Provisional Assembly Metrics

Upon submission of annotation to Genbank, the DACC runs the following set of quality control metrics on every HMP reference genome, to ensure ensure accuracy, completeness and continuity of draft and improved assemblies.

I. Preliminary metrics

Centers will submit to NCBI all contigs >= 500 bp and containing at least two reads (for all platforms).

Genome Size = Sum(submitted scaffold spans) for scaffolded assemblies with read pairs; Genome Size = Sum(submitted ctgs) for fragment assemblies without read pairs.

II. Provisional Assembly Metrics

[1] > 90% of the genome included in ctgs. [Scaffold completeness]

Comments:
1) The size of intra-scaffold gaps will be estimated either by assembly software or by a method selected by the center. The assembly software or method should be described in the assembly README.
2) This statistic is not meaningful for assemblies that are not scaffolded (e.g. of fragment reads), and is noted as N/A for such assemblies.
3) When joining scaffolds to produce a linear sequence, 100 Ns should be used to indicate inter-scaffold gaps, in accord with the system used at GenBank. These will not be counted in genome size estimates.

[2] >90% of the bases greater then 5x read coverage [Accuracy]

Method: Sum of all bases [A,T,C,G] in the consensus that have greater than 5 reads supporting the consensus call of that base. Denominator is sum of all submitted contig bases.

Comments:
1) Read coverage should be calculated locally by each center.

[3] > 5 KB contig N50, N75, or N90 length [Continuity]

Method: N90 is the length of the shortest contig such that the sum of contigs of equal length or longer is at least 90% of the total genome size.

Comments:
1) We are still in the process of evaluating N50, N75, N90 to determine where to set the bar.

[4] > 20 KB scaffold N50, N75, or N90 length [Continuity]

Method: N90 is the length of the shortest scaffold such that the sum of scaffolds of equal length or longer is at least 90% of the total genome size.

Comments:
1) We are still in the process of evaluating N50, N75, N90 to determine where to set the bar.

[5] Average contig length > 5Kb [Choppiness]

Comments:
1) This is to flag assemblies that satisfy Continuity metrics by having one or a few very large contigs and many very small ones. They will have too many gaps and will need more work.
2) We are still in the process of evaluating average contig length to determine if we can increase the bar.
3) If N75/N90 can be used in [3] it may make this metric unnecessary.

[6] > 90% of core genes present in gene list [Completeness and annotatability]

See core gene selection SOPs on the DACC website for more details about this metric.

Comments:
1) Bacterial core genes: The 66 core genes are present in all (99.6%) 621 finished bacterial genomes available in GenBank @ cut- off 30%id and 30% length.
2) Archaeal core genes: The 104 core genes are present in all 52 finished archaeal genomes available in GenBank @ cut-off 50%id and 70% length.