Gene set completeness analysis

Most genomes that will be sequenced nowadays will remain in a draft status. In other words, individual chromosomes are not reconstructed over their full length, but instead the assembly will result in individual contigs and scaffolds, each of which - if we neglect assembly errors for the moment - represents a part of a chromosome. Moreover, in most instances specialized software, e.g. the maker pipeline, is used to annotate the genome, and to identify genes. Before starting any comparative genomics study, it is thus quite common to find at least an approximate answer to the question *How good is my reconstruction?* This is not at least, because quite a number of comparative studies not only focus on the shared presence of genes, but instead might be also interested in identifying genes, which have been either modified in their structure, e.g. by gene fusions and fission, or which have been lost entirely.

Assessing the quality of a reconstruction if the original status is not known is a considerably hard task. It is, thus, that most quality assessment approaches either stick with summary statistics, just remember the QUAST analysis when you do not upload a reference genome, or a use any kind of a user-defined quality measure. Busco is a software that aims at assessing the completeness and the quality of a genome assembly and/or the gene annotation by querying the presence particular genes. The catalog of used genes is compiled in such a way that for each gene prior evidence exists that it should be present in the genome under study, the so-called universal single-copy orthologs. Figure 1 lists the main BUSCO categories.

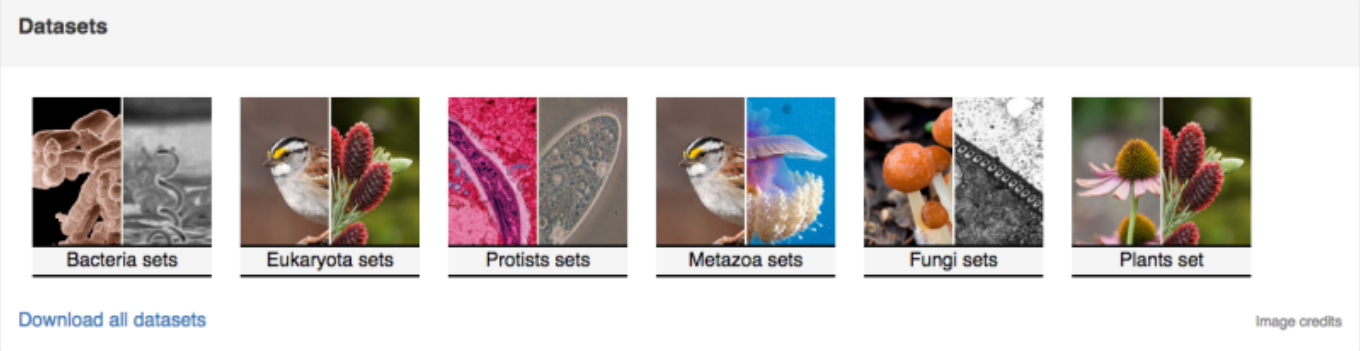

Figure 1: BUSCO set categories as provided via the BUSCO web pages. For each category several, more specialized BUSCO sets are available.

The way to compile these catalogs is straightforward: In brief, take a set of species, let's say fungi, such that the full phylogenetic diversity of this systematic group is represented. Use a standard ortholog search tool, such as OrthoDB or OMA, and identify orthologous groups. Subsequently, extract those orthologous groups which harbor for most (>90%) of the species in your taxon set exactly one sequence. For the corresponding genes you now conclude the following

- the gene is at least as old as the last common ancestor of your taxon collection
- since the gene is represented in most of the individual species, it seems to be so essential that it is typically not subjected to lineage-specific loss, hence you think its *universal*
- since most species are represented only by a single sequence, the gene is typically not subjected to a lineage specific duplication. In other words, you have identified a gene with *universal single copy orthologs*

Busco then performs a number of downstream post-processing steps, such as assessing the length variation of the proteins within each Busco group. On this basis, the tool later classifies whether a particular gene is represented completely or only partially in a test gene set.

For a more thorough introduction, please refer to the corresponding BUSCO webpages.

Outline

What you need

- The set of annotated proteins for your genome assembly[1)]:
  - /home/ubuntu/Share/Analysis/GeneAnnotation/Results/braker2/crypto_BCM2021_v2.braker2.proteins
  - /home/ubuntu/Share/Analysis/GeneAnnotation/Results/metaeuk/Crypto_Metaeuk.fas
- **optionally** the genome assembly in FASTA format: `/home/ubuntu/Share/Assemblies/crypto_BCM2021_v2.fasta`

What you get

- a brief summary file of the BUSCO results that informs about numbers/fractions genes that are
  - complete (single copy)
  - complete (duplicated)
  - fragmented
  - missing

Running BUSCO

Busco can be run on different input data

- directly on the genome assembly
- on transcript sets
- on protein sets

While the analysis on the genome sequence level has the advantage of being independent from the sensitivity of a preceding gene prediction, it is substantially more computationally intense. In this course, we will therefore restrict the analysis to the set of predicted proteins. We have preinstalled *Busco* for you in the **/home/ubuntu/miniconda3/envs/busco** environment.

1. make sure that Busco is installed by typing

```
conda deactivate
conda activate busco
busco -h
```

If Busco is not installed, please perform the installation via the conda package management system.

1. create a sub-directory ***busco*** in your project directory

```
mkdir -p $HOME/Analysis/busco
```

2. change into the new directory:

```
cd $HOME/Analysis/busco
```

3. make a directory ***data*** that will then take up all data for your subsequent analyses:

```
mkdir data
```

4. copy or better soft-link[2)] the following files into the data directory
    1. the genome sequence in FASTA format
    2. the gene set as a protein FASTA file
5. check the available datasets from Busco by typing

```
busco --list-datasets
```

and check for the sets *alveolata* and *eukaryota*

1. run the Busco analysis by typing for example

```
busco -i Crypto_Metaeuk.fas -c 1 -o Busco_metaeuk_eukaryota -m prot -l
eukaryota
```

Note, the option `-c 1` tells Busco to use only 1 cpu core for the analysis. On our cloud, the number of processors is limited, but feel free to increase this number on your own system.

Busco Results

1. Monitor the outcome of your analysis. The output directory is quite nested. The interesting files are located for the *ubuntu user* in `/home/ubuntu/Share/Analysis/busco/braker2/Busco_braker2_eukaryota/run_eukaryota_odb10/`
2. What do you conclude from the findings? Pay particular attention to the number of missing, partial and duplicated Busco genes, and **compare the results from the *eukaryota* and the *alveolata* data sets**.

```
    --------------------------------------------------
    |Results from dataset alveolata_odb10             |
    --------------------------------------------------
    |C:100.0%[S:100.0%,D:0.0%],F:0.0%,M:0.0%,n:171    |
    |171    Complete BUSCOs (C)                       |
    |171    Complete and single-copy BUSCOs (S)       |
    |0    Complete and duplicated BUSCOs (D)          |
    |0    Fragmented BUSCOs (F)                       |
    |0    Missing BUSCOs (M)                          |
    |171    Total BUSCO groups searched               |
    --------------------------------------------------

    --------------------------------------------------
    |Results from dataset eukaryota_odb10             |
```

```
    --------------------------------------------------
    |C:51.0%[S:50.6%,D:0.4%],F:8.2%,M:40.8%,n:255      |
    |130    Complete BUSCOs (C)                        |
    |129    Complete and single-copy BUSCOs (S)        |
    |1    Complete and duplicated BUSCOs (D)           |
    |21    Fragmented BUSCOs (F)                       |
    |104    Missing BUSCOs (M)                         |
    |255    Total BUSCO groups searched                |
    --------------------------------------------------
```

1. If you want to look up the orthologous group behind a BUSCO group, you have to look up at OrthoDB[3]. For looking up a gene in the assembly, check the fasta files in the directory `fragmented_busco_sequences`. You will find the deviating gene from your assembly there. Use its identifier to search in the web browser.
2. **Optional** repeat the analysis with the genome sequence as input, but  do you think that this makes sense for our purpose? Keep also in mind that BUSCO search in an unannotated genome sequence is computationally more demanding than the search in protein sequences[4]

Final remarks

Busco is common and valuable for assessing the completeness of a genome in a standardized manner. However, one should keep in mind that the results should not be over-interpreted.

- Busco sets can be considerably small, in other words you test for the presence of only a small set of the entire gene set. Thus, you should spend more than one thought about whether or not it is feasible to generalize the insights from the Busco analysis to the entire gene set.
- Busco checks for single copy orthologs only. So you deliberately avoiding the more difficult cases in genome annotation (and orthology inference)

The use of Busco is not limited to just assessing the completeness of gene set reconstructions (or genome assemblies). Instead, it can provide valuable information for the initial training of a gene prediction software, and thus shows up now and then in protocols that concentrate on genome annotation.

- Physalia main page
- Physalia gene set characterization page
- Proceed with fCAT

1) you can use any prediction

2)

ln -s
3)

see the example in the hidden section above
4)

Why?