Définitions :

Statistics without reference

 # scaffolds : total number of scaffolds of length >= 500bp in the assembly.

 # scaffolds (>= 0bp) : idem with length >= 0bp.

 # scaffolds (>= 1000bp) : idem with length >= 1000bp.

Largest scaffold : length of the longest scaffold in the assembly.

Total length : total number of bases in the assembly, counting the scaffolds of length >= 500bp.

Total length (>= 0bp) : idem with length >= 0bp.

Total length (>= 1000bp) : idem with length >= 1000bp.

N50 : scaffold length such that using longer or equal length scaffolds produces half (50%) of the bases of the assembly. Usually there is no value that produces exactly 50%, so the technical definition is the maximum length x such that using scaffolds of length at least x accounts for at least 50% of the total assembly length.

N75 : idem with 75%.

L50 : minimum number of scaffolds that produce half (50%) of the bases of the assembly. In other words, it's the number of scaffolds of length at least N50.

L75 : idem with 75%.

GC (%) : total number of G and C nucleotides in the assembly, divided by the total length of the assembly.

Misassemblies

 # misassemblies : number of positions in the assembled scaffolds where the left flanking sequence aligns over 1kbp away from the right flanking sequence on the reference (relocation) or they overlap on more than 1 kbp (relocation) or flanking sequences align on different strands (inversion) or different chromosomes (translocation).

 # relocations : number of relocation events among all misassembly events. Relocation is a misassembly where the left flanking sequence aligns over 1 kbp away from the right flanking sequence on the reference, or they overlap by more than 1 kbp and both flanking sequences align on the same chromosome.

 # translocations : number of translocation events among all misassembly events. Translocation is a misassembly where the flanking sequences align on different chromosomes.

 # inversions : number of inversion events among all misassembly events. Inversion is a misassembly where it is not a *relocation* and the flanking sequences align on opposite strands of the same chromosome.

 # misassembled scaffolds : number of scaffolds that contain misassembly events.

Misassembled scaffolds length : number of total bases contained in all scaffolds that have one or more misassemblies.

 # local misassemblies : number of local misassemblies. We define a local misassembly breakpoint as a breakpoint that satisfies these conditions:

1. Two or more distinct alignments cover the breakpoint.

2. The gap between left and right flanking sequences is less than 1 kbp.

3. The left and right flanking sequences both are the same strand of the same chromosome of the reference genome.

Unaligned

 # fully unaligned scaffolds : number of scaffolds that have no alignment to the reference sequence.

Fully unaligned length : total number of bases contained in all fully unaligned scaffolds.

 # partially unaligned scaffolds : number of scaffolds that are not fully unaligned ones but have fragments with no alignment to the reference.

 # with misassembly : number of partially unaligned scaffolds that contain misassembly events in their aligned fragment. Note that such misassemblies are not counted in *# misassemblies* and other *misassemblies* statistics.

 # both parts are significant : number of partially unaligned scaffolds that contain both aligned and unaligned fragments of length ≥ *min-scaffold threshold*.

Partially unaligned length : total number of unaligned bases in all partially unaligned scaffolds.

Mismatches

 # mismatches : number of mismatches in all aligned bases.

 # indels : number of indels in all aligned bases.

Indels length : total number of bases contained in all indels.

 # mismatches per 100 kbp : average number of mismatches per 100000 aligned bases.
 # indels per 100 kbp : average number of indels per 100000 aligned bases.
 # short indels : number of indels of length less or equal to 5bp.
 # long indels : number of indels of length greater than 5bp.
 # N's : total number of uncalled bases (N's) in the assembly.
 # N's per 100 kbp : average number of uncalled bases (N's) per 100000 assembly bases.

Genome statistics

Genome : total number of bases in the reference genome.
G+C content (%) : total number of bases being a G or a C in the reference genome.
Genome fraction (%) : total number of aligned bases in the reference, divided by the genome size. A base in the reference genome is counted as aligned if there is at least one scaffold with at least one alignment to this base. scaffolds from repeat regions may map to multiple places, and thus may be counted multiple times in this quantity.
Duplication ration : total number of aligned bases in the assembly (i.e. *Total length* - *Fully unaligned length* - *Partially unaligned length*), divided by the total number of aligned bases in the reference (see the **Genome fraction (%)** metric). If the assembly contains many scaffolds that cover the same regions of the reference, its **Duplication ratio** may be much larger than 1. This may occur due to overestimating repeat multiplicities and due to small overlaps between scaffolds, among other reasons.
Largest alignment : length of the largest continuous alignment in the assembly. This metric is always equal to the *Largest scaffold* metric but it can be smaller if the largest scaffold of the assembly contains a misassembly event.
NG50 : scaffold length such that using longer or equa length scaffolds produces half (50%) of the bases of the reference genome. This metric is computed only if a reference genome is provided.
NG75 : idem with 75%.
NA50 : N50 where the lengths of aligned blocks are counted instead of scaffold lengths. I.e., if a scaffold has a misassembly with respect to the reference, the scaffold is broken into smaller pieces. This metric is computed only if a reference genome is provided.
NA75 : idem with 75%.
NGA50 : NG50 where the lengths of aligned blocks are counted instead of scaffold lengths. i.e., if a scaffold has a misassembly with respect to the reference, the scaffold is broken into smaller pieces. This metric is computed only if a reference genome is provided.
NGA75 : idem with 75%.
LG50 : minimum number of scaffolds that produce half (50%) of the bases of the reference. genome. In other words, it's the number of scaffolds of length at least NG50. This metric is computed only if a reference genome is provided.
LG75 : idem with 75%.
LA50 : L50 where aligned blocks are counted instead of scaffolds. i.e., if a scaffold has a misassembly with respect to the reference, the scaffold is broken into smaller pieces.
LA75 : idem with 75%.
LGA50 : LG50 where aligned blocks are counted instead of scaffolds. i.e., if a scaffold has a misassembly with respect to the reference, the scaffold is broken into smaller pieces.
LGA75 : idem with 75%.

Predicted genes

 # predicted genes (unique) :
 # predicted genes (>= 0bp) :
 # predicted genes (>= 300bp) :
 # predicted genes (>= 1500bp) :
 # predicted genes (>= 3000bp) :