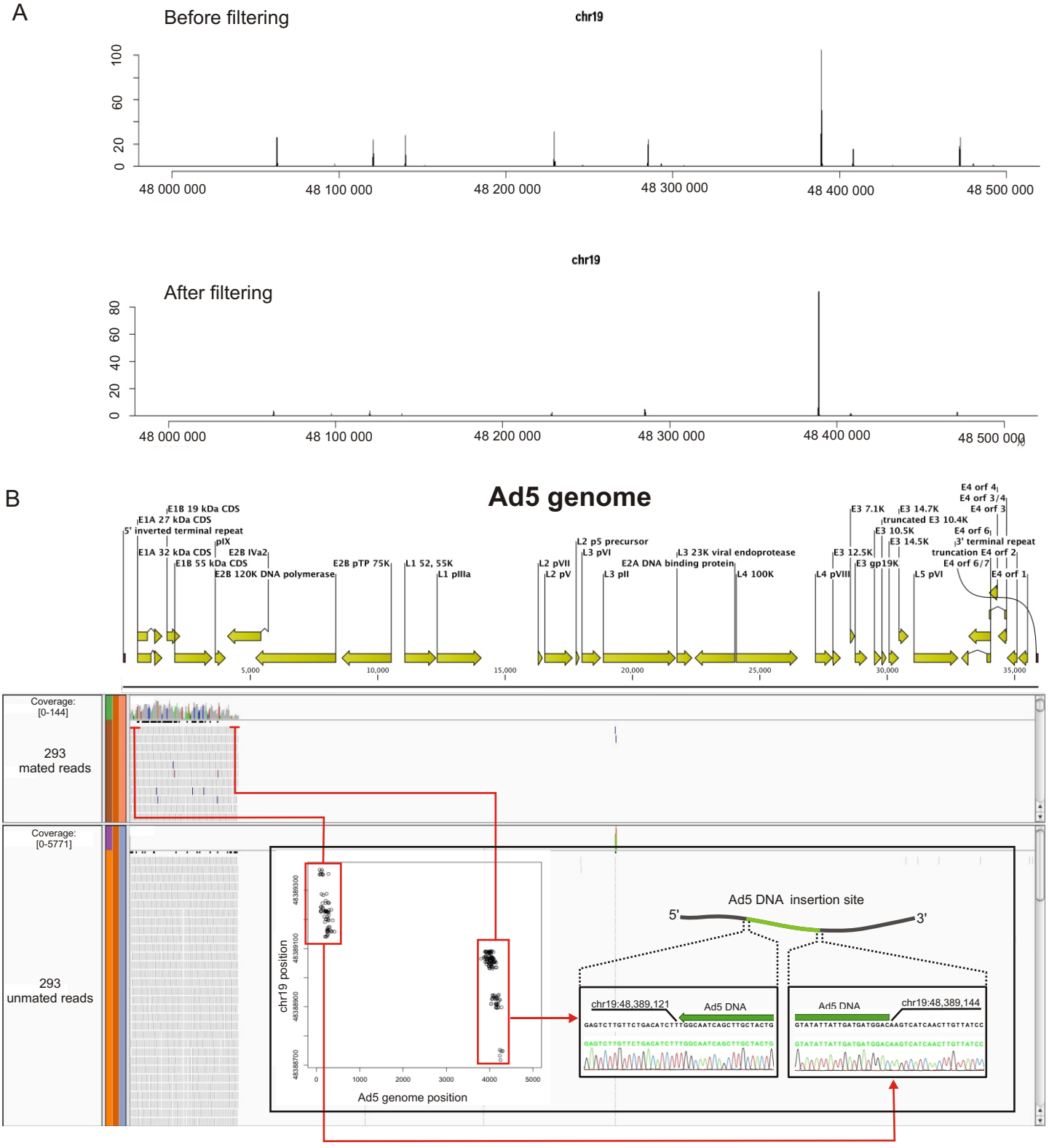
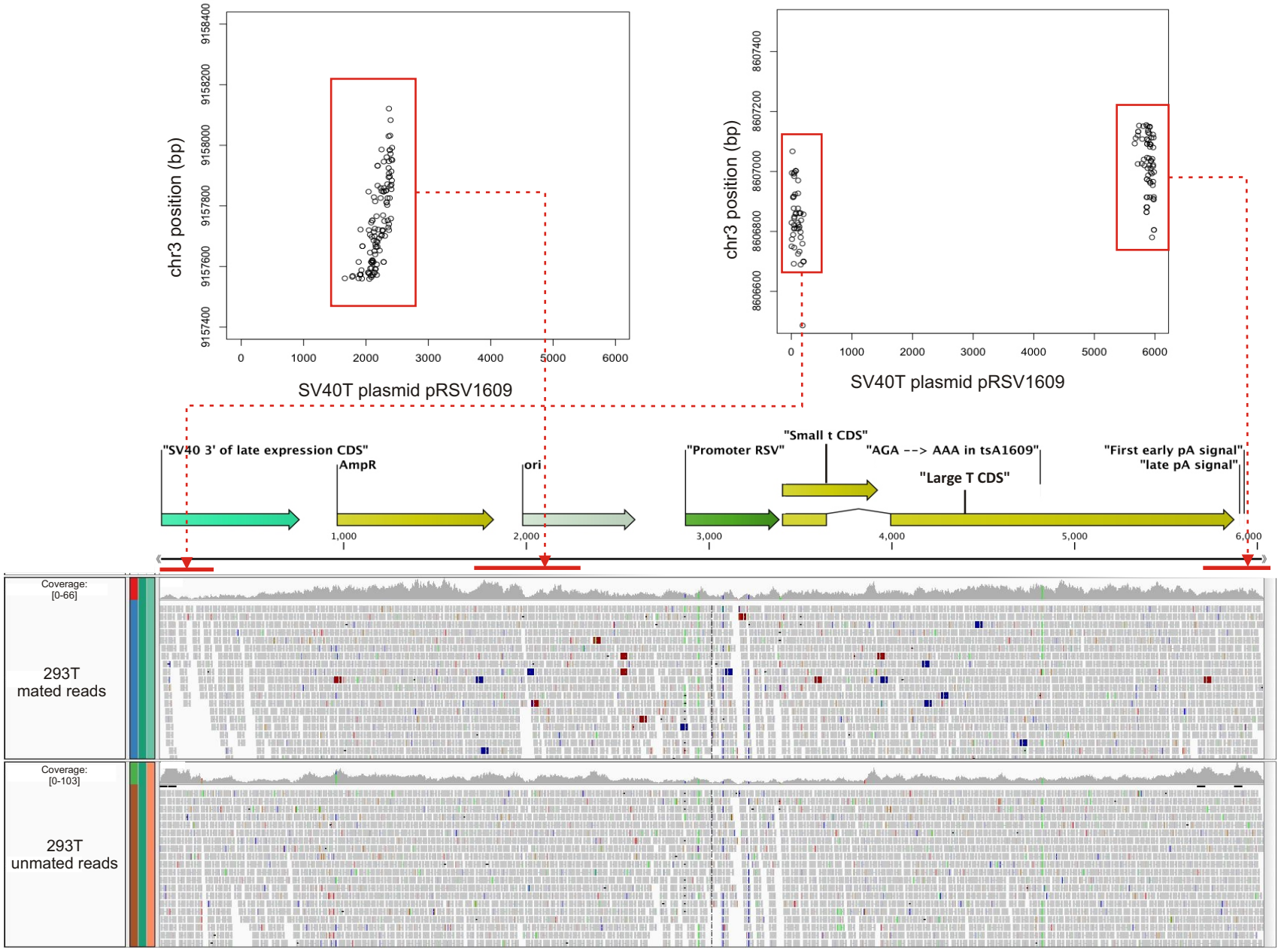
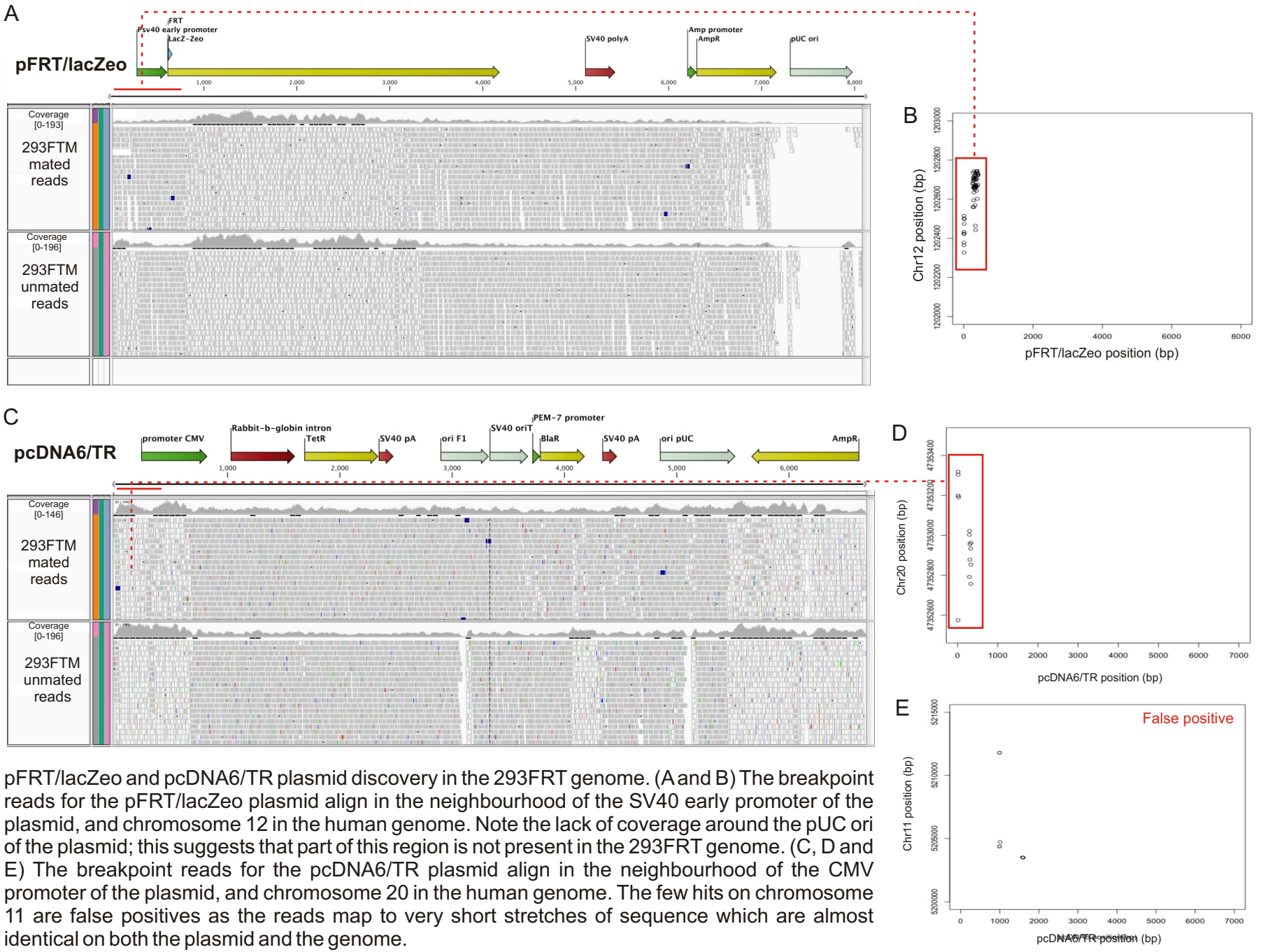


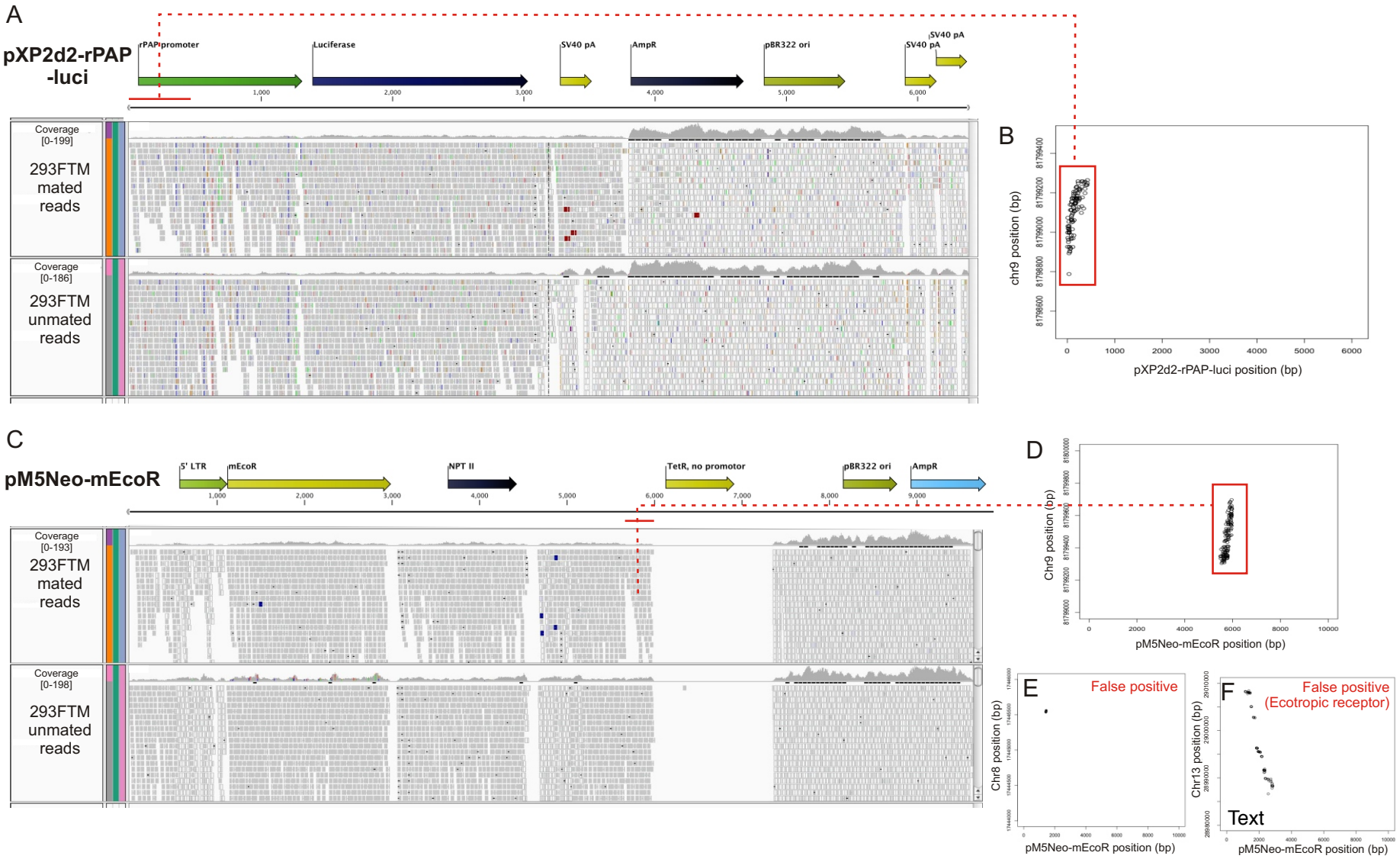
(A) Plasmid insertion site determination plot before and after filtering for reads mapping uniquely to the human genome sequence. The filtering significantly improves the specificity to find the insert site, as illustrated here for the adenoviral insertion in the 293 genome. (B) Adenoviral sequence segment present in the 293 genome. Both mated and unmated reads map to a ~4 kb region of the Ad5 viral genome that includes E1A and E1B coding sequences. Inset: the breakpoint-spanning reads map to a specific region on chromosome 19. The insertion site was confirmed on both sides by PCR and Sanger sequencing of the amplicons. All mapping coverage plots and scatter plots (including those on the next few pages) represent reads not filtered for unique mapping to the human genome, for maximal sensitivity. Coverage is calculated per bp (RTG).



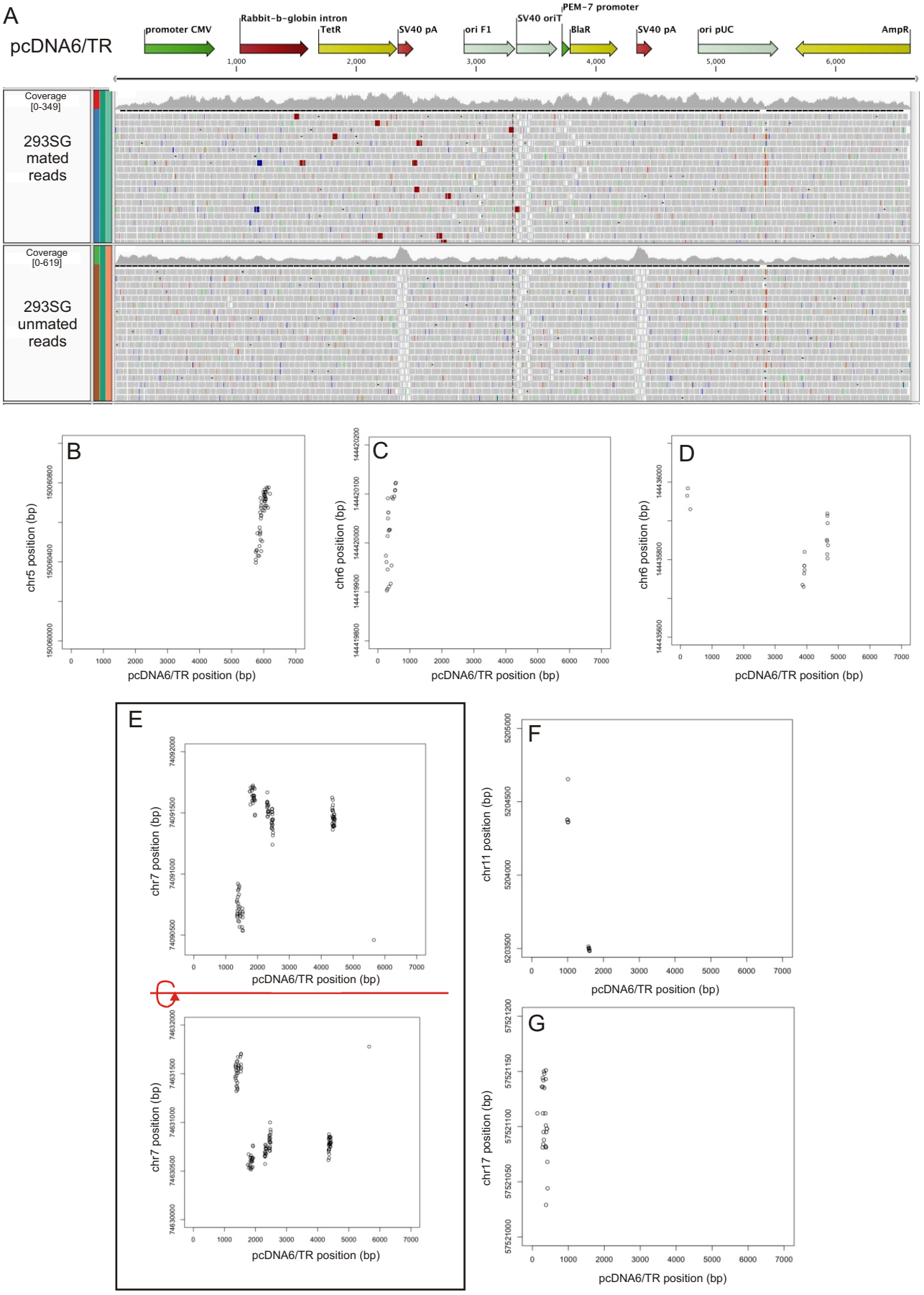
pRSV1609 plasmid insertion site discovery in the 293T genome. Both mated and unmated reads that align to the putative pRSV1609 sequence are illustrated here. We found 2 distinct sets of breakpoint reads, suggesting 2 different insertions on human chromosome 3: one located in the plasmid origin of replication, the other at the end of the Large T CDS and 3' region of its expression cassette.



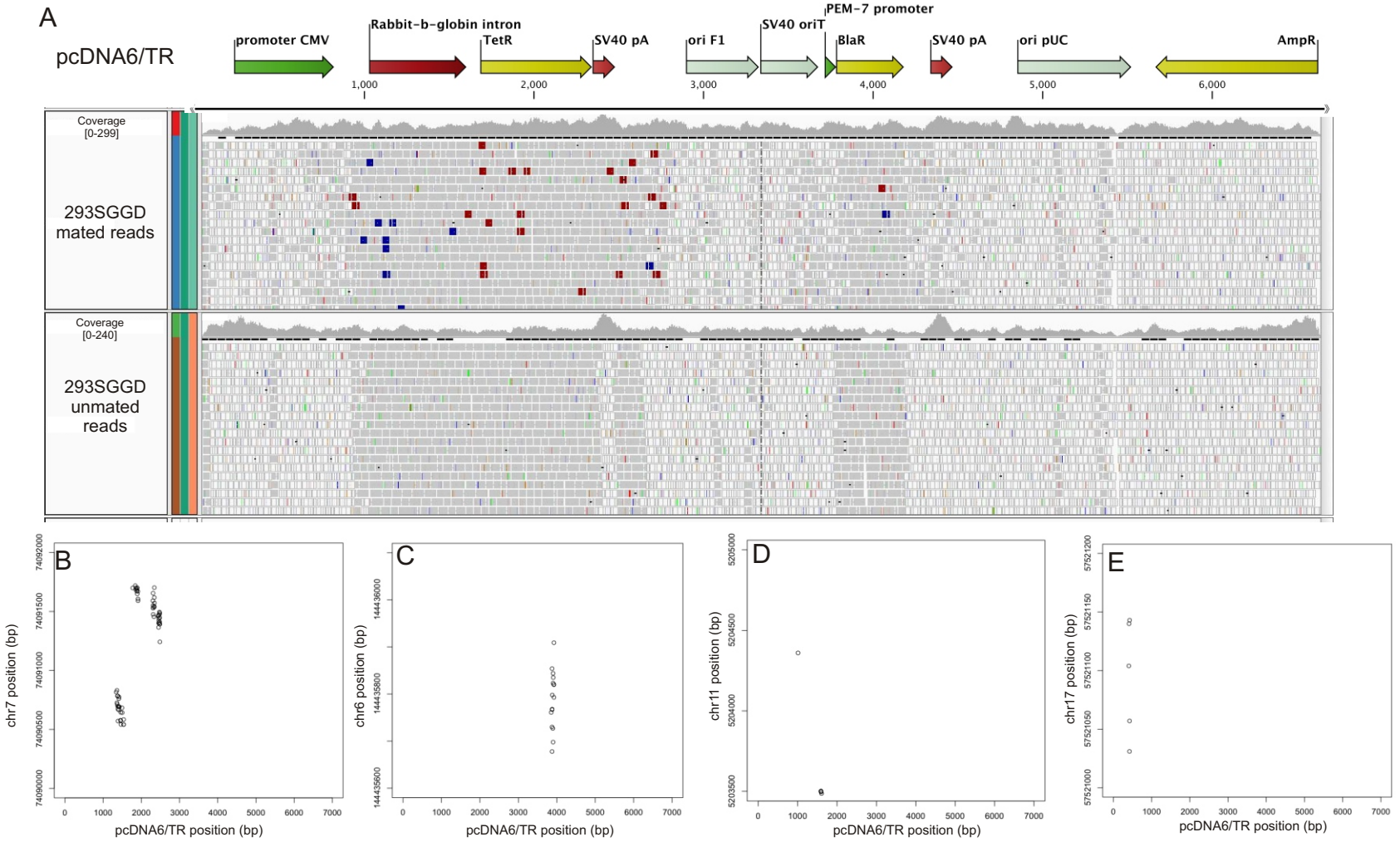
pFRT/lacZeo and pcDNA6/TR plasmid discovery in the 293FRT genome. (A and B) The breakpoint reads for the pFRT/lacZeo plasmid align in the neighbourhood of the SV40 early promoter of the plasmid, and chromosome 12 in the human genome. Note the lack of coverage around the pUC ori of the plasmid; this suggests that part of this region is not present in the 293FRT genome. (C, D and E) The breakpoint reads for the pcDNA6/TR plasmid align in the neighbourhood of the CMV promoter of the plasmid, and chromosome 20 in the human genome. The few hits on chromosome 11 are false positives as the reads map to very short stretches of sequence which are almost identical on both the plasmid and the genome.



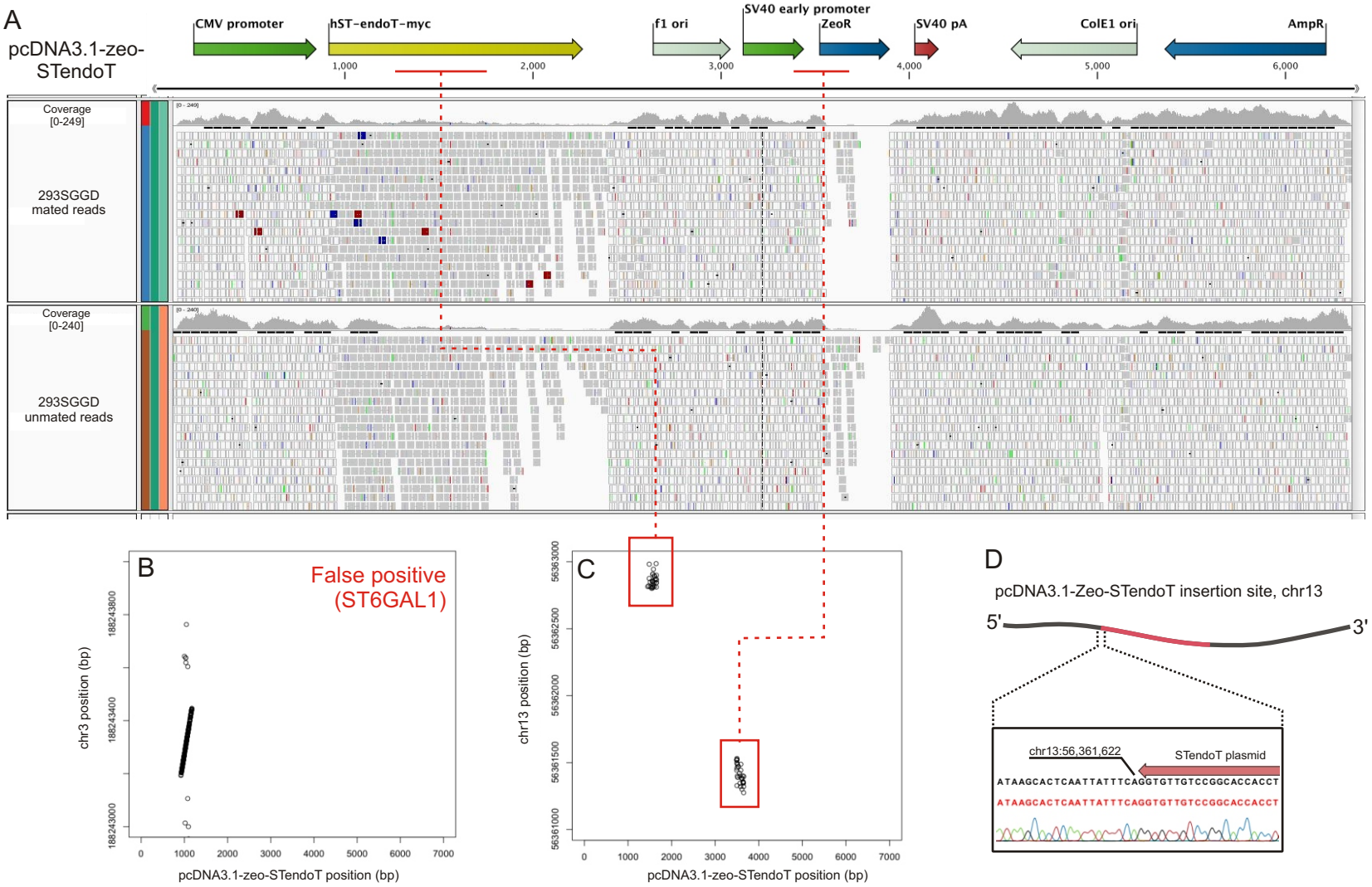
pXP2d2-rPAP-luci and pM5Neo-mEcoR plasmid discovery in the 293FRT genome. (A and B) The breakpoint reads for the pXP2d2-rPAP-luci plasmid align in the neighbourhood of the rPAP promoter of the plasmid, and chromosome 9 in the human genome. (C and D) Similarly, the breakpoint reads for the pM5Neo-mEcoR plasmid align to part of the backbone of the plasmid, and the same position as pXP2d2-rPAP-luci in the human genome. The absence of read coverage between position ~6000-7000 on the plasmid suggests that this part was deleted. The few hits on chromosome 8 are likely false positives (see panel E), while the hits on chromosome 13 are due to the presence of a homolog of the mouse ecotropic receptor in the human genome (panel F). The mEcoR coding sequence is present on this vector. The reads map to the exons of the human homolog in the genome (note the >20 kbp region to which these reads map in a discontinuous manner). See also Figure 5 for the results of the PCR and Sanger sequencing validation of these insertion sites.



pcDNA6/TR plasmid insertion site discovery in the 293SG genome. (A, B, C, D, E, F, G) The entire pcDNA6/TR plasmid sequence is well-covered by both mated and unmated reads. We found several sets of breakpoint reads, matching to positions on chromosome 5, 6, 7, 11 and 17, suggesting that this plasmid is present in multiple sites across the genome. One or more copies of the plasmid has integrated in a inverted repeat, reflected in scatter plots that are each other's mirror image. This region of chromosome 7 is known to be misassembled in the current builds of the human genome. Such complex integrations are difficult to resolve using short-read sequencing technologies.



pcDNA6/TR plasmid insertion site discovery in the 293SGGD genome. (A) The entire pcDNA6/TR plasmid sequence is well-covered by both mated and unmated reads. (B, C, D, E) Breakpoint reads match chromosome 6, 7, 11 and 17.



pcDNA3.1-zeo-STendoT plasmid insertion site discovery in the 293SGGD genome. (A and C) The breakpoint reads for the pcDNA3.1-zeo-STendoT plasmid suggest 2 different insertion sites on chromosome 13, one in which the plasmid breakpoint is located in the middle of the hST-endoT-myc cassette of the plasmid which would disrupt the function of this transgene. As 293SGGD cells express functional EndoT protein activity (manuscript in preparation), the other integrated plasmid copy must be responsible for the expression of hST-EndoT in this cell line. We were able to validate one breakpoint of that copy by PCR and Sanger sequencing (panel D). Additionally, we found one set of breakpoint reads that map to chromosome 3; these are false-positive reads associated with the sequence identity of the first section of the hST-EndoT CDS.