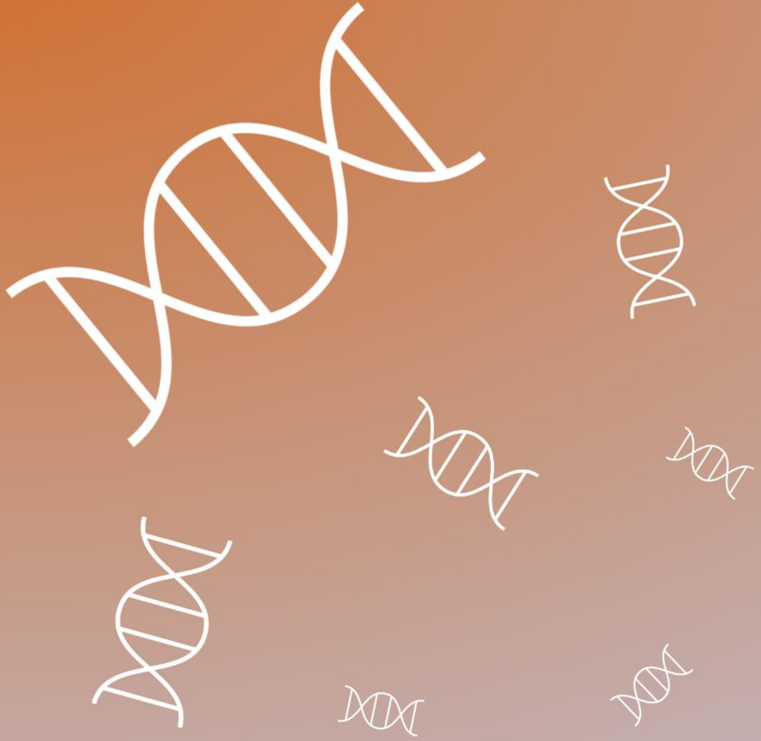
Home Tutorials Forum Databases ExpressAnalystR Updates Contact



Tutorial VI: Raw data processing

ExpressAnalyst

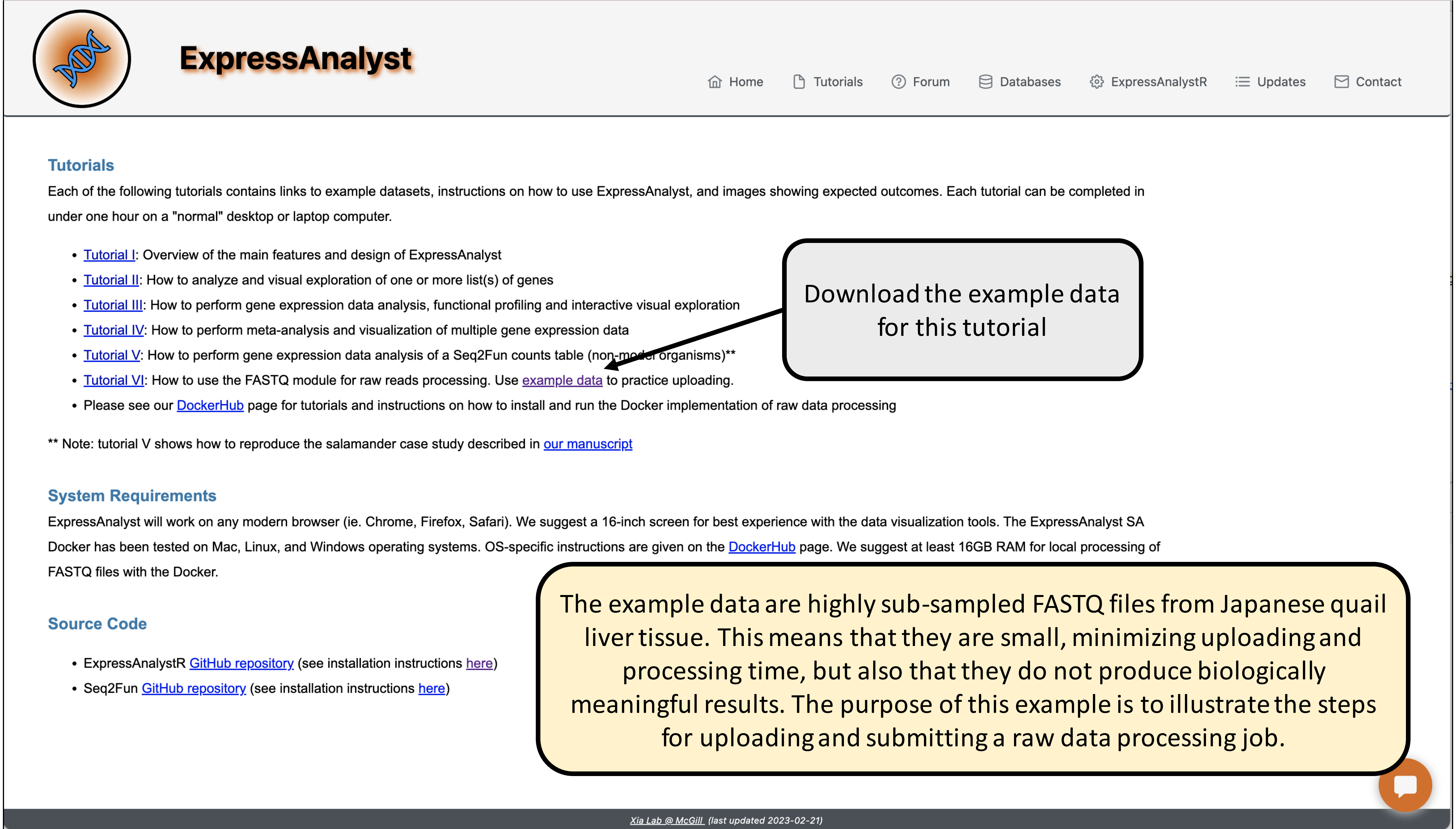
-- a unified platform for gene expression data analysis

Start Here

Xia Lab @ McGill (last updated 2023-02-21)

Overview

- In this tutorial, we will:
  - Learn how to upload files to ExpressAnalyst raw data processing module
  - Understand the differences between Seq2Fun and Kallisto
  - Learn about the results files produced after reads quantification



Tutorials

Each of the following tutorials contains links to example datasets, instructions on how to use ExpressAnalyst, and images showing expected outcomes. Each tutorial can be completed in under one hour on a "normal" desktop or laptop computer.

- Tutorial I: Overview of the main features and design of ExpressAnalyst
- Tutorial II: How to analyze and visual exploration of one or more list(s) of genes
- Tutorial III: How to perform gene expression data analysis, functional profiling and interactive visual exploration
- Tutorial IV: How to perform meta-analysis and visualization of multiple gene expression data
- Tutorial V: How to perform gene expression data analysis of a Seq2Fun counts table (non-model organisms)**
- Tutorial VI: How to use the FASTQ module for raw reads processing. Use example data to practice uploading.
- Please see our DockerHub page for tutorials and instructions on how to install and run the Docker implementation of raw data processing

** Note: tutorial V shows how to reproduce the salamander case study described in our manuscript

System Requirements

ExpressAnalyst will work on any modern browser (ie. Chrome, Firefox, Safari). We suggest a 16-inch screen for best experience with the data visualization tools. The ExpressAnalyst SA Docker has been tested on Mac, Linux, and Windows operating systems. OS-specific instructions are given on the DockerHub page. We suggest at least 16GB RAM for local processing of FASTQ files with the Docker.

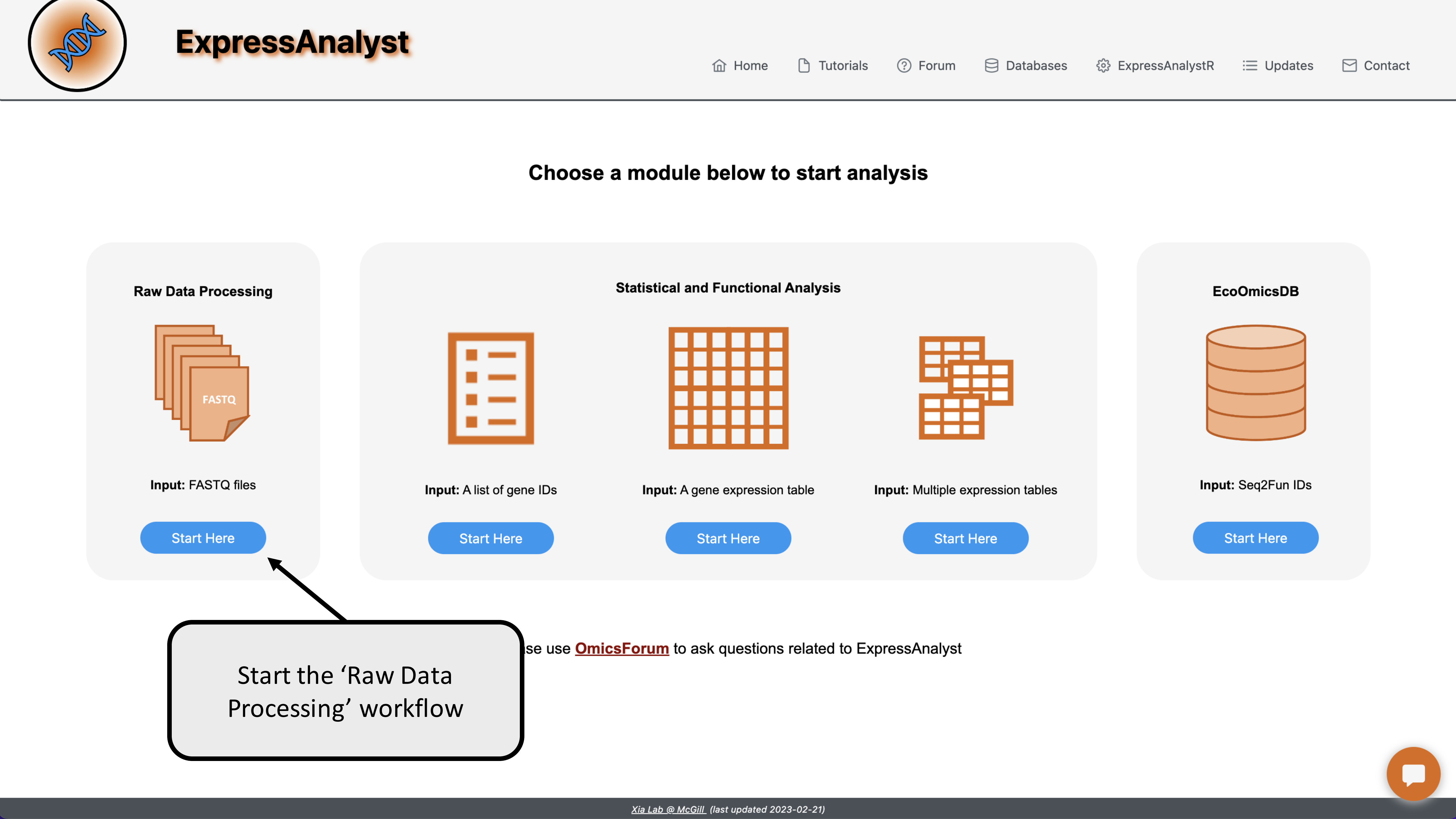
Source Code

- ExpressAnalystR GitHub repository (see installation instructions here)
- Seq2Fun GitHub repository (see installation instructions here)

Xia Lab @ McGill (last updated 2023-02-21)

Download the example data for this tutorial

The example data are highly sub-sampled FASTQ files from Japanese quail liver tissue. This means that they are small, minimizing uploading and processing time, but also that they do not produce biologically meaningful results. The purpose of this example is to illustrate the steps for uploading and submitting a raw data processing job.



Choose a module below to start analysis

Raw Data Processing

Input: FASTQ files

Statistical and Functional Analysis

Input: A list of gene IDs

Input: A gene expression table

Input: Multiple expression tables

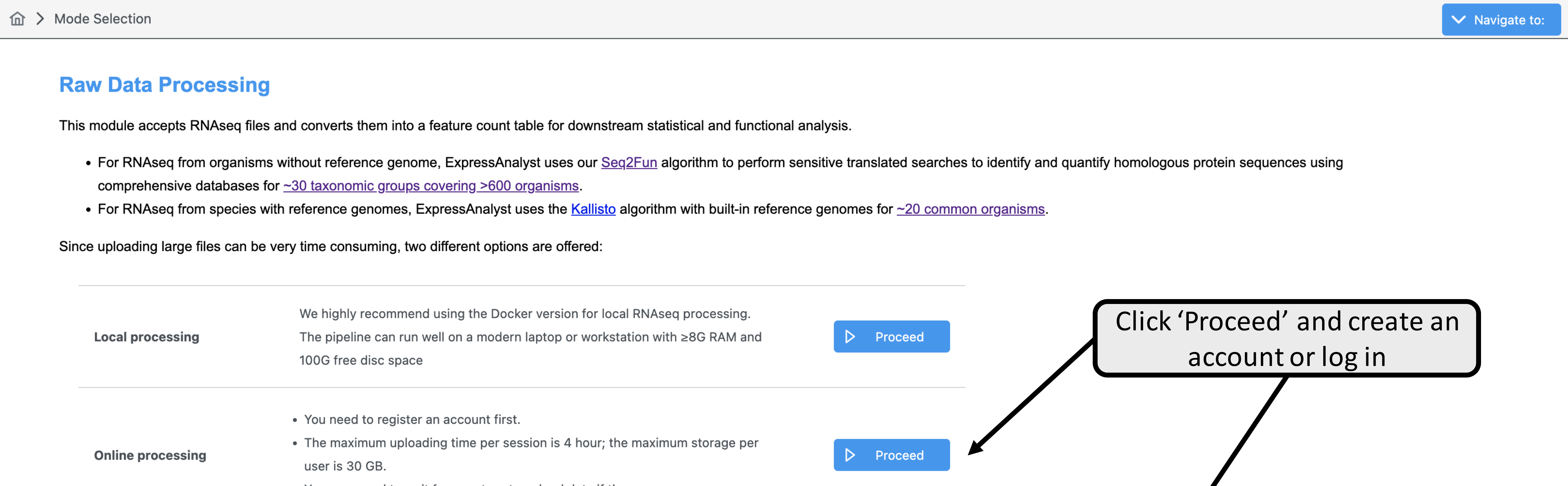
EcoOmicsDB

Input: Seq2Fun IDs

se use OmicsForum to ask questions related to ExpressAnalyst

Xia Lab @ McGill *(last updated 2023-02-21)*

Start the 'Raw Data Processing' workflow



Raw Data Processing

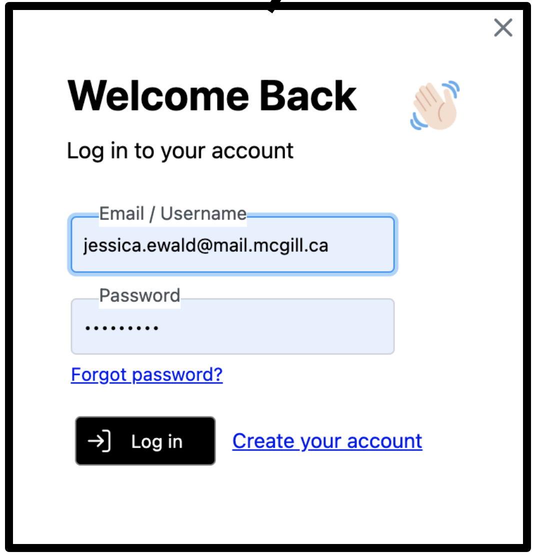
This module accepts RNAseq files and converts them into a feature count table for downstream statistical and functional analysis.

- For RNAseq from organisms without reference genome, ExpressAnalyst uses our Seq2Fun algorithm to perform sensitive translated searches to identify and quantify homologous protein sequences using comprehensive databases for ~30 taxonomic groups covering >600 organisms.
- For RNAseq from species with reference genomes, ExpressAnalyst uses the Kallisto algorithm with built-in reference genomes for ~20 common organisms.

Since uploading large files can be very time consuming, two different options are offered:

Local processing	We highly recommend using the Docker version for local RNAseq processing. The pipeline can run well on a modern laptop or workstation with ≥8G RAM and 100G free disc space	▷ Proceed
Online processing	• You need to register an account first. • The maximum uploading time per session is 4 hour; the maximum storage per user is 30 GB. • You may need to wait for your turn to upload data if there are many users.	▷ Proceed

Click 'Proceed' and create an account or log in



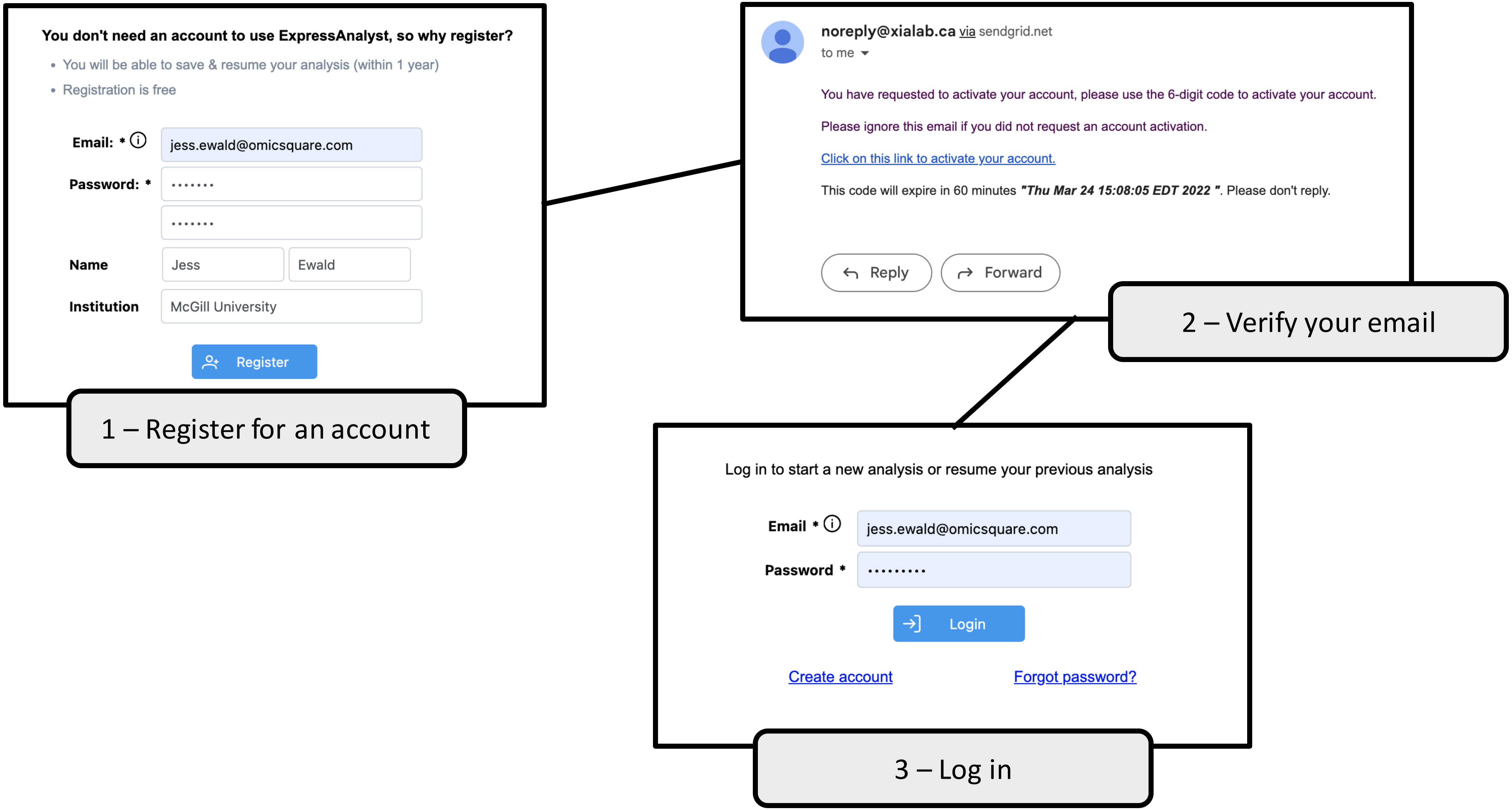
Welcome Back

Log in to your account

Email / Username: jessica.ewald@mail.mcgill.ca

Password: ·········

'Online processing' allows you to upload files to the Xia Lab server for remote processing. File upload can take a long time, and dataset size is limited to manage tool usage across many users. The alternative is to process your data with the 'Local processing' option, which requires installation of additional software. This tutorial is on the 'Online processing' option.



You don't need an account to use ExpressAnalyst, so why register?

- You will be able to save & resume your analysis (within 1 year)
- Registration is free

Email: * jess.ewald@omicsquare.com

Password: * •••••••

•••••••

Name: Jess Ewald

Institution: McGill University

1 – Register for an account

noreply@xialab.ca via sendgrid.net

to me

You have requested to activate your account, please use the 6-digit code to activate your account.

Please ignore this email if you did not request an account activation.

Click on this link to activate your account.

This code will expire in 60 minutes ***"Thu Mar 24 15:08:05 EDT 2022 "***. Please don't reply.

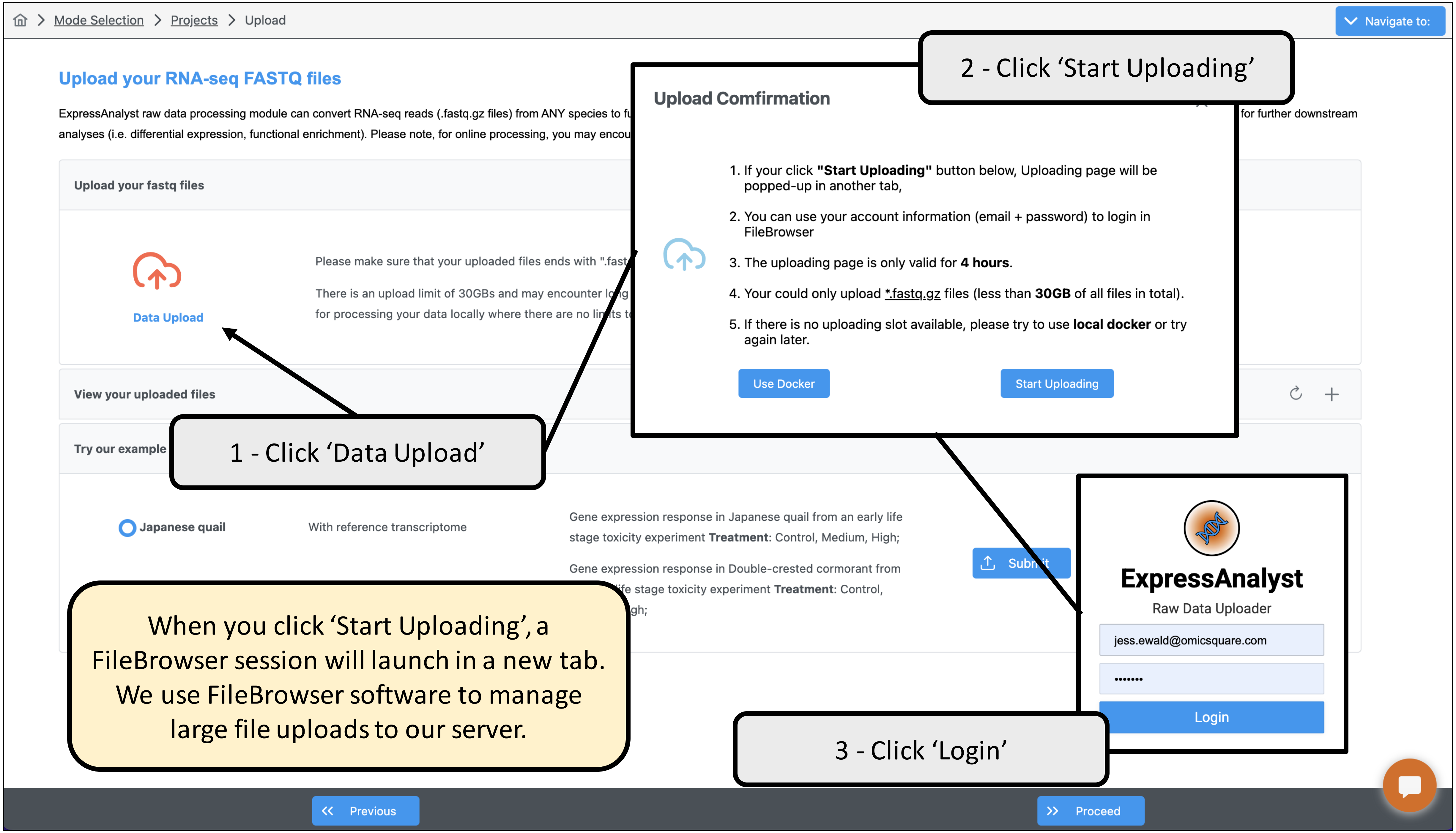
2 – Verify your email

Log in to start a new analysis or resume your previous analysis

Email * jess.ewald@omicsquare.com

Password * •••••••••

3 – Log in



Upload your RNA-seq FASTQ files

ExpressAnalyst raw data processing module can convert RNA-seq reads (.fastq.gz files) from ANY species to fu … for further downstream analyses (i.e. differential expression, functional enrichment). Please note, for online processing, you may encou

Upload your fastq files

Please make sure that your uploaded files ends with ".fast

There is an upload limit of 30GBs and may encounter long … for processing your data locally where there are no limits t

View your uploaded files

Try our example

○ Japanese quail — With reference transcriptome — Gene expression response in Japanese quail from an early life stage toxicity experiment **Treatment**: Control, Medium, High;

Gene expression response in Double-crested cormorant from … life stage toxicity experiment **Treatment**: Control, … gh;

Upload Comfirmation

1. If your click **"Start Uploading"** button below, Uploading page will be popped-up in another tab,
2. You can use your account information (email + password) to login in FileBrowser
3. The uploading page is only valid for **4 hours**.
4. Your could only upload *.fastq.gz files (less than **30GB** of all files in total).
5. If there is no uploading slot available, please try to use **local docker** or try again later.

ExpressAnalyst

Raw Data Uploader

jess.ewald@omicsquare.com

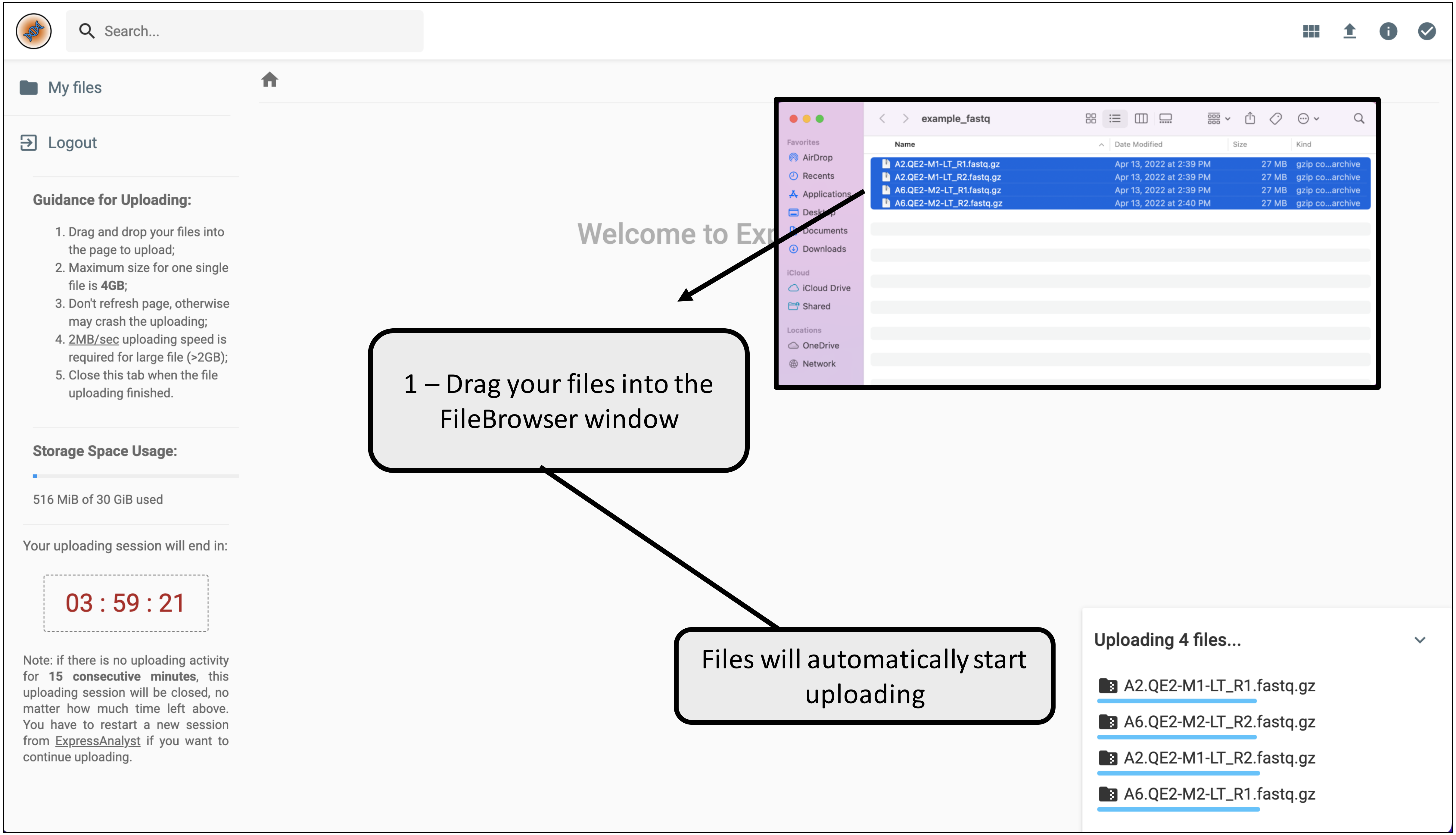
•••••••

1 - Click ‘Data Upload’

2 - Click ‘Start Uploading’

3 - Click ‘Login’

When you click ‘Start Uploading’, a FileBrowser session will launch in a new tab. We use FileBrowser software to manage large file uploads to our server.



Guidance for Uploading:

1. Drag and drop your files into the page to upload;
2. Maximum size for one single file is **4GB**;
3. Don't refresh page, otherwise may crash the uploading;
4. 2MB/sec uploading speed is required for large file (>2GB);
5. Close this tab when the file uploading finished.

Storage Space Usage:

516 MiB of 30 GiB used

Your uploading session will end in:

03 : 59 : 21

Note: if there is no uploading activity for **15 consecutive minutes**, this uploading session will be closed, no matter how much time left above. You have to restart a new session from ExpressAnalyst if you want to continue uploading.

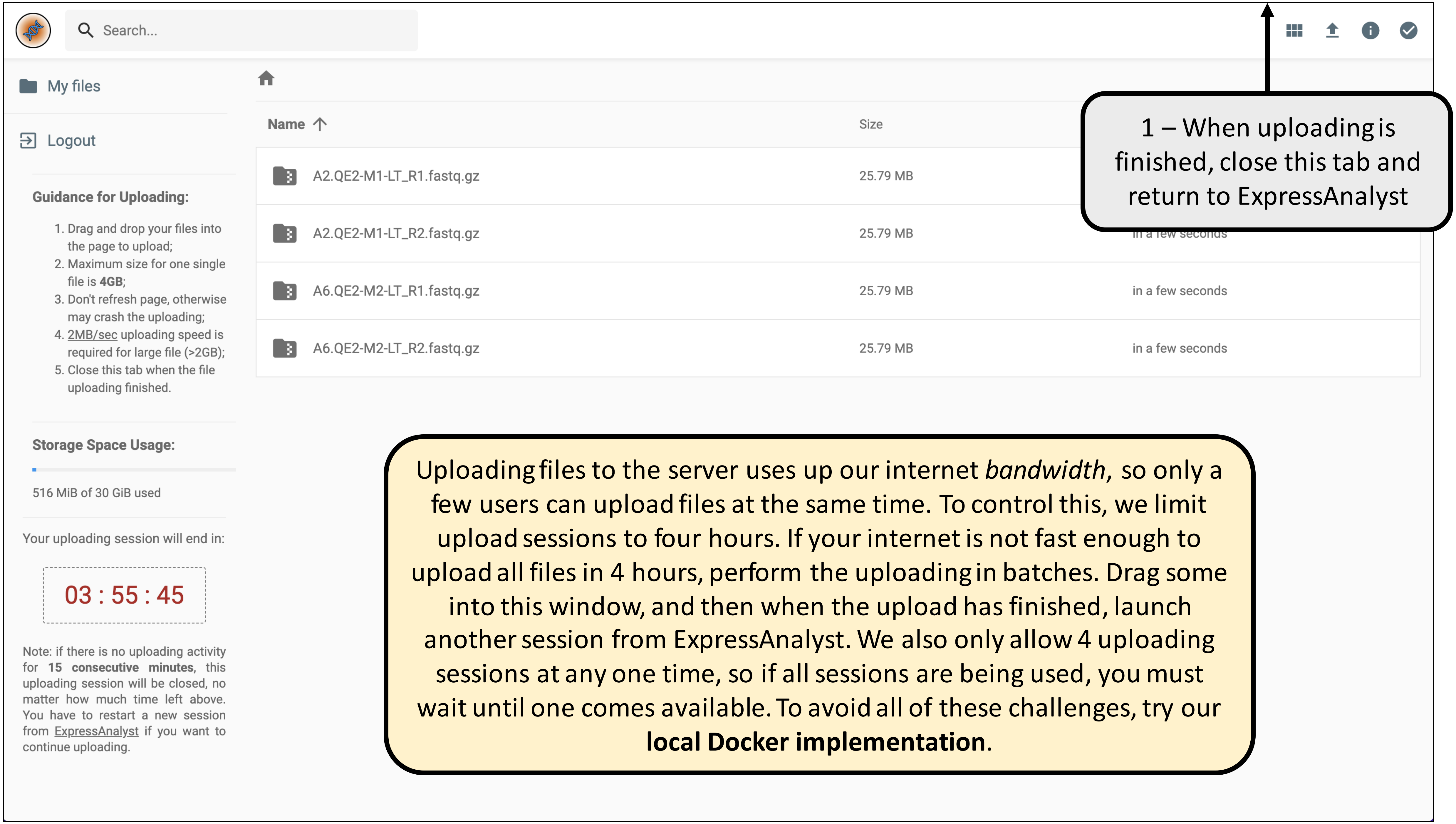
Name	Date Modified	Size	Kind
A2.QE2-M1-LT_R1.fastq.gz	Apr 13, 2022 at 2:39 PM	27 MB	gzip co...archive
A2.QE2-M1-LT_R2.fastq.gz	Apr 13, 2022 at 2:39 PM	27 MB	gzip co...archive
A6.QE2-M2-LT_R1.fastq.gz	Apr 13, 2022 at 2:39 PM	27 MB	gzip co...archive
A6.QE2-M2-LT_R2.fastq.gz	Apr 13, 2022 at 2:40 PM	27 MB	gzip co...archive

1 – Drag your files into the FileBrowser window

Files will automatically start uploading

Uploading 4 files...

- A2.QE2-M1-LT_R1.fastq.gz
- A6.QE2-M2-LT_R2.fastq.gz
- A2.QE2-M1-LT_R2.fastq.gz
- A6.QE2-M2-LT_R1.fastq.gz



Search...

My files

Logout

Guidance for Uploading:

1. Drag and drop your files into the page to upload;
2. Maximum size for one single file is **4GB**;
3. Don't refresh page, otherwise may crash the uploading;
4. 2MB/sec uploading speed is required for large file (>2GB);
5. Close this tab when the file uploading finished.

Storage Space Usage:

516 MiB of 30 GiB used

Your uploading session will end in:

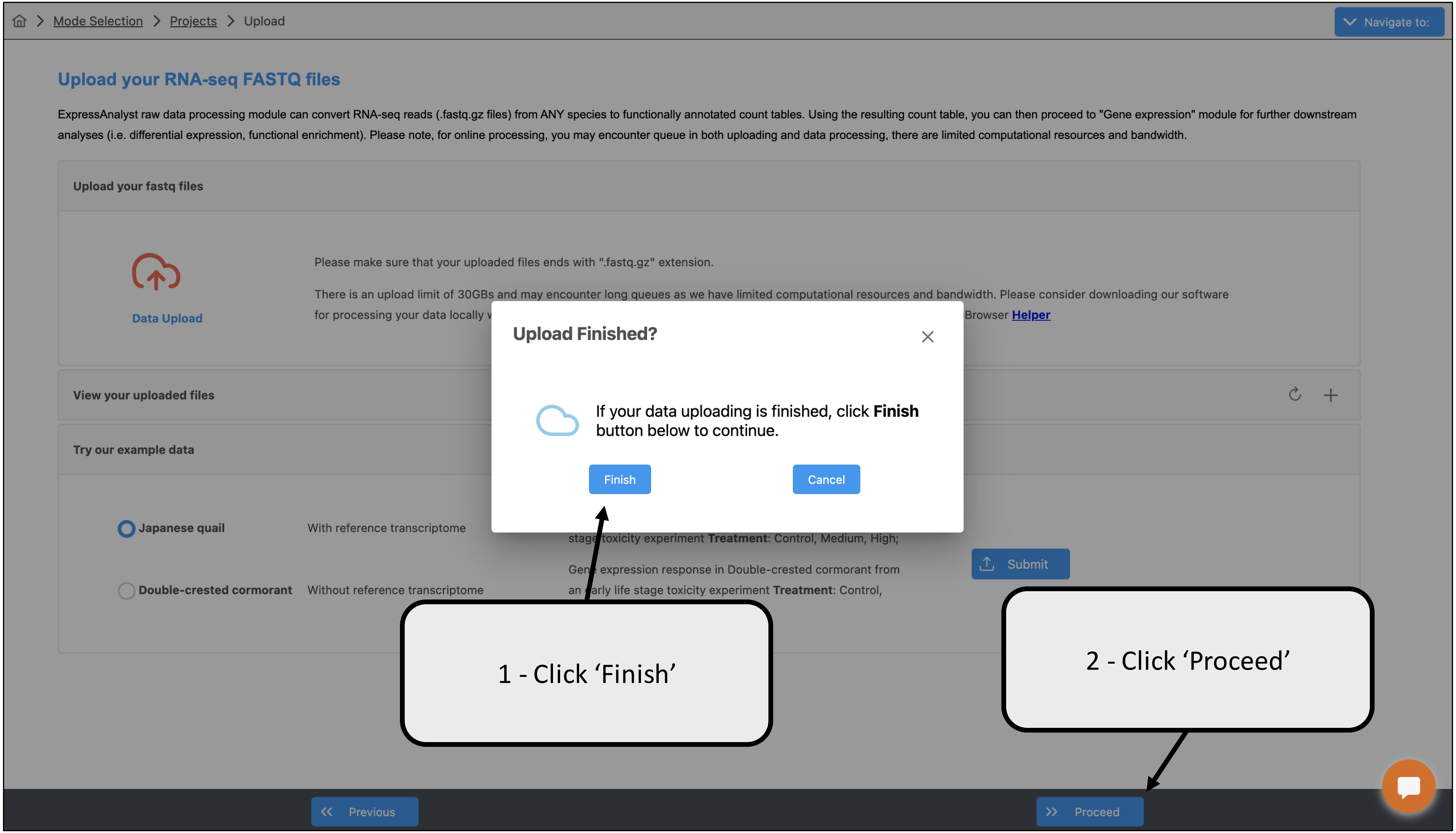
03 : 55 : 45

Note: if there is no uploading activity for **15 consecutive minutes**, this uploading session will be closed, no matter how much time left above. You have to restart a new session from ExpressAnalyst if you want to continue uploading.

Name ↑	Size	
A2.QE2-M1-LT_R1.fastq.gz	25.79 MB	
A2.QE2-M1-LT_R2.fastq.gz	25.79 MB	in a few seconds
A6.QE2-M2-LT_R1.fastq.gz	25.79 MB	in a few seconds
A6.QE2-M2-LT_R2.fastq.gz	25.79 MB	in a few seconds

1 – When uploading is finished, close this tab and return to ExpressAnalyst

Uploading files to the server uses up our internet *bandwidth*, so only a few users can upload files at the same time. To control this, we limit upload sessions to four hours. If your internet is not fast enough to upload all files in 4 hours, perform the uploading in batches. Drag some into this window, and then when the upload has finished, launch another session from ExpressAnalyst. We also only allow 4 uploading sessions at any one time, so if all sessions are being used, you must wait until one comes available. To avoid all of these challenges, try our **local Docker implementation**.



Upload your RNA-seq FASTQ files

ExpressAnalyst raw data processing module can convert RNA-seq reads (.fastq.gz files) from ANY species to functionally annotated count tables. Using the resulting count table, you can then proceed to "Gene expression" module for further downstream analyses (i.e. differential expression, functional enrichment). Please note, for online processing, you may encounter queue in both uploading and data processing, there are limited computational resources and bandwidth.

Upload your fastq files

Please make sure that your uploaded files ends with ".fastq.gz" extension.

There is an upload limit of 30GBs and may encounter long queues as we have limited computational resources and bandwidth. Please consider downloading our software for processing your data locally v Browser **Helper**

View your uploaded files

Try our example data

○ **Japanese quail** With reference transcriptome stage toxicity experiment **Treatment**: Control, Medium, High;

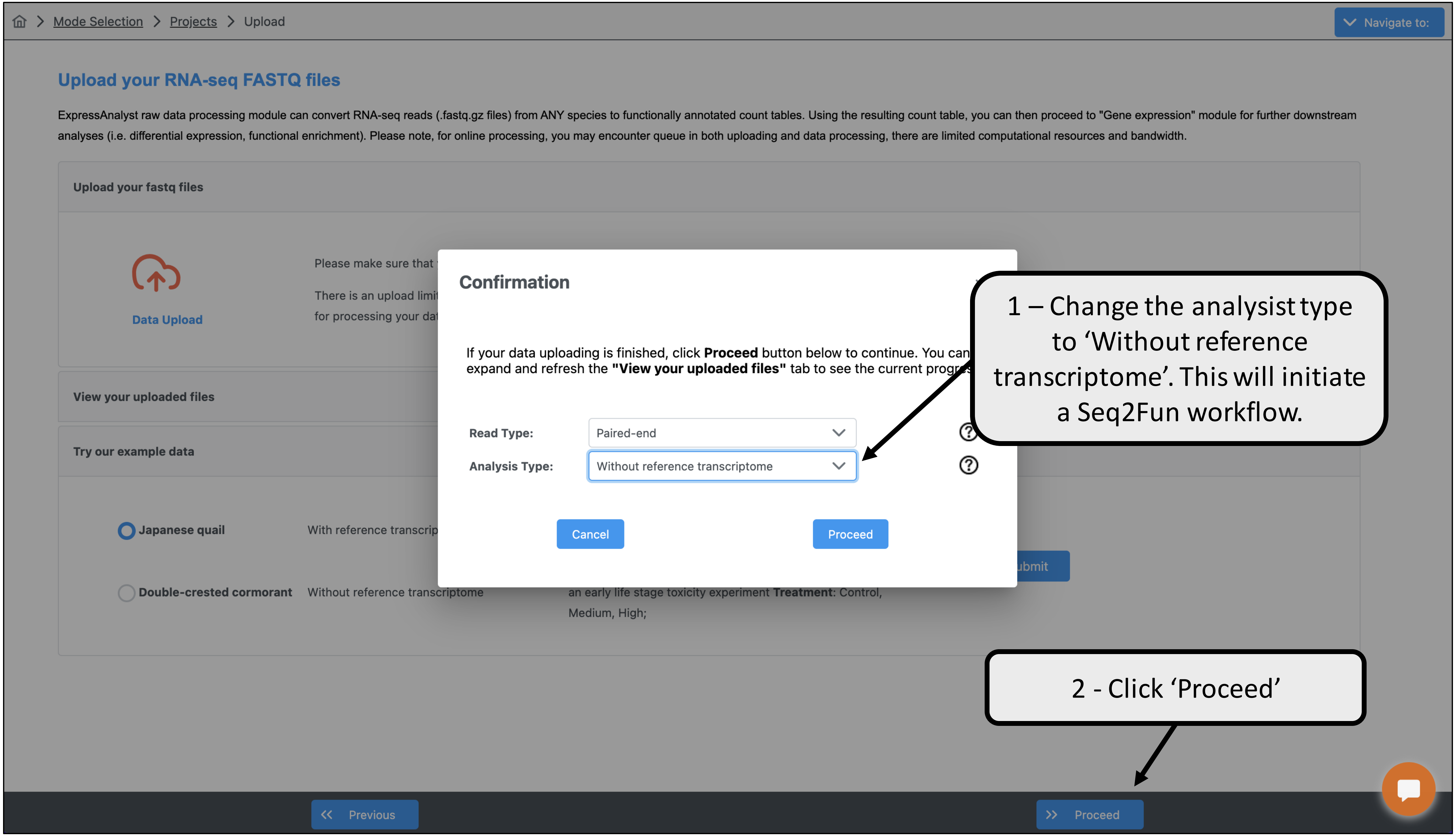
○ **Double-crested cormorant** Without reference transcriptome Gene expression response in Double-crested cormorant from an early life stage toxicity experiment **Treatment**: Control,

Upload Finished?

If your data uploading is finished, click **Finish** button below to continue.

1 - Click ‘Finish’

2 - Click ‘Proceed’



Upload your RNA-seq FASTQ files

ExpressAnalyst raw data processing module can convert RNA-seq reads (.fastq.gz files) from ANY species to functionally annotated count tables. Using the resulting count table, you can then proceed to "Gene expression" module for further downstream analyses (i.e. differential expression, functional enrichment). Please note, for online processing, you may encounter queue in both uploading and data processing, there are limited computational resources and bandwidth.

Upload your fastq files

Please make sure that

There is an upload limi
for processing your da

View your uploaded files

Try our example data

○ **Japanese quail** With reference transcrip

○ **Double-crested cormorant** Without reference transcriptome an early life stage toxicity experiment **Treatment**: Control, Medium, High;

Confirmation

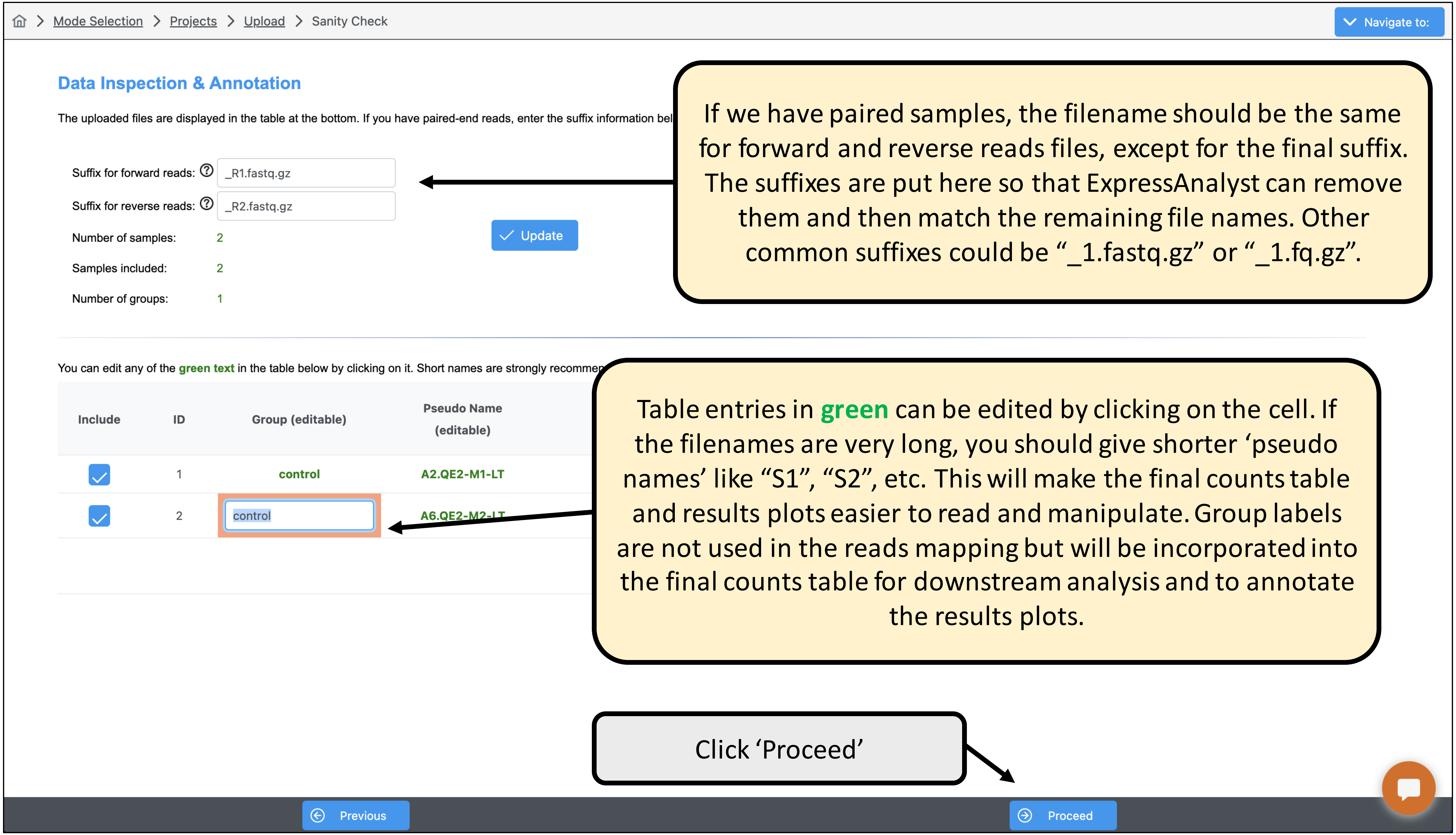
If your data uploading is finished, click **Proceed** button below to continue. You can expand and refresh the **"View your uploaded files"** tab to see the current progr

Read Type: Paired-end

Analysis Type: Without reference transcriptome

1 – Change the analysist type to ‘Without reference transcriptome’. This will initiate a Seq2Fun workflow.

2 - Click ‘Proceed’



Data Inspection & Annotation

The uploaded files are displayed in the table at the bottom. If you have paired-end reads, enter the suffix information bel

Suffix for forward reads:	_R1.fastq.gz
Suffix for reverse reads:	_R2.fastq.gz
Number of samples:	2
Samples included:	2
Number of groups:	1

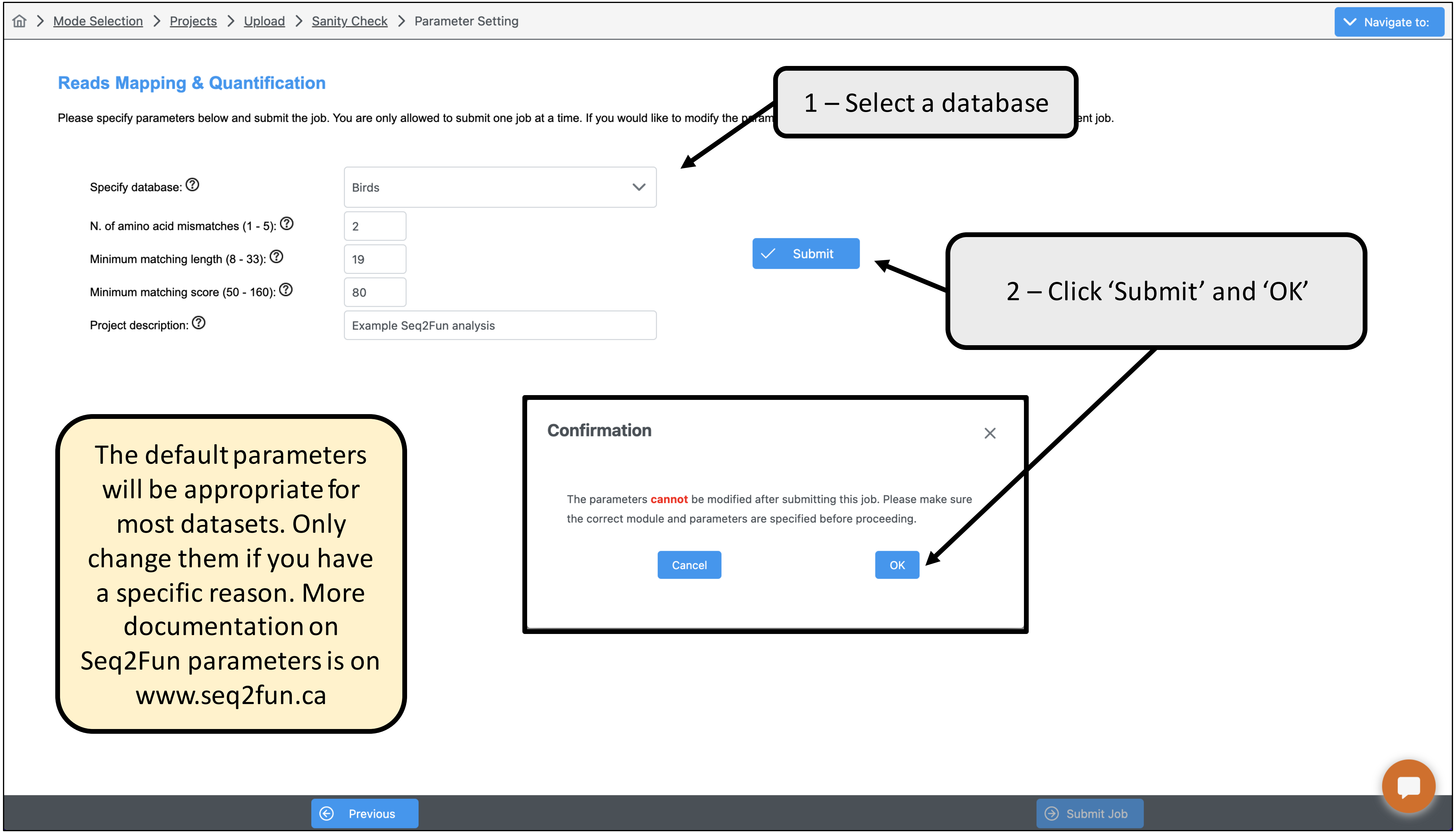
If we have paired samples, the filename should be the same for forward and reverse reads files, except for the final suffix. The suffixes are put here so that ExpressAnalyst can remove them and then match the remaining file names. Other common suffixes could be "_1.fastq.gz" or "_1.fq.gz".

You can edit any of the **green text** in the table below by clicking on it. Short names are strongly recommen

Include	ID	Group (editable)	Pseudo Name (editable)
☑	1	control	A2.QE2-M1-LT
☑	2	control	A6.QE2-M2-LT

Table entries in **green** can be edited by clicking on the cell. If the filenames are very long, you should give shorter 'pseudo names' like "S1", "S2", etc. This will make the final counts table and results plots easier to read and manipulate. Group labels are not used in the reads mapping but will be incorporated into the final counts table for downstream analysis and to annotate the results plots.

Click 'Proceed'


Mode Selection > Projects > Upload > Sanity Check > Parameter Setting
Navigate to:
Submit
Cancel
OK
Previous
Submit Job

Reads Mapping & Quantification

Please specify parameters below and submit the job. You are only allowed to submit one job at a time. If you would like to modify the param ent job.

Specify database: Birds

N. of amino acid mismatches (1 - 5): 2

Minimum matching length (8 - 33): 19

Minimum matching score (50 - 160): 80

Project description: Example Seq2Fun analysis

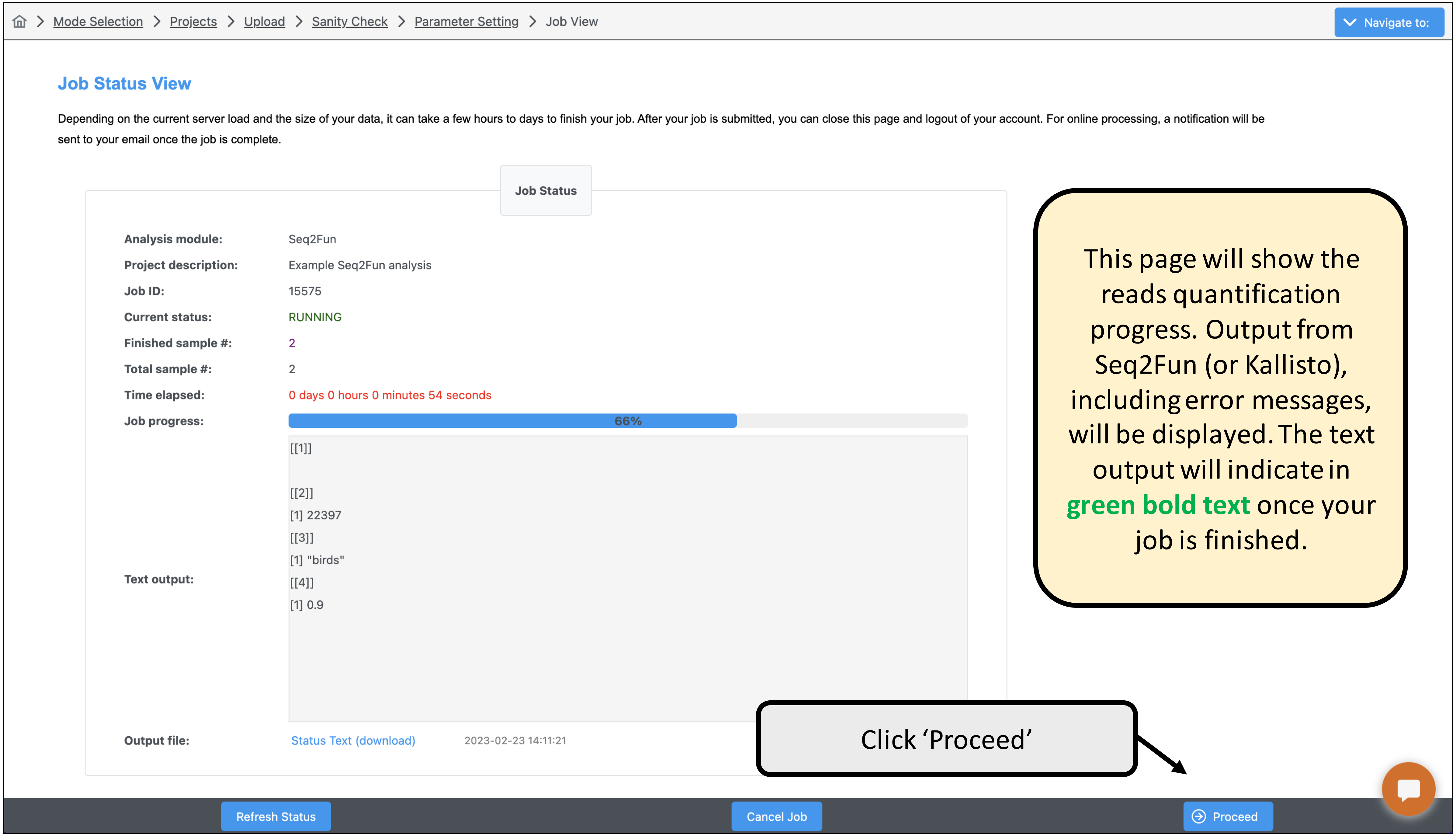
Confirmation

The parameters **cannot** be modified after submitting this job. Please make sure the correct module and parameters are specified before proceeding.

1 – Select a database

2 – Click ‘Submit’ and ‘OK’

The default parameters will be appropriate for most datasets. Only change them if you have a specific reason. More documentation on Seq2Fun parameters is on www.seq2fun.ca


Mode Selection > Projects > Upload > Sanity Check > Parameter Setting > Job View
Navigate to:
Job Status
Refresh Status
Cancel Job
Proceed

Job Status View

Depending on the current server load and the size of your data, it can take a few hours to days to finish your job. After your job is submitted, you can close this page and logout of your account. For online processing, a notification will be sent to your email once the job is complete.

Analysis module:	Seq2Fun
Project description:	Example Seq2Fun analysis
Job ID:	15575
Current status:	RUNNING
Finished sample #:	2
Total sample #:	2
Time elapsed:	0 days 0 hours 0 minutes 54 seconds
Job progress:	66%

Text output:

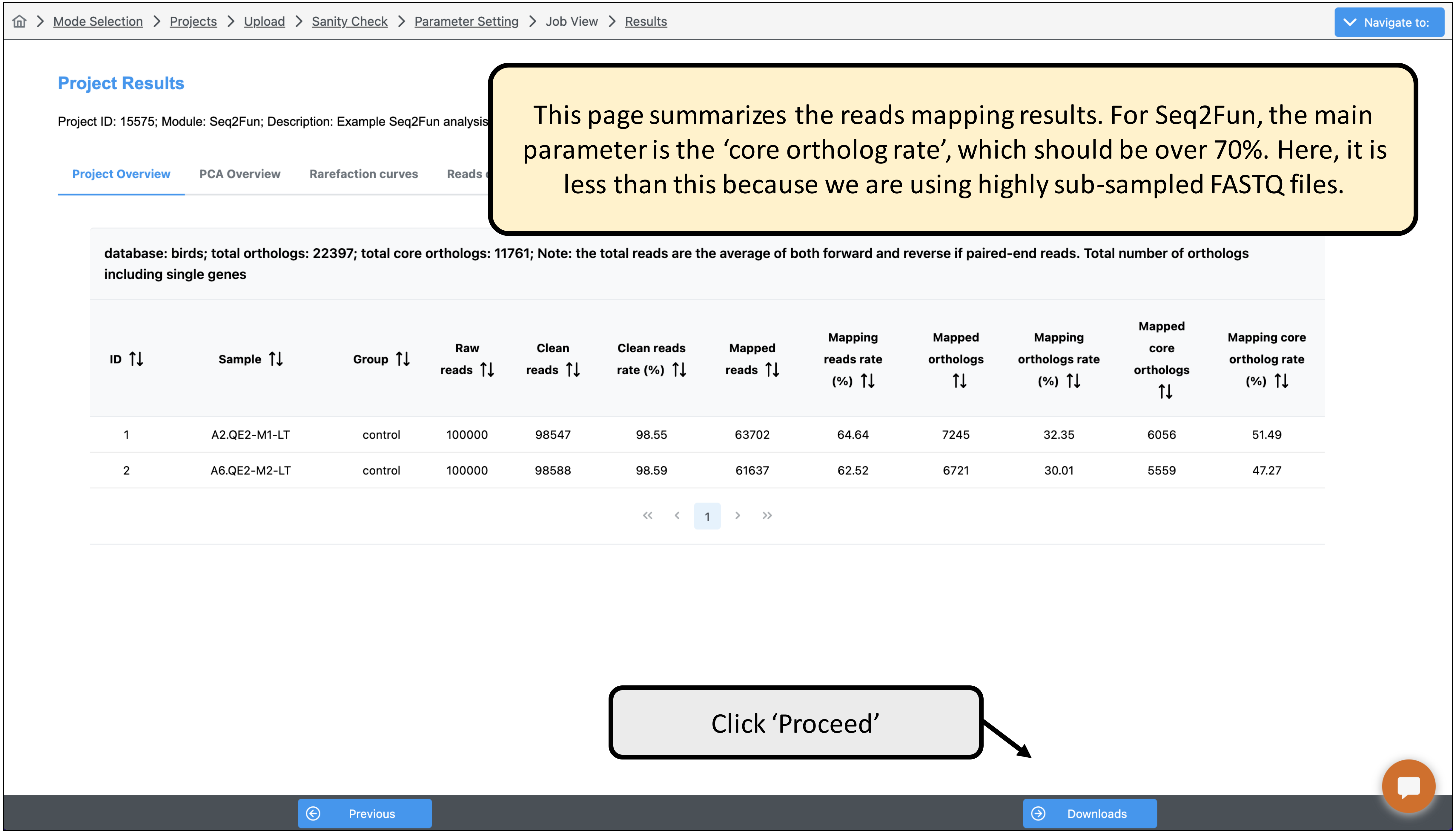
```
[[1]]


[[2]]
[1] 22397
[[3]]
[1] "birds"
[[4]]
[1] 0.9
```

Output file: Status Text (download) 2023-02-23 14:11:21

This page will show the reads quantification progress. Output from Seq2Fun (or Kallisto), including error messages, will be displayed. The text output will indicate in **green bold text** once your job is finished.

Click ‘Proceed’


Mode Selection > Projects > Upload > Sanity Check > Parameter Setting > Job View > Results
Navigate to:

Project Results

Project ID: 15575; Module: Seq2Fun; Description: Example Seq2Fun analysis

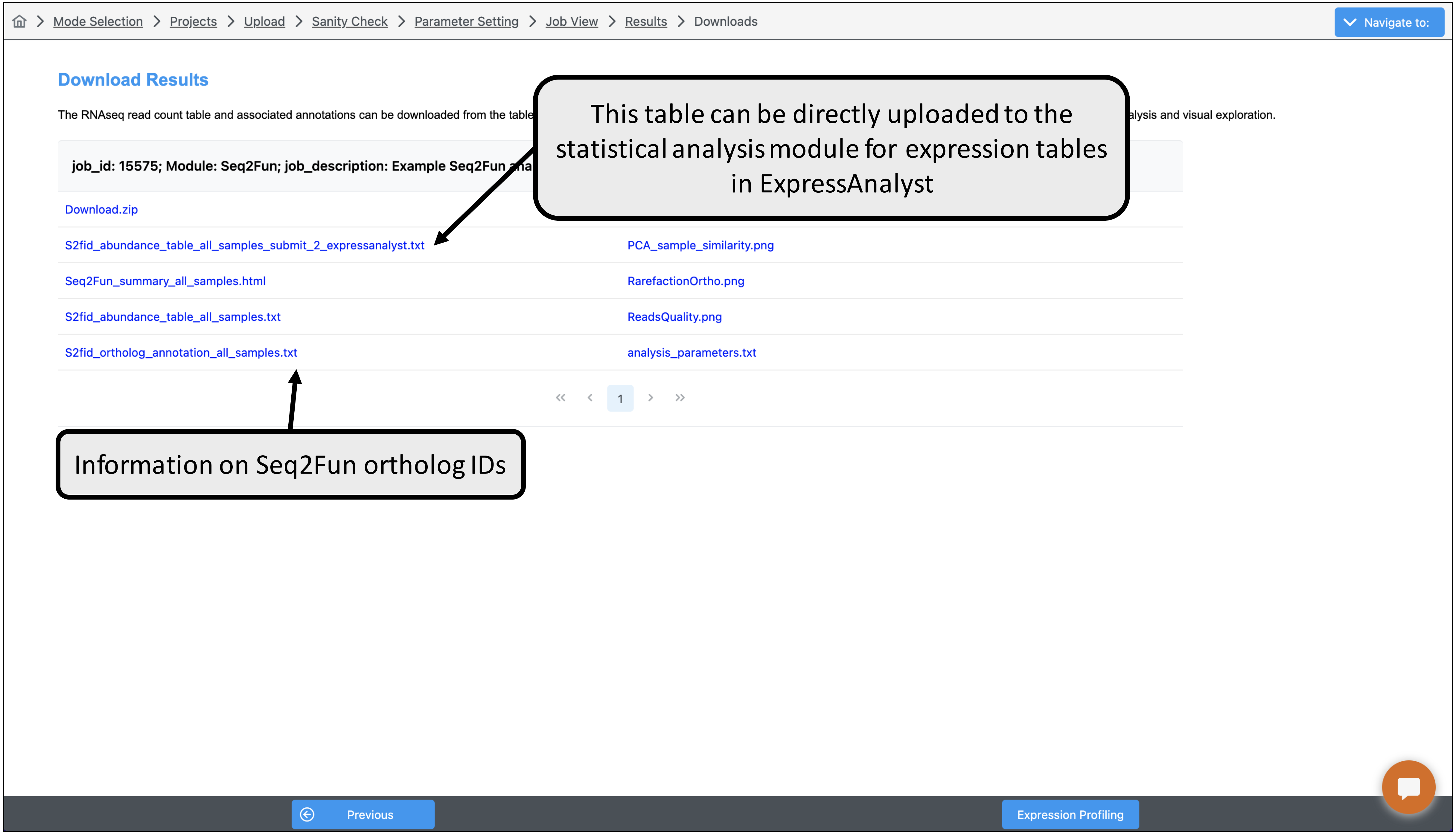
Project Overview
PCA Overview
Rarefaction curves
Reads

This page summarizes the reads mapping results. For Seq2Fun, the main parameter is the ‘core ortholog rate’, which should be over 70%. Here, it is less than this because we are using highly sub-sampled FASTQ files.

database: birds; total orthologs: 22397; total core orthologs: 11761; Note: the total reads are the average of both forward and reverse if paired-end reads. Total number of orthologs including single genes

ID	Sample	Group	Raw reads	Clean reads	Clean reads rate (%)	Mapped reads	Mapping reads rate (%)	Mapped orthologs	Mapping orthologs rate (%)	Mapped core orthologs	Mapping core ortholog rate (%)
1	A2.QE2-M1-LT	control	100000	98547	98.55	63702	64.64	7245	32.35	6056	51.49
2	A6.QE2-M2-LT	control	100000	98588	98.59	61637	62.52	6721	30.01	5559	47.27

Click ‘Proceed’


Mode Selection > Projects > Upload > Sanity Check > Parameter Setting > Job View > Results > Downloads
Navigate to:
« ‹ 1 › »
Previous
Expression Profiling

Download Results

The RNAseq read count table and associated annotations can be downloaded from the table alysis and visual exploration.

job_id: 15575; Module: Seq2Fun; job_description: Example Seq2Fun ana	
Download.zip	
S2fid_abundance_table_all_samples_submit_2_expressanalyst.txt	PCA_sample_similarity.png
Seq2Fun_summary_all_samples.html	RarefactionOrtho.png
S2fid_abundance_table_all_samples.txt	ReadsQuality.png
S2fid_ortholog_annotation_all_samples.txt	analysis_parameters.txt

This table can be directly uploaded to the statistical analysis module for expression tables in ExpressAnalyst

Information on Seq2Fun ortholog IDs

The End!!