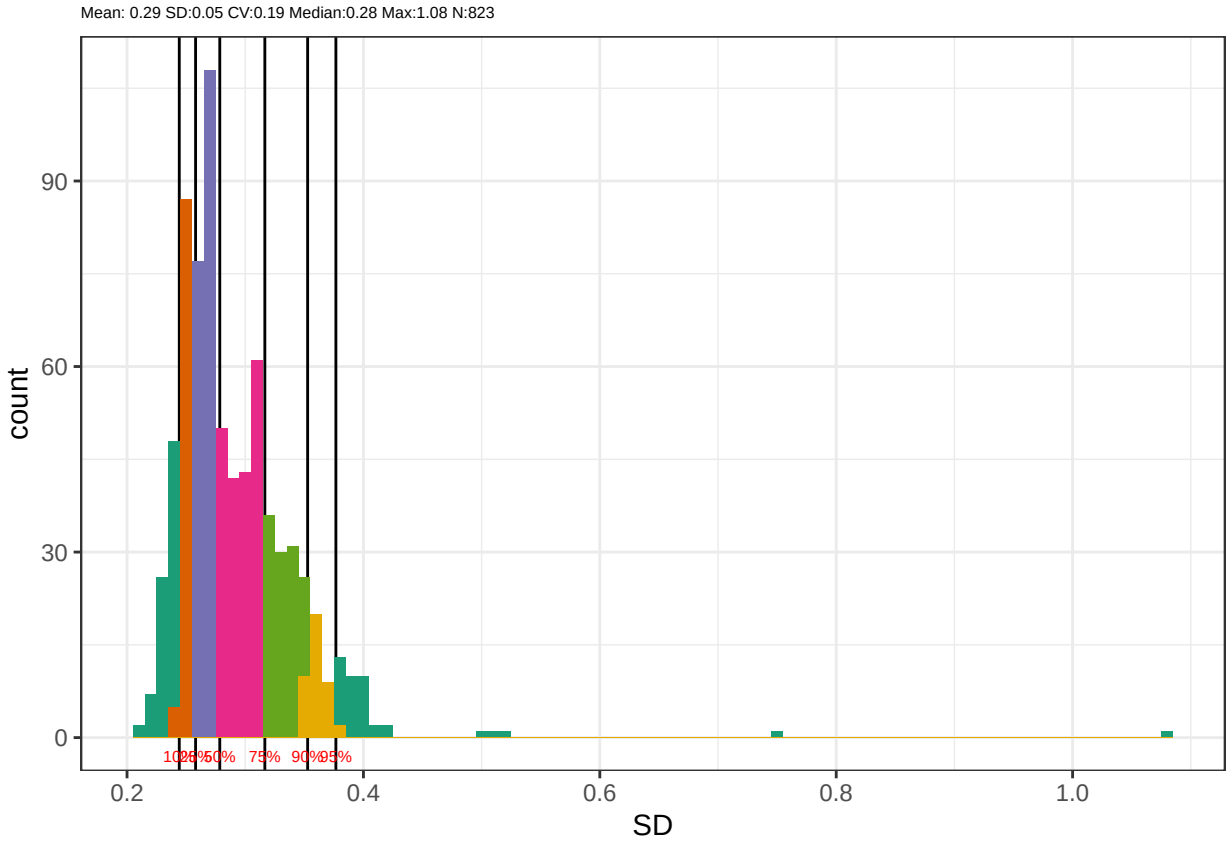
Analysis of CNV around RHT for QC

Ricardo H. Ramirez-Gonzalez

This notebook is to look at some examples of lines where calling the CNV may be troublesome, based on the dispersion

Out of 823 lines, 5 have an $\sigma > 0.45$. We consider lines above the threshold as noisy.

```
## Warning: Use of `quantiles$value` is discouraged. Use `value` instead.
```



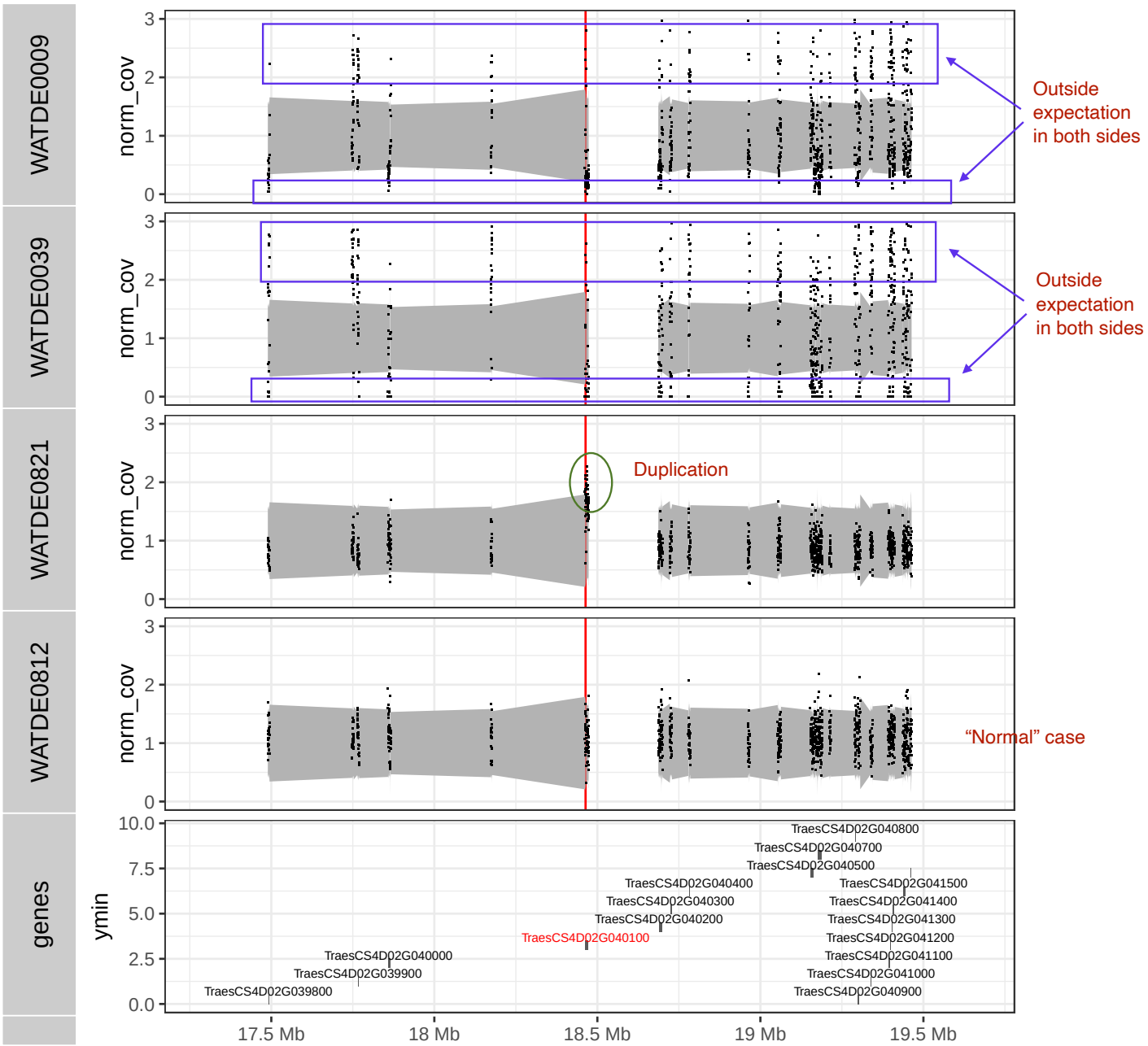
	line	SD
25	WATDE0009	0.7529014
54	WATDE0039	1.0792827
71	WATDE0056	0.5134964
75	WATDE0060	0.5173301
104	WATDE0090	0.4956258

Details for TraesCS4D02G040100

To showcase the issues with the `5` noisy lines, we have a closer look at `TB1-4D`. The ribbon is showing the area $\pm 3\sigma$ across each particular window. The dots that are outside the ribbon represent windows that are candidates to be considered a CNV, as tehy are outside the 99% expectation, assuming a normal distribution.

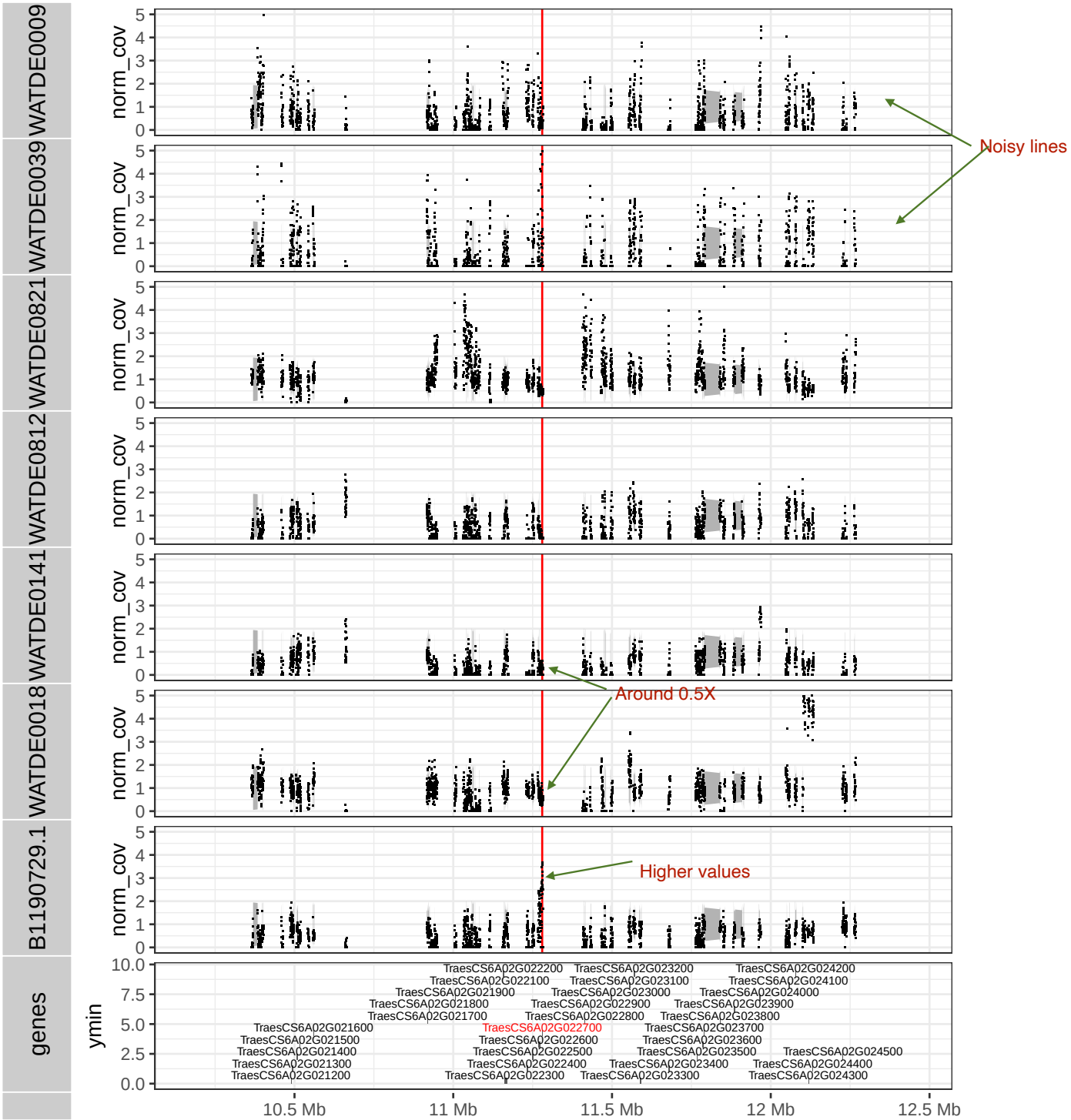
In this example, lines `WATDE0009` and `WATDE0039` are considered as noisy. Naivley, one would think that there are plenty of candidates for CNV on those lines. However, the fact that the values oscilate above and bellow the region of the expected distribution suggests that this calls are noisy. Line `WATDE0821` seem to have some duplications. We can call this because the normalised coverage is, in generalover the expected dispersion, withut any value close to 0. Furthermore, the windows outside the gene seem quite stable. Line `WATFE812` doesn't seem to have anything abnormal, which should be how most of hte lines look.

This is relevant because in some of the previous heatmaps this line are shown as having CNV.



Some examples around TraesCS6A02G022700

The gene `TraesCS6A02G022700` is reported has missing copies in the Chinese Spring reference. We picked this because we wanted to see if we observe this gene has some variation. Again, the lines `WATDE0009` and `WATDE0039` show a lot of variation, so this is consistent with the noise being noisy. Line `B1190729.1` seems to have higher coverage, which is expected. Lines 'WATDE0018 and `WATDE0141` have a lower coverage, but it doesn't go to 0. This may be a case where there is a single copy, in a region that is collapsed in the assembly.



Details for TraesCS6A02G022700, to show change in real coverage

To look if the change in global coverage in regions that are repetitve can be detected only from the raw coverage, but not from the normalised values, we explored in more detail lines `B1190729.1`, 'WATDE0018 and `WATDE0141`.

`WATDE0141` ann `WATDE0018` seems like having the same raw dispersion of raw coverage in the higlighted region, but a normalised coverage around and below 0.5.
This may indicate that the assembly is collapsing two copies, but this samples only have one copy of the gene.

`B1190729.1` Seems to have more coverage in that region, so it may be that the line has more than the 2 expected collapsed copies from the reference.