RNA-Seq data processing pipeline

Yizhou Wang
1/31/2019

Mapping_QC_Auto_v3 for RNA-seq data processing

- Perl based script
- Run on HPC
- 6 samples (20M reads/sample) take ~2h
- Get email notification when jobs finished
- Ready-to-deliver results

Command Line:

Mapping_QC_Auto_v3.pl -t <sequencing_type> -o <Species> -p <project_ID> -n <nodes number> -qc -gb

-t : single end or paired end
-o : Human/Mouse/other species
-p : project ID for these set of samples
-n : which nodes the jobs will run on
-qc: perform QC (if not, only do alignment)
-gb: perform genebody test to check 5’ or 3’ mapping bias

Integrated tools

- Aligner: STAR
- Quantification: RSEM
- Quality Control
  - Raw reads: FastQC
  - Mapping: RSeQC
- Files organization/format
  - customized perl/R/bash scripts

Input:

- FASTQ files only

Output:

- **final_results**:
  - Final QC report (deliver)
  - TPM/Count/FPKM expression matrix (deliver)

- **fastq**: original FASTQ files (deliver)

- **bam**: sorted bam and bam index files

- **others**: other intermediate files

- **RseQC_results**: QC report for each sample and intermediate files from RSeQC

- **genes_isoforms_results**: quantification results for each sample generated by RSEM

- **log && node_log** : log files from running nodes and tools

Example for QC report

- MultiQC HTML report:

http://10.220.239.17/demo.html

- Excel table

Sample	#_raw_reads	#_unique_reads	%_unique_reads	#_multi_mapping_reads	%_multi_map	total%_mapping	CDS	UTR	intron	mtRNA	rRNA	tRNA	ERCC
11--ILC2-N-900cells_S11_R1_001	28859600	20994112	72.75%	6073959	21.05%	93.79%	42.34%	19.67%	14.92%	2.74%	1.32%	0.63%	0.01%
14--ILC2-N-900cells_S6_R1_001	26292501	19594511	74.53%	4907889	18.67%	93.19%	38.13%	19.19%	17.82%	3.41%	2.67%	0.61%	0.01%
15--ILC2-P-900cells_S10_R1_001	27138370	20502734	75.55%	4568117	16.83%	92.38%	37.54%	19.88%	18.22%	2.87%	1.21%	0.78%	0.01%
16--ILC2-P-900cells_S5_R1_001	31291537	21805706	69.69%	7354772	23.50%	93.19%	52.76%	22.75%	8.93%	3.59%	2.17%	0.87%	0.01%
17--ILC2-P-900cells_S2_R1_001	30149984	21830004	72.40%	6278190	20.82%	93.23%	48.12%	21.53%	12.62%	3.07%	2.65%	0.75%	0.02%
19--ILC3-N-9000cells_S1_R1_001	35876236	24854664	69.28%	6674862	18.61%	87.88%	18.03%	3.31%	33.68%	1.20%	0.07%	0.17%	0.00%
1--ILC1-N-9000cells_S4_R1_001	29790674	22341290	74.99%	5422404	18.20%	93.20%	35.93%	20.19%	21.51%	3.14%	0.61%	0.54%	0.00%

Support

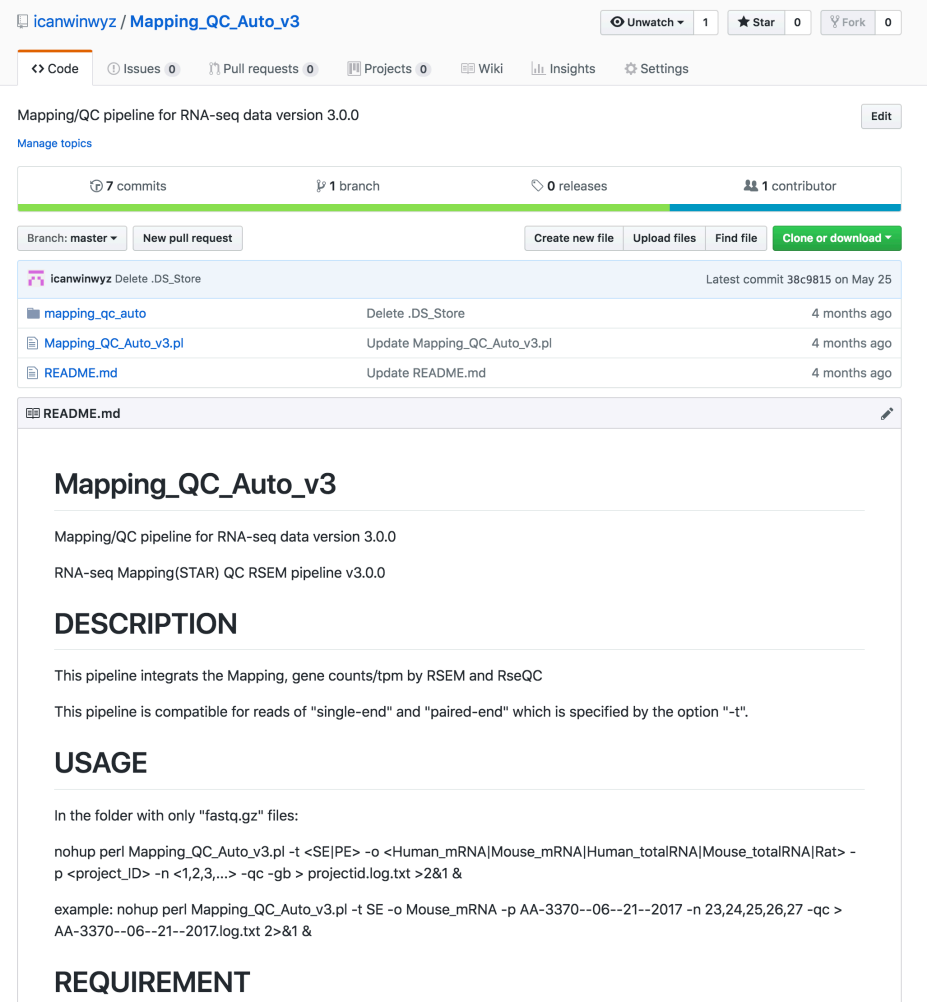
- Detailed help manual:

Mapping_QC_Auto.pl -h

```
NAME
    RNA-seq Mapping(STAR) QC RSEM pipeline v3

DESCRIPTION
    This pipeline integrats the Mapping, gene counts/tpm by RSEM and RseQC

    This pipeline is compatible for reads of "single-end" and "paired-end"
    which is specified by the option "-t".

USAGE
    In the folder with only "fastq.gz" files:

    nohup perl Mapping_QC_Auto_v3.pl -t <SE|PE> -o
    <Human_mRNA|Mouse_mRNA|Human_totalRNA|Mouse_totalRNA|Rat> -p <project_ID>
    -n <1,2,3,...> -qc -gb > projectid.log.txt >2&1 &

    example: nohup perl Mapping_QC_Auto_v3.pl -t SE -o Mouse_mRNA -p
    AA-3370--06--21--2017 -n 23,24,25,26,27 -qc >
    AA-3370--06--21--2017.log.txt 2>&1 &

REQUIREMENT
    - Perl 5

    - perl module: Getopt::Long

OPTIONS
    Running options:

    #-e or --email [optional] # Provide your email address if you would like
    to be notified after jobs completed.

    -t or --type [required if no samplesheet supplied] The sequencing type is
    "single end (SE)" or "paired end (PE)"

    -o or --organism the reference genoem: Human or Mouse

    -p or --project project ID for this run

    -qc or --qualitycontrol whether run quality contorl (RSeQC) or not after
    mapping

    -gb or --genebody whether run genebody test to check 3' or 5' bias
```

Support

- Deposited on GitHub (https://github.com/icanwinwyz/Mapping_QC_Auto_v3)



Deployed on Titan Server Portal for wet lab

- For small sample sets (< 10sampes), wet lab can do Mapping/QC by clicking button.
- Jobs run on local server instead of HPC

Path to Fastq files:

Organism:
- (•) Human
- ( ) Mouse
- ( ) Rat
- ( ) Others

Sequencing Type:
- (•) SE
- ( ) PE
- ( ) 10X_scRNA
- ( ) Others

Project Name:

Email:

Perfomr QC?
( ) Yes (•) No

Submit