

SRA-Explorer+Aspera才是打开SRA的正确方式！

自从去年NCBI取消SRA对FTP下载的支持后，下载SRA公共数据变的麻烦起来，也慢了很多，虽然prefecth直接通过SRR号直接下载数据，但是也取消了对FTP的支持，只能通过HTTPS方式下载，速度巨慢。虽然NCBI提供了AWS和google下载方式，但是两者注册比较麻烦，基本上是收费的。

以前的NCBI SRA支持Aspera方式高速下载，速度基本维持在100Mb/s左右。但是目前的SRA取消对FTP支持后，直接无法通过anonftp@ftp.ncbi.nlm.nih.gov获取到SRA的数据。

那么问题来了，到底还有没有好的SRA下载方式呢？答案是有的，而且很方便。

我们以一篇最新的Scientific Data文章为例，演示怎么快速下载SRA数据。

www.nature.com/scientificdata

SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

Highly accurate long-read HiFi sequencing data for five complex genomes

Ting Hon¹, Kristin Mars¹, Greg Young¹, Yu-Chih Tsai¹, Joseph W. Karalius¹, Jane M. Landolin², Nicholas Maurer³, David Kudrna⁴, Michael A. Hardigan⁵, Cynthia C. Steiner⁶, Steven J. Knapp⁵, Doreen Ware^{7,8}, Beth Shapiro^{3,9}, Paul Peluso¹ & David R. Rank¹✉

文章测了五个物种的基因组数据，而且是通过PacBio HiFi 测序方法。

Organism	HiFi library size (kb)	Sequel II Runs (number)	Bases > RQ20 (Gb)	Average RL (kb)	Reads (Millions)	Quality Value* (avg)	Data Record
<i>M. musculus</i>	15.9	2	66.5	16.4	4.1	31	SRR11606870 ³⁷
<i>Z. mays</i>	15.0	2	48.1	15.6	3.1	30	SRR11606869 ³⁸
<i>E. × ananassa</i>	23.0	1	29.7	21.7	1.4	28	SRR11606867 ³⁹
<i>R. muscosa</i>	15.8	8	189.1	15.7	12.1	31	SRR11606868 ⁴⁰
ATCC MSA-1003	14.1	2	59.1	10.5	5.6	35	SRR11606871 ⁴¹

文章提供的中SRA编号是：**SRP258341**

我们请上今天的主角：[SRA Explorer https://sra-explorer.info/](https://sra-explorer.info/)

SRA Explorer

This tool aims to make datasets within the Sequence Read Archive more accessible.

Search for:

Enter accession number

Q

Max Results

100

Start At Record

0

Need inspiration? Try [GSE30567](#) , [SRP043510](#) , [PRJEB8073](#) , [ERP009109](#) or [human liver miRNA](#) .

直接输入SRP258341

SRA Explorer

This tool aims to make datasets within the Sequence Read Archive more accessible.

Search for:

SRP258341[All Fields]

+

Q

Max Results

100

Start At Record

0

Need inspiration? Try [GSE30567](#) , [SRP043510](#) , [PRJEB8073](#) , [ERP009109](#) or [human liver miRNA](#) .

得到19条记录：

Showing 19 results.

Filter results:

Enter search term

All Fields

Add 0 to collection

<input type="checkbox"/>	Title	Accession	Instrument	Total Bases (Mb)	Date Created
<input type="checkbox"/>	Metagenomic Std ATCC MSA-1003: PacBio HiFi Reads FASTQ	SRR11606871	Sequel II	591457	24 Apr 2020
<input type="checkbox"/>	Mus musculus C57/BL6J: PacBio HiFi Reads FASTQ	SRR11606870	Sequel II	664929	26 Apr 2020
<input type="checkbox"/>	Zea mays B73: PacBio HiFi Reads FASTQ	SRR11606869	Sequel II	480749	24 Apr 2020
<input type="checkbox"/>	Rana muscosa: PacBio HiFi Reads FASTQ	SRR11606868	Sequel II	1890756	24 Apr 2020
<input type="checkbox"/>	Fragaria x ananassa 'Royal Royce': PacBio HiFi Reads FASTQ	SRR11606867	Sequel II	296763	26 Apr 2020
<input type="checkbox"/>	Mus musculus C57/BL6J: PacBio Raw Subreads BAM	SRR12358174	Sequel II	4883747	04 Aug 2020
<input type="checkbox"/>	Mus musculus C57/BL6J: PacBio Raw Subreads BAM	SRR12371718	Sequel II	4660027	04 Aug 2020
<input type="checkbox"/>	Zea mays B73: PacBio Raw Subreads BAM	SRR12358173	Sequel II	7821107	02 Aug 2020
<input type="checkbox"/>	Rana muscosa: PacBio Raw Subreads BAM	SRR12371727	Sequel II	3154543	04 Aug 2020
<input type="checkbox"/>	Rana muscosa: PacBio Raw Subreads BAM	SRR12371726	Sequel II	3433734	04 Aug 2020
<input type="checkbox"/>	Rana muscosa: PacBio Raw Subreads BAM	SRR12371725	Sequel II	3281117	04 Aug 2020
<input type="checkbox"/>	Rana muscosa: PacBio Raw Subreads BAM	SRR12371724	Sequel II	4192754	04 Aug 2020
<input type="checkbox"/>	Rana muscosa: PacBio Raw Subreads BAM	SRR12371723	Sequel II	3176854	04 Aug 2020

依次操作，加入购物车：

SRA-Explorer

Step319 saved datasets

Showing 19 results.

Filter results:

Enter search term

All Fields

Add 19 to collection

Step2

Step1

<input checked="" type="checkbox"/>	Title	Accession	Instrument	Total Bases (Mb)	Date Created
<input checked="" type="checkbox"/>	Metagenomic Std ATCC MSA-1003: PacBio HiFi Reads FASTQ	SRR11606871	Sequel II	591457	24 Apr 2020
<input checked="" type="checkbox"/>	Mus musculus C57/BL6J: PacBio HiFi Reads FASTQ	SRR11606870	Sequel II	664929	26 Apr 2020
<input checked="" type="checkbox"/>	Zea mays B73: PacBio HiFi Reads FASTQ	SRR11606869	Sequel II	480749	24 Apr 2020
<input checked="" type="checkbox"/>	Rana muscosa: PacBio HiFi Reads FASTQ	SRR11606868	Sequel II	1890756	24 Apr 2020
<input checked="" type="checkbox"/>	Fragaria x ananassa 'Royal Royce': PacBio HiFi Reads FASTQ	SRR11606867	Sequel II	296763	26 Apr 2020
<input checked="" type="checkbox"/>	Mus musculus C57/BL6J: PacBio Raw Subreads BAM	SRR12358174	Sequel II	4883747	04 Aug 2020
<input checked="" type="checkbox"/>	Mus musculus C57/BL6J: PacBio Raw Subreads BAM	SRR12371718	Sequel II	4660027	04 Aug 2020
<input checked="" type="checkbox"/>	Zea mays B73: PacBio Raw Subreads BAM	SRR12358173	Sequel II	7821107	02 Aug 2020

然后可以直接选择下载fastq格式的文件，而无需下载SRA格式，还得通过fasterq-dump转化格式。很方便，是不是？

19 Saved Datasets

Remove all from collection and send to search results

FastQ Downloads SRA Downloads Full Metadata

To download FastQ files directly, sra-explorer queries the ENA for each SRA run accession number.

Raw FastQ Download URLs

Bash script for downloading FastQ files

Aspera commands for downloading FastQ files

Cluster Flow FastQ download file (nice filenames)

bcbio project file for FastQ downloads (nice filenames)

点进去，发现SRA Explorer已经给出了下载脚本，还能直接修改下载的数据名字，添加了物种，数据类型等信息。非常地人性化。

Aspera commands for downloading FastQ files

This list of bash `ascp` commands to download each FastQ file from the ENA using the Aspera download tool.

ascp openssl key path \$HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh

Linux OSX

Rename files? ☒ Append `mv` command to rename downloaded files ☐ Don't rename files

Copy Download

```
#!/usr/bin/env bash
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR116/067/SRR11606867/SRR11606867
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR116/068/SRR11606868/SRR11606868
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR116/070/SRR11606870/SRR11606870
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh . && mv SRR12358174_Mus_musculus_C57_BL6J_PacBio_Raw_Subreads_BAM
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR116/069/SRR11606869/SRR11606869
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR116/071/SRR11606871/SRR11606871
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR123/018/SRR12371718/SRR12371718
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR123/025/SRR12371725/SRR12371725
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR123/027/SRR12371727/SRR12371727
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR123/024/SRR12371724/SRR12371724
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR123/073/SRR12358173/SRR12358173
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR123/071/SRR12358171/SRR12358171
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh . && mv SRR12371721_Rana_muscosa_PacBio_Raw_Subreads_BAM
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh . && mv SRR12358170_Metagenomic_Std_ATCC_MSA-1003_PacBio_Raw_Subreads_B
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR123/072/SRR12358172/SRR12358172
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac.uk:vol1/fastq/SRR123/019/SRR12371719/SRR12371719
```

通过Copy或者Download就可以得到下载脚本，竟然同时支持Linux和MacOS系统，后者需要打开终端操作。

当然，如果只对HiFi数据感兴趣，可以在加入购物车之前就选择，或者在脚本中选择亦可。其它的14个是HiFi数据的原始数据。

```
RR11606867_subreads.fastq.gz . && mv SRR11606867_subreads.fastq.gz SRR11606867_Fragaria_x_ananassa_Royal_Royce_PacBio_HiFi_Reads_FASTQ_subreads.fastq.gz
RR11606868_subreads.fastq.gz . && mv SRR11606868_subreads.fastq.gz SRR11606868_Rana_muscosa_PacBio_HiFi_Reads_FASTQ_subreads.fastq.gz
RR11606870_subreads.fastq.gz . && mv SRR11606870_subreads.fastq.gz SRR11606870_Mus_musculus_C57_BL6J_PacBio_HiFi_Reads_FASTQ_subreads.fastq.gz
_BAM
RR11606869_subreads.fastq.gz . && mv SRR11606869_subreads.fastq.gz SRR11606869_Zea_mays_B73_PacBio_HiFi_Reads_FASTQ_subreads.fastq.gz
RR11606871_subreads.fastq.gz . && mv SRR11606871_subreads.fastq.gz SRR11606871_Metagenomic_Std_ATCC_MSA-1003_PacBio_HiFi_Reads_FASTQ_subreads.fastq.gz
```

笔者实测下载玉米HiFi数据，压缩文件大小40Gb，耗时大约1个半小时，放在后台运行即可。

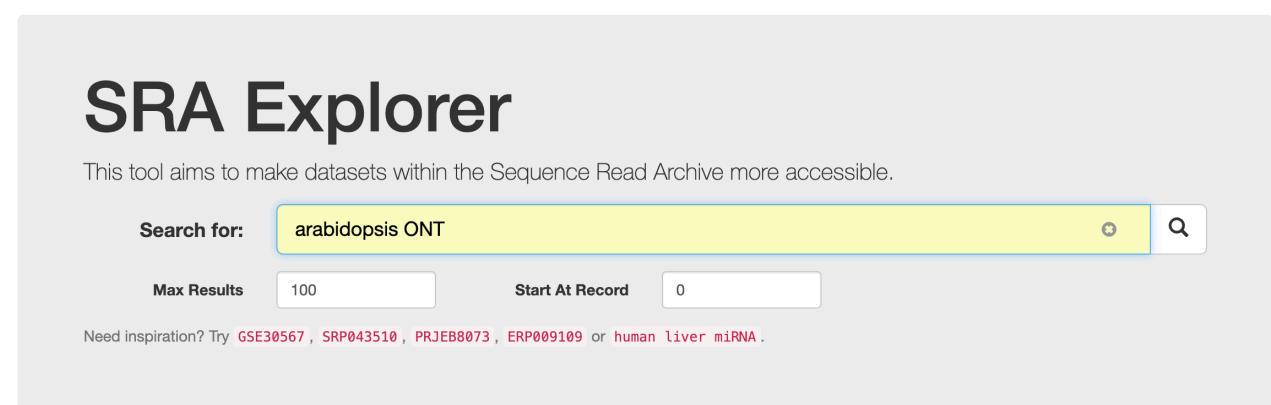
感兴趣的读者可以去github查看SRA-Explorer的源码：<https://github.com/ewels/sra-explorer>

还可以将index.html下载到本地运行。

Aspera有了SRA-Explorer，才香。

SRA-Explorer的fastq数据其实是通过European Nucleotide Archive（ENA）数据库获取的。

如果不知道具体的SRA编号，还可以根据自己的需求模糊搜索数据。比如。想下载拟南芥的ONT数据：



获得12条记录，其中有一个sequel数据误入，其它的都是ONT数据哦。当然还可以重新调整关键词，获得更精准的搜索！

Showing 12 results.

Filter results: All Fields ▾ Add 0 to collection

<input type="checkbox"/> Title	Accession	Instrument	Total Bases (Mb)	Date Created
<input type="checkbox"/> MinION sequencing	ERR2097528	MinION	0	22 Aug 2017
<input type="checkbox"/> GSM3688597: VLP DNA long read wild-type; Arabidopsis thaliana; OTHER	SRR8792549	GridION	18054	14 Mar 2020
<input type="checkbox"/> GSM3688598: VLP DNA long read ddm1; Arabidopsis thaliana; OTHER	SRR8792550	GridION	22061	14 Mar 2020
<input type="checkbox"/> GSM3688599: VLP DNA long read ddm1rrd6; Arabidopsis thaliana; OTHER	SRR8792551	GridION	14824	14 Mar 2020
<input type="checkbox"/> GSM4210380: Sample 1_Ath_mRNA_PacBio; Arabidopsis thaliana; RNA-Seq	SRR10611192	Sequel	439507	16 Dec 2019
<input type="checkbox"/> GSM4210381: Sample 2_Ath_mRNA_Nanopore; Arabidopsis thaliana; RNA-Seq	SRR10611193	PromethION	100919	16 Dec 2019
<input type="checkbox"/> GSM4210382: Sample 3_Ath_mRNA_Nanopore; Arabidopsis thaliana; RNA-Seq	SRR10611194	PromethION	97845	16 Dec 2019
<input type="checkbox"/> GSM4210383: Sample 4_Ath_mRNA_Nanopore; Arabidopsis thaliana; RNA-Seq	SRR10611195	PromethION	86081	16 Dec 2019
<input type="checkbox"/> Arabidopsis thaliana, Ped-0, PED_AA_01_ONT, leaf, Oxford Nanopore Ligation Sequencing Kit, SQK-LSK109	SRR12136400	MinION	104200	08 Jul 2020
<input type="checkbox"/> MinION sequencing	ERR4375138	MinION	103	29 Jul 2020
<input type="checkbox"/> MinION sequencing	ERR4375139	MinION	178	29 Jul 2020
<input type="checkbox"/> MinION sequencing	ERR4375140	MinION	114711	29 Jul 2020

