

Genome 540 Discussion

Feb. 19th 2026



HW6 Questions?

Input files:

Read count file:

chromosome (should stay const)

position

of reads

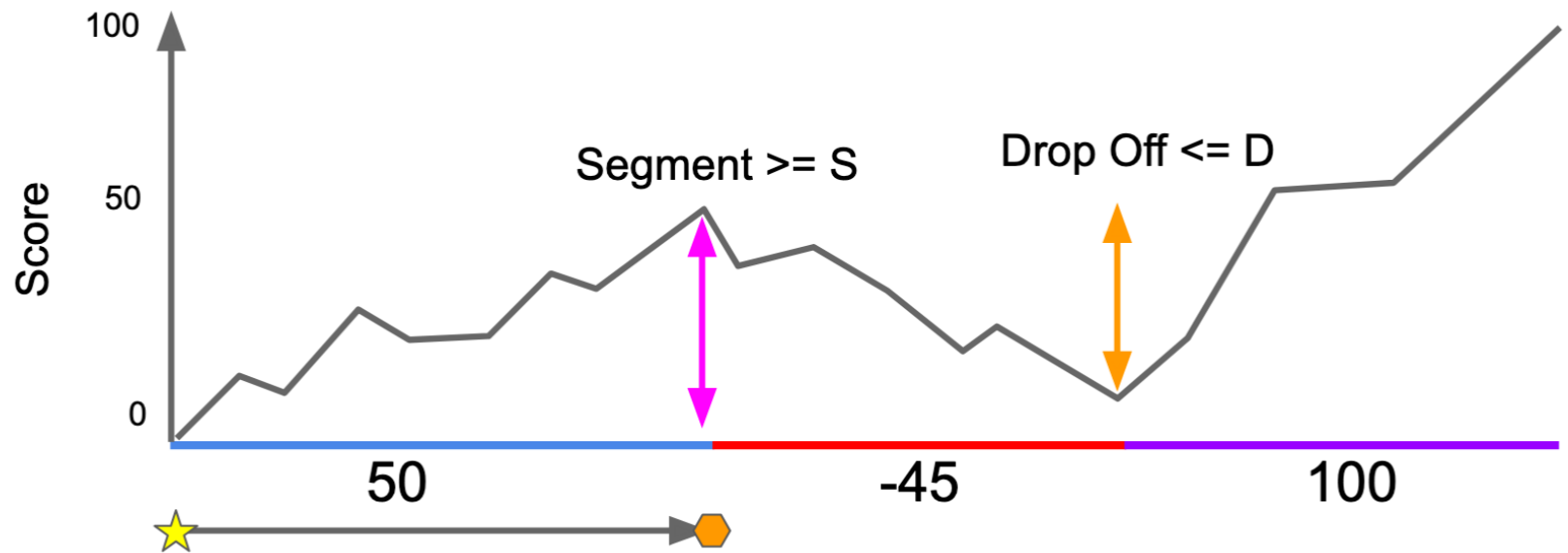
Score file:

scores, D and S

```
0 -0.3464
1 0.2488
2 0.8439
>=3 1.5337
D -33.219
S 33.219
```

17	1	0
17	2	0
17	3	0
17	4	0
17	5	2
17	6	0
17	7	0
17	8	0
17	9	0
17	10	0
17	11	0
17	12	1
17	13	0
17	14	3
17	15	0
17	16	0
17	17	2
17	18	0
17	19	0
17	20	0
17	21	5
17	22	0
17	23	1
17	24	0
17	25	0
17	26	3

D-segments



D-segments

$O(N)$ algorithm to find all maximal D-segs:

```
cumul = max = 0; start = 1;
for (i = 1; i ≤ N; i++) {
    cumul += s[i];
    if (cumul ≥ max)
        {max = cumul; end = i;}
    if (cumul ≤ 0 or cumul ≤ max + D or i == N) {
        if (max ≥ S)
            {print start, end, max; }
        max = cumul = 0; start = end = i + 1; /* NO BACKTRACKING
        NEEDED! */
    }
}
```

Read mapping

Different tools for different needs!

- Short read DNA aligners (BWA, Bowtie2)
- Long read DNA aligners (Minimap2)
- RNA-seq aligners (HISAT2, STAR)

Read mapping

Different tools for different needs!

- Short read DNA aligners (BWA, Bowtie2)
- Long read DNA aligners (Minimap2)
- RNA-seq aligners (HISAT2, STAR)

File Formats: SAM

[illegible]

Col	Field	Type	Regex/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2 ¹⁶ -1]	bitwise FLAG
3	RNAME	String	* [!-()+<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 ³¹ -1]	1-based leftmost mapping POSITION
5	MAPQ	Int	[0,2 ⁸ -1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	* = [!-()+<>-~] [!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2 ³¹ -1]	Position of the mate/next read
9	TLEN	Int	[-2 ³¹ +1,2 ³¹ -1]	observed Template LENGTH
10	SEQ	String	* [A-Za-z=.]+	segment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

<https://learn.gencore.bio.nyu.edu/ngs-file-formats/sambam-format/>

IGV Demo!

