Supplemental Data for

CAFE: an R package for the detection of gross chromosomal abnormalities from gene expression microarray data

Sander Bollen[1,2,*], Mathias Leddin[3], Miguel A. Andrade-Navarro[1] and Nancy Mah[1]

[1]Computational Biology and Data Mining Group, Max Delbruck Center for Molecular Medicine, 13125 Berlin, Germany. [2]Graduate School of Life Sciences, Utrecht University, Universiteitsweg 98, 3584 CG, Utrecht, the Netherlands. [3]Roche Diagnostics GmbH, 82377 Penzberg, Germany.

Table of Contents

1 Test datasets and summary of results

Table S1. Test datasets for detection of chromosomal aberrations

Dataset	GEO Accession	Sample ID	Number of Samples	Cell Type	Chromosome Status
1	GSE15148	GSM378882-GSM378835, GSM378837-38	15	iPS cells from episomal vectors	normal
	GSE15148	GSM378836	1	iPS cells from episomal vectors	partial deletion on Chr 15
2	GSE10809	GSM272914-21	10	HESC expressing Sox7/17	normal
	GSE6561	GSM151739, 41	2	HESC H14, passage 3, 6	normal
	GSE6561	GSM151738, 40	2	HESC H14, passage 29, 25	48,XY,+12, +der(17)del(17)(p12p13.3)hsr(17)(p11.2)

Table S2. Summary of results from CAFE and other packages

Dataset	Chromosomal Aberration	CAFE	aCGH	snapCGH	MACAT
1	partial deletion on chr 15	detected	detected	detected	detected
2	+12	detected	detected	detected	detected*
	+der(17)	detected	detected	detected by all 4 models in early passage sample; detected by 2 models in late passage sample	detected*
	hsr(17)(p11.2)	detected	only detected in early passage sample	detected by all 4 models in both samples	detected
	del(17)(p12p13.3)	visible in discontinuous plot but p-values not significant at p_{sig} < 1e-04	not detected	detected by 1 model in the late passage sample	partly detected

*MACAT package is not designed to detect whole chromosome duplications, but this information can be inferred from the plot

2 Results from CAFE and three different Bioconductor packages

2.1 Dataset 1: Detection of partial deletion on chromosome 15

2.1.1 Result from CAFE

The function bandStats() detects an enrichment of under-expressed genes at 15q15.1 ($p_{corrected}$ = 1.006399e-09), which is visually confirmed by the discontinuous plot (Figure S1). Note that there are no probesets located on 15p because the chromosome is acrocentric.

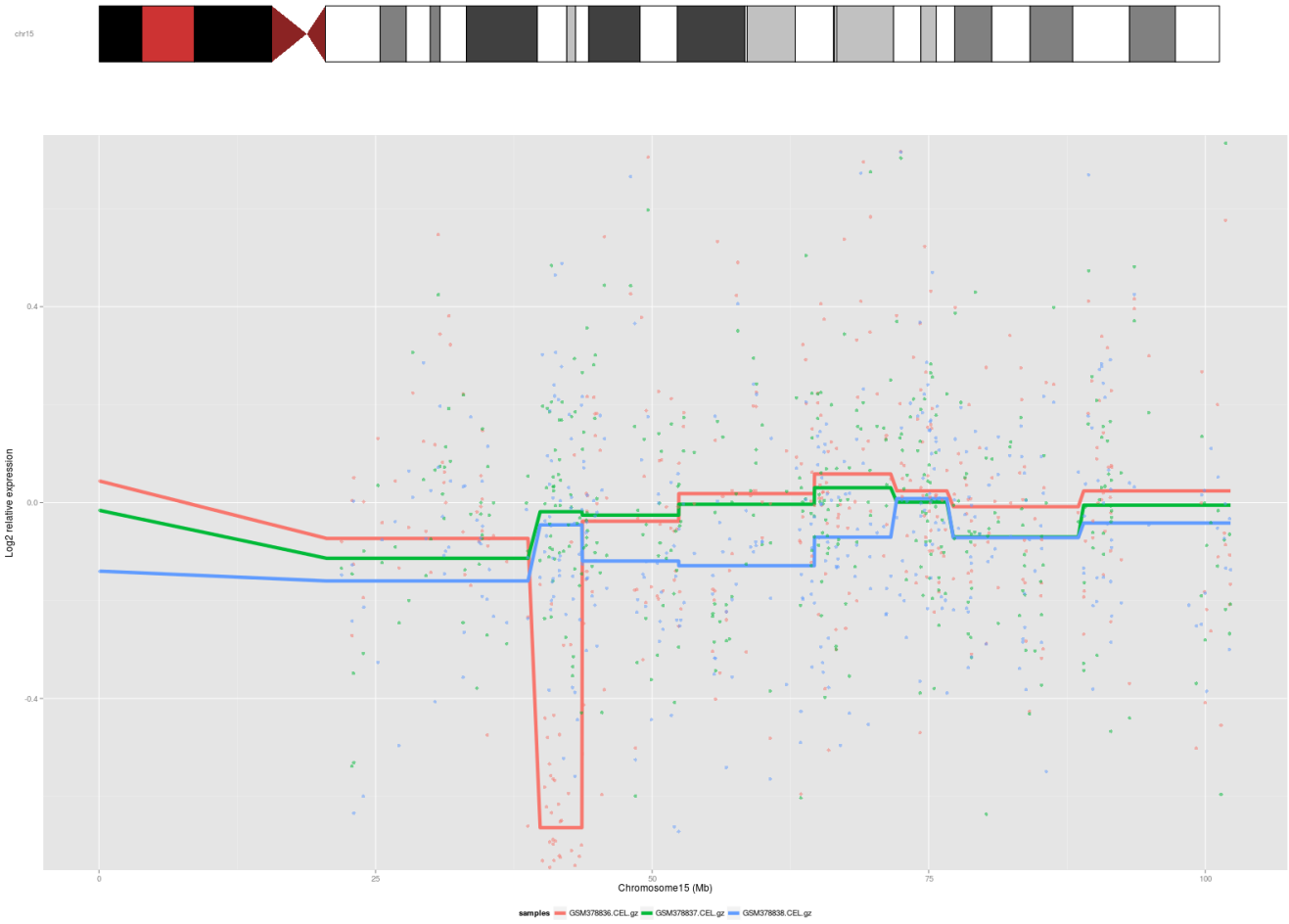


Figure S1. Discontinuous plot of two normal samples (GSM378837 and GSM378838; green and blue, respectively) and the sample containing the 15q deletion (GSM378836, orange). Log2 ratios for each probeset are plotted in the background as dots.

2.1.2 Result from aCGH

This Bioconductor package is designed for the analysis of array-based comparative genomic hybridization. The output of ProcessCels() from CAFE, was re-formatted for input into an aCGH object. aCGH analysis was run with default paramters. aCGH fits an unsupervised hidden Markov model to each chromosome for a varying number of states. The HMM states for the abnormal sample (GSM378836) and one of the normal controls (GSM378837) are shown in Figure S2 A and B, respectively. The deletion on chromosome 15 is clearly indicated by a state change on 15q (Figure S2A).

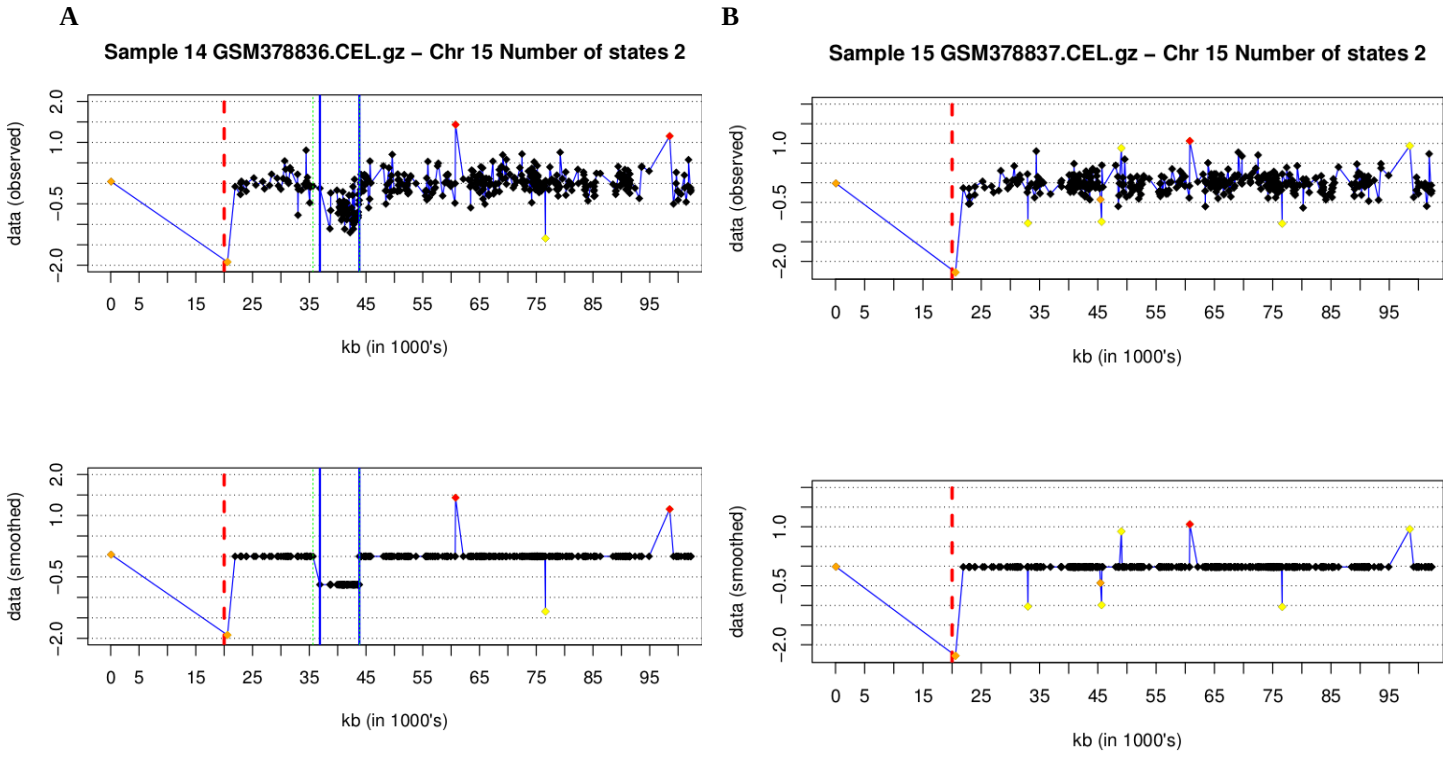


Figure S2. HMM states plot from Bioconductor package aCGH. A. Sample GSM378836 contains a deletion on chr 15. B. Sample GSM378837 has a normal karyotype. Top plots show the observed log2ratios and bottom plots show predicted values for all probesets, except for outliers, which show observed values. Blue line indicates the first probeset after transition and dotted green line indicates the last probeset after transition. Focal aberrations (orange), amplifications (red) and outliers (yellow) are indicated at circles. Dashed vertical line indicates the position of the centromere.

2.1.3 Result from snapCGH

This Bioconductor package is designed for the analysis of array-based comparative genomic hybridization. The output of ProcessCels() from CAFE, was re-formatted for input into a compatible object (MA List) for snapCGH. snapCGH provides a variety of segmentation methods, including homogeneous and heterogeneous HMMs. Unless stated otherwise in the plot legend, all methods were used with default parameters. All segmentation methods except the heterogeneous HMM method detect the deletion on 15q.

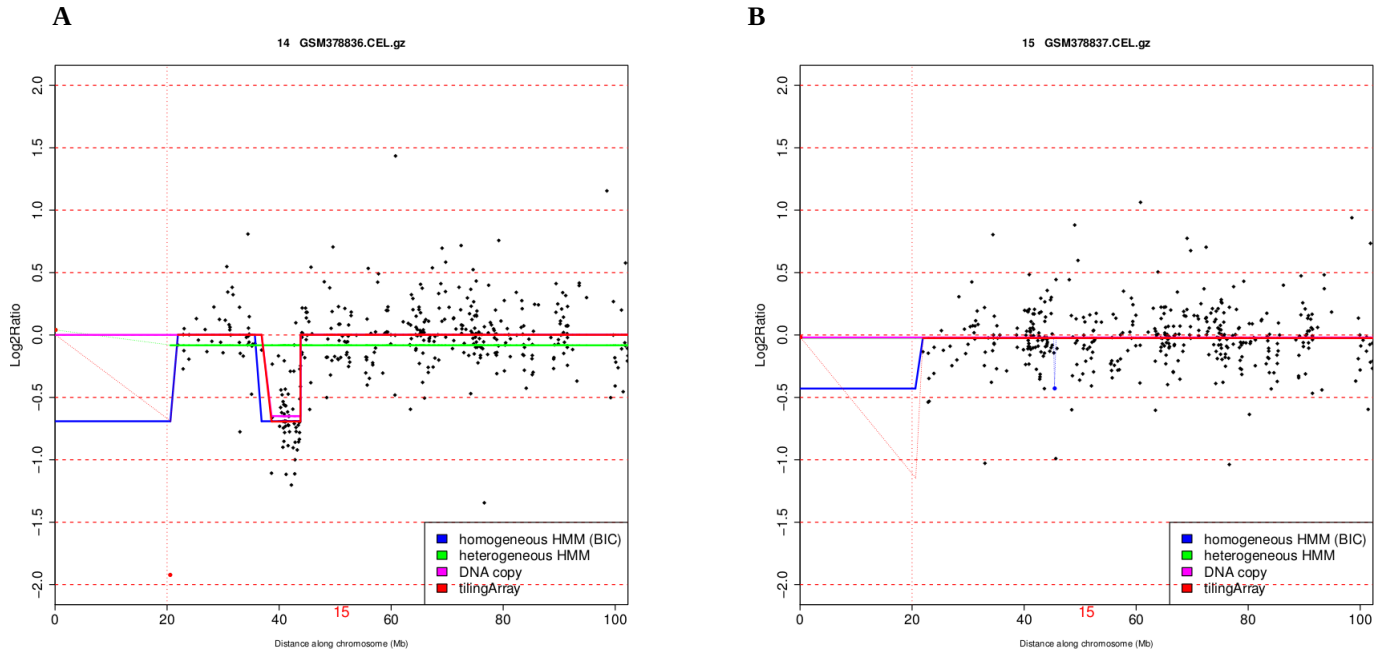


Figure S3. Plot of predicted states from Bioconductor package snapCGH. A. Sample GSM378836 contains a deletion on chr 15q. B. Sample GSM378837 has a normal karyotype. Dotted vertical line indicates centromere. Results from four different segmentation methods in snapCGH are shown. Observed log2 ratios are shown as black dots.

2.1.4 Result from MACAT

This Bioconductor package was designed to detect significantly differentially regulated regions on a chromosome, from expression microarray data. MACAT takes the normalized expression matrix (not log2 ratios) as input. Annotations were fetched by the MACAT functions. MACAT was run with default parameters. Differential expression between two classes of samples is determined by a modified t-statistic and permutation. The partial deletion on chromosome 15q is detected as a under-expression with the strongest signal deviating from quartile range. However, over-expressed regions also detected, which were not detected by any other method.

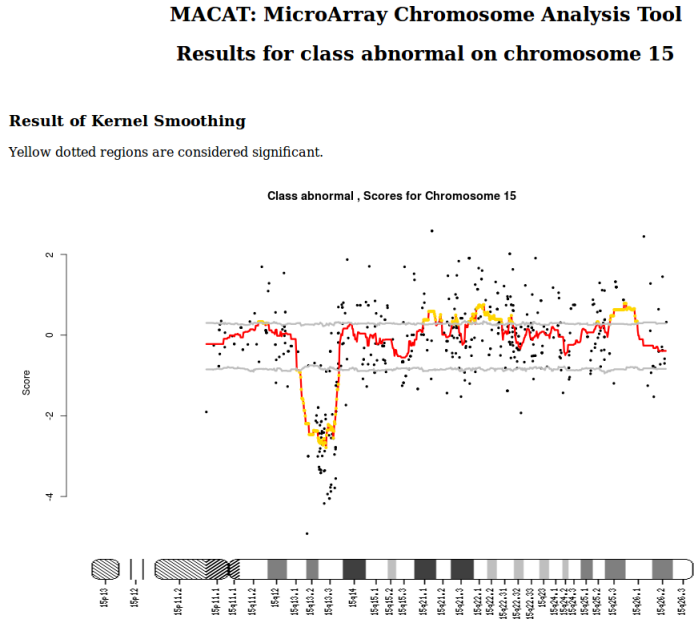


Figure S4. Scores for differential regulation between two classes. In this case, sample GSM378836, the only sample with a deletion on chr 15q, is the only sample in the "abnormal" class. Normalized expression values are shown in the background as black dots. Grey lines show the 0.025 and 0.975 quantiles of the permuted scores. The sliding average score is shown in red, except where the scores exceed the quantile boundaries (yellow).

2.2 Dataset 2: Detection of trisomy on chr 12 and derivative chr 17 in two samples

2.2.1 Result from CAFE

Trisomy on chr 12. The function chromosomeStats() detects enriched over-expression on chromosome 12 in two samples (GSM151740 (passage 25) and GSM151738 (passage 29); $p_{corrected}$ = 7.517695e-33 and $p_{corrected}$ = 1.070577e-10, respectively), indicating a trisomy on chromosome 12.

Derivative chromosome 17. There are a number of abnormalities on chromosome 17 according to the reported karyotype (+der(17)del(17)(p12p13.3)hsr(17)(p11.2)). At a significance level of p_{sig} < 1e-04, chromosome 17 was found to be enriched in over-expressed probesets in GSM151740 ($p_{corrected}$ = 1.351613e-45), but not in GSM151738 ($p_{corrected}$ = 1.243494e-02). This could be due to the fact that chromosome gains were lost in later passages. A section of the facetPlot (Figure S5) suggests that the detected chromosome gains are indeed trisomies on chromosome 12 and 17. At the band level, over-expression was detected at 17p11.2 in both samples (GSM151740 and GSM151738, $p_{corrected}$ = 5.614754e-11 and $p_{corrected}$ = 1.891353e-06, respectively), corresponding to hsr(17)(p11.2), which is clearly visible in the discontinuous plot (Figure S6). The known deletion at del(17)(p12p13.3) could not be detected in either sample at p_{sig} < 1e-04, but the discontinuous plot (Figure S6) hints that there could be a deletions in both GSM151740 and GSM151738, with the deletion more visible in the sample from the later passage (GSM151738).

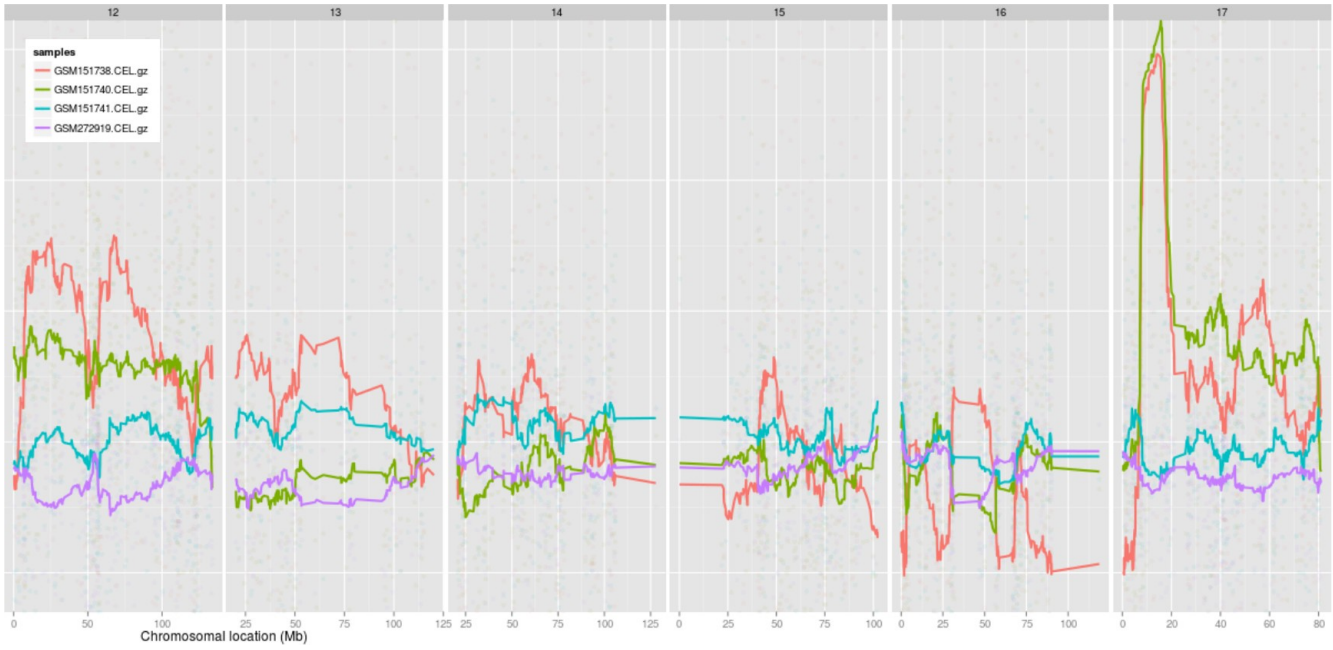


Figure S5. Section of the facet plot showing four samples on chromosomes 12 to 17. A sliding window is used to plot a smoothed line representing the log2 ratio (y-axis) along the chromosome coordinates (x-axis). Individual log2 ratios for each probeset are plotted in the background as dots. Samples GSM151738 and GSM151740 are aberrant (red and green, respectively), while GSM151741 and GSM272919 are normal (blue and purple, respectively).

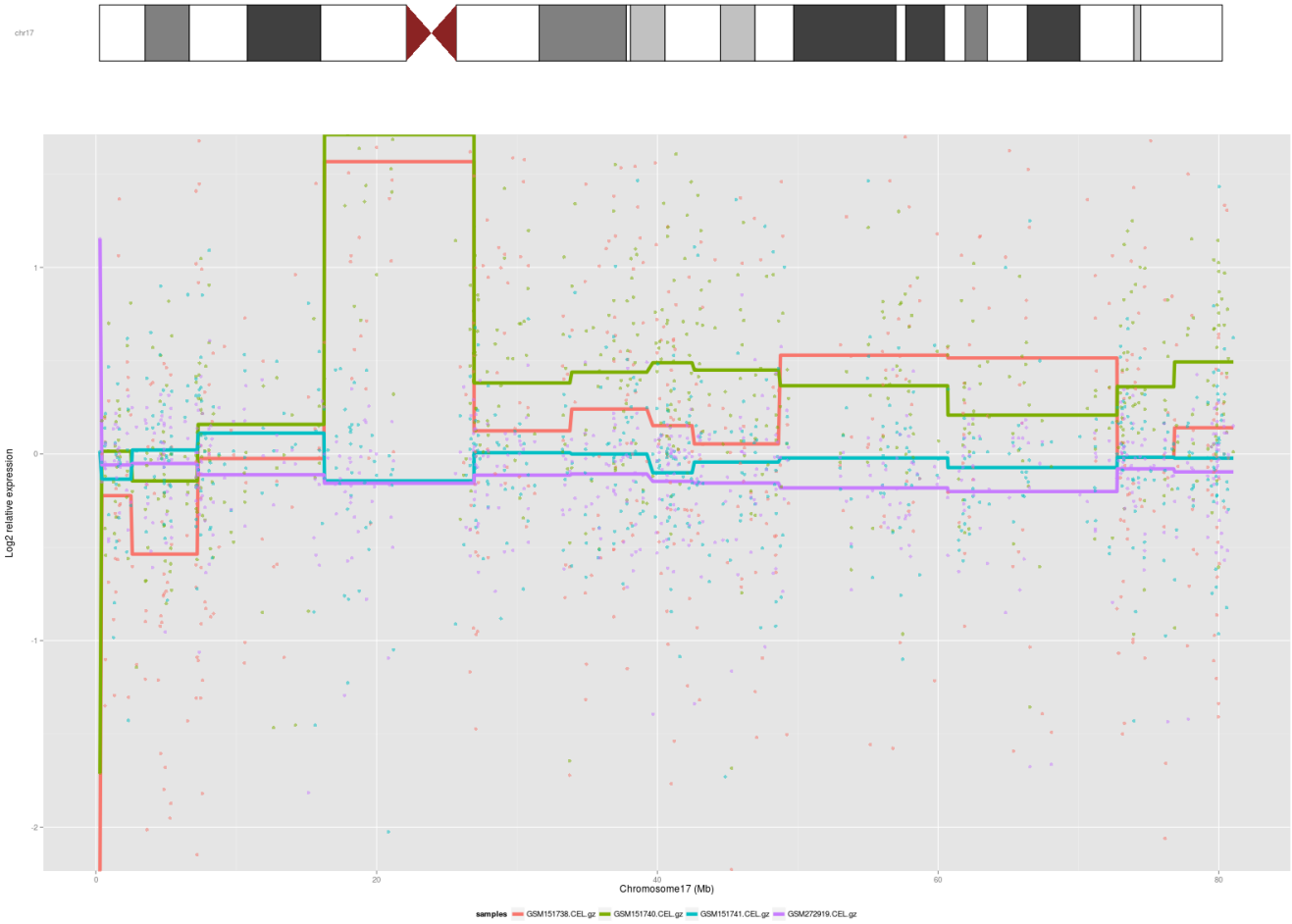
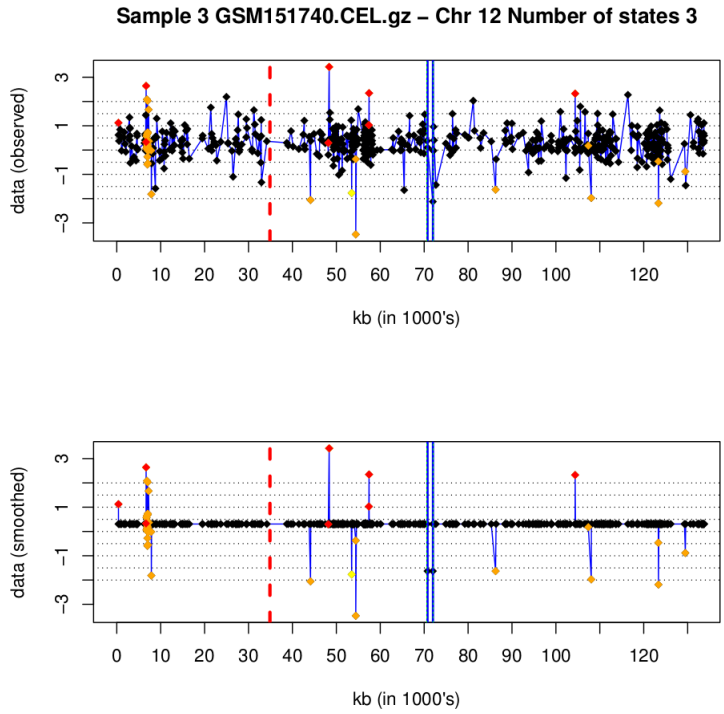


Figure S6. Discontinuous plot for chromosome 17. Individual log2 ratios for each probeset are plotted in the background as dots. Samples GSM151738 and GSM151740 are aberrant (red and green, respectively), while GSM151741 and GSM272919 are normal (blue and purple, respectively).

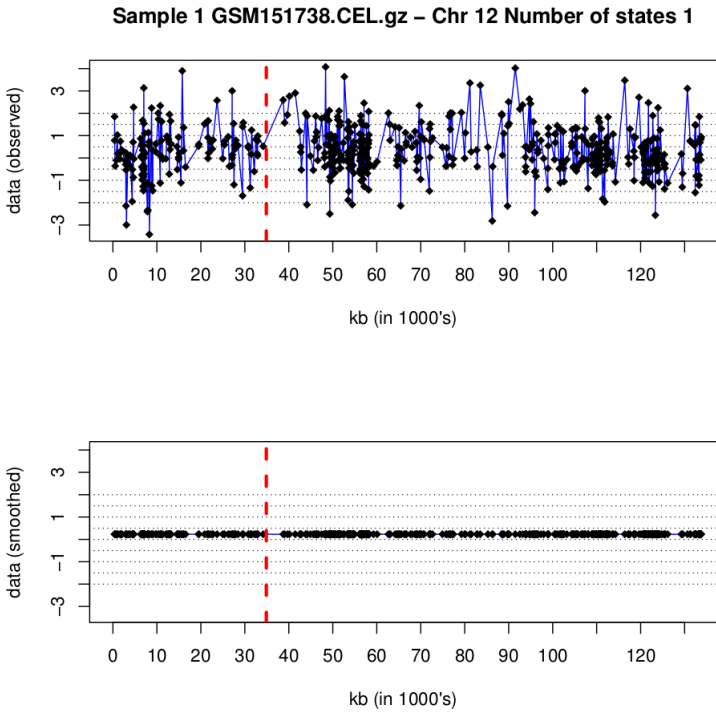
2.2.2 Result from aCGH

Trisomy on chr 12. This is clearly detected by in the plot of the HMM states, for both (GSM151740 (passage 25) and GSM151738 (passage 29), Figure S7 A and B, respectively. Log2 ratios for predicted (smoothed) data in the abnormal samples are located in a straight line at y~0.25 (Figure S7 A and B, respectively), indicating a higher copy number than normal controls. Controls (GSM151741, GSM272919, Figure S7 C and D respectively) have log2 ratios at y=0, indicating no change in copy number.

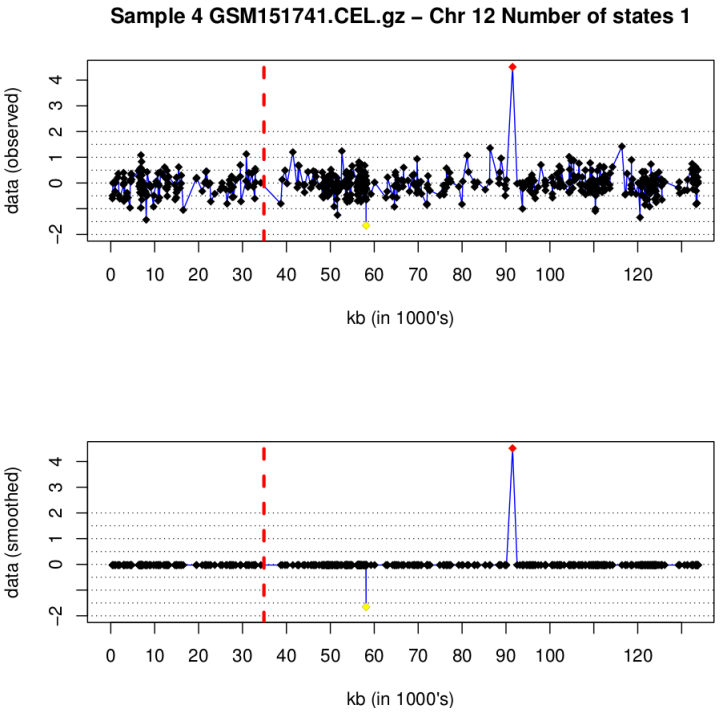
A



B



C



D

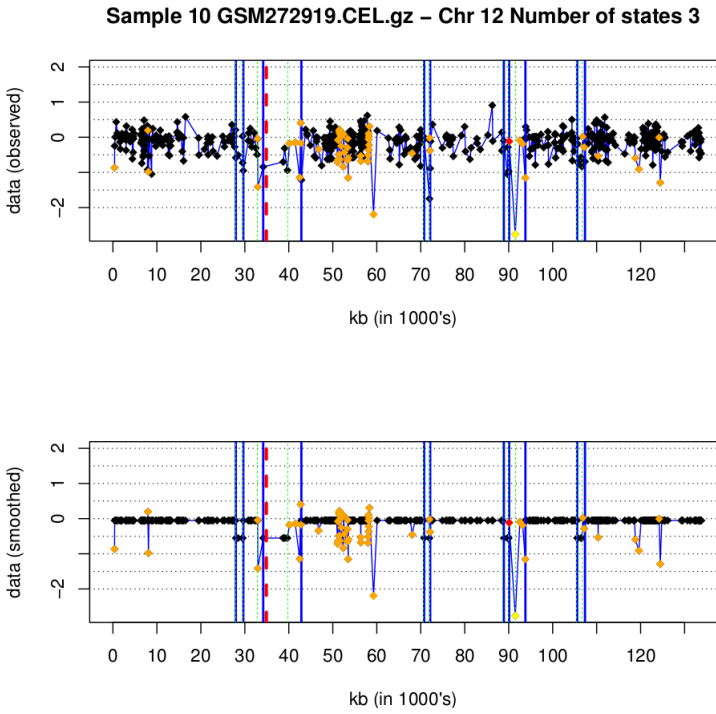


Figure S7. HMM states plot from Bioconductor package aCGH for chromosome 12. A, B. Samples GSM151740 and GSM151738 have the reported karyotype: 48,XY,+12, +der(17)del(17)(p12p13.3)hsr(17)(p11.2). C,D. Samples GSM151741 and GSM272919 are normal controls. Top plots show the observed log2ratios and bottom plots show predicted values for all probesets, except for outliers, which show observed values. Blue line indicates the first probeset after transition and dotted green line indicates the last probeset after transition. Focal aberrations (orange), amplifications(red) and outliers (yellow) are indicated at circles. Dashed vertical line indicates the position of the centromere.

Derivative chromosome 17. The extra chromosome 17 derivative is only detected in GSM151740, the sample taken at an earlier passage than GSM151738, as shown by the higher predicted values (y ~ 0.25) in GSM151740 for all of 17q (Figure S8A). The deletion at 17p12p13.3 is not detected at all. In contrast, no gross aberrations whatsoever are detected in GSM151738 (Figure S8B), which was taken at a later passage. Both controls (Figure S8 C,D) generally have predicted log2 ratios of zero, although there are apparent sporadic outliers.

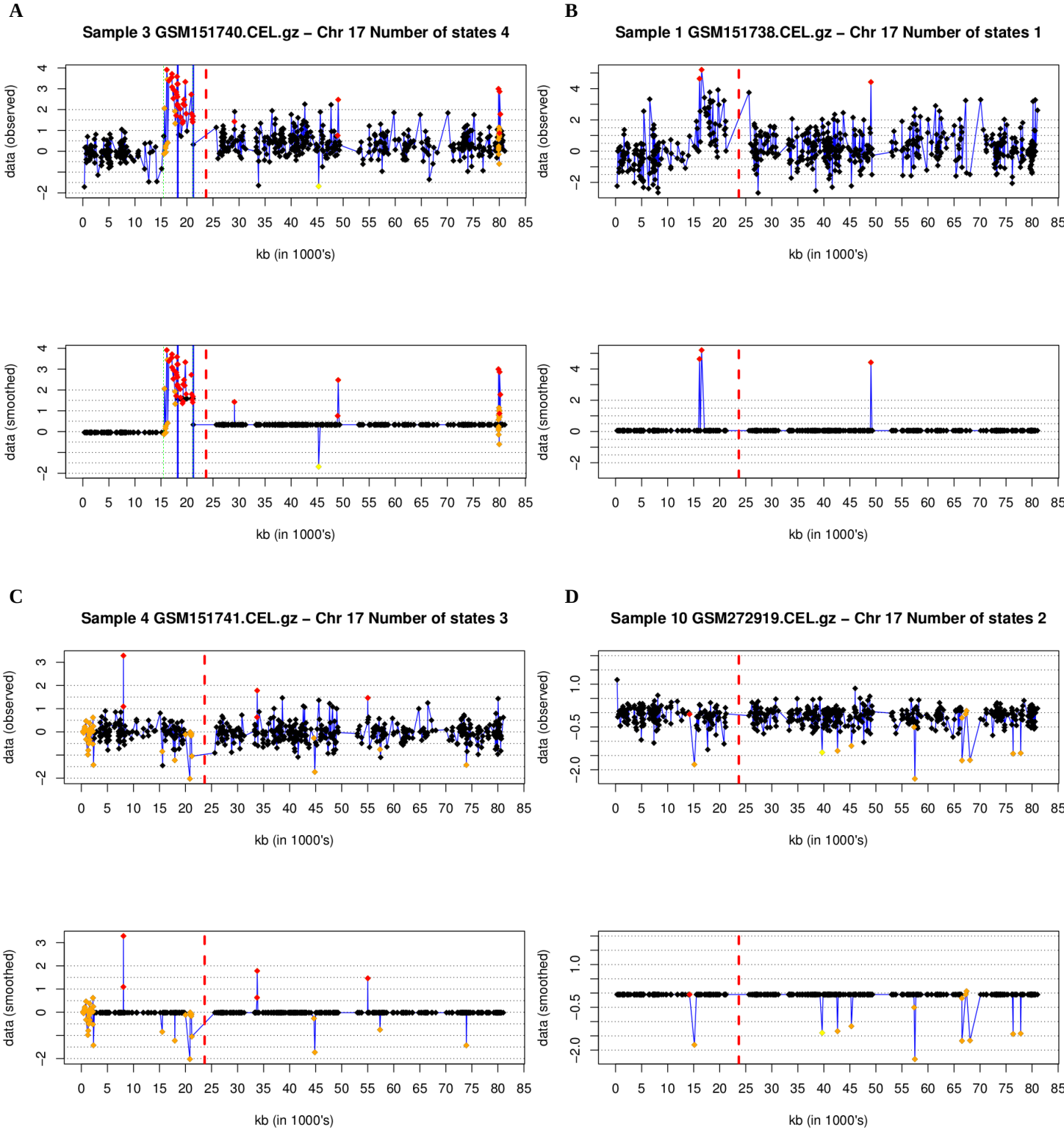


Figure S8. HMM states plot from Bioconductor package aCGH for chromosome 17. A, B. Samples GSM151740 and GSM151738 have the reported karyotype: 48,XY,+12, +der(17)del(17)(p12p13.3)hsr(17)(p11.2). C,D. Samples GSM151741 and GSM272919 are normal controls. Top plots show the observed log2ratios and bottom plots show predicted values for all probesets, except for outliers, which show observed values. Blue line indicates the first probeset after transition and dotted green line indicates the last probeset after transition. Focal aberrations (orange), amplifications(red) and outliers (yellow) are indicated at circles. Dashed vertical line indicates the position of the centromere.

2.2.3 Result from snapCGH

Trisomy on chr 12. There is a clear and constant increase in the log2 ratio over the entire chromosome for the aberrant samples (GSM151740 and GSM151738; Figure S9A, B), which indicates a chromosome gain. Conversely, samples with normal karyotype (GSM151741 and GSM272919; Figure S9 C,D) have a constant log2 ratio at or around zero, with some outliers.

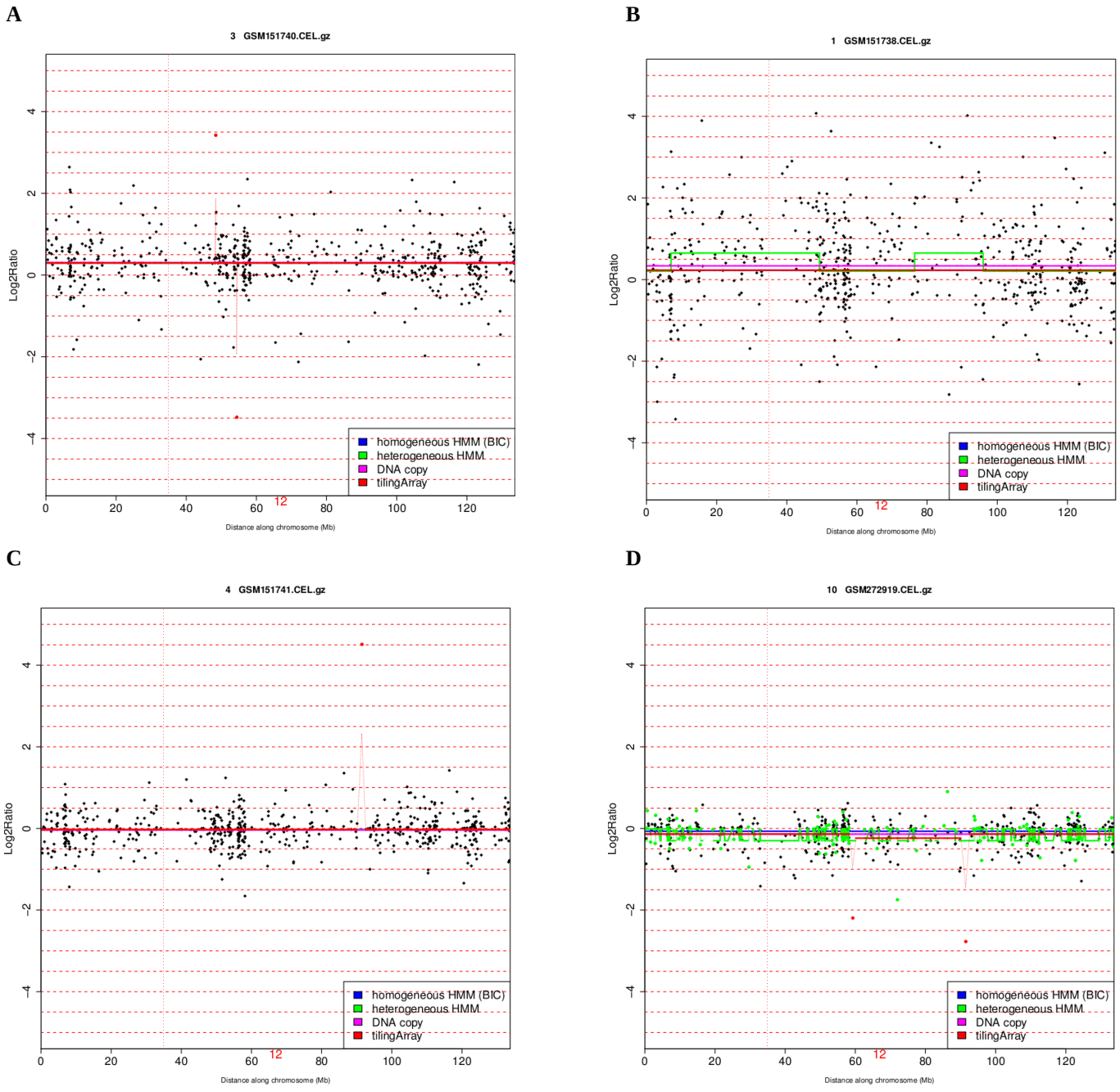


Figure S9. Plot of predicted states from Bioconductor package snapCGH for chromosome 12. A, B. Samples GSM151740 and GSM151738 have the reported karyotype: 48,XY,+12, +der(17)del(17)(p12p13.3)hsr(17)(p11.2). C,D. Samples GSM151741 and GSM272919 are normal controls. Dotted vertical line indicates centromere. Results from four different segmentation methods in snapCGH are shown. Observed log2 ratios are shown as black dots.

Derivative chromosome 17. The extra chromosome der(17) is present in the aberrant sample taken at an earlier passage (GSM1517140), as shown by all four models that predict increased log2 ratios (y~0.4) for 17q. The other aberrant sample (GSM1517138) is only predicted by one method (tiling array) to have increased copy number. The hsr region on 17p11.2 is predicted by all models on GSM1517140, but predicted only by three models on GSM1517138. The deletion on 17p12p13.3 is only detected on the later passage sample (GSM1517138). The two normal controls (GSM151741 and GSM272919) maintain more or less a constant log2 ratio about zero, indicating no detectable copy number changes.

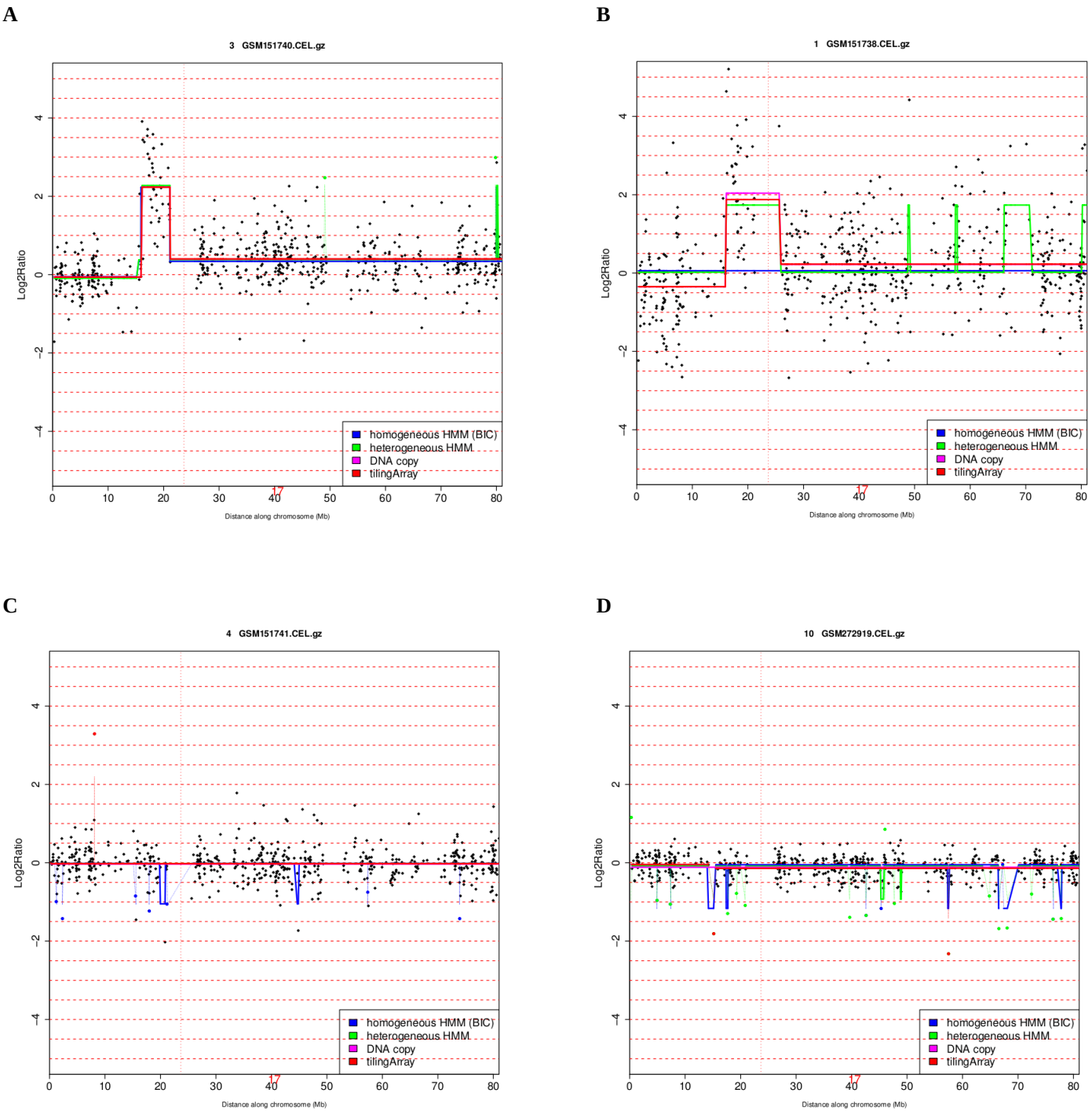


Figure S10. Plot of predicted states from Bioconductor package snapCGH for chromosome 17. A, B. Samples GSM151740 and GSM151738 have the reported karyotype: 48,XY,+12, +der(17)del(17)(p12p13.3)hsr(17)(p11.2). C,D. Samples GSM151741 and GSM272919 are normal controls. Dotted vertical line indicates centromere. Results from four different segmentation methods in snapCGH are shown. Observed log2 ratios are shown as black dots.

2.2.4 Result from MACAT

Trisomy on chr 12. MACAT is not specifically designed to detect gross aberrations at the whole chromosome level. However, if a horizontal line were drawn through the 0.5 quantile of the permuted scores, this line would lie above score=0 and therefore indicate a chromosome gain in GSM151740 and GSM151738 (Figure S11A,B). For comparison, the plot for chromosome 2 would have a 0.5 quantile at score=0, therefore indicating no copy number change. (Figure S11C).

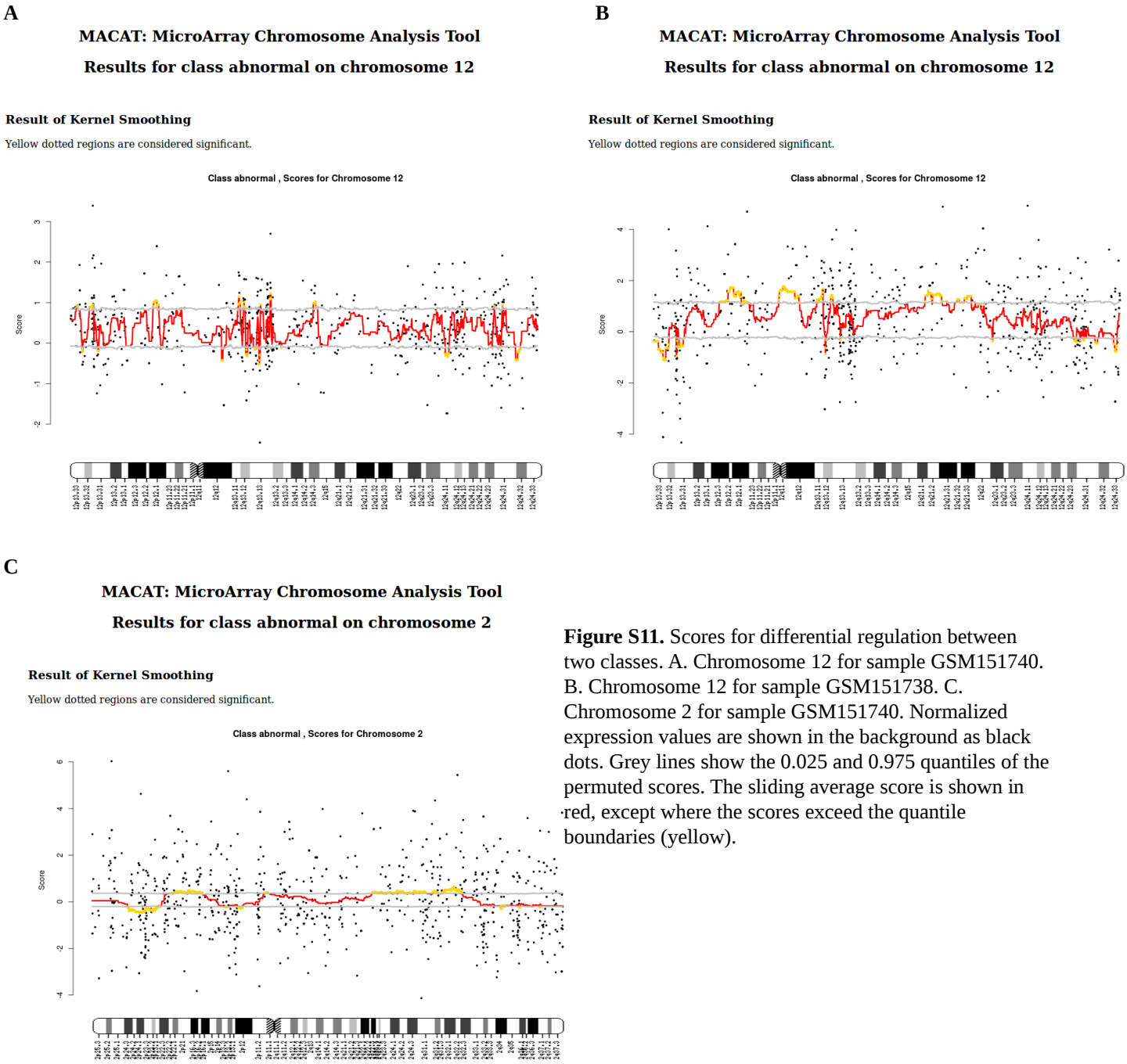


Figure S11. Scores for differential regulation between two classes. A. Chromosome 12 for sample GSM151740. B. Chromosome 12 for sample GSM151738. C. Chromosome 2 for sample GSM151740. Normalized expression values are shown in the background as black dots. Grey lines show the 0.025 and 0.975 quantiles of the permuted scores. The sliding average score is shown in red, except where the scores exceed the quantile boundaries (yellow).

Derivative chromosome 17. The chromosome gain can be inferred from the 0.5 quantile of the permuted scores, which would have a score of about 0.5, thereby indicating a chromosome gain in GSM151740. The whole chromosome gain in GSM151738 is not pronounced, since the 0.5 quantile of the permuted scores appears to lie just above score=0. The hsr region on 17p11.2 is clearly over-expressed in GSM151740 and to a lesser extent in GSM151378. Part of the deletion on 17p12p13.3 is suggested by significantly under-expressed genes in this region in both samples (Figure S12). These effects appear to be stronger in GSM151740 at the earlier passage.

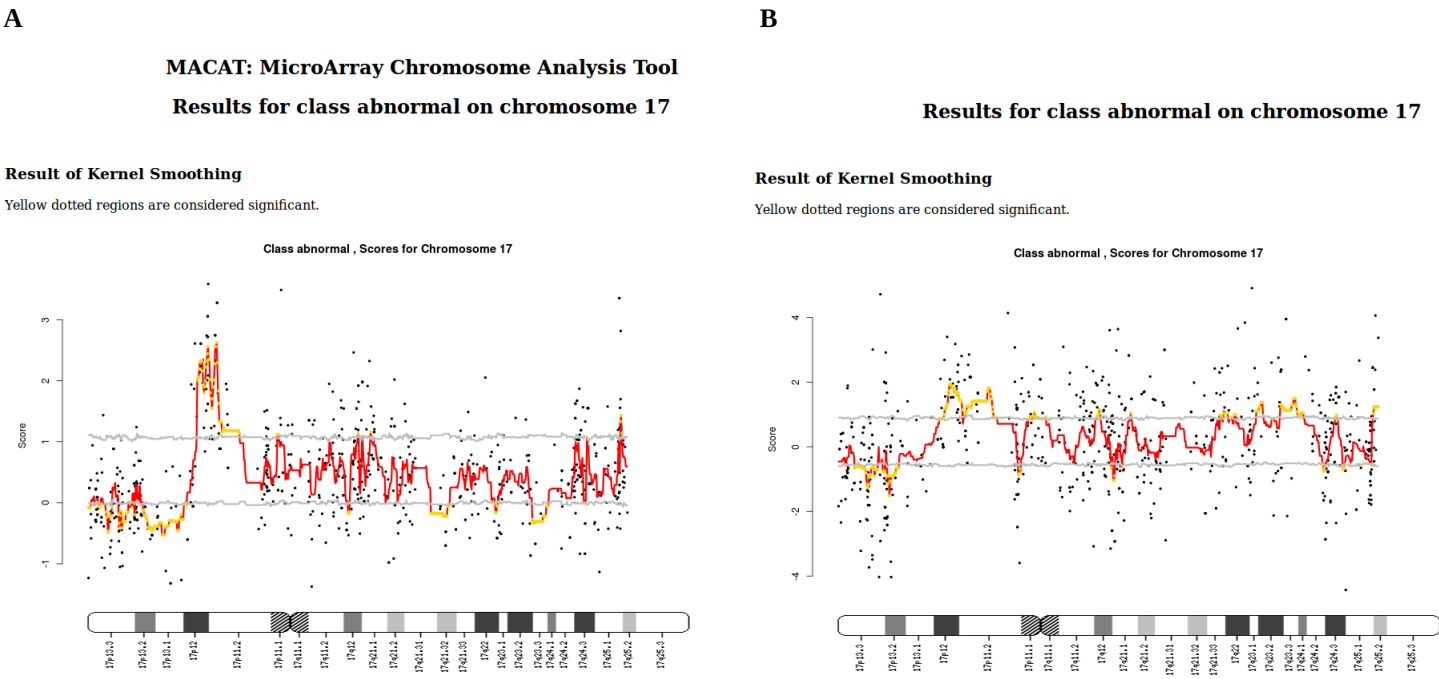


Figure S12. Scores for differential regulation between two classes. A. Sample GSM151740 (passage 25). B. Sample GSM151738 (passage 29). Normalized expression values are shown in the background as black dots. Grey lines show the 0.025 and 0.975 quantiles of the permuted scores. The sliding average score is shown in red, except where the scores exceed the quantile boundaries (yellow).