Supplementary Note: Assessment of TumorBoost based on tumor/normal pair TCGA-23-1027 in the Illumina Human1M-Duo data set TCGA,OV,BeadStudio,BAF using the SNPs with 90% highest confidence scores

Henrik Bengtsson, Pierre Neuvial, Terence P. Speed

March 9, 2010

Contents

1 Introduction

This report, which is automatically generated, assesses the performance of the TumorBoost method based on a few change points in a particular tumor/normal pair. For more details on the evaluation methods, see the main TumorBoost manuscript.

2 Data set

The evaluation is this report is based on the tumor/normal pair (01A,10A) for individual TCGA-23-1027 in the data set TCGA,OV,BeadStudio,BAF.

2.1 Preprocessing methods

The data was generated on the Illumina Human1M-Duo chip type. Each array was preprocessed using Birdseed's "BAF" normalization method Peiffer *et al.* (2006), which is a multi-array (population-based) method.

2.2 Stratification on genotype confidence scores

We focus on the SNPs in which we are the most confident that they are heterozygous: the evaluation will involve the 90% SNPs with highest genotype confidence scores.

2.3 List of change points

For this data set, we have selected a few regions for which one can safely assume that there exists a single copy number change point. By definition, each change point separates two sets of genomic loci such that the true Decrease in Heterozygosity (DH) is the same within one set of loci but differs between the two sets. These regions were selected visually. For each region we chose a large enough safety margin to make our evaluation independent of the uncertainty on the true location of the change point.

Chr	Region	Change point	Margin	Before	After
2	108-140	124	0.5	'normal' (1,1)	'gain' (1,2)
2	125-157	141	0.5	'gain' (1,2)	'copy neutral LOH' (0,2)
10	80-109	94	0.5	'normal' (1,1)	'deletion' (0,1)
10	106.5-113.5	110	0.5	'deletion' (0,1)	'copy neutral LOH' (0,2)
2	55-75	65	0.5	'normal' (1,1)	'gain' (1,2)

Table 1: Regions in TCGA-23-1027 used for the evaluation and that each contain a single changepoint. All positions and lengths are in units of Mb.

We next compare how well each of these change points is detected using the above preprocessed signals followed or not by TumorBoost normalization using the ROC analysis described in the main TumorBoost manuscript at the full resolution as well as smoothed resolution with bin sizes $h = \{1, 2, 4\}$. Specifically, we compare the following three methods: (1) **"raw"**: preprocessed signals without TumorBoost normalization. (2) **"TBN,NGC"**: preprocessed signals followed by TumorBoost normalization using NGC genotype calls. For completeness we also include an evaluation of Total copy numbers (TCN).

3 Region: TCGA-23-1027:Chr2@108-140,cp=124+/-0.5,s=0/1

3.1 Decrease in Heterozygosity (DH) and total copy-number tracks

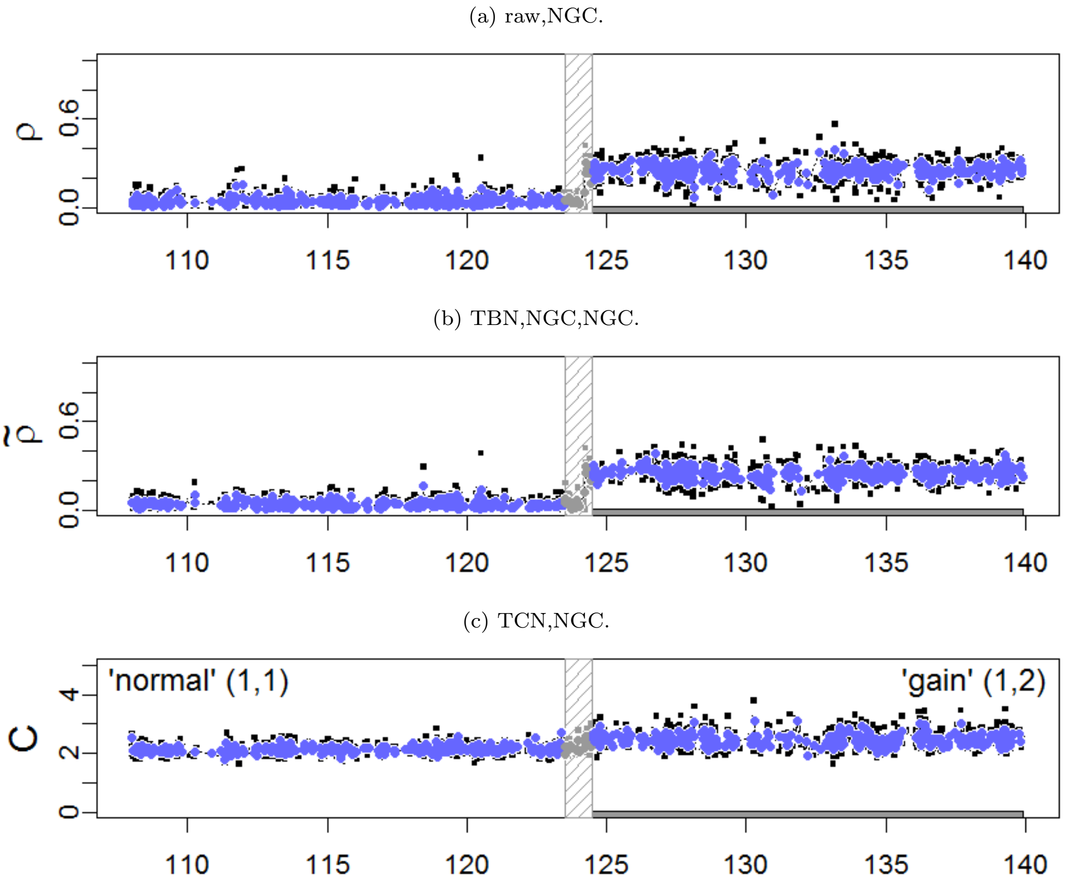
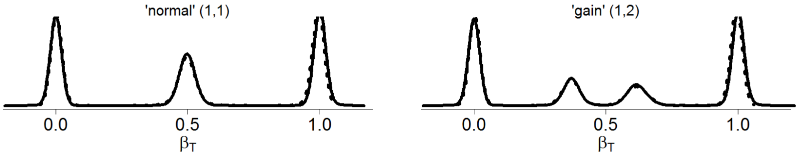


Figure 1: Decrease in Heterozygosity (DH) and total copy numbers for region TCGA-23-1027:Chr2@108-140,cp=124+/-0.5,s=0/1. Only heterozygous SNPs are plotted. There are 1444 loci of state 'normal' (1,1) ("negatives") and 1444 loci of state 'gain' (1,2) ("positives"), where the latter are highlighted with a solid bar beneath. In total 80 loci within the safety margin were excluded.

3.2 Allele B fraction density plots



(a) 'normal' (1,1).

(b) 'gain' (1,2).

Figure 2: Density of raw (dashed lines) and TumorBoost-normalized (solid lines) allele B fractions for region TCGA-23-1027:Chr2@108-140,cp=124+/-0.5,s=0/1.

3.3 ROC curves

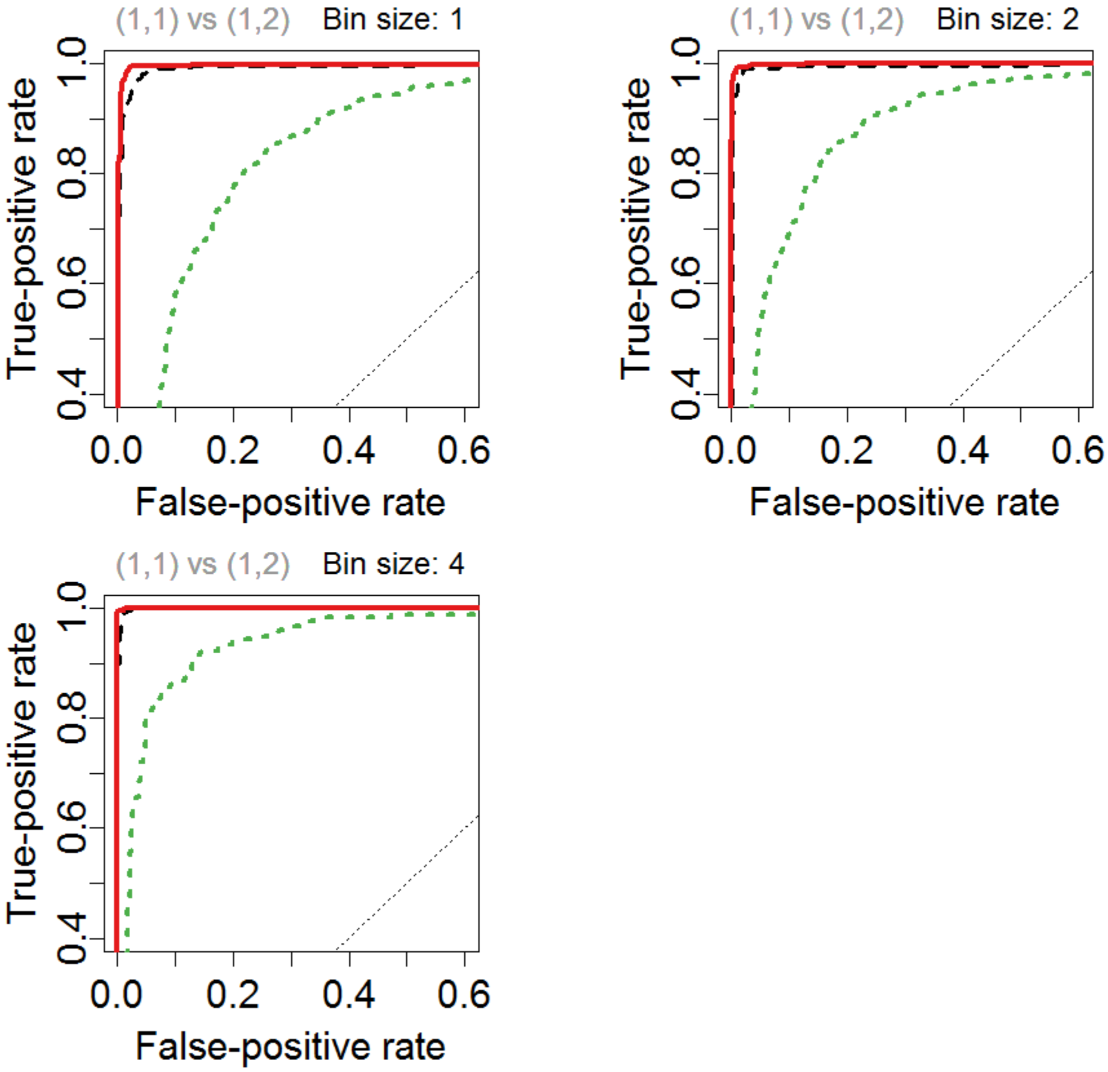


Figure 3: ROC curves for each preprocessing method at the full resolution as well as 2 different amounts of smoothing (using the mean() function) for region TCGA-23-1027:Chr2@108-140,cp=124+/-0.5,s=0/1. Legend: raw,NGC (dashed; #000000), TBN,NGC,NGC (solid; #E41A1C) and TCN,NGC (dotted; #4DAF4A).

3.4 (β_N, β_T) plots

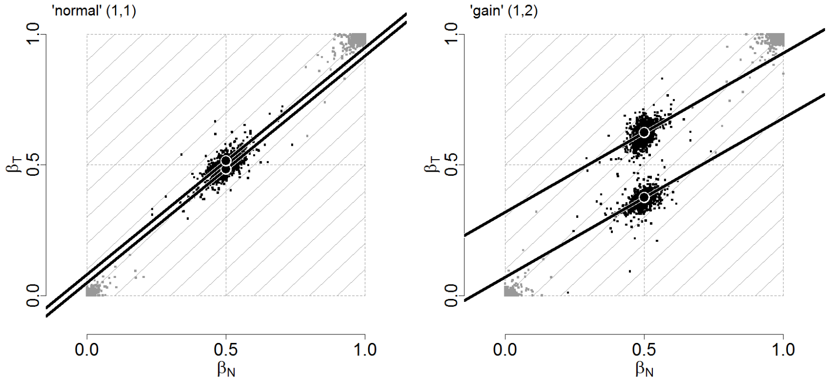


Figure 4: raw,NGC for region TCGA-23-1027:Chr2@108-140,cp=124+/-0.5,s=0/1.

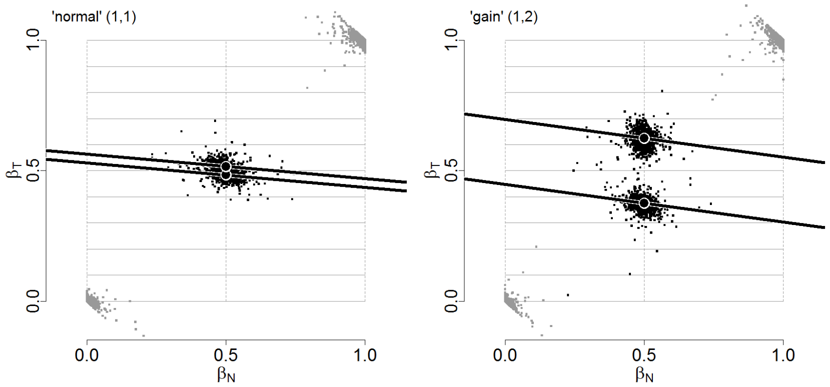


Figure 5: TBN,NGC,NGC for region TCGA-23-1027:Chr2@108-140,cp=124+/-0.5,s=0/1.

3.5 Allele-specific copy number estimates

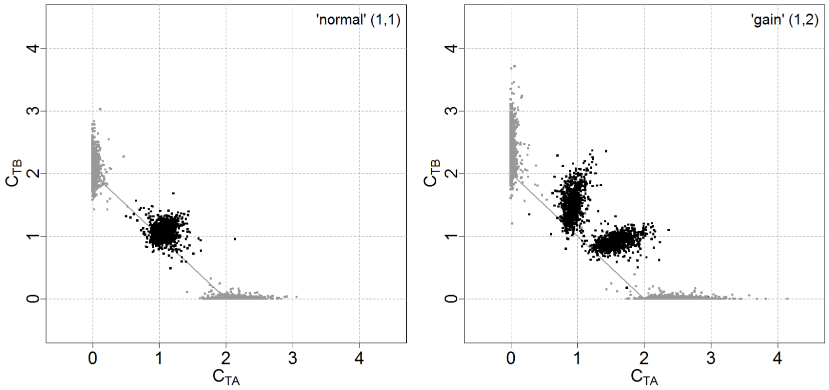


Figure 6: raw,NGC for region TCGA-23-1027:Chr2@108-140,cp=124+/-0.5,s=0/1.

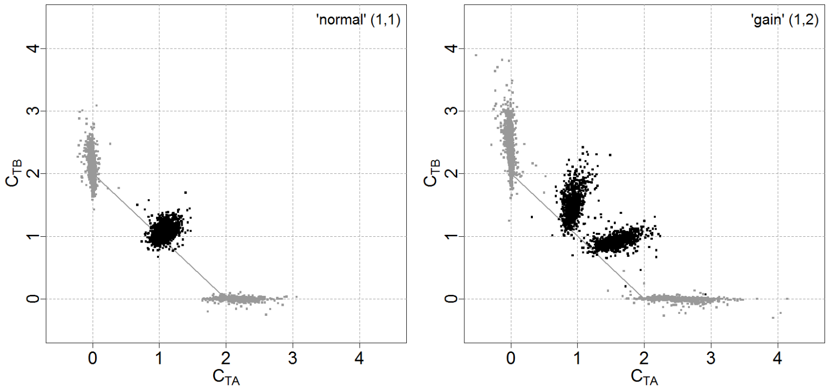
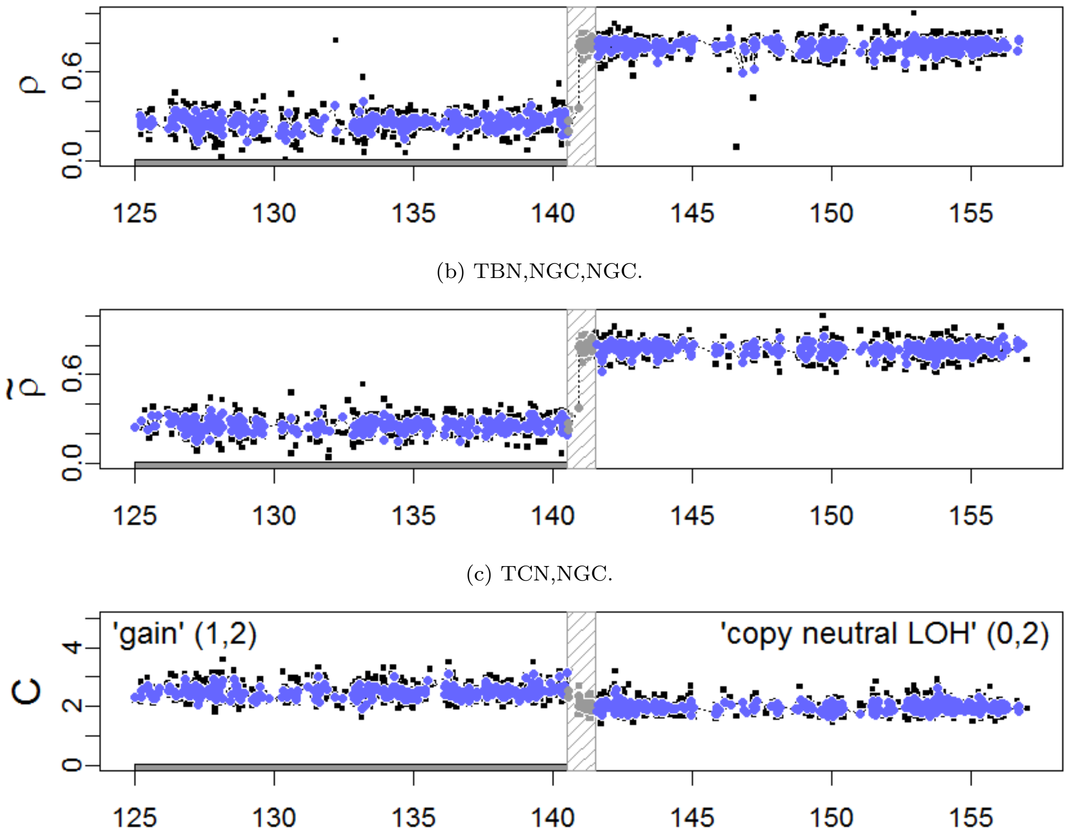


Figure 7: TBN,NGC,NGC for region TCGA-23-1027:Chr2@108-140,cp=124+/-0.5,s=0/1.

4 Region: TCGA-23-1027:Chr2@125.0-157.0,cp=141.0+/-0.5,s=1/3

4.1 Decrease in Heterozygosity (DH) and total copy-number tracks

(a) raw,NGC.

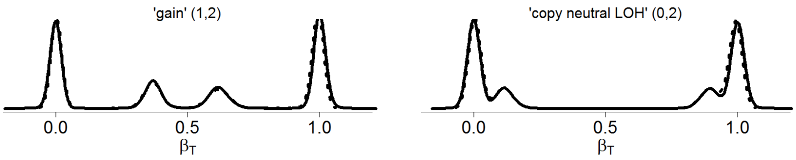


(b) TBN,NGC,NGC.

(c) TCN,NGC.

Figure 8: Decrease in Heterozygosity (DH) and total copy numbers for region TCGA-23-1027:Chr2@125.0-157.0,cp=141.0+/-0.5,s=1/3. Only heterozygous SNPs are plotted. There are 1145 loci of state 'gain' (1,2) ("negatives") and 1145 loci of state 'copy neutral LOH' (0,2) ("positives"), where the latter are highlighted with a solid bar beneath. In total 68 loci within the safety margin were excluded.

4.2 Allele B fraction density plots



(a) 'gain' (1,2).

(b) 'copy neutral LOH' (0,2).

Figure 9: Density of raw (dashed lines) and TumorBoost-normalized (solid lines) allele B fractions for region TCGA-23-1027:Chr2@125.0-157.0,cp=141.0+/-0.5,s=1/3.

4.3 ROC curves

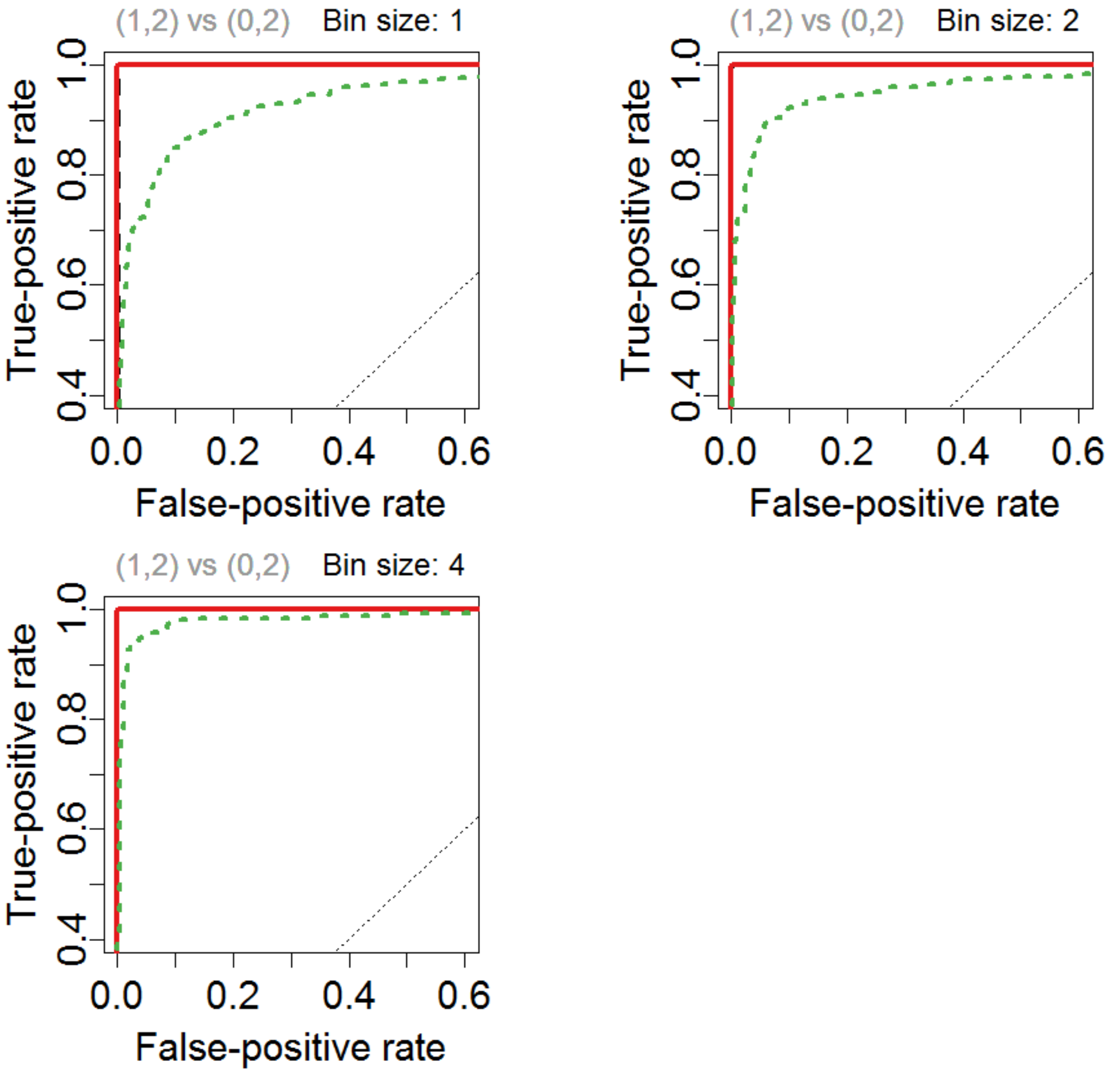


Figure 10: ROC curves for each preprocessing method at the full resolution as well as 2 different amounts of smoothing (using the mean() function) for region TCGA-23-1027:Chr2@125.0-157.0,cp=141.0+/-0.5,s=1/3. Legend: raw,NGC (dashed; #000000), TBN,NGC,NGC (solid; #E41A1C) and TCN,NGC (dotted; #4DAF4A).

4.4 (β_N, β_T) plots

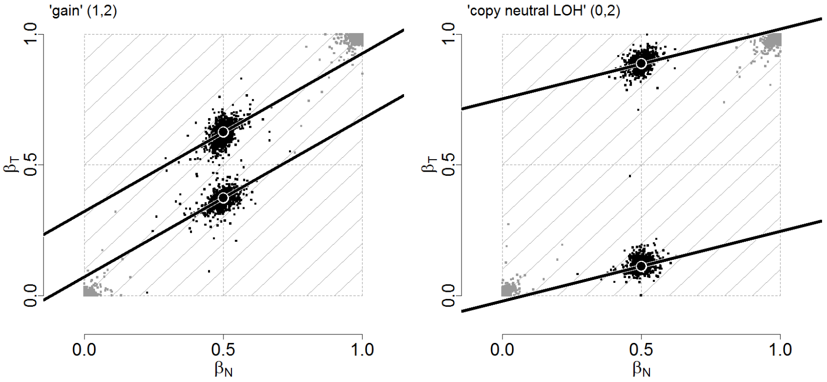


Figure 11: raw,NGC for region TCGA-23-1027:Chr2@125.0-157.0,cp=141.0+/-0.5,s=1/3.

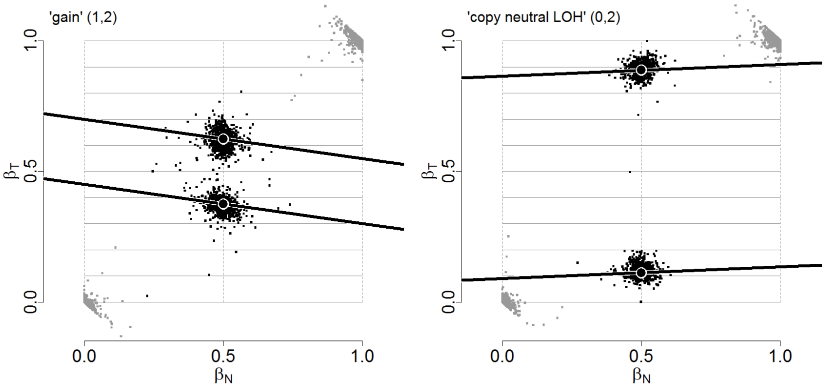


Figure 12: TBN,NGC,NGC for region TCGA-23-1027:Chr2@125.0-157.0,cp=141.0+/-0.5,s=1/3.

4.5 Allele-specific copy number estimates

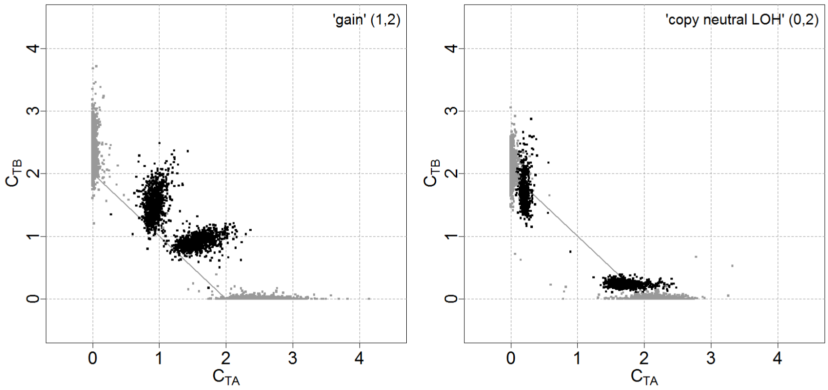


Figure 13: raw,NGC for region TCGA-23-1027:Chr2@125.0-157.0,cp=141.0+/-0.5,s=1/3.

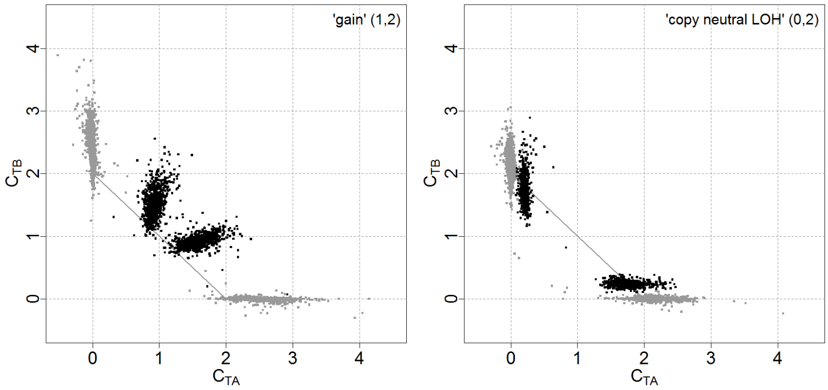
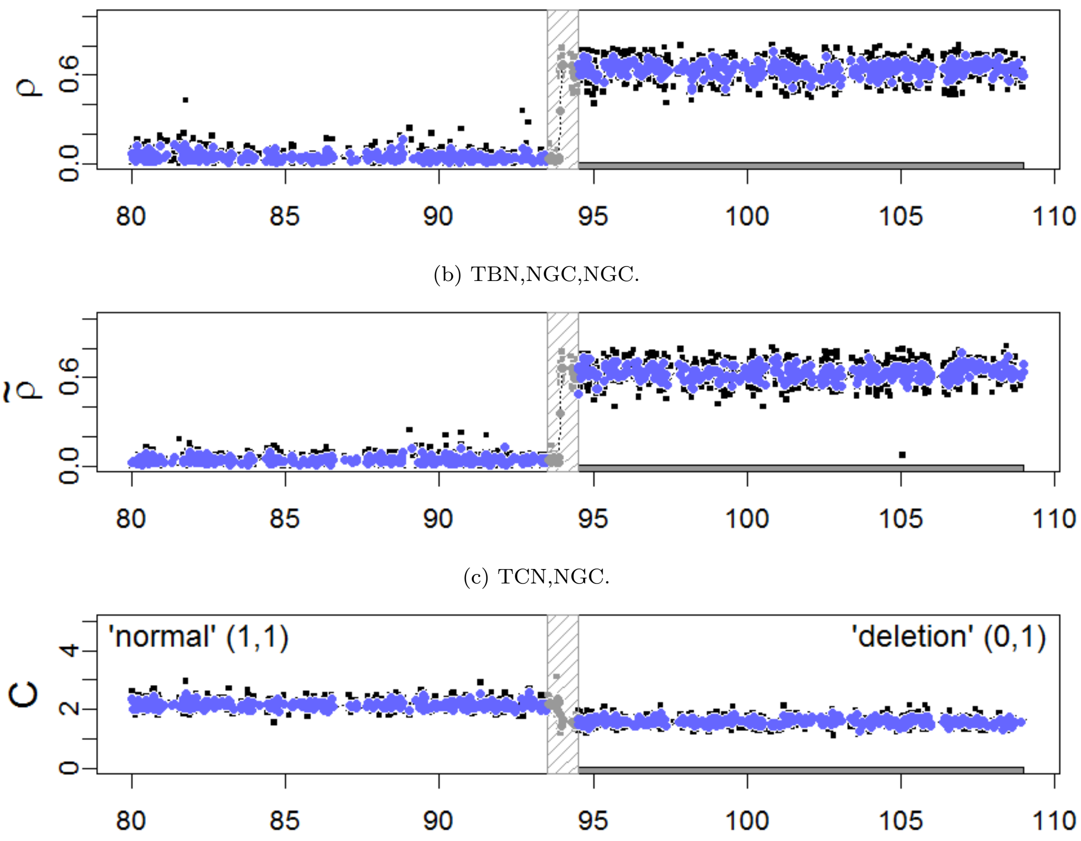


Figure 14: TBN,NGC,NGC for region TCGA-23-1027:Chr2@125.0-157.0,cp=141.0+/-0.5,s=1/3.

5 Region: TCGA-23-1027:Chr10@80-109,cp=94+/-0.5,s=0/2

5.1 Decrease in Heterozygosity (DH) and total copy-number tracks



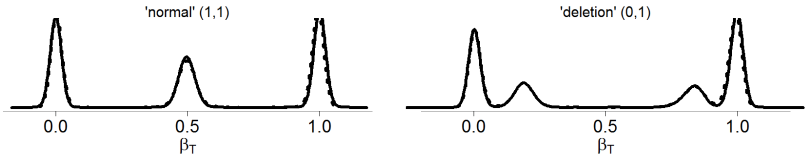
(a) raw,NGC.

(b) TBN,NGC,NGC.

(c) TCN,NGC.

Figure 15: Decrease in Heterozygosity (DH) and total copy numbers for region TCGA-23-1027:Chr10@80-109,cp=94+/-0.5,s=0/2. Only heterozygous SNPs are plotted. There are 1374 loci of state 'normal' (1,1) ("negatives") and 1374 loci of state 'deletion' (0,1) ("positives"), where the latter are highlighted with a solid bar beneath. In total 91 loci within the safety margin were excluded.

5.2 Allele B fraction density plots



(a) 'normal' (1,1).

(b) 'deletion' (0,1).

Figure 16: Density of raw (dashed lines) and TumorBoost-normalized (solid lines) allele B fractions for region TCGA-23-1027:Chr10@80-109,cp=94+/-0.5,s=0/2.

5.3 ROC curves

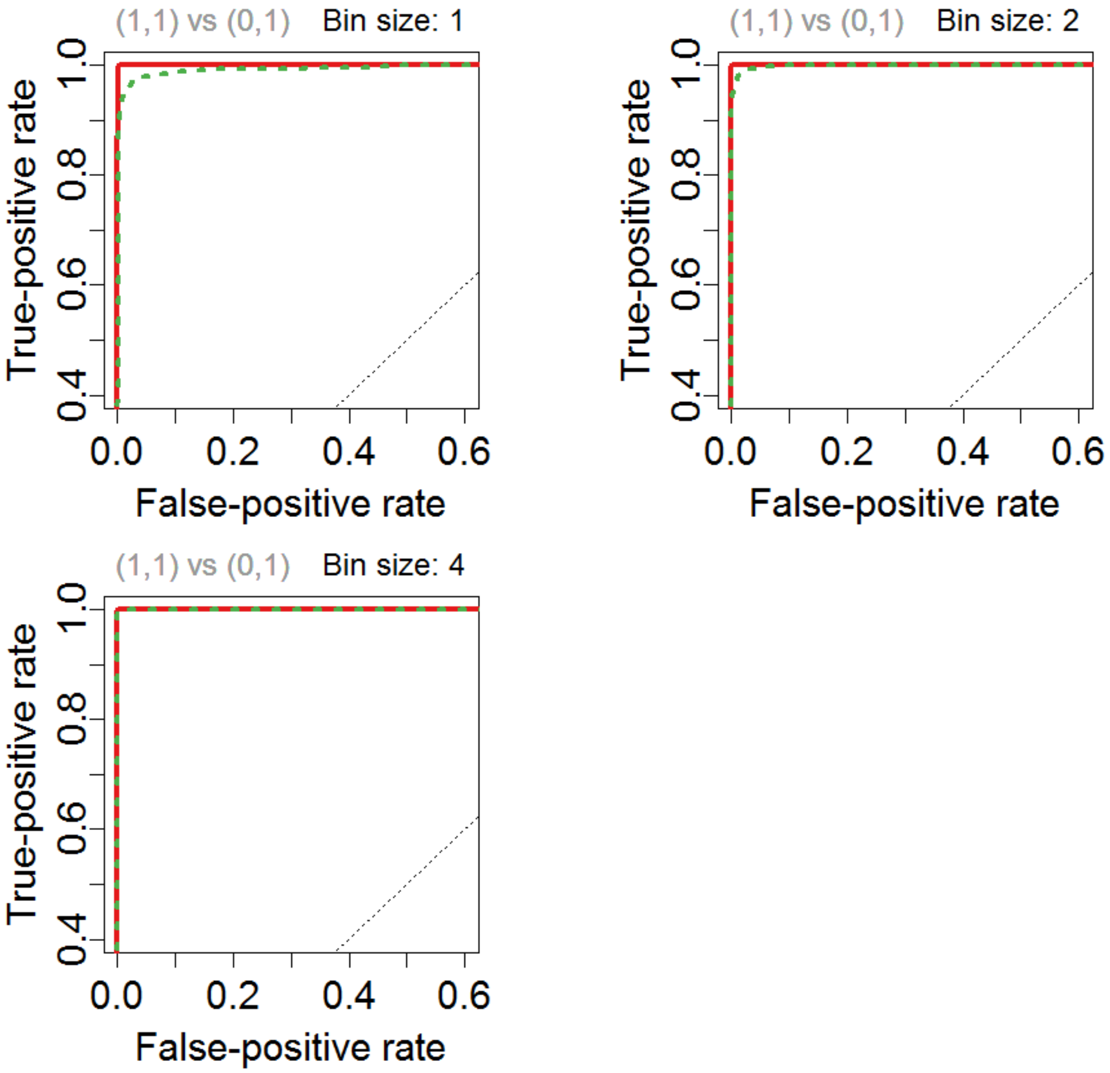


Figure 17: ROC curves for each preprocessing method at the full resolution as well as 2 different amounts of smoothing (using the mean() function) for region TCGA-23-1027:Chr10@80-109,cp=94+/-0.5,s=0/2. Legend: raw,NGC (dashed; #000000), TBN,NGC,NGC (solid; #E41A1C) and TCN,NGC (dotted; #4DAF4A).

5.4 (β_N, β_T) plots

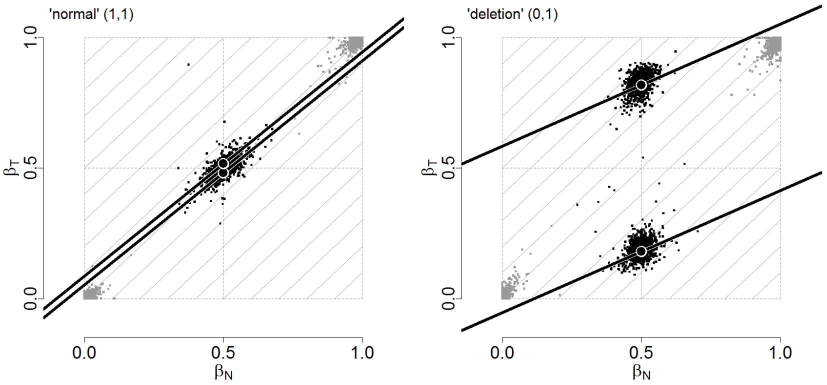


Figure 18: raw,NGC for region TCGA-23-1027:Chr10@80-109,cp=94+/-0.5,s=0/2.

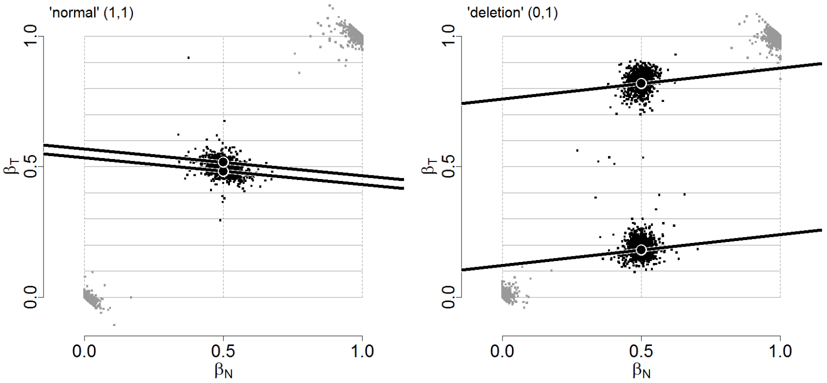


Figure 19: TBN,NGC,NGC for region TCGA-23-1027:Chr10@80-109,cp=94+/-0.5,s=0/2.

5.5 Allele-specific copy number estimates

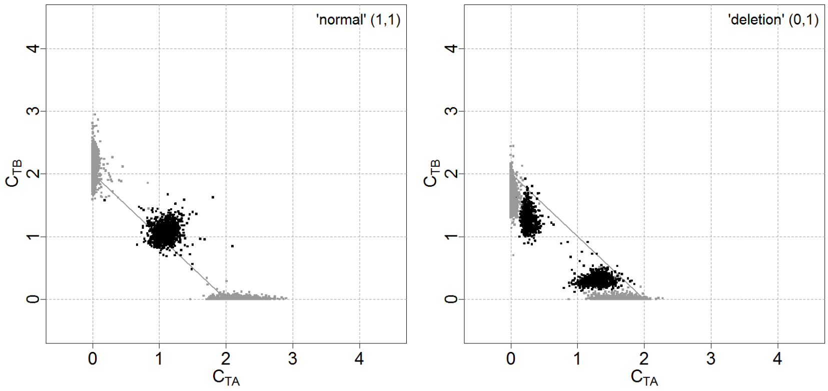


Figure 20: raw,NGC for region TCGA-23-1027:Chr10@80-109,cp=94+/-0.5,s=0/2.

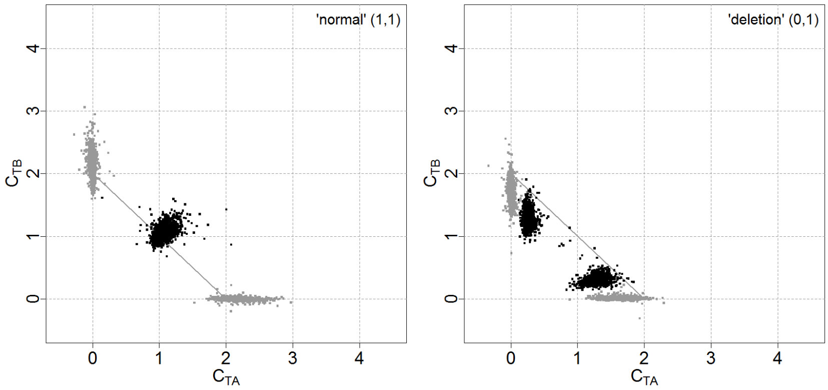
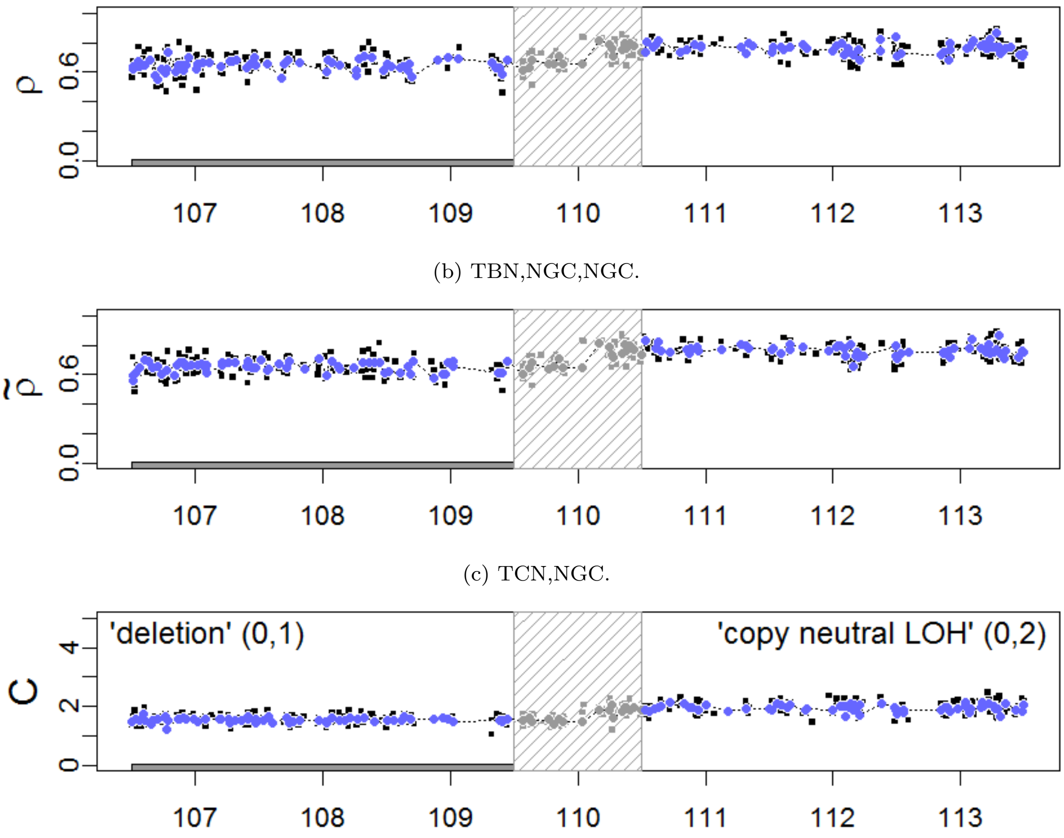


Figure 21: TBN,NGC,NGC for region TCGA-23-1027:Chr10@80-109,cp=94+/-0.5,s=0/2.

6 Region: TCGA-23-1027:Chr10@106.5-113.5,cp=110+/-0.5,s=2/3

6.1 Decrease in Heterozygosity (DH) and total copy-number tracks

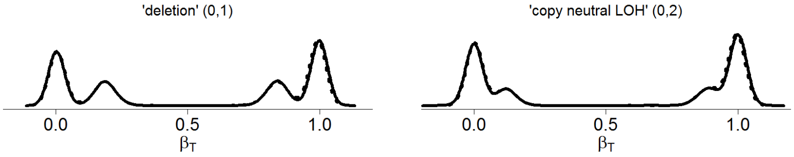

(a) raw,NGC.

(b) TBN,NGC,NGC.

(c) TCN,NGC.

Figure 22: Decrease in Heterozygosity (DH) and total copy numbers for region TCGA-23-1027:Chr10@106.5-113.5,cp=110+/-0.5,s=2/3. Only heterozygous SNPs are plotted. There are 241 loci of state 'deletion' (0,1) ("negatives") and 241 loci of state 'copy neutral LOH' (0,2) ("positives"), where the latter are highlighted with a solid bar beneath. In total 72 loci within the safety margin were excluded.

6.2 Allele B fraction density plots



(a) 'deletion' (0,1).

(b) 'copy neutral LOH' (0,2).

Figure 23: Density of raw (dashed lines) and TumorBoost-normalized (solid lines) allele B fractions for region TCGA-23-1027:Chr10@106.5-113.5,cp=110+/-0.5,s=2/3.

6.3 ROC curves

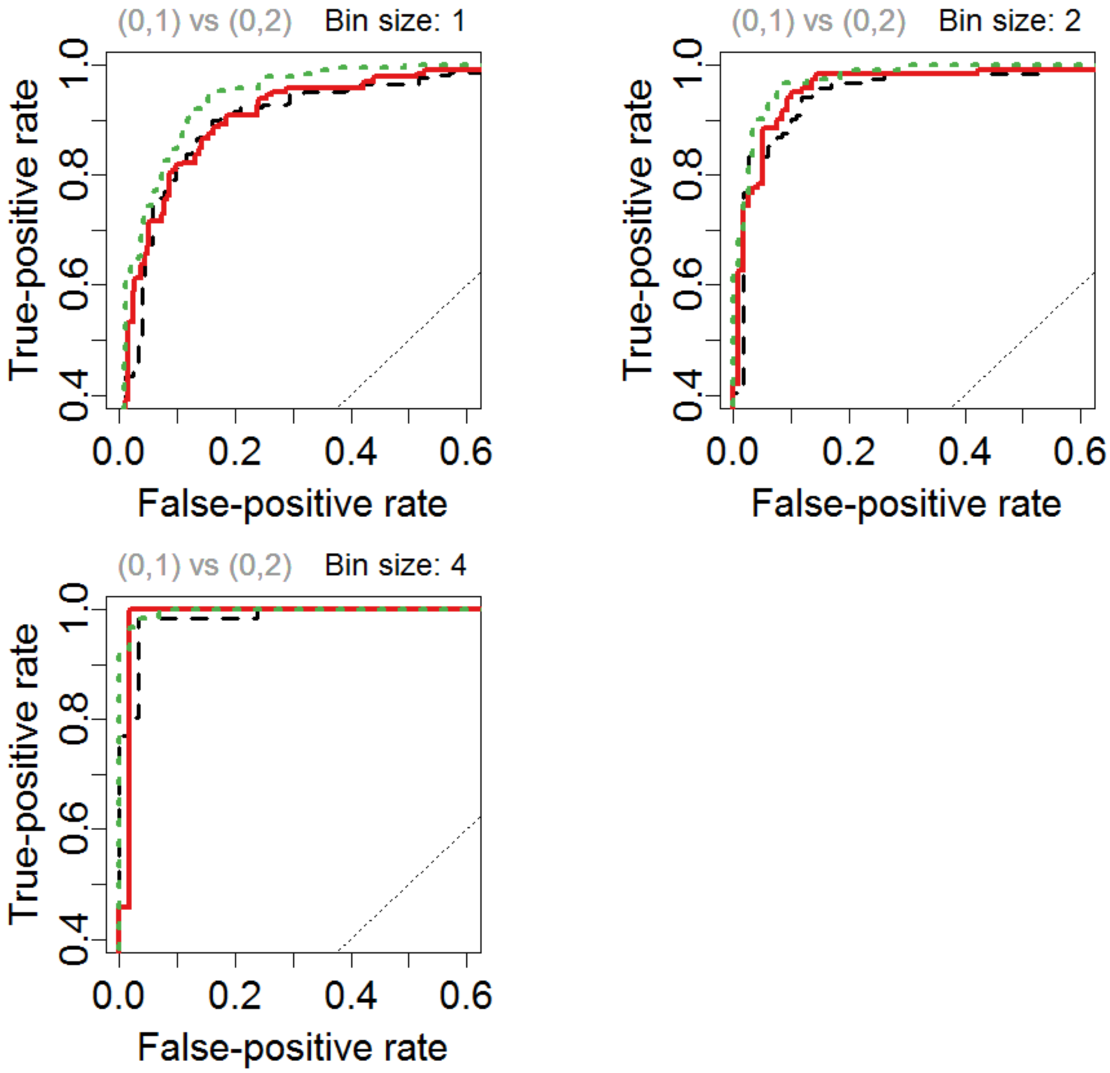


Figure 24: ROC curves for each preprocessing method at the full resolution as well as 2 different amounts of smoothing (using the mean() function) for region TCGA-23-1027:Chr10@106.5-113.5,cp=110+/-0.5,s=2/3. Legend: raw,NGC (dashed; #000000), TBN,NGC,NGC (solid; #E41A1C) and TCN,NGC (dotted; #4DAF4A).

6.4 (β_N, β_T) plots

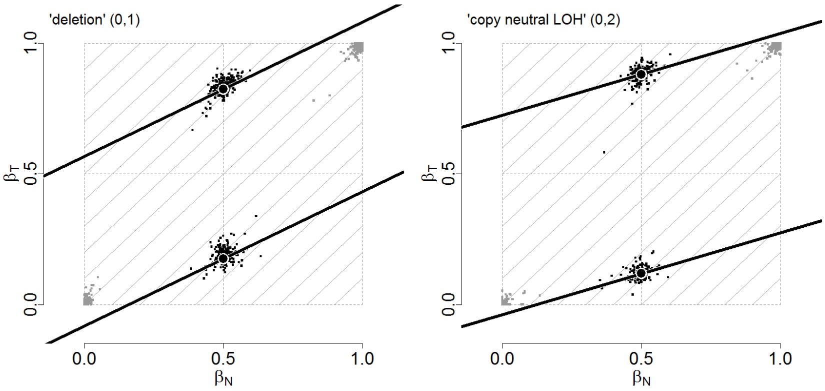


Figure 25: raw,NGC for region TCGA-23-1027:Chr10@106.5-113.5,cp=110+/-0.5,s=2/3.

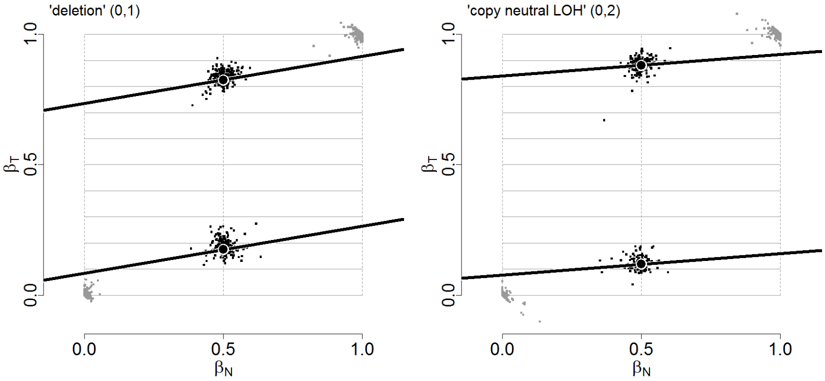


Figure 26: TBN,NGC,NGC for region TCGA-23-1027:Chr10@106.5-113.5,cp=110+/-0.5,s=2/3.

6.5 Allele-specific copy number estimates

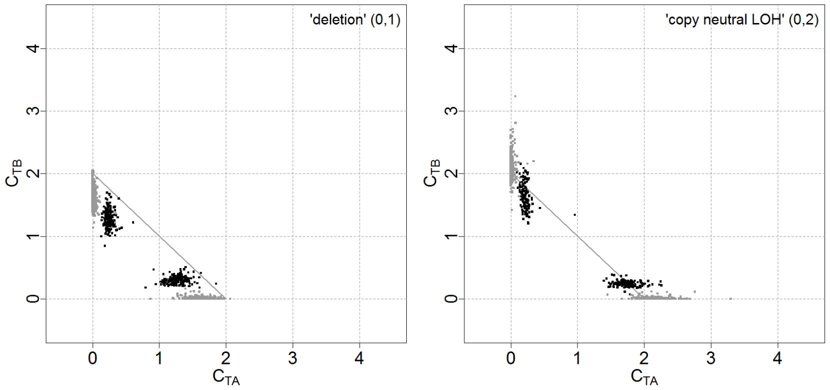


Figure 27: raw,NGC for region TCGA-23-1027:Chr10@106.5-113.5,cp=110+/-0.5,s=2/3.

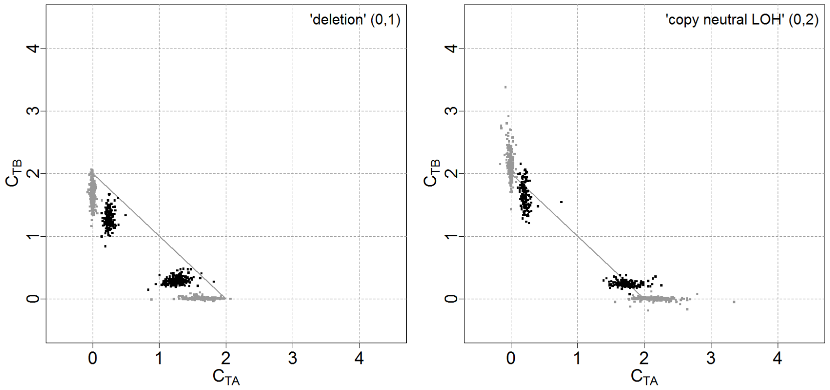
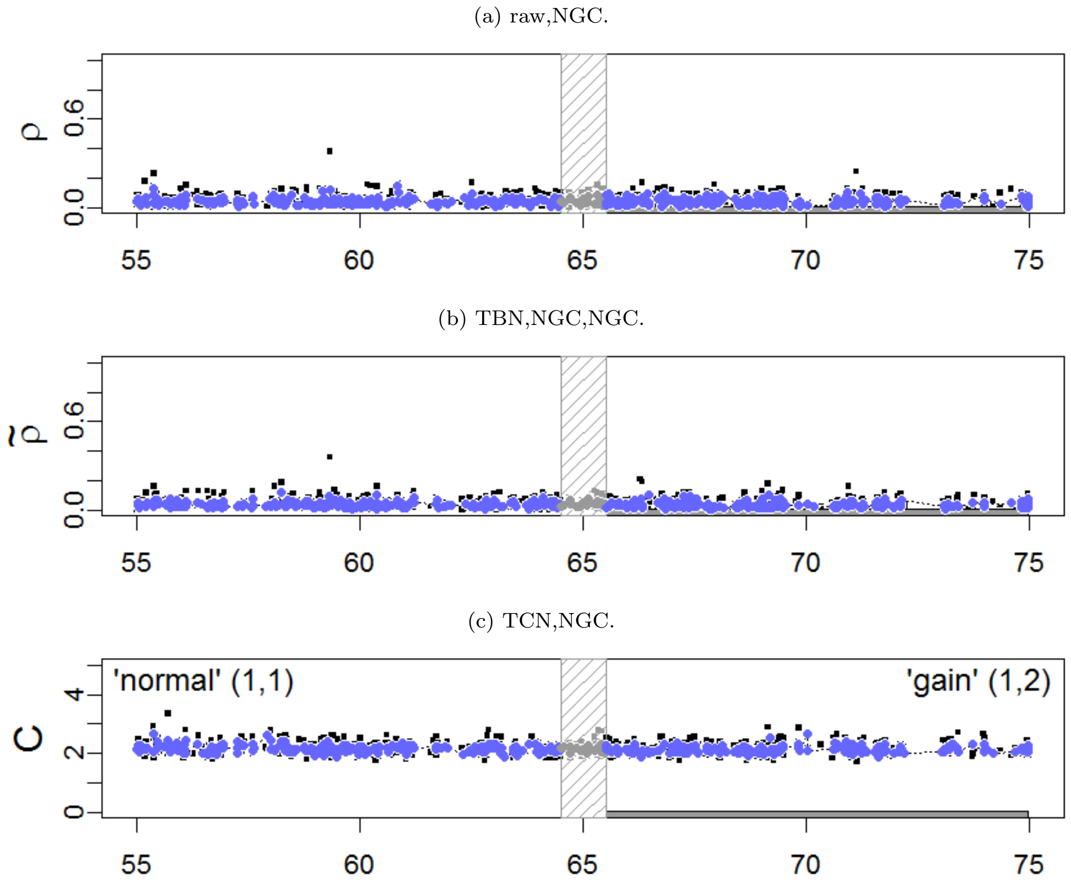


Figure 28: TBN,NGC,NGC for region TCGA-23-1027:Chr10@106.5-113.5,cp=110+/-0.5,s=2/3.

7 Region: TCGA-23-1027:Chr2@55-75.0,cp=65.0+/-0.5,s=0/1

7.1 Decrease in Heterozygosity (DH) and total copy-number tracks

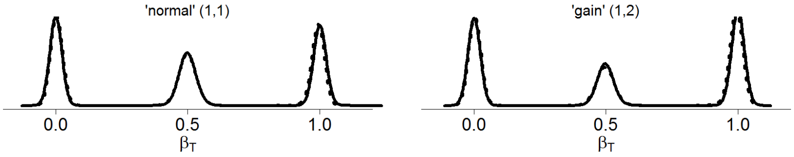

(a) raw,NGC.

(b) TBN,NGC,NGC.

(c) TCN,NGC.

Figure 29: Decrease in Heterozygosity (DH) and total copy numbers for region TCGA-23-1027:Chr2@55-75.0,cp=65.0+/-0.5,s=0/1. Only heterozygous SNPs are plotted. There are 868 loci of state 'normal' (1,1) ("negatives") and 868 loci of state 'gain' (1,2) ("positives"), where the latter are highlighted with a solid bar beneath. In total 82 loci within the safety margin were excluded.

7.2 Allele B fraction density plots



(a) 'normal' (1,1).

(b) 'gain' (1,2).

Figure 30: Density of raw (dashed lines) and TumorBoost-normalized (solid lines) allele B fractions for region TCGA-23-1027:Chr2@55-75.0,cp=65.0+/-0.5,s=0/1.

7.3 ROC curves

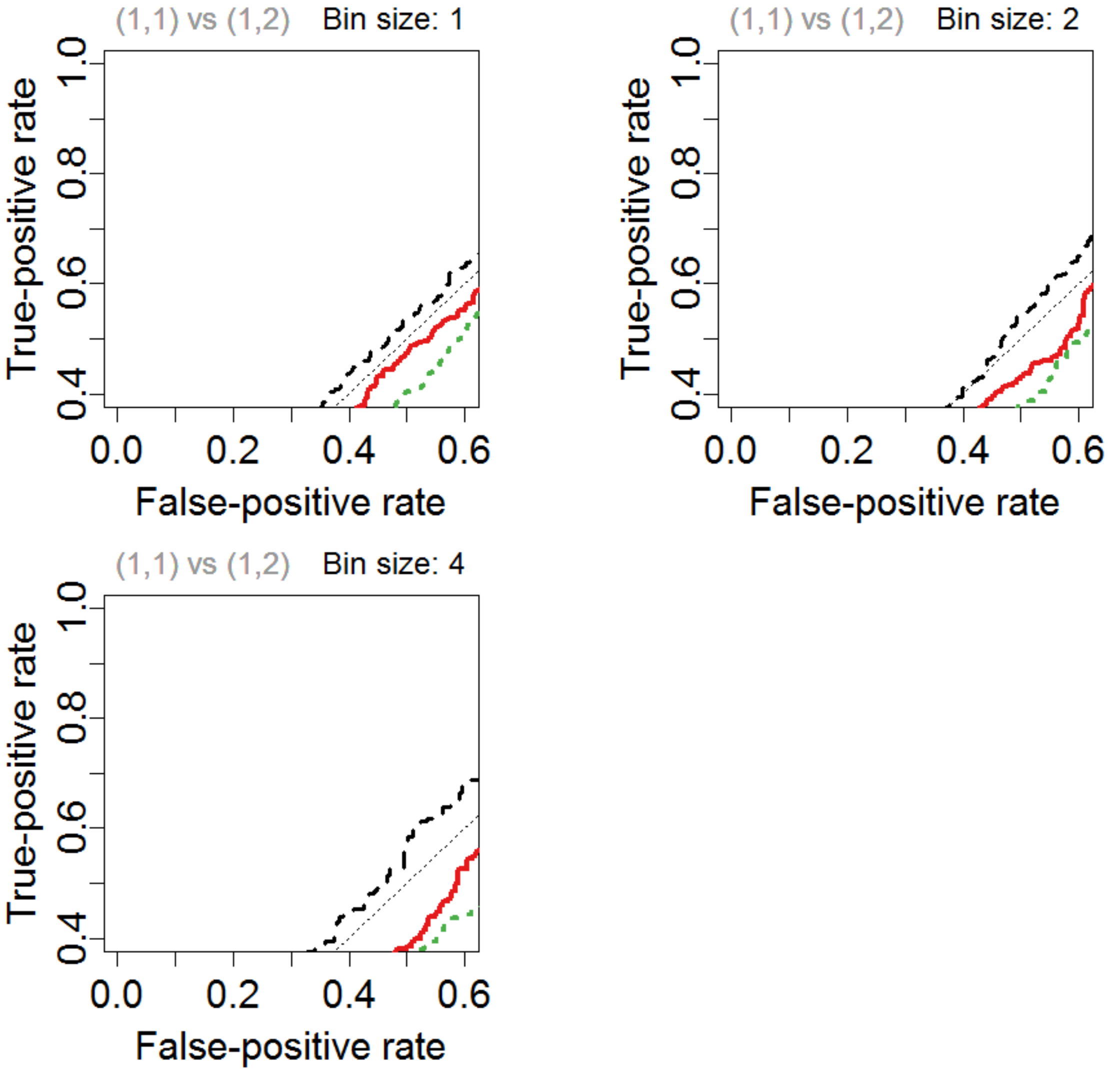


Figure 31: ROC curves for each preprocessing method at the full resolution as well as 2 different amounts of smoothing (using the mean() function) for region TCGA-23-1027:Chr2@55-75.0,cp=65.0+/-0.5,s=0/1. Legend: raw,NGC (dashed; #000000), TBN,NGC,NGC (solid; #E41A1C) and TCN,NGC (dotted; #4DAF4A).

7.4 (β_N, β_T) plots

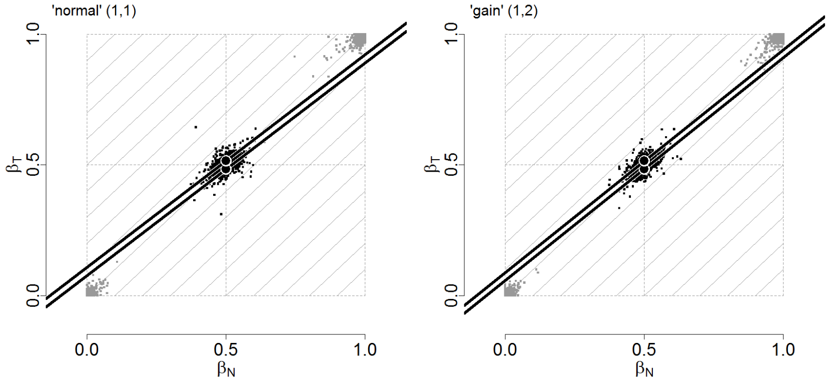


Figure 32: raw,NGC for region TCGA-23-1027:Chr2@55-75.0,cp=65.0+/-0.5,s=0/1.

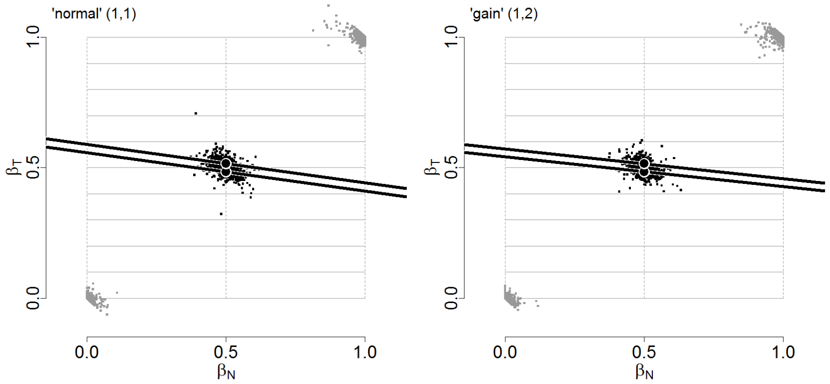


Figure 33: TBN,NGC,NGC for region TCGA-23-1027:Chr2@55-75.0,cp=65.0+/-0.5,s=0/1.

7.5 Allele-specific copy number estimates

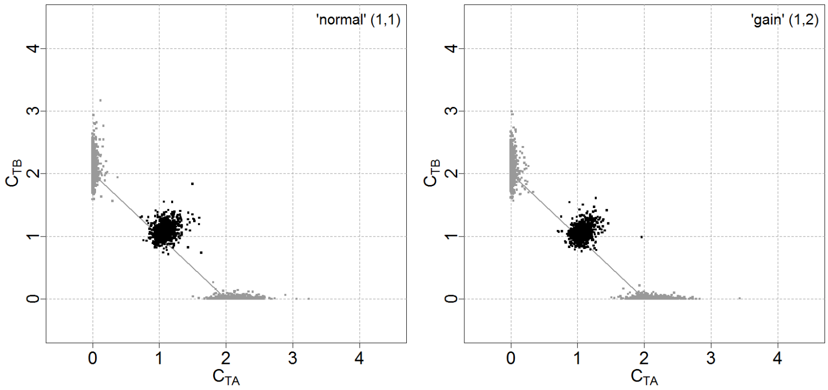


Figure 34: raw,NGC for region TCGA-23-1027:Chr2@55-75.0,cp=65.0+/-0.5,s=0/1.

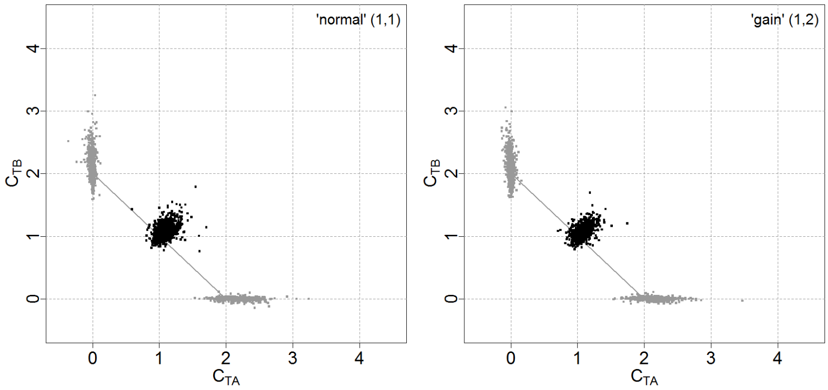


Figure 35: TBN,NGC,NGC for region TCGA-23-1027:Chr2@55-75.0,cp=65.0+/-0.5,s=0/1.

8 Bootstrap estimates of test statistics for all regions

	0/1	1/3	0/2	2/3	0/1
raw,NGC	41.428±3.084	90.583±6.583	111.277±6.927	21.343±1.421	1.029±0.678
TBN,NGC,NGC	47.045±3.772	95.000±8.958	117.973±7.758	22.082±1.259	1.089±0.766
TCN,NGC	16.040±1.121	21.895±1.293	39.165±1.817	23.926±1.275	1.810±1.000

Table 2: Student test statistics of the null hypothesis of equal mean before and after each PCN change point (heterozygous SNPs): raw or TumorBoost-normalized DH, and total copy number (last line). Mean ± standard deviation across 100 samplings of 225 points (for each PCN state) from the original data set. The larger value, the more different the true means are.

References

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J., Cheung, S. W., Shen, R. M., Barker, D. L., and Gunderson, K. L. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, **16**(9), 1136–1148.

A Data files

A.1 Total copy numbers

```
$‘TCGA,OV,BeadStudio,BAF‘
AromaUnitTotalCnBinarySet:
Name: TCGA
Tags: OV,BeadStudio,BAF
Full name: TCGA,OV,BeadStudio,BAF
Number of files: 1
Names: TCGA-23-1027
Path (to the first file): rawCnData/TCGA,OV,BeadStudio,BAF/Human1M-Duo
Total file size: 4.58 MB
RAM: 0.00MB
```

A.2 Allele B fractions

```
$raw
AromaUnitFracBCnBinarySet:
Name: TCGA
Tags: OV,BeadStudio,BAF
Full name: TCGA,OV,BeadStudio,BAF
Number of files: 1
Names: TCGA-23-1027
Path (to the first file): totalAndFracBData/TCGA,OV,BeadStudio,BAF/Human1M-Duo
Total file size: 4.58 MB
RAM: 0.00MB

$‘TBN,NGC‘
AromaUnitFracBCnBinarySet:
Name: TCGA
Tags: OV,BeadStudio,BAF,TBN,NGC
Full name: TCGA,OV,BeadStudio,BAF,TBN,NGC
Number of files: 1
Names: TCGA-23-1027
Path (to the first file): totalAndFracBData/TCGA,OV,BeadStudio,BAF,TBN,NGC/Human1M-Duo
Total file size: 4.58 MB
RAM: 0.00MB
```

A.3 Genotype calls

```
$NGC
AromaUnitGenotypeCallSet:
Name: TCGA
Tags: OV,BeadStudio,BAF,NGC
Full name: TCGA,OV,BeadStudio,BAF,NGC
Number of files: 1
Names: TCGA-23-1027
Path (to the first file): callData/TCGA,OV,BeadStudio,BAF,NGC/Human1M-Duo
Total file size: 2.29 MB
RAM: 0.00MB
```

A.4 Genotype confidence scores

```
$NGC
AromaUnitSignalBinarySet:
```

```
Name: TCGA
Tags: OV,BeadStudio,BAF,NGC
Full name: TCGA,OV,BeadStudio,BAF,NGC
Number of files: 1
Names: TCGA-23-1027
Path (to the first file): callData/TCGA,OV,BeadStudio,BAF,NGC/Human1M-Duo
Total file size: 4.57 MB
RAM: 0.00MB
```

B Session information

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

- R version 2.10.0 Patched (2009-11-21 r50532), `i386-pc-mingw32`
- Locale: `LC_COLLATE=English_United States.1252`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: aroma.cn.eval 0.1.1, aroma.core 1.3.5, aroma.light 1.15.1, digest 0.4.1, MASS 7.3-3, matrixStats 0.1.8, R.cache 0.2.0, R.filesets 0.6.5, R.menu 0.0.5, R.methodsS3 1.1.0, R.oo 1.6.6, R.rsp 0.3.6, R.utils 1.2.4, RColorBrewer 1.0-2, xtable 1.5-5
- Loaded via a namespace (and not attached): affxparser 1.18.0