Demultiplexing, Assignment-the-third

Files located on talapas here:

```
/projects/bgmp/camk/bioinfo/Bi622/Part3_Demultiplex
```

Test input files:

```
test_input_R1.fq.gz
test_input_R2.fq.gz
test_input_R3.fq.gz
test_input_R4.fq.gz
```

My test-output files are on github, and the ones created by the program looked identical to them, so I'm happy with the results.

Main demultiplex python script:

```
Demultiplex.py
```

Bash script for running in slurm:

```
run_demultiplex.sh
```

Threshold was set at 30, this gets data thats very high quality, only 0.1% chance of an incorrect base call, without further knowledge of what the data is for, I think setting a reasonably high threshold is better so that the downstream data will have higher call accuracy and fewer errors.

Script output:

```
Number of unknown records: 57748853
Number of index hopped records: 517612
Number of TACCGGAT records: 69307073
Number of CTCTGGAT records: 32163349
Number of AGAGTCCA records: 10378366
Number of GTAGCGTA records: 7450201
Number of ATCATGCG records: 9264615
```

```
Number of AACAGCGA records: 8178191
Number of TCGACAAG records: 3548541
Number of TCGAGAGT records: 10658212
Number of CGGTAATC records: 4498136
Number of TAGCCATG records: 9852258
Number of TCTTCGAC records: 39149148
Number of CTAGCTCA records: 16162895
Number of TATGGCAC records: 10195805
Number of ACGATCAG records: 7441721
Number of GATCTTGC records: 3425453
Number of AGGATAGC records: 8078057
Number of TGTTCCGT records: 14786868
Number of CACTTCAC records: 3833640
Number of GCTACTCT records: 6610857
Number of ATCGTGGT records: 6357656
Number of CGATCGAT records: 5225776
Number of GTCCTAAG records: 8164223
Number of GATCAAGG records: 6085915
Number of TCGGATTC records: 4163314

Command being timed: "python Demultiplex.py -1
/projects/bgmp/shared/2017_sequencing/1294_S1_L008_R1_001.fastq.gz -2
/projects/bgmp/shared/2017_sequencing/1294_S1_L008_R2_001.fastq.gz -3
/projects/bgmp/shared/2017_sequencing/1294_S1_L008_R3_001.fastq.gz -4
/projects/bgmp/shared/2017_sequencing/1294_S1_L008_R4_001.fastq.gz"

User time (seconds): 51510.04
System time (seconds): 10.47
Percent of CPU this job got: 98%
Elapsed (wall clock) time (h:mm:ss or m:ss): 14:30:03
Average shared text size (kbytes): 0
Average unshared data size (kbytes): 0
Average stack size (kbytes): 0
Average total size (kbytes): 0
Maximum resident set size (kbytes): 263328
Average resident set size (kbytes): 0
Major (requiring I/O) page faults: 0
Minor (reclaiming a frame) page faults: 192128
Voluntary context switches: 13109
Involuntary context switches: 120504
```

```
Swaps: 0
File system inputs: 0
File system outputs: 0
Socket messages sent: 0
Socket messages received: 0
Signals delivered: 0
Page size (bytes): 4096
Exit status: 0
```

Record table:

Name	# of Records	% of Total Records
Unknown	57748853	15.89796891
Index Hopped	517612	0.1424959814
TACCGGAT	69307073	19.07988877
CTCTGGAT	32163349	8.854408285
AGAGTCCA	10378366	2.857111985
GTAGCGTA	7450201	2.051002881
ATCATGCG	9264615	2.550501934
AACAGCGA	8178191	2.251414868
TCGACAAG	3548541	0.9768954978
TCGAGAGT	10658212	2.934152182
CGGTAATC	4498136	1.238314227
TAGCCATG	9852258	2.712277097
TCTTCGAC	39149148	10.77756363
CTAGCTCA	16162895	4.449563738
TATGGCAC	10195805	2.806853859
ACGATCAG	7441721	2.04866838
GATCTTGC	3425453	0.9430099902
AGGATAGC	8078057	2.223848481
TGTTCCGT	14786868	4.070750423
CACTTCAC	3833640	1.055381819
GCTACTCT	6610857	1.819935697
ATCGTGGT	6357656	1.750230735

CGATCGAT	5225776	1.438629861
GTCCTAAG	8164223	2.247569548
GATCAAGG	6085915	1.675421804
TCGGATTC	4163314	1.146139414
Total	363246735	100

So overall index hopping was quite low, about 0.15% of records showed signs of index hopping.