

Start	1	3	5	7	9	13	15	19	22	24	26	32	34	38
End	2	8	7	20	10	15	16	30	24	25	28	38	36	40
MaxE	2	8	8	20	20	20	20	30	30	30	30	38	38	40

(a)



(b)

Start	1	3	5	7	9	13	15	19	22	24	26	32	34	38
Index	1	2	3	3	4	5	6	6	7	8	9	10	11	12

(c)

Figure 1. (a) An interval list sorted by the *Start* and augmented with *MaxE*, the maximum *End* counting from the first interval. (b) Intervals in the above list with *End* value larger than that of the three following intervals are put into a separated list. The two split interval lists *L1* and *L2* are both ‘smoothed’. (c) Optional structure can be used to obtain the two location indexes of a query in *L1* and *L2* with a single binary search. Two queries [9,12) and [17,21) are discussed in the text.

Table 1: Genomic interval datasets used as database for performance evaluation. The query dataset is chainRn4 with 2,351 thousand intervals. Datasets fBrain-DS14718 and exons are from BedTools site, and others are from UCSC site.

Genomic Dataset Name	fBrain-DS14718	exons	chainOrnAna1	chainVicPac2	chainXenTro3 Link	chainMonDom5Link
No. of intervals (x1000)	199	439	1,957	7,684	50,981	128,187
No. of total overlaps (x1000)	321	2,633	1,086,692	3,892,116	18,432,255	27,741,145

Table 2: Run time of AIList, AITree, NCList and BEDTools for 6 different datasets listed in Table 1. The query dataset is chainRn4 with 2,351 thousand intervals. The time includes the data-loading, data structure construction, searching and result output. BEDTools was terminated for the largest dataset chainMonBom4Link by the system.

Run time (second)	fBrain- DS14718	exons	chainOrnAna1	chainVicPac2	chainXenTro3 Link	chainMonDo m5Link
AIList	0.898	0.995	14.985	27.403	79.387	135.635
AITree	1.235	1.532	24.053	73.670	368.177	581.189
NCList	1.080	1.192	26.094	101.796	419.106	661.759
BEDTools	1.741	2.073	46.533	139.846	1,430.620	---

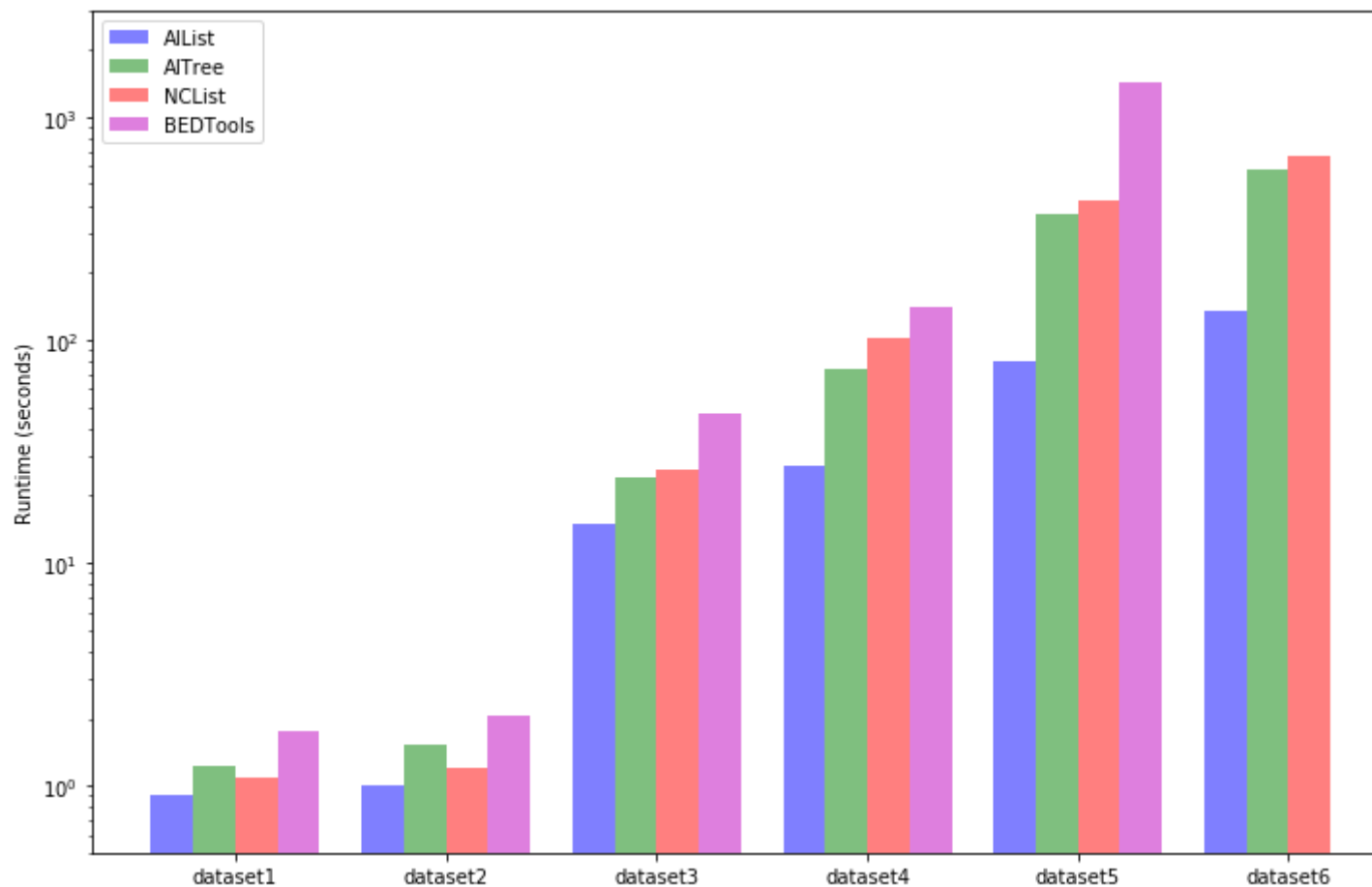


Figure 2. Performance comparison of AIList with AITree, NCList and BEDTools. Six datasets of size ~200K to 128M are used (see Table 1). The time is the wall time including dataset loading, data structure construction and searching. The query dataset is chainRn4 with size of 2,351 thousand. No result for BEDTools for the largest dataset since it took nearly all of the machine memory (16GB) and was terminated.