Analysis of CNV for PPD genes

This notebook is to include the analysis of:

1. The SD of the 823 lines
2. Comparing the deletions from the raw coverages and from the normalized coverage (PPD1-2A)
3. Details of clusters 5 of TraesCS4D01G040100

The input includes `6,652,249` out of ‘`6,719,173`. The excluded regions have no coverage across any of the lines.

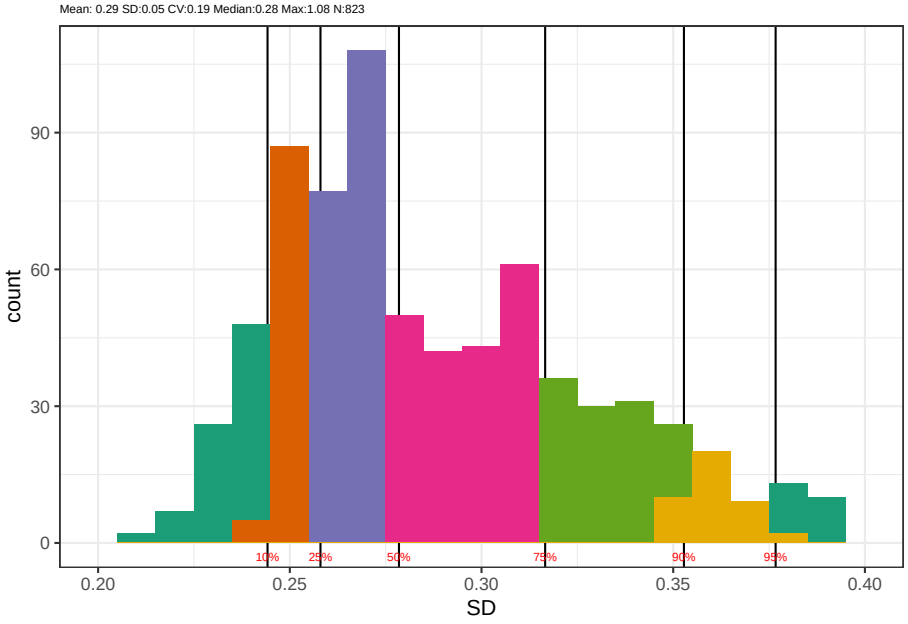
```
ruby meltMat.rb -f /Users/ramirezr/Documents/WatSeq/Deletions/20200724/200bp
Done iteration
Mat: 6652249
DF: 6652249
Cov: 6719173
```

Distribution of the SD across all the lines.

To prepare the `libSD.csv.gz` I used the following command:

```
gunzip -c libSD.csv.gz |
  sed -e s/\"\",\"x\"/\"line\",\"SD\"/ -e s/merge\.rmdup\.// > libSD_short.csv
```

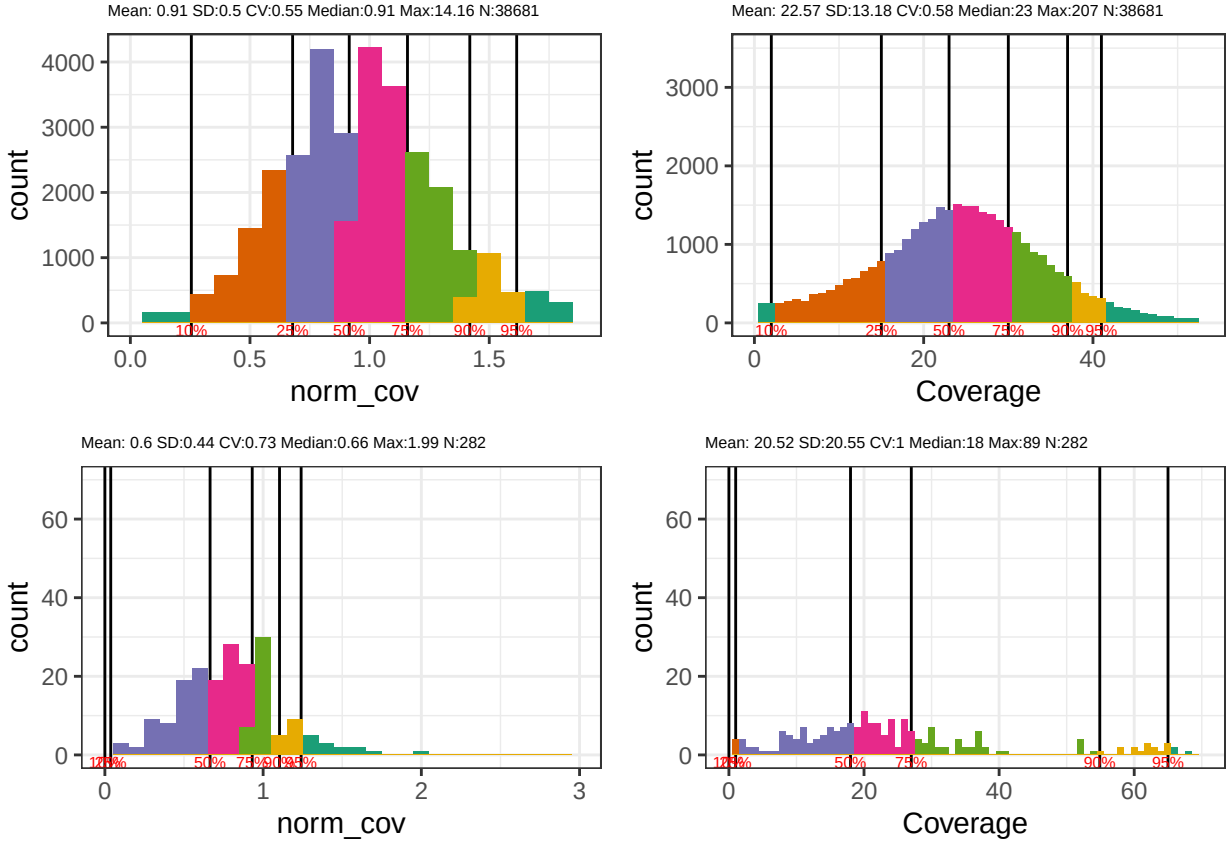
The following plot shows the distirbution of the Standard Deviation after the normalization of all the samples. As none of the samples goes above 0.5, we can find coverages changes in leves of 0x, 1X, 2X, 3X.



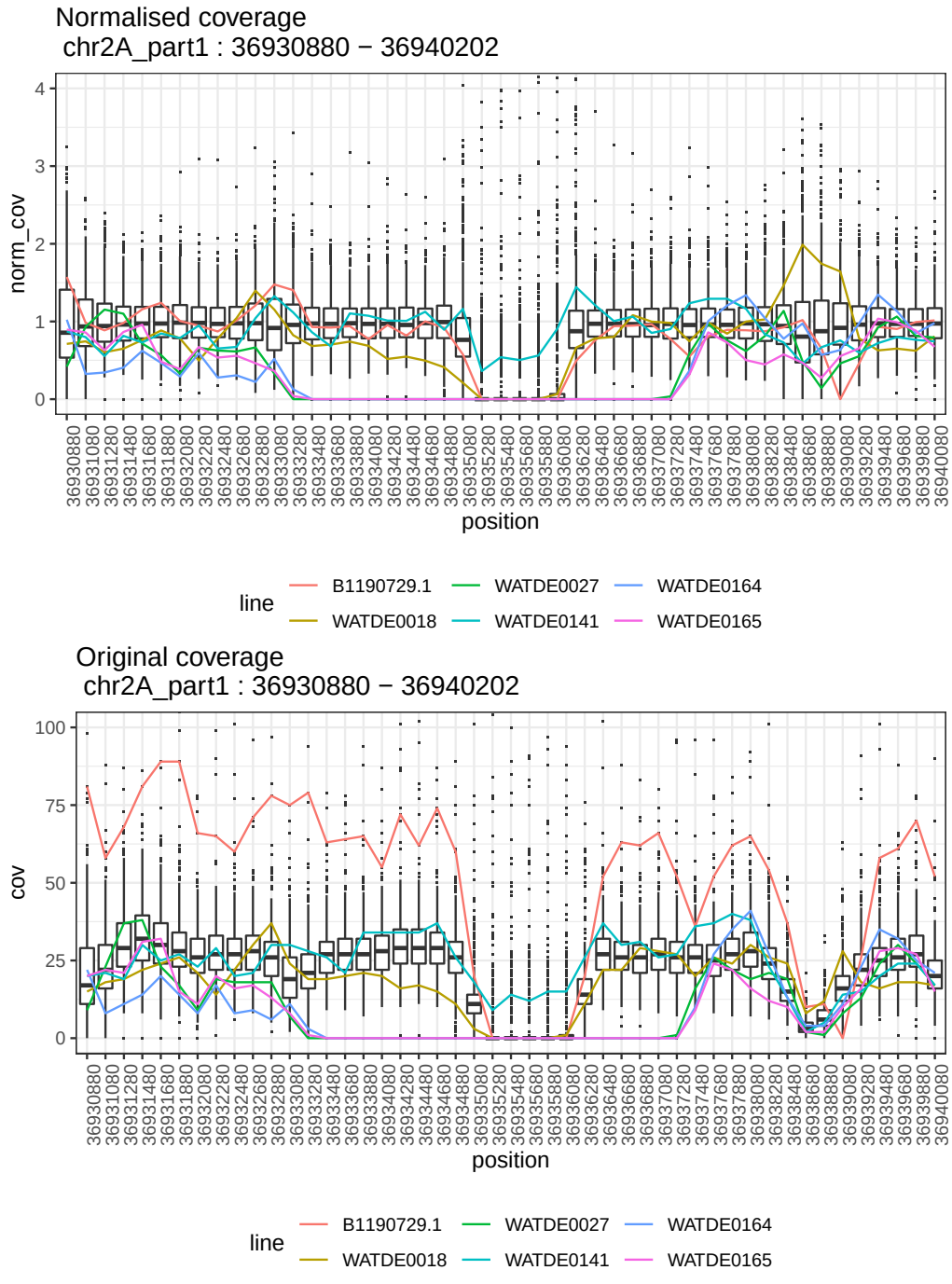
Comparing the deletions from the raw coverages and from the normalized coverage (PPD1-2A)

This is the distribution of all coverages in the region `chr2A_part1:36931000-36943300`. The windows on the normalised table are not the same as in the ppd example in the XLS file, as the window size is different and we only consider +/- 2kpb around the gene model. The plots in the top contains all the normalised values for all the windows in the range. The plots on the bottom is the distribution of the values on lines `WATDE0027`, `WATDE0164` and `WATDE0165`, which are in the cluster 1 of TraesCS2A01G081900.Ppd-1.png, but have a longer deletion than the rest of of the values. We are including also `WATDE0141` as it was previously requested by Simon. However, the new candidate is `B1190729.1` (also called `WATDE0100`, `1190729-1`) On the left, the values normalized by line and window. On the right, the raw data from the excel file with PPD1-2A. In the case of line `WATDE0141`, has some coverage. Before the latest normalisation, this case was going up to 3X. However, the method removing all the 0 values make it be close to `0.5` coverage, not quite a deletion. This may need validation in the lab.

As for line `B1190729.1` , there is a secondary deletion around `36,939,080`, just affter the region with lowe coverage in the rest of the lines.



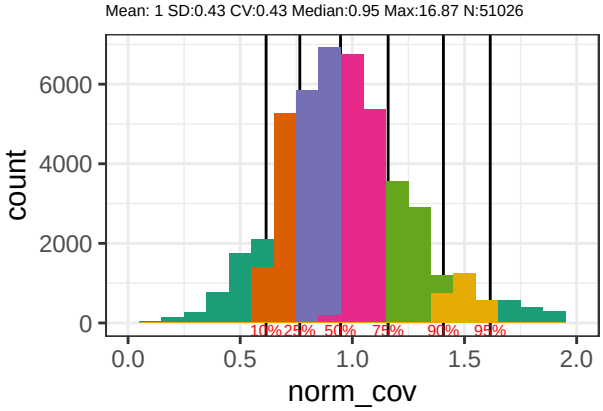
To explore in more detail those three lines, we had a look at three lines that seemed to have a longer deletion than the rest of the lines. On the normalised example, the second, smaller deletion, is reduces. The coverage distribution of the secondary deletion suggests that across all the lines there is a low amplification, but consistent across all the lines. This suggests that this deletion is likely to be artificial. However, it would be intresting to validate it.

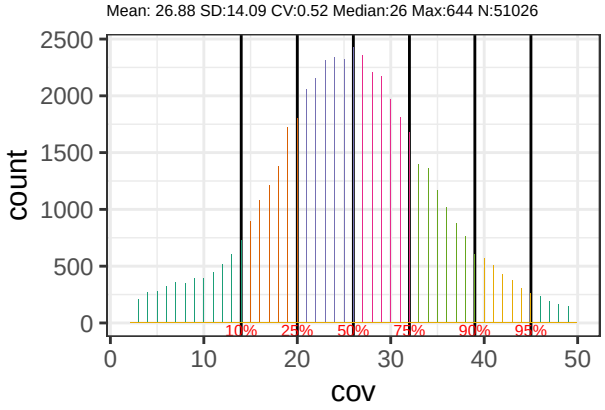


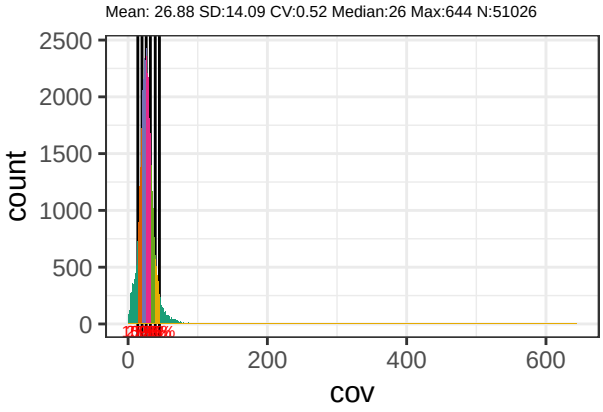
Details of cluster 5 of TraesCS4D01G040100

Finally, looking at the plots we found that for `TB1-4D`, clusters 4 and 5 seem to have a duplication. For those, I don't have the values from the 100bp bins as with `PPD1-2A`.
Hence, I only plotted the values for the normalisations we'd ben producing. From the 6 lines in cluster 5 that passed the filters (discussed above, to be relaxed), only `WATDE0821` seems to be consistently around the 2X coverage.

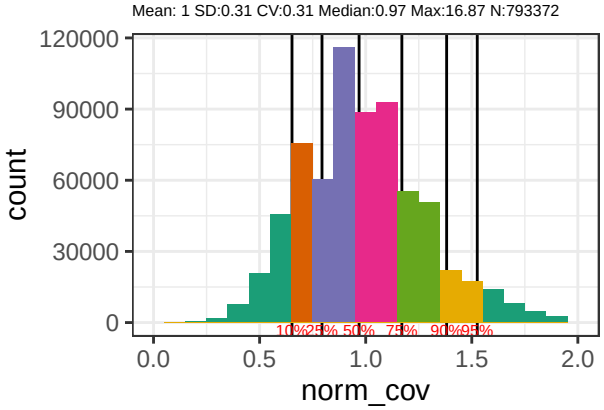
```
## Warning: Closing open result set, pending rows

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

## Warning: Removed 947 rows containing non-finite values (stat_bin).

## Warning: Removed 12 rows containing missing values (geom_bar).

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

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

## Warning: Removed 1673 rows containing non-finite values (stat_bin).

## Warning: Removed 12 rows containing missing values (geom_bar).
```

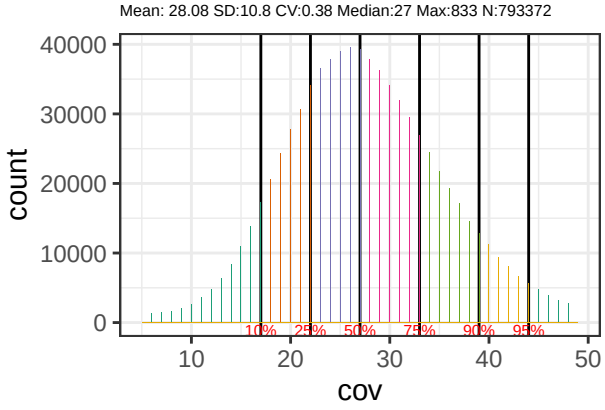


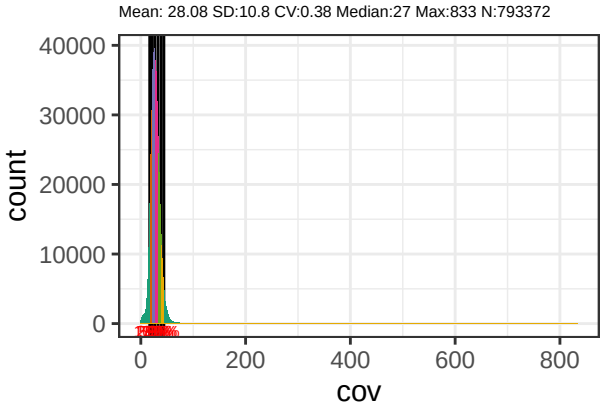




```
## Warning: Closing open result set, pending rows
## Warning: Use of `quantiles$value` is discouraged. Use `value` instead.
## Warning: Removed 3893 rows containing non-finite values (stat_bin).
## Warning: Removed 12 rows containing missing values (geom_bar).
## Warning: Use of `quantiles$value` is discouraged. Use `value` instead.
## Warning: Use of `quantiles$value` is discouraged. Use `value` instead.
## Warning: Removed 21499 rows containing non-finite values (stat_bin).
## Warning: Removed 12 rows containing missing values (geom_bar).
```

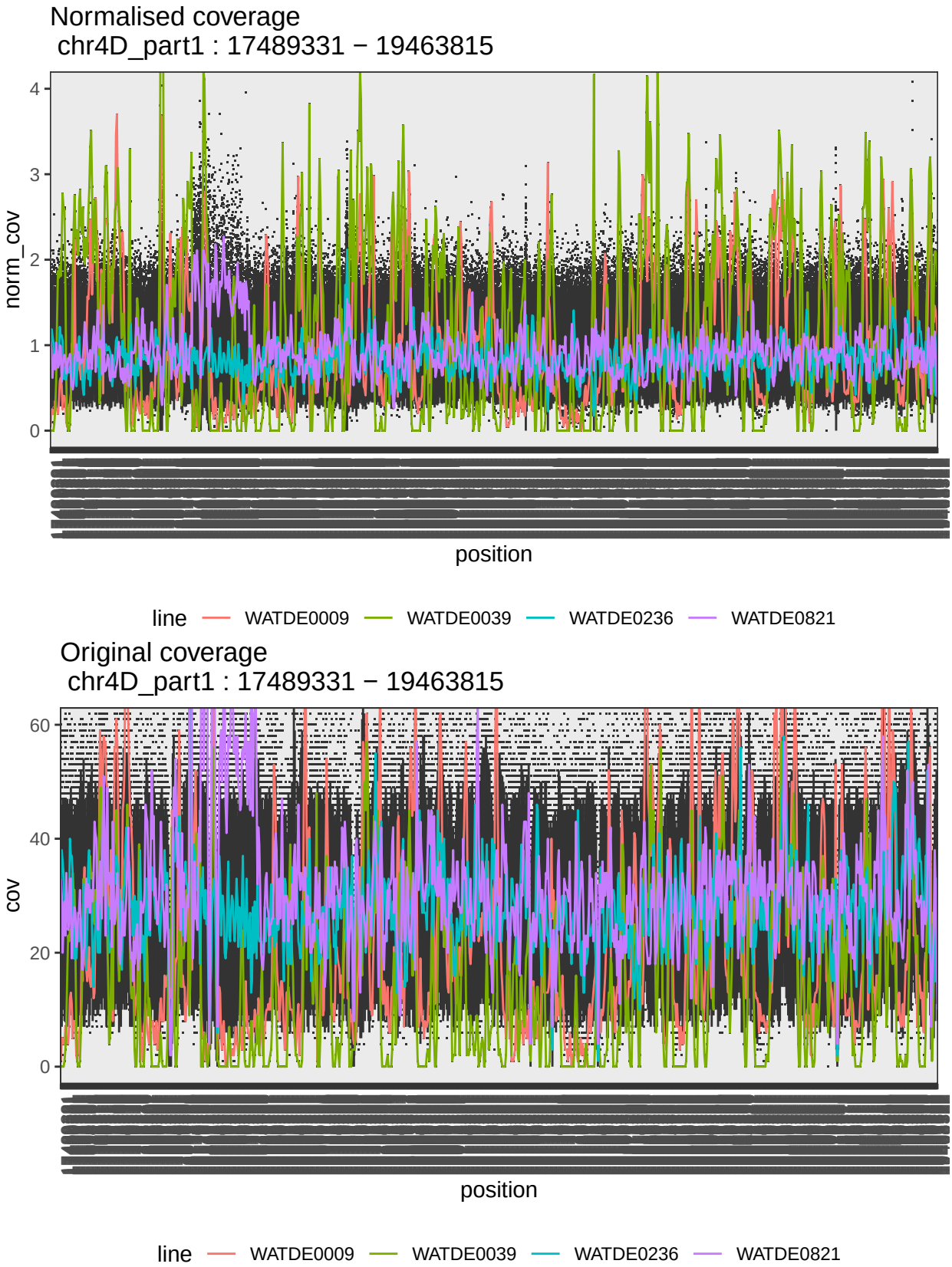






There are a few points that are otliers. as follows:

	chrom	chromStart	chromEnd	line	cov	norm_cov
117941	chr4D_part1	18175168	18175368	WATDE0278	778	2.022759
117961	chr4D_part1	18175168	18175368	WATDE0305	833	2.103121
172910	chr4D_part1	18473238	18473438	WATDE0066	644	13.931116
173214	chr4D_part1	18473238	18473438	WATDE0453	503	16.874856
173288	chr4D_part1	18473238	18473438	WATDE0562	617	16.360005
173733	chr4D_part1	18473438	18473638	WATDE0066	519	11.899795
176202	chr4D_part1	18474038	18474238	WATDE0066	630	11.257215
176506	chr4D_part1	18474038	18474238	WATDE0453	509	14.105232
176580	chr4D_part1	18474038	18474238	WATDE0562	581	12.725204


Normalised coverage
chr4D_part1 : 17489331 − 19463815
norm_cov
0
1
2
3
4
position
line
WATDE0009
WATDE0039
WATDE0236
WATDE0821
Original coverage
chr4D_part1 : 17489331 − 19463815
cov
0
20
40
60
position
line
WATDE0009
WATDE0039
WATDE0236
WATDE0821