**Bioinformatics Final Project Workflow**

Acquire dataset from this publication: Vydra, Natalia et al. “Heat shock factor 1 (HSF1) cooperates with estrogen receptor α (ERα) in the regulation of estrogen action in breast cancer cells.” eLife vol. 10 e69843. 16 Nov. 2021, doi:10.7554/eLife.69843

This data was found in NIH GEO database: GSE186004

**Make a final\_project directory:**

mkdir final\_project

cd final\_project

**First, we will download the required datasets into final\_project directory.**

Script to download dataset (control): geo\_dataset\_download.sh

#!/bin/bash

#SBATCH -t 0-10:00

#SBATCH --cpus-per-task=5

#SBATCH --mem=10Gb

#SBATCH --output=aquire\_data.txt

/project/stuckert/aspeidl/final\_project/sratoolkit.3.1.1-ubuntu64/bin/prefetch SRR16356004

/project/stuckert/aspeidl/final\_project/sratoolkit.3.1.1-ubuntu64/bin/fasterq-dump SRR16356004

Script to download dataset (estrogen): geo\_dataset\_download\_estrogen.sh

#!/bin/bash

#SBATCH -t 0-10:00

#SBATCH --cpus-per-task=5

#SBATCH --mem=10Gb

#SBATCH --output=acquire\_data2.txt

/project/stuckert/aspeidl/final\_project/sratoolkit.3.1.1-ubuntu64/bin/prefetch SRR16356007

/project/stuckert/aspeidl/final\_project/sratoolkit.3.1.1-ubuntu64/bin/fasterq-dump SRR16356007

**Next step will be to run FASTQC on the files to ensure quality control.**

Downloading FASTQC:

wget <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.12.1.zip>

unzip fastqc\_v0.12.1.zip

Script to run FASTQC: fastqc\_dataset\_6004.sh

#!/bin/bash

#SBATCH -t 0-10:00

#SBATCH --cpus-per-task=5

#SBATCH --mem=10Gb

#SBATCH --output=fastqc\_data.txt

/project/stuckert/aspeidl/final\_project/FastQC/fastqc /project/stuckert/aspeidl/final\_project/SRR16356004\_1.fastq

/project/stuckert/aspeidl/final\_project/FastQC/fastqc /project/stuckert/aspeidl/final\_project/SRR16356004\_2.fastq

Script to run FASTQC: fastqc\_dataset\_6007.sh

#!/bin/bash

#SBATCH -t 0-10:00

#SBATCH --cpus-per-task=5

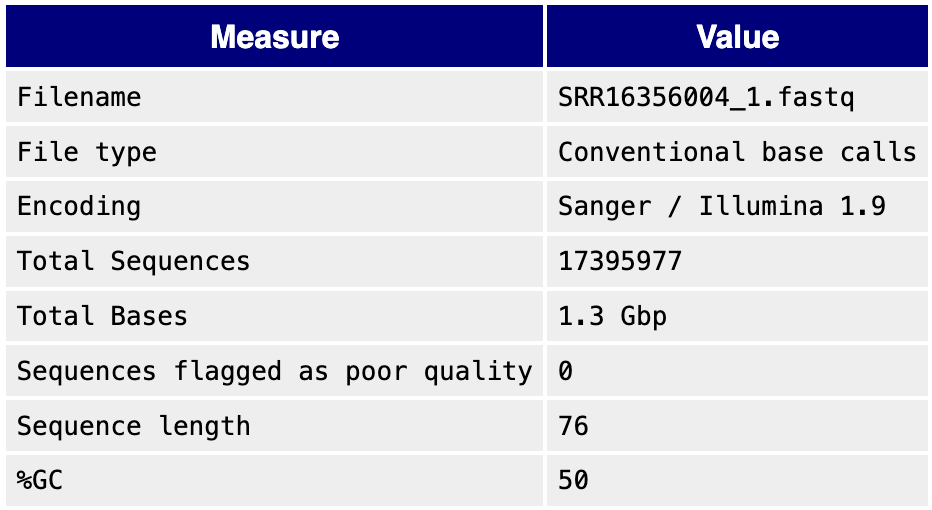
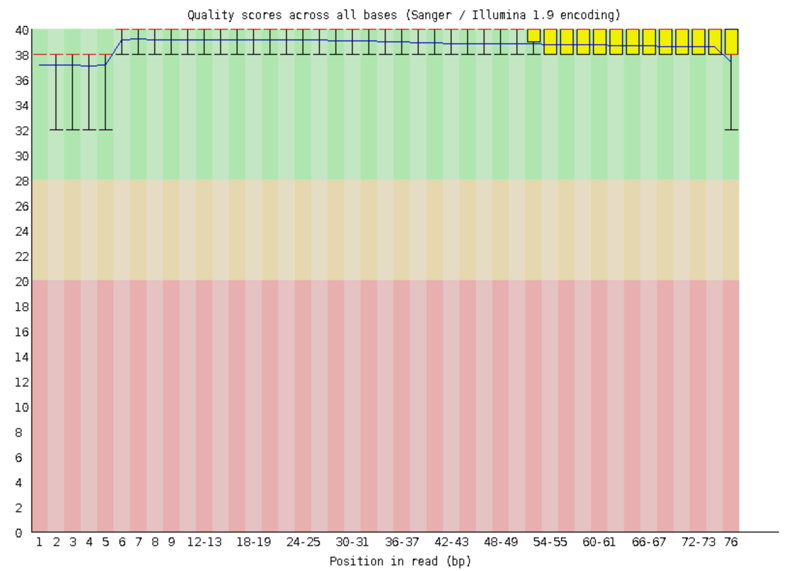
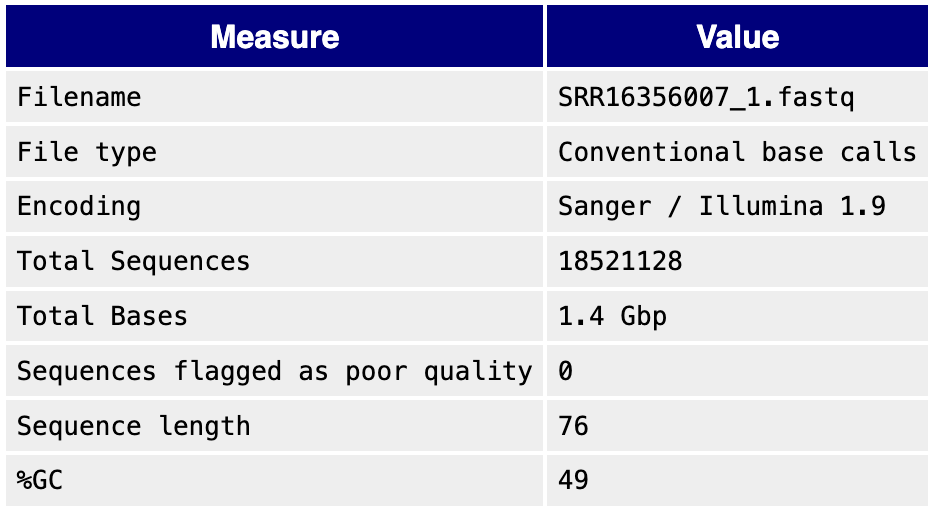
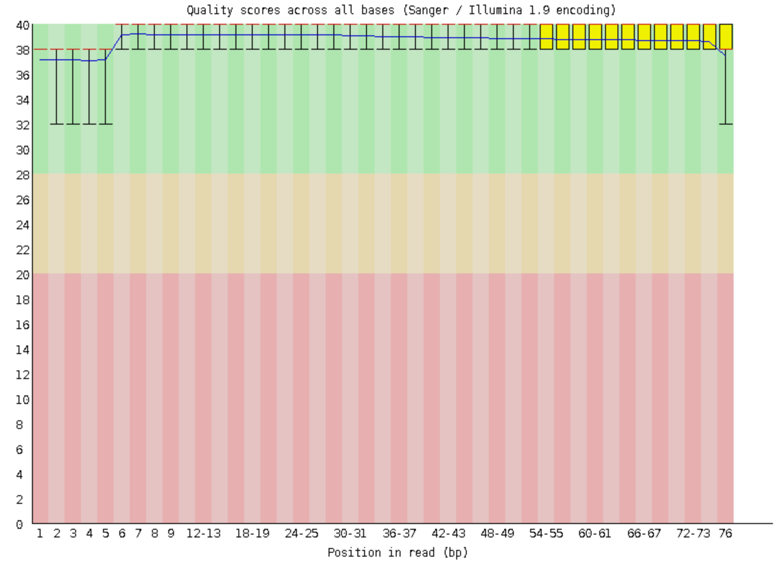
#SBATCH --mem=10Gb

#SBATCH --output=fastqc\_data.txt

/project/stuckert/aspeidl/final\_project/FastQC/fastqc /project/stuckert/aspeidl/final\_project/SRR16356007\_1.fastq

/project/stuckert/aspeidl/final\_project/FastQC/fastqc /project/stuckert/aspeidl/final\_project/SRR16356007\_2.fastq

**Results from FASTQ:**



Left represents cells with no exogenous estrogen treatment, right represents the estrogen-treated cells. FASTQC results look good, let’s move on to transcriptome assembly.

**To assemble the transcriptome, combined the forward reads and the reverse reads into one file each, respectively.**

**Control:** cat SRR16356004\_1.fastq SRR16356007\_1.fastq > combined\_01.fastq

**Estrogen:** cat SRR16356004\_2.fastq SRR16356007\_2.fastq > combined\_02.fastq

**The next step will be to assemble the transcriptome. We are going to use SPAdes, which was previously installed during class.**

Script to run SPAdes assembly: spades\_assembly.sh

#!/bin/bash

#SBATCH -t 0-20:00

#SBATCH --cpus-per-task=20

#SBATCH --mem=60Gb

/project/stuckert/aspeidl/SPAdes-4.0.0-Linux/bin/spades.py --rna --pe1-1 /project/stuckert/aspeidl/final\_project/combined\_01.fq --pe1-2 /project/stuckert/aspeidl/final\_project/combined\_02.fq --threads 16 -o /project/stuckert/aspeidl/final\_project/spades\_output

**The output file for the assembly is transcripts.fasta. Let’s analyze the results using assemblathon.**

**Code:**

/project/stuckert/software/assemblathon\_stats.pl transcripts.fasta

**Results from the assembly:**

Number of scaffolds 141700

Total size of scaffolds 102111744

Longest scaffold 20241

Shortest scaffold 49

Number of scaffolds > 1K nt 25747 18.2%

Number of scaffolds > 10K nt 152 0.1%

Number of scaffolds > 100K nt 0 0.0%

Number of scaffolds > 1M nt 0 0.0%

Number of scaffolds > 10M nt 0 0.0%

Mean scaffold size 721

Median scaffold size 252

N50 scaffold length 2038

L50 scaffold count 13704

scaffold %A 25.03

scaffold %C 24.72

scaffold %G 24.82

scaffold %T 25.39

scaffold %N 0.04

scaffold %non-ACGTN 0.00

Number of scaffold non-ACGTN nt 0

Percentage of assembly in scaffolded contigs 0.7%

Percentage of assembly in unscaffolded contigs 99.3%

Average number of contigs per scaffold 1.0

Average length of break (>25 Ns) between contigs in scaffold 97

Number of contigs 141925

Number of contigs in scaffolds 442

Number of contigs not in scaffolds 141483

Total size of contigs 102089767

Longest contig 20241

Shortest contig 6

Number of contigs > 1K nt 25775 18.2%

Number of contigs > 10K nt 148 0.1%

Number of contigs > 100K nt 0 0.0%

Number of contigs > 1M nt 0 0.0%

Number of contigs > 10M nt 0 0.0%

Mean contig size 719

Median contig size 252

N50 contig length 2028

L50 contig count 13756

contig %A 25.03

contig %C 24.73

contig %G 24.83

contig %T 25.39

contig %N 0.01

contig %non-ACGTN 0.00

Number of contig non-ACGTN nt 0

**Assembly is decent, let’s move on to assembling our BLAST database. First, we need to install BLAST+:**

wget https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.16.0+-x64-linux.tar.gz

tar -xzvf ncbi-blast-2.16.0+-x64-linux.tar.gz

**Next, make a BLAST database using our transcriptome:**

/project/stuckert/aspeidl/ncbi-blast-2.16.0+/bin/makeblastdb -in /project/stuckert/aspeidl/final\_project/spades\_output/transcripts.fasta -dbtype nucl -out transcripts\_estrogen\_db

**Create hsp27.fasta file with sequence using nano:** https://www.ncbi.nlm.nih.gov/nuccore/NG\_008995.1?from=5001&to=6740&report=fasta

**Query the database for HSP27 using new FASTA file:**

/project/stuckert/aspeidl/ncbi-blast-2.16.0+/bin/blastn -query /project/stuckert/aspeidl/final\_project/hsp27.fasta -db /project/stuckert/aspeidl/final\_project/transcripts\_estrogen\_db -out /project/stuckert/aspeidl/final\_project/transcript\_estrogen\_results.out -outfmt 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gene ID** | **Subject Seq ID** | **% Identity** | **Length of Alignment** | **E-value** |
| NG\_008995.1:5001-6740 | NODE\_17613\_length\_1615\_cov\_1859.696451\_g10461\_i0 | 100 | 1266 | 2.17e-164 |
| NG\_008995.1:5001-6740 | NODE\_27972\_length\_884\_cov\_3655.200708\_g10461\_i1 | 99.486 | 317 | 2.17e-164 |
| NG\_008995.1:5001-6740 | NODE\_27972\_length\_884\_cov\_3655.200708\_g10461\_i1 | 100 | 66 | 1.58e-26 |
| NG\_008995.1:5001-6740 | NODE\_29207\_length\_826\_cov\_80.665399\_g18146\_i0 | 93.023 | 86 | 1.22e-27 |
| NG\_008995.1:5001-6740 | NODE\_21493\_length\_1288\_cov\_74.782574\_g12749\_i0 | 91.086 | 86 | 5.70e-26 |

Some of the top hits from the query. HSP27 transcripts are found through the samples. Full file is in Github in transcripts\_estrogen\_results.output

**Finally, we will take our gene of interest located in our sample and place the gene ID into KEGG to do pathway analysis.**

NCBI ID:  NG\_008995.1

**Results:**

A diagram of a flowchart

Description automatically generated**A diagram of a network

Description automatically generated**

KEGG pathway analysis shows that HSP27 is highly regulated in the MAPK pathway which feeds into the estrogen response signaling. Literature on the topic highlights the phosphorylation of HSP27 by MAPK pathway which regulates estrogen signaling and promotes cancer cell survival.

**Overall workflow:**

1. **Download data from GEO and perform FASTQC to assess quality.**
2. **Assemble the transcriptome using SPAdes.**
3. **Build NCBI BLAST+ database using assembled transcriptome and query HSP27.**
4. **KEGG Pathway analysis.**