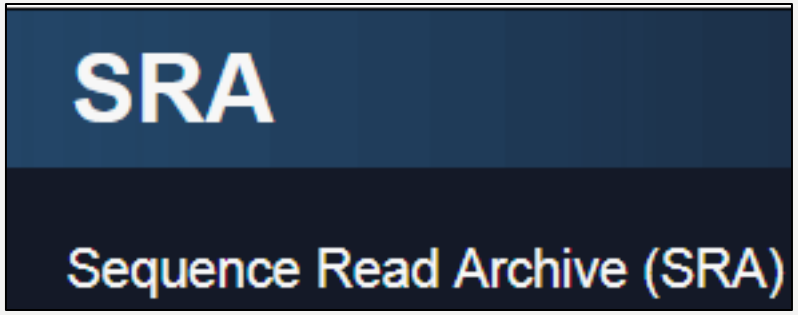


UNIT 2
LECTURE 2: DOWNLOAD SEQUENCE READS

Princess Rodriguez, PhD

MMG 232
Spring 2023

TODAY YOU WILL LEARN:

- What is GEO?
- Navigate through GEO and SRA
- Make sense of the accession numbers, the data they hold, and how they fit together
- How to download sequence reads (.fastq files) from SRA
  - Download and configure SRA Toolkit

WHAT IS GEO?

- GEO (Gene Expression Omnibus) is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community
- User-friendly
- 100's of organisms and thousands of different expression analysis platforms

SUBMITTING TO GEO

- Is a pre-requisite for publication of peer-reviewed research articles
- Step-by-step process that requires sample data spreadsheets, raw data, and other submission forms

MIAME COMPLIANT

- Minimum information about a microarray experiment
- Effort to standardize publicly available data

- MAIME Checklist
  - Experimental Design
  - Samples used, extract preparation, and labeling
  - Measurement data and specifications

GEO ARCHITECTURE

There are three types of GEO submitter records:

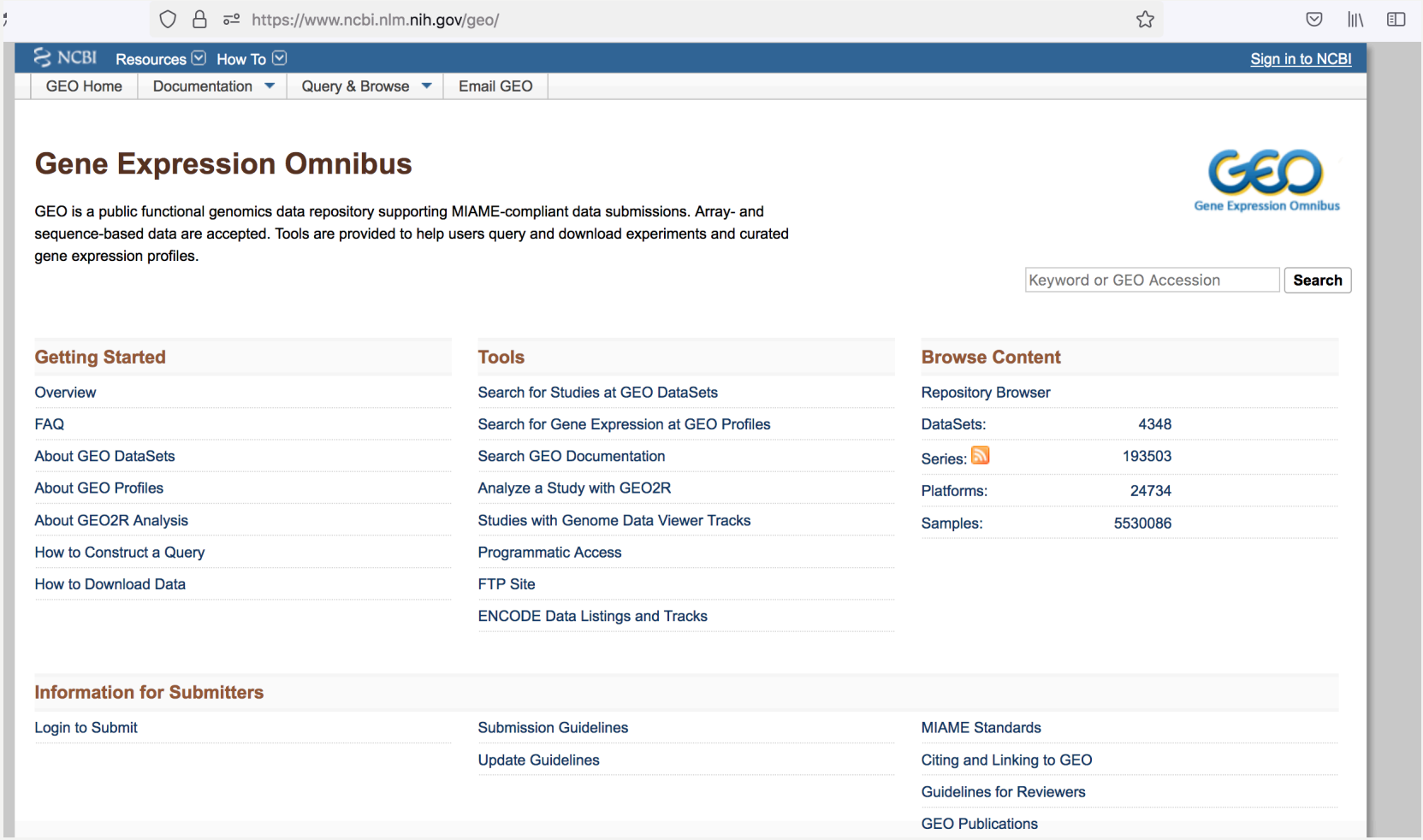
- A **Platform** record describes an array or sequencer and, for array-based platforms, a data table defining the array template.
- A **Sample** record describes the sample source, the protocols used in its analysis, and the expression data derived from it.
- A **Series** record links together a group of related Samples and describes a whole study.

Together, this information makes up a GEO record that is assembled by the GEO staff. These records provide a coherent synopsis about the experiment and data collected.

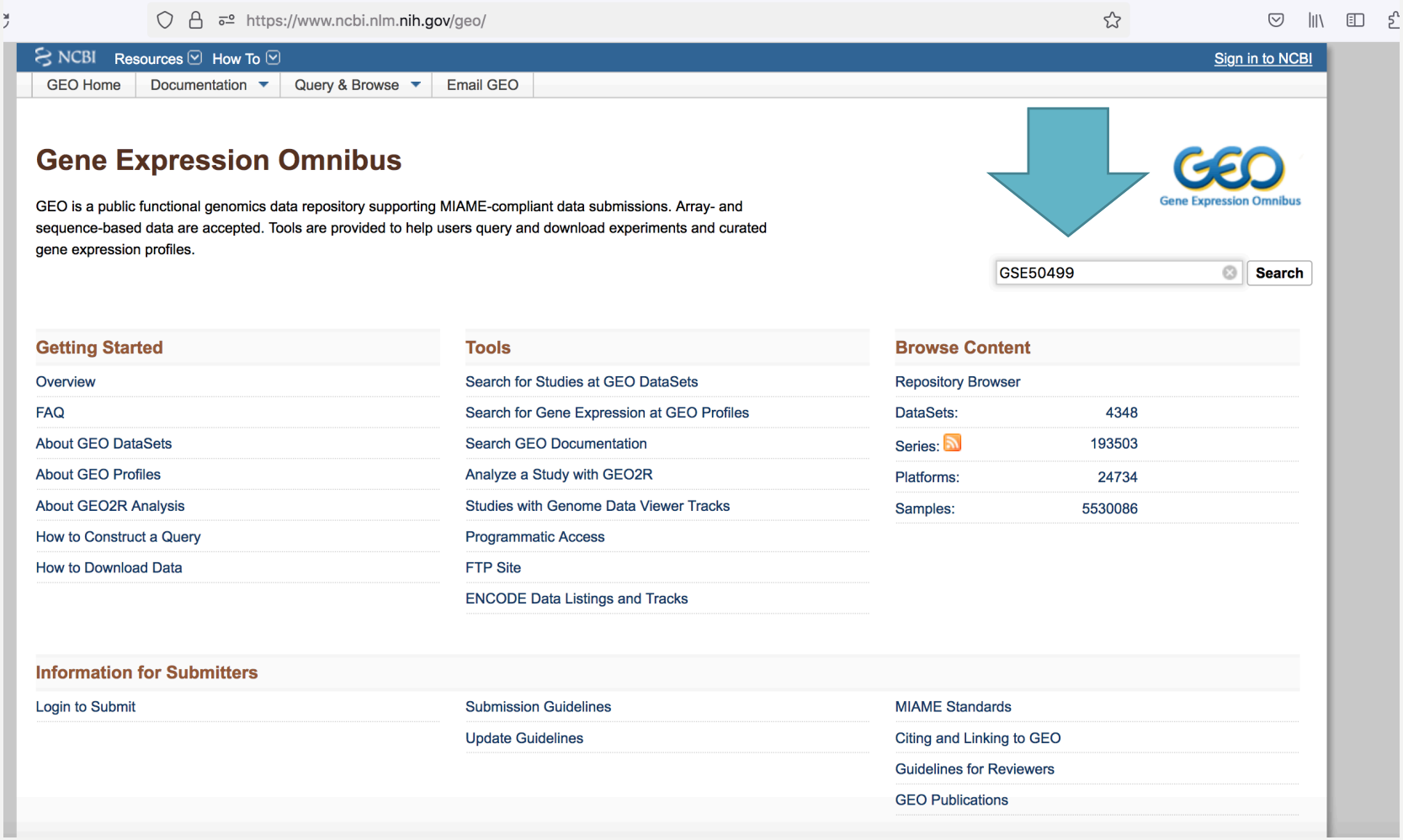
GEO DATA RETRIEVAL

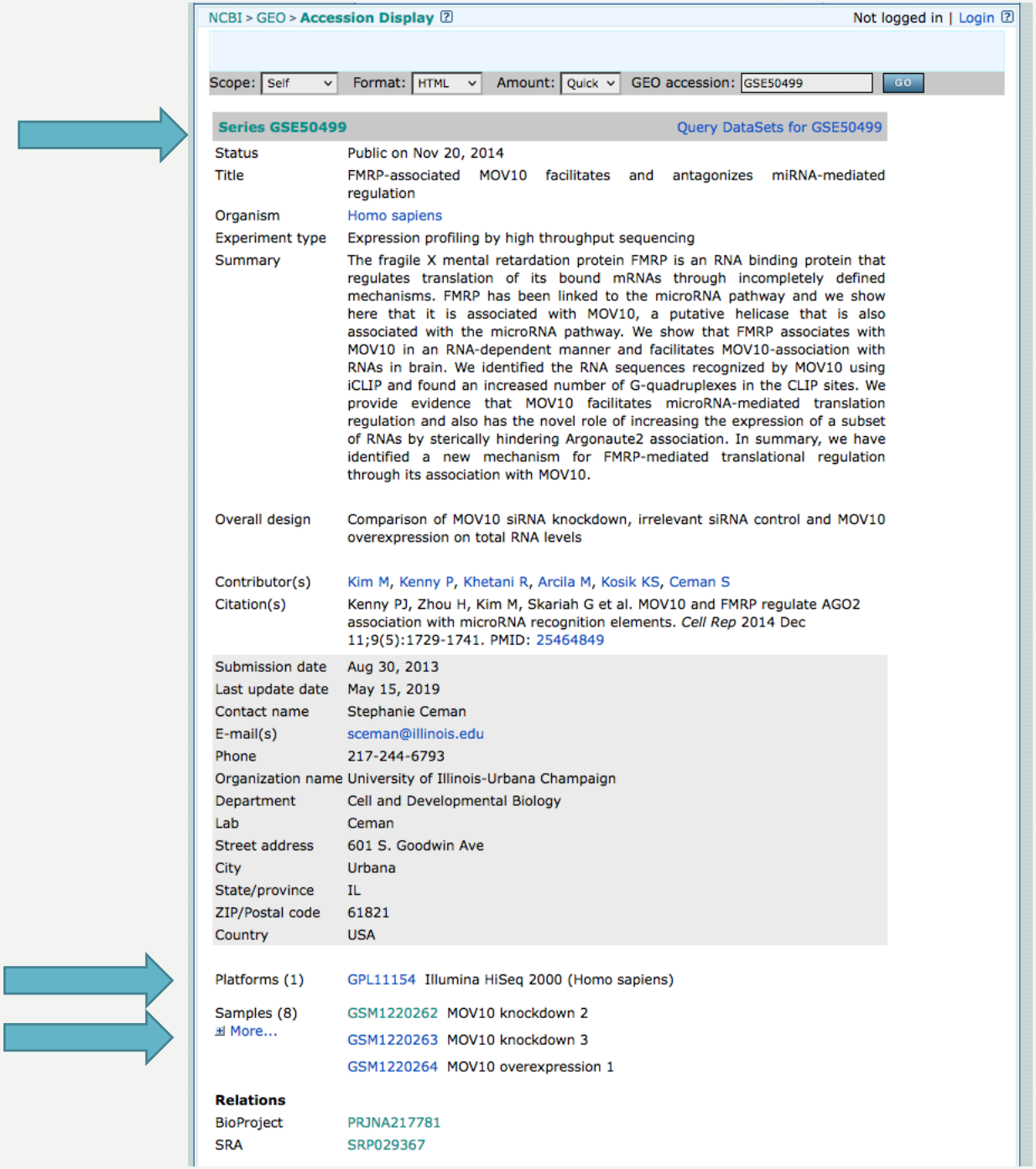
GEO data can be retrieved and analyzed in several ways:

- To look at a particular GEO record for which you have the accession number, use the GEO accession box located on the GEO homepage or at the top of each GEO record.
- To download data, see the various options described on the Download GEO data page.
- To quickly locate data relevant to your interests, search GEO DataSets and GEO Profiles



https://www.ncbi.nlm.nih.gov/geo/


https://www.ncbi.nlm.nih.gov/geo/
NCBI
Resources
How To
Sign in to NCBI
GEO Home
Documentation
Query & Browse
Email GEO
Gene Expression Omnibus
GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.
GEO
Gene Expression Omnibus
GSE50499
Search
Getting Started
Overview
FAQ
About GEO DataSets
About GEO Profiles
About GEO2R Analysis
How to Construct a Query
How to Download Data
Tools
Search for Studies at GEO DataSets
Search for Gene Expression at GEO Profiles
Search GEO Documentation
Analyze a Study with GEO2R
Studies with Genome Data Viewer Tracks
Programmatic Access
FTP Site
ENCODE Data Listings and Tracks
Browse Content
Repository Browser
DataSets: 4348
Series: 193503
Platforms: 24734
Samples: 5530086
Information for Submitters
Login to Submit
Submission Guidelines
Update Guidelines
MIAME Standards
Citing and Linking to GEO
Guidelines for Reviewers
GEO Publications



NCBI > GEO > **Accession Display** Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GSE50499 GO

Series GSE50499 Query DataSets for GSE50499

Status	Public on Nov 20, 2014
Title	FMRP-associated MOV10 facilitates and antagonizes miRNA-mediated regulation
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	The fragile X mental retardation protein FMRP is an RNA binding protein that regulates translation of its bound mRNAs through incompletely defined mechanisms. FMRP has been linked to the microRNA pathway and we show here that it is associated with MOV10, a putative helicase that is also associated with the microRNA pathway. We show that FMRP associates with MOV10 in an RNA-dependent manner and facilitates MOV10-association with RNAs in brain. We identified the RNA sequences recognized by MOV10 using iCLIP and found an increased number of G-quadruplexes in the CLIP sites. We provide evidence that MOV10 facilitates microRNA-mediated translation regulation and also has the novel role of increasing the expression of a subset of RNAs by sterically hindering Argonaute2 association. In summary, we have identified a new mechanism for FMRP-mediated translational regulation through its association with MOV10.
Overall design	Comparison of MOV10 siRNA knockdown, irrelevant siRNA control and MOV10 overexpression on total RNA levels
Contributor(s)	Kim M, Kenny P, Khetani R, Arcila M, Kosik KS, Ceman S
Citation(s)	Kenny PJ, Zhou H, Kim M, Skariah G et al. MOV10 and FMRP regulate AGO2 association with microRNA recognition elements. *Cell Rep* 2014 Dec 11;9(5):1729-1741. PMID: 25464849
Submission date	Aug 30, 2013
Last update date	May 15, 2019
Contact name	Stephanie Ceman
E-mail(s)	sceman@illinois.edu
Phone	217-244-6793
Organization name	University of Illinois-Urbana Champaign
Department	Cell and Developmental Biology
Lab	Ceman
Street address	601 S. Goodwin Ave
City	Urbana
State/province	IL
ZIP/Postal code	61821
Country	USA

Platforms (1) GPL11154 Illumina HiSeq 2000 (Homo sapiens)

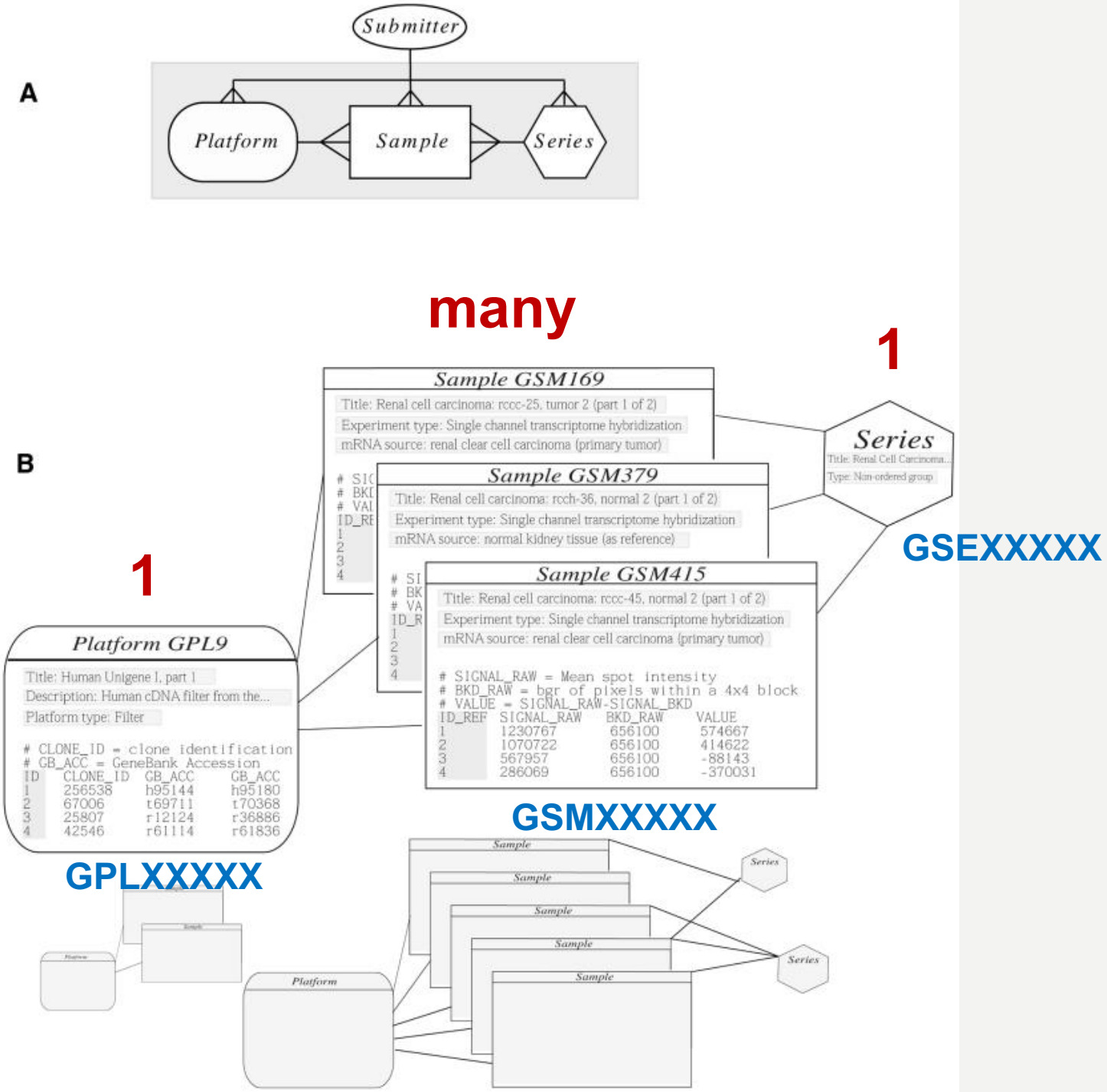
Samples (8) GSM1220262 MOV10 knockdown 2

More... GSM1220263 MOV10 knockdown 3

GSM1220264 MOV10 overexpression 1

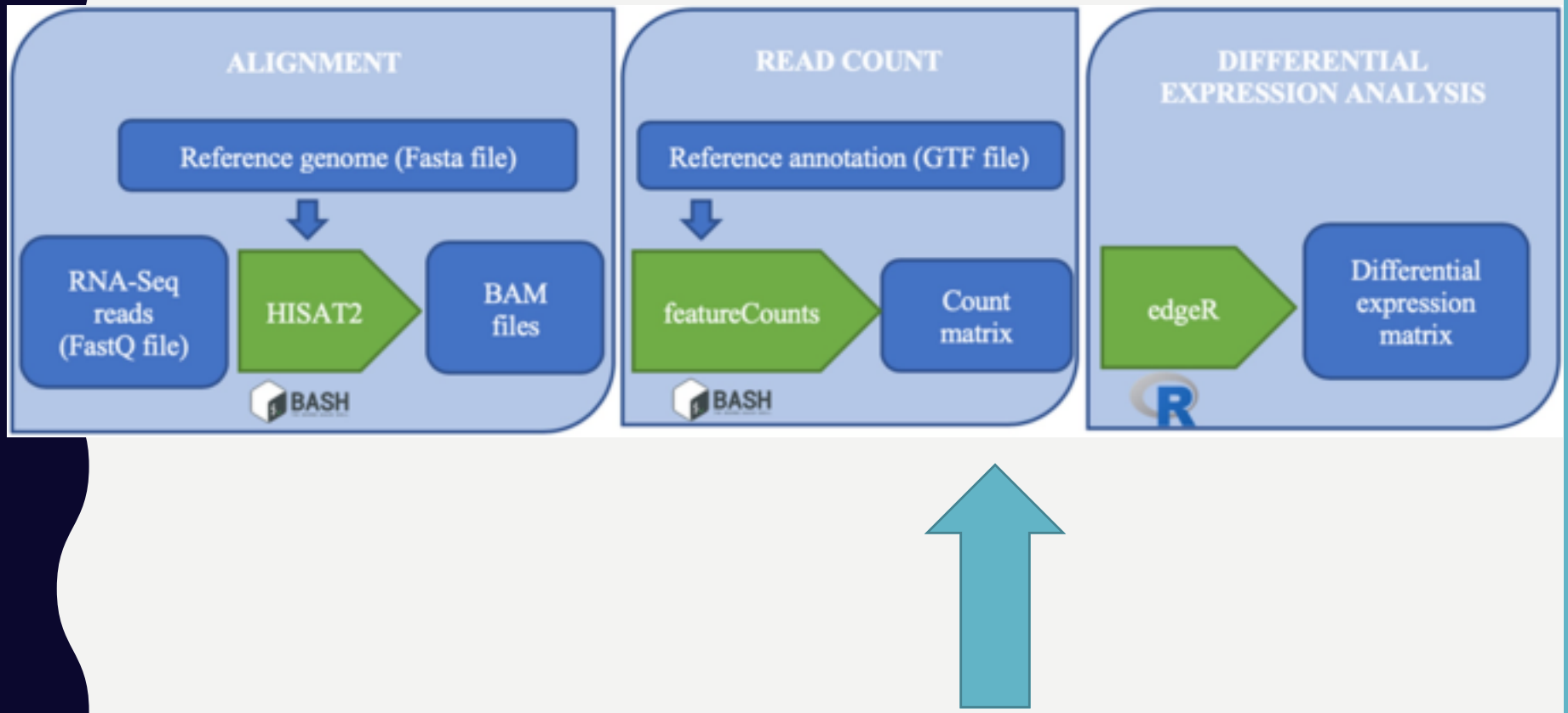
Relations

BioProject	PRJNA217781
SRA	SRP029367


A
Submitter
Platform
Sample
Series
B
many
1
Sample GSM169
Title: Renal cell carcinoma: rccc-25, tumor 2 (part 1 of 2)
Experiment type: Single channel transcriptome hybridization
mRNA source: renal clear cell carcinoma (primary tumor)
Sample GSM379
Title: Renal cell carcinoma: rcch-36, normal 2 (part 1 of 2)
Experiment type: Single channel transcriptome hybridization
mRNA source: normal kidney tissue (as reference)
Sample GSM415
Title: Renal cell carcinoma: rccc-45, normal 2 (part 1 of 2)
Experiment type: Single channel transcriptome hybridization
mRNA source: renal clear cell carcinoma (primary tumor)
SIGNAL_RAW = Mean spot intensity
BKD_RAW = bgr of pixels within a 4x4 block
VALUE = SIGNAL_RAW-SIGNAL_BKD
ID_REF SIGNAL_RAW BKD_RAW VALUE
1 1230767 656100 574667
2 1070722 656100 414622
3 567957 656100 -88143
4 286069 656100 -370031
Series
Title: Renal Cell Carcinoma...
Type: Non-ordered group
GSEXXXXX
1
Platform GPL9
Title: Human Unigene I, part 1
Description: Human cDNA filter from the...
Platform type: Filter
CLONE_ID = clone identification
GB_ACC = GeneBank Accession
ID CLONE_ID GB_ACC GB_ACC
1 256538 h95144 h95180
2 67006 t69711 t70368
3 25807 r12124 r36886
4 42546 r61114 r61836
GSMXXXXX
GPLXXXXX
Sample
Series
Platform

Sample GSM1220267 Query DataSets for GSM1220267

Status	Public on Nov 20, 2014
Title	irrelevant siRNA 1
Sample type	SRA
Source name	Human Embryonic Kidney cell lines
Organism	Homo sapiens
Characteristics	cell type: Human Embryonic Kidney cells cell line: HEK293F treatment: control mov expression: normal
Treatment protocol	MOV10 and irrelevant siRNA treatments were performed at 24 hr intervals three times, overexpression studies involved a single myc-MOV10 transfection
Growth protocol	Cells were grown in serum containing DMEM at 37C
Extracted molecule	total RNA
Extraction protocol	Cells were lysed and RNA was extracted using Trizol Illumina's TruSeq Stranded RNAseq Sample Prep kit was used with 1 ug of total RNA for the construction of sequencing libraries. Indices (barcodes) were included to be able to differentiate the sequences from each sample. The adapter sequence used was AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG (NNNNNN = 6 nt barcode index in .fastq file name)
Library strategy	RNA-Seq
Library source	transcriptomic
Library selection	cDNA
Instrument model	Illumina HiSeq 2000
Data processing	The libraries were quantitated by qPCR, pooled all together and sequenced on two lanes for 101 cycles from one end of the cDNA fragments on a HiSeq2000 using a TruSeq SBS sequencing kit version 3 Base calling and de-multiplexing of samples within each lane were done with Casava 1.8.2. Each of the two fastq files per sample were individually trimmed for quality using trimmomatic 0.22 with parameters TRAILING:20 MINLEN:30 (minimum PHRED quality score accepted 20, read length accepted 30). Each pair of technical replicates had similar quality and percent reads removed, so they were concatenated into one fastq file per sample before alignment. Alignments to UCSC hg19 were done using Tophat2 2.0.8 with parameters --coverage-search -p 8 -N 7 --read-edit-dist 7 --library-type fr-firststrand. Read counts for each gene were generated using htseq-count (version 0.5.3p9 from samtools 0.1.18) with parameters -s reverse -m intersection-nonempty and combined into one tab-delimited text file for all 20 samples Genome_build: UCSC hg19 Supplementary_files_format_and_content: a single tab-delimited text file of raw read counts were used directly in edgeR 3.2.1for statistical analysis and thus are the "normalized" data.


ALIGNMENT
Reference genome (Fasta file)
RNA-Seq reads (FastQ file)
HISAT2
BAM files
BASH
READ COUNT
Reference annotation (GTF file)
featureCounts
Count matrix
BASH
DIFFERENTIAL EXPRESSION ANALYSIS
edgeR
Differential expression matrix
R

Supplementary file	Size	Download	File type/resource
GSE50499_GEO_Ceman_counts.txt.gz	320.2 Kb	(ftp)(http)	TXT

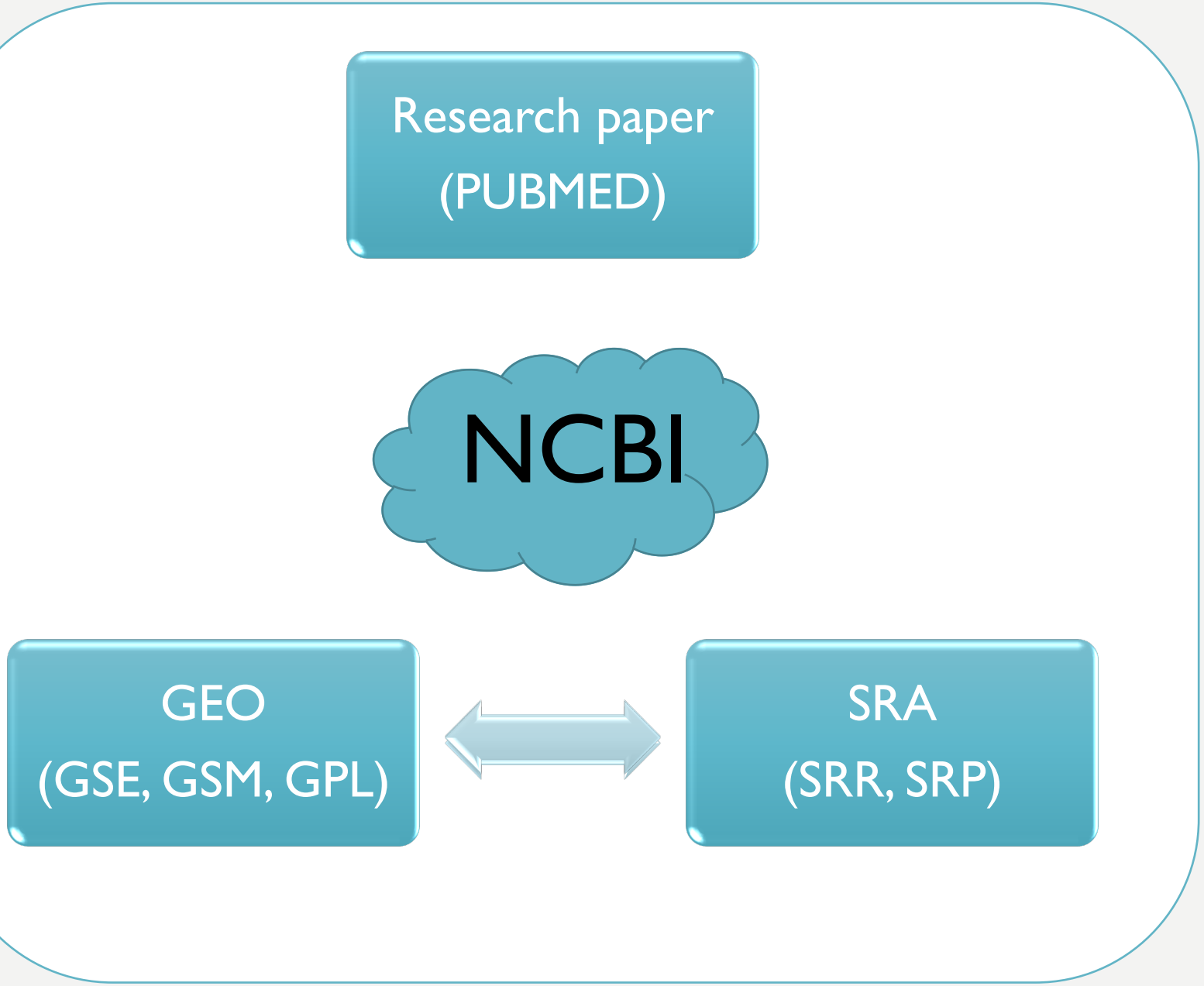
SRA Run Selector

Raw data are available in SRA

Processed data are available on Series record

- This assumes you completely understand the bioinformatic pipeline used and that it is still acceptable to current standards

  Was the latest reference genome used?
  Some aligners are outdated (ex. TopHat)


Research paper
(PUBMED)
NCBI
GEO
(GSE, GSM, GPL)
SRA
(SRR, SRP)

WHAT IS SRA?

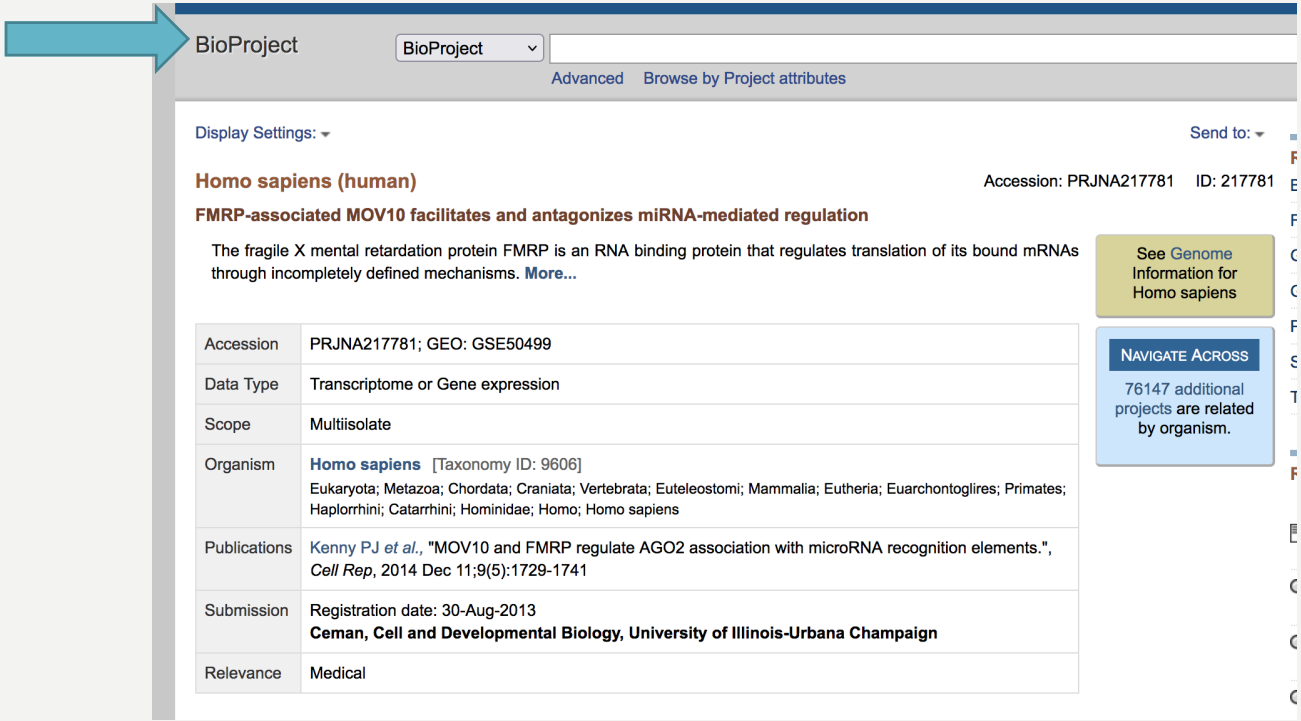
- The Sequence Read Archive (SRA) is an archive for high throughput sequencing data, publicly accessible, for the purpose of enhancing reproducibility in the scientific community.

THERE ARE LEVELS OF SRA ENTITIES AND THEIR ACCESSIONS:

- **STUDY** with accessions in the form of SRP, ERP, or DRP
- **SAMPLE** with accessions in the form of SRS, ERS, or DRS
- **EXPERIMENT** with accessions in the form of SRX, ERX, or DRX
- **RUN** with accessions in the form of SRR, ERR, or DRR

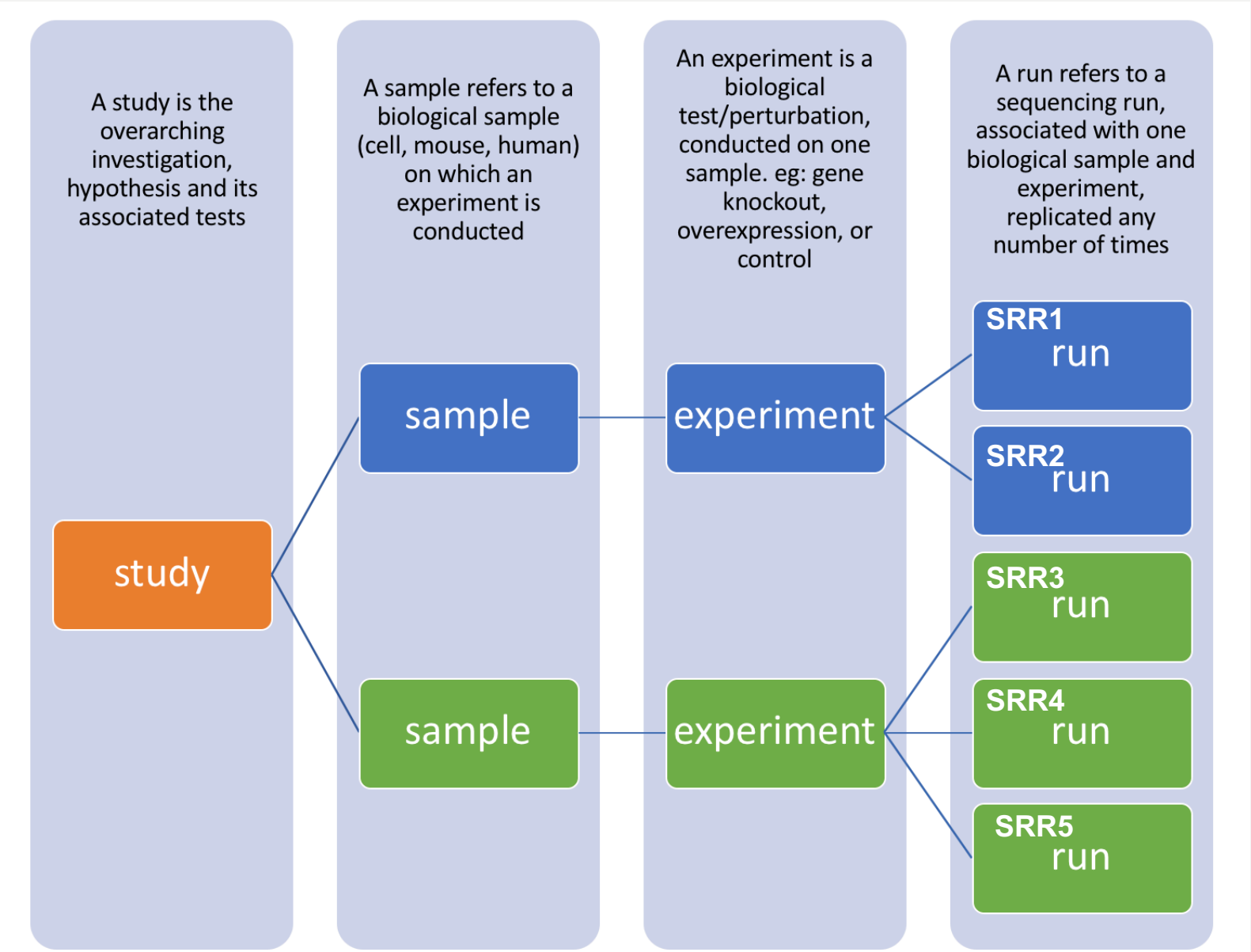
SRA HIERARCHY

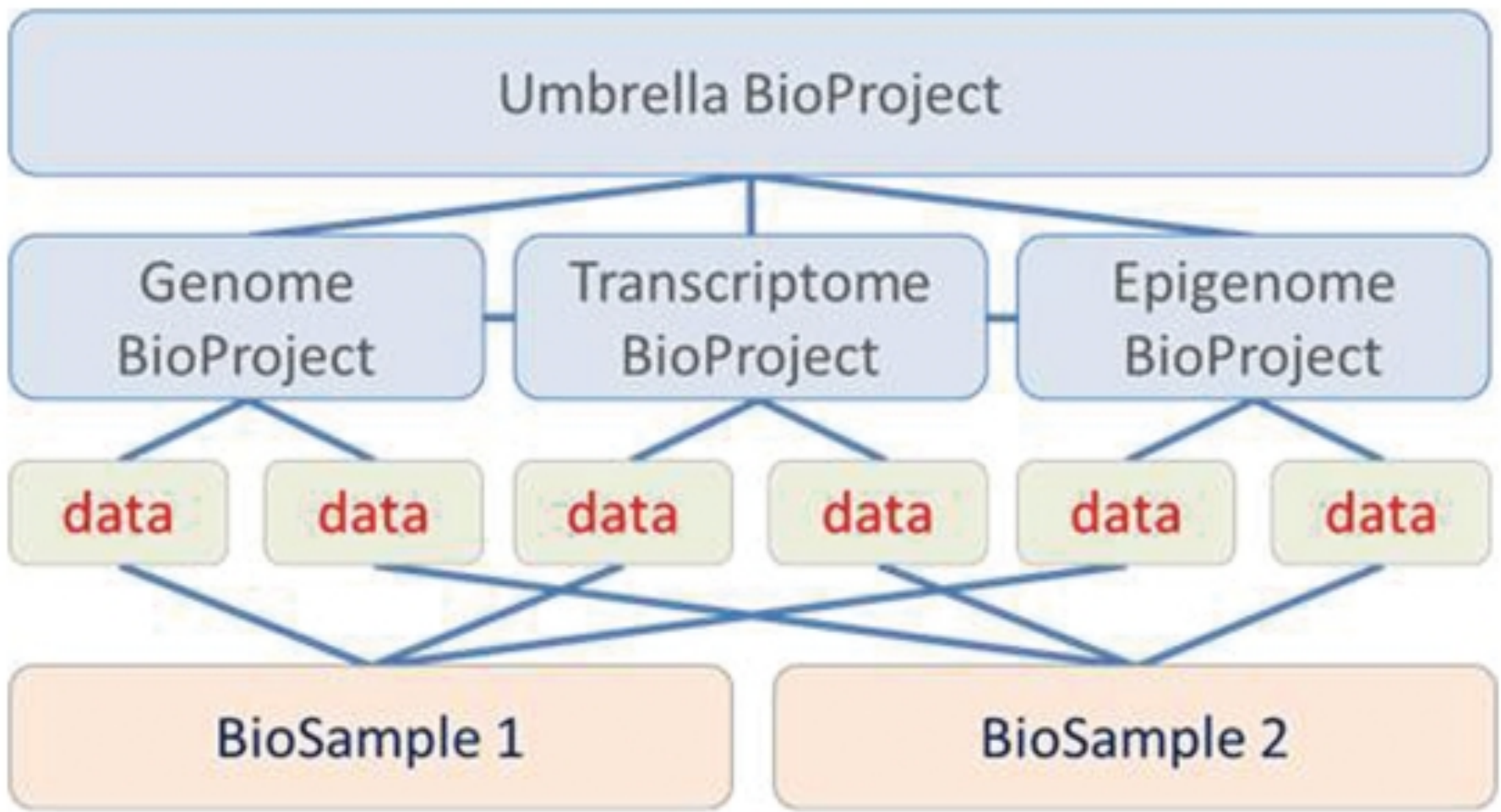
Stands for	NCBI Prefix
BioProject	PRJNA123456



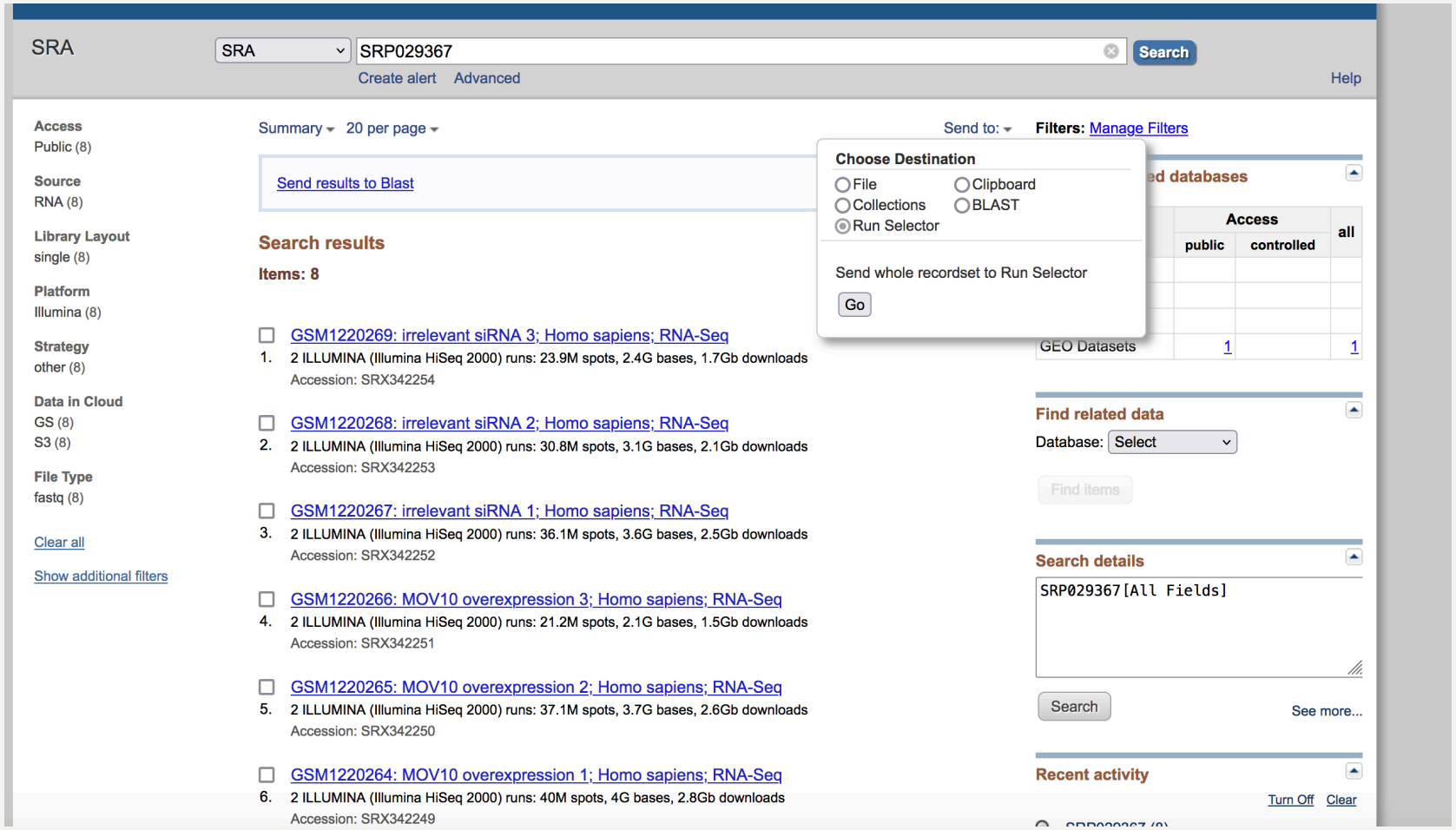
SRA HIERARCHY

Stands for	NCBI Prefix
BioProject	PRJNA123456
Sequence Read Archive Study	SRP123456
Sequence Read Archive Experiment	SRX123456
Sequence Read Archive Run	SRR123456


A study is the overarching investigation, hypothesis and its associated tests
A sample refers to a biological sample (cell, mouse, human) on which an experiment is conducted
An experiment is a biological test/perturbation, conducted on one sample. eg: gene knockout, overexpression, or control
A run refers to a sequencing run, associated with one biological sample and experiment, replicated any number of times
study
sample
sample
experiment
experiment
SRR1 run
SRR2 run
SRR3 run
SRR4 run
SRR5 run


Umbrella BioProject
Genome BioProject
Transcriptome BioProject
Epigenome BioProject
data
data
data
data
data
data
BioSample 1
BioSample 2

STEP 1: COLLECT SRR NUMBERS USING RUN SELECTOR



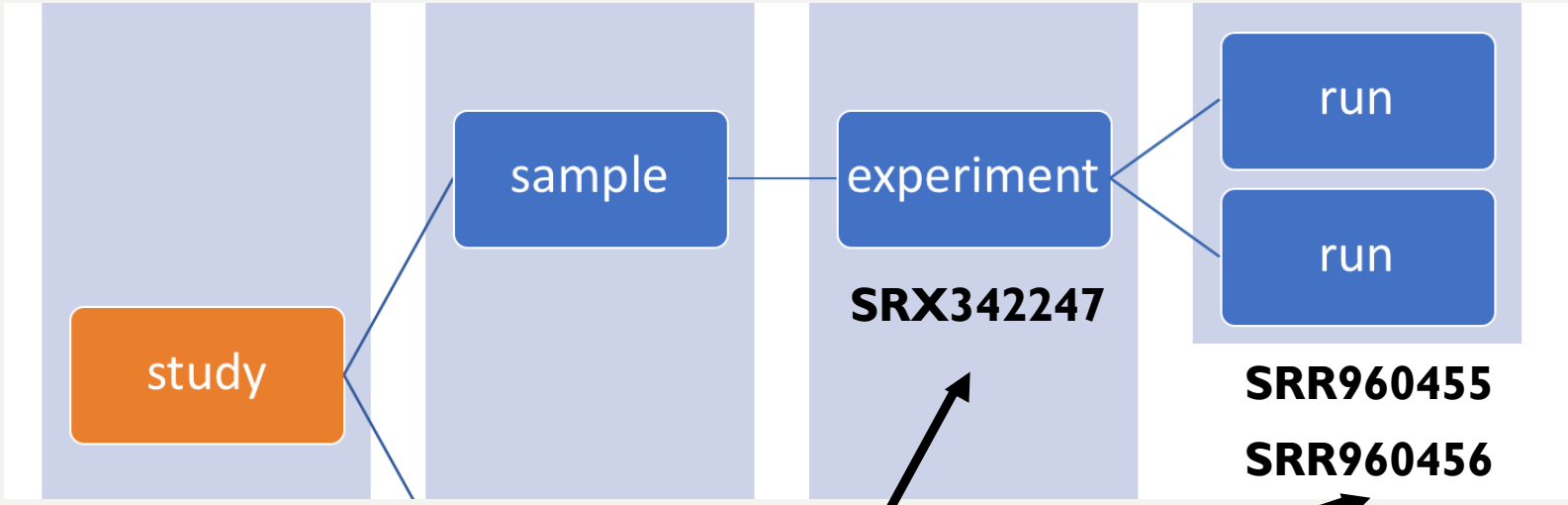
SRA Run Selector

Log in to NIH

Found 16 Items

	Run	BioSample	Bases	Bytes	Experiment	GEO_Accession	mov_expression	create_date	Sample Name	TREATMENT
1	SRR960455	SAMN02340011	2.74 G	1.90 Gb	SRX342247	GSM1220262	low	2013-08-30 13:30:00Z	GSM1220262	MOV10 knockdown
2	SRR960456	SAMN02340011	2.53 G	1.74 Gb	SRX342247	GSM1220262	low	2013-08-30 13:29:00Z	GSM1220262	MOV10 knockdown
3	SRR960457	SAMN02340009	1.62 G	1.12 Gb	SRX342248	GSM1220263	low	2013-08-30 13:38:00Z	GSM1220263	MOV10 knockdown
4	SRR960458	SAMN02340009	1.49 G	1.03 Gb	SRX342248	GSM1220263	low	2013-08-30 13:30:00Z	GSM1220263	MOV10 knockdown
5	SRR960459	SAMN02340010	2.08 G	1.44 Gb	SRX342249	GSM1220264	high	2013-08-30 13:32:00Z	GSM1220264	MOV10 overexpression
6	SRR960460	SAMN02340010	1.92 G	1.32 Gb	SRX342249	GSM1220264	high	2013-08-30 13:28:00Z	GSM1220264	MOV10 overexpression
7	SRR960461	SAMN02340016	1.93 G	1.34 Gb	SRX342250	GSM1220265	high	2013-08-30 13:32:00Z	GSM1220265	MOV10 overexpression
8	SRR960462	SAMN02340016	1.78 G	1.23 Gb	SRX342250	GSM1220265	high	2013-08-30 13:30:00Z	GSM1220265	MOV10 overexpression
9	SRR960463	SAMN02340013	1.10 G	778.10 Mb	SRX342251	GSM1220266	high	2013-08-30 13:26:00Z	GSM1220266	MOV10 overexpression
10	SRR960464	SAMN02340013	1.02 G	721.91 Mb	SRX342251	GSM1220266	high	2013-08-30 13:27:00Z	GSM1220266	MOV10 overexpression
11	SRR960465	SAMN02340014	1.88 G	1.30 Gb	SRX342252	GSM1220267	normal	2013-08-30 13:26:00Z	GSM1220267	control
12	SRR960466	SAMN02340014	1.73 G	1.20 Gb	SRX342252	GSM1220267	normal	2013-08-30 13:26:00Z	GSM1220267	control
13	SRR960467	SAMN02340012	1.61 G	1.11 Gb	SRX342253	GSM1220268	normal	2013-08-30 13:24:00Z	GSM1220268	control
14	SRR960468	SAMN02340012	1.48 G	1.02 Gb	SRX342253	GSM1220268	normal	2013-08-30 13:26:00Z	GSM1220268	control
15	SRR960469	SAMN02340015	1.24 G	879.83 Mb	SRX342254	GSM1220269	normal	2013-08-30 13:25:00Z	GSM1220269	control
16	SRR960470	SAMN02340015	1.15 G	815.22 Mb	SRX342254	GSM1220269	normal	2013-08-30 13:24:00Z	GSM1220269	control

Notice, there are 8 samples however for each sample there were (2) runs making the final total 16

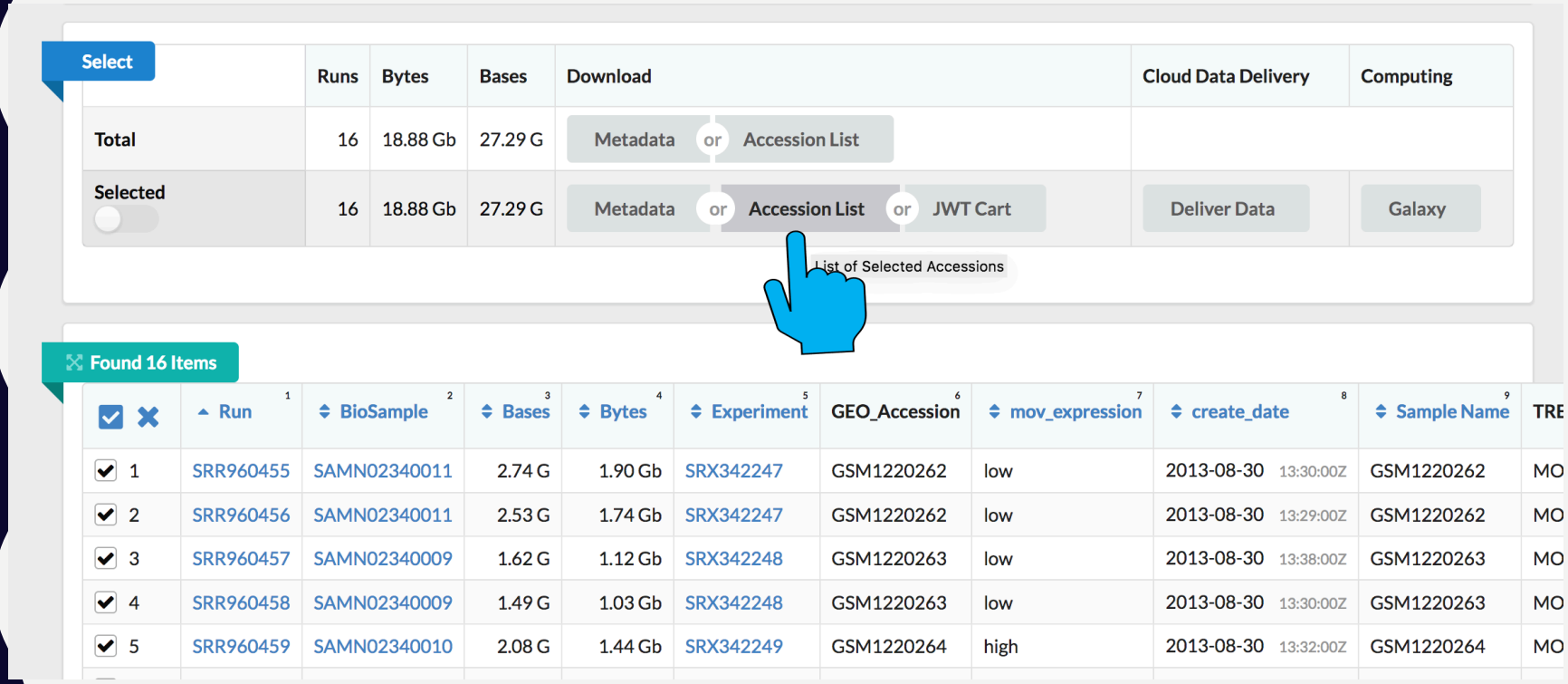


SRA Run Selector

Log in to NIH

Found 16 Items

	Run	BioSample	Bases	Bytes	Experiment	GEO_Accession	mov_expression	create_date	Sample Name	TREATMENT
1	SRR960455	SAMN02340011	2.74 G	1.90 Gb	SRX342247	GSM1220262	low	2013-08-30 13:30:00Z	GSM1220262	MOV10 knockdown
2	SRR960456	SAMN02340011	2.53 G	1.74 Gb	SRX342247	GSM1220262	low	2013-08-30 13:29:00Z	GSM1220262	MOV10 knockdown



Select only the desired samples!

SRR960455
SRR960456
SRR960457
SRR960458
SRR960459
SRR960460
SRR960461
SRR960462
SRR960463
SRR960464
SRR960465
SRR960466
SRR960467
SRR960468
SRR960469
SRR960470

OUTPUT .TXT FILE WITH LIST OF SRR

Step #1 is complete!

STEP 2: DOWNLOAD SRA-TOOLKIT

- SRA-Toolkit was developed by HPC at the NIH
- Using SRA-Toolkit you will be able to download SRR FASTQ files using the **fasterq-dump** tool

HOW DO YOU RUN A PROGRAM ON THE VACC?

1) Install it in your personal VACC account and then be able to call it

2) Or load it from the shared computing cluster using `module` package

HOW DO YOU RUN A PROGRAM ON THE VACC?

1) Install it in your personal VACC account and then be able to call it

2) Or load it from the shared computing cluster using `module` package

We will do this for Alignment: HiSAT2

STEP 2: DOWNLOAD SRA-TOOLKIT

- Follow Steps 1-3 under ## Download SRA-toolkit
- *Continue to step 4 only if you are confident*

ENVIRONMENTAL VARIABLES

- What are they?
- How do we use them?
- Why do we use them?

USER ENVIRONMENTAL VARIABLES

- User environmental variables are those that are local to the user profile
- These variables are used to store user-specific information such as the PATH to a local installation of libraries
- You do not need the system administrator (staff of VACC) to make changes to these variables, you can do this yourself as the user

BASH PROFILE

- Bash shell uses a few startup files to set up the environment
- Bash profile is a file that Bash runs every time a new Bash session is created
- This is useful because it will run certain code specific to your environment before you begin coding
- Users can *personalize their* environment by modifying the Bash profile
- In Step #4 you are adding the PATH to sratoolkit to your environment

PATH

- PATH is an environmental variable
- PATH contains a list of file system paths where the operating system can find programs to run
- The operating system will look for the program in each of the paths that PATH contains, starting with the first path listed
- *If the operating system can't find the program in the first path, it will look in the second path, and so on – therefore order matters!*

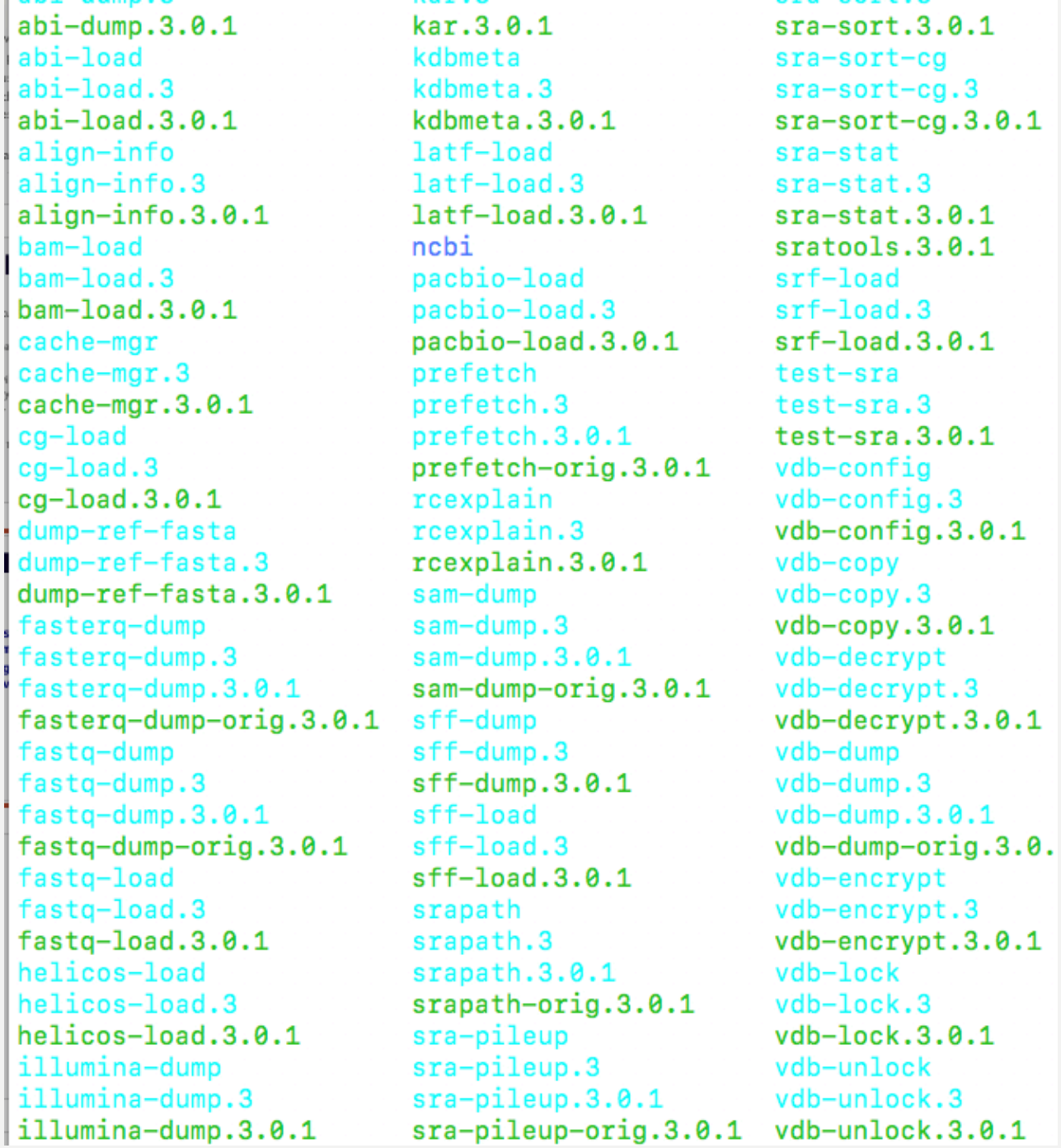
```
echo $PATH
```

```
/users/m/m/mmg232in/miniconda3/condabin:/gpfs1/arch/spack-0.14.2/bin:/usr/local/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/var/cfengine/bin:/usr/lpp/mmfs/bin:/opt/env-switcher/bin:/users/m/m/mmg232in/.local/bin:/users/m/m/mmg232in/bin:/users/m/m/mmg232in/software/sratoolkit.3.0.1-ubuntu64/bin
```

Each path is delimited by a colon (:)

WHY BIN FOLDER?

- The path `/bin` is usually where executable programs are stored
- We will be running the program called “fastq-dump” however there are many more!



By downloading sratoolkit, we now have access to many executable programs, each with their own function.

This will be a reoccurring theme!

WHY DO THIS?

- Every time you run a program you will need to designate the **FULL PATH** to the program to use it!

```
/users/m/m/mmg232in/software/sratoolkit.
3.0.1-ubuntu64/bin/fastq-dump
<inputfile>
```

- **Imagine typing that every time!**

WHY DO THIS?

Instead, now all you have to do is type the program name!

Fastq-dump <inputfile>

Proceed to steps 4 & 5

I expect there will be questions at Step 6

STEPS TO DOWNLOADING DATA FROM GEO:

1. Collect SRR Numbers using RUN SELECTOR
2. Download SRA-ToolKit and set up

   Note: Once this is done, you do not need to do this again!
3. Run sra_fqdump.sh script using output from step #1

   But…in order to run this script we need to submit a job using SLURM!

WHY SUBMIT A JOB?

- It will be annoying to run each sample, one by one:

htseq-count -f bam -s yes -i gene_name -m union sample1.bam annotation.gtf > sample1.count.txt

WAIT FOR IT TO RUN

htseq-count -f bam -s yes -i gene_name -m union sample2.bam annotation.gtf > sample2.count.txt

WAIT FOR IT TO RUN

- Also, some programs…. take a very long time to run!

SUBMITTING A JOB

- Submitting a job to an HPC machine (such as Bluemoon) is done using the batch system.

As of March 2022, the VACC provides three Clusters:

- BlackDiamond
- Bluemoon
- DeepGreen

We will primarily use the **Bluemoon** cluster for any downstream analysis.

SUBMITTING A JOB

- The batch system allows users to submit jobs requesting the resources (nodes, processors, memory, GPUs) that they need.

- The jobs are queued and then run as resources become available.

SUBMITTING A JOB

- All jobs (scripts with file extension .sh) are submitted to a batch system called SLURM

BASIC STEPS:

1. Log in to VACC
2. Write job script
3. Submit Job
4. Monitor job and wait for it to run
5. Retrieve your output

BASIC STEPS:

1. Log in to VACC
2. **Write job script**
3. **Submit Job**
4. Monitor job and wait for it to run!
5. Retrieve your output

WHAT ARE SLURM DIRECTIVES?

- At the top of the job script will always be several lines that start with #SBATCH.
- The Slurm directives provide the job setup information used by Slurm.
- This information is then followed by the commands to be executed in the script.

COMMON SLURM DIRECTIVES

```
#!/bin/bash
#SBATCH --partition=bluemoon
#SBATCH --nodes=1
#SBATCH --ntasks=4
#SBATCH --mem=50G
#SBATCH --time=30:00:00
#SBATCH --job-name=RNAseq
# %x=job-name %j=jobid
#SBATCH --output=%x_%j.out
```

COMMON SLURM DIRECTIVES

```
#!/bin/bash
#SBATCH --partition=bluemoon
#SBATCH --nodes=1
#SBATCH --ntasks=2
#SBATCH --mem=50G
#SBATCH --time=30:00:00
#SBATCH --job-name=RNAseq
# %x=job-name %j=jobid
#SBATCH --output=%x_%j.out
```

shebang = used to tell the linux OS which interpreter to use

COMMON SLURM DIRECTIVES

```
#!/bin/bash
#SBATCH --partition=bluemoon
#SBATCH --nodes=1
#SBATCH --ntasks=2
#SBATCH --mem=50G
#SBATCH --time=30:00:00
#SBATCH --job-name=RNAseq
# %x=job-name %j=jobid
#SBATCH --output=%x_%j.out
```

partition = default is bluemoon if not specified

Other partitions are available.
This is important to know as some jobs will take longer to run!

Other Partitions:

Partition	Intended Use	Max Runtime
bluemoon	General computing – default partition	30 hours
short	General computing with short runtime	3 hours
week	General computing with longer runtime	7 days
bigmem	Large memory requirements computing	30 hours
bigmemwk	Large memory requirements with longer runtime	7 days

You can check partition usage using the following command:

```
sinfo -p partition_name
sinfo -p bluemoon
```

COMMON SLURM DIRECTIVES

```
#!/bin/bash
#SBATCH --partition=bluemoon
#SBATCH --nodes=1
#SBATCH --ntasks=2
#SBATCH --mem=50G
#SBATCH --time=30:00:00
#SBATCH --job-name=RNAseq
# %x=job-name %j=jobid
#SBATCH --output=%x_%j.out
```

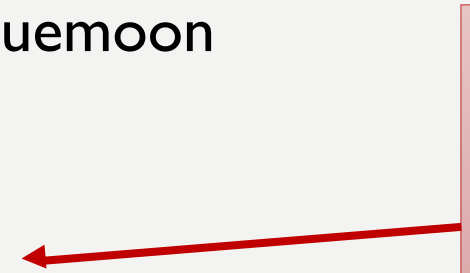
Node: A "node" is a server in the cluster. Each node has is configured with a certain number of cores (CPUs).

Task: A "task" is a process sent to a core. By default, 1 core is assigned per 1 task

Recommend that you begin with 1 node and 2 processes

COMMON SLURM DIRECTIVES

```
#!/bin/bash
#SBATCH --partition=bluemoon
#SBATCH --nodes=1
#SBATCH --ntasks=2
#SBATCH --mem=50G
#SBATCH --time=30:00:00
#SBATCH --job-name=RNAseq
# %x=job-name %j=jobid
#SBATCH --output=%x_%j.out
```



If your job requires more than 1G of memory – need to specify this

COMMON SLURM DIRECTIVES

```
#!/bin/bash
#SBATCH --partition=bluemoon
#SBATCH --nodes=1
#SBATCH --ntasks=2
#SBATCH --mem=50G
#SBATCH --time=30:00:00
#SBATCH --job-name=RNAseq
# %x=job-name %j=jobid
#SBATCH --output=%x_%j.out
```

Walltime is the maximum amount of time your job will run.

STEPS TO DOWNLOADING DATA FROM GEO:

1. Collect SRR Numbers using RUN SELECTOR
2. Download SRA-ToolKit and set up

   Note: Once this is done, you do not need to do this again!

3. Run sra_fqdump.sh script using output from step #1

PLEASE DOWNLOAD THE SCRIPT FROM THIS LOCATION:

```
cp -r /gpfs1/cl/mmg232/course_materials/download_from_SRA .
```

RUN SCRIPT USING:

```
sbatch sra_fqdump.sh
```

CHECK SIZE:

```
du –h SRR17379677.fastq.gz
524M SRR17379677.fastq.gz
```