TGAC miRNA Workshop

Day 1 Morning Session 1

Day 1 Morning Session 1

Raw Data Processing - Quality Control - Normalisation - Annotation

Sequence Data File Formats

FASTQ

- Raw sequencing information from the instrumentation and quality scores

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Sequence Data File Formats

FASTA

- Arbitrary header and processed sequences

```
>HEADER 1
TCCTCTCTCACAAGCCAGGCTAT
>HEADER 2
TTTCAGATAATAGAAGTGAAGCGT
>HEADER 3
TCGGTTCCTGAATGGTTTTT
```

Sequence Data File Formats

GFF

- Sequence Annotations (more on this later…)
- (up to) 9 Tab separated values

```
Chr1    TAIR10  CDS 30902   31079   .   +   1   Parent=AT1G01040.2,AT1G01040.2-Protein;
Chr1    TAIR10  three_prime_UTR 31080   31120   .   +   .   Parent=AT1G01040.2
Chr1    TAIR10  gene    28500   28706   .   +   .   ID=AT1G01046;Note=miRNA;Name=AT1G01046
Chr1    TAIR10  miRNA   28500   28706   .   +   .   ID=AT1G01046.1;Parent=AT1G01046;Name=AT1G01046.1;Index=1
Chr1    TAIR10  exon    28500   28706   .   +   .   Parent=AT1G01046.1
Chr1    TAIR10  gene    31170   33153   .   -   .   ID=AT1G01050;Note=protein_coding_gene;Name=AT1G01050
Chr1    TAIR10  mRNA    31170   33153   .   -   .   ID=AT1G01050.1;Parent=AT1G01050;Name=AT1G01050.1;Index=1
Chr1    TAIR10  protein 31382   32670   .   -   .   ID=AT1G01050.1-Protein;Name=AT1G01050.1;Derives_from=AT1G01050.1
Chr1    TAIR10  five_prime_UTR  33029   33153   .   -   .   Parent=AT1G01050.1
Chr1    TAIR10  exon    33029   33153   .   -   .   Parent=AT1G01050.1
Chr1    TAIR10  CDS 32547   32670   .   -   0   Parent=AT1G01050.1,AT1G01050.1-Protein;
Chr1    TAIR10  exon    32547   32670   .   -   .   Parent=AT1G01050.1
Chr1    TAIR10  CDS 32431   32459   .   -   2   Parent=AT1G01050.1,AT1G01050.1-Protein;
Chr1    TAIR10  exon    32431   32459   .   -   .   Parent=AT1G01050.1
```

FASTQC

Initial Quality Control

- FASTQC
    - Most sequencers will give a basic overview of the run but this is not enough
- Check for technical quality of sequencing run
    - Looking for flow cell errors
    - machine errors
    - overall sequencing quality
- ‘Normal’ is of course dependant on what you might expect to see, certain samples will be bias for good reason

FASTQC

- Basic Statistics
  - Basic information about total read counts
- Per base sequence quality
  - PHRED scores
- Per tile sequence quality
- Per sequence quality scores

FASTQC

- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution

FASTQC

- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

FASTQC EXERCISE

- Download the FASTQC program from the Babraham institute website
    - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Run the program by typing cd ~/Desktop/fastqc
    - Type perl fastqc
- Load the data found in ~/Desktop/data/FASTQ/SRR873382.fastq

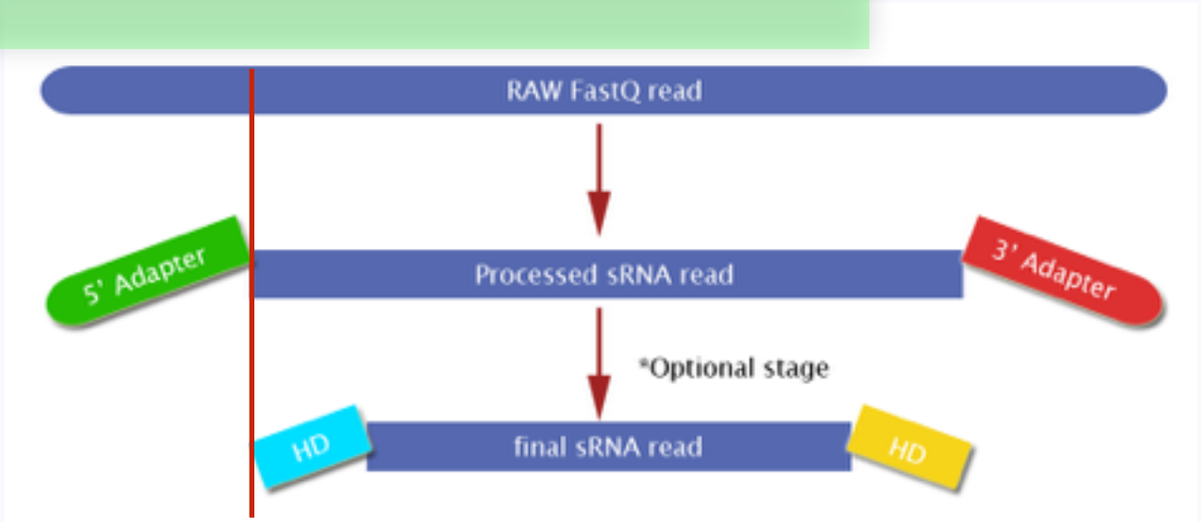
What can we tell?

- Per base sequence quality ok(ish)
- Per tile is not, why?
    - Is this a disaster?
- Per sequence quality scores are telling us that the majority of sequences are ok, but as we saw in the per tile, a problem occurred with a number of sequences
- Why is the nucleotide composition so spiky?

What can we tell?

- How about the Sequence Duplication level, FASTQC indicates a fail, but is this really a problem?
- KMER content, overrepresented motifs, what would we expect to see here?

Trimming Adapters



- sRNA sequences range from 18-40nt, however raw sequence reads from the illumina platform are generally selected as 50bp and therefore will potentially contain traces of the adapter sequence

- For the Illumina instruments these sequences will only be found at the 3' end of a read as the 5' end is used as a primer

- Adapter fragments must be removed prior to initial processing

- In addition to Illumina adapters certain protocols may introduce further inserts that require removal

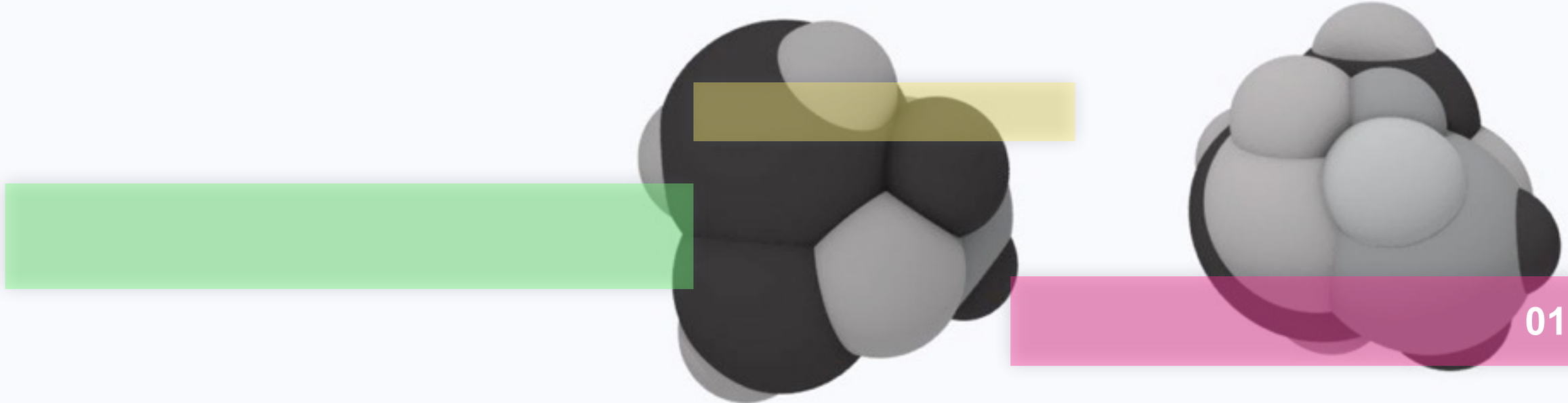
  - HD adapter protocol [Sorefan K et al. Reducing ligation bias of small RNAs in libraries for next generation sequencing. Silence. 2012;3: 4]

SOFTWARE

THE UEA SRNA WORKBENCH

- A suite of tools for various data preparation and analysis of sRNA-seq datasets
- Developed at the University of East Anglia and The Genome Analysis Centre
- Freely available from http://srna-workbench.cmp.uea.ac.uk
- Visit the web page and download the software
- Extract the file and place it on the Desktop

Sequence Data Preparation



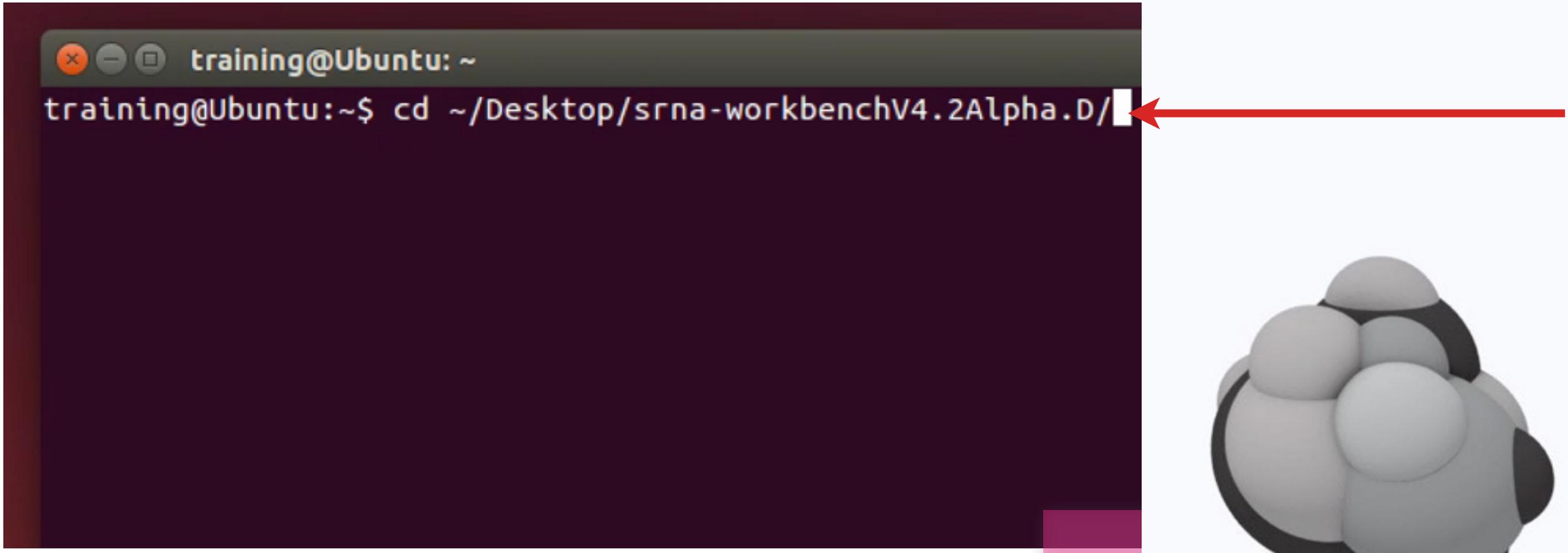
Trimming Adapters

- Using the sRNA Workbench exercise
- Adapter Sequence: ATCTCGTATGCCGTCT
- Sample investigating the effects of hypoxic conditions on MCF7 cells in H.Sapiens
- Load the Normoxia (time point 0) FASTQ sample

Adapter Removal Exercise

Strip the Adapters and Filter

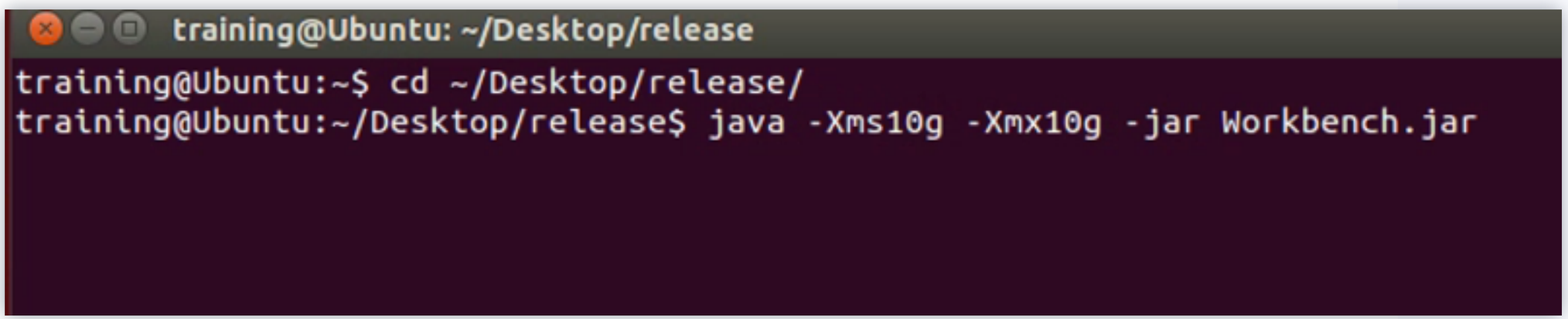
1. Load the Terminal App. Type: cd ~/Desktop/srna-workbenchV4.2.1Alpha.D



ADAPTER REMOVAL EXERCISE

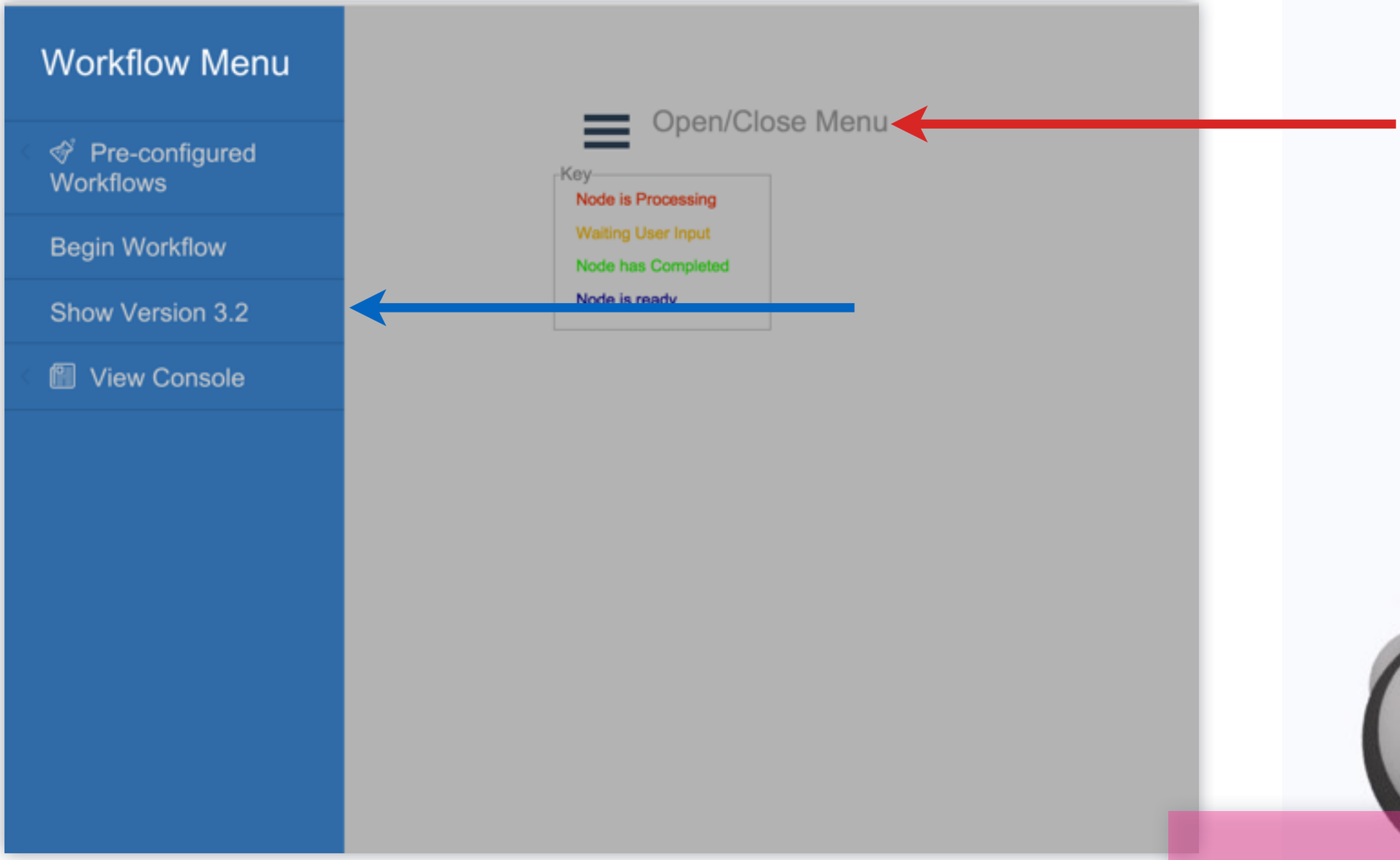
RUN THE sRNA WORKBENCH

1. Type java -Xms10g -Xmx10g -jar Workbench.jar



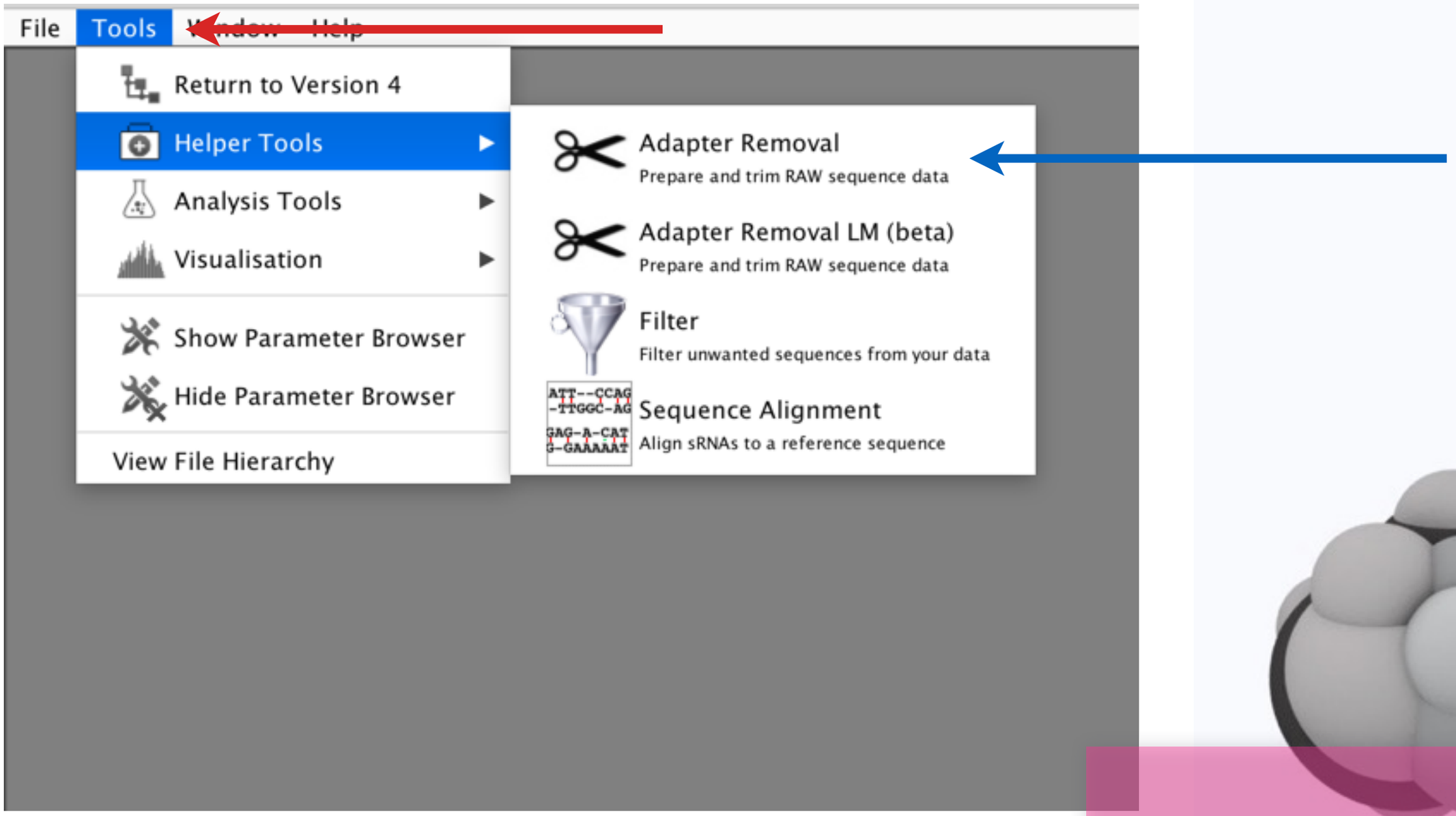
Adapter Removal Exercise

Load the V3.2 Workbench

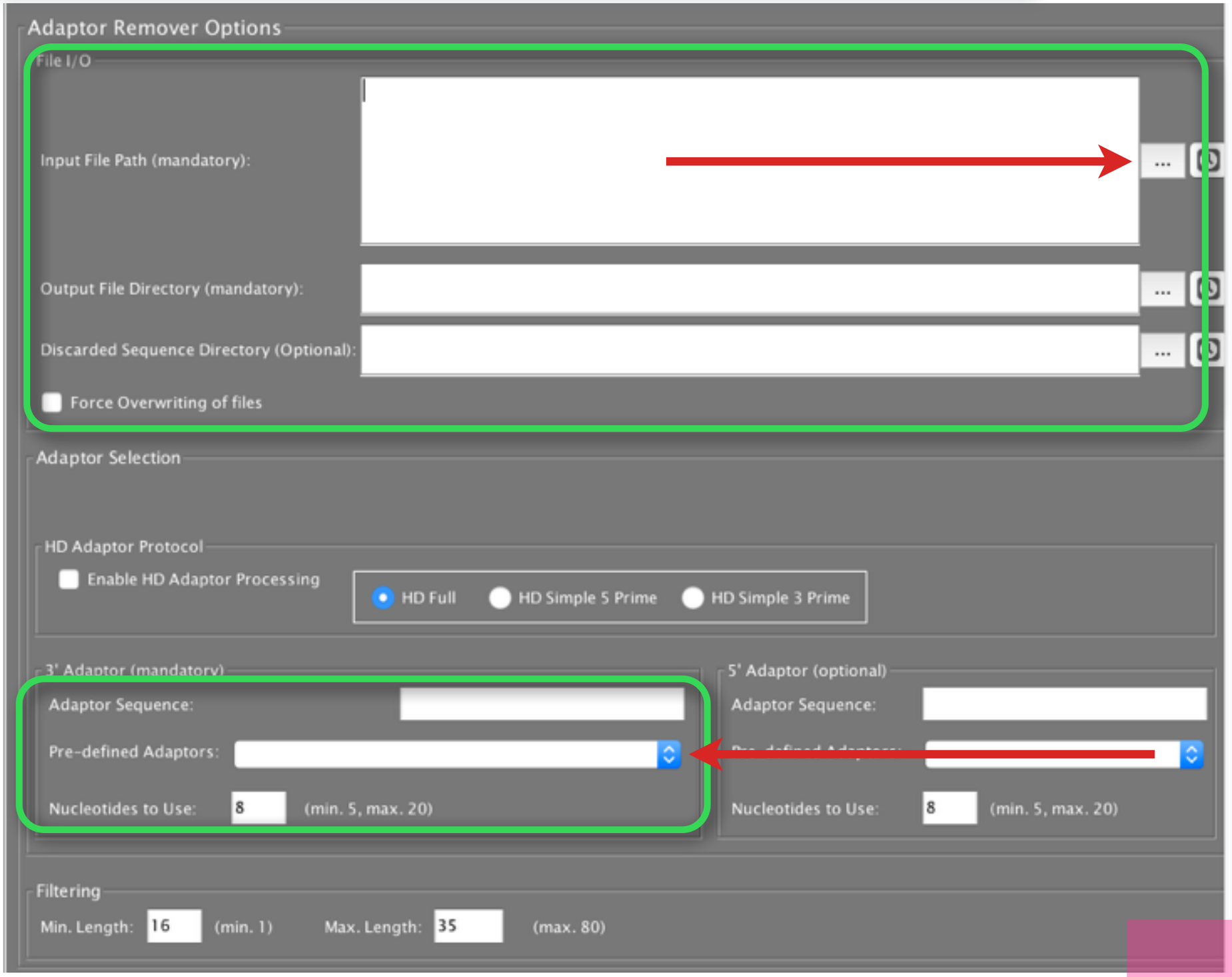


Adapter Removal Exercise

Load the Adapter Removal Tool



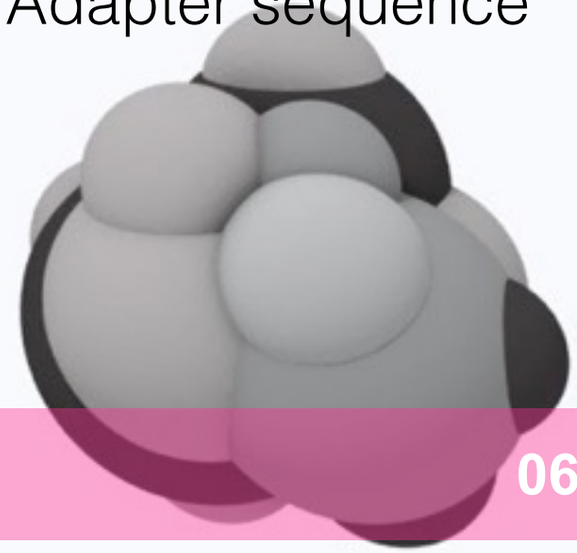
Removing Adapters



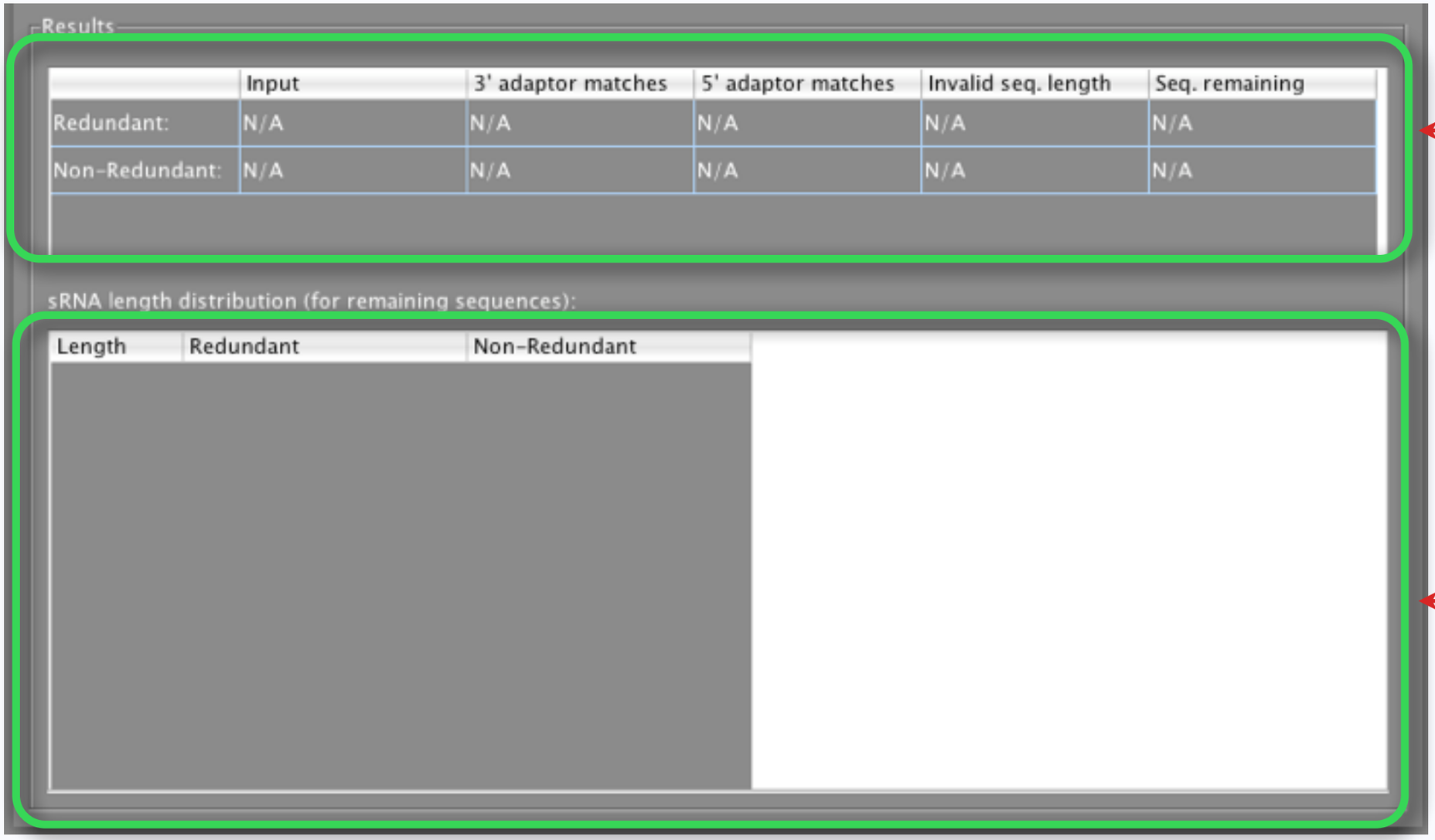
Open:
~/Desktop/data/FASTQ
/SRR873382.fastq

Output:
~/Desktop/data/results

Select the LMN2 3’
(Illumina Tru Seq 1.5)
Adapter sequence



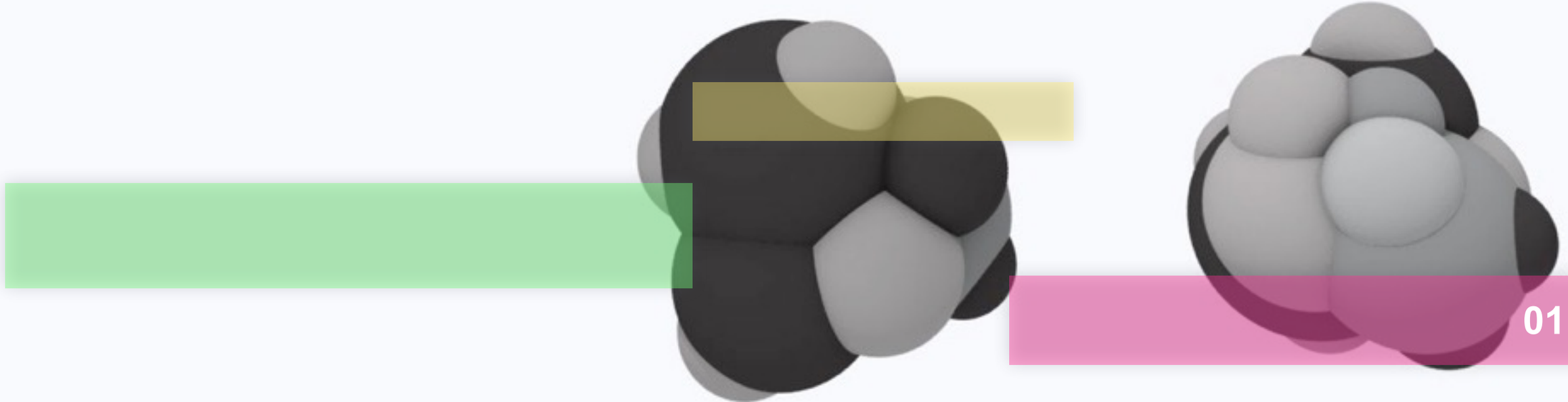
Results



Process Overview

Length Distributions

Further Quality Checking



SETUP PIPELINE

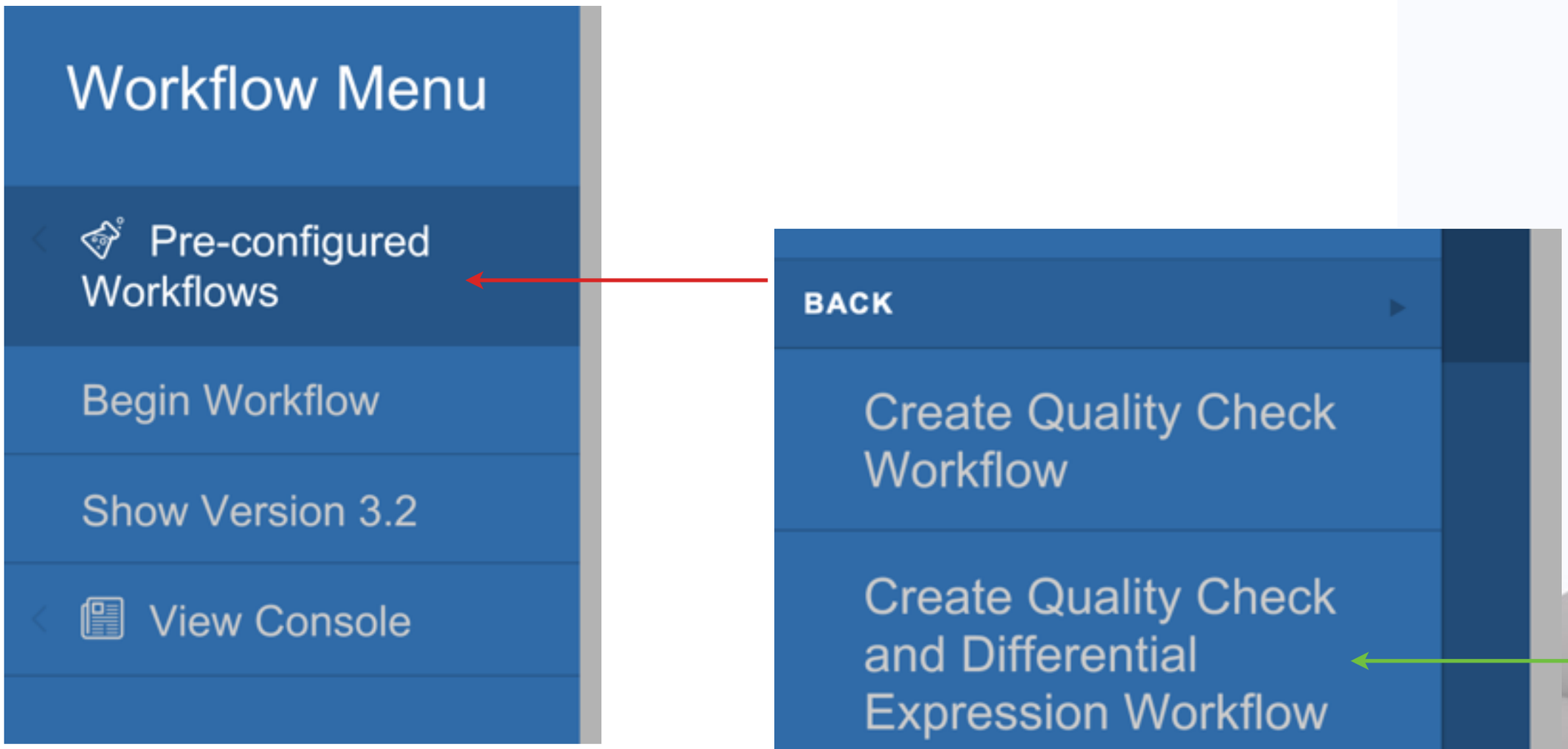
LOAD THE QUALITY CHECK PIPELINE

1. Load the Terminal App. Type: ‘cd ~/Desktop/srna-workbench

2. Type java -Xms10g -Xmx10g -jar Workbench.jar

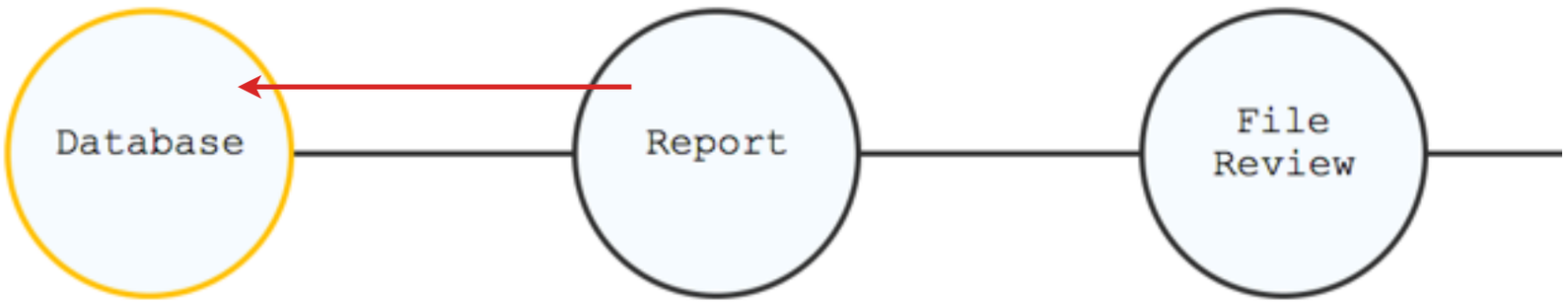
Setup Pipeline

Load the QC/DE Pipeline

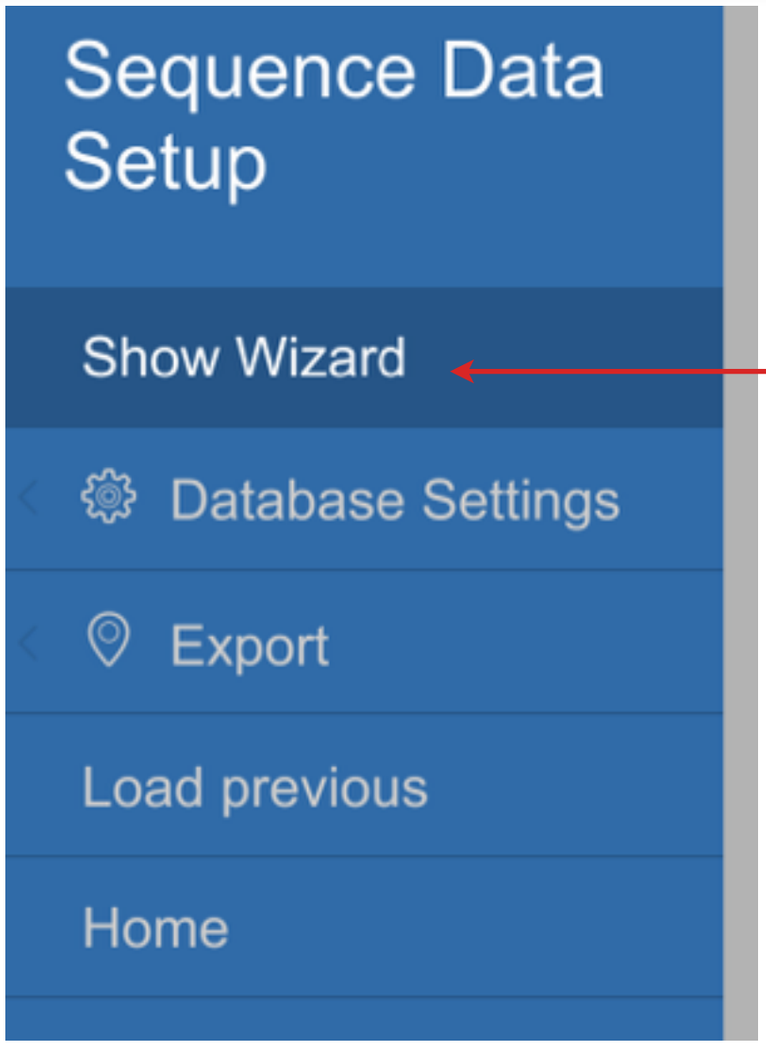


Setup Pipeline

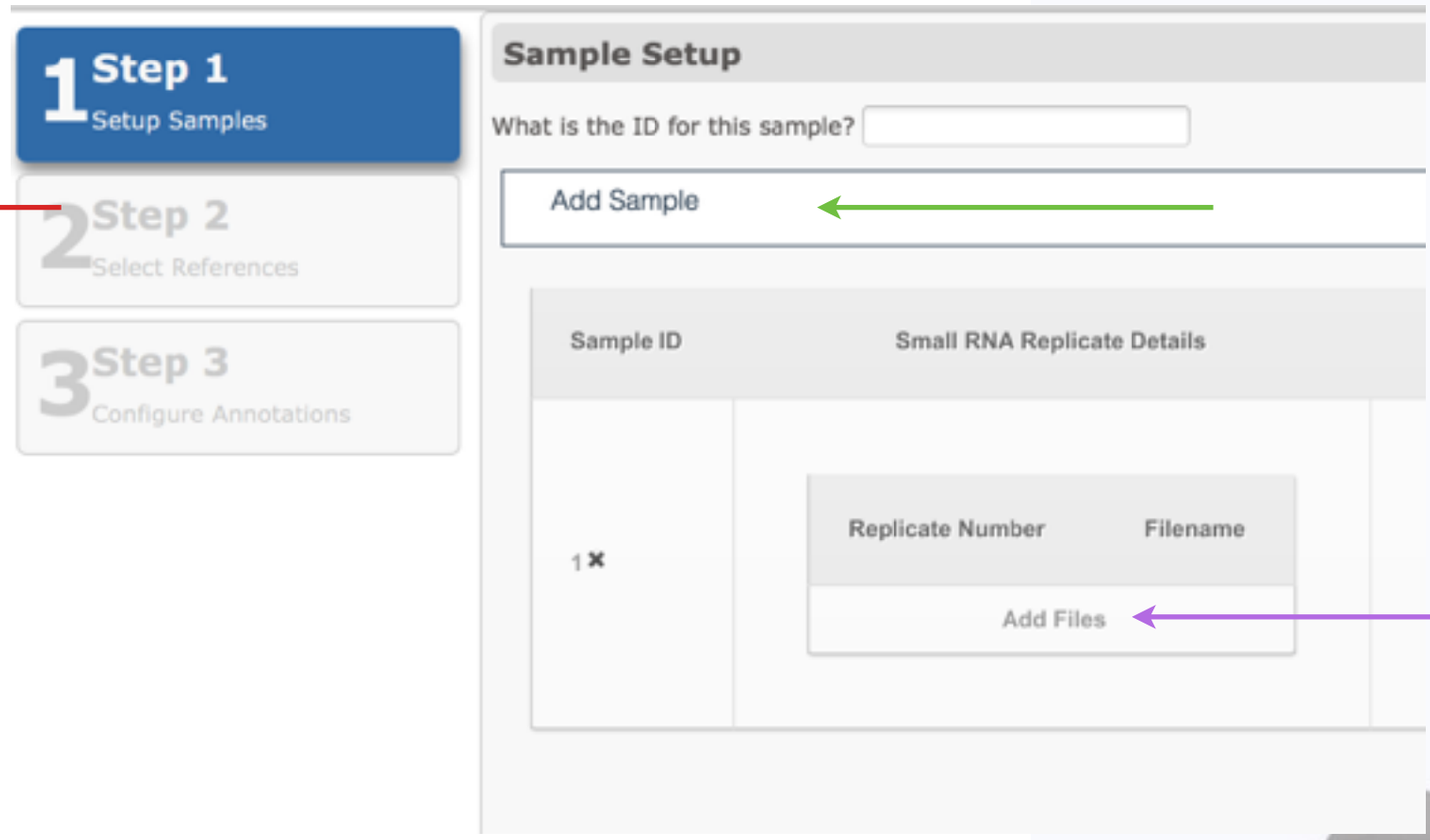
Load the Wizard



SETUP PIPELINE



CONFIGURE THE SAMPLES



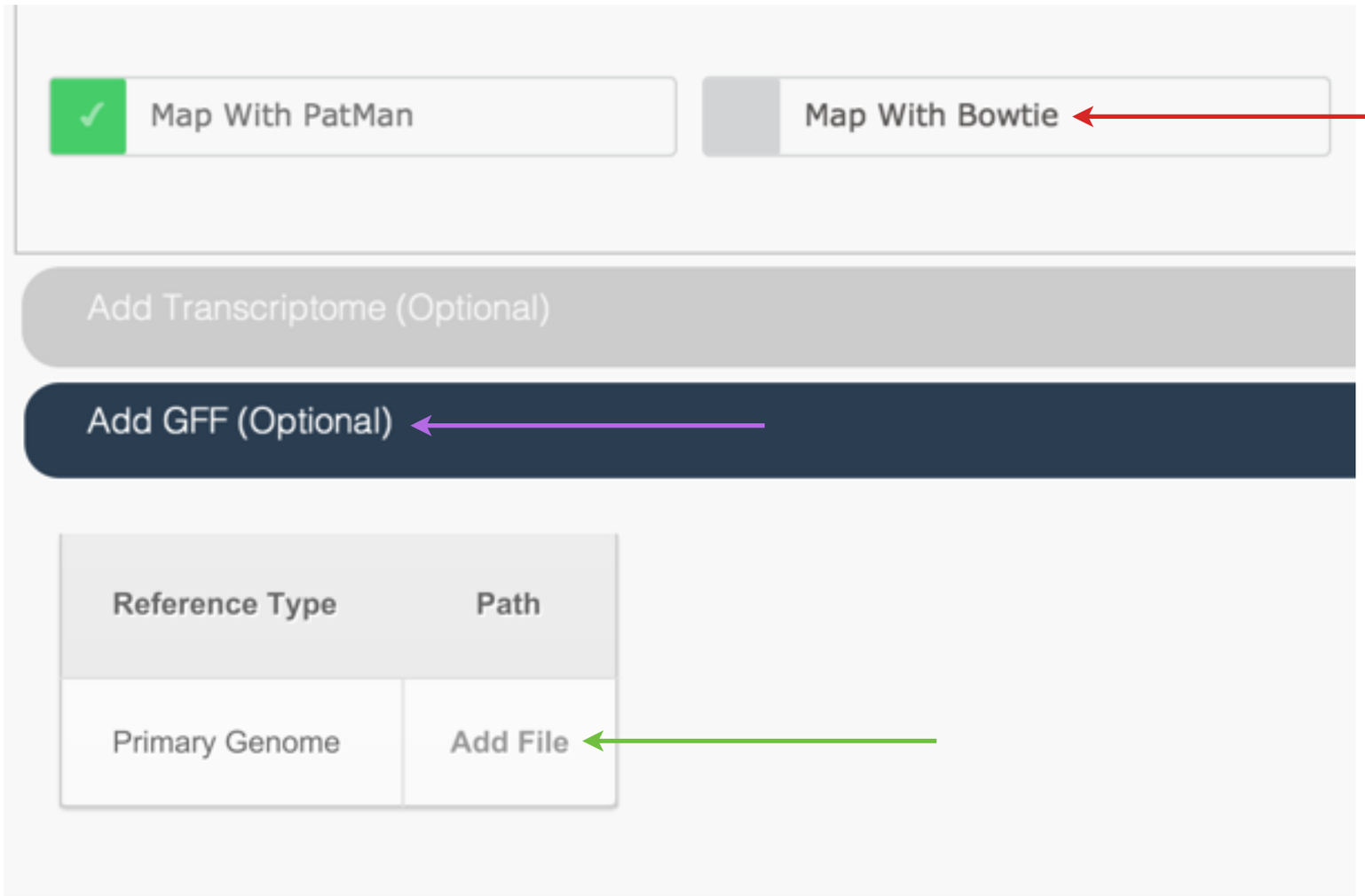
Load the files found in ~Desktop/data/FASTA

Sample 1: SRR873382_na_sampled_0.1.fasta, SRR873383_na_sampled_0.1.fasta

Sample 2: SRR873384_na_sampled_0.1.fasta, SRR873385_na_sampled_0.1.fasta

Setup Pipeline

Configure the References

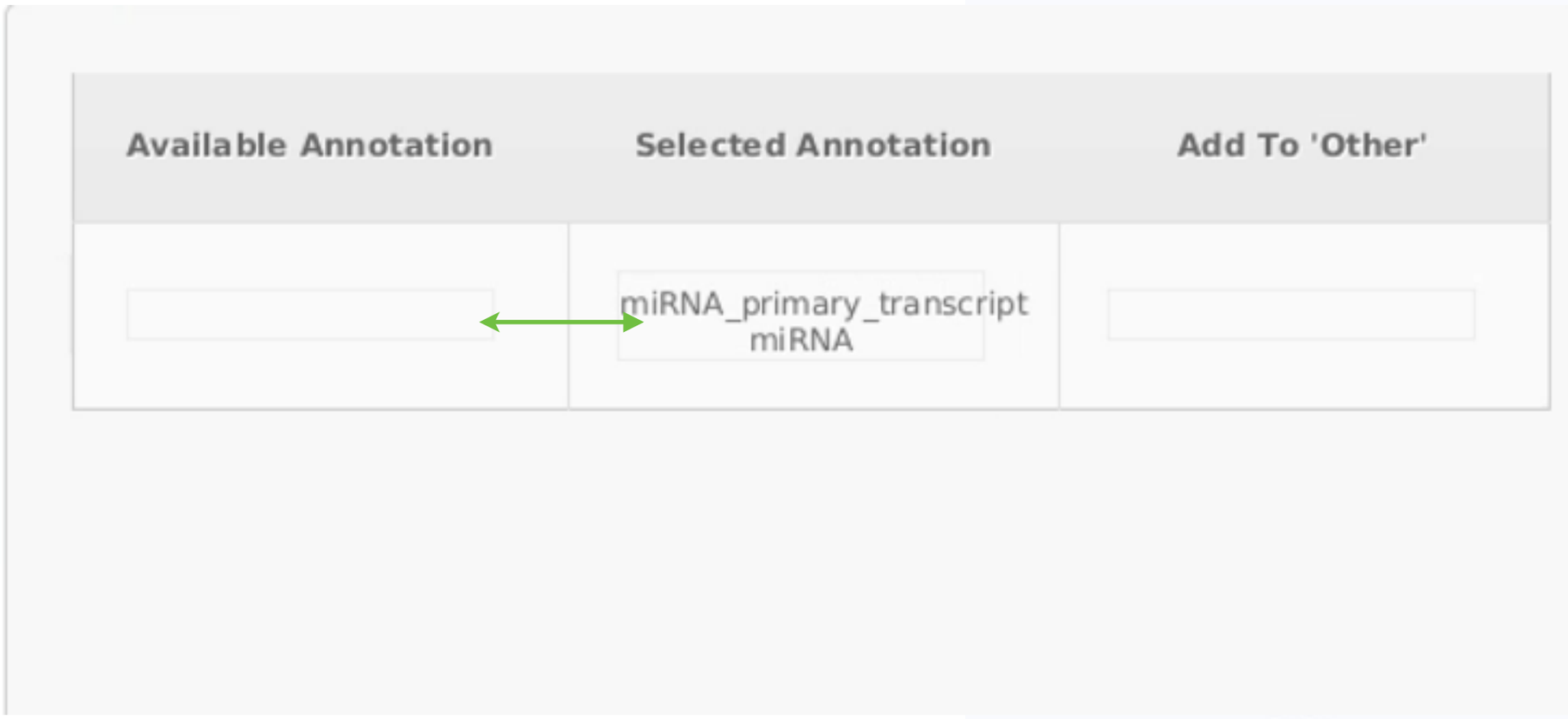


Select “Map with Bowtie”

Load the genome file found in ~Desktop/data/FASTA/genome/GRCh38_no_alt

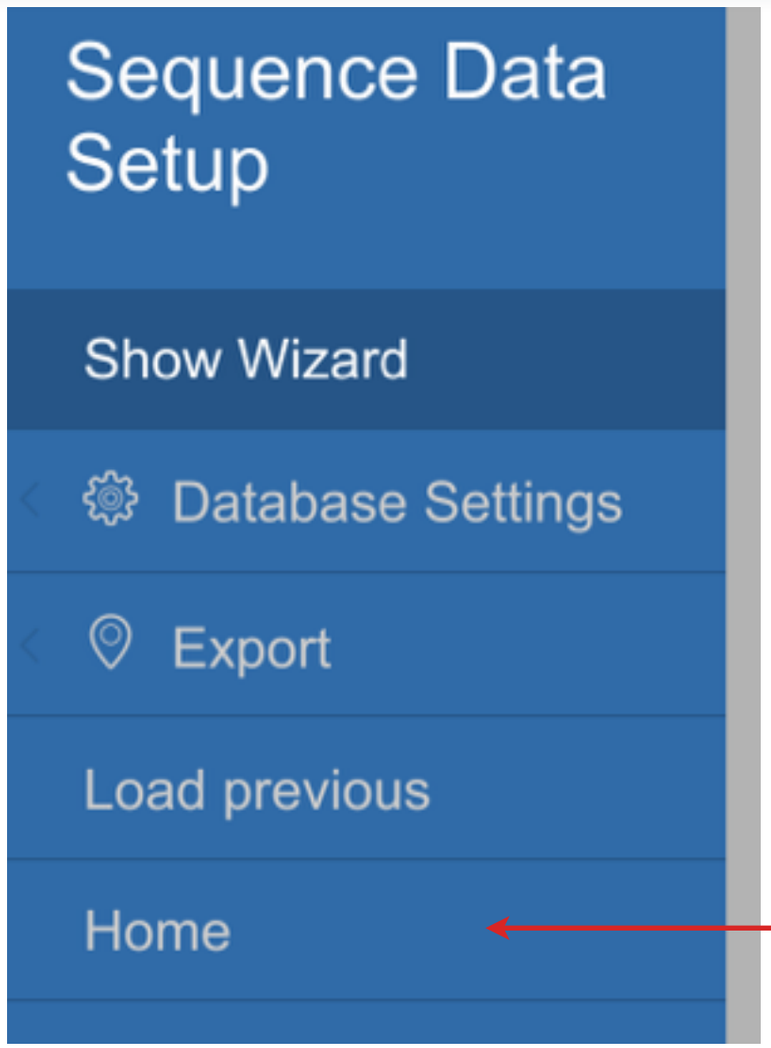
and the GFF file found in ~/Desktop/data/GFF

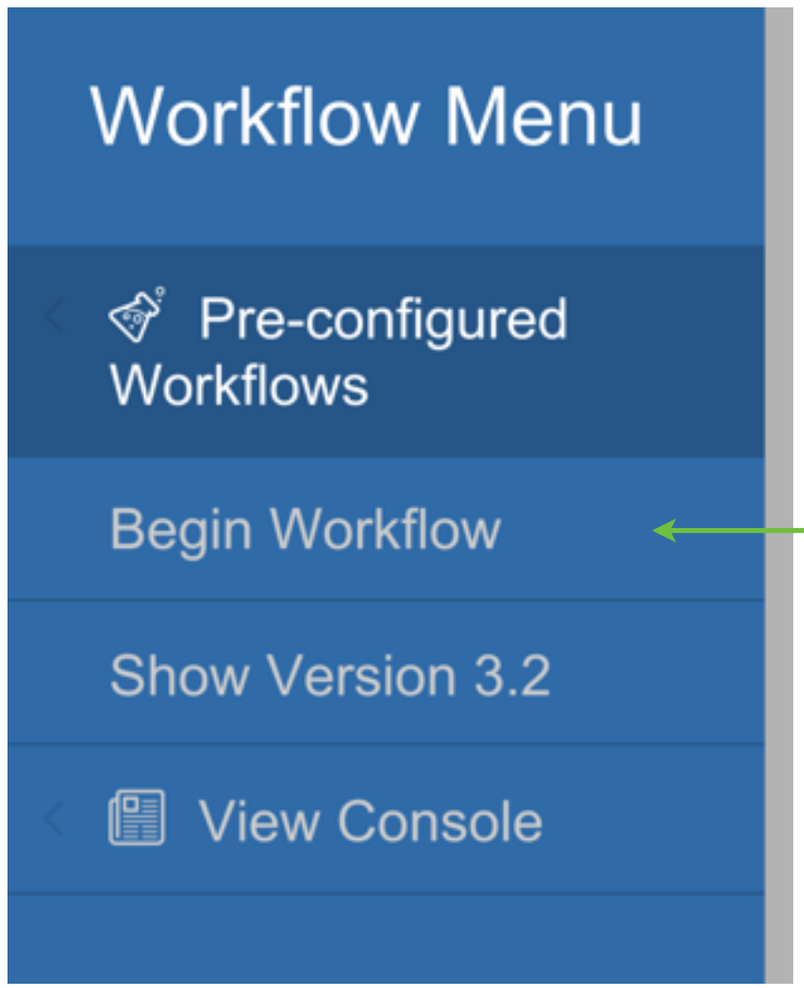
Setup Pipeline



Drag the miRNA and annotation into the selected options

Begin Pipeline





Quality Control

- An often overlooked step…
- Checking the sequencing information for issues that may impact downstream analysis
  - Fails: Technical
    - Library preparation
  - Fails: Biological
    - DNA or other contamination to RNA?
    - Sample infection?
- Various indications on the quality of both biological sample and sequencing run can be made through a set of diagnostic plots
  - Summarise, Visualise but not necessarily discarding!

Switch back to software

Quality Control

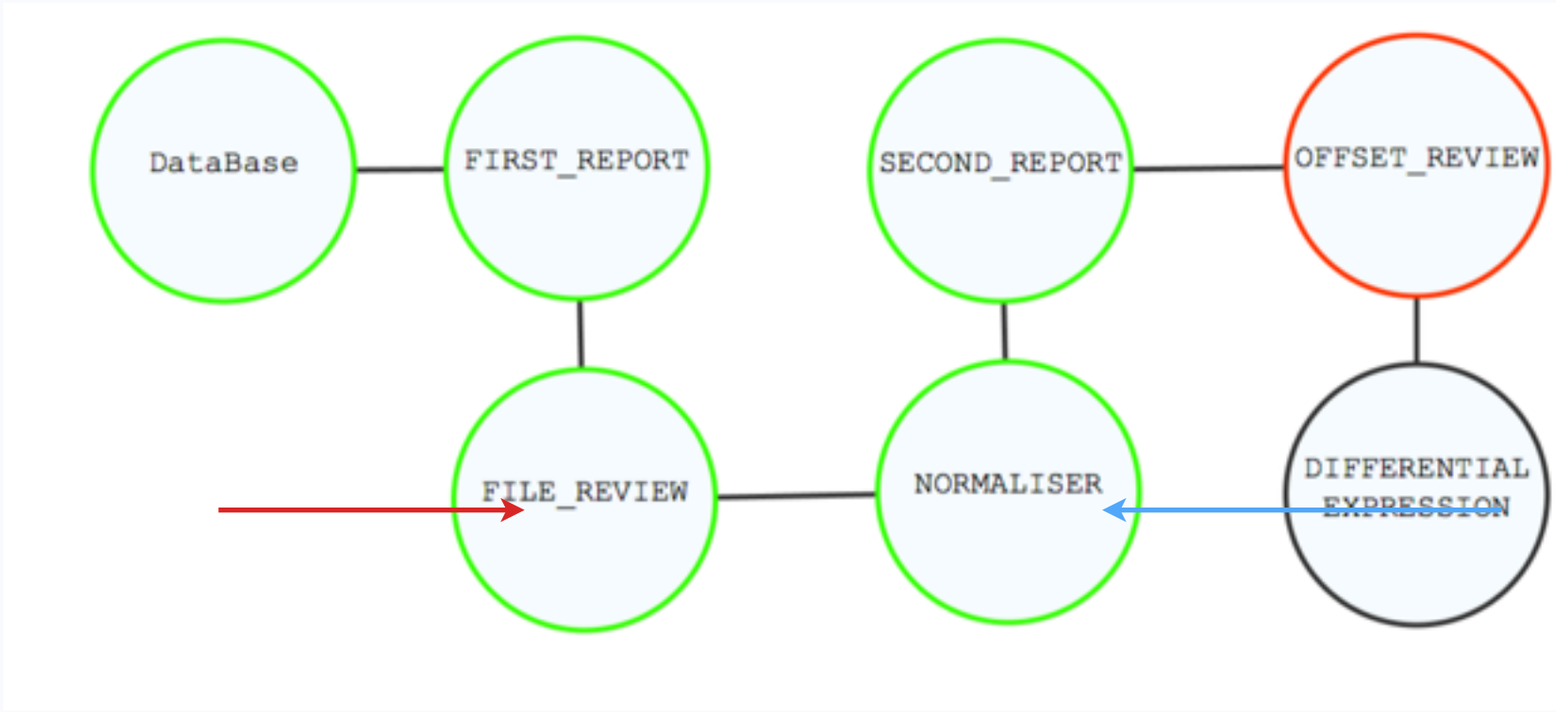
- Investigate the plots found in the first report node of the workflow
- Do the plots indicate comparable sets of replicates and samples?
- How about the mapping quality?

NORMALISATION

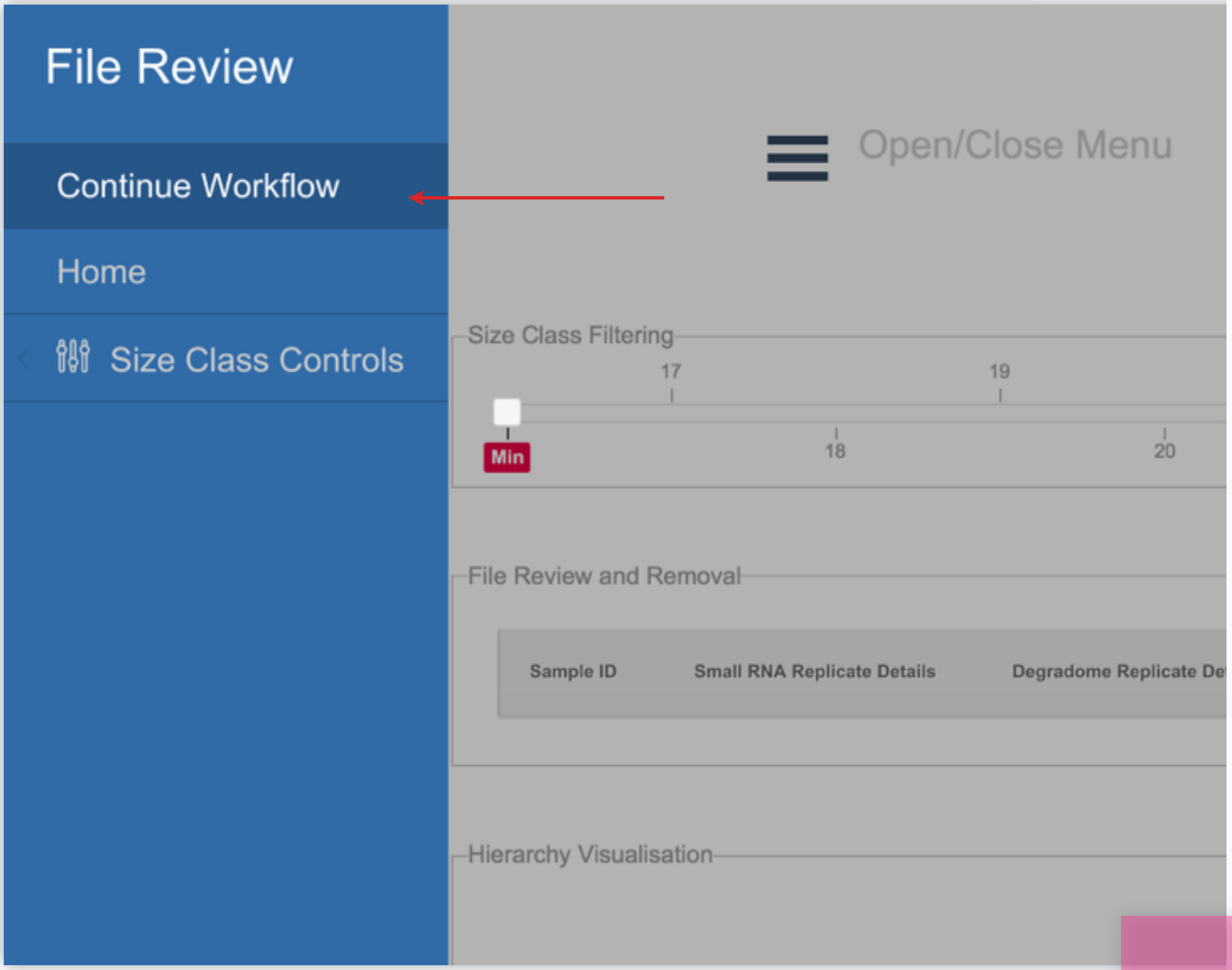
Continue Pipeline

Review Replication Quality

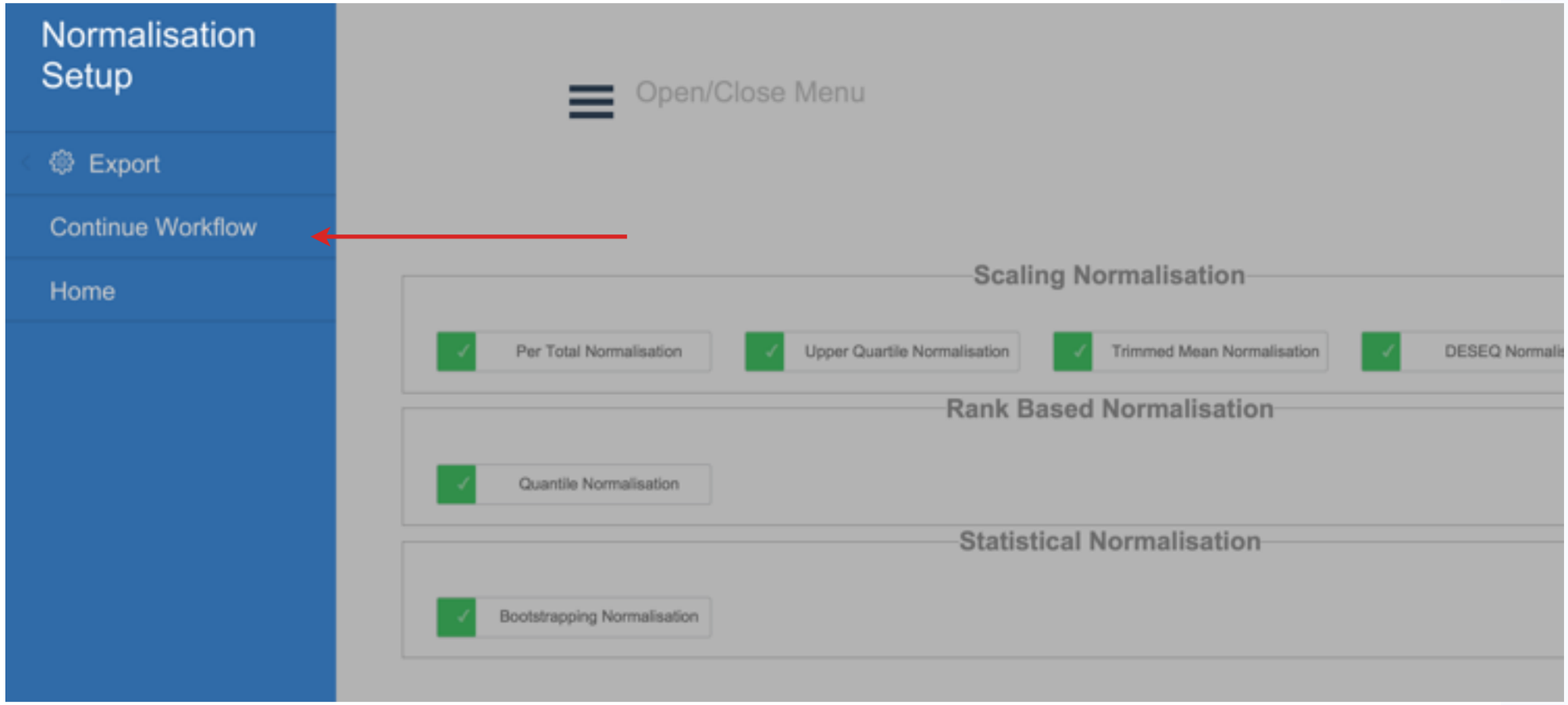
1. Go into the file review node, and hit continue
2. Continue to Normalisation
3. Select All Normalisations and Continue



Continue Pipeline



Continue Pipeline



Normalisation

Why do we do it?

- To attempt to correct issues found during the initial quality control investigation such as highly differentially expressed reads swallowing up the sequencing space

- Sequencing instruments do not have pre-determined sequencing depths

- When sequencing multiple samples expression levels should be adjusted to take this into account

- So if attempting to identify differentially expressed reads over a set of reads the false positive rate is reduced

Normalisation

Three distinct categories

- Scaling
  - RPM (Reads per million*)
  - TMM Trimmed mean of means (edgeR)
  - DESeq 2
- Rank based
  - Quantile
- Statistical
  - Subsampling and Bootstrapping

Normalisation QC

Why look at Quality Control again?

- No single normalisation is suitable for all datasets

- Return to the issues that may have been discovered during QC and investigate the effect made by each normalisation

Switch back to software

Normalisation QC

- Investigate the box plots, fold change box plots and MA plot by normalisation, which normalisation is suitable for this data?
-

Sequence Annotation

Annotation

What did we sequence?

- Several Sources of annotations
  - GFF
  - BLAST
  - RFAM
  - miRBase

Annotated Quality Reports

GFF Annotation

- Align reads to a ‘known’ set of sRNAs
- Annotation information can be used in later analysis for example in differential expression analysis
- Certain Quality Checks should exhibit certain properties
    - miRNA should show a tight cone distribution between replicates
    - t/r RNA are expected to show dispersed distribution between replicates (large amounts of degradation)
- Requires all possible* alignments of an sRNA sequence

Returning to the Report

Annotated QC

- From the plots, is the proportion of non-redundant miRNAs that map to the genome what you may expect from this dataset?
    - If something has gone very wrong with the data, it may be worth looking at this proportion for other classes of sRNA such as t and rRNA
- Have a look at the MA plots for miRNAs, what shape would you expect?

Tea and Coffee

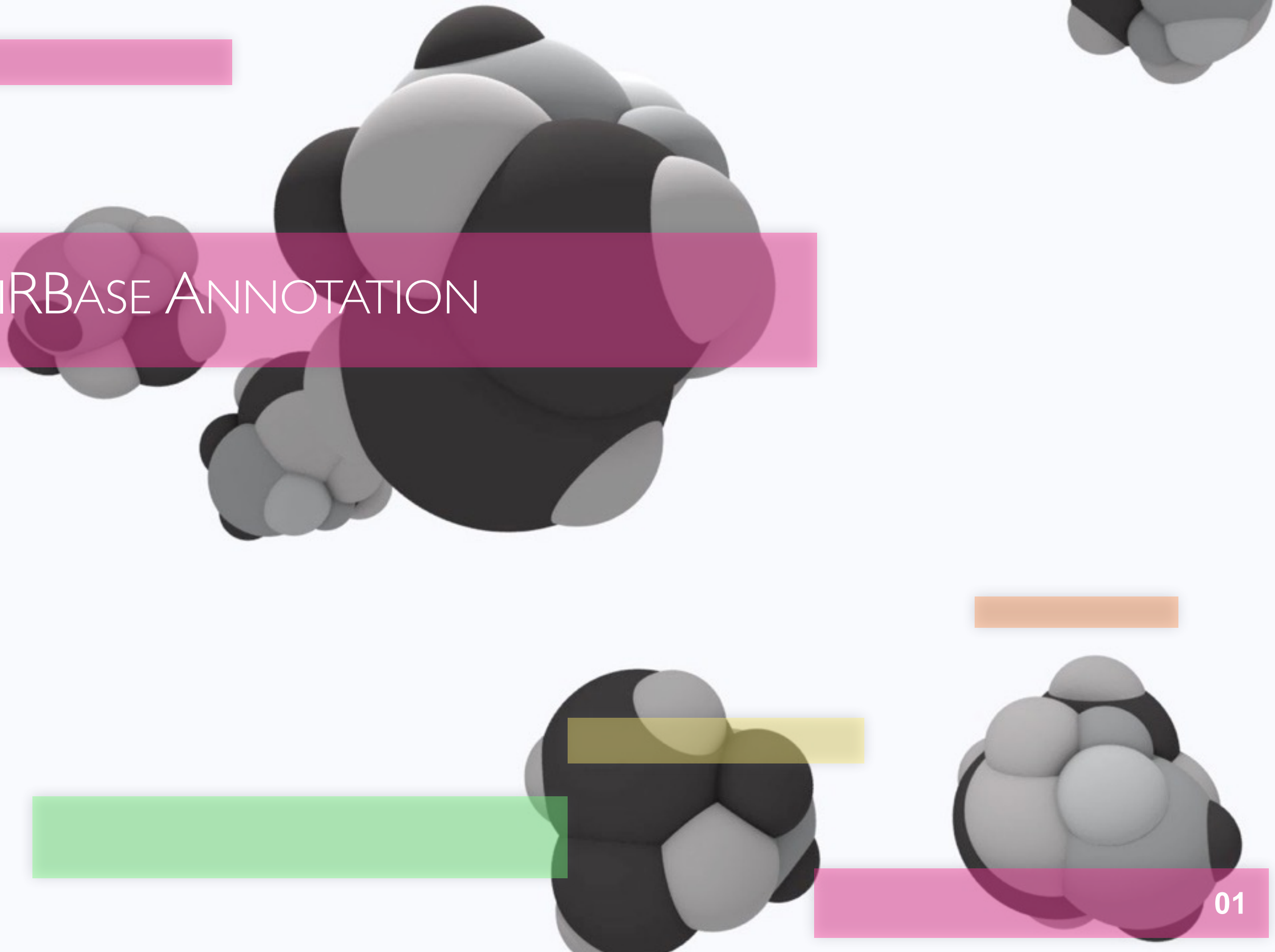
Day 1 Morning Session 2

Day 1 Morning Session 2

miRNA analysis

- Overview of miRNAs
- Annotating known miRNAs from miRBase
- Profiling miRNAs in deep sequencing data
- How to analyse miRNA prediction data
- How to spot false positives

miRBase Annotation



miRBase Annotation

- http://www.mirbase.org/
- Central repository of annotated miRNAs
- Contains all (or the vast majority) of published miRNAs
- Provides a search function
- Provides direct experimental evidence for miRNA expression from NGS data
- Is downloadable and machine parsable

miRBase Annotation

Stem-loop sequence hsa-mir-23a

Accession	MI0000079
Symbol	HGNC:MIR23A
Description	Homo sapiens miR-23a stem-loop
Gene family	MIPF0000027; mir-23
Stem-loop	c c - g g cuuc gg cgg ugggg uuccugg gaug gauuug c cc gcc accuu agggacc uuac cuaaac u a a u g a acug Get sequence
Deep sequencing	89501 reads, 57 experiments reads GGCCGGCUGGGGUUCCUGGGGAUGGGAUUUGCUUCCUGUCACAAAUCACAUUGCCAGGGAUUUCCAACCGACC
Comments	This miRNA was previously named miR-23 [1,2] but is renamed here to avoid confusion with the more recently described miR-23b (MI0000439). Kawasaki and Taira reported that miR-23 regulates the transcriptional repressor Hairy enhancer of split (HES1) [3]. This finding was later retracted after the discovery that the regulated gene was human homolog of ES1 (HES1), whose function is unknown.
Genome context	*Coordinates (GRCh37)*: 19: 13947401-13947473 [-] *Overlapping transcripts*: intergenic
Clustered miRNAs	*< 10kb from hsa-mir-23a* hsa-mir-23a — 19: 13947401-13947473 [-] hsa-mir-27a — 19: 13947254-13947331 [-] hsa-mir-24-2 — 19: 13947101-13947173 [-]

miRBase Annotation

Mature sequence hsa-miR-23a-5p	
Accession	MIMAT0004496
Previous IDs	hsa-miR-23a*
Sequence	9 - gggguuccuggggaugggauuu - 30 Get sequence
Deep sequencing	330 reads, 19 experiments
Evidence	experimental; cloned [6-7]
Predicted targets	DIANA-MICROT: hsa-miR-23a-5p MICRORNA.ORG: hsa-miR-23a-5p MIRDB: hsa-miR-23a-5p PICTAR-VERT: hsa-miR-23a

Mature sequence hsa-miR-23a-3p	
Accession	MIMAT0000078
Previous IDs	hsa-miR-23a
Sequence	45 - aucacauugccagggauuucc - 65 Get sequence
Deep sequencing	89229 reads, 52 experiments
Evidence	experimental; cloned [1,4-7], Northern [1]
Validated targets	TARBASE: hsa-miR-23a-3p
Predicted targets	DIANA-MICROT: hsa-miR-23a-3p MICRORNA.ORG: hsa-miR-23a-3p MIRDB: hsa-miR-23a-3p TARGETSCAN-VERT: hsa-miR-23a PICTAR-VERT: hsa-miR-23a

miRBase Annotation

References

1. PMID:11679670
"Identification of novel genes coding for small expressed RNAs"
Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T
Science. 294:853-858(2001).

2. PMID:11914277
"miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs"
Mourelatos Z, Dostie J, Paushkin S, Sharma A, Charroux B, Abel L, Rappsilber J, Mann M, Dreyfuss G
Genes Dev. 16:720-728(2002).

3. PMID:12808467
"Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells"
Kawasaki H, Taira K
Nature. 423:838-842(2003).

4. PMID:14573789
"Reduced accumulation of specific microRNAs in colorectal neoplasia"
Michael MZ, O' Connor SM, van Holst Pellekaan NG, Young GP, James RJ
Mol Cancer Res. 1:882-891(2003).

5. PMID:15325244
"Altered expression profiles of microRNAs during TPA-induced differentiation of HL-60 cells"
Kasashima K, Nakamura Y, Kozu T
Biochem Biophys Res Commun. 322:403-410(2004).

6. PMID:17604727
"A mammalian microRNA expression atlas based on small RNA library sequencing"
Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foa R, Schliwka J, Fuchs U, Novosel A, Muller RU, Schermer B, Bissels U, Inman J, Phan Q, Chien M
Cell. 129:1401-1414(2007).

7. PMID:17616659
"Patterns of known and novel small RNAs in human cervical cancer"
Lui WO, Pourmand N, Patterson BK, Fire A
Cancer Res. 67:6031-6043(2007).

miRBase Annotation

Deep sequencing reads for stem-loop sequence MI0000079

Stem-loop ID hsa-mir-23a

Reads

hsa-miR-23a-5p	hsa-miR-23a-3p	Count	RPM
.....UGGGGUUCCUGGGGAUGGGAUU.....		29	0.332
.....UGGGGUUCCUGGGGAUGGGAU.....		10	0.0923
.....GGGGUUCCUGGGGAUGGGAUUU.....		215	2.81
.....GGGGUUCCUGGGGAUGGGAUU.....		34	0.316
.....GGGGUUCCUGGGGAUGGGAU.....		10	1.97
	.....AAAUCACAUUGCCAGGGAUUUCCA.....	403	678
	.....AAAUCACAUUGCCAGGGAUUUCC.....	348	708
	.....AAAUCACAUUGCCAGGGAUUUCCAA.....	77	103
	.....AAAUCACAUUGCCAGGGAUUUCCAAC.....	75	106
	.....AAAUCACAUUGCCAGGGAUUUC.....	67	166
	.....AAAUCACAUUGCCAGGGAUUU.....	46	107
	.....AAAUCACAUUGCCAGGGAUU.....	23	79.1
	.....AAUCACAUUGCCAGGGAUUUCCA.....	148	2.33
	.....AAUCACAUUGCCAGGGAUUUCC.....	86	60.5
	.....AAUCACAUUGCCAGGGAUUU.....	69	27.6
	.....AAUCACAUUGCCAGGGAUUUC.....	45	11.8
	.....AAUCACAUUGCCAGGGAUUUCCAA.....	23	0.265
	.....AUCACAUUGCCAGGGAUUUCCA.....	28827	404
	.....AUCACAUUGCCAGGGAUUUCCAA.....	13214	164
	.....AUCACAUUGCCAGGGAUUUCC.....	6167	2.29e+03
	.....AUCACAUUGCCAGGGAUUUC.....	5293	541
	.....AUCACAUUGCCAGGGAUUU.....	4248	783
	.....AUCACAUUGCCAGGGAUU.....	838	249
	.....AUCACAUUGCCAGGGAU.....	129	45.1
	.....AUCACAUUGCCAGGGAUUUCCAAC.....	46	0.479
	.....AUCACAUUGCCAGGGA.....	35	46.3
	.....UCACAUUGCCAGGGAUUUCCAA.....	128	2.04
	.....UCACAUUGCCAGGGAUUUCCA.....	118	4.77
	.....UCACAUUGCCAGGGAUUUCCAAC.....	62	1.76
	.....UCACAUUGCCAGGGAUUUCC.....	48	22.1
	.....ACAUUGCCAGGGAUUUCCA.....	66	0.748
	.....ACAUUGCCAGGGAUUUCCAA.....	48	0.568
	.....ACAUUGCCAGGGAUUUCC.....	21	0.26
	.....ACAUUGCCAGGGAUUU.....	13	0.217
	.....ACAUUGCCAGGGAUUUC.....	13	0.211

GGCCGGCUGGGGUUCCUGGGGAUGGGAUUUGCUUCCUGUCACAAAUCACAUUGCCAGGGAUUUCCAACCGACC
((.(((.((((((((((((.((((.((((((..........)))))).)))).)))))))).)))))).))).)) (-33.20)

miRBase Exercise 1

- Go to miRBase http://www.mirbase.org/ and have a look at the kinds of data that miRBase contains.
- Find the miRBase entry for mir-100 in mouse.
- What other microRNAs are located close to mir-100 in mouse?
- What other microRNAs are related to mir-100?
- In what range of species are mir-100 family members found?

miRBase Exercise 2

- Find mouse mir-196-2
    - Look at the mature sequences that derive from the mir-196-2 hairpin.
    - Have a look at the predicted targets of miR-196-2 sequences
    - Are there are any validated target sites?
    - Click the genomic coordinates of the miRNA to link to the EnsEMBL genome annotation database.
    - Is there any genomic relationship between the miRNA and its validated targets?

EXPECTED PROFILE

Deep sequencing reads for stem-loop sequence MI0000079

Stem-loop ID hsa-mir-23a

Reads

hsa-miR-23a-5p	hsa-miR-23a-3p	Count	RPM
.....UGGGGUUCCUGGGGAUGGGAUU.....		29	0.332
.....UGGGGUUCCUGGGGAUGGGAU.....		10	0.0923
.....GGGGUUCCUGGGGAUGGGAUUU.....		215	2.81
.....GGGGUUCCUGGGGAUGGGAUU.....		34	0.316
.....GGGGUUCCUGGGGAUGGGAU.....		10	1.97
	.....AAAUCACAUUGCCAGGGAUUUCCA.....	403	678
	.....AAAUCACAUUGCCAGGGAUUUCC.....	348	708
	.....AAAUCACAUUGCCAGGGAUUUCCAA.....	77	103
	.....AAAUCACAUUGCCAGGGAUUUCCAAC.....	75	106
	.....AAAUCACAUUGCCAGGGAUUUC.....	67	166
	.....AAAUCACAUUGCCAGGGAUUU.....	46	107
	.....AAAUCACAUUGCCAGGGAUU.....	23	79.1
	.....AAUCACAUUGCCAGGGAUUUCCA.....	148	2.33
	.....AAUCACAUUGCCAGGGAUUUCC.....	86	60.5
	.....AAUCACAUUGCCAGGGAUUU.....	69	27.6
	.....AAUCACAUUGCCAGGGAUUUC.....	45	11.8
	.....AAUCACAUUGCCAGGGAUUUCCAA.....	23	0.265
	.....AUCACAUUGCCAGGGAUUUCCA.....	28827	404
	.....AUCACAUUGCCAGGGAUUUCCAA.....	13214	164
	.....AUCACAUUGCCAGGGAUUUCC.....	6167	2.29e+03
	.....AUCACAUUGCCAGGGAUUUC.....	5293	541
	.....AUCACAUUGCCAGGGAUUU.....	4248	783
	.....AUCACAUUGCCAGGGAUU.....	838	249
	.....AUCACAUUGCCAGGGAU.....	129	45.1
	.....AUCACAUUGCCAGGGAUUUCCAAC.....	46	0.479
	.....AUCACAUUGCCAGGGA.....	35	46.3
	.....UCACAUUGCCAGGGAUUUCCAA.....	128	2.04
	.....UCACAUUGCCAGGGAUUUCCA.....	118	4.77
	.....UCACAUUGCCAGGGAUUUCCAAC.....	62	1.76
	.....UCACAUUGCCAGGGAUUUCC.....	48	22.1
	.....ACAUUGCCAGGGAUUUCCA.....	66	0.748
	.....ACAUUGCCAGGGAUUUCCAA.....	48	0.568
	.....ACAUUGCCAGGGAUUUCC.....	21	0.26
	.....ACAUUGCCAGGGAUUU.....	13	0.217
	.....ACAUUGCCAGGGAUUUC.....	13	0.211

GGCCGGCUGGGGUUCCUGGGGAUGGGAUUUGCUUCCUGUCACAAAUCACAUUGCCAGGGAUUUCCAACCGACC
((.(((.((((((((((((.((((.((((((..........)))))).)))).)))))))).)))))).))).)) (-33.20)

Potential for Issues

Deep sequencing reads for mature sequence MIMAT0014139

Stem-loop ID: ath-MIR2935

Mature ID: ath-miR2935

Reads (ath-miR2935)

Read	Count
CUUGUCUCUACUUAAGGCUUGGAA	2
UGUCUCUACUUAAGGCUUGG	2
UCUCUACUUAAGGCUUGGAAU	2
CUACUUAAGGCUUGGAAUCACACG	5
ACUUAAGGCUUGGAAUCACACGGU	7
GCUUGGAAUCACACGGUCGUCAUU	2
CUUGGAAUCACACGGUCGUCAUUC	7
UUGGAAUCACACGGUCGUC	2
UGGAAUCACACGGUCGUCAUUCAU	2
AAUCACACGGUCGUCAUUCAUCUU	3
UCACACGGUCGUCAUUCAUCUUUU	4
ACACGGUCGUCAUUCAUCUUUU	19
ACACGGUCGUCAUUCAUCUUUUGA	2
CACGGUCGUCAUUCAUCUUUUGAU	10
ACGGUCGUCAUUCAUCUUUUGAU	5
CGGUCGUCAUUCAUCUUUUGAUU	2
CUUUUGAUUUGGUAGUCUGGCGAA	2
GAUUUGGUAGUCUGGCGAAGUUAG	7
AUUUGGUAGUCUGGCGAAGUUAGA	15
UAGUCUGGCGAAGUUAGAAAUCAA	2
AUAUCCUUCUAAAUGGGUGGUAGC	2
CUUCUAAAUGGGUGGUAGCCUA	5
CUUCUAAAUGGGUGGUAGCCU	2
UUCUAAAUGGGUGGUAGCCUAGU	2
UCUAAAUGGGUGGUAGCCUAGUGA	2
CUAAAUGGGUGGUAGCCUAGUGAU	7
AAAUGGGUGGUAGCCUAGUGAU	2
UGGGUGGUAGCCUAGUGAUAU	2
GUGGUAGCCUAGUGAUAUUAAGAG	2
GUAGCCUAGUGAUAUUAAGAGUGA	2
CCUAGUGAUAUUAAGAGUGAGUAU	4
AGUGAUAUUAAGAGUGAGUAUUCA	7

```
CUUGUCUCUACUUAAGGCUUGGAAUCACACGGUCGUCAUUCAUCUUUUGAUUUGGUAGUCUGGCGAAGUUAGAAAUCAAUUUUGAGUUGGUGGAUAUCCUUCUAAAUGGGUGGUAGCCUAGUGAUAUUAAGAGUGAGUAUUCAAG
((((....((((((...(((((.(((((..(((..((((((((.....((...((..(((((.(((..((((((.....)))))).))).)))))..)).))...))))))))..)))..))))).)))))..))))))..)))) (-39.10)
```

miRBase Exercise 3

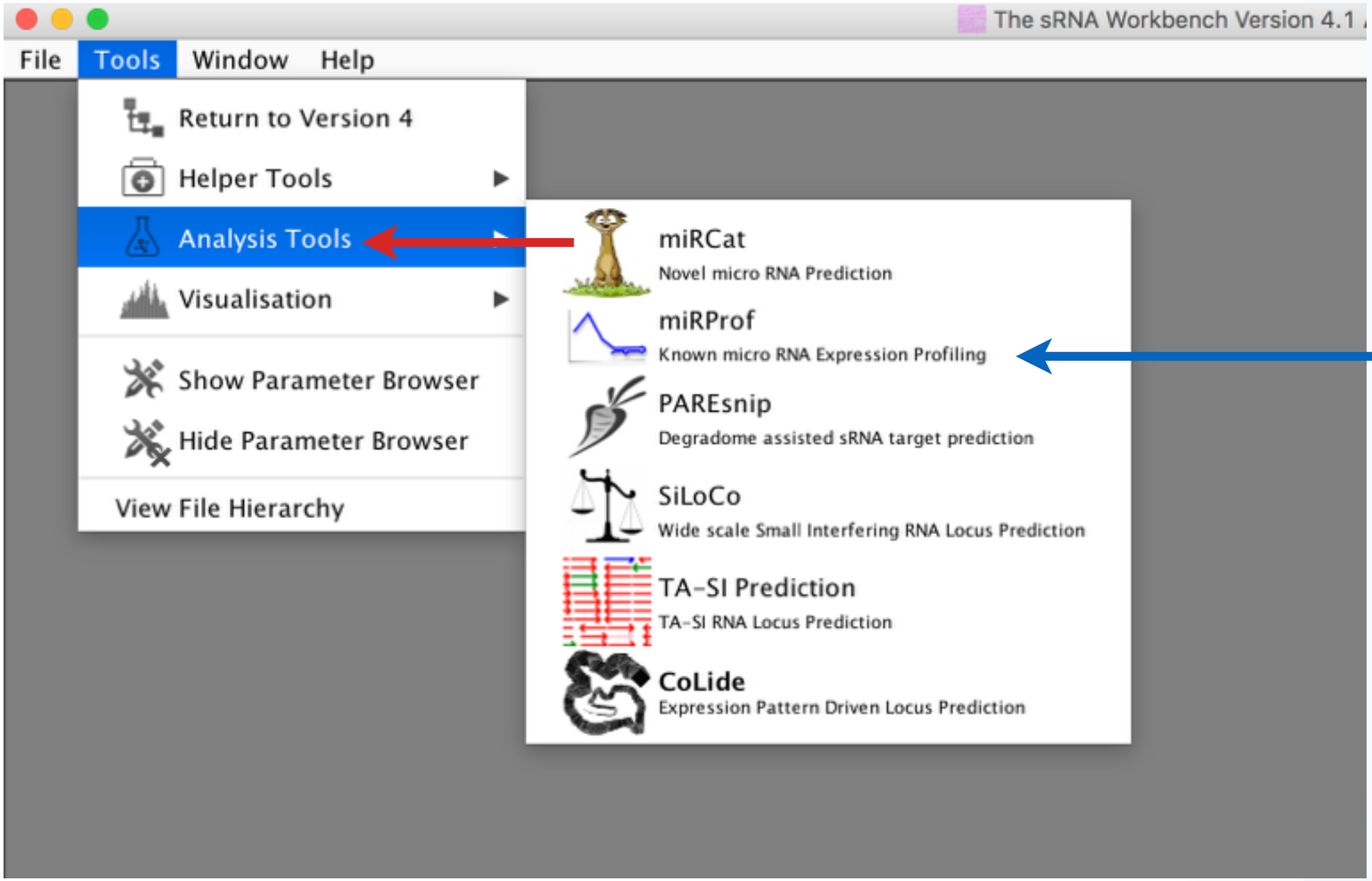
- Find the miRBase entry for miR-317 in Drosophila melanogaster.
    - Look at the pattern of reads that map to the locus from deep-sequencing experiments. Do the read counts support the mature sequence annotations?
    - Which mature sequence is most abundant?
- Now search for the microRNA osa-MIR439c.
    - Explore the expression pattern of the mature miRNAs given by the NGS data (how many reads? what is the most abundant read? etc.)
    - Considering the miRNA annotation guidelines, what can you say about this miRBase entry?
    - Look at other miRNAs in the same gene family - are they likely to be real miRNAs?
    - How many other microRNA were discovered in the same study?

miRBase Exercise 4

- Use the browse and search pages for this exercise.
    - From the browse page, how many microRNAs are annotated in Drosophila melanogaster?
    - From the "tissue expression" section of the search page, find the list of all deep sequencing experiments in D. melanogaster.
- Compare the miRNAs found in a head dataset with an embryonic dataset.
- Identify miRNAs that are highly expressed in the head, but not in the embryo.

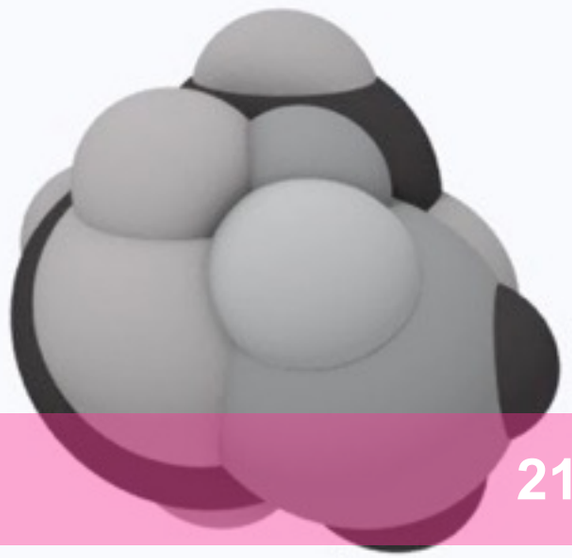
miRProf Exercise

Load the miRProf Tool

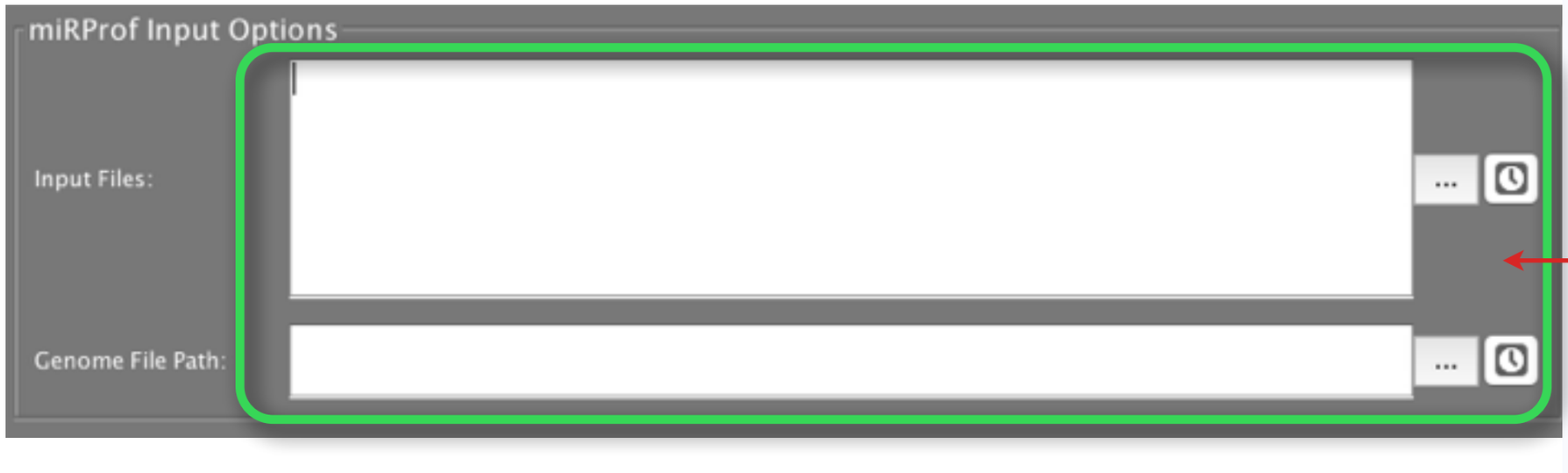


miRProf Exercise 1

- Visit the UCSC Genome Browser website
  - http://genome.ucsc.edu/
- Locate Chromosome 1 of the human genome and download it, then extract it to any location



miRProf Exercise I



Add the files found in ~/Desktop/data/FASTA/

1.SRR873382_na_sampled_0.1.fasta

2.SRR873384_na_sampled_0.1.fasta

and the single chromosome file you downloaded

miRProf Exercise

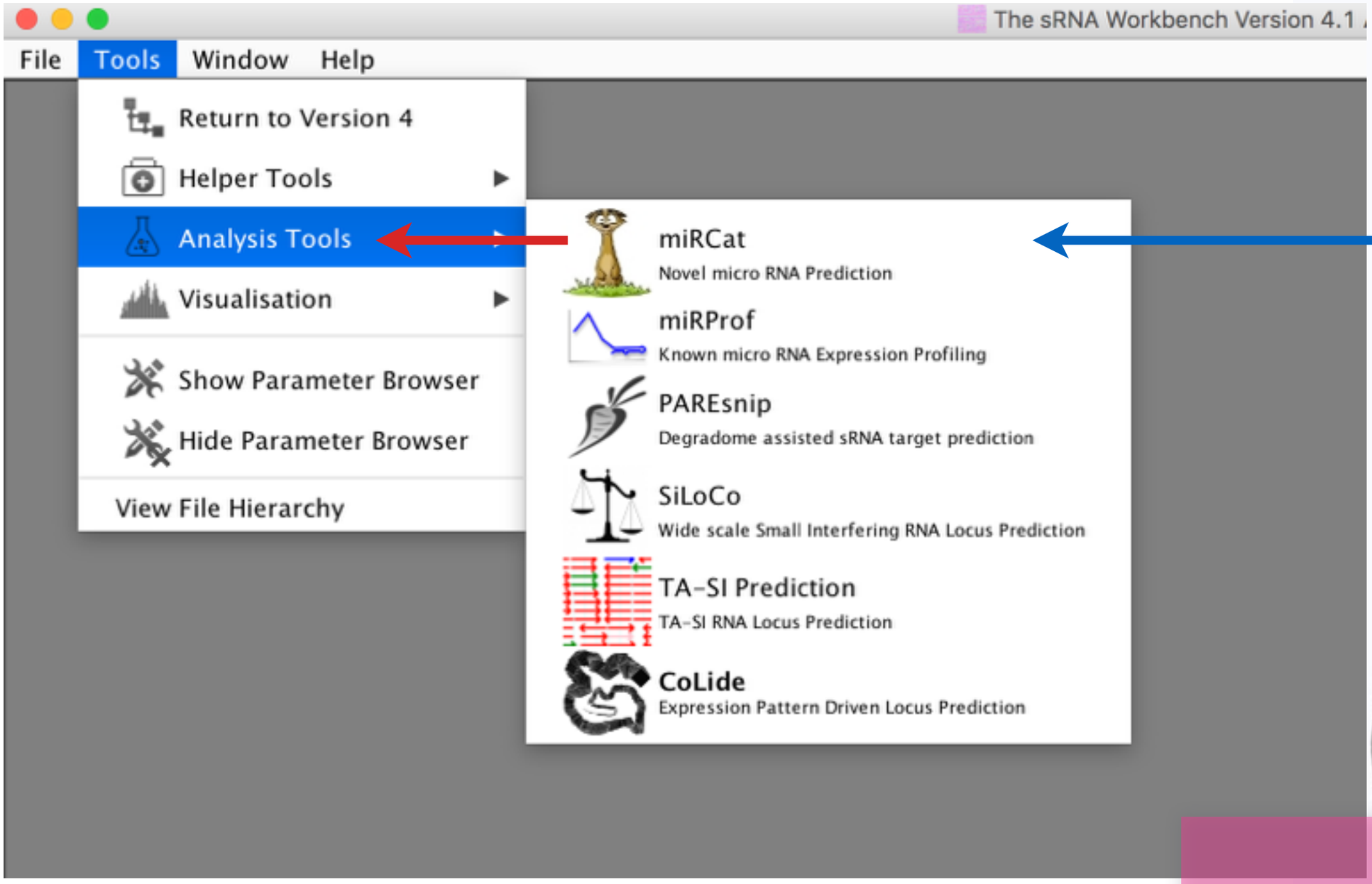
- Use the default parameters and run the program
- Export the results to csv and open them in the spreadsheet editor
- Calculate the simple fold change for each miRNA using the total counts
  - (Expression Sample 2 - Expression Sample 1)/Expression Sample 1)
- Are there differences in the miRNA Expression for each sample

miRCat Exercise

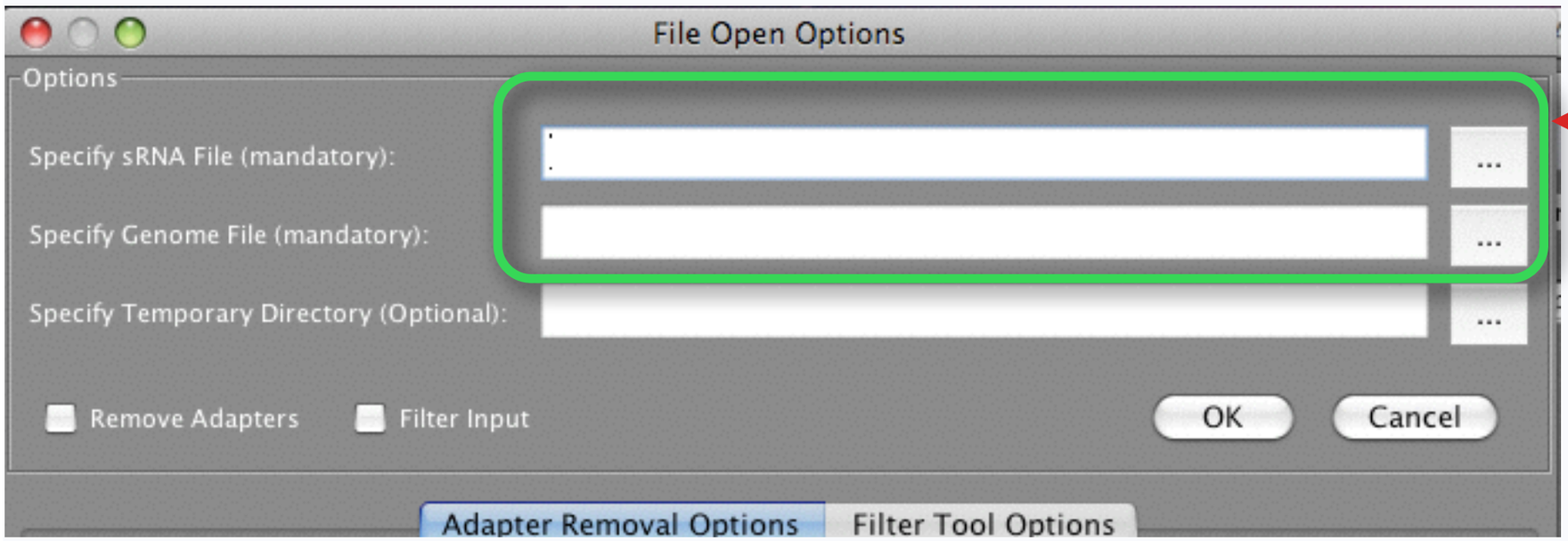
- Trim the adapters from the reads.fa file found in ~/Desktop/miRDeep2/tutorial_dir

miRCat Exercise

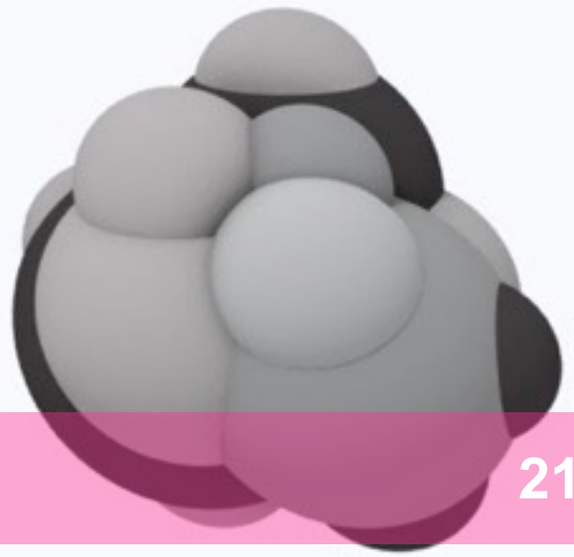
Load the miRCat Tool



miRCat Exercise



- Start a new project from the menu
- load the trimmed file you created into the sRNA input box
- and the genome file found in the original location called cel_cluster.fa
- Use the default animal parameters and run the program



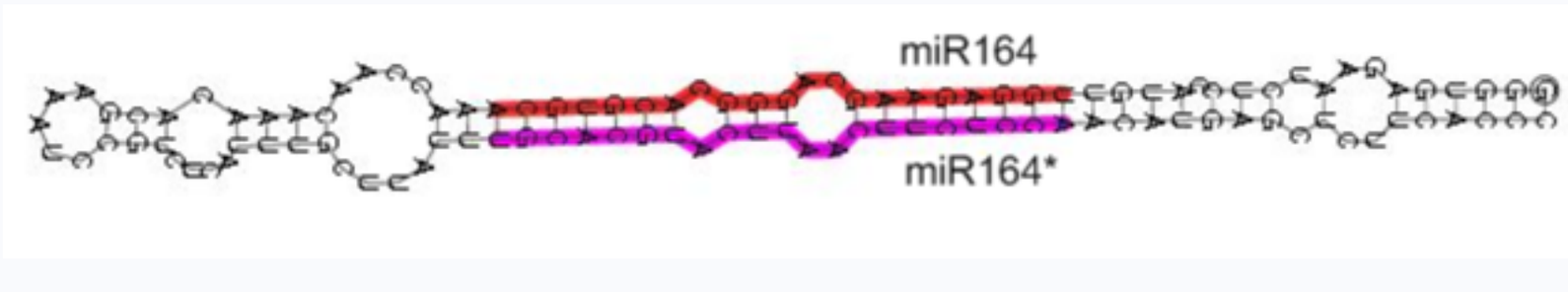
miRCat Exercise

Options

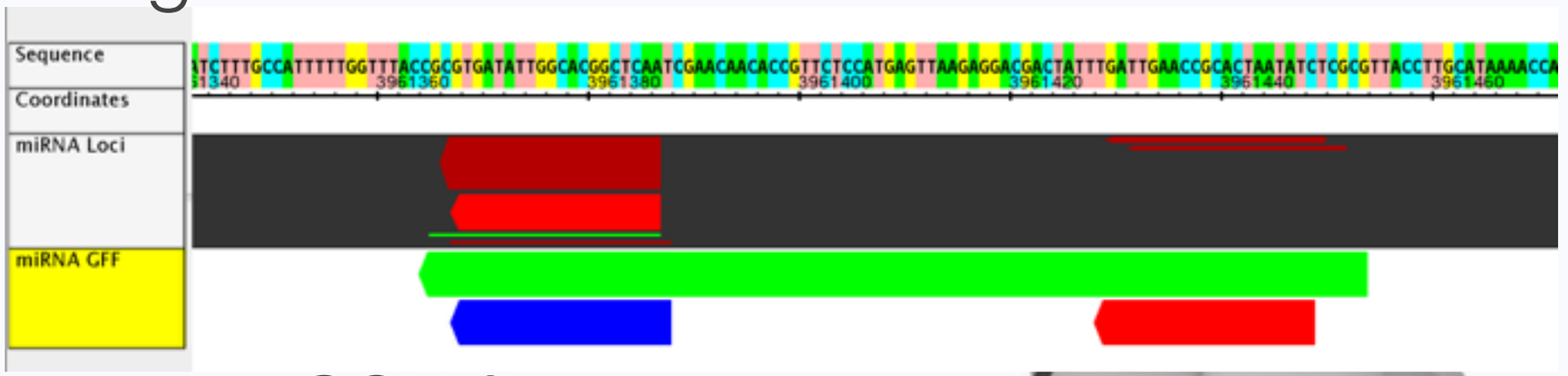
Output table

miRCat Exercise 1

- Pick one and generate the pre cursor plot for it (right click menu)

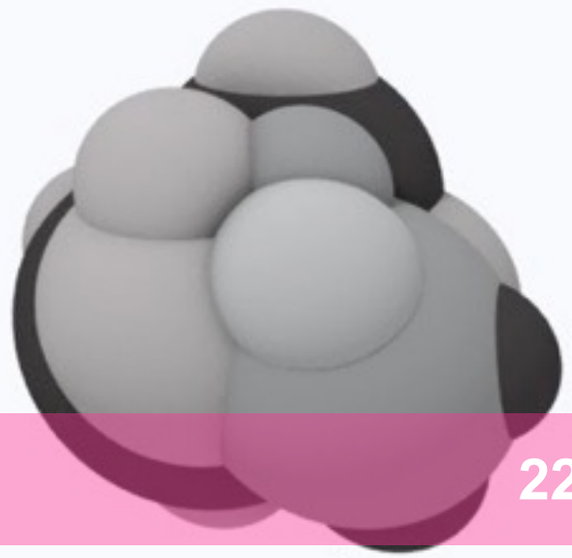


- Look at the locus using the VisSR tool (right click menu)



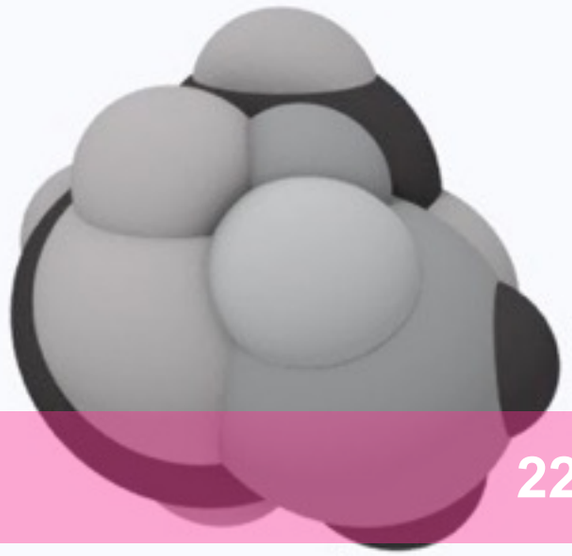
- Export all of the results to CSV file, we will use them later

miRDeep 2 Exercise 1

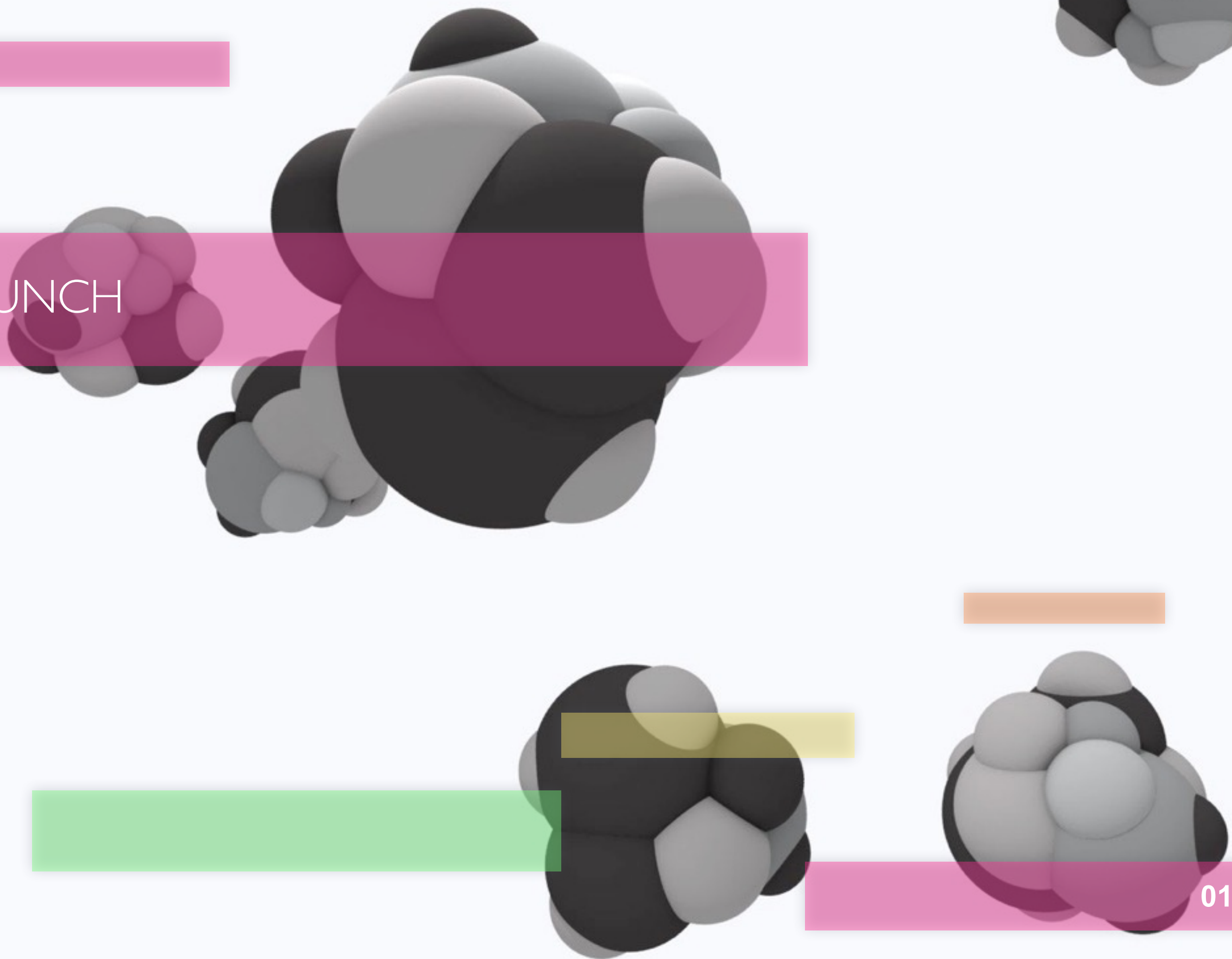


miRCat Exercise 2

- Examine the results from both the miRCat and miRDeep runs
- Are there any differences between the two?
- What can we infer from this?



LUNCH



Day 1 Afternoon Session 1

Differential Expression

DE EXERCISE 1

- From the second report, with a fill dataset we would normally select the normalisation that best suits the data

- Double click in the Offset Review Node and choose the bootstrapping normalisation from the menu

- Select continue workflow

DE EXERCISE 1

- Look at the KL graph produced from the dataset
- Does the curve line up with the global minimum for average abundances across the dataset?

DE Exercise 1

- Open the Differential Expression node. Compare Sample 1 with Sample 2 by dragging their references into the reference and observed columns of the interface
- Use the options menu to set the fold change cut off to 0.5
- Investigate the result set