Легковой Станислав, GSE72573
/home/slegkovoi/

1) обрезаем адаптеры у ридов (всего 4 набора парных ридов) с помощью **Trimmomatic**:

java -jar ../../../scratch/tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE -phred33 SRR2225457_1.fastq.gz SRR2225457_2.fastq.gz out_f_p_1.fastq.gz out_f_unp_1.fastq.gz out_r_p_1.fastq.gz out_r_unp_1.fastq.gz LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20

java -jar ../../../scratch/tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE -phred33 SRR2225458_1.fastq.gz SRR2225458_2.fastq.gz out_f_p_2.fastq.gz out_f_unp_2.fastq.gz out_r_p_2.fastq.gz out_r_unp_2.fastq.gz LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20

java -jar ../../../scratch/tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE -phred33 SRR2225459_1.fastq.gz SRR2225459_2.fastq.gz out_f_p_3.fastq.gz out_f_unp_3.fastq.gz out_r_p_3.fastq.gz out_r_unp_3.fastq.gz LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20

java -jar ../../../scratch/tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE -phred33 SRR2225460_1.fastq.gz SRR2225460_2.fastq.gz out_f_p_4.fastq.gz out_f_unp_4.fastq.gz out_r_p_4.fastq.gz out_r_unp_4.fastq.gz LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20

2а) воспользуемся **rnaSPAdes**, чтобы собрать все риды в один транскриптом (первая сборка):

../../scratch/tools/SPAdes-3.13.1-Linux/bin/rnaspades.py --pe1-1 aft_trim/1_PE/out_f_p_1.fastq --pe1-2 aft_trim/1_PE/out_r_p_1.fastq --pe1-1 aft_trim/2_PE/out_f_p_2.fastq --pe1-2 aft_trim/2_PE/out_r_p_2.fastq --pe1-1 aft_trim/3_PE/out_f_p_3.fastq --pe1-2 aft_trim/3_PE/out_r_p_3.fastq --pe1-1 aft_trim/4_PE/out_f_p_4.fastq --pe1-2 aft_trim/4_PE/out_r_p_4.fastq -o rna_spades_output

2б) воспользуемся **Trinity,** чтобы собрать все риды в один транскриптом (вторая сборка);

../../../scratch/tools/trinityrnaseq-v2.8.6/Trinity --seqType fq
--left aft_trim/1_PE/out_f_p_1.fastq,aft_trim/2_PE/out_f_p_2.fastq,aft_trim/3_PE/out_f_p_3.fastq,aft_trim/4_PE/out_f_p_4.fastq
--right aft_trim/1_PE/out_r_p_1.fastq,aft_trim/2_PE/out_r_p_2.fastq,aft_trim/3_PE/out_r_p_3.fastq, aft_trim/4_PE/out_r_p_4.fastq --CPU 8 --max_memory 64G

3а) с помощью **BUSCO** оценивался сам транскриптомы обоих сборок (*transcripts.fasta* и *Trinity.fasta*):

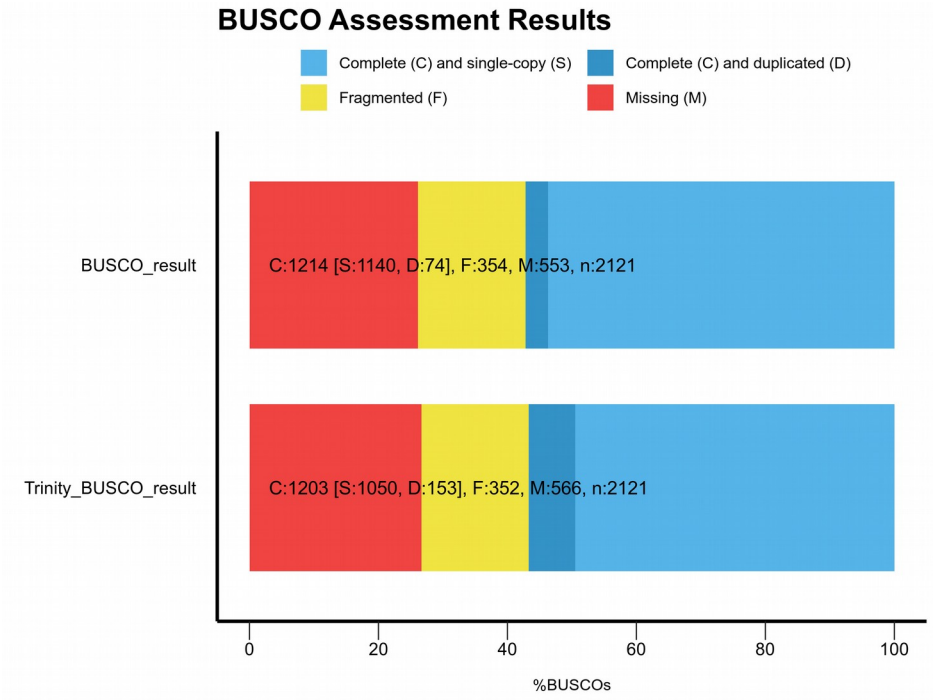
python3 scripts/run_BUSCO.py -i ../transcripts.fasta -o BUSCO_result -l ../eudicotyledons_odb10/ -m tran

python3 scripts/run_BUSCO.py -i ../Trinity.fasta -o Trinity_BUSCO_result -l ../eudicotyledons_odb10/ -m tran

для ключа -l скачивался отдельный файл *eudicotyledons_odb10.tar.gz с сайта BUSCO*, т.к. модельный организм статьи- *Pisum sativum L.* - двудольное растение

также для визуального отображения состояния транскриптомов (файлы *short_summary_BUSCO_result.txt* и *short_summary_Trinity_BUSCO_result.txt* копировались в отдельную папку my_summary) выполнялась следующая команда из BUSCO:

python3 scripts/generate_plot.py –wd my_summary/



3б) помимо **BUSCO** для сравнения сборок транскриптомов использовался **Transrate**:

*../../../scratch/tools/transrate-1.0.3-linux-x86_64/transrate *
--assembly rna_spades_output/transcripts.fasta,trinity_out_dir/Trinity.fasta –threads 8

rnaSPAdes		**Trinity**	
n seqs	42781	n seqs	36332
smallest	49	smallest	201
largest	11531	largest	14207
n bases	28725804	n bases	31049966
mean len	**633.52**	mean len	**854.62**
n under 200	9265	n under 200	0
n over 1k	9234	n over 1k	10697
n over 10k	4	n over 10k	7
n with orf	13853	n with orf	15878
mean orf percent	74.26	mean orf percent	73.15
n90	340	n90	360
n70	818	n70	807
n50	**1343**	n50	**1298**
n30	1997	n30	1875
n10	3829	n10	3086
gc	0.41	gc	0.41
bases n	4576	bases n	0
proportion n	0.0	proportion n	0.0

4) полученный на этапе (2а) транскриптом из **rnaSPAdes** (*transcripts.fasta*) подвергался процедуре удаления похожих транскриптов (коллапсирование) с помощью программы **cd-hit**:

./cd-hit-est -i ../../BUSCO/transcripts.fasta -o ~/result_cd_hit.fasta -c 0.95 -n 10 -d 0 -M 0M -T 0

5) полученный после работы **cd-hit-est** файл (*result_cd_hit.fasta*) аннотировался с помощью blastx, где в качестве референса использовался файл *Uniprot SwissProt;*

сначала из *Uniprot SwissProt* создаем базу данных:

../../../scratch/tools/ncbi-blast-2.9.0+/bin/makeblastdb -in ~/uniprot_sprot.fasta -parse_seqids -dbtype prot -out ~/my_db

затем уже **blastx**:

./../../../scratch/tools/ncbi-blast-2.9.0+/bin/blastx -query result_cd_hit.fasta -out blastx_output.txt -db /home/slegkovoi/my_db/my_db -num_threads 8

Фрагмент *blastx_output.txt:*

Query= NODE_1_length_11531_cov_7.257359_g0_i0

Length=11531

Sequences producing significant alignments:	Score (Bits)	E Value
Q8H0T4 E3 ubiquitin-protein ligase UPL2 OS=Arabidopsis thaliana O...	1103	0.0
Q8GY23 E3 ubiquitin-protein ligase UPL1 OS=Arabidopsis thaliana O...	1103	0.0
Q9P4Z1 E3 ubiquitin-protein ligase TOM1-like OS=Neurospora crassa...	498	2e-139
O13834 E3 ubiquitin-protein ligase ptr1 OS=Schizosaccharomyces po...	456	5e-127
Q756G2 Probable E3 ubiquitin-protein ligase TOM1 OS=Ashbya gossyp...	446	5e-124
Q03280 E3 ubiquitin-protein ligase TOM1 OS=Saccharomyces cerevisi...	441	2e-122
Q7Z6Z7 E3 ubiquitin-protein ligase HUWE1 OS=Homo sapiens OX=9606 ...	430	7e-119
Q7TMY8 E3 ubiquitin-protein ligase HUWE1 OS=Mus musculus OX=10090...	430	7e-119
P51593 E3 ubiquitin-protein ligase HUWE1 (Fragment) OS=Rattus nor...	359	1e-110
F8W2M1 E3 ubiquitin-protein ligase HACE1 OS=Danio rerio OX=7955 G...	365	3e-105
O14326 E3 ubiquitin-protein ligase pub3 OS=Schizosaccharomyces po...	360	8e-105
Q92462 E3 ubiquitin-protein ligase pub1 OS=Schizosaccharomyces po...	351	4e-102
D3ZBM7 E3 ubiquitin-protein ligase HACE1 OS=Rattus norvegicus OX=...	355	1e-101
Q3U0D9 E3 ubiquitin-protein ligase HACE1 OS=Mus musculus OX=10090...	354	1e-101
Q8IYU2 E3 ubiquitin-protein ligase HACE1 OS=Homo sapiens OX=9606 ...	354	2e-101
Q28BK1 E3 ubiquitin-protein ligase HACE1 OS=Xenopus tropicalis OX...	353	2e-101
Q6DCL5 E3 ubiquitin-protein ligase HACE1 OS=Xenopus laevis OX=835	353	5e-101
F1N6G5 E3 ubiquitin-protein ligase HACE1 OS=Bos taurus OX=9913 GN...	352	5e-101
E1C656 E3 ubiquitin-protein ligase HACE1 OS=Gallus gallus OX=9031...	352	2e-100
Q5BDP1 E3 ubiquitin-protein ligase RSP5 OS=Emericella nidulans (s...	332	5e-95
A1D3C5 Probable E3 ubiquitin-protein ligase hulA OS=Neosartorya f...	332	7e-95
Q0CCL1 Probable E3 ubiquitin-protein ligase hulA OS=Aspergillus t...	331	9e-95
Q4WTF3 Probable E3 ubiquitin-protein ligase hulA OS=Neosartorya f...	331	1e-94
B0XQ72 Probable E3 ubiquitin-protein ligase hulA OS=Neosartorya f...	331	1e-94
G0S9J5 E3 ubiquitin-protein ligase RSP5 OS=Chaetomium thermophilu...	330	4e-94
A2QQ28 Probable E3 ubiquitin-protein ligase hulA OS=Aspergillus n...	330	4e-94
P39940 E3 ubiquitin-protein ligase RSP5 OS=Saccharomyces cerevisi...	329	6e-94
A1CQG2 Probable E3 ubiquitin-protein ligase hulA OS=Aspergillus c...	328	9e-94

…

p.s.: был бы поумнее, поставил бы флаги *-evalue 1e-5 -max_target_seqs 5*, тогда посчиталось бы быстрее + также было и в статье сделано

p.p.s: в папке */home/slegkovoi/fastqc_dir* лежат HTML-отчёты по качеству ридов от fastqc, но их вставлять было уже лень