Novoalign / BWA MEM Differences

Terms

ALT	From BWA refers to alternate scaffold sequences
REF	Main reference sequences
REF region	A region on the main sequences that has one or more alternative scaffolds
Alternate mapping	Refers to any mapping on an ALT sequence or a REF region with alternative scaffolds.
Surrogate mapping	Refers to any mapping that has an alternate mapping in the same REF region with a better alignment score.

Key Differences

1. Supplementary flag is set differently.
    - BWA sets supplementary on all mappings to ALT when there is a mapping to REF
    - For Novoalign the best mapping is not supplementary even if on ALT. Surrogate mappings are supplementary.
2. MAPQ is calculated differently
    - BWA the MAPQ of a REF hit is computed across REF hits only. The MAPQ of an ALT hit is computed across all hits.
    - In Novoalign the MAPQ is computed excluding surrogate mappings. MAPQ of surrogate mappings exclude any alternate mappings for the same REF region.
3. Novoalign adds a tag ZA:I: with a new MAPQ calculated from all mappings.
4. Primary flag is set differently. In Novoalign primary flag is set for the best alignment, all surrogates to the best alignment and on the best REF mapping (except for improper pairs). It is possible to have primary and supplementary set.
5. What mappings get reported is likely different. Novoalign’s reporting options for multi-mapped reads with ALT mappings have changed. For purpose of counting reported alignments and identifying multi-mapped reads surrogate mappings are not counted.

Illustrations

One significant difference is how the supplementary flag is set. BWA sets supplementary for any alignment to the alternate scaffolds and resets for all mappings to the main sequences.
Novoalign can set supplementary on mappings to the main sequence or to alternate scaffolds. When we have a mapping for a read to an alternative scaffold and to a main sequence the one with the better alignment score has supplementary flag reset and higher scoring alignments are supplementary. This can be seen in the two IGV screen shots below. These are for one gene in Chr 7 and it's alternate sequence KI270803V1.

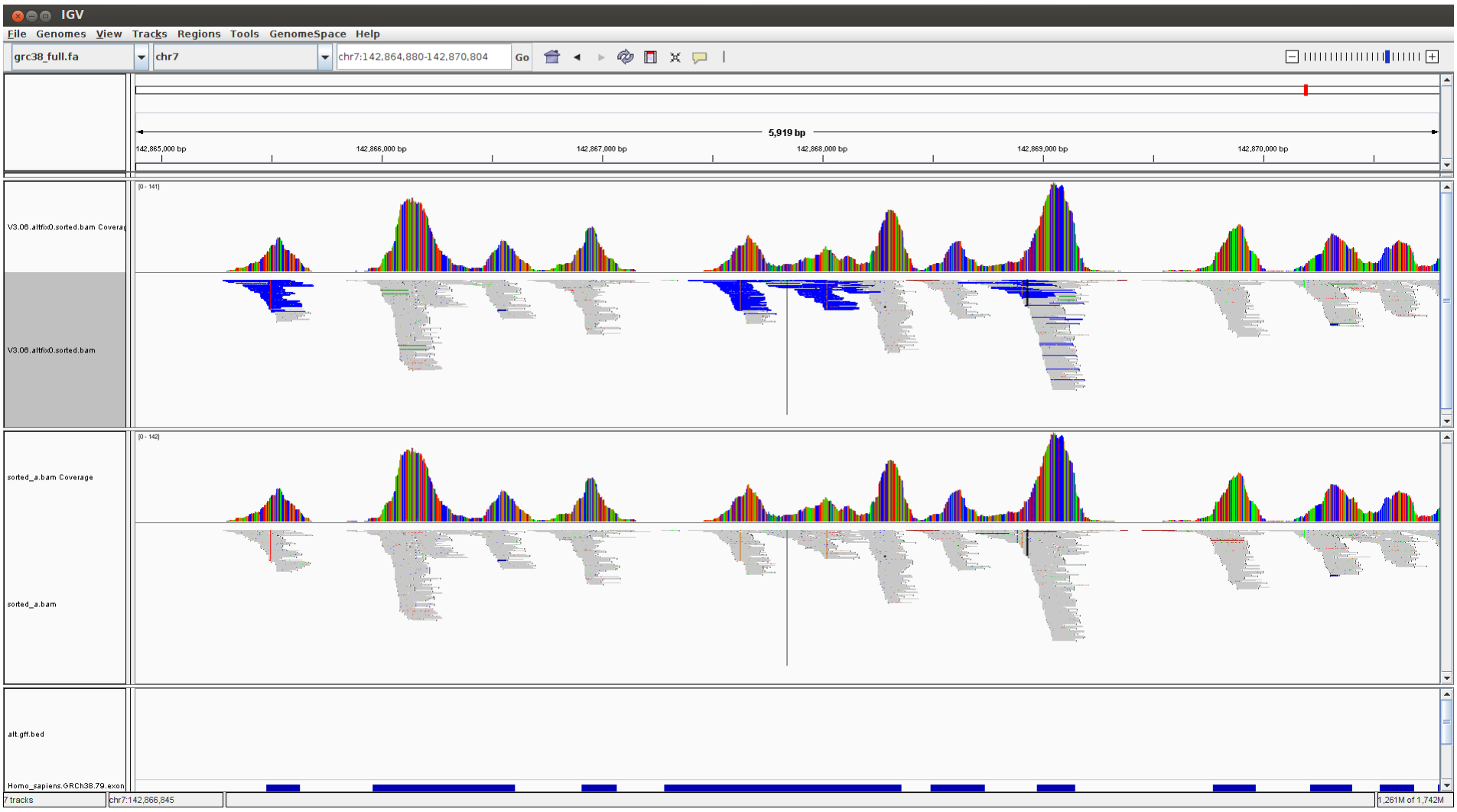

Illustration 1: Mappings to main sequences which have a better alignment on an alternate scaffold are flagged as supplementary (blue) by Novoalign. Novoalign is upper tracks, BWA MEM the lower tracks.

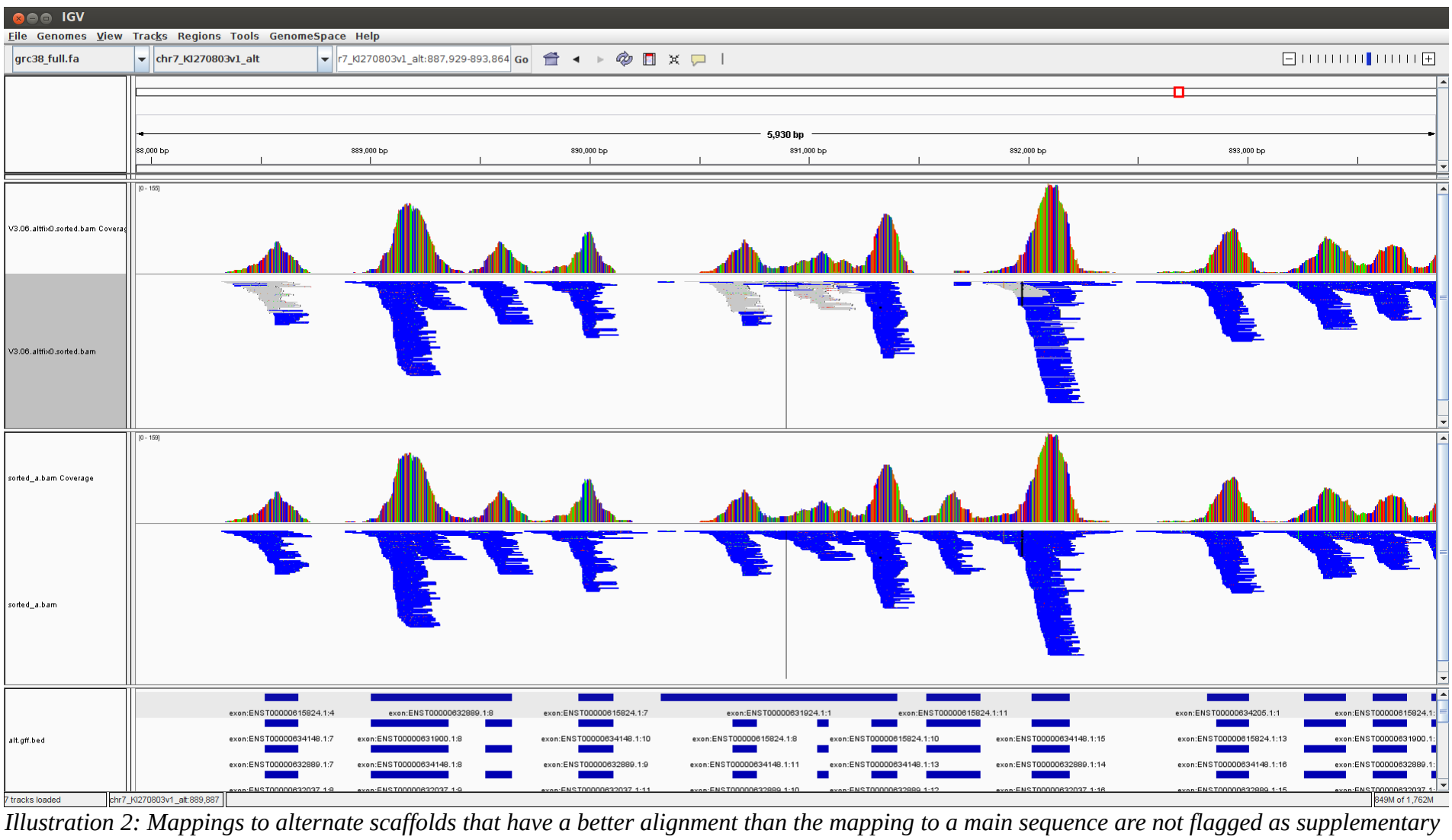

Illustration 2: Mappings to alternate scaffolds that have a better alignment than the mapping to a main sequence are not flagged as supplementary (grey) by Novoalign. Novoalign is upper tracks, BWA MEM the lower tracks.

The next example is an exon from chromosome 6 at chr6:32,521,700-32,522,308 where a large portion of reads map to an alternate scaffold with a much better alignment score. It appears the sample is heterozygous to the main sequence exon and the alternate scaffold exon. Due to the large difference in alignment score Novoalign has not created supplementary alignments to the main sequence copy of the exon.

Looking at the mappings on the alternate scaffold for this exon we can see Novoalign has mapped the reads as primary proper pairs while BWA has mapped the reads as supplementary and each read has it's mate set to the mapping location on the main chromosome.

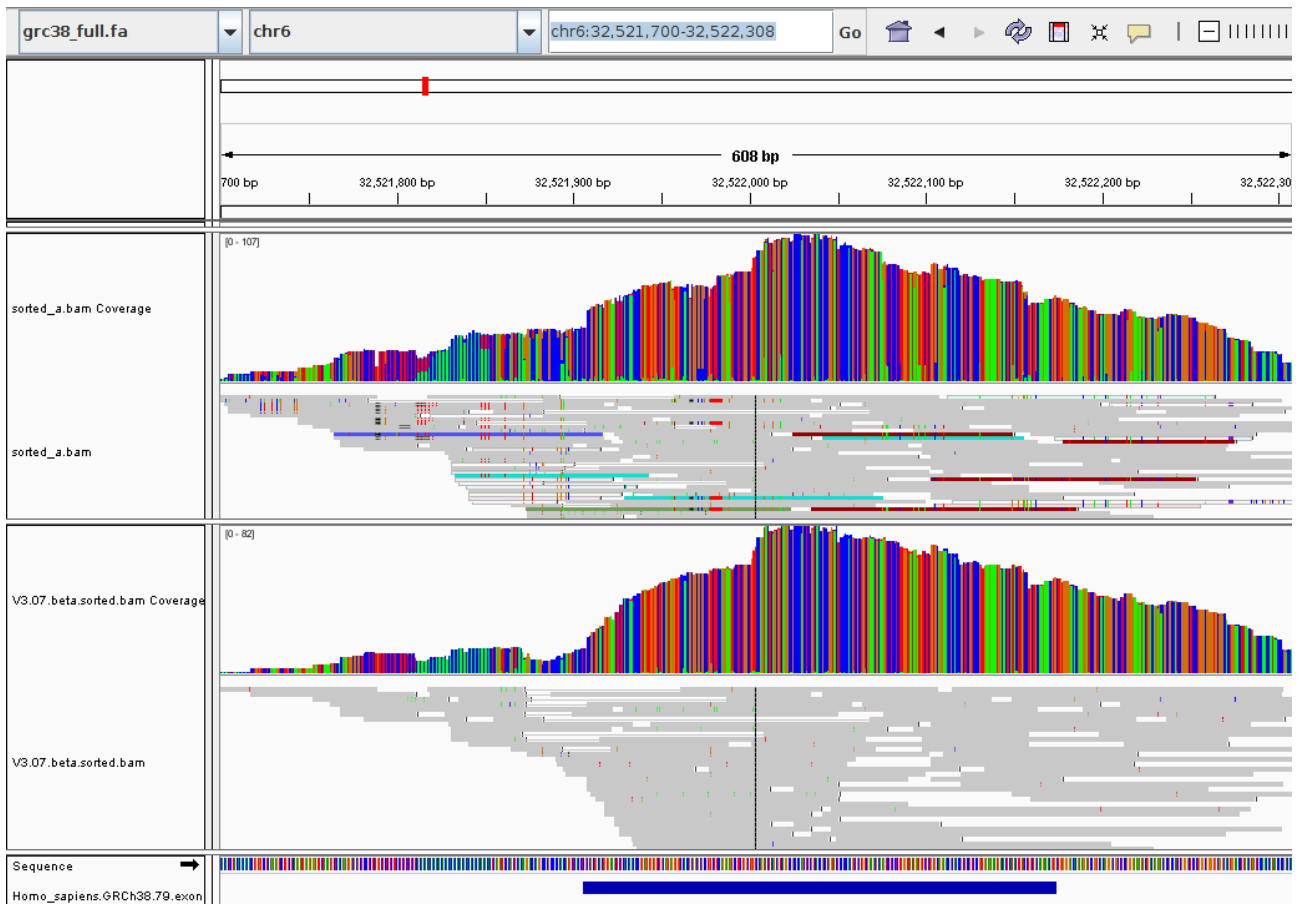

Illustration 3: For Novoalign, reads that mapped to the alternate scaffold do not have a main sequence mapping due to the large difference in alignment score. Upper tracks are BWA mappings, lower are Novoalign.

```
NS500134:53:XXXXXXXXX:3:21612:25474:9191  2209  chr6_GL000251v2_alt     3939048 60  148M  chr6  32584416    0 ...
NS500134:53:XXXXXXXXX:3:21612:25474:9191  2129  chr6_GL000251v2_alt     3939307     60    151M  chr6  32521982    0 ...
```

So for BWA reads are flagged as Primary, Proper Pair and Supplementary with a TLEN of zero.

Novoalign SAM recorded the mapping to the alternate sequence as primary, proper pair and with mate alignment also on the alternate scaffold and with TLEN set appropriately.

```
NS500134:53:XXXXXXXXX:3:21612:25474:9191  163   chr6_GL000251v2_alt     3939048 70  148M  =     3939307     410 ...
NS500134:53:XXXXXXXXX:3:21612:25474:9191  83    chr6_GL000251v2_alt     3939307     70    151M  =     3939048     -410 ...
```

Any analysis of alignments will require establishing which alternate sequences are present. A naive approach is to run…

```
samtools view  -F 2304 alignments.sam |
grep  _alt.*= | cut -f 3 | sort | uniq -
c
```

Which is just counting how many primary non-supplementary alignments each alternate scaffold has.

For WES we suggest either counting these mappings by exons or genes or using the ZA tag to calculate a Bayesian posterior probability that the alternate form is present.

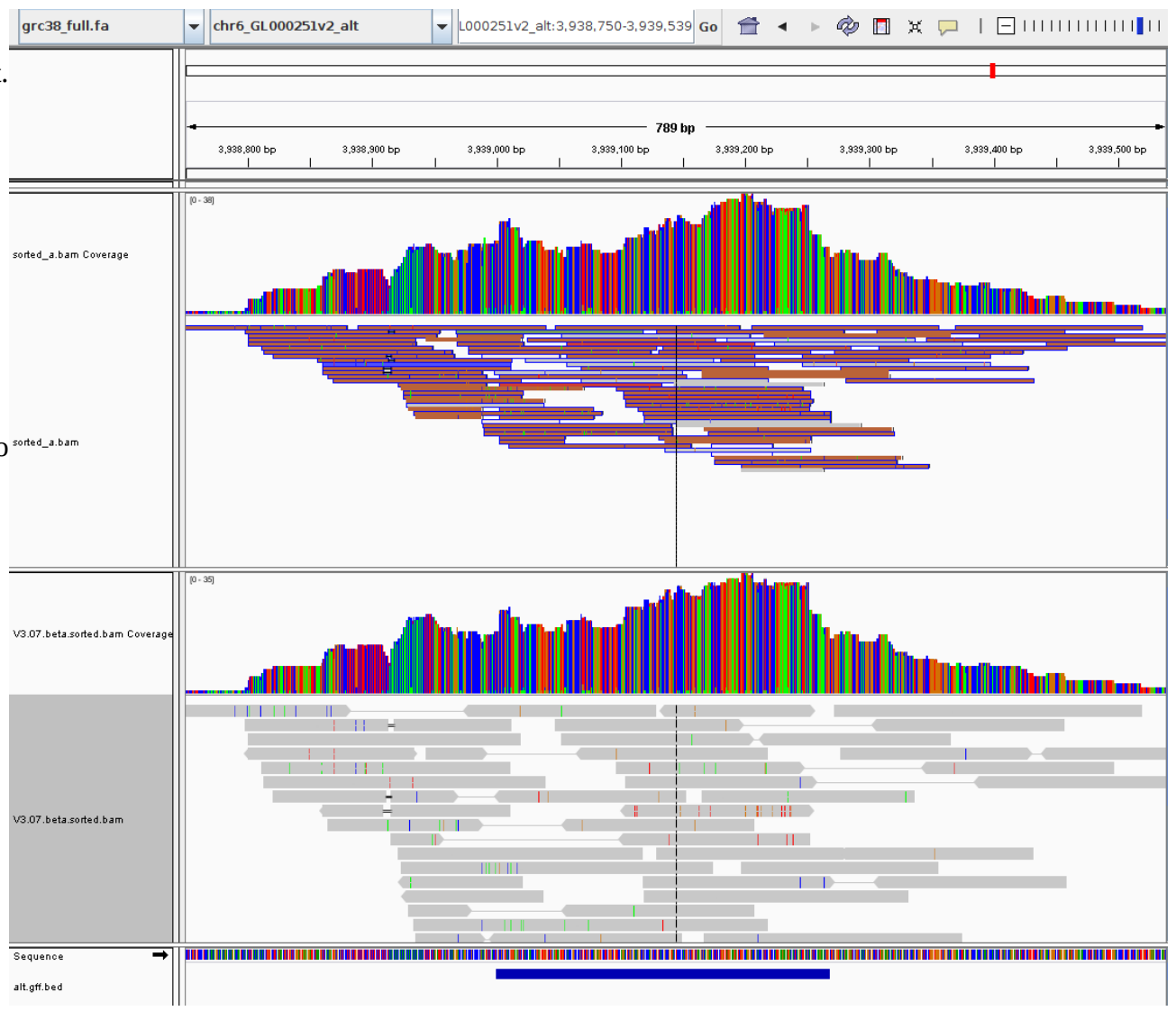

Illustration 4: Mappings to the alternate scaffold for this region. BWA is upper track, Novoalign the lower one.

Program: bwa
Version: 0.7.12-r1039

Program: novoalign
Version: V3.07.beta