Sequencing the *Phaseolus vulgaris* G19833 Genome

Scott Jackson, Purdue University

Phil McClean, North Dakota State University

Perry Cregan, USDA-ARS, Beltsville

Dan Rokhsar, DOE Joint Genome Institute

Jeremy Schmutz, HudsonAlpha Institute

David Hyten, USDA-ARS, Beltsville

Project overview

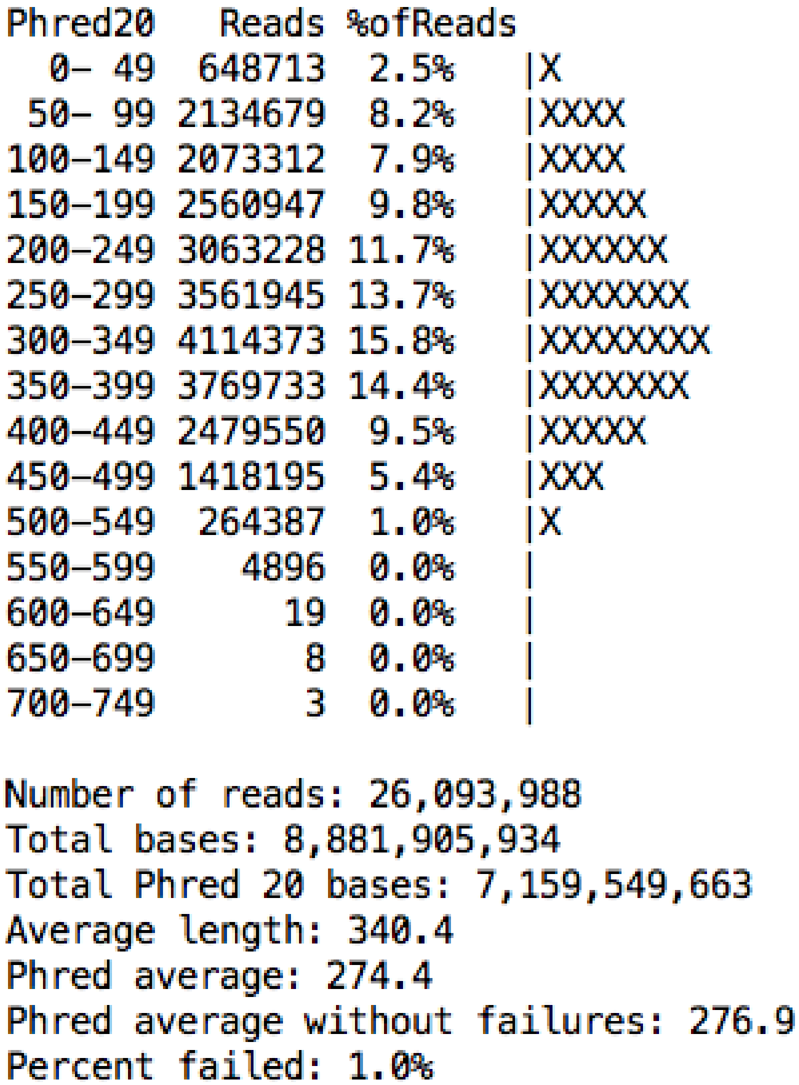
- Produce as good as possible reference sequence for PV with available funds that will serve as a directly useful tool for common bean improvement (when combined with ongoing Bean-CAP efforts) and a reference comparator for soybean
- The DOE Joint Genome Institute is producing a 454/Illumina based draft sequence for soybean comparative work
- USDA has funded the project "**A sequence map of the common bean genome for bean improvement"** to add longer range linking data and genetic mapping data to the draft in order to produce a reference PV genome sequence
- The overlap between the DOE-JGI portion and the USDA portion is coordinated by Jeremy Schmutz at HudsonAlpha, who is also funded as part of the JGI Plant Genome Program

Expected genome data for PV V1.0

- 12x 454 linear data
- 4-6x 454 paired data from multiple sized libraries (4kb and 8kb)
- 50x of Illumina data
- 200k fosmid Sanger sequenced pairs
- 80k BAC Sanger sequenced pairs (includes 40k from BeanMap efforts)
- Up to 1400 marker genetic map (Hyten & Cregan)
- BAC physical map (phaseolus.genomics.purdue.edu)

454 linear sequencing

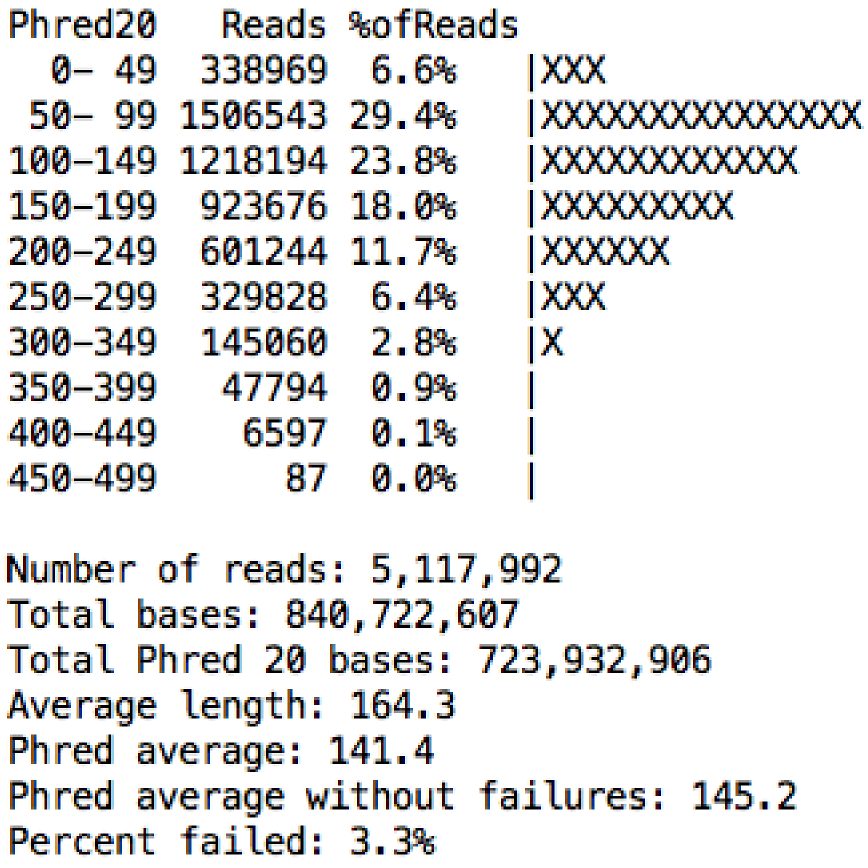
- Collected 7.1 GB of linear data (~11x HQ bases)
- Includes 26 runs from 4 libraries
- Average HQ read length of 274 bps



```
Phred20    Reads %ofReads
  0- 49   648713  2.5%    |X
 50- 99  2134679  8.2%    |XXXX
100-149  2073312  7.9%    |XXXX
150-199  2560947  9.8%    |XXXXX
200-249  3063228 11.7%    |XXXXXX
250-299  3561945 13.7%    |XXXXXXX
300-349  4114373 15.8%    |XXXXXXXX
350-399  3769733 14.4%    |XXXXXXX
400-449  2479550  9.5%    |XXXXX
450-499  1418195  5.4%    |XXX
500-549   264387  1.0%    |X
550-599     4896  0.0%    |
600-649       19  0.0%    |
650-699        8  0.0%    |
700-749        3  0.0%    |

Number of reads: 26,093,988
Total bases: 8,881,905,934
Total Phred 20 bases: 7,159,549,663
Average length: 340.4
Phred average: 274.4
Phred average without failures: 276.9
Percent failed: 1.0%
```

454 paired sequencing

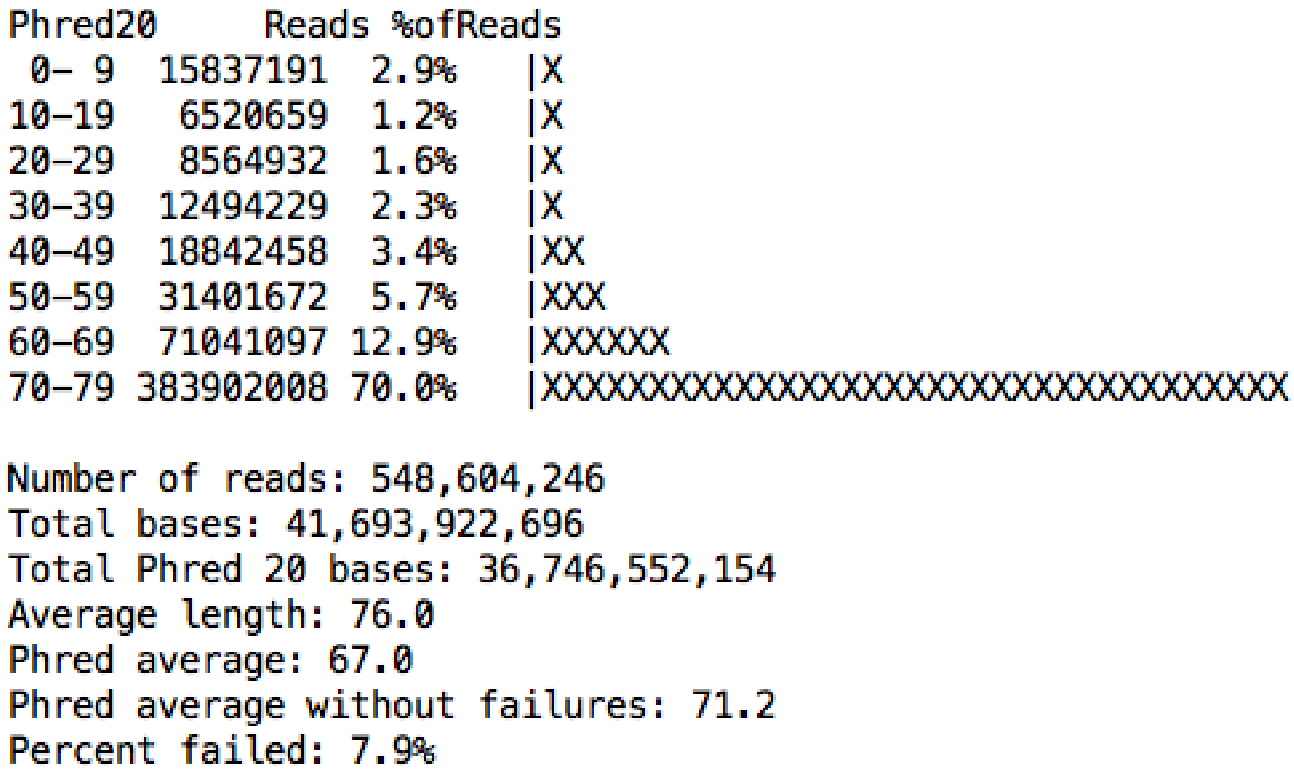
- Collected 3.4 Gb of total HQ raw data
- Includes 14 runs from 8 libraries
- Once true pairs are determined we only collect 724 Mb or ~1.1x of paired data
- Making good paired libraries has been the most difficult part of this project
- We have 4 more library attempts in progress



```
Phred20   Reads %ofReads
  0- 49  338969  6.6%    |XXX
 50- 99 1506543 29.4%    |XXXXXXXXXXXXXXX
100-149 1218194 23.8%    |XXXXXXXXXXXX
150-199  923676 18.0%    |XXXXXXXXX
200-249  601244 11.7%    |XXXXXX
250-299  329828  6.4%    |XXX
300-349  145060  2.8%    |X
350-399   47794  0.9%    |
400-449    6597  0.1%    |
450-499      87  0.0%    |

Number of reads: 5,117,992
Total bases: 840,722,607
Total Phred 20 bases: 723,932,906
Average length: 164.3
Phred average: 141.4
Phred average without failures: 145.2
Percent failed: 3.3%
```

Illumina sequencing

- We have collected several different GA2x flow cells for PV, most of the reads are paired 76mers that will be used to correct 454 read errors



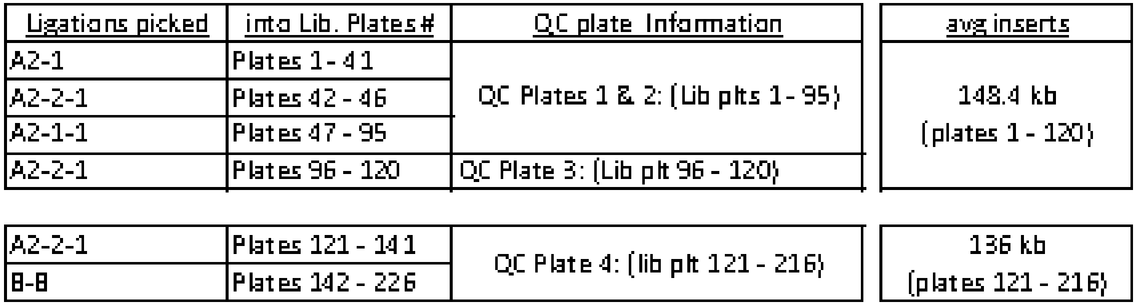
```
Phred20      Reads %ofReads
 0- 9  15837191  2.9%    |X
10-19   6520659  1.2%    |X
20-29   8564932  1.6%    |X
30-39  12494229  2.3%    |X
40-49  18842458  3.4%    |XX
50-59  31401672  5.7%    |XXX
60-69  71041097 12.9%    |XXXXXX
70-79 383902008 70.0%    |XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Number of reads: 548,604,246
Total bases: 41,693,922,696
Total Phred 20 bases: 36,746,552,154
Average length: 76.0
Phred average: 67.0
Phred average without failures: 71.2
Percent failed: 7.9%
```

Fosmid libraries

- Goal is to produce about 400k reads or 200k pairs from fosmid length libraries
- 2 libraries have been constructed
  - VUL 33,465 bp insert; 200 plates
  - VUK 35,558 bp insert; 260 plates
- We have sequenced a combined 152k reads, 100k more are in progress now at HudsonAlpha
- We are sequencing more BES from the new BAC library to replace some of the FES to beef up long range linking

New BAC library

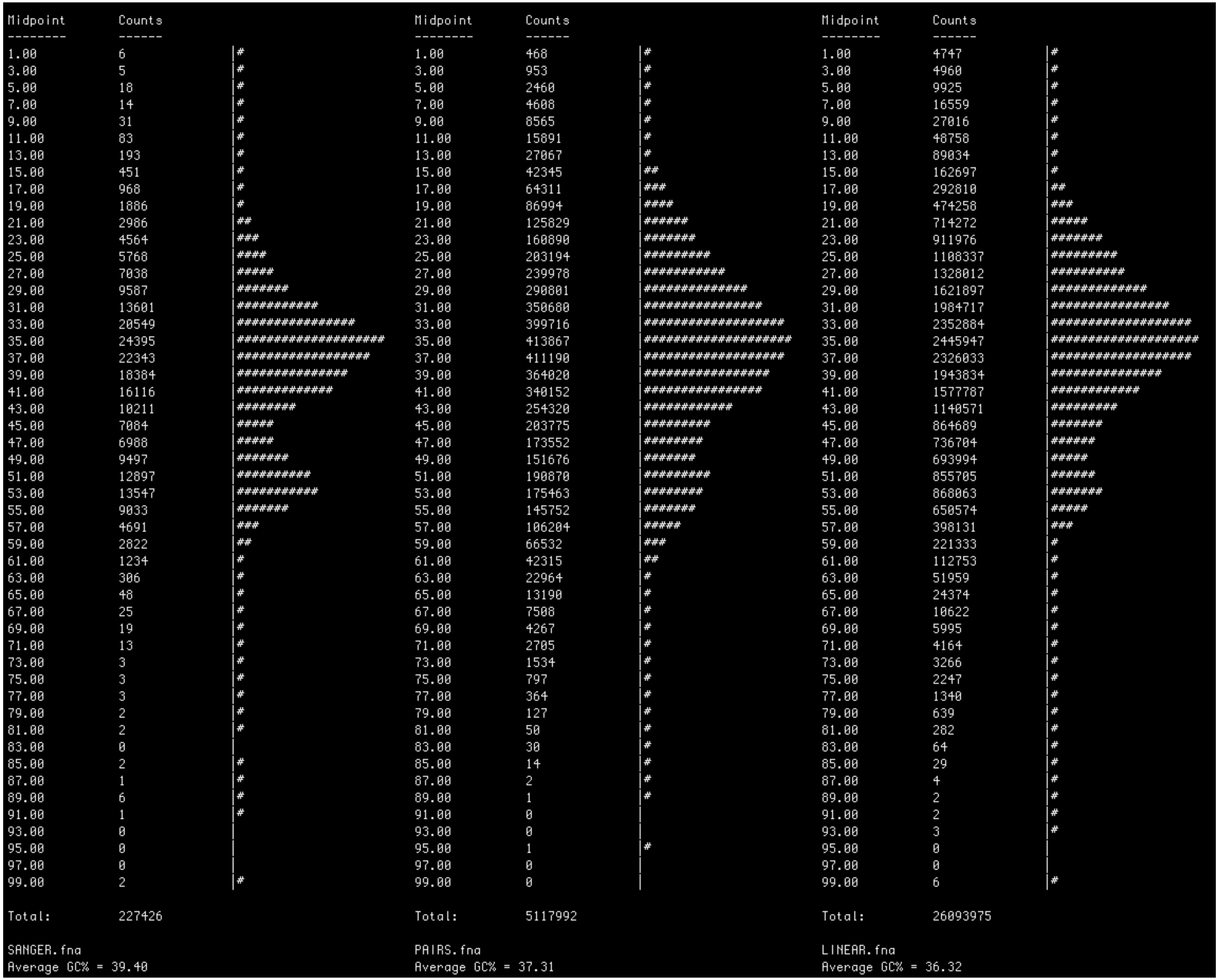
- Dave Kudrna at Arizona Genomics created a new BAC library called PV_GBb with the goal of maximizing the insert length
- The first section is 120 plates with 149kb average
- The second section is 105 plates with a 136kb insert
- We are sequencing the entire first section and filling in additional inserts from the second section
- These will add to the 89K reads from PV_Gba and give us decent BAC coverage across the genome



Ligations picked	into Lib. Plates #	QC plate Information	avg inserts
A2-1	Plates 1 - 41	QC Plates 1 & 2: (Lib plts 1 - 95)	148.4 kb (plates 1 - 120)
A2-2-1	Plates 42 - 46		
A2-1-1	Plates 47 - 95		
A2-2-1	Plates 96 - 120	QC Plate 3: (Lib plt 96 - 120)	

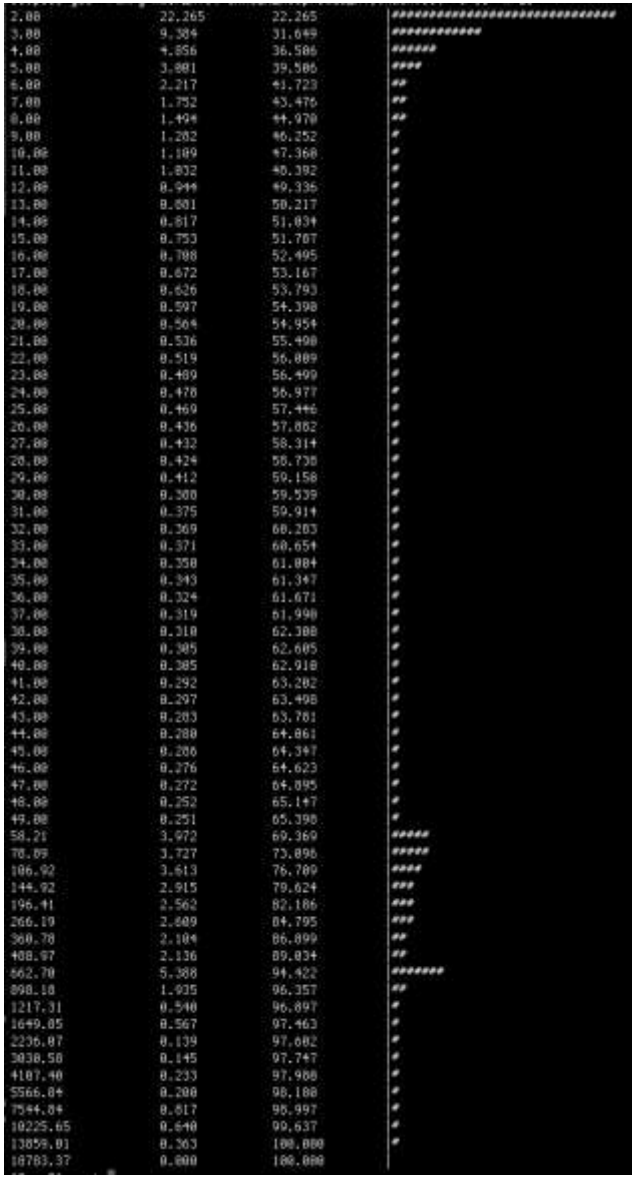
A2-2-1	Plates 121 - 141	QC Plate 4: (lib plt 121 - 216)	136 kb (plates 121 - 216)
B-B	Plates 142 - 226		

GC profiles of data sets



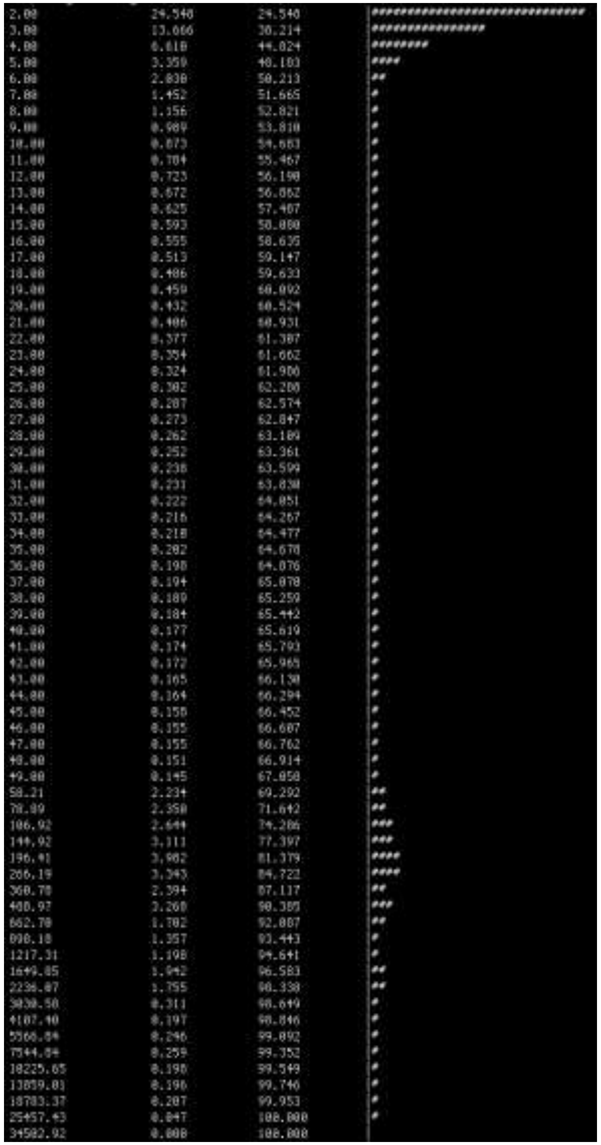
```
Midpoint       Counts                                   Midpoint       Counts                                   Midpoint       Counts
--------       ------                                   --------       ------                                   --------       ------
1.00           6               |#                       1.00           468             |#                       1.00           4747            |#
3.00           5               |#                       3.00           953             |#                       3.00           4960            |#
5.00           18              |#                       5.00           2460            |#                       5.00           9925            |#
7.00           14              |#                       7.00           4608            |#                       7.00           16559           |#
9.00           31              |#                       9.00           8565            |#                       9.00           27016           |#
11.00          83              |#                       11.00          15891           |#                       11.00          48758           |#
13.00          193             |#                       13.00          27067           |#                       13.00          89034           |#
15.00          451             |#                       15.00          42345           |##                      15.00          162697          |#
17.00          968             |#                       17.00          64311           |###                     17.00          292810          |##
19.00          1886            |#                       19.00          86994           |####                    19.00          474258          |###
21.00          2986            |##                      21.00          125829          |######                  21.00          714272          |#####
23.00          4564            |###                     23.00          160890          |#######                 23.00          911976          |#######
25.00          5768            |####                    25.00          203194          |#########               25.00          1108337         |#########
27.00          7038            |#####                   27.00          239978          |###########             27.00          1328012         |##########
29.00          9587            |#######                 29.00          290801          |##############          29.00          1621897         |#############
31.00          13601           |###########             31.00          350680          |################        31.00          1984717         |################
33.00          20549           |################        33.00          399716          |###################     33.00          2352884         |###################
35.00          24395           |####################    35.00          413867          |####################    35.00          2445947         |####################
37.00          22343           |##################      37.00          411190          |###################     37.00          2326033         |###################
39.00          18384           |###############         39.00          364020          |#################       39.00          1943834         |###############
41.00          16116           |#############           41.00          340152          |################        41.00          1577787         |############
43.00          10211           |########                43.00          254320          |############            43.00          1140571         |#########
45.00          7084            |#####                   45.00          203775          |#########               45.00          864689          |#######
47.00          6988            |#####                   47.00          173552          |########                47.00          736704          |######
49.00          9497            |#######                 49.00          151676          |#######                 49.00          693994          |#####
51.00          12897           |##########              51.00          190870          |#########               51.00          855705          |######
53.00          13547           |###########             53.00          175463          |########                53.00          868063          |#######
55.00          9033            |#######                 55.00          145752          |#######                 55.00          650574          |#####
57.00          4691            |###                     57.00          106204          |#####                   57.00          398131          |###
59.00          2822            |##                      59.00          66532           |###                     59.00          221333          |#
61.00          1234            |#                       61.00          42315           |##                      61.00          112753          |#
63.00          306             |#                       63.00          22964           |#                       63.00          51959           |#
65.00          48              |#                       65.00          13190           |#                       65.00          24374           |#
67.00          25              |#                       67.00          7508            |#                       67.00          10622           |#
69.00          19              |#                       69.00          4267            |#                       69.00          5995            |#
71.00          13              |#                       71.00          2705            |#                       71.00          4164            |#
73.00          3               |#                       73.00          1534            |#                       73.00          3266            |#
75.00          3               |#                       75.00          797             |#                       75.00          2247            |#
77.00          3               |#                       77.00          364             |#                       77.00          1340            |#
79.00          2               |#                       79.00          127             |#                       79.00          639             |#
81.00          2               |#                       81.00          50              |#                       81.00          282             |#
83.00          0               |                        83.00          30              |#                       83.00          64              |#
85.00          2               |#                       85.00          14              |#                       85.00          29              |#
87.00          1               |#                       87.00          2               |#                       87.00          4               |#
89.00          6               |#                       89.00          1               |#                       89.00          2               |#
91.00          1               |#                       91.00          0               |                        91.00          2               |#
93.00          0               |                        93.00          0               |                        93.00          3               |#
95.00          0               |                        95.00          1               |#                       95.00          0               |
97.00          0               |                        97.00          0               |                        97.00          0               |
99.00          2               |#                       99.00          0               |                        99.00          6               |#

Total:         227426                                   Total:         5117992                                  Total:         26093975

SANGER.fna                                              PAIRS.fna                                               LINEAR.fna
Average GC% = 39.40                                     Average GC% = 37.31                                     Average GC% = 36.32
```

Kmer distributions



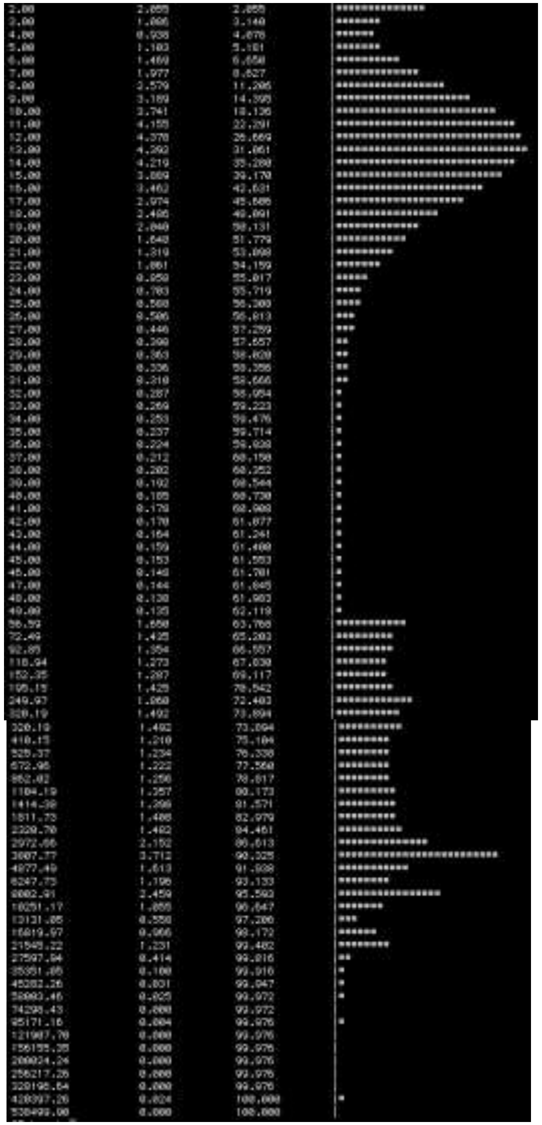
```
2.00          22.265     22.265     |###############################
3.00          9.384      31.649     |############
4.00          4.856      36.506     |######
5.00          3.001      39.506     |####
6.00          2.217      41.723     |##
7.00          1.752      43.476     |##
8.00          1.494      44.970     |##
9.00          1.282      46.252     |#
10.00         1.109      47.360     |#
11.00         1.032      48.392     |#
12.00         0.944      49.336     |#
13.00         0.881      50.217     |#
14.00         0.817      51.034     |#
15.00         0.753      51.787     |#
16.00         0.708      52.495     |#
17.00         0.672      53.167     |#
18.00         0.626      53.793     |#
19.00         0.597      54.390     |#
20.00         0.564      54.954     |#
21.00         0.536      55.490     |#
22.00         0.519      56.009     |#
23.00         0.489      56.499     |#
24.00         0.478      56.977     |#
25.00         0.469      57.446     |#
26.00         0.436      57.882     |#
27.00         0.432      58.314     |#
28.00         0.424      58.738     |#
29.00         0.412      59.150     |#
30.00         0.389      59.539     |#
31.00         0.375      59.914     |#
32.00         0.369      60.283     |#
33.00         0.371      60.654     |#
34.00         0.350      61.004     |#
35.00         0.343      61.347     |#
36.00         0.324      61.671     |#
37.00         0.319      61.990     |#
38.00         0.310      62.300     |#
39.00         0.305      62.605     |#
40.00         0.305      62.910     |#
41.00         0.292      63.202     |#
42.00         0.297      63.498     |#
43.00         0.283      63.781     |#
44.00         0.280      64.061     |#
45.00         0.286      64.347     |#
46.00         0.276      64.623     |#
47.00         0.272      64.895     |#
48.00         0.252      65.147     |#
49.00         0.251      65.398     |#
58.21         3.972      69.369     |#####
78.89         3.727      73.096     |#####
106.92        3.613      76.709     |####
144.92        2.915      79.624     |###
196.41        2.562      82.186     |###
266.19        2.609      84.795     |###
360.78        2.104      86.899     |##
488.97        2.136      89.034     |##
662.78        5.388      94.422     |#######
898.18        1.935      96.357     |##
1217.31       0.540      96.897     |#
1649.85       0.567      97.463     |#
2236.87       0.139      97.602     |#
3030.58       0.145      97.747     |#
4107.40       0.233      97.980     |#
5566.84       0.208      98.188     |#
7544.84       0.817      98.997     |#
10225.65      0.640      99.637     |#
13859.81      0.363      100.000    |#
18783.37      0.000      100.000    |
```

Sanger



```
2.00          24.540     24.540     |##############################
3.00          13.666     38.214     |################
4.00          6.610      44.824     |########
5.00          3.359      48.183     |####
6.00          2.030      50.213     |##
7.00          1.452      51.665     |#
8.00          1.156      52.821     |#
9.00          0.989      53.810     |#
10.00         0.873      54.683     |#
11.00         0.784      55.467     |#
12.00         0.723      56.190     |#
13.00         0.672      56.862     |#
14.00         0.625      57.487     |#
15.00         0.593      58.080     |#
16.00         0.555      58.635     |#
17.00         0.513      59.147     |#
18.00         0.486      59.633     |#
19.00         0.459      60.092     |#
20.00         0.432      60.524     |#
21.00         0.406      60.931     |#
22.00         0.377      61.307     |#
23.00         0.354      61.662     |#
24.00         0.324      61.986     |#
25.00         0.302      62.288     |#
26.00         0.287      62.574     |#
27.00         0.273      62.847     |#
28.00         0.262      63.109     |#
29.00         0.252      63.361     |#
30.00         0.238      63.599     |#
31.00         0.231      63.830     |#
32.00         0.222      64.051     |#
33.00         0.216      64.267     |#
34.00         0.210      64.477     |#
35.00         0.202      64.678     |#
36.00         0.198      64.876     |#
37.00         0.194      65.070     |#
38.00         0.189      65.259     |#
39.00         0.184      65.442     |#
40.00         0.177      65.619     |#
41.00         0.174      65.793     |#
42.00         0.172      65.965     |#
43.00         0.165      66.130     |#
44.00         0.164      66.294     |#
45.00         0.158      66.452     |#
46.00         0.155      66.607     |#
47.00         0.155      66.762     |#
48.00         0.151      66.914     |#
49.00         0.145      67.058     |#
58.21         2.234      69.292     |##
78.89         2.350      71.642     |##
106.92        2.644      74.286     |###
144.92        3.111      77.397     |###
196.41        3.982      81.379     |####
266.19        3.343      84.722     |####
360.78        2.394      87.117     |##
488.97        3.268      90.385     |###
662.78        1.702      92.087     |##
898.18        1.357      93.443     |#
1217.31       1.198      94.641     |#
1649.85       1.942      96.583     |##
2236.87       1.755      98.338     |##
3030.58       0.311      98.649     |#
4107.40       0.197      98.846     |#
5566.84       0.246      99.092     |#
7544.84       0.259      99.352     |#
10225.65      0.198      99.549     |#
13859.81      0.196      99.746     |#
18783.37      0.207      99.953     |#
25457.43      0.047      100.000    |#
34582.92      0.000      100.000    |
```

454 Pairs

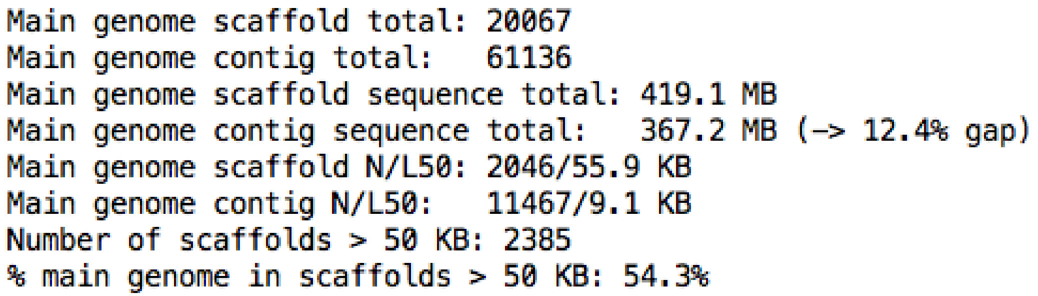


```
2.00          2.055      2.055      |**************
3.00          1.086      3.140      |*******
4.00          0.938      4.078      |******
5.00          1.103      5.181      |*******
6.00          1.469      6.650      |**********
7.00          1.977      8.627      |*************
8.00          2.579      11.206     |*****************
9.00          3.189      14.395     |*********************
10.00         3.741      18.136     |*************************
11.00         4.155      22.291     |****************************
12.00         4.378      26.669     |*****************************
13.00         4.392      31.061     |******************************
14.00         4.219      35.280     |****************************
15.00         3.889      39.170     |**************************
16.00         3.462      42.631     |***********************
17.00         2.974      45.606     |********************
18.00         2.486      48.091     |****************
19.00         2.040      50.131     |*************
20.00         1.648      51.779     |***********
21.00         1.319      53.098     |*********
22.00         1.061      54.159     |*******
23.00         0.858      55.017     |*****
24.00         0.703      55.719     |****
25.00         0.580      56.300     |****
26.00         0.506      56.813     |***
27.00         0.446      57.259     |***
28.00         0.398      57.657     |**
29.00         0.363      58.020     |**
30.00         0.336      58.356     |**
31.00         0.310      58.666     |**
32.00         0.287      58.954     |*
33.00         0.269      59.223     |*
34.00         0.253      59.476     |*
35.00         0.237      59.714     |*
36.00         0.224      59.938     |*
37.00         0.212      60.150     |*
38.00         0.202      60.352     |*
39.00         0.192      60.544     |*
40.00         0.185      60.730     |*
41.00         0.178      60.908     |*
42.00         0.170      61.077     |*
43.00         0.164      61.241     |*
44.00         0.159      61.400     |*
45.00         0.153      61.553     |*
46.00         0.148      61.701     |*
47.00         0.144      61.845     |*
48.00         0.138      61.983     |*
49.00         0.135      62.119     |*
56.59         1.650      63.768     |***********
72.49         1.435      65.203     |*********
92.85         1.354      66.557     |*********
118.94        1.273      67.830     |********
152.35        1.287      69.117     |********
195.15        1.425      70.542     |*********
249.97        1.860      72.403     |************
320.19        1.492      73.894     |**********
320.19        1.492      73.894     |**********
410.15        1.210      75.104     |********
528.37        1.234      76.338     |********
672.96        1.222      77.560     |********
862.02        1.256      78.817     |********
1104.19       1.357      80.173     |*********
1414.38       1.398      81.571     |*********
1811.73       1.408      82.979     |*********
2320.70       1.482      84.461     |**********
2972.66       2.152      86.613     |**************
3807.77       3.712      90.325     |**************************
4877.49       1.613      91.938     |***********
6247.75       1.196      93.133     |********
8002.91       2.459      95.592     |*****************
10251.17      1.055      96.647     |*******
13131.05      0.558      97.206     |***
16819.97      0.966      98.172     |******
21545.22      1.231      99.402     |********
27597.94      0.414      99.816     |**
35351.05      0.100      99.916     |*
45282.26      0.031      99.947     |*
58003.46      0.025      99.972     |*
74298.43      0.000      99.972     |
95171.16      0.004      99.976     |*
121907.70     0.000      99.976     |
156155.35     0.000      99.976     |
200024.24     0.000      99.976     |
256217.26     0.000      99.976     |
328190.64     0.000      99.976     |
420397.26     0.024      100.000    |*
538499.00     0.000      100.000    |
```

454 Linear

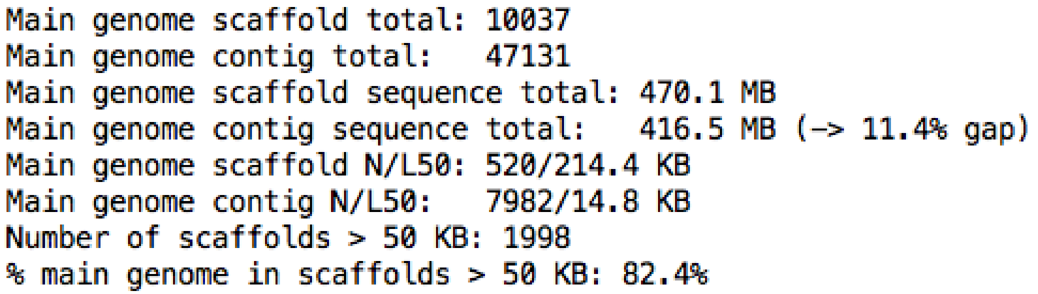
Initial Newbler assemblies 454 only

- ~8x assembly



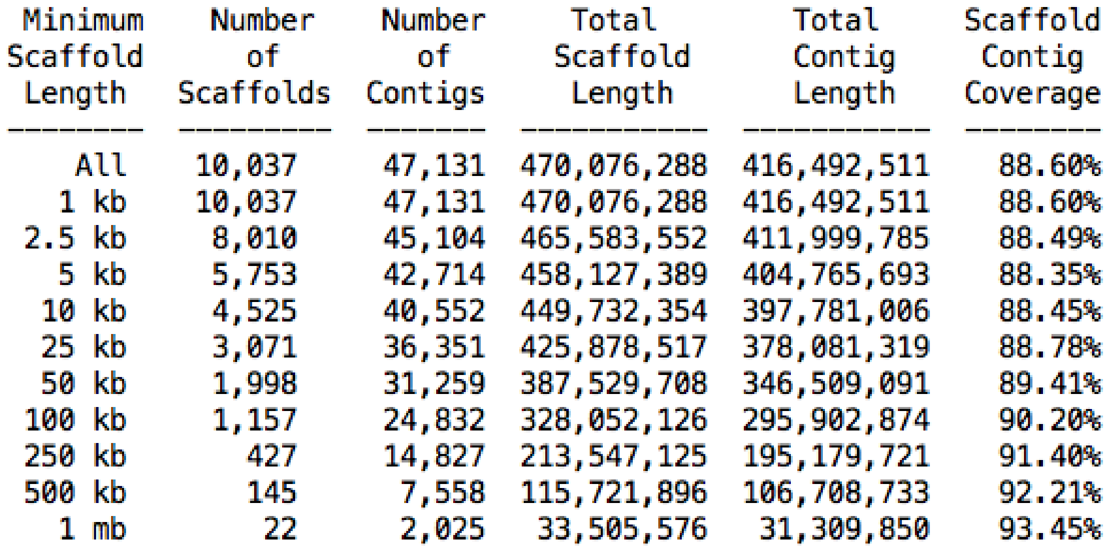
```
Main genome scaffold total: 20067
Main genome contig total:   61136
Main genome scaffold sequence total: 419.1 MB
Main genome contig sequence total:   367.2 MB (-> 12.4% gap)
Main genome scaffold N/L50: 2046/55.9 KB
Main genome contig N/L50:   11467/9.1 KB
Number of scaffolds > 50 KB: 2385
% main genome in scaffolds > 50 KB: 54.3%
```

- ~12x assembly



```
Main genome scaffold total: 10037
Main genome contig total:   47131
Main genome scaffold sequence total: 470.1 MB
Main genome contig sequence total:   416.5 MB (-> 11.4% gap)
Main genome scaffold N/L50: 520/214.4 KB
Main genome contig N/L50:   7982/14.8 KB
Number of scaffolds > 50 KB: 1998
% main genome in scaffolds > 50 KB: 82.4%
```

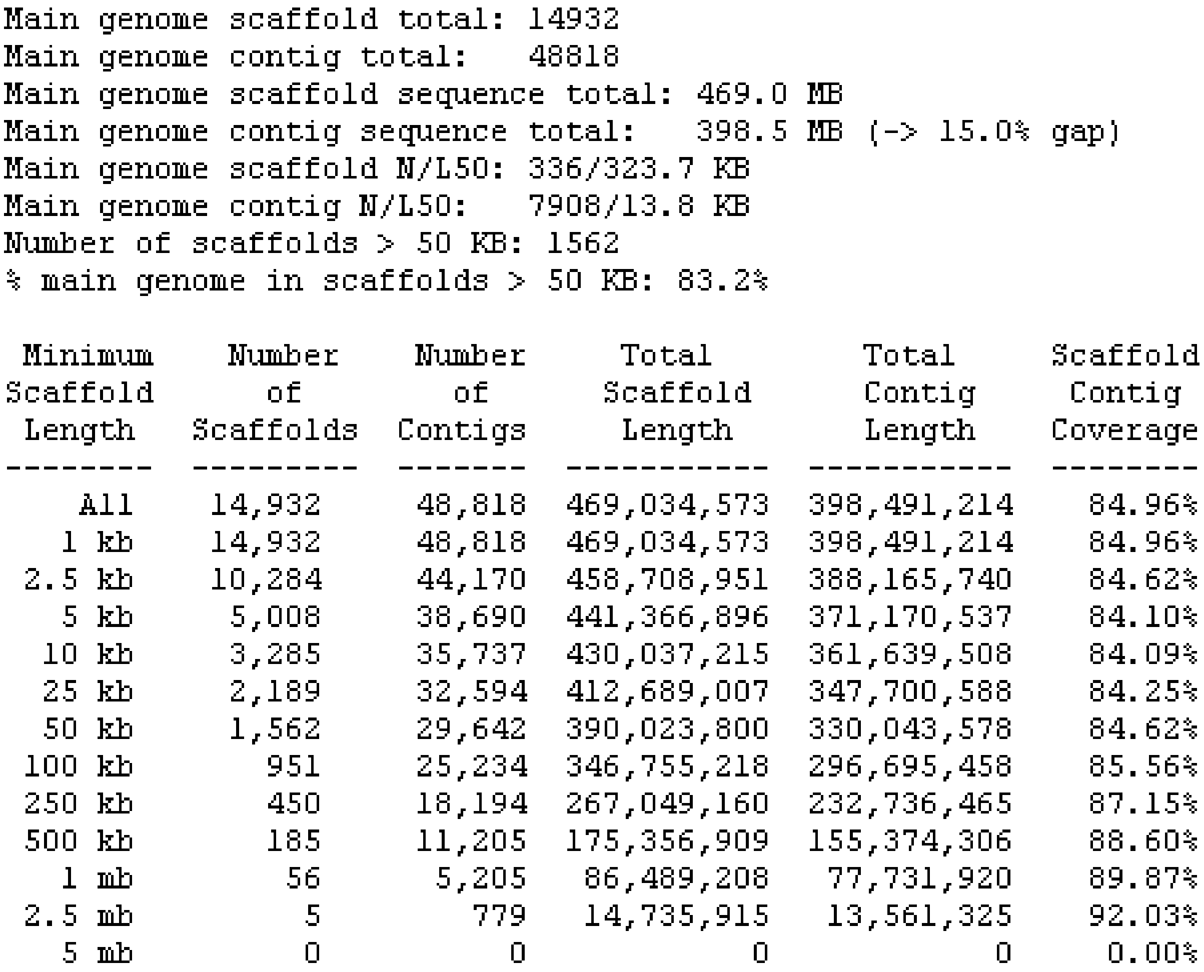
- These builds are the basis for the genetic mapping effort



```
 Minimum     Number     Number       Total          Total       Scaffold
Scaffold       of         of       Scaffold        Contig        Contig
 Length     Scaffolds  Contigs      Length         Length       Coverage
--------    ---------  -------   -----------    -----------     --------
    All      10,037     47,131   470,076,288    416,492,511      88.60%
   1 kb      10,037     47,131   470,076,288    416,492,511      88.60%
 2.5 kb       8,010     45,104   465,583,552    411,999,785      88.49%
   5 kb       5,753     42,714   458,127,389    404,765,693      88.35%
  10 kb       4,525     40,552   449,732,354    397,781,006      88.45%
  25 kb       3,071     36,351   425,878,517    378,081,319      88.78%
  50 kb       1,998     31,259   387,529,708    346,509,091      89.41%
 100 kb       1,157     24,832   328,052,126    295,902,874      90.20%
 250 kb         427     14,827   213,547,125    195,179,721      91.40%
 500 kb         145      7,558   115,721,896    106,708,733      92.21%
   1 mb          22      2,025    33,505,576     31,309,850      93.45%
```

Latest Newbler assembly

- We are running several version of the compete data set to date, including BES and FES



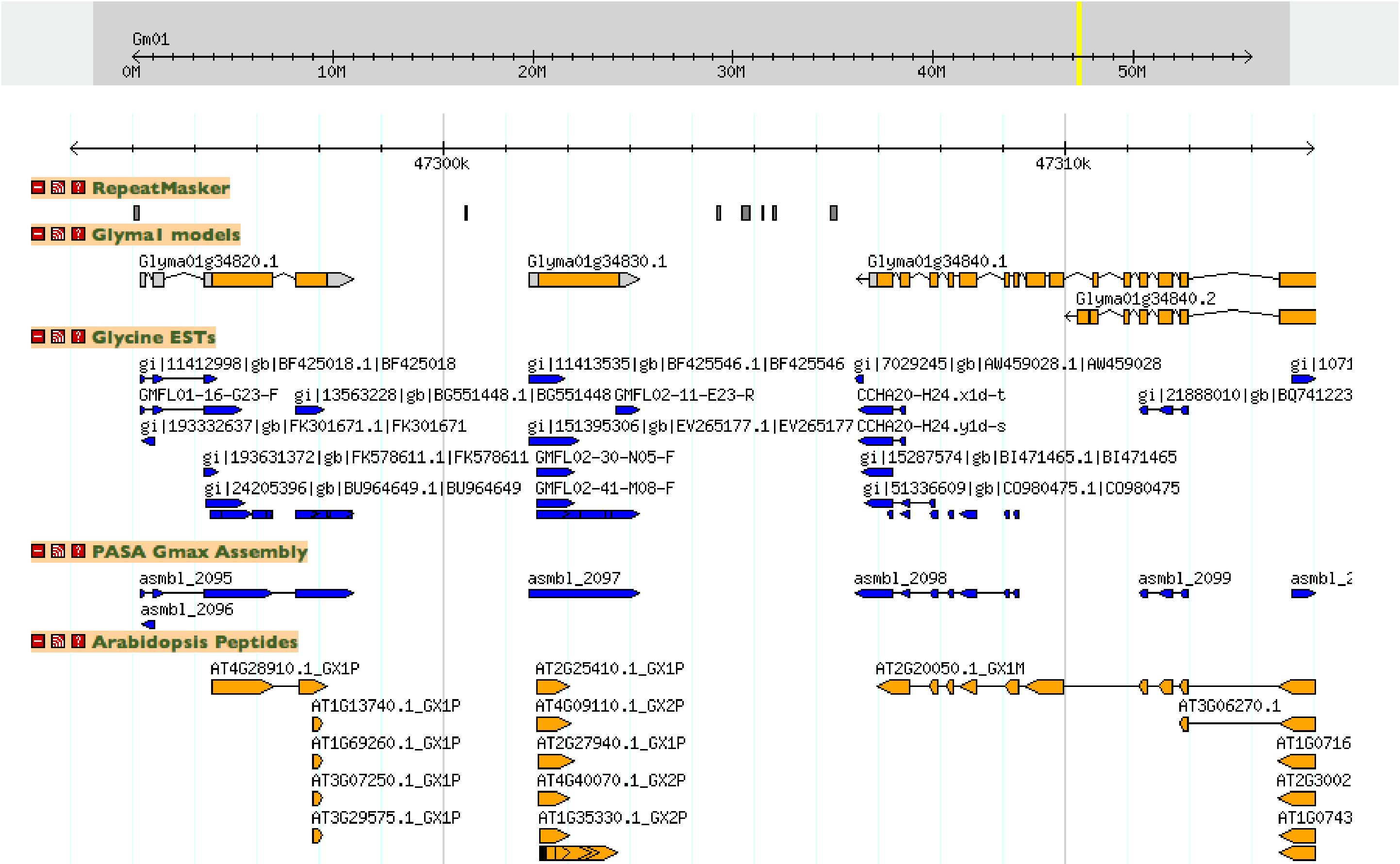
```
Main genome scaffold total: 14932
Main genome contig total:   48818
Main genome scaffold sequence total: 469.0 MB
Main genome contig sequence total:   398.5 MB (-> 15.0% gap)
Main genome scaffold N/L50: 336/323.7 KB
Main genome contig N/L50:   7908/13.8 KB
Number of scaffolds > 50 KB: 1562
% main genome in scaffolds > 50 KB: 83.2%

 Minimum    Number     Number     Total         Total       Scaffold
Scaffold      of         of      Scaffold       Contig       Contig
 Length    Scaffolds  Contigs     Length        Length      Coverage
--------   ---------  -------  -----------   -----------   --------
    All     14,932     48,818  469,034,573   398,491,214    84.96%
   1 kb     14,932     48,818  469,034,573   398,491,214    84.96%
 2.5 kb     10,284     44,170  458,708,951   388,165,740    84.62%
   5 kb      5,008     38,690  441,366,896   371,170,537    84.10%
  10 kb      3,285     35,737  430,037,215   361,639,508    84.09%
  25 kb      2,189     32,594  412,689,007   347,700,588    84.25%
  50 kb      1,562     29,642  390,023,800   330,043,578    84.62%
 100 kb        951     25,234  346,755,218   296,695,458    85.56%
 250 kb        450     18,194  267,049,160   232,736,465    87.15%
 500 kb        185     11,205  175,356,909   155,374,306    88.60%
   1 mb         56      5,205   86,489,208    77,731,920    89.87%
 2.5 mb          5        779   14,735,915    13,561,325    92.03%
   5 mb          0          0            0             0     0.00%
```

New assembly pipeline

- Remove duplicate 454 pairs and unpaired reads
- Correct 454 bp and indel errors using Illumina reads
- Assemble using HA Arachne version including BES and FES and all of the 454 corrected sequence data
- Make scaffold breaks on misjoins by comparing to genetic map
- Order and orientate scaffolds based on genetic map and build chromosome scale assembly
- This same pipeline is being used to produce multiple plant genomes for the JGI Plant Genome Program

Annotation

- PV will be annotated using the same Phytozome pipeline as soybean, our goal is to have a comparable gene annotation between the two genomes



(Mitros, Shu, Goodstein, Rokhsar)

Timeline goals

- April 2011: Data collection complete
- July 2011: Chromosome scale assembly
- September 2011: Annotation complete and V1 PV publicly available