

# Python Workshop #2



# Recap

- launching the python interpreter
  - use the interactive interpreter to help you!
- datatypes:
  - int: 1234
  - float: 1.23e6
  - string: "Hello World!"
- string formatting:
  - 'Your score is: {}'.format(123)

# Today: Scripting!

- Two steps:
  - write the script in a **suitable editor**
  - run the script from command line
- Text Editor / Development Environment:
  - not a word processor!
  - syntax highlighting, code indentation
  - Windows: Notepad++
  - Mac: gedit
  - Linux: geany, gedit, kate, nano, vim, emacs
  - Eclipse + PyDev
- indentation: 4 spaces per tab!
- line endings!

# Installing Software

- CAUTION: it's easy to install malware
  - sourceforge Filezilla example
- Linux: managed repositories
- Mac: Homebrew

# Scripting Hello World

- Step one:
  - `print "Hello World!"`
- Step two:
  - `python hello_world.py`
- Linux/Mac alternative:
  - `#!/usr/bin/python`
  - `chmod u+x hello_world_alternative.py`

# Example Problem: Reikou



## Search NCBI databases

PRJDB1445

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	74002
WGS master	6
SRA Experiments	20
PUBLICATIONS	
PubMed	1
PMC	1
OTHER DATASETS	
BioSample	10
Assembly	6

Send to: ☒ Filters: [Manage](#)

### Choose Destination

- ☒ File
 ☐ Clipboard  
☐ Collections
 ☐ BLAST  
☐ Run Selector

Download 20 items.

Format

Accession List

Create File

1 DRR013866  
 2 DRR013867  
 3 DRR013868  
 4 DRR013869  
 5 DRR013870  
 6 DRR013871  
 7 DRR013873  
 8 DRR013874  
 9 DRR013875  
 10 DRR013876  
 11 DRR013877  
 12 DRR013878  
 13 DRR013879  
 14 DRR013880  
 15 DRR013881  
 16 DRR013882  
 17 DRR013883  
 18 DRR013884  
 19 DRR013885  
 20 DRR013886

Search

### Spot descriptor:

1 forward 102 reverse

Total: 1 run, 106.9M spots, 21.6G bases, [11.7Gb](#)

#	Run	# of Spots	# of Bases	Size	Published
1.	<a href="#">DRR013876</a>	106,874,998	21.6G	<a href="#">11.7Gb</a>	2014-04-02

ID: 698914

Index of <ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013876/>

[Up to higher level directory](#)

Name

[DRR013876.sra](#)

Size

12281612 KB

Last Modified

27/03/14 00:00:00

Display Settings: ☐ Summary, 20 per page

Results: 20

- Whole-genome sequencing data for *Fraxinus x ananassa*  
1. ILLUMINA (Barrera-HSeq 3000) run: 106.9M spots, 21.6G bases, 11.7Gb download  
Accession: DRR013876
- Whole-genome sequencing data for *Fraxinus x ananassa*  
2. ILLUMINA (Barrera-HSeq 3000) run: 122.8M spots, 24.9G bases, 12.3Gb download  
Accession: DRR013877
- Whole-genome sequencing data for *Fraxinus x ananassa*  
3. ILLUMINA (Barrera-HSeq 3000) run: 125.4M spots, 106.1G bases, 85G download  
Accession: DRR013878
- Whole-genome sequencing data for *Fraxinus x ananassa*  
4. ILS4H (4H GS FLX+) run: 917.328 spots, 285.7M bases, 948.7Mb download  
Accession: DRR013879
- Whole-genome sequencing data for *Fraxinus x ananassa*  
5. ILLUMINA (Barrera-HSeq 3000) run: 538.4M spots, 109G bases, 70.1Gb download  
Accession: DRR013880
- Whole-genome sequencing data for *Fraxinus x ananassa*  
6. ILLUMINA (Barrera-HSeq 3000) run: 232.7M spots, 22.8G bases, 14.2Gb download  
Accession: DRR013881
- Whole-genome sequencing data for *Fraxinus x ananassa*  
7. ILS4H (4H GS FLX+) run: 975.408 spots, 517.5M bases, 1.4Gb download  
Accession: DRR013882
- Whole-genome sequencing data for *Fraxinus x ananassa*  
8. ILLUMINA (Barrera-HSeq 3000) run: 125.7M spots, 32G bases, 18.1Gb download  
Accession: DRR013883
- Whole-genome sequencing data for *Fraxinus x ananassa*  
9. ILS4H (4H GS FLX+) run: 1.3M spots, 1.1G bases, 2.8Gb download  
Accession: DRR013884
- Whole-genome sequencing data for *Fraxinus x ananassa*  
10. ILLUMINA (Barrera-HSeq 3000) run: 537.7M spots, 109.5G bases, 69.5Gb download  
Accession: DRR013885
- Whole-genome sequencing data for *Fraxinus x ananassa*  
11. ILS4H (4H GS FLX+) run: 975.282 spots, 383.3M bases, 1.1Gb download  
Accession: DRR013886
- Whole-genome sequencing data for *Fraxinus x ananassa*  
12. ILS4H (4H GS FLX+) run: 1.3M spots, 1.3G bases, 2.7Gb download  
Accession: DRR013887
- Whole-genome sequencing data for *Fraxinus x ananassa*  
13. ILLUMINA (Barrera-HSeq 3000) run: 537.9M spots, 108.7G bases, 71.1Gb download  
Accession: DRR013888
- Whole-genome sequencing data for *Fraxinus x ananassa*  
14. ILLUMINA (Barrera-HSeq 3000) run: 177.5M spots, 25.9G bases, 22.1Gb download  
Accession: DRR013889
- Whole-genome sequencing data for *Fraxinus x ananassa*  
15. ILLUMINA (Barrera-HSeq 3000) run: 146.1M spots, 29.9G bases, 12.1Gb download  
Accession: DRR013890
- Whole-genome sequencing data for *Fraxinus x ananassa*  
16. ILLUMINA (Barrera-HSeq 3000) run: 267.1M spots, 54G bases, 31.2Gb download  
Accession: DRR013891
- Whole-genome sequencing data for *Fraxinus x ananassa*  
17. ILLUMINA (Barrera-HSeq 3000) run: 145.8M spots, 28.7G bases, 17.8Gb download  
Accession: DRR013892
- Whole-genome sequencing data for *Fraxinus x ananassa*  
18. ILS4H (4H GS FLX+) run: 798.938 spots, 342.1M bases, 3G download  
Accession: DRR013893
- Whole-genome sequencing data for *Fraxinus x ananassa*  
19. ILLUMINA (Barrera-HSeq 3000) run: 245.1M spots, 25.9G bases, 20.6Gb download  
Accession: DRR013894
- Whole-genome sequencing data for *Fraxinus x ananassa*  
20. ILLUMINA (Barrera-HSeq 3000) run: 154.8M spots, 21.3G bases, 14.4Gb download  
Accession: DRR013895

Display Settings: ☐ Summary, 20 per page

wget <http://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013876/DRR013876.sra>

# Example Problem: Reikou

Send to: ☒ Filters: [Manage F](#)

**Choose Destination**

☒ File ☐ Clipboard

☐ Collections ☐ BLAST

☐ Run Selector

Download 20 items.

Format

RunInfo

Create File

	A	B	C	D	E	F	G	H	I	J		K	L	M	N	O	P	Q	R	S	T
	Run	ReleaseDate	LoadDate	spots	bases	spots_with_mates	avgLength	size_MB	AssemblyName	download_path		Experiment	LibraryName	LibraryStrategy	LibrarySelection	LibrarySource	LibraryLayout	InsertSize	InsertDev	Platform	Model
1	DRR013884	Apr 02, 2014	Mar 27, 2014	539412216	108961267632	0	202	71743		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013884/DRR013884/DRX012432.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
2	DRR013886	Apr 02, 2014	Mar 27, 2014	525407773	106132370146	0	202	67582		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013886/DRR013886/DRX012434.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
3	DRR013882	Apr 02, 2014	Mar 27, 2014	153682508	31043866616	0	202	19587		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013882/DRR013882/DRX012430.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
4	DRR013885	Apr 02, 2014	Mar 27, 2014	537920510	108659943020	0	202	72818		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013885/DRR013885/DRX012433.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
5	DRR013878	Apr 02, 2014	Mar 27, 2014	141836049	2865081898	141836049	202	18177		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013878/DRR013878/DRX012426.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
6	DRR013890	Apr 02, 2014	Mar 27, 2014	177512764	35857578328	0	202	22587		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013890/DRR013890/DRX012428.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
7	DRR013866	Apr 02, 2014	Mar 27, 2014	1456315	1121627892	0	770	2843		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013866/DRR013866/DRX012414.101PE		WGS	RANDOM	GENOMIC	SINGLE	0	0	LS454	454 GS FLX+		
8	DRR013867	Apr 02, 2014	Mar 27, 2014	1122425	1266346021	0	1128	2733		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013867/DRR013867/DRX012415.101PE		WGS	RANDOM	GENOMIC	SINGLE	0	0	LS454	454 GS FLX+		
9	DRR013868	Apr 02, 2014	Mar 27, 2014	768919	342097262	298149	444	1059		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013868/DRR013868/DRX012416.3kbPE		WGS	RANDOM	GENOMIC	PAIRED	0	0	LS454	454 GS FLX+		
10	DRR013869	Apr 02, 2014	Mar 27, 2014	973292	363328471	397872	373	1126		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013869/DRR013869/DRX012417.5kbPE		WGS	RANDOM	GENOMIC	PAIRED	0	0	LS454	454 GS FLX+		
11	DRR013870	Apr 02, 2014	Mar 27, 2014	917338	380733769	322856	415	949		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013870/DRR013870/DRX012418.8kbPE		WGS	RANDOM	GENOMIC	PAIRED	0	0	LS454	454 GS FLX+		
12	DRR013871	Apr 02, 2014	Mar 27, 2014	975406	517506128	365739	530	1474		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013871/DRR013871/DRX012419.20kbPE		WGS	RANDOM	GENOMIC	PAIRED	0	0	LS454	454 GS FLX+		
13	DRR013873	Apr 02, 2014	Mar 27, 2014	245059949	35778752554	0	146	21044		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013873/DRR013873/DRX012421.73PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
14	DRR013874	Apr 02, 2014	Mar 27, 2014	267097179	53953630158	0	202	33964		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013874/DRR013874/DRX012422.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
15	DRR013875	Apr 02, 2014	Mar 27, 2014	231661590	23629482180	231661590	102	14588		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013875/DRR013875/DRX012423.51PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
16	DRR013876	Apr 02, 2014	Mar 27, 2014	106874998	21588749596	106874998	202	11993		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013876/DRR013876/DRX012424.2kbMP		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
17	DRR013877	Apr 02, 2014	Mar 27, 2014	154774733	31264496066	0	202	19896		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013877/DRR013877/DRX012425.101OF		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
18	DRR013881	Apr 02, 2014	Mar 27, 2014	146076200	29507392400	146076200	202	18652		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013881/DRR013881/DRX012429.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
19	DRR013879	Apr 02, 2014	Mar 27, 2014	121636111	24570494422	121636111	202	15632		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013879/DRR013879/DRX012427.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		
20	DRR013883	Apr 02, 2014	Mar 27, 2014	537176805	108509714610	0	202	71169		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013883/DRR013883/DRX012431.101PE		WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA/HiSeq	1		

# Our First Script

```
wget_sra_files.py ✕
1#!/usr/bin/python
2# -*- coding: utf-8 -*-
3
4'''
5download all the Reikou SRA files
6
7PRJDB1445
8http://www.ncbi.nlm.nih.gov/bioproject/268159
9http://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=268159
10sendtofile => accession list => SraAccList.txt
11example HTTP download URL
12wget http://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013876/DRR013876.sra
13'''
14
15import os
16
17#file listing SRA ids saved from ncbi through browser
18inp = 'SraAccList.txt'
19
20f = open(inp)
21for line in f:
22    uid = line.strip()
23    #print uid
24
25    id_list = [uid[:3], uid[:6], uid, uid + '.sra']
26
27    cmd = 'wget http://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/'
28    cmd += '/'.join(id_list)
29
30    print cmd
31    os.system(cmd)
32f.close()
33
```



# Line by Line

- triple quote string / comment
- module import
- `inp = 'filename'` assignment
- `f = open(inp)` open file / `f.close()` close file
- for line in f:
- indentation
- `uid = line.strip()`
- `id_list = [...]`
- `cmd += '/'.join(...)`
- `os.system()`

# For loop

```
f = open('SraAccList.txt')  
for line in f:  
    print line  
f.close()
```