# 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

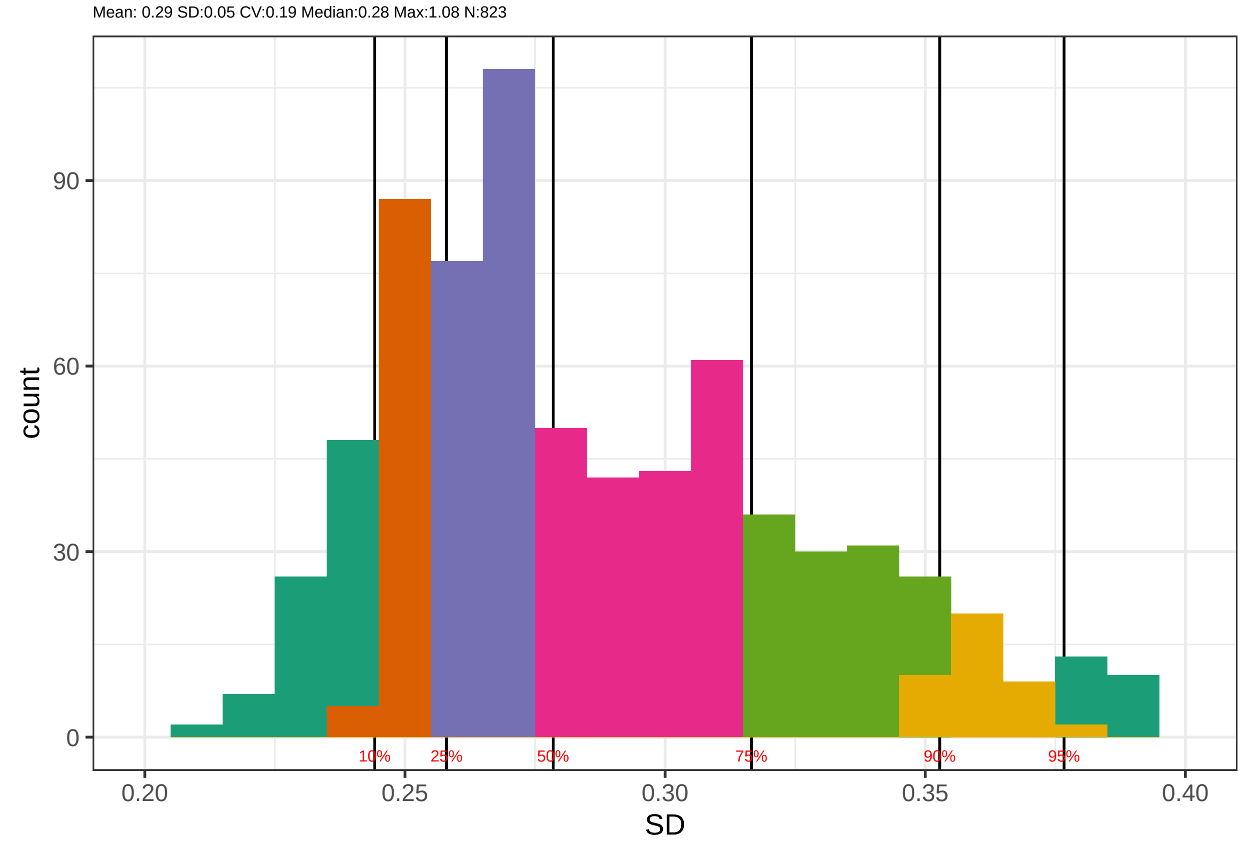
The input includes `6,665,337` out of ‘`6,719,174`. The excluded regions have no coverage across any of the lines. Cov: 6719174

## Distribution of the SD across all the lines.

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

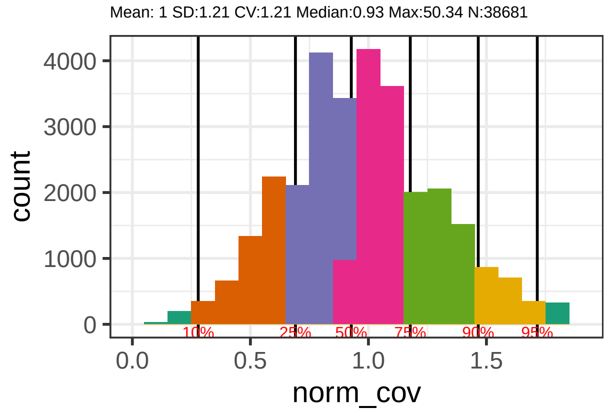
```
gunzip -c libSD.csv.gz |
  sed -e s/\"\",\"x\"/\"line\",\"SD\"/ -e s/merge\.rmdup\.// > libSD_short.csv
```

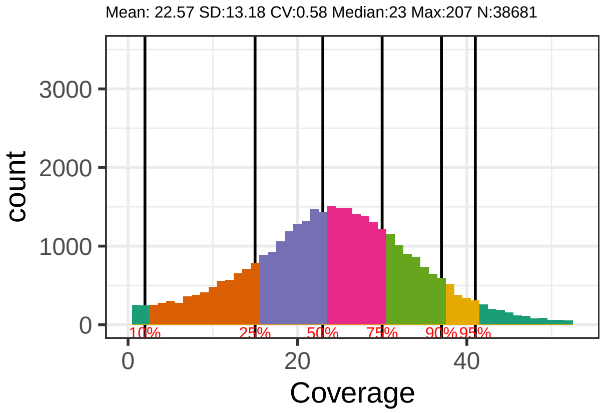
The following plot shows the distirbution of the Standard Deviation after the normalization of all the samples..

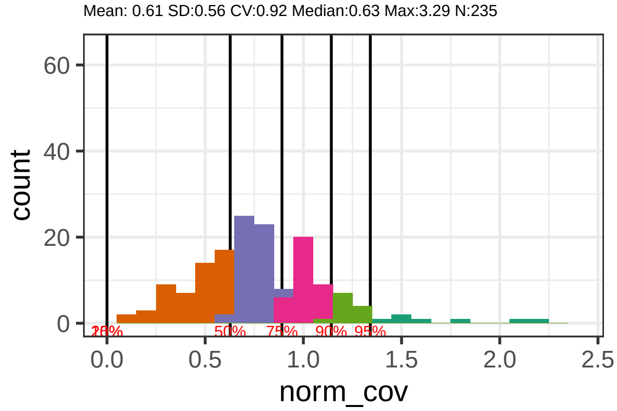


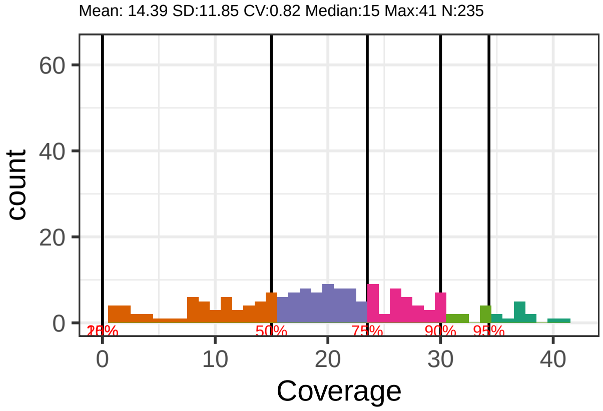
## Comparing the deletions from the raw coverages and from the normalized coverage (PPD1-2A)

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 requested by Simon 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`, the absolute coverage seems to be low, but after normalisation it is clear that deletion looks more like a copied region. This may be real, or it may be an issue of the region having a low mapping in general. This mau need validation in the lab.

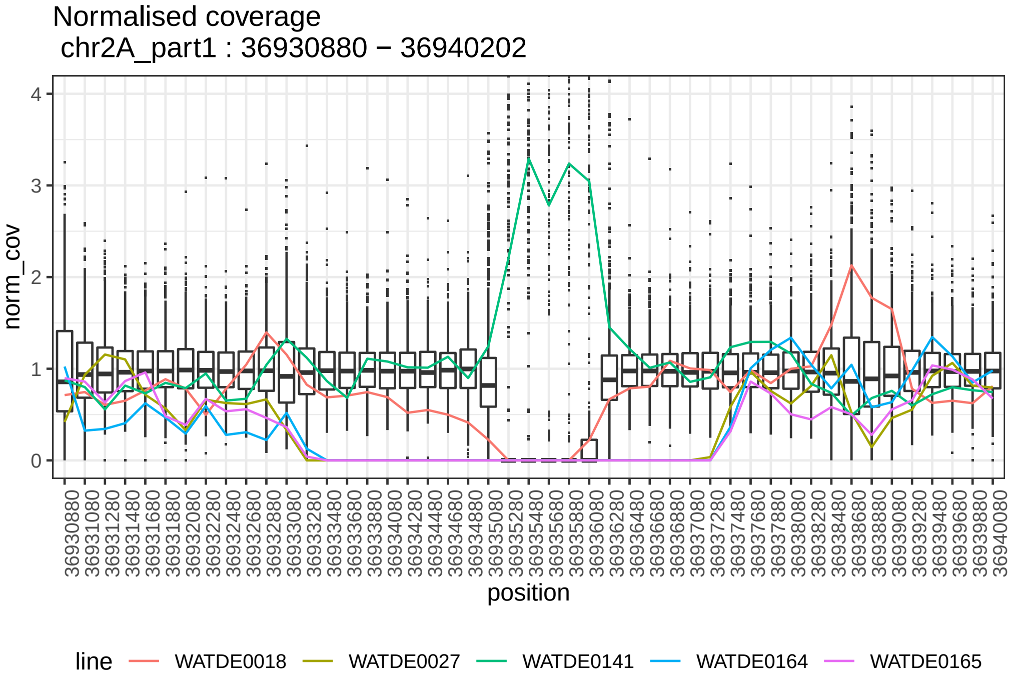


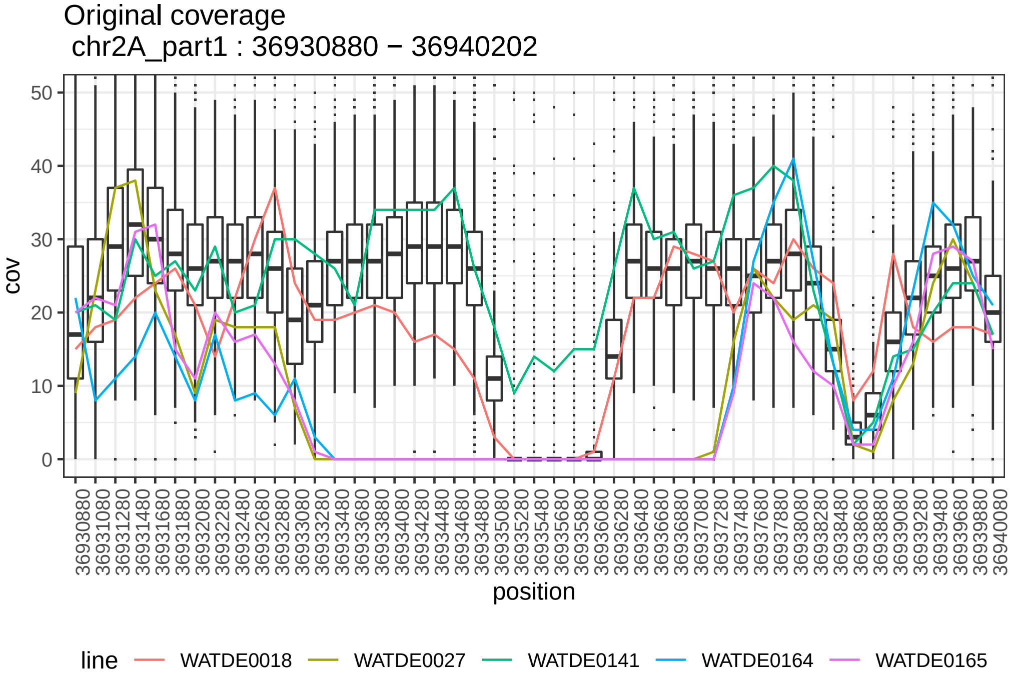






To explore in more detail those three lines, we had a look at three lines that seemed to have a longer deletion than the rest of the lines. On the normalised example, the second, smaller deletion, is reduces. The coverage distribution of the secondary deletion suggests that across all the lines there 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.



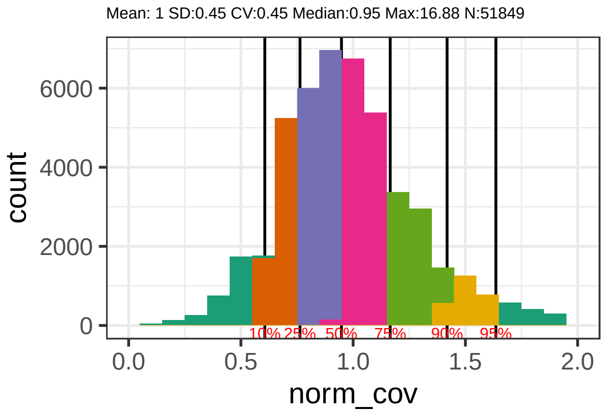


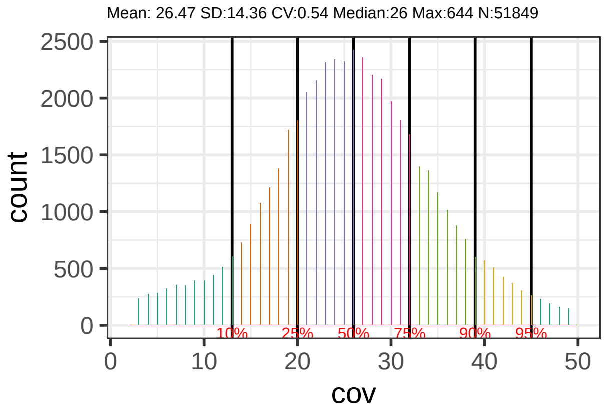
### Details of cluster 5 of TraesCS4D01G040100

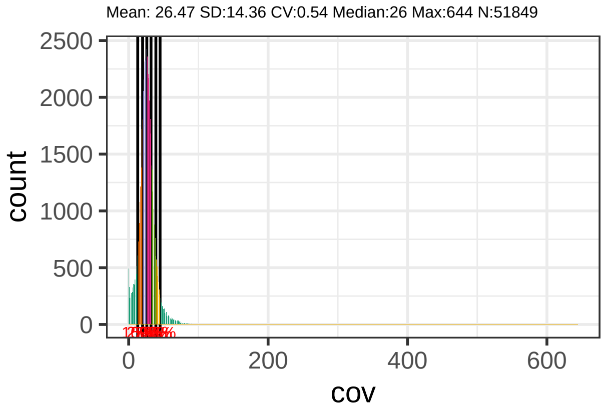
Finally, looking at the plots we found that for `TB1-4D`, clusters 4 and 5 seem to have a duplication. For those, I don't have the values from the 100bp bins as with `PPD1-2A`.
Hence, I only plotted the values for the normalisations we'd ben producing. From the 6 lines in cluster 5 that passed the filters (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 1098 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 2345 rows containing non-finite values (stat_bin).

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

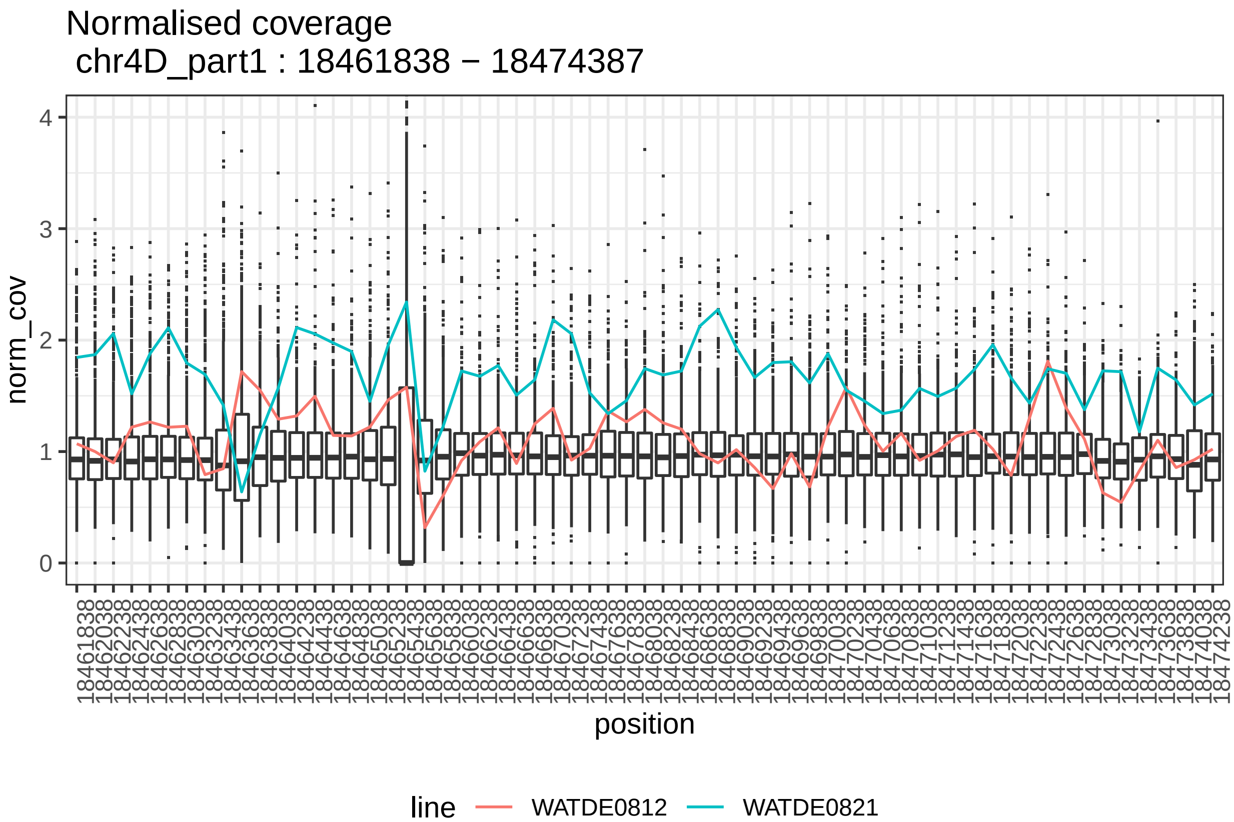




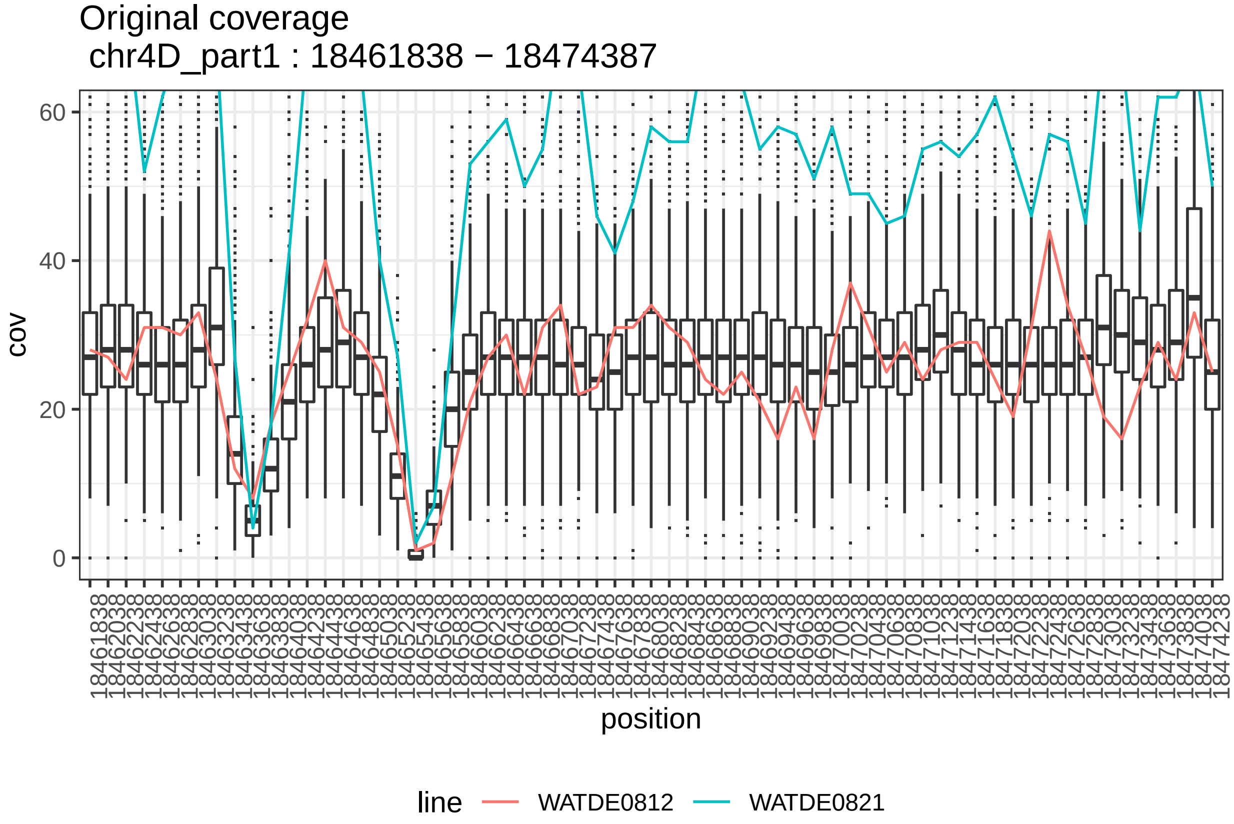


There are a few points that are otliers. as follows:

| chrom | chromStart | chromEnd | line | cov | norm_cov |
|---|---|---|---|---|---|



Normalised coverage
chr4D_part1 : 18461838 − 18474387



Original coverage
chr4D_part1 : 18461838 − 18474387

This means