BedFileParser

BedFileParser is a tool to read BED files and allows you to subset or summarize it. It includes following features

- Read BED file and parse according to specifications.
- Subset BED file using chromosome only or using chromosome and genomic position.
- Subset BED using feature names.
- Calculate summary statistics on whole BED file
- Output to specified text file

Installation

BedFileParser requires pandas library to run. Use pip to install dependencies and package

Start with downloading the zip file from github. Once downloaded, unzip to a directory called BedFileParser. Then follow below prompts:

```
$ cd BedFileParser
$ pip install -e .
```

This will build the package and install it on system-wide level.

Testing

The package comes with pre-written tests to ensure everything is working as expected. These tests can be run using the following command:

```
$ python setup.py test
```

Below is the expected result:

```
running test
running egg_info
creating BedFileParser.egg-info
writing requirements to BedFileParser.egg-info/requires.txt
writing BedFileParser.egg-info/PKG-INFO
writing top-level names to BedFileParser.egg-info/top_level.txt
writing dependency_links to BedFileParser.egg-info/dependency_links.txt
writing entry points to BedFileParser.egg-info/entry_points.txt
writing manifest file 'BedFileParser.egg-info/SOURCES.txt'
reading manifest file 'BedFileParser.egg-info/SOURCES.txt'
reading manifest template 'MANIFEST.in'
writing manifest file 'BedFileParser.egg-info/SOURCES.txt'
running build_ext
```

```
test_fileload_chr (tests.test_BedFileParser.MyTestCase) ... Offending Line: 1	1	2999	gene1	+
ok
test_fileload_coord (tests.test_BedFileParser.MyTestCase) ... Offending Line: chr1	2999	1	gene1	+
ok
test_fileload_feature (tests.test_BedFileParser.MyTestCase) ... Offending Line: chr2	8000	74000	gene2!	+
ok
test_fileload_strand (tests.test_BedFileParser.MyTestCase) ... Offending Line: chr1	3	2442	genex	_
ok
test_fileload_working (tests.test_BedFileParser.MyTestCase) ... ok
test_subset_Chr_working (tests.test_BedFileParser.MyTestCase) ... ok
test_subset_Feature_working (tests.test_BedFileParser.MyTestCase) ... ok
test_subset_Pos_working (tests.test_BedFileParser.MyTestCase) ... ok
test_subset_chr (tests.test_BedFileParser.MyTestCase) ... Search Input: 1
ok
test_subset_coord (tests.test_BedFileParser.MyTestCase) ... Search start and end input: 3000 1
ok
test_subset_feature (tests.test_BedFileParser.MyTestCase) ... Search feature name input: genex!
```

```
ok
test_summary_working (tests.test_BedFileParser.MyTestCase) ... ok

----------------------------------------------------------------------
Ran 12 tests in 0.024s

OK

```

Usage

Once installed BedFileParser can be called commandline using `BedFileParser` command directly.

Using `-h` will show help:

```
$ BedFileParser -h
usage: BedFileParser [-h] -f FILE
                     (--summary | --chrom CHROM | --feature FEATURE)
                     [--pos start end] [--outfile OUTFILE]

A script to process a bedfile and return a subset or summary statistics

```

```
optional arguments:
  -h, --help            show this help message and exit
  -f FILE, --file FILE  Name of the file (Required)
  --summary             Output statistics across all chromosomes to screen
  --chrom CHROM         Chromosome name for subset. Providing this option
                        without start/end will result in all entries of that
                        chromosome to be outputted
  --feature FEATURE     Search by feature name
  --pos start end       Start and end position on chromosome to subset
                        (Optional)
  --outfile OUTFILE     name of output file. printed to cmd-line if not
                        specified (Optional)

```

Notes

- filename is a required option
- Only 1 option can be specified at one time: `--summary`, `--chrom`, `--feature`. This allows to either show the summary, subset using chromosome name or subset using feature name.
- `--pos` must specify **both** *start* and *end*. integer value between 1 and 2^32. Start < End.
- `--chrom` option must be named as chrXX, where XX is 1-22.

- `--featureName` can include alphanumeric, underscore, hyphen, parenthesis
- strand must be ‘+’ or ‘-’ in BED file.

Respective errors will be thrown if any of these specifications are violated.

Examples (w/ expected output)

The below commands assume you have data folder with file `test_bed_working.bed`

Summarize all information from bed file

```
$ BedFileParser -f data/test_bed_working.bed --summary
       TotalFeatures  NumFeaturesPos  NumFeaturesNeg    mi
n    max     mean
chrom

1                  2               1               1   243
9   2998   2718.5
2                  1               1               0  6600
0  66000  66000.0
3                  1               0               1   600
0   6000   6000.0
```

Subset BED file by chr1

```
$ BedFileParser -f data/test_bed_working.bed --chrom chr1
1   1   2999    gene1   +
1   3   2442    genex   -
```

Subset BED file by chr1 and genomic coordinates 1-2500

```
$ BedFileParser -f data/test_bed_working.bed --chrom chr1 --pos 1 2500
1   3   2442    genex   -
```

Subset BED file by feature name (e.g. gene name)

```
$ BedFileParser -f data/test_bed_working.bed --feature genex
1   3   2442    genex   -
```

Output the results to a text file instead of printing to command line.

```
$ BedFileParser -f data/test_bed_working.bed --feature genex --outfile outfile.txt
$
```

Summary statistics of hg19 transcriptome (chr1-chr22) downloaded from UCSC genome browser. Assumes `data/hg19_UCSC.bed` file:

```
$ BedFileParser -f data/hg19_UCSC.bed --summary
```

```
       TotalFeatures  NumFeaturesPos  NumFeaturesNeg  min
     max  \
chrom

1               6529            3424            3105   41
 2320934
11              3791            1976            1815   20
 2173326
12              3471            1761            1710   60
 1249864
13              1346             656             690   48
 1468615
14              2146            1131            1015   21
 1697918
15              2361            1259            1102   33
  949245
16              2673            1509            1164   51
 1694321
17              3534            1748            1786   47
 1143720
18              1104             517             587   50
 1195732
19              4029            2076            1953   47
  341194
2               4614            2323            2291   49
 1900275
21               872             448             424   60
  836140
```

```
22              1492            755             737   52
  701852
3               4030           2088            1942   21
 1504132
4               2522           1350            1172   44
 1474687
5               3178           1634            1544   43
 1527247
6               3450           1715            1735   20
 1987243
7               3220           1621            1599   53
 2304636
8               2808           1326            1482   21
 2059454
9               2646           1307            1339   54
 2298478

                mean
chrom
1       57137.405116
11      54168.833553
12      63945.829444
13      86430.567608
14      62322.189189
15      58661.018636
16      40246.624392
17      36380.125071
18     108710.716486
```

```
19        20436.010921
2         83889.263112
21        71915.842890
22        42571.552279
3         87247.528040
4         89656.109437
5         78112.614223
6         69087.652174
7         87092.219565
8        105115.475427
9         62937.773243
```

License

MIT