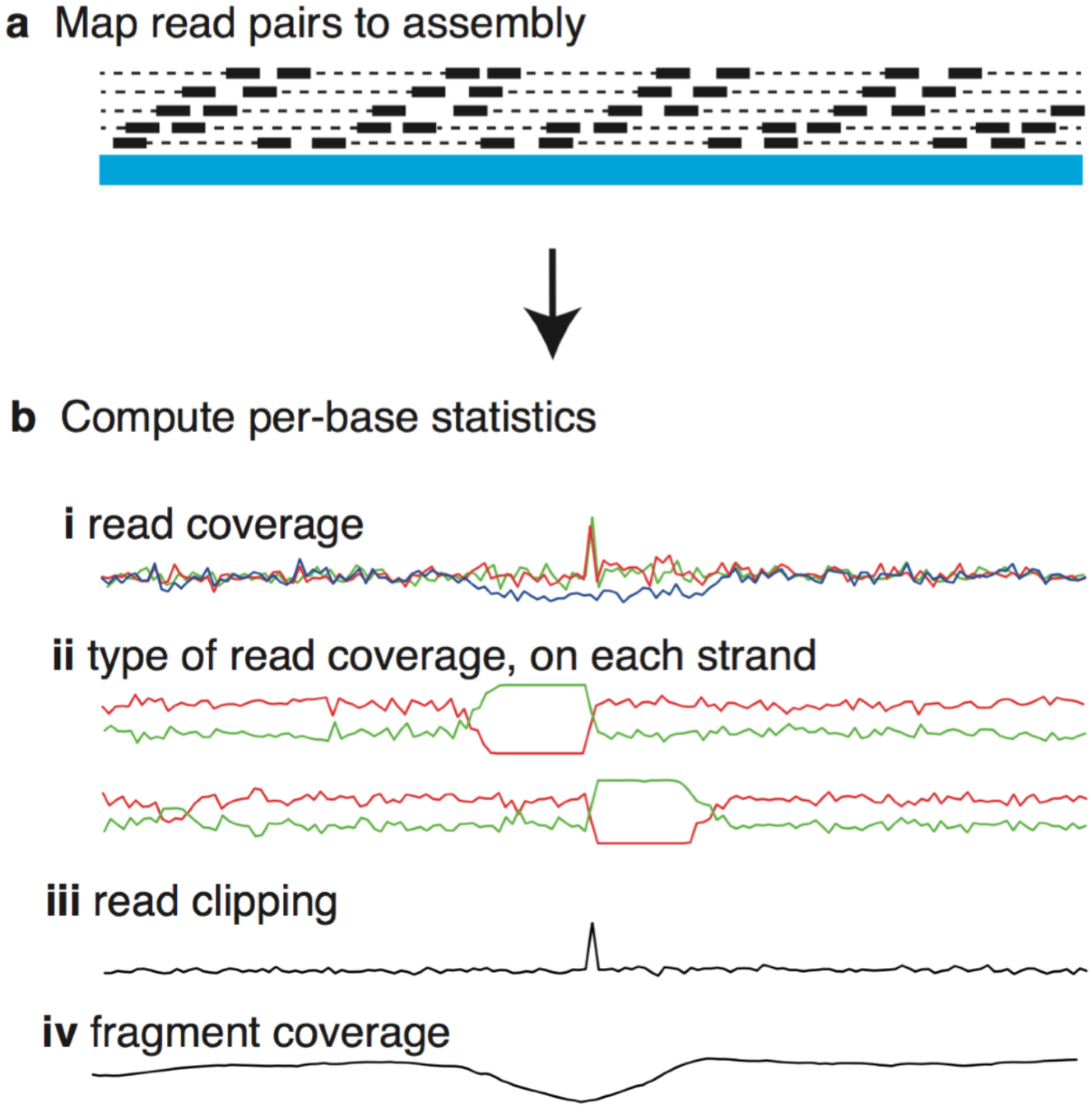
- also K-mer coverage vs. GC
- added taxonomic classification (blast against e.g. nr database)

BlobTools: Interrogation of genome assemblies. DR Laetsch, ML Blaxter, F1000 Research, 2017, 6:1287.

Read mapping

REAPR

https://www.sanger.ac.uk/science/tools/reapr



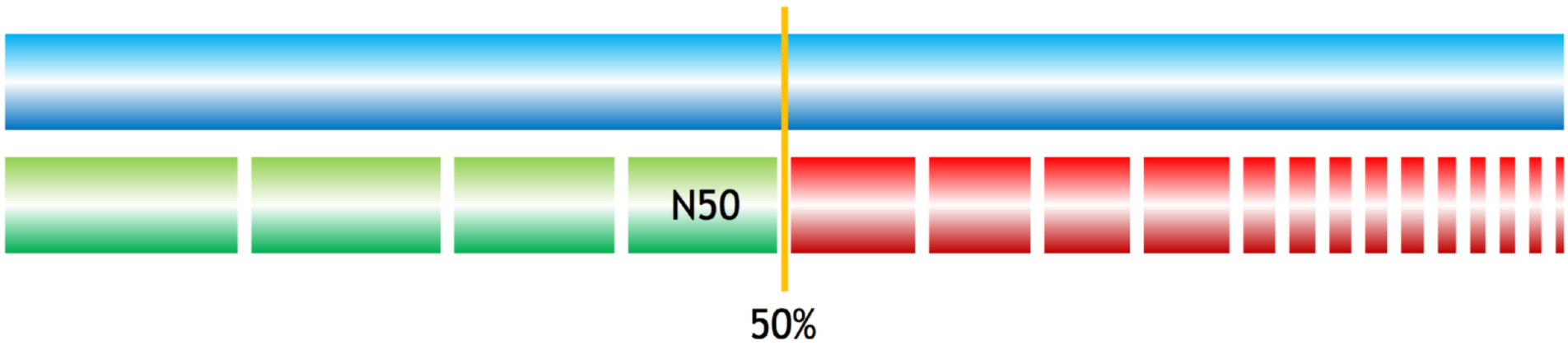
- strand specific coverage
- systematic clipping or do reads align completely
- is fragment coverage following theoretical shape

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., & Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation., 14(5), R47

How to Measure Assembly Quality?

Easy

1. Does length match expectation (c-Value)?
2. How fragmented is the assembly?



3. Are all genes there?

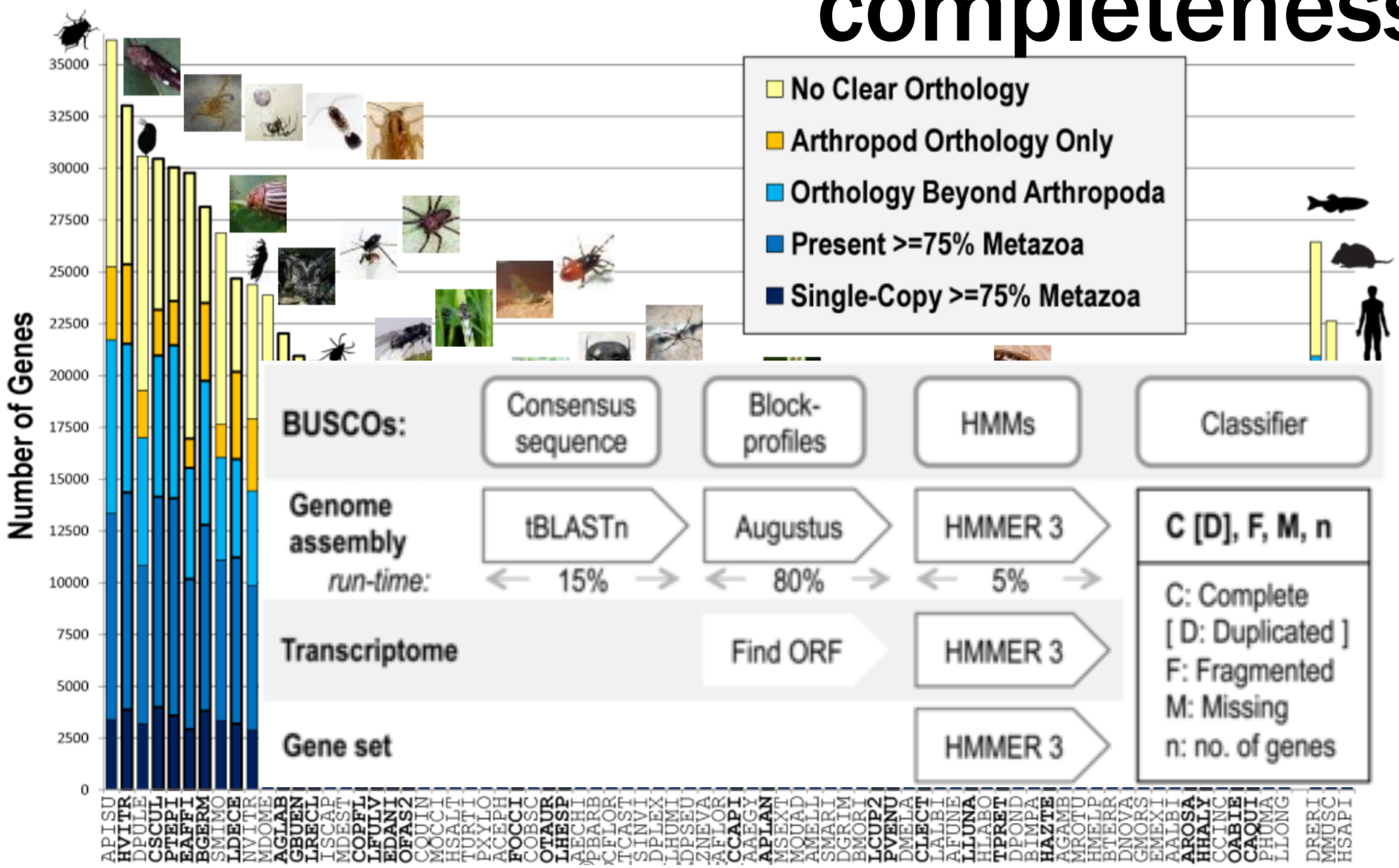
Hard

4. How large are the gaps*? (Linkage Map)

BUSCO v3

Benchmark Universal Single Copy Orthologs

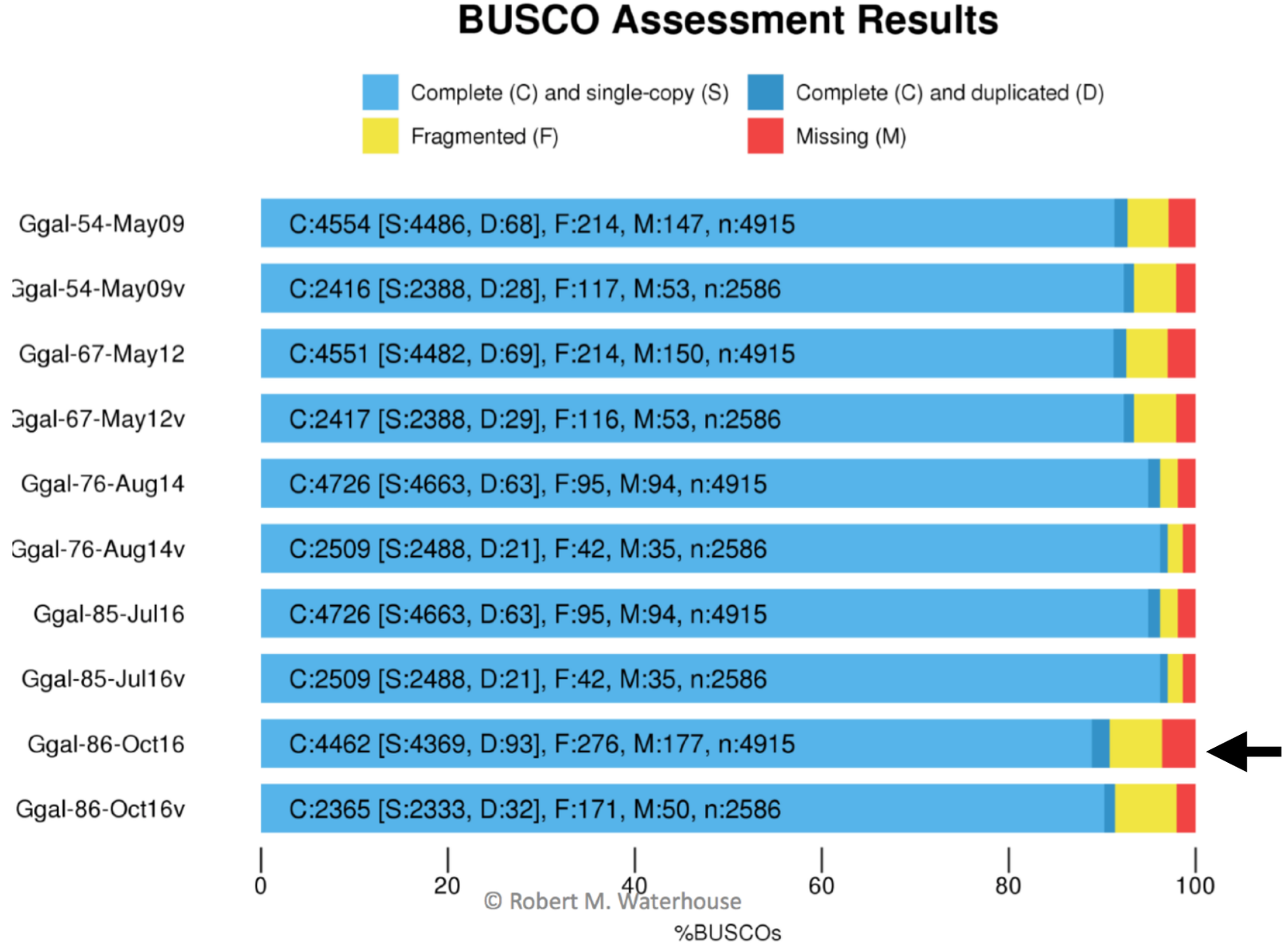
How to measure gene completeness?



Orthologous groups with single gene per species in >90% of considered species

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. FA Simão, RM Waterhouse, P Ioannidis EV Kriventseva, EM Zdobnov. 2015 Bioinformatics, 31-19-1, 3210–3212

Assembly completeness



- not every new assembly is better
- hybrid assembly - long reads, finished BACs, improved physical maps

A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3. Nov. 2016

Questions?