Lecture 17

Next Generation Sequence mapping - HPC

MCB 416A/516A
Statistical Bioinformatics and Genomic Analysis

Prof. Lingling An
Univ of Arizona

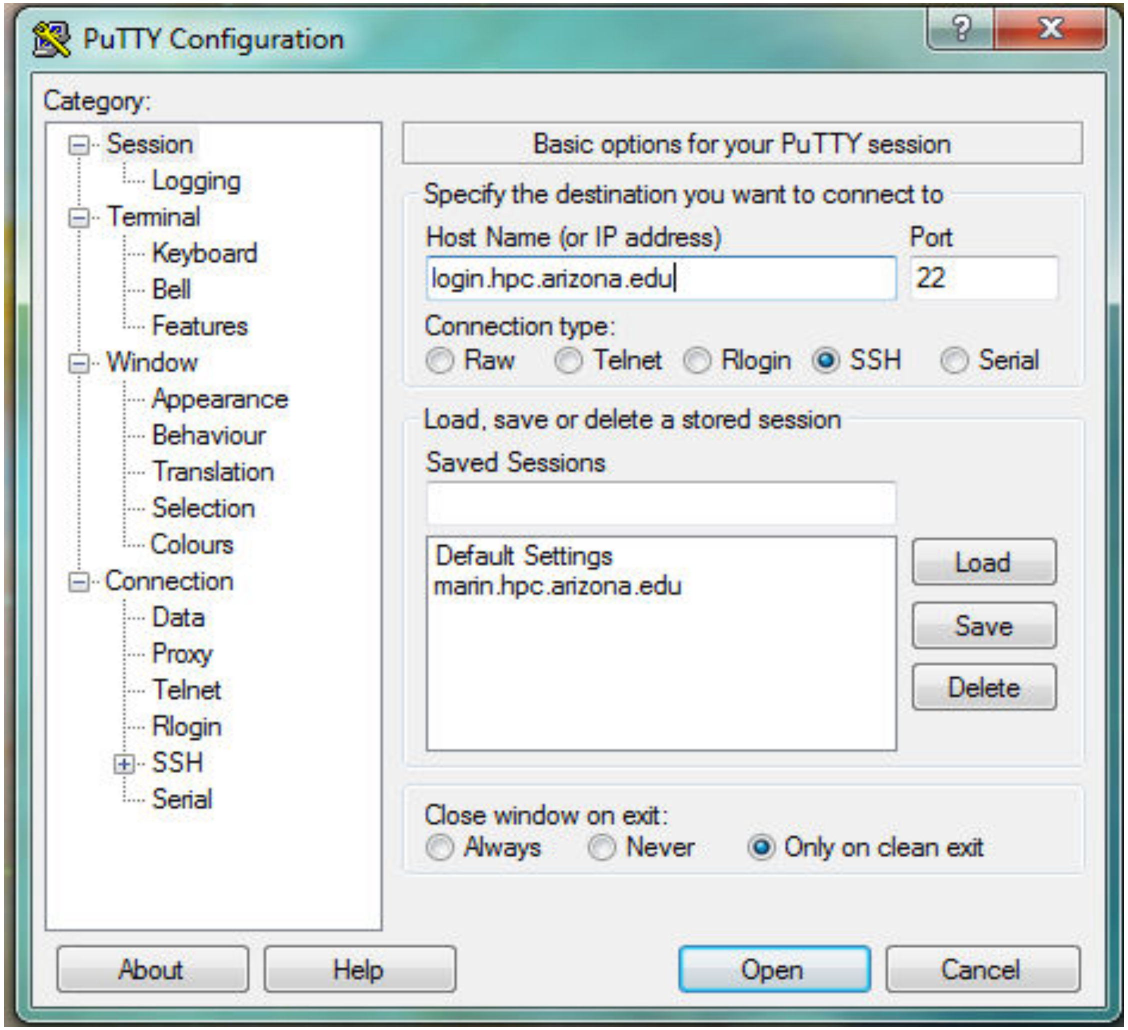
Schedule

Mon	Wed
	4/4 NGS - alignment
4/9 NSG- DE (R code)	4/11 metagenomics - Intr
4/16 metagenomics - code	4/18 pathway
4/23 project 2 presentation	4/25 project 2 presentation
4/30 QA	5/2 project 3 report

Connect to HPC: SSH Clients for Windows users

Microsoft Windows Users:

- Example using PuTTY to connect to the UA HPC login nodes.
  http://softwarelicense.arizona.edu/ssh-clients-windows-and-mac
- Host Name:
  hpc.arizona.edu



PuTTY Configuration

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): login.hpc.arizona.edu

Port: 22

Connection type:
○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions:

Default Settings
marin.hpc.arizona.edu

Close window on exit:
○ Always ○ Never ● Only on clean exit

Connect to HPC: Unix/Linux users

- At the shell prompt type: ssh then the host name destination
  Example: **ssh hpc.arizona.edu**

- If your workstation username is not your NetID you may need to type ssh username@host name destination
  Example: **ssh username@hpc.arizona.edu**

Connect to HPC: Mac users

- Open the Terminal application

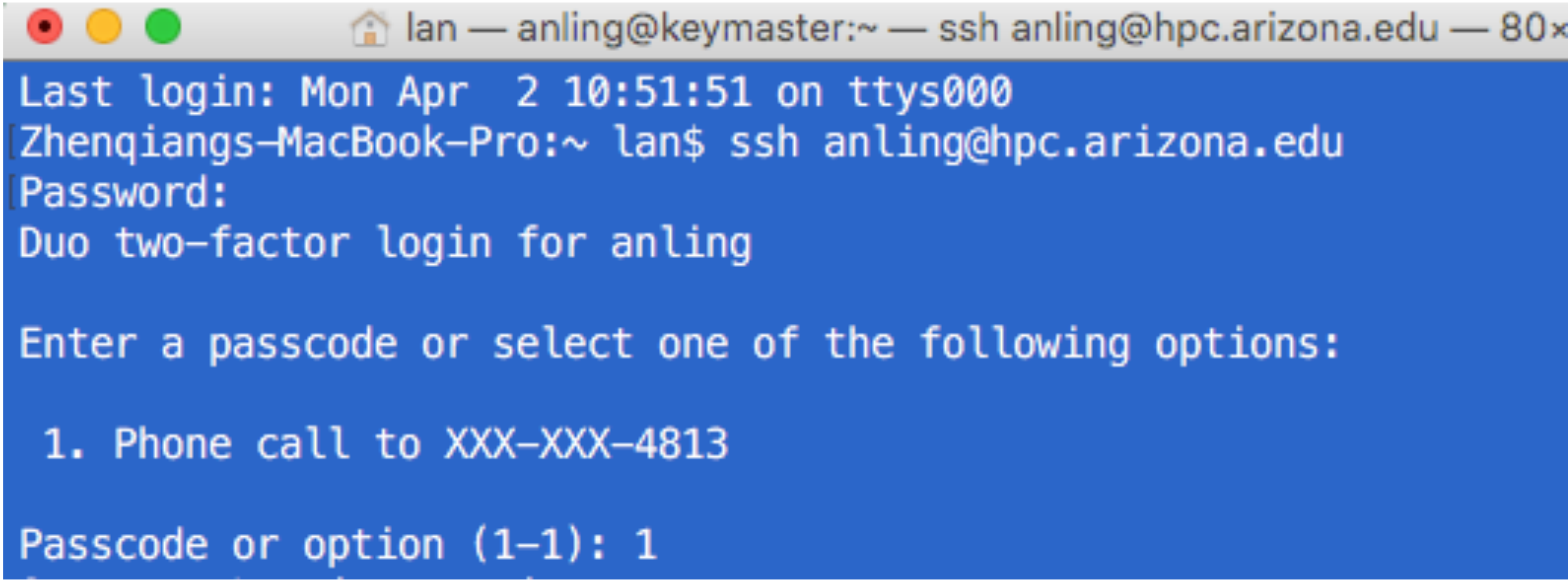
- At the shell prompt type: ssh then the host name destination
  Example: **ssh hpc.arizona.edu**

- If your workstation username is not your NetID you may need to type ssh username@host name destination
  Example: **ssh username@hpc.arizona.edu**

- **After accessing the terminal,** there will be a prompt for NetID+ authentication.

- SSH login session with NetID authentication, then NetID+ 2nd factor authentication.

(details can be found at https://docs.hpc.arizona.edu/display/UAHPC/System+Access

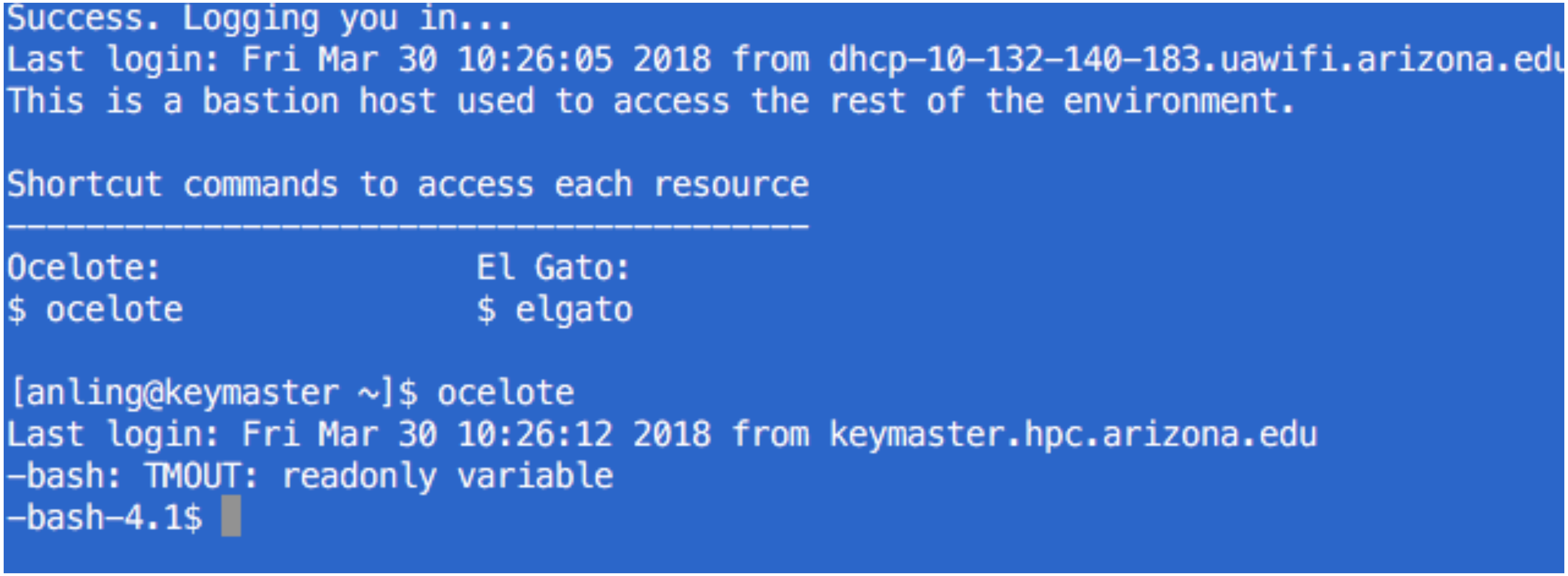


```
Last login: Mon Apr  2 10:51:51 on ttys000
[Zhenqiangs-MacBook-Pro:~ lan$ ssh anling@hpc.arizona.edu
[Password:
Duo two-factor login for anling

Enter a passcode or select one of the following options:

 1. Phone call to XXX-XXX-4813

Passcode or option (1-1): 1
```



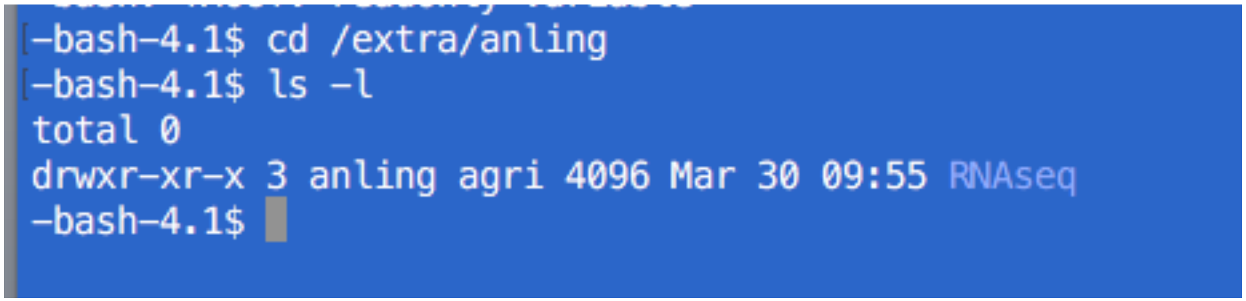
```
Success. Logging you in...
Last login: Fri Mar 30 10:26:05 2018 from dhcp-10-132-140-183.uawifi.arizona.edu
This is a bastion host used to access the rest of the environment.

Shortcut commands to access each resource
-----------------------------------------
Ocelote:                El Gato:
$ ocelote               $ elgato

[anling@keymaster ~]$ ocelote
Last login: Fri Mar 30 10:26:12 2018 from keymaster.hpc.arizona.edu
-bash: TMOUT: readonly variable
-bash-4.1$
```

Take a look at my hpc account …

ls -l	directory listing
cd	change directory
cp file1 file2	copy file 1 to file 2
mv file1 file2/dir	move or rename file1 to file2 or a directory “dir”
rm <-option> file/dir	remove file or directory “dir”
mkdir dir	create a directory “dir”
head <-option> file	output the first 10 lines of file
tail <-option> file	output the last 10 lines of file
pwd	show current directory

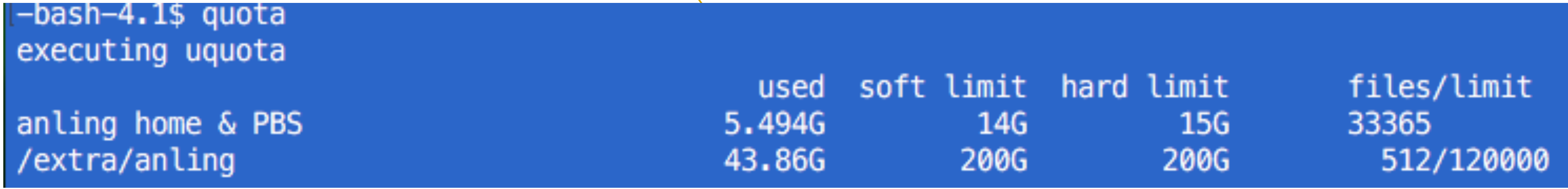


```
-bash-4.1$ cd /extra/anling
-bash-4.1$ ls -l
total 0
drwxr-xr-x 3 anling agri 4096 Mar 30 09:55 RNAseq
-bash-4.1$
```

To check the usage of your spaces

- `quota`

My home folder space



```
-bash-4.1$ quota
executing uquota
                                           used  soft limit  hard limit      files/limit
anling home & PBS                        5.494G         14G         15G      33365
/extra/anling                            43.86G        200G        200G        512/120000
```

My allocated extra space

Again!
Go to your own home folder: `cd ~`
Go to your allocated space: `cd /extra/anling`

Create a new folder: `mkdir check`

Download a dataset and transfer it to HPC

Assume “example.fastq” is downloaded from D2L and saved to your own computer.

Transfer example.fastq file from your computer to hpc: (use another window)

```
scp example.fastq anling@filexfer.hpc.arizona.edu:~/temp
```

Now I want to copy the file in my home folder to extra space

```
cd temp
ll

cp example.fastq /extra/anling/check/
cd /extra/anling/check/
ll
```

Remember to add / after the folder name, otherwise your original file will be renamed

iRODS for transfering big data;

(download files from online directly into HPC):

```
wget website/filename
```

Raw data formats of 454, Illumina and SOLiD

Can be recognized by the file suffix (extension).

Illumina: *.fastq or *.fq (one file per lane or barcoded sample)

```
ZmB73_6DAP_RNA.fastq (10 Gb)
```

SOLiD: a pair of *.csfasta and *.qual (per lane)

```
solid309_20100923_FRAG_BC_yadegari_F3_6DAP.csfasta (6.2 Gb)
solid309_20100923_FRAG_BC_yadegari_F3_6DAP.qual (14 Gb)
solid309_20100923_FRAG_BC_yadegari_F3_6DAP.stats (78 kb)
```

454: a pair of *.fasta and *.qual (per sample)

```
CFGU.fasta (200 Mb)
CFGU.qual
```

If you see suffix like *.tar or *.tar.gz or *.gz, decompress them first.

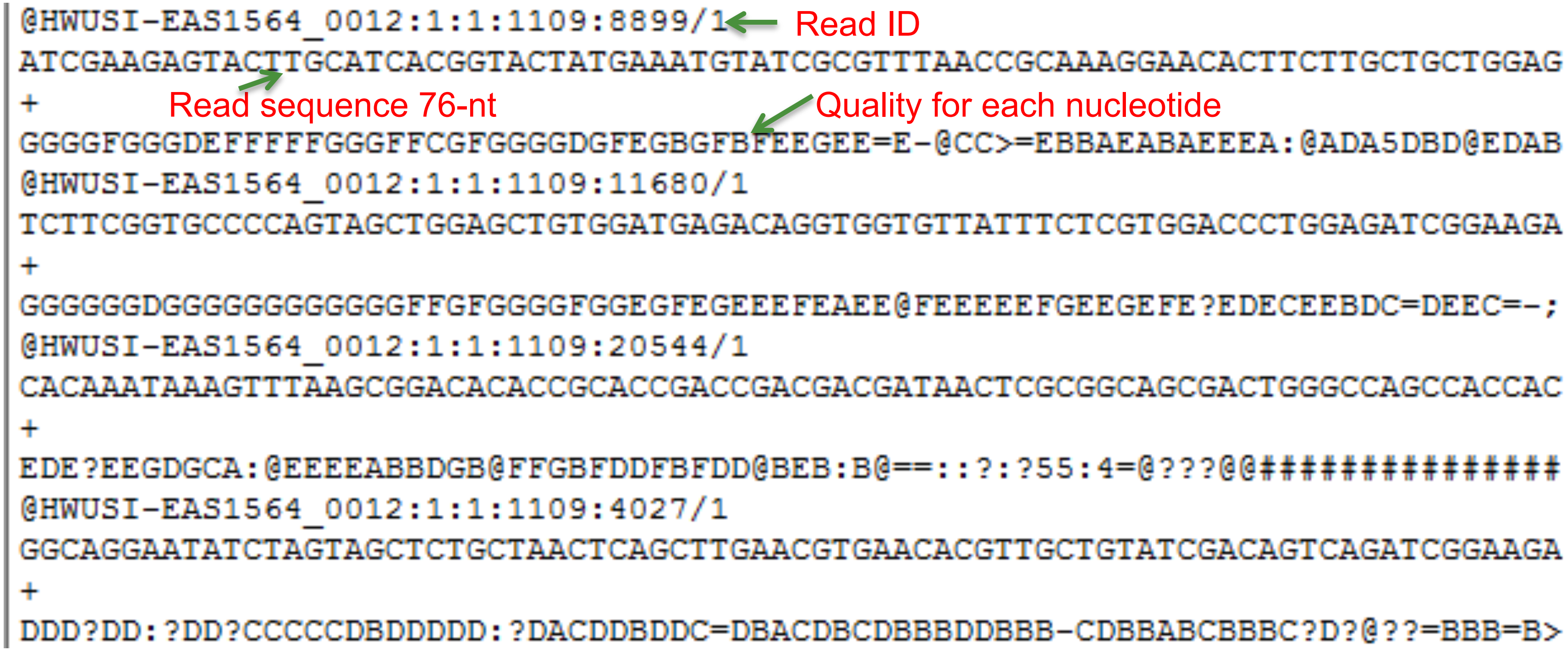
```
tar xvzf file.tar.gz
```

If you download raw data from GEO’s short read archive (SRA).
You need to use sra toolkit to convert the format.

What information inside the file (Illumina fastq):

Four lines of information for each read

wc to count how many lines



```
@HWUSI-EAS1564_0012:1:1:1109:8899/1
ATCGAAGAGTACTTGCATCACGGTACTATGAAATGTATCGCGTTTAACCGCAAAGGAACACTTCTTGCTGCTGGAG
+
GGGFGGGDEFFFFFGGGFFCGFGGGGDGFEGBGFBFEEGEE=E-@CC>=EBBAEABAEEEA:@ADA5DBD@EDAB
@HWUSI-EAS1564_0012:1:1:1109:11680/1
TCTTCGGTGCCCCAGTAGCTGGAGCTGTGGATGAGACAGGTGGTGTTATTTCTCGTGGACCCTGGAGATCGGAAGA
+
GGGGGGDGGGGGGGGGGFFGFGGGGFGGEGFEGEEEFEAEE@FEEEEEFGEEGEFE?EDECEEBDC=DEEC=-;
@HWUSI-EAS1564_0012:1:1:1109:20544/1
CACAAATAAAGTTTAAGCGGACACACCGCACCGACCGACGACGATAACTCGCGGCAGCGACTGGGCCAGCCACCAC
+
EDE?EEGDGCA:@EEEEABBDGB@FFGBFDDFBFDD@BEB:B@==::?:?55:4=@???@@###############
@HWUSI-EAS1564_0012:1:1:1109:4027/1
GGCAGGAATATCTAGTAGCTCTGCTAACTCAGCTTGAACGTGAACACGTTGCTGTATCGACAGTCAGATCGGAAGA
+
DDD?DD:?DD?CCCCCDBDDDDD:?DACDDBDDC=DBACDBCDBBBDDBBB-CDBBABCBBBC?D?@??=BBB=B>
```

Read ID

Read sequence 76-nt

Quality for each nucleotide

vi example.fastq
wc example.fastq

vi/vim text editor

vi file	open file with vi editor
:wq	exit vi and save changes
:q!	exit vi without saving changes
[esc]	enter vi command mode
i	insert before cursor
h	move left
j	move down
k	move up
l	move right

Check sequence quality in HPC: fastqc

```
module ava
module load fastqc
fastqc example.fastq
#################
 cp *fastqc* ~/temp/
```

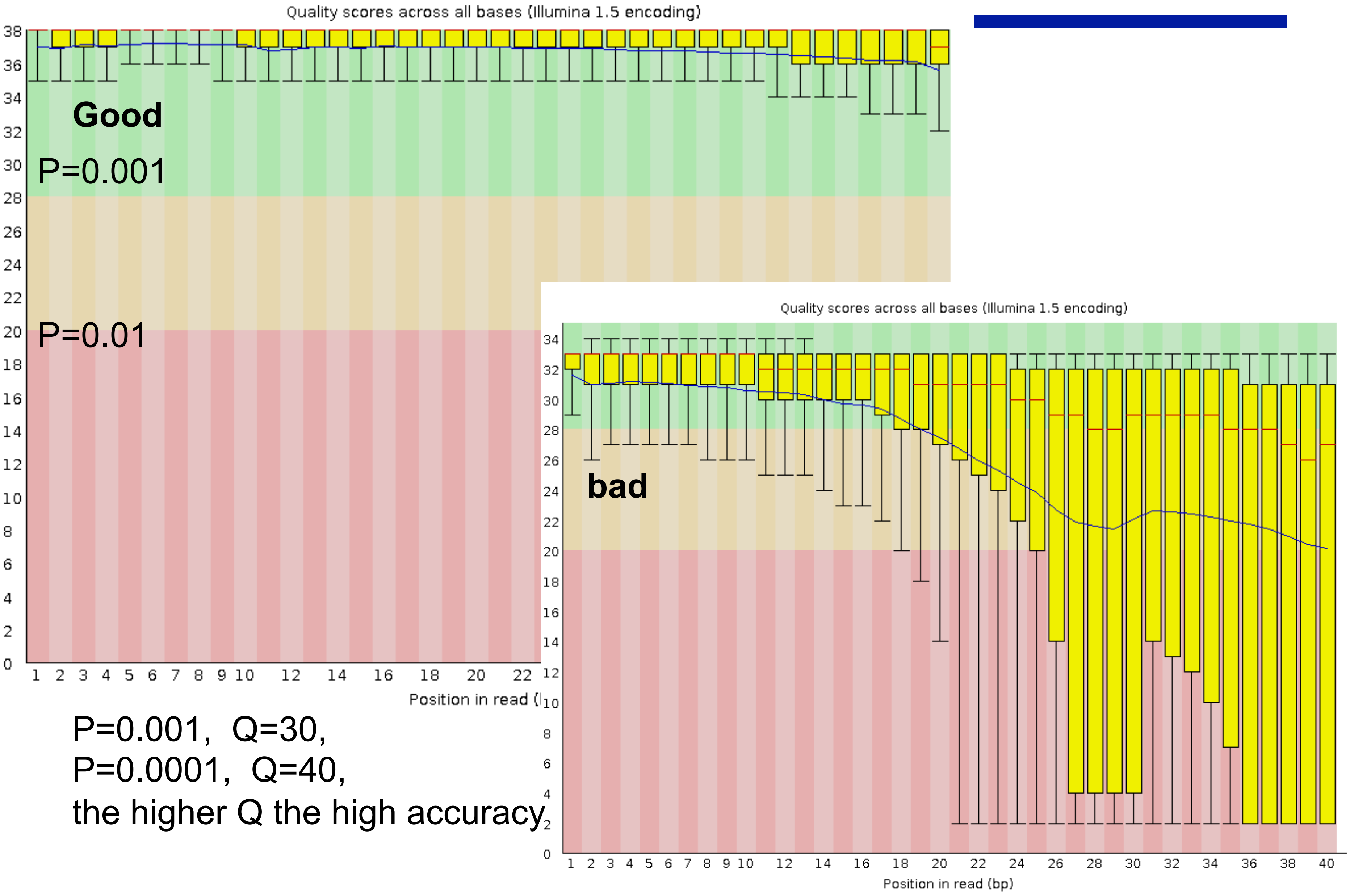
Transfer files from HPC to local computer

```
scp anling@filexfer.hpc.arizona.edu:~/temp/
*fastqc* .
```

(**then** take a look at the .html file!)

Good or bad quality per base?

Quality drops from start to end.



Good

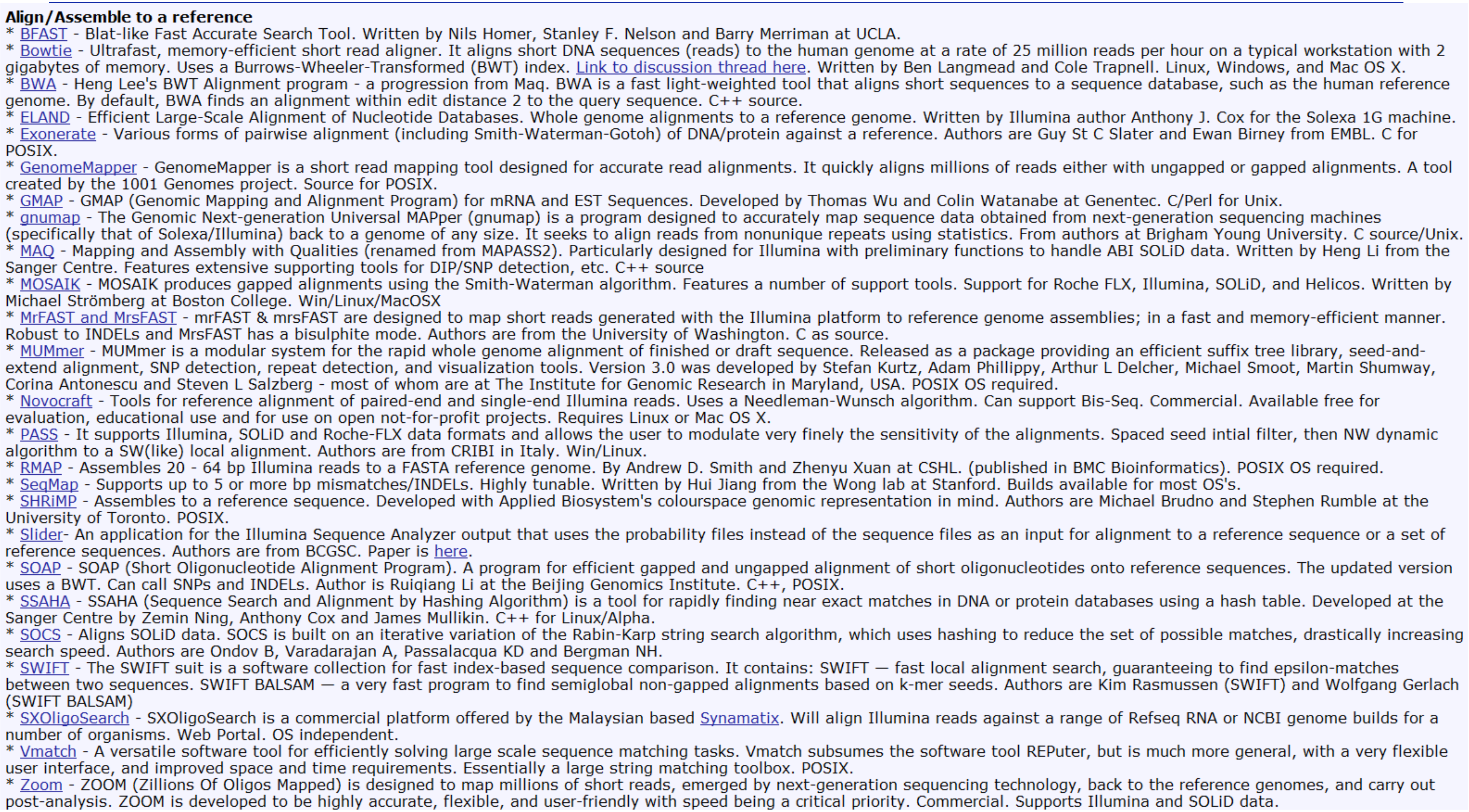
P=0.001

P=0.01

bad

P=0.001, Q=30,
P=0.0001, Q=40,
the higher Q the high accuracy

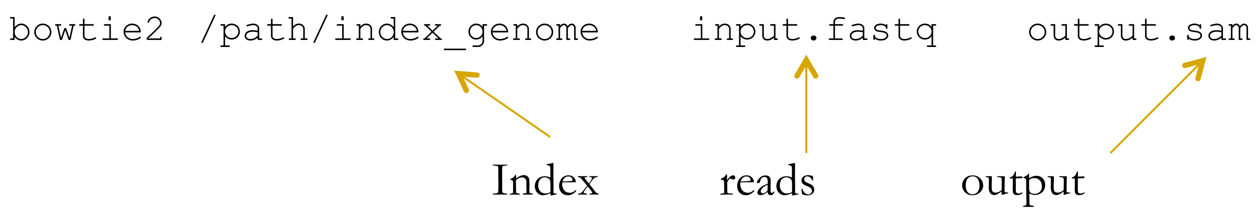
Tools for mapping short reads



Align/Assemble to a reference
* BFAST - Blat-like Fast Accurate Search Tool. Written by Nils Homer, Stanley F. Nelson and Barry Merriman at UCLA.
* Bowtie - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hour on a typical workstation with 2 gigabytes of memory. Uses a Burrows-Wheeler-Transformed (BWT) index. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell. Linux, Windows, and Mac OS X.
* BWA - Heng Lee's BWT Alignment program - a progression from Maq. BWA is a fast light-weighted tool that aligns short sequences to a sequence database, such as the human reference genome. By default, BWA finds an alignment within edit distance 2 to the query sequence. C++ source.
* ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony J. Cox for the Solexa 1G machine.
* Exonerate - Various forms of pairwise alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Birney from EMBL. C for POSIX.
* GenomeMapper - GenomeMapper is a short read mapping tool designed for accurate read alignments. It quickly aligns millions of reads either with ungapped or gapped alignments. A tool created by the 1001 Genomes project. Source for POSIX.
* GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C/Perl for Unix.
* gnumap - The Genomic Next-generation Universal MAPper (gnumap) is a program designed to accurately map sequence data obtained from next-generation sequencing machines (specifically that of Solexa/Illumina) back to a genome of any size. It seeks to align reads from nonunique repeats using statistics. From authors at Brigham Young University. C source/Unix.
* MAQ - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina with preliminary functions to handle ABI SOLiD data. Written by Heng Li from the Sanger Centre. Features extensive supporting tools for DIP/SNP detection, etc. C++ source
* MOSAIK - MOSAIK produces gapped alignments using the Smith-Waterman algorithm. Features a number of support tools. Support for Roche FLX, Illumina, SOLiD, and Helicos. Written by Michael Strömberg at Boston College. Win/Linux/MacOSX
* MrFAST and MrsFAST - mrFAST & mrsFAST are designed to map short reads generated with the Illumina platform to reference genome assemblies; in a fast and memory-efficient manner. Robust to INDELs and MrsFAST has a bisulphite mode. Authors are from the University of Washington. C as source.
* MUMmer - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficient suffix tree library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. POSIX OS required.
* Novocraft - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Can support Bis-Seq. Commercial. Available free for evaluation, educational use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
* PASS - It supports Illumina, SOLiD and Roche-FLX data formats and allows the user to modulate very finely the sensitivity of the alignments. Spaced seed intial filter, then NW dynamic algorithm to a SW(like) local alignment. Authors are from CRIBI in Italy. Win/Linux.
* RMAP - Assembles 20 - 64 bp Illumina reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinformatics). POSIX OS required.
* SeqMap - Supports up to 5 or more bp mismatches/INDELs. Highly tunable. Written by Hui Jiang from the Wong lab at Stanford. Builds available for most OS's.
* SHRiMP - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Brudno and Stephen Rumble at the University of Toronto. POSIX.
* Slider- An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a reference sequence or a set of reference sequences. Authors are from BCGSC. Paper is here.
* SOAP - SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The updated version uses a BWT. Can call SNPs and INDELs. Author is Ruiqiang Li at the Beijing Genomics Institute. C++, POSIX.
* SSAHA - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases using a hash table. Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
* SOCS - Aligns SOLiD data. SOCS is built on an iterative variation of the Rabin-Karp string search algorithm, which uses hashing to reduce the set of possible matches, drastically increasing search speed. Authors are Ondov B, Varadarajan A, Passalacqua KD and Bergman NH.
* SWIFT - The SWIFT suit is a software collection for fast index-based sequence comparison. It contains: SWIFT — fast local alignment search, guaranteeing to find epsilon-matches between two sequences. SWIFT BALSAM — a very fast program to find semiglobal non-gapped alignments based on k-mer seeds. Authors are Kim Rasmussen (SWIFT) and Wolfgang Gerlach (SWIFT BALSAM)
* SXOligoSearch - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Refseq RNA or NCBI genome builds for a number of organisms. Web Portal. OS independent.
* Vmatch - A versatile software tool for efficiently solving large scale sequence matching tasks. Vmatch subsumes the software tool REPuter, but is much more general, with a very flexible user interface, and improved space and time requirements. Essentially a large string matching toolbox. POSIX.
* Zoom - ZOOM (Zillions Of Oligos Mapped) is designed to map millions of short reads, emerged by next-generation sequencing technology, back to the reference genomes, and carry out post-analysis. ZOOM is developed to be highly accurate, flexible, and user-friendly with speed being a critical priority. Commercial. Supports Illumina and SOLiD data.

Bowtie best for Illumina; **PerM** is best for SOLiD; **BLAT** is good for 454 reads

mapping reads to reference genome



```
bowtie2  /path/index_genome    input.fastq    output.sam
```

Provide the full path for the index

The simplest syntax using default parameters

A on-screen message will be popped out after mapping is done

Total reads

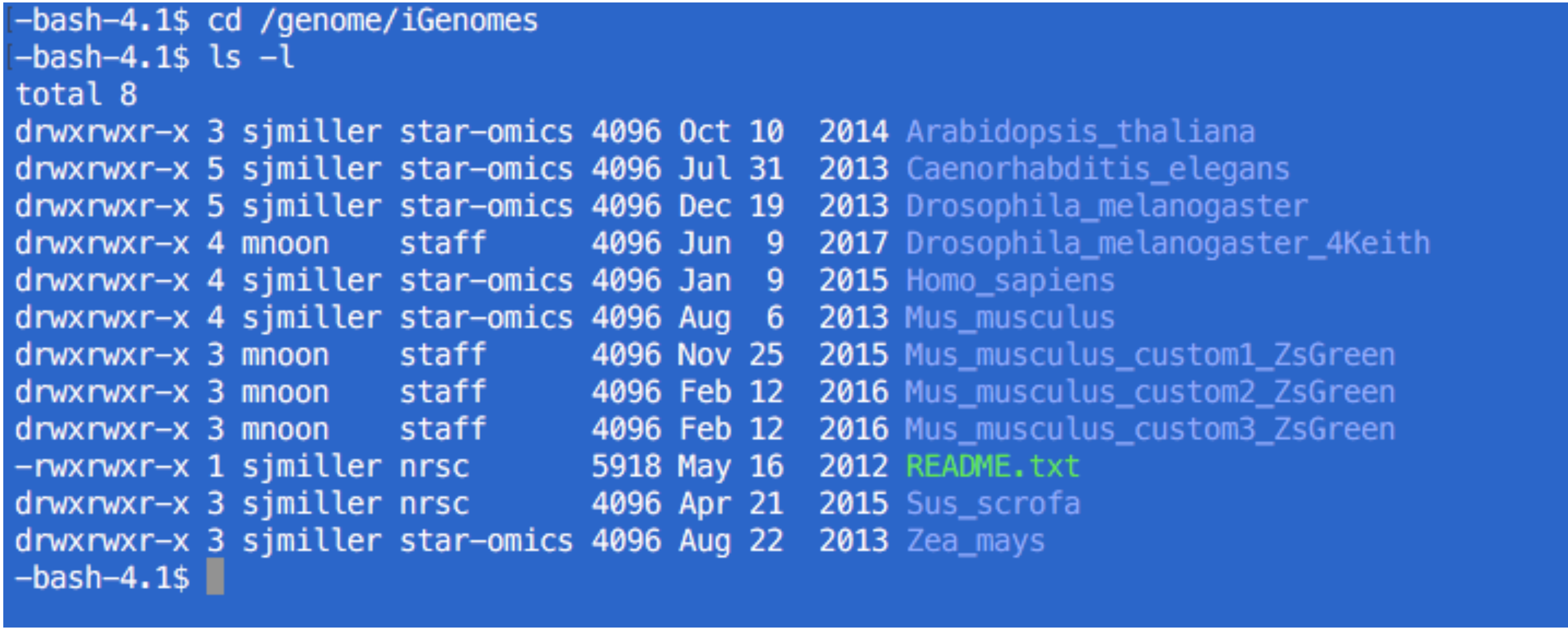
Mapped reads

```
reads processed: 91359186
# reads with at least one reported alignment: 43366753 (47.47%)
# reads that failed to align: 45809573 (50.14%)
# reads with alignments suppressed due to -m: 2182860 (2.39%)
Reported 43366753 alignments to 1 output stream(s)
```

Unmapped reads

Multiply mapped reads

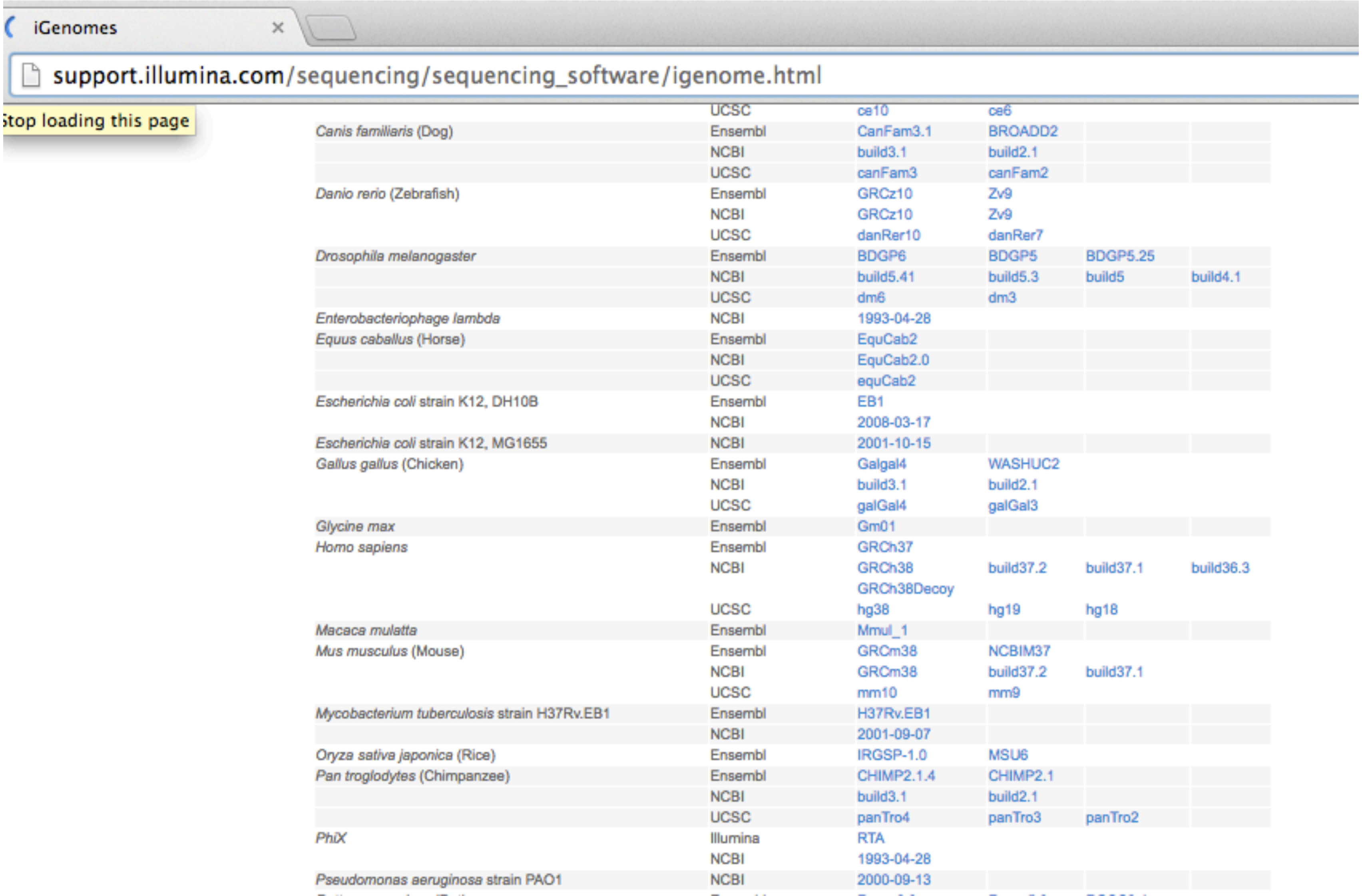
Index files in HPC



```
-bash-4.1$ cd /genome/iGenomes
-bash-4.1$ ls -l
total 8
drwxrwxr-x 3 sjmiller star-omics 4096 Oct 10  2014 Arabidopsis_thaliana
drwxrwxr-x 5 sjmiller star-omics 4096 Jul 31  2013 Caenorhabditis_elegans
drwxrwxr-x 5 sjmiller star-omics 4096 Dec 19  2013 Drosophila_melanogaster
drwxrwxr-x 4 mnoon    staff      4096 Jun  9  2017 Drosophila_melanogaster_4Keith
drwxrwxr-x 4 sjmiller star-omics 4096 Jan  9  2015 Homo_sapiens
drwxrwxr-x 4 sjmiller star-omics 4096 Aug  6  2013 Mus_musculus
drwxrwxr-x 3 mnoon    staff      4096 Nov 25  2015 Mus_musculus_custom1_ZsGreen
drwxrwxr-x 3 mnoon    staff      4096 Feb 12  2016 Mus_musculus_custom2_ZsGreen
drwxrwxr-x 3 mnoon    staff      4096 Feb 12  2016 Mus_musculus_custom3_ZsGreen
-rwxrwxr-x 1 sjmiller nrsc       5918 May 16  2012 README.txt
drwxrwxr-x 3 sjmiller nrsc       4096 Apr 21  2015 Sus_scrofa
drwxrwxr-x 3 sjmiller star-omics 4096 Aug 22  2013 Zea_mays
-bash-4.1$
```

Illumina index files:

- http://support.illumina.com/sequencing/sequencing_software/igenome.html


iGenomes
support.illumina.com/sequencing/sequencing_software/igenome.html
top loading this page

	UCSC	ce10	ce6		
Canis familiaris (Dog)	Ensembl	CanFam3.1	BROADD2		
	NCBI	build3.1	build2.1		
	UCSC	canFam3	canFam2		
Danio rerio (Zebrafish)	Ensembl	GRCz10	Zv9		
	NCBI	GRCz10	Zv9		
	UCSC	danRer10	danRer7		
Drosophila melanogaster	Ensembl	BDGP6	BDGP5	BDGP5.25	
	NCBI	build5.41	build5.3	build5	build4.1
	UCSC	dm6	dm3		
Enterobacteriophage lambda	NCBI	1993-04-28			
Equus caballus (Horse)	Ensembl	EquCab2			
	NCBI	EquCab2.0			
	UCSC	equCab2			
Escherichia coli strain K12, DH10B	Ensembl	EB1			
	NCBI	2008-03-17			
Escherichia coli strain K12, MG1655	NCBI	2001-10-15			
Gallus gallus (Chicken)	Ensembl	Galgal4	WASHUC2		
	NCBI	build3.1	build2.1		
	UCSC	galGal4	galGal3		
Glycine max	Ensembl	Gm01			
Homo sapiens	Ensembl	GRCh37			
	NCBI	GRCh38	build37.2	build37.1	build36.3
		GRCh38Decoy			
	UCSC	hg38	hg19	hg18	
Macaca mulatta	Ensembl	Mmul_1			
Mus musculus (Mouse)	Ensembl	GRCm38	NCBIM37		
	NCBI	GRCm38	build37.2	build37.1	
	UCSC	mm10	mm9		
Mycobacterium tuberculosis strain H37Rv.EB1	Ensembl	H37Rv.EB1			
	NCBI	2001-09-07			
Oryza sativa japonica (Rice)	Ensembl	IRGSP-1.0	MSU6		
Pan troglodytes (Chimpanzee)	Ensembl	CHIMP2.1.4	CHIMP2.1		
	NCBI	build3.1	build2.1		
	UCSC	panTro4	panTro3	panTro2	
PhiX	Illumina	RTA			
	NCBI	1993-04-28			
Pseudomonas aeruginosa strain PAO1	NCBI	2000-09-13			

```
module ava   (get a list of available
              modules in HPC)
module load bowtie2
module load samtools

bowtie2 -x /genome/iGenomes/Homo_sapiens/
Ensembl/GRCh37/Sequence/Bowtie2Index/genome
-U example.fastq -S my1.sam

samtools view -bS my1.sam > my1.bam
```

The file is called pp.pbs

```
#!/bin/csh
#PBS -N bowtie_ex
#PBS -m bea
#PBS -M anling@email.arizona.edu
#PBS -W group_list=anling
#PBS –q standard
#PBS -l select=1:ncpus=28:mem=168gb
#PBS -l cput=56:0:0
#PBS -l walltime=2:0:0

###source /usr/share/modules/init/csh
module load bowtie
module load samtools

cd /extra/anling/check/

bowtie2 -x /genome/iGenomes/Homo_sapiens/Ensembl/GRCh37/Sequence/Bowtie2Index/genome -U example.fastq -S my1.sam

samtools view -bS my1.sam > my1.bam

cd ..
```

More details can be found at:

https://docs.hpc.arizona.edu/display/UAHPC/PBS+Examples+for+Life+Sciences#PBSExamplesforLifeSciences-bowtie2/tophat/cufflinks

```
qsub pp.pbs
```

(submit a job)

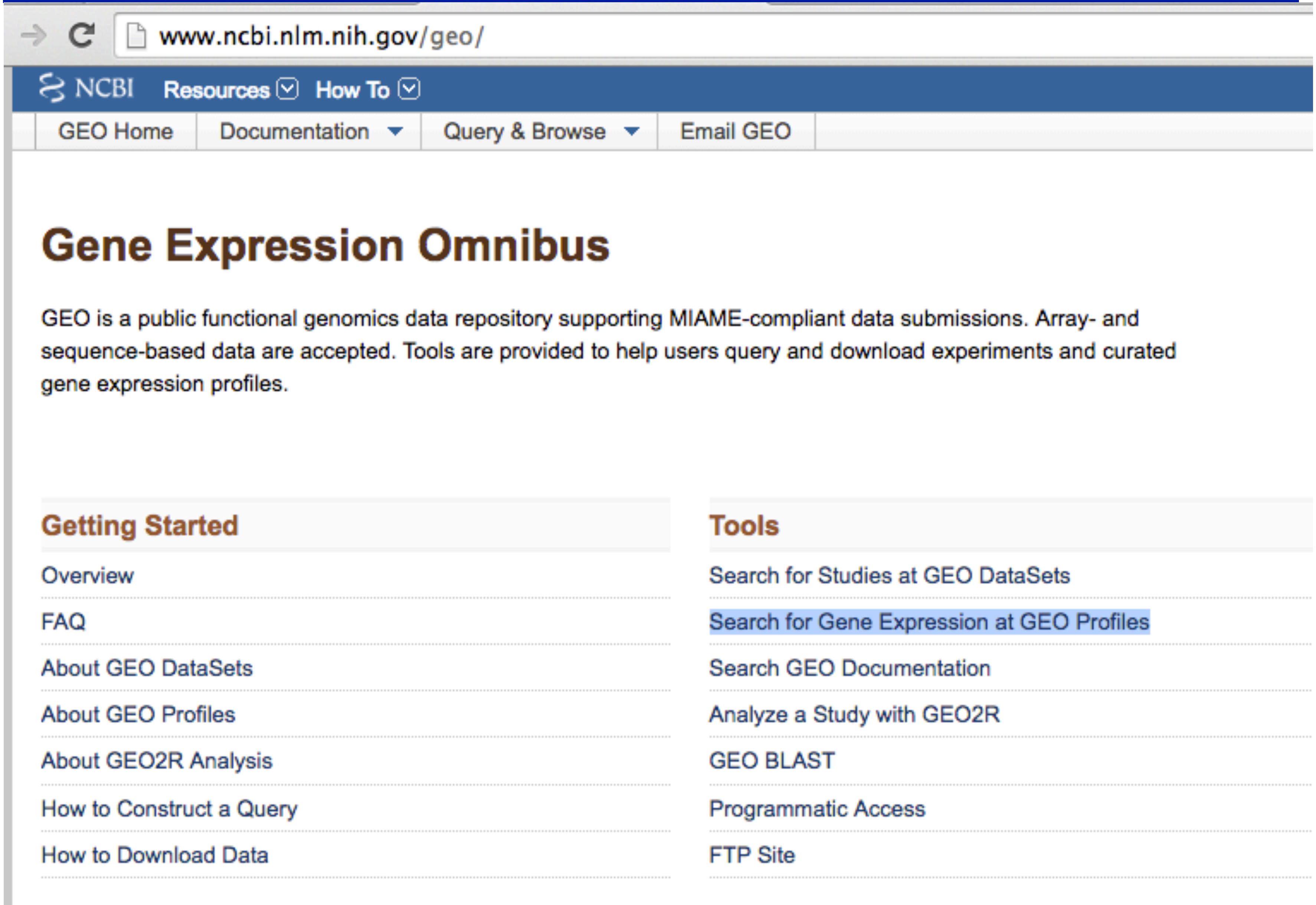
```
qstat -u yourNetId
```

(check the process)

Check the .o file, .e file, and the output result file.

```
ls -l
vi filename
```

How to find a dataset from 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.

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

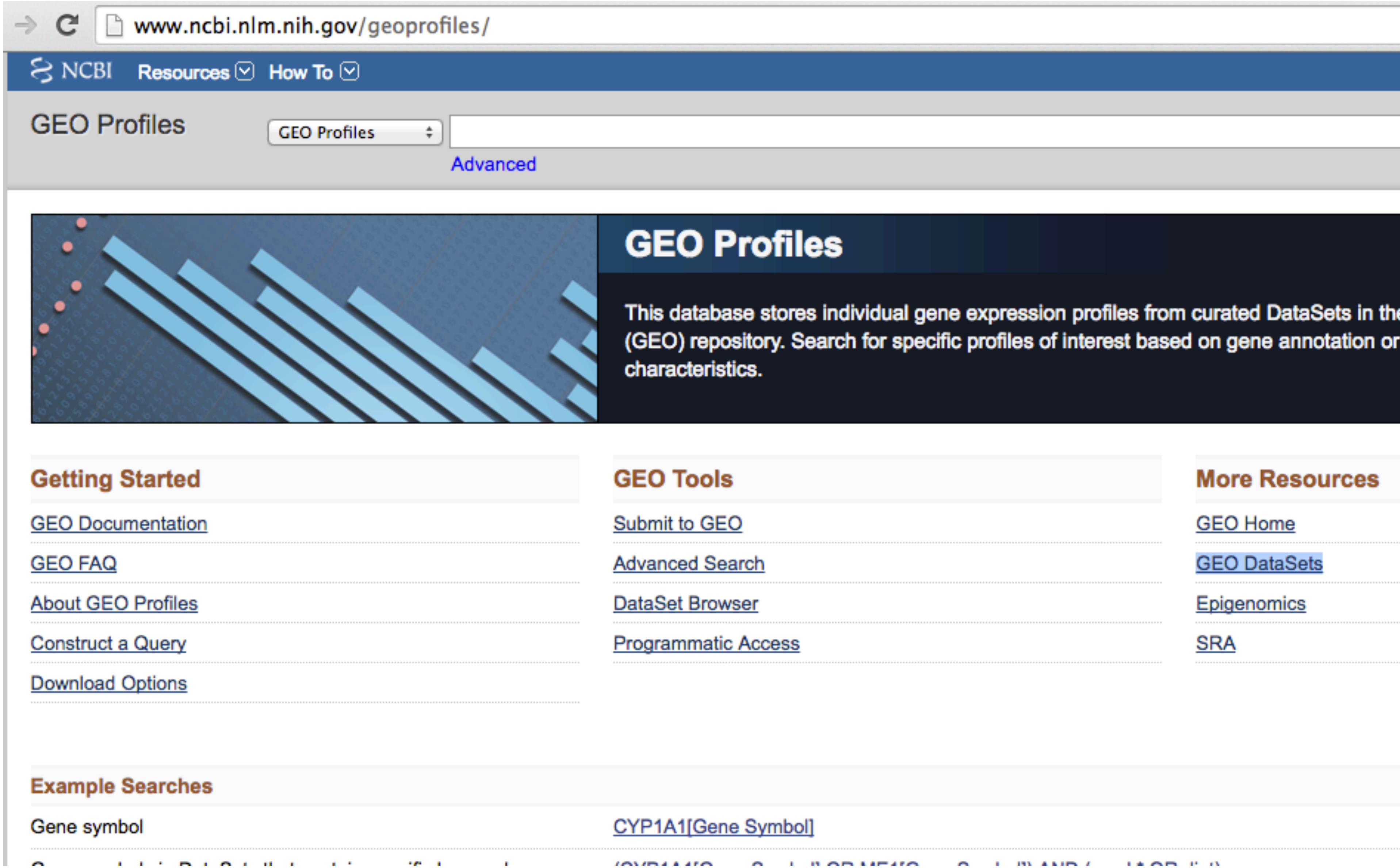
Search GEO Documentation

Analyze a Study with GEO2R

GEO BLAST

Programmatic Access

FTP Site



GEO Profiles

This database stores individual gene expression profiles from curated DataSets in the (GEO) repository. Search for specific profiles of interest based on gene annotation or characteristics.

Getting Started

- GEO Documentation
- GEO FAQ
- About GEO Profiles
- Construct a Query
- Download Options

GEO Tools

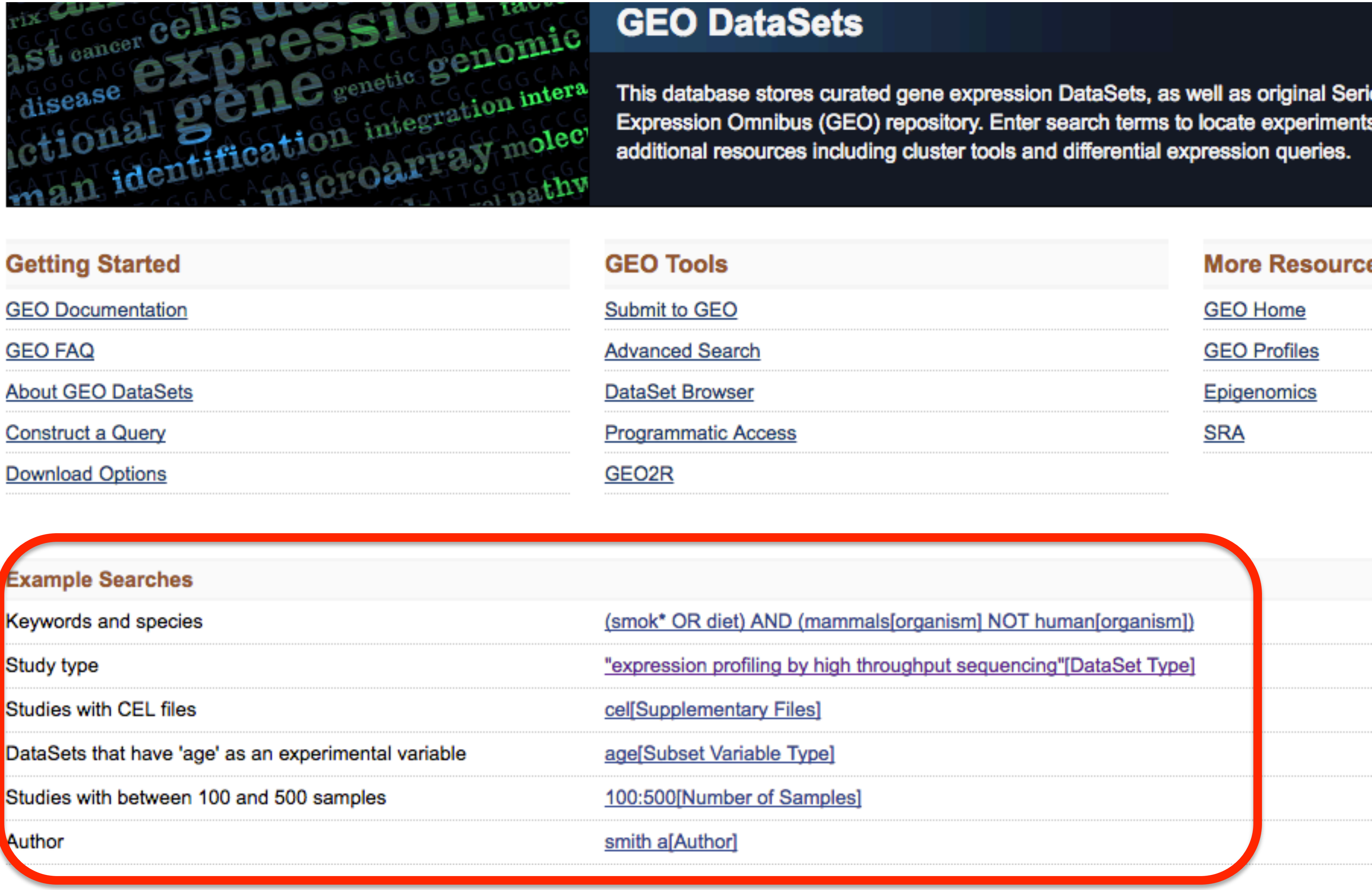
- Submit to GEO
- Advanced Search
- DataSet Browser
- Programmatic Access

More Resources

- GEO Home
- GEO DataSets
- Epigenomics
- SRA

Example Searches

Gene symbol	CYP1A1[Gene Symbol]



GEO DataSets

This database stores curated gene expression DataSets, as well as original Seri
Expression Omnibus (GEO) repository. Enter search terms to locate experiments
additional resources including cluster tools and differential expression queries.

Getting Started

- GEO Documentation
- GEO FAQ
- About GEO DataSets
- Construct a Query
- Download Options

GEO Tools

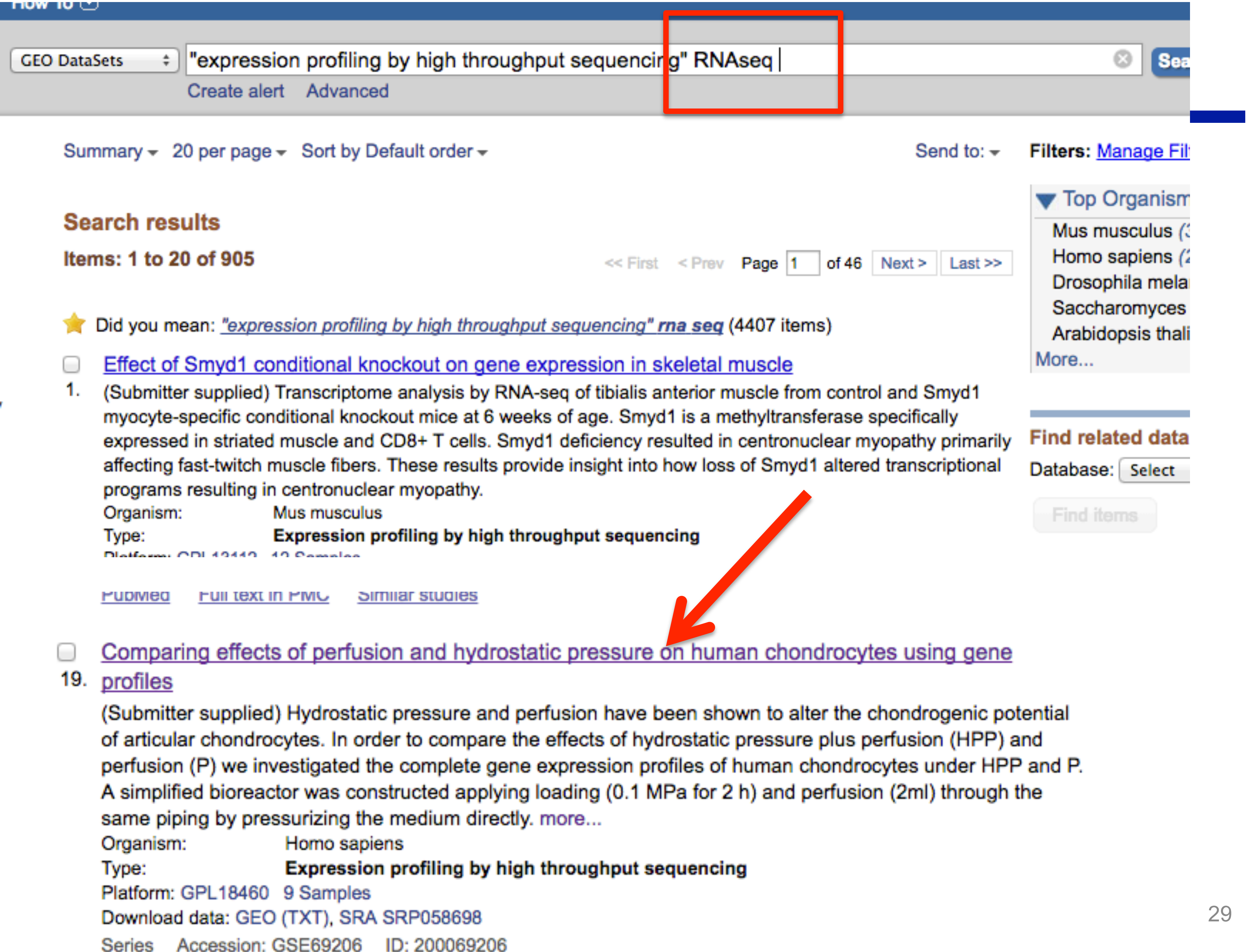
- Submit to GEO
- Advanced Search
- DataSet Browser
- Programmatic Access
- GEO2R

More Resource

- GEO Home
- GEO Profiles
- Epigenomics
- SRA

Example Searches

Keywords and species	(smok* OR diet) AND (mammals[organism] NOT human[organism])
Study type	"expression profiling by high throughput sequencing"[DataSet Type]
Studies with CEL files	cel[Supplementary Files]
DataSets that have 'age' as an experimental variable	age[Subset Variable Type]
Studies with between 100 and 500 samples	100:500[Number of Samples]
Author	smith a[Author]



"expression profiling by high throughput sequencing" RNAseq

Search results

Items: 1 to 20 of 905

Did you mean: *"expression profiling by high throughput sequencing" **rna seq*** (4407 items)

1. Effect of Smyd1 conditional knockout on gene expression in skeletal muscle

(Submitter supplied) Transcriptome analysis by RNA-seq of tibialis anterior muscle from control and Smyd1 myocyte-specific conditional knockout mice at 6 weeks of age. Smyd1 is a methyltransferase specifically expressed in striated muscle and CD8+ T cells. Smyd1 deficiency resulted in centronuclear myopathy primarily affecting fast-twitch muscle fibers. These results provide insight into how loss of Smyd1 altered transcriptional programs resulting in centronuclear myopathy.

Organism: Mus musculus

Type: **Expression profiling by high throughput sequencing**

Platform: GPL13112 12 Samples

PubMed Full text in PMC Similar studies

19. Comparing effects of perfusion and hydrostatic pressure on human chondrocytes using gene profiles

(Submitter supplied) Hydrostatic pressure and perfusion have been shown to alter the chondrogenic potential of articular chondrocytes. In order to compare the effects of hydrostatic pressure plus perfusion (HPP) and perfusion (P) we investigated the complete gene expression profiles of human chondrocytes under HPP and P. A simplified bioreactor was constructed applying loading (0.1 MPa for 2 h) and perfusion (2ml) through the same piping by pressurizing the medium directly. more...

Organism: Homo sapiens

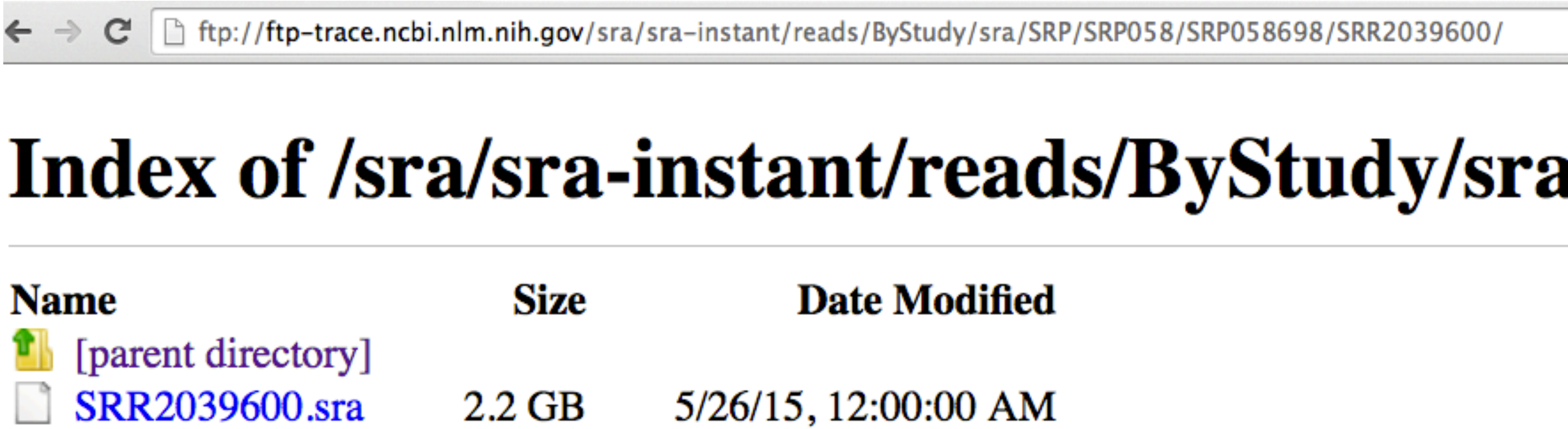
Type: **Expression profiling by high throughput sequencing**

Platform: GPL18460 9 Samples

Download data: GEO (TXT), SRA SRP058698

Series Accession: GSE69206 ID: 200069206

How to download .SRA files to HPC



ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP058/SRP058698/SRR2039600/

Index of /sra/sra-instant/reads/ByStudy/sra

Name	Size	Date Modified
[parent directory]		
SRR2039600.sra	2.2 GB	5/26/15, 12:00:00 AM

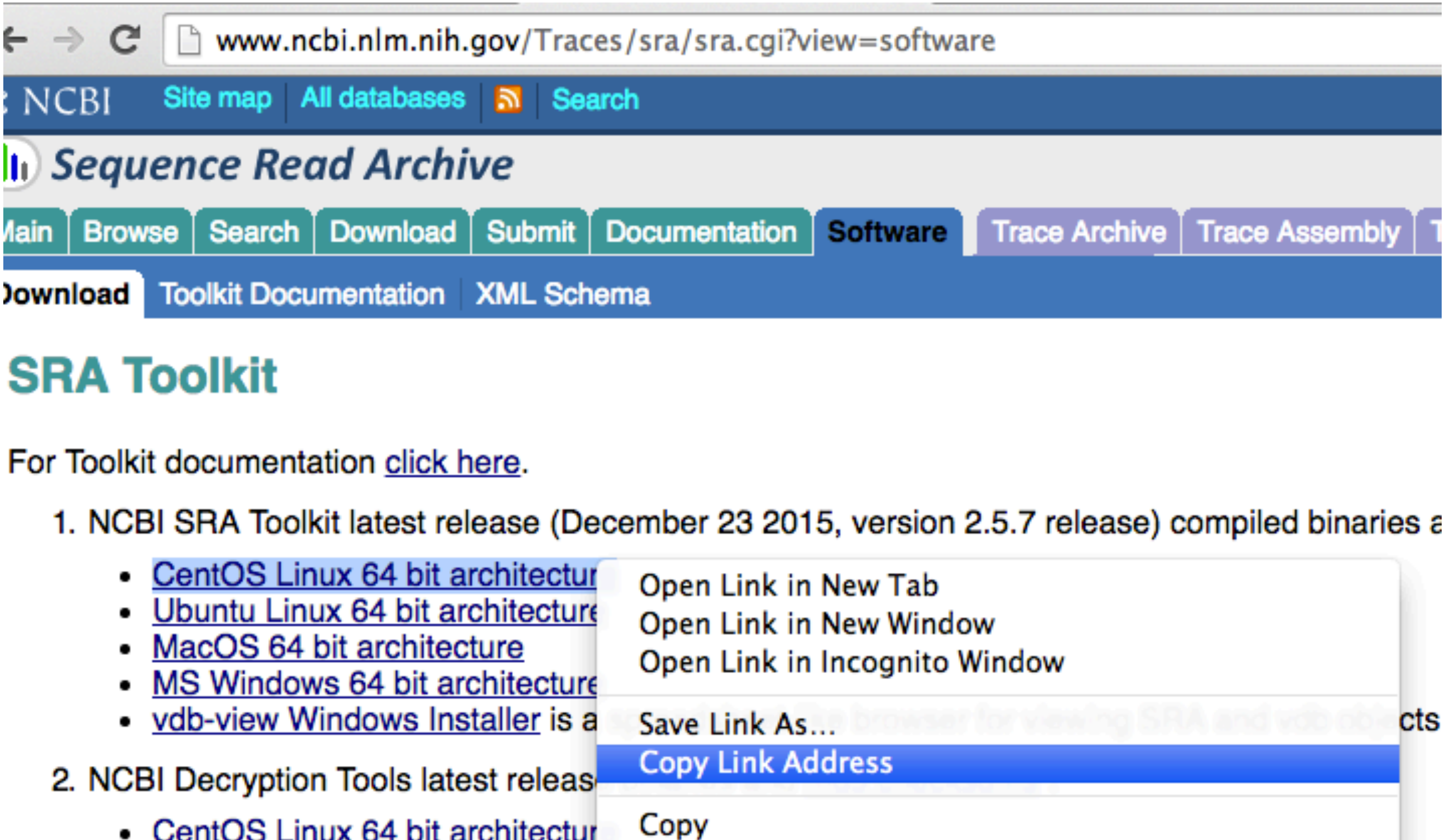
In HPC type:

```
wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP058/SRP058698/SRR2039600/SRR2039600.sra
```

How to convert .SRA to .fastq files

Download SRA Toolkit:

(http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software)



SRA Toolkit

For Toolkit documentation click here.

1. NCBI SRA Toolkit latest release (December 23 2015, version 2.5.7 release) compiled binaries a
   - CentOS Linux 64 bit architectur
   - Ubuntu Linux 64 bit architecture
   - MacOS 64 bit architecture
   - MS Windows 64 bit architecture
   - vdb-view Windows Installer is a cts
2. NCBI Decryption Tools latest releas
   - CentOS Linux 64 bit architectur

In HPC type:

```
wget (paste the link here) <return>
```

And then unzip the .tar.gz file:

```
tar xvzf sratoolkit.2.9.0-centos_linux64.tar.gz
```

Then type the following to get .fastq file:

```
sratoolkit.2.9.0-centos_linux64/bin/fastq-dump
SRR2039600.sra
```

Note: it will take a while …

Take a subset of .fastq file

How many lines in the .fastq file?

```
wc -l SRR2039600.fastq
```

Take a small subset of it:

```
sed -n "1, 100000p" SRR2039600.fastq > mysmall.fastq
```

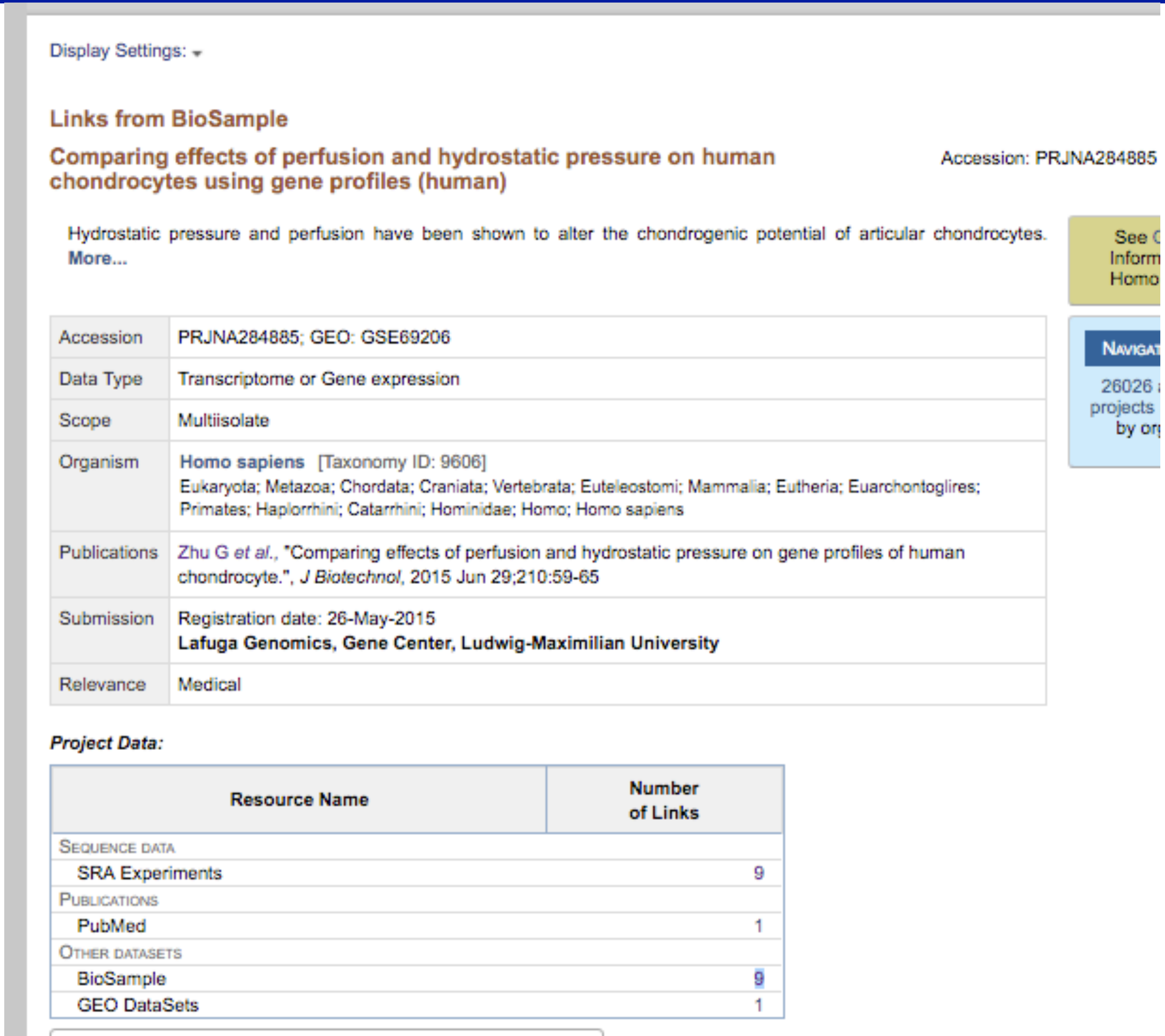
Use the following Bioconductor packages for next class

- source("http://bioconductor.org/biocLite.R")
- biocLite("GenomicRanges")
- biocLite("GenomicFeatures")
- biocLite("Rsamtools")
- biocLite("DESeq")
- biocLite("edgeR") ### you may already have it.
- ## biocLite("org.Mm.eg.db")   ### mouse sequence
- biocLite("org.Hs.eg.db")   ### human sequence
- biocLite("limma") ### you may already have it.
- biocLite("Rsubread")
- biocLite("readGAlignmentsFromBam")
- biocLite("GenomicAlignments")

- library(GenomicFeatures)
- library(GenomicRanges)
- library(Rsamtools)
- library(Rsubread)
- library(limma)
- library(edgeR)
- library(DESeq)
- library(readGAlignmentsFromBam)
- library(GenomicAlignments)

- Download 9 .bam files from D2L

Metadata for the downloaded dataset



Display Settings: ▾

Links from BioSample

Comparing effects of perfusion and hydrostatic pressure on human chondrocytes using gene profiles (human)

Accession: PRJNA284885

Hydrostatic pressure and perfusion have been shown to alter the chondrogenic potential of articular chondrocytes. **More...**

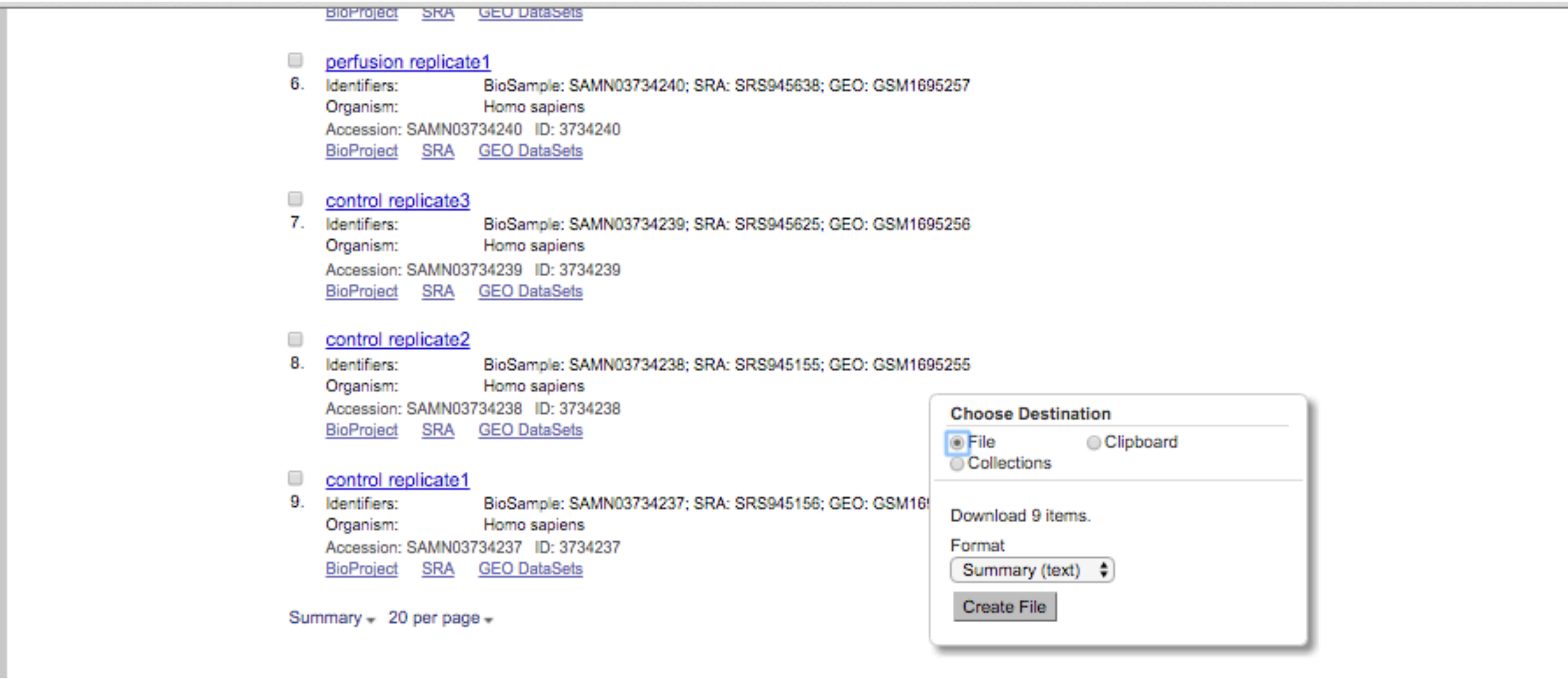
See G
Inform
Homo

NAVIGAT
26026
projects
by or

Accession	PRJNA284885; GEO: GSE69206
Data Type	Transcriptome or Gene expression
Scope	Multiisolate
Organism	**Homo sapiens** [Taxonomy ID: 9606] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens
Publications	Zhu G *et al.*, "Comparing effects of perfusion and hydrostatic pressure on gene profiles of human chondrocyte.", *J Biotechnol*, 2015 Jun 29;210:59-65
Submission	Registration date: 26-May-2015 **Lafuga Genomics, Gene Center, Ludwig-Maximilian University**
Relevance	Medical

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	9
PUBLICATIONS	
PubMed	1
OTHER DATASETS	
BioSample	9
GEO DataSets	1



BioProject SRA GEO DataSets

□ perfusion replicate1

6. Identifiers: BioSample: SAMN03734240; SRA: SRS945638; GEO: GSM1695257
Organism: Homo sapiens
Accession: SAMN03734240 ID: 3734240
BioProject SRA GEO DataSets

□ control replicate3

7. Identifiers: BioSample: SAMN03734239; SRA: SRS945625; GEO: GSM1695256
Organism: Homo sapiens
Accession: SAMN03734239 ID: 3734239
BioProject SRA GEO DataSets

□ control replicate2

8. Identifiers: BioSample: SAMN03734238; SRA: SRS945155; GEO: GSM1695255
Organism: Homo sapiens
Accession: SAMN03734238 ID: 3734238
BioProject SRA GEO DataSets

□ control replicate1

9. Identifiers: BioSample: SAMN03734237; SRA: SRS945156; GEO: GSM16
Organism: Homo sapiens
Accession: SAMN03734237 ID: 3734237
BioProject SRA GEO DataSets

Summary ▾ 20 per page ▾

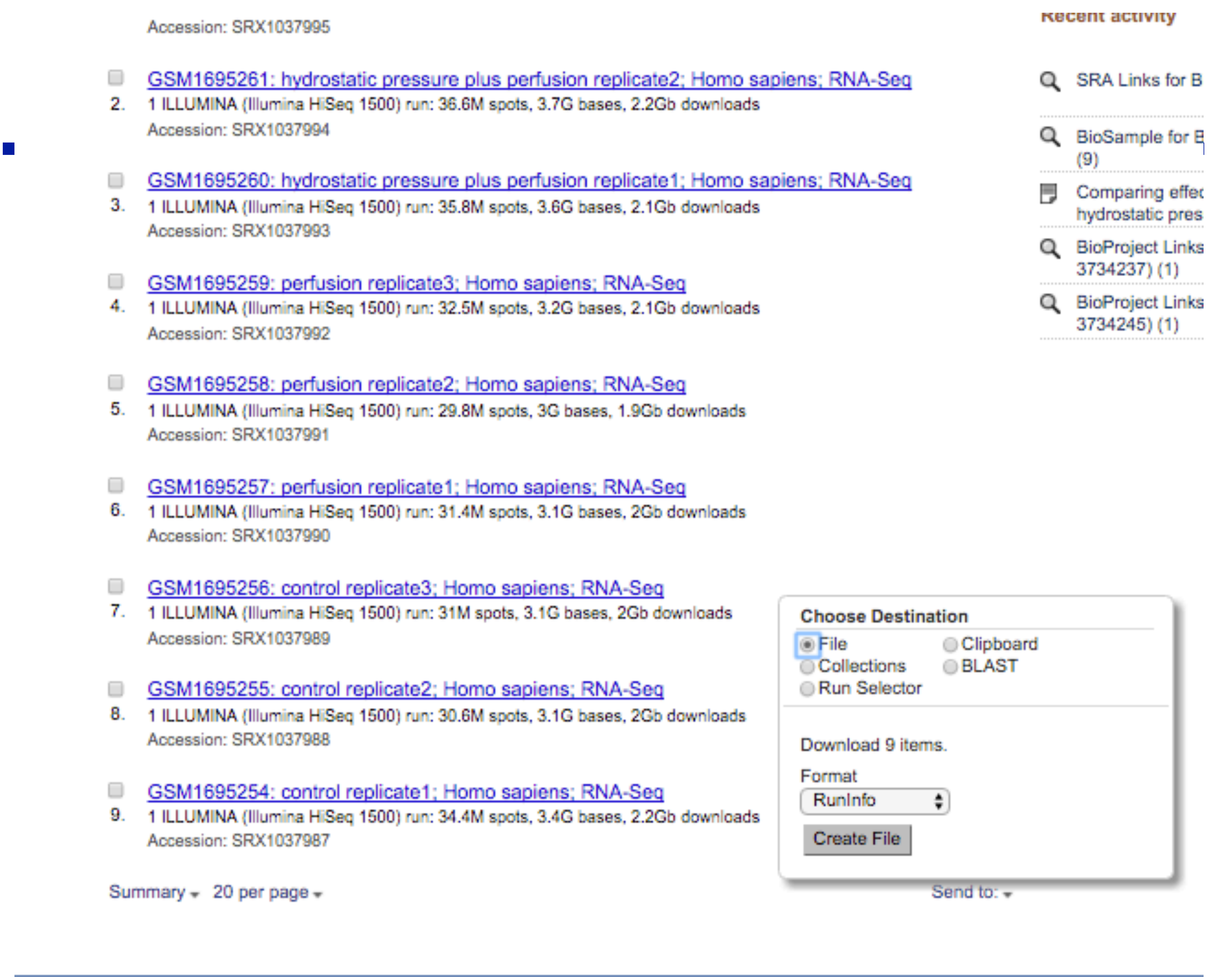
Choose Destination

● File ○ Clipboard
○ Collections

Download 9 items.

Format
Summary (text)

Create File



Accession: SRX1037995

□ 2. GSM1695261: hydrostatic pressure plus perfusion replicate2; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 1500) run: 36.6M spots, 3.7G bases, 2.2Gb downloads
Accession: SRX1037994

□ 3. GSM1695260: hydrostatic pressure plus perfusion replicate1; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 1500) run: 35.8M spots, 3.6G bases, 2.1Gb downloads
Accession: SRX1037993

□ 4. GSM1695259: perfusion replicate3; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 1500) run: 32.5M spots, 3.2G bases, 2.1Gb downloads
Accession: SRX1037992

□ 5. GSM1695258: perfusion replicate2; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 1500) run: 29.8M spots, 3G bases, 1.9Gb downloads
Accession: SRX1037991

□ 6. GSM1695257: perfusion replicate1; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 1500) run: 31.4M spots, 3.1G bases, 2Gb downloads
Accession: SRX1037990

□ 7. GSM1695256: control replicate3; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 1500) run: 31M spots, 3.1G bases, 2Gb downloads
Accession: SRX1037989

□ 8. GSM1695255: control replicate2; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 1500) run: 30.6M spots, 3.1G bases, 2Gb downloads
Accession: SRX1037988

□ 9. GSM1695254: control replicate1; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 1500) run: 34.4M spots, 3.4G bases, 2.2Gb downloads
Accession: SRX1037987

Recent activity

SRA Links for B
BioSample for B (9)
Comparing effec hydrostatic pres
BioProject Links 3734237) (1)
BioProject Links 3734245) (1)

Choose Destination

● File ○ Clipboard
○ Collections ○ BLAST
○ Run Selector

Download 9 items.

Format: RunInfo