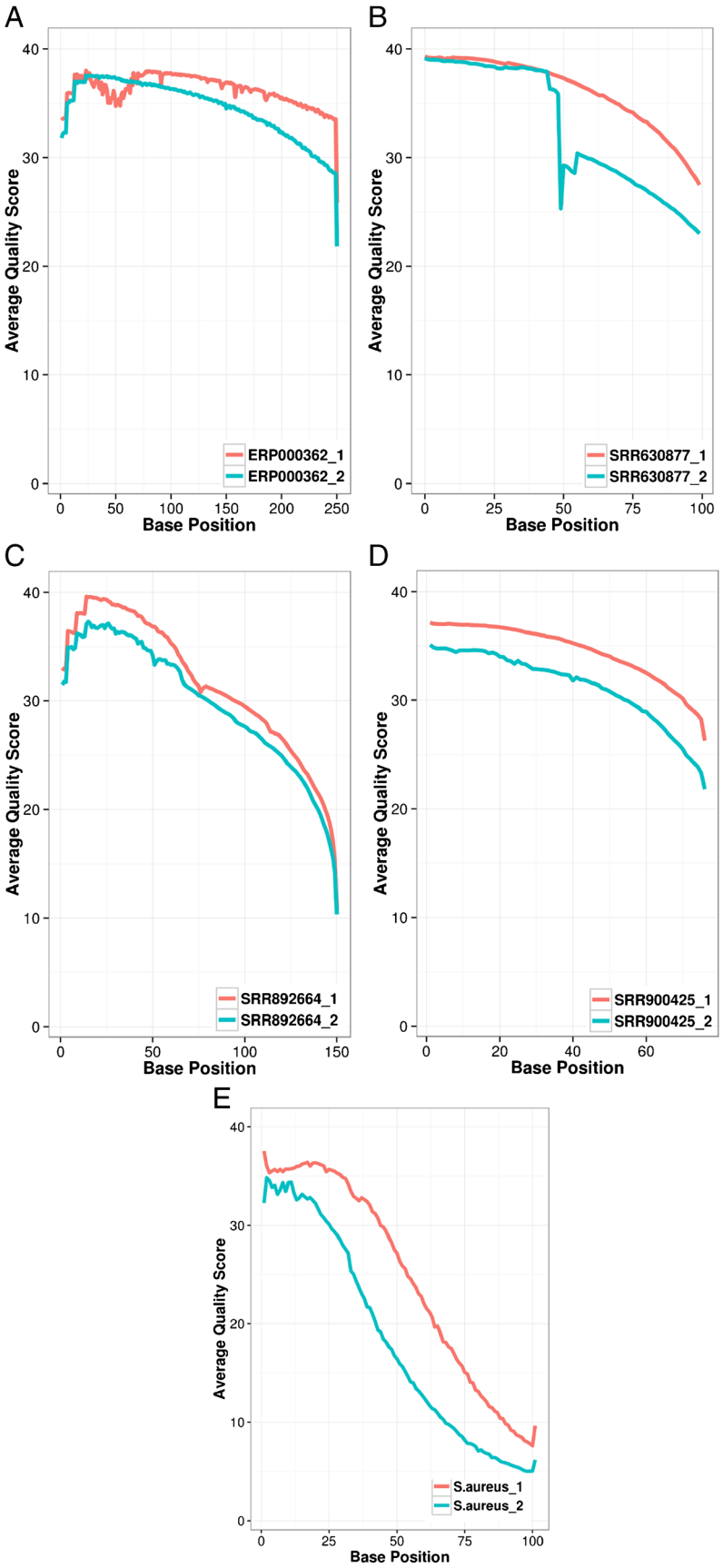
Identifying wrong assemblies in *de novo* short read primary sequence assembly contigs

Vandna Chawla, Rajnish Kumar and Ravi Shankar

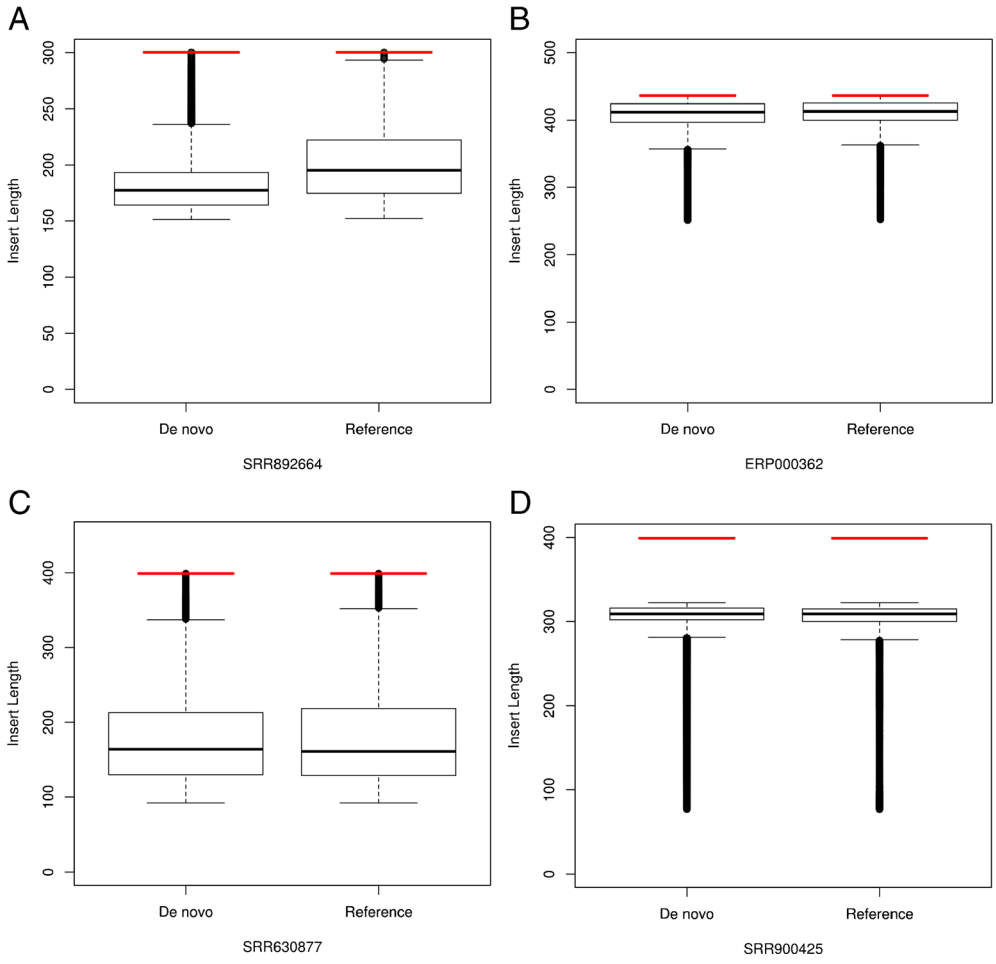
Supplementary material

Supplementary file 1. Summary of datasets collected for Illumina Sequencers' Variants.

Supplementary file 2. b a. HiSeq dataset with SRA accession number SRR892664 and read length 150 bases (*H. sapiens*), b. MiSeq dataset with SRA accession number ERP000362 and read length 250 bases (*H. sapiens*), c. GA IIx dataset with SRA accession number SRR660877 and read length 91 bases (*H. sapiens*), d. HiSeq dataset with SRA accession number SRR900425 and read length 76 bases (*D. melanogaster*), e. HiSeq dataset with SRA accession number SRR892664 and read length 150 bases and GA IIx dataset with SRA accession number SRR660877 and read length 91 bases (*H. sapiens*), f. MiSeq dataset with SRA accession number ERP000362 and read length 250 bases and GA IIx dataset with SRA accession number SRR660877 and read length 91 bases (*H. sapiens*), g. HiSeq dataset with SRA accession number SRR892664 and read length 150 bases and MiSeq dataset with SRA accession number ERP000362 and read length 250 bases (*H. sapiens*), h. HiSeq dataset with SRA accession number SRR892664 and read length 150 bases, MiSeq dataset with SRA accession number ERP000362 and read length 250 bases, GA IIx dataset with SRA accession number SRR660877 and read length 91 bases (*H. sapiens*) and i. GA IIx dataset with SRA accession number SRR022868 and read lengths 63 bases (R1), 42 bases (R2) (*S. aureus*).



Supplementary file 3. Average quality score distribution along the read's length.



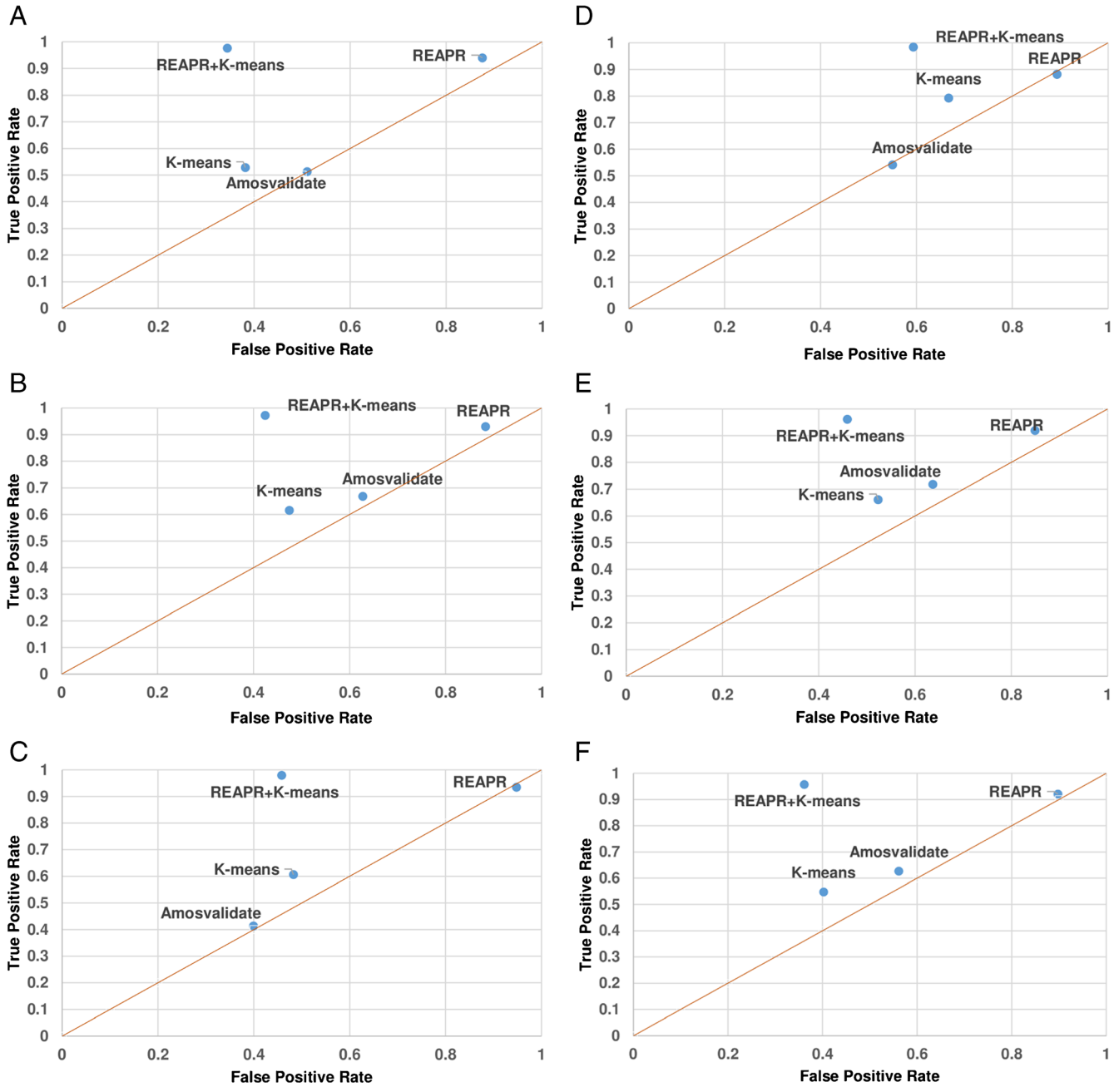
Supplementary file 4. Box-plot representing insert size distribution of paired-end reads. Comparative PE reads distribution with varying insert-size on reference sequences and on *de novo* assembled sequences for each of the four sets: A). *H. sapiens* SRR892664; B) *H. sapiens* ERP000362; C). *H. sapiens* SRR630877; and D). *D. melanogaster* SRR900425 assembled by SOAPdenovo2. Red line represents the given insert size. Dark horizontal lines represent the median, with the box representing the 25th and 75th percentiles, the whiskers, the 5th and 95th percentiles, and outliers represented by dots.

Supplementary file 5. Performance evaluation of assembly validation tools on assemblies from different assemblers. The table represents the values of Sn (Sensitivity), Sp (Specificity), ACC (Accuracy) and MCC for a.) Amosvalidate, and b.) REAPR.

Supplementary file 6. Performance evaluation of assembly validation tools with the set of contigs participating in *K*-means lengthwise clusters from assemblies of different assemblers. The table represents the values of Sn (Sensitivity), Sp (Specificity), ACC (Accuracy) and MCC for a. Amosvalidate, b. REAPR, c. *K*-means and d. *K*-means + REAPR.

Supplementary file 7. Comparative error fraction plot for Amosvalidate, REAPR and *K*-means clustering along variable contig lengths. Comparative error fraction plot showing error rate variation along different lengths' of contigs for *H. sapiens* SRR892664 dataset assembled using five different assemblers namely, ABySS, JR-Assembler, Ray, SGA, Velvet and SOAPdenovo2.

Supplementary file 8. Error fraction values for Amosvalidate, REAPR and *K*-means clustering along variable contig lengths. Error fraction values showing error rate variation along length for A). *H. sapiens* SRR892664, B). *D. melanogaster* SRR900425.



Supplementary file 9. Performance plots for False positive rate (FPR) vs True positive rate (TPR) in ROC space. Results from five different assemblers, A. ABySS, B. Ray, C. JR-Assembler, D. SGA, E. Velvet, and F. SOAPdenovo on *H. sapiens* (SRR892664) are represented. The results from *K*-means, REAPR, Amosvalidate and REAPR+*K*-means from contingency table are plotted as points. A ROC space is defined by FPR and TPR as x-axis and y-axis, respectively. Diagonal line (Orange color) divides the ROC space. Points above the diagonal represent better performance results.