Supplementary information: Detection of differentially expressed segments in tiling array data

Christian Otto [1,2], Kristin Reiche [3,1,4], Jörg Hackermüller [3,1,4]*

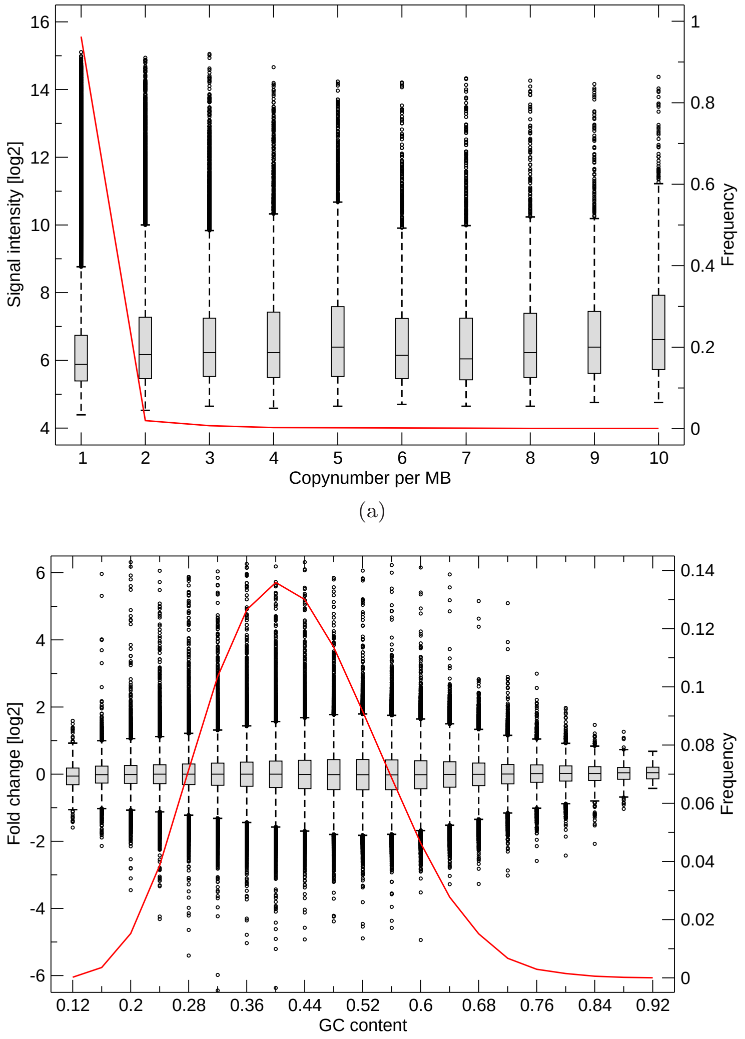
July 10, 2012

[1]Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, 04107 Leipzig, Germany
[2]LIFE Leipzig Research Center for Civilization Diseases, Universität Leipzig, Germany
[3]Young Investigators Group Bioinformatics and Transcriptomics, Department Proteomics, Helmholtz Centre for Environmental Research – UFZ, 04318 Leipzig, Germany
[4]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany

List of Figures

*to whom correspondence should be addressed

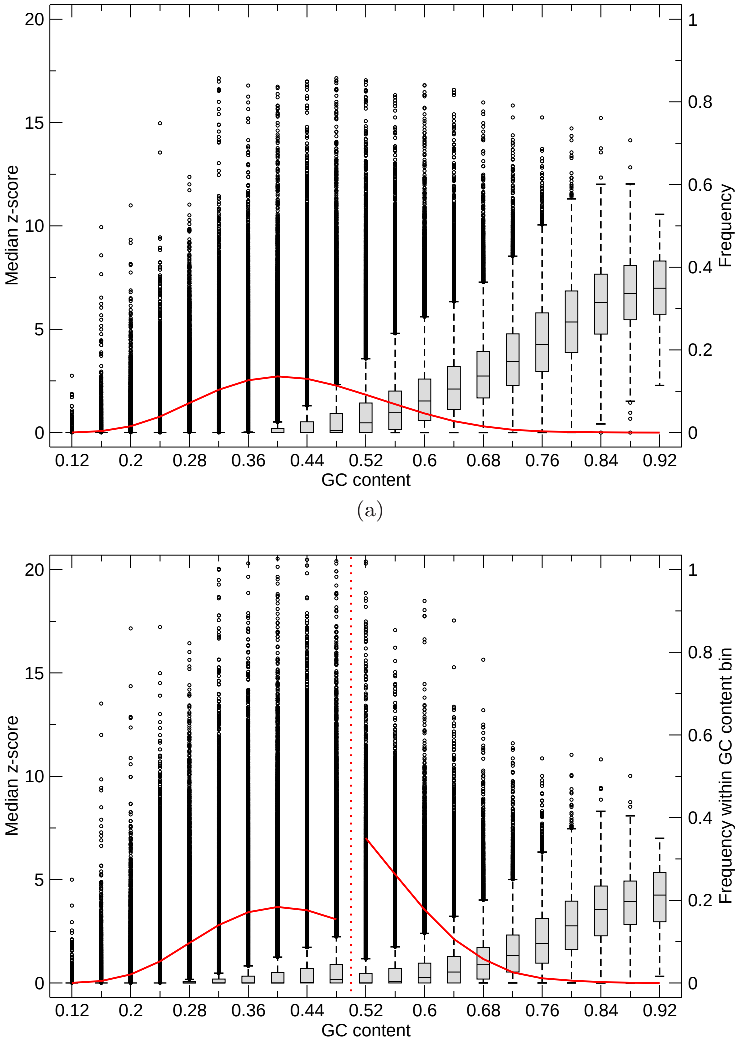
List of Tables



(a)

(b)

Figure S1: **Effect of copy number and GC-content on signal intensities and fold changes.** (a) Boxplot of probe signal intensities on a tiling array for different copy numbers, i.e., number of perfect matches of a probe sequence per megabase of genomic sequence. A comprehensive set of copy numbers for different tiling array designs is provided by `MAT` on their web page and taken here as reference. The relative frequency of each copy number is shown in the overlay graph (red solid line). (b) Boxplot of log-fold changes on a tiling array between two cellular states for different GC content in the probe sequence. The relative frequency of the GC content on the tiling array is shown in the overlay graph (red solid line).



(a)

(b)

Figure S2: **Boxplot of probe median z-scores according to one and two distinct GC content bins.** The probe median z-score is defined as the median over the z-scores of all windows enclosing the probe while z-scores are estimated by `TileShuffle` on a tiling array without GC content binning (a) and with two different GC content bins (b). Vertical dotted red lines display the boundaries of different bins while solid red lines indicate the relative frequency of the GC content in its corresponding bin.

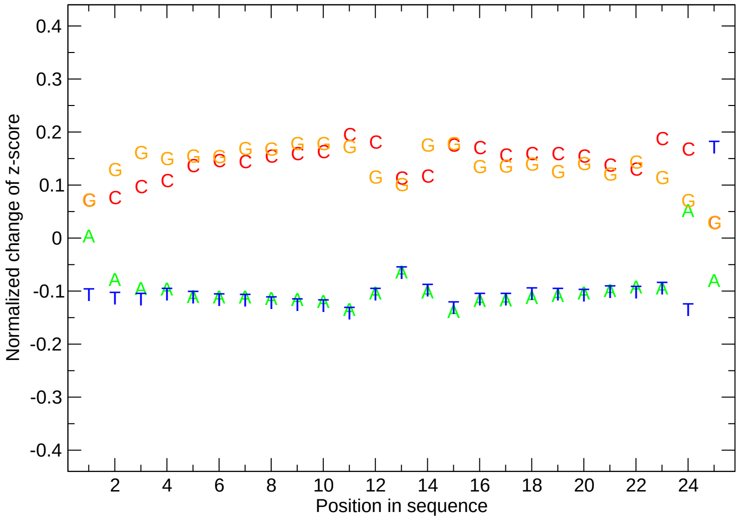


Figure S3: **Position-specific nucleotide bias using TileShuffle with one GC content bin.** The effect was calculated on probe median z-scores for every nucleotide in each of the 25 positions within the probe by use of the Starr R package. The probe median z-scores are further normalized by dividing them by the standard deviation of the intensity and median z-score distribution, respectively. The probe median z-score is calculated as the median over the z-scores of all windows enclosing the probe where z-scores were estimated by `TileShuffle` using one GC content bin.

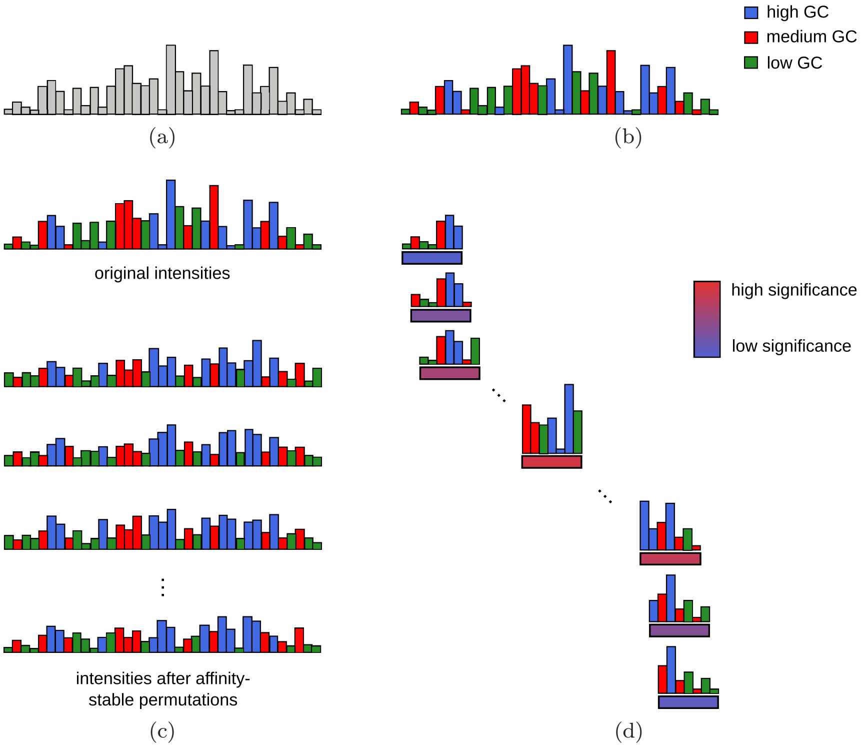


Figure S4: **Detailed outline of `TileShuffle` (detection of expressed segments).** (a) Signal intensities of probes in a short region. (b) Classification of probes into affinity bins according to the GC content of their sequences (high GC in blue, medium GC in red, low GC in green). (c) Original intensities and intensities after affinity-stable permutations where intensities of probes belonging to different affinity bins must not be interchanged. (d) Assignment of significances in terms of corrected empirical p-values to each window of given width. The empirical p-values are estimated by comparing the score of the original window to the scores of the permuted ones. In the end, windows with a corrected empirical p-value (q-value) below a given threshold are merged and then reported.

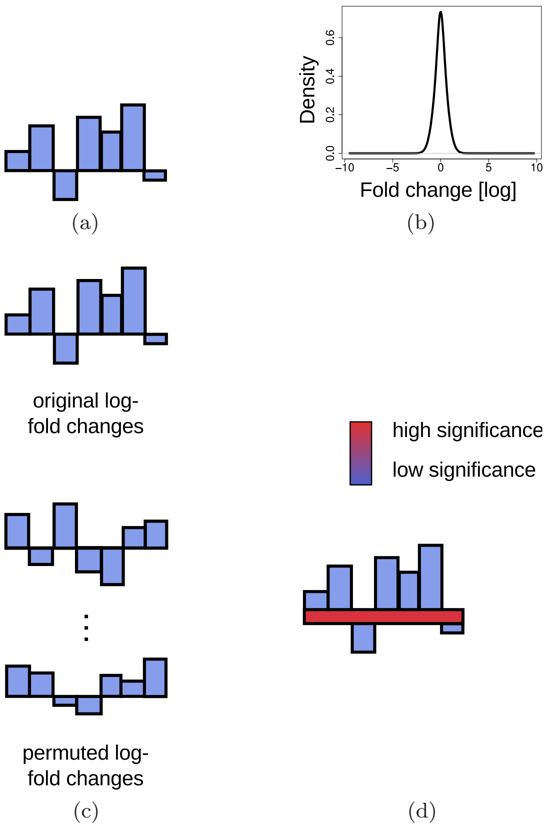
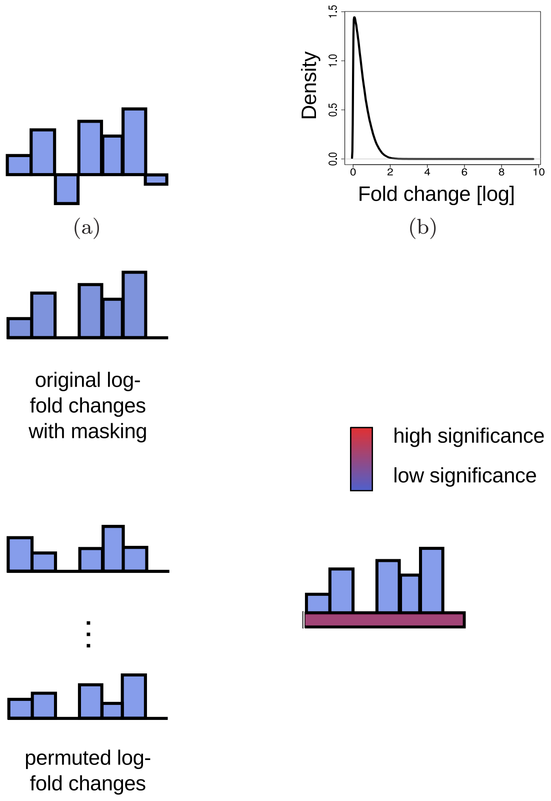
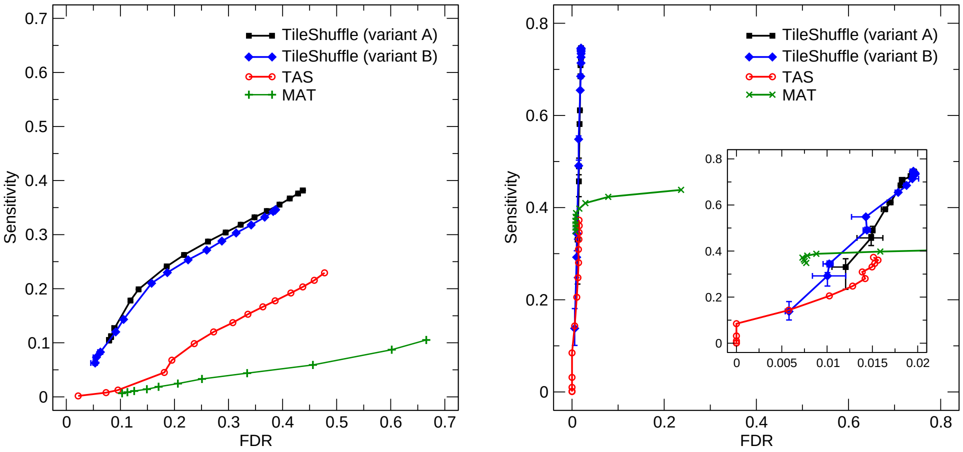


Figure S5: **Detailed outline of `TileShuffle` (detection of differentially expressed segments - variant A).** (a) Window of probes with log-fold changes between two different cellular states. (b) Density of the distribution of positive and negative log-fold changes on a tiling array. (c) Original log-fold changes and permuted log-fold changes with entire background. (d) Assignment of the significance in terms of a corrected empirical two-tailed p-value to the window by comparing the original window score to the permuted ones. Note that the multiple testing correction is applied to all window p-values on the tiling array. Furthermore, it is adjusted to account for the additional comparisons in case of the two-tailed p-value estimations.



(a) (b) (c) (d)

Figure S6: **Detailed outline of `TileShuffle` (detection of differentially expressed segments - variant B).** (a) Positive window of probes with log-fold changes between two different cellular states. (b) Density of the distribution of positive log-fold changes on a tiling array. (c) Log-fold changes of non-masked probes, i.e., probes with positive log-fold changes in case of a positive window, and permuted log-fold changes of non-masked probes with positive background. (d) Assignment of the significance in terms of a corrected empirical one-tailed p-value to the window by comparing the original window score to the permuted ones. Note that the multiple testing correction is applied to all window p-values on the tiling array.



(a)

(b)

Figure S7: **Comparison of TileShuffle with TAS and MAT: Detection of *highdiff* segments in the G0/G1 transition of the cell cycle dataset (a) and in the spike-in tiling array dataset between the concentrations of 0.0055μg and 0.055μg (b).** Sensitivities as function of FDR after evaluating the outcome TAS, MAT, and TileShuffle with both variants with a range of different p/q-value cutoffs in the differential analysis. In the cell cycle dataset, the positive set is obtained by conducting and evaluating verification experiments using a custom-designed microarray with triplicates while in the spike-in dataset it is comprised of regions covered by the 162 full-length cDNA clones which were spiked in. Note that the whiskers express the variation in the outcome of TileShuffle after five repetitions, i.e., smallest and highest value on the x-axis (or y-axis) for each differential significance threshold, with the median result shown on the solid line. The inlay magnifies the area in the right panel where the x-coordinate is close to zero (same units on axes).

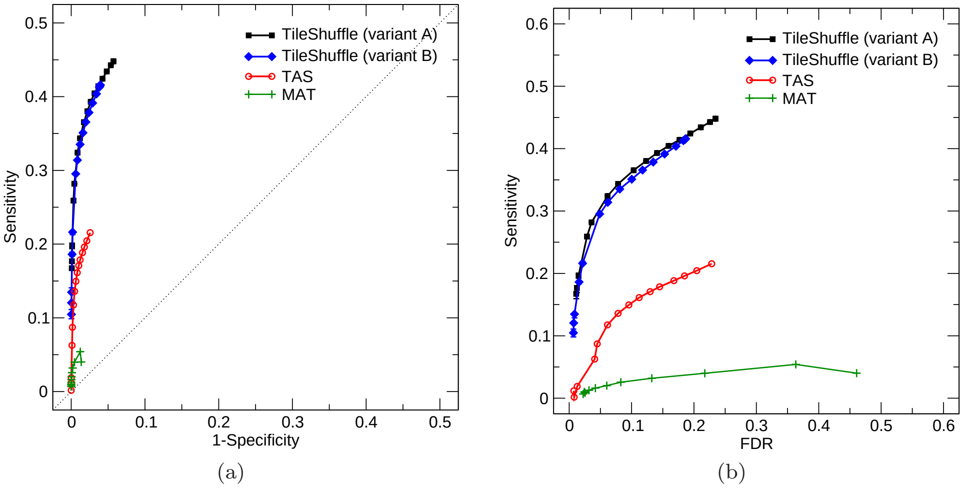
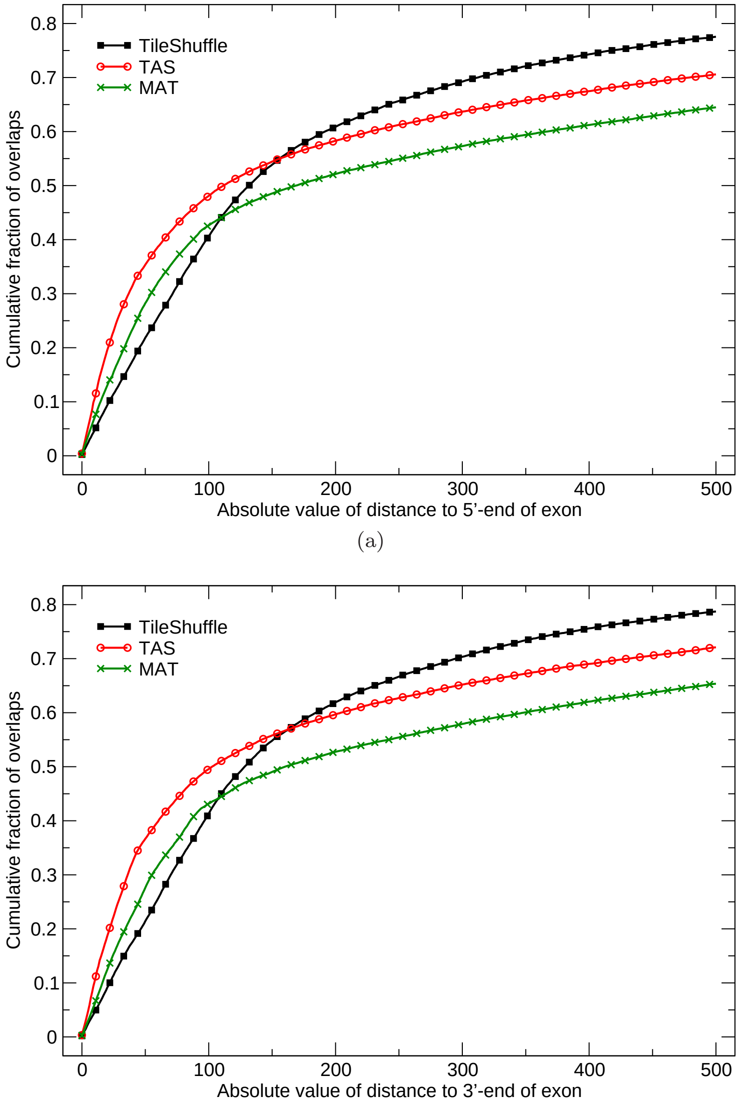


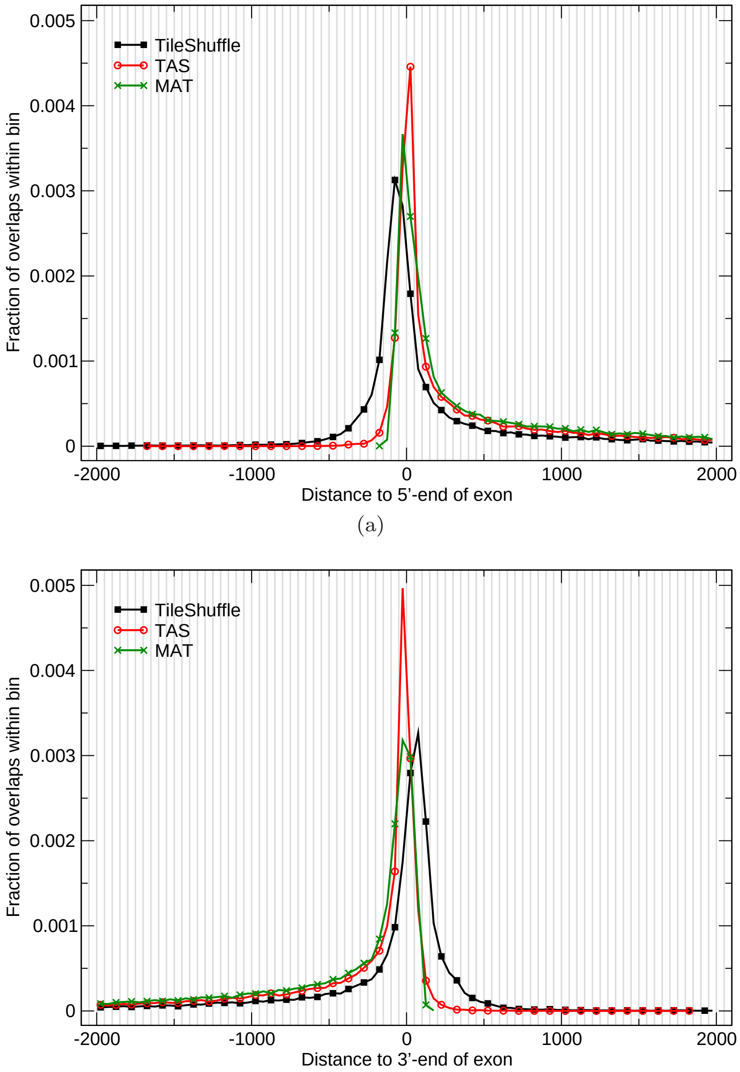
Figure S8: **Count-based comparison of `TileShuffle` with `TAS` and `MAT`: Detection of *highdiff* segments in the G0/G1 transition of the cell cycle tiling array dataset.** ROC curve (a) and sensitivity as function of FDR (b) after evaluating the outcome `TAS`, `MAT`, and `TileShuffle` with both variants with a range of different p/q-value cutoffs in the differential analysis. Here, the evaluations are based on counts rather than on nucleotides. Then, TP corresponds to the number of tiling array regions that are *highdiff* and contain at least one probe on the custom microarray that was found significantly differentially expressed. The number of false positives (FP) is defined as the number of *highdiff* regions that do not contain a probe that is significantly differentially expressed in the custom microarray experiment. P is defined as the number of probes that are significantly differentially expressed in the custom microarray experiment ($FDR < 0.05$), and N as the number of probes that are not significantly differentially expressed, respectively. Note that the whiskers express the variation in the outcome of `TileShuffle` after five repetitions, i.e., smallest and highest value on the x-axis (or y-axis) for each differential significance threshold, with the median result shown on the solid line.



(a)

(b)

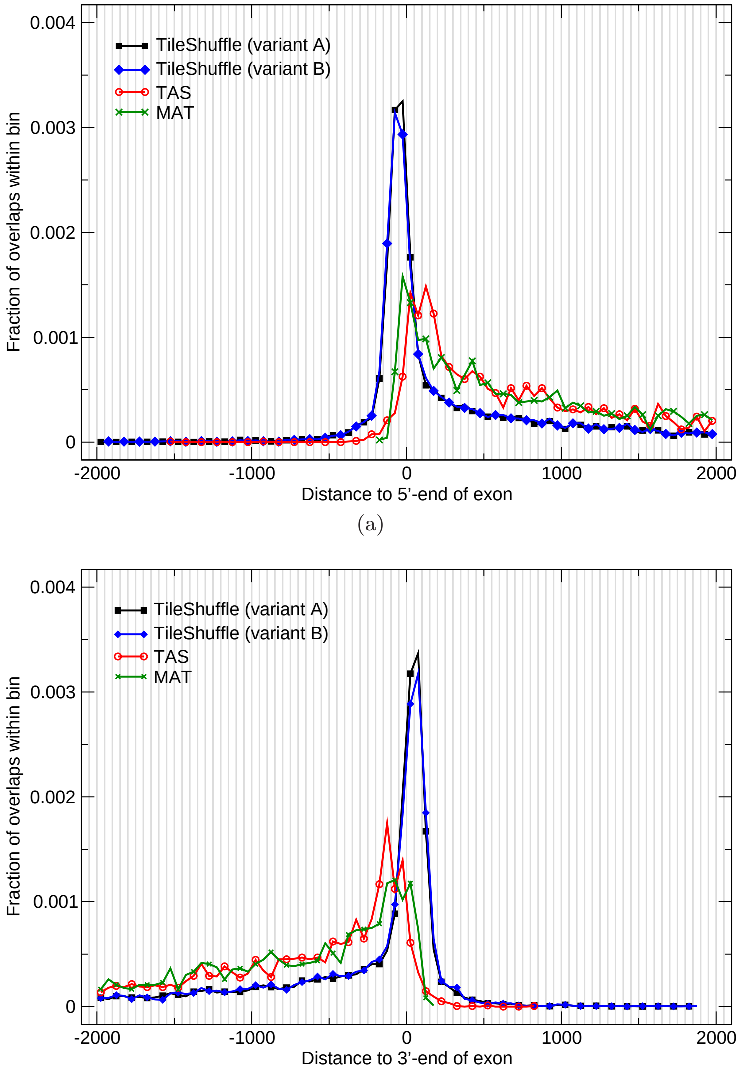
Figure S9: **Comparison of `TileShuffle` with `TAS` and `MAT`: Detection of transcript structures on the basis of highly expressed regions in the G0 phase of the cell cycle tiling array dataset.** Empirical cumulative distribution function of the absolute distances between 5′- (a) and 3′-end (b), respectively, of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. Only every 10th data point is drawn as a symbol.



(a)

(b)

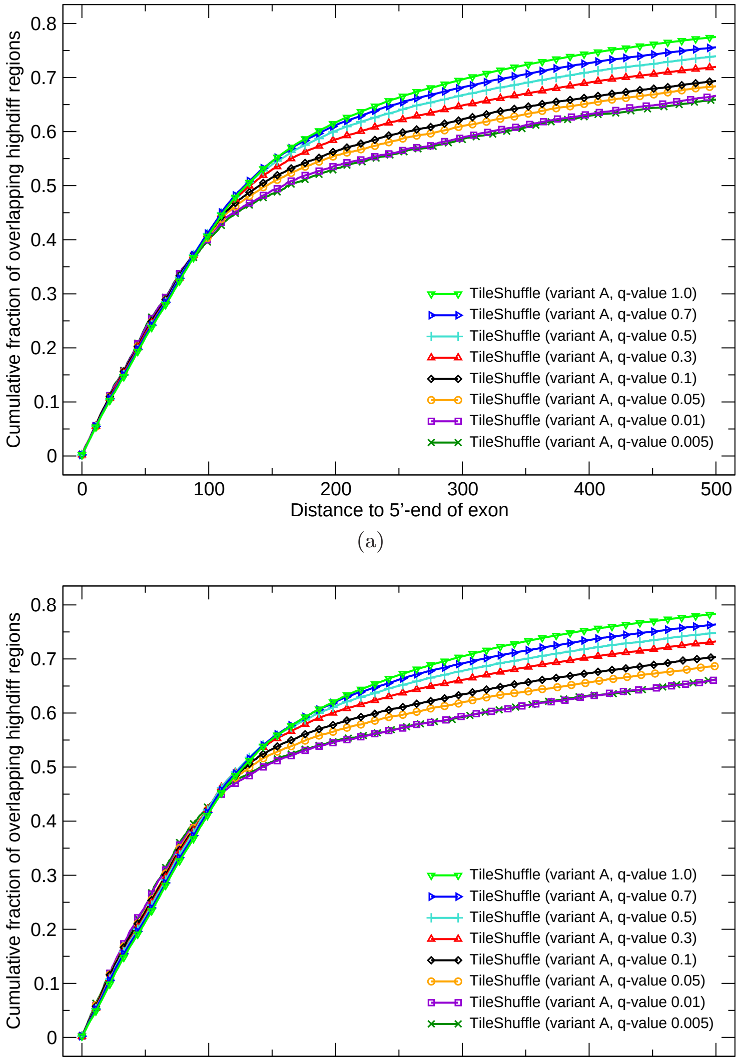
Figure S10: **Comparison of `TileShuffle` with `TAS` and `MAT`: Distribution of distances between annotated exons and highly expressed regions in the G0 phase of the cell cycle tiling array dataset.** Frequency polygons with bin size of 50nt on the distribution of distances between 5′- (a) and 3′-end (b) of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. A frequency polygon simply is a density estimator based on a histogram where the mid points of the histogram bars are connected by straight lines. The breaks in the histogram are illustrated as gray vertical lines. Only every second point is drawn as a symbol.



(a)

(b)

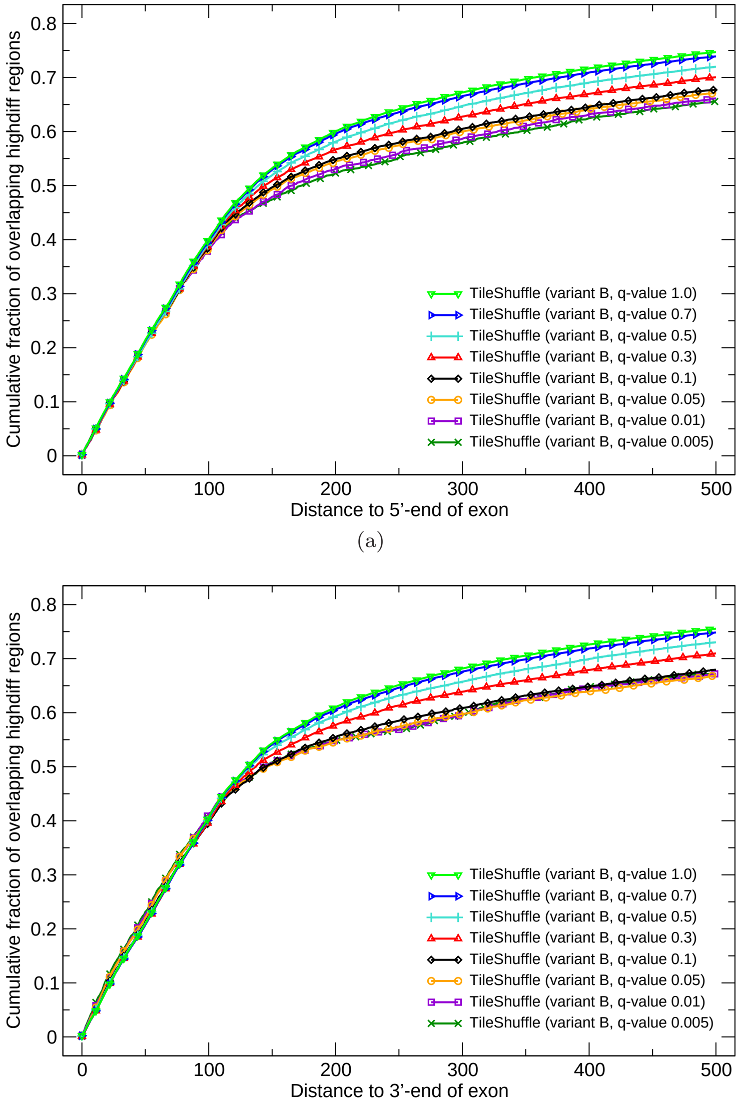
Figure S11: **Comparison of `TileShuffle` with `TAS` and `MAT`: Distribution of distances between annotated exons and *highdiff* regions in the G0/G1 transition of the cell cycle tiling array dataset.** Frequency polygons with a bin size of 50nt on the distribution of distances between 5′- (a) and 3′-end (b) of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. A frequency polygon simply is a density estimator based on a histogram where the mid points of the histogram bars are connected by straight lines. The breaks in the histogram are illustrated as gray vertical lines. Only every second point is drawn as a symbol. The significance thresholds in the differential analyses of the methods were adjusted to obtain similar FDRs as estimated before using the custom microarray, i.e., 18% FDR in case of `TAS` (q=0.05), 17% in case of `MAT` (p=1e-6), and 19% and 18% in case of `TileShuffle` with variant A (q=0.05) and variant B (q=0.1), respectively. The absolute number of overlaps is 15 835 and 13 479 with `TileShuffle` and variant A and B, respectively, 4337 with `TAS`, and 2381 with `MAT`.



(a)

(b)

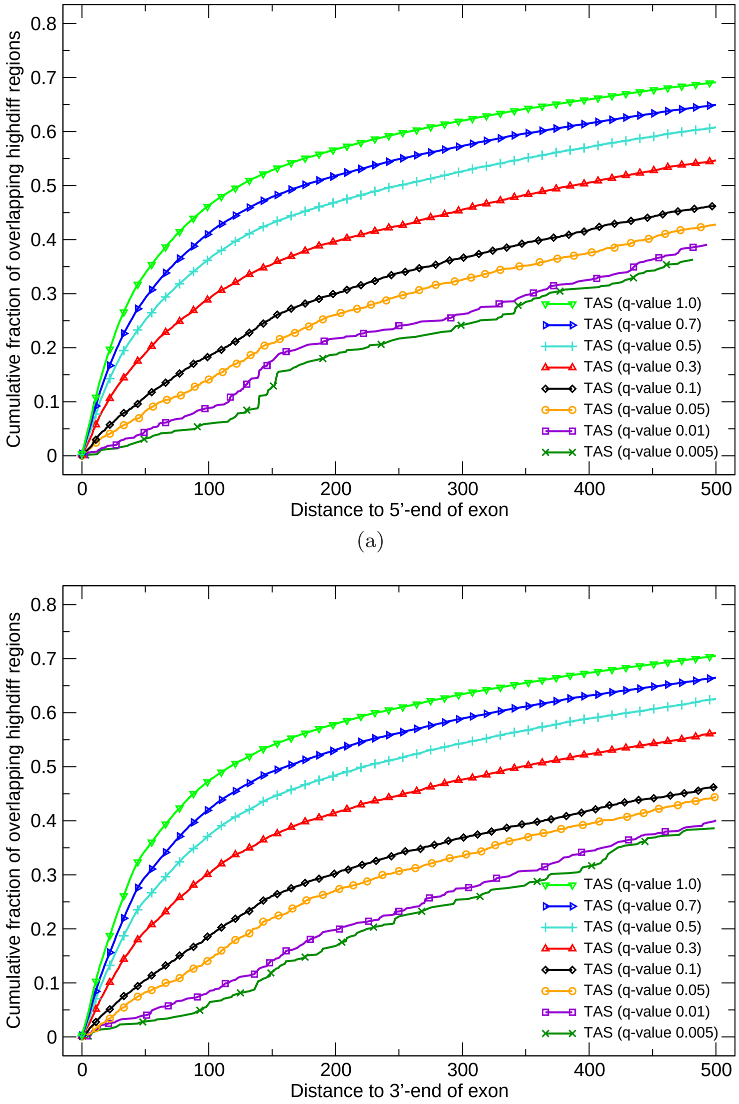
Figure S12: **Comparison of `TileShuffle` (variant A) with different q-values: Detection of transcript structures on the basis of *highdiff* regions in the G0/G1 transition of the cell cycle tiling array dataset.** Empirical cumulative distribution function of the absolute distances between 5′- (a) and 3′-end (b), respectively, of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. Only every 10th data point is drawn as a symbol.



(a)

(b)

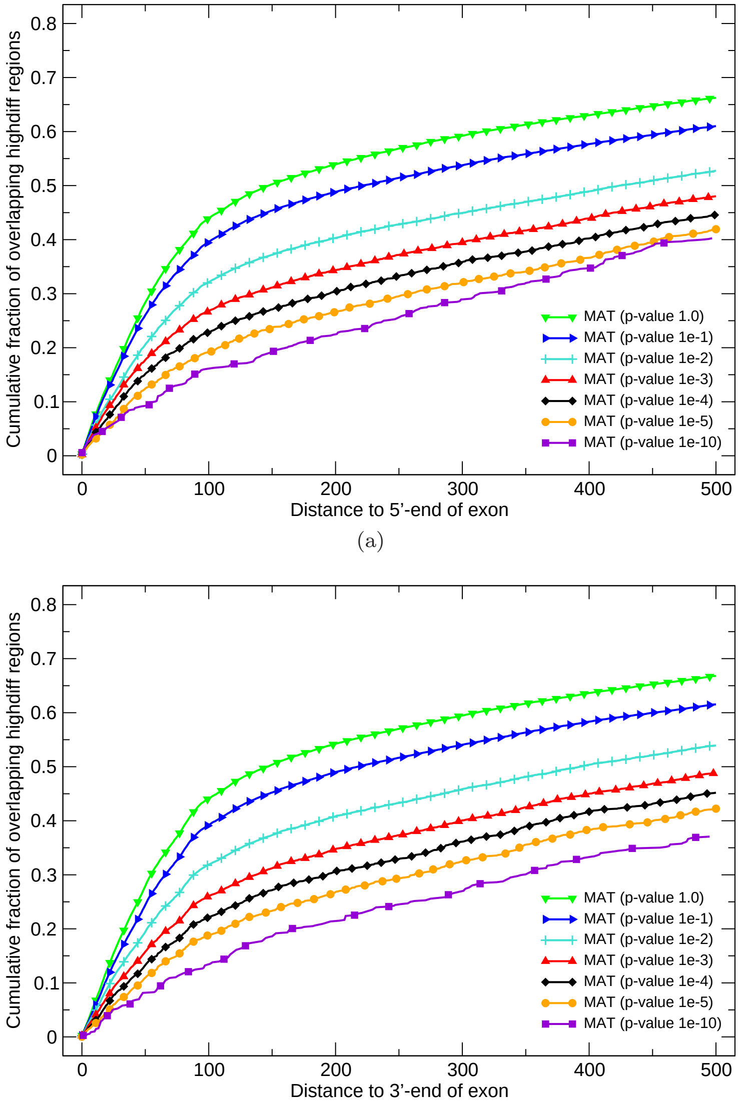
Figure S13: **Comparison of `TileShuffle` (variant B) with different q-values: Detection of transcript structures on the basis of *highdiff* regions in the G0/G1 transition of the cell cycle tiling array dataset.** Empirical cumulative distribution function of the absolute distances between 5′- (a) and 3′-end (b), respectively, of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. Only every 10th data point is drawn as a symbol.



(a)

(b)

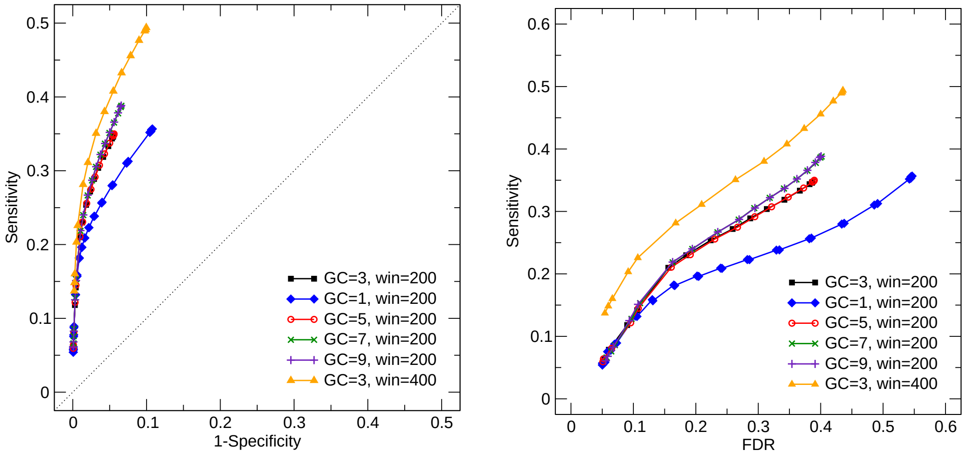
Figure S14: **Comparison of `TAS` with different q-values: Detection of transcript structures on the basis of *highdiff* regions in the G0/G1 transition of the cell cycle tiling array dataset.** Empirical cumulative distribution function of the absolute distances between 5′- (a) and 3′-end (b), respectively, of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. Only every 10th data point is drawn as a symbol.



(a)

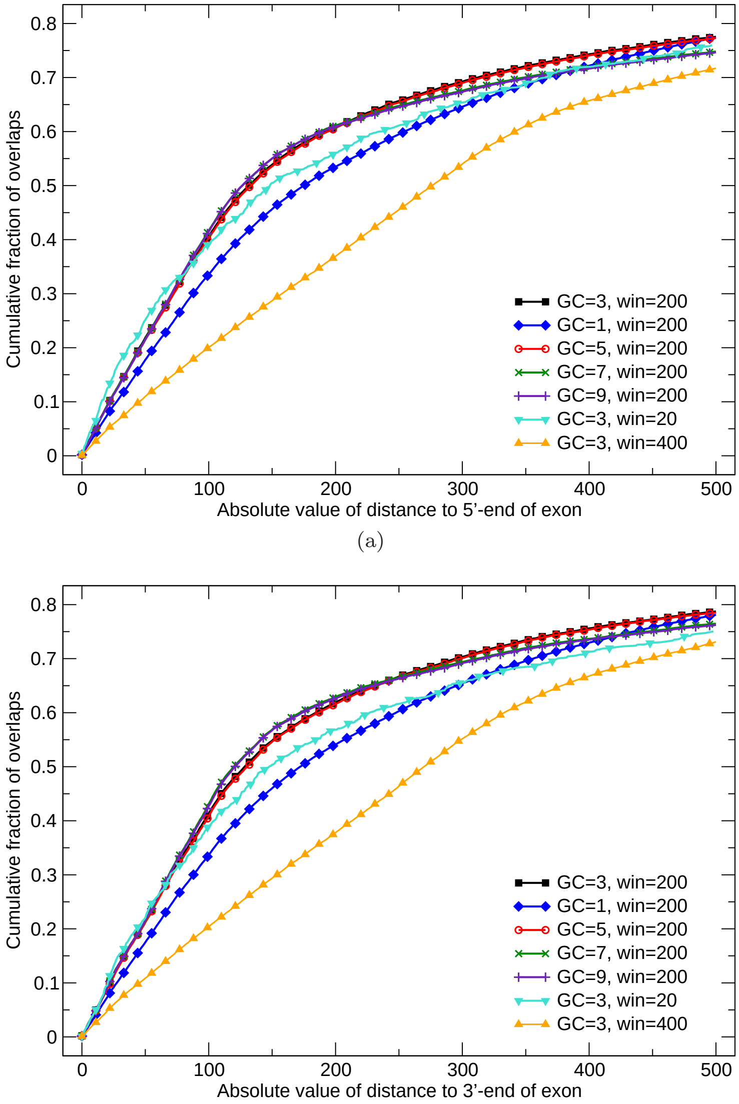
(b)

Figure S15: **Comparison of `MAT` with different q-values: Detection of transcript structures on the basis of *highdiff* regions in the G0/G1 transition of the cell cycle tiling array dataset.** Empirical cumulative distribution function of the absolute distances between 5′- (a) and 3′-end (b), respectively, of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. Only every 10th data point is drawn as a symbol.



(a) (b)

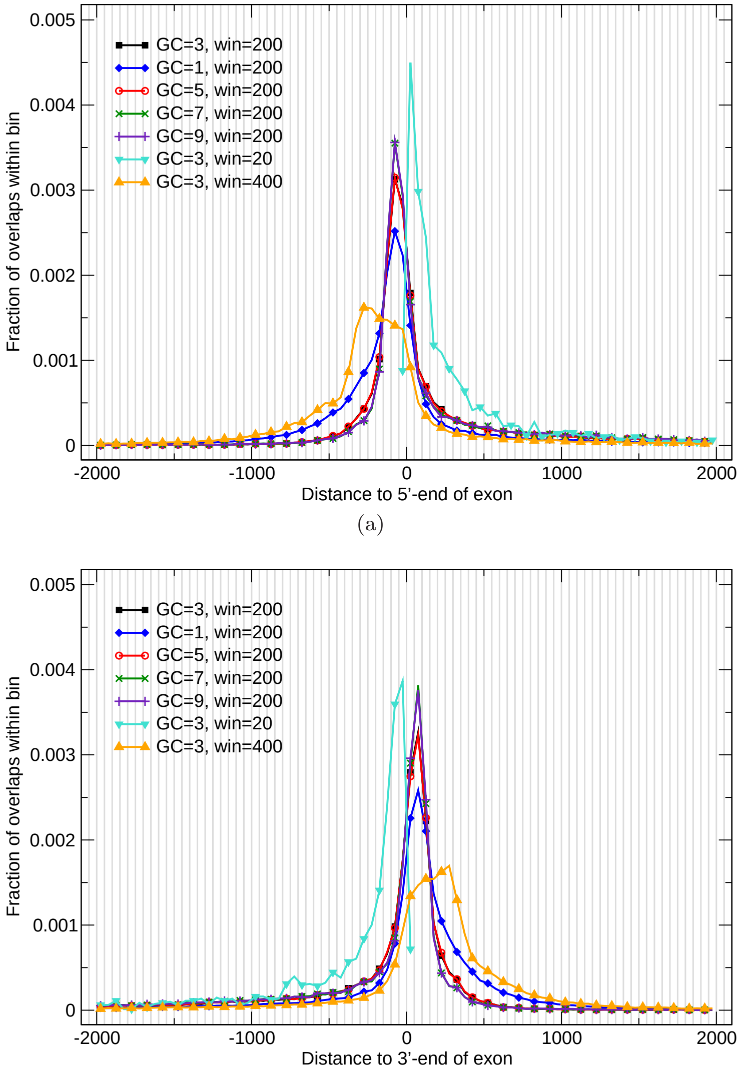
Figure S16: **Comparison of `TileShuffle` (variant B) with different number of GC bins and window sizes: Detection of *highdiff* segments in the G0/G1 transition of the cell cycle tiling array dataset with a range of different q-value cutoffs in the differential analysis.** ROC curve (a) and sensitivity as function of FDR (b) after evaluating the outcome of `TileShuffle` (variant B) with different number of GC bins and window sizes with a range of different p/q-value cutoffs in the differential analysis. The positive set is obtained by conducting and evaluating verification experiments using a custom-designed microarray with triplicates. Note that TileShuffle with GC=3 and win=20 could not be evaluated since it was not represented on the custom microarray.



(a)

(b)

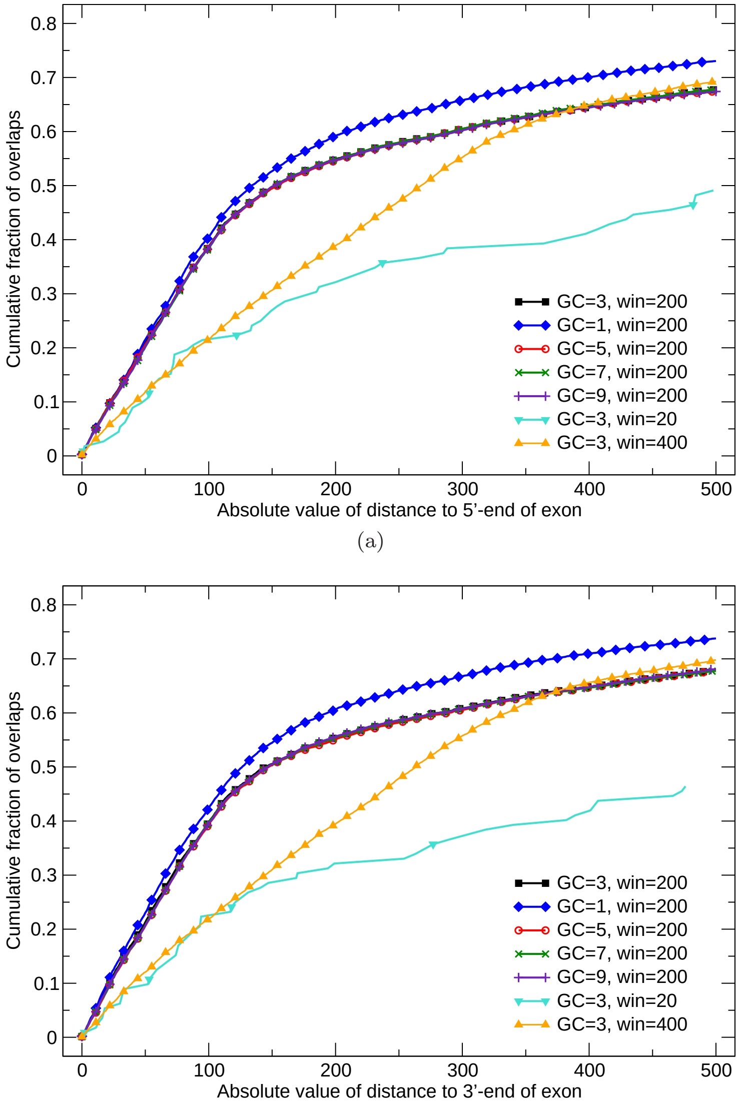
Figure S17: **Comparison of `TileShuffle` with different number of GC bins and window sizes: Detection of transcript structures on the basis of highly expressed regions in the G0 phase of the cell cycle tiling array dataset.** Empirical cumulative distribution function of the absolute distances between 5′- (a) and 3′-end (b), respectively, of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. Only every 10th data point is drawn as a symbol.



(a)

(b)

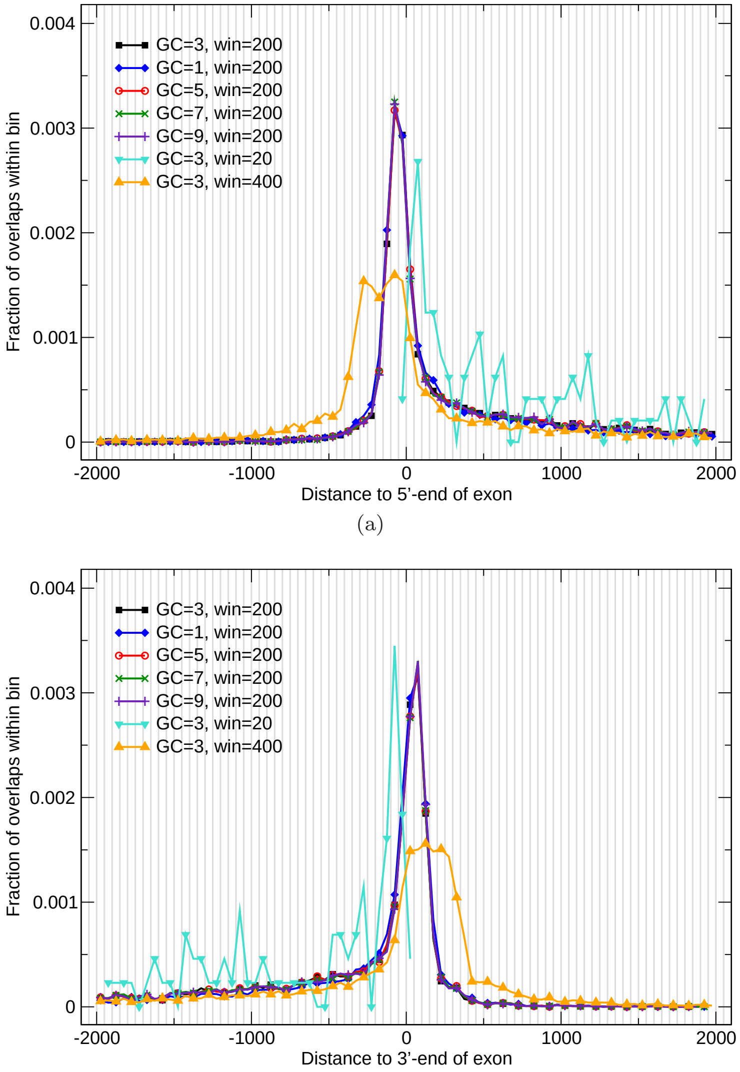
Figure S18: **Comparison of `TileShuffle` with different number of GC bins and window sizes: Distribution of distances between annotated exons and highly expressed regions in the G0 phase of the cell cycle tiling array dataset.** Frequency polygons with bin size of 50nt on the distribution of distances between 5′- (a) and 3′-end (b) of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. A frequency polygon simply is a density estimator based on a histogram where the mid points of the histogram bars are connected by straight lines. The breaks in the histogram are illustrated as gray vertical lines. Only every second point is drawn as a symbol.



(a)

(b)

Figure S19: **Comparison of `TileShuffle` (variant B) with different number of GC bins and window sizes: Detection of transcript structures on the basis of *highdiff* regions in the G0/G1 transition of the cell cycle tiling array dataset.** Empirical cumulative distribution function of the absolute distances between 5′- (a) and 3′-end (b), respectively, of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. Only every 10th data point is drawn as a symbol. The significance thresholds in the differential analyses of the methods were adjusted to obtain similar FDRs as estimated before using the custom microarray.



(a)

(b)

Figure S20: **Comparison of `TileShuffle` (variant B) with different number of GC bins and window sizes: Distribution of distances between annotated exons and *highdiff* regions in the G0/G1 transition of the cell cycle tiling array dataset.** Frequency polygons with a bin size of 50nt on the distribution of distances between 5′- (a) and 3′-end (b), respectively, of exon and reported interval for all overlapping pairs of unique GENCODE annotated exons and reported intervals. Overlapping here means any overlap in genomic coordinates ignoring strand. A frequency polygon simply is a density estimator based on a histogram where the mid points of the histogram bars are connected by straight lines. The breaks in the histogram are illustrated as gray vertical lines. Only every second point is drawn as a symbol. The significance thresholds in the differential analyses of the methods were adjusted to obtain similar FDRs as estimated before using the custom microarray.

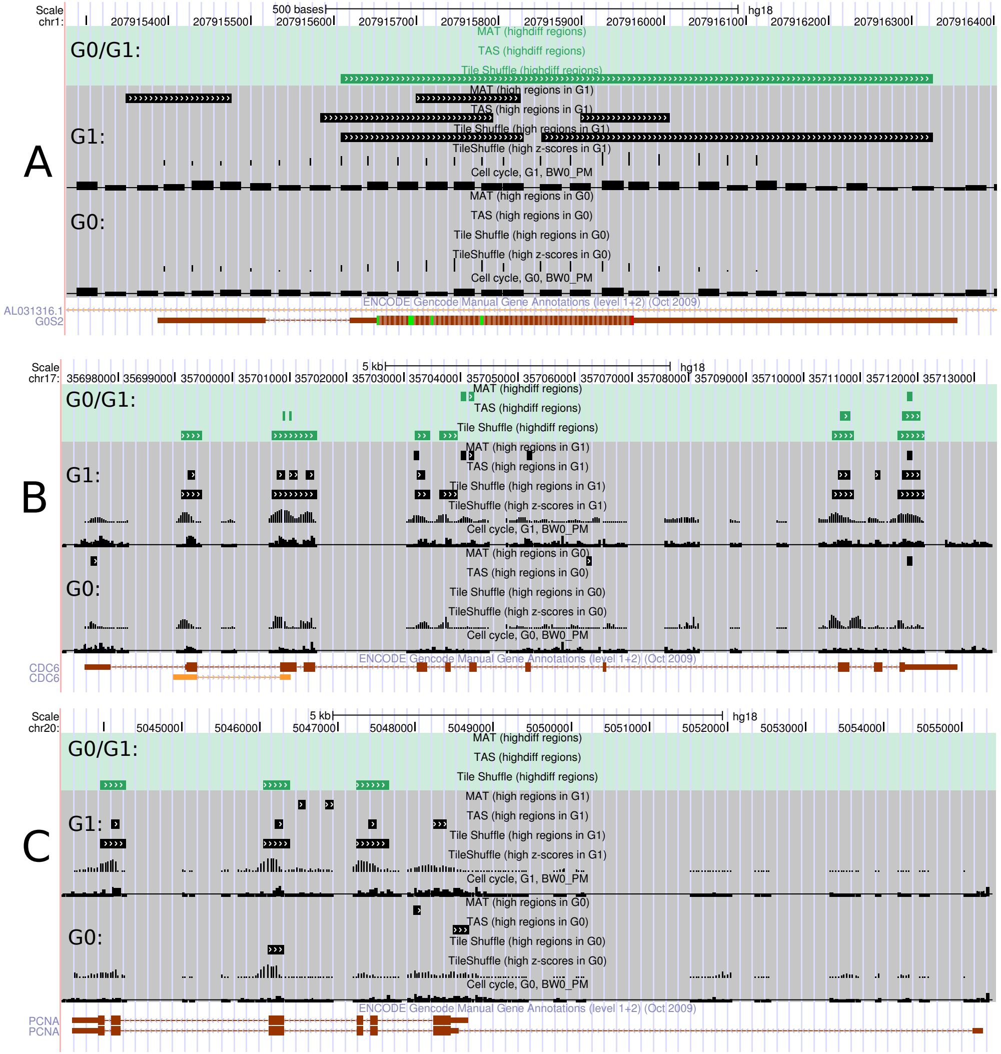


Figure S21: **Examples of identified transcript structures of two protein-coding genes known to change expression from cell cycle state G0 to state G1 for `TileShuffle` variant B, `TAS`, and `MAT` in the cell cycle tiling array dataset.** A: Detection of exons of the gene GOS2 (G0/G1switch 2), B: detection of exons of the gene CDC6 (cell division cycle 6 homolog). G0S2 and CDC6 are known to be upregulated in G1 phase [1, 2]. PCNA (proliferating cell nuclear antigen) is known to be expressed in G1 phase [3].The transcript structures of both genes are defined according to the GENCODE version 3c. Parameters for the one-state analyses (identification of highly expressed segments) are chosen as follows: `TileShuffle` with $q < 0.05$, `MAT` with $p < 0.05$, and `TAS` with MM corrected PM probe intensities above the threshold of 150. Parameters for the two-state analyses (identification of *highdiff* regions) are chosen as follows: `TileShuffle` with $q < 0.1$ (variant B), `MAT` with a $p < 10^{-6}$, and `TAS` with $q < 0.05$ in order to yield comparable FDR values.

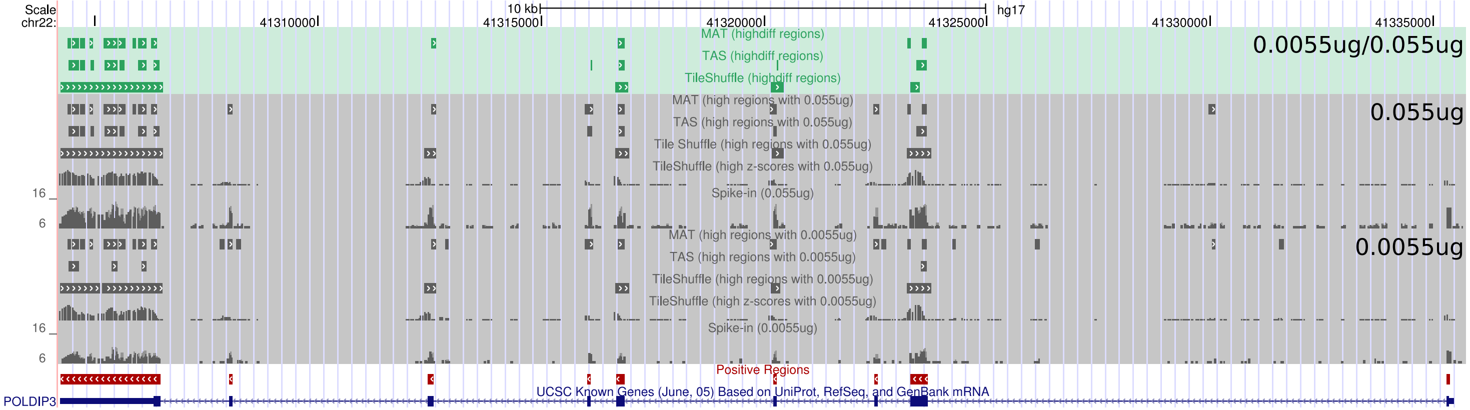


Figure S22: **Examples of identified transcript structures of a positive regions for `TileShuffle` variant B, `TAS`, and `MAT` in the spike-in tiling array dataset between the concentrations 0.0055μg and 0.055μg.** Detection of positive regions in the spike-in tiling array dataset between the concentrations 0.0055μg and 0.055μg. Parameters for the one-state analyses (identification of highly expressed segments) are chosen as follows: `TileShuffle` with $q < 0.05$, `MAT` with $p < 0.05$, and `TAS` with MM corrected PM probe intensities above the threshold of 150. Parameters for the two-state analyses (identification of *highdiff* regions) are chosen as follows: `TileShuffle` with $q < 0.1$ (variant B), `MAT` with a $p < 10^{-4}$, and `TAS` with $q < 0.8$ in order to yield comparable FDR values.

Type of experiment	MAT	TAS	TileShuffle	Custom Array
cell cycle	GSE36187	GSE36189	GSE36190	GSE29792

Table S1: **GEO accession IDs for human tiling array datasets used in this study and the custom microarray used for validation.** For each algorithm, a Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) superseries has been created including the datasets of highly expressed regions as well as *highdiff* regions. The last column gives the GEO accession IDs for the custom microarray data that has been used to validate the outcome of all three algorithms.

	regions	nucleotides	mean length
TileShuffle (GC=3, win=200)	65 680	23 085 575	351.5
TileShuffle (GC=1, win=200)	220 014	88 797 882	403.6
TileShuffle (GC=5, win=200)	67 000	23 708 839	353.9
TileShuffle (GC=7, win=200)	56 775	20 487 147	360.8
TileShuffle (GC=9, win=200)	56 804	20 527 250	361.4
TileShuffle (GC=3, win=20)	4 310	110 074	25.5
TileShuffle (GC=3, win=400)	77 867	58 138 624	746.6
TAS	95 840	13 924 718	145.3
MAT	280 548	27 708 046	98.8

Table S2: **Comparison of `TileShuffle` using different number of GC bins and window sizes with `TAS` and `MAT`:** Quantity, nucleotides, and average length of highly expressed in the G0 phase of the cell cycle tiling array dataset. The default settings for the number of GC bins and window size are indicated in bold.

	regions	nucleotides	mean length
TileShuffle (Variant A, GC=3, win=200)	18 299	6 185 940	338
TileShuffle (Variant B, GC=3, win=200)	16 043	5 182 484	323
TileShuffle (Variant B, GC=1, win=200)	15 898	4 849 800	305.1
TileShuffle (Variant B, GC=5, win=200)	17 255	5 546 853	321.5
TileShuffle (Variant B, GC=7, win=200)	18 946	6 128 117	323.5
TileShuffle (Variant B, GC=9, win=200)	19 090	6 172 643	323.3
TileShuffle (Variant B, GC=3, win=20)	703	17 673	25.1
TileShuffle (Variant B, GC=3, win=400)	15 866	10 612 052	668.9
TAS	5 470	746 184	136.4
MAT	3 020	284 518	94.2

Table S3: **Comparison of `TileShuffle` using different number of GC bins and window sizes with `TAS` and `MAT`:** Quantity, nucleotides, and average length of *highdiff* regions in the G0/G1 transition of the cell cycle tiling array dataset. The default settings for the number of GC bins and window size are indicated in bold. The significance thresholds in the differential analyses of the methods were adjusted to obtain similar FDRs as estimated before using the custom microarray.

	overall	exons	introns	5'-UTR	3'-UTR	CDS	intergenic	non-exonic (novel)
TileShuffle (GC=3, win=200)	23 085 575	9 630 623	9 976 153	1 401 154	4 002 576	4 958 849	5 134 922	13 454 952
TileShuffle (GC=1, win=200)	88 797 882	21 181 872	42 560 546	4 089 095	5 921 145	13 065 024	29 252 823	67 616 010
TileShuffle (GC=5, win=200)	23 708 839	9 828 747	10 355 747	1 401 146	4 203 847	4 952 081	5 204 545	13 880 092
TileShuffle (GC=7, win=200)	20 487 147	8 842 836	9 029 187	823 543	4 258 464	4 411 174	4 054 238	11 644 311
TileShuffle (GC=9, win=200)	20 527 250	8 861 036	9 057 023	823 841	4 287 601	4 403 656	4 055 591	11 666 214
TileShuffle (GC=3, win=20)	110 074	30 230	38 262	7 437	7 461	17 750	49 424	79 844
TileShuffle (GC=3, win=400)	58 138 624	18 669 097	29 657 851	2 920 280	8 051 262	9 009 867	13 212 891	39 469 527
TAS	13 924 718	5 882 732	5 586 961	581 220	2 129 572	3 714 287	3 424 662	8 019 604
MAT	27 707 039	3 657 494	13 540 120	385 967	1 580 965	2 000 879	11 104 453	24 047 546

Table S4: **Comparison of `TileShuffle` using different number of GC bins and window sizes with `TAS` and `MAT`:** Base pair overlap between highly expressed regions in the G0 phase of the cell cycle tiling array dataset and GENCODE version 3c annotations.

	overall	exons	introns	5’-UTR	3’-UTR	CDS	intergenic	non-exonic (novel)
TileShuffle (Variant A, GC=3, win=200)	6 185 648	3 128 775	2 613 639	161 654	1 716 580	1 421 648	881 018	3 056 873
TileShuffle (Variant B, GC=3, win=200)	5 180 272	2 514 956	2 254 373	137 856	1 377 062	1 138 181	763 634	2 665 316
TileShuffle (Variant B, GC=1, win=200)	4 845 447	2 139 056	2 077 488	163 346	981 066	1 129 788	962 093	2 706 391
TileShuffle (Variant B, GC=5, win=200)	5 544 073	2 619 939	2 475 285	142 267	1 455 302	1 165 872	816 444	2 924 134
TileShuffle (Variant B, GC=7, win=200)	6 124 282	2 858 239	2 764 521	137 025	1 601 331	1 278 510	901 613	3 266 043
TileShuffle (Variant B, GC=9, win=200)	6 168 836	2 875 126	2 795 081	140 457	1 609 812	1 286 798	901 863	3 293 710
TileShuffle (Variant B, GC=3, win=20)	17 673	1 942	7 316	31	1 394	573	8 965	15 731
TileShuffle (Variant B, GC=3, win=400)	10 600 047	3 894 467	5 633 565	197 834	2 329 656	1 559 489	1 656 765	6 705 580
TAS	631 950	347 219	247 969	10 497	215 849	135 465	87 731	284 335
MAT	284 518	147 436	125 786	3 377	107 509	41 108	32 119	137 082

Table S5: **Comparison of `TileShuffle` using different number of GC bins and window sizes with `TAS` and `MAT`:** Base pair overlap between *highdiff* regions in the G0/G1 transition of the cell cycle tiling array dataset and Gencode version 3c annotations. The significance thresholds in the differential analyses of the methods were adjusted to obtain similar FDRs as estimated before using the custom microarray.

	Number of tiling array regions
TileShuffle (Variant A, GC=3, win=200)	20 599 (56%)
TileShuffle (Variant B, GC=3, win=200)	20 103 (63%)
TileShuffle (Variant B, GC=1, win=200)	14 898 (47%)
TileShuffle (Variant B, GC=3, win=20)	0 (0%)
TileShuffle (Variant B, GC=3, win=400)	18 080 (57%)
TileShuffle (Variant B, GC=5, win=200)	20 249 (59%)
TileShuffle (Variant B, GC=7, win=200)	20 894 (55%)
TileShuffle (Variant B, GC=9, win=200)	20 845 (55%)
TAS	5 144 (56%)
MAT	2 276 (38%)

Table S6: **Representation of *highdiff* tiling array regions on custom microarray:** Number and fraction of *highdiff* intervals that are represented by at least one probe on the custom microarray. Numbers base upon all tiling array regions identified by either `TileShuffle`, `TAS`, or `MAT` to be significantly differentially expressed between G0/G1 transition of the cell cycle tiling array dataset. A tiling array region is represented on the custom microarray if the custom microarray contains at least one probe overlapping completely with the tiling array region. The significance thresholds in the differential analyses of the methods were adjusted to obtain similar FDRs as estimated before using the custom microarray.

References

[1] L. Russell and D. R. Forsdyke. A human putative lymphocyte G0/G1 switch gene containing a CpG-rich island encodes a small basic protein with the potential to be phosphorylated. *DNA Cell Biol*, 10(8):581–591, Oct 1991.

[2] Tong Zhou, Jeff W. Chou, Dennis A. Simpson, Yingchun Zhou, Thomas E. Mullen, Margarida Medeiros, Pierre R. Bushel, Richard S. Paules, Xuebin Yang, Patrick Hurban, Edward K. Lobenhofer, and William K. Kaufmann. Profiles of global gene expression in ionizing-radiation-damaged human diploid fibroblasts reveal synchronization behind the G1 checkpoint in a G0-like state of quiescence. *Environ Health Perspect*, 114(4):553–559, Apr 2006.

[3] M L Whitfield, G Sherlock, A J Saldanha, J I Murray, C A Ball, K E Alexander, J C Matese, C M Perou, M M Hurt, P O Brown, and D Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell*, 13(6):1977–2000, Jun 2002.