Final Project Report

Jacob Halle

Introduction

Differential gene expression analysis is a powerful tool that allows researchers to study the transcriptomic profile of cells and tissue. By sequencing mRNA and quantifying the number of unique transcripts, researchers are able to compare gene expression levels between different experimental conditions and cell/tissue types. This technique is applied here to study the effect of the removal of the human NRDE2 gene on the transcriptome. The human breast cancer cell line MDA-MB-231 was transfected in triplicate with either NRDE2-targeting siRNA or a control then allowed to incubate for 48 hrs. This transfection will effectively remove NRDE2 from the transcriptome, allowing the researchers to study the transcriptome when the NRDE2 gene is not active. After the incubation, the cells were harvested, and RNA-sequencing was performed. The below analysis was performed on the raw data generated from this experiment.

Methods

The raw data was processed using the nf-core/rna-seq pipeline version 3.14.0. The executed workflow took fastq files as inputs and performed extensive quality control measures, including adaptor trimming using the "Trim Galore!" tool, duplicated read removal, and trimming overrepresented bases at the beginning of reads. These samples had a fairly high duplication rate (~65%). This can likely be attributed to PCR amplification bias and should not affect downstream analysis because duplicates are removed from the final counts. There was also CG bias detected at the beginning of all the reads, which is likely a result of the library preparation methods and common in RNA-seq libraries. The beginnings of the reads with CG bias were trimmed. After the trimming of these low quality or adaptor sections of the reads, all reads passed quality metrics and were considered in downstream analysis.

The pseudo-aligner 'salmon' was used to quantify the reads as opposed to a full alignment onto the transcriptome through tools such as 'star'. Salmon works by aligning the reads to known transcripts, but not to the exact location in the genome. This approach is more computationally efficient, and since the goal of the project is DGE analysis, it is an appropriate choice. Salmon was run with the "—gcbias" option active.

Statistical analysis was performed in Rstudio 4.3.3. Read count data was imported using the "tximport" package and a DESeqDataSet object was created using the design "~ condition". This formula is appropriate because there is only one variable in this experiment; whether the sample was a 'control' or 'treated' sample. Next, very low expression genes were removed, since it is difficult to be statistically confidant in samples that have counts near zero. Thus, any gene that had less than 8 counts were removed. Next, the log-fold changes and associated p-values for each gene were calculated using the 'results' function. To correct for multiple hypothesis testing, an FDR threshold of 0.05 was chosen to reduce the probability of type I errors. Next, the log fold change (LFC) values were shrunk to stabilize the variance across genes with varying expression

levels. This was done by using the 'lfcshrink' function and the 'apeglm' option. This option is suitable for making the LFCs between high expression and low expression more comparable.

Results and Discussion

Figure 1. Sample Aligning Statistics

Sample Name	% Aligned	M Aligned	M Seqs
control1	90.9%	55.7	61.4
control2	92.1%	58.7	63.9
control3	93.1%	52.1	56.1
treated1	92.4%	53.3	57.8
treated2	92.7%	54.6	58.9
treated3	92.7%	43.0	46.4

This table describes the total number of reads present in each sample, as well as the number and percentage that aligned to known transcripts. The reads that did not align to transcripts, could be the result of using a pseudo-aligner instead of a full alignment tool. These tools sacrifice possible alignments for computational speed.

Figure 2. Top ten most significantly different genes.

feature_id <chr>	baseMean <dbl>	log2FoldChange <dbl>	lfcSE <dbl>	pvalue <dbl>	padj <dbl>
ENSG00000175334	6423.02696	1.6558541	0.06586739	1.026772e-140	1.627433e-136
ENSG00000163041	7971.70070	1.6643853	0.07193897	1.124175e-119	8.909084e-116
ENSG00000196396	6618.42328	1.1493915	0.05311471	6.075417e-105	3.209845e-101
ENSG00000105976	9565.62118	1.5682550	0.07597062	6.698841e-96	2.654416e-92
ENSG00000128595	22966.87111	1.4809691	0.07212181	5.506210e-95	1.745468e-91
ENSG00000101384	11784.31488	1.3116920	0.06635210	3.716152e-88	8.414431e-85
ENSG00000124333	2741.54405	1.4826359	0.07502744	3.658551e-88	8.414431e-85
ENSG00000117632	16768.29952	1.3400885	0.07060932	1.456710e-81	2.886106e-78
ENSG00000180398	20587.62952	0.9084392	0.05191299	1.056673e-69	1.860918e-66
ENSG00000213281	6961.64814	1.1806417	0.06782013	4.418120e-69	7.002720e-66

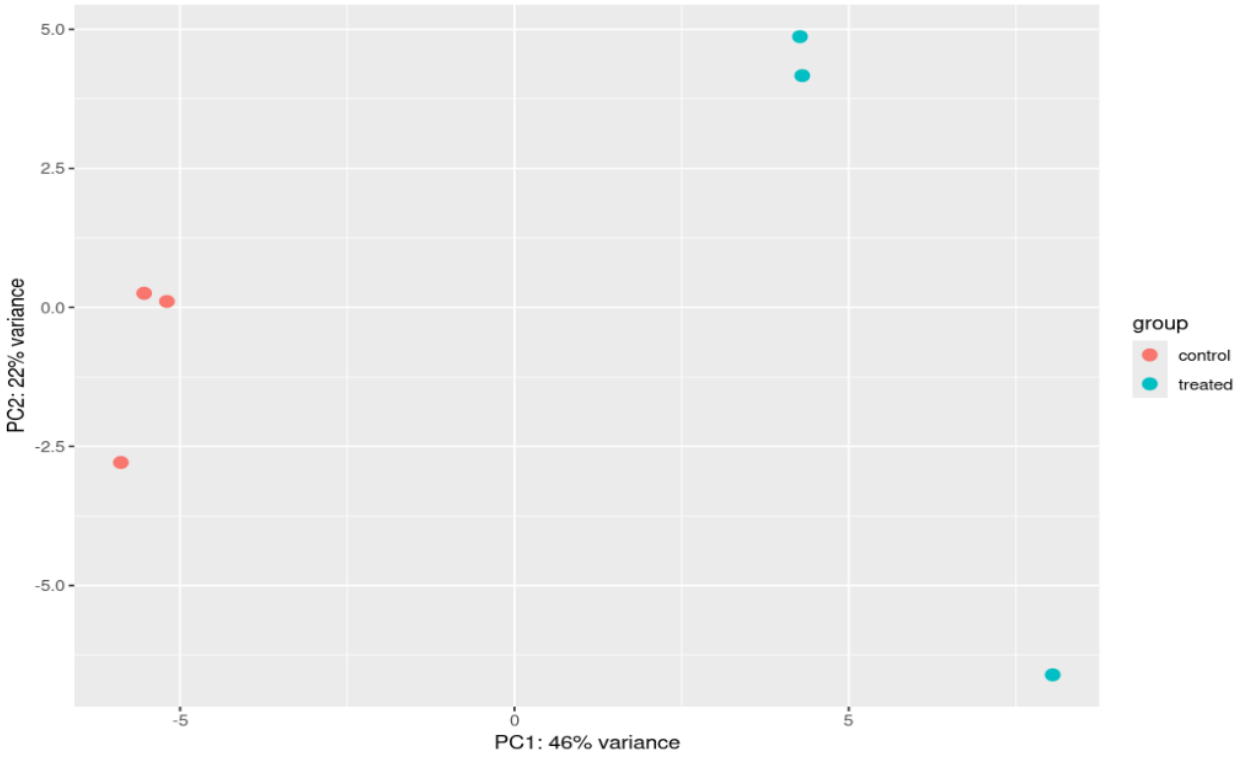
This table summarizes the most statistically significantly different genes after an FDR of 0.05 adjustment. This can be seen by the increasing adjusted p-values. These genes have been most affected by the removal of NRDE2 gene from the transcriptome.

Figure 3. Statistically relevant differentially expressed genes

	FDR < 0.05	LFC < 0	LFC > 0
NUMBER OF GENES	3596	1666	1930

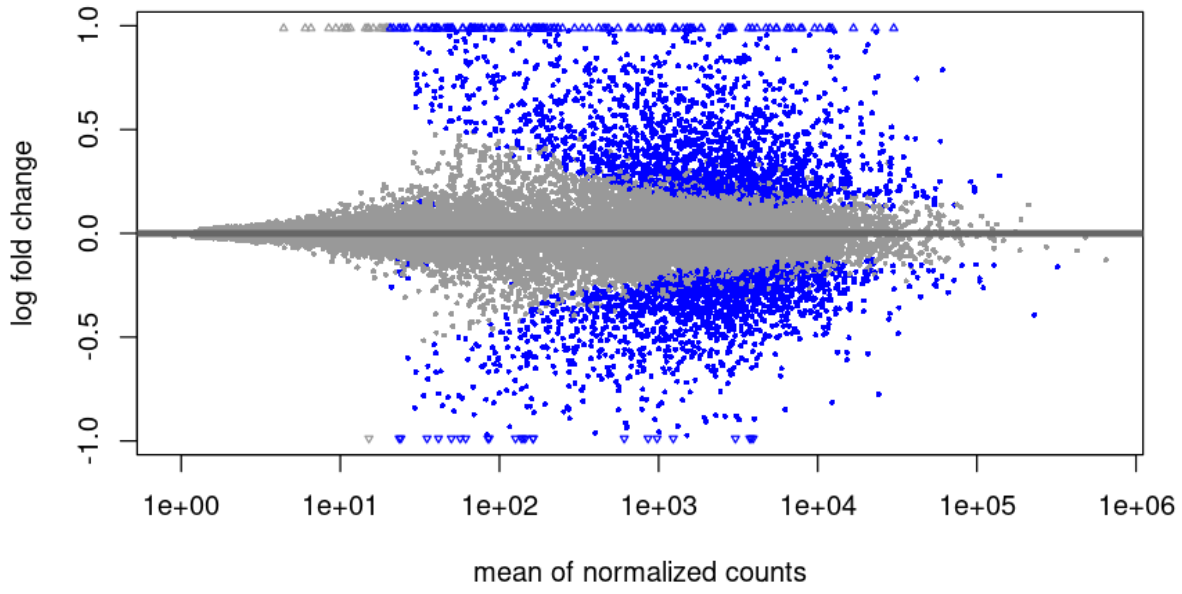
Of the 23669 genes that were significantly expressed in the dataset, 3596 were significantly affected by the absence of NRDE2. Of these genes, slightly more were upregulated than downregulated. These results indicate that without NRDE2, a widespread biological response was triggered to accommodate for the lack of function for this gene.

Figure 4. PCA plot



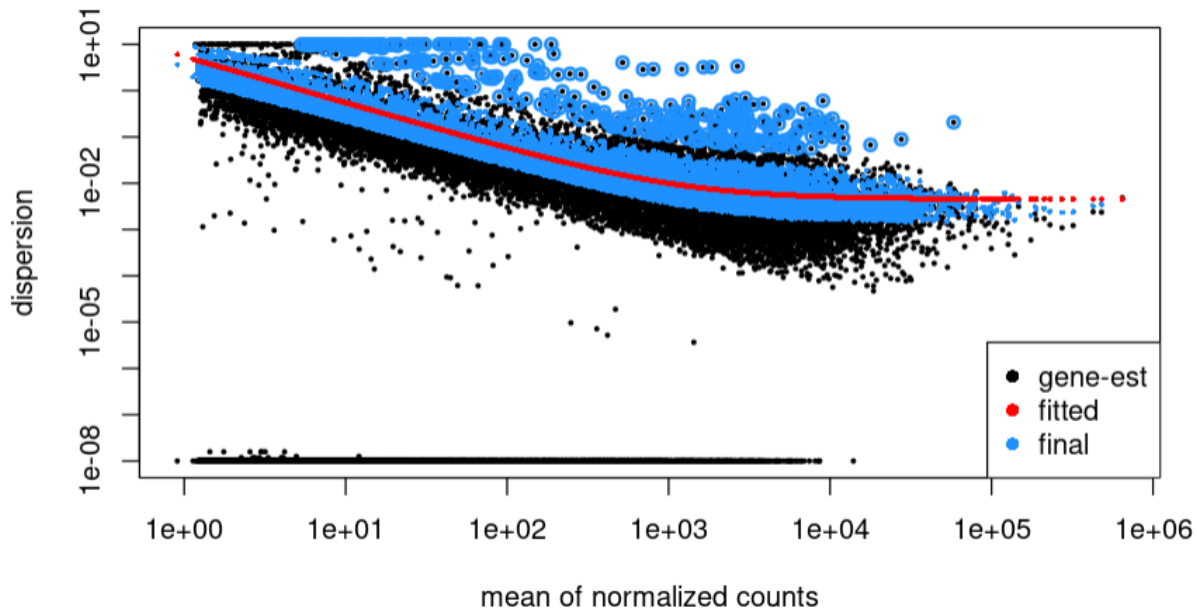
The PCA plot indicates that a considerable amount of variance (46%) can be described by one principal component alone. This PC is likely capturing the differences introduced by the removal of NRDE2, because the control and treated samples are separated greatly on the x-axis.

Figure 5. MA plot



The MA plot describes the relationship between average expression and log fold change. From this graph, it can be determined that most genes have expression levels ranging from 10 to 10,000, and most significantly different genes have expression levels from 100 to 10,000. There is also good symmetry about the x-axis, implying that there is no bias in this library.

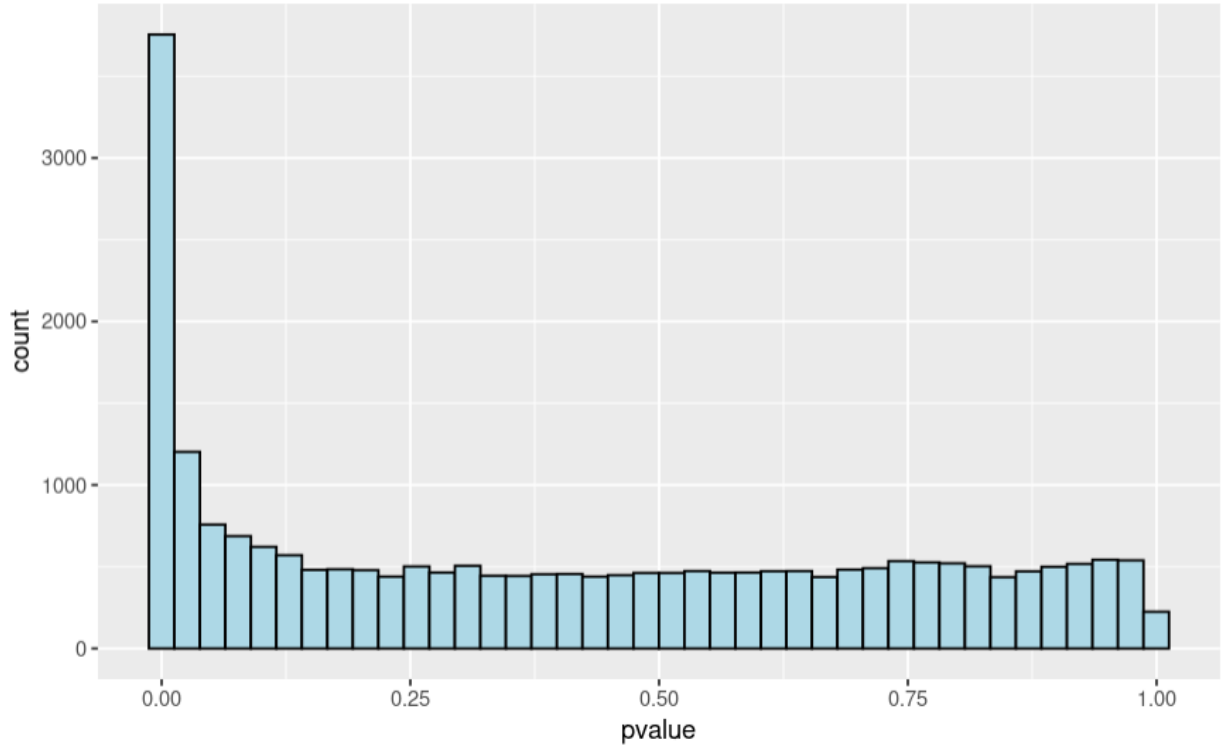
Figure 6. Dispersion by mean plot



The dispersion decreases with increasing mean expression levels. This pattern and shape of the plot

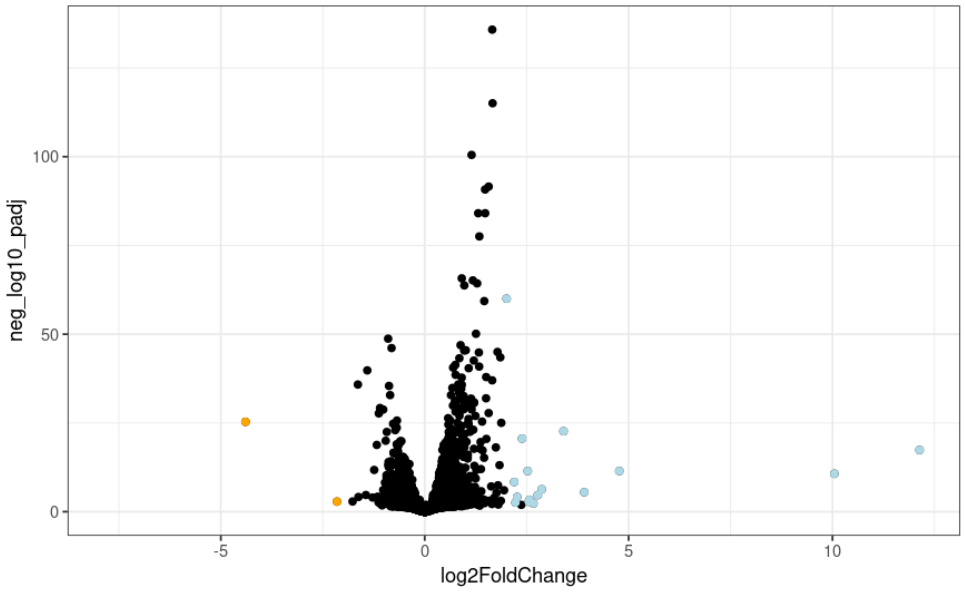
indicates that the data is suitable for DESeq2. It is expected that dispersion will decrease as expression goes up.

Figure 7. p-value histogram



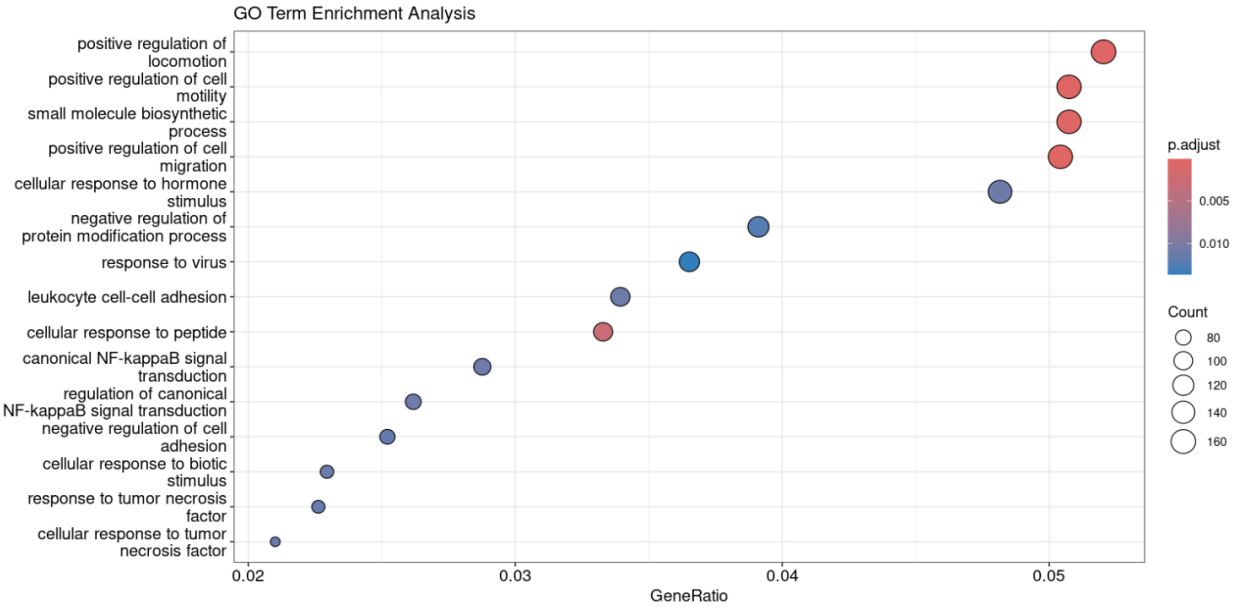
This histogram indicates that there is an overepresentation of low p-values. This is to be expected in an experiment with treated samples. It is expected that the groups will have some genes that are significantly different, thus the overepresentation of low p-values.

Figure 8. Volcano plot



This plot visualizes log-fold changes and p-values. There are not many genes that have LFCs over 2 or under -2, but there are a considerable number of genes with smaller LFCs and low adjusted p-values. A reason for this could be a large-scale system of adaptation that involves slight up or down regulation across many different genes in response to the lack of NRDE2.

Figure 9. GO term enrichment analysis



Some of the most enriched GO terms relate to cell mobility. This indicates that one of main functions of NRDE2 is to regulate cell movement. Without this crucial biological process, it is unlikely that the treated cells will be able to properly grow and proliferate.

The library used in this experiment was of high quality. After the reads had been trimmed, all were of high enough quality to be used in this analysis. However, due to the use of a pseudo-aligner, not all reads were mapped to a known transcript. The relationship between the log fold change and mean expression indicates that there is little bias influencing the LFCs. Comparing the dispersion to the mean expression reveals that this dataset is an appropriate fit for DESeq2 analysis. The removal of the NRDE2 gene had wide reaching yet modest effects on the transcriptome. There were many genes that had a significantly low adjusted p-value, yet only a few genes had a log-fold change above 2. Of the effected genes, the most enriched go terms related to cell movement.

Appendix

Software Versions:

multiqc v1.19
Rstudio 4.3.3

Process Name	Software	Version
CUSTOM_DUMPSOFTWAREVERSIONS	python	3.11.7
	yaml	5.4.1
CUSTOM_GETCHROMSIZES	getchromsizes	1.16.1
DESEQ2_QC_PSEUDO	bioconductor-deseq2	1.28.0
	r-base	4.0.3
FASTQC	fastqc	0.12.1
FQ_SUBSAMPLE	fq	0.9.1 (2022-02-22)
GTF2BED	perl	5.26.2
GTF_FILTER	python	3.9.5
GUNZIP_FASTA	gunzip	1.10
GUNZIP_GTF	gunzip	1.10
MAKE_TRANSCRIPTS_FASTA	rsem	1.3.1

	star	2.7.10a
SALMON_INDEX	salmon	1.10.1
SALMON_QUANT	salmon	1.10.1
SE_GENE	bioconductor-summarizedexperiment	1.24.0
	r-base	4.1.1
TRIMGALORE	cutadapt	3.4
	trimgalore	0.6.7
TX2GENE	python	3.9.5
TXIMPORT	bioconductor-tximeta	1.12.0
	r-base	4.1.1
Workflow	Nextflow	23.04.1
	nf-core/rnaseq	3.14.0

Nf-core/rna-seq script

```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=8
#SBATCH --time=12:00:00
#SBATCH --mem=4GB
#SBATCH --job-name=Align
#SBATCH --mail-type=FAIL
#SBATCH --mail-user=jh8755@nyu.edu

module purge
module load nextflow/23.04.1

nextflow run nf-core/rnaseq -r 3.14.0 \
--input /scratch/jh8755/ngs.final/samplesheet.csv \
--outdir res \
--fasta /scratch/jh8755/ngs.final/Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa.gz \
--gtf /scratch/jh8755/ngs.final/Homo_sapiens.GRCh38.111.gtf.gz \
--extra_salmon_quant_args "--gcBias " \
--skip_trimming false \
--skip_alignment true \
--pseudo_aligner salmon \
--save_reference true \
-profile nyu_hpc \
```

Fastq file download script

```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=24:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=download_fastqs
#SBATCH --mail-type=FAIL
#SBATCH --mail-user=jmf11@nyu.edu

module purge

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR781/000/SRR7819990/SRR7819990.fastq.gz
echo _ESTATUS_ [ wget SRR7819990 ]: $?
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR781/001/SRR7819991/SRR7819991.fastq.gz
echo _ESTATUS_ [ wget SRR7819991 ]: $?
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR781/002/SRR7819992/SRR7819992.fastq.gz
echo _ESTATUS_ [ wget SRR7819992 ]: $?
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR781/003/SRR7819993/SRR7819993.fastq.gz
echo _ESTATUS_ [ wget SRR7819993 ]: $?
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR781/004/SRR7819994/SRR7819994.fastq.gz
echo _ESTATUS_ [ wget SRR7819994 ]: $?
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR781/005/SRR7819995/SRR7819995.fastq.gz
echo _ESTATUS_ [ wget SRR7819995 ]: $?
echo _END_ [ download.slurm ]: $(date)
```