QC on DNAscent and other analyses from nanopore data

Sathish Thiyagarajan
sathish.thiyagarajan@earlham.ac.uk

October 3, 2024

Contents

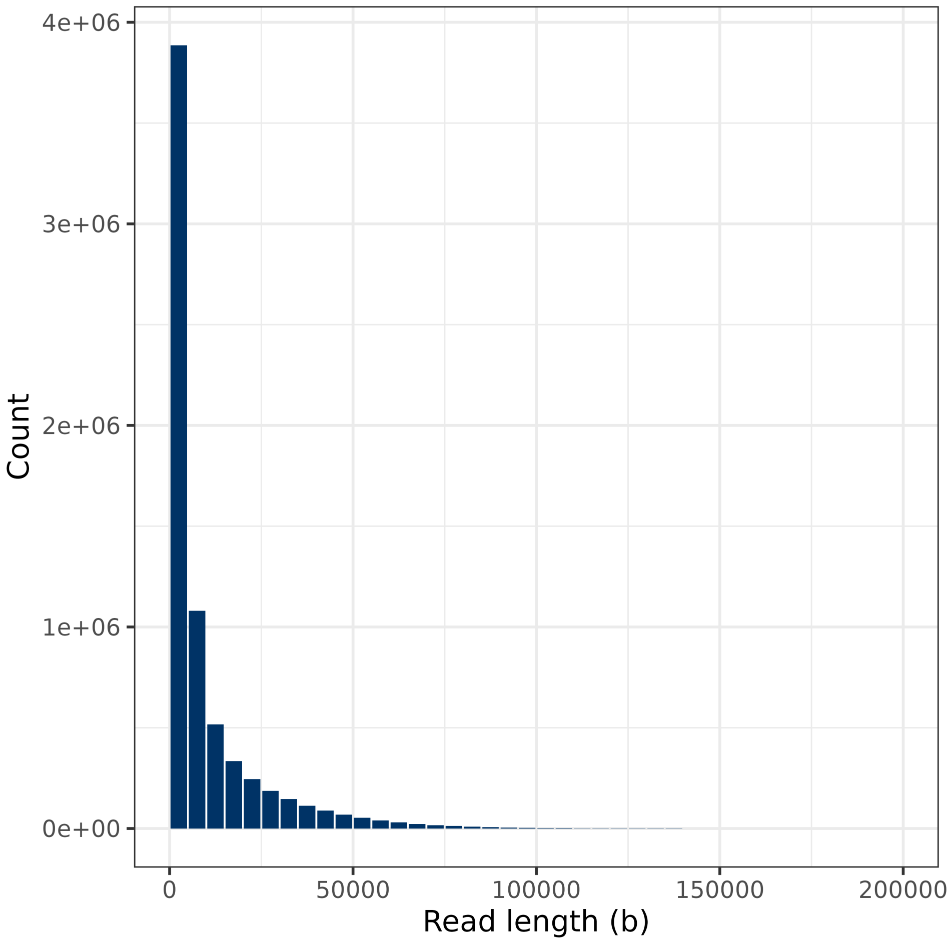
1 Source files/directories

```
Sequencing summary: /ei/projects/8/8e1c31fd-918e-4bd7-8fb8-2de97ff7675f/data/raw
/CN_AR_ARY017-30uM_202031109/CN_AR_ARY017-30uM_202031109/20231109_1524_P2S-00659
-B_PAM35249_14fe2872/sequencing_summary_PAM35249_14fe2872_890bff3c.txt
Mod bam file: /ei/projects/8/8e1c31fd-918e-4bd7-8fb8-2de97ff7675f/scratch/202311
09_AR_ONT_SC_ARY01730uMpromethion_e50860f/dnascent_sam_minimap2_fastq_20231109_A
R_ONT_SC_ARY01730uMpromethion_e50860f.detect.mod.sorted.bam
Forksense directory: /ei/projects/8/8e1c31fd-918e-4bd7-8fb8-2de97ff7675f/scratch
/20231109_AR_ONT_SC_ARY01730uMpromethion_e50860f/forkSenseOverallBedgraphs
```

2 Read lengths

2.1 Data for histogram of read lengths from sequencing summary file

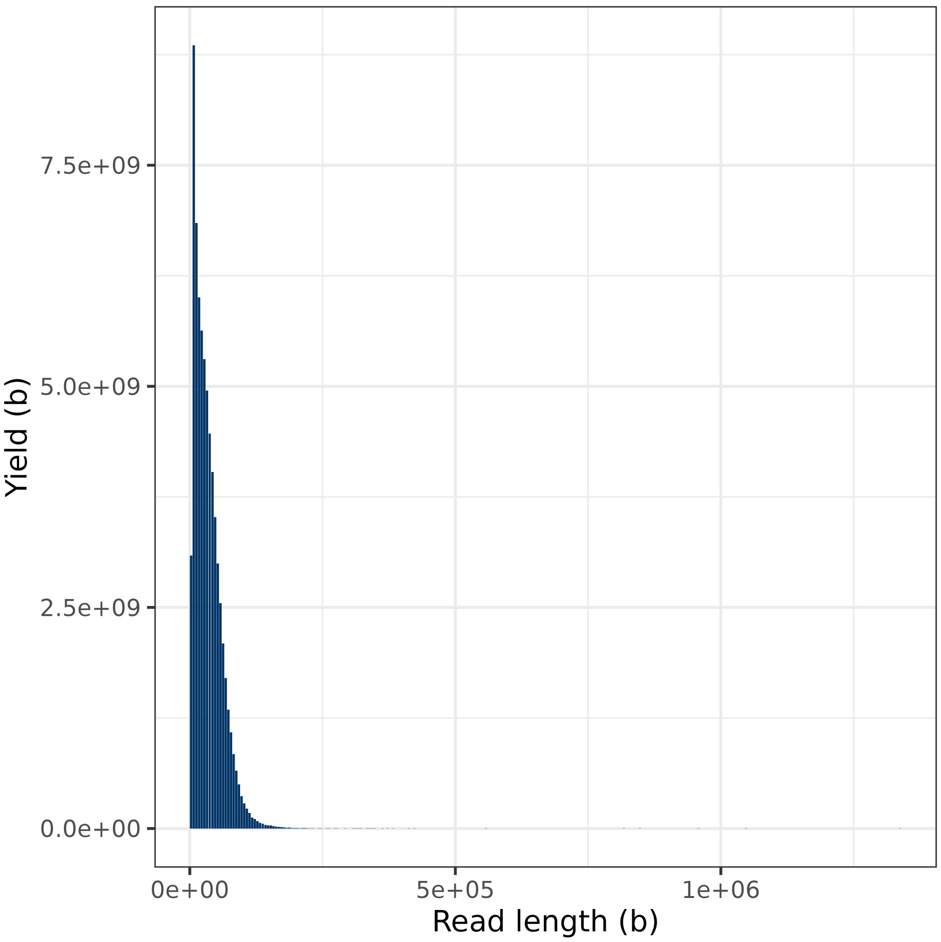
```
bin_lo bin_hi sequence_length_template_count
0      5000   3885935
5000   10000  1080688
10000  15000  516588
15000  20000  333945
20000  25000  244670
25000  30000  187189
30000  35000  145832
35000  40000  113489
40000  45000  89066
45000  50000  68952
50000  55000  52793
55000  60000  40355
60000  65000  30240
65000  70000  22655
70000  75000  16480
75000  80000  12468
80000  85000  9059
85000  90000  6558
90000  95000  4637
95000  100000 3294
100000 105000 2434
105000 110000 1860
110000 115000 1322
115000 120000 960
120000 125000 818
125000 130000 557
130000 135000 437
135000 140000 355
140000 145000 260
145000 150000 225
150000 155000 199
155000 160000 150
160000 165000 128
165000 170000 96
170000 175000 83
175000 180000 72
180000 185000 37
185000 190000 48
190000 195000 27
195000 200000 25
```


4e+06
3e+06
2e+06
1e+06
0e+00
Count
0
50000
100000
150000
200000
Read length (b)

2.2 Data for yield in bases vs binned read length from sequencing summary file

bin_lo	bin_hi	num_bases
0	5000	3086365592
5000	10000	8856293478
10000	15000	6847690284
15000	20000	6004369087
20000	25000	5631755792
25000	30000	5308275503
30000	35000	4952210742
35000	40000	4465801162
40000	45000	4030086364
45000	50000	3521690306
50000	55000	2997205982
55000	60000	2547735888
60000	65000	2091567313
65000	70000	1702288603
70000	75000	1344252481
75000	80000	1088439366
80000	85000	841662042
85000	90000	653333530
90000	95000	499890247
95000	100000	365636234
100000	105000	284367051
105000	110000	225622956
110000	115000	175012929
115000	120000	125118489
120000	125000	107950044
125000	130000	83104581
130000	135000	64595167
135000	140000	54516211
140000	145000	39594189
145000	150000	35932906
150000	155000	33770576
155000	160000	25886388
160000	165000	21923034
165000	170000	17646519
170000	175000	15476133
175000	180000	12944423
180000	185000	8623345
185000	190000	10935759
190000	195000	5515713
195000	200000	4873689
200000	205000	5407286
205000	210000	3268220
210000	215000	3782006
215000	220000	4511280
220000	225000	3306195
225000	230000	1571059
230000	235000	1612853
235000	240000	471628
240000	245000	1200809
245000	250000	1465857
250000	255000	251920
255000	260000	1274439
260000	265000	1559094
265000	270000	527597

270000	275000	1351688
275000	280000	1376002
280000	285000	280682
285000	290000	285994
290000	295000	1158689
295000	300000	294923
300000	305000	299493
305000	310000	1519544
310000	315000	930687
315000	320000	946167
320000	325000	1600136
325000	330000	325966
330000	335000	1320977
335000	340000	1004406
340000	345000	1019991
345000	350000	1036983
350000	355000	695870
355000	360000	354867
360000	365000	2152931
365000	370000	727052
370000	375000	1106753
380000	385000	1141996
395000	400000	393316
400000	405000	402421
410000	415000	1224392
415000	420000	416408
420000	425000	839950
430000	435000	431046
435000	440000	436018
440000	445000	437962
445000	450000	443189
455000	460000	452504
465000	470000	464372
555000	560000	1110032
565000	570000	564763
575000	580000	577334
580000	585000	581179
590000	595000	588151
595000	600000	596784
620000	625000	620696
640000	645000	638921
815000	820000	816199
845000	850000	845639
955000	960000	954260
1045000	1050000	1044870
1335000	1340000	1334163



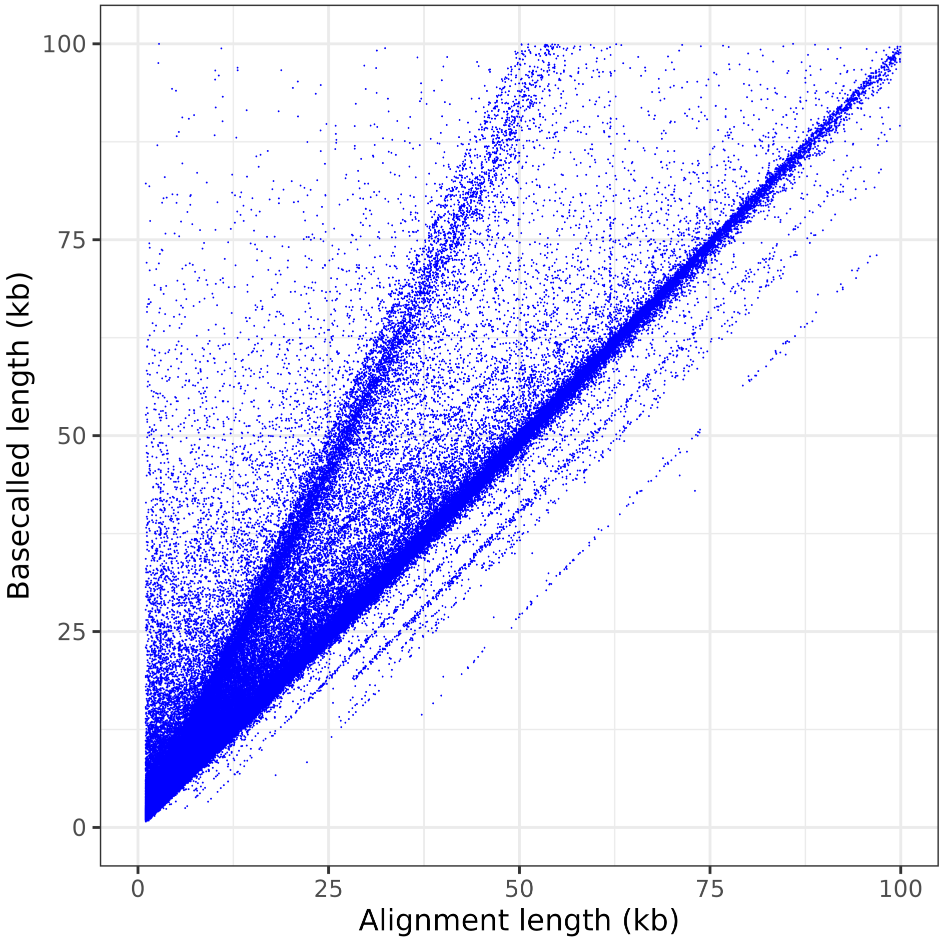
Yield (b)
7.5e+09
5.0e+09
2.5e+09
0.0e+00
0e+00
5e+05
1e+06
Read length (b)

2.3 Statistics of read lengths from sequencing summary file

```
sequence_length_template_count  6875207
sequence_length_template_sum    68257316707
sequence_length_template_min    19
sequence_length_template_p10    660
sequence_length_template_p50    3935
sequence_length_template_mean   9928.038051363399
sequence_length_template_p90    28229
sequence_length_template_max    1334163
sequence_length_template_stddev 14701.612137295344
```

```
 N50 (sampling 100,000 reads randomly)

sequence_length_template                     25779
sequence_length_template_cumulative_fraction 0.5000060942981484
```

2.4 Read length from sequencing summary file vs alignment length from mod bam file



```
NOTE: The plot runs from x = 0 to x = 100 and same for y. Data outside this range are not shown.
```

3 Fork and origin statistics

3.1 Numbers of different features

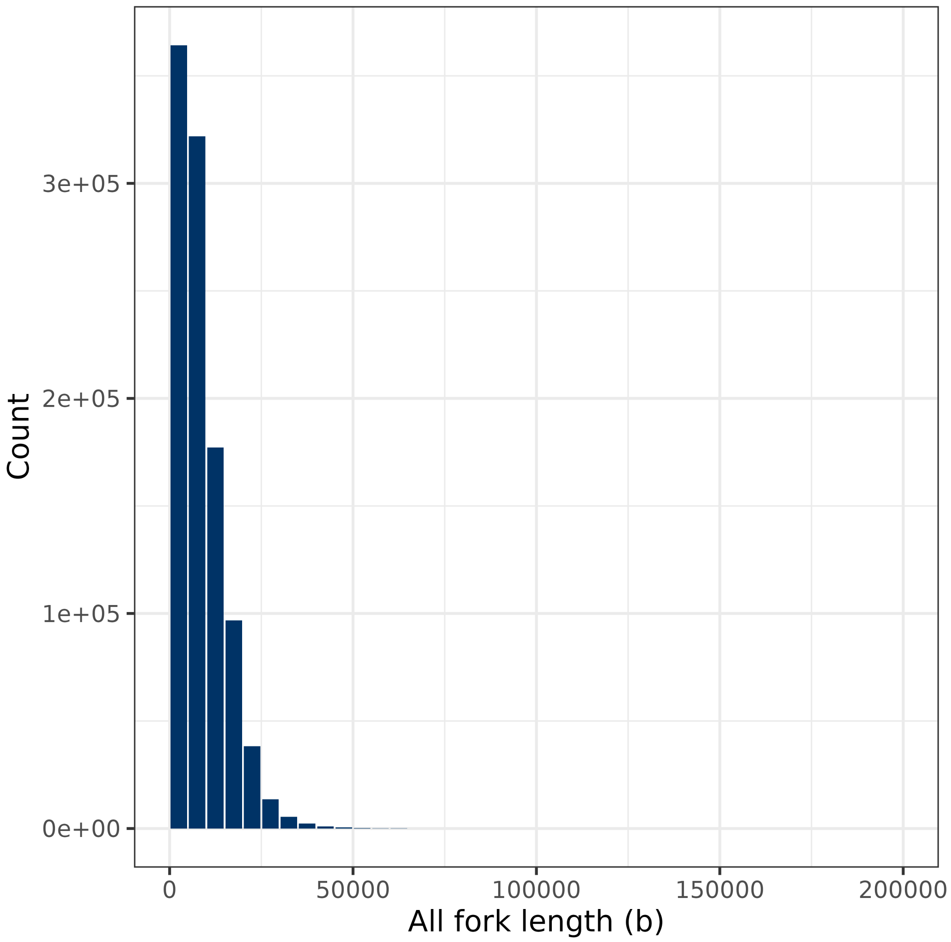
```
Number of left forks
540908
Number of right forks
480687
Number of origins
242634
Number of terminations
128930
Number of molecules with left forks
449932
Number of molecules with right forks
397934

Considering only fwd reads
Number of left forks
262145
Number of right forks
247579
Number of origins
116320
Number of terminations
66269
Number of molecules with left forks
218055
Number of molecules with right forks
204989

Considering only rev reads
Number of left forks
278763
Number of right forks
233108
Number of origins
126314
Number of terminations
62661
Number of molecules with left forks
231877
Number of molecules with right forks
192945
```

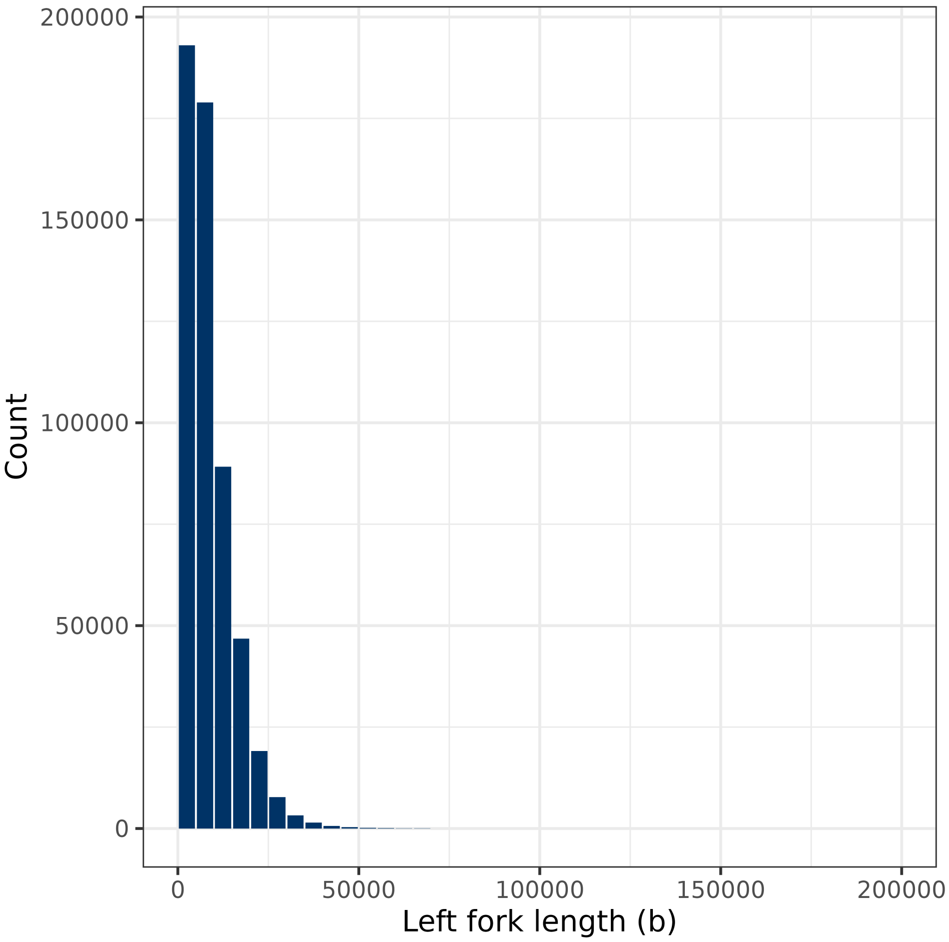
3.2 Raw data for histogram of fork lengths

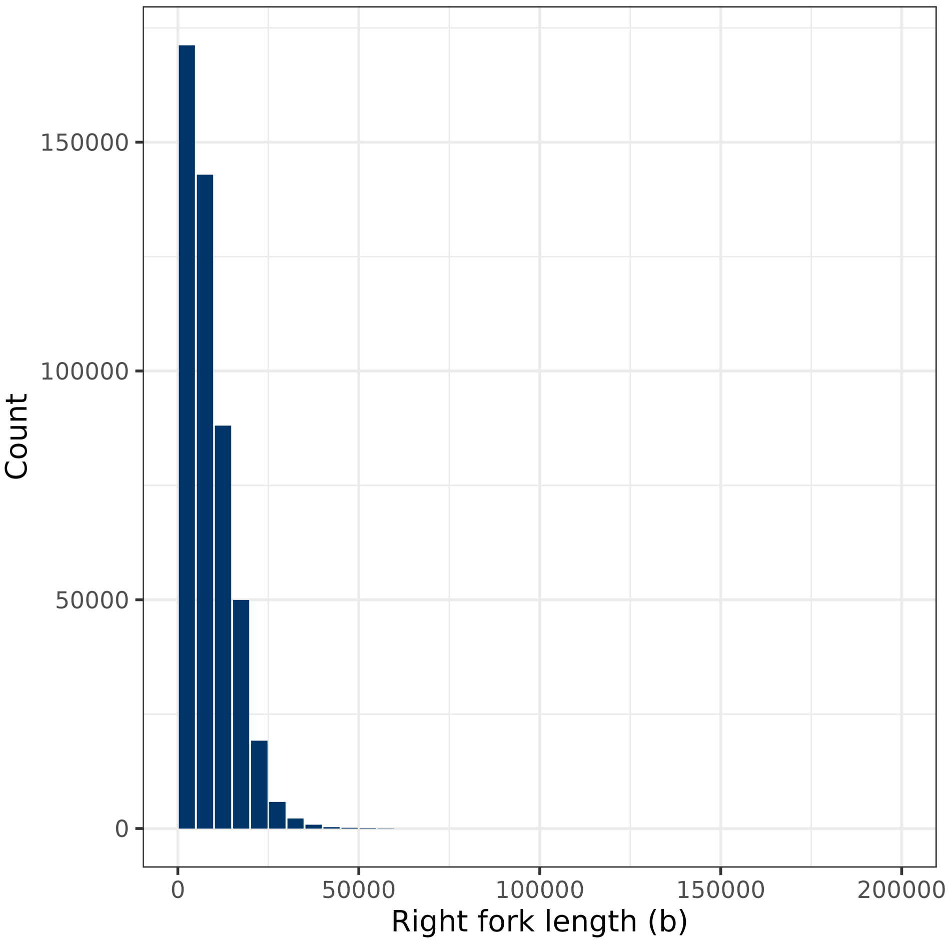
```
 All forks

bin_lo bin_hi fork_length_count
0      5000   364213
5000   10000  321843
10000  15000  177184
15000  20000  96751
20000  25000  38286
25000  30000  13563
30000  35000  5447
35000  40000  2346
40000  45000  990
45000  50000  492
50000  55000  253
55000  60000  113
60000  65000  58
65000  70000  24
70000  75000  18
75000  80000  6
80000  85000  3
85000  90000  4
90000  95000  1
95000  100000 0
100000 105000 0
105000 110000 0
110000 115000 0
115000 120000 0
120000 125000 0
125000 130000 0
130000 135000 0
135000 140000 0
140000 145000 0
145000 150000 0
150000 155000 0
155000 160000 0
160000 165000 0
165000 170000 0
170000 175000 0
175000 180000 0
180000 185000 0
185000 190000 0
190000 195000 0
195000 200000 0

 Left forks

bin_lo bin_hi fork_length_count
0      5000   193029
5000   10000  178960
10000  15000  89157
15000  20000  46804
20000  25000  19115
25000  30000  7763
30000  35000  3258
35000  40000  1492
40000  45000  662
```

```
45000  50000  336
50000  55000  162
55000  60000  82
60000  65000  47
65000  70000  17
70000  75000  14
75000  80000  4
80000  85000  2
85000  90000  3
90000  95000  1
95000  100000 0
100000 105000 0
105000 110000 0
110000 115000 0
115000 120000 0
120000 125000 0
125000 130000 0
130000 135000 0
135000 140000 0
140000 145000 0
145000 150000 0
150000 155000 0
155000 160000 0
160000 165000 0
165000 170000 0
170000 175000 0
175000 180000 0
180000 185000 0
185000 190000 0
190000 195000 0
195000 200000 0

 Right forks

bin_lo bin_hi fork_length_count
0      5000   171184
5000   10000  142883
10000  15000  88027
15000  20000  49947
20000  25000  19171
25000  30000  5800
30000  35000  2189
35000  40000  854
40000  45000  328
45000  50000  156
50000  55000  91
55000  60000  31
60000  65000  11
65000  70000  7
70000  75000  4
75000  80000  2
80000  85000  1
85000  90000  1
90000  95000  0
95000  100000 0
100000 105000 0
105000 110000 0
110000 115000 0
```

```
115000 120000 0
120000 125000 0
125000 130000 0
130000 135000 0
135000 140000 0
140000 145000 0
145000 150000 0
150000 155000 0
155000 160000 0
160000 165000 0
165000 170000 0
170000 175000 0
175000 180000 0
180000 185000 0
185000 190000 0
190000 195000 0
195000 200000 0
```


Count
3e+05
2e+05
1e+05
0e+00
0
50000
100000
150000
200000
All fork length (b)


200000
150000
100000
50000
0
Count
0
50000
100000
150000
200000
Left fork length (b)


Count
150000
100000
50000
0
0
50000
100000
150000
200000
Right fork length (b)

3.3 Statistics of fork lengths

```
 All forks

fork_length_count  1021595
fork_length_sum    8651068411
fork_length_min    6
fork_length_p10    1384
fork_length_p50    6892
fork_length_mean   8468.197682056001
fork_length_p90    17467
fork_length_max    92252
fork_length_stddev 6654.351076884262

 N50

fork_length                        12368
fork_length_cumulative_fraction 0.5000000964620739

 Left forks

fork_length_count  540908
fork_length_sum    4534688355
fork_length_min    6
fork_length_p10    1394
fork_length_p50    6695
fork_length_mean   8383.47437087268
fork_length_p90    17294
fork_length_max    92252
fork_length_stddev 6709.4800414018055

 N50

fork_length                        11993
fork_length_cumulative_fraction 0.5000018906481171

 Right forks

fork_length_count  480687
fork_length_sum    4116380056
fork_length_min    6
fork_length_p10    1378
fork_length_p50    7126
fork_length_mean   8563.535223544635
fork_length_p90    17633
fork_length_max    87320
fork_length_stddev 6590.468916696038

 N50

fork_length                        12737
fork_length_cumulative_fraction 0.5000014102196398
```

3.4 Statistics of origin, termination uncertainty intervals

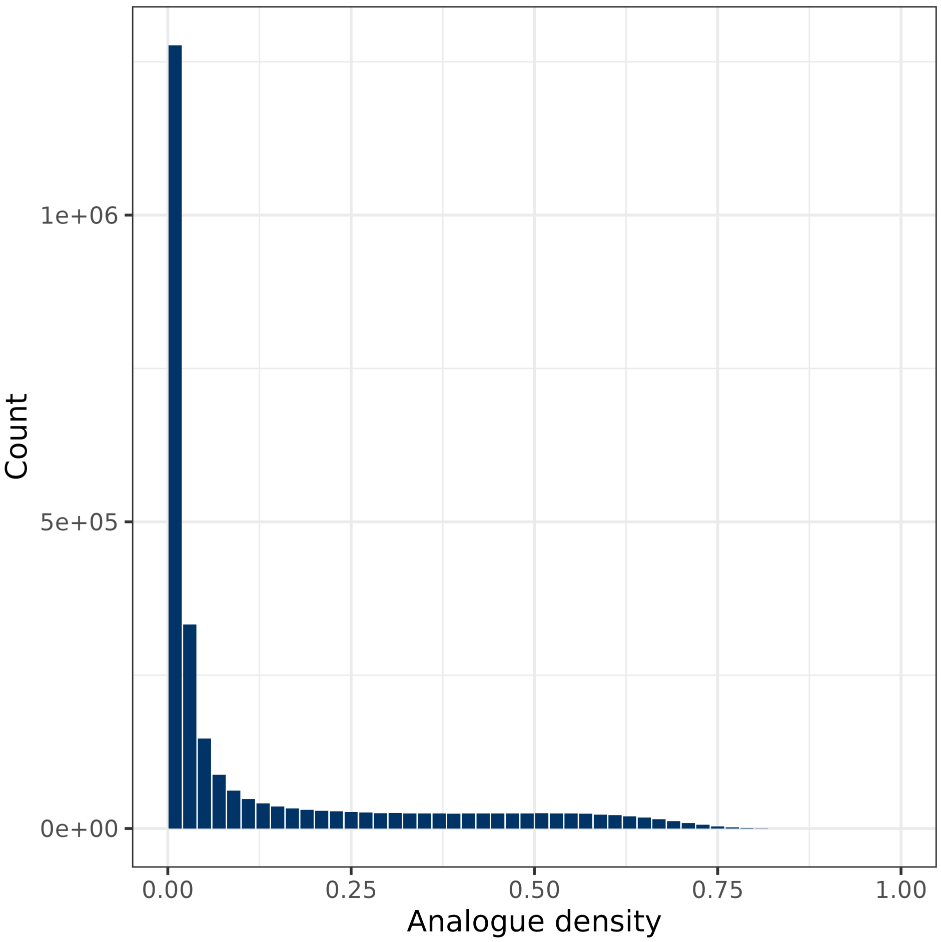
DNAscent associates one genomic window per called origin/termination. Here are the statistics of those windows.

```
 All origins

origin_length_count  242634
origin_length_sum    269404819
origin_length_min    94
origin_length_p10    351
origin_length_p50    696
origin_length_mean   1110.334161741553
origin_length_p90    2082
origin_length_max    80042
origin_length_stddev 1526.9092110956542

 All terminations

termination_length_count  128930
termination_length_sum    771919387
termination_length_min    535
termination_length_p10    1761
termination_length_p50    3169
termination_length_mean   5987.120041883192
termination_length_p90    14485
termination_length_max    97686
termination_length_stddev 7252.06108587167
```

4 Whole read analogue density statistics

4.1 Raw data for histogram of densities

```
 All molecules

bin_lo bin_hi mean_brdU_count
0      0.02   1276852
0.02   0.04   332469
0.04   0.06   146544
0.06   0.08   87524
0.08   0.1    61749
0.1    0.12   48143
0.12   0.14   40909
0.14   0.16   36010
0.16   0.18   32795
0.18   0.2    30512
0.2    0.22   29049
0.22   0.24   27918
0.24   0.26   26884
0.26   0.28   26124
0.28   0.3    25141
0.3    0.32   25391
0.32   0.34   24484
0.34   0.36   24491
0.36   0.38   24436
0.38   0.4    24135
0.4    0.42   24395
0.42   0.44   24501
0.44   0.46   24596
0.46   0.48   24684
0.48   0.5    24631
0.5    0.52   25082
0.52   0.54   24730
0.54   0.56   24391
0.56   0.58   24136
0.58   0.6    22781
0.6    0.62   21922
0.62   0.64   19846
0.64   0.66   17905
0.66   0.68   15155
0.68   0.7    12142
0.7    0.72   9075
0.72   0.74   6122
0.74   0.76   3596
0.76   0.78   1897
0.78   0.8    835
0.8    0.82   356
0.82   0.84   94
0.84   0.86   18
0.86   0.88   7
0.88   0.9    0
0.9    0.92   0
0.92   0.94   0
0.94   0.96   0
0.96   0.98   0
0.98   1      0
```


1e+06
5e+05
0e+00
Count
0.00
0.25
0.50
0.75
1.00
Analogue density

4.2 Analogue density statistics

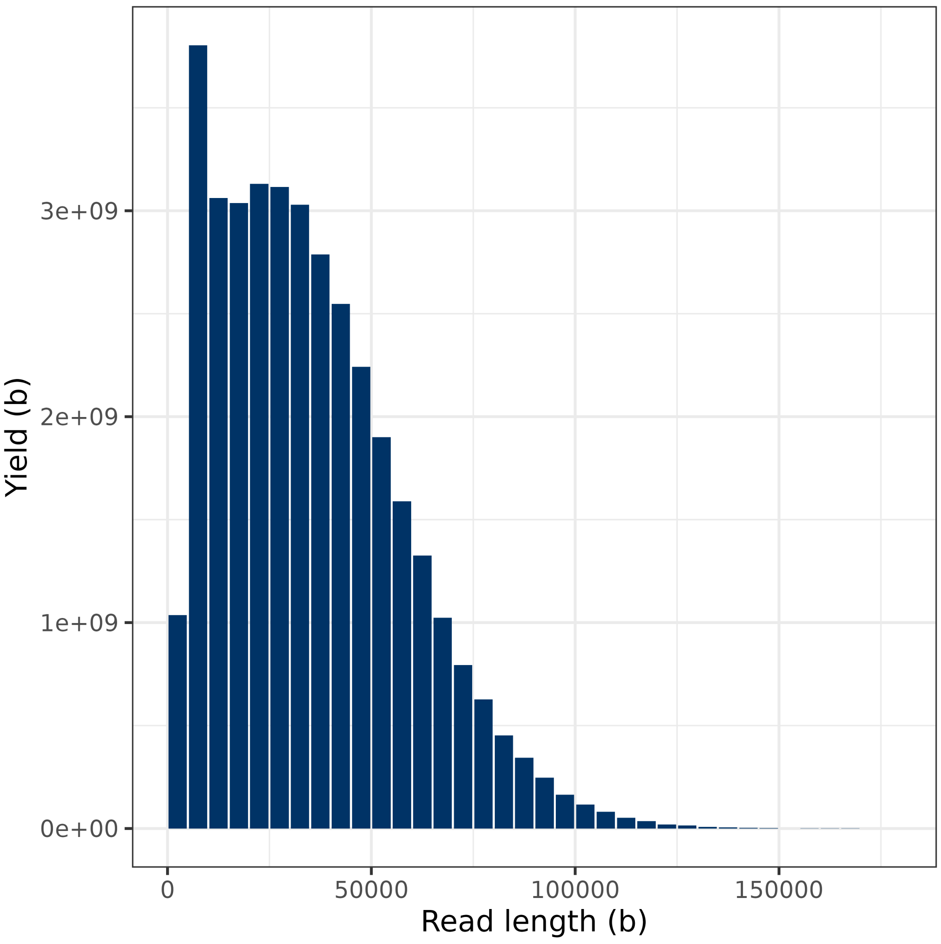
```
 All molecules

mean_brdU_count  2704457
mean_brdU_sum    337090.0156380542
mean_brdU_min    0
mean_brdU_p10    0.004207
mean_brdU_p50    0.022837
mean_brdU_mean   0.12464240164959331
mean_brdU_p90    0.467262
mean_brdU_max    0.878603
mean_brdU_stddev 0.1893193246108163
```

4.3 Statistics of reads and read lengths in modbam file

```
l_count  2704457
l_sum    36599799906
l_min    1000
l_p10    1624
l_p50    6314
l_mean   13533.14173824912
l_p90    36501
l_max    173311
l_stddev 16122.577919425146

 N50 of reads within modbam file

l     l_cumulative_fraction
29316 0.5000125765441664

 Number of forward and reverse reads within modbam file

orientation count
+ 1351948
-1352509
```

4.4 Raw data for yield in bases vs binned read length from mod bam file

```
bin_lo bin_hi n_bases
0      5000   1036209747
5000   10000  3803571347
10000  15000  3062316275
15000  20000  3037740986
20000  25000  3130586436
25000  30000  3115164106
30000  35000  3029010634
35000  40000  2787565420
40000  45000  2548236361
45000  50000  2241694773
50000  55000  1900944395
55000  60000  1589415082
60000  65000  1326116867
65000  70000  1023486786
70000  75000  793814999
75000  80000  626964310
80000  85000  452081955
85000  90000  344260623
90000  95000  246779022
95000  100000 164591251
100000 105000 115792010
105000 110000 81613329
110000 115000 51687552
115000 120000 35799285
120000 125000 19157188
125000 130000 14931630
130000 135000 7913259
135000 140000 5637881
140000 145000 3069735
145000 150000 1588023
150000 155000 148176
155000 160000 927006
160000 165000 480293
165000 170000 329853
175000 180000 173311
```

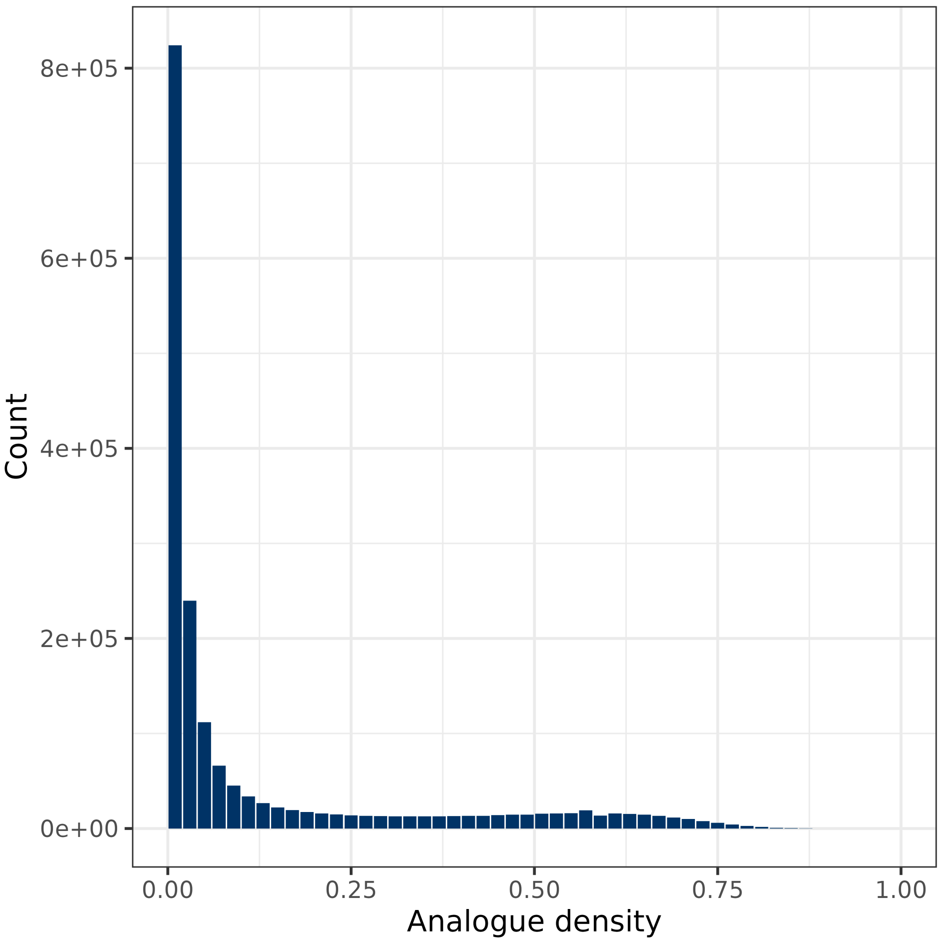

Yield (b)
3e+09
2e+09
1e+09
0e+00
0
50000
100000
150000
Read length (b)

5 Windowed analogue density statistics

5.1 Raw data for histogram of densities

```
Using a window size of 300 thymidines.
NOTE: As this calculation is compute-intensive, we choose 5% of the reads at ran
dom and calculate the windowed analogue density along them.

 Subset of molecules

bin_lo bin_hi mean_brdU_count
0      0.02   824118
0.02   0.04   239813
0.04   0.06   111903
0.06   0.08   66056
0.08   0.1    45195
0.1    0.12   33753
0.12   0.14   26827
0.14   0.16   22065
0.16   0.18   19516
0.18   0.2    17383
0.2    0.22   15871
0.22   0.24   14902
0.24   0.26   13875
0.26   0.28   13348
0.28   0.3    13153
0.3    0.32   12858
0.32   0.34   12903
0.34   0.36   12727
0.36   0.38   12939
0.38   0.4    13080
0.4    0.42   13252
0.42   0.44   13409
0.44   0.46   14144
0.46   0.48   14560
0.48   0.5    14722
0.5    0.52   15569
0.52   0.54   15968
0.54   0.56   16077
0.56   0.58   19246
0.58   0.6    13668
0.6    0.62   15972
0.62   0.64   15268
0.64   0.66   14547
0.66   0.68   13324
0.68   0.7    11686
0.7    0.72   9959
0.72   0.74   7814
0.74   0.76   5974
0.76   0.78   4236
0.78   0.8    2713
0.8    0.82   1586
0.82   0.84   764
0.84   0.86   358
0.86   0.88   132
0.88   0.9    41
0.9    0.92   10
0.92   0.94   2
0.94   0.96   0
```

```
0.96   0.98   0
0.98   1      0
```

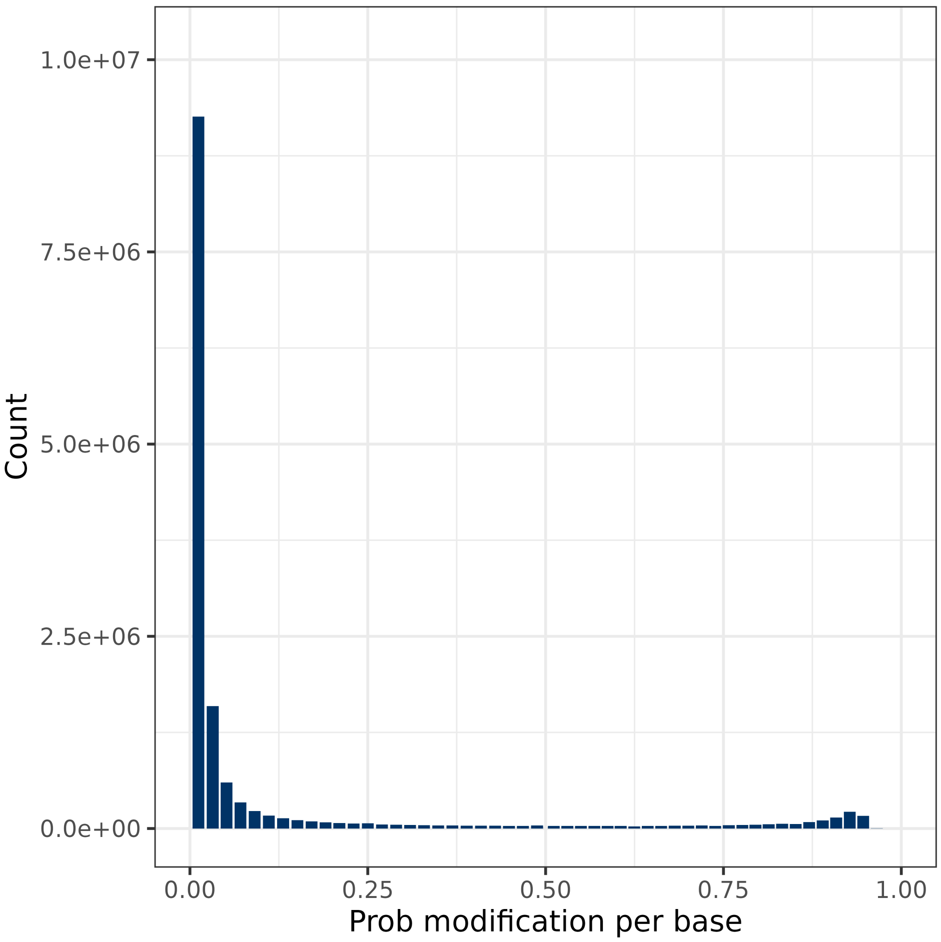

Count
8e+05
6e+05
4e+05
2e+05
0e+00
0.00
0.25
0.50
0.75
1.00
Analogue density

6 Raw analogue probability statistics

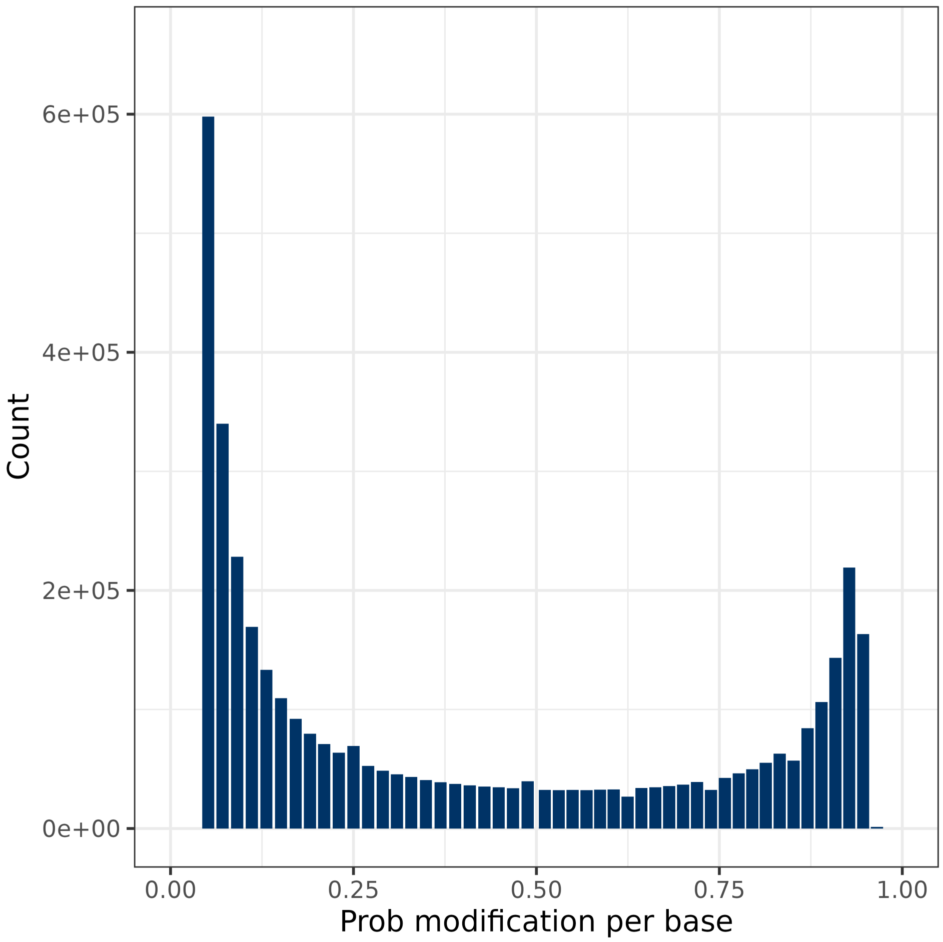
6.1 Raw data for histogram of probabilities

```
NOTE: As this calculation is compute-intensive, we choose 10000 reads at random
NOTE: This calculation may count bases whose modification status is unknown as u
nmodified

 Subset of molecules

count bin_lo bin_hi
9261387 0.0020000000000000018 0.022000000000000002
1593654 0.022000000000000002 0.042000000000000004
598077 0.042000000000000004 0.061000000000000054
340057 0.061000000000000054 0.08099999999999996
228297 0.08099999999999996 0.10099999999999998
169329 0.10099999999999998 0.121
133291 0.121 0.14100000000000001
109407 0.14100000000000001 0.16100000000000003
92199 0.16100000000000003 0.18100000000000005
79686 0.18100000000000005 0.19999999999999996
70993 0.19999999999999996 0.21999999999999997
63790 0.21999999999999997 0.24
69286 0.24 0.26
52527 0.26 0.28
48540 0.28 0.30000000000000004
45607 0.30000000000000004 0.31899999999999995
43284 0.31899999999999995 0.33899999999999997
40775 0.33899999999999997 0.359
38871 0.359 0.379
37506 0.379 0.399
36164 0.399 0.41900000000000004
35289 0.41900000000000004 0.43899999999999995
34584 0.43899999999999995 0.45799999999999996
33854 0.45799999999999996 0.478
39695 0.478 0.498
32509 0.502 0.521
32229 0.521 0.540
32417 0.540 0.559
32294 0.559 0.578
32650 0.578 0.596
32848 0.596 0.615
26722 0.615 0.634
34088 0.634 0.653
34632 0.653 0.672
35577 0.672 0.691
36810 0.691 0.710
39020 0.710 0.729
32347 0.729 0.748
42550 0.748 0.767
46261 0.767 0.786
49834 0.786 0.804
55251 0.804 0.823
62858 0.823 0.842
57127 0.842 0.861
84334 0.861 0.880
106173 0.880 0.899
143449 0.899 0.918
219237 0.918 0.937
```

```
163298 0.937 0.956
1462 0.956 0.975
```


1.0e+07
7.5e+06
5.0e+06
2.5e+06
0.0e+00
Count
0.00
0.25
0.50
0.75
1.00
Prob modification per base

Zoom in to probabilities greater than 0.04


6e+05
4e+05
2e+05
0e+00
Count
Prob modification per base