7BBG1002
Applied Bioinformatics

Introduction to the group project

Dr Aleksej Zelezniak
Randall Centre for Cell & Molecular Biophysics

October 2025

Group project – Reproducing published research

Aims

- Reproduce the analysis of a published bioinformatics study
- Identify potential problems in the technical and scientific reproducibility
- Critically reflect on your findings

How it works

- Work collaboratively in groups of 3
- Work on a recent bioinformatics study (published since 2011) using NGS
- Analyse the data from the study starting from the raw data to reproduce the study's findings

Timeline and support

Introduction session - Today

Overview of the group project and outline of what is expected

Project workshop – 24th November

Three-hour workshop where you can ask for help and support

Submission deadline – 19th December 16:00

Submit your report on Keats (Hard deadline!)

Additional support:

Post questions on Keats, we or GTAs will try to answer them

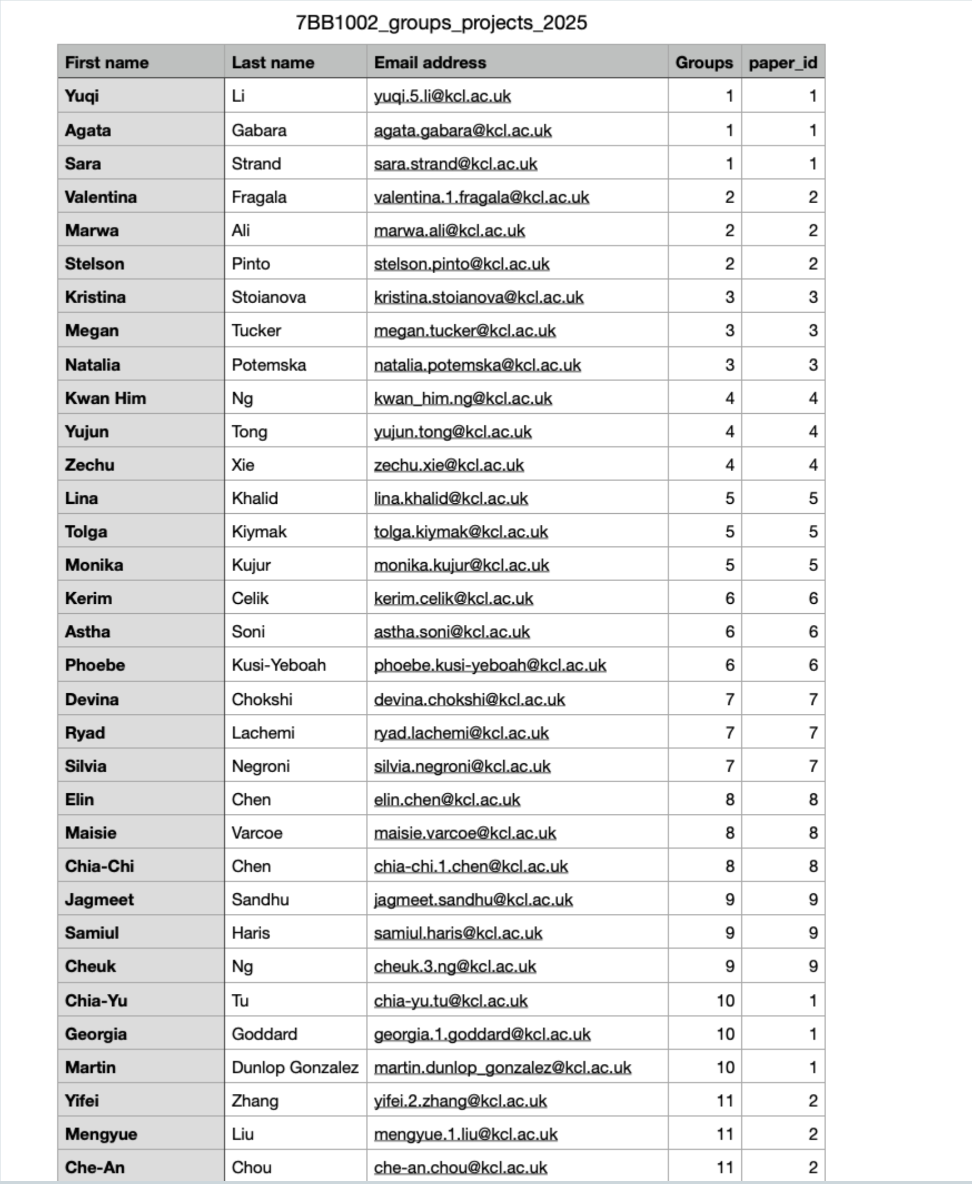
Project basic rules

Do

- Reproduce the analysis of a published study that uses NGS technology
- Start your analysis from raw data (you may start from BAM files)
- Follow the methods in the published paper
- Critically reflect on what you are doing, especially when things don't work

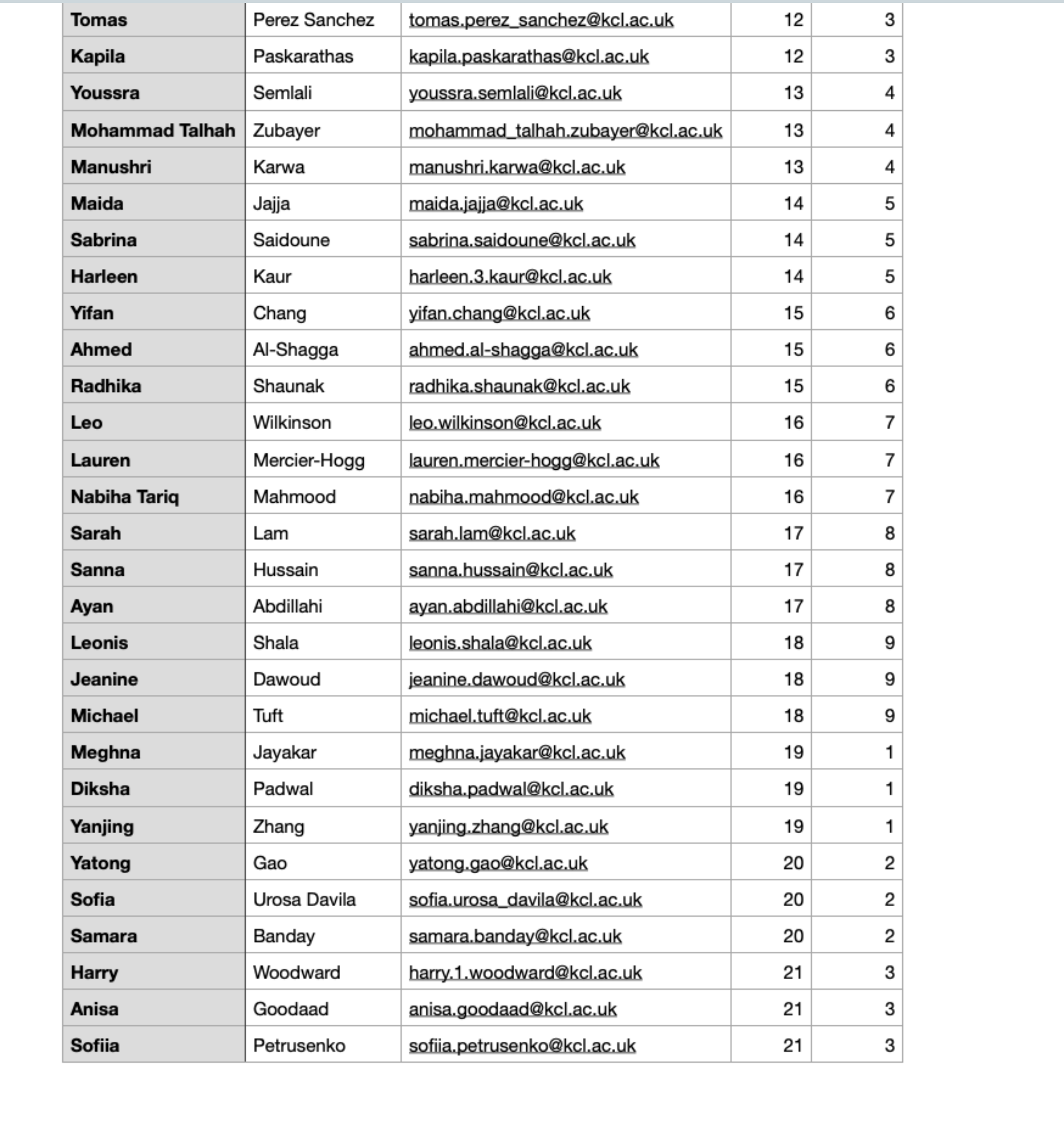
Don't

- Copy analysis code from the original publications
- Use supplementary processed data



7BB1002_groups_projects_2025

First name	Last name	Email address	Groups	paper_id
Yuqi	Li	yuqi.5.li@kcl.ac.uk	1	1
Agata	Gabara	agata.gabara@kcl.ac.uk	1	1
Sara	Strand	sara.strand@kcl.ac.uk	1	1
Valentina	Fragala	valentina.1.fragala@kcl.ac.uk	2	2
Marwa	Ali	marwa.ali@kcl.ac.uk	2	2
Stelson	Pinto	stelson.pinto@kcl.ac.uk	2	2
Kristina	Stoianova	kristina.stoianova@kcl.ac.uk	3	3
Megan	Tucker	megan.tucker@kcl.ac.uk	3	3
Natalia	Potemska	natalia.potemska@kcl.ac.uk	3	3
Kwan Him	Ng	kwan_him.ng@kcl.ac.uk	4	4
Yujun	Tong	yujun.tong@kcl.ac.uk	4	4
Zechu	Xie	zechu.xie@kcl.ac.uk	4	4
Lina	Khalid	lina.khalid@kcl.ac.uk	5	5
Tolga	Kiymak	tolga.kiymak@kcl.ac.uk	5	5
Monika	Kujur	monika.kujur@kcl.ac.uk	5	5
Kerim	Celik	kerim.celik@kcl.ac.uk	6	6
Astha	Soni	astha.soni@kcl.ac.uk	6	6
Phoebe	Kusi-Yeboah	phoebe.kusi-yeboah@kcl.ac.uk	6	6
Devina	Chokshi	devina.chokshi@kcl.ac.uk	7	7
Ryad	Lachemi	ryad.lachemi@kcl.ac.uk	7	7
Silvia	Negroni	silvia.negroni@kcl.ac.uk	7	7
Elin	Chen	elin.chen@kcl.ac.uk	8	8
Maisie	Varcoe	maisie.varcoe@kcl.ac.uk	8	8
Chia-Chi	Chen	chia-chi.1.chen@kcl.ac.uk	8	8
Jagmeet	Sandhu	jagmeet.sandhu@kcl.ac.uk	9	9
Samiul	Haris	samiul.haris@kcl.ac.uk	9	9
Cheuk	Ng	cheuk.3.ng@kcl.ac.uk	9	9
Chia-Yu	Tu	chia-yu.tu@kcl.ac.uk	10	1
Georgia	Goddard	georgia.1.goddard@kcl.ac.uk	10	1
Martin	Dunlop Gonzalez	martin.dunlop_gonzalez@kcl.ac.uk	10	1
Yifei	Zhang	yifei.2.zhang@kcl.ac.uk	11	2
Mengyue	Liu	mengyue.1.liu@kcl.ac.uk	11	2
Che-An	Chou	che-an.chou@kcl.ac.uk	11	2



Tomas	Perez Sanchez	tomas.perez_sanchez@kcl.ac.uk	12	3
Kapila	Paskarathas	kapila.paskarathas@kcl.ac.uk	12	3
Youssra	Semlali	youssra.semlali@kcl.ac.uk	13	4
Mohammad Talhah	Zubayer	mohammad_talhah.zubayer@kcl.ac.uk	13	4
Manushri	Karwa	manushri.karwa@kcl.ac.uk	13	4
Maida	Jajja	maida.jajja@kcl.ac.uk	14	5
Sabrina	Saidoune	sabrina.saidoune@kcl.ac.uk	14	5
Harleen	Kaur	harleen.3.kaur@kcl.ac.uk	14	5
Yifan	Chang	yifan.chang@kcl.ac.uk	15	6
Ahmed	Al-Shagga	ahmed.al-shagga@kcl.ac.uk	15	6
Radhika	Shaunak	radhika.shaunak@kcl.ac.uk	15	6
Leo	Wilkinson	leo.wilkinson@kcl.ac.uk	16	7
Lauren	Mercier-Hogg	lauren.mercier-hogg@kcl.ac.uk	16	7
Nabiha Tariq	Mahmood	nabiha.mahmood@kcl.ac.uk	16	7
Sarah	Lam	sarah.lam@kcl.ac.uk	17	8
Sanna	Hussain	sanna.hussain@kcl.ac.uk	17	8
Ayan	Abdillahi	ayan.abdillahi@kcl.ac.uk	17	8
Leonis	Shala	leonis.shala@kcl.ac.uk	18	9
Jeanine	Dawoud	jeanine.dawoud@kcl.ac.uk	18	9
Michael	Tuft	michael.tuft@kcl.ac.uk	18	9
Meghna	Jayakar	meghna.jayakar@kcl.ac.uk	19	1
Diksha	Padwal	diksha.padwal@kcl.ac.uk	19	1
Yanjing	Zhang	yanjing.zhang@kcl.ac.uk	19	1
Yatong	Gao	yatong.gao@kcl.ac.uk	20	2
Sofia	Urosa Davila	sofia.urosa_davila@kcl.ac.uk	20	2
Samara	Banday	samara.banday@kcl.ac.uk	20	2
Harry	Woodward	harry.1.woodward@kcl.ac.uk	21	3
Anisa	Goodaad	anisa.goodaad@kcl.ac.uk	21	3
Sofiia	Petrusenko	sofiia.petrusenko@kcl.ac.uk	21	3

Submission and assessment

Assessment outline

Everyone must submit a report by the deadline. You will be assessed individually on the quality of your personal submission. Most of your report can be written as a group, but you will need to add your own personal reflections.

What your submission must include

1. A scientific report (2000 words and 2+ figures) in article format (i.e. your report must include an introduction, a results, a discussion and a methods section)

2. The code you used for analysis annotated for example with R Markdown. If you use github, make sure the code is provided with the submission document, **not just the link**

3. A reflection on the reproducibility of the research both on a technical and a scientific level (up to 1000 words)

4. A statement of how the work was shared in the group

Submit these as a single pdf by the deadline.

Some pointers for the report

Scientific report

This should be in the format of a scientific paper (see here for a guide) and include 2-4 scientific figures on your results with appropriate figure captions. You may use any appropriate layout (for example, journal templates). This part of the assessment and the code annotation should be completed as a group.

You must not copy from the original article, and your write up should focus on the **question of reproducibility.** For example, your introduction may sound like this: *"XYZ , in their 2017 study, concluded that during anaerobic growth, yeast shutdowns its TCA cycle by showing downregulation of TCA cycle genes and upregulation of 28 other genes based on RNA-Seq data. Here, we investigate whether we can reproduce this finding by reanalysing of the data."*

Reflection on reproducibility

This part of the submission must be completed individually. You should reflect on two main points:

a) How easy was the technical replication of the research? (Was the information in the publication sufficient? Did you encounter any technical problems?)

b) Were you able to reproduce the scientific findings? (Are the findings statistically significant? Are there choices in algorithms that where not detailed in the paper but change the outcome of the analysis?)

Assessment criteria

- **Scientific report (60% - 20% for figures, 20% for writing and 20% for content)**

  Content: Factual accuracy and understanding of the methodology used and results shown

  Figures: Clear and well-annotated figures highlighting key findings (We want to see figures produced by you, not just replicas of what is in the publication – Remember the focus on reproducibility!)

  Writing: Clear writing at an academic standard

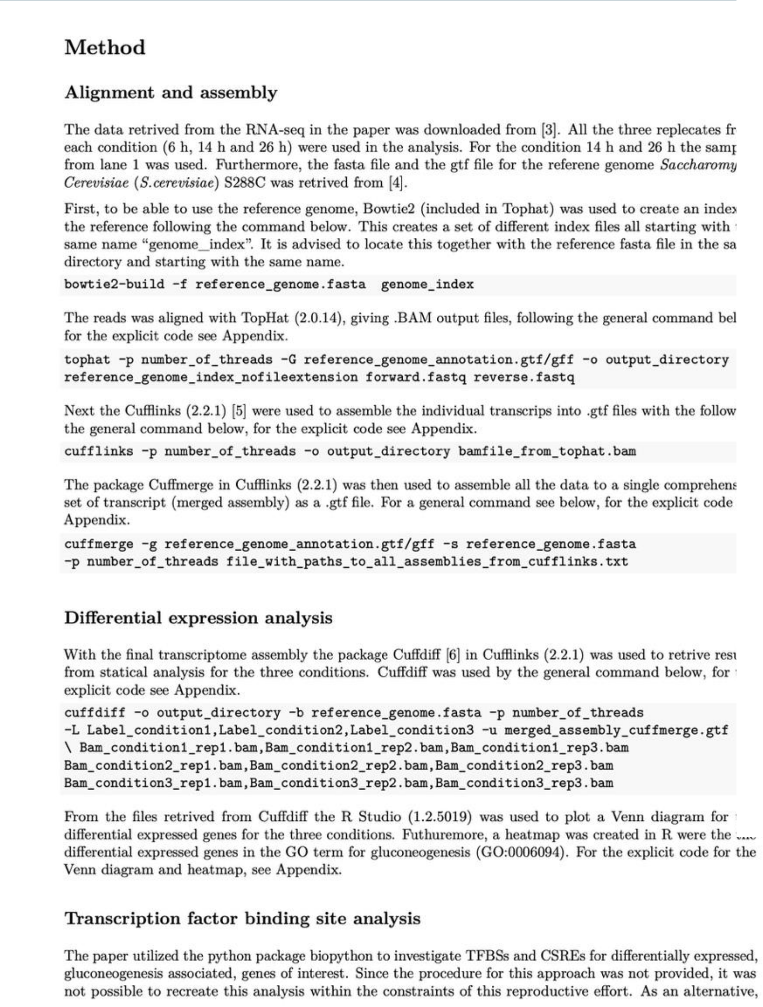
- **Reflection (20%)**

  Insightful commentary on the process of reproducing research results

- **Analysis code and markdown (20%)**

  Clear annotations and commentary for using sensible analysis methods following the original methodology, feel free to use Git for this.

Examples



Method

Alignment and assembly

The data retrived from the RNA-seq in the paper was downloaded from [3]. All the three replecates fr each condition (6 h, 14 h and 26 h) were used in the analysis. For the condition 14 h and 26 h the samp from lane 1 was used. Furthermore, the fasta file and the gtf file for the referene genome *Saccharomy Cerevisiae* (*S.cerevisiae*) S288C was retrived from [4].

First, to be able to use the reference genome, Bowtie2 (included in Tophat) was used to create an index the reference following the command below. This creates a set of different index files all starting with same name “genome_index”. It is advised to locate this together with the reference fasta file in the sa directory and starting with the same name.

```
bowtie2-build -f reference_genome.fasta  genome_index
```

The reads was aligned with TopHat (2.0.14), giving .BAM output files, following the general command bel for the explicit code see Appendix.

```
tophat -p number_of_threads -G reference_genome_annotation.gtf/gff -o output_directory
reference_genome_index_nofileextension forward.fastq reverse.fastq
```

Next the Cufflinks (2.2.1) [5] were used to assemble the individual transcrips into .gtf files with the follow the general command below, for the explicit code see Appendix.

```
cufflinks -p number_of_threads -o output_directory bamfile_from_tophat.bam
```

The package Cuffmerge in Cufflinks (2.2.1) was then used to assemble all the data to a single comprehens set of transcript (merged assembly) as a .gtf file. For a general command see below, for the explicit code Appendix.

```
cuffmerge -g reference_genome_annotation.gtf/gff -s reference_genome.fasta
-p number_of_threads file_with_paths_to_all_assemblies_from_cufflinks.txt
```

Differential expression analysis

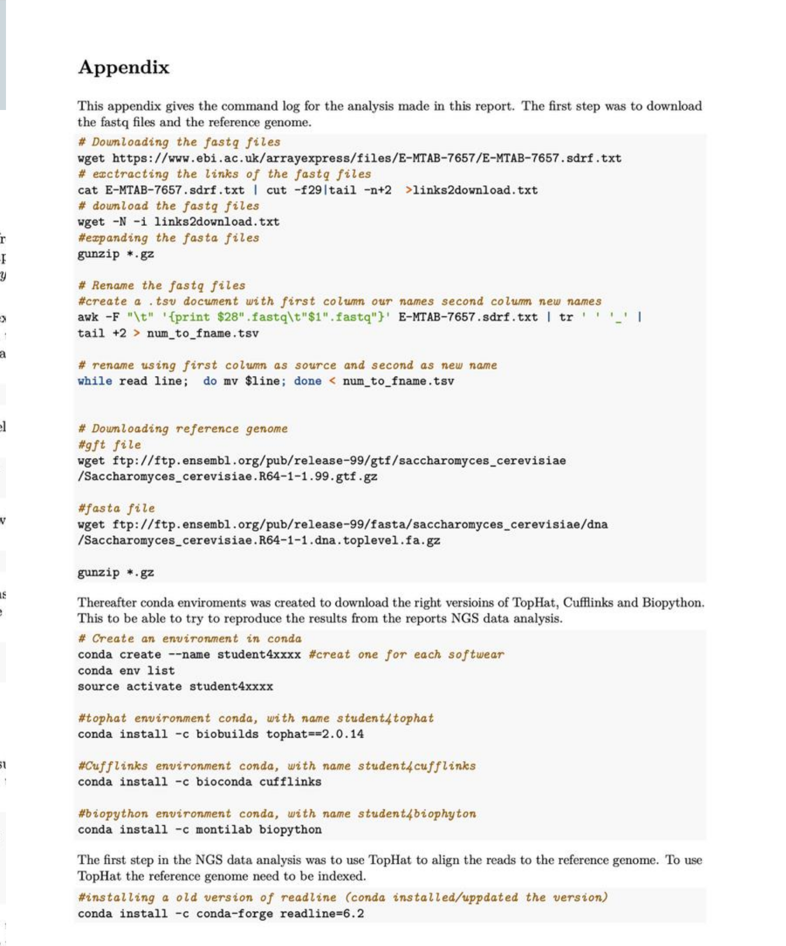
With the final transcriptome assembly the package Cuffdiff [6] in Cufflinks (2.2.1) was used to retrive resu from statical analysis for the three conditions. Cuffdiff was used by the general command below, for explicit code see Appendix.

```
cuffdiff -o output_directory -b reference_genome.fasta -p number_of_threads
-L Label_condition1,Label_condition2,Label_condition3 -u merged_assembly_cuffmerge.gtf
\ Bam_condition1_rep1.bam,Bam_condition1_rep2.bam,Bam_condition1_rep3.bam
Bam_condition2_rep1.bam,Bam_condition2_rep2.bam,Bam_condition2_rep3.bam
Bam_condition3_rep1.bam,Bam_condition3_rep2.bam,Bam_condition3_rep3.bam
```

From the files retrived from Cuffdiff the R Studio (1.2.5019) was used to plot a Venn diagram for differential expressed genes for the three conditions. Futhuremore, a heatmap was created in R were the differential expressed genes in the GO term for gluconeogenesis (GO:0006094). For the explicit code for the Venn diagram and heatmap, see Appendix.

Transcription factor binding site analysis

The paper utilized the python package biopython to investigate TFBSs and CSREs for differentially expressed, gluconeogenesis associated, genes of interest. Since the procedure for this approach was not provided, it was not possible to recreate this analysis within the constraints of this reproductive effort. As an alternative,



Appendix

This appendix gives the command log for the analysis made in this report. The first step was to download the fastq files and the reference genome.

```
# Downloading the fastq files
wget https://www.ebi.ac.uk/arrayexpress/files/E-MTAB-7657/E-MTAB-7657.sdrf.txt
# exctracting the links of the fastq files
cat E-MTAB-7657.sdrf.txt | cut -f29|tail -n+2  >links2download.txt
# download the fastq files
wget -N -i links2download.txt
#expanding the fasta files
gunzip *.gz

# Rename the fastq files
#create a .tsv document with first column our names second column new names
awk -F "\t" '{print $28".fastq\t"$1".fastq"}' E-MTAB-7657.sdrf.txt | tr ' ' '_' |
tail +2 > num_to_fname.tsv

# rename using first column as source and second as new name
while read line;  do mv $line; done < num_to_fname.tsv


# Downloading reference genome
#gft file
wget ftp://ftp.ensembl.org/pub/release-99/gtf/saccharomyces_cerevisiae
/Saccharomyces_cerevisiae.R64-1-1.99.gtf.gz

#fasta file
wget ftp://ftp.ensembl.org/pub/release-99/fasta/saccharomyces_cerevisiae/dna
/Saccharomyces_cerevisiae.R64-1-1.dna.toplevel.fa.gz

gunzip *.gz
```

Thereafter conda enviroments was created to download the right versioins of TopHat, Cufflinks and Biopython. This to be able to try to reproduce the results from the reports NGS data analysis.

```
# Create an environment in conda
conda create --name student4xxxx #creat one for each softwear
conda env list
source activate student4xxxx

#tophat environment conda, with name student4tophat
conda install -c biobuilds tophat==2.0.14

#Cufflinks environment conda, with name student4cufflinks
conda install -c bioconda cufflinks

#biopython environment conda, with name student4biophyton
conda install -c montilab biopython
```

The first step in the NGS data analysis was to use TopHat to align the reads to the reference genome. To use TopHat the reference genome need to be indexed.

```
#installing a old version of readline (conda installed/uppdated the version)
conda install -c conda-forge readline=6.2
```

KING'S College LONDON