

Introduction to Biocomputing Tutorial

Week 8






Debartolo 319

Review of quiz and homework

- `str.strip([chars])`
 - All chars are stripped from the beginning and the end of the string
 - Default whitespace characters
- `InFile=open("Lecture11.fasta","r"). # read in file`
- `for Line in InFile:`

Today's tutorial

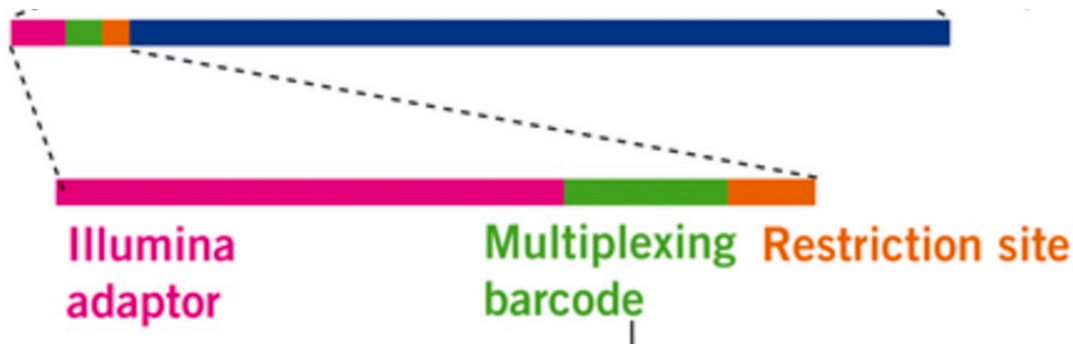
- `import re`
- `re.sub(regexString,replacement,searchString)`
- `re.search(regexString,searchString)`
- `match.group(index)` # 1 for the first match in parenthesis
- `var.start(index)` # Starting position for the match

 Cflorida.vcf	vcf file for Q1
 Exercise08_1_Pseudo.py	Pseudo code for Q1
 Exercise08_2_Pseudo.py	Pseudo code for Q2
 Exercise08_Functions_Python.py	Useful functions
 Exercise08_Python.pdf	Handout
 indivIDs.txt	ID file for Q2
 seqFastq.fq	fq file for Q2

Dictionaries in Python

```
dictionary = {}  
for line in file:  
    line = line.strip()  
    cols = line.split()  
    if cols[0] in dictionary:  
        print("Duplicate: " + cols[1])  
        break  
    else:  
        dictionary[cols[0]] = cols[1]  
  
dictionary[key]
```

RADseq Sequence Structure



- Adaptor structure
 - green barcode unique to individual
- Sequencing primer sits on pink adaptor
- Resulting data sequence

Raw Sequence Data (fastq format)

Info from sequencer (header) Sequence we care about

@FCC638CACXX:5:1101:1207:1875#ATCNCGATC/1

NATCCAGACAAATTCGCAAACATGTGTGGGTCAGCTGGGGTTTACGTAAACC
ATCGTTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCT

+ Another header (don't worry about this)

BP\cceeegcegghiihiiiiihfhhihiiiiihigghiiiiihifghiighigggggeecddcccc
ccc^bcccccccbcccY]_bbcccc_W`

Quality score for each bp (don't worry about this!)

Challenge #2 Goal

> v14.osb.014

CGCTGCTTGCTAAATATGCATGCTACCACTTTGCAATGGGGGTGCTGT
CGTACTCTTCCGATTTTGGAAGAGAATTGCATTTCGTGCA

> V10.F.016

TACTGCCTGAGGGTTATAATCCCTAAGTGCGGGCCAAAGGCCGCCTA
CGCAGAAAGGTGTTTCAGCCCCCCTTCTTTCGCAACACAGAA

> V10.F.013

ACGTCGATGACACACCAAATCCGACGGCGAATAATCCCAAATACT
CGGAGTAACATTCGACAGCCTGCTCTCCGTCTCTGCGCACG

> V10.F.035

AGGAGCGACACGCGCTCGTAGGCATAATGCAACAATTGCAACAAT
TTGTGCTTTAGATCGGAAGAGCACACGTCTGAACTCCAGTCA

SNP Data (VCF format)

Sample names we care about (line starts with #)

Header (usually many more lines starting with ##)

##Deleted a large header, all lines starting with ##

#CHROM	POS	INFO	FORMAT	CF.A.003	CF.A.004	CF.A.005	CF.A.006	CF.A.007	CF.A.008
Contig23	34	DeletedStuff	GT:AD:DP:GQ:PL	0/0:1,0:1:3:0,3,40			0/0:4,0:4:12:0,12,160	0/0:3,0:3:9:0,9,120	
Contig150	37	DeletedStuff	GT:AD:DP:GQ:PL	0/0:5,0:5:15:0,15,123			0/0:5,0:5:15:0,15,121	0/0:4,0:4:12:0,12,99	
Contig234	63	DeletedStuff	GT:AD:DP:GQ:PL	0/0:1,0:1:3:0,3,35			0/0:1,0:1:3:0,3,39	0/0:1,0:1:3:0,3,39	
Contig286	18	DeletedStuff	GT:AD:DP:GQ:PL	0/0:5,0:5:15:0,15,199			0/0:1,0:1:3:0,3,32	0/0:1,0:1:3:0,3,40	
Contig286	76	DeletedStuff	GT:AD:DP:GQ:PL	0/0:5,0:5:15:0,15,190			0/0:1,0:1:3:0,3,37	0/0:1,0:1:3:0,3,39	
Contig286	84	DeletedStuff	GT:AD:DP:GQ:PL	0/0:5,0:5:15:0,15,196			0/0:1,0:1:3:0,3,39	0/0:1,0:1:3:0,3,39	
Contig319	49	DeletedStuff	GT:AD:DP:GQ:PL	0/0:1,0:1:3:0,3,40			0/0:1,0:1:3:0,3,40	./.:.:.:.:.	0/0:2
Contig319	64	DeletedStuff	GT:AD:DP:GQ:PL	0/0:1,0:1:3:0,3,40			0/0:1,0:1:3:0,3,40	./.:.:.:.:.	0/0:2
Contig355	39	DeletedStuff	GT:AD:DP:GQ:PL	0/0:2,0:2:6:0,6,55			0/0:2,0:2:6:0,6,55	0/0:3,0:3:9:0,9,83	
Contig355	83	DeletedStuff	GT:AD:DP:GQ:PL	0/0:2,0:2:6:0,6,55			0/0:2,0:2:6:0,6,55	0/0:3,0:3:9:0,9,83	
Contig426	46	DeletedStuff	GT:AD:DP:GQ:PL	0/0:2,0:2:6:0,6,55			0/0:2,0:2:6:0,6,55	0/0:1,0:1:3:0,3,28	
Contig426	60	DeletedStuff	GT:AD:DP:GQ:PL	0/0:2,0:2:6:0,6,55			0/0:2,0:2:6:0,6,55	0/0:1,0:1:3:0,3,28	
Contig449	19	DeletedStuff	GT:AD:DP:GQ:PL	0/0:5,0:5:15:0,15,194			0/0:3,0:3:9:0,9,120	0/0:4,0:4:12:0,12,160	
Contig454	2	DeletedStuff	GT:AD:DP:GQ:PL	0/0:1,0:1:3:0,3,37			./.:.:.:.:.	0/0:1,0:1:3:0,3,40	0/0:5
Contig454	58	DeletedStuff	GT:AD:DP:GQ:PL	0/0:1,0:1:3:0,3,37			./.:.:.:.:.	0/0:1,0:1:3:0,3,34	0/0:5
Contig500	27	DeletedStuff	GT:AD:DP:GQ:PL	0/0:2,0:2:6:0,6,80			0/0:3,0:3:9:0,9,107	0/0:2,0:2:6:0,6,80	
Contig538	36	DeletedStuff	GT:AD:DP:GQ:PL	./.:.:.:.:.	0/0:2,0:2:6:0,6,80		0/0:5,0:5:15:0,15,200	0/0:2	
Contig588	70	DeletedStuff	GT:AD:DP:GQ:PL	0/0:1,0:1:3:0,3,39			0/0:1,0:1:3:0,3,39	0/0:2,0:2:6:0,6,77	
Contig588	78	DeletedStuff	GT:AD:DP:GQ:PL	0/0:1,0:1:3:0,3,39			0/0:1,0:1:3:0,3,39	0/0:2,0:2:6:0,6,77	
Contig588	84	DeletedStuff	GT:AD:DP:GQ:PL	0/0:1,0:1:3:0,3,39			0/0:1,0:1:3:0,3,39	0/0:2,0:2:6:0,6,77	

SNP positions

0/0:7,0:7:21:0,21,194

Per sample SNP data

Challenge #1 Goal

##Deleted a large header, all lines starting with ##

#CHROM	POS	INFO	FORMAT	Cf.Sfa.003	Cf.Sfa.004	Cf.Sfa.005	Cf.Sfa.006	
Contig23	34	DeletedStuff	GT:AD:DP:GQ:PL	1,0	4,0	3,0	1,0	4,0
Contig150	37	DeletedStuff	GT:AD:DP:GQ:PL	5,0	5,0	4,0	3,0	1,0
Contig234	63	DeletedStuff	GT:AD:DP:GQ:PL	1,0	1,0	1,0	NA	1,0
Contig286	18	DeletedStuff	GT:AD:DP:GQ:PL	5,0	1,0	1,0	1,0	NA
Contig286	76	DeletedStuff	GT:AD:DP:GQ:PL	5,0	1,0	1,0	1,0	NA
Contig286	84	DeletedStuff	GT:AD:DP:GQ:PL	5,0	1,0	1,0	1,0	NA
Contig319	49	DeletedStuff	GT:AD:DP:GQ:PL	1,0	1,0	NA	2,0	1,0
Contig319	64	DeletedStuff	GT:AD:DP:GQ:PL	1,0	1,0	NA	2,0	1,0
Contig355	39	DeletedStuff	GT:AD:DP:GQ:PL	2,0	2,0	3,0	2,0	4,0
Contig355	83	DeletedStuff	GT:AD:DP:GQ:PL	2,0	2,0	3,0	2,0	4,0
Contig426	46	DeletedStuff	GT:AD:DP:GQ:PL	2,0	2,0	1,0	2,0	3,0
Contig426	60	DeletedStuff	GT:AD:DP:GQ:PL	2,0	2,0	1,0	2,0	3,0
Contig449	19	DeletedStuff	GT:AD:DP:GQ:PL	5,0	3,0	4,0	4,0	7,0
Contig454	2	DeletedStuff	GT:AD:DP:GQ:PL	1,0	NA	1,0	5,0	3,0
Contig454	58	DeletedStuff	GT:AD:DP:GQ:PL	1,0	NA	1,0	5,0	3,0
Contig500	27	DeletedStuff	GT:AD:DP:GQ:PL	2,0	3,0	2,0	5,0	1,1
Contig538	36	DeletedStuff	GT:AD:DP:GQ:PL	NA	2,0	5,0	2,0	1,0
Contig588	70	DeletedStuff	GT:AD:DP:GQ:PL	1,0	1,0	2,0	1,0	3,1
Contig588	70	DeletedStuff	GT:AD:DP:GQ:PL	1,0	1,0	2,0	1,0	4,0

For After Break

- Monday (10/23)
- Read Kelly *et al.* 2014
 - Posted on Sakai
 - Quiz at the beginning of lecture!
- Friday (10/27)
- Exercise 8: String manipulation with regex
 - Work through the exercise together
 - One member of each team submits a pull request
 - Due by start of next tutorial

