United Nations
Educational, Scientific and
Cultural Organization

क्षेत्रीय जैव प्रौद्योगिकी केन्द्र
Regional Centre
for Biotechnology

Indian Biological
Data Centre

# Standard Operating Procedure (SOP)
# For Nucleotide Data Submission to
# Indian Nucleotide Data Archive (INDA)

**Version_0.1**

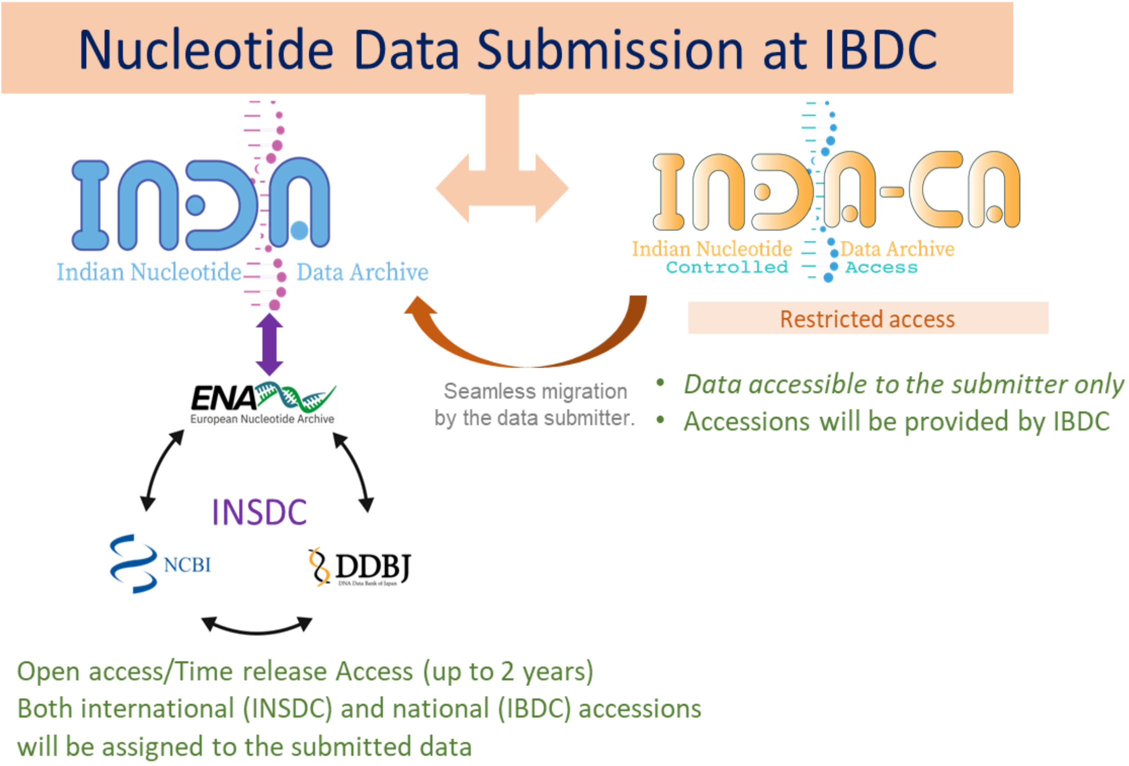
**2022**

INDA-CA
Indian Nucleotide Data Archive
Controlled Access

## INDA-CA: Table of Content

## Overview

Welcome to the data archive solutions covered by the Indian Biological Data Centre (IBDC). This guide will be helpful in understanding the standard operating procedure for the submission of the nucleotide data to IBDC. Users are requested to devote a moment towards understanding the structure and mandate of portals developed for dedicated nucleotide data archive before they proceed with the submissions. IBDC allows nucleotide data submission in two modes based on data accessibility i.e. open and controlled access (Figure 1). The portal for open-access data is "Indian Nucleotide Data Archive (INDA)" while controlled-access / private data is handled by "Indian Nucleotide Data Archive – Controlled Access (INDA-CA)".


Nucleotide Data Submission at IBDC
INDA
Indian Nucleotide Data Archive
INDA-CA
Indian Nucleotide Data Archive
Controlled Access
Restricted access
Seamless migration by the data submitter.
ENA
European Nucleotide Archive
INSDC
NCBI
DDBJ
DNA Data Bank of Japan

- *Data accessible to the submitter only*
- Accessions will be provided by IBDC

- Open access/Time release Access (up to 2 years)
- Both international (INSDC) and national (IBDC) accessions will be assigned to the submitted data

**Figure 1**. Nucleotide data submission Domains of IBDC based on data accessibility model.

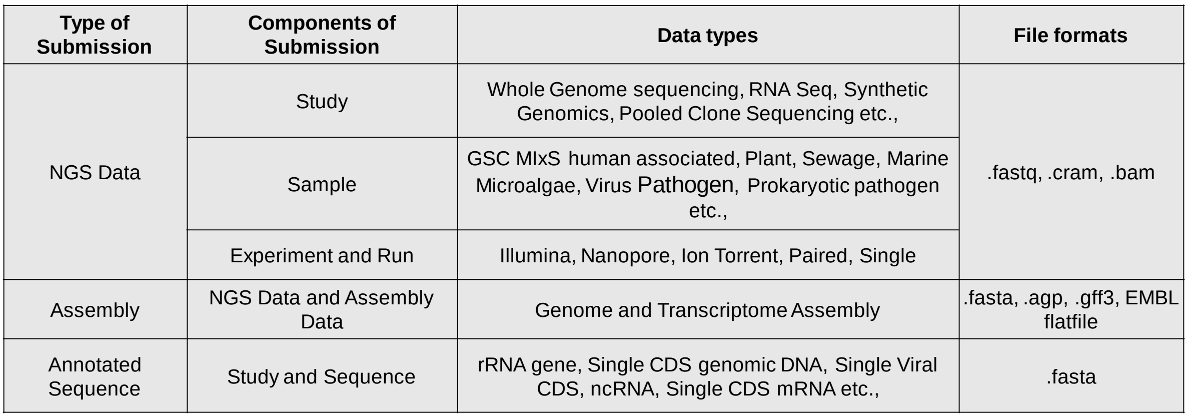
## Data sharing and accessibility model of INDA-CA

The Indian Nucleotide Data Archive- Controlled access (INDA-CA), is a controlled-access/ private platform for archiving and managing diverse types of nucleotide sequencing data generated across India. The portal can be accessed at https://inda.rcb.ac.in/indasecure .

The data submitted to INDA-CA is completely confidential and only user has the access to the data. Data submitted to INDA-CA will be stored in private or in a controlled-access manner with IBDC until the user decides to stay it the private way. The user will get accession from IBDC only and the data will not be shared with INSDC organizations.

## Data Type submission services offered by INDA-CA

INDA-CA also offers three types of nucleotide data submission i.e. Next generation sequencing data, assembly and annotated sequence submission (Table 1). The details about each data type will be discussed in detail in the upcoming sections.



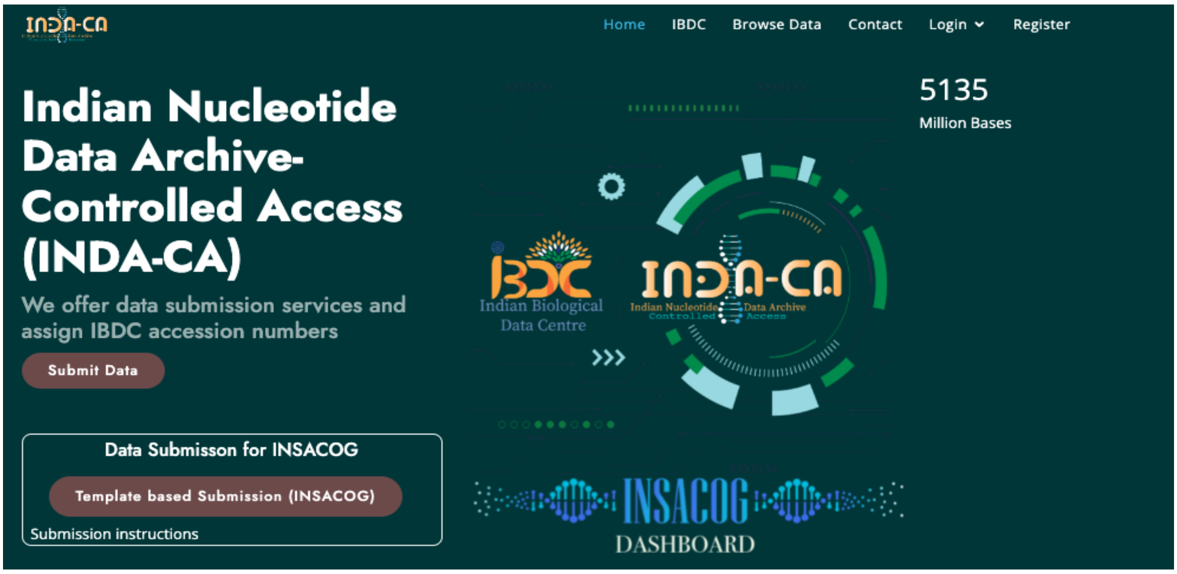
| Type of Submission | Components of Submission | Data types | File formats |
|---|---|---|---|
| NGS Data | Study | Whole Genome sequencing, RNA Seq, Synthetic Genomics, Pooled Clone Sequencing etc., | .fastq, .cram, .bam |
| | Sample | GSC MIxS human associated, Plant, Sewage, Marine Microalgae, Virus Pathogen, Prokaryotic pathogen etc., | |
| | Experiment and Run | Illumina, Nanopore, Ion Torrent, Paired, Single | |
| Assembly | NGS Data and Assembly Data | Genome and Transcriptome Assembly | .fasta, .agp, .gff3, EMBL flatfile |
| Annotated Sequence | Study and Sequence | rRNA gene, Single CDS genomic DNA, Single Viral CDS, ncRNA, Single CDS mRNA etc., | .fasta |

**Table 1**. Types of Data submission available on INDA-CA.

## A general guide on data submission

### Getting started on the submission

To submit data to INDA-CA, the user must register a submission account at INDA-CA portal (https://inda.rcb.ac.in/indasecure). Otherwise, also if one visits the parent IBDC website (https://ibdc.rcb.res.in/), clicking on the 'Submit Data' button or INDA-CA also navigates to the INDA-CA portal as presented below in the Figure 3. In the INDA-CA portal, click on the 'Submit Data' button and click on the INDA-CA link to initiate the general data submissions to INDA-CA. Specialized links of specific consortium are also shown at the INDA-CA portal, but that are only accessible to the network partners.



Indian Nucleotide Data Archive-Controlled Access (INDA-CA)

We offer data submission services and assign IBDC accession numbers

5135

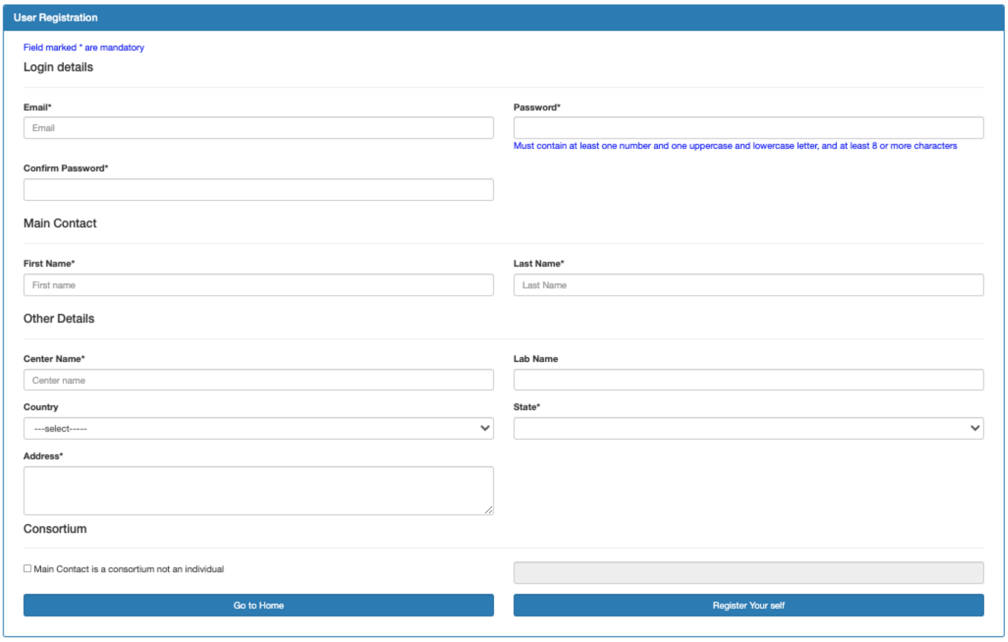
Million Bases

Data Submisson for INSACOG

Submission instructions

**Figure 3.** INDA-CA Home page

To register a submission account at INDA-CA, user has to click ‘Register’ button at the home page of the INDA-CA portal. User will be redirected to the user registration form as shown in the figure 4. User has to enter all the required details in the registration page. The email id will be considered as the primary identification of the user, based which the user will be given a unique user id. The user can set a secure password with format of one number, one uppercase, lowercase and at least 8 or more characters. After successful registration, the user account will be reviewed by IBDC for validity of the entered details and an approval will be sent for the activation of the account to the registered E-mail if all the details are valid.



User Registration

Field marked * are mandatory

Login details

Email*: Email

Password*:

Must contain at least one number and one uppercase and lowercase letter, and at least 8 or more characters

Confirm Password*:

Main Contact

First Name*: First name

Last Name*: Last Name

Other Details

Center Name*: Center name

Lab Name:

Country: ---select-----

State*:

Address*:

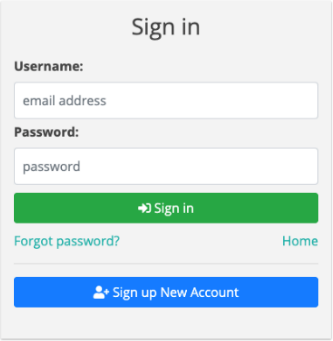
Consortium

□ Main Contact is a consortium not an individual

**Figure 4.** The snapshot of the registration page for registering a user account on INDA-CA portal.

## User Login

The user can login to the INDA-CA portal by clicking on the ‘Login’ button or ‘Submit Data’ button at the INDA-CA home page. User has to enter the registered email (Username) and the password (set by the user) in the Sign in page (Figure 5) to login into his INDA-CA account.



Sign in

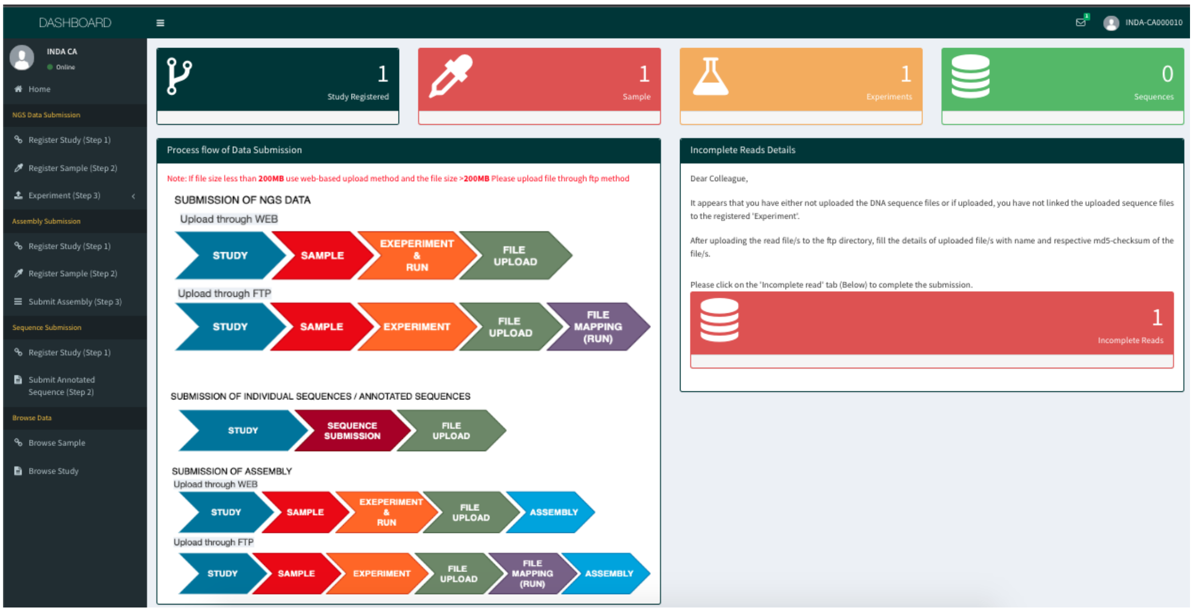
Username: email address

Password: password

**Figure 5.** Snapshot of login page of INDA-CA.

## User Dashboard

On successful login, the users are directed to their dashboard page, which provide the user with the various data submission services, summary of data flow to guide them through the submission and their data upload summary. On the top right corner of the user-dashboard page, user’s unique IBDC_ID button is present (Figure 6). On clicking the IBDC user ID, profile and sign out options are given. The profile page presents the personal and other important information required for data upload via FTP.



Note: If file size less than **200MB** use web-based upload method and the file size >**200MB** Please upload file through ftp method

Dear Colleague,

It appears that you have either not uploaded the DNA sequence files or if uploaded, you have not linked the uploaded sequence files to the registered 'Experiment'.

After uploading the read file/s to the ftp directory, fill the details of uploaded file/s with name and respective md5-checksum of the file/s.

Please click on the 'Incomplete read' tab (Below) to complete the submission.

**Figure 6.** Snapshot of user’s Dashboard page of INDA-CA.

## Metadata Model in INDA-CA

Users are advised to go through the metadata model adopted by INDA-CA before proceeding with the submission to understand which metadata object can represent what part of user's research project.

### Metadata model

**Study**: Study is defined as an entity/object, which helps in grouping the data submitted to the data archive and controls all the associated data. A study accession is a unique_id which used when citing data submitted to data archive.

**Sample**: Sample comprises of the information regarding the sequenced source of material. Samples are associated with group of list, which define the various parameters of the sample known as checklists. The parameters in the checklists will help in annotating the sample clearly. The registered samples are associated with an organism specifically called as taxon. The taxon is referenced from INSDC taxon identifiers.

**Experiment:** An experiment is an object, which contains all the information regarding a sequencing experiment including the library and instrument details.

**Run:** A run is part of experiment, which refers to data files containing sequencing reads.

**Targeted (Individual) Sequences:** Targeted sequences are the submission of individual sequences obtained from the sequencing experiments. Some of the examples of sequence types are cDNA, rRNA, Satellite DNA etc.

**Assembly:** The arrangement of nucleotide sequences in a correct order obtained from the sequencing raw data.

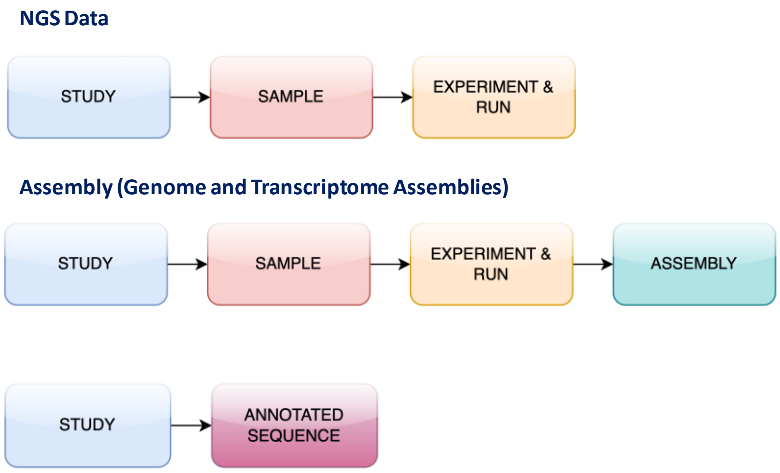


**Figure 7. Metadata objects used in different INDA-CA submission services**

Kindly go through the examples shown in the figure 8 to get an overview on how to use and register the different objects for your research.

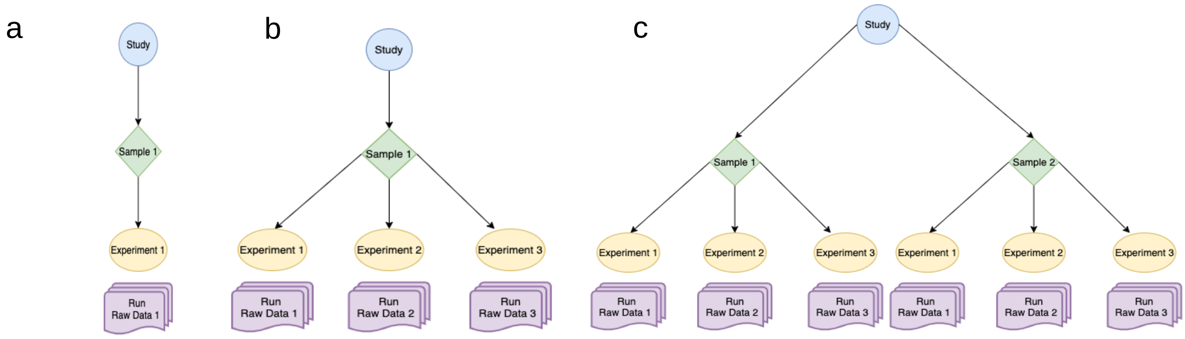
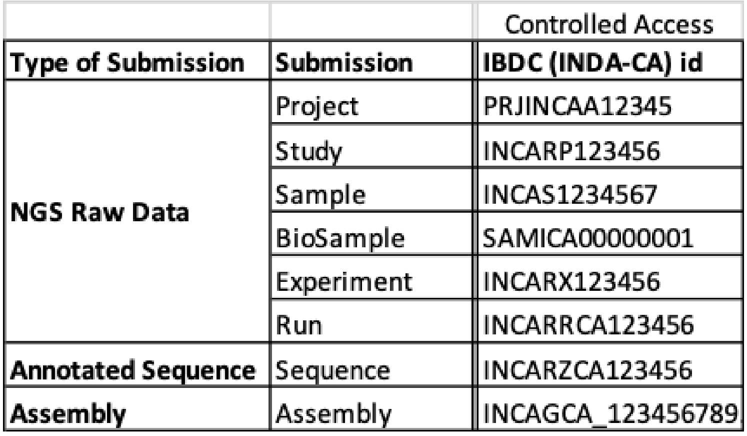


**Figure 8.** Examples showing the (a) study with single sample and experiments, (b) study with single sample but multiple experiments (c) study with multiple sample and experiments.

## Accessions

The accession for the submission will be provided individually. Every type of submission is given a specific format of accession. Data will be provided with IBDC-INDA-CA accessions.



| | | Controlled Access |
|---|---|---|
| **Type of Submission** | **Submission** | **IBDC (INDA-CA) id** |
| **NGS Raw Data** | Project | PRJINCAA12345 |
| | Study | INCARP123456 |
| | Sample | INCAS1234567 |
| | BioSample | SAMICA00000001 |
| | Experiment | INCARX123456 |
| | Run | INCARRCA123456 |
| **Annotated Sequence** | Sequence | INCARZCA123456 |
| **Assembly** | Assembly | INCAGCA_123456789 |

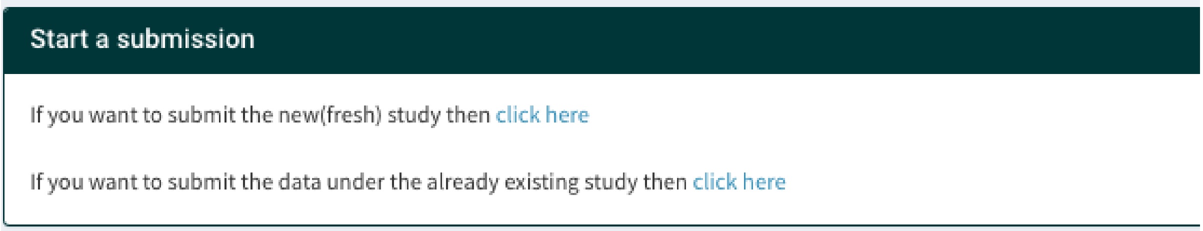
**Table 2.** The example showing the different types of accessions users will get after successful submission.

## Data Submission steps: NGS data submission

INDA-CA offers data submission via Web-Based mode, which can be completed by filling the web forms directly in your browser. Data submission steps required to submit nucleotide data to INDA-CA are given below

### Step 1: Study Registration

Study is defined as an entity/object, which helps in grouping the data submitted to the data archive. Every data submission requires registration of a study/project as the first step. This step is the most critical part of submission, so users are advised to provide sufficient details in the respective fields reflecting good overview of their research project.



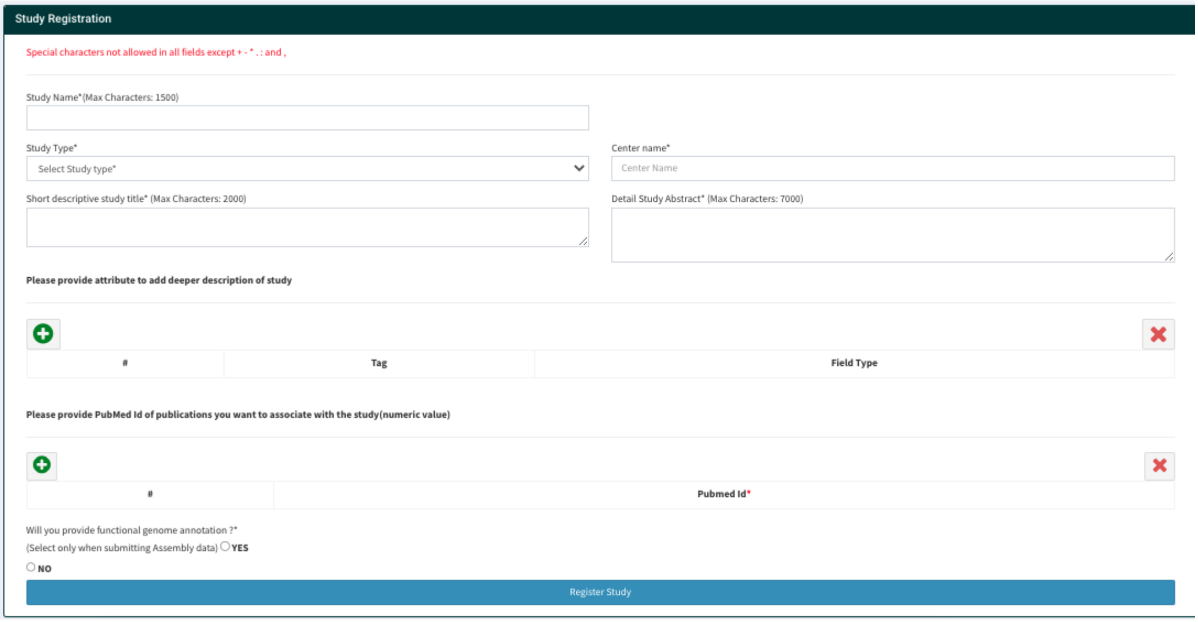
Start a submission

If you want to submit the new(fresh) study then click here

If you want to submit the data under the already existing study then click here

**Figure 9**. Pre-Study page options

Study registration page consists of fields which define the study type, title, description, Centre name, Study details and Study Abstract. It also has fields like tag entity and pubmed_id which helps in more description of the study and it can be of multiple numbers. If submitting assembly annotation, Annotation field has to be selected with YES option and provide a annotation term. The study registered and its associated data will not be made public until the user intends to.



Study Registration

Special characters not allowed in all fields except + - * . : and ,

Study Name*(Max Characters: 1500)

Study Type*

Select Study type*

Center name*

Center Name

Short descriptive study title* (Max Characters: 2000)

Detail Study Abstract* (Max Characters: 7000)

Please provide attribute to add deeper description of study

| # | Tag | Field Type |
|---|---|---|

Please provide PubMed Id of publications you want to associate with the study(numeric value)

| # | Pubmed Id* |
|---|---|

Will you provide functional genome annotation ?*
(Select only when submitting Assembly data) ○ YES
○ NO

**Figure 9.** Study registration page of INDA-CA

Study type field is a dropdown type of field which has multiple options to select from. Below are the different option and their details that user has to select from the dropdown.

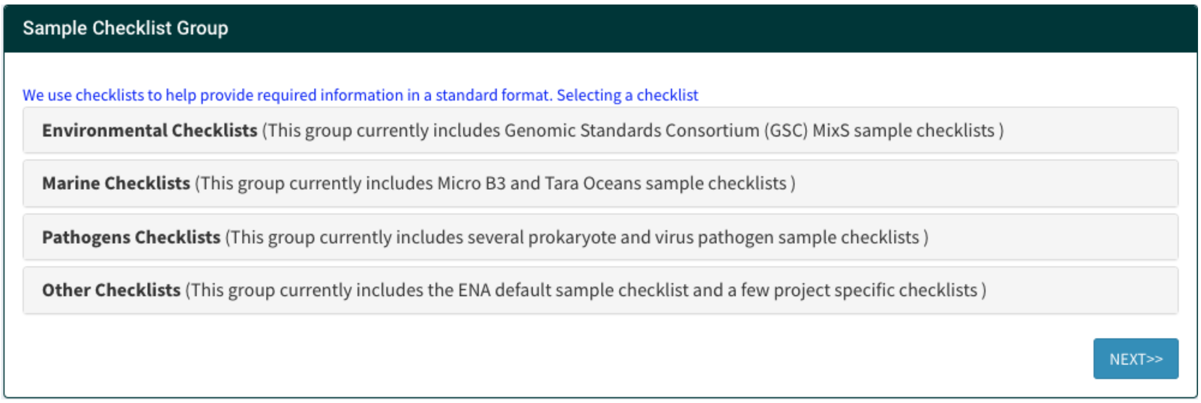
Study Type Options

| | |
|---|---|
| ♦ Whole Genome Sequencing | Sequencing of a single organism |
| ♦ Metagenomics | Sequencing of a community |
| ♦ Transcriptome Analysis | Sequencing and Characterisation of transcription elements |
| ♦ Resequencing | Sequencing of a sample with respect to a reference |
| ♦ Epigenetics | Cellular differentiation study |
| ♦ Synthetic Genomics | Sequencing of modified, synthetic, or transplanted genomes |
| ♦ Forensic or Paleo-genomics | Sequencing of recovered genomic material |
| ♦ Gene Regulation Study | Study of Gene Expression Regulation |
| ♦ Cancer Genomics | Study of cancer genomics |
| ♦ Population genomics | Study of populations and evolution through genomics |
| ♦ RNASeq | RNA Sequencing study |
| ♦ Exam Sequencing | The study investigates the axons of the genome |
| ♦ Pooled Clone Sequencing | The study is sequencing clone pools (BACs, fosmids, other constructs). |
| ♦ Transcriptome sequencing | Sequencing of transcription elements |

- Other Study type not listed

### Step 2: Sample Registration

Sample denotes the biomaterial, which is the source of origin for your sequencing data. Therefore, it is important to define the metadata associated with sample as extensive and accurate as possible. Ideally, user should register one sample for each biological replicate. To ease the metadata annotation simple 'sample checklist' are provided for users to select and tick the most relevant options. The detailed regarding the groups and their child classes are shown below in the snapshots.



Sample Checklist Group

We use checklists to help provide required information in a standard format. Selecting a checklist

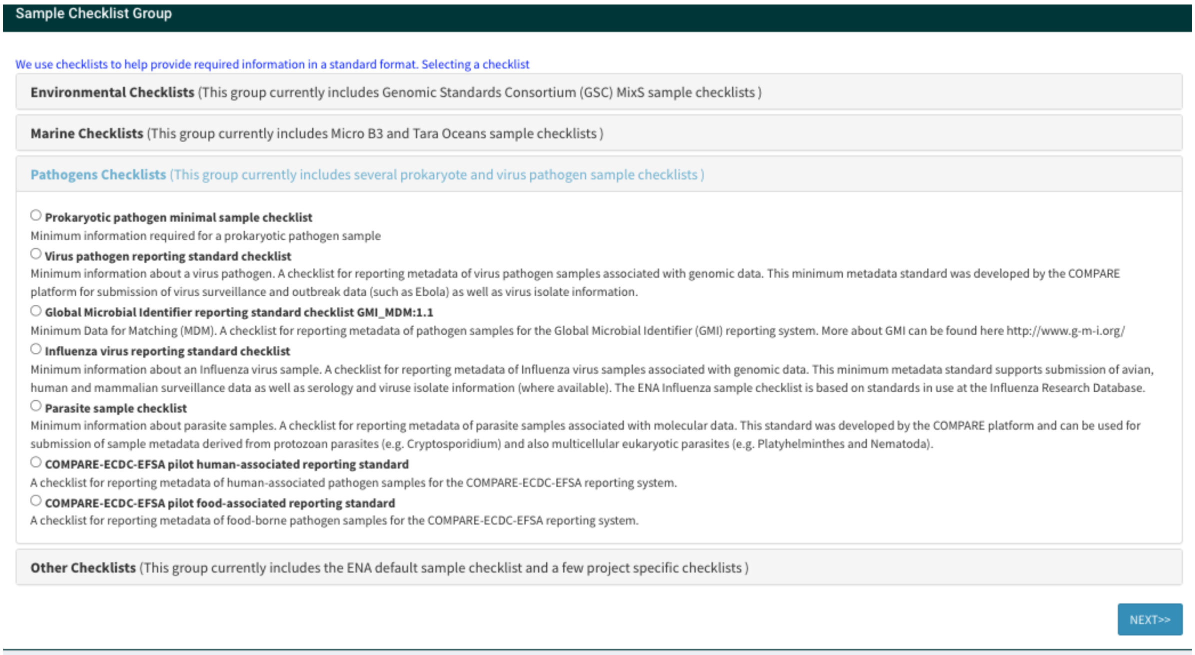
**Environmental Checklists** (This group currently includes Genomic Standards Consortium (GSC) MixS sample checklists )

**Marine Checklists** (This group currently includes Micro B3 and Tara Oceans sample checklists )

**Pathogens Checklists** (This group currently includes several prokaryote and virus pathogen sample checklists )

**Other Checklists** (This group currently includes the ENA default sample checklist and a few project specific checklists )

**Figure 10**. Snapshot showing the sample checklist groups.



Sample Checklist Group

We use checklists to help provide required information in a standard format. Selecting a checklist

**Environmental Checklists** (This group currently includes Genomic Standards Consortium (GSC) MixS sample checklists )

**Marine Checklists** (This group currently includes Micro B3 and Tara Oceans sample checklists )

**Pathogens Checklists** (This group currently includes several prokaryote and virus pathogen sample checklists )

○ **Prokaryotic pathogen minimal sample checklist**
Minimum information required for a prokaryotic pathogen sample

○ **Virus pathogen reporting standard checklist**
Minimum information about a virus pathogen. A checklist for reporting metadata of virus pathogen samples associated with genomic data. This minimum metadata standard was developed by the COMPARE platform for submission of virus surveillance and outbreak data (such as Ebola) as well as virus isolate information.

○ **Global Microbial Identifier reporting standard checklist GMI_MDM:1.1**
Minimum Data for Matching (MDM). A checklist for reporting metadata of pathogen samples for the Global Microbial Identifier (GMI) reporting system. More about GMI can be found here http://www.g-m-i.org/

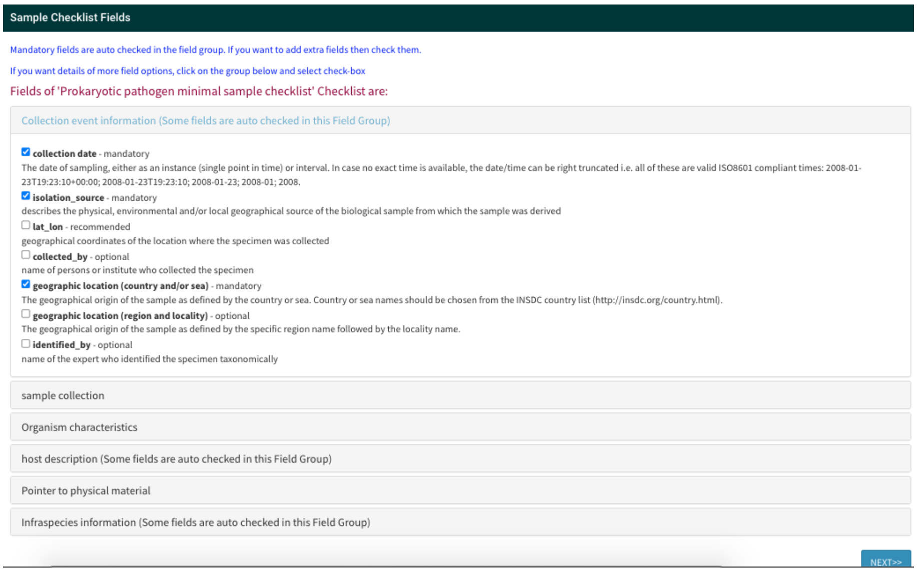
○ **Influenza virus reporting standard checklist**
Minimum information about an Influenza virus sample. A checklist for reporting metadata of Influenza virus samples associated with genomic data. This minimum metadata standard supports submission of avian, human and mammalian surveillance data as well as serology and viruse isolate information (where available). The ENA Influenza sample checklist is based on standards in use at the Influenza Research Database.

○ **Parasite sample checklist**
Minimum information about parasite samples. A checklist for reporting metadata of parasite samples associated with molecular data. This standard was developed by the COMPARE platform and can be used for submission of sample metadata derived from protozoan parasites (e.g. Cryptosporidium) and also multicellular eukaryotic parasites (e.g. Platyhelminthes and Nematoda).

○ **COMPARE-ECDC-EFSA pilot human-associated reporting standard**
A checklist for reporting metadata of human-associated pathogen samples for the COMPARE-ECDC-EFSA reporting system.

○ **COMPARE-ECDC-EFSA pilot food-associated reporting standard**
A checklist for reporting metadata of food-borne pathogen samples for the COMPARE-ECDC-EFSA reporting system.

**Other Checklists** (This group currently includes the ENA default sample checklist and a few project specific checklists )



Sample Checklist Fields

Mandatory fields are auto checked in the field group. If you want to add extra fields then check them.

If you want details of more field options, click on the group below and select check-box

Fields of 'Prokaryotic pathogen minimal sample checklist' Checklist are:

Collection event information (Some fields are auto checked in this Field Group)

☑ **collection date** - mandatory
The date of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid ISO8601 compliant times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008.

☑ **isolation_source** - mandatory
describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived

□ **lat_lon** - recommended
geographical coordinates of the location where the specimen was collected

□ **collected_by** - optional
name of persons or institute who collected the specimen

☑ **geographic location (country and/or sea)** - mandatory
The geographical origin of the sample as defined by the country or sea. Country or sea names should be chosen from the INSDC country list (http://insdc.org/country.html).

□ **geographic location (region and locality)** - optional
The geographical origin of the sample as defined by the specific region name followed by the locality name.

□ **identified_by** - optional
name of the expert who identified the specimen taxonomically

sample collection

Organism characteristics

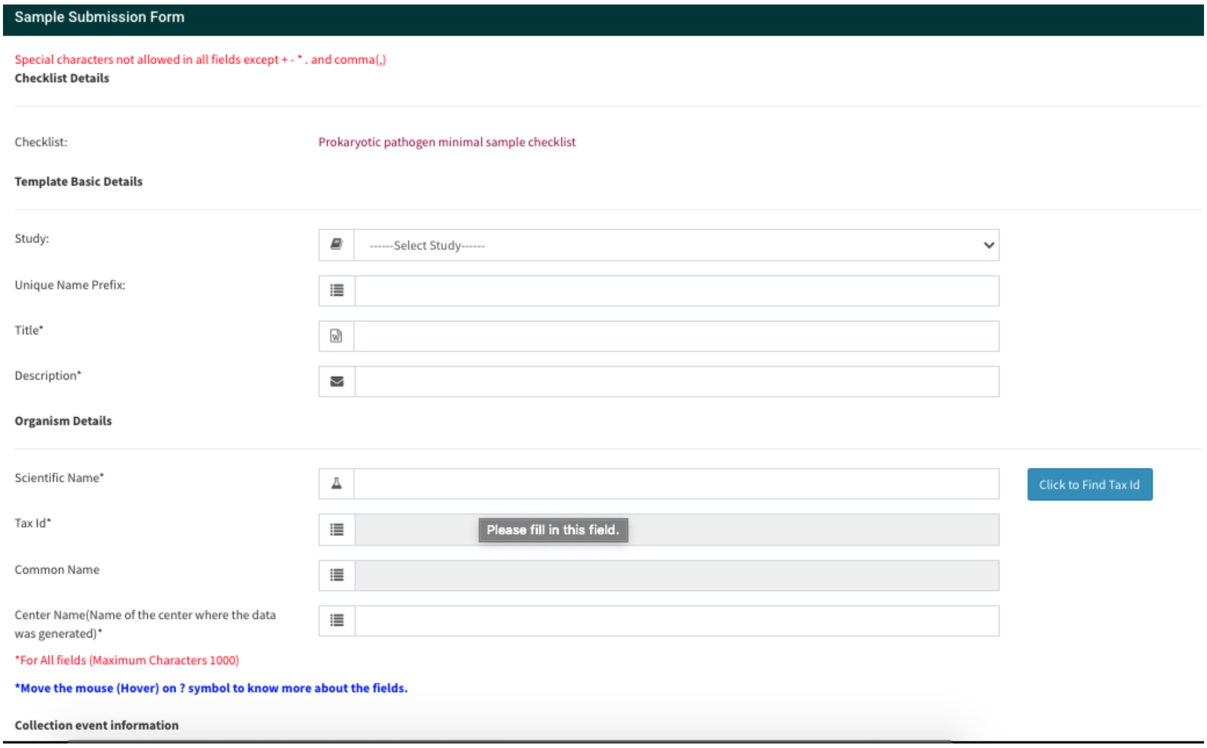
host description (Some fields are auto checked in this Field Group)

Pointer to physical material

Infraspecies information (Some fields are auto checked in this Field Group)

**Figure 11.** Selection of sample checklist classes.

Please use default sample checklist if no other options are relevant to your sample. Once the sample checklist is selected, sample entry form will be generated based on your selection. Carefully fill out all the fields and proceed with the submission. Once the samples are registered, users will be informed by email confirmation and now user will receive accessions for the samples.



Sample Submission Form

Special characters not allowed in all fields except + - * . and comma(,)

**Checklist Details**

Checklist: Prokaryotic pathogen minimal sample checklist

**Template Basic Details**

Study: ------Select Study------

Unique Name Prefix:

Title*

Description*

**Organism Details**

Scientific Name*

Tax Id*

Common Name

Center Name(Name of the center where the data was generated)*

*For All fields (Maximum Characters 1000)

***Move the mouse (Hover) on ? symbol to know more about the fields.**

**Collection event information**

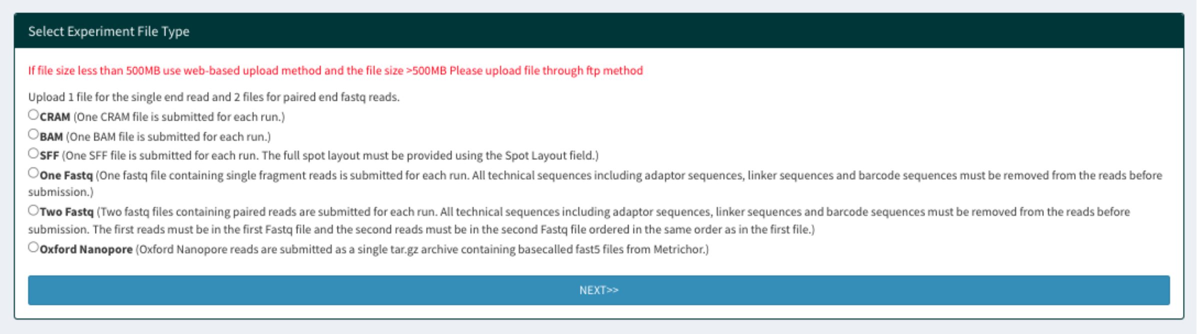
**Figure 12.** Sample field entry page of sample registration.

## Step 3: Experiment Registration

The final step of the NGS data submission is registering the experiments. But before that its advisable to prepare your files as per requirement e.g. correct format, md5 checksum for each file, adapter trimming, file compression etc. Therefore, follow the steps below to process your files before submission:

- Ensure the reads are good quality and free of adapter sequences
- Compress the file using gzip or bzip
- Calculate the md5 checksum of file

Click on the ‘Experiment’ and select either ‘register and upload via WEB’ or ‘Register then upload via FTP’ depending upon the size of your data files. After that the interface showing the ‘Experiment file type’ will be displayed and user has to select the applicable file format depending upon their study (Figure 12).



Select Experiment File Type

If file size less than 500MB use web-based upload method and the file size >500MB Please upload file through ftp method

Upload 1 file for the single end read and 2 files for paired end fastq reads.

○ **CRAM** (One CRAM file is submitted for each run.)

○ **BAM** (One BAM file is submitted for each run.)

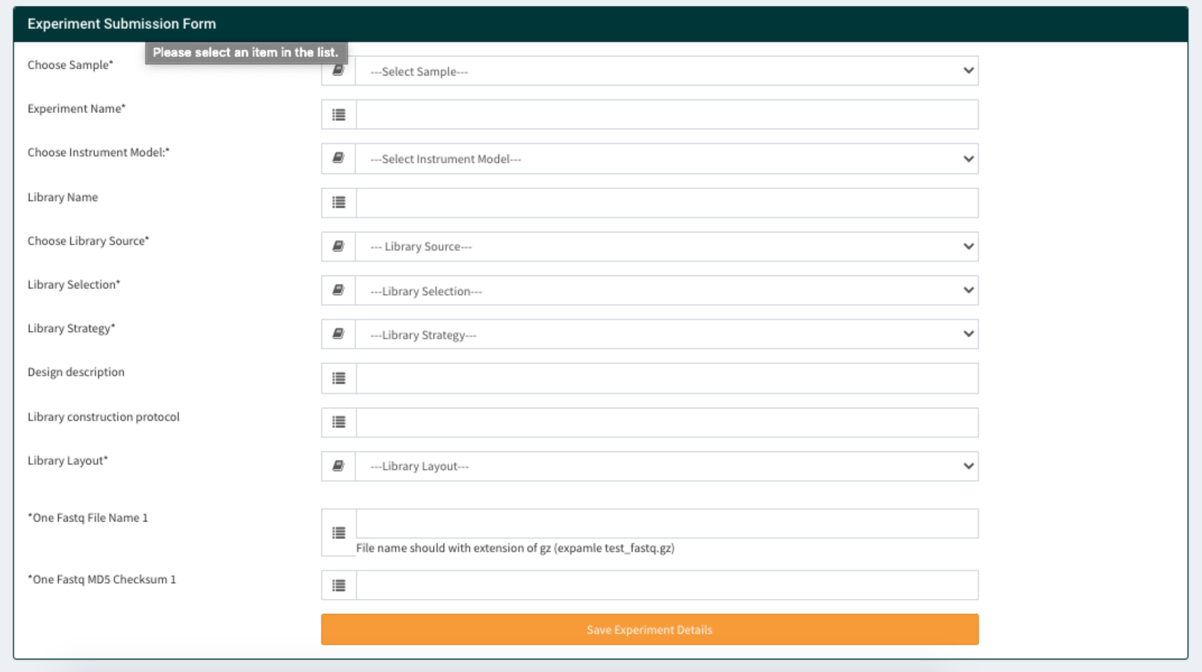
○ **SFF** (One SFF file is submitted for each run. The full spot layout must be provided using the Spot Layout field.)

○ **One Fastq** (One fastq file containing single fragment reads is submitted for each run. All technical sequences including adaptor sequences, linker sequences and barcode sequences must be removed from the reads before submission.)

○ **Two Fastq** (Two fastq files containing paired reads are submitted for each run. All technical sequences including adaptor sequences, linker sequences and barcode sequences must be removed from the reads before submission. The first reads must be in the first Fastq file and the second reads must be in the second Fastq file ordered in the same order as in the first file.)

○ **Oxford Nanopore** (Oxford Nanopore reads are submitted as a single tar.gz archive containing basecalled fast5 files from Metrichor.)

**Figure 13.** Pre experiment stage to select different file types for the raw read submission.



Experiment Submission Form

| | |
|---|---|
| Choose Sample* | ---Select Sample--- |
| Experiment Name* | |
| Choose Instrument Model:* | ---Select Instrument Model--- |
| Library Name | |
| Choose Library Source* | --- Library Source--- |
| Library Selection* | ---Library Selection--- |
| Library Strategy* | ---Library Strategy--- |
| Design description | |
| Library construction protocol | |
| Library Layout* | ---Library Layout--- |
| *One Fastq File Name 1 | File name should with extension of gz (expamle test_fastq.gz) |
| *One Fastq MD5 Checksum 1 | |

**Figure 14.** Experiment and Run field entry page.

A number of options in the experiment and run registration page have dropdown selection and the dropdown option for the fields are given below. The guide here will help user to understand different aspects required for the submission before reaching the actual submission page.

Instrument model options

454 GS
454 GS 20
454 GS FLX
454 GS FLX+
454 GS FLX Titanium
454 GS Junior
HiSeq X Five
HiSeq X Ten
Illumina Genome Analyzer
Illumina Genome Analyzer II
Illumina Genome Analyzer IIx
Illumina HiScanSQ
Illumina HiSeq 1000
Illumina HiSeq 1500
Illumina HiSeq 2000
Illumina HiSeq 2500
Illumina HiSeq 3000
Illumina HiSeq 4000
Illumina iSeq 100
Illumina MiSeq
Illumina MiniSeq
Illumina NovaSeq 6000
NextSeq 500
NextSeq 550
PacBio RS
PacBio RS II
Sequel
Ion Torrent PGM
Ion Torrent Proton
Ion Torrent S5
Ion Torrent S5 XL
AB 3730xL Genetic Analyzer
AB 3730 Genetic Analyzer

AB 3500xL Genetic Analyzer

AB 3500 Genetic Analyzer

AB 3130xL Genetic Analyzer

AB 3130 Genetic Analyzer

AB 310 Genetic Analyzer

MinION

GridION

PromethION

BGISEQ-500

DNBSEQ-T7

DNBSEQ-G400

DNBSEQ-G50

DNBSEQ-G400 FAST

unspecified

## Library source options

- GENOMIC: Genomic DNA (includes PCR products from genomic DNA).
- GENOMIC SINGLE CELL:
- TRANSCRIPTOMIC: Transcription products or non-genomic DNA (EST, cDNA, RT-PCR, screened libraries).
- TRANSCRIPTOMIC SINGLE CELL:
- METAGENOMIC: Mixed material from metagenome.
- METATRANSCRIPTOMIC: Transcription products from community targets
- SYNTHETIC: Synthetic DNA.
- VIRAL RNA: Viral RNA.
- OTHER: Other, unspecified, or unknown library source material.

## Library selection options

RANDOM: No Selection or Random selection

- PCR: target enrichment via PCR
- RANDOM PCR: Source material was selected by randomly generated primers.
- RT-PCR: target enrichment via
- HMPR: Hypo-methylated partial restriction digest
- MF: Methyl Filtrated
- repeat fractionation: Selection for less repetitive (and more gene rich) sequence through Cot filtration (CF) or other fractionation techniques based on DNA kinetics.
- size fractionation: Physical selection of size appropriate targets.
- MSLL: Methylation Spanning Linking Library
- cDNA: PolyA selection or enrichment for messenger RNA (mRNA); synonymize with PolyA
- cDNA_randomPriming:
- cDNA_oligo_dT:
- PolyA: PolyA selection or enrichment for messenger RNA (mRNA); should replace cDNA enumeration.
- Oligo-dT: enrichment of messenger RNA (mRNA) by hybridization to Oligo-dT.
- Inverse rRNA: depletion of ribosomal RNA by oligo hybridization.
- Inverse rRNA selection: depletion of ribosomal RNA by inverse oligo hybridization.
- ChIP: Chromatin immunoprecipitation
- ChIP-Seq: Chromatin immunoPrecipitation, reveals binding sites of specific proteins, typically transcription factors (TFs) using antibodies to extract DNA fragments bound to the target protein.
- MNase: Identifies well-positioned nucleosomes. uses Micrococcal Nuclease (MNase) is an endo-exonuclease that processively digests DNA until an obstruction, such as a nucleosome, is reached.
- DNase: DNase I endonuclease digestion and size selection reveals regions of chromatin where the DNA is highly sensitive to DNase I.
- Hybrid Selection: Selection by hybridization in array or solution.
- Reduced Representation: Reproducible genomic subsets, often generated by restriction fragment size selection, containing a manageable number of loci to facilitate re-sampling.
- Restriction Digest: DNA fractionation using restriction enzymes.

- 5-methylcytidine antibody: Selection of methylated DNA fragments using an antibody raised against 5-methylcytosine or 5-methylcytidine (m5C).
- MBD2 protein methyl-CpG binding domain: Enrichment by methyl-CpG binding domain.
- CAGE: Cap-analysis gene expression.
- RACE: Rapid Amplification of cDNA Ends.
- MDA: Multiple Displacement Amplification, a non-PCR based DNA amplification technique that amplifies a minute quantifies of DNA to levels suitable for genomic analysis.
- padlock probes capture method: Targeted sequence capture protocol covering an arbitrary set of nonrepetitive genomics targets. An example is capture bisulfite sequencing using padlock probes (BSPP).
- other: Other library enrichment, screening, or selection process.
- unspecified: Library enrichment, screening, or selection is not specified

## Library strategy options

- WGS: Whole Genome Sequencing - random sequencing of the whole genome (see pubmed 10731132 for details)
- WGA: Whole Genome Amplification followed by random sequencing. (see pubmed 1631067,8962113 for details)
- WXS: Random sequencing of exonic regions selected from the genome. (see pubmed 20111037 for details)
- RNA-Seq: Random sequencing of whole transcriptome, also known as Whole Transcriptome Shotgun Sequencing, or WTSS). (see pubmed 18611170 for details)
- ssRNA-seq: Strand-specific RNA sequencing.
- miRNA-Seq: Micro RNA sequencing strategy designed to capture post-transcriptional RNA elements and include non-coding functional elements. (see pubmed 21787409 for details)
- ncRNA-Seq: Capture of other non-coding RNA types, including post-translation modification types such as snRNA (small nuclear RNA) or snoRNA (small nucleolar RNA), or expression regulation types such as siRNA (small interfering RNA) or piRNA/piwi/RNA (piwi-interacting RNA).
- FL-cDNA: Full-length sequencing of cDNA templates

- EST: Single pass sequencing of cDNA templates
- Hi-C: Chromosome Conformation Capture technique where a biotin-labeled nucleotide is incorporated at the ligation junction, enabling selective purification of chimeric DNA ligation junctions followed by deep sequencing.
- ATAC-seq: Assay for Transposase-Accessible Chromatin (ATAC) strategy is used to study genome-wide chromatin accessibility. alternative method to DNase-seq that uses an engineered Tn5 transposase to cleave DNA and to integrate primer DNA sequences into the cleaved genomic DNA.
- WCS: Random sequencing of a whole chromosome or other replicon isolated from a genome.
- RAD-Seq:
- CLONE: Genomic clone based (hierarchical) sequencing.
- POOLCLONE: Shotgun of pooled clones (usually BACs and Fosmids).
- AMPLICON: Sequencing of overlapping or distinct PCR or RT-PCR products. For example, metagenomic community profiling using SSU rRNA.
- CLONEEND: Clone end (5’, 3’, or both) sequencing.
- FINISHING: Sequencing intended to finish (close) gaps in existing coverage.
- ChIP-Seq: ChIP-seq, Chromatin ImmunoPrecipitation, reveals binding sites of specific proteins, typically transcription factors (TFs) using antibodies to extract DNA fragments bound to the target protein.
- MNase-Seq: Identifies well-positioned nucleosomes. uses Micrococcal Nuclease (MNase) is an endo-exonuclease that processively digests DNA until an obstruction, such as a nucleosome, is reached.
- DNase-Hypersensitivity: Sequencing of hypersensitive sites, or segments of open chromatin that are more readily cleaved by DNaseI.
- Bisulfite-Seq: MethylC-seq. Sequencing following treatment of DNA with bisulfite to convert cytosine residues to uracil depending on methylation status.
- CTS: Concatenated Tag Sequencing
- MRE-Seq: Methylation-Sensitive Restriction Enzyme Sequencing.
- MeDIP-Seq: Methylated DNA Immunoprecipitation Sequencing.
- MBD-Seq: Methyl CpG Binding Domain Sequencing.
- Tn-Seq: Quantitatively determine fitness of bacterial genes based on how many times a purposely seeded transposon gets inserted into each gene of a colony after some time.

- VALIDATION: CGHub special request: Independent experiment to re-evaluate putative variants.
- FAIRE-seq: Formaldehyde Assisted Isolation of Regulatory Elements. Reveals regions of open chromatin.
- SELEX: Systematic Evolution of Ligands by Exponential enrichment
- RIP-Seq: Direct sequencing of RNA immunoprecipitates (includes CLIP-Seq, HITS-CLIP and PAR-CLIP).
- ChIA-PET: Direct sequencing of proximity-ligated chromatin immunoprecipitates.
- Synthetic-Long-Read: binning and barcoding of large DNA fragments to facilitate assembly of the fragment
- Targeted-Capture: Enrichment of a targeted subset of loci.
- Tethered Chromatin Conformation Capture:
- OTHER: Library strategy not listed.

### Library layout options

- Paired
- Single

After submitting the metadata form associated with the experiment details, now user has to upload the data files as per data upload option selected.

- Web Submission

  Web based raw data submission accepts file only up to 500 Mb and if the file size is more than 500 Mb it has to be submitted via ftp mode.

- FTP Submission

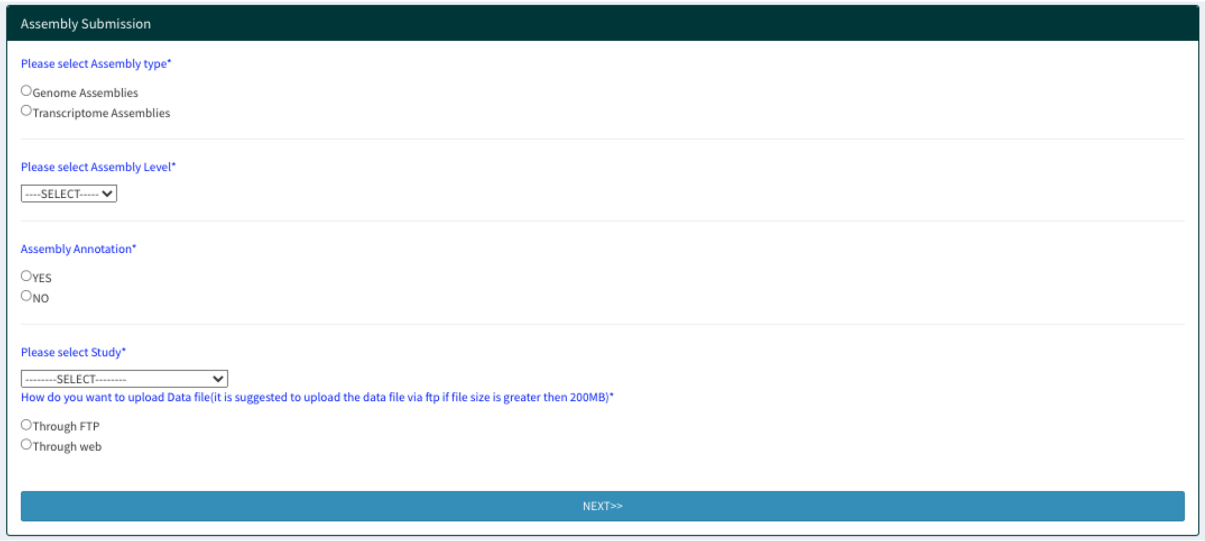
  FTP submission is tailor made for the raw data submission where the file size is greater than 500 Mb. The raw data file has to be uploaded in the directory corresponding to study registered, sample and experiment.

## Assembly Submission Steps

All genome and transcriptome assemblies are submitted through the ‘Assembly submission’ by following steps given below:

- Register Study: Same as step 1 of NGS data submission.
- Register Sample: Same as step 2 of NGS data submission.

- Submit Assembly: This step requires information regarding the assembly type, level and annotation. Here again depending upon the size of the assembly, users are provided with FTP or Web based submission of files.



Assembly Submission

Please select Assembly type*

○Genome Assemblies
○Transcriptome Assemblies

Please select Assembly Level*

----SELECT-----

Assembly Annotation*

○YES
○NO

Please select Study*

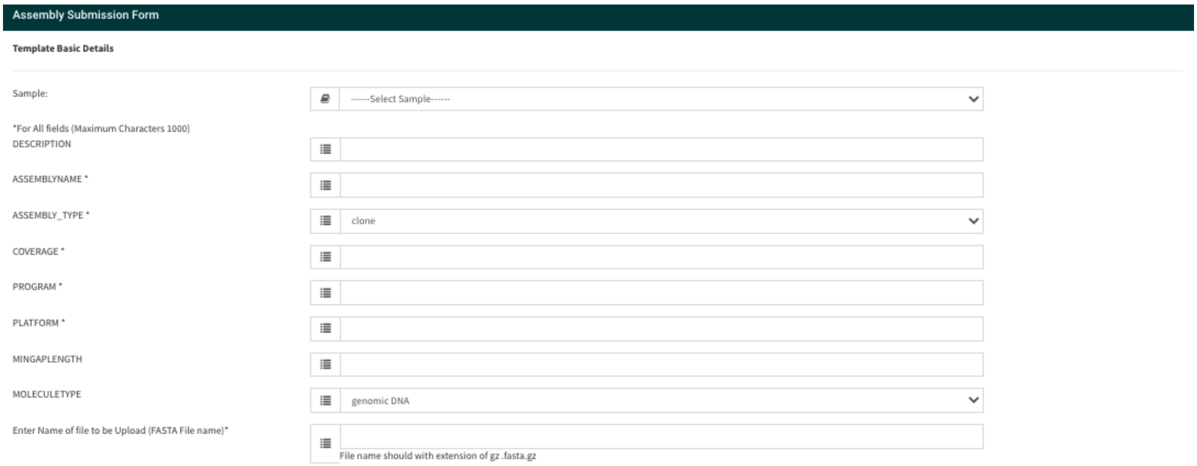
--------SELECT--------

How do you want to upload Data file(it is suggested to upload the data file via ftp if file size is greater then 200MB)*

○Through FTP
○Through web

**Figure 25.** Assembly submission types available for selection

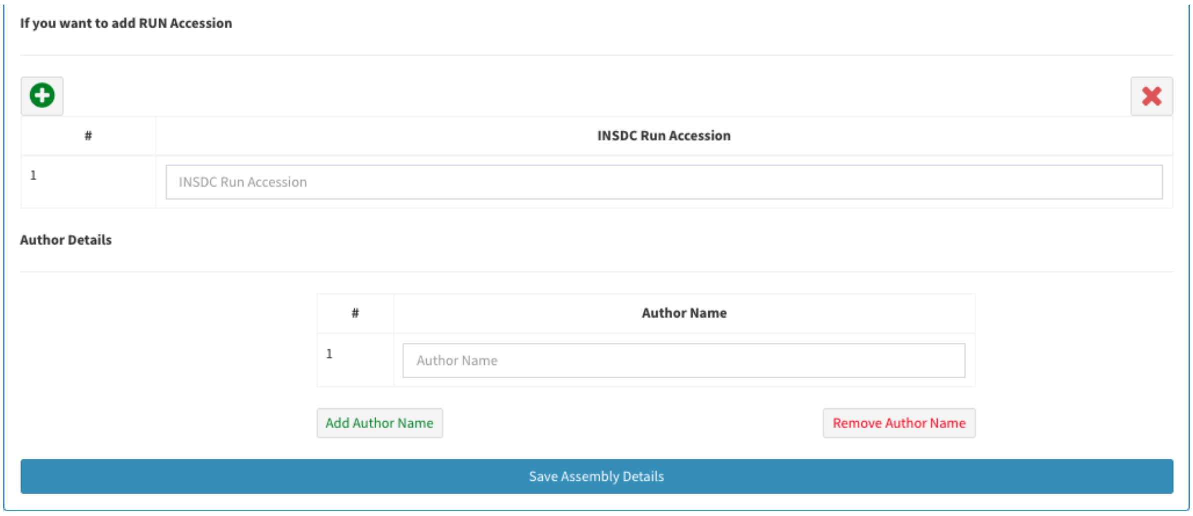
Then user has to provide the relevant details regarding the associated sample, description, name, type, coverage etc including the authors to extensively record the metadata. As you have noticed that the assembly submission requires the reference of study and sample objects, one must submit these before proceeding with assembly submission. But if user want to submit the assembly in an already submitted project, then he can directly proceed with step 3.



Assembly Submission Form

Template Basic Details

| | |
|---|---|
| Sample: | ------Select Sample------ |
| *For All fields (Maximum Characters 1000)<br>DESCRIPTION | |
| ASSEMBLYNAME * | |
| ASSEMBLY_TYPE * | clone |
| COVERAGE * | |
| PROGRAM * | |
| PLATFORM * | |
| MINGAPLENGTH | |
| MOLECULETYPE | genomic DNA |
| Enter Name of file to be Upload (FASTA File name)* | File name should with extension of gz .fasta.gz |



If you want to add RUN Accession

| # | INSDC Run Accession |
|---|---|
| 1 | INSDC Run Accession |

Author Details

| # | Author Name |
|---|---|
| 1 | Author Name |

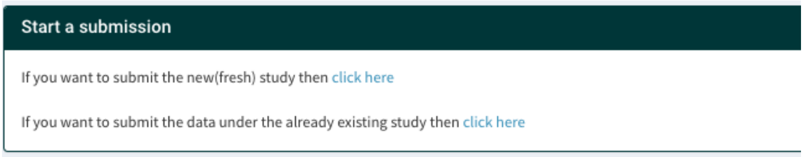
**Figure 16**. Assembly fields entry page.

## Sequence submission steps

This section deals with the submission of short assembled and annotated sequences eg single gene sequence. This submission contains two steps:

- Register Study: same as step 1 of NGS data submission
- Register Sample

To initiate the submission user has to click on to the 'Register Study' and then specify whether he is making new submission or want to contribute in already registered study.  Then to properly record the metadata associated with the sequence pre-defined sequence checklist group and fields are given and user has to select the most closely relatable options from the list.



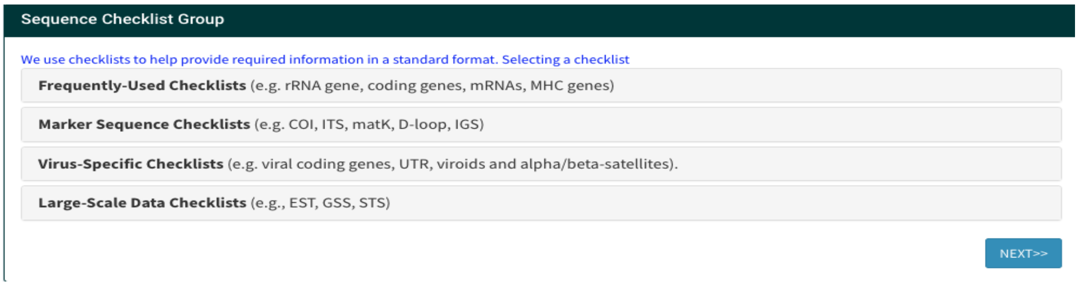
Start a submission

If you want to submit the new(fresh) study then click here

If you want to submit the data under the already existing study then click here

**Figure 17.**  Pre – Sequence Submission page

### Sequence checklist group



Sequence Checklist Group

We use checklists to help provide required information in a standard format. Selecting a checklist

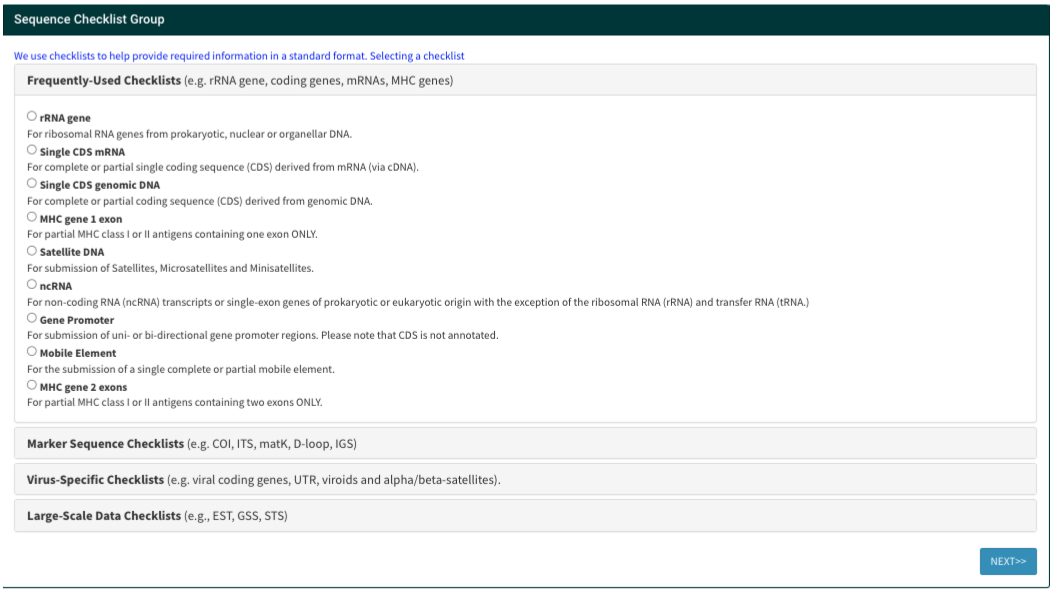
**Frequently-Used Checklists** (e.g. rRNA gene, coding genes, mRNAs, MHC genes)

**Marker Sequence Checklists** (e.g. COI, ITS, matK, D-loop, IGS)

**Virus-Specific Checklists** (e.g. viral coding genes, UTR, viroids and alpha/beta-satellites).

**Large-Scale Data Checklists** (e.g., EST, GSS, STS)

**Figure 38.** Sequence checklist group selection page



Sequence Checklist Group

We use checklists to help provide required information in a standard format. Selecting a checklist

**Frequently-Used Checklists** (e.g. rRNA gene, coding genes, mRNAs, MHC genes)

○ **rRNA gene**
For ribosomal RNA genes from prokaryotic, nuclear or organellar DNA.
○ **Single CDS mRNA**
For complete or partial single coding sequence (CDS) derived from mRNA (via cDNA).
○ **Single CDS genomic DNA**
For complete or partial coding sequence (CDS) derived from genomic DNA.
○ **MHC gene 1 exon**
For partial MHC class I or II antigens containing one exon ONLY.
○ **Satellite DNA**
For submission of Satellites, Microsatellites and Minisatellites.
○ **ncRNA**
For non-coding RNA (ncRNA) transcripts or single-exon genes of prokaryotic or eukaryotic origin with the exception of the ribosomal RNA (rRNA) and transfer RNA (tRNA.)
○ **Gene Promoter**
For submission of uni- or bi-directional gene promoter regions. Please note that CDS is not annotated.
○ **Mobile Element**
For the submission of a single complete or partial mobile element.
○ **MHC gene 2 exons**
For partial MHC class I or II antigens containing two exons ONLY.

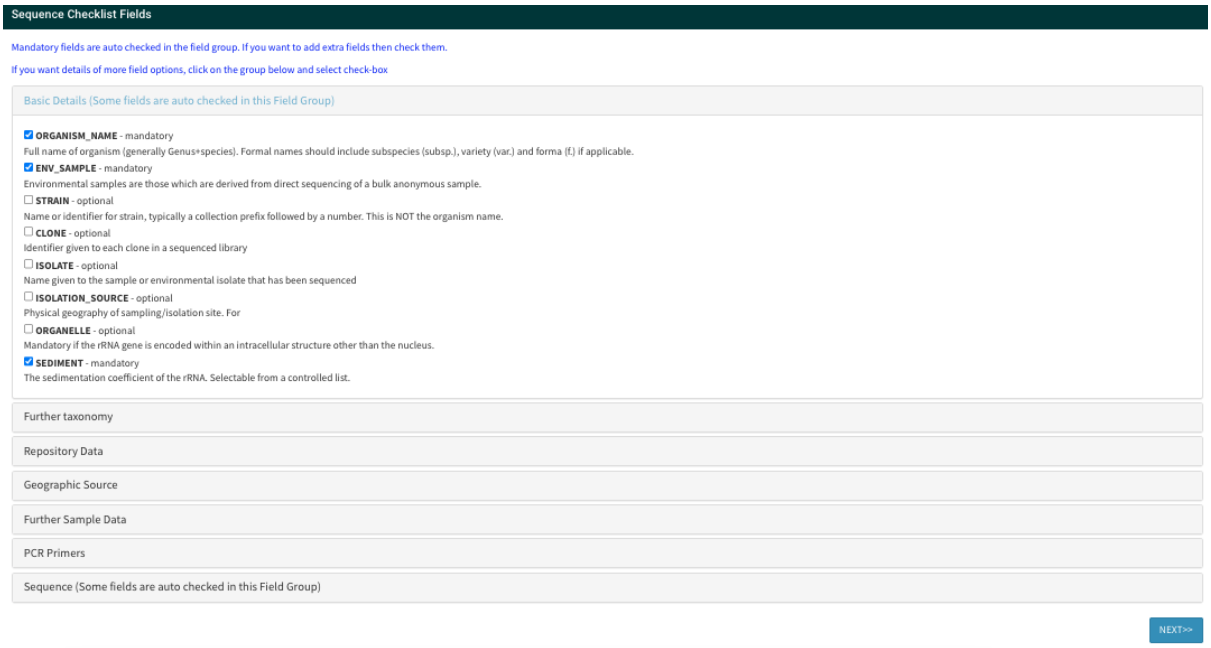
**Marker Sequence Checklists** (e.g. COI, ITS, matK, D-loop, IGS)

**Virus-Specific Checklists** (e.g. viral coding genes, UTR, viroids and alpha/beta-satellites).

**Large-Scale Data Checklists** (e.g., EST, GSS, STS)

**Figure 194.** Sequence checklist selection page

Mandatory fields will be auto selected and the other fields can be selected if the user wants the fields to be included for the submission.



Sequence Checklist Fields

Mandatory fields are auto checked in the field group. If you want to add extra fields then check them.

If you want details of more field options, click on the group below and select check-box

Basic Details (Some fields are auto checked in this Field Group)

☑ **ORGANISM_NAME** - mandatory
Full name of organism (generally Genus+species). Formal names should include subspecies (subsp.), variety (var.) and forma (f.) if applicable.
☑ **ENV_SAMPLE** - mandatory
Environmental samples are those which are derived from direct sequencing of a bulk anonymous sample.
□ **STRAIN** - optional
Name or identifier for strain, typically a collection prefix followed by a number. This is NOT the organism name.
□ **CLONE** - optional
Identifier given to each clone in a sequenced library
□ **ISOLATE** - optional
Name given to the sample or environmental isolate that has been sequenced
□ **ISOLATION_SOURCE** - optional
Physical geography of sampling/isolation site. For
□ **ORGANELLE** - optional
Mandatory if the rRNA gene is encoded within an intracellular structure other than the nucleus.
☑ **SEDIMENT** - mandatory
The sedimentation coefficient of the rRNA. Selectable from a controlled list.

Further taxonomy

Repository Data

Geographic Source

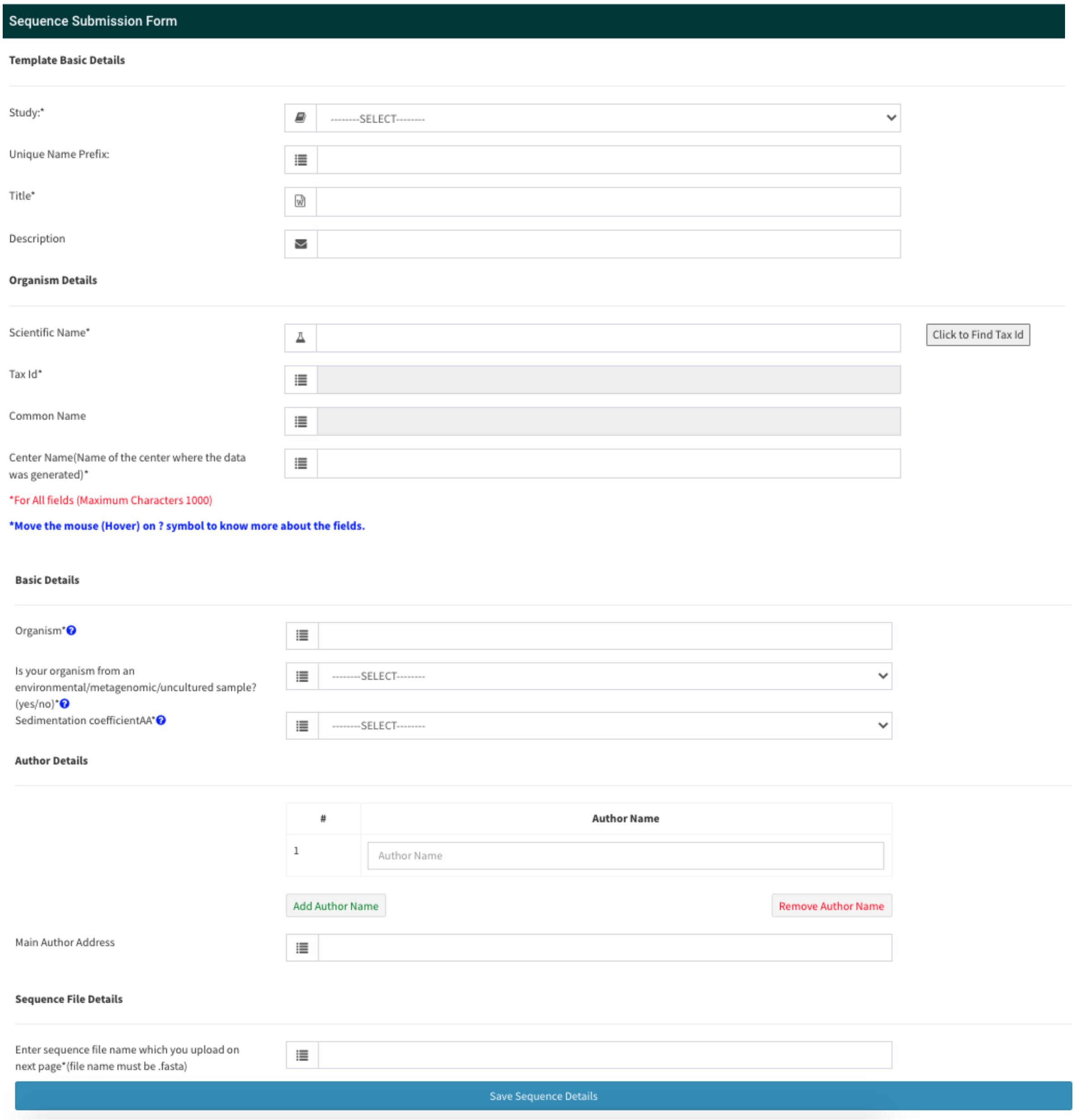
Further Sample Data

PCR Primers

Sequence (Some fields are auto checked in this Field Group)

**Figure 20.** Sequence field selection page

Then user need fill out the auto-generated form, based on his selection from the checklist, with relevant details along with other information and proceed with sequence upload in fasta format once the data is validated accessions will be provided by IBDC.



Sequence Submission Form

Template Basic Details

Study:* --------SELECT--------

Unique Name Prefix:

Title*

Description

Organism Details

Scientific Name* Click to Find Tax Id

Tax Id*

Common Name

Center Name(Name of the center where the data was generated)*

*For All fields (Maximum Characters 1000)

*Move the mouse (Hover) on ? symbol to know more about the fields.

Basic Details

Organism*?

Is your organism from an environmental/metagenomic/uncultured sample? (yes/no)*? --------SELECT--------

Sedimentation coefficientAA*? --------SELECT--------

Author Details

| # | Author Name |
|---|---|
| 1 | Author Name |

Add Author Name Remove Author Name

Main Author Address

Sequence File Details

Enter sequence file name which you upload on next page*(file name must be .fasta)

Save Sequence Details

**Figure 21**5. Sequence field entry page.