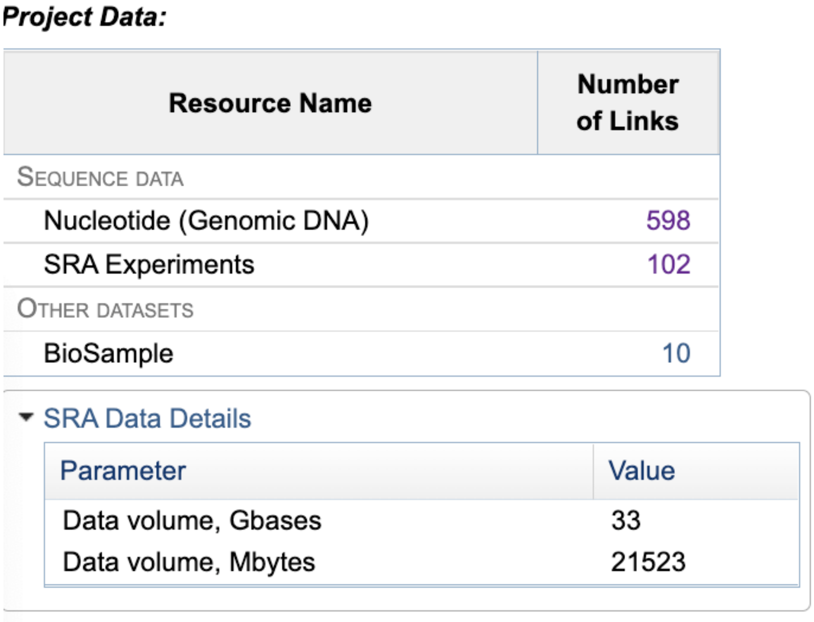
I have picked a paper entitled "Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software". This paper addresses the lack of consensus about assessing the methods for assembly, taxonomic profiling and binning of metagenomic data. The plasmid assemblies, raw data and metadata in this publication have been deposited in the European Nucleotide Archive (ENA) under accession number PRJEB20380. As seen in the figure below, the project summary included 102 experiments that produced 598 nucleotide.



Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (Genomic DNA)	598
SRA Experiments	102
OTHER DATASETS	
BioSample	10

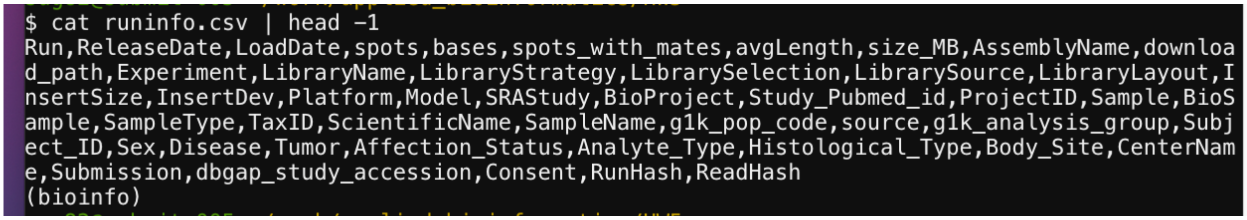
▾ SRA Data Details

Parameter	Value
Data volume, Gbases	33
Data volume, Mbytes	21523

The commands below were run successfully:

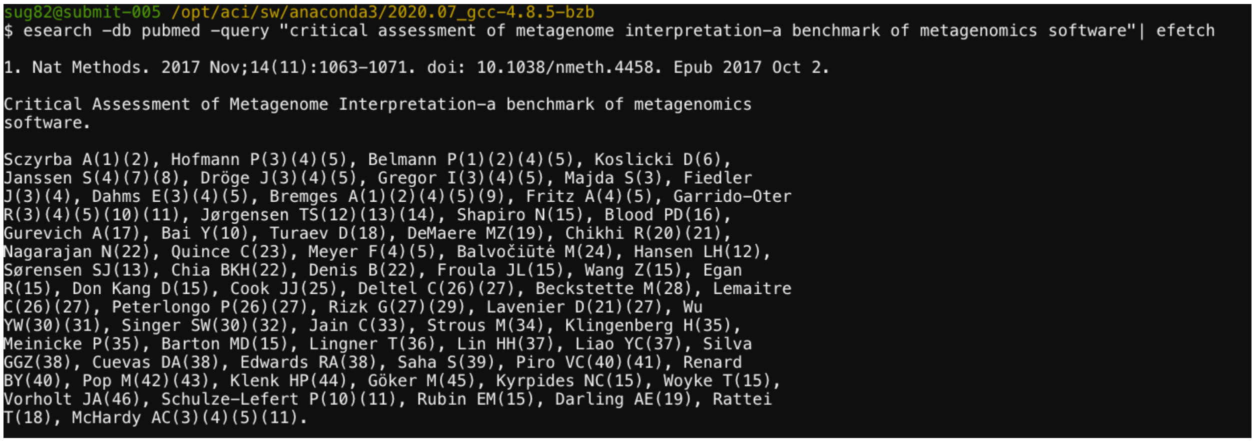
```
# Get the sequencing run information
esearch -db sra -query  PRJEB20380| efetch -format runinfo > runinfo.csv

# Download the assembled genome information
esearch -db nucleotide -query PRJEB20380| efetch -format fasta > genomes.fa
```

The first line of runinfo.csv is observed below:

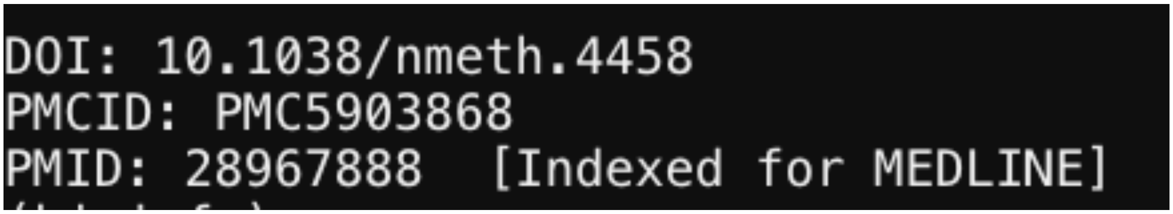


```
$ cat runinfo.csv | head -1
Run,ReleaseDate,LoadDate,spots,bases,spots_with_mates,avgLength,size_MB,AssemblyName,downloa
d_path,Experiment,LibraryName,LibraryStrategy,LibrarySelection,LibrarySource,LibraryLayout,I
nsertSize,InsertDev,Platform,Model,SRAStudy,BioProject,Study_Pubmed_id,ProjectID,Sample,BioS
ample,SampleType,TaxID,ScientificName,SampleName,g1k_pop_code,source,g1k_analysis_group,Subj
ect_ID,Sex,Disease,Tumor,Affection_Status,Analyte_Type,Histological_Type,Body_Site,CenterNam
e,Submission,dbgap_study_accession,Consent,RunHash,ReadHash
(bioinfo)
```

esearch -db pubmed -query "critical assessment of metagenome interpretation-a benchmark of metagenomics software"| efetch

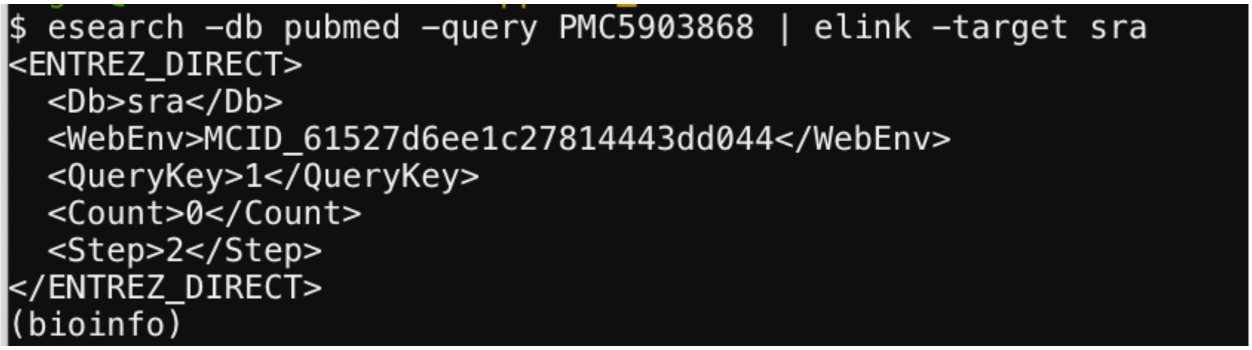


```
sug82@submit-005 /opt/aci/sw/anaconda3/2020.07_gcc-4.8.5-bzb
$ esearch -db pubmed -query "critical assessment of metagenome interpretation—a benchmark of metagenomics software"| efetch

1. Nat Methods. 2017 Nov;14(11):1063-1071. doi: 10.1038/nmeth.4458. Epub 2017 Oct 2.

Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics
software.

Sczyrba A(1)(2), Hofmann P(3)(4)(5), Belmann P(1)(2)(4)(5), Koslicki D(6),
Janssen S(4)(7)(8), Dröge J(3)(4)(5), Gregor I(3)(4)(5), Majda S(3), Fiedler
J(3)(4), Dahms E(3)(4)(5), Bremges A(1)(2)(4)(5)(9), Fritz A(4)(5), Garrido-Oter
R(3)(4)(5)(10)(11), Jørgensen TS(12)(13)(14), Shapiro N(15), Blood PD(16),
Gurevich A(17), Bai Y(10), Turaev D(18), DeMaere MZ(19), Chikhi R(20)(21),
Nagarajan N(22), Quince C(23), Meyer F(4)(5), Balvočiūtė M(24), Hansen LH(12),
Sørensen SJ(13), Chia BKH(22), Denis B(22), Froula JL(15), Wang Z(15), Egan
R(15), Don Kang D(15), Cook JJ(25), Deltel C(26)(27), Beckstette M(28), Lemaitre
C(26)(27), Peterlongo P(26)(27), Rizk G(27)(29), Lavenier D(21)(27), Wu
YW(30)(31), Singer SW(30)(32), Jain C(33), Strous M(34), Klingenberg H(35),
Meinicke P(35), Barton MD(15), Lingner T(36), Lin HH(37), Liao YC(37), Silva
GGZ(38), Cuevas DA(38), Edwards RA(38), Saha S(39), Piro VC(40)(41), Renard
BY(40), Pop M(42)(43), Klenk HP(44), Göker M(45), Kyrpides NC(15), Woyke T(15),
Vorholt JA(46), Schulze-Lefert P(10)(11), Rubin EM(15), Darling AE(19), Rattei
T(18), McHardy AC(3)(4)(5)(11).
```



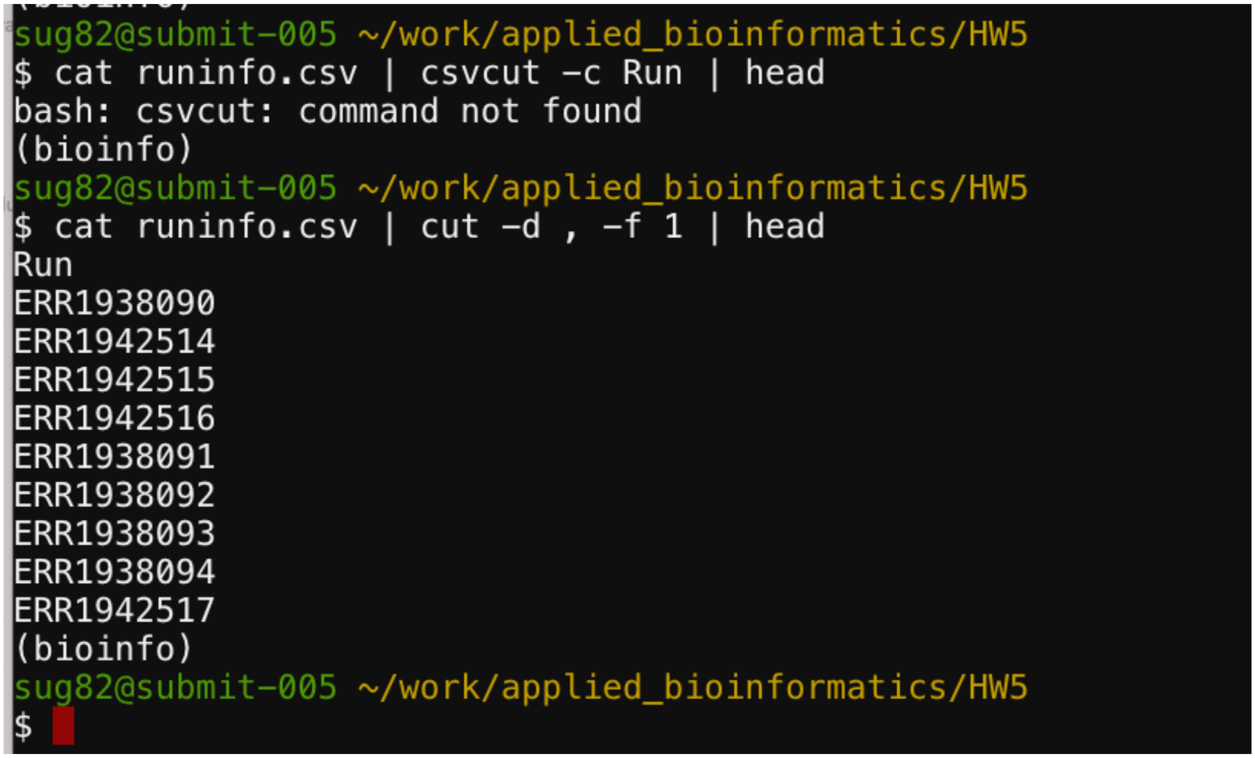
```
DOI: 10.1038/nmeth.4458
PMCID: PMC5903868
PMID: 28967888  [Indexed for MEDLINE]
```

Running `esearch -db pubmed -query PMC5903868 | elink -target sra` did not give us any meaningful results.



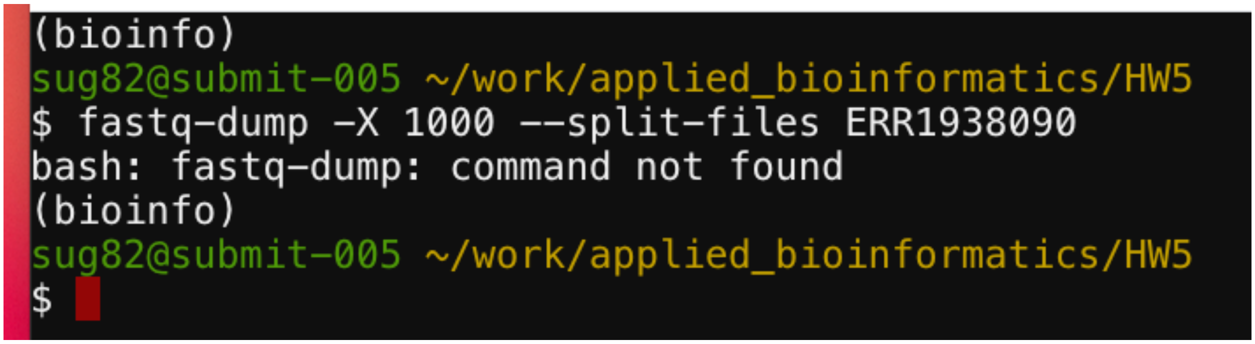
```
$ esearch -db pubmed -query PMC5903868 | elink -target sra
<ENTREZ_DIRECT>
  <Db>sra</Db>
  <WebEnv>MCID_61527d6ee1c27814443dd044</WebEnv>
  <QueryKey>1</QueryKey>
  <Count>0</Count>
  <Step>2</Step>
</ENTREZ_DIRECT>
(bioinfo)
```

As you see in the screenshot below, **csvcut** is not found. I replaced it with **cut -d , -f 1.**



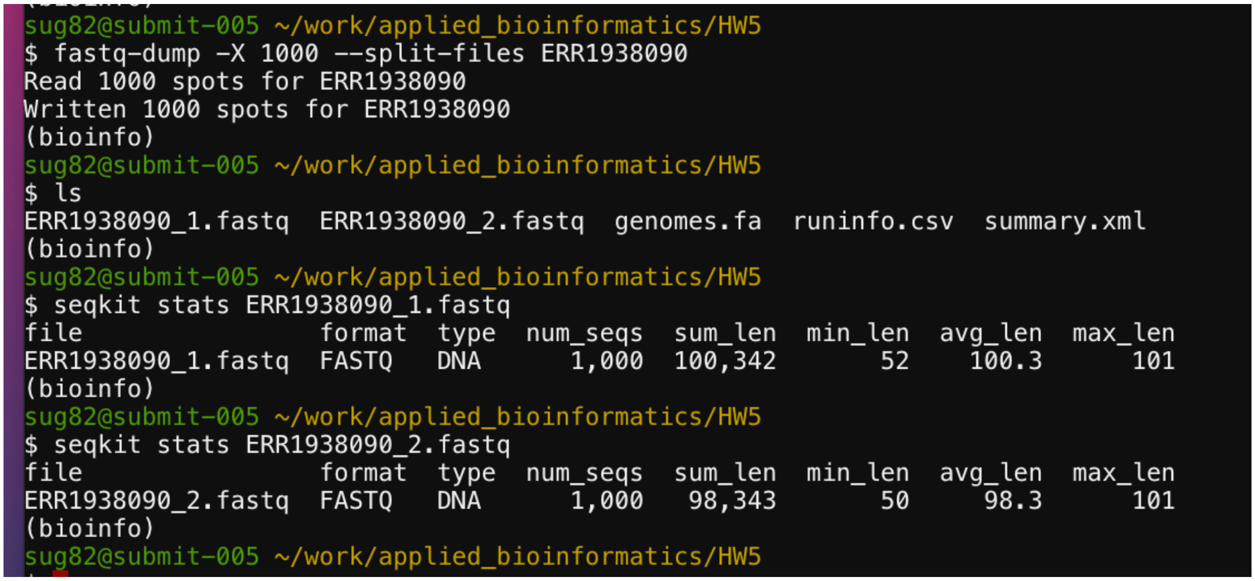
```
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$ cat runinfo.csv | csvcut -c Run | head
bash: csvcut: command not found
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$ cat runinfo.csv | cut -d , -f 1 | head
Run
ERR1938090
ERR1942514
ERR1942515
ERR1942516
ERR1938091
ERR1938092
ERR1938093
ERR1938094
ERR1942517
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$ 
```

I faced the error stating that fastq-dump command is not found. I solved this issue by running command **mamba install sra-tools==2.10.1**



```
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$ fastq-dump -X 1000 --split-files ERR1938090
bash: fastq-dump: command not found
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$ 
```

Finally, I ran **seqkit stats** on the fast file of one particular sequencing as shown below:



```
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$ fastq-dump -X 1000 --split-files ERR1938090
Read 1000 spots for ERR1938090
Written 1000 spots for ERR1938090
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$ ls
ERR1938090_1.fastq  ERR1938090_2.fastq  genomes.fa  runinfo.csv  summary.xml
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$ seqkit stats ERR1938090_1.fastq
file                format  type  num_seqs  sum_len  min_len  avg_len  max_len
ERR1938090_1.fastq  FASTQ   DNA      1,000  100,342       52    100.3      101
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$ seqkit stats ERR1938090_2.fastq
file                format  type  num_seqs  sum_len  min_len  avg_len  max_len
ERR1938090_2.fastq  FASTQ   DNA      1,000   98,343       50     98.3      101
(bioinfo)
sug82@submit-005 ~/work/applied_bioinformatics/HW5
$
```