Python Workshop #2

Recap

- launching the python interpretter
  - 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 Search

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

```
 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
```

Send to: Filters: Manage

Choose Destination

File Clipboard
Collections BLAST
Run Selector

Download 20 items.

Format

Accession List

Create File

Spot descriptor:

1 forward 102 reverse

Total: 1 run, 106.9M spots, 21.6G bases, 11.7Gb

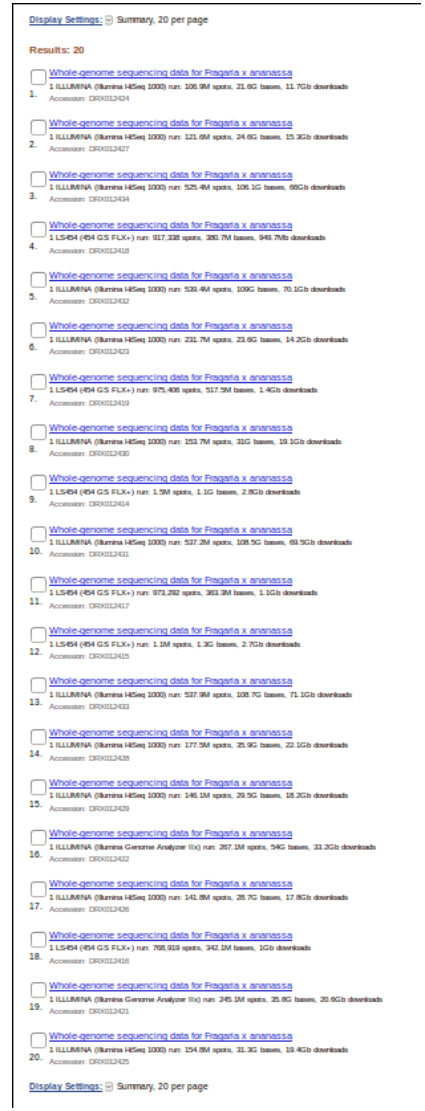
#	Run	# of Spots	# of Bases	Size	Published
1.	DRR013876	106,874,998	21.6G	11.7Gb	2014-04-02

ID: 698914

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

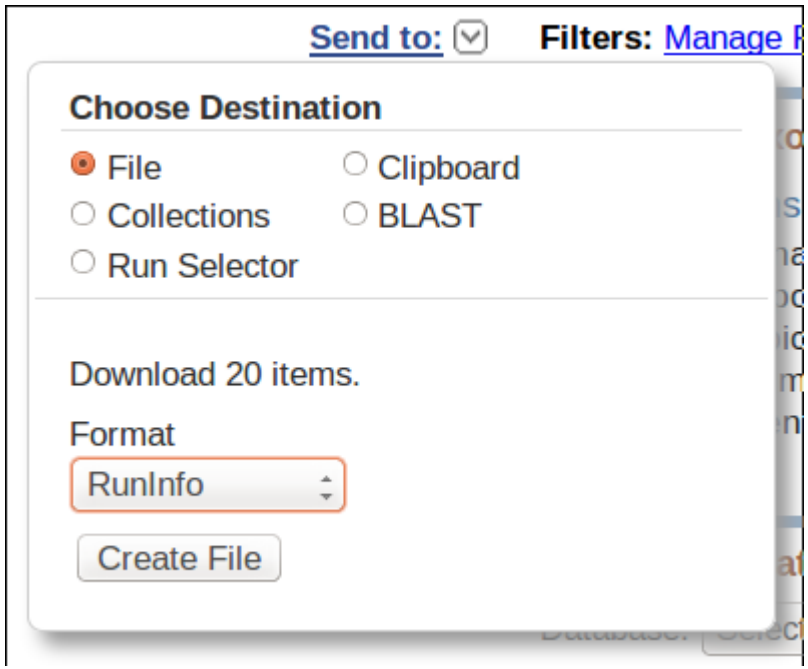
Up to higher level directory

Name	Size	Last Modified
DRR013876.sra	12281612 KB	27/03/14 00:00:00



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

Example Problem: Reikou



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Run	ReleaseDate	LoadDate	spots	bases	spots_with_mates	avgLength	size_MB	AssemblyName	download_path	Experiment	LibraryName	LibraryStrategy	LibrarySelection	LibrarySource	LibraryLayout	InsertSize	InsertDev	Platform	Model
2	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/DRR0138	DRX012432	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
3	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/DRR0138	DRX012434	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
4	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/DRR0138	DRX012430	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
5	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/DRR0138	DRX012433	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
6	DRR013878	Apr 02, 2014	Mar 27, 2014	141836049	28650881898	141836049	202	18177		ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013878/DRR0138	DRX012426	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
7	DRR013880	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/DRR013880/DRR0138	DRX012428	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
8	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/DRR0138	DRX012414	SE	WGS	RANDOM	GENOMIC	SINGLE	0	0	LS454	454 GS FLX+
9	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/DRR0138	DRX012415	SE	WGS	RANDOM	GENOMIC	SINGLE	0	0	LS454	454 GS FLX+
10	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/DRR0138	DRX012416	3kbPE	WGS	RANDOM	GENOMIC	PAIRED	0	0	LS454	454 GS FLX+
11	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/DRR0138	DRX012417	5kbPE	WGS	RANDOM	GENOMIC	PAIRED	0	0	LS454	454 GS FLX+
12	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/DRR0138	DRX012418	8kbPE	WGS	RANDOM	GENOMIC	PAIRED	0	0	LS454	454 GS FLX+
13	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/DRR0138	DRX012419	20kbPE	WGS	RANDOM	GENOMIC	PAIRED	0	0	LS454	454 GS FLX+
14	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/DRR0138	DRX012421	73PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina Genome
15	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/DRR0138	DRX012422	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina Genome
16	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/DRR0138	DRX012423	51PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
17	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/DRR0138	DRX012424	2kbMP	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
18	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/DRR0138	DRX012425	101OF	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
19	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/DRR0138	DRX012429	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
20	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/DRR0138	DRX012427	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1
21	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/DRR0138	DRX012431	101PE	WGS	RANDOM	GENOMIC	PAIRED	0	0	ILLUMINA	Illumina HiSeq 1

Our First Script

wget_sra_files.py

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

'''
download all the Reikou SRA files

PRJDB1445
http://www.ncbi.nlm.nih.gov/bioproject/268159
http://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=268159
sendtofile => accession list => SraAccList.txt
example HTTP download URL
wget http://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR/DRR013/DRR013876/DRR013876.sra
'''

import os

#file listing SRA ids saved from ncbi through browser
inp = 'SraAccList.txt'

f = open(inp)
for line in f:
    uid = line.strip()
    #print uid

    id_list = [uid[:3], uid[:6], uid, uid + '.sra']

    cmd = 'wget http://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/'
    cmd += '/'.join(id_list)

    print cmd
    os.system(cmd)
f.close()

```

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()
```