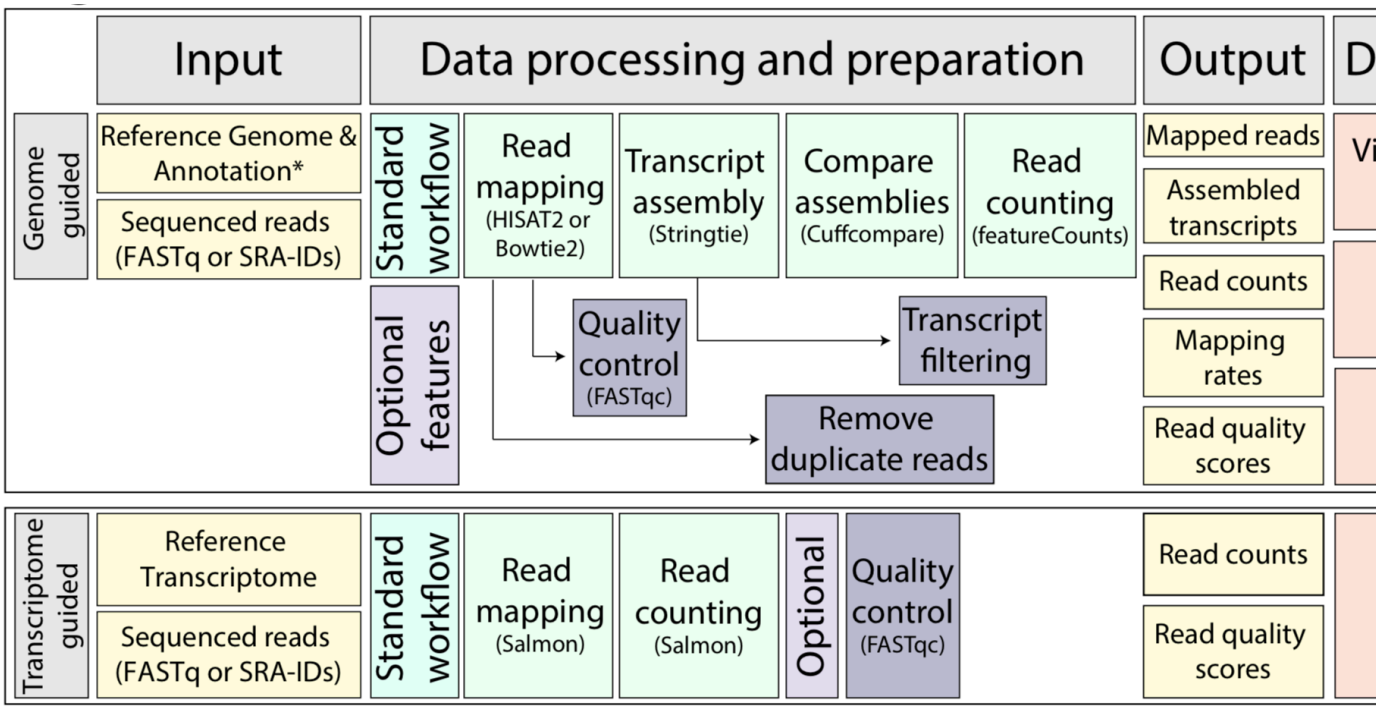
RMTA v2.6.3

The DE Quick Start tutorial provides an introduction to basic DE functionality and navigation.
Please work through the documentation and add your comments on the bottom of this page, or email comments to support@cyverse.org.

Rationale and background:

- RMTA is a high throughput RNA-seq read mapping and transcript assembly workflow. RMTA incorporates the standard RNA-seq analysis programs traditionally used one at a time into a single, easy to use workflow that can rapidly assemble and process any amount of local (FASTq) or NCBI-stored RNA-seq (SRA) data.
- RMTA maps reads to user-provided reference genome using either HISAT2 (transcript analysis) or Bowtie2 (SNP analysis), assembles transcripts using StringTie, and then performs read quantification using FeatureCounts.
- RMTA also supports for read alignment directly to a transcriptome using the quasi-aligner and transcript abundance quantifier Salmon (Rob et al., 2017; Srivastava et al., 2019). Salmon maps reads to the provided transcript assembly and then counts the number of reads associated with each transcript, generating an output file (quant.sf) that can immediately be used for differential expression. **Note:** The utilization of Salmon is only appropriate when the user is wanting to rapidly test for differential expression and cannot facilitate the identification of novel genes or data visualization in a genome browser.
- Beyond read mapping and assembly, RMTA has a number of additional features that automate onerous data transformation and quality control steps, thus producing outputs that can be directly used for differential expression analysis, data visualization, or novel gene identification - data analyses that can all be performed in the DE or at CoGe.



Pre-Requisites:

1. **A CyVerse account (Register for a free CyVerse account at https://user.cyverse.org).**
2. **An up-to-date Java-enabled web browser.**

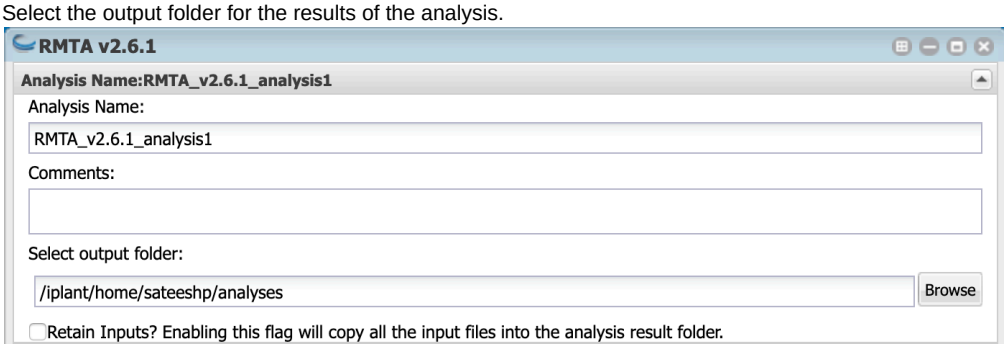
Genome-guided mapping:

Input data requirements:

1. Reference Genome (FASTA) or HISAT2 Indexed Reference Genome (in a subdirectory)
2. Reference Transcriptome (GFF3/GTF/GFF)
3. RNA-Seq reads (FASTQ) - Single end or Paired-end (compressed or uncompressed) or multiple NCBI SRA id's (each SRA ID on a separate row in the text file).

1. **Mandatory fields**

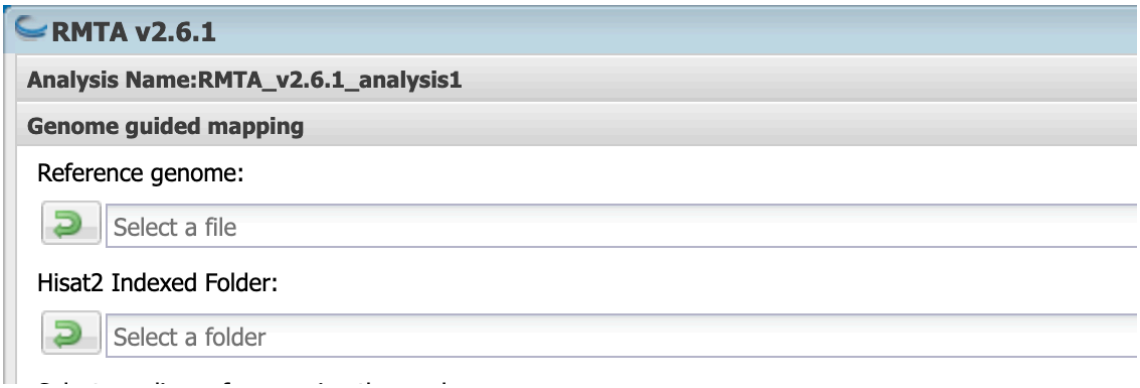
a. **Analysis Name**
   i. Choose an appropriate name for your analysis and make comments if you wish. Default name is shown in the figure below.
   ii. Select the output folder for the results of the analysis.



b. **Genome guided mapping**

> ⚠ *****Select at least one of the below two options for the indexing of the Reference Genome*****

   i. Custom genome (required)
   ii. HISAT2 Indexed folder (for indexed genomes)



c. **Select an aligner**

> ⚠ *****Only one of the below two options needs to be selected. Both cannot be selected.*****

i) Hisat2
ii) Bowtie2

> ⓘ **When to use HISAT2 or Bowtie2**
>
> HISAT2 is a splice-aware algorithm used to perform reference genome-based read mapping. Stringtie is then used to assemble transcripts based on this read mapping.
>
> The read aligner Bowtie2 has been included as an optional aligner in the RMTA workflow for users wishing to call single nucleotide polymorphisms (SNPs) from their RNA-seq (or DNA-seq) data in a high throughput manner. When the Bowtie2 option is selected, HISAT2 and Stringtie are both removed from the workflow, but the additional option to remove duplicate reads (important for population level analyses) becomes available.

d. **Reference annotation**

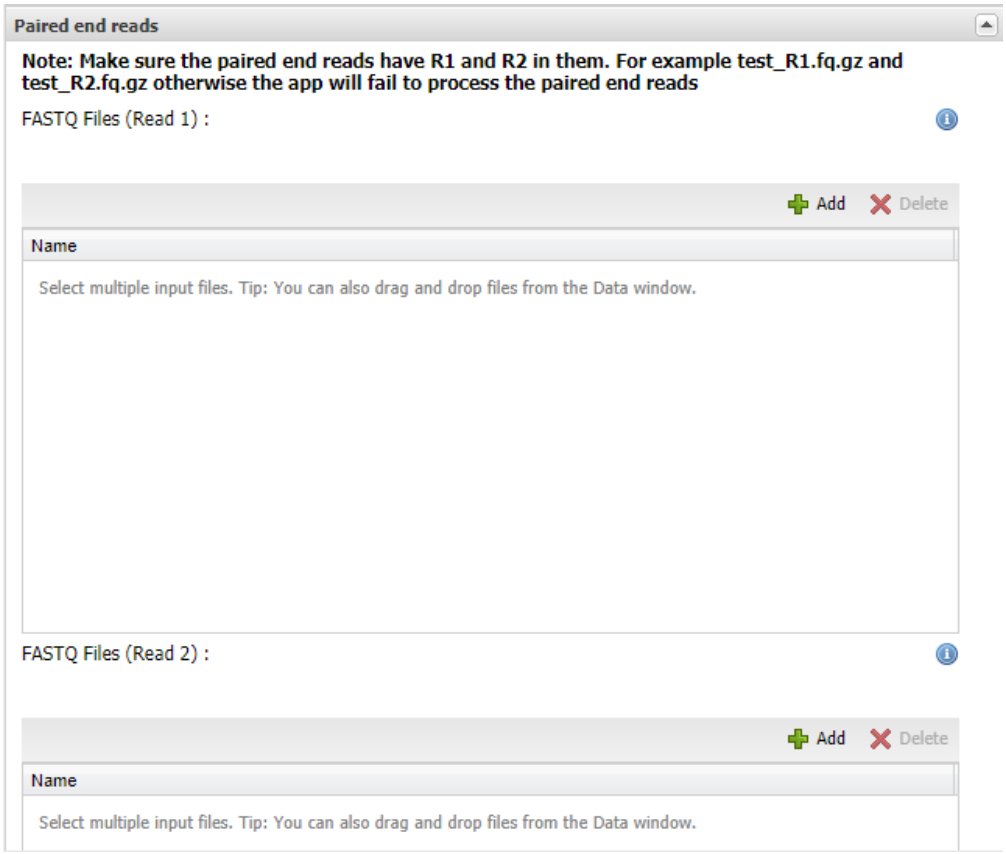
Providing reference annotation is optional and not required when doing *de novo* transcriptome assembly

e. **Feature Count Options**
   i. Choose a Feature Type. The default option will be "exon"
   ii. Choose a Gene Attribute. The default option will be "gene_id"
   iii. Select the Type of Strandedness. The three options include unstranded, stranded, and reversely stranded.
   iv. Please refer to your Genome Annotation File (.gtf), and confirm that these settings match your data. For Gene Attribute, be sure that gene_id is written before the name of each gene.

f. **Input reads**

Only one of the following three read options (d, e, or f) may be selected per job.

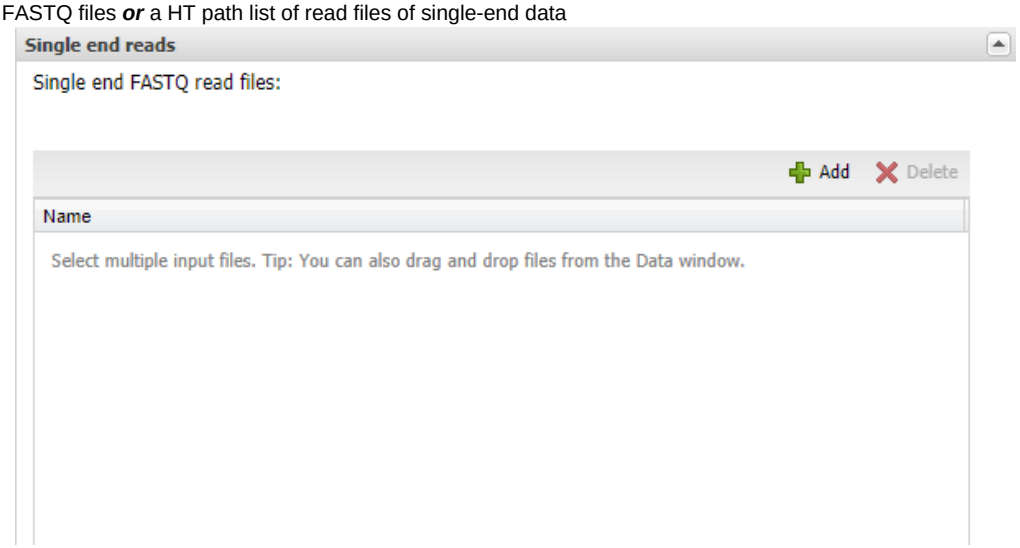
Paired-end reads
   i. FASTQ Files (Read 1): HT path list of read 1 files of paired-end data
   ii. FASTQ Files (Read 2): HT path list of read 2 files of paired-end data



⚠ ***When inserting multiple paired end FASTQ files, be sure that reads 1 and 2 have matching patterns***

Single-end reads

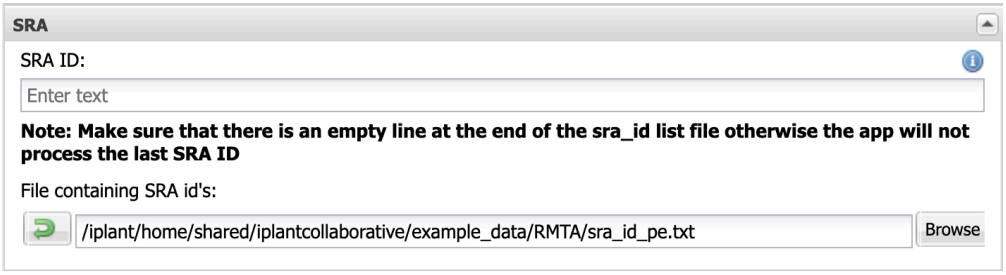
i. Single end FASTQ files ***or*** a HT path list of read files of single-end data



SRA

i. Enter the SRA id, ***or***

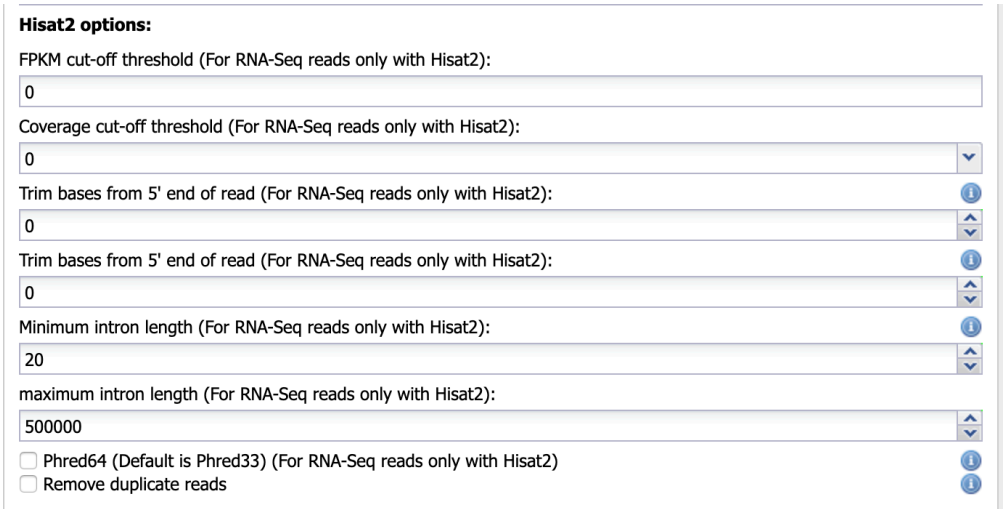
ii. Select a file containing a list of SRA ids (one per line) *or* a HT path list of multiple SRA ids list files



If you have many files to process through the Discovery Environment, an HT Analysis Path List File may prove useful, as this app takes only 1 file at a time. For information on how to create an HT path list, click here

g. **Parameters**
   i. Type of Sequence: Choose either Single End or Paired End
   ii. Choose RNA strandedness (default is unstranded)
   iii. Number of threads (Default is 4)
   iv. Run FastQC

h. **Advanced options:**



Phred64, Fastqc, and Remove Duplicate reads options
Phred gives a quality score of how confidently nucleotides were identified during sequencing. Here, Phred33 is default. Phred 64 should be used if sequencing was performed using Illumina technology 1.3 - 1.8. An error will occur and the run will fail if the wrong quality score has been selected.

FastQC provides the user with both an overview of potential issues with the data, as well as summary graphs highlighting issues such as per base sequence quality and Kmer content. When the FastQC option has been selected,

BAM files are converted back into FASTq, with mapped and unmapped reads, along with their associated quality score retained. This FASTq file is then used as input for FastQC. If issues are detected at the 5' or 3' of sequencing reads, RMTA includes additional options for specifically trimming bases off of either end during the next analysis. Sequencing reads of overall poor quality will simply not be mapped and therefore do not need to be trimmed. FASTq files are removed following FastQC analysis.

If the user chose Bowtie2 as the read aligner and "remove duplicate reads" as an additional option, then the RMTA_Output folder will only contain a sorted BAM file with duplicates removed for each SRA/FASTq input file, as well as a mapped.txt file. No additional files will be generated.

When using Bowtie2
When using Bowtie2, be sure to check the box labeled "Remove duplicate reads," as shown in the figure below. Duplicate removal is suggested when performing SNP analysis

i. ***RMTA_Output:***
Name of the output folder (Default is RMTA_Output)

Example Runs:

The following test data using Arabidopsis are provided for testing RMTA in here - /iplant/home/shared/iplantcollaborative/example_data/RMTA (this path can be copied and pasted into the navigation bar in a data window within the DE)

*****Note that when testing SRA IDs, only one of steps 3, 4, or 5 may be used at a time per test run.*****

1. Reference Genome: genome_chr1.fa
2. Reference Annotation: genome_chr1.gtf
3. Paired End Reads:
    a. Left End Reads: SRR2037320_R1.fastq.gz and SRR2932454_R1.fast.gz
    b. Right End Reads: SRR2037320_R2.fastq.gz and SRR2932454_R2.fastq.gz
4. Single End Reads: SRR3464102.fastq.gz and SRR3464103.fastq.gz
5. List of SRA IDs:
    a. Paired End: sra_id_pe.txt
    b. Single End: sra_id_se.txt
6. Aligners: HISAT2
7. Feature Count: leave as default
8. Advanced Options:
    a. Type of Sequence: select Single for Single End and Paired for Paired End
    b. Leave the rest as default
9. RMTA Output: leave as default

All other settings should be left as default.

Results

Successful execution of RMTA will generate three output folders:

1. Index: This folder consists of the index of the genome
2. Output: This folder consists of the output from HISAT2, Stringtie and Cuffcompare as well as the Feature_counts and FASTqc (optional) folders.
    a. In turn, this folder will consist of five files associated with each SRA or FASTq:
        i. A filtered GTF (if either the read/base or FPKM filter was set).
        ii. A sorted.bam file of all mapped and unmapped reads.
        iii. An index associated with the above BAM file.
        iv. a GTF that represents the unprocessed file straight out of Stringtie if the user would like to investigate their data further.
        v. A combined.gtf that is the output of the cuffmerge step of RMTA (comparing all identified transcripts back to the reference to identify novel transcripts). This file is useful for novel transcript identification.
    b. The Feature_counts folder contains one file with all of the read count data for each gene across all SRA/FASTq files examined in the run.
    c. The FASTqc_out folder contains subfolders associated with each SRA/FASTq input file. In each folder is an html file with all of

the details from the FASTqc run for each set of reads (1 for SE, 2 for PE).

3. Logfiles: This folder consists of stout and sterr (information written to standard out or standard error log files) files as well as logs specific to running within the DE

Transcriptome-guided Mapping:

Input data requirements:

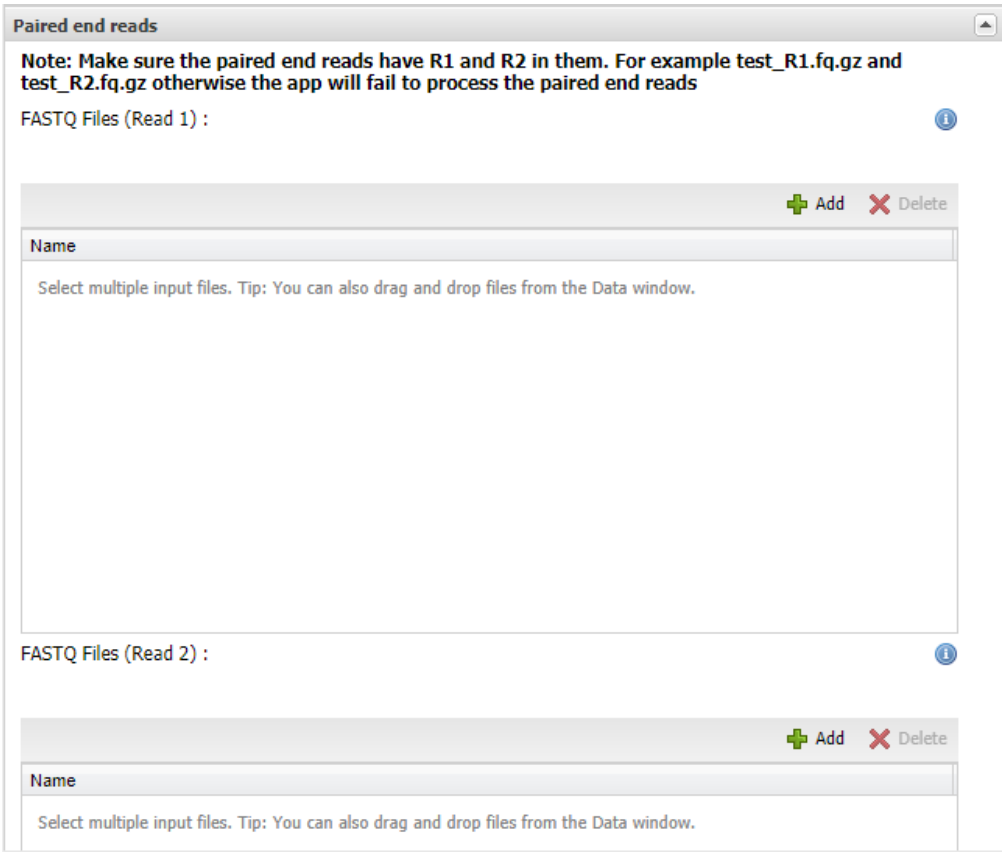
1. Reference Genome (FASTA) or HISAT2 Indexed Reference Genome (in a subdirectory)
2. Reference Transcriptome (GFF3/GTF/GFF)
3. RNA-Seq reads (FASTQ) - Single end or Paired-end (compressed or uncompressed) or multiple NCBI SRA id's (each SRA ID on a separate row in the text file).

1. **Mandatory fields**
    a. **Reference transcriptome**

    b. **Input data**

> ⚠ *****Only one of the following three read options (d, e, or f) may be selected per job.*****

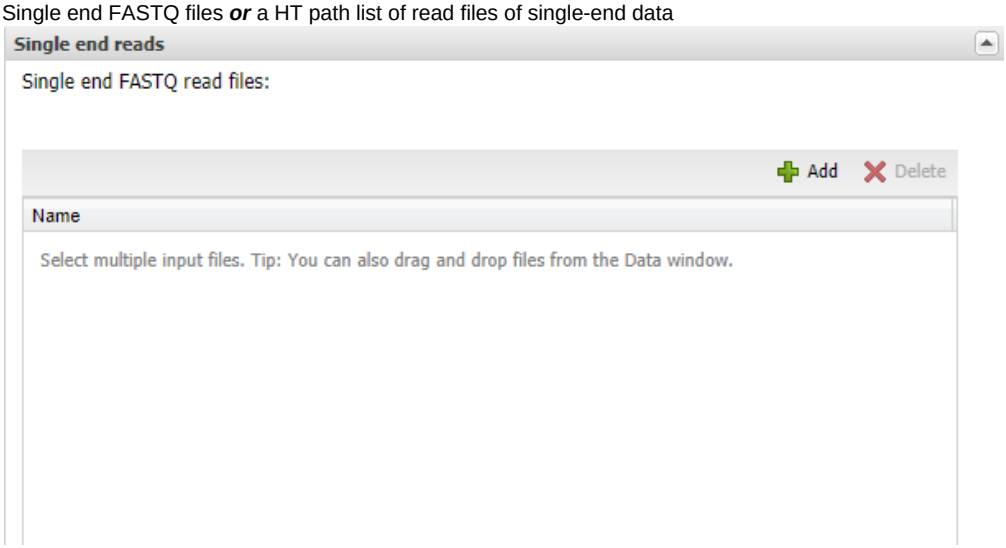
        i. **Paired-end reads**
            1. FASTQ Files (Read 1): HT path list of read 1 files of paired-end data
            2. FASTQ Files (Read 2): HT path list of read 2 files of paired-end data



*****When inserting multiple paired end FASTQ files, be sure that reads 1 and 2 have matching patterns*****

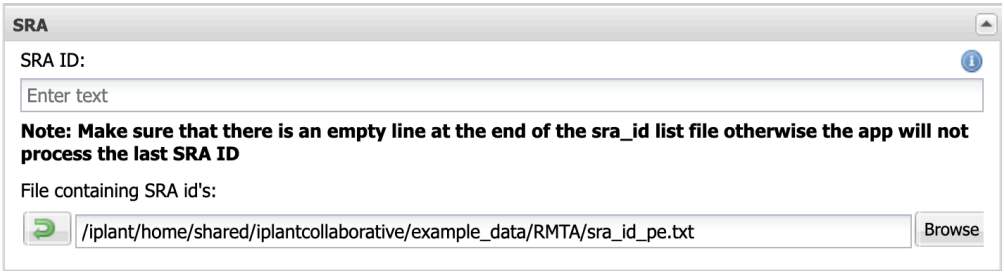
ii. **Single-end reads**

1. Single end FASTQ files ***or*** a HT path list of read files of single-end data



iii. **SRA**

1. Enter the SRA id, ***or***
2. Select a file containing a list of SRA ids (one per line) ***or*** a HT path list of multiple SRA ids list files



c. ***Parameters:***

i. Type of Sequence: Choose either Single End or Paired End

ii. Choose RNA strandedness (default is unstranded)

iii. Number of threads (Default is 4)

iv. Run FastQC

Example Runs:

The following test data using Arabidopsis are provided for testing RMTA in here - /iplant/home/shared/iplantcollaborative/example_data/RMTA (this path can be copied and pasted into the navigation bar in a data window within the DE)

*****Note that when testing SRA IDs, only one of steps 2, 3, or 4 may be used at a time per test run.*****

1. Reference Transcriptome: athal.fa.gz
2. Paired End Reads:
    a. Left End Reads: SRR2037320_R1.fastq.gz and SRR2932454_R1.fast.gz
    b. Right End Reads: SRR2037320_R2.fastq.gz and SRR2932454_R2.fastq.gz
3. Single End Reads: SRR3464102.fastq.gz and SRR3464103.fastq.gz
4. List of SRA IDs:
    a. Paired End: sra_id_pe.txt
    b. Single End: sra_id_se.txt
5. Advanced Options:
    a. Type of Sequence: select Single for Single End and Paired for Paired End
    b. Leave the rest as default
6. RMTA Output: leave as default

All other settings should be left as default.

Results

Successful execution of RMTA will generate three output folders:

1. Index: This folder consists of the salmon index of the transcriptome
2. Output: This folder consists of the output from salmon quasi-mapping for each sample and FASTqc (optional) folders.
    a. aux_info: Auxiliary files
    b. libParams: This JSON file reports the number of fragments that had at least one mapping compatible with the designated library format, as well as the number that didn't.
    c. cmd_info.json: file that records the main command line parameters with which Salmon was invoked
    d. lib_format_counts.json:
    e. quant.sf: Salmon's main output quantification file
3. Logfiles: This folder consists of stout and sterr (information written to standard out or standard error log files) files as well as logs specific to running within the DE