How to store your sequence read data at TSL

The TSL Bioinformatics team try to make sure that all sequence data from TSL projects is handled and archived properly. This includes safe and secure data storage in the HPC environment, initial quality control, and registration and deposition in public archives, specifically to the European Nucleotide Archive (ENA).

We created a web interface (DataHog) to the TSL data storage structure ("the cluster"). It can be reached at http://data.tsl.ac.uk.

We mimic ENA's project/sample/read data structure. Each set of read data belongs to a sample (NCBI/ENA:BioSample) from which it was sequenced from. A sample can have multiple sets of read data (e.g. Illumina HiSeq and PacBio) associated.
In turn, each sample belongs to a project (NCBI/ENA:BioProject) and a project can have multiple samples associated (e.g. different timepoints, infections with different pathogens, environmental conditions etc.)

Why do this?

Simple. Think "research integrity",or in particular, "reproducibility". A key element of being able to reproduce the work you have done is your data. This is not just important for any potential colleague who wants to work on your data in the future, but also part of the publication process. Increasingly journals require data submitted to public repositories. Furthermore, think about what usually happens when you are about to submit a manuscript. Your mind is likely to be full of things that have to be done and all of a sudden the submission guidelines tell you to publish your data. This may or may not lead to a certain panic with people running to the Bioinformatics team, yelling "Where is my data?". And then imagine us saying: "What data?", because we may have never seen your data in the first place.

To help us help you manage your data, or if you want to work with your data on the cluster, then we ask you to use the web interface at http://data.tsl.ac.uk and register and deposit the data.

This generally only requires a very short amount of your time, but remember, if you want to publish your results, it is highly likely that you will have to (and want to) publish your data as well. To put it briefly, at some point you will need to provide all this information anyway, so the best way is to register it now and save yourself some extra hassle later.

Before you start

1. **Attention: By default this system will deposit your data at ENA - a public repository**
   Read data archived by DataHog will automatically be deposited in ENA and kept confidential for 2 years after submission. There will be an email reminder a month before this period ends, in case the project is not done yet and needs to be kept confidential for an extended period of time.

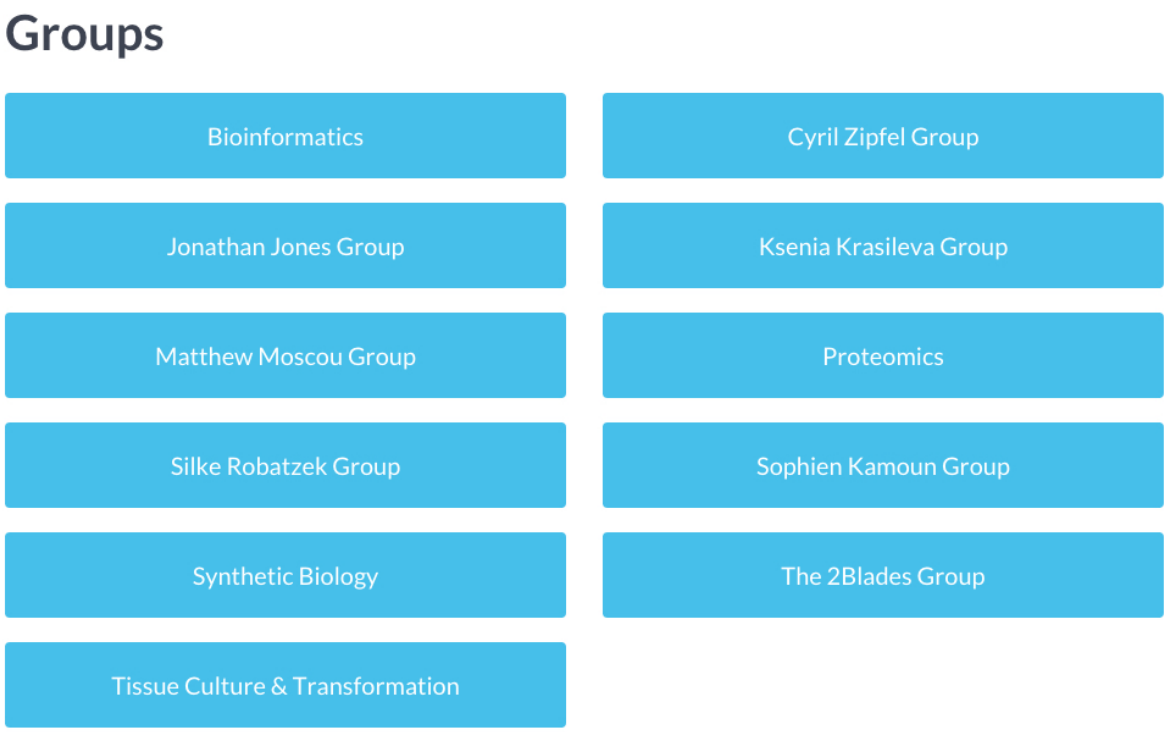
   If you are a member of the 2Blades group or if your data has to be kept out of the public archives for whatever reason, please seek out a member of the Bioinformatics team before archiving your data.

2. **Raw data means *raw* data**
   Since we need to store the raw data of all in-house sequencing projects, we need to ask you to adhere to the following policy. When you receive your sequencing data from your provider (e.g. via download or hard disk), you should take a moment to register and submit your data to the TSL Read Data Archives. Please upload the data files exactly as you obtained them from the sequencing company (e.g., do not decompress them or annotate them prior to this archiving step).

3. **Data formats**
   The TSL read archives and will store next generation sequencing data in Fastq-formatted read files (allowed file extensions are txt, fq, fastq). These files can be compressed with gzip (gz) or bzip2 (bz2). If your data does not comply with either of these formats, please see a member of the Bioinformatics team.

Getting Started

Please navigate your web browser to http://data.tsl.ac.uk, admire the pig snout logo, and click on “Show Groups”. You will be asked to sign in with your NBI username and password, after which you will be transferred to the “Groups” page. TSL’s sequencing projects are grouped by the various research and support groups.

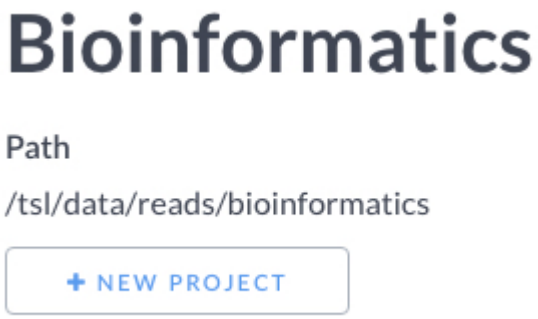


If you click on the field with your group's name, you will be presented with a list of projects that belong to your group.

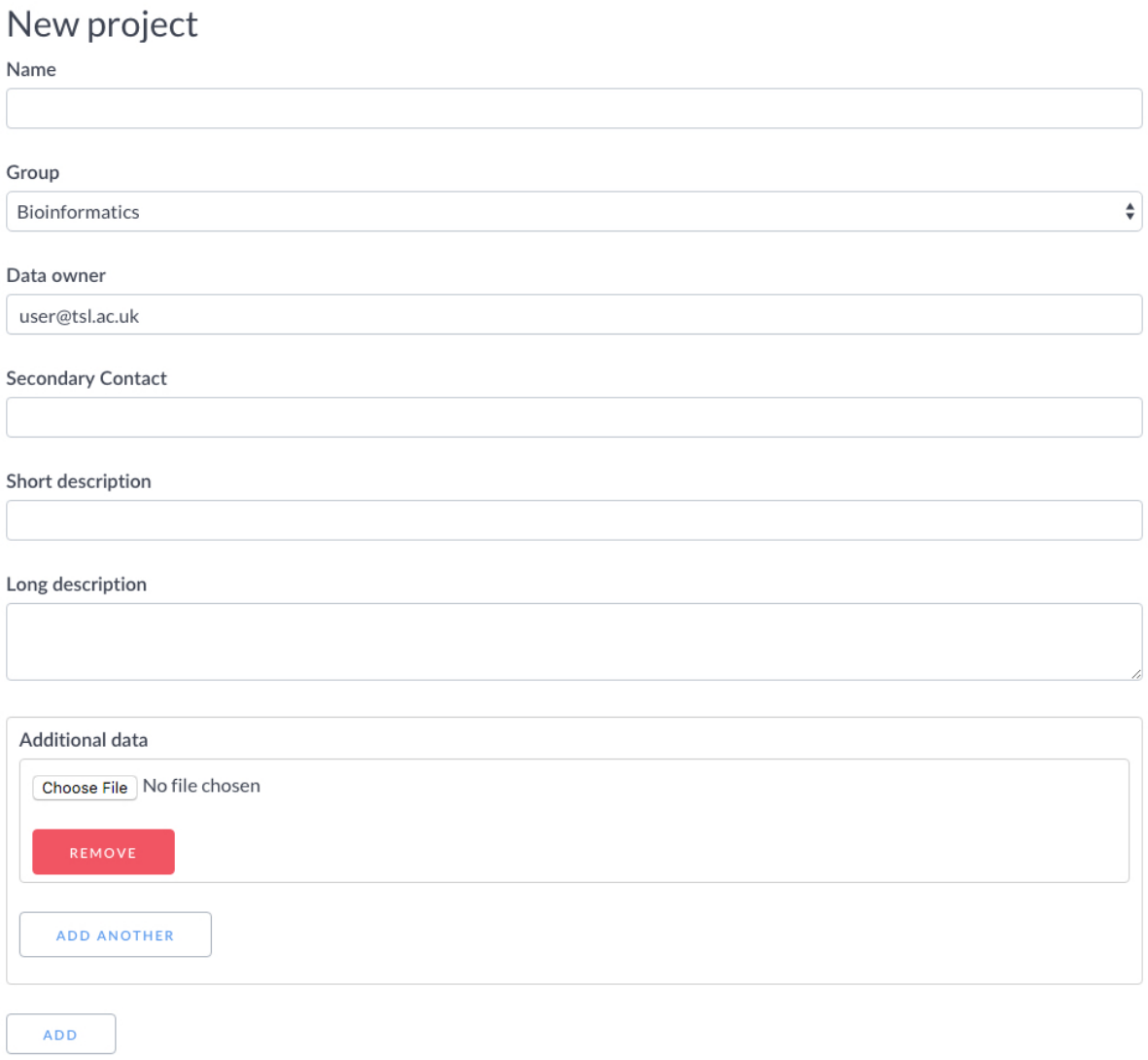
If you want to add your own data it is important to know whether you start your own project (see **"Adding a new project"**) or are contributing to an existing study (see **"Adding data to an existing project"**). If you are not sure about this, please discuss it with your colleagues, supervisor, or someone from the Bioinformatics team.

Adding a new project

If you are starting a completely new project, then you will have to create a new project in DataHog. This can simply be achieved by clicking on the New Project button and filling in the information described below in the following form.



The information that you provide to generate projects (and later samples) in DataHog will be used for registering your project as a BioProject in the European Nucleotide Archive (ENA). Input fields are in general required. Information from input fields marked with an asterisk (*) is submitted to ENA.



*Name** - Find a suitable short name for your project, something that you can memorise and that also works reasonably well to present your study to the public.

Data owner - Please provide your TSL email (i.e. firstname.lastname@tsl.ac.uk or firstname.lastname@sainsbury-laboratory.ac.uk). If/when you leave the lab and if your data is property of TSL, please seek out a member of the Bioinformatics team to update the data owner records. We will require a contact email address in case there are uncertainties about how the data was processed.

Secondary contact - Please provide a second contact from within your group. This should be someone who is familiar with your data as well, e.g. your group's lab manager / RA or your

main collaborator/supervisor (if you are a student) in the group, if that person is likely to stay in the lab longer than you. It should not be your group leader.

The next 2 fields are required by ENA. Here is an example of a published TSL-BioProject with a comprehensive description (http://www.ebi.ac.uk/ena/data/view/ERP002644) and one submitted from TGAC with a somewhat less comprehensive description (http://www.ebi.ac.uk/ena/data/view/PRJEB3149). Try to be somewhere in between.

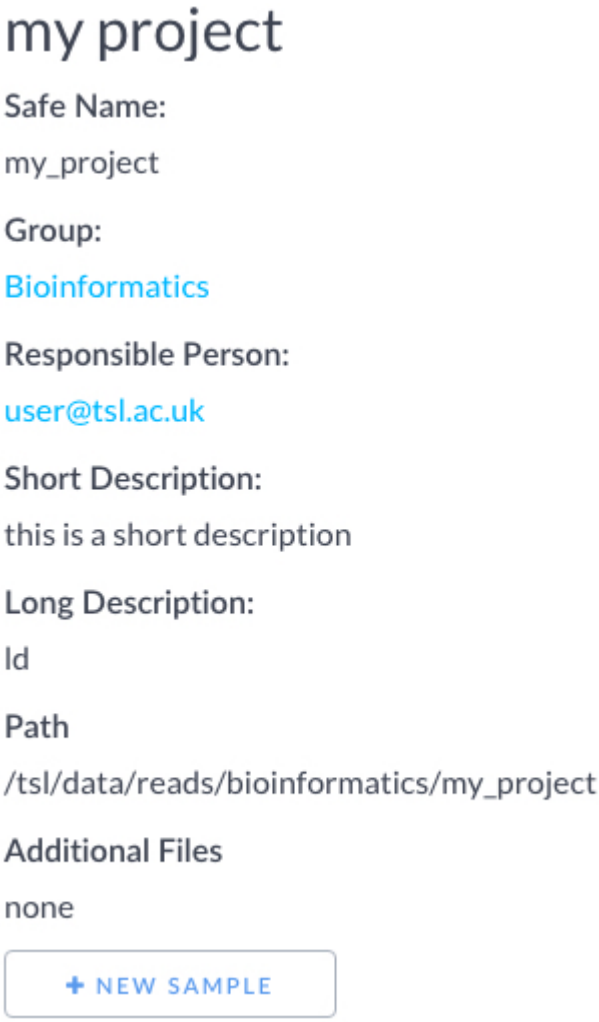
*Short description** - One to three short descriptive sentences that provide information about the study.

*Long description** - The purpose of this field is to provide an abstract about the study. It is a required field for ENA and if you already have an abstract for a publication ready, then by all means use it. If not, simply copy the short description and paste it here.

Supporting files - We ask you to upload any files (data sheets, quality reports, contract pdfs that you obtained from the sequencing provider. This means that you should download/copy *everything* from the weblink/hard disk that has been provided to you. Multiple files can be added using the “ADD ANOTHER” button.

When you are done filling the form, simply click “ADD”.

If everything went fine, you should be redirected to the project page of the newly created project.

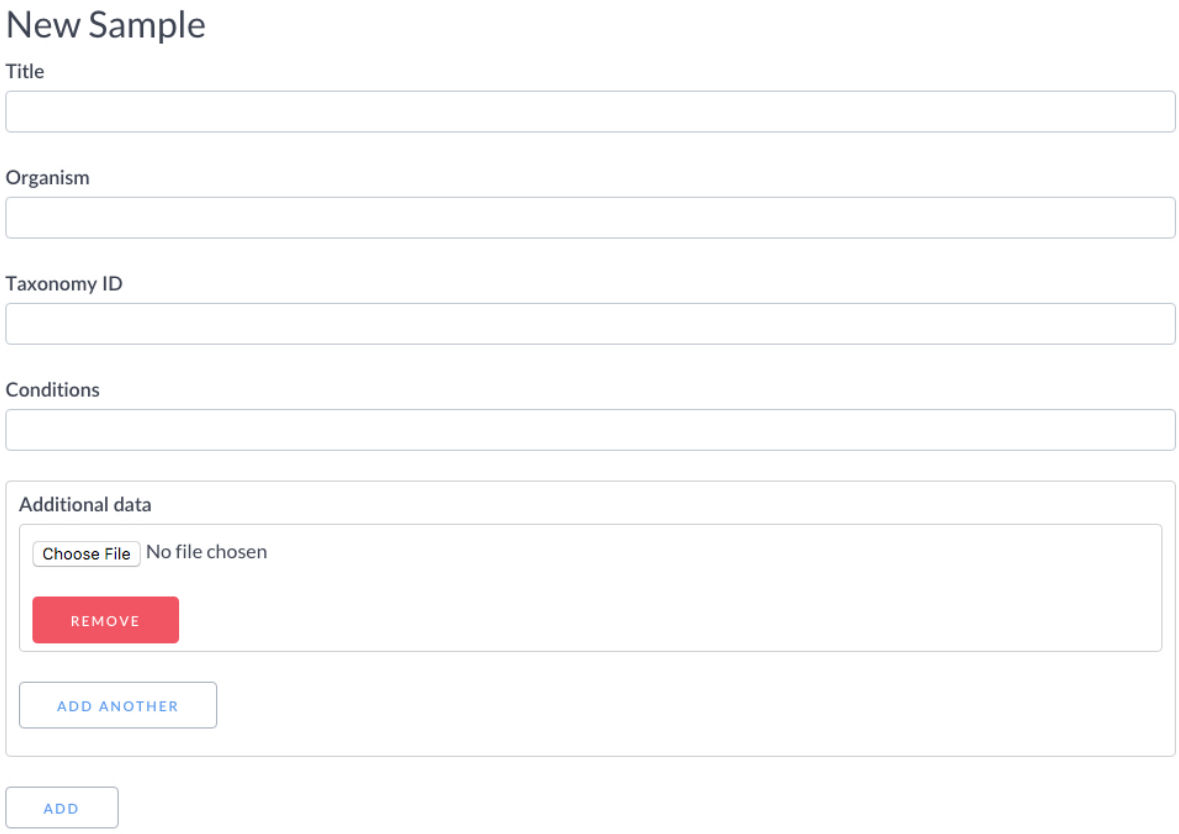


Adding data to an existing project

Click on the button representing a project to which you want to add data from the project list. If you have completely new data, then click on the New Sample button (see **"Adding a sample"**). Otherwise, i.e. if you are adding data that were sequenced from an existing sample (e.g. PacBio and Illumina data or multiple rounds of Illumina data that were obtained from sequencing the same sample), click on the respective sample (see **"Adding read data to a sample"**).

Adding a sample

Sample creation is performed similarly to the creation of a project (and the corresponding BioProject) by filling out the following form. Again, fields marked with an asterisk are submitted to ENA. Samples created in this way will be registered in ENA as BioSamples and associated with the corresponding BioProject.



*Title** - Please provide an informative title for the sample. Something like “Renseq-s_americanum_some-accessionID”.

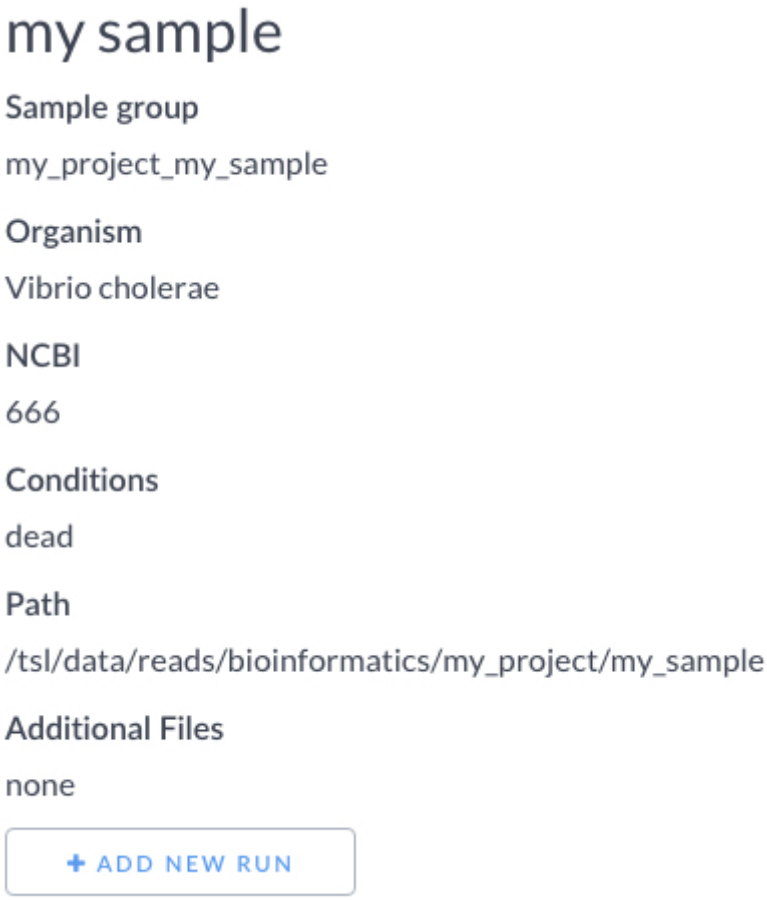
*Organism/Taxonomy ID** - Choose the organism your sample was extracted from either by typing the name into the *Organism* field (the form will suggest the proper organism if you type enough letters, you can then simply click on the suggested title and the *Taxonomy ID* field will update automatically) or if you know the NCBI taxonomy ID (you can find out at http://www.ncbi.nlm.nih.gov/taxonomy) then simply enter it into the *Taxonomy ID* field and the *Organism* field will update.

If you have an infected sample, e.g. Phakopsora on soy or P. infestans in potato, then choose the host as organism since the sample will likely contain more host than pathogen. Unfortunately, the ENA data model (and therefore DataHog data model) does not allow multi-organism samples.

Conditions - Enter the sample conditions here. E.g. wildtype, strain information, infection information, etc. Anything that helps as metadata to understand the sample better. The contents of this field are for in-house use and will not be transmitted to ENA.

Additional Data - Similar to additional data on the project level, if you have any information that only pertains to a certain sample, please add it here.

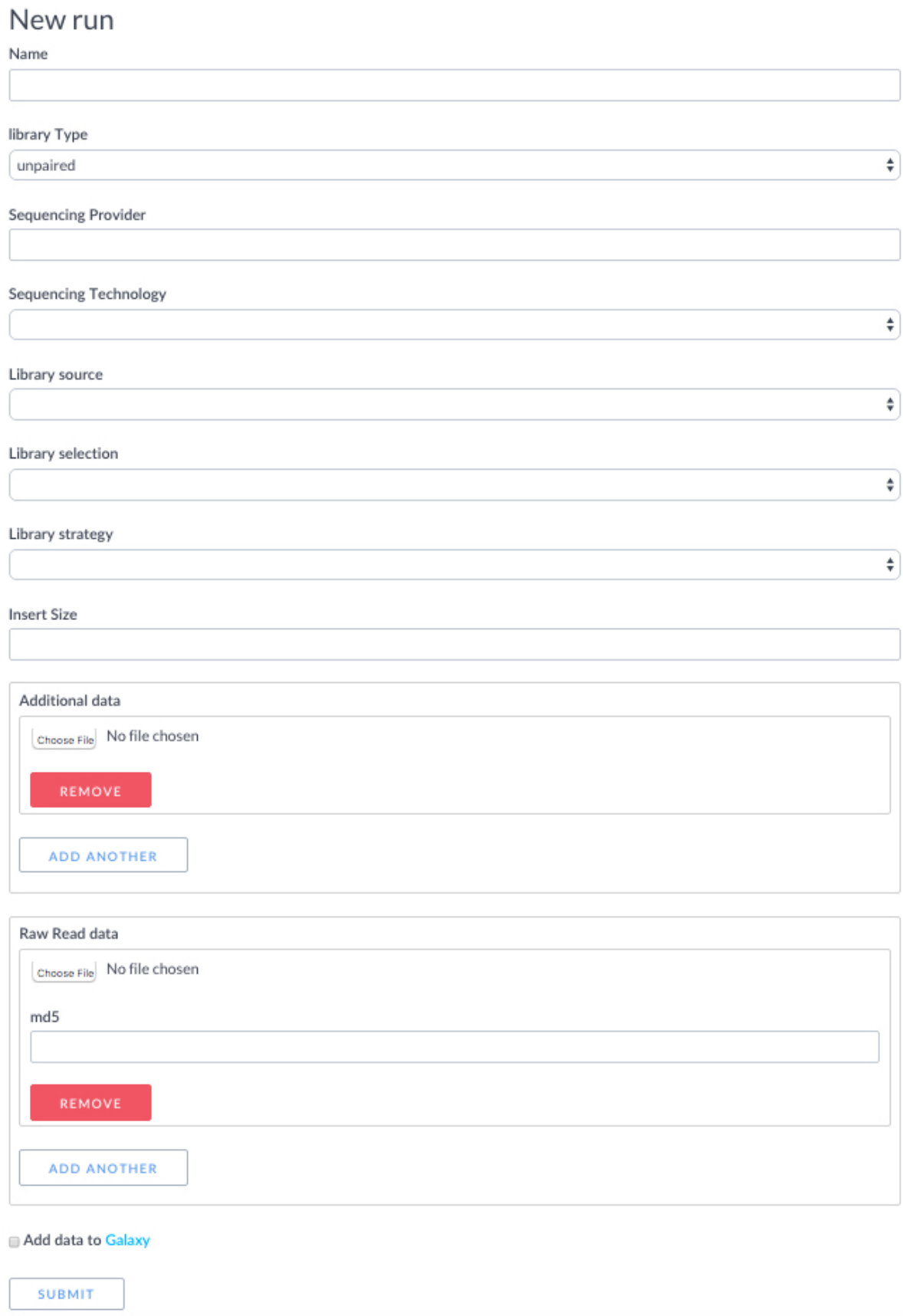
Finally, when you are done filling the form, click on the ADD button. This should create a new sample page and redirect your browser to it.



Adding read data to a sample

Read data will be submitted in batches or “runs”. Data generated at the same time with the same technology and library parameters can be submitted in the same batch. For instance, a data set of 3 x 2 paired-end Illumina MiSeq files with insert sizes of 500, 500, and 800 and 2 PacBio files will be submitted in 3 batches (4 files MiSeq/500, 2 files MiSeq/800, 2 files PacBio).

Read data can be added to a sample by clicking on the Add New Run button. Again, you will have to fill a number of fields (with fields marked with a ‘*’ denoting ENA required fields) with information regarding your data.



Name - Provide an informative and concise name for your sample.

*Library Type** - This can be either *paired, mate,* or *unpaired*. If you choose *paired* or *mate* then you will have to provide an even number of read files and an insert size.

Sequencing Provider - provide at least the company name (TGAC, BGI) and maybe the contact email from the company representative.

For the following 4 fields, you may have to refer to the reports/data sheets/contract sent to you by the sequencing provider or contact them. If it is impossible to obtain that information, then please select *OTHER* or *unspecified*.

*Sequencing Technology** - Select one from the list.

*Library Source** - Select one from the list.

*Library Strategy** - Select one from the list.

*Insert Size** - Enter the mean/average insert size (hopefully) reported by the sequencing provider.

Additional Data - Similar to additional data on the project and sample levels, if you have any information that only pertains to a certain run, please add it here.

This way, we will provide a record of all project metadata and you won't have to frantically dig through your emails at the end of the project when the submission deadline is due and you are stressed out enough already.

We further ask you to save any email exchange with details about the project (problems that occurred, clarifications, etc.) in a text or pdf file and upload this as well.

To upload a file, click the Choose file button in the Additional Data field. Multiple files can be added consecutively by clicking the Add More Additional Data button.

Raw Read Data - Fastq files containing read data will be uploaded in groups. If you have selected a *paired* or *mate* library type then you will have to provide Fastq files in groups of two. The first field is for the forward (.R1/_1/etc.) file, the second for the reverse (.R2/_2/etc.) file. Otherwise, it will be one Fastq file per group. This field is only available upon run creation. You will not be able to add raw reads to a run later on (but you can always create a new run at no extra cost.)

Processed Read Data - In general, DataHog will only store raw sequencing data. However, in certain situations, you will obtain e.g. already cleaned, i.e. 'processed' read data from your sequencing provider. In such cases you can store processed data after you created the run and added the raw data. The Processed Read Data field will only be available after run creation.

To upload a file, click the Choose file button in the Read Data field. Multiple files can be added consecutively by clicking the Add More Additional Data button.

In addition, the upload form requires you to provide the MD5 sum (https://en.wikipedia.org/wiki/MD5, or briefly: the fingerprint of a file) for each file. This is needed so we can be certain that we store exactly the same data as generated by the sequencing provider. You should have obtained the MD5 sums for your files together with your read data. If not and if you downloaded your data from the company's website then you should inquire with them.

If the MD5 sums are for some reason not available, then you have to generate them yourself.

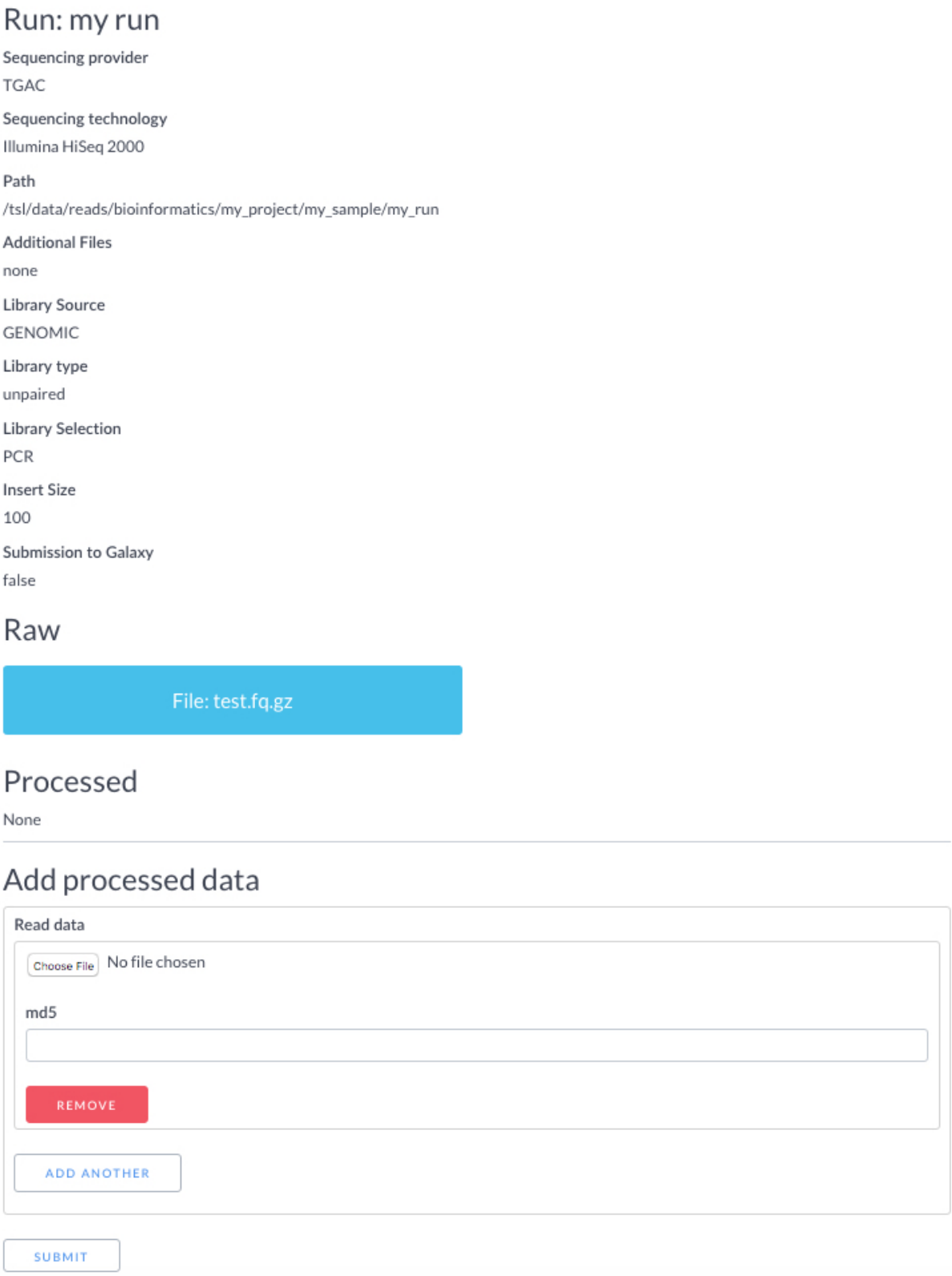
If you are on a windows machine without Cygwin (or something similar) then you will have to come down to us to find a solution. Maybe Neil can install this (https://support.microsoft.com/en-us/kb/841290) for you, but so far we don't know whether this works or not.

If you are on a Mac or Linux system or know how to access the cluster and have stored the read files in your cluster home, then you can generate the MD5 sums like this:

1. Open a terminal.
2. Navigate to the directory that contains your read data with the 'cd' command. On Mac you could drag the folder from the Finder into the terminal in order to get the proper directory name.
3. Type 'md5 filename' (or 'md5sum filename') without the '' and filename being the name of the read file and hit enter. This will take a moment but then it will give you a string that looks something like this 0dd888df2350f72ca5def246b080cbfb. Copy that string and paste it into the MD5 field for the file.

Add Data to Galaxy - If you intend to work with this data using the inhouse Galaxy server right away, then please tick this box. It will notify the TSL Galaxy administrator to import the data into Galaxy and make them available to you. This is optional. You can always ask later to import your data into Galaxy.

If you are done adding files and filling the form, please click on the Submit button. Congratulations, you have just registered your first read data set in the TSL Read Data Archives.



Accessing Read Data

DataHog web interface allows you to easily get access to your read data. Read data files are accessible via the blue buttons in the sections “Raw” and “Processed”. Single end read data is represented by a single button, whereas the two files of a paired end data run is represented by a split button. Clicking on a Read data button, will show the MD5 sum as well as its location on the cluster. It also allows to download the file and view an automatically generated FastQC report for the data.