Supplementary data to:

Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions

Wu Xia[1,2,11], Kabalane Hadi[3,11], Kahli Malik[1], Petryk Nataliya[1], Laperrousaz Bastien[3,4], Jaszczyszyn Yan[5], Drillon Guenola[3], Nicolini Frank-Emmanuel[4,6], Perot Gaëlle[7], Robert Aude[8], Fund Cédric[9], Chibon Frédéric[7], Xia Ruohong[2], Wiels Joëlle[8], Argoul Françoise[10], Maguer-Satta Véronique[4], Arneodo Alain[10], Audit Benjamin[3*] and Hyrien Olivier[1*]

(1) *Institut de Biologie de l'École Normale Supérieure (IBENS), Département de Biologie, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, F-75005 Paris, France.* (2) *Physics Department, East China Normal University, Shanghai, China.* (3) *Univ Lyon, ENS de Lyon, Univ Claude Bernard Lyon 1, CNRS, Laboratoire de Physique, F-69342, Lyon, France.* (4) *CNRS UMR5286, INSERM U1052, Centre de Recherche en Cancérologie de Lyon, F- 69008 Lyon, France.* (5) *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, France.* (6) *Centre Léon Bérard, Lyon, France.* (7) *INSERM U1218, Institut Bergonié, F-33000, Bordeaux, France.* (8) *UMR 8126, Université Paris-Sud Paris-Saclay, CNRS, Institut Gustave Roussy, Villejuif, France.* (9) *École Normale Supérieure, PSL Research University, CNRS, Inserm, IBENS, Plateforme Génomique, 75005 Paris, France.* (10) *LOMA, Université de Bordeaux, CNRS, UMR 5798, Talence, F-33405, France.*

[11] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

*To whom correspondance should be addressed. Tel :33 1 4432 3920; Fax: 33 1 4432 3941; Email: hyrien@biologie.ens.fr. Correspondance may also be addressed to Benjamin Audit. Tel: 33 4 7272 8691; Email: Benjamin.Audit@ens-lyon.fr.

Present address: [Malik Kahli] Department of Biology, New York University, New York, NY 10003, USA.
Present address: [Nataliya Petryk] BRIC - Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark.
Present address: [Frédéric Chibon] INSERM U1037 - CRCT, Institut Claudius Regaud, 31037 Toulouse.

Contents:

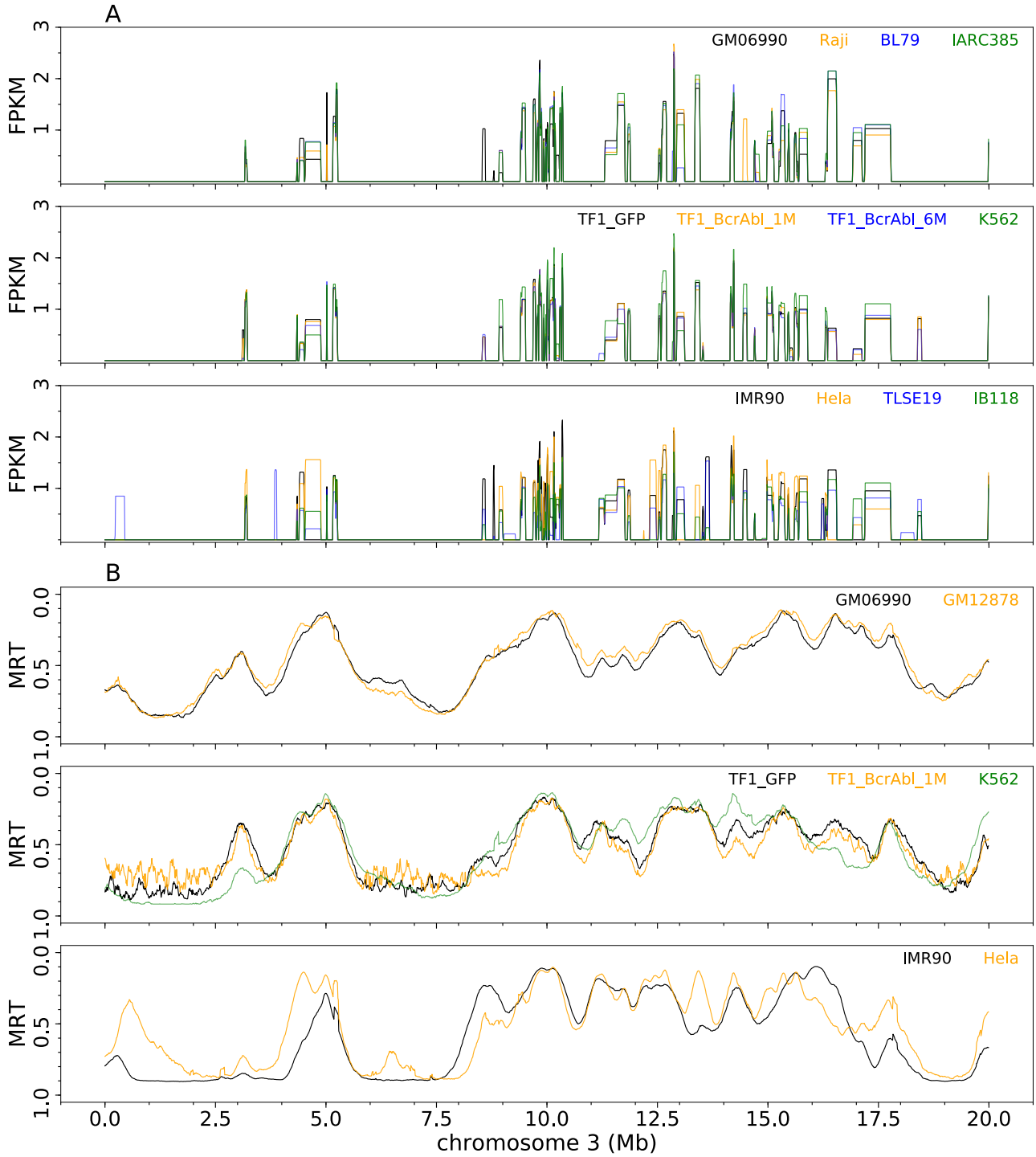


Figure S1. Genome-wide profiling of gene expression (A) and replication timing (B). Shown are RNA-seq (A) and MRT (B) data for the same genomic region and the same cell lines as in Figure 1A.

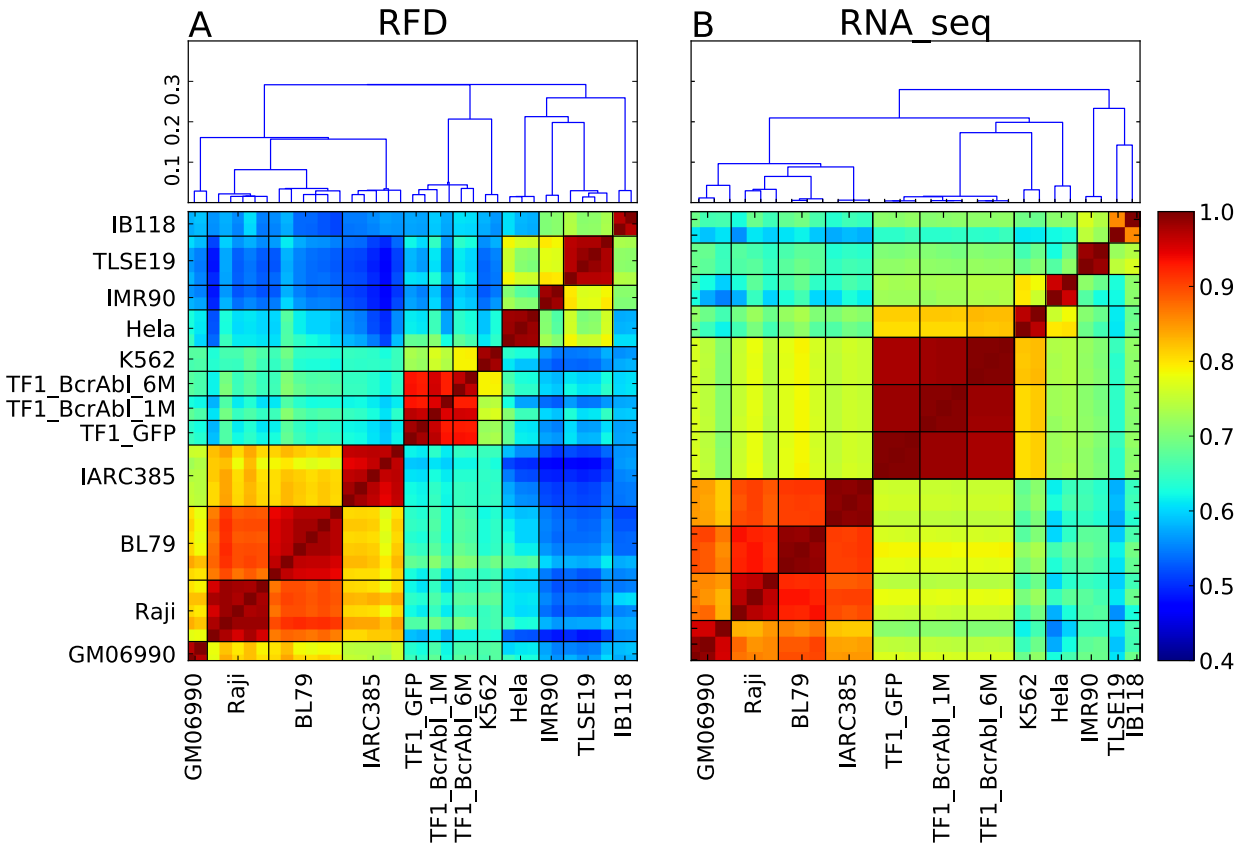


Figure S2. Cell line classification based on correlations between replication and gene expression profiles for all biological replicates. (Bottom) Correlation matrices between RFD profiles (C_{RFD}; A) and RNA-seq ($C_{RNA\text{-}seq}$; B) profiles. Pearson correlation coefficient values are colour-coded from blue (0.4) to red (1) using the colour bar on the right (Materials and Methods). Dark lines separate cell line groups of biological replicates. (Top) Dendrogram representation of the hierarchical classification of cell lines based on the corresponding correlation matrix; ordinate is the correlation distance.

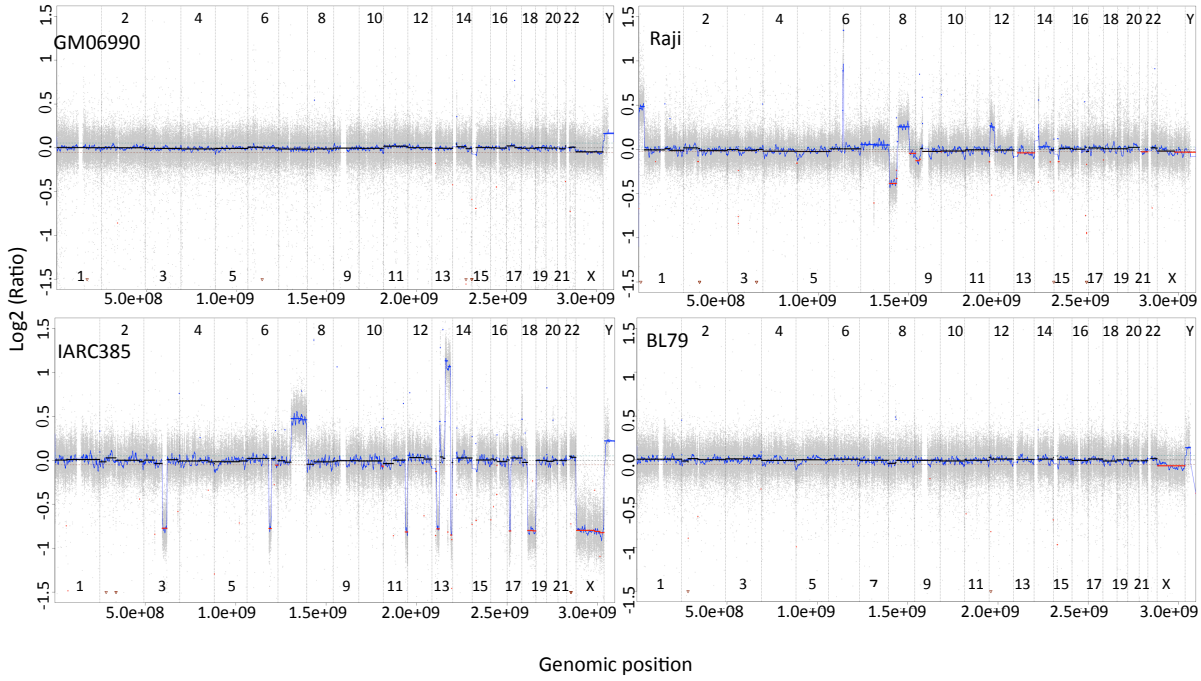


Figure S3. CGH array analysis of GM06990, Raji, BL79, and IARC385. The log2 ratio of copy number is plotted along the length of each chromosome. Red and blue horizontal segments indicate copy number losses and gains, respectively.

Methods : Total genomic DNA was extracted from 3-5x10^6 exponentially growing GM06990, Raji, IARC385 or BL79 cells by using QIAamp DNA mini Kit (Qiagen) according to the manufacturer's manuals. In all experiments, sex-matched normal DNA from a pooled human female or male (Promega, Madison, WI, USA) was used as a reference. Oligonucleotide aCGH processing was performed as detailed in the manufacturer's protocol (version 7.5; http://www.agilent.com). Equal amounts (500 ng) of tumour and normal DNAs were fragmented with AluI and RsaI (Fermentas, Euromedex, France). The fragmented DNAs were labelled with cyanine Cy3-deoxyuridine triphosphate (dUTP) or Cy5-dUTP. Hybridization was carried out on SurePrint G3 Human CGH Microarray 4x180K (Agilent Technologies, Santa Clara, CA, USA) arrays for 40 h at 65°C in a rotating oven (Robbins Scientific, Mountain View, CA) at 20 rpm. The hybridization was followed by appropriate washing steps. Scanning of glass microarrays was performed with an Agilent G2505C DNA Microarray scanner at 100% PMT with 3 µm resolution at 20°C in low ozone concentration environment. Data were extracted from scanned TIFF images using the Feature Extraction software (v11.5.1.1, Agilent), along with protocol CGH_1105_Oct12. All further data manipulation were performed under the R statistical environment in v3.4 (http://cran.r-project.org). Acquired raw intensities were normalized by subtracting the respective labelling dye differences, and pre-computed local GC% content through a lowess regression. Test and reference intensities were then combined as log2(test/ref) and segmented using the CBS [refCBS] algorithm implementation from the 'DNAcopy' package (v.1.50.1) with default parameters. Centering of profiles was performed using a in-house method selecting the most-centered mode in the distribution density of probes' log2(ratio) values. Aberration status calling was automatically performed for each profile according to its internal noise (measured as the half of the median absolute variation of log2(ratio) values across consecutive probes on the genome).

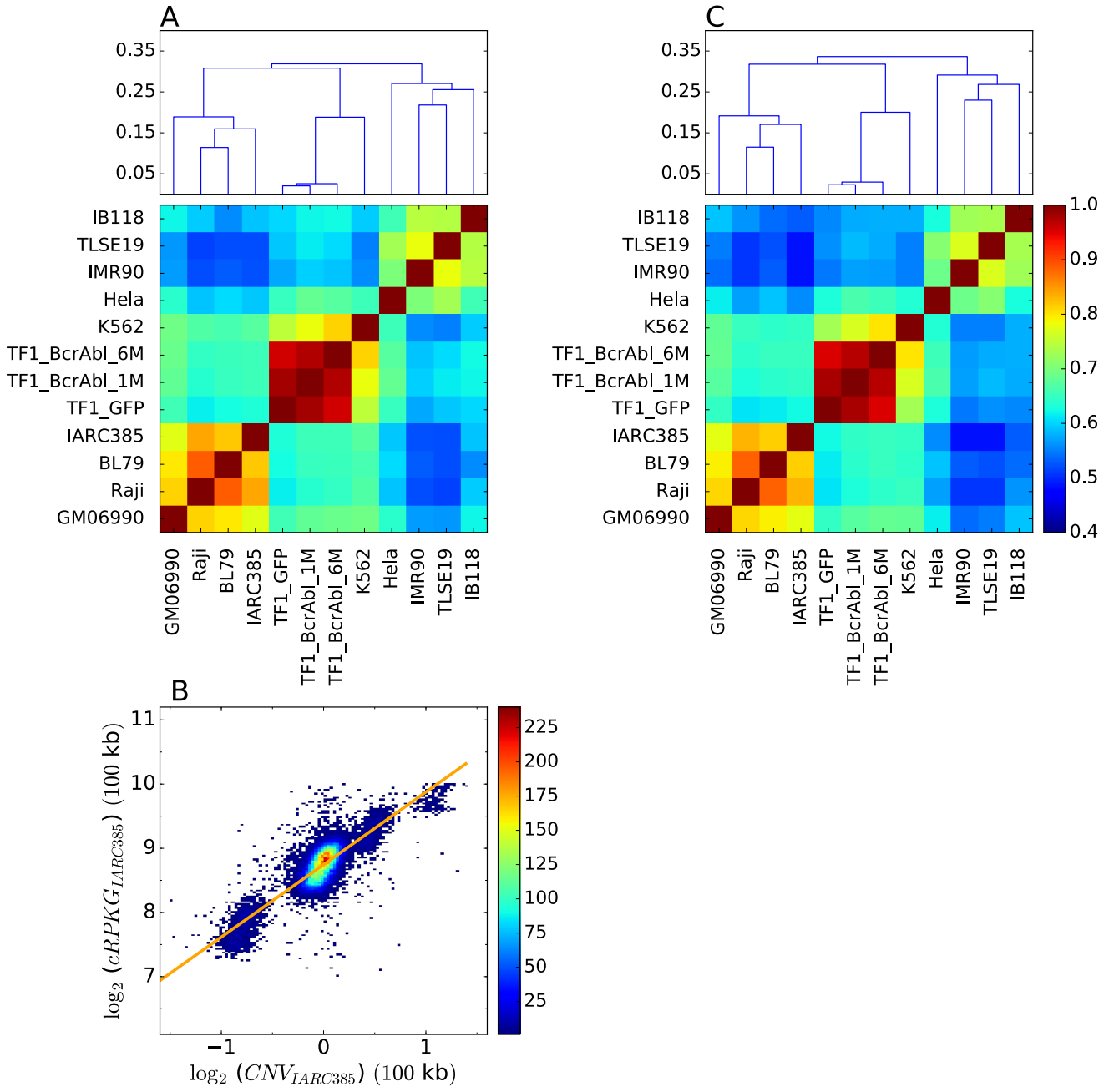
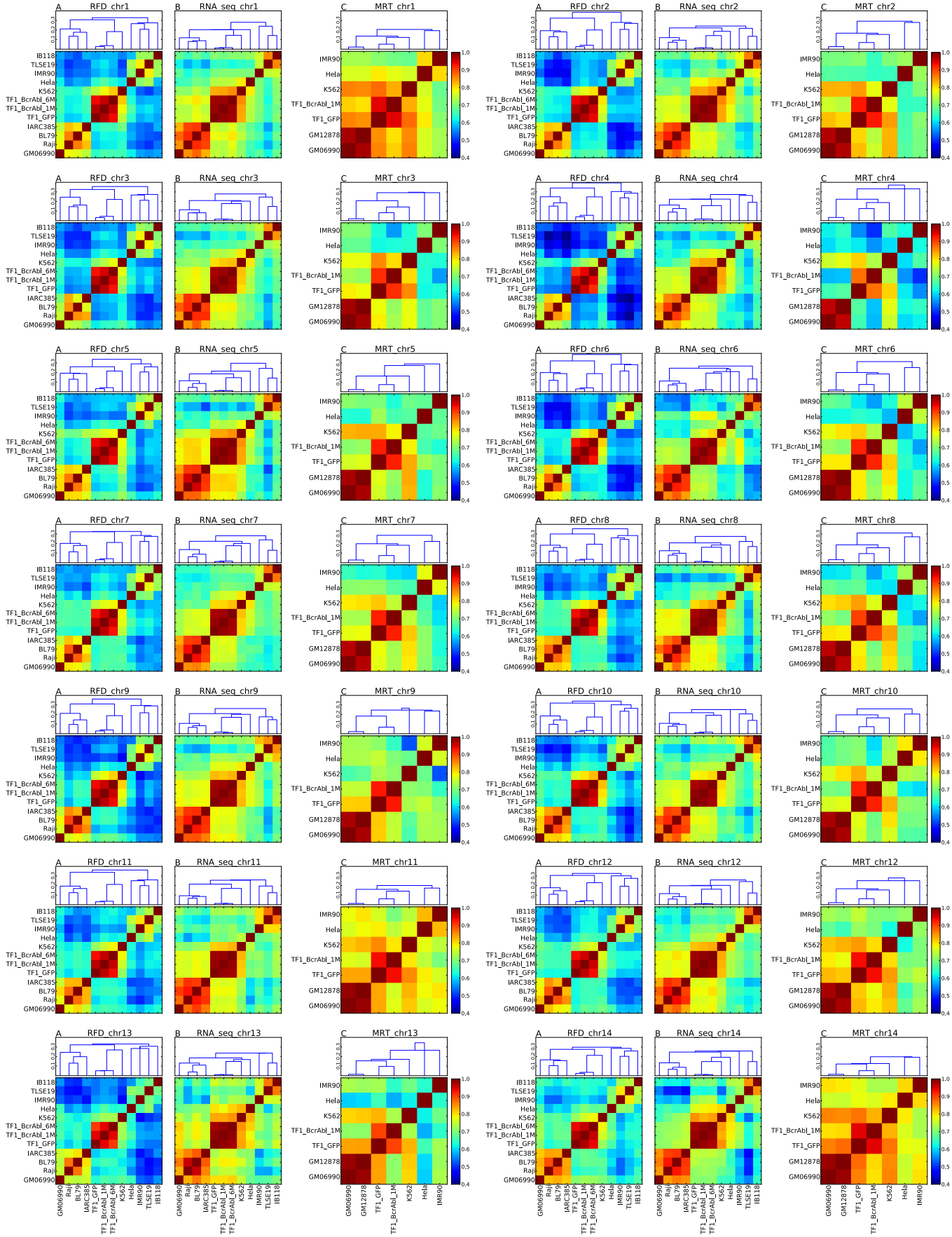


Figure S4. RFD correlation analysis is robust to CNVs determined by CGH arrays and OK-seq coverage. (A) Same analysis as in Figure 2A after filtering out aneuploid regions determined by CGH array analysis in lymphoid cell lines (Figure S3). (B) Correlation of OK-seq coverage normalized by mappability with log2 ratio of copy number determined by CGH array signal in IARC385. The orange line is a linear regression fit. (C) Same analysis as in (A) after further filtering out aneuploid regions in TF1 cell lines based on OK-seq coverage.


A RFD_chr1
B RNA_seq_chr1
C MRT_chr1
A RFD_chr2
B RNA_seq_chr2
C MRT_chr2
A RFD_chr3
B RNA_seq_chr3
C MRT_chr3
A RFD_chr4
B RNA_seq_chr4
C MRT_chr4
A RFD_chr5
B RNA_seq_chr5
C MRT_chr5
A RFD_chr6
B RNA_seq_chr6
C MRT_chr6
A RFD_chr7
B RNA_seq_chr7
C MRT_chr7
A RFD_chr8
B RNA_seq_chr8
C MRT_chr8
A RFD_chr9
B RNA_seq_chr9
C MRT_chr9
A RFD_chr10
B RNA_seq_chr10
C MRT_chr10
A RFD_chr11
B RNA_seq_chr11
C MRT_chr11
A RFD_chr12
B RNA_seq_chr12
C MRT_chr12
A RFD_chr13
B RNA_seq_chr13
C MRT_chr13
A RFD_chr14
B RNA_seq_chr14
C MRT_chr14
IB118
TLSE19
IMR90
Hela
K562
TF1_BcrAbl_6M
TF1_BcrAbl_1M
TF1_GFP
IARC385
BL79
Raji
GM06990
GM12878
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.1
0.2
0.3

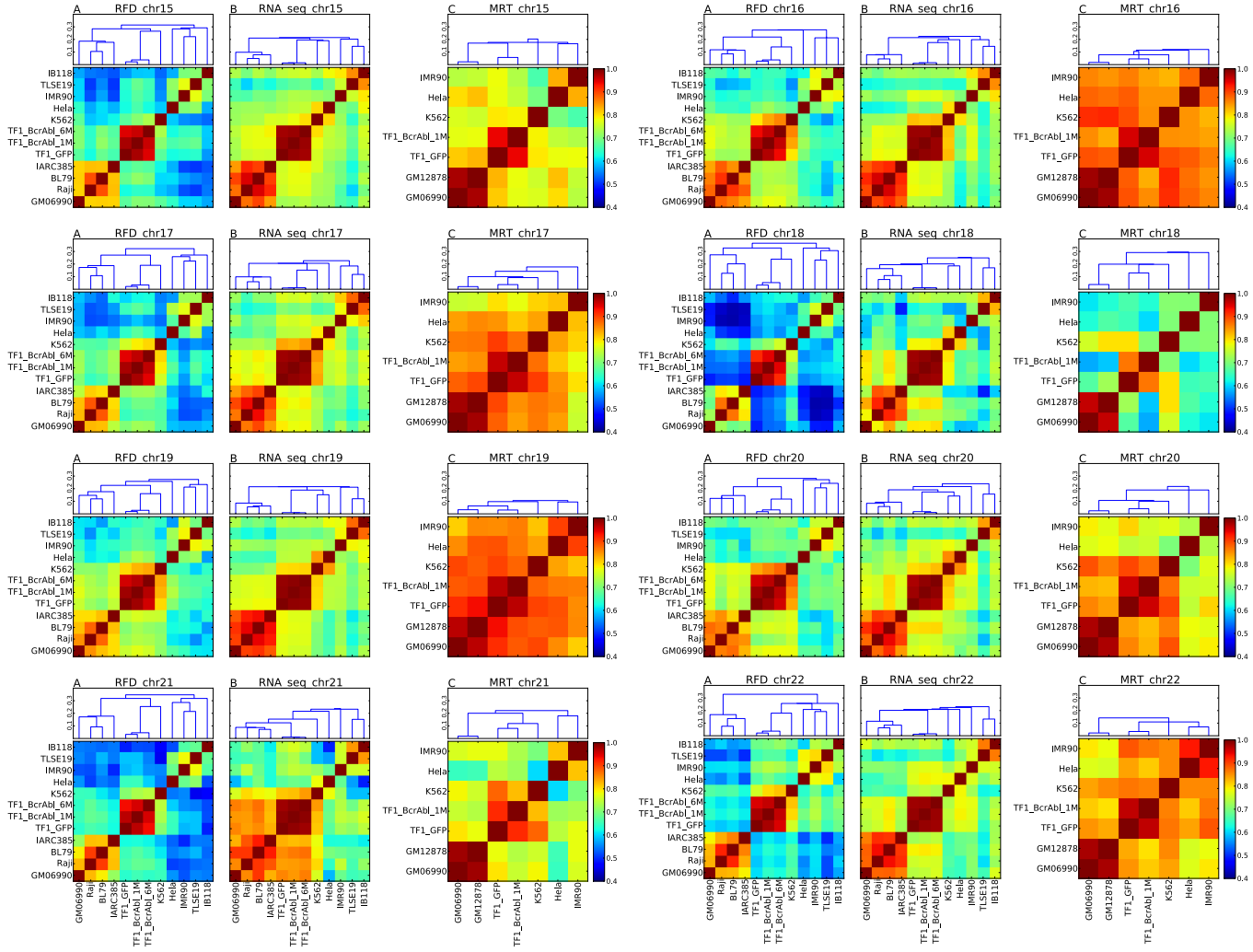


Figure S5. Cell line classification based on correlations between replication and gene expression profiles for each chromosome. Correlation matrices between RFD profiles (C_{RFD} ; A), RNA-seq ($C_{RNA-seq}$; B) and MRT profiles (C_{MRT} ; C); Pearson correlation coefficient values are colour-coded from blue (0.4) to red (1) using the colour bar on the right. A corresponding dendrogram representation of the hierarchical classification of cell lines is shown on top of each correlation matrix; ordinate is the correlation distance.

The results obtained for the entire genome were recapitulated for each chromosome, with the following minor exceptions. Lymphoid, myeloid and adherent cells formed three separate RFD clusters, except that for chromosome 19, HeLa clustered with myeloid rather than adherent cells, and for chromosome 16, HeLa was equally distant from both groups. Distances within and between groups were conserved, except that for chromosome 22, myeloid cells were closer to adherent than to lymphoid cells, and for chromosome 17, IARC385, rather than GM06990, was the most distant among lymphoid cells. As to RNA-seq data, the clustering of HeLa with myeloid rather than adherent cells was again observed for each chromosome, although for chromosomes 10 and 11 HeLa appeared equidistant from myeloid and adherent cells. In addition, IMR90 clustered with HeLa and myeloid cells rather than adherent cells for chromosomes 6 and 20, and was equidistant from both groups for chromosomes 13, 14, 18 and 21. As to MRT data, the global classification was reproduced for each chromosome except that the branching point of HeLa and K562 was somewhat unstable.

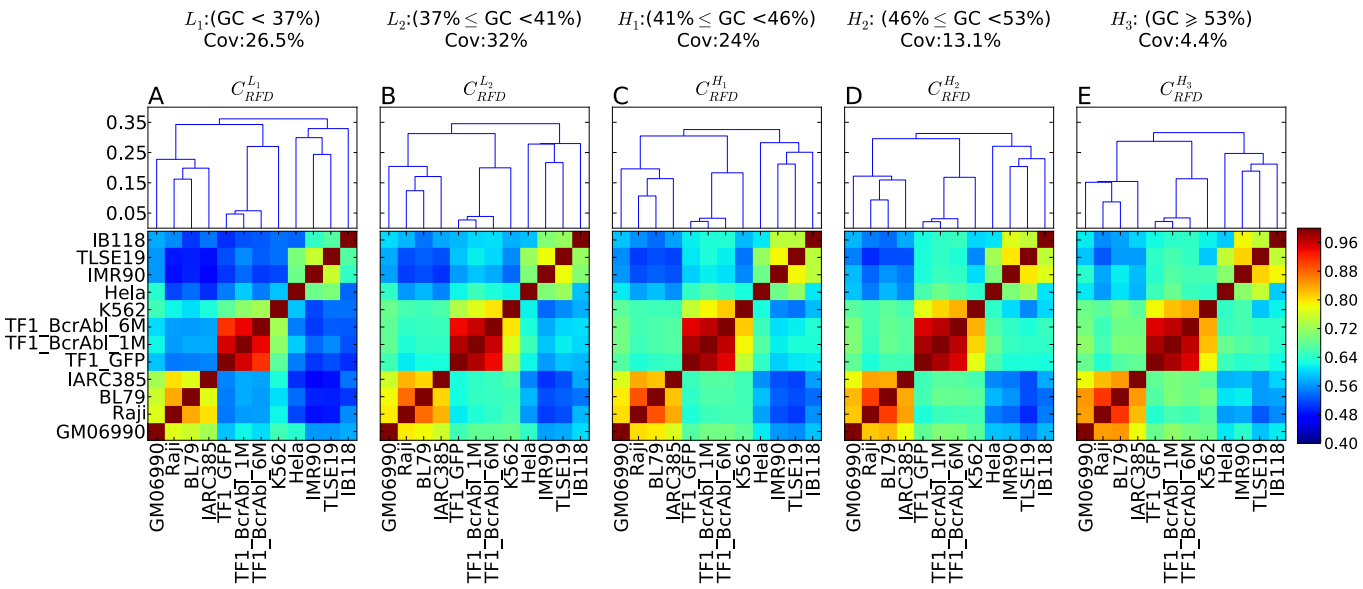


Figure S6. Hierarchical clustering of the 12 RFD profiles depending on GC-content. Dendrogram representations of the hierarchical classification of cell lines is shown on top of the same correlation matrices as in Figure 4A-E.

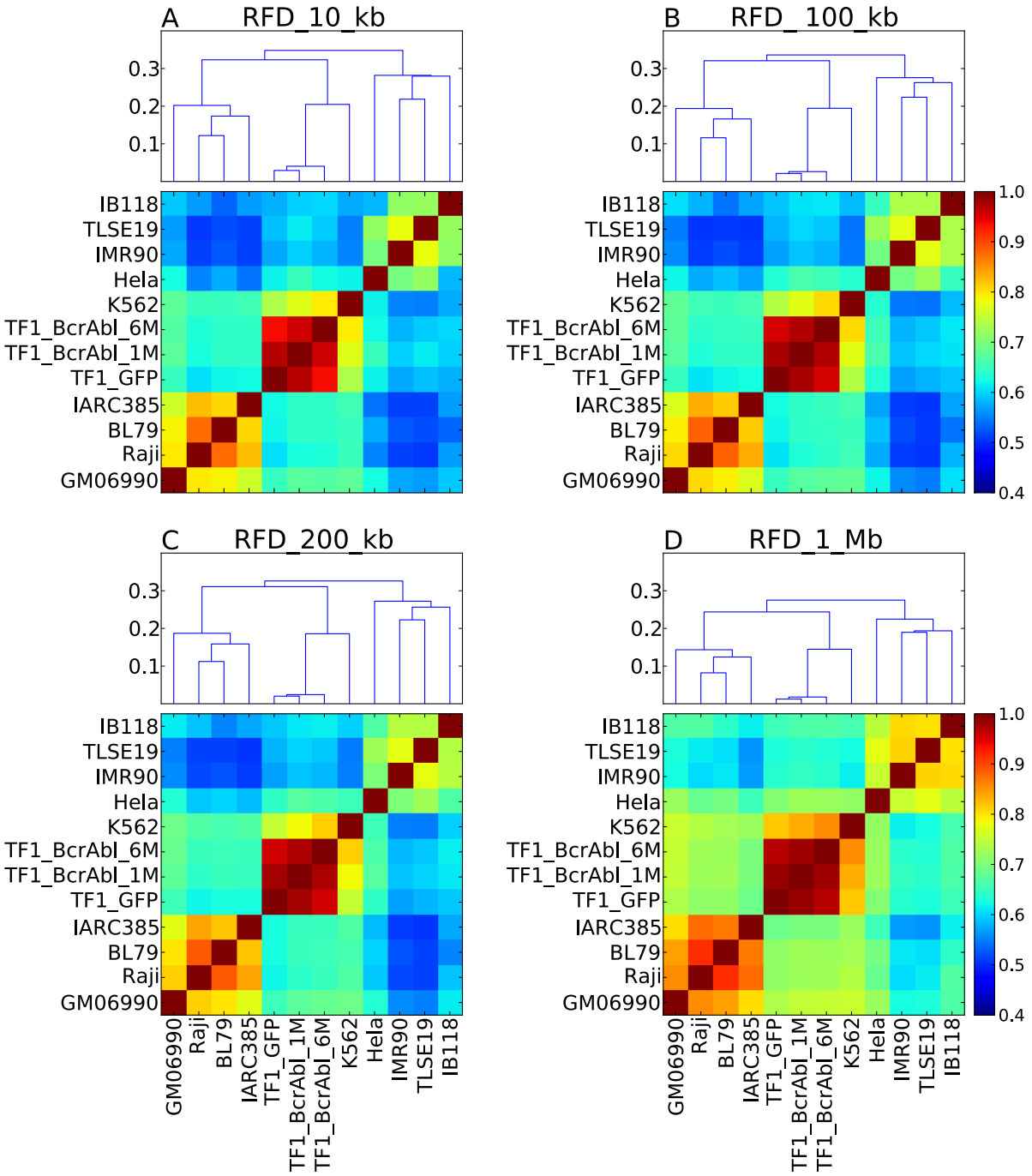


Figure S7. Analysis of the scale dependence of the cell line classification based on RFD profiles. Correlation analysis and hierarchical classification of RFD profiles was performed as in Figure 2A at 10kb, 100 kb, 200 kb and 1 Mb scales. The same classification of cell lines was obtained at all scales.

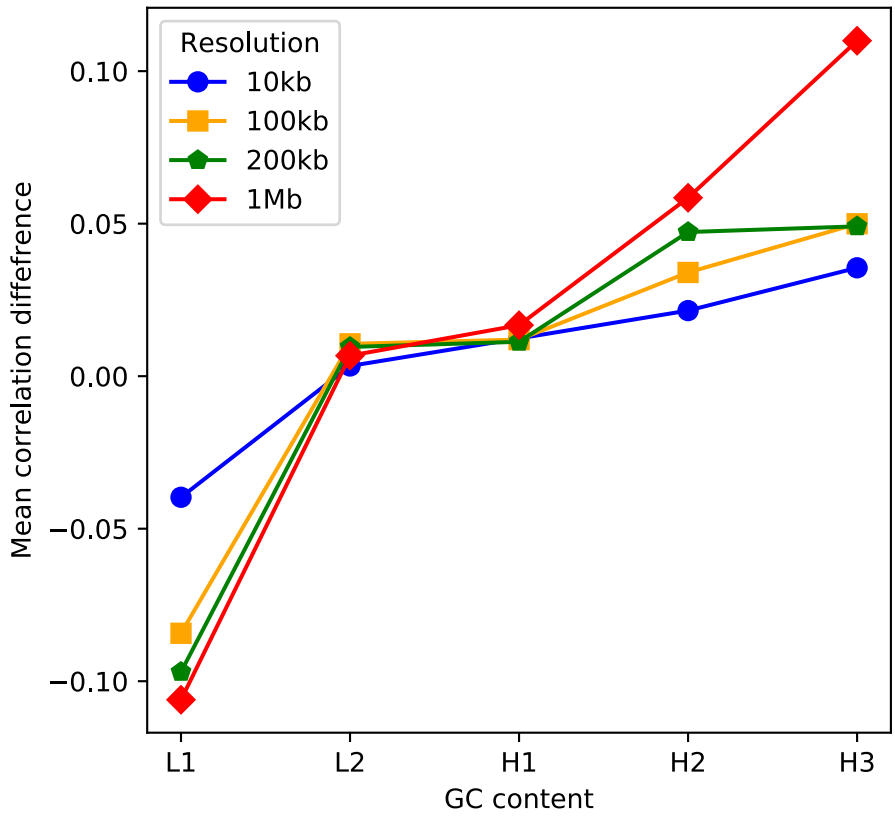


Figure S8. Analyses of the scale dependence of the mean correlation difference between isochore- and global genome-based RFD correlation matrices. The mean values of correlation differences $\Delta C^{I}_{RFD} = C^{I}_{RFD} - C_{RFD}$ (where I can be L1, L2, H1, H2 or H3) were computed for each isochore as in Figure 4F-J at 10kb, 100 kb, 200 kb and 1 Mb scales. The mean correlation differences increase with GC-content at all analysis scales.

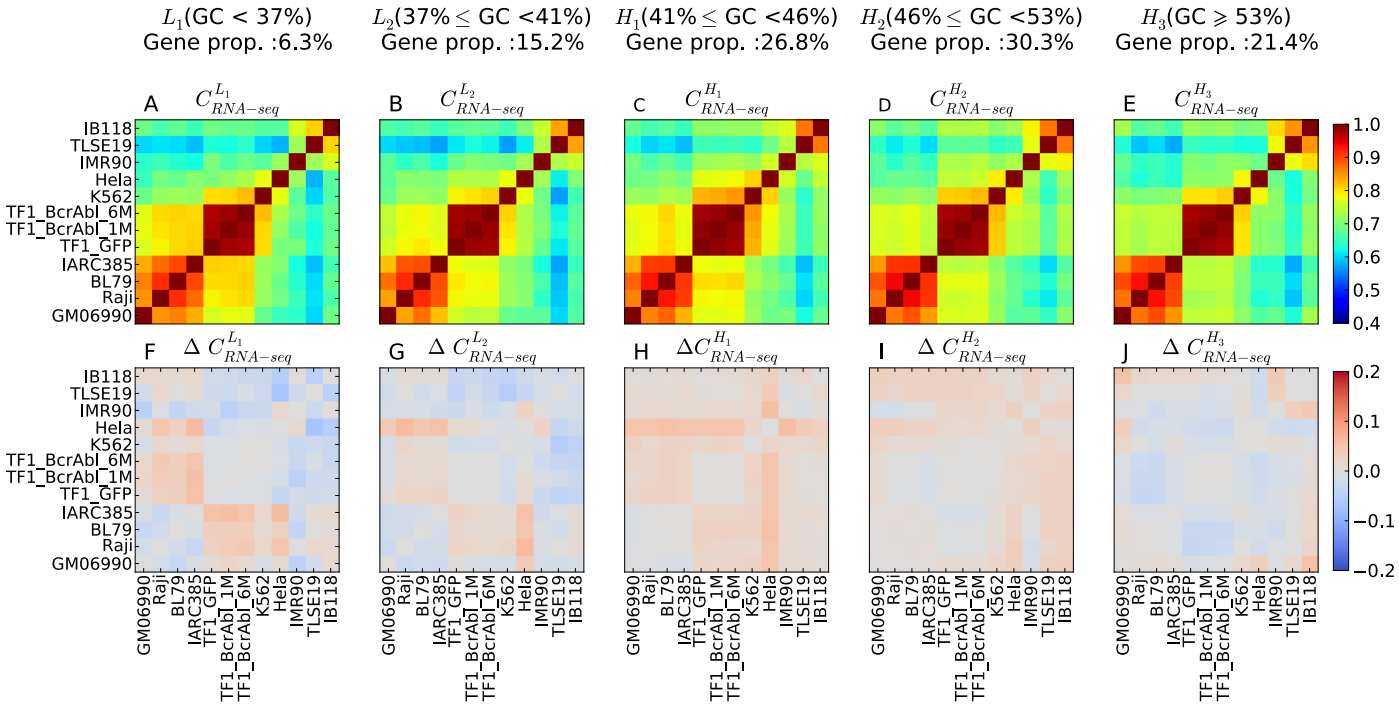


Figure S9. Transcription program conservation for regions with different GC-content. Same as Figure 4 but using RNA-seq data instead of RFD data.

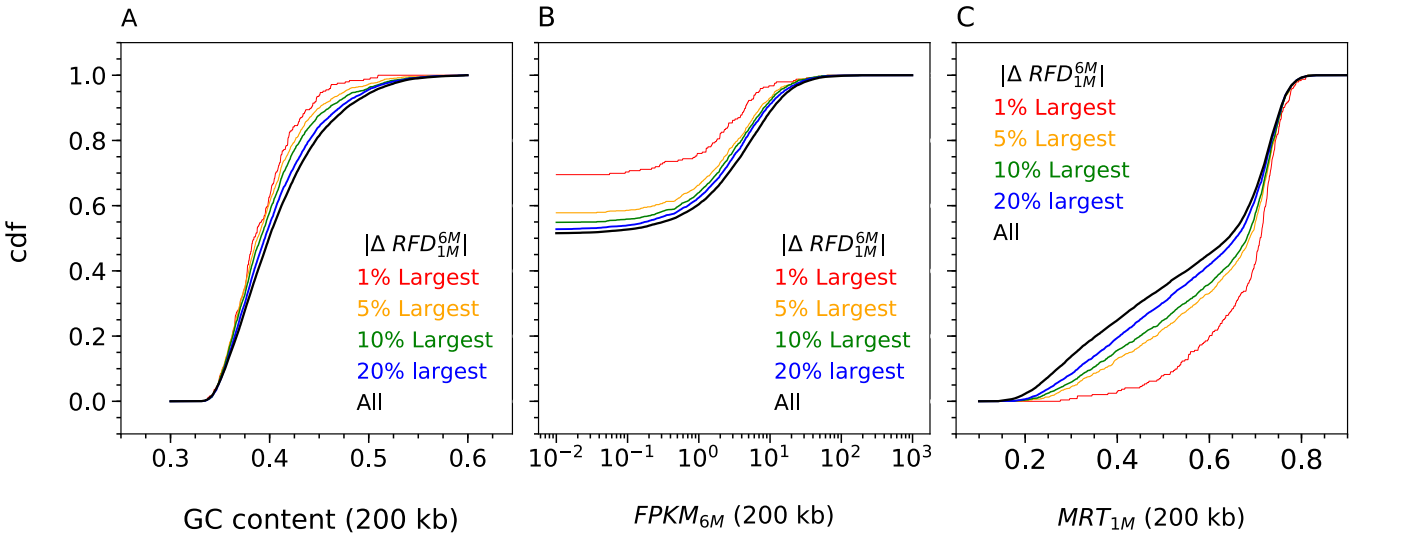


Figure S10. The largest RFD changes between 1 month and 6 months of BCR-ABL1 expression are observed in GC-poor, lowly-expressed and late replicating regions. Same as Figure 5 but for the largest RFD changes between TF1-BCRABL-1M and TF1-BCRABL-6M.

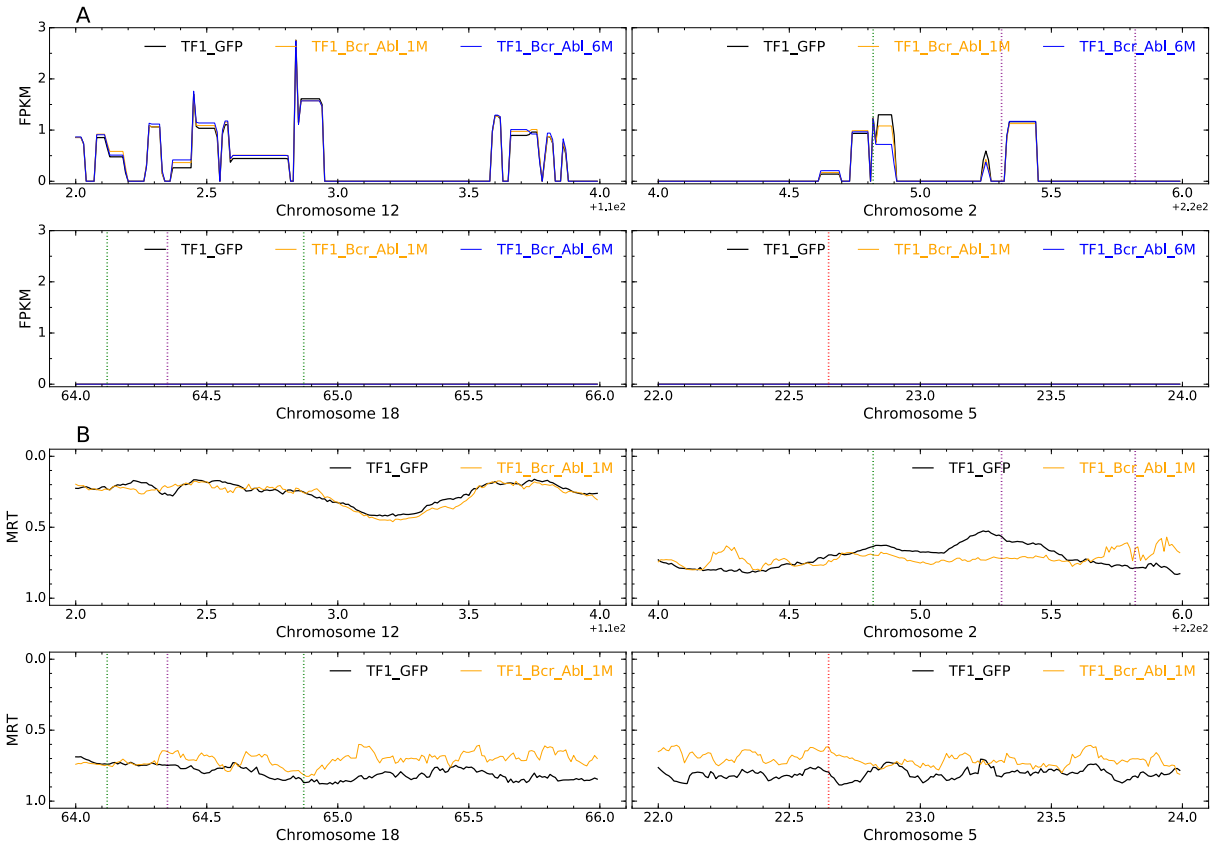


Figure S11. RNA-seq (A) and MRT (B) profiles of the same 2Mb regions of the same cell lines as in Figure 6.

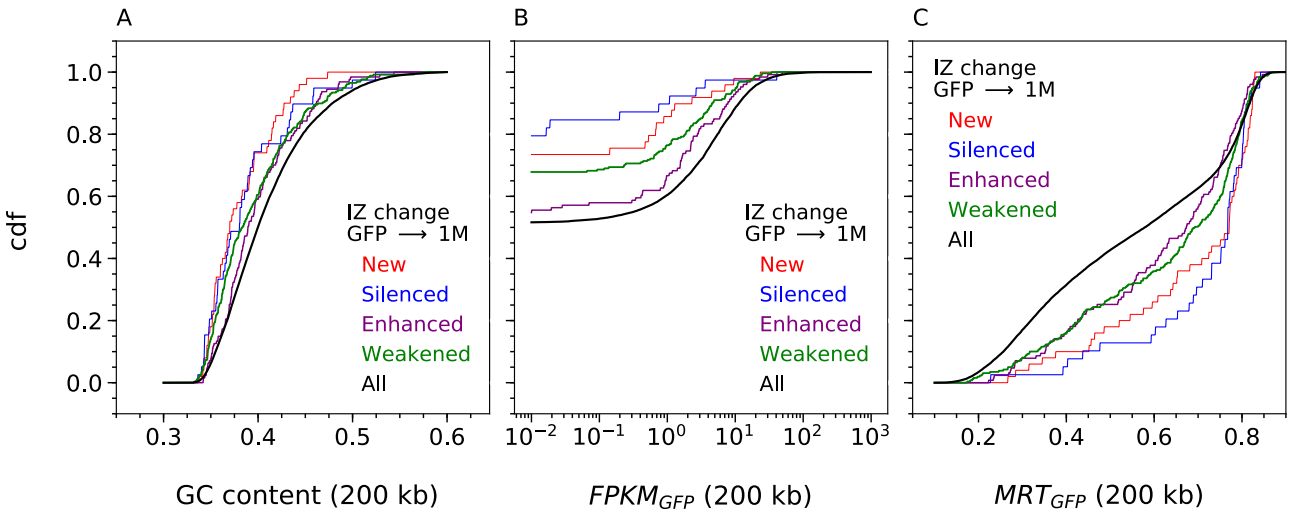


Figure S12. Initiation zones efficiency changes in response to 1 month of *BCR-ABL1* expression are observed in GC-poor, lowly-transcribed and late replicating regions. Same as Figure 8 but using transcription (B) and MRT (C) data from TF1-GFP instead of TF1-BCRABL-1M.

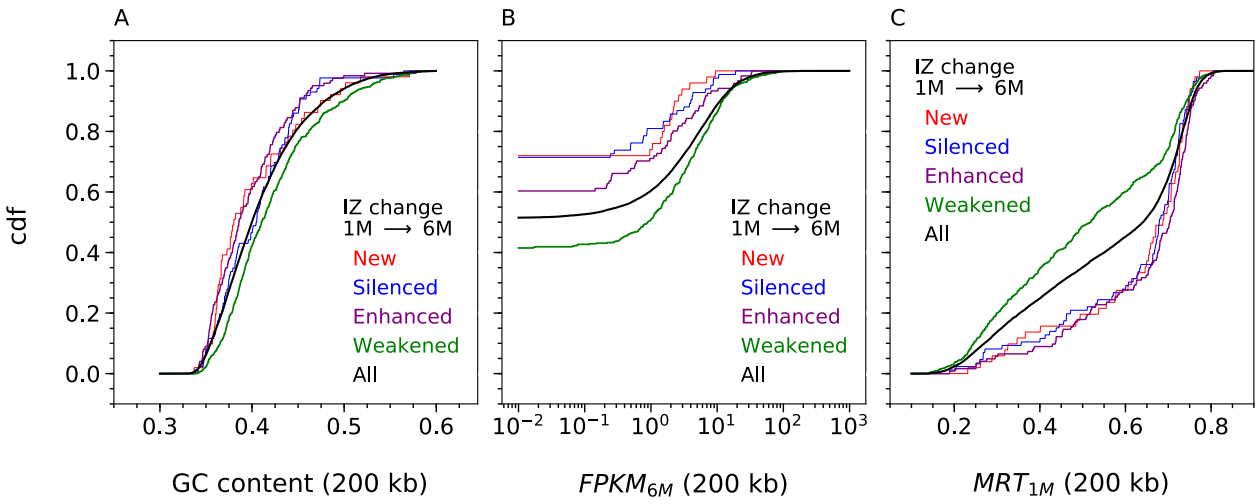


Figure S13. Initiation zones efficiency changes between 1 month and 6 months of *BCR-ABL1* expression are observed in GC-poor, lowly-transcribed and late-replicating regions, except for weakened IZs which show the opposite tendency. Cumulative distribution functions (cdf) of GC content (A), transcription in TF1-BCRABL-6M (FPKM6M) (B) and MRT in TF1-BCRABL-1M (C) computed in non-overlapping 200 kb windows of the 22 autosomes (Materials and Methods). Cdfs were determined for all windows (all, black) or limited to windows with Silenced (blue), Weakened (green), Enhanced (violet) and New (red) IZ between 1 month and 6 months BCR-ABL1 expression.

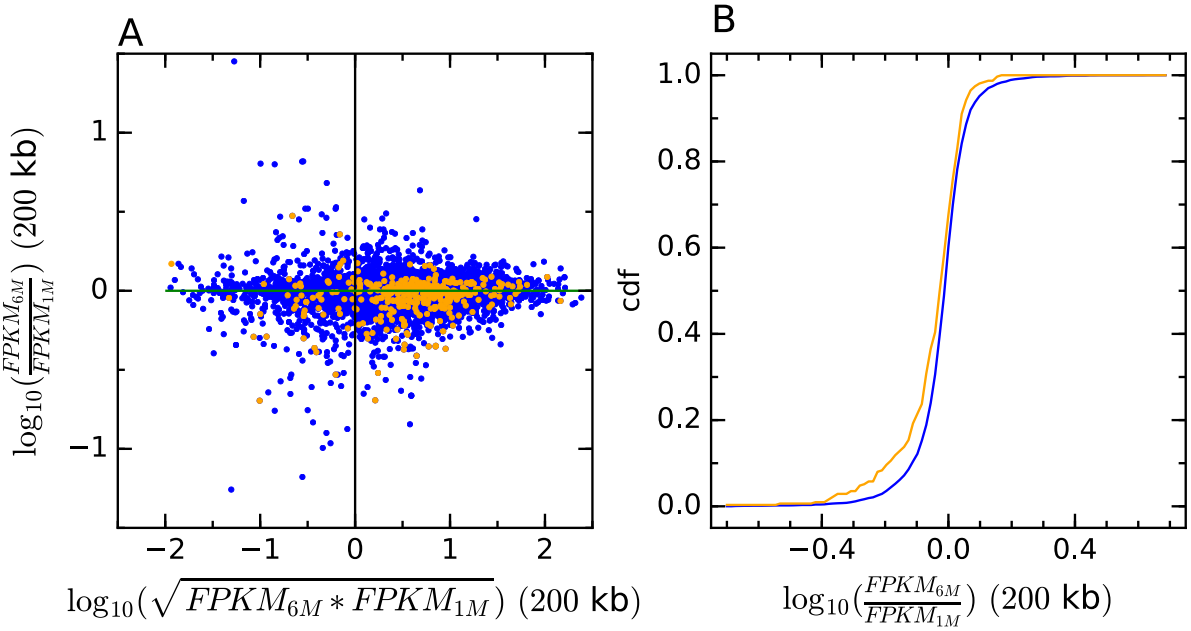


Figure S14. Weakened initiation zones between 1 month and 6 months of BCR-ABL1 expression are associated with transcription repression. (A) Log_{10}-ratio of FPKM values in TF1-BCRABL-6M over TF1-BCRABL-1M as a function of their log_{10} (geometric) mean. FPKM values were computed in non-overlapping 200 kb windows. Only windows expressed in both cell lines were considered (FPKM > 0.01). Windows containing at least one weakened IZ at Step 2 are in *orange*, the rest is in *blue*. (B) Cdf of the FPKM log_{10}-ratios for windows with FPKM (geometric) mean > 0. *Blue* and *orange,* as in (A).

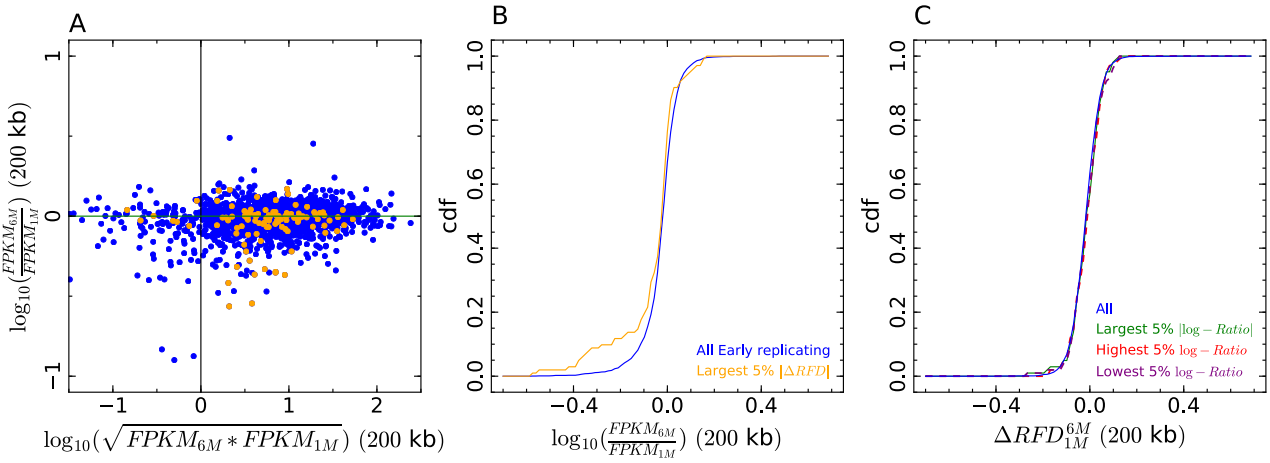


Figure S15. The largest RFD changes in early replicating regions between 1 month and 6 months of BCR-ABL1 expression are associated with transcription repression, but transcription changes do not predict RFD changes. (A,B), Same as Figure S14A,B but for all early replicating regions (MRT < 0: 33). The 5% with the largest RFD changes are in *orange*, the rest is in *blue*. (C) Cdf of ΔRFD_{6M}^{1M} between TF1-BCRABL-6M and TF1-BCRABL-1M for windows with FPKM (geometric) mean > 0. Results in green, red and purple correspond to the 5% of the regions plotted in blue with the largest absolute FPKM ratio,the largest FPKM ratio and the lowest FPKM ratio, respectively.

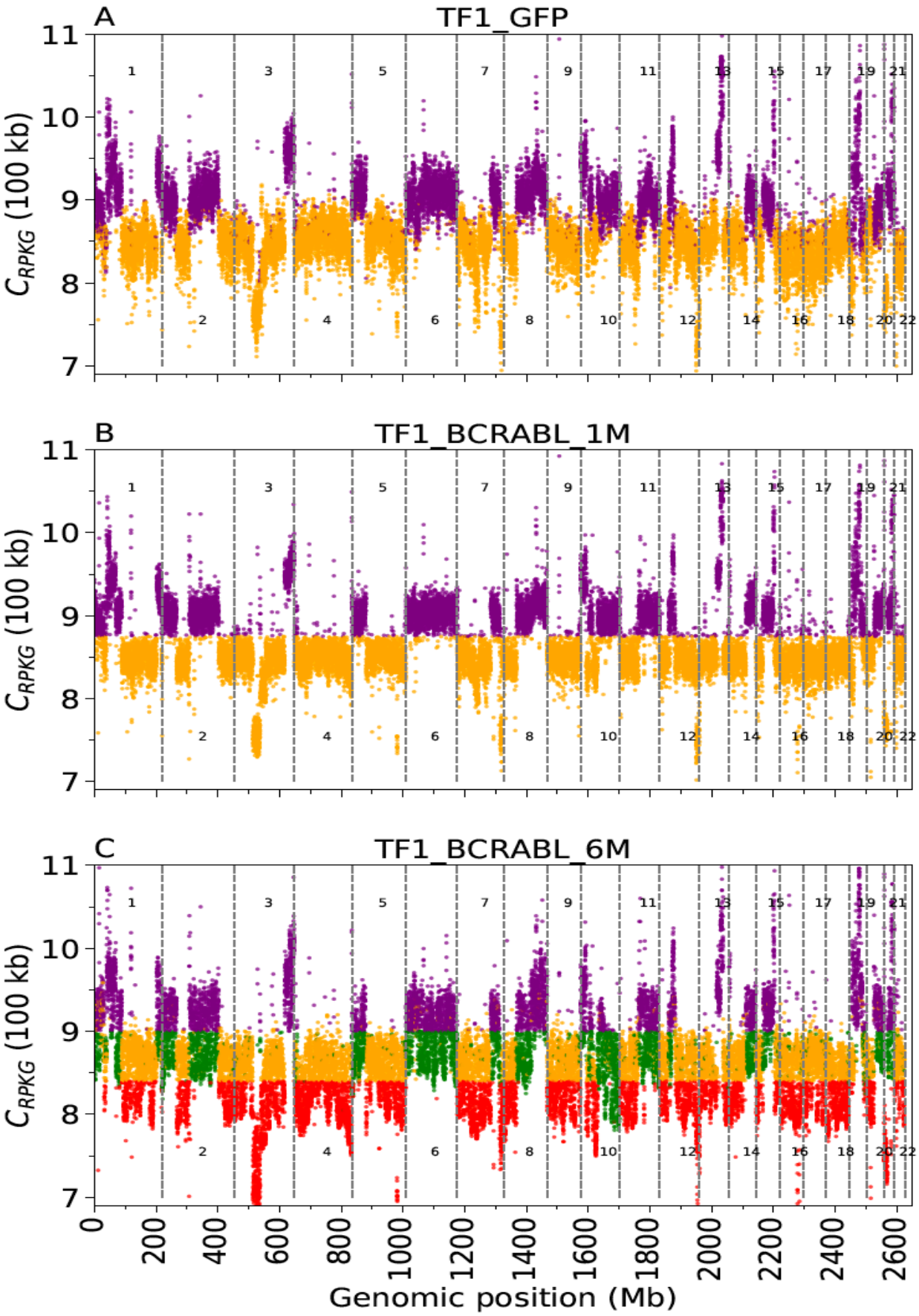


Figure S16. Copy number chromosomal profiles of TF1 cell lines determined by OK-seq coverage. OK-seq coverage (log2(cRPKG)) is plotted along the length of each chromosome. Stable and unstable diploid and triploid regions are colour-coded (by 100 kb bins) as in Figure 9.

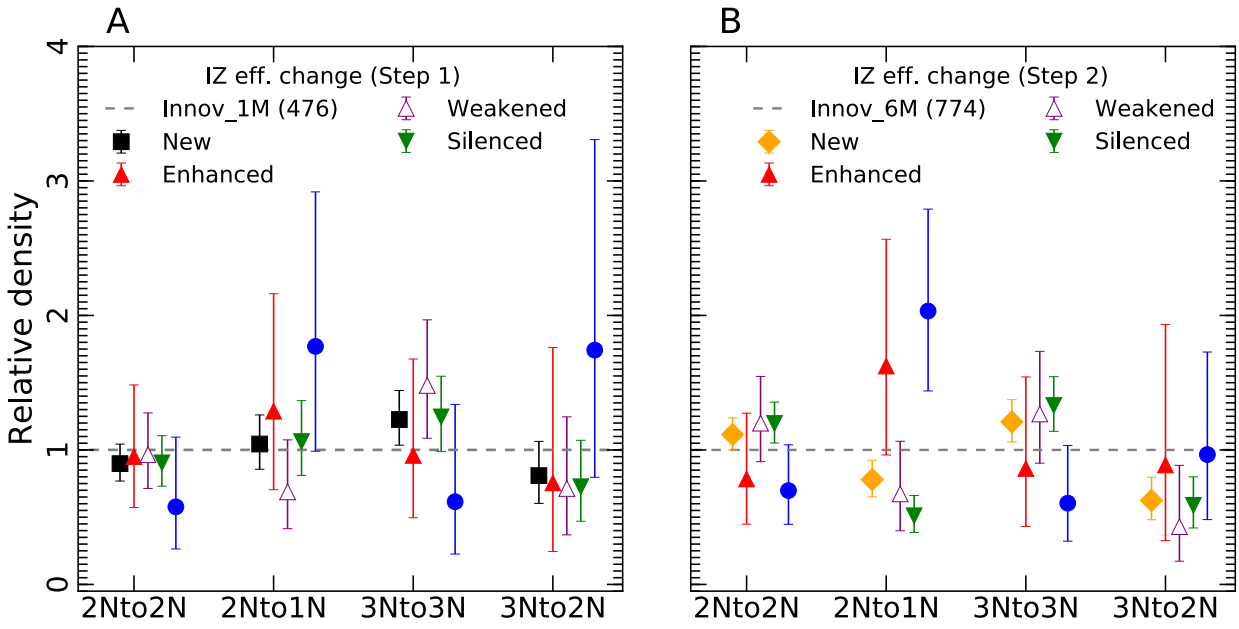


Figure S17. Relative densities of the different types of IZ efficiency changes at Step 1 (A) and Step 2 (B) of CML progression according to ploidy evolution at Step 2. Same as Figure 11C,D but after decomposing Stable and Unstable regions according to initial ploidy.

		TF1-BCRABL-1M			
		Active IZ		Inactive IZ	Total
TF1-GFP		Enhanced	Weakened	Silenced	
	Active IZ	127	260	39	426
		New			
	Inactive IZ	50		-	50
	Total	437		39	476

		TF1-BCRABL-6M			
		Active IZ		Inactive IZ	Total
TF1-BCRABL-1M		Enhanced	Weakened	Silenced	
	Active IZ	123	514	86	723
		New			
	Inactive IZ	51		-	51
	Total	688		86	774

		K562			
		Active IZ		Inactive IZ	Total
TF1-BCRABL-6M		Enhanced	Weakened	Silenced	
	Active IZ	157	230	315	702
		New			
	Inactive IZ	14		-	14
	Total	401		315	716

Table S1. Database of replication initiation zone efficiency changes. Summary of the manual annotation of RFD profiles for changes in IZ efficiency between TF1-GFP and TF1-BCRABL-1M (top), TF1-BCRABL-1M and TF1-BCRABL-6M (middle) and, TF1-BCRABL-6M and K562 cell lines (bottom). Note that for the latter case, only the 1027 loci presenting an IZ efficiency change in at least one of the first two comparisons were analysed. IZ change types (New, Enhanced, Weakened and Silenced) were organised to highlight the active or inactive status of IZ in each cell line.