

Genome 540 Discussion

Feb. 24th 2026



HW7 Overview

HW6 Review

Read counts:

16	16507014	0
16	16507015	0
16	16507016	0
16	16507017	0
16	16507018	0
16	16507019	0
16	16507020	0
16	16507021	2
16	16507022	2
16	16507023	0
16	16507024	0
16	16507025	0
16	16507026	0
16	16507027	0
16	16507028	0
16	16507029	1
16	16507030	0
16	16507031	0
16	16507032	0
16	16507033	0
16	16507034	0
16	16507035	0
16	16507036	0
16	16507037	1
16	16507038	0
16	16507039	0
16	16507040	0
16	16507041	1
16	16507042	1
16	16507043	0
16	16507044	0
16	16507045	0
16	16507046	0

Scoring Scheme:

- score for 0 reads: -0.1077
- score for 1 read: 0.4772
- score for 2 reads: 1.0622
- score for ≥ 3 reads: 1.6748
- $D = -20$
- $S = -D = 20$

How many maximal D-segments were there? Is that more or less than you expected? Is the provided scoring scheme a good discriminator?

HW7 Overview

- Take the D-segments identified from HW6
 - a) Generate a new scoring scheme based on log-likelihood ratios
 - b) Generate a simulated read count file
- Run D-segments again, using updated scoring scheme and $S = -D = 5$
- Answer some questions about the comparison of simulated and real data

HW7 Generating Frequencies:

Read counts:

Non-elevated regions:

0 = 10000000

1 = 200000

2 = 30000

>3 = 4000

Elevated regions:

0 = 1000

1 = 200

2 = 30

>3 = 4

All regions:

0 = 10001000

1 = 200200

2 = 30030

>3 = 4004

Total ne: 10234000

Total elev: 1234

Total total: 10235234

HW7 Generating Frequencies:

Background frequencies:

Target frequencies:

$$0 = 10001000 / 10235234 = 0.9771$$

$$0 = 1000 / 1234 = 0.8104$$

$$1 = 200200 / 10235234 = 0.0195$$

$$1 = 200 / 1234 = 0.1621$$

$$2 = 30030 / 10235234 = 0.0029$$

$$2 = 30 / 1234 = 0.0243$$

$$>3 = 4004 / 10235234 = 0.0003$$

$$>3 = 4 / 1234 = 0.0032$$

Formatting: {#.####}

HW7 Simulating BG Sequence:

Background frequencies:

$$0 = 10001000 / 10235234 = 0.9771$$

$$1 = 200200 / 10235234 = 0.0195$$

$$2 = 30030 / 10235234 = 0.0029$$

$$>3 = 4004 / 10235234 = 0.0003$$

```
for each i = 1...N
  x = random number between 0 and 1 (uniform distributi
  if x < bkgd[0]
    sim_seq[i] = 0
  else if x < bkgd[0] + bkgd[1]
    sim_seq[i] = 1
  else if x < bkgd[0] + bkgd[1] + bkgd[2]
    sim_seq[i] = 2
  else
    sim_seq[i] = 3
```

HW7 Scoring Scheme:

New Scheme:

$$0 = \log_2(?) - \log_2 (?) = ?$$

$$1 = \log_2(?) - \log_2 (?) = ?$$

$$2 = \log_2(?) - \log_2 (?) = ?$$

$$>3 = \log_2(?) - \log_2 (?) = ?$$

Tiny Gotcha: Writing four decimal places to a file truncates scores, slightly changes results.

HW7 Running D-Segments:

New Scheme:

0 = ?

1 = ?

2 = ?

>3 = ?

S = -D = 5

real sequence:

5 {# of segments with score ≥ 5 }

6 {# of segments with score ≥ 6 }

7 {# of segments with score ≥ 7 }

.

.

.

list all the segment score counts for scores
between 5 and 30

(only first/last 3 shown here)

.

.

.

28 {# of segments with score ≥ 28 }

29 {# of segments with score ≥ 29 }

30 {# of segments with score ≥ 30 }

simulated sequence:

5 {# of segments with score ≥ 5 }

6 {# of segments with score ≥ 6 }

7 {# of segments with score ≥ 7 }

.

.

.

list all the segment score counts for scores
between 5 and 30

(only first/last 3 shown here)

.

.

.

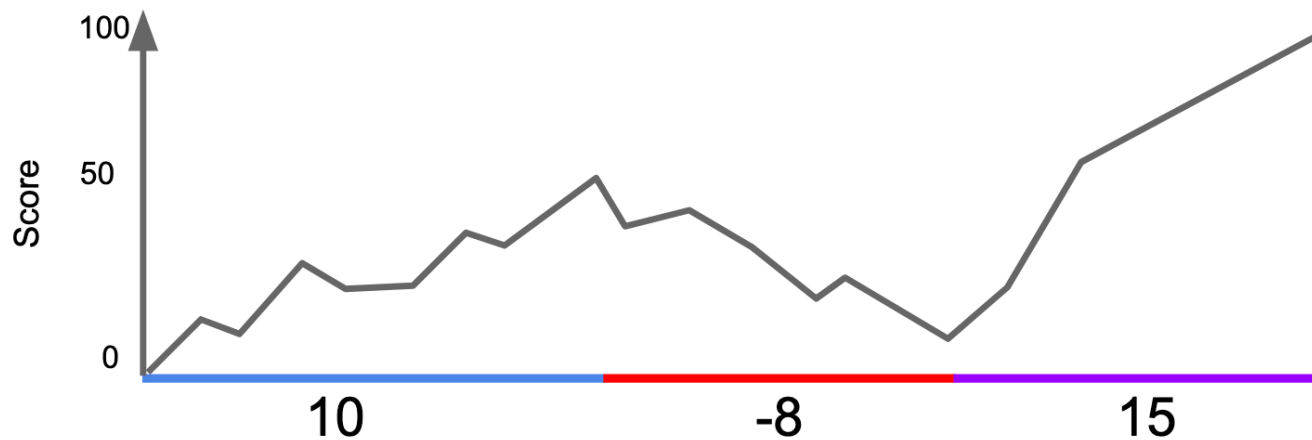
28 {# of segments with score ≥ 28 }

29 {# of segments with score ≥ 29 }

30 {# of segments with score ≥ 30 }

HW7 Running D-Segments:

As you find D-segments, you can add to the histogram:



5: 2
6: 2
7: 2
8: 2
9: 2
10: 2
11: 1
12: 1
13: 1
14: 1
15: 1
16: 0
17: 0
...
30: 0

HW7: BG Distribution

Finally, we want to understand the shape of the BG distribution:

Ratios of simulated data:

$N_seg(5)/N_seg(6)$ {# of segments with score ≥ 5 / # of segments with score ≥ 6 }

$N_seg(6)/N_seg(7)$ {# of segments with score ≥ 6 / # of segments with score ≥ 7 }

$N_seg(7)/N_seg(8)$ {# of segments with score ≥ 7 / # of segments with score ≥ 8 }

.

.

.

list all ratios

(only first/last 3 shown here)

.

.

.

$N_seg(27)/N_seg(28)$ {# of segments with score ≥ 27 / # of segments with score ≥ 28 }

$N_seg(28)/N_seg(29)$ {# of segments with score ≥ 28 / # of segments with score ≥ 29 }

$N_seg(29)/N_seg(30)$ {# of segments with score ≥ 29 / # of segments with score ≥ 30 }

HW7 Comparing Real Data to Sim

Does the D-segment distribution match the expectation from Karlin-Altshul theory? In the simulated case? In the real data?

What is a good selection for a score threshold? About what would you expect the false-positive rate would be?