[BUSCO][https://busco.ezlab.org]

BUSCO 采用benchmarking universal single-copy orthologs数据集来评估完整性(www.orthodb.org), 提供基因组组装, 注释基因集, 期待基因的转录本的完整性的量化信息. Genes that make up the BUSCO sets for each major lineage are selected from orthologous groups with genes present as single-copy orthologs in at least 90% of the species. While allowing for rare gene duplications for losses, this establishes an evolutionarily-informed expectation that these genes should be found as single-copy othologs in any newly-sequenced genome.

quick start BUSCO assessments

-m / --mode 评估模式: genome/ proteins/ transcriptome

基因组组装评估: `busco –i sequence_file –o output_name –l lineage –m geno`

对应其他模式为:proteins/prot; transcriptiome/tran

lineage: 为所使用的BUSCO lineage数据(http://busco.ezlab.org/ ; eukaryote_odb9/vertebrata_odb9/fungi_odb9)

运行时间:

> Human genome (3.1 Gbp), assessed with 4'104 mammalian BUSCOs: 6 days 15 hours
> Human gene set (20'398 proteins), assessed with 4'104 mammalian BUSCOs: ~20 minutes
> Human genome (3.1 Gbp), assessed with 978 metazoan BUSCOs: ~21 hours
> Human gene set (20'398 proteins), assessed with 978 metazoan BUSCOs: ~3 minutes
> *Drosophila* genome (140 Mbp), assessed with 2'799 dipteran BUSCOs: ~1 hour 45 minutes
> *Drosophila* gene set (13'954 proteins), assessed with 2'799 dipteran BUSCOs: ~14 minutes
> *Drosophila* genome (140 Mbp), assessed with 978 metazoan BUSCOs: ~19 minutes
> *Drosophila* gene set (13'954 proteins), assessed with 978 metazoan BUSCOs: ~2 minutes
>
> **NB:** more fragmented genomes will take longer as second round searches and gene predictions are performed for BUSCOs found to be fragmented or missing after the first round.

-c N / --cpu N 指定线程或核, 默认1

-e N / --evalue N 指定blast搜索的E-value, 默认0.001/1e-3

-f / --force 强制写入现存文件或目录

-sp SEPCIES / --species SPECIES 现存Augustus 物种基因查询参数名称, 每个leneage拥有一默认物种, 推荐选择最相关的物种

-t PATH / --tmp PATH 临时文件存储位置, 默认: ./tmp

-z / --tarzip 输出文件tarzipped

-r / --restart 从最近完成的步骤开始重新运行BUSCO

--limit REGION_LIMIT 考虑多少候选区域, 整数, 默认为3

--long 开启augustus最佳运行模式, 用于自我训练, 会增加运行时间, 默认为off

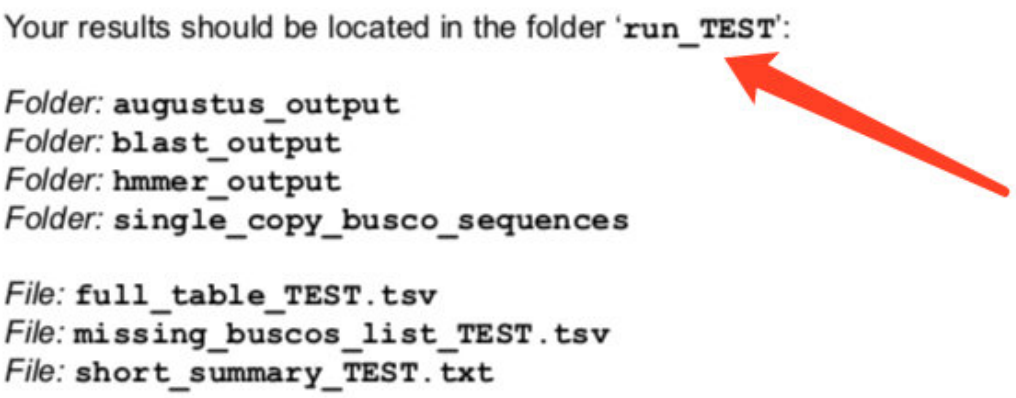
Output

short_summary_XXXX.txt 包含BUSCO注释文件

full_table_XXXX.tsv 表格形式完整结果

missing_buscos_list_XXXX.tsv 包含一系列缺失BUSCOs信息

其他略



例如: short_summary_TEST.txt

```
# BUSCO version is: 3.0.0
# The lineage dataset is: sample dataset BUSCO 2.0 (Creation date:
07.10.2016, number of species: 23, number of BUSCOs: 10)
# To reproduce this run: python scripts/run_BUSCO.py -i
sample_data/target.fa -o TEST -l sample_data/example -m genome -c 1 -f
#
# Summarized benchmarking in BUSCO notation for file
sample_data/target.fa
# BUSCO was run in mode: genome

    C:80.0%[S:80.0%,D:0.0%],F:0.0%,M:20.0%,n:10

    8    Complete BUSCOs (C)
    8    Complete and single-copy BUSCOs (S)
    0    Complete and duplicated BUSCOs (D)
    0    Fragmented BUSCOs (F)
    2    Missing BUSCOs (M)
    10   Total BUSCO groups searched
```

[QUAST][http://quast.bioinf.spbau.ru/manual.html]

QUAST为Quality Assessment Tool. 通过计算不同的metrics评估基因组组装. QUAST用于基因组组装, MetaQUAST为metagenomic数据组装, QUAST-LG为大基因组组装评估(例如, 哺乳动物), Icarus, interactive visualizer for these tools.

QUAST default pipeline utilizes Minimap2. Functional elements prediction modules use GeneMarkS, GeneMark-ES, GlimmerHMM,Barrnap, and BUSCO. QUAST module for finding structural variations applies BWA, Sambamba, and GRIDSS. Also we use bedtools for calculating raw and physical read coverage, which is shown in Icarus contig alignment viewer. Icarus also can use Circos if it is installed in PATH. QUAST-LG introduced modules requiring KMC and Red. In addition, MetaQUAST uses MetaGeneMark, Krona tools, BLAST, and SILVA 16S rRNA database.

Running QUAST

```
./quast.py test_data/contigs_1.fasta \
               test_data/contigs_2.fasta \
               -r test_data/reference.fasta.gz \
               -g test_data/genes.gff
```

查看输出:

```
less quast_results/latest/report.txt
```

Input data

1. sequences, 组装序列和fasta格式的参考基因组, 可以是zip, gzip, bzip2压缩格式
2. reads, fastq格式的Illumina, PacBio, Nanopore reads, 或者SAM/BAM比对格式文件
3. Genes and Operons, 可指定包含参考基因组中的基因和operon位置文件, QUAST将会计算全部和部分比对区域, 输出total values as well as cumulative plots

The following file formats are supported:

- GFF, versions 2 and 3;
- BED: sequence name, start position, end position, gene/operon id, optional fields;
- the format used by NCBI for genes ("Summary (text)");
- four tab-separated columns: sequence name, gene/operon id, start position, end position.

Note that the sequence name has to fully match a name in the reference file.

Coordinates are 1-based, i.e. the first nucleotide in the reference genome has position 1, not 0.

Command line options

-o <output_dir> 输出目录, 默认为quast_results/results_<data_time>

-r <path> 参考基因组文件, 可选参数, metrics在缺少参考基因组时无法实现

--features (or -g) <path> 参考基因组的feature位置信息文件: 若仅计算gff文件中的指定feature信息:

--features CDS:~/data/my_genome_annotation.gff

--features gene:./test_data/genes.gff

默认为所有features

--min-contig (or -m) <int> 最短contig长度阈值(bp), 默认为500

--threads (or -t) <int> 线程

其他略

QUAST output

输出包含:

QUAST output contains:

report.txt	assessment summary in plain text format,
report.tsv	tab-separated version of the summary, suitable for spreadsheets (Google Docs, Excel, etc),
report.tex	LaTeX version of the summary,
icarus.html	Icarus main menu with links to interactive viewers. See section 3.4 for details,
report.pdf	all other plots combined with all tables (file is created if matplotlib python library is installed),
report.html	HTML version of the report with interactive plots inside,
contigs_reports/	(only if a reference genome is provided)
misassemblies_report	detailed report on misassemblies. See section 3.1.2 for details,
unaligned_report	detailed report on unaligned and partially unaligned contigs. See section 3.1.3 for details,
k_mer_stats/	(only if `--k-mer-stats` option is specified)
kmers_report	detailed report on k-mer-based metrics,
reads_stats/	(only if reads are provided)
reads_report	detailed report on mapped reads statistics.

Note:

- metrics based on a reference genome are computed only if a reference is provided (see section 2.3),
- metrics based on genes and operons are computed only if proper annotations are provided (see section 2.3).

其他略