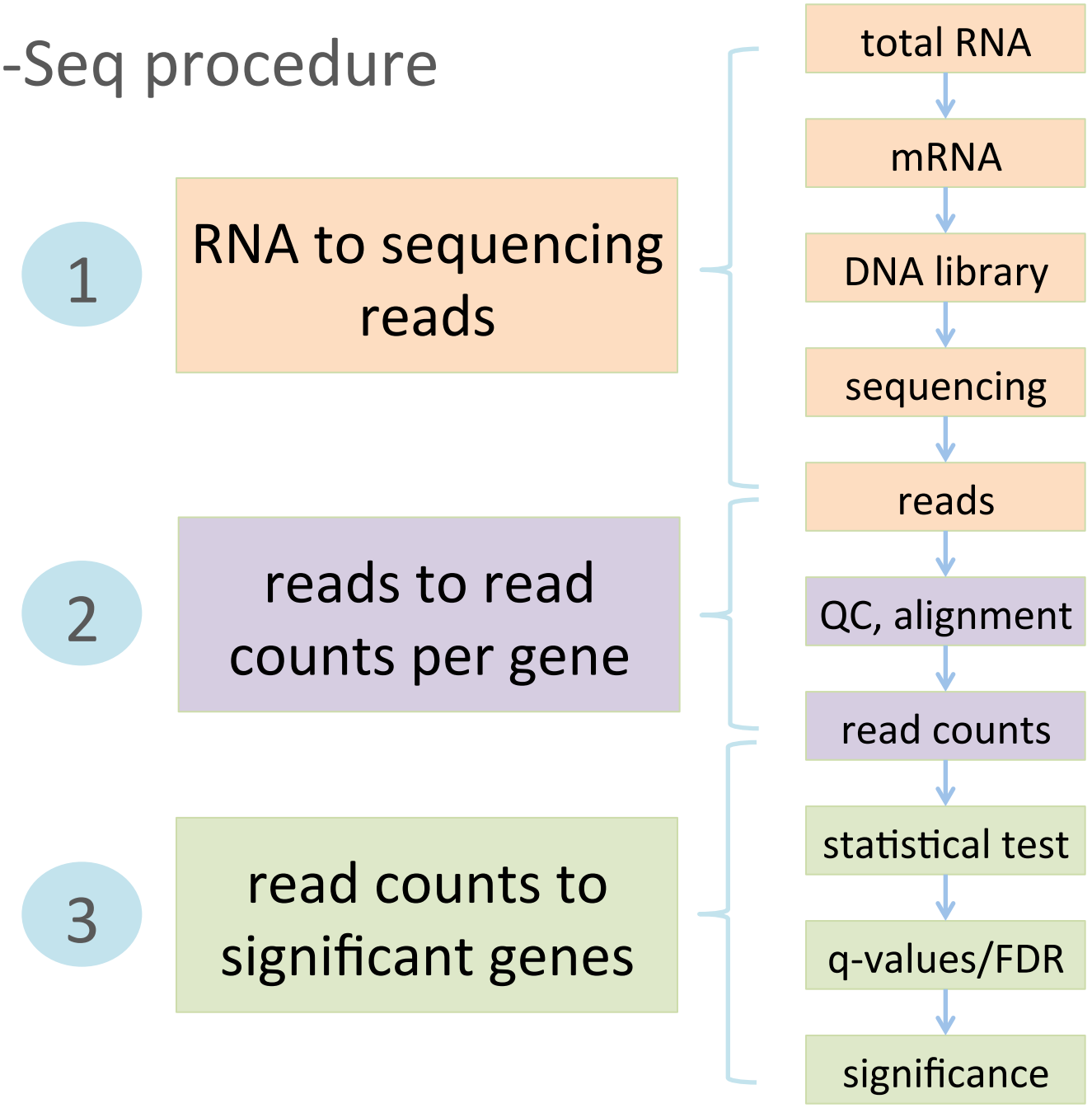
In-class project – DE

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

4/25/2017

RNA-Seq procedure



data information

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) **84**, 491–503

doi: 10.1111/tpj.13014

Genomic limitations to RNA sequencing expression profiling

Cory D. Hirsch[1], Nathan M. Springer[1] and Candice N. Hirsch[2,*]

[1]*Department of Plant Biology, University of Minnesota, St Paul, MN 55108, USA, and*

[2]*Department of Agronomy and Plant Genetics, University of Minnesota, St Paul, MN 55108, USA*

Sequence Read Archive (SRA; B73 control accessions SRR1238718, SRR1819617, SRR1819621; B73 cold accessions SRR1238717, SRR1819204, SRR1819205). Sequence adapters were

SRR1238718

Makarevitch *et al*., 2015

Hirsch *et al*., 2016

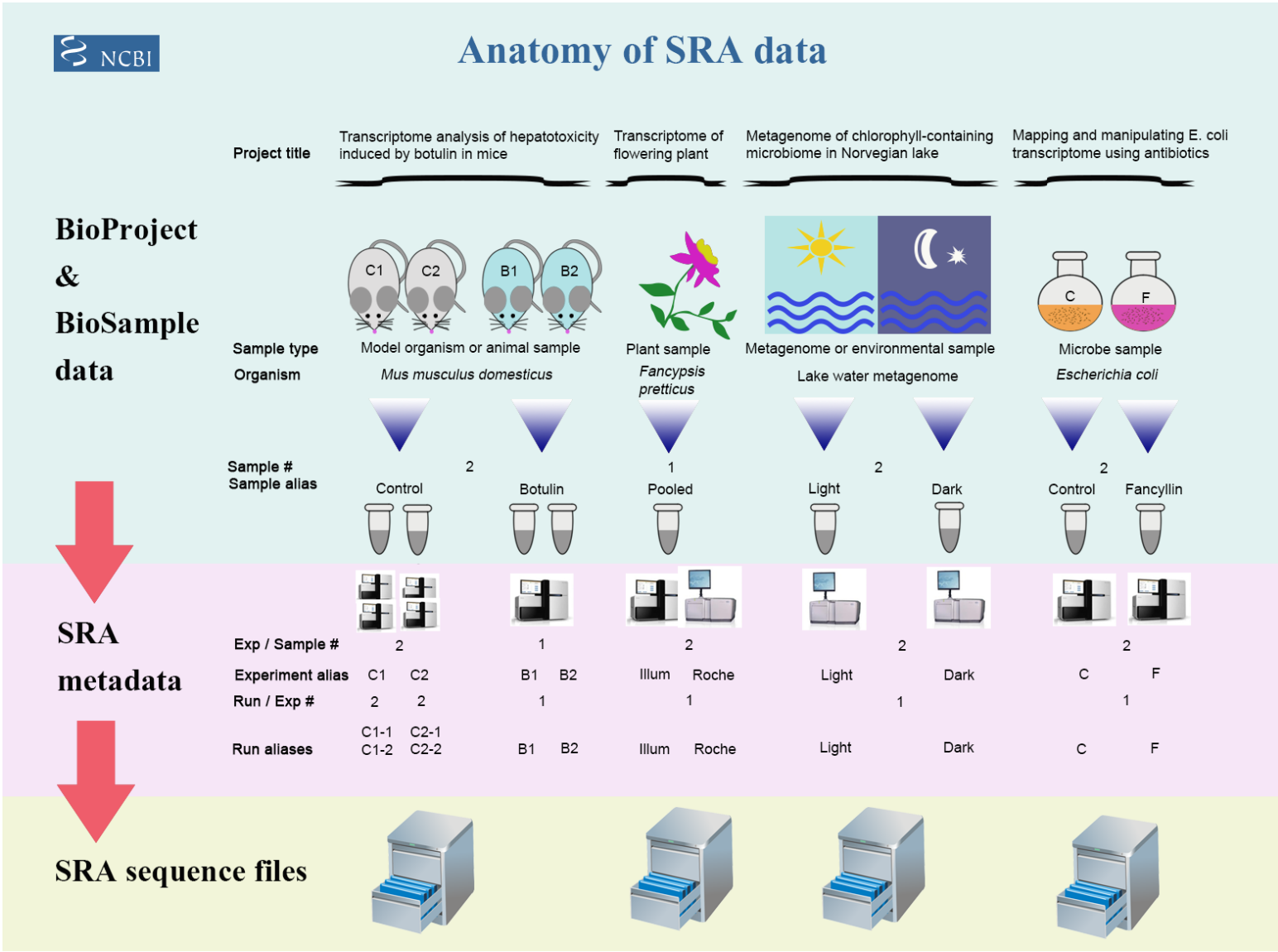
Part I: Data downloading

- Introduction of Sequence Read Archive (SRA) (2007)

- Data download (SRA toolkit)

- fastq-dump

Framework of data submission



Metadata and sequence data

- **Study** – a set of experiments with an overall goal

SRA Study accessions – SRP, DRP, or ERP

- **Experiment** –laboratory operations on input material

SRA Experiment accessions – SRE, DRE, or ERE

- **Sample** – An experiment targets one or more samples

SRA Sample accessions – SRS, DRS, or ERS

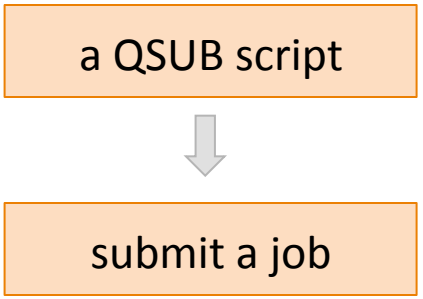
- **Run** –the data gathered for a sample or sample bundle

SRA Run accessions – SRR, DRR, or ERR

format conversion - fastq-dump in Beocat

fastq-dump [options] <accession>

```
#!/bin/bash
#$ -cwd
#$ -l mem=16G,h_rt=16:00:00
#$ -pe single 1
#$ -j y
/homes/liu3zhen/local/bin/fastq-dump \
      --split-spot --split-3 \
      --defline-seq '@$sn/$ri' \
      --defline-qual '+' \
      --gzip -A <accession>
```



Beocat pipeline to download SRA in batch - I

step 1: Prepare data: dataset.txt

Order	ID	Genotype	Sample	Treatment	Citation
1	SRR1238718	B73	norm1	norm	Hirsch2015TPJ
2	SRR1819617	B73	norm2	norm	Hirsch2015TPJ
3	SRR1819621	B73	norm3	norm	Hirsch2015TPJ
4	SRR1238717	B73	cold1	cold	Hirsch2015TPJ
5	SRR1819204	B73	cold2	cold	Hirsch2015TPJ
6	SRR1819205	B73	cold3	cold	Hirsch2015TPJ

Makarevitch *et al*., 2015
Hirsch *et al*., 2016

General procedure

a QSUB script

submit a job

a script to generate and submit QSUB scripts

run the script to generate QSUB scripts

Organize the running script in a shell script

```
perl srr.qsub.pl \
      --mem 16 \
      --time 16:00:00 \
      --list dataset.txt \
      --srrcol 2 \
      --path /homes/liu3zhen/local/bin/
```

dataset.txt

Order	ID	Genotype	Sample	Treatment	Citation
1	SRR1238718	B73	norm1	norm	Hirsch2015TPJ
2	SRR1819617	B73	norm2	norm	Hirsch2015TPJ
3	SRR1819621	B73	norm3	norm	Hirsch2015TPJ
4	SRR1238717	B73	cold1	cold	Hirsch2015TPJ
5	SRR1819204	B73	cold2	cold	Hirsch2015TPJ
6	SRR1819205	B73	cold3	cold	Hirsch2015TPJ

* path to the command: fastq-dump

Beocat pipeline to download SRA in batch - II

step 2: Run "bash 1c-download.sh"

must check
meta_file=dataset.txt
srr_col=2
rename_col=4

might need to change
max_mem_size=16 ### requested memory
max_time=16:00:00 ### requested running time
fdpath=/homes/liu3zhen/local/bin/ ### fastq-dump path
srr_script_path=/homes/liu3zhen/local/pipelines/SRA/
rename_script=2c-rename.sh

running
perl $srr_script_path/**srr.qsub.pl** --mem $max_mem_size \
    --time $max_time --list $meta_file --srrcol $srr_col --path $fdpath
create a script for renaming downloaded files
cut $meta_file -f $srr_col,$rename_col | grep "^[EDS]RR" | sed 's/^/rename /g' | sed 's/\t/ /g'| sed 's/$/ *gz/g' > $rename_script

Note: 1c-download.sh is copied from **/homes/liu3zhen/local/pipelines/SRA**

Beocat pipeline to download SRA in batch - III

step 3: Run "2c-rename.sh" to change names (optional)

```
bash 2c-rename.sh
```

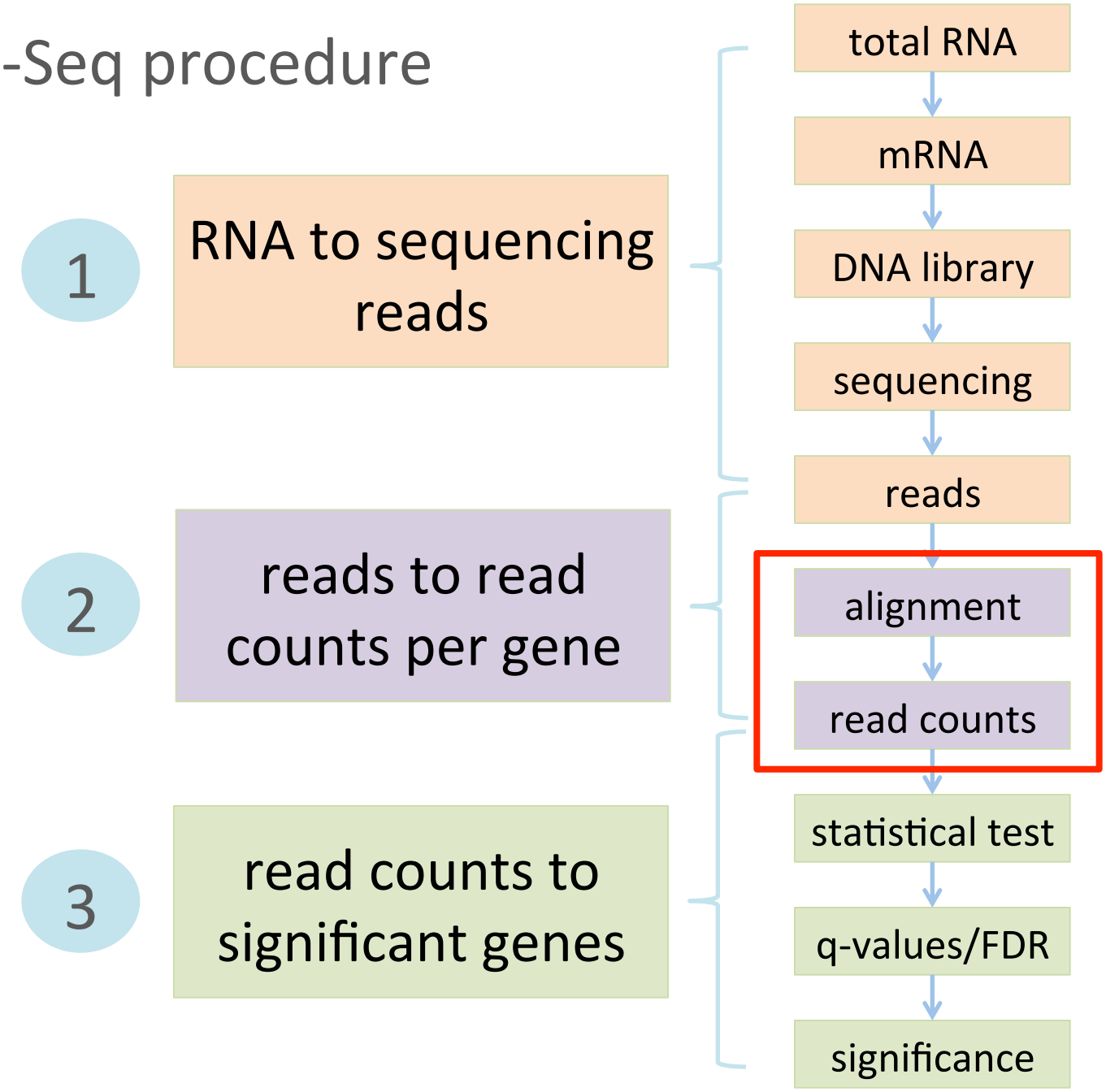
Part II. Trimming: generating qsub script

```
perl /homes/liu3zhen/local/pipelines/trimmomatic/trimmomatic.qsub.pl \
--mem 16 \
--time 12:00:00 \
--trim_shell "/homes/liu3zhen/local/pipelines/trimmomatic/trimmomatic.pe.sh" \
--trimmomatic "/homes/liu3zhen/local/jars/trimmomatic-0.36.jar" \
--adaptor_file "/homes/liu3zhen/local/pipelines/trimmomatic/trimmomatic_adaptDB/TruSeq3-PE.fa" \
--indir "../1-raw" \
--outdir "." \
--fq1feature "_1.fastq.gz" \
--fq2feature "_2.fastq.gz" \
--threads 4 \
--min_len 40
```

Part II. Trimming: qsub command

```
#!/bin/bash
#$ -cwd
#$ -l mem=16G,h_rt=12:00:00
#$ -pe single 4
#$ -j y
bash /homes/liu3zhen/local/pipelines/trimmomatic/trimmomatic.pe.sh \
/homes/liu3zhen/local/jars/trimmomatic-0.36.jar \
/homes/liu3zhen/local/pipelines/trimmomatic/trimmomatic_adaptDB/TruSeq3-PE.fa \
../1-raw \
. \
_1.fastq.gz \
_2.fastq.gz \
4 \
40 \
cold1_1.fastq.gz
```

RNA-Seq procedure



Part IV. STAR: qsub script (one sample)

```
#!/bin/bash
#$ -cwd
#$ -l mem=48G,h_rt=12:00:00
#$ -pe single 1
#$ -j y
/homes/liu3zhen/local/bin/STAR --runThreadN 1 \
--genomeDir ../0-ref \
--readFilesIn ../2-trim/cold1.R1.pair.fq ../2-trim/cold1.R2.pair.fq \
--alignIntronMax 100000 \
--alignMatesGapMax 100000 \
--outFileNamePrefix cold1 \
--outSAMattrIHstart 0 \
--outSAMmultNmax 1 \
--outSAMstrandField intronMotif \
--outFilterIntronMotifs RemoveNoncanonicalUnannotated \
--outSAMtype BAM SortedByCoordinate \
--quantMode GeneCounts \
--outFilterMismatchNmax 2 \
--outFilterMismatchNoverLmax 0.05 \
--outFilterMatchNmin 40 \
--outSJfilterReads Unique \
--outFilterMultimapNmax 1 \
--outSAMmapqUnique 60 \
--outFilterMultimapScoreRange 2
```

Part III. STAR: Download and index the reference genome

```
wget ftp://ftp.ensemblgenomes.org/pub/release-35/plants/fasta/zea_mays/dna/Zea_mays.AGPv4.dna.toplevel.fa.gz
wget ftp://ftp.ensemblgenomes.org/pub/release-35/plants/gtf/zea_mays/Zea_mays.AGPv4.35.gtf.gz
gunzip *gz
qsub STAR.index.qsub
```

```
#!/bin/bash
#$ -cwd
#$ -l mem=3G,h_rt=12:00:00
#$ -pe single 1
#$ -j y #
/homes/liu3zhen/local/bin/STAR --runThreadN 4 \
        --runMode genomeGenerate \
        --genomeDir . \
        --genomeFastaFiles Zea_mays.AGPv4.dna.toplevel.fa \
        --sjdbGTFfile Zea_mays.AGPv4.35.gtf \
        --sjdbOverhang 100
```

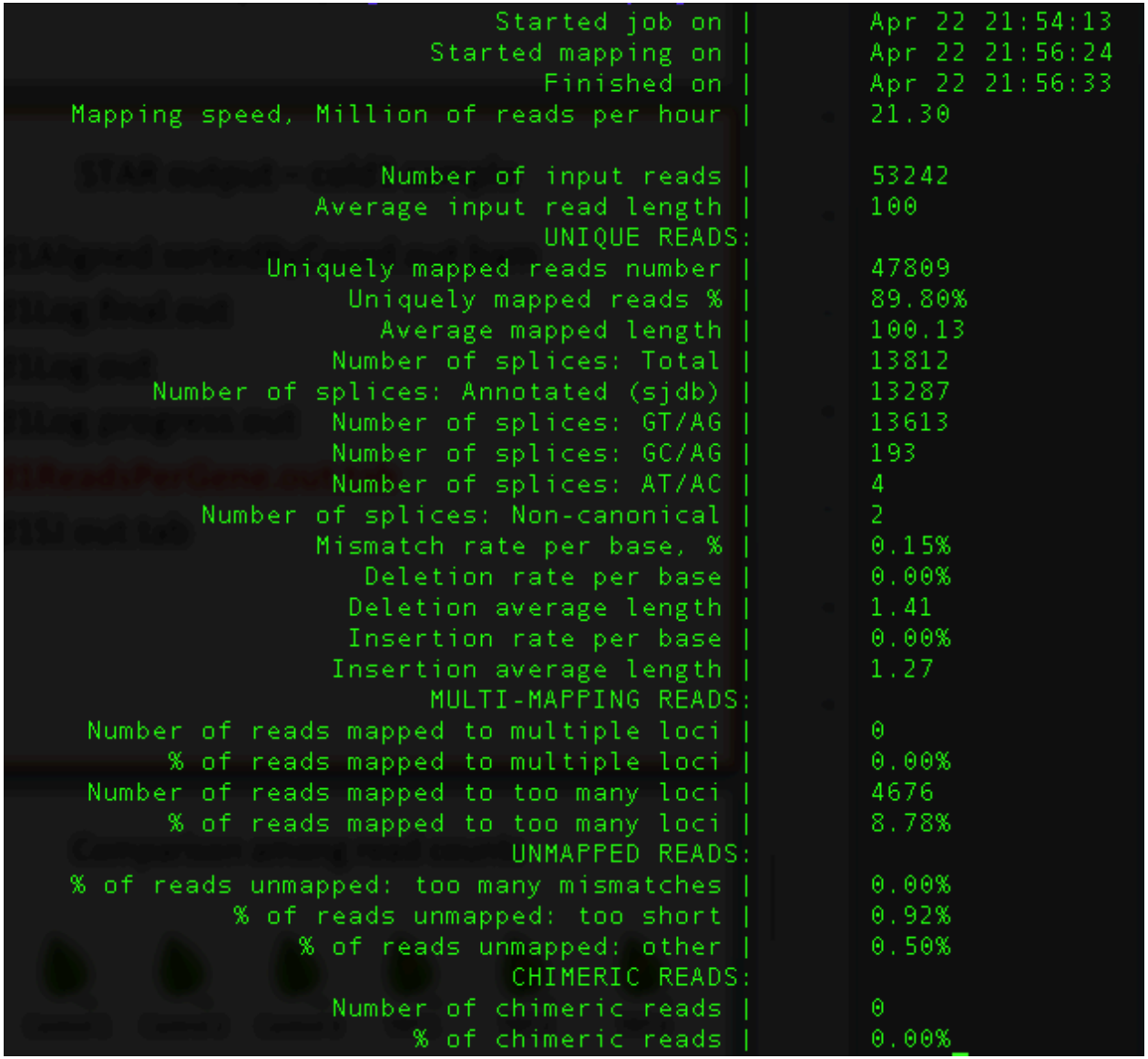
Part IV. STAR: generate qsub script and submit jobs

```
dbdir=../0-ref
perl /homes/liu3zhen/local/pipelines/STAR/STAR.qsub.pl \
--mem 48 --threads 1 --time 12:00:00 \
--star_cmd /homes/liu3zhen/local/bin/STAR \
--indir ../2-trim \
--dbdir $dbdir \
--fq1feature .R1.pair.fq \
--fq2feature .R2.pair.fq \
--alignIntronMax 100000 \
--alignMatesGapMax 100000 \
--outSAMattrIHstart 0 \
--outSAMmultNmax 1 \
--outSAMstrandField intronMotif  \
--outFilterIntronMotifs RemoveNoncanonicalUnannotated \
--outSAMtype "BAM SortedByCoordinate" \
--quantMode GeneCounts \
--outFilterMismatchNmax 2 \
--outFilterMismatchNoverLmax 0.05 \
--outFilterMatchNmin 40 \
--outSJfilterReads Unique \
--outFilterMultimapNmax 1 \
--outSAMmapqUnique 60 \
--outFilterMultimapScoreRange 2
```

STAR output – cold1 sample

- cold1Aligned.sortedByCoord.out.bam
- cold1Log.final.out
- cold1Log.out
- cold1Log.progress.out
- **cold1ReadsPerGene.out.tab**
- cold1SJ.out.tab

cold1Log.final.out



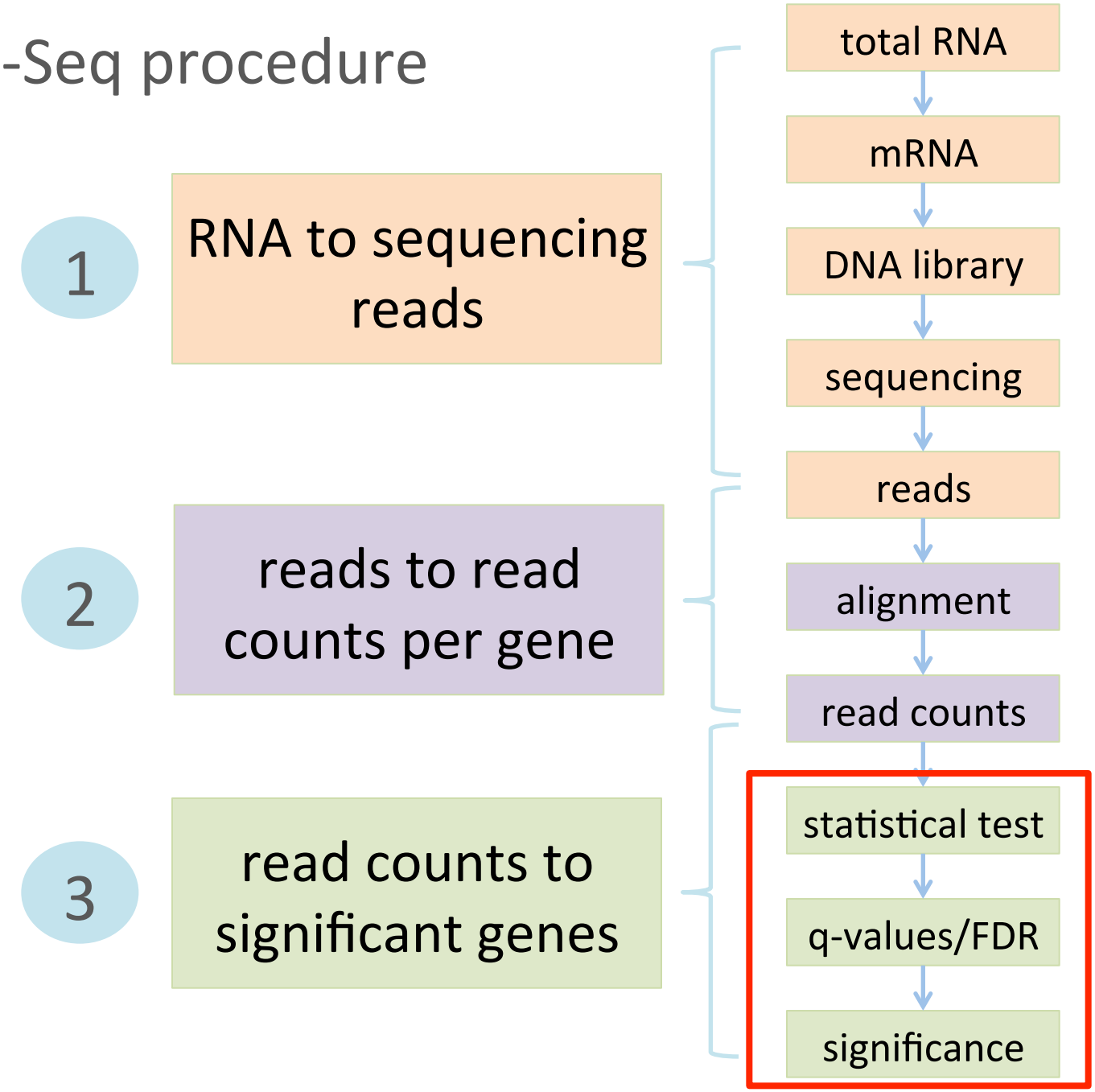
Started job on	Apr 22 21:54:13
Started mapping on	Apr 22 21:56:24
Finished on	Apr 22 21:56:33
Mapping speed, Million of reads per hour	21.30
Number of input reads	53242
Average input read length	100
UNIQUE READS:	
Uniquely mapped reads number	47809
Uniquely mapped reads %	89.80%
Average mapped length	100.13
Number of splices: Total	13812
Number of splices: Annotated (sjdb)	13287
Number of splices: GT/AG	13613
Number of splices: GC/AG	193
Number of splices: AT/AC	4
Number of splices: Non-canonical	2
Mismatch rate per base, %	0.15%
Deletion rate per base	0.00%
Deletion average length	1.41
Insertion rate per base	0.00%
Insertion average length	1.27
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	0
% of reads mapped to multiple loci	0.00%
Number of reads mapped to too many loci	4676
% of reads mapped to too many loci	8.78%
UNMAPPED READS:	
% of reads unmapped: too many mismatches	0.00%
% of reads unmapped: too short	0.92%
% of reads unmapped: other	0.50%
CHIMERIC READS:	
Number of chimeric reads	0
% of chimeric reads	0.00%

cold1**ReadsPerGene.out.tab**

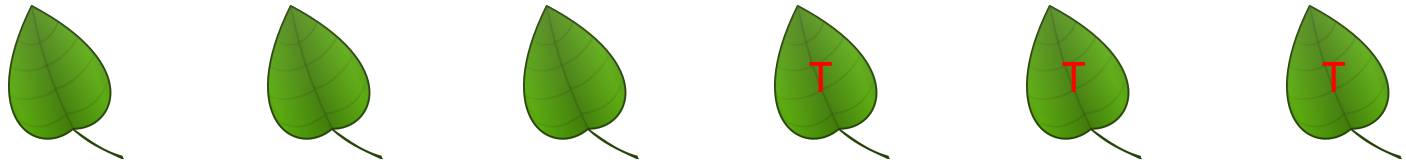
N_unmapped	5433	5433	5433
N_multimapping	0	0	0
N_noFeature	2927	25138	25275
N_ambiguous	566	125	119
Zm00001d027230	3	1	2
Zm00001d027231	2	1	1
Zm00001d027232	0	0	0
Zm00001d027233	0	0	0
Zm00001d027234	0	0	0
Zm00001d027235	0	0	0

column 1: gene ID
column 2: counts for unstranded RNA-seq
column 3: counts for the 1st read strand aligned with RNA (htseq-count option -s yes)
column 4: counts for the 2nd read strand aligned with RNA (htseq-count option -s reverse)

RNA-Seq procedure



Comparison among read counts



	Control 1	Control 2	Control 3	TRT 1	TRT 2	TRT 3
Gene 1	C1	C2	C3	T1	T2	T3
…						
Sum	N1	N2	N3	N4	N5	N6

Sequence depth (total read number) influences read counts.

Can we generate some comparable numbers among samples?

Statistical test for differential expression

- Statistical test to discover differential expression (DE)
  - **Count data**: Generalized Linear Model (GLM) to deal with count data

    e.g., Poisson GLM could handle count data but overdispersion exits

  - **Dispersion issue**: Using negative binomial GLM to incorporate dispersion into the model

  - **Small n problem**: a few number of replication

    Borrowing information across all the genes to estimate gene-specific variation

edgeR (Robinson and Smyth, 2007), DESeq (Anders and Huber, 2010), NBPSeq (Di et al., 2011), and QuasiSeq (Lund 2012)

Part V. DE: merge counting data

```
setwd("/homes/liu3zhen/teaching/BA17/in-class.project/DE/4-DE/")
library("DESeq2")

### Parameters - Subject to change
datapath <- "/homes/liu3zhen/teaching/BA17/in-class.project/DE/3-STAR/"
suffix <- "ReadsPerGene.out.tab"
count.files <- dir(path = datapath, pattern = suffix)

### merge all counts
allcounts <- NULL
for (cf in count.files) {
  counts <- read.delim(paste0(datapath, "/", cf), header = F, stringsAsFactors = F, skip = 4)
  base <- gsub(suffix, "", cf)
  counts <- counts[, 1:2]
  colnames(counts) <- c("Gene", base)

  ### merge data
  if (is.null(allcounts)) {
    allcounts <- counts
  } else {
    allcounts <- merge(allcounts, counts, by = "Gene")
  }
}
```

Part VI. DE

```
### load modules
source("/homes/liu3zhen/local/share/LiuLabScripts/DESeq2.single.trt.R")
source("/homes/liu3zhen/local/share/LiuLabScripts/DE.summary.R")
### DE parameters
fdr.cutoff <- 0.05
# data reformat:
input <- allcounts[, 2:7]
rownames(input) <- allcounts[, 1]
# DE statistical analysis:
DE.out <- DESeq2.single.trt(input.matrix = input,
                            min.mean.reads = 5,
                            group1.col = 1:3,
                            group2.col = 4:6,
                            comparison = c("norm", "cold"),
                            geneID = rownames(input),
                            fdr = fdr.cutoff,
                            logpath = ".",
                            logfile = "cold-norm.log.md")

# merge DE with counts and output DE result:
DE.out <- data.frame(DE.out)
final.out <- merge(allcounts, DE.out, by.x = "Gene", by.y = "GeneID")
write.table(final.out, "cold-norm.DESeq2.txt", sep="\t", quote=F,
row.names=F )
```

Part VI. DE summary

```
de.summary <- DE.summary(DE.path=".",
    DE.files="cold-norm.DESeq2.txt",
    qval.feature=".qval",
    log2FC.feature=".log2FC",
    fdr=fdr.cutoff,
    out.path=".",
    out.file="cold-norm.DESeq2.summary.txt")
```

your turn

Order	Runs	Samples	Tissue	Replicate
1	SRR3466605	RH1	root hairs	rep1
2	SRR3466606	RH2	root hairs	rep2
3	SRR3466607	RH3	root hairs	rep3
4	SRR3466608	RH4	root hairs	rep4
5	SRR3466609	root1	root hair less roots	rep1
6	SRR3466610	root2	root hair less roots	rep2
7	SRR3466611	root3	root hair less roots	rep3
8	SRR3466612	root4	root hair less roots	rep4

root hair (RH) vs. root without hair (root)