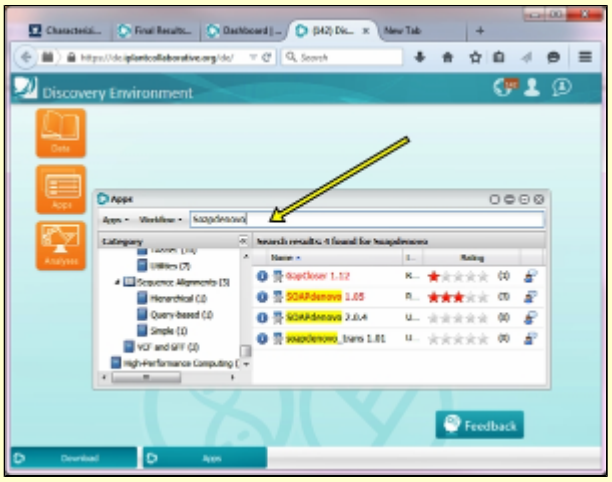
rnaQUAST 1.2.0 (denovo based) using DE

Alert:



The iPlant App Store is currently being restructured, and apps are being moved to an HPC environment. During this transition, users may occasionally be unable to locate or use apps that are listed in our tutorials. In many cases, these apps can be located by searching them using the search bar at the top of the Apps window in the DE. To increase the chance for search success, try not searching the entire app name and version number but only the portion that refers to the app's function or origin (e.g. 'SOAPdenovo' instead of 'SOAPdenovo-Trans 1.01'). In critical cases, please report your concern to the iPlant Ask forum or to support@iplantcollaborative.org. Thank you for your patience.

The DE Quick Start tutorial provides an introduction to basic DE functionality and navigation.
Please work through the tutorial and add your comments on the bottom of this page. Or send comments per email to upendra@cyverse.org. Thank you.

Rationale and background:

rnaQUAST is a tool for evaluating RNA-Seq assemblies using reference genome and gene data database. In addition, rnaQUAST is also capable of estimating gene database coverage by raw reads and *de novo* quality assessment using third-party software. The following tutorial is denovo based quality assessement of transcripts using rnaQUAST 1.2.0. If you have reference genome you can use reference based rnaQUAST 1.2.0 ap p.

Pre-Requisites:

1. A CyVerse account. (Register for an CyVerse account here - user.cyverse.org)
2. Input/Outputs
    a. Transcript file(s) in FASTA format (Mandatory)
    b. Output directory to store all results.
3. Options/Parameters
    a. Run with GeneMarkS-T gene prediction tool. Use `--prokaryote` option if the genome is prokaryotic. Eukaryote is default.
    b. Run BUSCO tool, which detects core genes in the assembly. BUSCO lineage data (Eukaryota, Metazoa, Arthropoda, Vertebrata or Fungi). You can select the BUSCO profile files for your species of interest from here : **/iplant/home/shared/iplantcollaborative/example_data/BUSCO.sample.data**
    c. Run **disable_infer_genes** option if your GTF file already contains genes records, otherwise gffutils will fix it. Note that gffutils may work for quite a long time.
    d. Run **disable_infer_transcripts** if your GTF file already contains transcripts records, otherwise gffutils will fix it. Note that gffutils may work for quite a long time.
    e. Name(s) of assemblies that will be used in the reports separated by space and given in the same order as files with transcripts / alignments.
    f. Set if transcripts were assembled using strand-specific RNA-Seq data in order to benefit from knowing whether the transcript originated from the + or - strand.
    g. Do not draw plots (makes rnaQUAST run a bit faster).

Test/sample data:

The following test data are provided for testing rnaQUAST 1.2.0 in here - /iplant/home/shared/iplantcollaborative/example_data/rnaQUAST.sample.data:

1. idba.fasta
2. spades.311.fasta and
3. Trinity.fasta

de novo quality assessment:

a. Using rnaQUAST 1.1.0 tool with GeneMarkS-T

1. Input file(s): idba.fasta, spades.311.fasta and Trinity.fasta (transcript files)
2. Output folder name - rnaQUAST_output_GM

and leave the rest of the options as default

b. Using rnaQUAST 1.1.0 tool with BUSCO

1. Input file(s): idba.fasta, spades.311.fasta and Trinity.fasta (transcript files)
2. lineage data - Select the BUSCO profile folder "**arthropoda**" from here : **/iplant/home/shared/iplantcollaborative/example_data/BUSCO.sample.data**
3. Output folder name - rnaQUAST_output_arthropoda_BUSCO

and leave the rest of the options as default

Output Reports

The following text files with reports are contained in `comparison_output` directory and include results for all input assemblies. In addition, these reports are contained in`<assembly_label>_output` directories for each assembly separately.

`basic_mertics.txt`
Basic transcripts metrics are calculated without reference genome and gene database.

- ***Transcripts*** – total number of assembled transcripts.
- ***Transcripts > 500 bp***
- ***Transcripts > 1000 bp***

BUSCO metrics. The following metrics are calculated only when `--busco` and `--clade` options are used (see options for details).

- ***Complete*** – percentage of completely recovered genes.
- ***Partial*** – percentage of partially recovered genes.

GeneMarkS-T metrics. The following metrics are calculated when reference and gene database are not provided.

- ***Genes*** – number of predicted genes in transcripts.

`alignment_metrics.txt`
Alignment metrics are calculated with reference genome but without using gene database. To calculate the following metrics rnaQUAST filters all short partial alignments (see `--min_alignmentoption`) and attempts to select the best hits for each transcript.

- *Transcripts* – total number of assembled transcripts.
- *Aligned* – the number of transcripts having at least 1 significant alignment.
- *Uniquely aligned* – the number of transcripts having a single significant alignment.
- *Multiply aligned* – the number of transcripts having 2 or more significant alignments. Multiply aligned transcripts are stored in `<assembly_label>.paralogs.fasta` file.
- Misassembly candidates reported by GMAP (or BLAT) – transcripts that have discordant best-scored alignment (partial alignments that are either mapped to different strands / different chromosomes / in reverse order / too far away).
- *Unaligned* – the number of transcripts without any significant alignments. Unaligned transcripts are stored in `<assembly_label>.unaligned.fasta` file.

Number of assembled transcripts = Unaligned + Aligned = Unaligned + (Uniquely aligned + Multiply aligned + Misassembly candidates reported by GMAP (or BLAT)).

Alignment metrics for non-misassembled transcripts

- *Average aligned fraction.* Aligned fraction for a single transcript is defined as total number of aligned bases in the transcript divided by the total transcript length.
- *Average alignment length.* Aligned length for a single transcript is defined as total number of aligned bases in the transcript.
- Average blocks per alignment. A block is defined as a continuous alignment fragment without indels.
- Average block length (see above).
- *Average mismatches per transcript* – average number of single nucleotide differences with reference genome per transcript.
- NA50 – N50 for alignments.

`misassemblies.txt`

- *Transcripts* – total number of assembled transcripts.
- Misassembly candidates reported by GMAP (or BLAT) – transcripts that have discordant best-scored alignment (partial alignments that

are either mapped to different strands / different chromosomes / in reverse order / too far away).
- Misassembly candidates reported by BLASTN – transcripts are aligned to the isoform sequences extracted from the genome using gene database with BLASTN and then transcripts that have partial alignments to multiple isoforms are selected.
- *Misassemblies* – misassembly candidates confirmed by both methods described above. Using both methods simultaneously allows to avoid considering misalignments that can be caused, for example, by paralogous genes or genomic repeats. Misassembled transcripts are stored in `<assembly_label>.misassembled.fasta` file.

`sensitivity.txt`
Assembly completeness (sensitivity). For the following metrics (calculated with reference genome and gene database) rnaQUAST attempts to select best-matching database isoforms for every transcript. Note that a single transcript can contribute to multiple isoforms in the case of, for example, paralogous genes or genomic repeats. At the same time, an isoform can be covered by multiple transcripts in the case of fragmented assembly or duplicated transcripts in the assembly.

- *Database coverage* – the total number of bases covered by transcripts (in all isoforms) divided by the total length of all isoforms.
- Duplication ratio – total number of aligned bases in assembled transcripts divided by the total number of isoform covered bases. This metric does not count neither paralogous genes nor shared exons, only real overlaps of the assembled sequences that are mapped to the same isoform.
- Average number of transcripts mapped to one isoform.
- *x%-assembled genes / isoforms* / exons – number of genes / isoforms / exons from the database that have at least x% captured by a single assembled transcript, where x is specified with `--lower_threshold / --upper_threshold` options (50% / 95% by default). 95%-assembled isoforms are stored in `<assembly_label>.95%assembled.fasta` file.
- *x%-covered genes / isoforms* – number of genes / isoforms from the database that have at least x% of bases covered by all alignments, where x is specified with `--lower_threshold / --upper_threshold` options (50% / 95% by default).
- *Mean isoform assembly* – assembled fraction of a single isoform is calculated as the largest number of its bases captured by a single assembled transcript divided by its length; average value is computed for isoforms with > 0 bases covered.
- *Mean isoform coverage* – coverage of a single isoform is calculated as the number of its bases covered by all assembled transcripts divided by its length; average value is computed for isoforms with > 0 bases covered.
- Mean exon coverage – coverage of a single exon is calculated as the number of its bases covered by all assembled transcripts divided by its length; average value is computed for exons with > 0 bases covered.
- Average percentage of isoform x%-covered exons, where x is specified with `--lower_threshold / --upper_threshold` options (50% / 95% by default). For each isoform rnaQUAST calculates the number of x%-covered exons divided by the total number of exons. Afterwards it computes average value for all covered isoforms.

`specificity.txt`

Assembly specificity. To compute the following metrics we use only transcripts that have at least one significant alignment and are not misassembled.

- *Unannotated* – total number of transcripts that do not cover any isoform from the database. Unannotated transcripts are stored in `<assembly_label>.unannotated.fasta` file.
- *x%-matched* – total number of transcripts that have at least x% covering an isoform from the database, where x is specified with `--lower_threshold / --upper_threshold` options (50% / 95% by default).
- *Mean fraction of transcript matched* – matched fraction of a single transcript is calculated as the number of its bases covering an isoform divided by the transcript length; average value is computed for transcripts with > 0 bases matched.
- Mean fraction of block matched – matched fraction of a single block is calculated as the number of its bases covering an isoform divided by the block length; average value is computed for blocks with > 0 bases matched.
- x%-matched blocks – percentage of blocks that have at least x% covering an isoform from the database, where x is specified with `--lower_threshold / --upper_threshold` options (50% / 95% by default).
- Matched length – total number of transcript bases covering isoforms from the database.
- Unmatched length – total alignment length - Matched length.

Plots

The following plots are similarly contained in both `comparison_output` directory and `<assembly_label>_output` directories. Please note, that most of the plots represent cumulative distributions and some plots are given in logarithmic scale.

- `Nx.png` – Nx plot for transcripts. Nx is a maximal number N, such that the total length of all transcripts longer than N bp is at least x% of the total length of all transcripts.
- *`transcript_length.png`* – assembled transcripts length distribution (+ database isoforms length distribution).