Guia de Submissão para BioProject, BioSample, SRA e GEO

> Este documento explica como preparar os metadados necessários para submissão de experimentos de RNA-seq ao NCBI. Ele inclui exemplos práticos e instruções claras para compilar cada tabela.

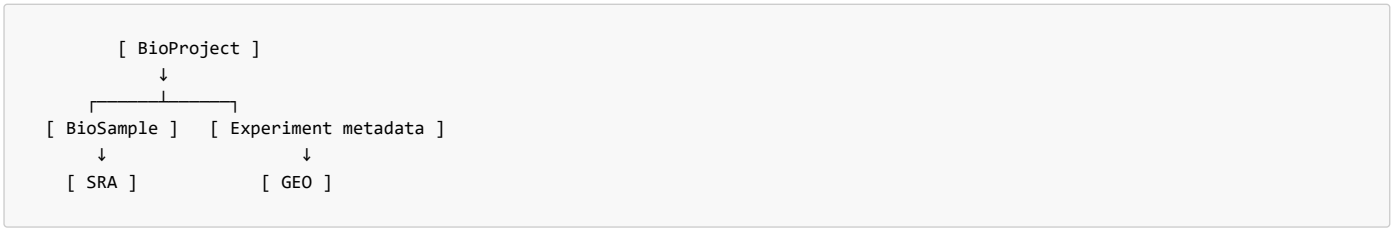
> Você pode montar os metadados usando Excel, Google Sheets, LibreOffice Calc ou até um script em R/Python

Por que metadados organizados são essenciais?

- Reprodutibilidade: outros pesquisadores podem validar e reutilizar seus dados com segurança.
- Interoperabilidade: ao seguir padrões internacionais, seus dados ficam visíveis em portais como GEO, SRA e ENA.
- Buscabilidade: bons metadados permitem que seus dados sejam facilmente encontrados por palavras-chave.
- Contextualização: descrevem o experimento, condições, grupos e protocolos com riqueza de detalhes.
- Compliance: repositórios públicos exigem certas informações obrigatórias.

---

1. Worflow



---

1. Organize todos os metadados em .tsv para BioSample, SRA e GEO
2. Crie seu BioProject primeiro
3. Submeta os BioSamples individualmente ou via planilha
4. Envie os arquivos brutos (FASTQ) pelo SRA
5. Envie os arquivos processados (matriz de contagem) pelo GEO
6. Vincule todos os IDs entre si (BioProject → BioSample → SRA/GEO)
7. Finalize o envio e aguarde os validadores do NCBI

O que é um arquivo .tsv?

TSV significa Tab-Separated Values — ou seja, um arquivo de texto onde cada coluna é separada por tabulação (TAB), e cada linha representa uma entrada (como uma amostra ou arquivo).

É muito parecido com um .csv (Comma-Separated Values), mas:

.csv	**.tsv**
Usa vírgulas entre os campos	Usa tabulação (TAB)
Pode causar conflitos com vírgulas no texto	Mais seguro para metadados textuais
Extensão: .csv	Extensão: .tsv

Se estiver no Excel:

1. Clique em Arquivo > Salvar como
2. Escolha o tipo "Texto (separado por tabulação) (.txt)"*
3. Renomeie a extensão de .txt para .tsv (se necessário)

No Google Sheets:

1. Vá em Arquivo > Fazer download > Valores separados por tabulação (.tsv)

---

2. BioProject

> O BioProject agrupa todas as amostras e dados de um estudo. *Crie apenas um BioProject por estudo.*

> É necessário criar um login no NCBI para isso.

- **Exemplo**:
  - **project_title**: Transcriptomic profiling of PBMCs during CHIKV infection
  - **description**: RNA-seq of CHIKV-infected PBMCs to study differential gene expression across timepoints.
  - **organism**: *Homo sapiens*
  - **data_type**: Transcriptome data
- Crie pelo site: https://submit.ncbi.nlm.nih.gov/subs/bioproject
- Quando enviado, ele gera um código PRJNAxxxxx (ex: `PRJNA123456`)

---

3. BioSample

Cada amostra biológica recebe uma entrada única. Os metadados aqui descrevem o material de onde vieram os dados.

- **Campos principais**:

sample_name	organism	tissue	collection_date	disease	geo_loc_name	bioproject_accession
CHIKV_01	Homo sapiens	PBMC	2023-03-05	Chikungunya fever	Brazil: Bahia	PRJNA123456

- Crie BioSamples vinculados ao seu BioProject pelo mesmo portal (https://submit.ncbi.nlm.nih.gov/subs/bioproject), ou prepare o upload em lote com .tsv.

Dê preferência a tabelas mais completas, como:

sample_name	organism	isolate	sex	age	tissue	cell_type	disease	treatment	time_point	geo_loc_name	collection_date	d
CHIKV_01	Homo sapiens	CHIKV_Ba01	F	35	Peripheral blood	PBMC	Chikungunya fever	CHIKV infection	3 dpi	Brazil: Bahia	2023-03-05	P is d C ir
CHIKV_02	Homo sapiens	CHIKV_Ba02	M	42	Peripheral blood	PBMC	Chikungunya fever	CHIKV infection	5 dpi	Brazil: Bahia	2023-03-07	P is d C ir

> Essa tabela pode ser salva como `biosample_metadata.tsv`

Após o envio, o sistema retorna accessions para cada amostra como:

```
SAMN45678901
SAMN45678902
```

---

4. Metadados SRA

Aqui você descreve os arquivos `.fastq.gz`, tipo de biblioteca, plataforma usada, etc.

O SRA aceita:

- Arquivos .fastq (brutos) ou .bam (alinhados)
- Metadados completos da biblioteca:

sample_name	biosample_accession	library_strategy	library_layout	filename
CHIKV_01	SAMN45678901	RNA-Seq	PAIRED	CHIKV_01_R1.fastq.gz;CHIKV_01_R2.fastq.gz

Atenção: os arquivos `.fastq.gz` devem estar nomeados de forma consistente com este campo.

Quanto mais detalhado, melhor para quem vai reutilizar seus dados ou revisar seu estudo.

sample_name	biosample_accession	library_ID	title	library_strategy	library_source	library_selection	library_layout	platform	instrum
CHIKV_01	SAMN45678901	LIB01	RNA-seq of CHIKV-infected PBMCs 3dpi	RNA-Seq	TRANSCRIPTOMIC	RANDOM	PAIRED	ILLUMINA	Illumina 6000

sample_name	biosample_accession	library_ID	title	library_strategy	library_source	library_selection	library_layout	platform	instrum
CHIKV_02	SAMN45678902	LIB02	RNA-seq of CHIKV-infected PBMCs 5dpi	RNA-Seq	TRANSCRIPTOMIC	RANDOM	PAIRED	ILLUMINA	Illumina 6000

Adicione outros campos se quiser mais detalhes (ex: basecaller, software de alinhamento). Salve como `sra_metadata.tsv`

---

5. Metadados GEO

O GEO (Gene Expression Omnibus) é um repositório público do NCBI focado em dados de expressão gênica processados, incluindo:

Tipo de arquivo	Extensão	Exemplo
Matriz de contagem	`.tsv`, `.csv`	`counts_matrix.tsv`
Contagens por amostra	`.tsv`, `.txt`	`counts_CHIKV_01.tsv`
Arquivos normalizados	`.tsv`, `.rds`	`normalized_counts.rds`
Scripts ou pipelines	`.R`, `.sh`, `.ipynb`	`deseq2_analysis.R`
Metadados das amostras	`.tsv`	`geo_sample_metadata.tsv`

- Você também pode incluir diagramas de fluxo experimental, fatores de batch, e até RIN e concentração do RNA.

Tabela simples:

title	biosample_accession	source_name	organism	treatment	time_point	file_type	file_name	BioProject
Expression of PBMCs CHIKV 3dpi	SAMN45678901	PBMC	Homo sapiens	CHIKV	3dpi	Counts	counts_CHIKV_01.tsv	PRJNA123456

Prepare com base em GEO submission templates: https://www.ncbi.nlm.nih.gov/geo/info/submission.html?form=MG0AV3

Exemplo mais completo:

sample_title	biosample_accession	source_name	organism	characteristics_ch1	time_point	treatment	protocol_ch1	data_processing	file_nam
CHIKV_01	SAMN45678901	PBMC	Homo sapiens	disease: Chikungunya fever	3 dpi	CHIKV infection	rRNA depletion + TruSeq	alignment with HISAT2, counts with StringTie and prepDE	counts_C
CHIKV_02	SAMN45678902	PBMC	Homo sapiens	disease: Chikungunya fever	5 dpi	CHIKV infection	rRNA depletion + TruSeq	alignment with HISAT2, counts with StringTie and prepDE	counts_C

> O campo `characteristics_ch1` no GEO é extremamente flexível e poderoso — ele permite descrever várias características biológicas, clínicas ou técnicas da sua amostra, além da doença.

Outro exemplo de como pode ser mais completo:

sample_title	biosample_accession	source_name	organism	characteristics_ch1	characteristics_ch1	characteristics_ch1	characteristics_ch1	tim
CHIKV_01_3dpi	SAMN45678901	PBMC	Homo sapiens	disease: Chikungunya fever	sex: female	age: 35	RIN: 8.5	3 d

---

6. Ferramentas e Sites úteis

- `pandoc` → conversão para PDF/HTML:

```
pandoc metadados_submissao.md -o metadados_submissao.pdf
pandoc metadados_submissao.md -o metadados_submissao.html
```

- https://submit.ncbi.nlm.nih.gov/?form=MG0AV3

- https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/?form=MG0AV3

Fluxograma de como submeter

```
PRJNA123456 (BioProject)
  ├─ SAMN45678901 (BioSample)
  │     ├─ SRRxxxxxxx (SRA - dados brutos)
  │     └─ GSMxxxxxxx (GEO - expressão)
```

1. Crie seu BioProject

- Acesse: https://submit.ncbi.nlm.nih.gov/subs/bioproject
- Preencha as informações sobre o estudo (título, organismo, tipo) como foi explicado acima
- Quando enviado, ele gera um código como: `PRJNA123456`

2. Submeta seus BioSamples

Vá para: https://submit.ncbi.nlm.nih.gov/subs/biosample

Ao preencher cada linha (via formulário ou .tsv), inclua o campo:

```
bioproject_accession
PRJNA123456
```

Cada amostra recebe um código como: `SAMN45678901`

Cada BioSample deve ter um nome único (ex: CHIKV_01) e esse mesmo nome será usado nos metadados do SRA e GEO.

3. Submissão ao SRA (dados brutos)

Acesse: https://submit.ncbi.nlm.nih.gov/subs/sra

Faça upload dos arquivos `.fastq.gz`

No seu .tsv ou formulário, inclua:

sample_name	biosample_accession
CHIKV_01	SAMN45678901

O SRA usará isso para fazer o vínculo entre seu `.fastq` e a amostra correta

4. Submissão ao GEO

Acesse: https://submit.ncbi.nlm.nih.gov/subs/geo

No `sample_metadata.tsv`, inclua:

```
BioSample         BioProject
SAMN45678901      PRJNA123456
```

Revise e submeta para revisão. Após o envio, você recebe um GSE ID temporário (ex: `GSE123456`), e o time do NCBI faz a curadoria.

Quando tudo está vinculado corretamente, qualquer pessoa (ou revisor!) poderá:

Entrar no BioProject → Ver as BioSamples → Acessar os dados no SRA → Ver os arquivos processados no GEO — como se fosse um só estudo interligado.

> Para submeter arquivos processados

ex: contagens por gene

Entre em: https://submit.ncbi.nlm.nih.gov/subs/geo/

1. Crie uma nova submissão
2. Escolha: Processed Data Submission (GSE)
3. Faça upload:
    - Dos arquivos processados (.tsv, .rds, etc.)

    - Da planilha de metadados
    - Dos scripts ou suplementares

4. Preencha a descrição do estudo, protocolo, objetivos, etc.