Spreadsheet Data Processing in Python

University of Illinois at Urbana-Champaign

Christina T. Ruiz-Rodriguez

Objectives

- The objective of this project was to learn programming in Python.
- I created my own project using a dataset with results from my work with koala DNA single tandem repeats (STRs).
- An example of the dataset used for this project is shown in Table 1.

←AGAGAGAGAGAGAG→
Example of a STR (8 AG repeats)

DNA sequencing methods

Figure 1. Workflow to obtain DNA STRs from koala DNA

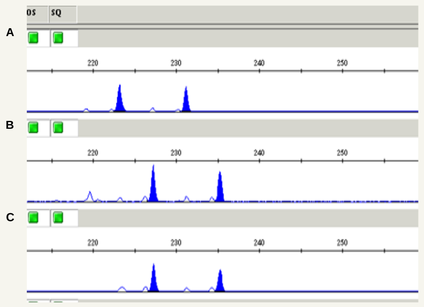
DNA isolated from whole blood from one koala

Shotgun sequencing using Roche 454 technology

PRIMER3 was used to design suitable primers for koala STRs

Genotyping samples and testing for polymorphism

- Results for STR genotypes in three different individual samples (A,B,C). The peaks represent the STRs of three individuals.



Project Data

Table 1. Excel spreadsheet with STR data from DNA sequencing.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Sample Name	Marker	Dye	Allele 1	Allele 2	Size 1	Size 2	Height 1	Height 2	Peak Area 1	Peak Area 2	Data Point 1	Data Point 2	ADO	AE	OS	SHP	OBA
2	11120P17-NA	Phci_17	G	158	158	157.71	157.71	1391	1391	8827	8827	2461	2461	FALSE	FALSE	0	0	0
3	11407P17-NA	Phci_17	G	160	163	159.56	163.43	978	792	5116	4423	2464	2508	FALSE	TRUE	-4	-4	-4
4	1217P17-NA	Phci_17	G	160	167	159.47	167.49	717	240	4250	1630	2395	2485	FALSE	FALSE	0	0	0
5	1218P17-NA	Phci_17	G	160	160	159.46	159.46	464	464	2733	2733	2379	2379	FALSE	FALSE	0	0	0
6	1219P17-NA	Phci_17	G	156	156	155.61	155.61	1470	1470	8983	8983	2388	2388	FALSE	FALSE	0	0	0
7	1220P17-NA	Phci_17	G	156	156	155.61	155.61	877	877	5292	5292	2412	2412	FALSE	FALSE	0	0	0
8	1224P17-NA	Phci_17	G	156	156	155.56	155.56	910	910	5342	5342	2372	2372	FALSE	FALSE	0	0	0
9	1225P17-NA	Phci_17	G	160	160	159.56	159.56	778	778	4635	4635	2392	2392	FALSE	FALSE	0	0	0
10	1226P17-NA	Phci_17	G	156	156	155.48	155.48	1378	1378	7901	7901	2305	2305	FALSE	TRUE	-4	-4	-1
11	1231P17-NA	Phci_17	G	156	160	155.57	159.46	920	609	5181	3224	2311	2354	FALSE	FALSE	0	0	0
12	1232P17-NA	Phci_17	G	156	160	155.48	159.46	878	666	5110	3447	2324	2368	FALSE	FALSE	0	0	0
13	1239P17-NA	Phci_17	G	156	156	155.57	155.57	1375	1375	8411	8411	2341	2341	FALSE	TRUE	-4	-4	-1
14	1241P17-NA	Phci_17	G	156	158	155.48	157.48	637	457	3312	2220	2358	2380	FALSE	FALSE	0	0	0
15	1242P17-NA	Phci_17	G	154	156	154.62	155.53	1553	1386	8337	8366	2335	2345	FALSE	TRUE	-4	-4	-4
16	1243P17-NA	Phci_17	G	156	156	155.57	155.57	934	934	5669	5669	2393	2393	FALSE	FALSE	0	0	0
17	1244P17-NA	Phci_17	G	160	160	159.46	159.46	1183	1183	6929	6929	2370	2370	FALSE	TRUE	-4	-4	-1
18	1249P17-NA	Phci_17	G	154	156	154.56	155.57	1859	1642	10496	9846	2349	2360	FALSE	FALSE	0	0	1
19	1250P17-NA	Phci_17	G	156	160	155.48	159.46	1092	889	5941	4515	2310	2354	FALSE	TRUE	-4	-4	-4
20	1252P17-NA	Phci_17	G	154	156	154.56	155.48	2905	2217	15191	13171	2305	2315	FALSE	FALSE	0	0	1
21	1253P17-NA	Phci_17	G	156	160	155.53	159.47	1140	845	6550	4384	2307	2351	FALSE	TRUE	-4	-4	-4
22	1254P17-NA	Phci_17	G	156	160	155.56	159.65	678	714	4063	4219	2416	2462	FALSE	FALSE	0	0	0
23	1258P17-NA	Phci_17	G	156	156	155.56	155.56	1087	1087	6997	6997	2476	2476	FALSE	TRUE	-4	-4	-1
24	1259P17-NA	Phci_17	G	154	156	154.57	155.49	1561	1280	8302	7650	2304	2314	FALSE	FALSE	0	0	1
25	1262P17-NA	Phci_17	G	158	160	157.5	159.47	964	776	5219	3767	2374	2396	FALSE	FALSE	0	0	0
26	1266P17-NA	Phci_17	G	156	156	155.48	155.48	1400	1400	8457	8457	2302	2302	FALSE	FALSE	0	0	0
27	1267P17-NA	Phci_17	G	156	160	155.57	159.46	954	886	5045	4594	2324	2367	FALSE	FALSE	0	0	0

Phci_17

Python Pseudo Code

Figure 2. Summary of python code. This program reads multiple large excel spreadsheets and outputs a new spreadsheet with a small subset of the all the data. The output is a .csv file with just the sample name of the koala and the STR data associated with each sample (Table 2).

```
I. Create list of koala samples ()
    {
        open_and_parse_file (koala _dictionary.csv)
        for every entry in the file
        {
            add to koala list
        }
    }
II. Create sample dictionary ()
    {
        Declare_raw_data_files (Phci_17. csv, Phci_18.csv)
        for each raw data file
        {
            for each row in the row file
            {
                match_columns_to _variables ([0],[1],[3],[4])
                if sample exists in dictionary
                {
                    append to exisiting data
                }
                else
                {
                    create new dictionary entry
                }
            }
        }
    }
III. Generate_output_file ()
     {
        Declare output file ()
        For koala in the koala list
        {
            If dictionary data exist(koala)
            {
                Output_data_as_a_row(koala_dictionary(koala))
            }
        }
     }
```

Results

Table 2. Output data spreadsheet. The data shown in the table consist of the sample name (column A) and STR data for each koala (columns B-E) at two different repeats (columns B and C; D and E). Question mark indicates missing data from that repeat.

	A	B	C	D	E
1	Sample Name	Phci 17	Phci 17	Phci 18	Phci 18
2	1217	160	167	?	?
3	1218	160	160	?	?
4	1224	156	156	?	?
5	1231	156	160	?	?
6	1232	156	160	?	?
7	1234	?	?	?	?
8	1239	156	156	?	?
9	1241	156	158	?	?
10	1242	154	156	?	?
11	1243	156	156	171	175
12	1249	154	156	?	?
13	1254	156	160	171	171
14	1258	156	156	?	?
15	1262	158	160	?	?
16	1267	156	160	?	?
17	1270	156	156	?	?
18	1272	158	160	171	171
19	1278	156	156	?	?
20	1288	156	160	?	?
21	1290	154	156	?	?
22	1331	154	156	?	?
23	1334	154	156	?	?
24	1335	154	160	?	?
25	1216	?	?	?	?
26	1219	156	156	?	?
27	1220	156	156	?	?
28	1225	160	160	?	?
29	1226	156	156	?	?
30	1238	?	?	?	?

Phci_str_out

Acknowledgements

Many thanks to Halie Rando and Diana Byrne for teaching me their amazing programming skills. Thanks to Heidi Imker and Ayla Stein for their support during this project. Many thanks to Simon for his help and input. This work was part of a Focal Point grant funded by the Graduate College at the University of Illinois at Urbana-Champaign.