Bioinformatics Advance Access published June 9, 2015

Genome analysis

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs

Felipe A. Simão[†], Robert M. Waterhouse[†], Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov[*]

Department of Genetic Medicine and Development, University of Geneva Medical School and Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland.

Associate Editor: Dr. John Hancock

ABSTRACT

Motivation: Genomics has revolutionised biological research, but quality assessment of the resulting assembled sequences is complicated and remains mostly limited to technical measures like N50.

Results: We propose a measure for quantitative assessment of genome assembly and annotation completeness based on evolutionarily informed expectations of gene content. We implemented the assessment procedure in open-source software, with sets of Benchmarking Universal Single-Copy Orthologs, named BUSCO.

Availability and Implementation: Software implemented in Python and datasets available for download from http://busco.ezlab.org.

Contact: Evgeny.Zdobnov@unige.ch

Genomics data acquisition continues to accelerate, however, the short lengths of sequencing reads make their assembly into full-length chromosomes extremely challenging. To guage potential limitations and implement improvements it is thus important to assess the quality of the resulting data. Proposed measures (Simpson, 2014; Clark et al., 2013; Gurevich et al., 2013; Hunt et al., 2013) reflect methodologies, e.g. per-base error rates, insert size distributions; or genome biases, e.g. k-mer distributions; or fragment (contig) length distributions, e.g. N50, which summarises assembly contiguity in a single number: half the genome is assembled on contigs of length N50 or longer. However, such measures do not assess assembly completeness in terms of gene content: an important consideration that also affects data interpretation and helps to guide improved assembly and annotation strategies.

With the growing number of available sequenced genomes, knowledge of their gene content is consolidating and can be used to develop an evolutionary measure of genome completeness. Here we revisit the idea of using known genes to measure genome assembly and annotation completeness (Parra, et al., 2009, Mende, et al., 2013), by introducing a citable notation for well-defined measures, compiling the comprehensive datasets to support such assessments, and offering these as an off-the-shelf software.

As proposed previously (Waterhouse et al., 2013), Benchmarking Universal Single-Copy Orthologs (BUSCO) are ideal for such quantifications of completeness, as the expectations for these genes to be found in a genome and to be found only in single-copy are evolutionarily sound. We used our OrthoDB database of orthologs (www.orthodb.org) to define BUSCO sets for six major phylogenetic clades. Sampling hundreds of genomes, orthologous groups with single-copy orthologs in more than 90% of species were selected. Importantly, this threshold accommodates the fact that even well-conserved genes can be lost in some lineages, as well as allowing for incomplete gene annotations and rare gene duplications. Subsequent filtering, e.g. on sequence uniqueness and conservation levels [see Supplementary Online Material (SOM) for details], resulted in BUSCO sets representing 3,023 genes for vertebrates, 2,675 for arthropods, 843 for metazoans, 1,438 for fungi, and 429 for eukaryotes. We also adopted 40 universal marker genes proposed for the assessment of prokaryotic genomes (Mende et al., 2013). The clades spanning many phyla offer comprehensive coverage of the tree of life, while the more narrowly-defined clades provide a much greater resolution with much larger BUSCO sets. These are applicable not only to the assessment of genome assemblies, but also to annotated gene sets, as well as assembled transcriptomes (Figure 1). Additionally, as near-universal single-copy markers, the recovered genes are ideal for species phylogeny reconstructions.

We propose intuitive metrics to describe genome, gene set, or transcriptome completeness in BUSCO notation - C:complete [D:duplicated], F:fragmented, M:missing, n:number of genes used (Figure 1). The recovered genes are classified as 'complete' when their lengths are within two standard deviations of the BUSCO group mean length (i.e. within ~95% expectation, Figure S1). 'Complete' genes found with more than one copy are classified as 'duplicated'. These should be rare, as BUSCOs are evolving under single-copy control (Waterhouse et al., 2011), and the recovery of many duplicates may therefore indicate erroneous assembly of haplotypes. Genes only partially recovered are classified as 'fragmented', and genes not recovered are classified as 'missing'. Finally, the 'number of genes used' indicates the resolution and hence is informative of the confidence of these assessments.

Using HMMER 3 (Eddy, 2011) hidden Markov model (HMM) profiles from amino acid alignments, the core of the analysis workflow (Figure 1) assesses whether BUSCO gene matches are orthologous or not (i.e. satisfy BUSCO group-specific bitscore cut-offs; detailed in SOM), and classifies positive matches as complete or fragmented. This core analysis is the same for assessing genomes, transcriptomes, or gene sets. However, additional analyses are

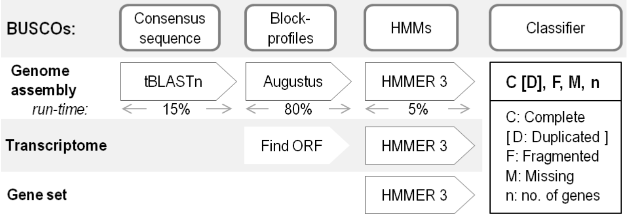

Fig. 1. BUSCO assessment workflow and relative run-times.

[†] Equal contribution. [*] To whom correspondence should be addressed.

required to first annotate genes from transcriptomes and genomes. The simple longest open reading frame approach performs well for transcriptomes. For genomes, gene annotation is performed with Augustus (Keller et al., 2011), guided by amino acid BUSCO group block-profiles, on genomic loci detected with tBLASTn searches using BUSCO group consensus sequences (detailed in SOM). While this gene prediction approach may have its limitations and biases, they are consistent across different species, making for fair comparisons. Conveniently, the thousands of confident BUSCO gene models provide an excellent gene predictor training set for use as part of genome annotation pipelines.

Table 1 reports BUSCO notation assessments of five diverse species for both their genome assemblies and their annotated gene sets. Assessing 70 genomes, 163 gene sets, and 96 transcriptomes revealed substantial variability of completeness (Table S1). Poor correlation with scaffold N50 (Figure S2) highlights how completeness provides important complementary information for quality assessment. Nevertheless, the fact that some genome assemblies appear less complete than their corresponding gene sets (e.g. *H. sapiens* Table 1) reveals limitations of the BUSCO gene prediction step. On the other hand, a reversal of this trend (e.g. *A. nidulans* Table 1) suggests that the annotated gene set may be missing some BUSCO gene matches that are in fact present in the genome. Thus, it should be noted that while BUSCO assessments aim to robustly estimate completeness of the data sets, technical limitations (particularly gene prediction) may inflate proportions of 'fragmented' and 'missing' BUSCOs, especially for large genomes.

Table 1. Assessment of fruitfly (*D.mela*,), nematode worm (*C.eleg*,), human (*H.sapi*,), owl limpet (*L.giga*,), and fungus (*A.nidu*,) genome assemblies (upper row) and gene sets (lower row) in BUSCO notation (C:complete [D:duplicated], F:fragmented, M:missing, n: gene number).

Species	Size	BUSCO notation assessment results
D.mela	139 Mbp	C:98% [D:6.4%], F:0.6%, M:0.3%, n:2,675
	13,918	C:99% [D:3.7%], F:0.2%, M:0.0%, n:2,675
C.eleg	100 Mbp	C:85% [D:6.9%], F:2.8%, M:11%, n:843
	20,447	C:90% [D:11%], F:1.7%, M:7.5%, n:843
H.sapi	3,381 Mbp	C:89% [D:1.5%], F:6.0%, M:4.5%, n:3,023
	20,364	C:99% [D:1.7%], F:0.0%, M:0.0%, n:3,023
L.giga	359 Mbp	C:89% [D:2.3%], F:4.3%, M:5.8%, n:843
	23,349	C:90% [D:13%], F:7.8%, M:2.1%, n:843
A.nidu	30 Mbp	C:98% [D:1.8%], F:0.9%, M:0.2%, n:1,438
	10,534	C:95% [D:7.3%], F:3.8%, M:0.9%, n:1,438

Comparing genome to gene set completeness of 40 species using a 250-BUSCO eukaryotic subset reveals generally consistent assessments across highly divergent lineages from fungi to human (Figure 2). Employing the 248 genes of the Core Eukaryotic Gene Mapping Approach (CEGMA) (Parra, et al., 2007) in a like-for-like comparison (i.e. implementing gene set assessments using CEGMA HMMs, see SOM for details) appears somewhat less consistent (Figure 2, BUSCO linear regression is closer to the diagonal). Additionally, in comparable 250-BUSCO and 248-CEGMA assessments BUSCO run-times are substantially faster, ~2x for small genomes and ~10x for large genomes, but of course the higher resolutions achievable with the thousands of vertebrate, arthropod, and fungal BUSCO sets do require longer run-times (Table S2). Run-times are generally proportional to the size of the BUSCO set used and the sizes of the genomes being assessed, e.g., on 4 CPU cores with up to 8 GB memory: the 180 Mbp fruit fly genome ran for 3.2h and 7.6h, while the 3,381 Mbp human genome ran for 13h and 29h with 843 metazoan and 2,675 arthropod or 3,023 vertebrate BUSCOs, respectively (Table S2).

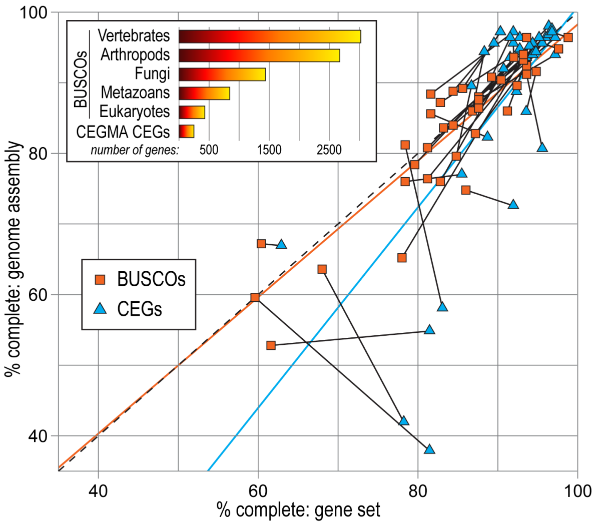

Fig. 2. BUSCOs (eukaryotic subset) and CEGMA CEGs recovered from 40 representative genome assemblies and their respective gene sets. Inset: number of genes in each BUSCO set and the CEGMA CEGs.

BUSCO quality assessments provide high-resolution quantifications citeable in the simple C[D],F,M,n notation for genomes, gene sets, and transcriptomes. This facilitates informative comparisons, e.g. of newly-sequenced draft genome assemblies to those of gold-standard models, or to quantify iterative improvements to assemblies or annotations. BUSCO assessments therefore offer intuitive metrics, based on evolutionarily informed expectations of gene content from hundreds of species, to gauge completeness of rapidly accumulating genomic data and satisfy an Iberian's quest for quality - *"Busco calidad/qualidade"*.

ACKNOWLEDGEMENTS

The authors thank colleagues and *i*5K members for user feedback. Some computations were performed at Vital-IT. *Funding*: This work benefitted from Swiss National Science Foundation grants 31003A_143936 and PDFMP3_137064 to E.M.Z. and PIOF-GA-2011-303312 to R.M.W. *Conflicts of Interest*: none declared.

REFERENCES

Clark, S.C., *et al.* (2013) ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies, *Bioinformatics*, **29**, 435-443.

Eddy SR. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol. 2011 Oct;7(10):e1002195

Gurevich, A., *et al.* (2013) QUAST: quality assessment tool for genome assemblies, *Bioinformatics*, **29**, 1072-1075.

Hunt, M., *et al.* (2013) REAPR: a universal tool for genome assembly evaluation, *Genome Biol*, **14**, R47.

Keller O., *et al.* (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. Mar 15:27(6):757-63

Mende DR, *et al.* (2013). Accurate and universal delineation of prokaryotic species. Nat Methods. 2013 Sep;10(9):881-4.

Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, Bioinformatics, **23**, 1061-1067.

Parra, G., *et al.* (2009) Assessing the gene space in draft genomes, *Nucleic Acids Res*, **37**, 289-297.

Simpson, J.T. (2014) Exploring genome characteristics and sequence quality without a reference, *Bioinformatics*.

Waterhouse, R.M., *et al.* (2011). Correlating traits of gene retention, sequence divergence, duplicability and essentiality. Genome Biol Evol. 2011;3:75-86.

Waterhouse, R.M., *et al.* (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs, *Nucleic Acids Research*, **41**, D358-D365.