Partitioning the HMP mock data set

Partitioning breaks your data set into many smaller chunks so that they can be assembled more easily. See the k-mer percolation paper (at http://ivory.idyll.org/blog/dec-11/kmer-percolation-posted.html) for more information.

Here we are starting with the 'keep.abundfilt' files from the previous tutorial.

In [1]:
```
cd /test/tutorial-files/
```

```
/test/tutorial-files
```

In [2]:
```
!ls
```

```
mock-pass1.kh  SRR172903.fastq.gz.keep
pass1.report   SRR172903.fastq.gz.keep.below
```

The *keep file is the diginormed file. The *below file is the high abundance filtered file (previously diginormed). We'll work with the *below file for partitioning.

Initial round of partitioning

In [3]:
```
!/usr/local/src/khmer/scripts/load-graph.py -k 32 -N 4 -x 1e9 /test/tutori
```

```
PARAMETERS:
 - kmer size =    32            (-k)
 - n hashes =     4             (-N)
 - min hashsize = 1e+09         (-x)

Estimated memory usage is 5e+08 bytes (n_hashes x min_hashsize / 8)
--------
Saving hashtable to /test/tutorial-files/mock.part1
Loading kmers from sequences in ['SRR172903.fastq.gz.keep.below']
We WILL build the tagset (for partitioning/traversal).
making hashtable
consuming input SRR172903.fastq.gz.keep.below
n tags: 173339
... consume_fasta_and_tag 100000 3741212
n tags: 331081
... consume_fasta_and_tag 200000 6999381
n tags: 478735
... consume_fasta_and_tag 300000 9934364
n tags: 618123
... consume_fasta_and_tag 400000 12608998
n tags: 750486
... consume_fasta_and_tag 500000 15075380
n tags: 878322
```

```
... consume_fasta_and_tag 600000 17402849
n tags: 1000104
... consume_fasta_and_tag 700000 19562100
n tags: 1118520
... consume_fasta_and_tag 800000 21640685
n tags: 1233477
... consume_fasta_and_tag 900000 23642413
n tags: 1345752
... consume_fasta_and_tag 1000000 25572425
n tags: 1453688
... consume_fasta_and_tag 1100000 27401883
n tags: 1555629
... consume_fasta_and_tag 1200000 29083819
n tags: 1657186
... consume_fasta_and_tag 1300000 30781061
n tags: 1761866
... consume_fasta_and_tag 1400000 32598915
n tags: 1856796
... consume_fasta_and_tag 1500000 34133496
n tags: 1949962
... consume_fasta_and_tag 1600000 35644956
n tags: 2054990
... consume_fasta_and_tag 1700000 37524893
n tags: 2146884
... consume_fasta_and_tag 1800000 39023347
n tags: 2246848
... consume_fasta_and_tag 1900000 40794224
n tags: 2349898
... consume_fasta_and_tag 2000000 42651231
n tags: 2475005
... consume_fasta_and_tag 2100000 45207089
n tags: 2573712
... consume_fasta_and_tag 2200000 46949732
n tags: 2662785
... consume_fasta_and_tag 2300000 48418135
n tags: 2753173
... consume_fasta_and_tag 2400000 49912426
n tags: 2862516
... consume_fasta_and_tag 2500000 52019582
n tags: 2958757
... consume_fasta_and_tag 2600000 53709507
n tags: 3076083
... consume_fasta_and_tag 2700000 56066157
n tags: 3182492
... consume_fasta_and_tag 2800000 58049025
n tags: 3278634
... consume_fasta_and_tag 2900000 59731112
n tags: 3374429
... consume_fasta_and_tag 3000000 61400767
n tags: 3468657
... consume_fasta_and_tag 3100000 63026253
n tags: 3561344
... consume_fasta_and_tag 3200000 64618636
n tags: 3655717
```

```
... consume_fasta_and_tag 3300000 66264730
n tags: 3762370
... consume_fasta_and_tag 3400000 68248082
n tags: 3872647
... consume_fasta_and_tag 3500000 70337348
n tags: 3998433
... consume_fasta_and_tag 3600000 73025549
n tags: 4104092
... consume_fasta_and_tag 3700000 75018507
n tags: 4207519
... consume_fasta_and_tag 3800000 76938328
n tags: 4310757
... consume_fasta_and_tag 3900000 78846812
n tags: 4416067
... consume_fasta_and_tag 4000000 80809556
n tags: 4520778
... consume_fasta_and_tag 4100000 82758527
n tags: 4626628
... consume_fasta_and_tag 4200000 84738555
n tags: 4733922
... consume_fasta_and_tag 4300000 86756232
n tags: 4843321
... consume_fasta_and_tag 4400000 88844404
n tags: 4955072
... consume_fasta_and_tag 4500000 91010330
n tags: 5074409
... consume_fasta_and_tag 4600000 93393271
saving hashtable in /test/tutorial-files/mock.part1.ht
saving tagset in /test/tutorial-files/mock.part1.tagset
fp rate estimated to be 0.000
```

In [4]:
```
!/usr/local/src/khmer/scripts/partition-graph.py --threads 8 -s 1e5 /test/
```

```
--
SUBSET SIZE 100000.0
N THREADS 8
--
loading ht /test/tutorial-files/mock.part1.ht
** Traverse all the things: stop_big_traversals is false.
enqueued 51 subset tasks
starting 8 threads
---
starting: /test/tutorial-files/mock.part1 0
starting: /test/tutorial-files/mock.part1 1
starting: /test/tutorial-files/mock.part1 2
starting: /test/tutorial-files/mock.part1 3
starting: /test/tutorial-files/mock.part1 4
starting: /test/tutorial-files/mock.part1 5
starting: /test/tutorial-files/mock.part1 6
done starting threads
starting: /test/tutorial-files/mock.part1 7
...subset-part 608325372573204944-875280618668052438: 100000 <- 88135
saving: /test/tutorial-files/mock.part1 4
```

```
starting: /test/tutorial-files/mock.part1 8
...subset-part 16777984-94137232714629893: 100000 <- 81154
saving: /test/tutorial-files/mock.part1 0
...subset-part 94137232714629893-260207897496116208: 100000 <- 86167
saving: /test/tutorial-files/mock.part1 1
starting: /test/tutorial-files/mock.part1 9
starting: /test/tutorial-files/mock.part1 10
...subset-part 260207897496116208-409737121305374142: 100000 <- 84020
saving: /test/tutorial-files/mock.part1 2
starting: /test/tutorial-files/mock.part1 11
...subset-part 1110096598908202681-1297623081413842401: 100000 <- 84884
saving: /test/tutorial-files/mock.part1 6
...subset-part 1297623081413842401-1531300088663273228: 100000 <- 85728
saving: /test/tutorial-files/mock.part1 7
starting: /test/tutorial-files/mock.part1 12
...subset-part 875280618668052438-1110096598908202681: 100000 <- 88032
saving: /test/tutorial-files/mock.part1 5
...subset-part 409737121305374142-608325372573204944: 100000 <- 85647
saving: /test/tutorial-files/mock.part1 3
 starting: /test/tutorial-files/mock.part1 13
starting: /test/tutorial-files/mock.part1 14
starting: /test/tutorial-files/mock.part1 15
...subset-part 1531300088663273228-1732089634563558140: 100000 <- 86171
saving: /test/tutorial-files/mock.part1 8
starting: /test/tutorial-files/mock.part1 16
...subset-part 1732089634563558140-1859649043415500255: 100000 <- 87256
saving: /test/tutorial-files/mock.part1 9
starting: /test/tutorial-files/mock.part1 17
...subset-part 2306518342785591497-2644758356675095272: 100000 <- 87789
saving: /test/tutorial-files/mock.part1 12
starting: /test/tutorial-files/mock.part1 18
...subset-part 2644758356675095272-3004050199097019834: 100000 <- 88717
saving: /test/tutorial-files/mock.part1 13
...subset-part 2049400903460208816-2306518342785591497: 100000 <- 87383
saving: /test/tutorial-files/mock.part1 11
...subset-part 3261770092864709950-3537229543903104406: 100000 <- 88267
saving: /test/tutorial-files/mock.part1 15
starting: /test/tutorial-files/mock.part1 19starting: /test/tutorial-
files/mock.part1 20

starting: /test/tutorial-files/mock.part1 21
...subset-part 1859649043415500255-2049400903460208816: 100000 <- 86683
saving: /test/tutorial-files/mock.part1 10
starting: /test/tutorial-files/mock.part1 22
...subset-part 3004050199097019834-3261770092864709950: 100000 <- 86968
saving: /test/tutorial-files/mock.part1 14
starting: /test/tutorial-files/mock.part1 23
...subset-part 3537229543903104406-3883333354027009776: 100000 <- 88367
saving: /test/tutorial-files/mock.part1 16
starting: /test/tutorial-files/mock.part1 24
...subset-part 3883333354027009776-4209144081561864332: 100000 <- 87991
saving: /test/tutorial-files/mock.part1 17
...subset-part 4684895333112905664-4995697692860048572: 100000 <- 83550
saving: /test/tutorial-files/mock.part1 20
```

```
starting: /test/tutorial-files/mock.part1 25
starting: /test/tutorial-files/mock.part1 26
...subset-part 4995697692860048572-5357121369695934211: 100000 <- 86501
saving: /test/tutorial-files/mock.part1 21
starting: /test/tutorial-files/mock.part1 27
...subset-part 5357121369695934211-5800693503835845052: 100000 <- 88304
saving: /test/tutorial-files/mock.part1 22
starting: /test/tutorial-files/mock.part1 28
...subset-part 5800693503835845052-6107027715926537746: 100000 <- 84684
saving: /test/tutorial-files/mock.part1 23
...subset-part 4462552133586894742-4684895333112905664: 100000 <- 88154
saving: /test/tutorial-files/mock.part1 19
starting: /test/tutorial-files/mock.part1 29
starting: /test/tutorial-files/mock.part1 30
...subset-part 4209144081561864332-4462552133586894742: 100000 <- 86335
saving: /test/tutorial-files/mock.part1 18
starting: /test/tutorial-files/mock.part1 31
...subset-part 6107027715926537746-6322925388513717670: 100000 <- 85185
saving: /test/tutorial-files/mock.part1 24
starting: /test/tutorial-files/mock.part1 32
...subset-part 6541038354198814700-6838569075732247502: 100000 <- 88618
saving: /test/tutorial-files/mock.part1 26
...subset-part 6322925388513717670-6541038354198814700: 100000 <- 88834
saving: /test/tutorial-files/mock.part1 25
starting: /test/tutorial-files/mock.part1 33
starting: /test/tutorial-files/mock.part1 34
...subset-part 6838569075732247502-7183080571868626843: 100000 <- 88294
saving: /test/tutorial-files/mock.part1 27
starting: /test/tutorial-files/mock.part1 35
...subset-part 7183080571868626843-7326490994813815436: 100000 <- 89291
saving: /test/tutorial-files/mock.part1 28
starting: /test/tutorial-files/mock.part1 36
...subset-part 7326490994813815436-7603579705978108979: 100000 <- 86898
saving: /test/tutorial-files/mock.part1 29
starting: /test/tutorial-files/mock.part1 37
...subset-part 7603579705978108979-7795999829424067896: 100000 <- 86160
saving: /test/tutorial-files/mock.part1 30
starting: /test/tutorial-files/mock.part1 38
...subset-part 7795999829424067896-8063092306009078686: 100000 <- 81920
saving: /test/tutorial-files/mock.part1 31
starting: /test/tutorial-files/mock.part1 39
...subset-part 8063092306009078686-8441249233996454970: 100000 <- 88578
saving: /test/tutorial-files/mock.part1 32
starting: /test/tutorial-files/mock.part1 40
...subset-part 8441249233996454970-8840279918210633727: 100000 <- 88411
saving: /test/tutorial-files/mock.part1 33
starting: /test/tutorial-files/mock.part1 41
...subset-part 9137393177526664591-9588858429440543234: 100000 <- 89144
saving: /test/tutorial-files/mock.part1 35
starting: /test/tutorial-files/mock.part1 42
...subset-part 8840279918210633727-9137393177526664591: 100000 <- 86330
saving: /test/tutorial-files/mock.part1 34
starting: /test/tutorial-files/mock.part1 43
...subset-part 10072415962488957630-10579129017951714586: 100000 <- 88105
```

```
saving: /test/tutorial-files/mock.part1 37
starting: /test/tutorial-files/mock.part1 44
...subset-part 9588858429440543234-10072415962488957630: 100000 <- 87681
saving: /test/tutorial-files/mock.part1 36
starting: /test/tutorial-files/mock.part1 45
...subset-part 10579129017951714586-11032074342873574171: 100000 <- 88812
saving: /test/tutorial-files/mock.part1 38
starting: /test/tutorial-files/mock.part1 46
...subset-part 11032074342873574171-11054958194102887515: 100000 <- 90164
saving: /test/tutorial-files/mock.part1 39
starting: /test/tutorial-files/mock.part1 47
...subset-part 11054958194102887515-11433525864876977338: 100000 <- 84111
saving: /test/tutorial-files/mock.part1 40
starting: /test/tutorial-files/mock.part1 48
...subset-part 11433525864876977338-11928641720686732059: 100000 <- 86210
saving: /test/tutorial-files/mock.part1 41
starting: /test/tutorial-files/mock.part1 49
...subset-part 11928641720686732059-12298392609422803627: 100000 <- 81400
saving: /test/tutorial-files/mock.part1 42
starting: /test/tutorial-files/mock.part1 50
...subset-part 12298392609422803627-12518823802717580798: 100000 <- 80128
saving: /test/tutorial-files/mock.part1 43
exiting
...subset-part 12797720555158153786-13323740612976716198: 100000 <- 83265
saving: /test/tutorial-files/mock.part1 45
...subset-part 12518823802717580798-12797720555158153786: 100000 <- 81438
saving: /test/tutorial-files/mock.part1 44
exiting
exiting
...subset-part 13689469228424413878-14292920736027679690: 100000 <- 86515
saving: /test/tutorial-files/mock.part1 47
exiting
...subset-part 13323740612976716198-13689469228424413878: 100000 <- 79773
saving: /test/tutorial-files/mock.part1 46
exiting
...subset-part 14292920736027679690-15472918990353706926: 100000 <- 87308
saving: /test/tutorial-files/mock.part1 48
exiting
...subset-part 15472918990353706926-16635852473598717694: 100000 <- 86802
saving: /test/tutorial-files/mock.part1 49
exiting
saving: /test/tutorial-files/mock.part1 50
exiting
---
done making subsets! see /test/tutorial-files/mock.part1.subset.*.pmap
```

In [5]:
```
ls
```

```
mock.part1.ht              mock.part1.subset.34.pmap
mock.part1.info            mock.part1.subset.35.pmap
mock.part1.subset.0.pmap   mock.part1.subset.36.pmap
mock.part1.subset.10.pmap  mock.part1.subset.37.pmap
mock.part1.subset.11.pmap  mock.part1.subset.38.pmap
mock.part1.subset.12.pmap  mock.part1.subset.39.pmap
```

```
mock.part1.subset.13.pmap  mock.part1.subset.3.pmap
mock.part1.subset.14.pmap  mock.part1.subset.40.pmap
mock.part1.subset.15.pmap  mock.part1.subset.41.pmap
mock.part1.subset.16.pmap  mock.part1.subset.42.pmap
mock.part1.subset.17.pmap  mock.part1.subset.43.pmap
mock.part1.subset.18.pmap  mock.part1.subset.44.pmap
mock.part1.subset.19.pmap  mock.part1.subset.45.pmap
mock.part1.subset.1.pmap   mock.part1.subset.46.pmap
mock.part1.subset.20.pmap  mock.part1.subset.47.pmap
mock.part1.subset.21.pmap  mock.part1.subset.48.pmap
mock.part1.subset.22.pmap  mock.part1.subset.49.pmap
mock.part1.subset.23.pmap  mock.part1.subset.4.pmap
mock.part1.subset.24.pmap  mock.part1.subset.50.pmap
mock.part1.subset.25.pmap  mock.part1.subset.5.pmap
mock.part1.subset.26.pmap  mock.part1.subset.6.pmap
mock.part1.subset.27.pmap  mock.part1.subset.7.pmap
mock.part1.subset.28.pmap  mock.part1.subset.8.pmap
mock.part1.subset.29.pmap  mock.part1.subset.9.pmap
mock.part1.subset.2.pmap   mock.part1.tagset
mock.part1.subset.30.pmap  mock-pass1.kh
mock.part1.subset.31.pmap  pass1.report
mock.part1.subset.32.pmap  SRR172903.fastq.gz.keep
mock.part1.subset.33.pmap  SRR172903.fastq.gz.keep.below
```

In [6]:
```
!/usr/local/src/khmer/scripts/merge-partitions.py mock.part1
```

```
loading 51 pmap files (first one: mock.part1.subset.4.pmap)
merging mock.part1.subset.4.pmap
merging mock.part1.subset.0.pmap
merging mock.part1.subset.1.pmap
merging mock.part1.subset.2.pmap
merging mock.part1.subset.6.pmap
merging mock.part1.subset.7.pmap
merging mock.part1.subset.5.pmap
merging mock.part1.subset.3.pmap
merging mock.part1.subset.8.pmap
merging mock.part1.subset.9.pmap
merging mock.part1.subset.12.pmap
merging mock.part1.subset.13.pmap
merging mock.part1.subset.11.pmap
merging mock.part1.subset.15.pmap
merging mock.part1.subset.10.pmap
merging mock.part1.subset.14.pmap
merging mock.part1.subset.16.pmap
merging mock.part1.subset.17.pmap
merging mock.part1.subset.20.pmap
merging mock.part1.subset.21.pmap
merging mock.part1.subset.22.pmap
merging mock.part1.subset.23.pmap
merging mock.part1.subset.19.pmap
merging mock.part1.subset.18.pmap
merging mock.part1.subset.24.pmap
merging mock.part1.subset.26.pmap
```

```
merging mock.part1.subset.25.pmap
merging mock.part1.subset.27.pmap
merging mock.part1.subset.28.pmap
merging mock.part1.subset.29.pmap
merging mock.part1.subset.30.pmap
merging mock.part1.subset.31.pmap
merging mock.part1.subset.32.pmap
merging mock.part1.subset.33.pmap
merging mock.part1.subset.35.pmap
merging mock.part1.subset.34.pmap
merging mock.part1.subset.37.pmap
merging mock.part1.subset.36.pmap
merging mock.part1.subset.38.pmap
merging mock.part1.subset.39.pmap
merging mock.part1.subset.40.pmap
merging mock.part1.subset.41.pmap
merging mock.part1.subset.42.pmap
merging mock.part1.subset.43.pmap
merging mock.part1.subset.45.pmap
merging mock.part1.subset.44.pmap
merging mock.part1.subset.47.pmap
merging mock.part1.subset.46.pmap
merging mock.part1.subset.48.pmap
merging mock.part1.subset.49.pmap
merging mock.part1.subset.50.pmap
saving merged to mock.part1.pmap.merged
removing pmap files
```

In [7]:
```
!/usr/local/src/khmer/scripts/annotate-partitions.py mock.part1 *.keep.bel
```

```
loading partition map from: mock.part1.pmap.merged
outputting partitions for SRR172903.fastq.gz.keep.below
... output_partitions 100000 0
... output_partitions 200000 0
... output_partitions 300000 0
... output_partitions 400000 0
... output_partitions 500000 0
... output_partitions 600000 0
... output_partitions 700000 0
... output_partitions 800000 0
... output_partitions 900000 0
... output_partitions 1000000 0
... output_partitions 1100000 0
... output_partitions 1200000 0
... output_partitions 1300000 0
... output_partitions 1400000 0
... output_partitions 1500000 0
... output_partitions 1600000 0
... output_partitions 1700000 0
... output_partitions 1800000 0
... output_partitions 1900000 0
... output_partitions 2000000 0
... output_partitions 2100000 0
```

```
... output_partitions 2200000 0
... output_partitions 2300000 0
... output_partitions 2400000 0
... output_partitions 2500000 0
... output_partitions 2600000 0
... output_partitions 2700000 0
... output_partitions 2800000 0
... output_partitions 2900000 0
... output_partitions 3000000 0
... output_partitions 3100000 0
... output_partitions 3200000 0
... output_partitions 3300000 0
... output_partitions 3400000 0
... output_partitions 3500000 0
... output_partitions 3600000 0
... output_partitions 3700000 0
... output_partitions 3800000 0
... output_partitions 3900000 0
... output_partitions 4000000 0
... output_partitions 4100000 0
... output_partitions 4200000 0
... output_partitions 4300000 0
... output_partitions 4400000 0
... output_partitions 4500000 0
... output_partitions 4600000 0
output 477730 partitions for SRR172903.fastq.gz.keep.below
partitions are in SRR172903.fastq.gz.keep.below.part
```

In [8]:
```
!/usr/local/src/khmer/scripts/extract-partitions.py mock.part1 *keep.below
```

```
---
reading partitioned files: ['SRR172903.fastq.gz.keep.below.part']
outputting to files named "mock.part1.groupN.fa"
min reads to keep a partition: 5
max size of a group file: 1000000
partition size distribution will go to mock.part1.dist
---
... 0
... 100000
... 200000
... 300000
... 400000
... 500000
... 600000
... 700000
... 800000
... 900000
... 1000000
... 1100000
... 1200000
... 1300000
... 1400000
... 1500000
```

```
... 1600000
... 1700000
... 1800000
... 1900000
... 2000000
... 2100000
... 2200000
... 2300000
... 2400000
... 2500000
... 2600000
... 2700000
... 2800000
... 2900000
... 3000000
... 3100000
... 3200000
... 3300000
... 3400000
... 3500000
... 3600000
... 3700000
... 3800000
... 3900000
... 4000000
... 4100000
... 4200000
... 4300000
... 4400000
... 4500000
2 groups
...x2 0
...x2 100000
...x2 200000
...x2 300000
...x2 400000
...x2 500000
...x2 600000
...x2 700000
...x2 800000
...x2 900000
...x2 1000000
...x2 1100000
...x2 1200000
...x2 1300000
...x2 1400000
...x2 1500000
...x2 1600000
...x2 1700000
...x2 1800000
...x2 1900000
...x2 2000000
...x2 2100000
...x2 2200000
```

```
...x2 2300000
...x2 2400000
...x2 2500000
...x2 2600000
...x2 2700000
...x2 2800000
...x2 2900000
...x2 3000000
...x2 3100000
...x2 3200000
...x2 3300000
...x2 3400000
...x2 3500000
...x2 3600000
...x2 3700000
...x2 3800000
...x2 3900000
...x2 4000000
...x2 4100000
...x2 4200000
...x2 4300000
...x2 4400000
...x2 4500000
```

Done!

You can see now that you have some grouped files which you can assemble in parallel. This was actually a very small dataset, so you have only 2 of these files. However, for larger metagenomes, you could have millions of partitions and thus want to create hundreds of these grouped files.

Now take the '*.group*.fa' files and proceed to assembly. If you're looking for more resources on how we do this, check out these tutorials.

http://ged.msu.edu/angus/metag-assembly-2011/index.html http://ged.msu.edu/angus/nih-hmp-2012/index.html