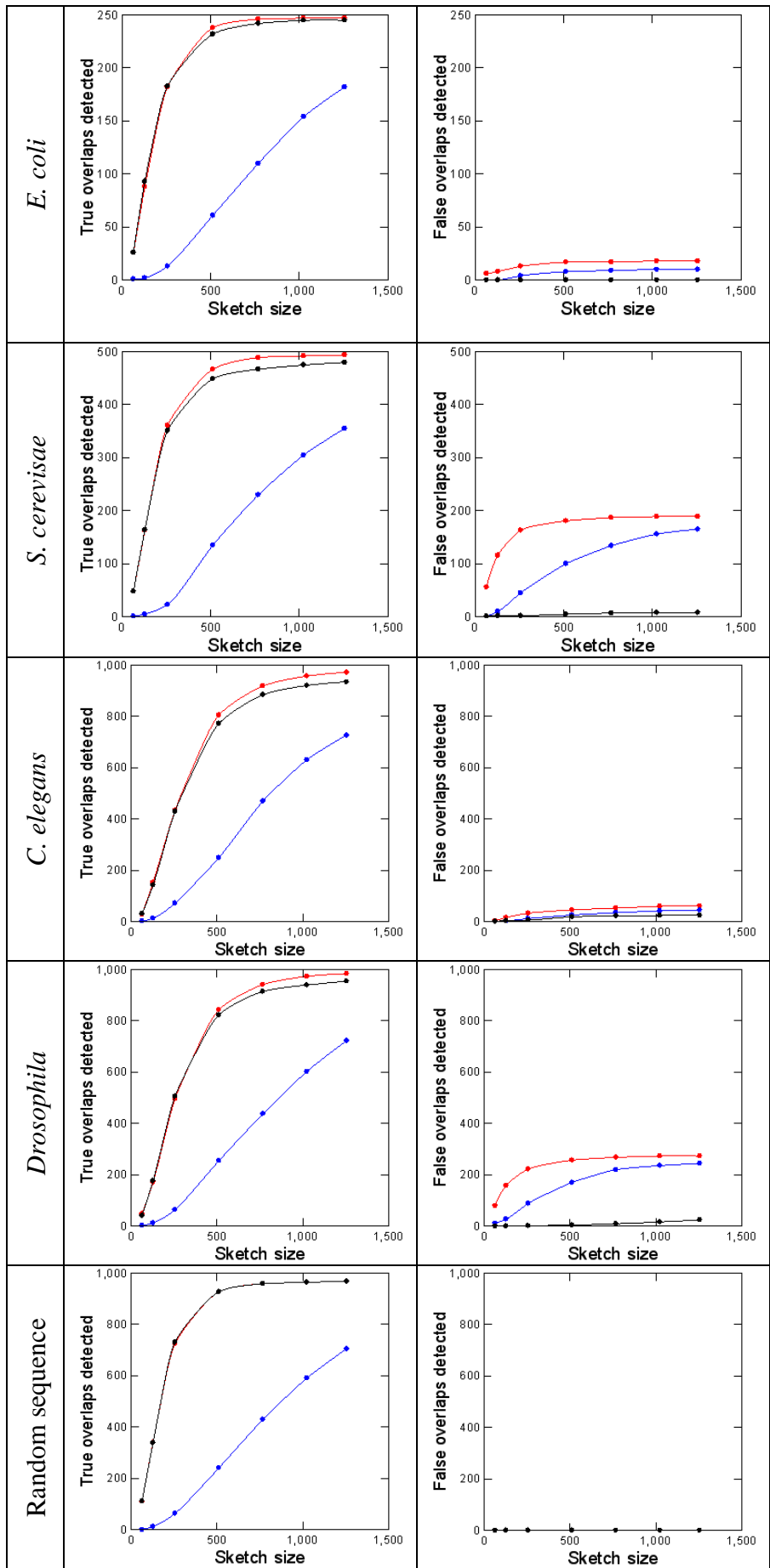


Supplemental Fig. S4



Supplemental Figure S4. Sensitivity and specificity of read overlap detection with masking of repetitive *k*-mers. Simulated PacBio reads from diverse sources were subjected to standard MHAP (blue), MHAP with masking of low-frequency *k*-mers (red), or MHAP with masking of low-frequency and high-frequency *k*-mers (black). Note that masking of low and high-frequency *k*-mers cause a huge improvement in specificity (right panel) with minimal losses in sensitivity (left panel). Note also that with random sequence, which does not contain repeats, false positives are nearly absent (with the three methods). The valid *k*-mer lists came from Illumina reads (top panels), or from the original random sequence data (bottom panel). In the random data, true overlaps with the two types of masking (red and black lines) are nearly identical.