CAMI metagenome assembly evaluation, guidelines for strain-specific assemblies

Definitions:

We consider two genomes to be different strains of the same species if they have >95% average nucleotide identity.

- A *consensus* (or *strain-unresolved*) assembly is one where each contig may correspond to 1 or many strains (*strain-unresolved* contig), and reciprocally (and importantly), any genomic region that has one or more homologs across different strains is represented in only one contig[1]. Assemblers that create such assemblies are *strain-oblivious* assemblers. (Most assemblers, as of 2020, fall under this category.)

- A *strain-resolved, or strain-specific* assembly is one where each contig either i) maps equally likely to >1 strains (*core* contigs), or ii) maps unambiguously to only 1 strain (*strain-specific* contigs). Assemblers that create such assemblies are strain-aware or strain-resolved assemblers. In addition, core contigs should be present in as many copies as there are genomes containing such regions (see example below)[2].

Evaluation of strain-specific assemblies:

Consider the following reference genomes

Where R1 and R2 are two strains of the same species (%-identity higher or identical to our threshold set above). R3 is a different species, without homologous regions with R1/R2.

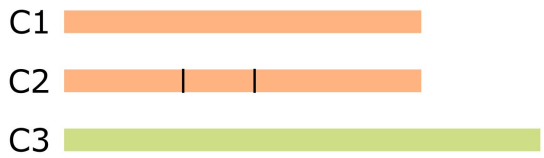
Case 1: two (extreme) examples of assemblies

Assembly **A1**

[1] Note that this definition purposely leaves the meaning of the term ‘homologs’ vague, as it depends on criteria set by what the assembler does, on the sequencing technology, etc.
[2] We realize this last condition might be difficult to satisfy (as it amounts to determining the number of strains) so we will adapt our assembly evaluation metrics depending on how well methods manage to satisfy it.

A1 is a consensus assembly. Here contig C1 corresponds to a consensus of R1 and R2, and contig C2 corresponds exactly to R3.

Assembly **A2**

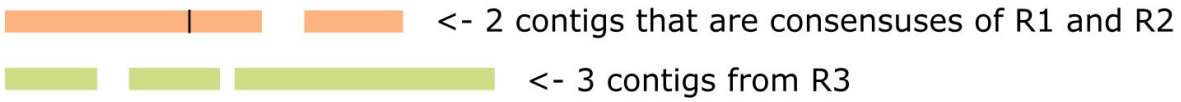


The contigs of A2 correspond exactly to the reference genomes. A2 is a strain-specific assembly. Contigs C1 and C2 are >= 95% identical.

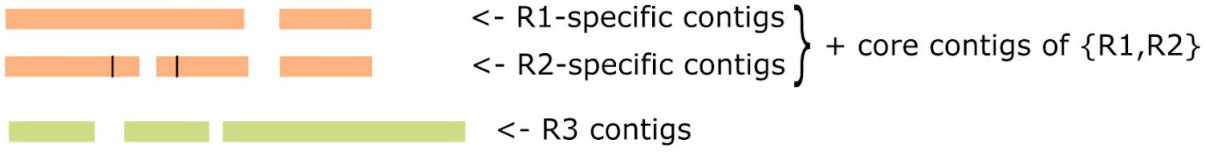
Between **A1** and **A2**, in the context of strain-aware evaluation, we will design metrics that favor assembly **A2** to **A1**.

Case 2: two more realistic assemblies

A3



A4



CAMI will perform strain-aware evaluation. Thus in that context we will prefer assemblies like **A4** to assemblies like **A3**. (Despite the apparent fragmentation in R2 contigs.) We will also prefer **A2** to **A4**.

Evaluation will involve common criteria such as contiguity, correctness, completeness of individual strain genomes.