Bcool_report

February 4, 2020

1 Data Source

URL: https://www.encodeproject.org/experiments/ENCSR749BAG/

```
wget -c https://www.encodeproject.org/files/ENCFF002EZY/@@download/ENCFF002EZY.fastq.gz
wget -c https://www.encodeproject.org/files/ENCFF002EZZ/@@download/ENCFF002EZZ.fastq.gz
```

Raw sequencing data

Isogenic replicate	Library	Accession	File type	Run type	Read	Lab	Date added	File size	Audit status	File status
1	ENCLB483KDW	ENCFF002EZY	fastq	PE101nt	1	Thomas Gingeras, CSHL	2014-07-31	6.44 GB	✔	released
		ENCFF002EZZ	fastq	PE101nt	2	Thomas Gingeras, CSHL	2014-07-31	6.36 GB	✔	released

2 Sample

[9]:
```
%%bash
bold=$(tput bold)
normal=$(tput sgr0)
echo "${bold}Read1:${normal}"
zcat data/ENCFF002EZY.fastq.gz | head -n 4
echo -e "\n${bold}-------------${normal}\n"
echo "${bold}Read2:${normal}"
zcat data/ENCFF002EZZ.fastq.gz | head -n 4
```

```
Read1:
@D2FC08P1:272:C4JVBACXX:5:1101:1415:1928 1:N:0:GTGTGT
NACCAACAGATTGGGAAAGGATCTTTACCAATCCTAAATCAGATAGGGGACTAATATCCAATATATATAAAGAACCCAAG
AAAGTGGACTCCAGAAAATCA
+
#1:DDDDDH;FHHIIIGGGIGIIIIIIIIIIIIIGIIIIGIIIIIIIIIID;FHHIIIIHIIIHIIIHHIIFHHEEE=A
BCCCACCCCCCCCCCCCCCCC

-------------

Read2:
```

```
@D2FC08P1:272:C4JVBACXX:5:1101:1415:1928 2:N:0:GTGTGT
GTTGGATCTCAGGGAAGTTTTGATTTGCATTTCCCTAATGATTAAGGATGCTGAACATTTTTTCAGGTGCTTCTCAGCCA
TTCAGTATTCCTAGGTGAGAA
+
@@BFFDFFHHHHHJEHHIGIJIEHGIIIIIJJJIII@GHGDGIGIIJEGIGHIJII@GIJJJJGHIJCHHHJIGHCCHHB
FFFFFEEEEEEC>@A@ACB>:
```

2.1 Stats

[10]:
```
%%bash
bold=$(tput bold)
normal=$(tput sgr0)
echo "${bold}ENCFF002EZY.fastq.gz Stats:${normal}"
seqkit stats data/ENCFF002EZY.fastq.gz
echo "${bold}ENCFF002EZZ.fastq.gz Stats:${normal}"
seqkit stats data/ENCFF002EZZ.fastq.gz
```

```
ENCFF002EZY.fastq.gz Stats:
file                        format  type    num_seqs        sum_len  min_len
avg_len  max_len
data/ENCFF002EZY.fastq.gz  FASTQ   DNA   79,763,453  8,056,108,753      101
101      101
ENCFF002EZZ.fastq.gz Stats:
Read2:
file                        format  type    num_seqs        sum_len  min_len
avg_len  max_len
data/ENCFF002EZZ.fastq.gz  FASTQ   DNA   79,763,453  8,056,108,753      101
101      101
```

2.2 Subsetting first 5m reads

```
seqkit head -n 5000000 data/ENCFF002EZY.fastq.gz | seqkit fq2fa -o ENCFF002EZY_5m.fa.gz
seqkit head -n 5000000 data/ENCFF002EZZ.fastq.gz | seqkit fq2fa -o ENCFF002EZZ_5m.fa.gz
```

2.3 Sample Stats

[12]:
```
%%bash
bold=$(tput bold)
normal=$(tput sgr0)
echo "${bold}ENCFF002EZY_5m.fa.gz Stats:${normal}"
seqkit stats ENCFF002EZY_5m.fa.gz
echo "${bold}ENCFF002EZZ_5m.fa.gz Stats:${normal}"
seqkit stats ENCFF002EZZ_5m.fa.gz
```

```
ENCFF002EZY_5m.fa.gz Stats:
file                  format  type   num_seqs      sum_len  min_len  avg_len
max_len
ENCFF002EZY_5m.fa.gz  FASTA   DNA   5,000,000  505,000,000      101      101
101
ENCFF002EZZ_5m.fa.gz Stats:
file                  format  type   num_seqs      sum_len  min_len  avg_len
max_len
ENCFF002EZZ_5m.fa.gz  FASTA   DNA   5,000,000  505,000,000      101      101
101
```

[14]:
```
%%bash
zcat ENCFF002EZY_5m.fa.gz | grep ">" | head
echo -e "\n--------------\n"
zcat ENCFF002EZZ_5m.fa.gz | grep ">" | head
```

```
>D2FC08P1:272:C4JVBACXX:5:1101:1415:1928 1:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1397:1937 1:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1462:1954 1:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1495:1984 1:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1708:1900 1:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1745:1902 1:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1608:1912 1:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1670:1917 1:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1711:1919 1:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1528:1940 1:N:0:GTGTGT

--------------

>D2FC08P1:272:C4JVBACXX:5:1101:1415:1928 2:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1397:1937 2:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1462:1954 2:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1495:1984 2:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1708:1900 2:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1745:1902 2:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1608:1912 2:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1670:1917 2:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1711:1919 2:N:0:GTGTGT
>D2FC08P1:272:C4JVBACXX:5:1101:1528:1940 2:N:0:GTGTGT
```

3 BCooooooool

```
conda install -c bioconda -y bcool
```

3.1 RUN #1 (k=25)

```
bcool -u merged_ENCFF002EZ_5m.fa -t 4 -k 25 -d 1 -o bcool_k25
```

-

3.2 Corrected reads: 291,098

- **Connected components (original reads k25):**
  - **Steps**:
    - ```
      bcalm -kmer-size 25 -in original_reads.fa -max-memory 10000 -out-dir cDBG_k25
      python convertToGFA.py original_reads.unitigs.fa original_reads_k25.GFA 25
      Bandage info original_reads_k25.GFA
      ```
    - **Connected components:** 649,720
    - **Largest component (bp):** 25,693,553
  - **Bandage**:
    - ```
      Connected components:             649,720
      Node count:                       1523382
      Edge count:                       1120807
      Smallest edge overlap (bp):       24
      Largest edge overlap (bp):        24
      Total length (bp):                108346839
      Total length no overlaps (bp):    86571039
      Dead ends:                        1616226
      Percentage dead ends:             53.0473%
      Largest component (bp):           25693553
      Total length orphaned nodes (bp): 25693553
      N50 (bp):                         101
      Shortest node (bp):               25
      Lower quartile node (bp):         28
      Median node (bp):                 49
      Upper quartile node (bp):         101
      Longest node (bp):                2483
      Median depth:                     2.28713
      Estimated sequence length (bp):   208933724
      ```
  - **Connected components (Corrected reads k25):**
  - **Steps**:

        * bcalm -kmer-size 25 -in reads_corrected.fa -max-memory 10000
          python convertToGFA.py reads_corrected.unitigs.fa reads_corrected_k25.GFA 25
          Bandage info reads_corrected_k25.GFA
        * **Connected components:** 646,104
        * **Largest component (bp):** 23,549,930
    - **Bandage**:
        *
```
Node count:                        1426893
Edge count:                        979372
Smallest edge overlap (bp):        24
Largest edge overlap (bp):         24
Total length (bp):                 105362219
Total length no overlaps (bp):     85815059
Dead ends:                         1577628
Percentage dead ends:              55.2819%
Connected components:              646104
Largest component (bp):            23549930
Total length orphaned nodes (bp): 23549930
N50 (bp):                          101
Shortest node (bp):                25
Lower quartile node (bp):          28
Median node (bp):                  49
Upper quartile node (bp):          101
Longest node (bp):                 2988
Median depth:                      2.28713
Estimated sequence length (bp):    221191836
```

3.3 RUN #2 (k=21)

```
bcool -u merged_ENCFF002EZ_5m.fa -t 4 -k 21 -d 1 -o bcool_k21
```

- **Corrected reads:** 348,434
- **Connected components (original reads k21):**
    - **Steps**:
        * bcalm -kmer-size 21 -in original_reads.fa -max-memory 10000
          python ../convertToGFA.py original_reads.unitigs.fa original_reads_k21.GFA 21
          Bandage info original_reads_k21.GFA
        * **Connected components:** 601,411
        * **Largest component (bp):** 46,548,499
    - **Bandage**:
        *
```
Node count:                        1910042
Edge count:                        1798172
Smallest edge overlap (bp):        20
Largest edge overlap (bp):         20
Total length (bp):                 112074285
Total length no overlaps (bp):     85242325
```

```
Dead ends:                          1653651
Percentage dead ends:               43.2883%
Connected components:               601411
Largest component (bp):             46548499
Total length orphaned nodes (bp): 46548499
N50 (bp):                           101
Shortest node (bp):                 21
Lower quartile node (bp):           22
Median node (bp):                   34
Upper quartile node (bp):           91
Longest node (bp):                  2479
Median depth:                       2.35
Estimated sequence length (bp):   213631714
```

- **Connected components (Corrected reads k21):**
- **Steps**:
  - `bcalm -kmer-size 21 -in reads_corrected.fa -max-memory 10000`
    `python convertToGFA.py reads_corrected.unitigs.fa reads_corrected_k21.GFA 21`
    `Bandage info reads_corrected_k21.GFA`
  - **Connected components:** 597,025
  - **Largest component (bp):** 43,406,980
- **Bandage**:
  -
```
Node count:                         1783239
Edge count:                         1615510
Smallest edge overlap (bp):         20
Largest edge overlap (bp):          20
Total length (bp):                  108809811
Total length no overlaps (bp):      84424811
Dead ends:                          1606712
Percentage dead ends:               45.0504%
Connected components:               597025
Largest component (bp):             43406980
Total length orphaned nodes (bp): 43406980
N50 (bp):                           101
Shortest node (bp):                 21
Lower quartile node (bp):           22
Median node (bp):                   36
Upper quartile node (bp):           99
Longest node (bp):                  2479
Median depth:                       2.37342
Estimated sequence length (bp):   226854509
```

RUN #3 (k=91 = 101-10)

PROMISING

`bcool -u merged_ENCFF002EZ_5m.fa -t 4 -k 91 -d 1 -o bcool_k91`

- **Corrected reads:** 228,914
- **Connected components (original reads k91):**
  - **Steps**:

    * bcalm -kmer-size 91 -in original_reads.fa -max-memory 10000
      python convertToGFA.py original_reads.unitigs.fa original_reads_k91.GFA 91
      Bandage info original_reads_k91.GFA
    * **Connected components:** 1,214,311
    * **Largest component (bp):** 478,977
- **Bandage**:
    * Node count: 1307157
      Edge count: 96392
      Smallest edge overlap (bp): 90
      Largest edge overlap (bp): 90
      Total length (bp): 133631258
      Total length no overlaps (bp): 123270998
      Dead ends: 2477639
      Percentage dead ends: 94.7721%
      Connected components: 1214311
      Largest component (bp): 478977
      Total length orphaned nodes (bp): 478977
      N50 (bp): 101
      Shortest node (bp): 91
      Lower quartile node (bp): 101
      Median node (bp): 101
      Upper quartile node (bp): 101
      Longest node (bp): 657
      Median depth: 0.267327
      Estimated sequence length (bp): 230522548
- **Connected components (Corrected reads k91):**
- **Steps**:
    * bcalm -kmer-size 91 -in reads_corrected.fa -max-memory 10000
      python convertToGFA.py reads_corrected.unitigs.fa reads_corrected_k91.GFA 91
      Bandage info reads_corrected_k91.GFA
    * **Connected components:** 1,201,258
    * **Largest component (bp):** 116,657
- **Bandage**:
    * Node count: 1276295
      Edge count: 75895
      Smallest edge overlap (bp): 90
      Largest edge overlap (bp): 90
      Total length (bp): 130680244
      Total length no overlaps (bp): 121932604
      Dead ends: 2440951
      Percentage dead ends: 95.6264%
      Connected components: 1201258
      Largest component (bp): 116657
      Total length orphaned nodes (bp): 116657
      N50 (bp): 101
      Shortest node (bp): 91
      Lower quartile node (bp): 101
      Median node (bp): 101

```
Upper quartile node (bp):          101
Longest node (bp):                 5462
Median depth:                      0.271845
Estimated sequence length (bp):    251704280
```

4 Summary