Supplementary data

Kristoffer Sahlin, Nathaniel Street, Joakim Lundeberg, Lars Arvestad

Illustrations of gap estimations by GapEst for 7 different assemblies on S. aureus are found in Figures 1-7 below. The plots shows gap estimations made by GapEst compared to gaps found by using MUMmer to align contigs to the reference genome.

Using our heuristic link filters, we observed outliers for some assemblies (on the order of 1-10%). These outliers suggests spurious links and we decided to investigate two of them, one over- and one underestimation.

1. **Velvet coordinate (1713,265):** This edge had 617 links with $o \in [2925, 3941]$. Given the true gap of 1713, the links would have a insert sizes between [4638,5654], i.e. $x \in [\mu + 3.5\sigma, \mu + 7\sigma]$. We also have $c_1 + d + c_2 = 65337 >> \mu + 3\sigma$

2. **Bambus2 coordinate (118, 1556):** This edge had 15 links with $o \in [1635, 2138]$. Given the true gap of 118, the links would have a insert sizes between [1753, 2256], i.e. $x \in [\mu - 7\sigma, \mu - 4.5\sigma]$. We also have $c_1 + d + c_2 = 2836 > \mu - 3\sigma$

Two possible explanations of outliers in gap estimation

a Multiple links of insert sizes far out in the tail placing within a region can occur if the library has high coverage and the tails of the insert size distribution is thick.

b Small repeat regions (repeat with respect to the length of the read we are mapping) has been resolved during the assembly is causing links to map wrongly within regions of some contigs.

For the above two gaps where we know the true gap size, we can see that no links with insert size in the range $x \in [\mu - 4.5\sigma, \mu + 3.5\sigma]$ are spanning the contigs. In the first case, the contigs are of sufficient length to contain observations falling within any point in this interval, this strongly advocates for explanation b). In the second case, observations with larger insert size than $\mu - 4.5\sigma$ should be likely to be seen spanning the contigs but it is not that obvious since the range of the contigs with the gap is only $\mu - 3\sigma$. Thus, this can be a cause of either a) or b).

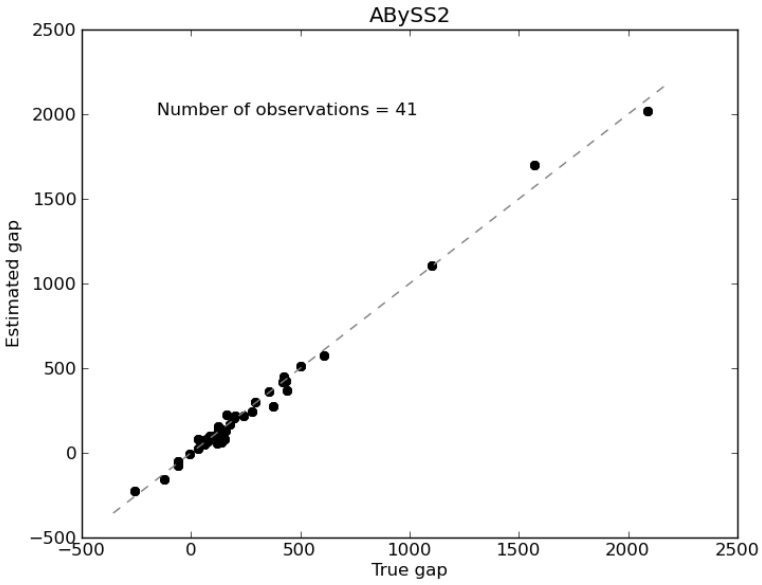

Figure 1: Dot plot of gap estimations for ABySS2 assembly on S. aurens.

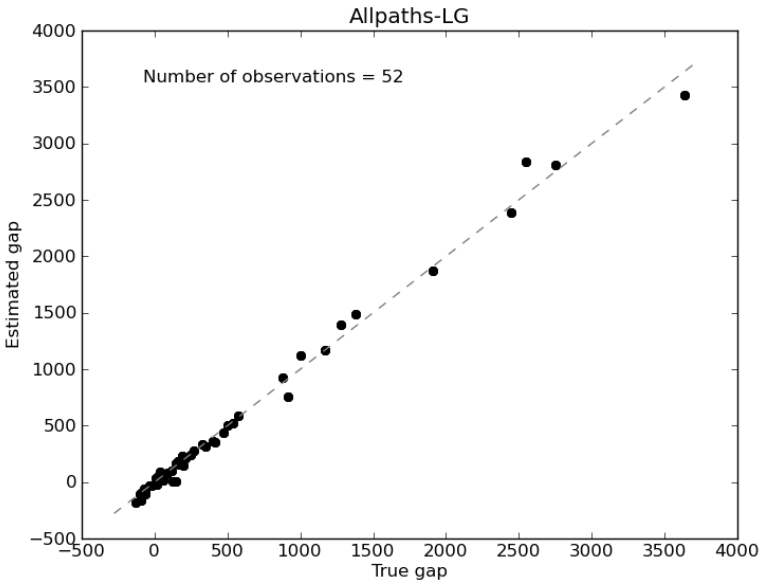

Figure 2: Dot plot of gap estimations for Allpaths-LG assembly on S. aurens.

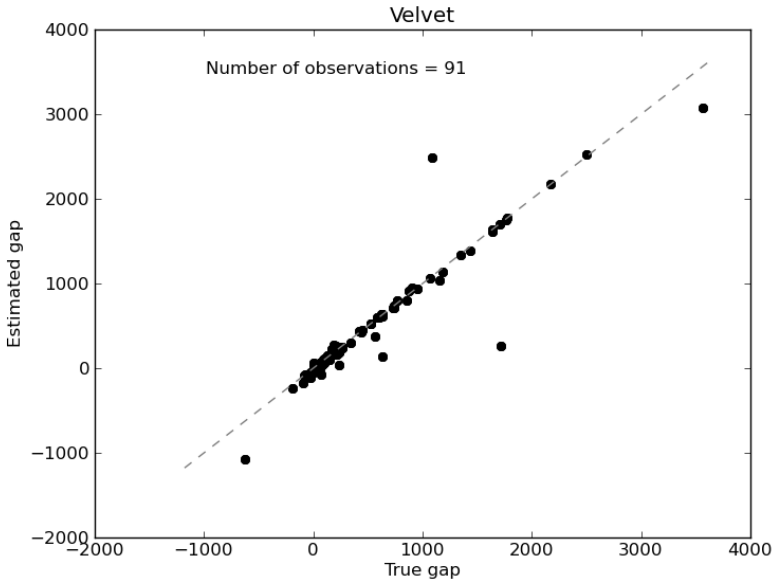

Figure 3: Dot plot of gap estimations for Velvet assembly on S. aurens.

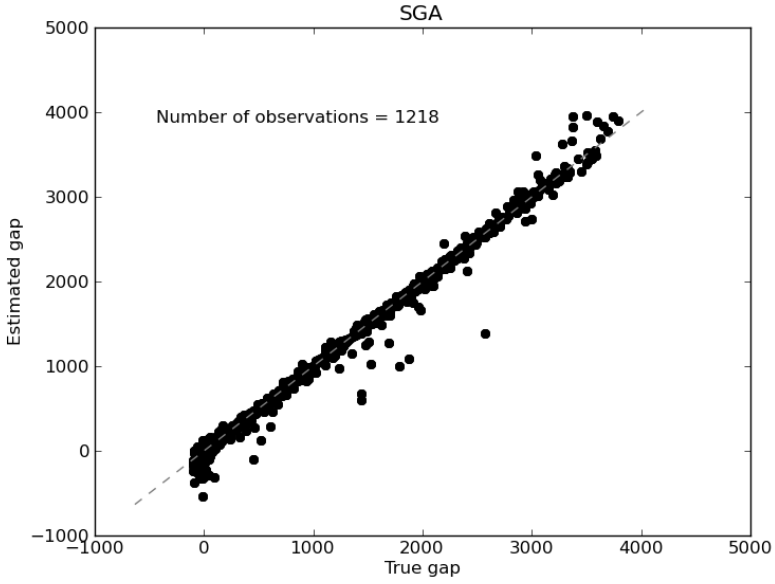

Figure 4: Dot plot of gap estimations for SGA assembly on S. aurens.

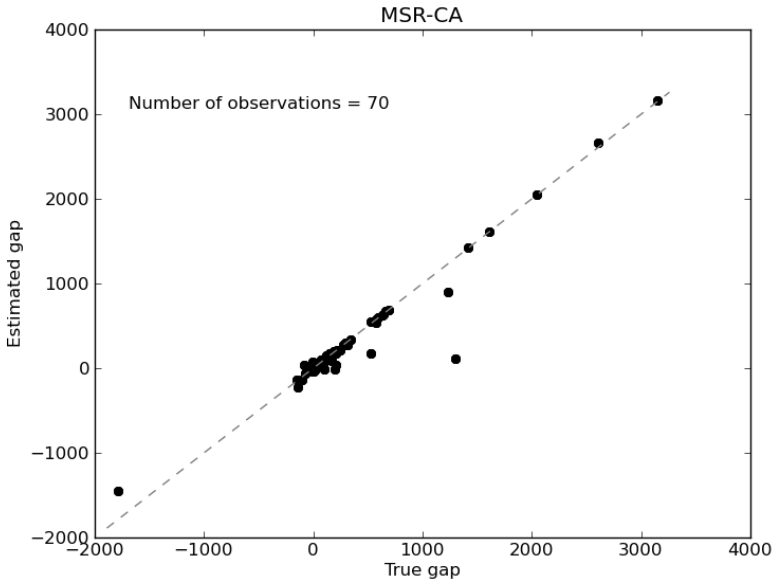

Figure 5: Dot plot of gap estimations for MSR-CA assembly on S. aurens.

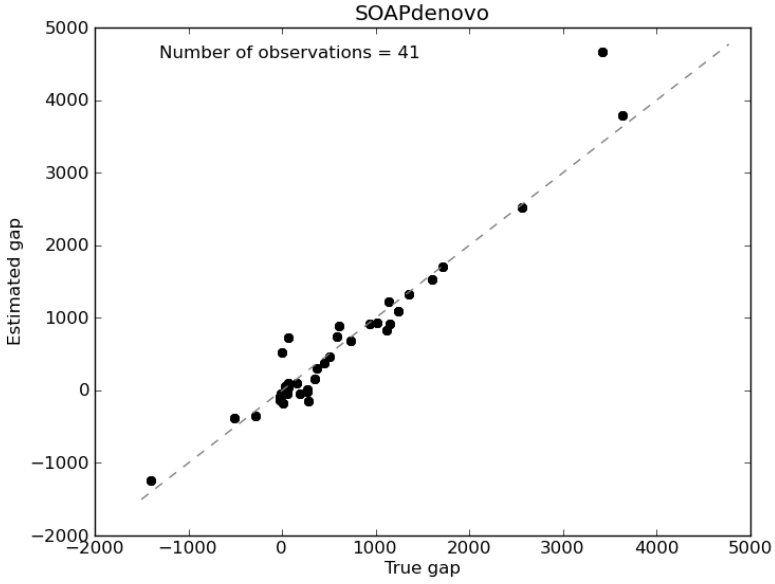

Figure 6: Dot plot of gap estimations for SOAPdenovo assembly on S. aurens.

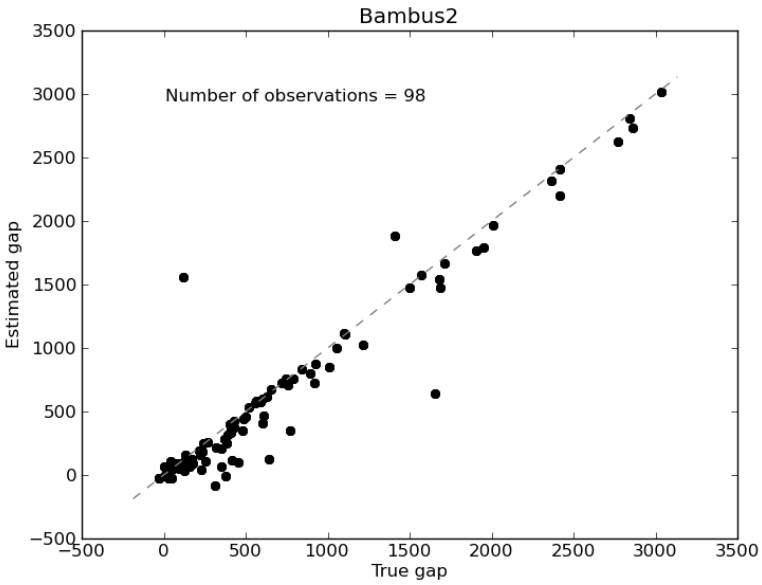

Figure 7: Dot plot of gap estimations for Bambus2 assembly on S. aurens.