How to interpret multi-sample plots?

CoNVex's multi-sample plots are used to visualise each CNV call and the underlying error-corrected log2 ratio and heuristic scores that are used to detect the CNV. The automated calling algorithm (based on the Smith-Waterman algorithm) is optimised to correctly merge/split nearby calls and to map the boundaries in an intuitive way. It is possible in some cases that the algorithm may merge two nearby calls, or split a single CNV into two. It is also possible that the boundaries are not estimated accurately, although every effort is made to fine-tune this process.

Multi-sample plots are essential to visualise the underlying scores manually and decide whether one of the cases described above are true. It is also important to confirm a real, correctly detected CNV by going through each of them manually for the clinical practice. These plots contains 3 panels:

1) The top panel contains the log2 ratio that may also include noise based on the number of sequence reads that hit the target region. Noise free regions (their log2 ratio) make it unambiguous to interpret the CNV calls. Multiple samples (all or a subset) that are in the analysis are denoted as grey lines – one for each sample – and the sample with that particular CNV is denoted with a red or blue line in case of a deletion or duplication respectively. Scores are expected to be high for duplications and low for deletions. Depending on the fluctuations of each probe's score, we can interpret the validity of each call easily in most cases. The inner dotted lines indicate the minimum extent of the CNV (start position of the first probe and end position of the last probe within the call as shown in the legend) and the outer dotted lines estimate the maximum possible extent (end position of the last probe before the call and start position of the first probe after the call).
2) The mid panel is derived from the log2 ratio in the top panel but is corrected for noise in the target region. This is – preferred – over the log2 ratio in many cases although a combination of both offers a better understanding of the CNV region.
3) The bottom panel is used to display information about the CNV as well as other helpful information from different sources if exist. The X-axis in this panel contains the chromosome co-ordinates that are common for all 3 panels.

CNV information include:

- The CNV region (chromosome co-ordinates) and the CNV type
- Sample ID
- #Probe regions within the call
- CoNVex score (CNVs with score ≥ 5 are selected)
- Genes
- Mother, father (optional)
- Known CNVs as lines in panel 3 (optional, under development)

UCSC genome browser links to probe regions:

- Agilent SureSelect V3 (hg19) – http://bit.ly/M8UDuJ
- Agilent SureSelect V1 (hg18) – http://bit.ly/NK16JM
- DDD Exome+ (hg19) – http://bit.ly/QjGkoM

I have listed a few examples below together with a brief description of the interpretation so that you get an idea of how to interpret the CNV calls from these multi-sample plots.

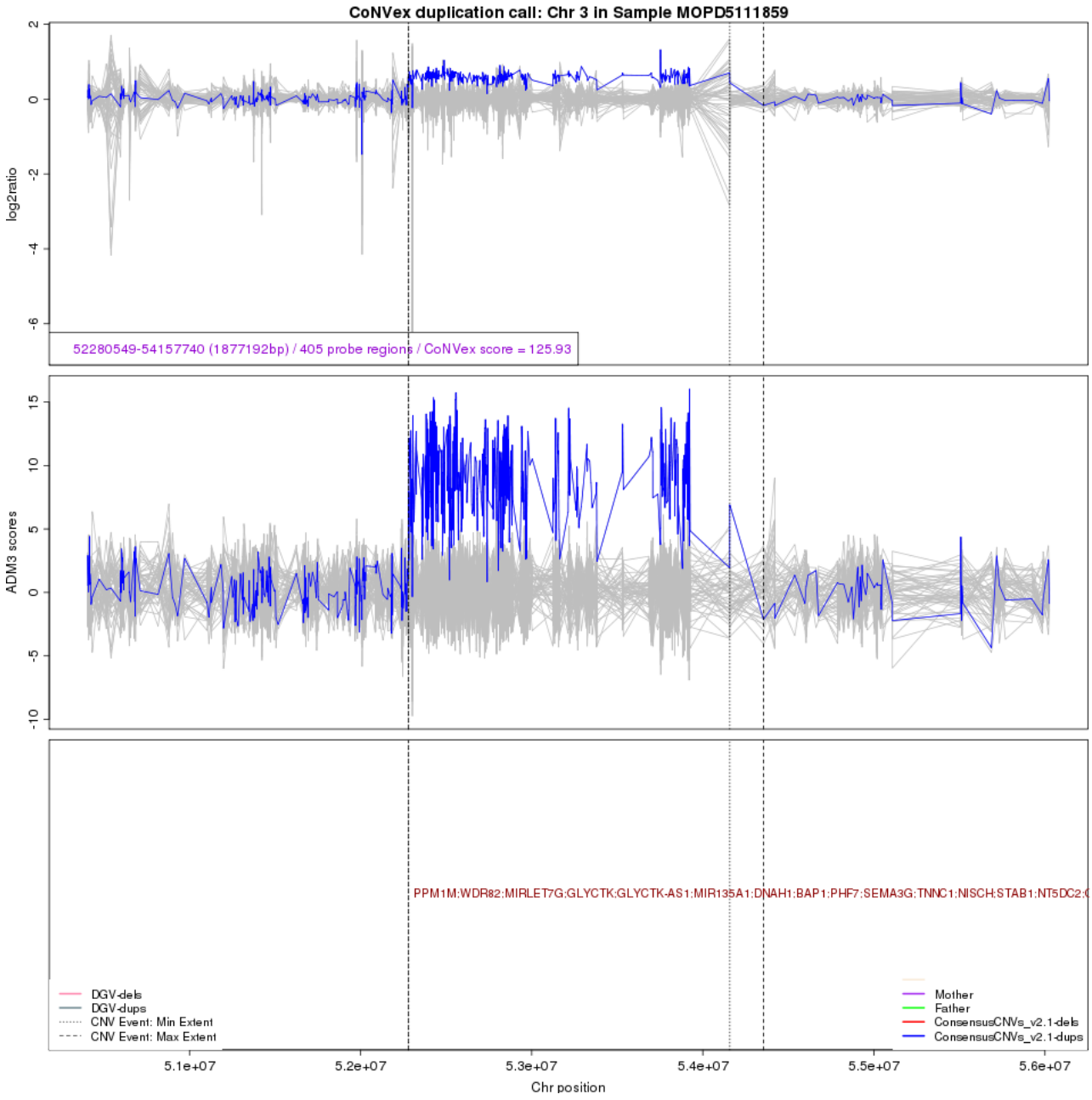

Figure 1: A duplication in sample MOPD5111859 shows a significant *increase* in heuristic scores (ADM3 scores) in the mid panel. This is also clear from the log2 ratio in the top panel.

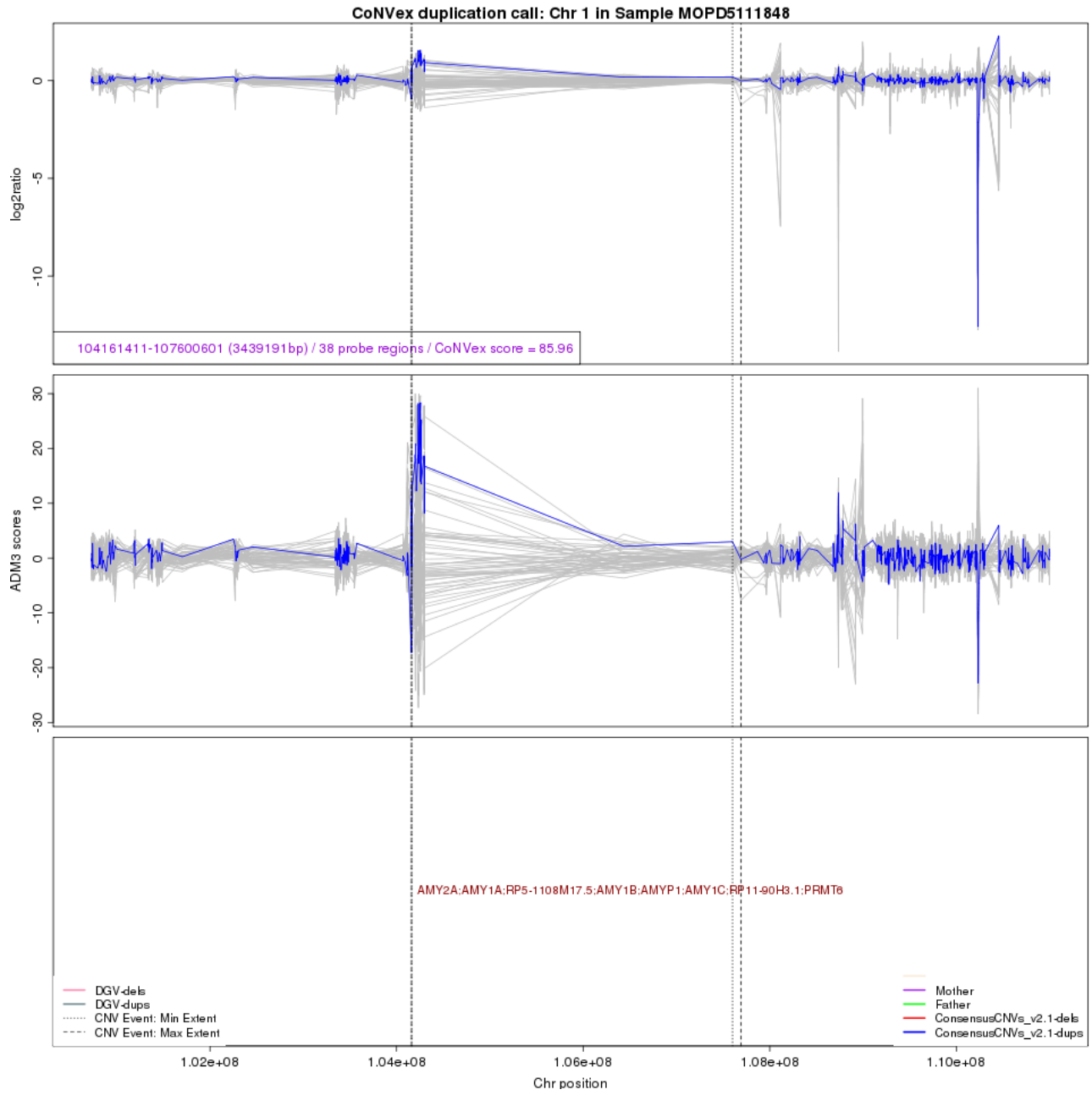

Figure 2: A duplication in sample MOPD5111848 shows a significant increase in heuristic scores (ADM3 scores) in the mid panel, but extended upto two more nearby probes. This is probably a common CNV region as other samples (grey lines) also shows increase/decrease in the scores.

In order to determine the boundary accurately, it will be required to look at the probe regions in the UCSC genomes browser together with these calls. One may also do this by looking at the probe regions in the text editor. In both cases, excluding the last two probes within the call (based on the 'score points' in the line) is preferred. Use the appropriate UCSC genome browser links provided in this document.

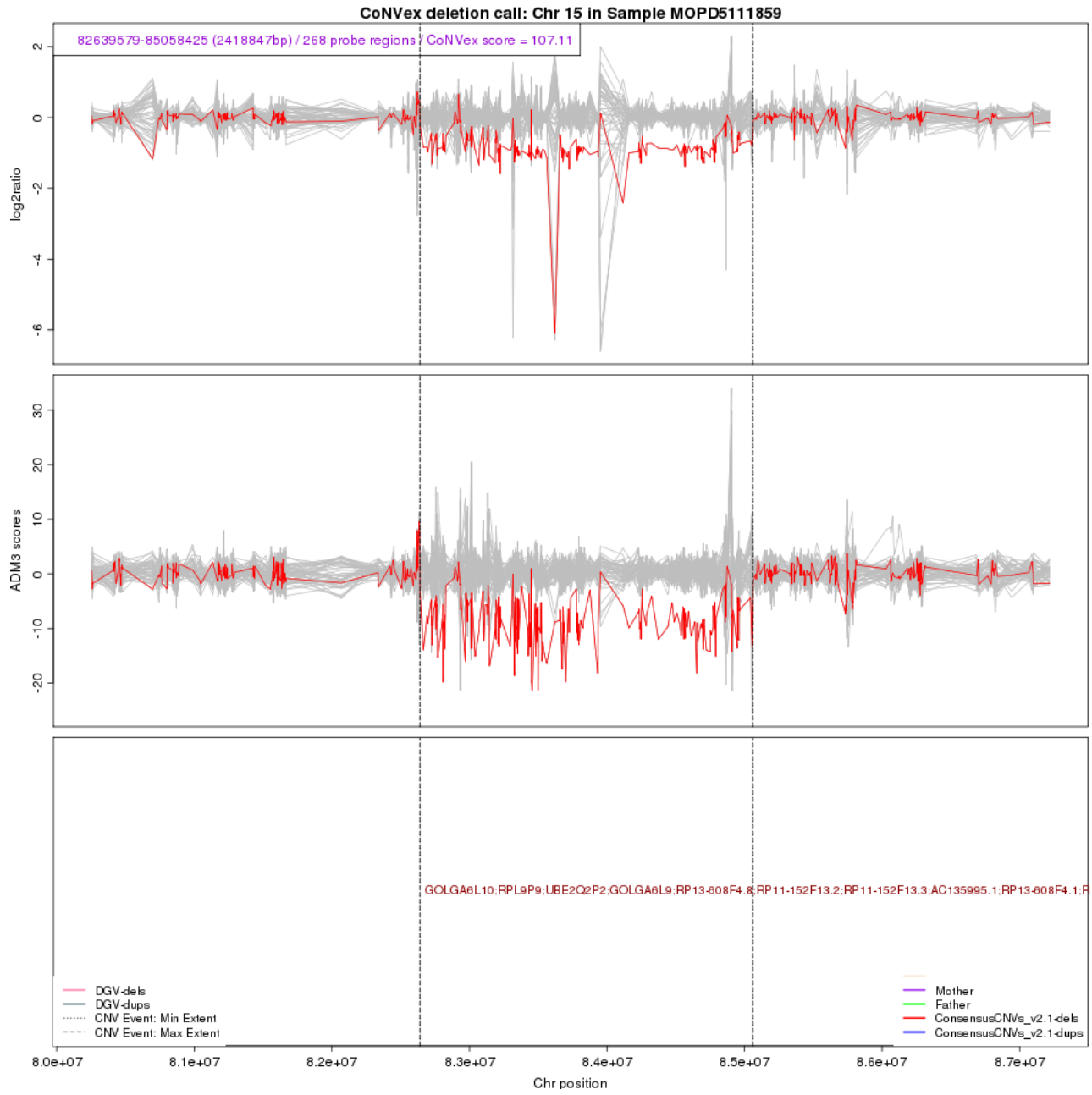

Figure 3: A deletion in sample MOPD5111859 shows a significant *decrease* in heuristic scores (ADM3 scores) in the mid panel. This is also clear from the log2 ratio in the top panel.

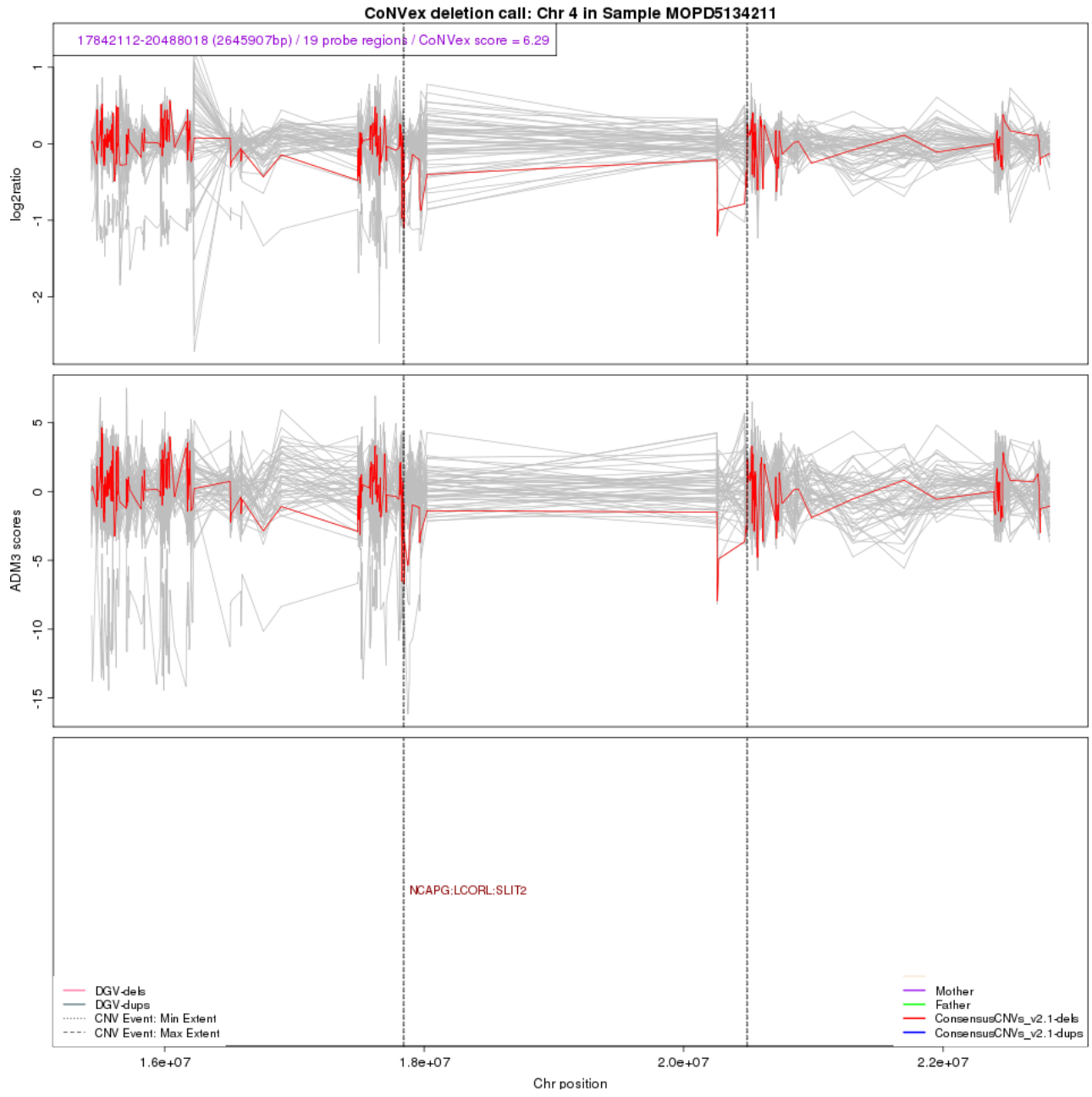

Figure 3: There are probably two small deletions at the periphery of the region called as a single deletion. As there are no probes in between in the ~1Mb region, the algorithm merged these deletions.