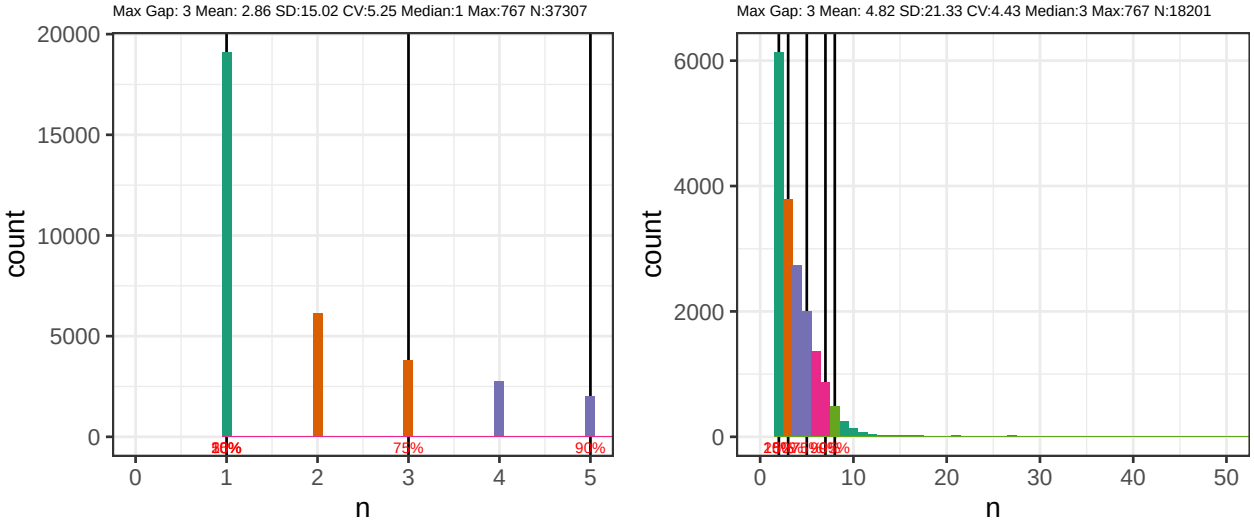
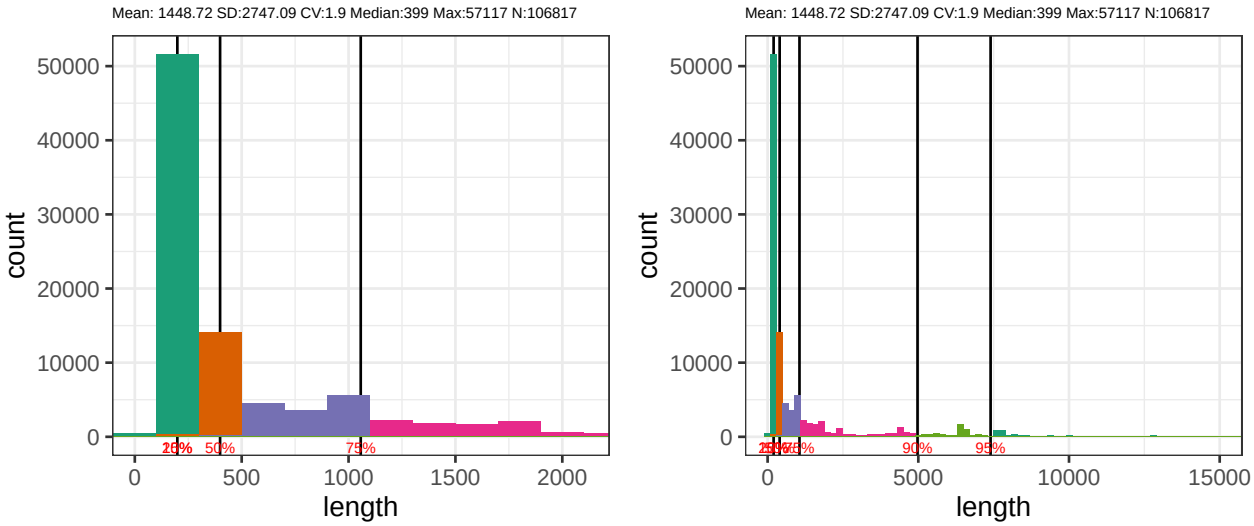
Analisis of CNV stiching improvments

Ricardo H. Ramirez-Gonzalez

The analysis corresponds only to the values for the second part of chromosome 6D. One of the assumptions that we have is that CNVs are more likely to be true if they appear on more than one line. Hence, I look at the distribution to see how many unique events we can observe. Half of the event are unique, this may be rare events or noise.



Another general validation is the length of the deletions. Very long deletions (over 20,000bp) are not detected yet. This is because the current version does not cross across genes that are farther than 2,000bp (but the next iteretion will). The expectation is that events will be longer than 200bp, as events of a single window are more prone to be noisy. Again, we have 75% of the events containing more than individual window.

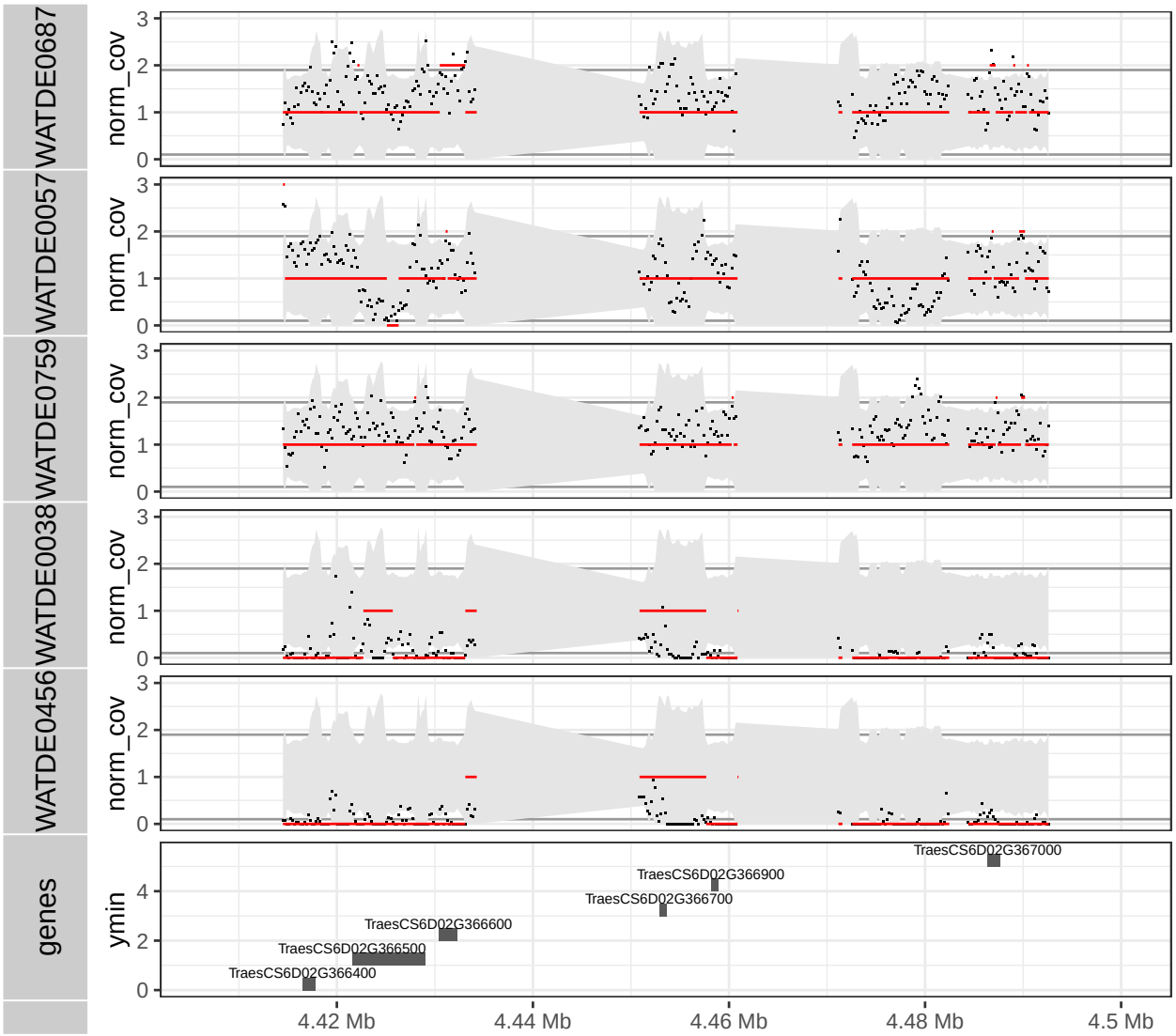


Exploring some events.

To understand detail how the CNV detection algorithm is behaving, we plotted the normalised coverage and the stiched deletion events. The plot contains a couple of horizontal lines, definding the maximum variation in a window that can be considered as reliable. This is a range between $(0.1, 0.9)$. The shadow area represents the 2.5σ confidence interval. The regions where the shadow is above the line represent windows with low confidence, hence those windows are not used in this version of the analysis. To improve the analysis, we need to get the alignments of only the reads that have single mapping.

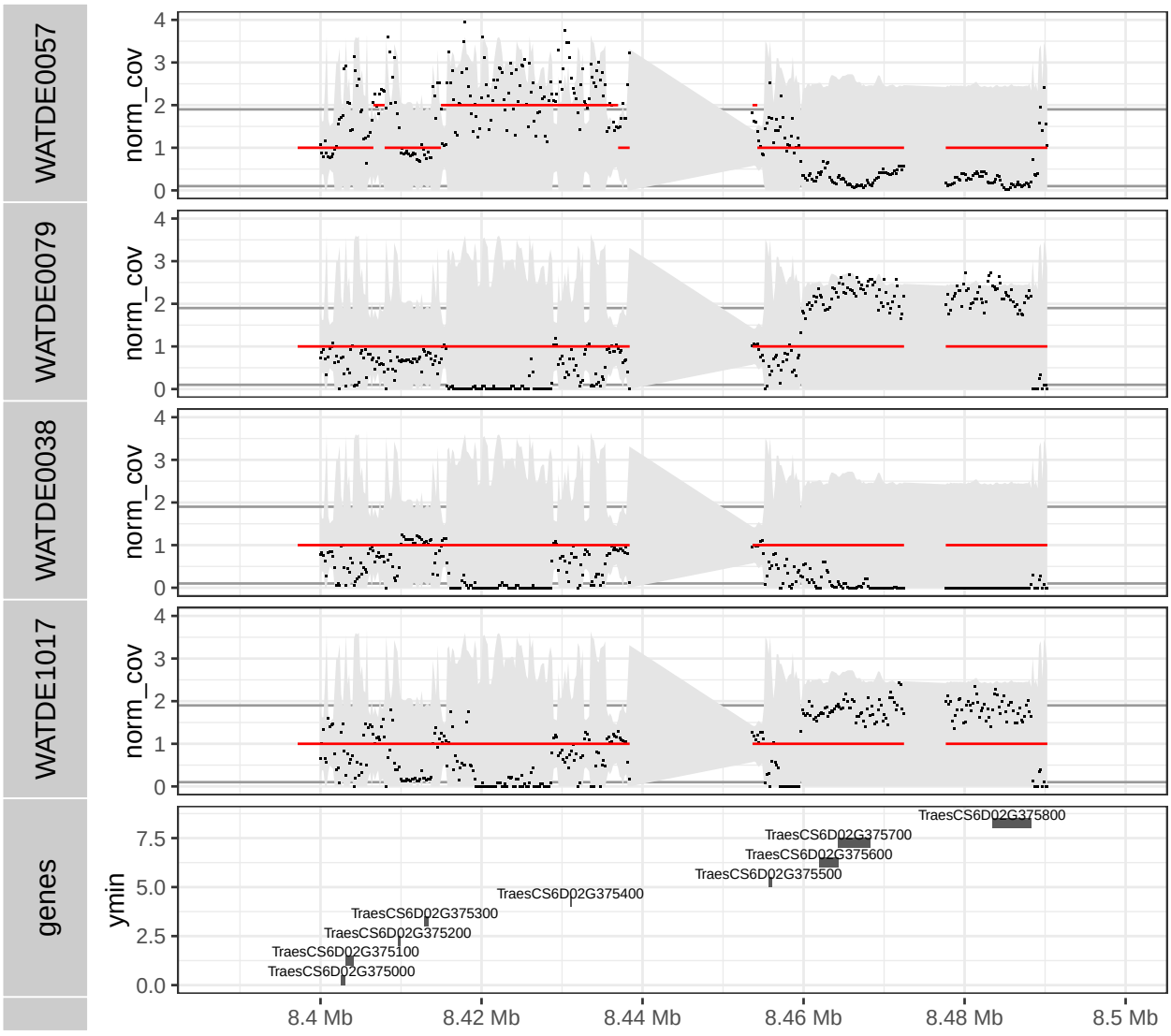
On this example, from the region between `4.4Mbp` to `4.5Mbp` of `chr6D_part2`, containing 6 genes, we can already observe some intresting examples.

1. `WATDE0687` Has some small duplications, but some candidates may be longer, but they go over the noisy areas.
2. `WATDE0057` There may be a longer deletion over `TraesCS6D02G366500` and just before `TraesCS6D02G367000`, but again,t hey fall in patches of nosiy coverage.
3. Lines `WATDE0038` and `WATDE0456` May be a single long deletion event. But some regions seem to have enough coverage to go over the threshold



The region from `8.4Mbp` to `8.5Mbp` is more noisy, most of the windows are excluded. However, we can still have a couple of conclusions from this reagion.

1. Line `WATDE0057` probably has a duplication of `TraesCS6D02G375400` and `TraesCS6D02G3753`. A few of the windows have enough quality.
2. Lines `WATDE1017`, `WATDE0038` and `WATDE1017` seem to have deletions.



Finally, the region between `20.1Mbp` and `20.2Mbp` has mostly windows that can be used to find CNVs.

1. `WATDE1023` Has a couple of small deletions on gene `TraesCS6D02G401700`
2. `WATDE1017` and `WATDE1018` are likelky to have a long duplication, but some noisy windows break it. Or it may be real small duplications
3. `WATDE0910` has a duplication over `TraesCS6D02G402000`.