TRv1.0 regions outside HPRC draft benchmark

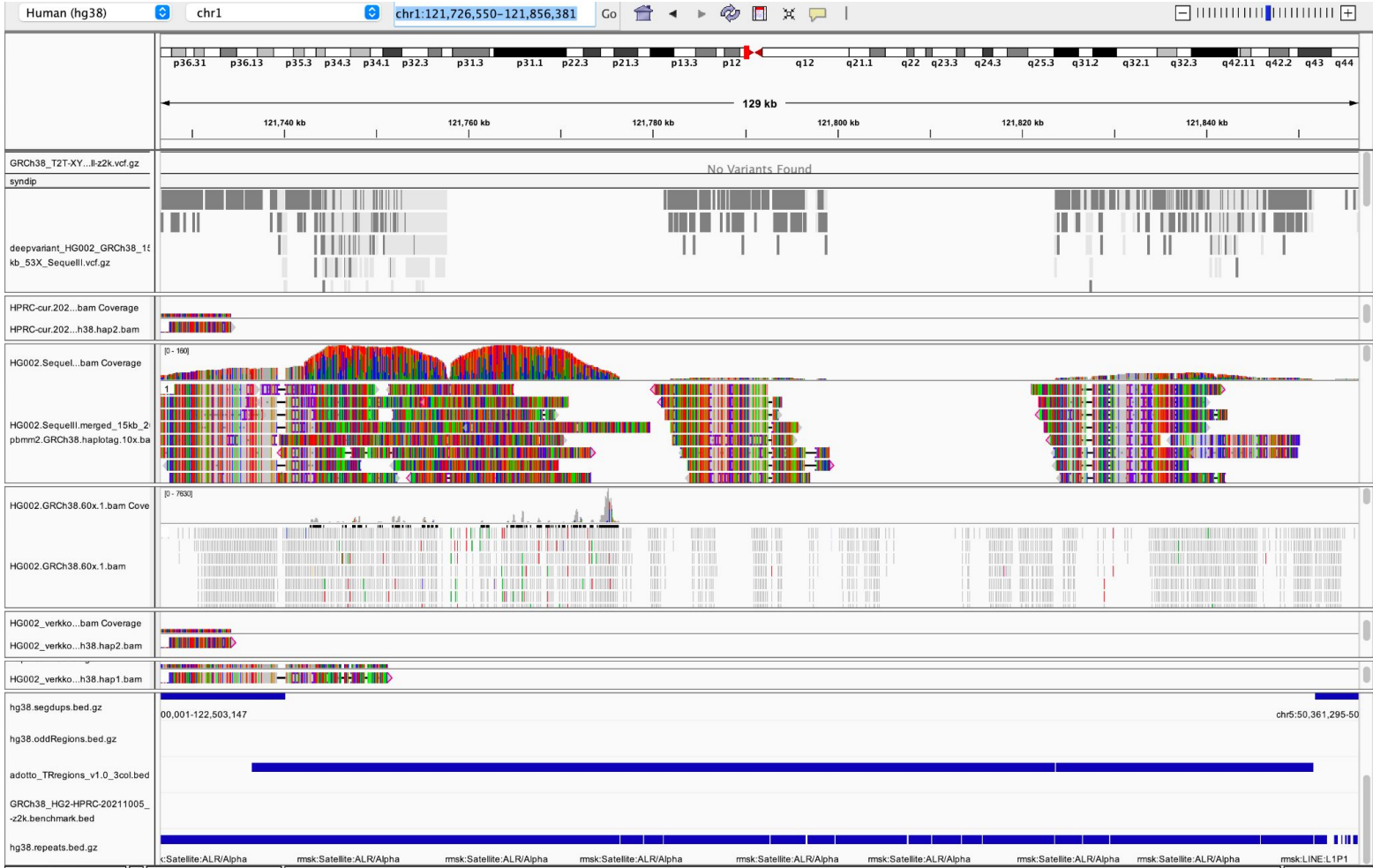
Email description

Essentially, I wanted to see how many of the TRv1.0 regions were included in our draft SV benchmark regions from the curated HPRC assembly. When developing our benchmark regions, we take the dip.bed file from dipcall (which only keeps regions with 1 contig from each haplotype), and then exclude any large repeats (i.e., segdups, TRs>10kb, satellites, gaps) + 15kb on each side, if there is a break in the dip.bed anywhere within the repeat+15kb. This excludes benchmark variants around very large SVs/CNVs/assembly errors that cause a break in the alignment or extra contigs to align. The result is the attached bed.

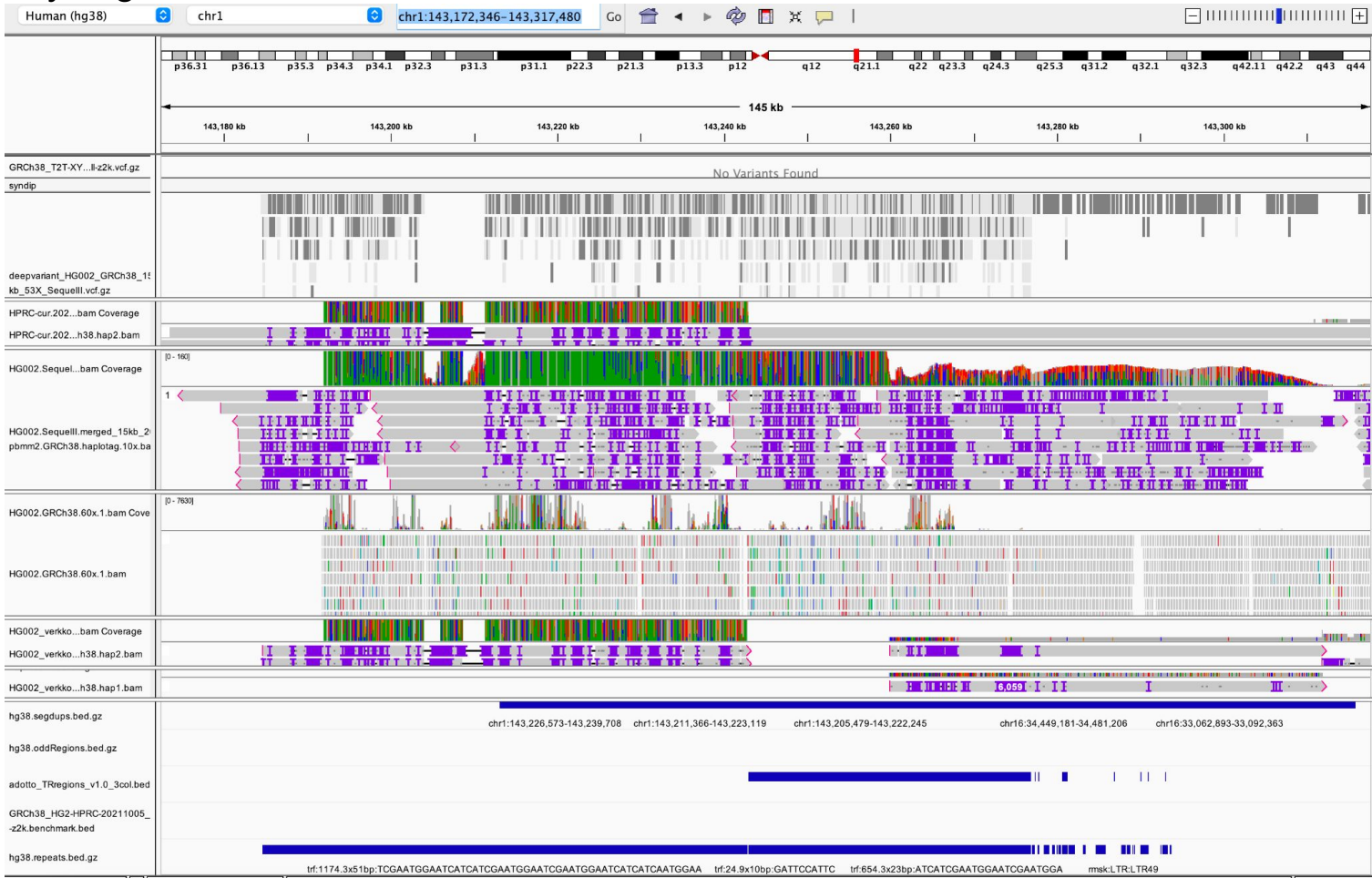
It looks like 1772496 out of the 1784804 (99.3%) TRv1.0 regions are inside our draft benchmark regions, and 219654072 out of the 237865075 (92%) bases in the TRv1.0 regions are inside our draft benchmark regions. Large TRs are preferentially excluded, with over half of the excluded bases in TRs longer than 2kb and ~1/4 longer than 20kb. Looking at a few of the large TRs excluded by our bed, it looks like most are near gaps in GRCh38 or in satellites in the centromere, where there are nearby breaks in the alignment, so I think they likely should be excluded from the benchmark. The smaller TRs that are excluded tend to be inside segdups where there is a break in the alignment due to large SVs or CNVs, so I also think it makes sense to exclude these. I'm guessing this will help with excluding some of the cases that are really hard to benchmark also, since there are sometimes adjacent VNTRs or gaps.

chr1:2,743,906-2,785,106 VNTR next to gap

chr1:121,726,550-121,856,381 satellite flanked by segdups



chr1:143,172,346-143,317,480 VNTR adjacent to another VNTR that doesn't seem to be covered by the assembly alignments



chr1:12,984,001-13,414,271 complex inverted segdups next to gap with many small TRs