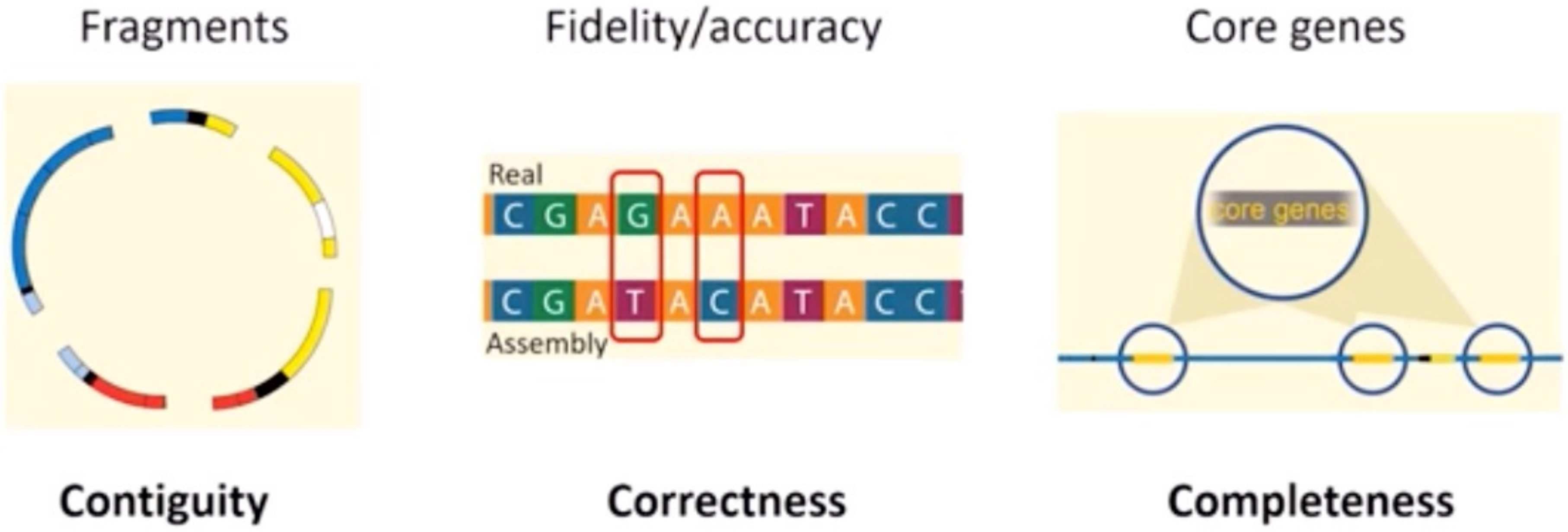
Genome quality control

Is my genome ready for annotation?

The 3C:

- Contiguity
- Completeness
- Correctness

Molina-Mora, J.A., Campos-Sánchez, R.,
Rodríguez, C. *et al.* (2020).
https://doi.org/10.1038/s41598-020-58319-6



https://youtu.be/N7oVyOTGfsk

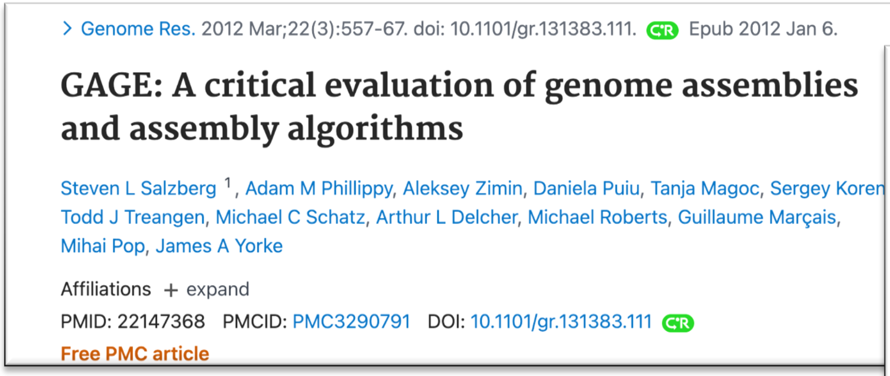
Genome quality control

Contiguity

Metrics:

- Number of contigs
- Average, min and max contigs length
- N50

Tools: QUAST, CLC, etc



> Genome Res. 2012 Mar;22(3):557-67. doi: 10.1101/gr.131383.111. Epub 2012 Jan 6.

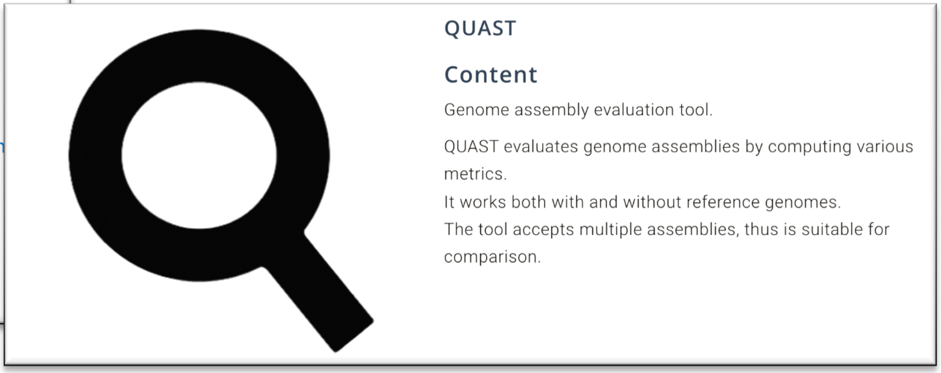
GAGE: A critical evaluation of genome assemblies and assembly algorithms

Steven L Salzberg [1], Adam M Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J Treangen, Michael C Schatz, Arthur L Delcher, Michael Roberts, Guillaume Marçais, Mihai Pop, James A Yorke

Affiliations + expand

PMID: 22147368 PMCID: PMC3290791 DOI: 10.1101/gr.131383.111

Free PMC article



QUAST

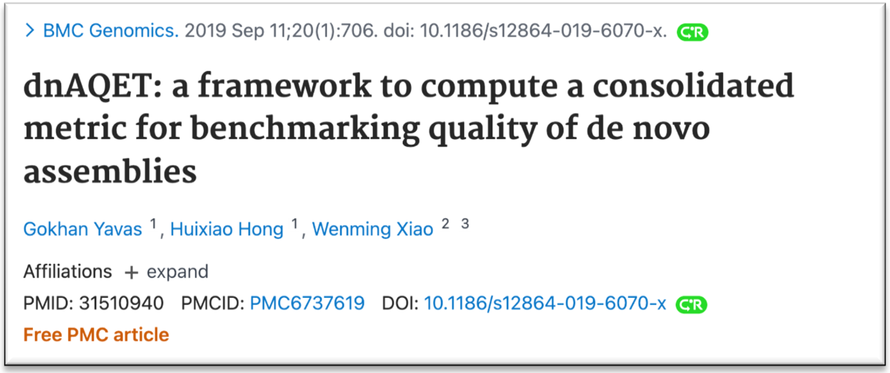
Content

Genome assembly evaluation tool.

QUAST evaluates genome assemblies by computing various metrics.

It works both with and without reference genomes.

The tool accepts multiple assemblies, thus is suitable for comparison.



> BMC Genomics. 2019 Sep 11;20(1):706. doi: 10.1186/s12864-019-6070-x.

dnAQET: a framework to compute a consolidated metric for benchmarking quality of de novo assemblies

Gokhan Yavas [1], Huixiao Hong [1], Wenming Xiao [2] [3]

Affiliations + expand

PMID: 31510940 PMCID: PMC6737619 DOI: 10.1186/s12864-019-6070-x

Free PMC article

Metrics

- The number of contigs/scaffolds in the assembly
- The size of the smallest contigs/scaffolds
- The size of the largest contigs/scaffolds
- The number of bases included in the assembly
- The mean length of the contigs/scaffolds
- The number of contigs <200 bases
- The number of contigs >1,000 bases
- The number of contigs >10,000 bases
- ~~The number of contigs that had an open reading frame~~
- ~~The mean % of the contig covered by the ORF~~
- NX (e.G. N50): the largest contig size at which at least X% of bases are contained in contigs at least this length

○ % Of bases that are G or C

○ GC skew

○ AT skew

○ The number of bases that are N

○ The proportion of bases that are N

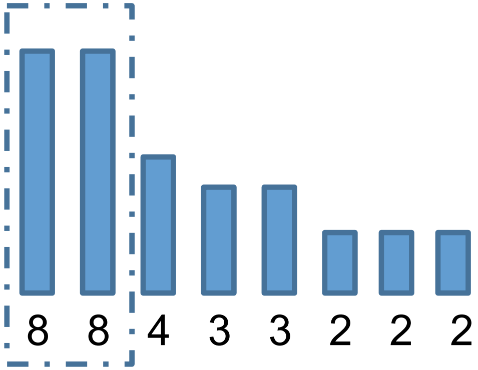
○ The total linguistic complexity of the assembly

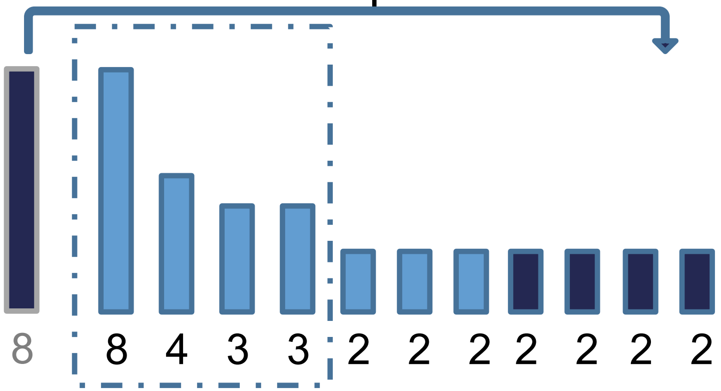
N50

- **N50**: given a set of contigs of varying lengths, the N50 length is defined as the length N for which 50% of all bases in the contigs are in contigs of length L < N

contig size list L = (8,8,4, 3, 3, 2, 2, 2) = 32

we have 50% of total length (16/32) above 4 -> **N50** is equal to 8





N50 = 8

Average : 32/8 = 4

Mediane = 3

N50 = 3

Average : 32/11 = 2.9

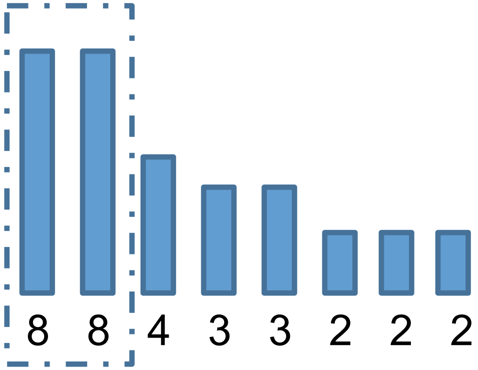
Mediane = 2

N50

N50 may not reflect some improvements to the assembly.

If we connect two contigs longer than N50 or connect two contigs shorter than N50, N50 is not changed; N50 is only improved if we connect a contig shorter than N50 and a contig longer than N50.

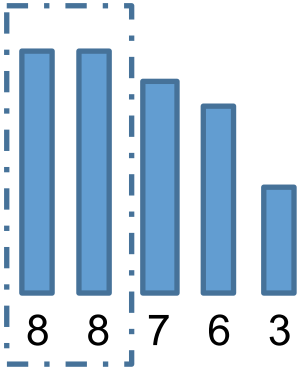
If we assembler testers solely target N50, we may be misled by it.



N50 = 8

Average : 32/8 = 4

Mediane = 3



N50 = 8

Average : 32/4 = 6.4

Mediane = 7

QUAST

QUAST tool for genome assemblies,
MetaQUAST, the extension for metagenomic datasets,
QUAST-LG, the extension for large genomes (e.g., mammalians),
rnaQUAST, the extension for RNAseq,
and **Icarus**, the interactive visualizer for these tools.

QUAST default pipeline utilizes Minimap2. Reads mapping on genome.
Functional elements prediction modules use GeneMarkS, GeneMark-ES, GlimmerHMM, Barrnap, and BUSCO.
QUAST module for finding structural variations applies BWA, Sambamba, and GRIDSS.

QUAST we use bedtools for calculating raw and physical read coverage, which is shown in Icarus contig alignment viewer.

Icarus also can use Circos
QUAST-LG introduced modules requiring KMC and Red.

MetaQUAST uses MetaGeneMark, Krona tools, BLAST, and SILVA 16S rRNA database.

Genome quality control

Contiguity

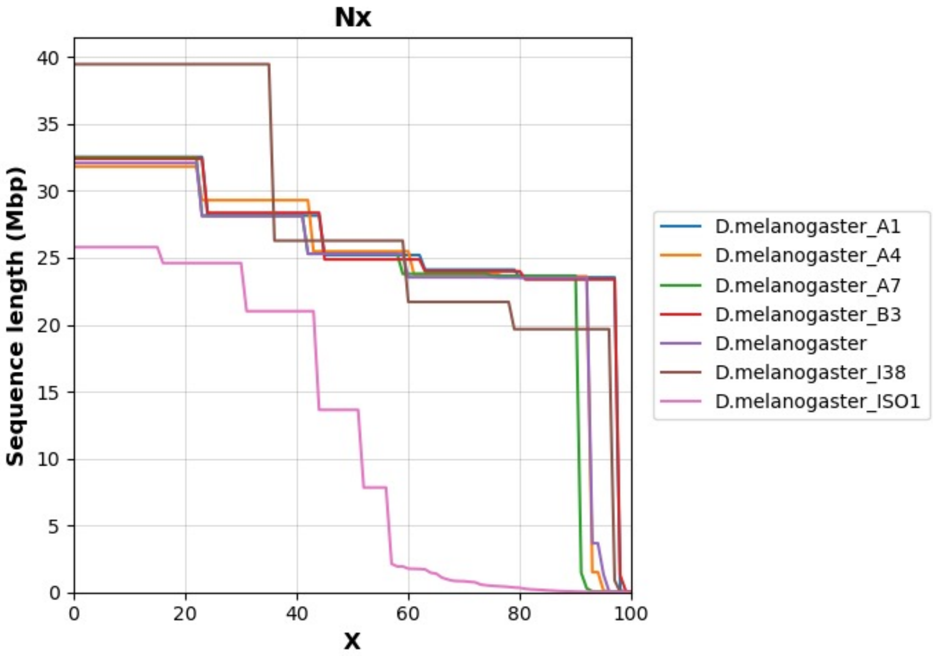
# contigs (≥ x bp) is total number of contigs of length ≥ x bp.
Total length (≥ x bp) is the total number of bases in contigs of length ≥ x bp.

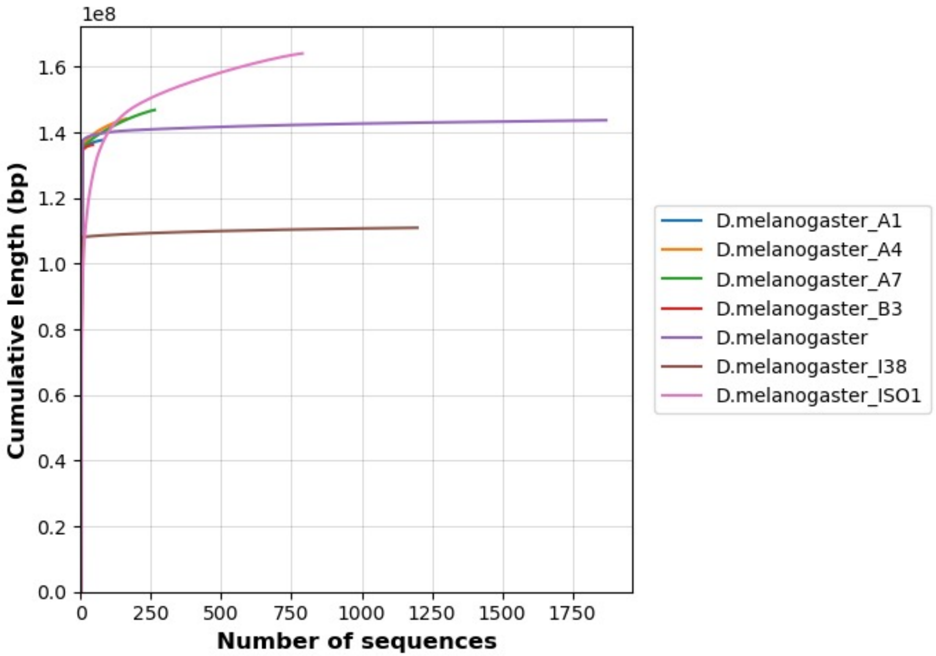
# contigs is the total number of contigs in the assembly.
Largest contig is the length of the longest contig in the assembly.
Total length is the total number of bases in the assembly.
Reference length is the total number of bases in the reference genome.
GC (%) is the total number of G and C nucleotides in the assembly, divided by the total length of the assembly.
Reference GC (%) is the percentage of G and C nucleotides in the reference genome.
N50 is the length for which the collection of all contigs of that length or longer covers at least half an assembly.

NG50 is the length for which the collection of all contigs of that length or longer covers at least half the reference genome.
This metric is computed only if the reference genome is provided.
N75 and NG75 are defined similarly to N50 but with 75 % instead of 50 %.
L50 (L75, LG50, LG75) is the number of contigs equal to or longer than N50 (N75, NG50, NG75)
In other words, L50, for example, is the minimal number of contigs that cover half the assembly

NX curve

« 50 » is a single point on the N*x* curve. The entire N*x* curve in fact gives us a better sense of contiguity.





Genome quality control

Completeness

Proportion of the original genome represented by the assembly

$$\frac{Assembled\ genome\ size}{Estimated\ genome\ size\ *}$$

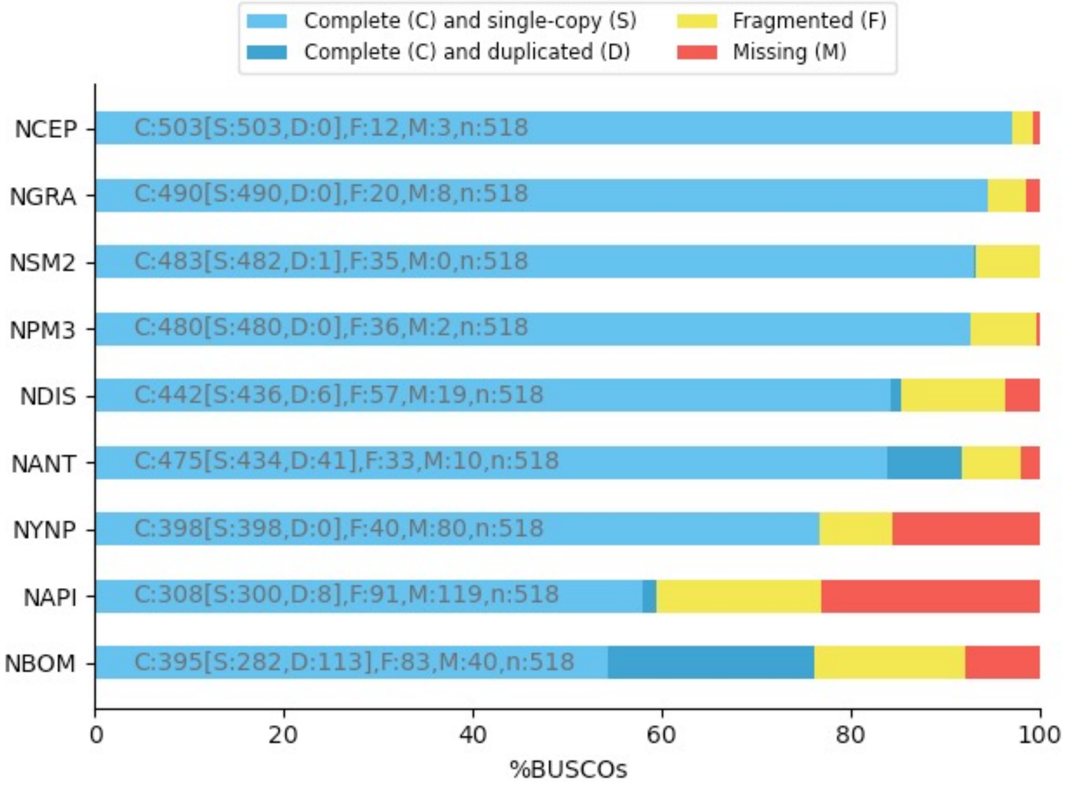
* it’s an estimation, so not perfect

Genome quality control

Completeness

Core genes (BUSCO) : quantitative assessment of genome assembly based on evolutionarily informed expectations of gene content from near-universal single-copy orthologs.

$$\frac{Core\ genes\ in\ assembly}{Core\ genes\ in\ reference\ database}$$



Tips: Reference databases are constructed using known genomes. Species with few/no close genomes available can have very bad scores.

BUSCO analysis

CEGMA : Core **E**ukaryotic **G**enes **M**apping **A**pproach : (http://korflab.ucdavis.edu/datasets/cegma/)

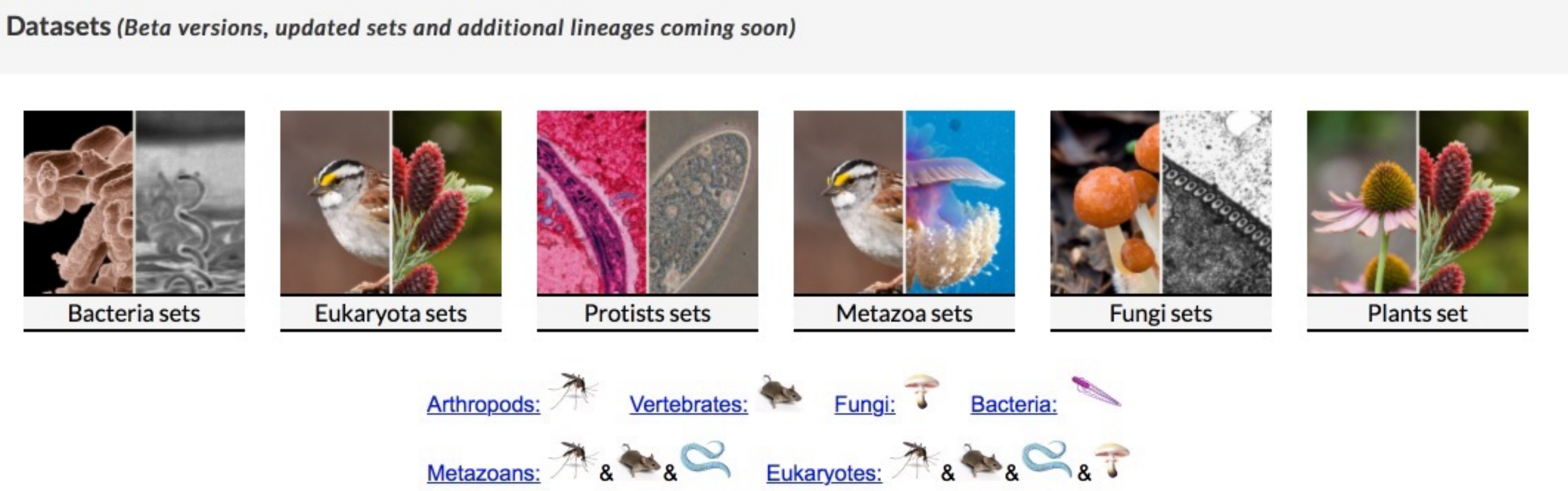
HMM:s for 248 core eukaryotic genes aligned to your assembly to assess completeness of gene space

“complete”: 70% aligned
“partial”: 30% aligned

A set of eukaryotic core proteins (KOG = euKaryotic Orthologous Groups) from 6 species: H. sapiens, D. melanogaster, C. elegans, A. thaliana, S. cerevisiae, S.pombe

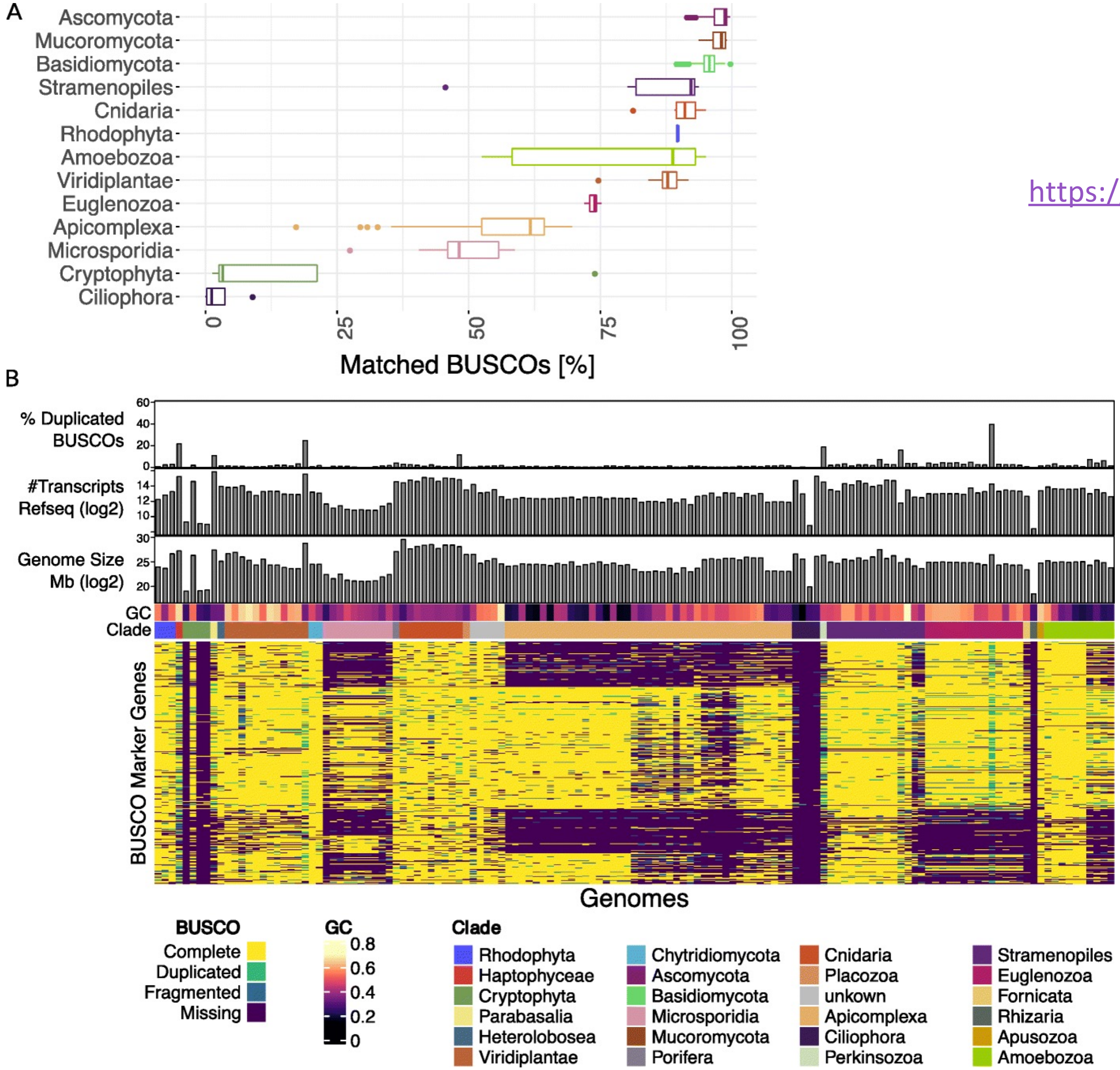
BUSCO (http://busco.ezlab.org/)

Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs



Datasets *(Beta versions, updated sets and additional lineages coming soon)*

BUSCO limitation



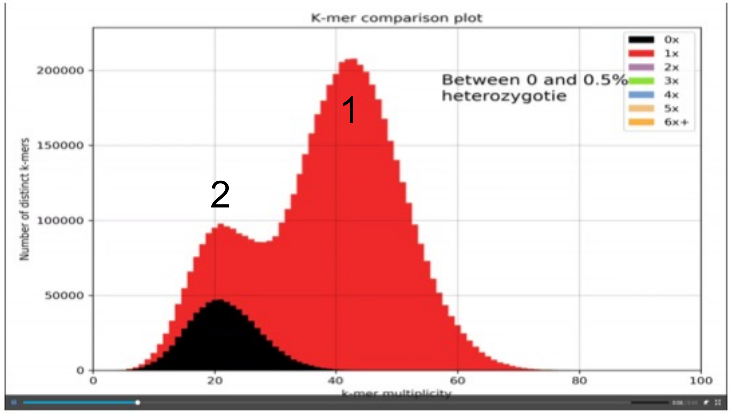
https://github.com/Finn-Lab/EukCC/

Saary, P., Mitchell, A.L. & Finn, R.D. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol* **21,** 244 (2020). https://doi.org/10.1186/s13059-020-02155-4

Genome quality control

Completeness

Kmer representation (Merquryl, YAK)



kat spectra-cn plot
> Histogram is build with read kmer content.

Colors come from assembly.
> Black = not in the assembly (heterozygous part, second haplotype).

➢ Red = once in the assembly.

1 K-mers on both chromosomes (homozygotes curves)
2 different k-mers on each chromosome (heterozygote curves)

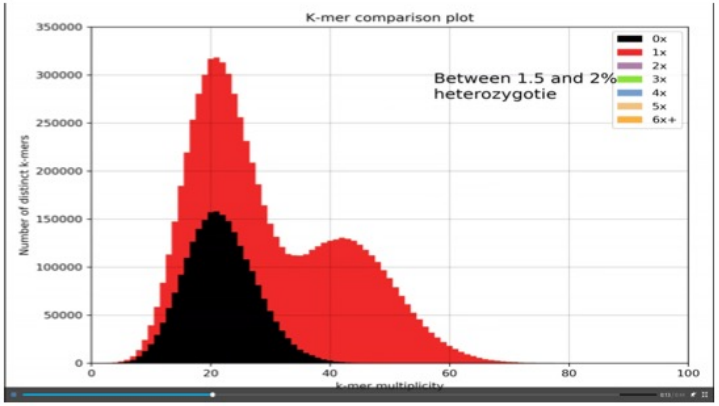


Figure: kat spectra-cn 1.5

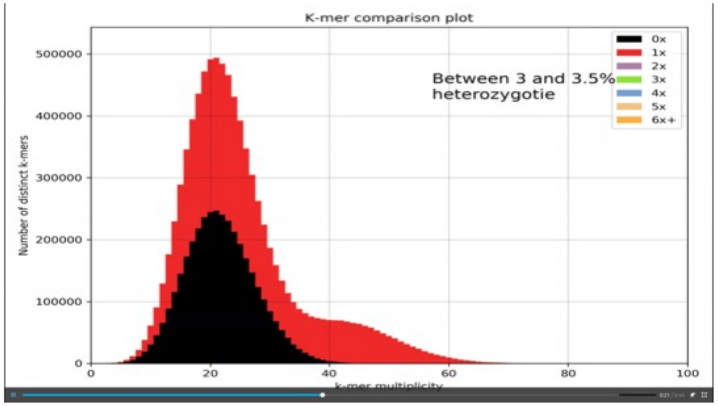


Figure: kat spectra-cn 3.5

Genome quality control

Completeness

Kmer representation (Merquryl, YAK)

kat spectra-cn plot on homozygous genomes

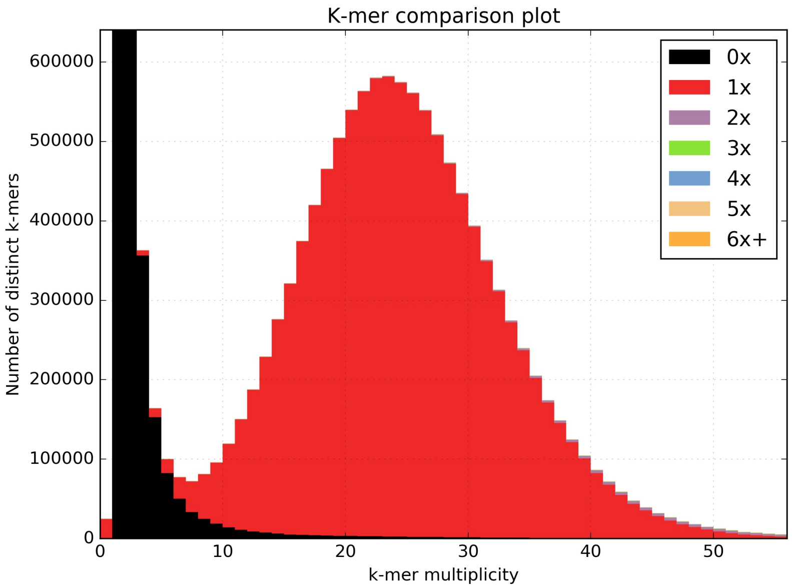
> Histogram is build with read kmer content.

Colors come from assembly.

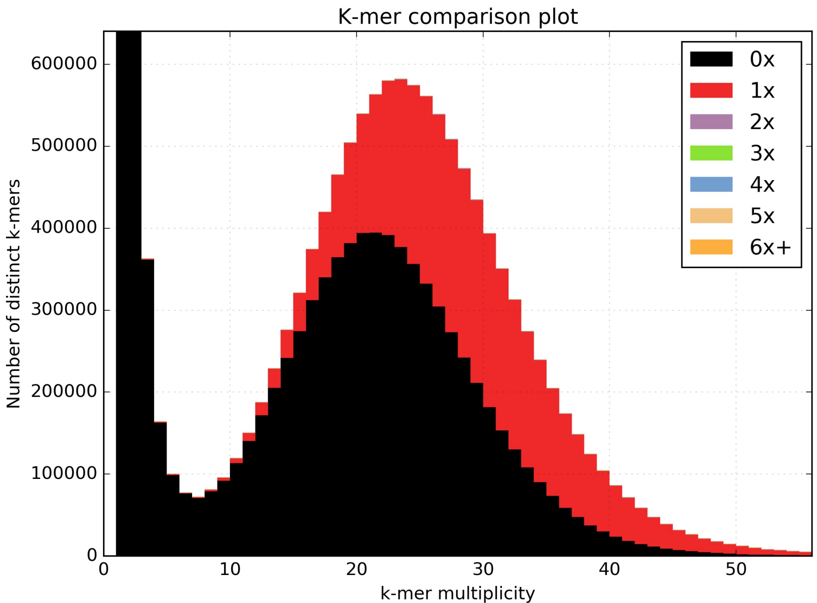
> Black = not in the assembly (errors).

- Red = once in the assembly.

Good assembly



Wrong assembly : too small k-value during assembly



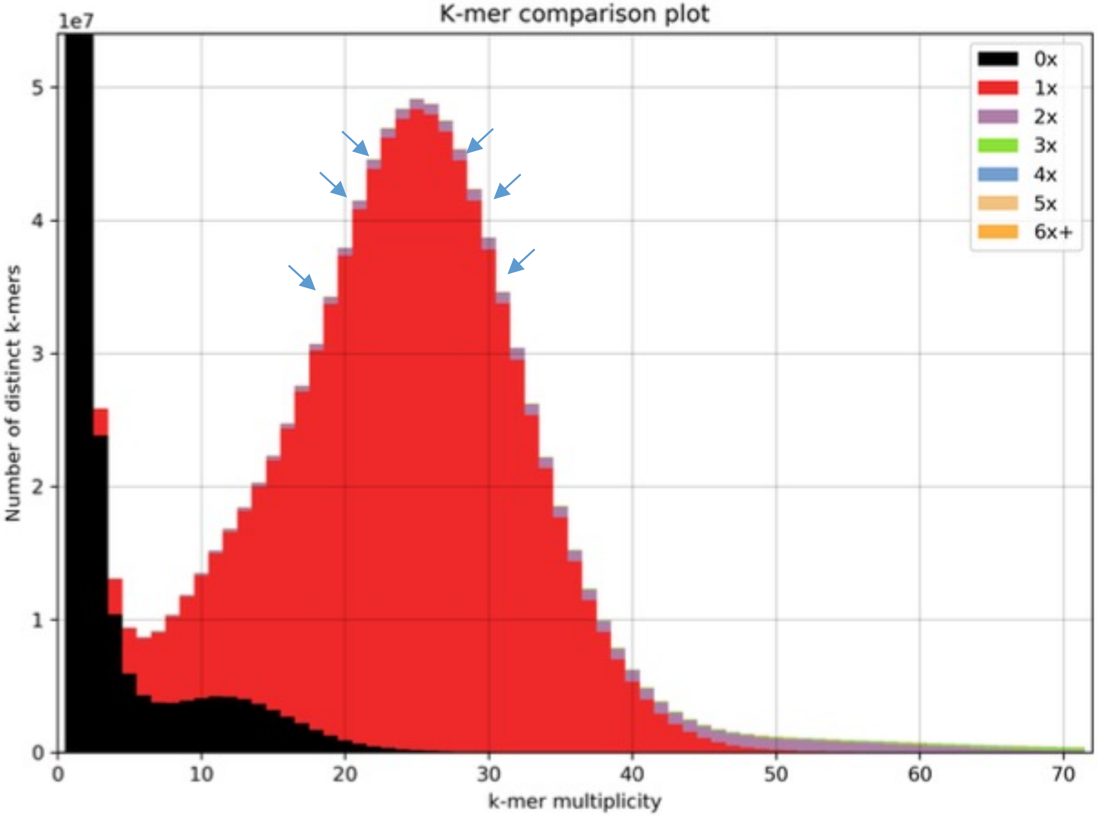
Genome quality control

Completeness

Colors come from assembly.

> Black = not in the assembly (heterozygous part, second haplotype).

- Red = once in the assembly.
- Purple = twice in the assembly



Sometimes assembler have problems to attribute contigs to the correct haplotype.

> These contigs stay in the main assembly
> This impacts the spectra-cn color profile, remaining purple on top of the red.

Genome quality control

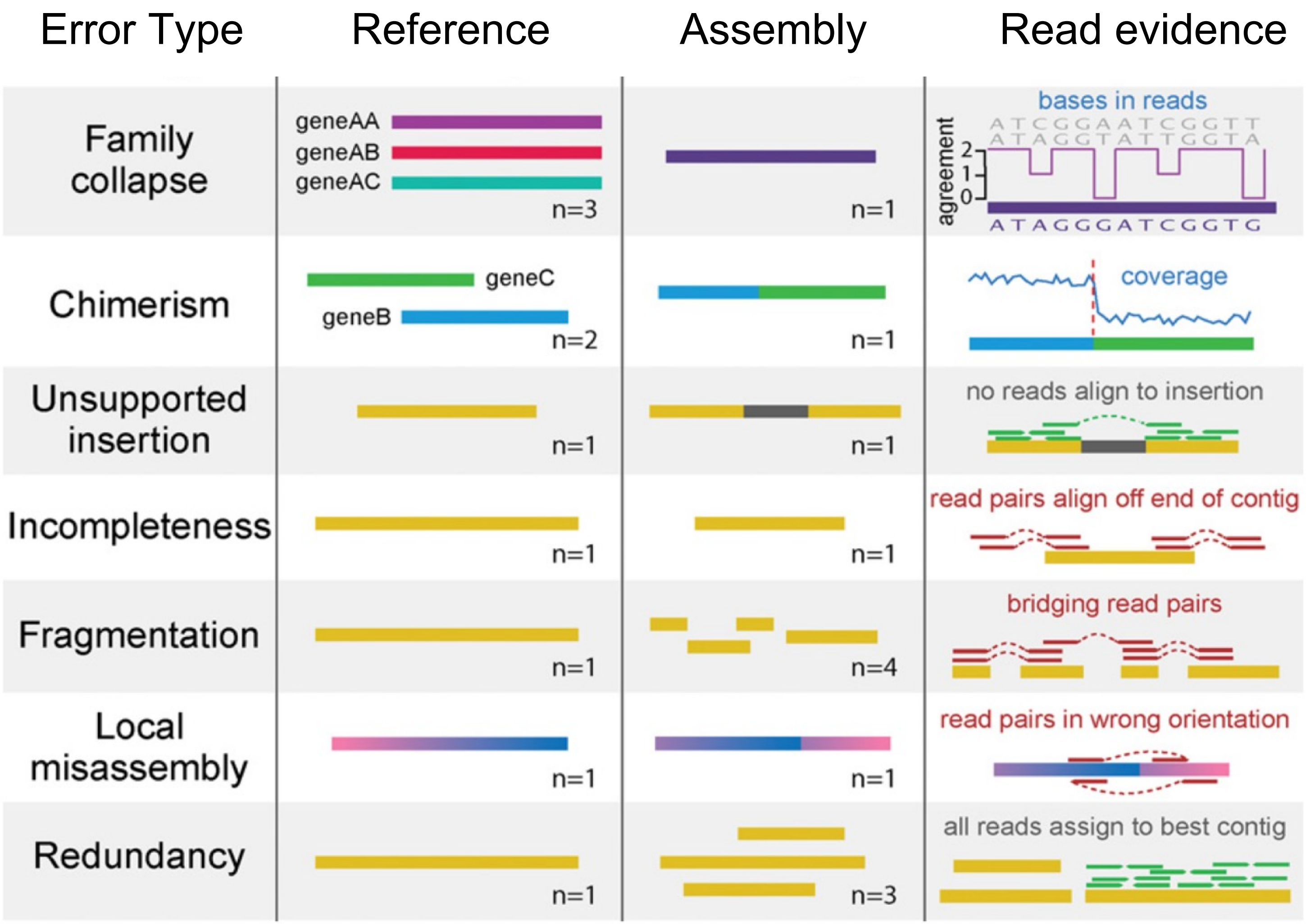
Correctness

Proportion of the assembly that is free from mistakes

- Mis-joins
- Repeat compressions
- Unnecessary duplications
- Indels / SNPs caused by assembler

Align back reads to the assembly and check for inconsistencies

Genome quality control



Error Type	Reference	Assembly	Read evidence
Family collapse	geneAA geneAB geneAC n=3	n=1	bases in reads ATCGGAATCGGTT ATAGGTATTGGTA agreement 2 1 0 ATAGGGATCGGTG
Chimerism	geneC geneB n=2	n=1	coverage
Unsupported insertion	n=1	n=1	no reads align to insertion
Incompleteness	n=1	n=1	read pairs align off end of contig
Fragmentation	n=1	n=4	bridging read pairs
Local misassembly	n=1	n=1	read pairs in wrong orientation
Redundancy	n=1	n=3	all reads assign to best contig