Supplementary to "Improved representation of sequence Bloom trees" by Harris and Medvedev

In this section, we propose and explore a culling technique for allowing the tree to be non-binary, something that was explored in the original Bloofi paper but was not supported by subsequent SBT approaches. The technique is applied during construction, right after the baseline binary tree topology is constructed but before the bitvector representations are computed.

For a node u, let $saturation(u)$ be the number of bits set to one and active in $B_{\text{det}}(u)$ divided by the number of bits that are active in $B_{\text{det}}(u)$. For example, in Figure 1, the left child of the root has saturation of $9/15$. Since at this stage B_{det} and B_{how} are not yet computed, we estimate the saturation by a sub-sampling technique, similar to that used to estimate node similarity First, we identify the *culling threshold* as two standard deviations below the mean of the saturation value in the internal nodes. This was 20% on our dataset. Second, we scan the baseline topology to identify internal nodes u for which $saturation(u)$ is below the threshold. These nodes are then removed from the topology, with their children reassigned as children of the removed node's parent; if no parent exists, these nodes become roots of a new tree in the forest. We call this process *culling*. The result of culling is that the binary tree potentially becomes a non-binary forest. This does not change the query algorithm in any substantial way.

We investigate the effect of culling on the tree, as a function of the culling threshold (Table S2). At the threshold of 20%, we remove about 4% of the nodes, with a negligible decrease in total index size. The query times fluctuate (Table S3), showing a slight increase or decrease depending on the batch size.

Overall, we conclude that culling did not have a substantial effect on the SBT. It is interesting to observe that when the threshold is high (40%), 45% of the nodes are removed but the index size actually increases by 62%. This is due to the fact that a single active bit in $B_{\text{det}}(u)$ is replaced by two active bits in u's children, if u is removed.

Table S1. Query times (seconds) using a warm cache. Values shown are the median over all the replicates.

	HowDe-SBT	AllSome-SBT	SSBT
single	1.2	27.3	24.9
ten	6.5	16.2	28.8
hundred	41.8	50.0	245.6
thousand	501.2	172.6	3226.6

Table S2. The effect of culling on the tree.

Culling threshold	total n. nodes	n. nodes culled	max depth	n. trees in forest	tree size (GiB)
0%	5169	0	26	1	14.3
10%	5157	12	25	1	14.3
20%	5056	113	23	2	14.3
30%	4679	490	19	8	14.9
40%	3997	1172	16	407	23.2

Table S3. The effect of culling on query time. A cold cache was used.

Culling threshold	median query time (s)			
	singles	tens	hundreds	thousands
0%	5.4	45	171	720
10%	5.1	43	170	713
20%	6.1	39	137	712
30%	7.7	44	155	636
40%	7.4	44	154	625