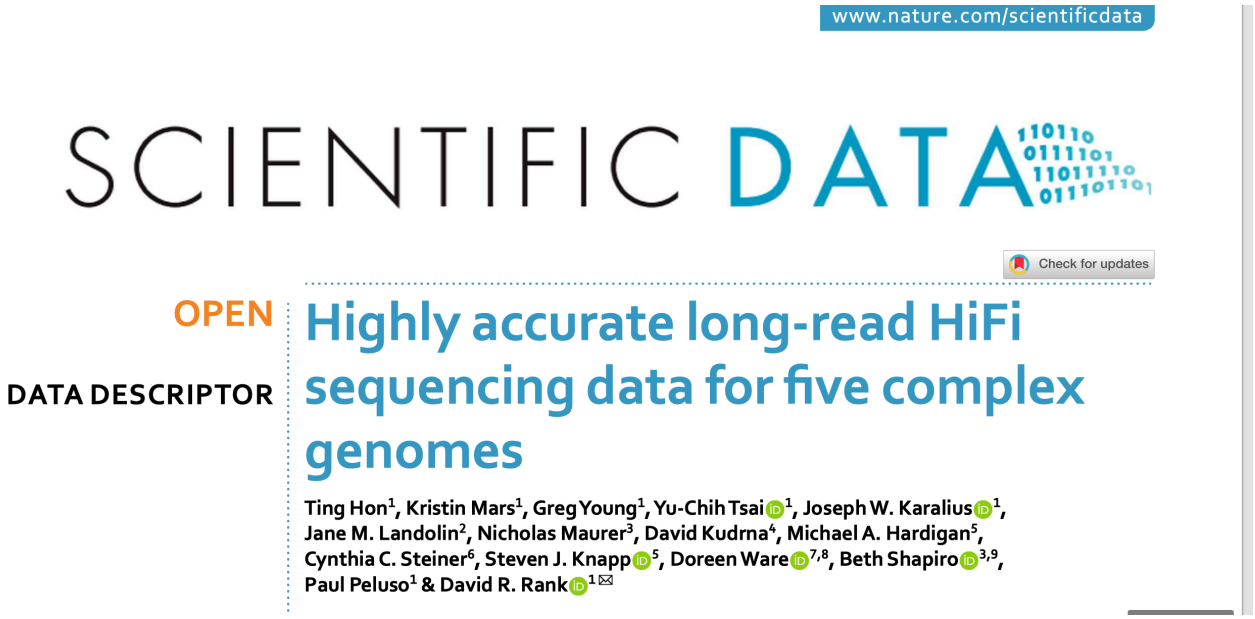
SRA-Explorer+Aspera才是打开SRA的正确方式！

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。

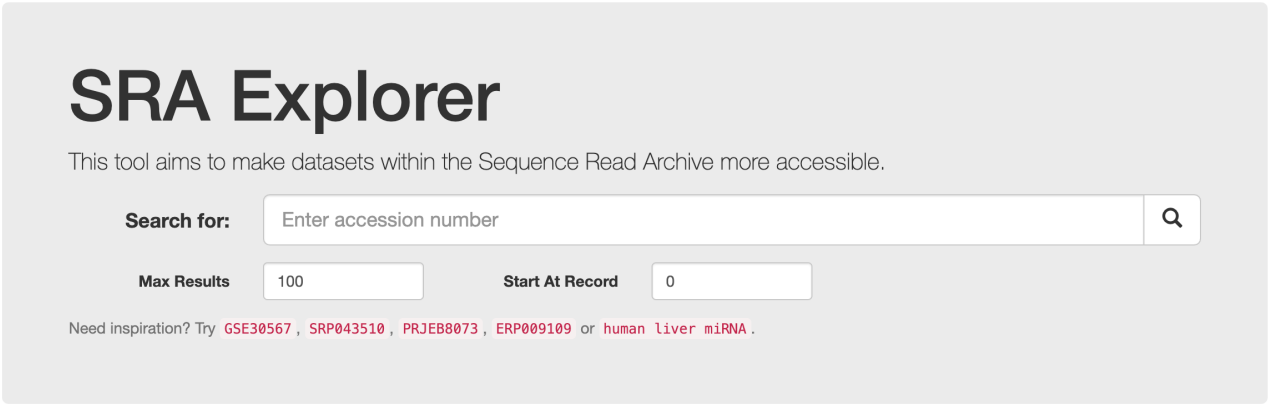


文章测了五个物种的基因组数据，而且是通过PacBio HIFI 测序方法。

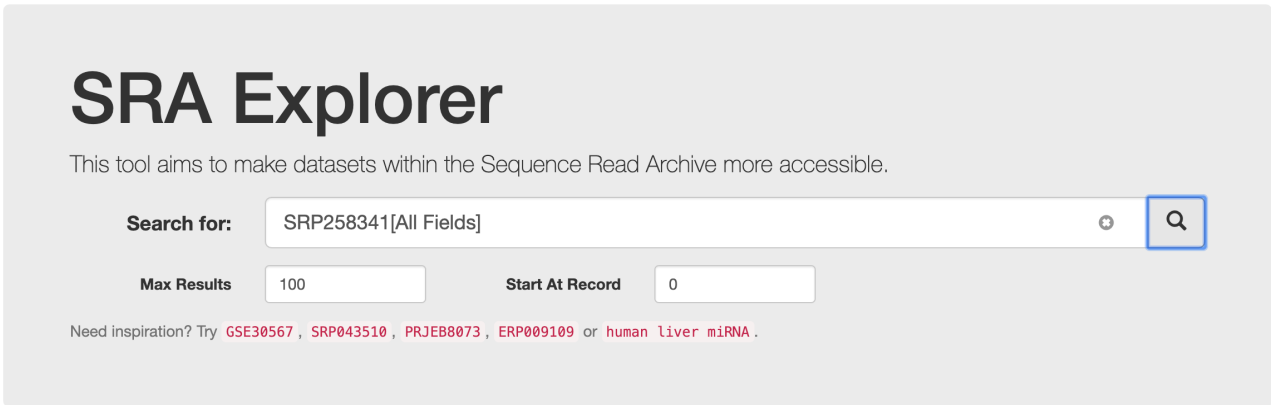
Organism	HiFi library size (kb)	Sequel II Runs (number)	Bases > RQ20 (Gb)	Average RL (kb)	Reads (Millions)	Quality Value* (avg)	Data Record
M. musculus	15.9	2	66.5	16.4	4.1	31	SRR11606870[37]
Z. mays	15.0	2	48.1	15.6	3.1	30	SRR11606869[38]
F. × ananassa	23.0	1	29.7	21.7	1.4	28	SRR11606867[39]
R. muscosa	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/



直接输入**SRP258341**



得到19条记录：

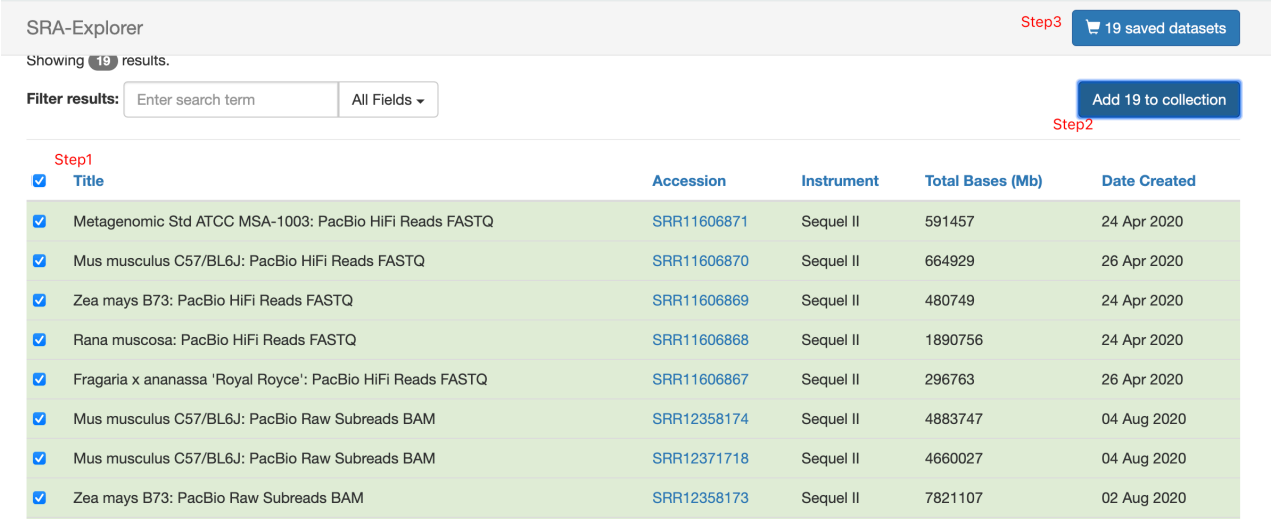
Showing 19 results.

Filter results: Enter search term All Fields

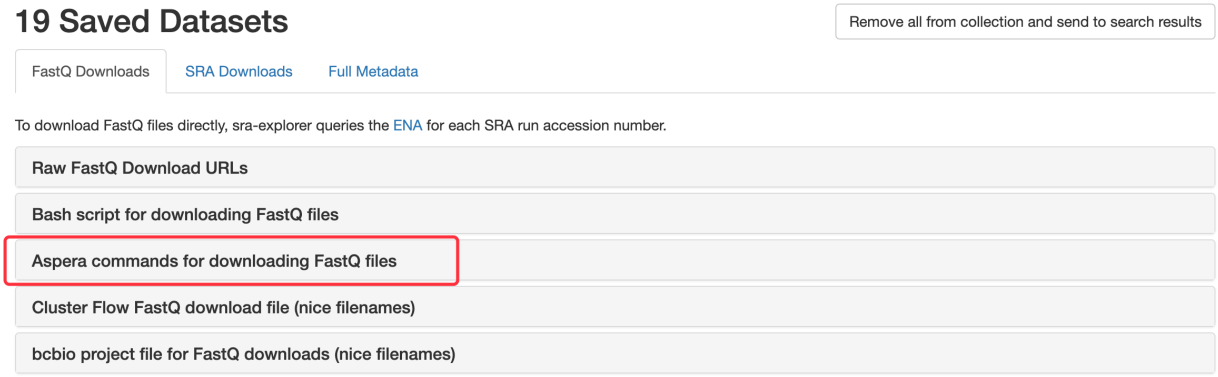
Add 0 to collection

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

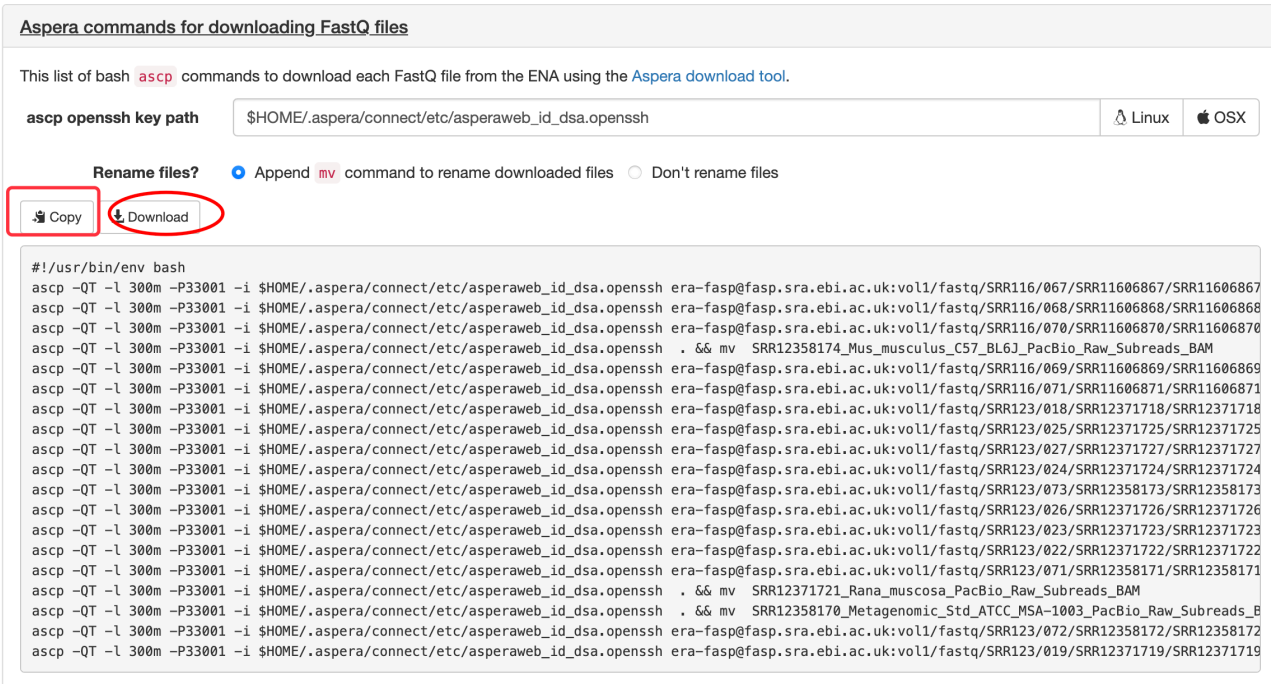
依次操作，加入购物车：



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?



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。



通过Copy或者Download就可以得到下载的脚本，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。

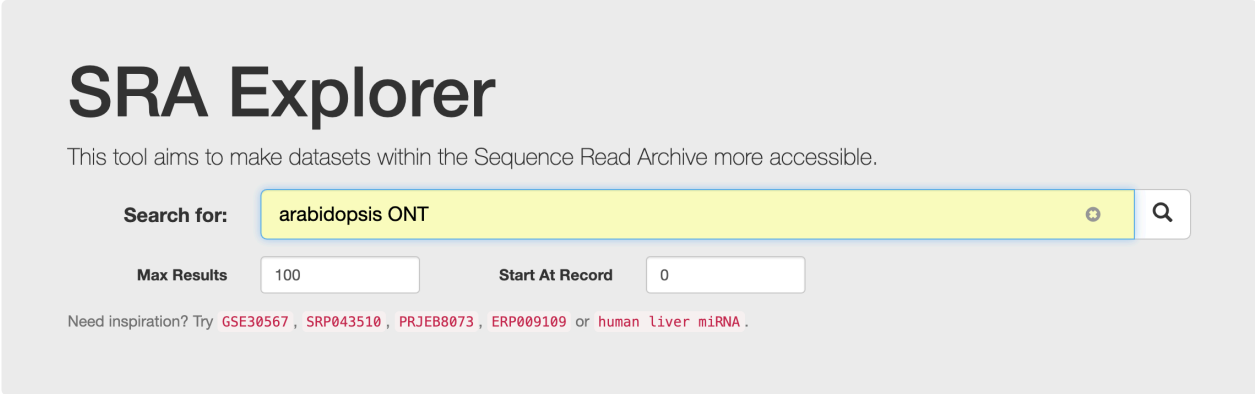
感兴趣的读者可以去github查看SRA-Explorer的源码：https://github.com/ewels/sra-explorer

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: Enter search term | All Fields | Add 0 to collection

Title	Accession	Instrument	Total Bases (Mb)	Date Created
MinION sequencing	ERR2097528	MinION	0	22 Aug 2017
GSM3688597: VLP DNA long read wild-type; Arabidopsis thaliana; OTHER	SRR8792549	GridION	18054	14 Mar 2020
GSM3688598: VLP DNA long read ddm1; Arabidopsis thaliana; OTHER	SRR8792550	GridION	22061	14 Mar 2020
GSM3688599: VLP DNA long read ddm1rdr6; Arabidopsis thaliana; OTHER	SRR8792551	GridION	14824	14 Mar 2020
GSM4210380: Sample 1_Ath_mRNA_PacBio; Arabidopsis thaliana; RNA-Seq	SRR10611192	Sequel	439507	16 Dec 2019
GSM4210381: Sample 2_Ath_mRNA_Nanopore; Arabidopsis thaliana; RNA-Seq	SRR10611193	PromethION	100919	16 Dec 2019
GSM4210382: Sample 3_Ath_mRNA_Nanopore; Arabidopsis thaliana; RNA-Seq	SRR10611194	PromethION	97845	16 Dec 2019
GSM4210383: Sample 4_Ath_mRNA_Nanopore; Arabidopsis thaliana; RNA-Seq	SRR10611195	PromethION	86081	16 Dec 2019
Arabidopsis thaliana, Ped-0, PED_AA_01_ONT, leaf, Oxford Nanopore Ligation Sequencing Kit, SQK-LSK109	SRR12136400	MinION	104200	08 Jul 2020
MinION sequencing	ERR4375138	MinION	103	29 Jul 2020
MinION sequencing	ERR4375139	MinION	178	29 Jul 2020
MinION sequencing	ERR4375140	MinION	114711	29 Jul 2020