

Supplementary Notes #3:

Single-sample assessment of copy-number preprocessing methods CRMA v2, Affymetrix CN5 and dChip based on tumor-normal sample HCC1143 in GEO data set GSE13372

Henrik Bengtsson et al.

April 11, 2009

1 Introduction

This report provides a single-sample approach to assess the relative performance of the CRMA v2, the Affymetrix CN5, and the dChip method. The results presented herein are intended to complement the results presented in the main CRMA v2 manuscript.

2 Method

Consider a local genomic region with loci where there exists exactly one copy-number (CN) changepoint (breakpoint) at position x_0 and that the true CN level at each side of the changepoint is constant (which is a common assumption in CN studies). Assume that we can identify this CN changepoint and with very high confidence locate it to be within $[x_0 - \delta, x_0 + \delta)$ where we refer to δ as the safety margin on each side of the changepoint. Then we can assess how well we can detect this changepoint given the two sets of data points one each side of changepoint.

2.1 Receiver Operator Characteristic performance

One approach is to use Receiver Operator Characteristic (ROC) analysis analogously to what is done in the (multi-sample) evaluation method presented in Bengtsson *et al.* (2008). Given the changepoint and excluding all loci within the safety margin, the remaining J loci are split up in two data sets $\mathcal{J}_A = \{x_j \leq x_0 - \delta; j = 1, \dots, J\}$ and $\mathcal{J}_B = \{x_j > x_0 + \delta; j = 1, \dots, J\}$, where x_j is the position of locus j and J is the total number of loci. Continuing, let $\mathcal{M}_s = \{M_j; j \in \mathcal{J}_s\}$ denote the (full-resolution) raw (relative) copy-number estimates of set $s \in \{A, B\}$ obtained by one of the preprocessing methods of interest.

Without loss of generality, assume the loci in \mathcal{J}_A are copy neutral and the ones in \mathcal{J}_B are deletions. By using a (global) threshold/cutoff τ we can then call the CN state for each locus. We say that locus j belongs to the CN state B if $M_j \leq \tau$. Let $\mathcal{J}_+ = \{M_j \leq \tau; j = 1, \dots, J\}$ be all such loci. Let $\mathcal{J}_- = \{M_j > \tau; j = 1, \dots, J\}$ be the remaining loci, which are said to belong to CN state A . Then $\mathcal{J}_{+|B} = \mathcal{J}_+ \cap \mathcal{J}_B$ represents the set of true positives (true deletions correctly called deletions) and $\mathcal{J}_{+|A} = \mathcal{J}_+ \cap \mathcal{J}_A$ represents the set of false positives (true copy neutral loci incorrectly called deletions). Finally, with $|\mathcal{J}|$ denoting the cardinality of the set \mathcal{J} , we define $\beta = \beta(\tau) = |\mathcal{J}_{+|B}|/|\mathcal{J}_B| \in [0, 1]$ and $\alpha = \alpha(\tau) = |\mathcal{J}_{+|A}|/|\mathcal{J}_A| \in [0, 1]$ to be the true-positive (TP) rate and the false-positive (FP) rate, respectively. The Receiver Operator Characteristic (ROC) performance is defined by the ROC curve $R(\tau) : \tau \rightarrow (\alpha, \beta) \in [0, 1]^2$.

2.2 Performance at different levels of resolution

By smoothing the data points in each set ($s = \{A, B\}$) by binning the loci in non-overlapping bins of width w , we can generate a new set of smoothed CN ratios on which we can do ROC analysis. Since the smoothed CN estimates are less noisy, the TP rate will increase at any given FP rate. The price for achieving this is that the resolution at which we can detect change points decreases.

2.3 Comparing preprocessing methods

For each preprocessing methods $p \in \{\text{CRMA v2, dChip, CN5}\}$ we will obtain one ROC curve $R_p(\cdot)$. With standard ROC analysis we can then compare the relative performance of the different preprocessing methods.

3 Data set

This report is based on the public GEO data set GSE13372 (Chiang *et al.*, 2009), which among other things contains several replicated tumor-normal pairs for sample HCC1143. All data is based on the Affymetrix GenomeWideSNP_6 chip type. This evaluation will be based on CN estimated from one such pair, more precisely the GSM337641 CEL file (tumor) and the GSM337662 CEL file (match normal). For CRMA v2 we processed the two CEL files separately (without any reference arrays) and then calculated the tumor-normal relative CN ratios. For both dChip and CN5, we calculated the ratios for this single pair but including a total of 68 CEL files from GSE13372 in the preprocessing and at the end extracting the pair of interest. For further details on the preprocessing, see the other Supplementary Note.

3.1 List of change points

For this data set, we have selected a few regions for which one safely can assume there exists a single changepoint and for which the CN ratios look constant. This selection was done visually. For each region we chose large enough safety margin such that the risk for the two sets \mathcal{J}_A and \mathcal{J}_B to contain loci from the other set is extremely small.

Tumor-normal pair	Chromosome	Region	Change point	Safety region
GSM337641 / GSM337662	1	100.10-107.50	103.80	0.25
GSM337641 / GSM337662	3	80.00-90.90	85.30	0.25
GSM337641 / GSM337662	4	60.50-65.75	63.40	0.25
GSM337641 / GSM337662	10	61.00-69.00	65.30	0.25
GSM337641 / GSM337662	11	78.20-83.00	80.20	0.25
GSM337641 / GSM337662	12	57.00-63.00	59.80	0.25

Table 1: Regions used for the evaluation and that each contain a single changepoint. All positions and lengths are in units of Mb.

4 Results

We compare the CRMA v2, the Affymetrix CN5, and the dChip preprocessing methods using the aforementioned ROC analysis at the full resolution as well as smoothed resolution with bin sizes $w = \{1.0, 2.0, 5.0, 10.0, 20.0\}$ kb.

4.1 Region: GSM337641:Chr1@100.1-107.5,cp=103.8+/-0.25,s=0/-1

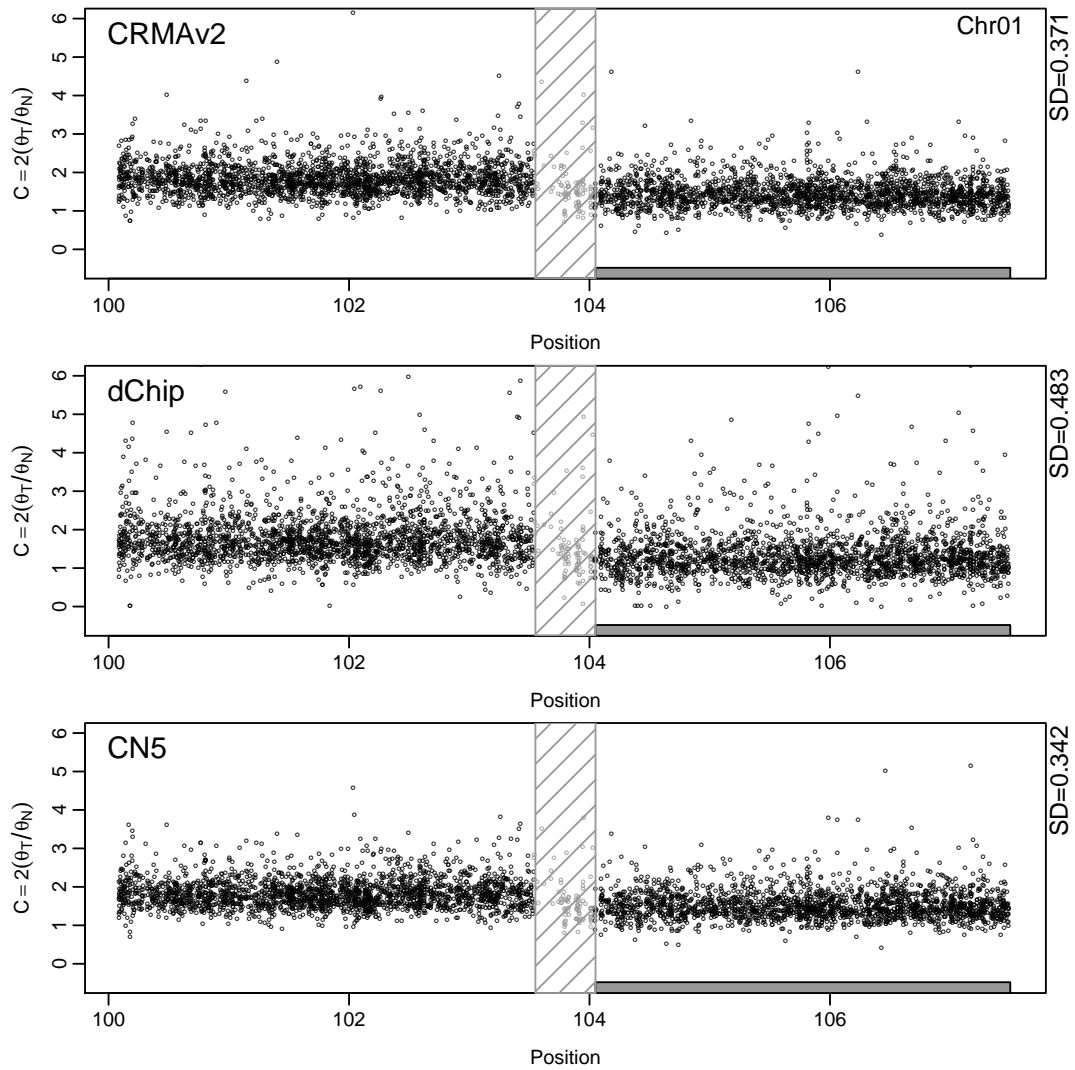


Figure 1: (GSM337641:Chr1@100.1-107.5,cp=103.8+/-0.25,s=0/-1) There are 2242 loci of state 'COPY NEUTRAL' ($s=0$) ("negatives") and 2074 loci of state 'LOSS' ($s=-1$) ("positives"), where the latter are highlighted with a solid bar beneath. In total 91 loci within the safety margin were excluded.

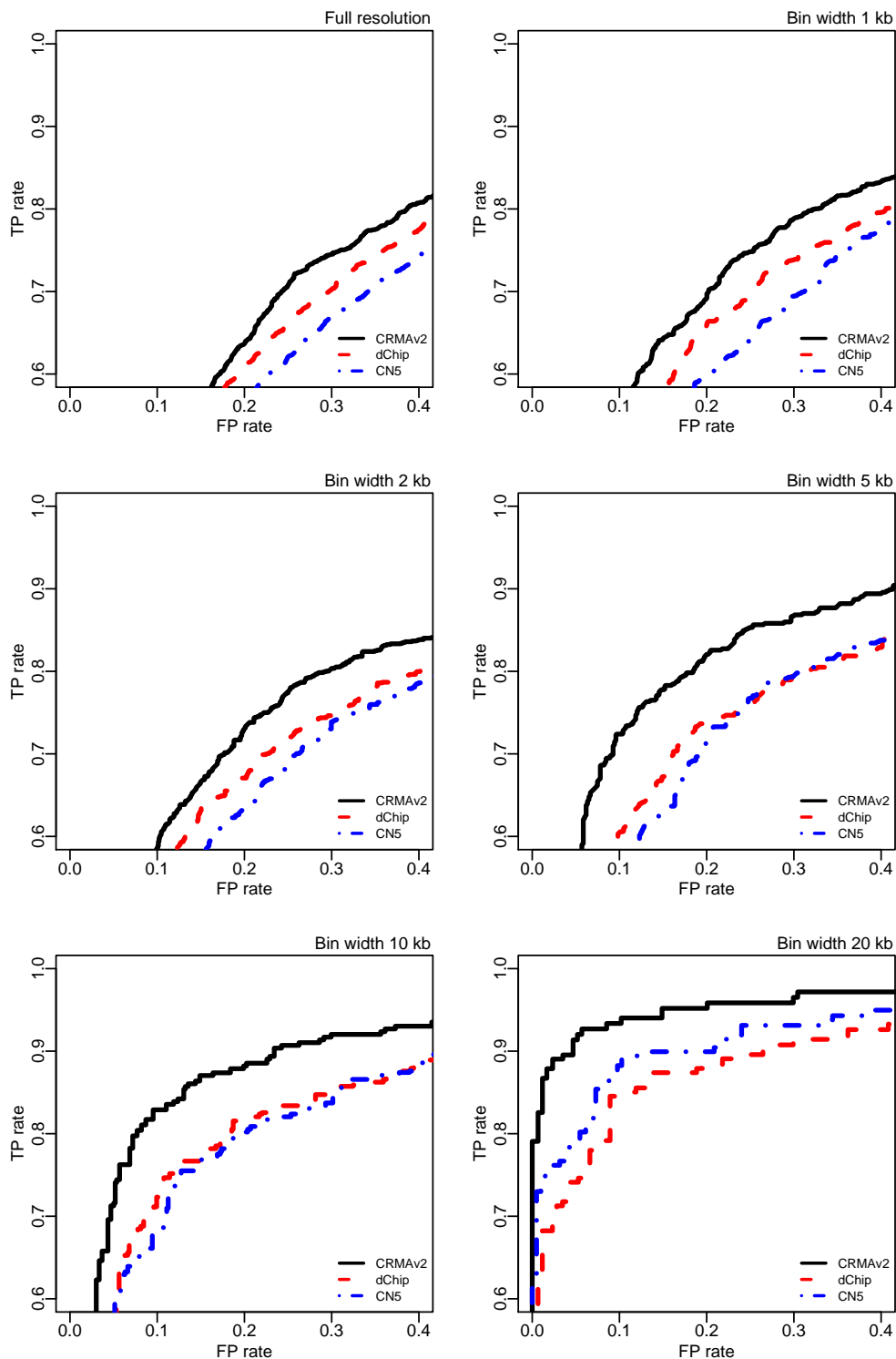


Figure 2: (GSM337641:Chr1@100.1-107.5,cp=103.8+/-0.25,s=0/-1) ROC curves for each of the 3 preprocessing methods at the full resolution as well as 5 different amounts of smoothing.

4.2 Region: GSM337641:Chr3@80-90.9,cp=85.3+/-0.25,s=-1/0

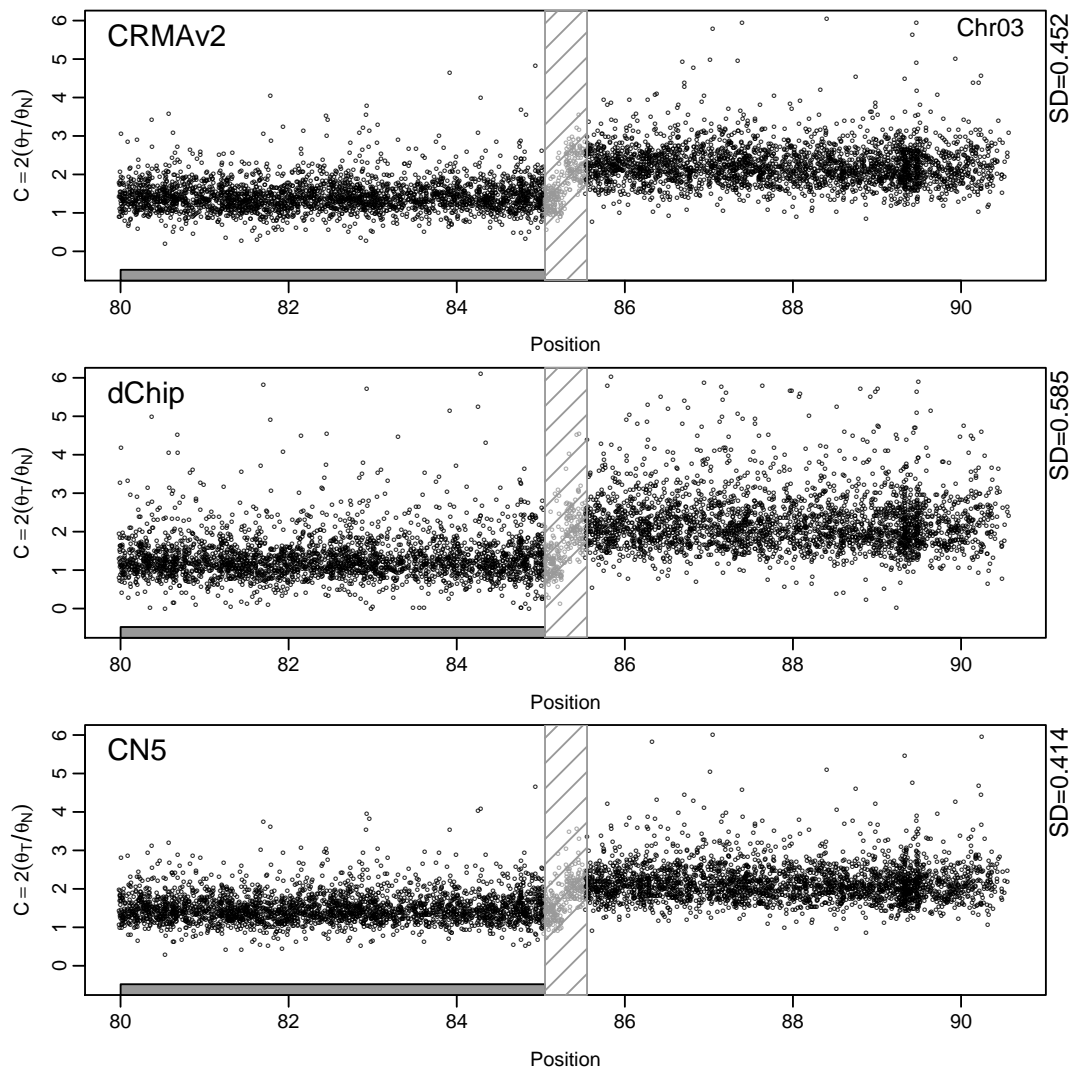


Figure 3: (GSM337641:Chr3@80-90.9,cp=85.3+/-0.25,s=-1/0) There are 2720 loci of state 'COPY NEUTRAL' ($s=0$) ("negatives") and 2764 loci of state 'LOSS' ($s=-1$) ("positives"), where the latter are highlighted with a solid bar beneath. In total 259 loci within the safety margin were excluded.

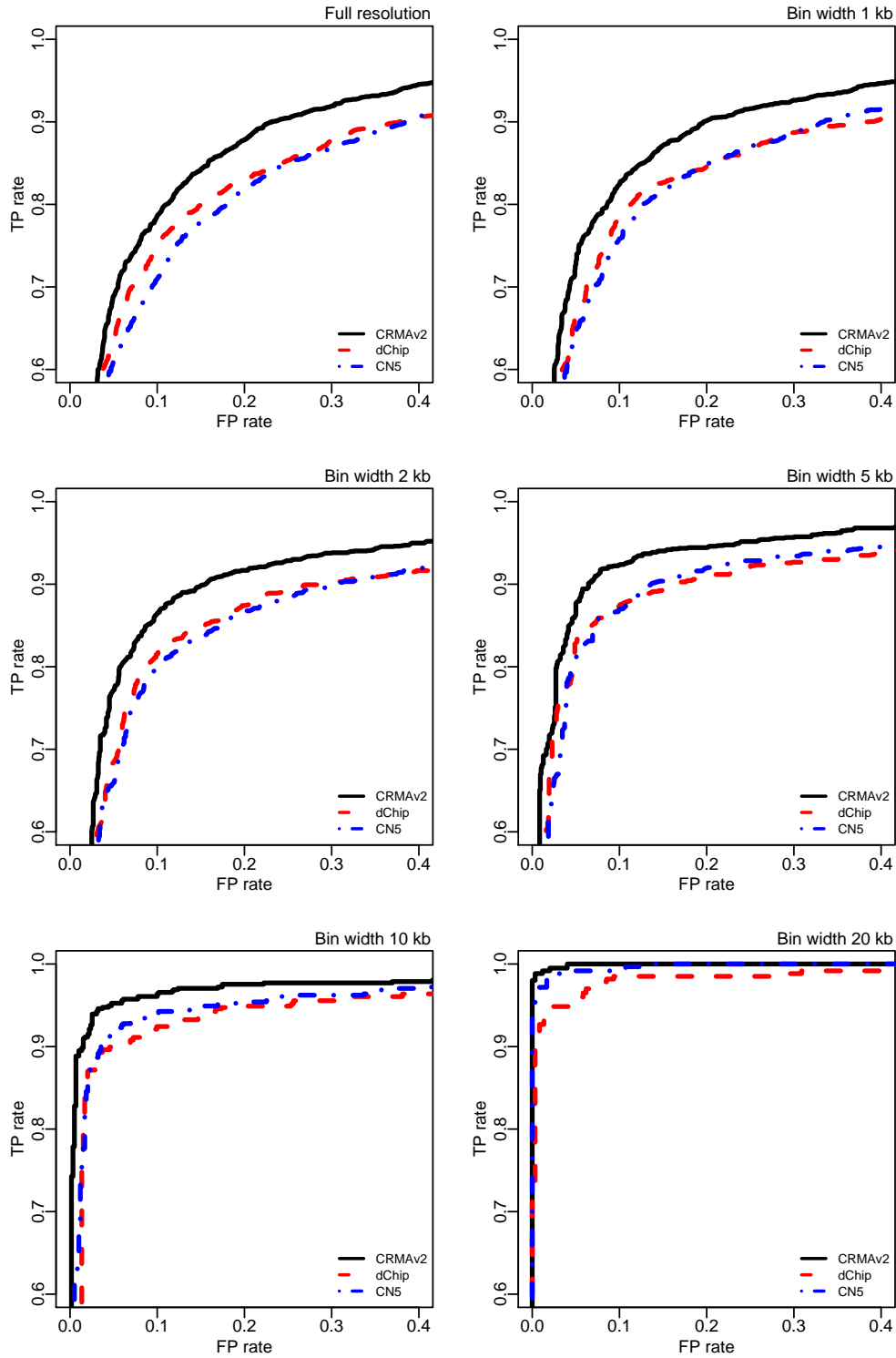


Figure 4: (GSM337641:Chr3@80-90.9,cp=85.3+/-0.25,s=-1/0) ROC curves for each of the 3 pre-processing methods at the full resolution as well as 5 different amounts of smoothing.

4.3 Region: GSM337641:Chr4@60.5-65.75,cp=63.40+/-0.25,s=0/+1

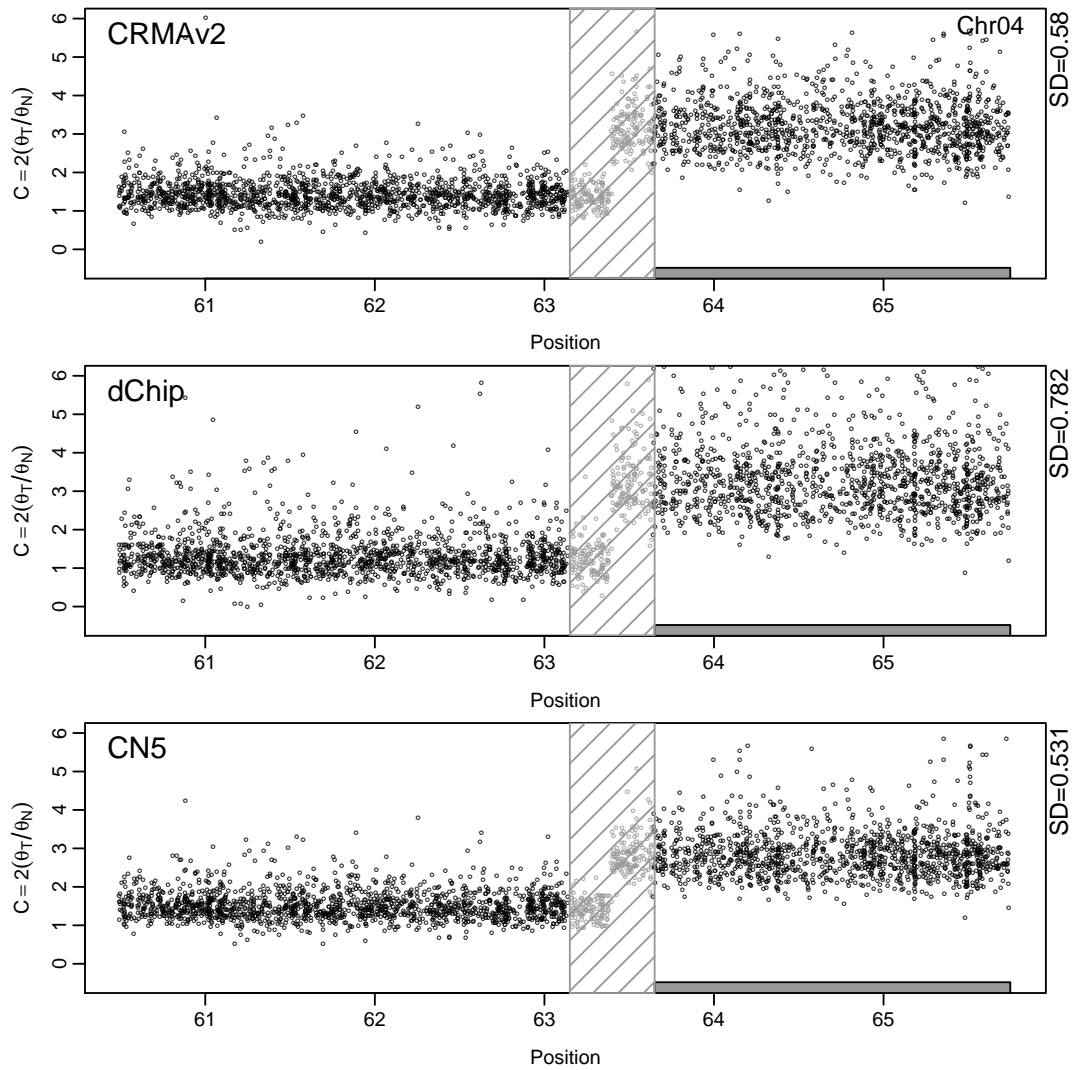


Figure 5: (GSM337641:Chr4@60.5-65.75,cp=63.40+/-0.25,s=0/+1) There are 1560 loci of state 'COPY NEUTRAL' ($s=0$) ("negatives") and 1288 loci of state 'GAIN' ($s=+1$) ("positives"), where the latter are highlighted with a solid bar beneath. In total 266 loci within the safety margin were excluded.

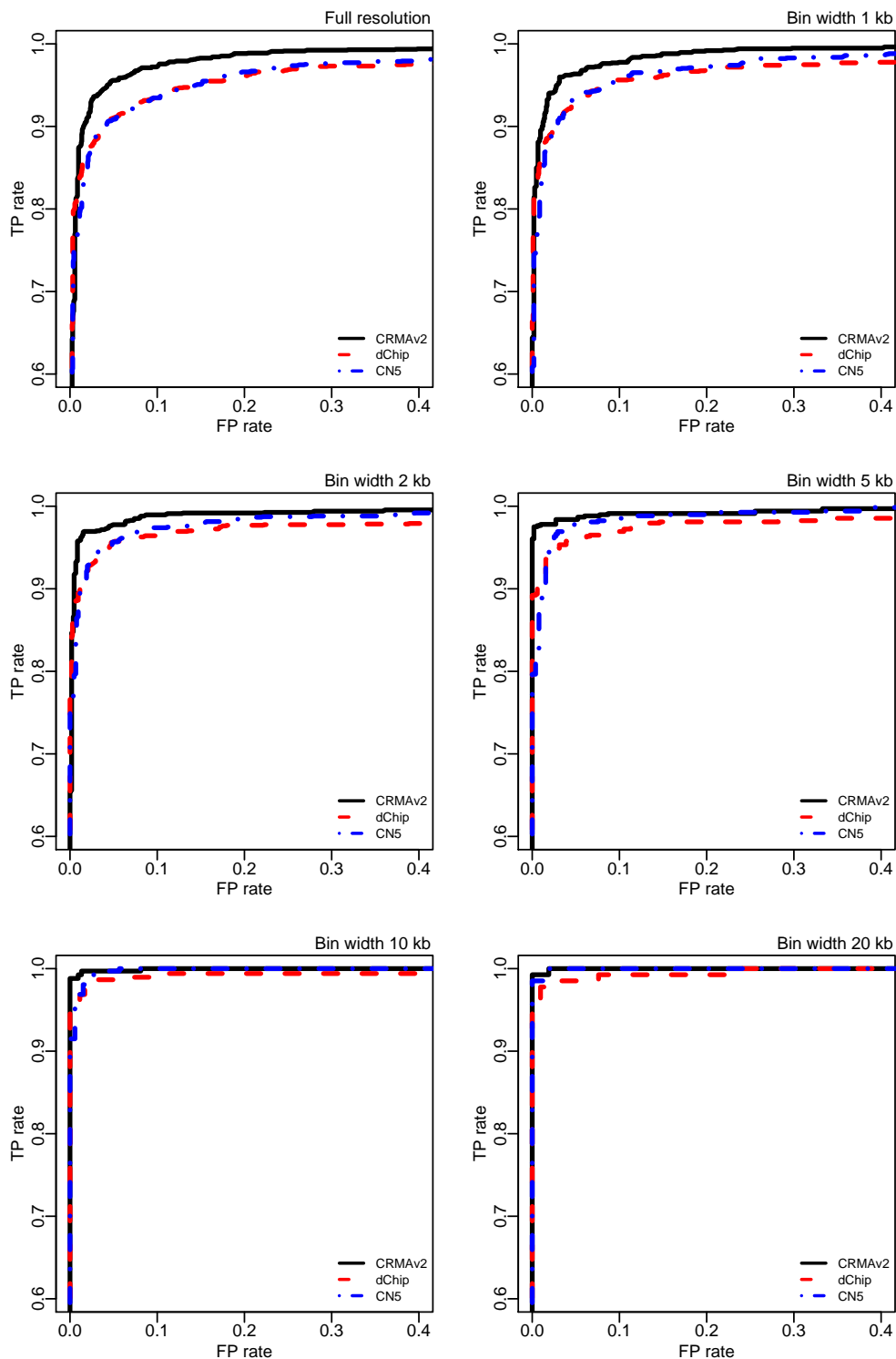


Figure 6: (GSM337641:Chr4@60.5-65.75,cp=63.40+/-0.25,s=0/+1) ROC curves for each of the 3 preprocessing methods at the full resolution as well as 5 different amounts of smoothing.

4.4 Region: GSM337641:Chr10@61-69,cp=65.30+/-0.25,s=+1/0

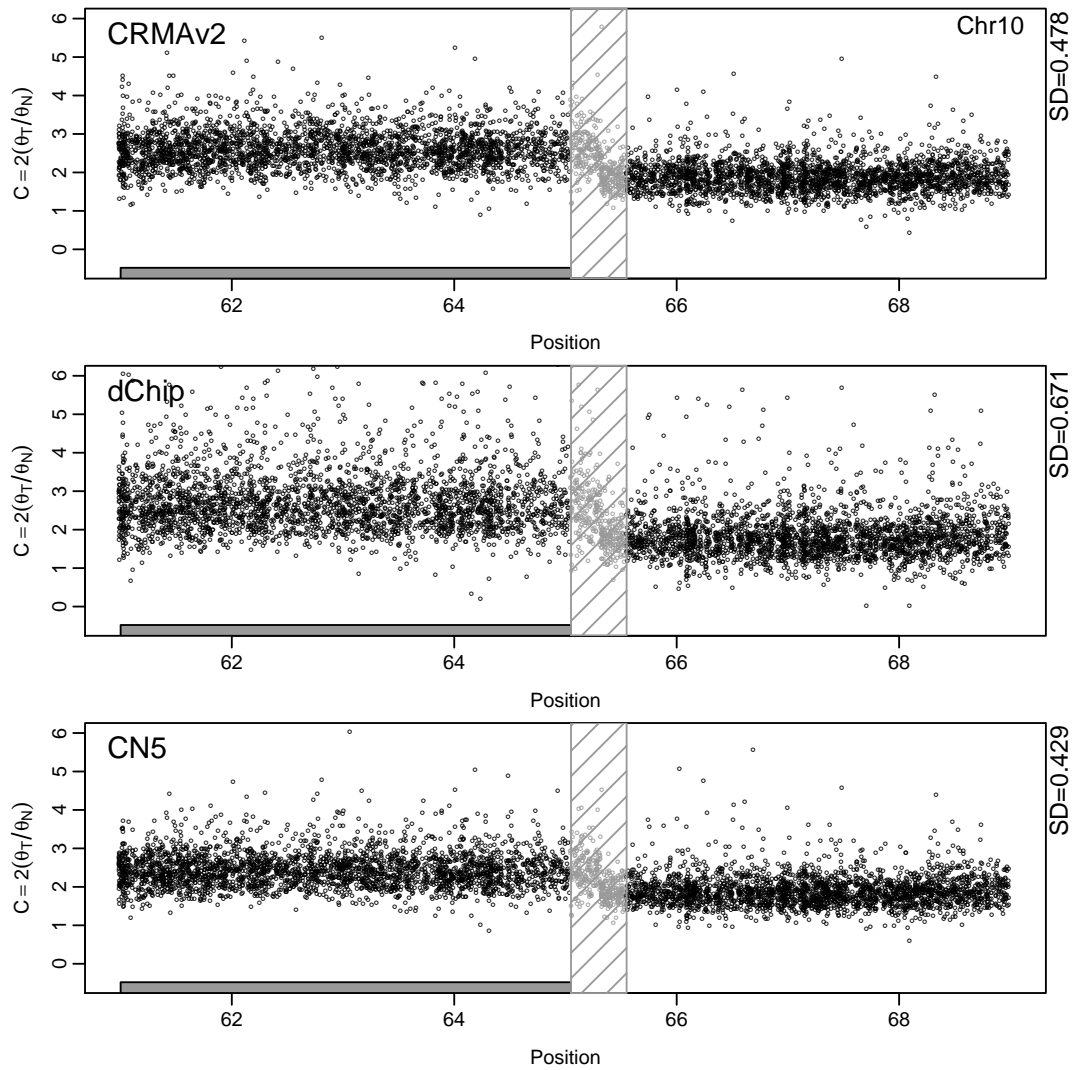


Figure 7: (GSM337641:Chr10@61-69,cp=65.30+/-0.25,s=+1/0) There are 2480 loci of state 'COPY NEUTRAL' ($s=0$) ("negatives") and 2805 loci of state 'GAIN' ($s=+1$) ("positives"), where the latter are highlighted with a solid bar beneath. In total 355 loci within the safety margin were excluded.

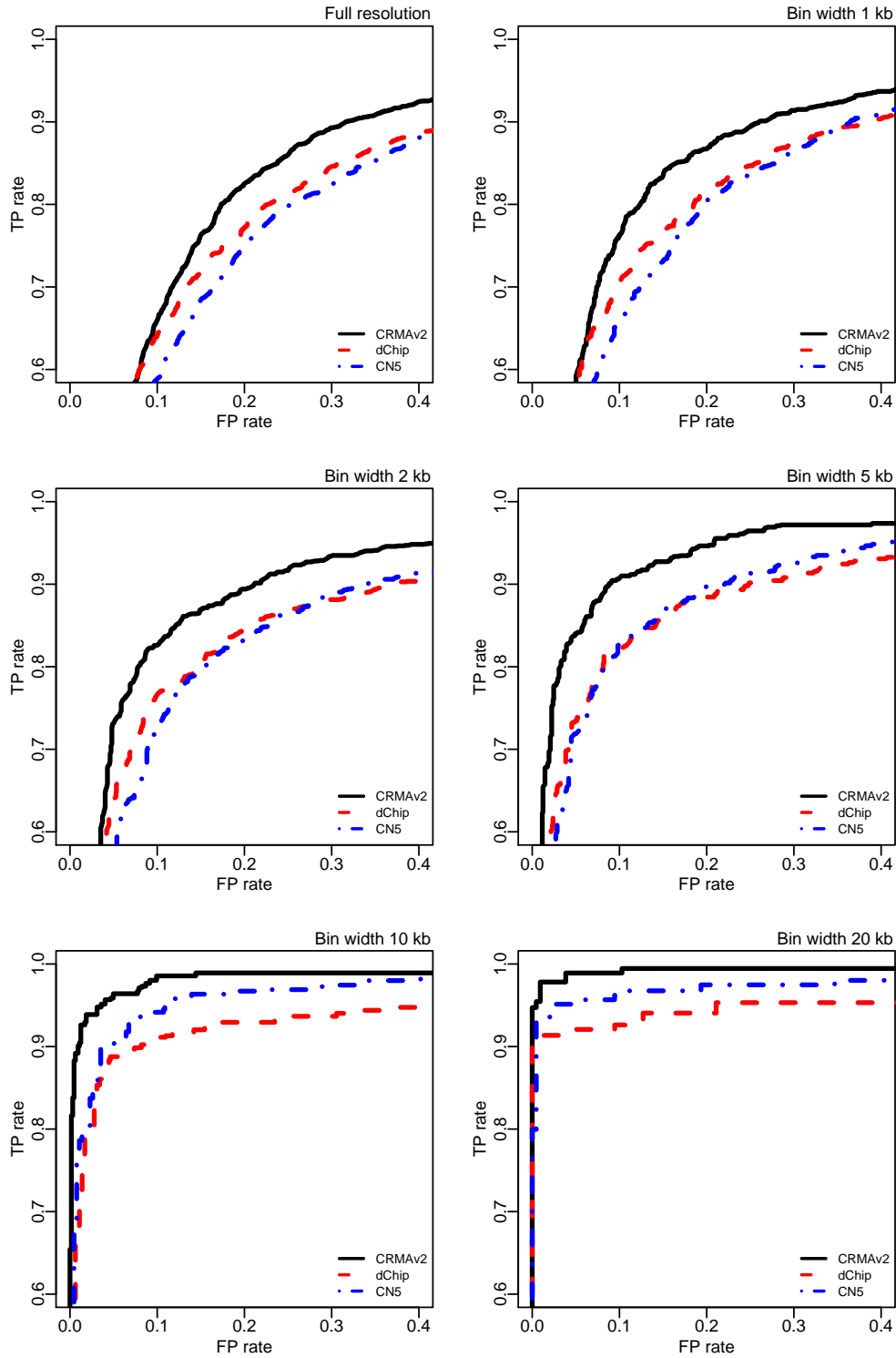


Figure 8: (GSM337641:Chr10@61-69,cp=65.30+/-0.25,s=+1/0) ROC curves for each of the 3 preprocessing methods at the full resolution as well as 5 different amounts of smoothing.

4.5 Region: GSM337641:Chr11@78.2-83,cp=80.2+/-0.25,s=0/-1

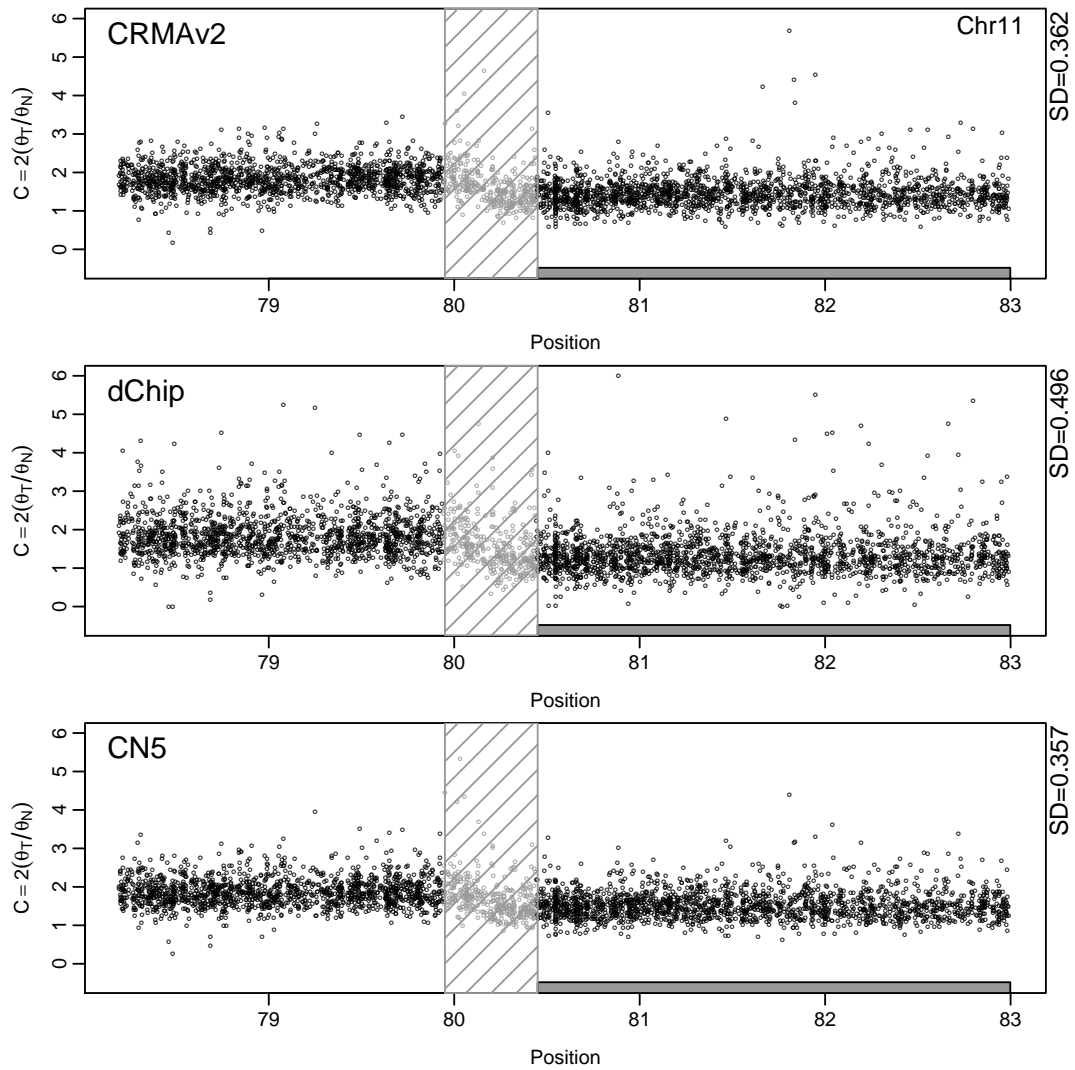


Figure 9: (GSM337641:Chr11@78.2-83,cp=80.2+/-0.25,s=0/-1) There are 1481 loci of state 'COPY NEUTRAL' ($s=0$) ("negatives") and 1885 loci of state 'LOSS' ($s=-1$) ("positives"), where the latter are highlighted with a solid bar beneath. In total 356 loci within the safety margin were excluded.

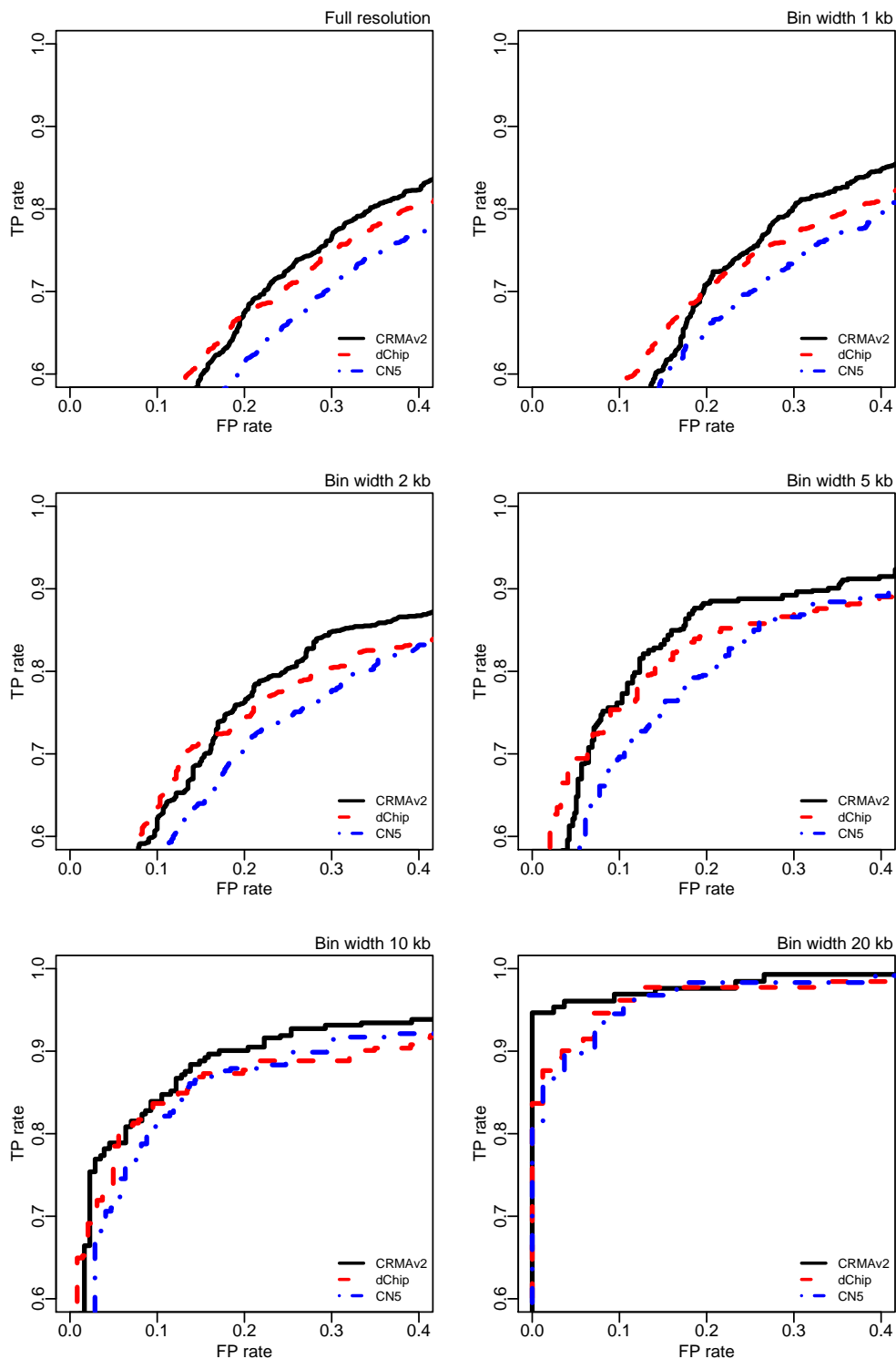


Figure 10: (GSM337641:Chr11@78.2-83,cp=80.2+/-0.25,s=0/-1) ROC curves for each of the 3 preprocessing methods at the full resolution as well as 5 different amounts of smoothing.

4.6 Region: GSM337641:Chr12@57-63,cp=59.8+/-0.25,s=+1/0

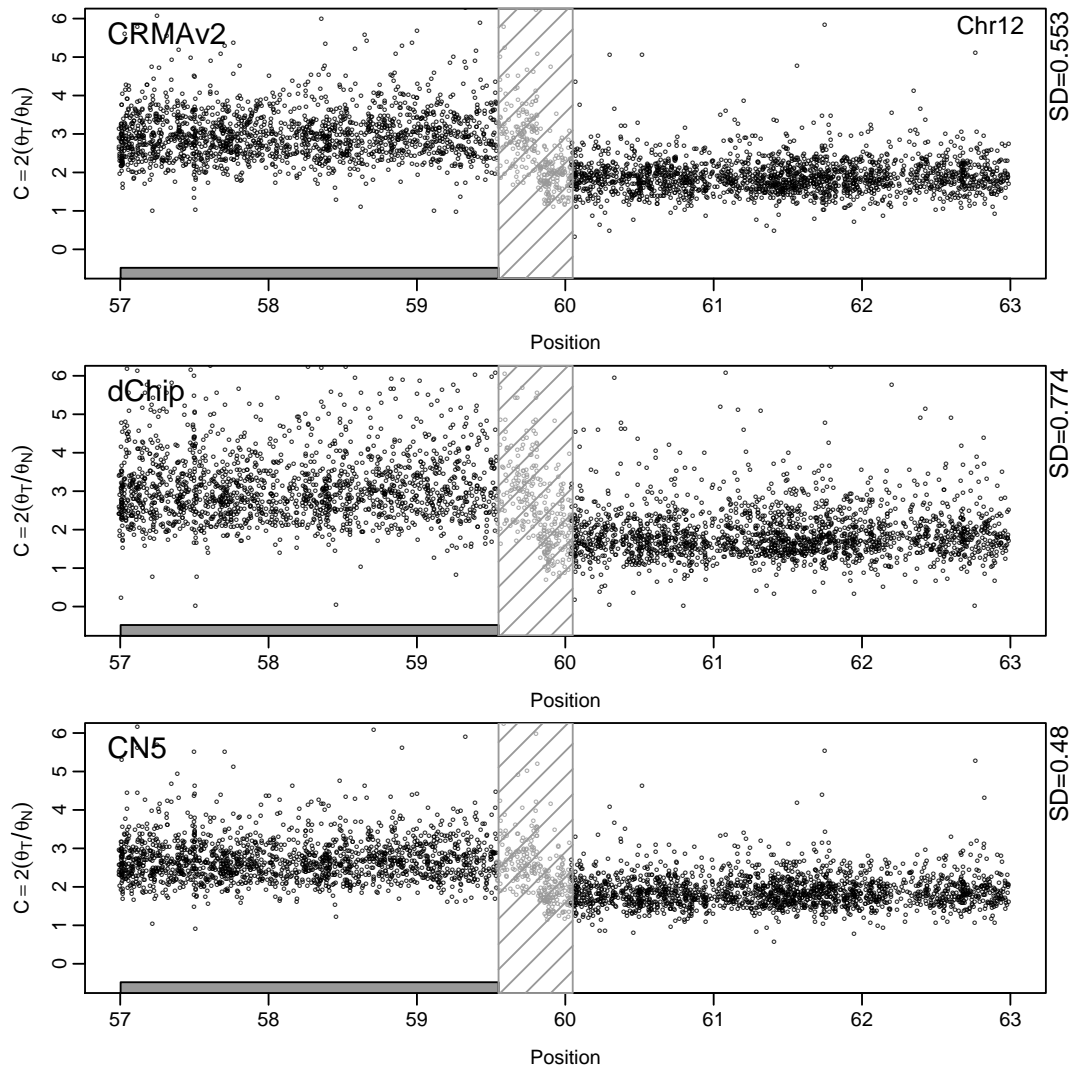


Figure 11: (GSM337641:Chr12@57-63,cp=59.8+/-0.25,s=+1/0) There are 1864 loci of state 'COPY NEUTRAL' (s=0) ("negatives") and 1716 loci of state 'GAIN' (s=+1) ("positives"), where the latter are highlighted with a solid bar beneath. In total 304 loci within the safety margin were excluded.

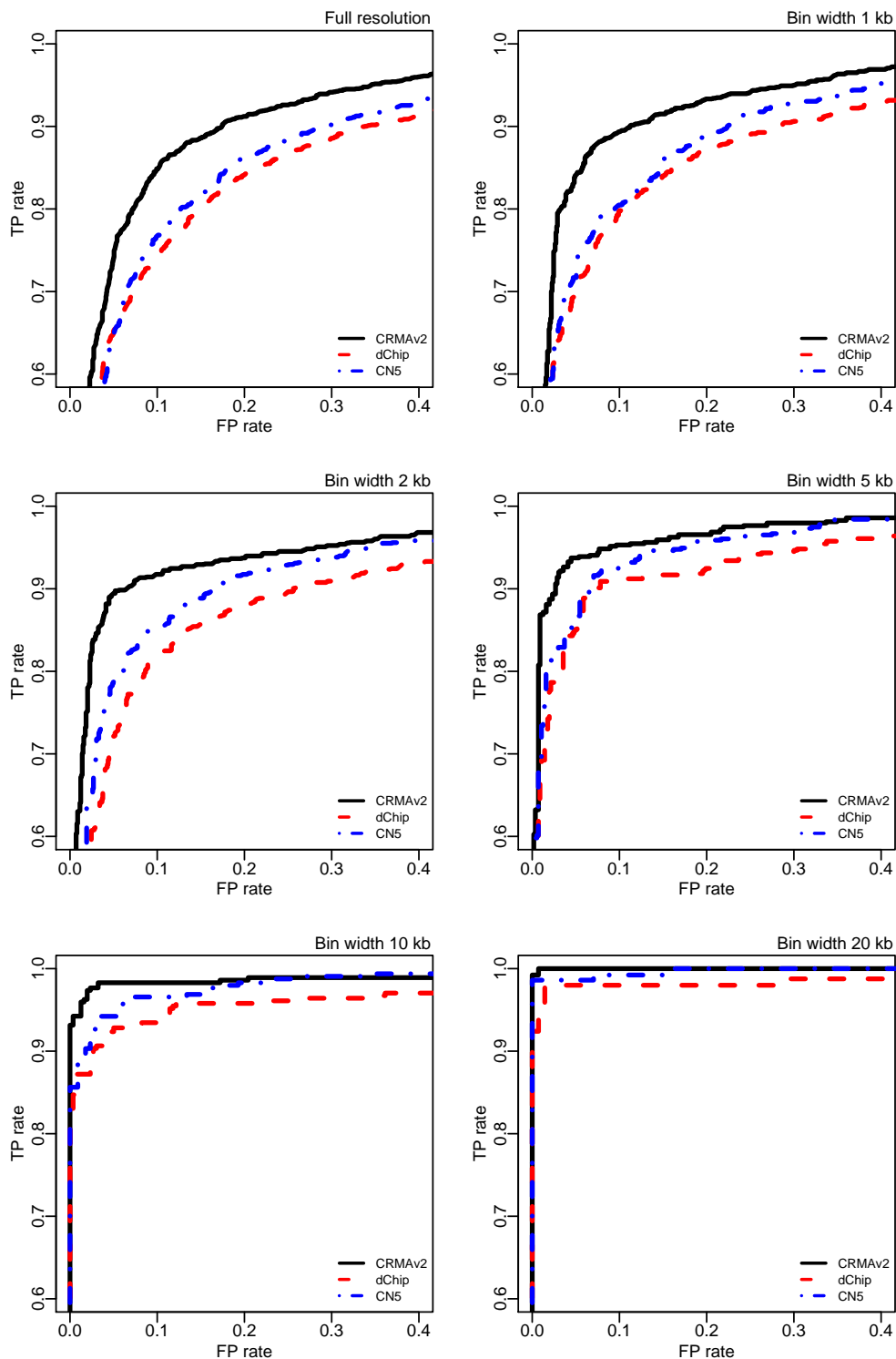


Figure 12: (GSM337641:Chr12@57-63,cp=59.8+/-0.25,s=+1/0) ROC curves for each of the 3 preprocessing methods at the full resolution as well as 5 different amounts of smoothing.

References

- Bengtsson, H., Irizarry, R. A., Carvalho, B., and Speed, T. P. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**(6), 759–767.
- Chiang, D. Y., Getz, G., Jaffe, D. B., O’Kelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Met.*, **6**(1), 99–103.

A Data files

\$CRMAv2

AromaUnitTotalCnBinarySet:

Name: GSE13372

Tags: HCC1143,onePair,CRMAv2,pairs

Full name: GSE13372,HCC1143,onePair,CRMAv2,pairs

Number of files: 1

Names: GSM337641

Path (to the first file): rawCnData/GSE13372,HCC1143,onePair,CRMAv2,pairs/GenomeWideSNP'6

Total file size: 7.18 MB

RAM: 0.00MB

\$dChip

AromaUnitTotalCnBinarySet:

Name: GSE13372

Tags: HCC1143,onePair,dChip,pairs

Full name: GSE13372,HCC1143,onePair,dChip,pairs

Number of files: 1

Names: GSM337641

Path (to the first file): rawCnData/GSE13372,HCC1143,onePair,dChip,pairs/GenomeWideSNP'6

Total file size: 7.18 MB

RAM: 0.00MB

\$CN5

AromaUnitTotalCnBinarySet:

Name: GSE13372

Tags: HCC1143,onePair,GTC,pairs

Full name: GSE13372,HCC1143,onePair,GTC,pairs

Number of files: 1

Names: GSM337641

Path (to the first file): rawCnData/GSE13372,HCC1143,onePair,GTC,pairs/GenomeWideSNP'6

Total file size: 7.08 MB

RAM: 0.00MB

B Session information

This report was automatically generated using the R.rsp package.

- R version 2.8.1 Patched (2008-12-22 r47296), i386-pc-mingw32
- Locale: LC_COLLATE=English_United States.1252;LC_CTYPE=English_United States.1252;LC_MONETARY=English_United States.1252
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: aroma.cn.eval 0.0.3, aroma.core 1.0.4, aroma.light 1.11.2, digest 0.3.1, matrixStats 0.1.4, R.cache 0.1.7, R.menu 0.0.4, R.methodsS3 1.0.3, R.oo 1.4.6, R.rsp 0.3.5, R.utils 1.1.4