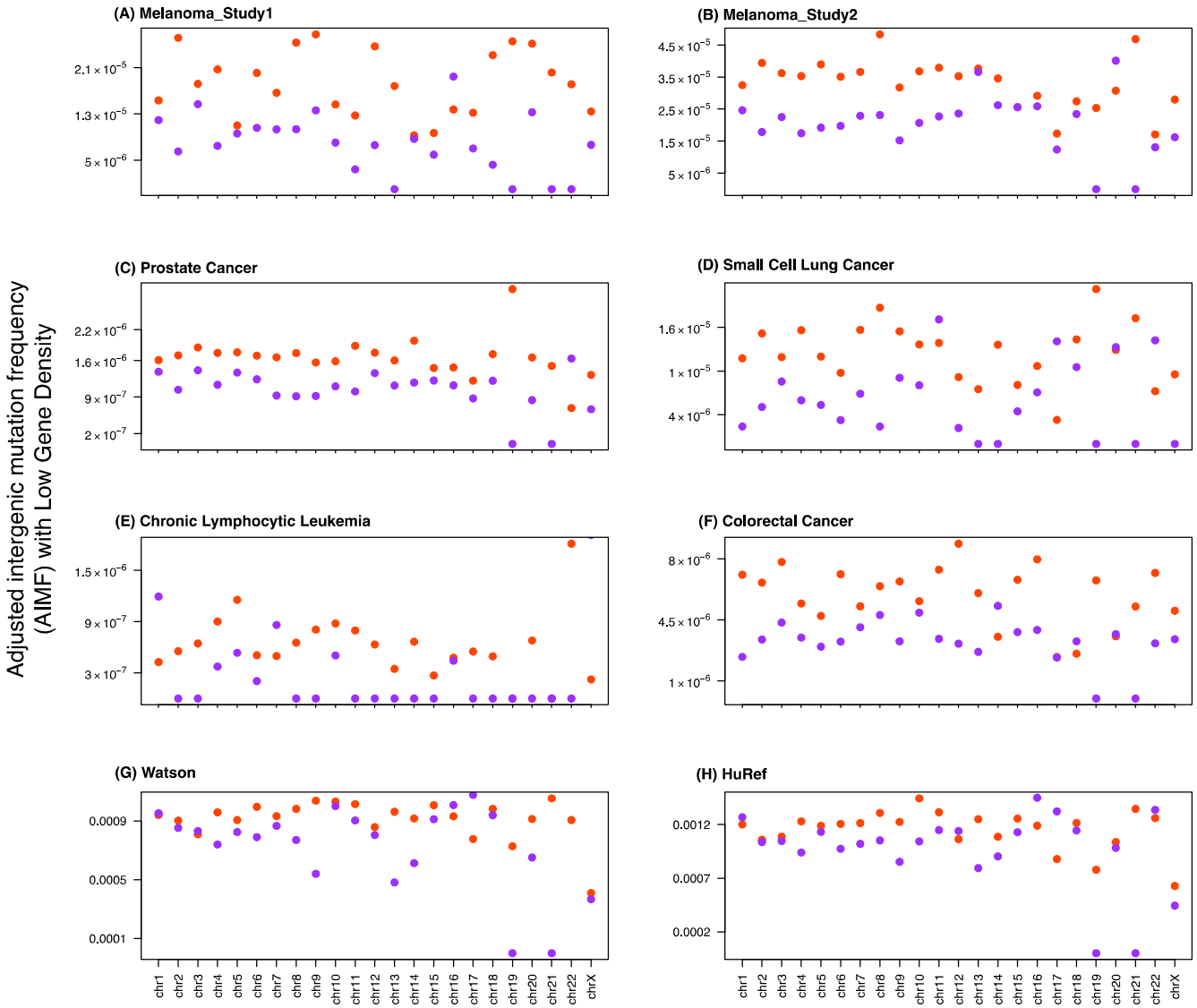
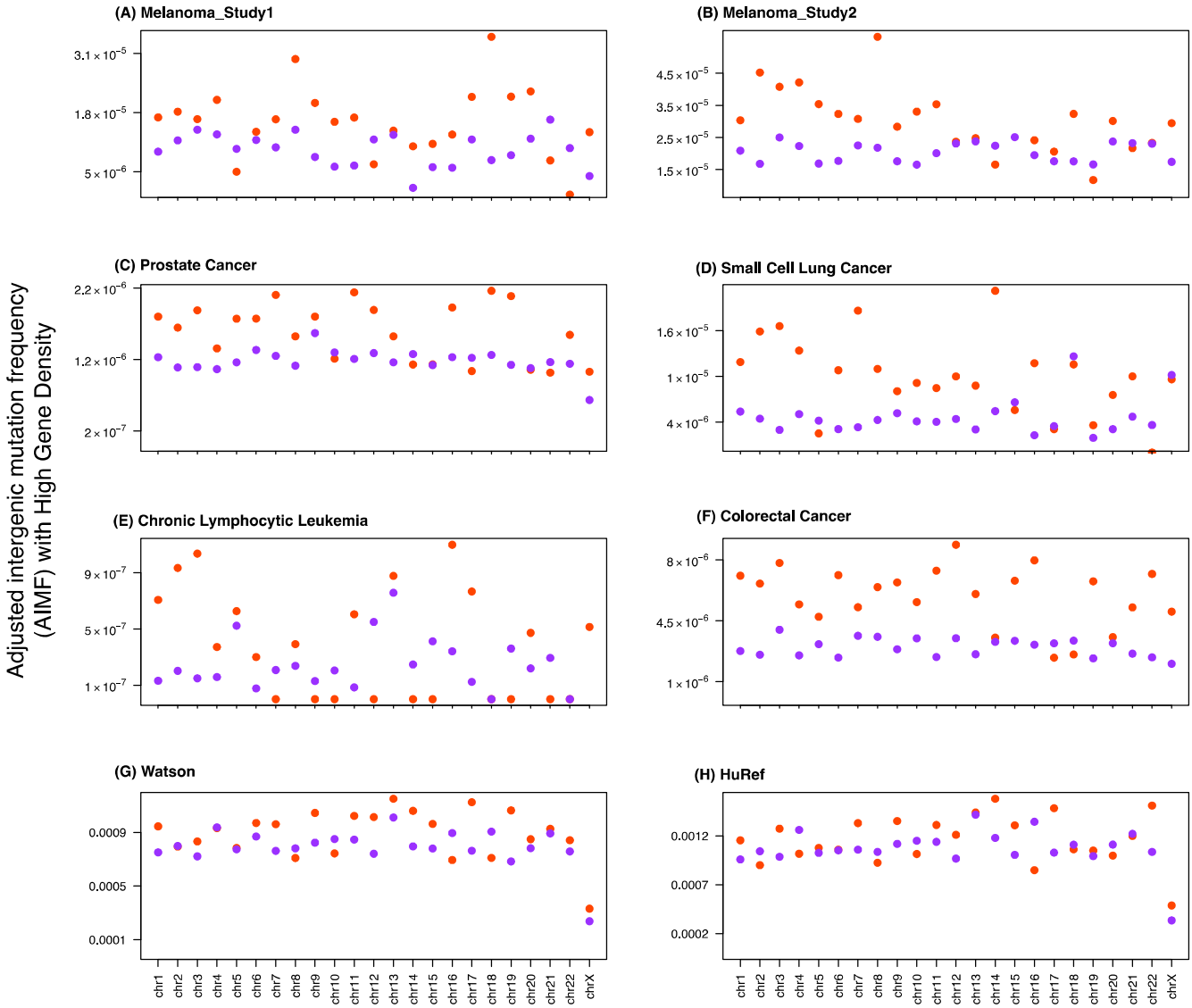
SUPPLEMENTARY FIGURES

Supplementary Figure S1: AIMF for constant late (orange) and constant early (purple) DNA replication timing zones for low gene density regions.



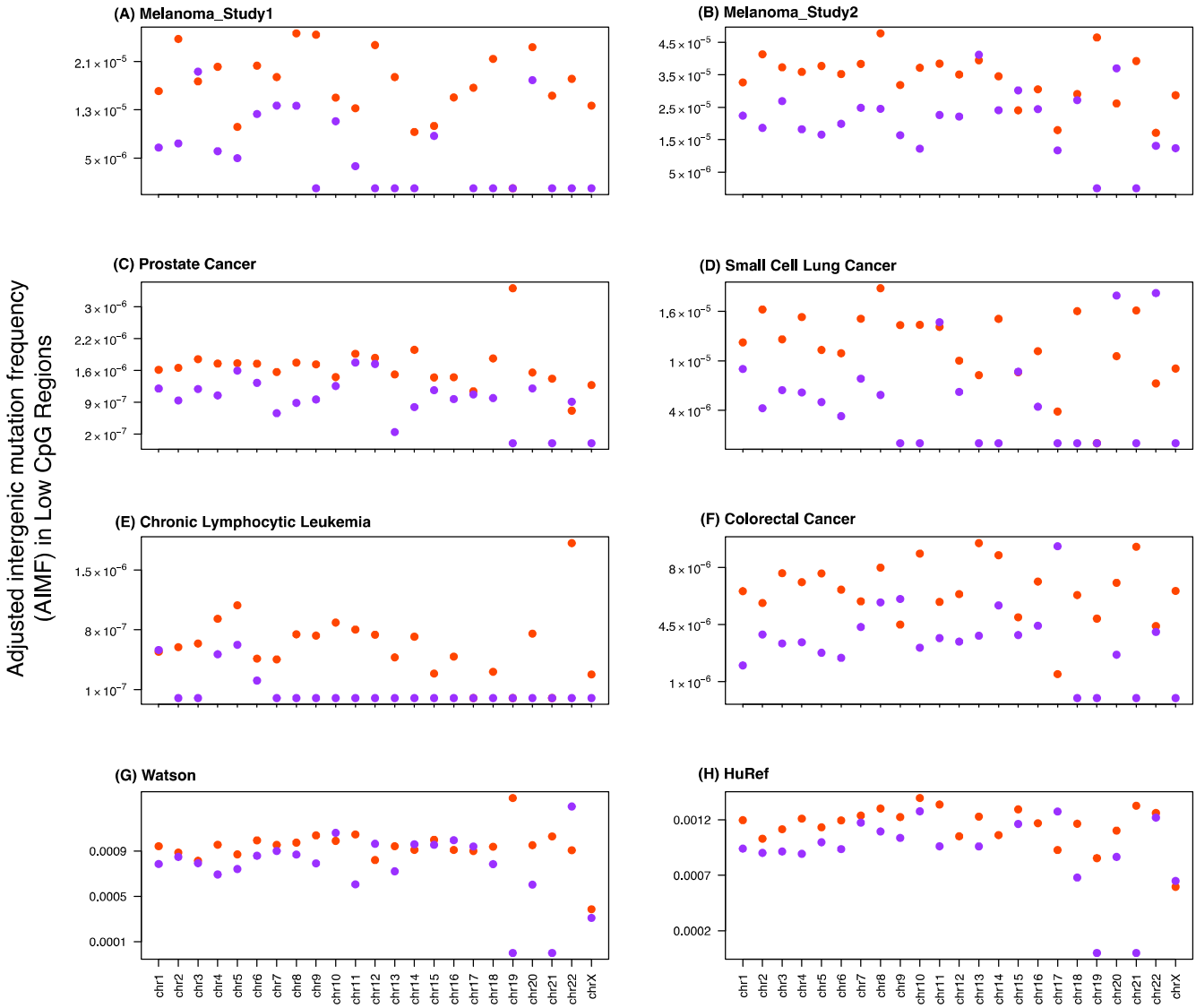
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S2: AIMF for constant late (orange) and constant early (purple) DNA replication timing zones for high gene density regions.



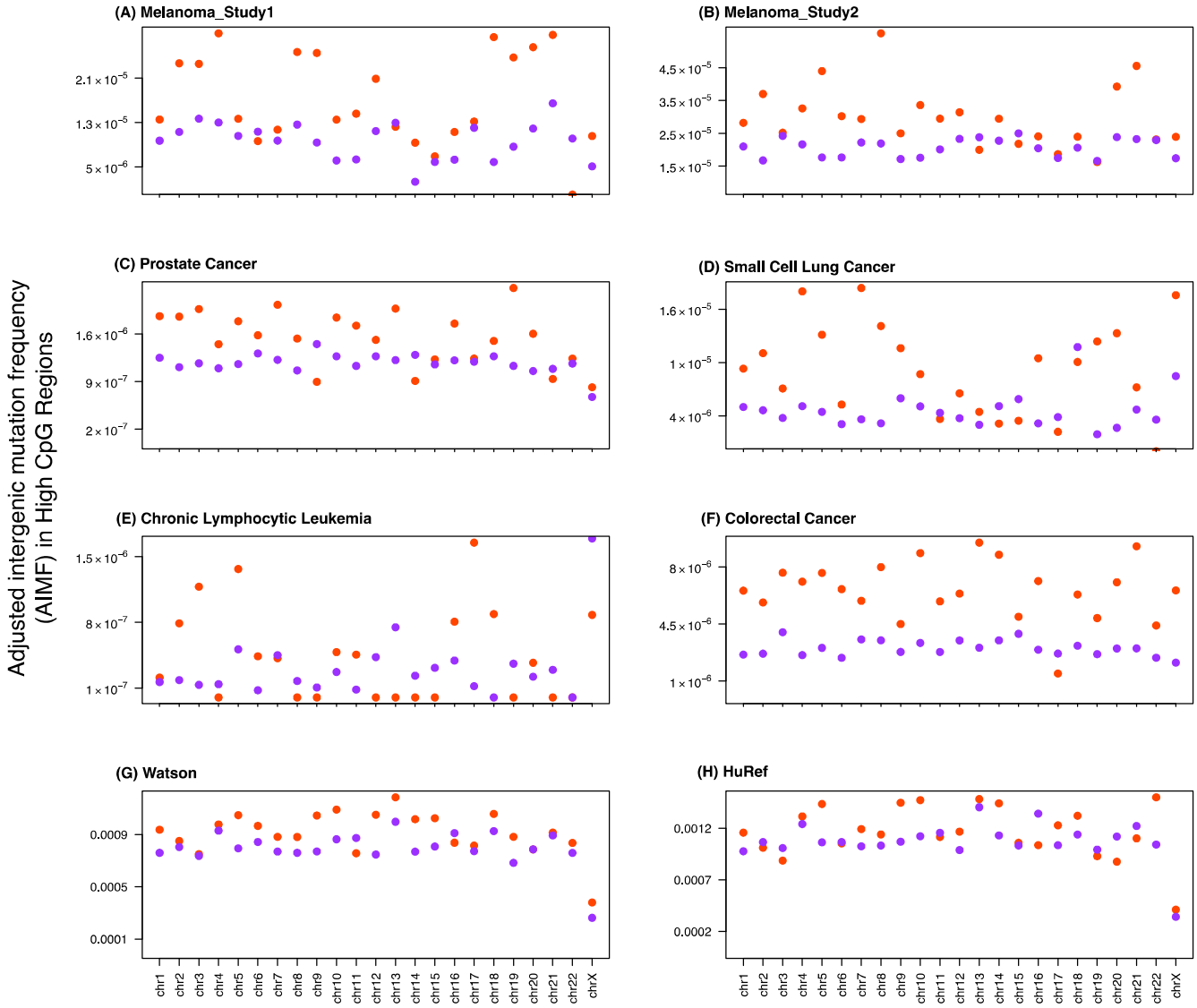
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S3: AIMF for constant late (orange) and constant early (purple) DNA replication timing zones for regions with a low fraction of bases in CpG islands.



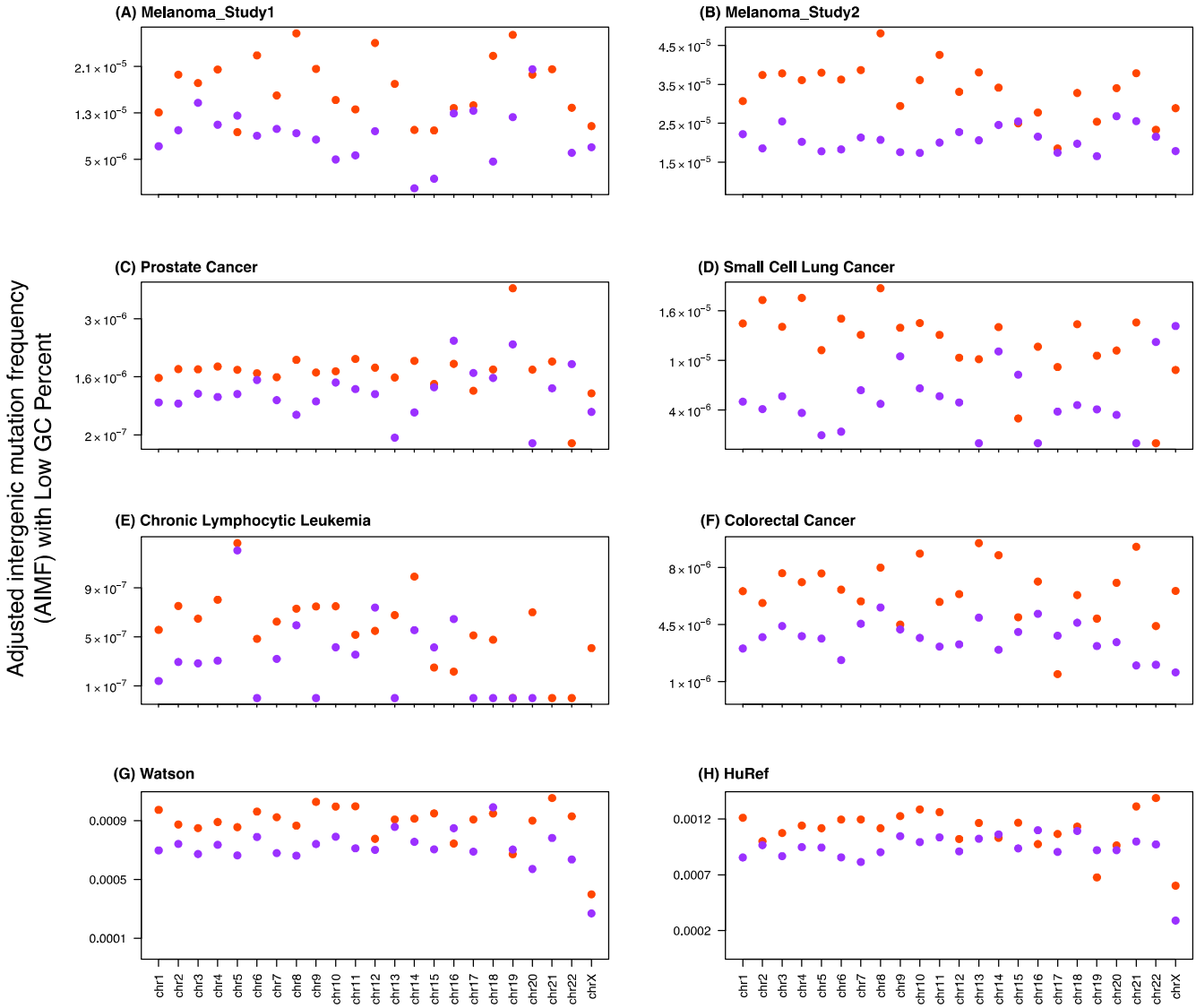
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S4: AIMF for constant late (orange) and constant early (purple) DNA replication timing zones for regions with a high fraction of bases in CpG islands.



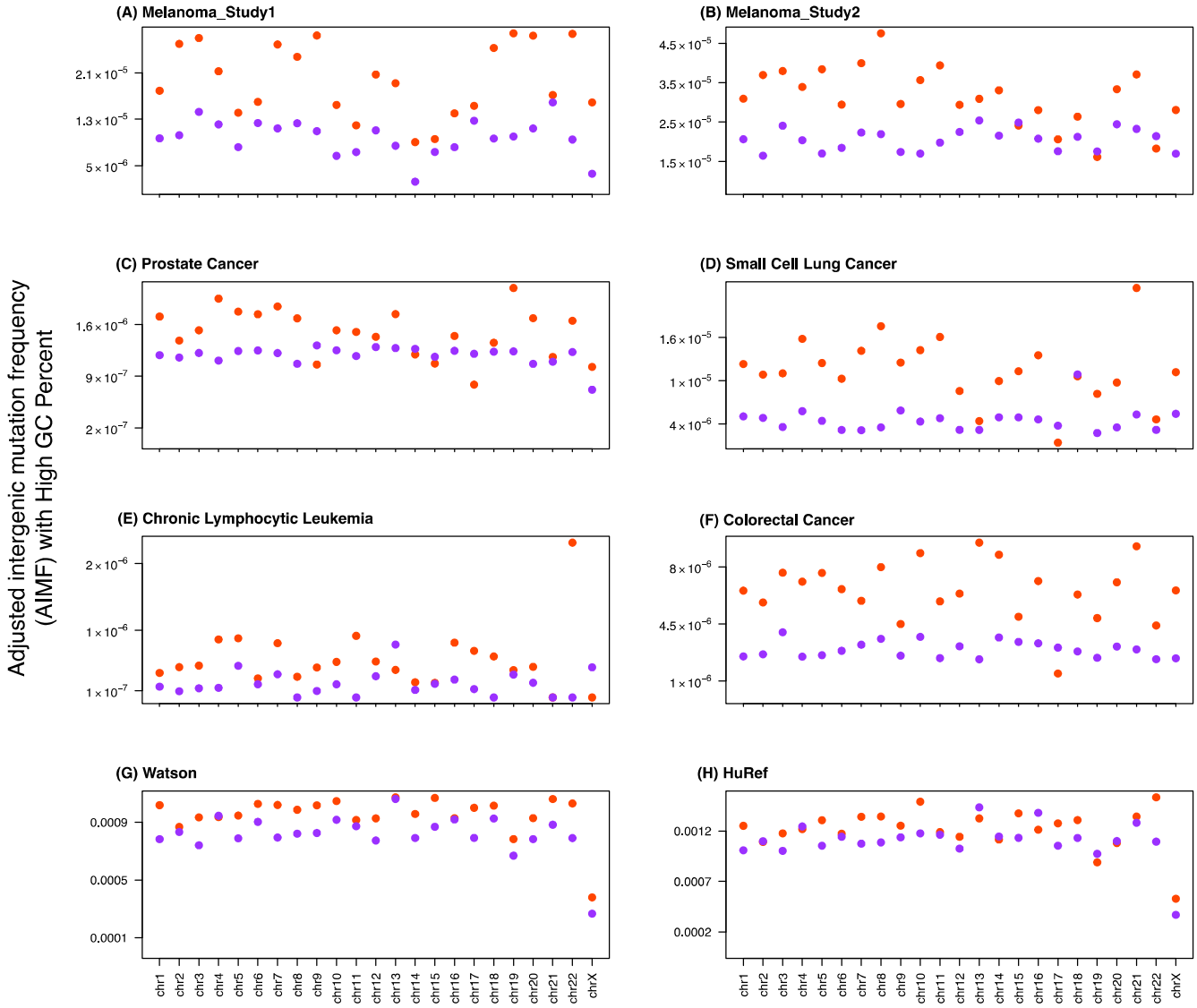
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S5: AIMF for constant late (orange) and constant early (purple) DNA replication timing for low GC percentage regions.



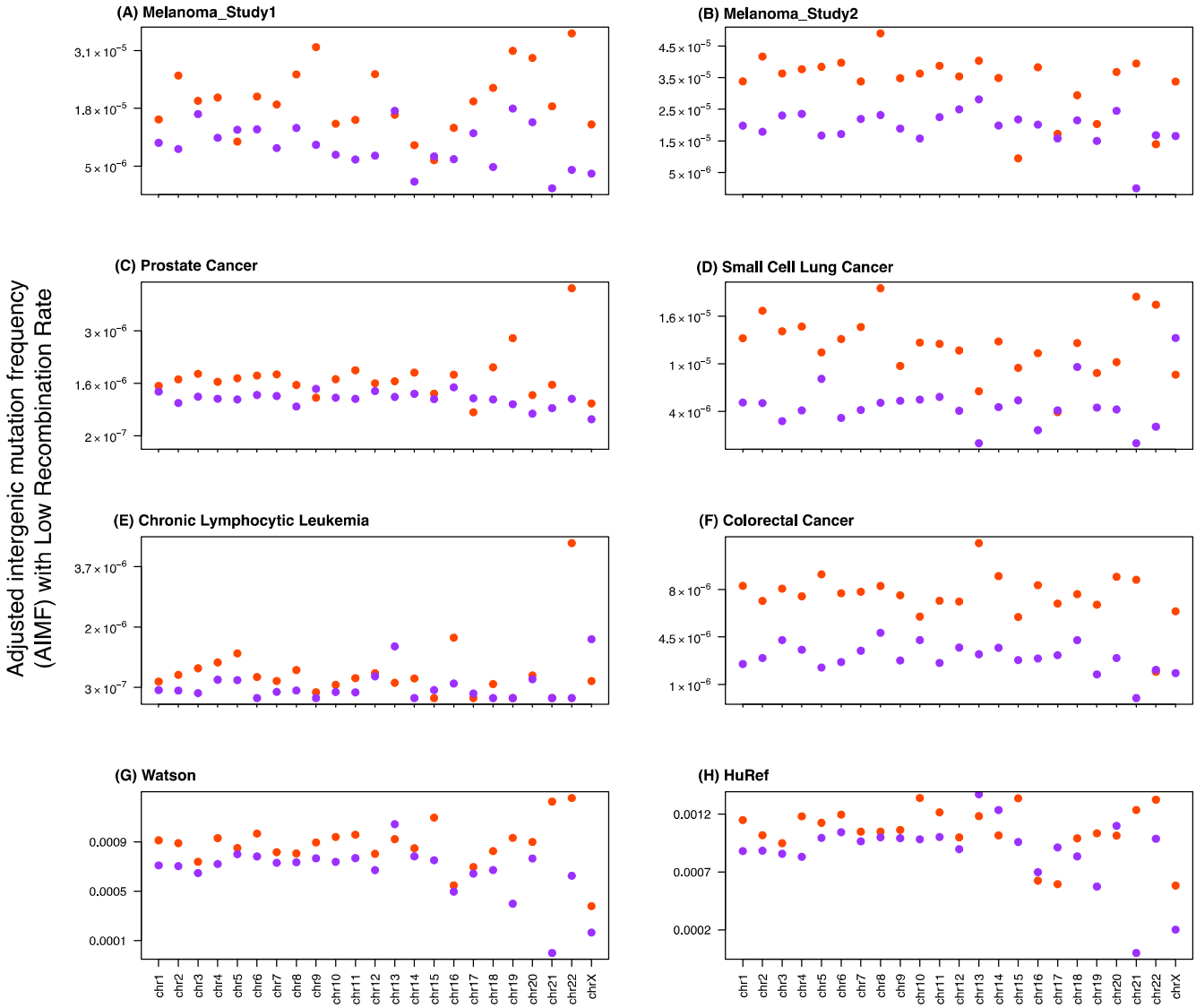
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S6: AIMF for constant late (orange) and constant early (purple) DNA replication timing for high GC percentage regions.



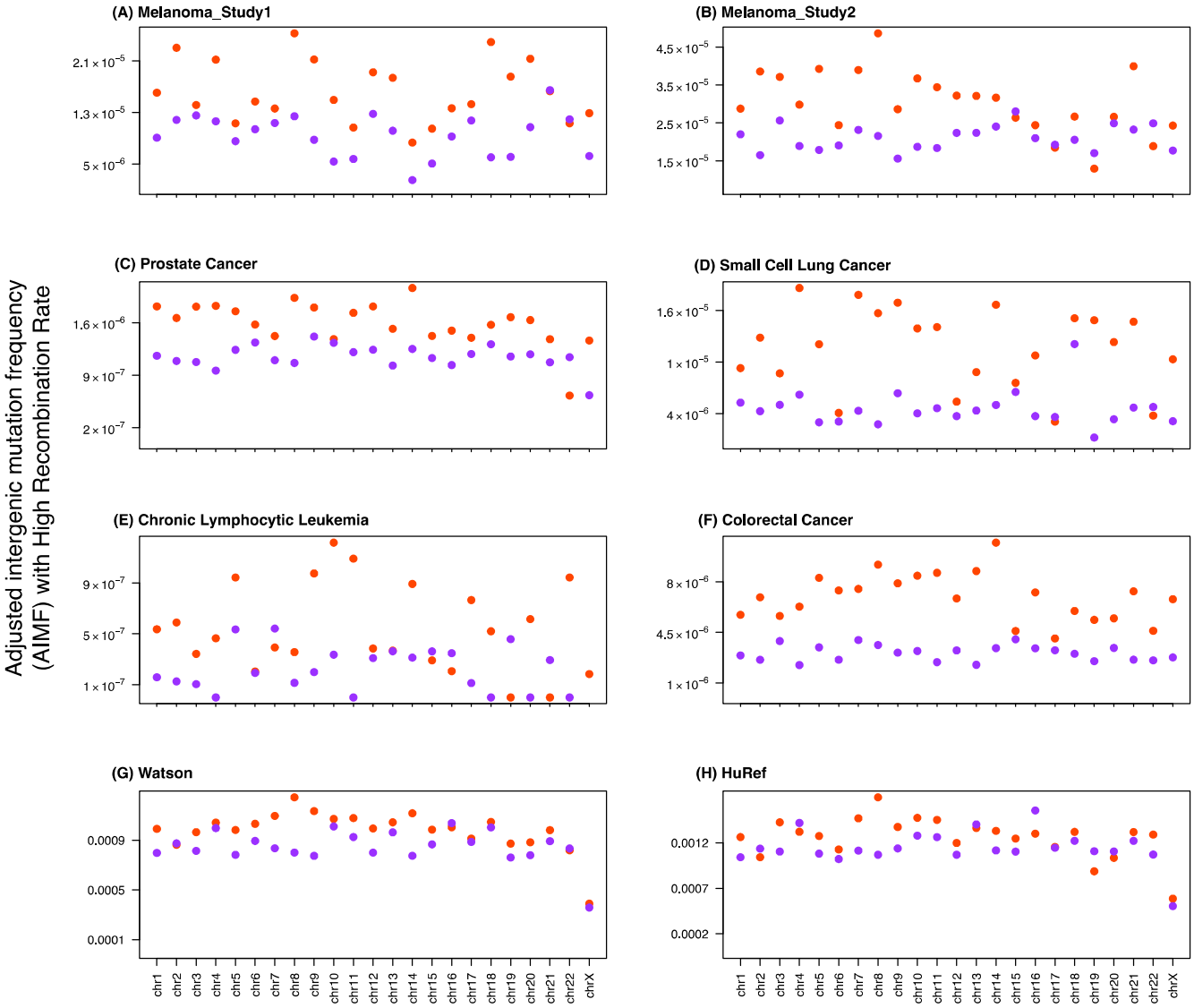
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S7: AIMF for constant late (orange) and constant early (purple) DNA replication timing for regions with low recombination rate.



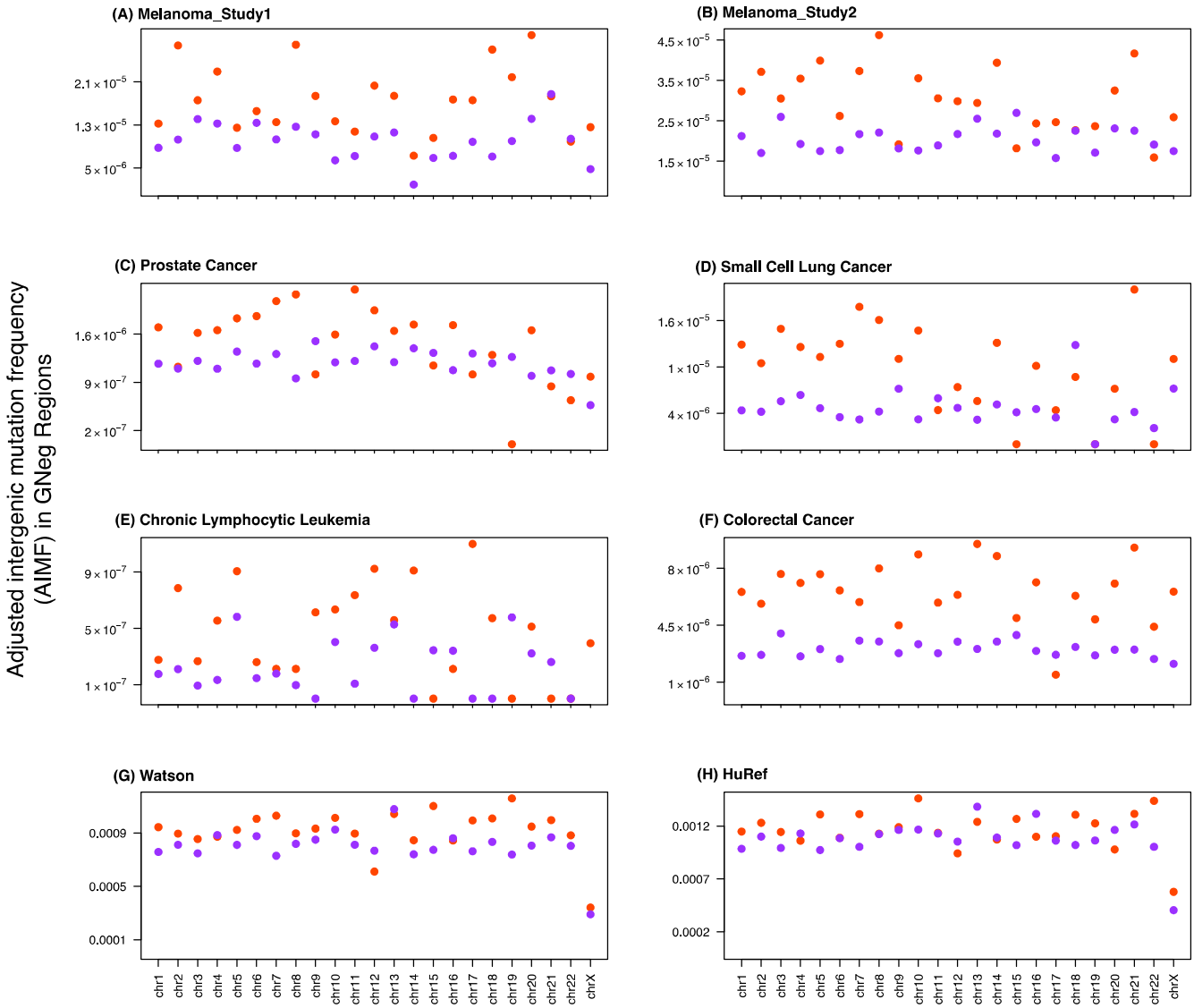
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S8: AIMF for constant late (orange) and constant early (purple) DNA replication timing for regions with high recombination rate.



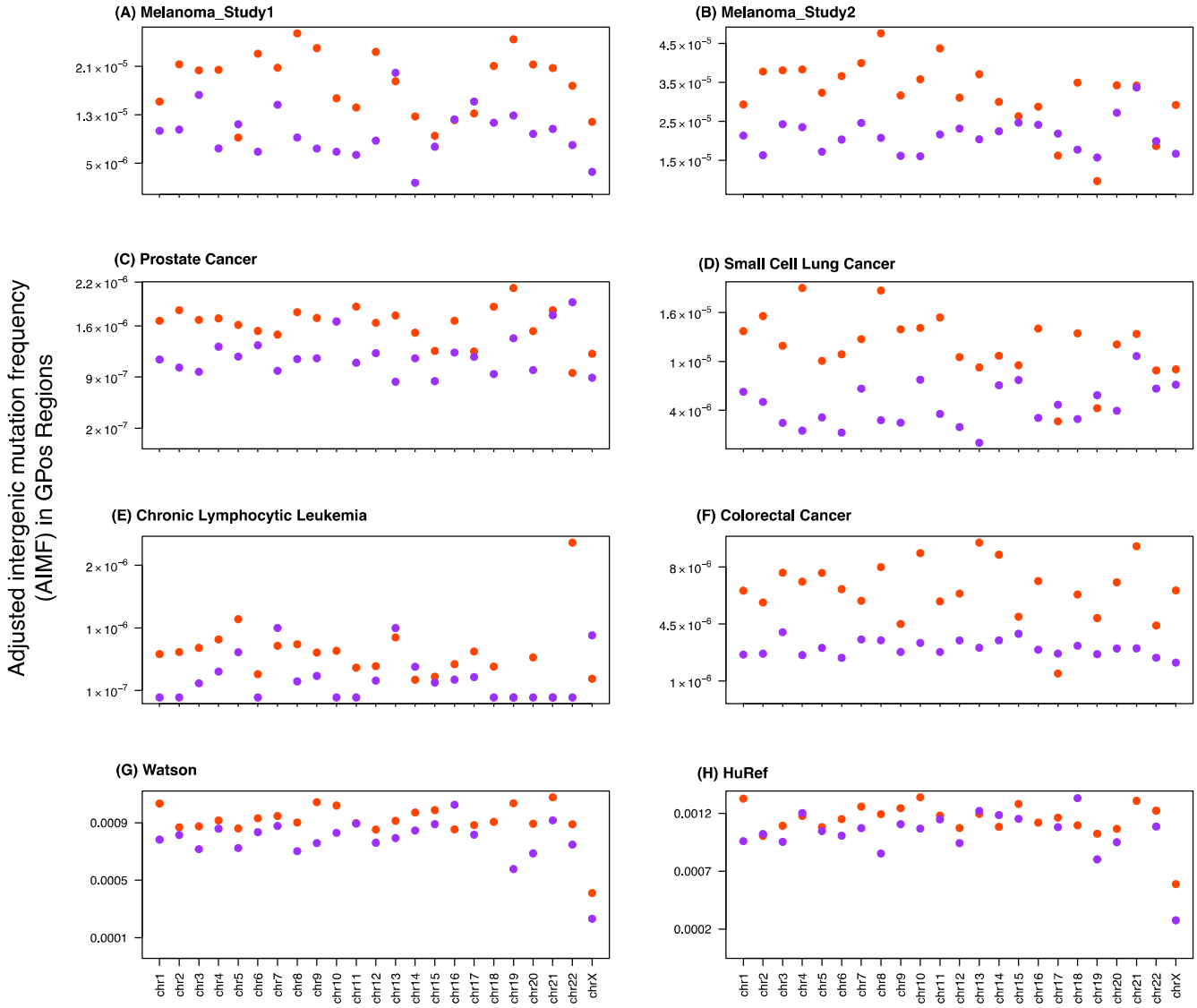
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S9: AIMF for constant late (orange) and constant early (purple) DNA replication timing for G-Neg regions.



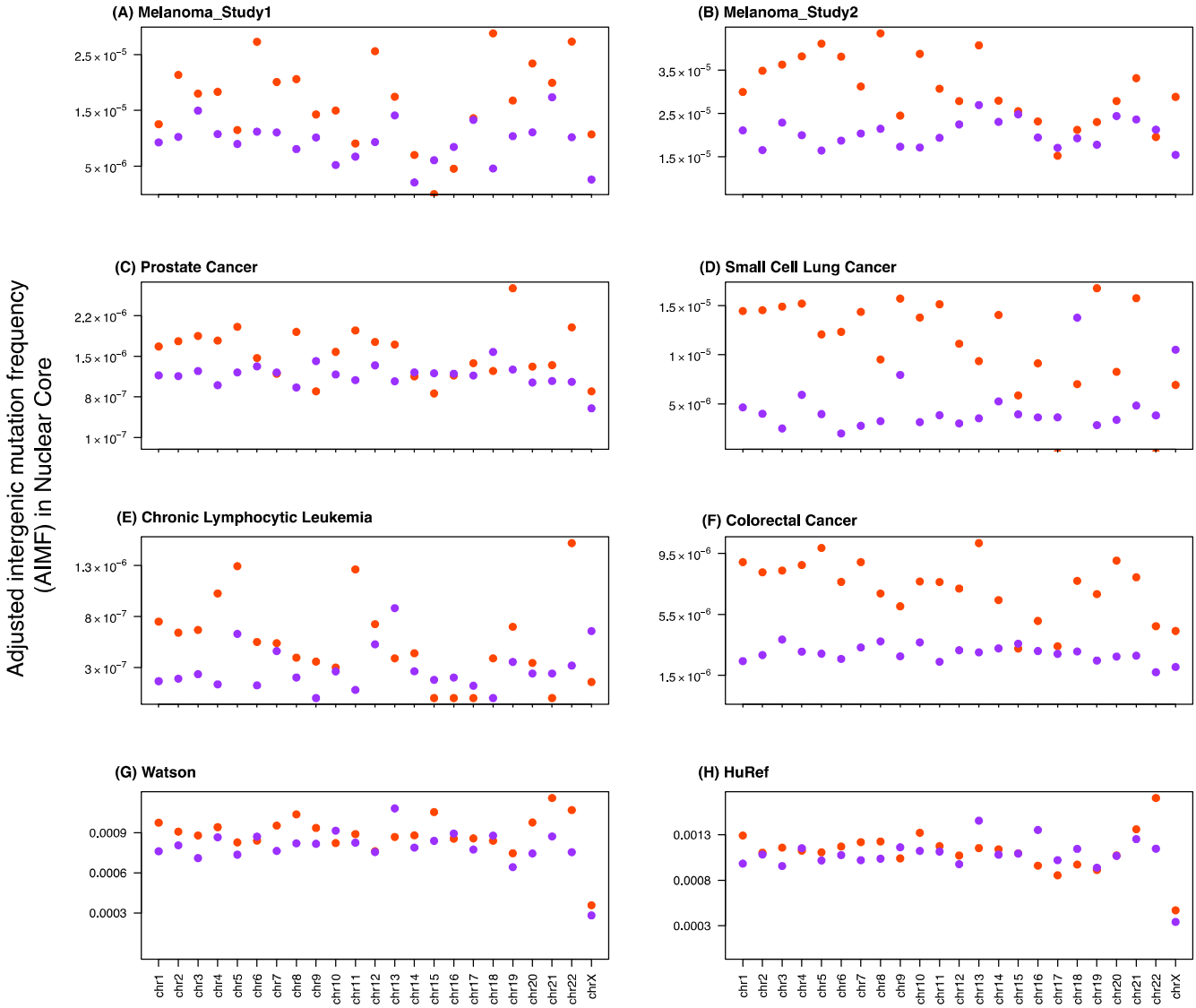
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S10: AIMF for constant late (orange) and constant early (purple) DNA replication timing for G-Pos regions.



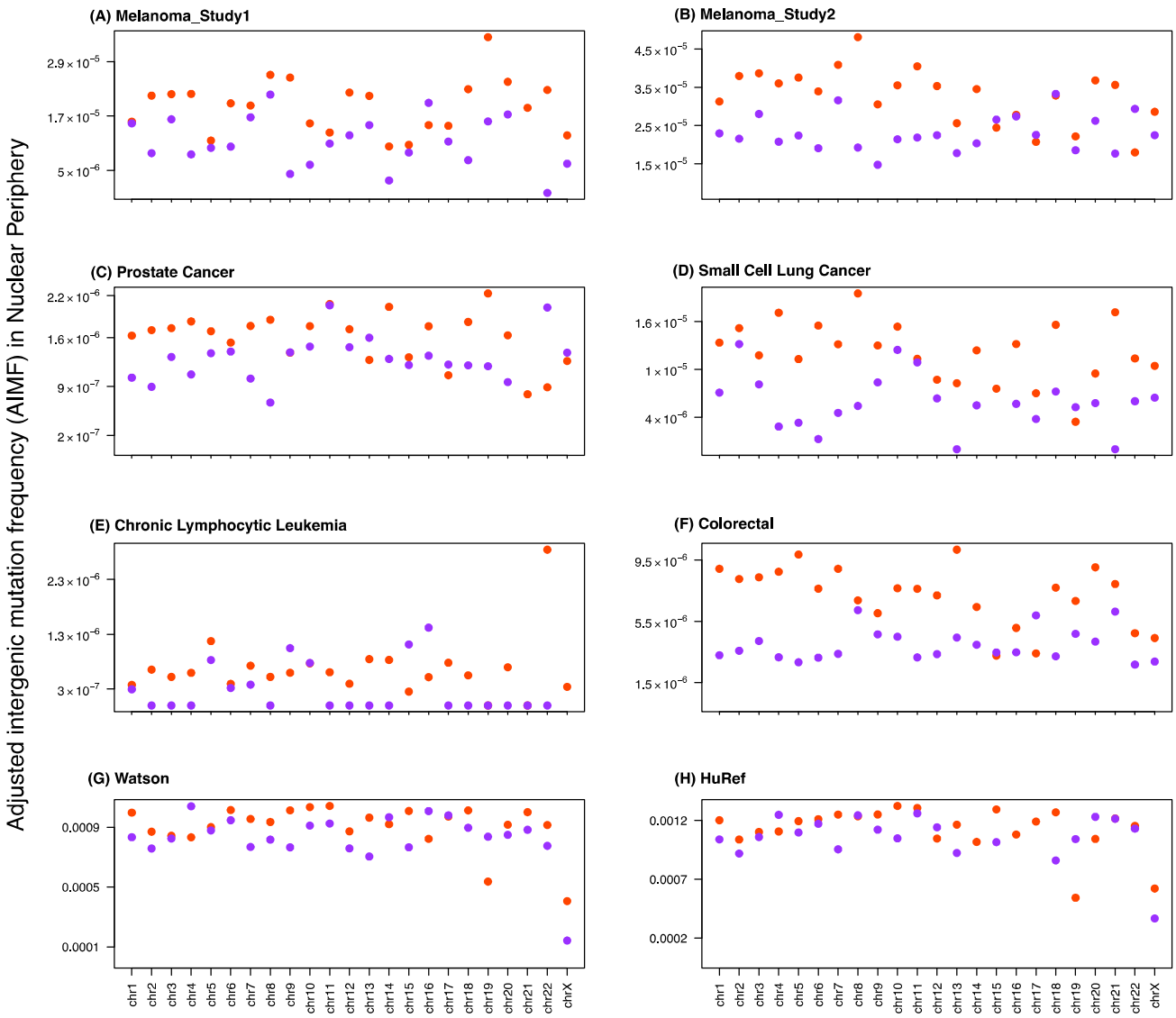
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S11: AIMF for constant late (orange) and constant early (purple) DNA replication timing for regions residing in the nuclear core.



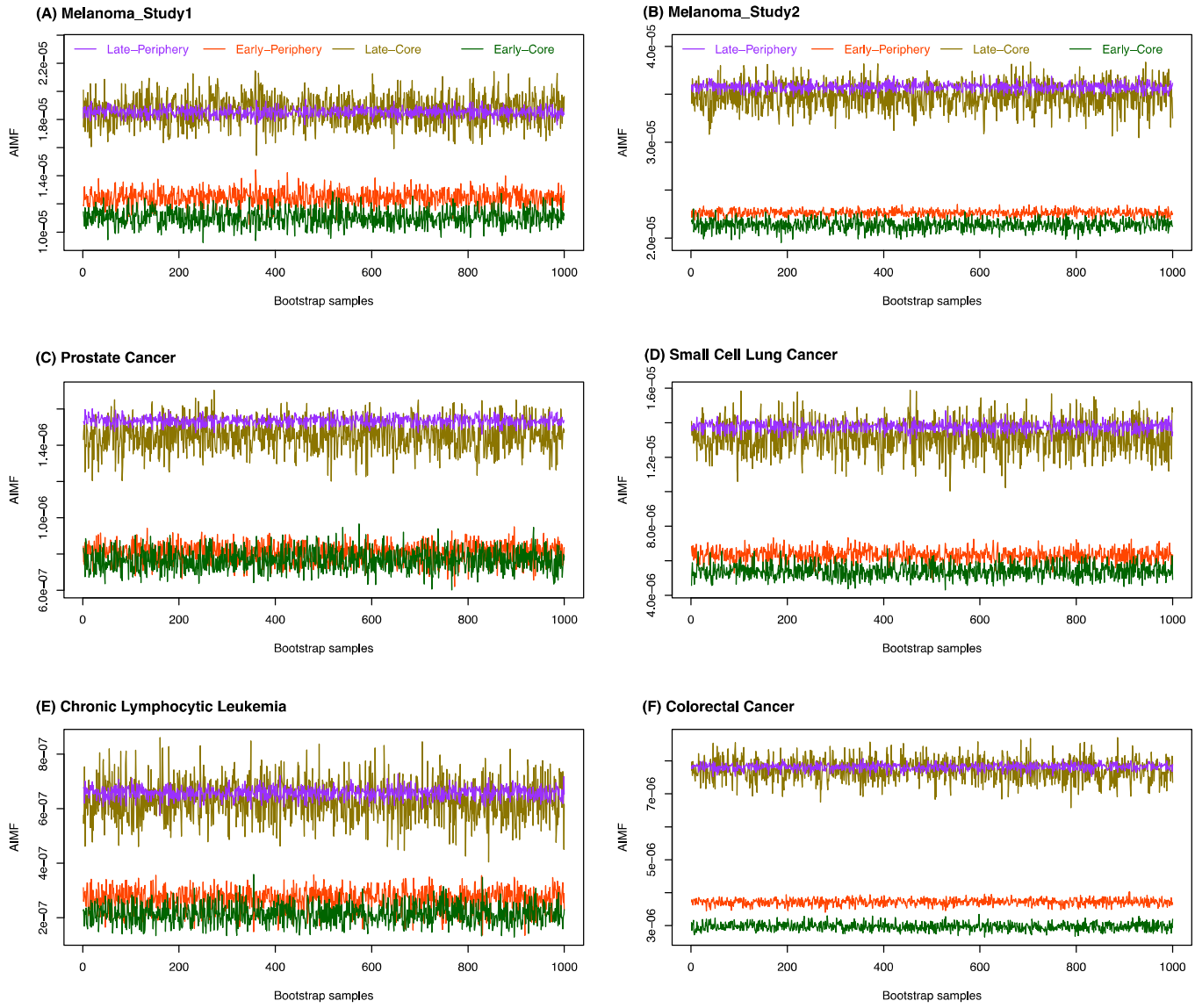
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S12: AIMF for constant late (orange) and constant early (purple) DNA replication timing for regions residing in the nuclear periphery.



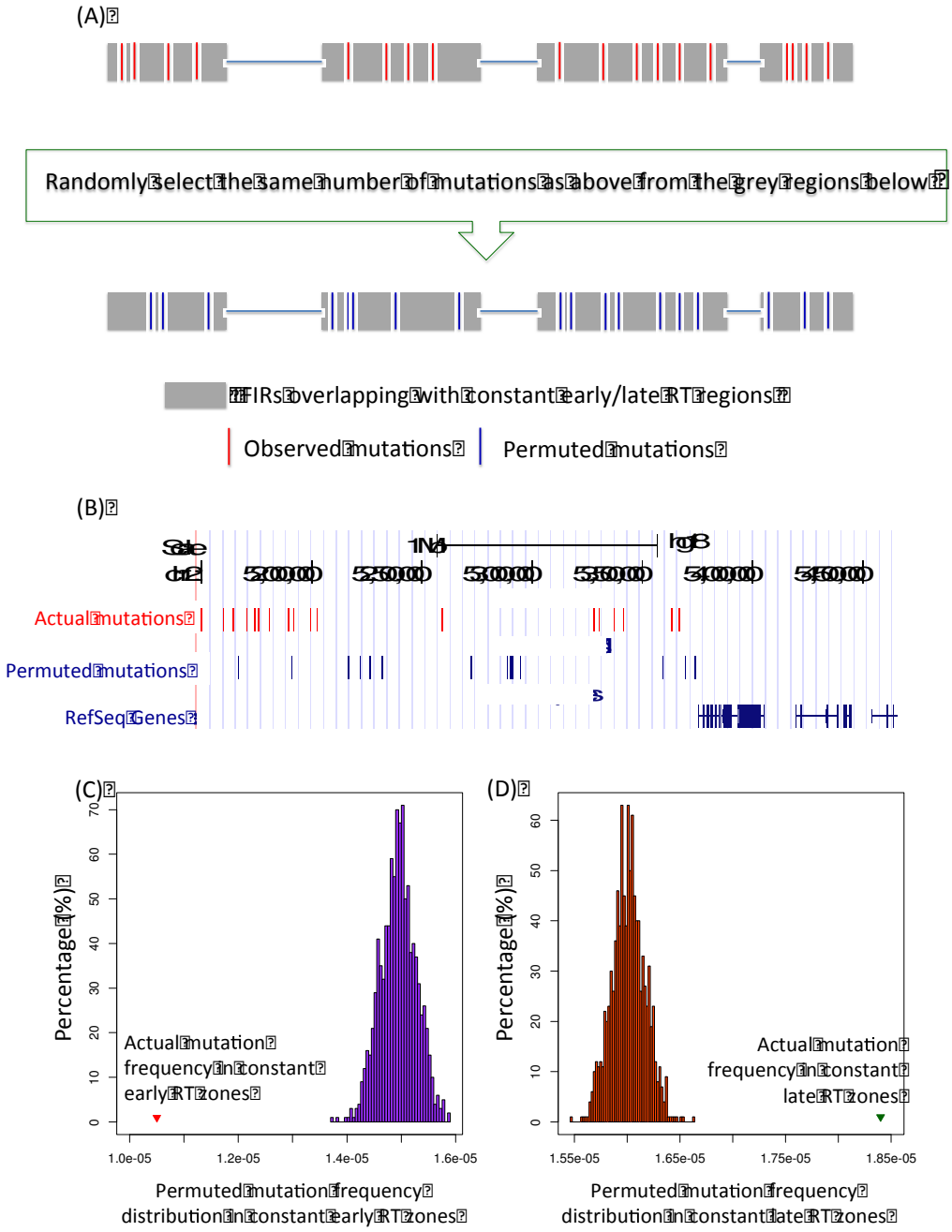
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The x-axes display results for chr1 – chr22 and chrX.

Supplementary Figure S13: AIMF for constant late (orange) and constant early (purple) DNA replication timing for regions residing in the nuclear periphery.



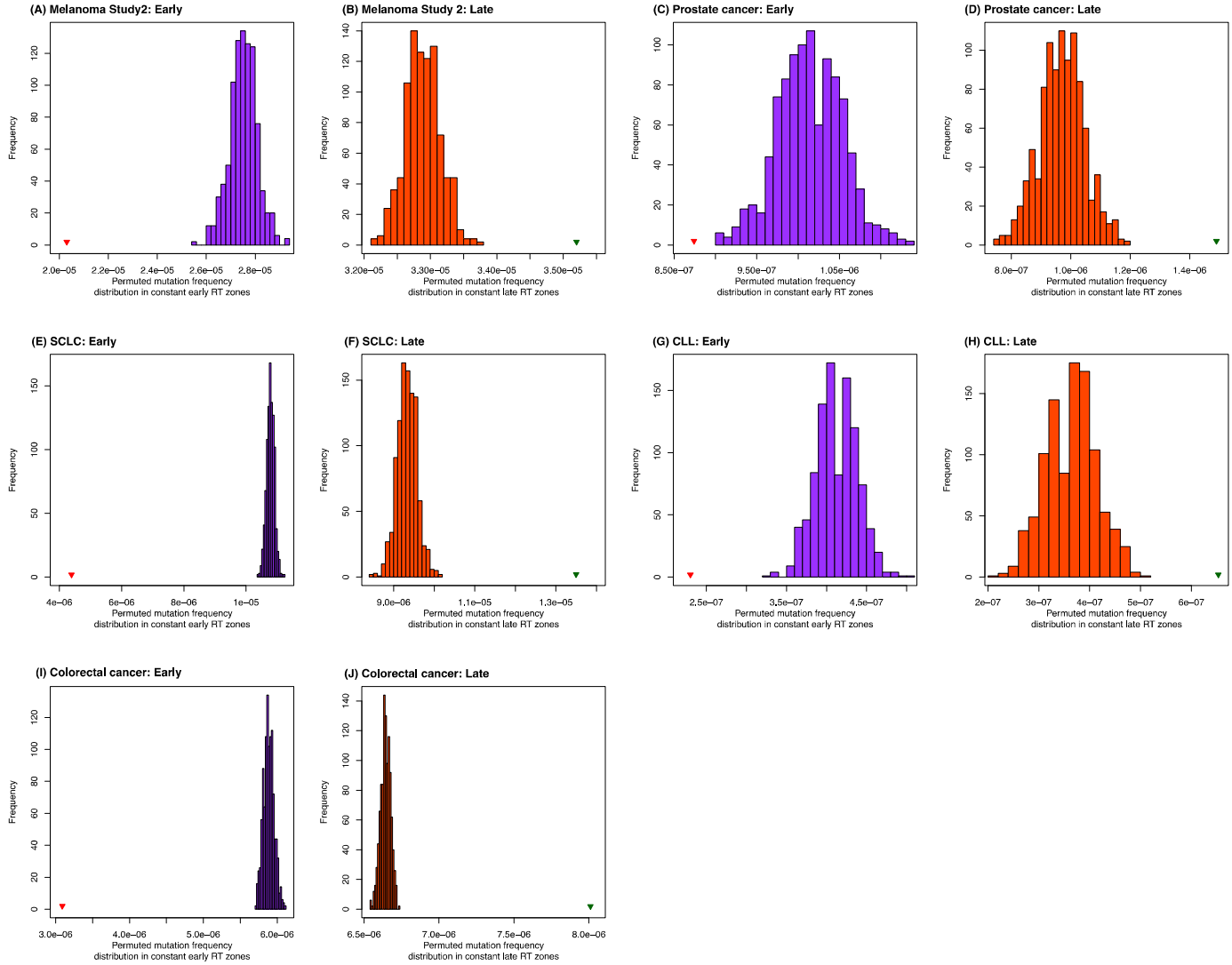
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S14: The permutation procedure.



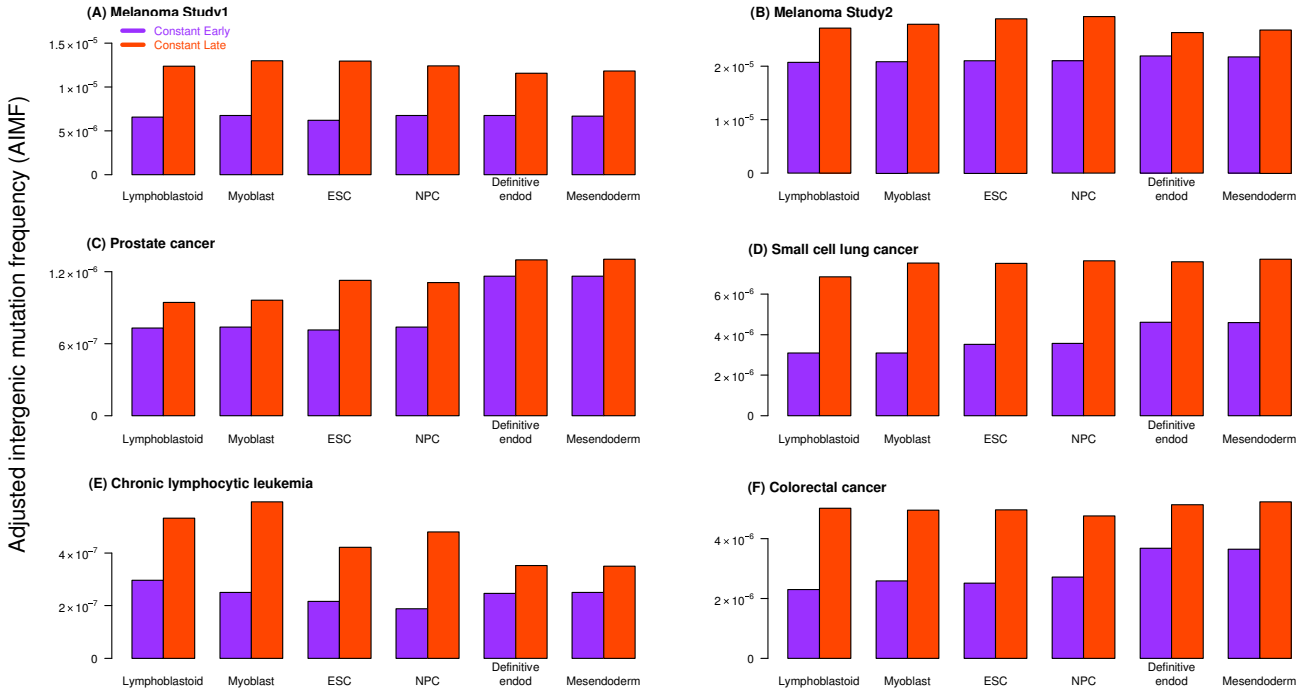
(A) Workflow of the permutation approach: we chose a certain number of mutations in the constant late or early replication timing zones overlapping with FIRs, and then randomly selected the same number of mutations (potentially at different positions) from these regions. (B) The panel displays the actual mutations (upper panel) and permuted mutations (lower panel) in a range of chr2 in melanoma. (C) - (D) The null distribution of permuted mutations for constant early and constant late DNA replication timing (RT) zones, with the actual mutation distribution marked by a triangle (red: constant early mutations; green: constant late mutations) for melanoma study 1.

Supplementary Figure S15: The permutation results.



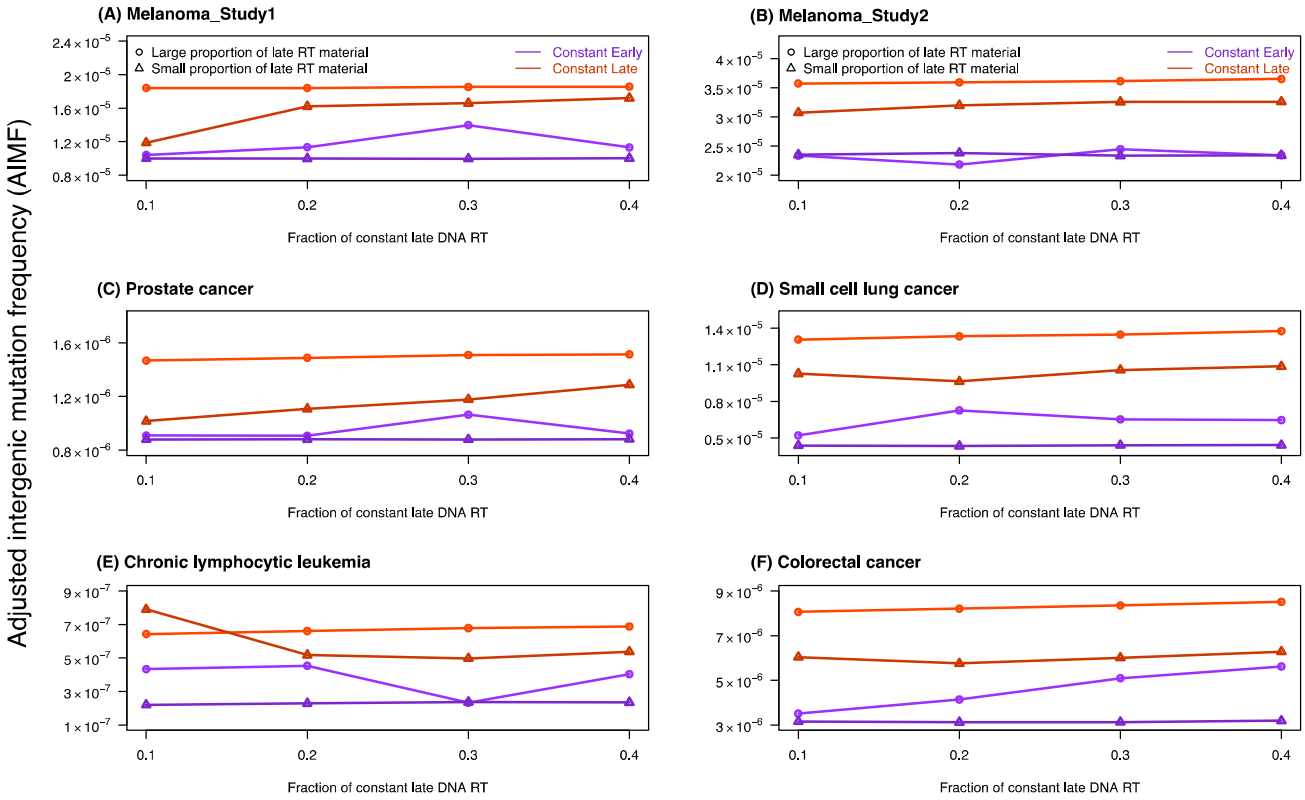
(A) – (J): The panels display the null distributions of permuted mutations for constant early and constant late DNA replication timing (RT) zones, with the actual mutation distribution marked by a triangle (red: constant early mutations; green: constant late mutations) for melanoma study 2[26], prostate cancer[27], small cell lung cancer[28], chronic lymphocytic leukemia[29], and colorectal cancer[30].

Supplementary Figure S16: AIMF for the six cancer sample sets using replication timing data from lymphoblastoid, myoblast, embryonic stem cells (ESC), neural precursor cells (NPC), definitive endoderm and mesendoderm cell lines.



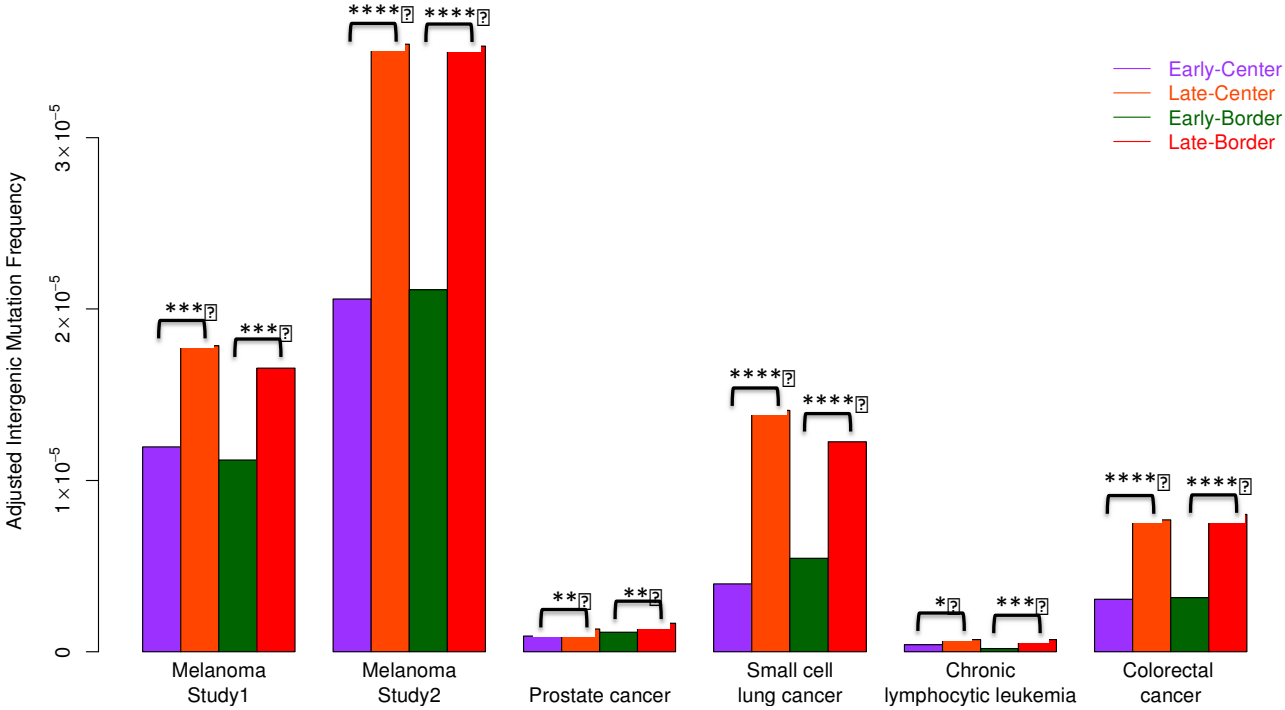
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], and (F) colorectal cancer samples[30].

Supplementary Figure S17: AIMF stratified by the proportion of constant late replication timing (RT) regions.



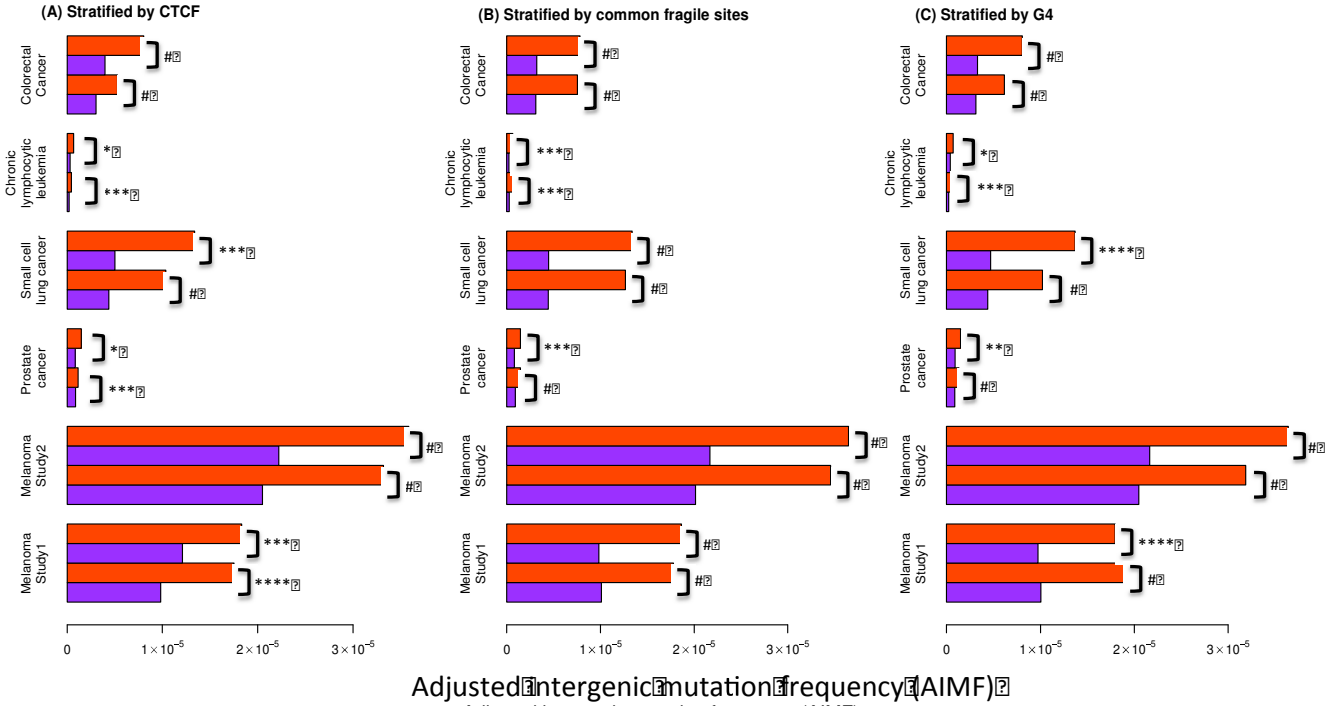
We used different cutoffs of fractions of constant late DNA replicating base pairs (0.1, 0.2, 0.3 and 0.4). The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], and (F) colorectal cancer samples[30].

Supplementary Figure S18: Genome-wide AIMF in melanoma, prostate cancer, small cell lung cancer, and chronic lymphocytic leukemia.



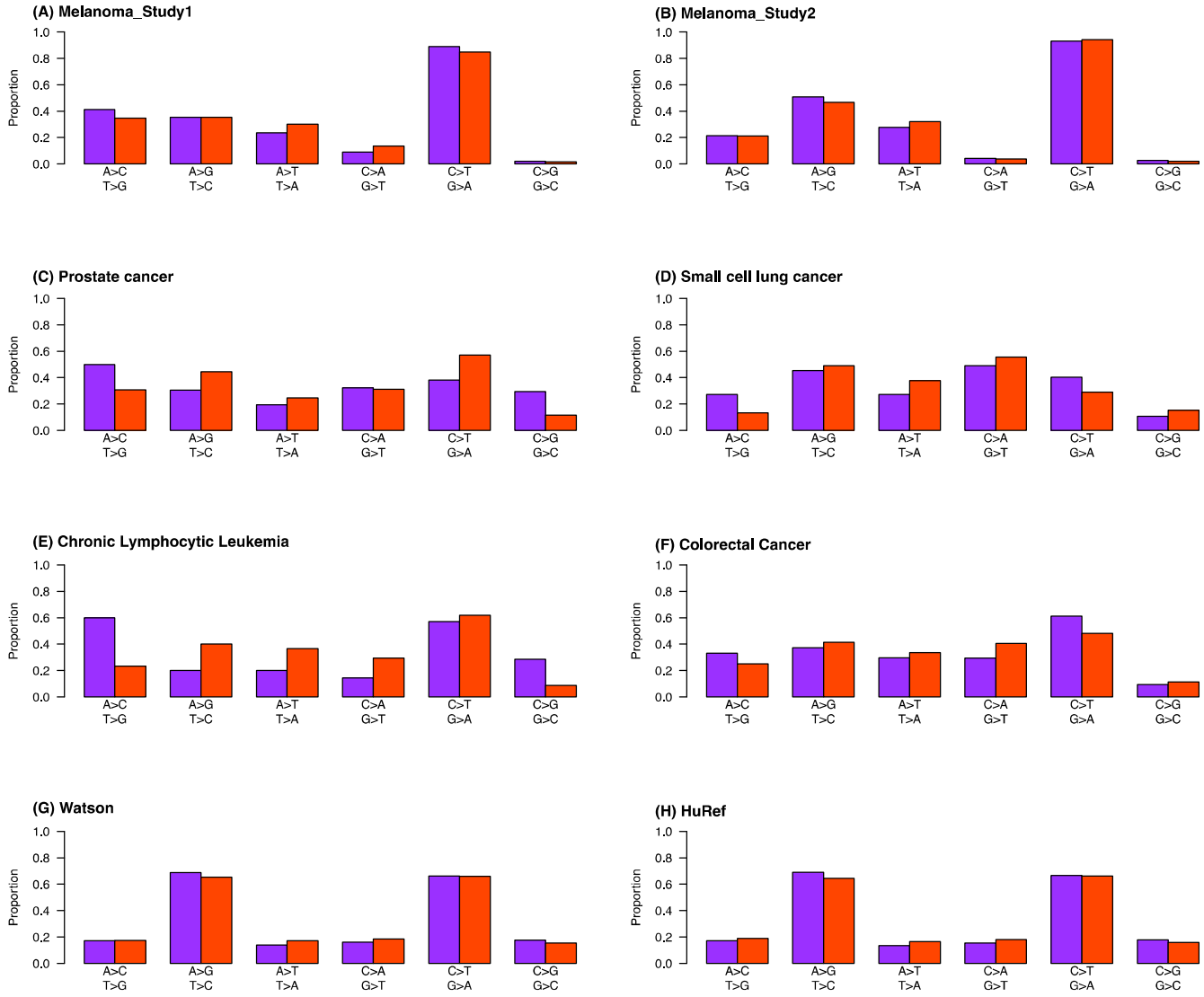
The figure shows regions in the following four categories, depending on where the mutation resided with respect to the region of a certain replication timing: in the center in constant late replication timing zones (mid 40Kb, orange), in the center in constant early replication timing zones (mid 40Kb, purple), at the border in constant late replication timing zones (40Kb from the boundary, red), and at the border in constant early replication timing zones (40Kb from the boundary, green). In each group, the differences in mutation frequencies were significant (Mann-Whitney U-Test, *: 0.01 < p-value < 0.05, **: 0.001 < p-value < 0.01, ***: 0.0001 < p-value < 0.001, and ****: p-value < 0.0001).

Supplementary Figure S19: Stratification analyses of the AIMF in early and late replication timing zones across five cancer types stratified by the prevalence of (A) CTCF-binding sites, (B) common fragile sites, and (C) G-quadruplex (G4) secondary structures.



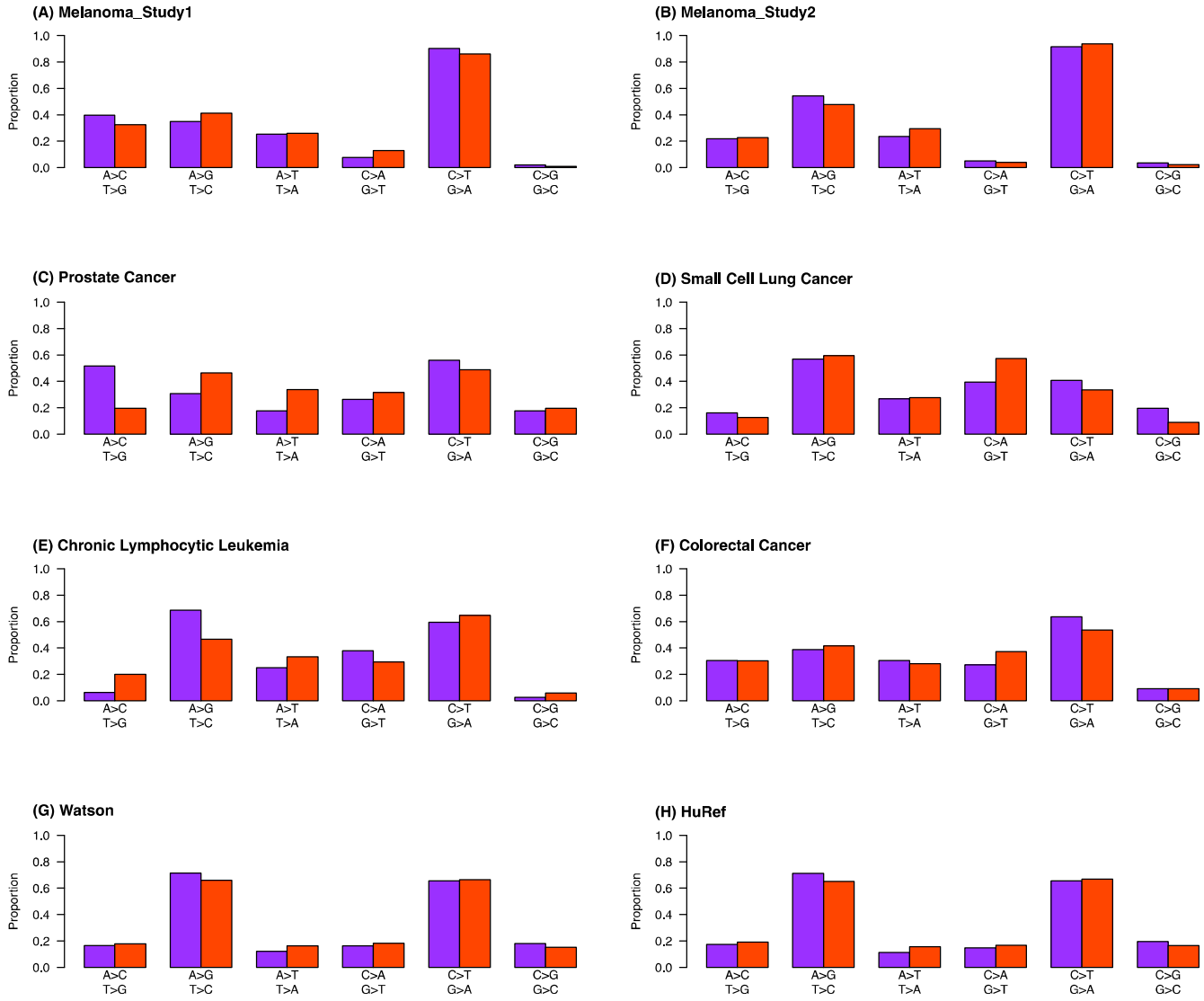
The stratification was based on the genome-wide median frequencies of these factors in each 1Mb non-overlapping window. For each factor and each cancer type, the upper two bars display results for the category 'large proportion of factor x' in the 1Mb windows, whereas the lower two bars display results for the category 'small proportions of factor x' in the 1Mb windows. P-values were calculated based on Mann-Whitney U-test and then adjusted for multiple comparisons (n = 2). *: 0.1 < p-value < 0.25 **: 0.01 < p-value < 0.1 ***: 0.001 < p-value < 0.01 ****: 0.0001 < p-value < 0.001 #: p-value < 0.0001.

Supplementary Figure S20: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for low gene density regions.



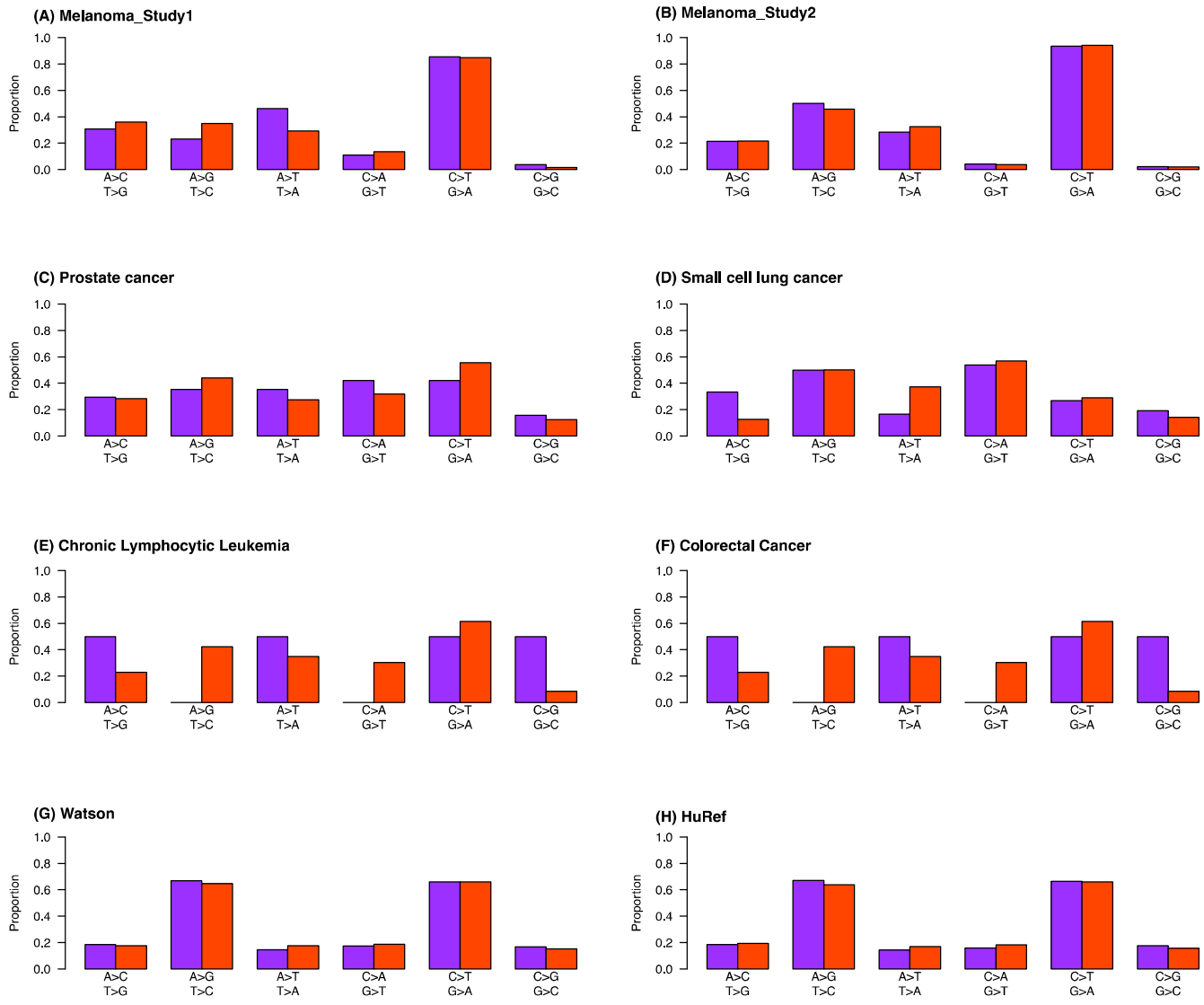
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S21: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for high gene density regions.



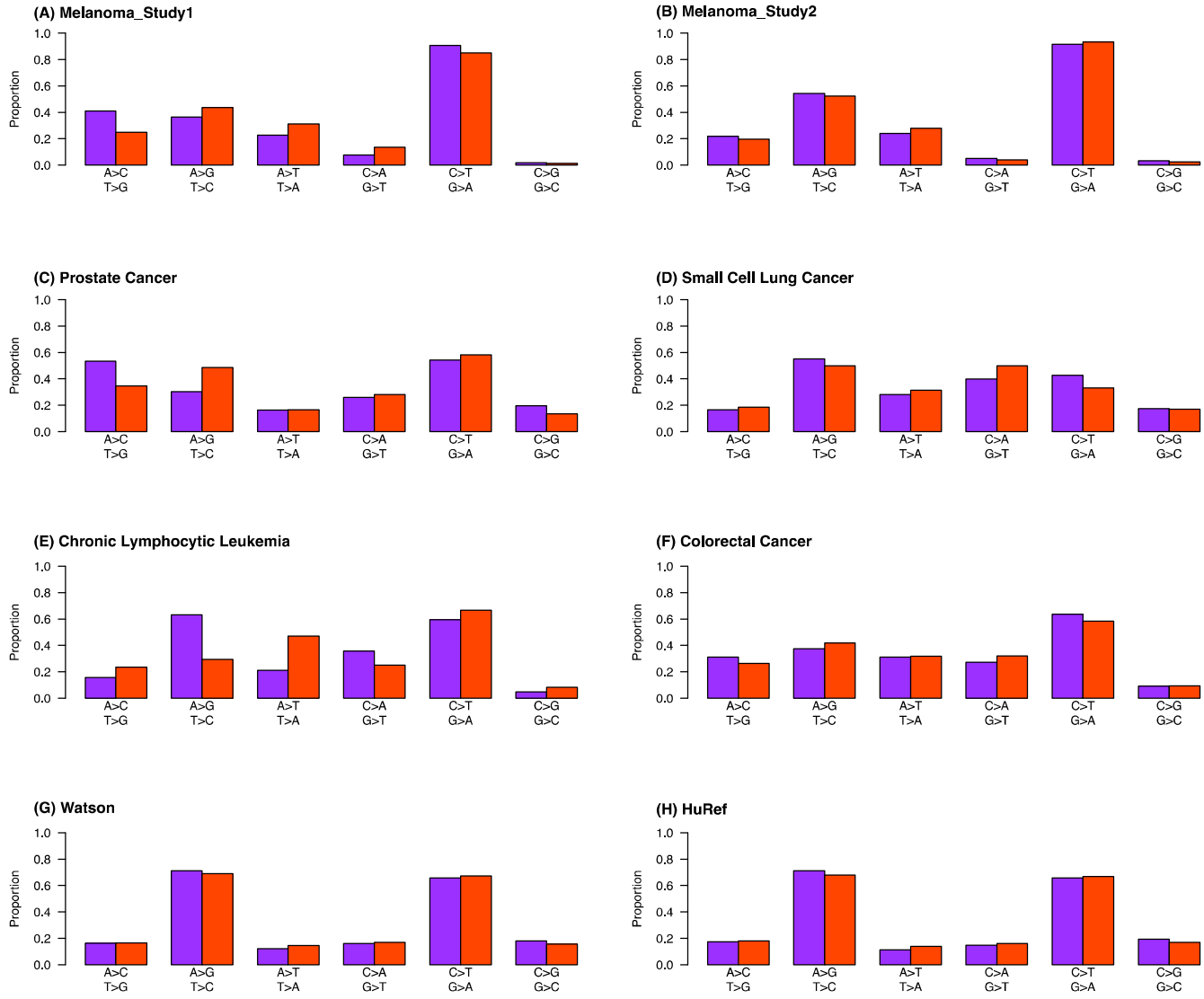
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S22: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for genomic regions with a low density of CpG islands.



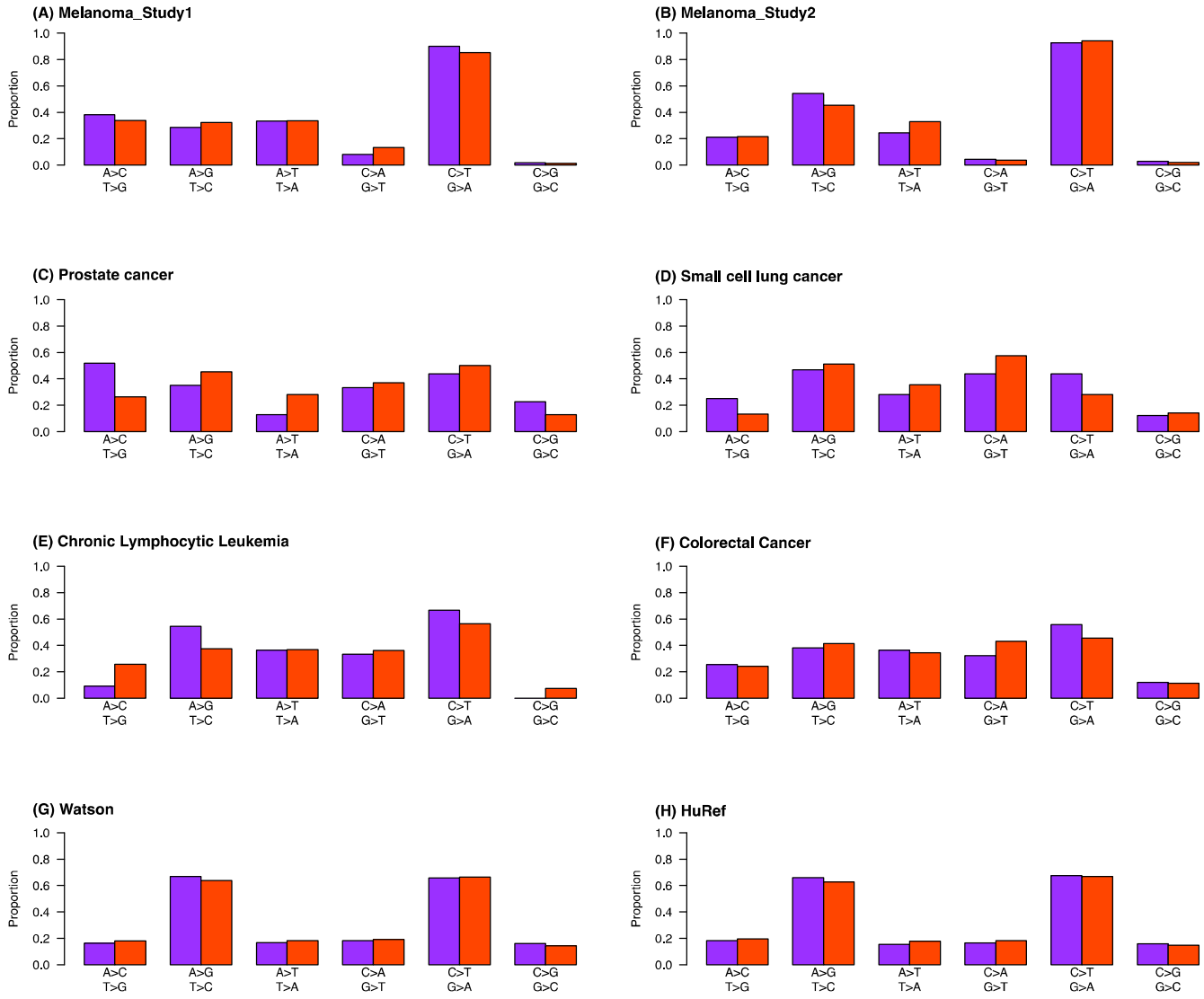
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S23: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for genomic regions with a high density of CpG islands.



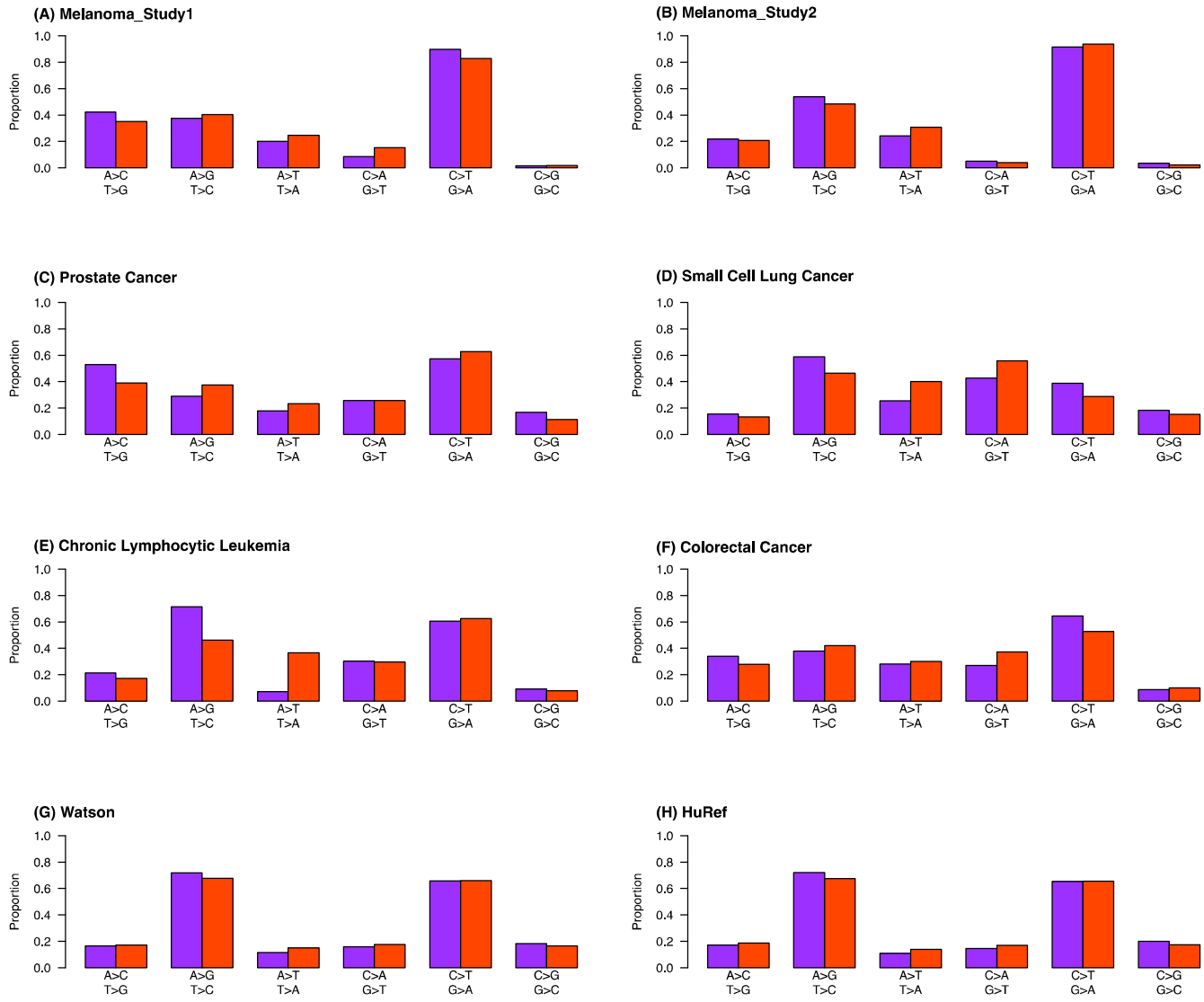
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S24: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for genomic regions with a low GC percentage.



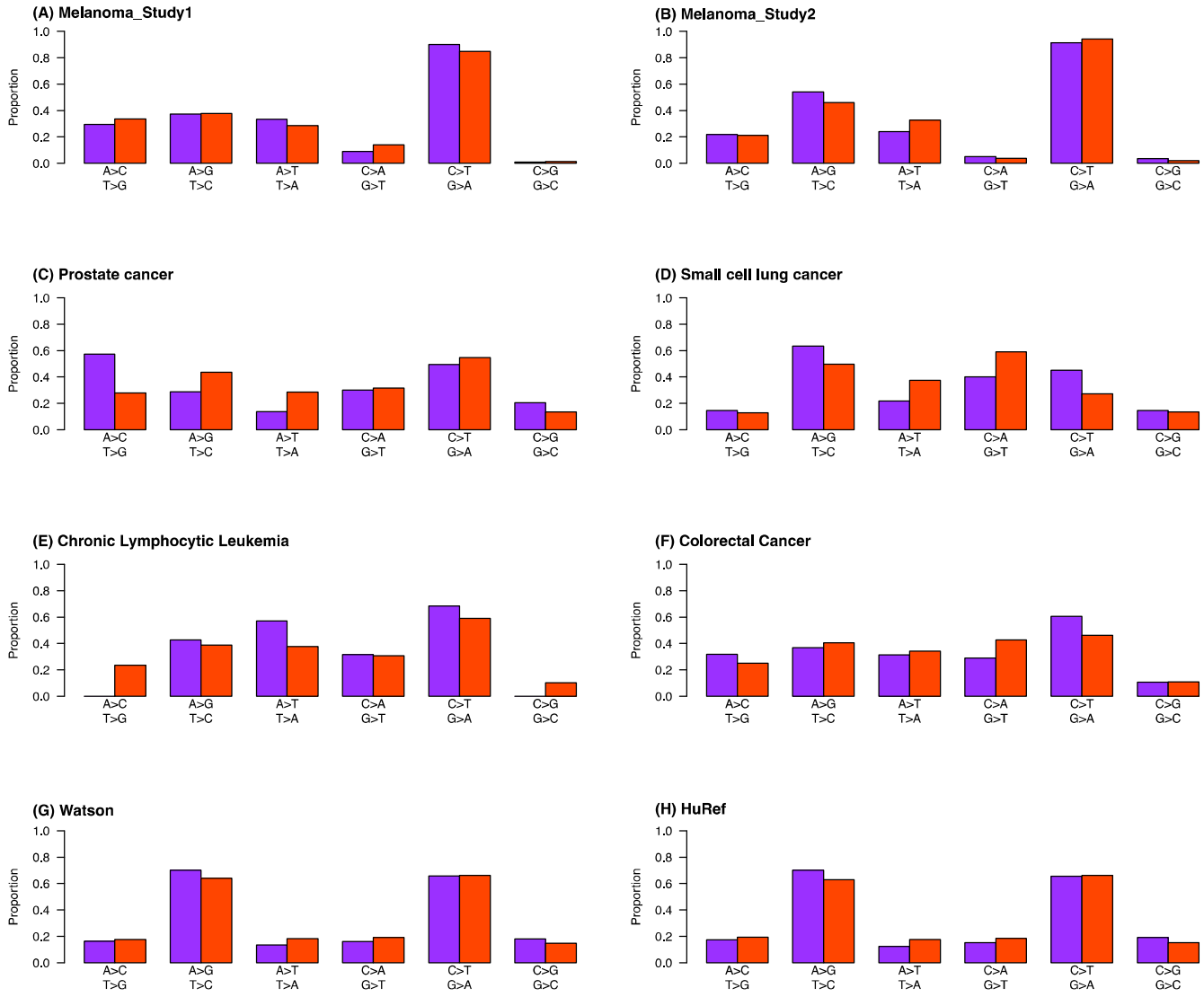
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S25: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for genomic regions with a high GC percentage.



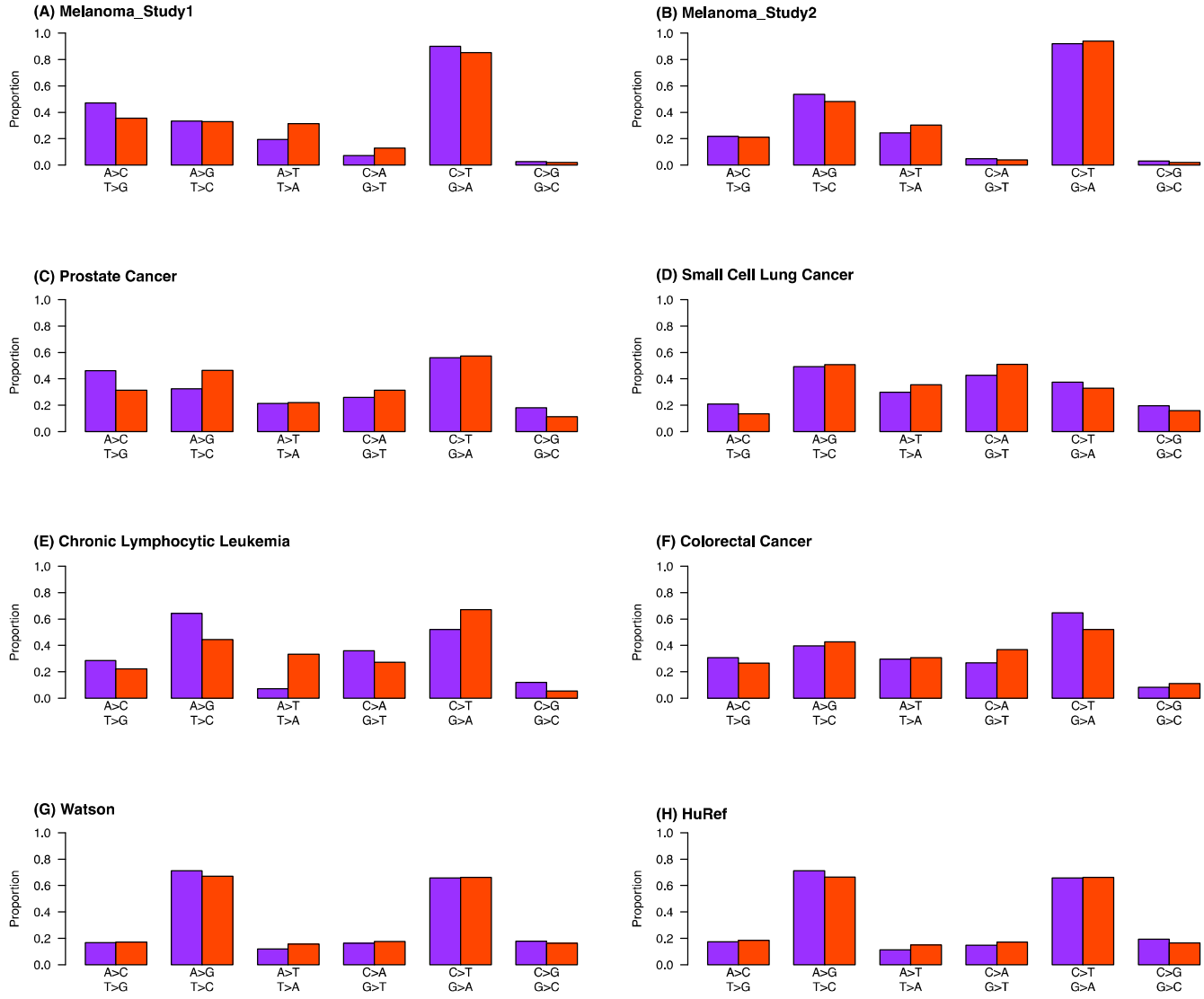
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S26: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for genomic regions with a low recombination rate.



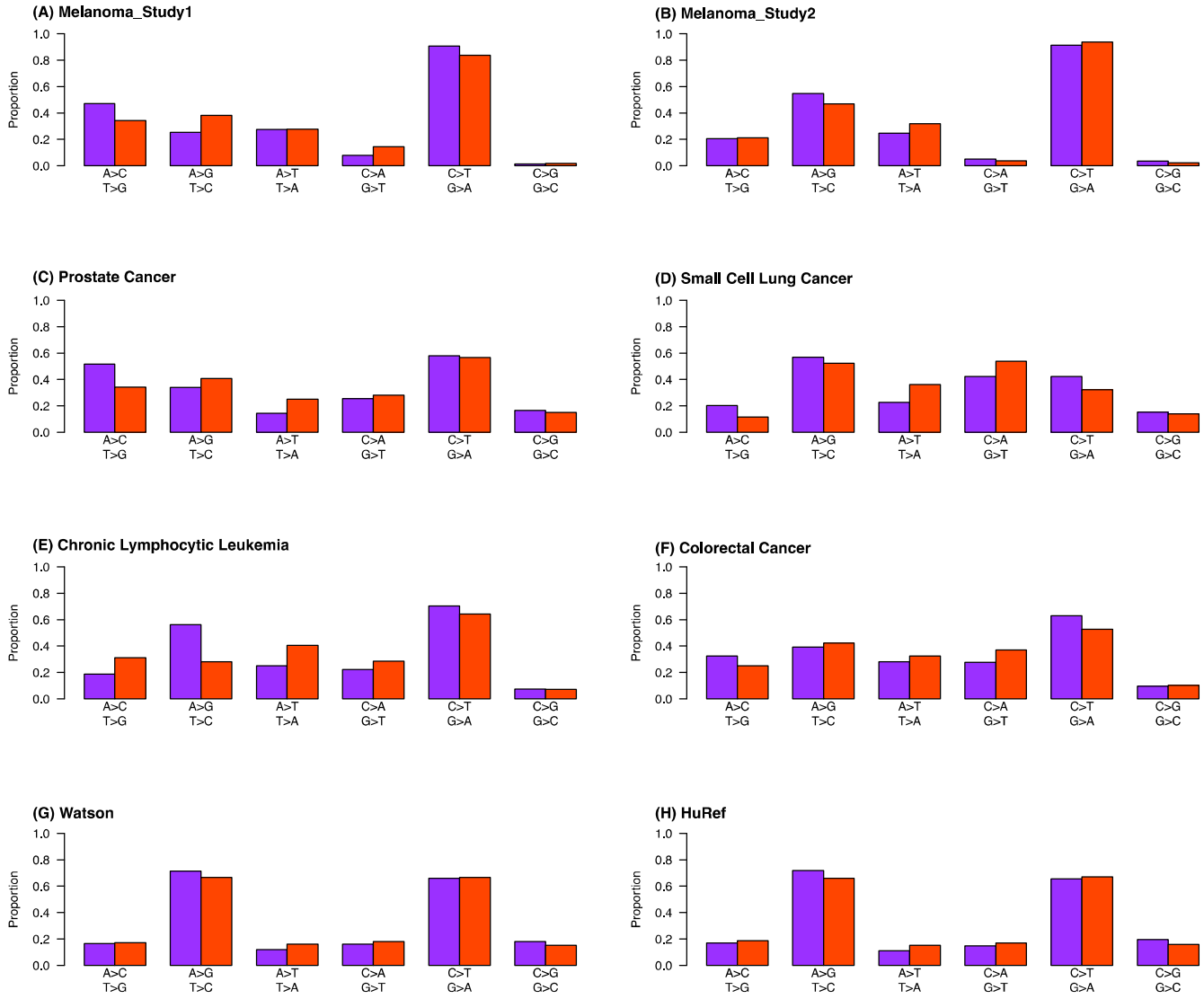
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S27: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for genomic regions with a high recombination rate.



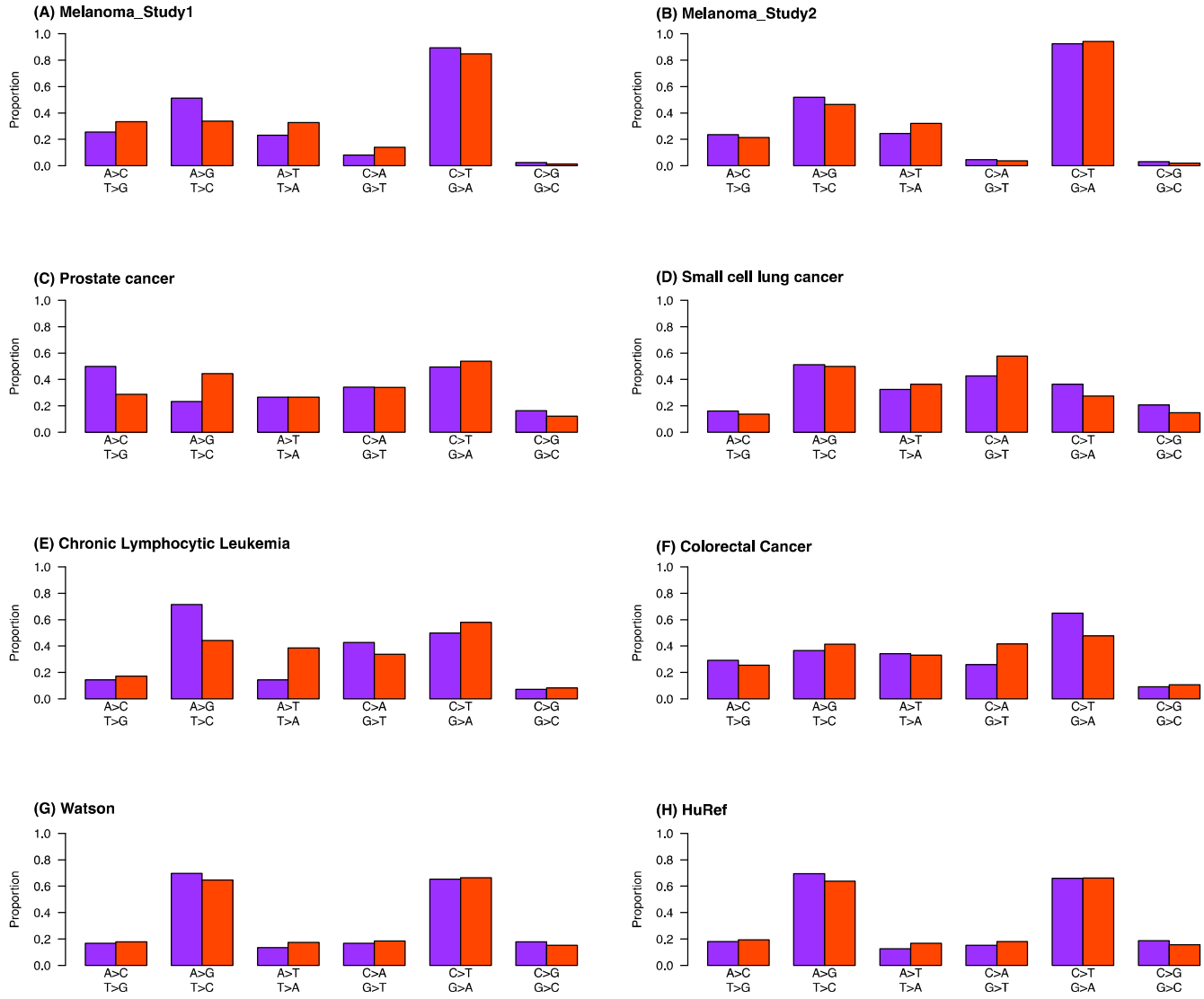
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S28: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for G-Neg genomic regions.



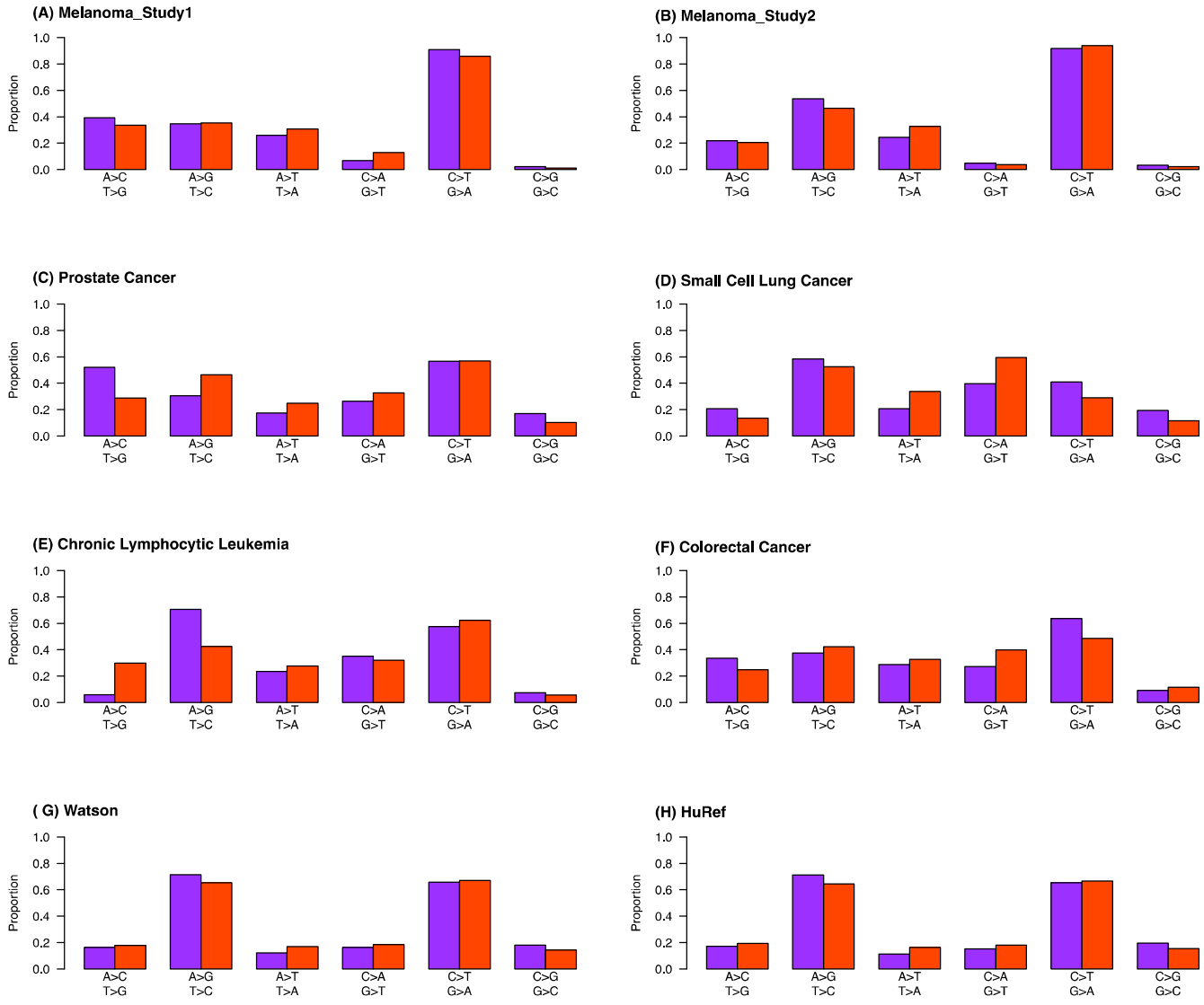
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S29: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for G-Pos genomic regions.



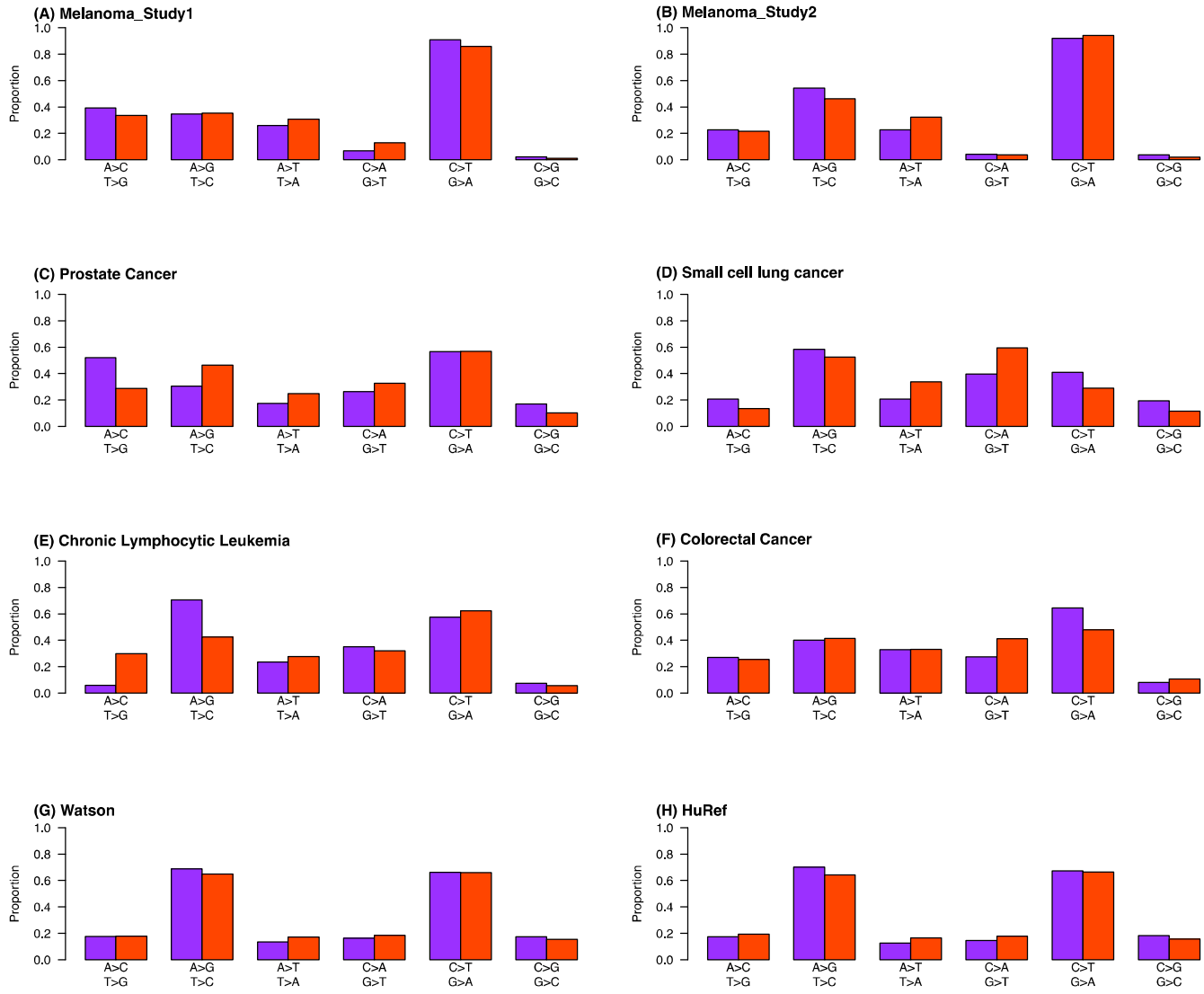
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S30: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for genomic regions residing in the nuclear core.



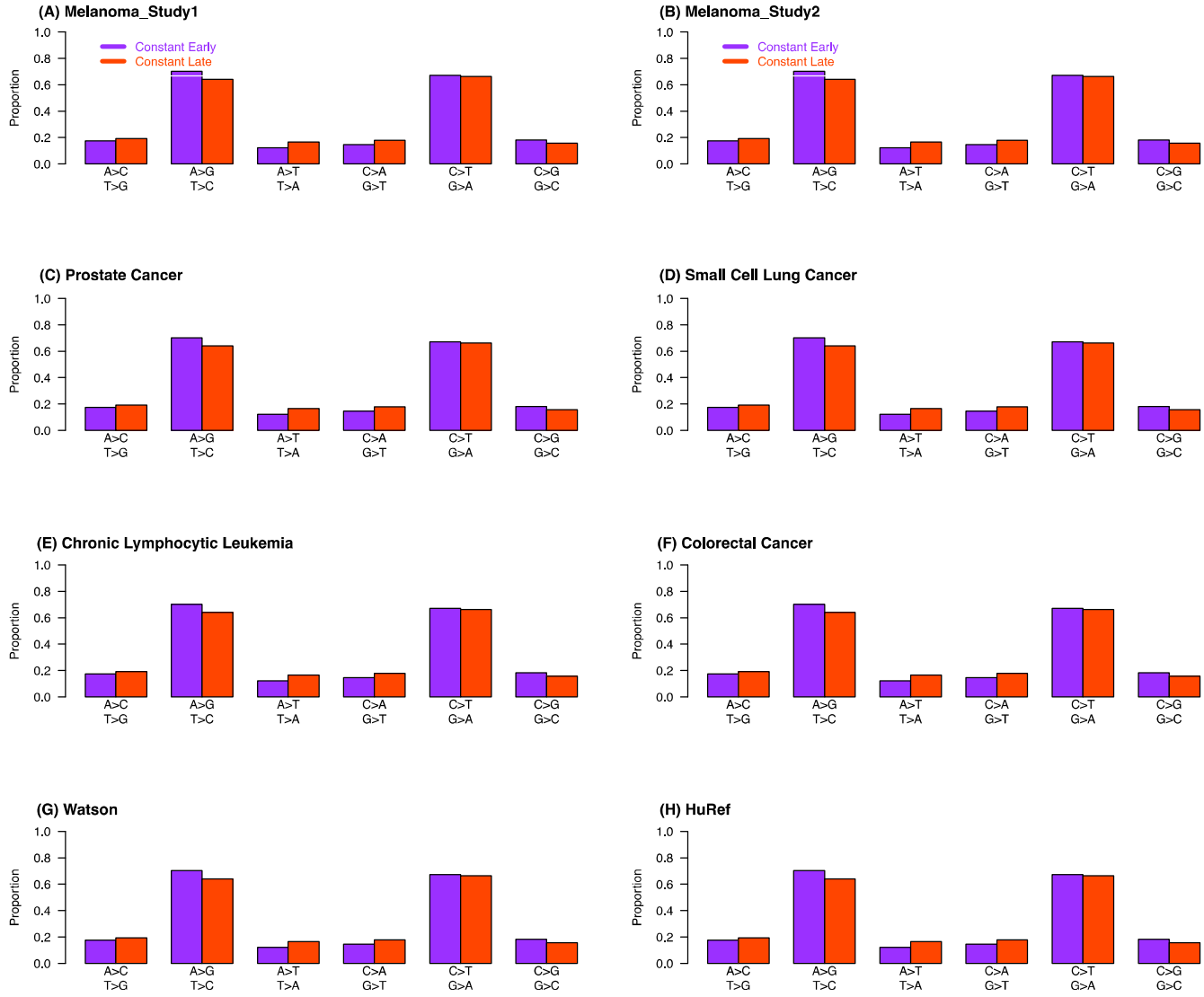
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S31: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing zones for genomic regions residing in the nuclear periphery.



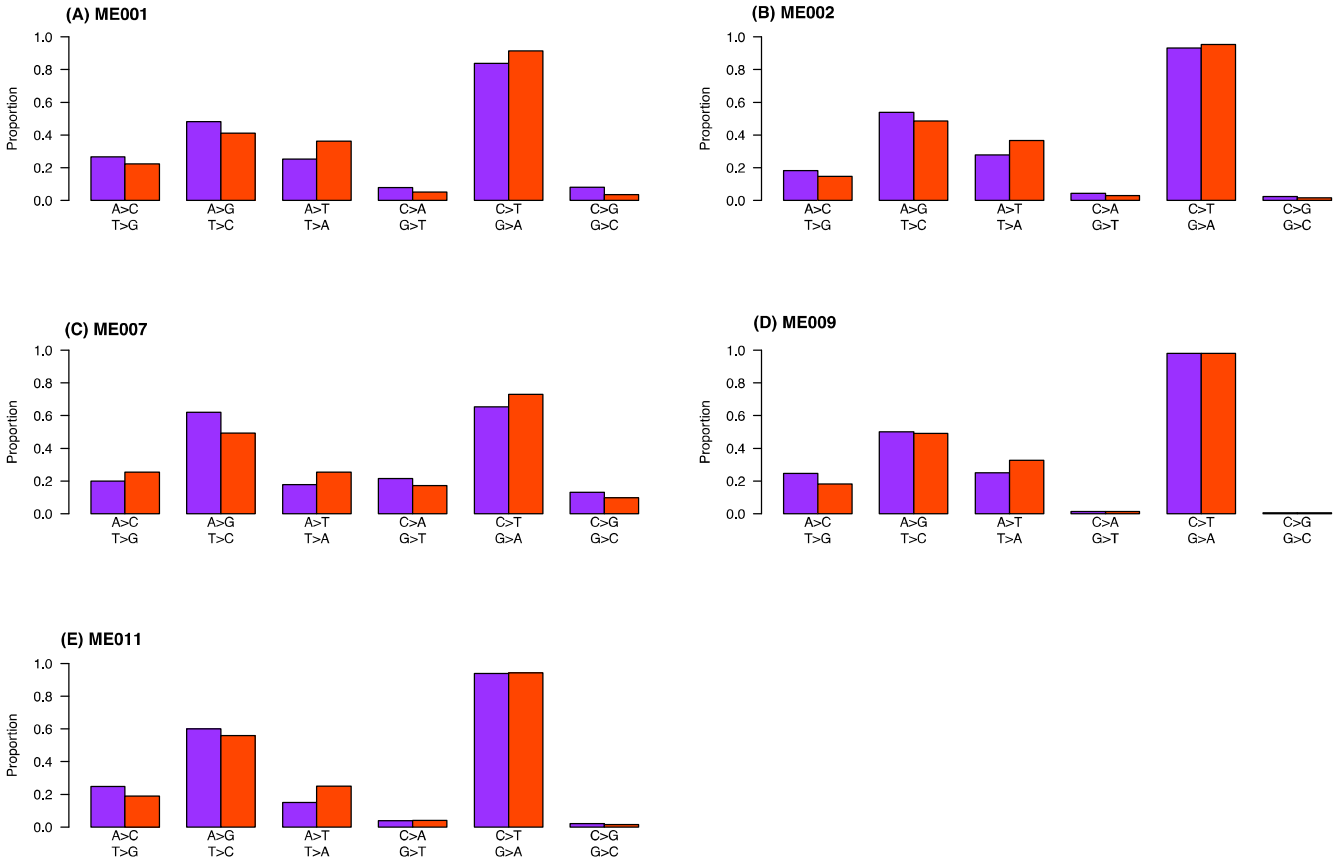
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32].

Supplementary Figure S32: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing within genes and promoters.



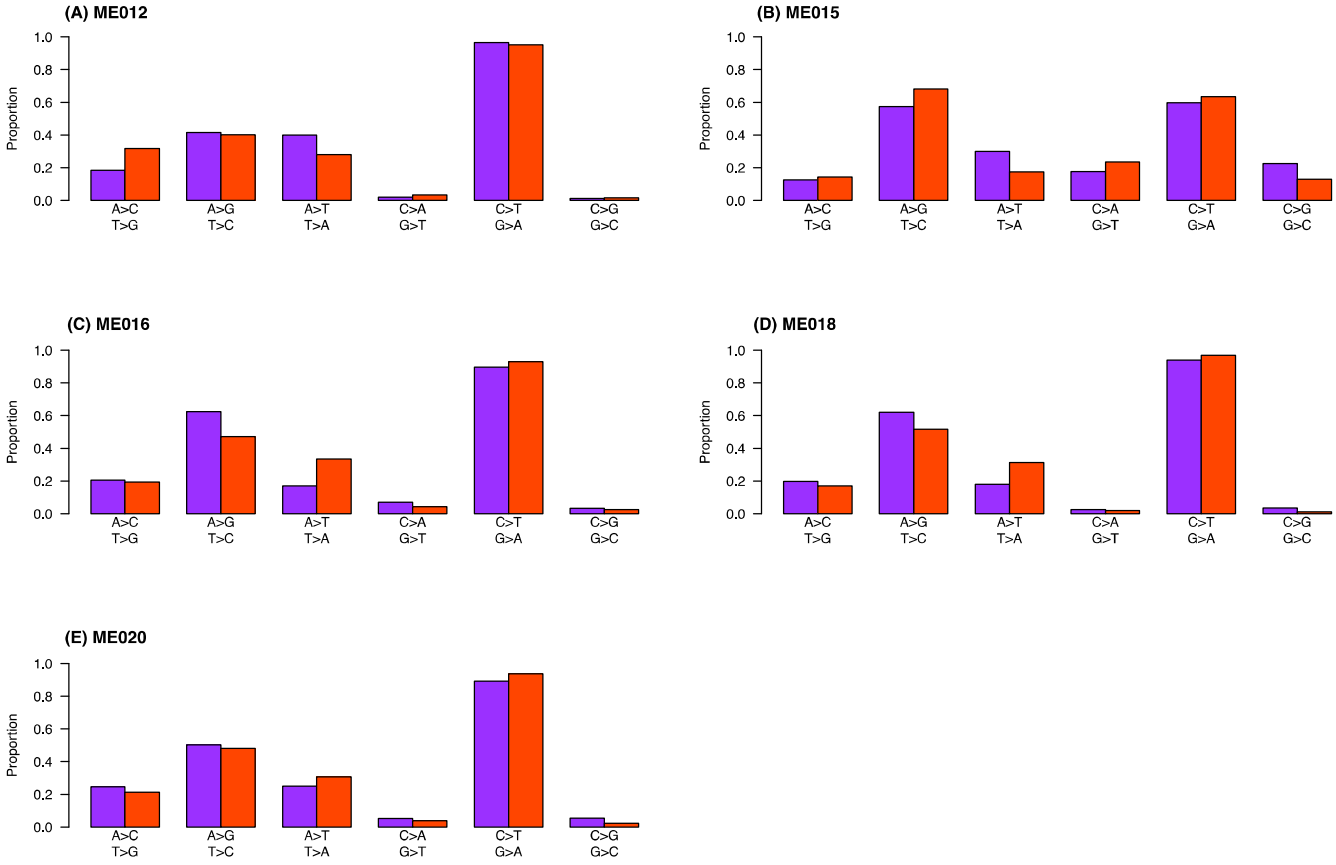
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early replication timing categories. The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. The correlations between mutation frequency in genes and promoters and AIMF were 0.9936674, 0.9999835, 0.9191375, 0.9824988, 0.9699376, 0.9884642, 0.9991288, and 0.9987917 for each of the six cases.

Supplementary Figure S33: 25 melanoma samples[26] Part I (ME001, ME002, ME007, ME009 and ME011): Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing.



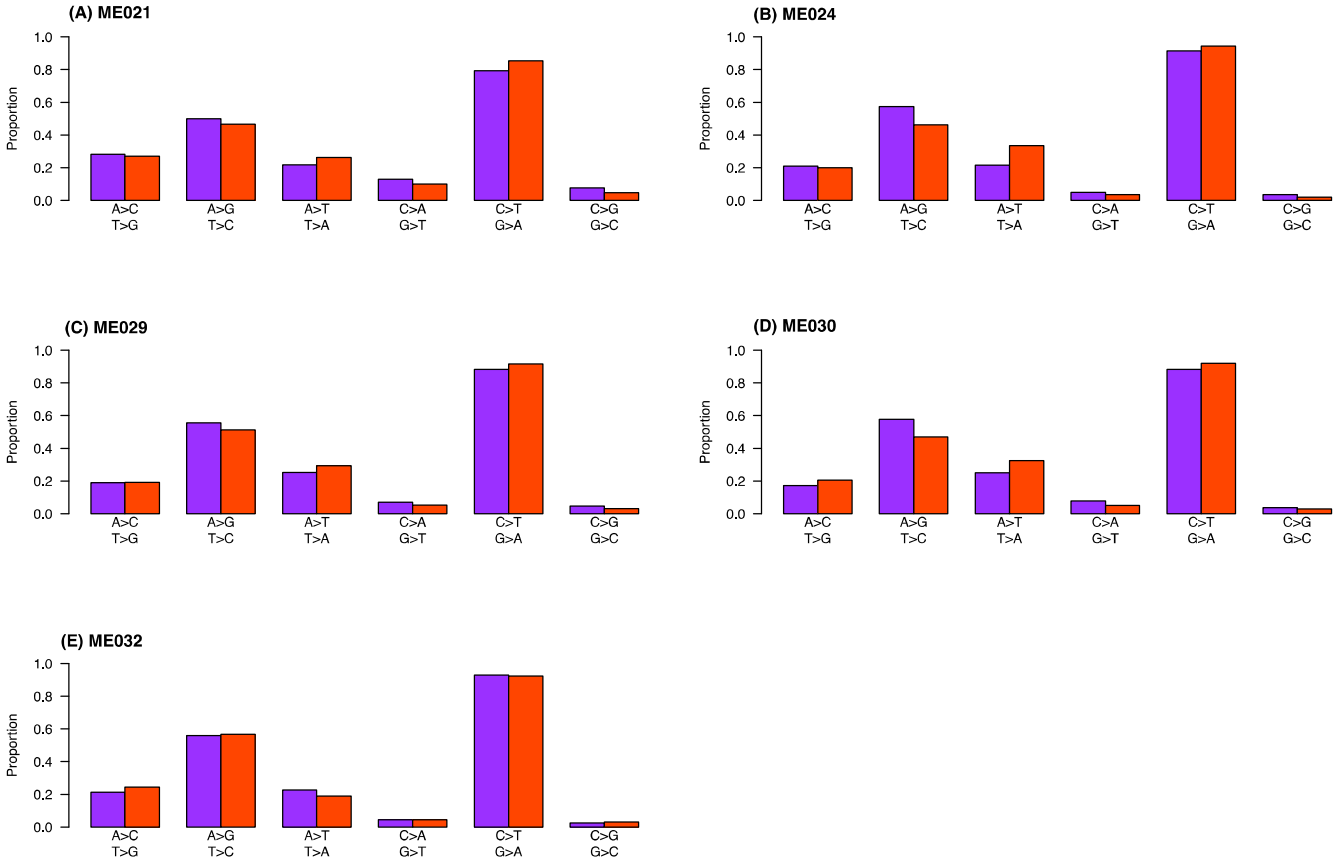
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. (A) ME001, (B) ME002, (C) ME007, (D) ME009 and (E) ME011.

Supplementary Figure S34: 25 melanoma samples[26] Part II (ME012, ME015, ME016, ME018 and ME020): Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing.



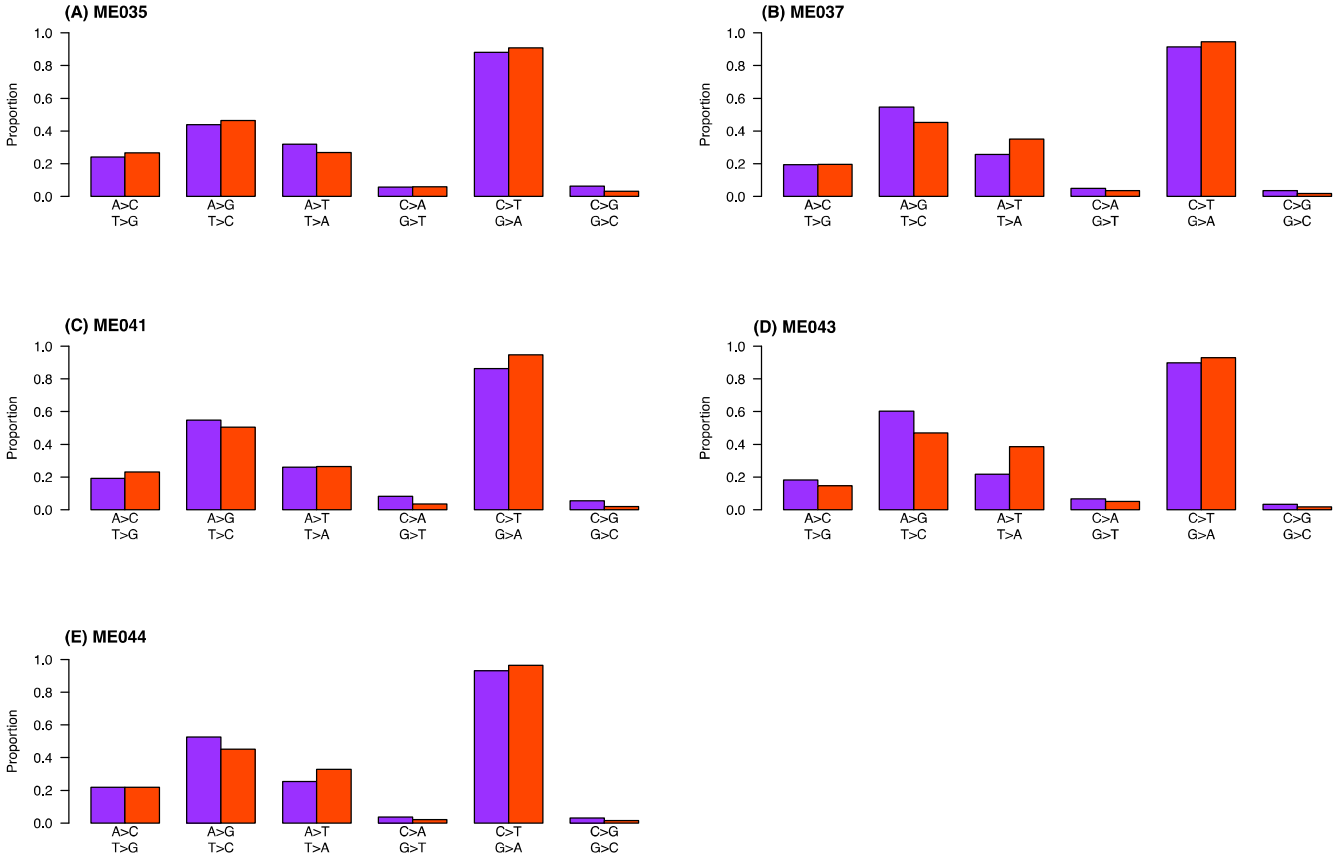
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. (A) ME012, (B) ME015, (C) ME016, (D) ME018 and (E) ME020.

Supplementary Figure S35: 25 melanoma samples[26] Part III (ME021, ME024, ME029, ME030 and ME032): Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing.



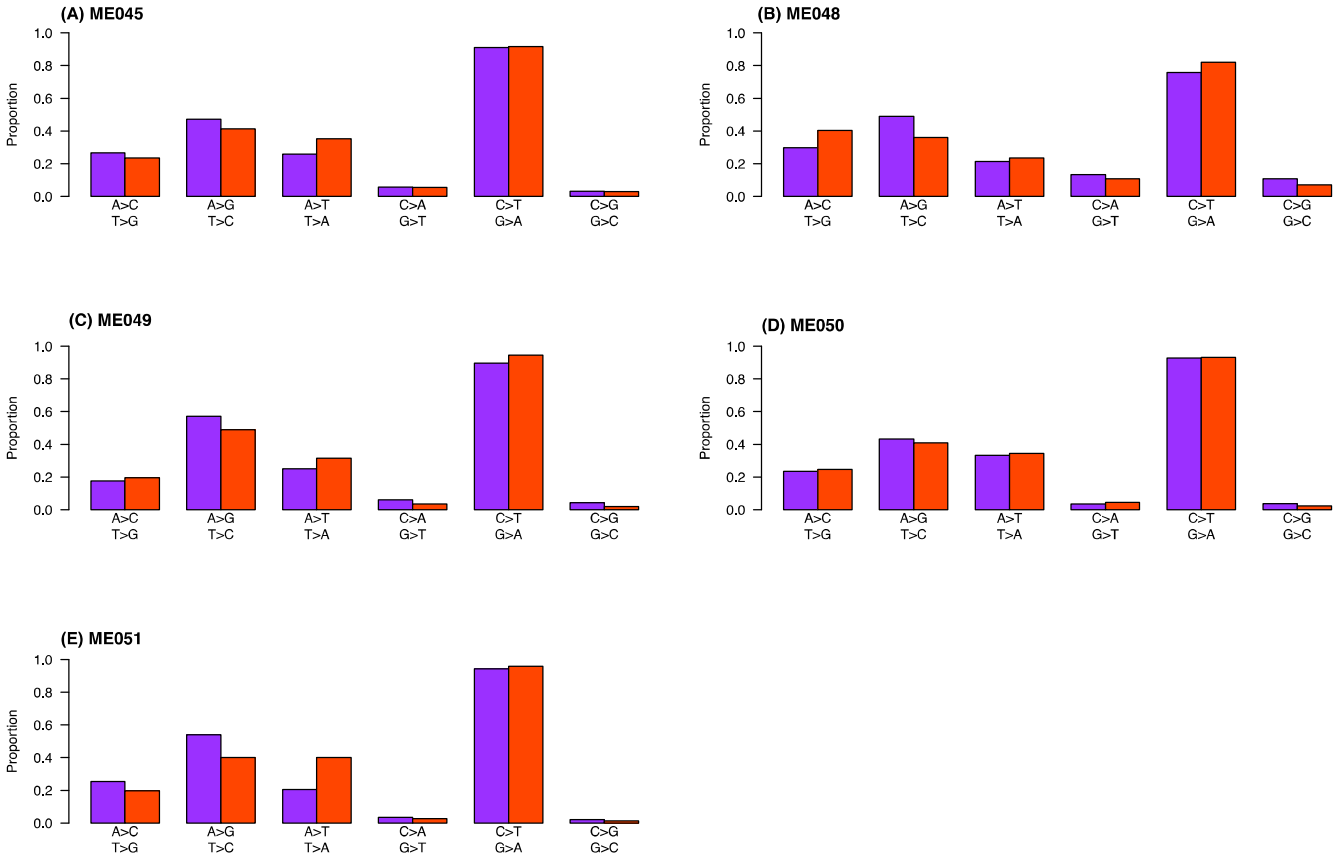
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. (A) ME021, (B) ME024, (C) ME029, (D) ME030 and (E) ME032.

Supplementary Figure S36: 25 melanoma samples[26] Part IV (ME035, ME037, ME041, ME043 and ME044): Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing.



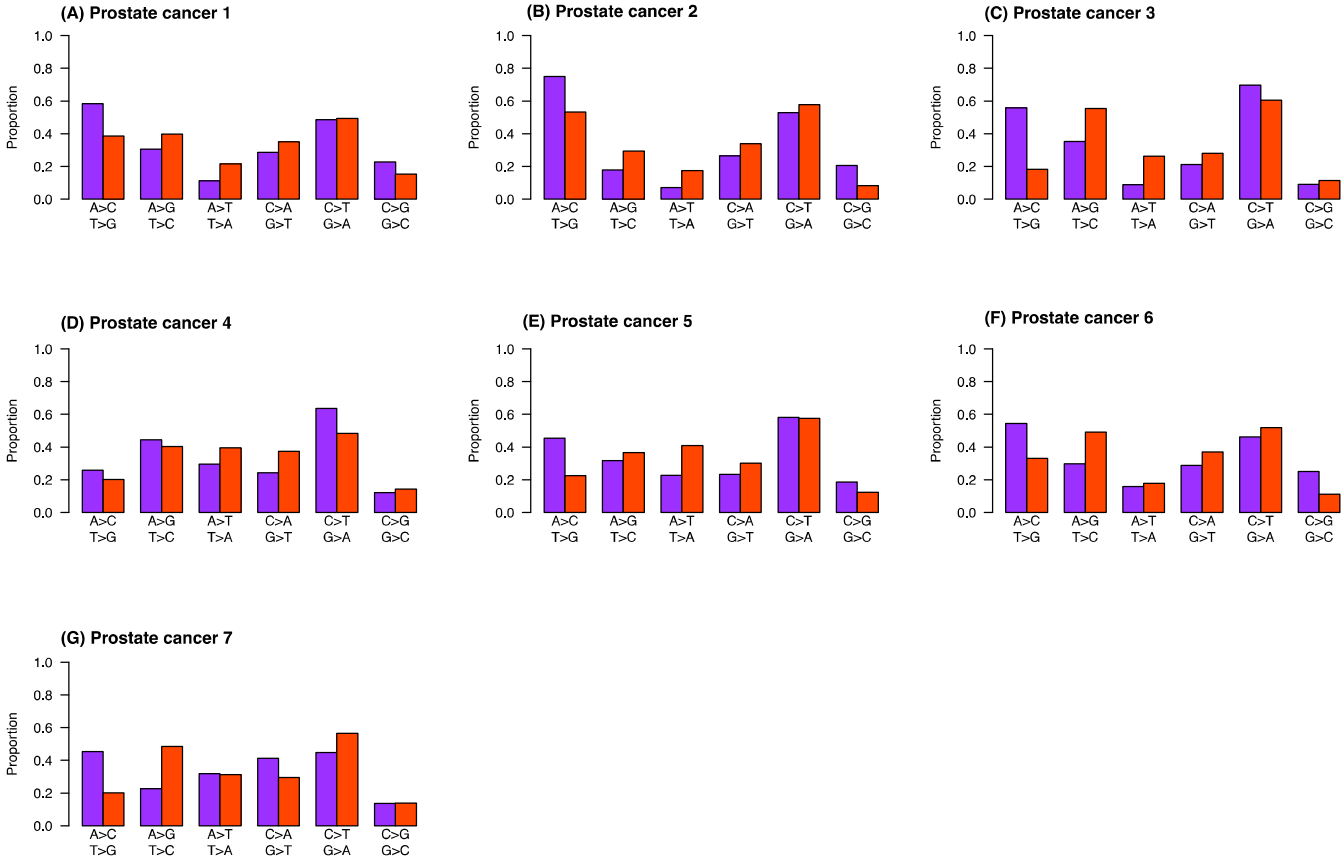
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. (A) ME035, (B) ME037, (C) ME041, (D) ME043 and (E) ME044.

Supplementary Figure S37: 25 melanoma samples[26] Part V (ME045, ME048, ME049, ME050 and ME051): Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing.



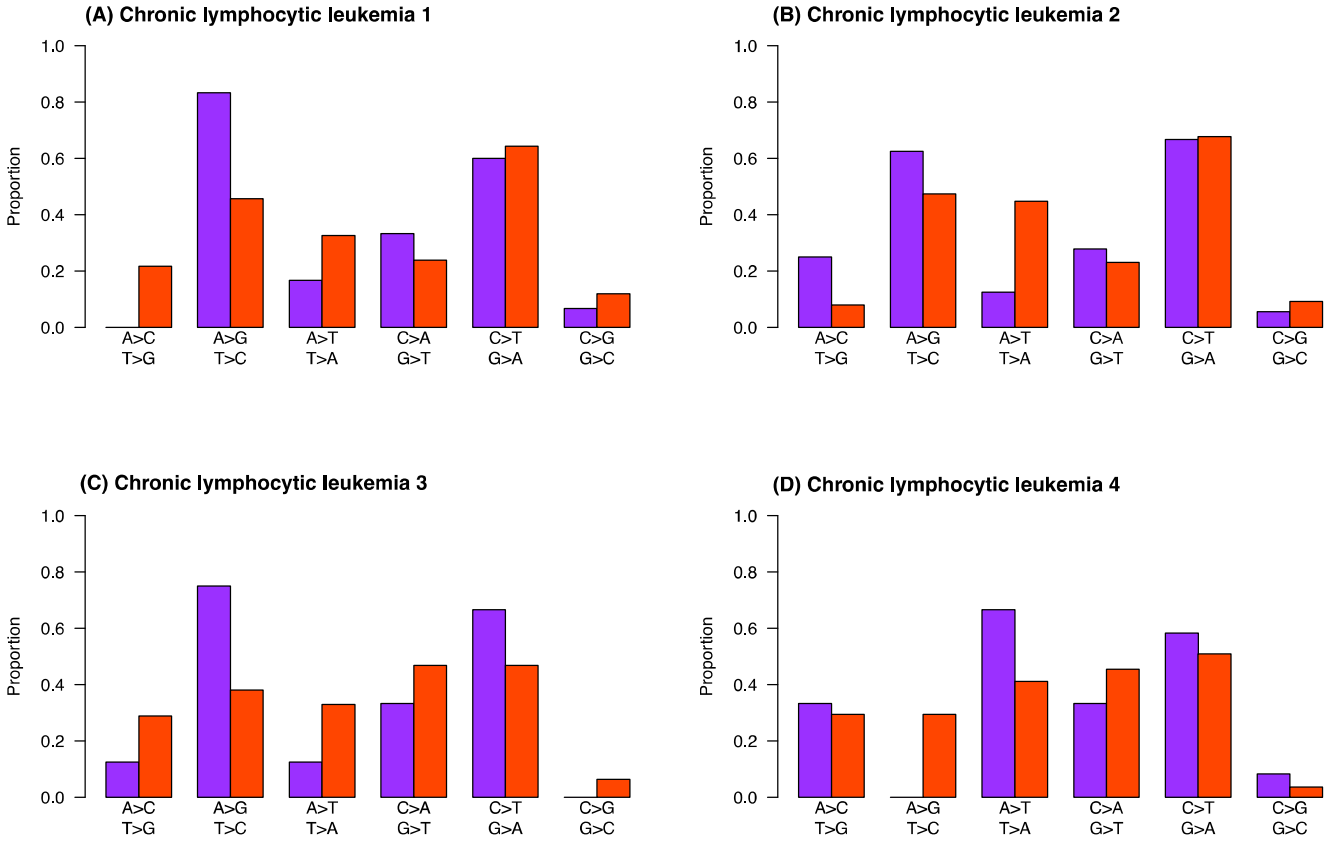
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. (A) ME045, (B) ME048, (C) ME049, (D) ME050 and (E) ME051.

Supplementary Figure S38: 7 prostate cancer samples[27]: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing.



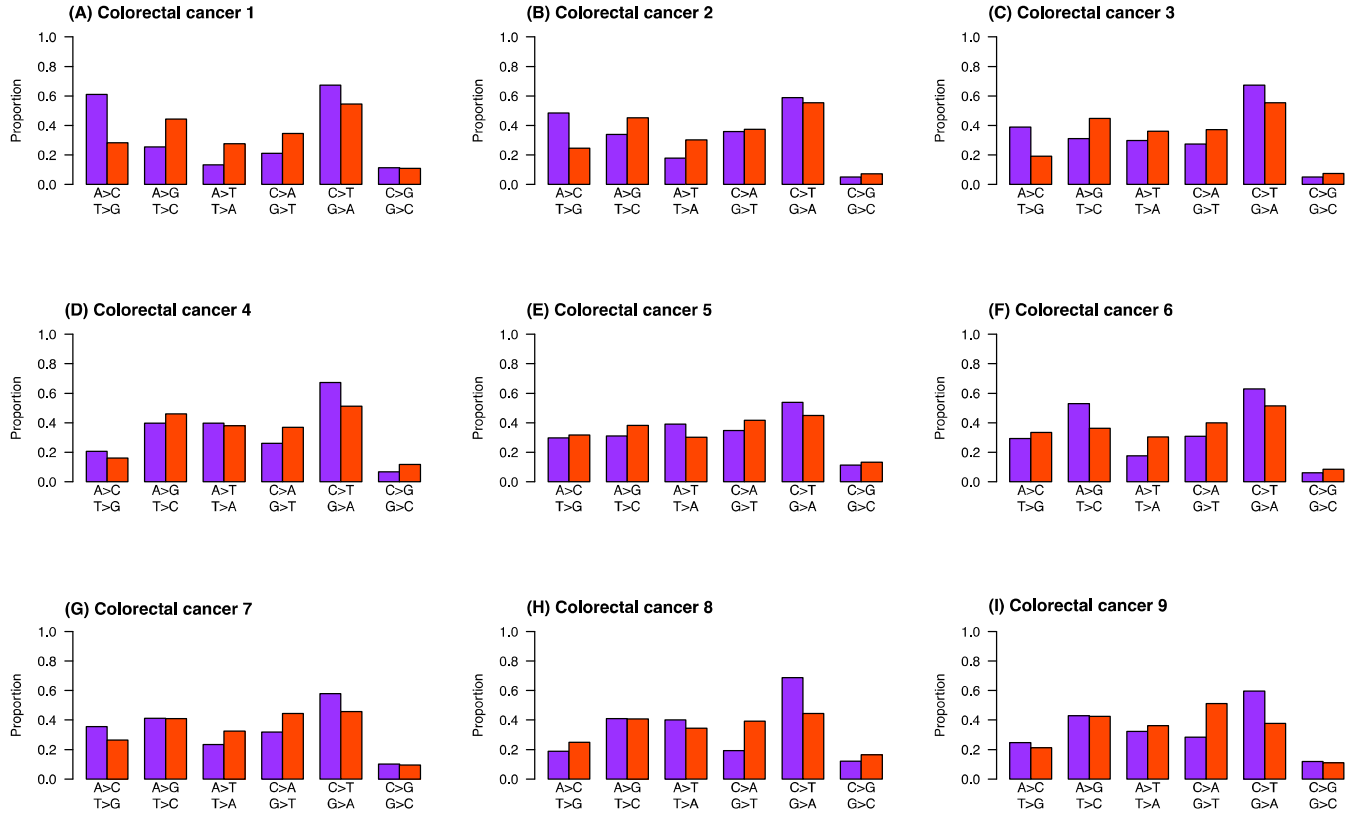
The proportions were calculated based on the hg18 reference allele so that Proportion(A→ C: T→ G) + Proportion(A→ G: T→ C) + Proportion(A→ T: T→ A) = 100%, and Proportion(C→ A: G→ T) + Proportion(C→ T: G→ A) + Proportion(C→ G: G→ C) = 100% for both constant late and constant early categories. (A) – (G): Prostate cancer samples 1 – 7.

Supplementary Figure S39: 4 chronic lymphocytic leukemia samples[29]: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing.



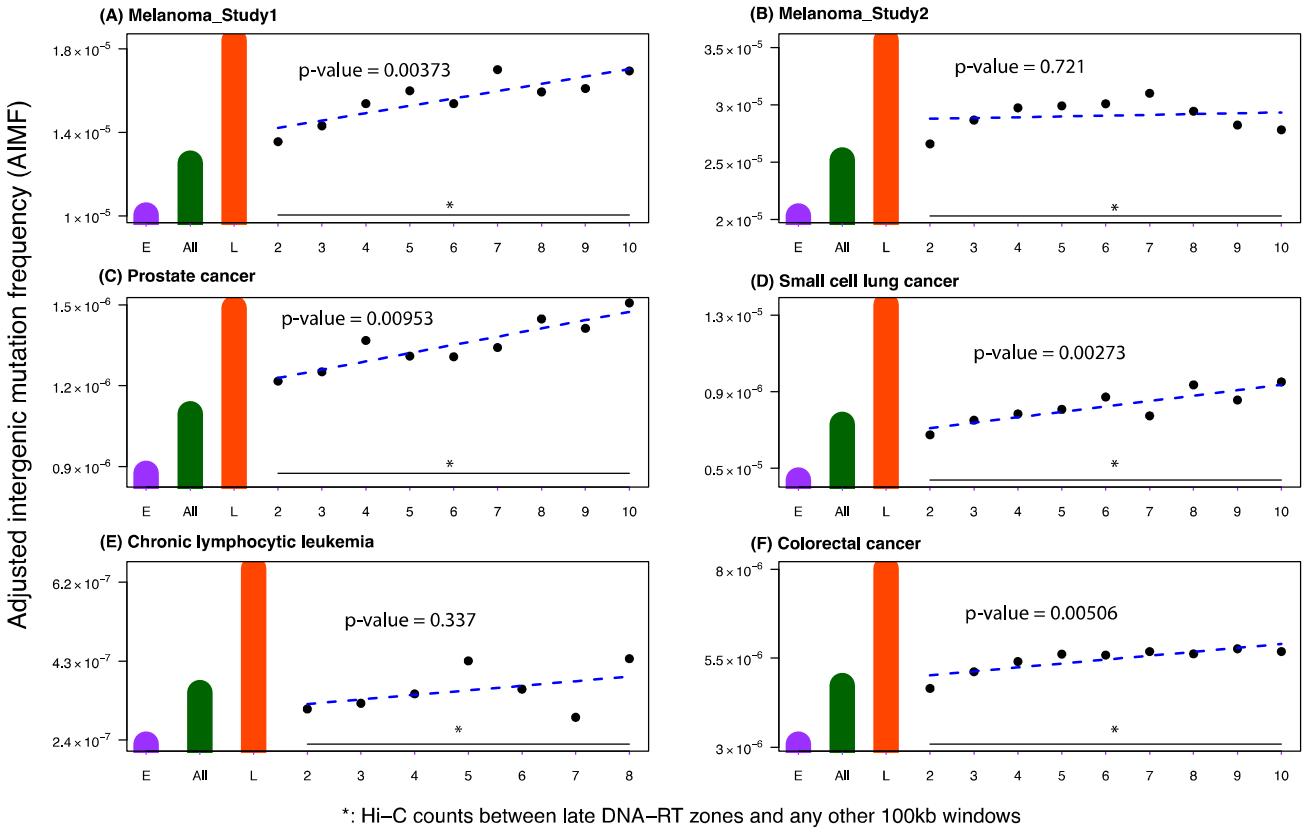
The proportions were calculated based on the hg18 reference allele so that Proportion(A→C: T→G) + Proportion(A→G: T→C) + Proportion(A→T: T→A) = 100%, and Proportion(C→A: G→T) + Proportion(C→T: G→A) + Proportion(C→G: G→C) = 100% for both constant late and constant early categories. (A) – (D): chronic lymphocytic leukemia samples 1 – 4.

Supplementary Figure S40: 9 colorectal cancer samples[30]: Proportion of different types of single nucleotide substitutions in the constant late (orange) and constant early (purple) DNA replication timing.



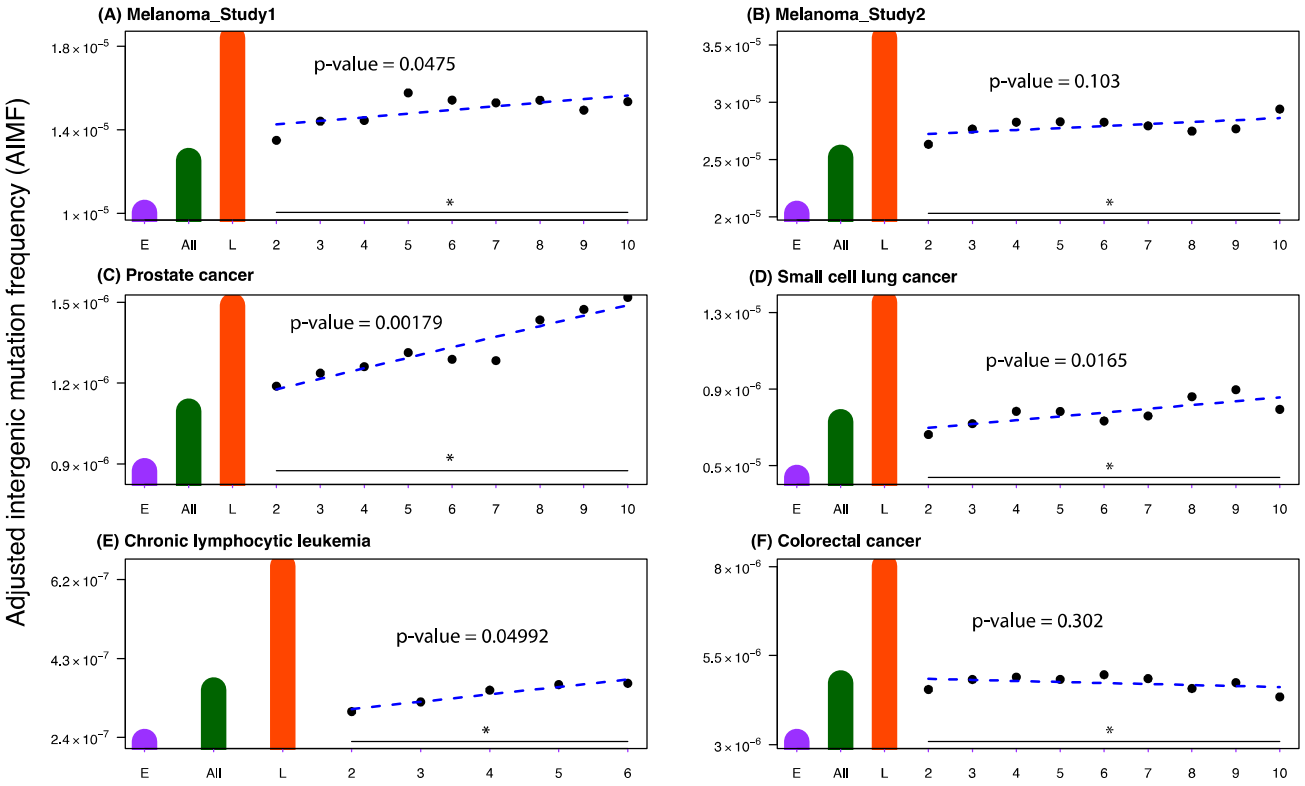
The proportions were calculated based on the hg18 reference allele so that Proportion(A→C: T→G) + Proportion(A→G: T→C) + Proportion(A→T: T→A) = 100%, and Proportion(C→A: G→T) + Proportion(C→T: G→A) + Proportion(C→G: G→C) = 100% for both constant late and constant early categories. (A) – (I): colorectal cancer samples 1 – 9.

Supplementary Figure S41: AIMF in the 'transition-to-late' regions defined by different numbers of Hi-C interaction counts from the HindIII dataset in the GM06990 cell line between regions inside and outside the 'constant late' DNA replication timing zones.



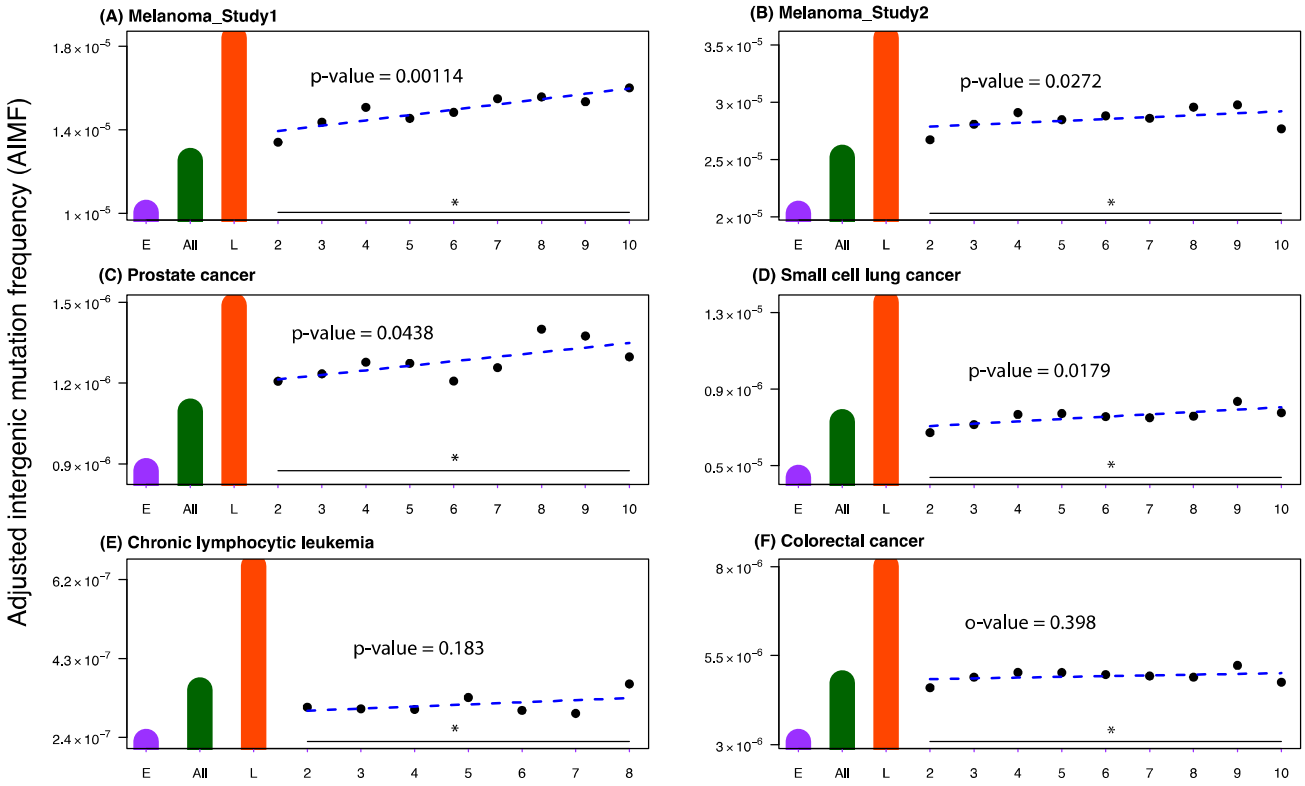
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. Statistical significance was evaluated using simple linear regression, and p-values were obtained. All p-values were less than 0.01 except chronic lymphocytic leukemia (D), probably due to the small number of mutations in this dataset. The green bar shows the genome-wide AIMF, the orange bar the AIMF in constant late DNA replication timing FIR, and the purple bar the AIMF in constant early DNA replication timing FIR. The blue dashed line, i.e. the fitted linear model, shows the positive association between the AIMF and the Hi-C counts. Due to the small number of mutations in the chronic lymphocytic leukemia genome, we only used Hi-C counts from 2 to 8 in panel D. The x-axes display the groups of regions stratified by the number of Hi-C interactions with constant late replication timing zones.

Supplementary Figure S42: AIMF in the 'transition-to-late' regions defined by different numbers of Hi-C interaction counts from the NCoI dataset in the GM06990 cell line between regions inside and outside the 'constant late' DNA replication timing zones.



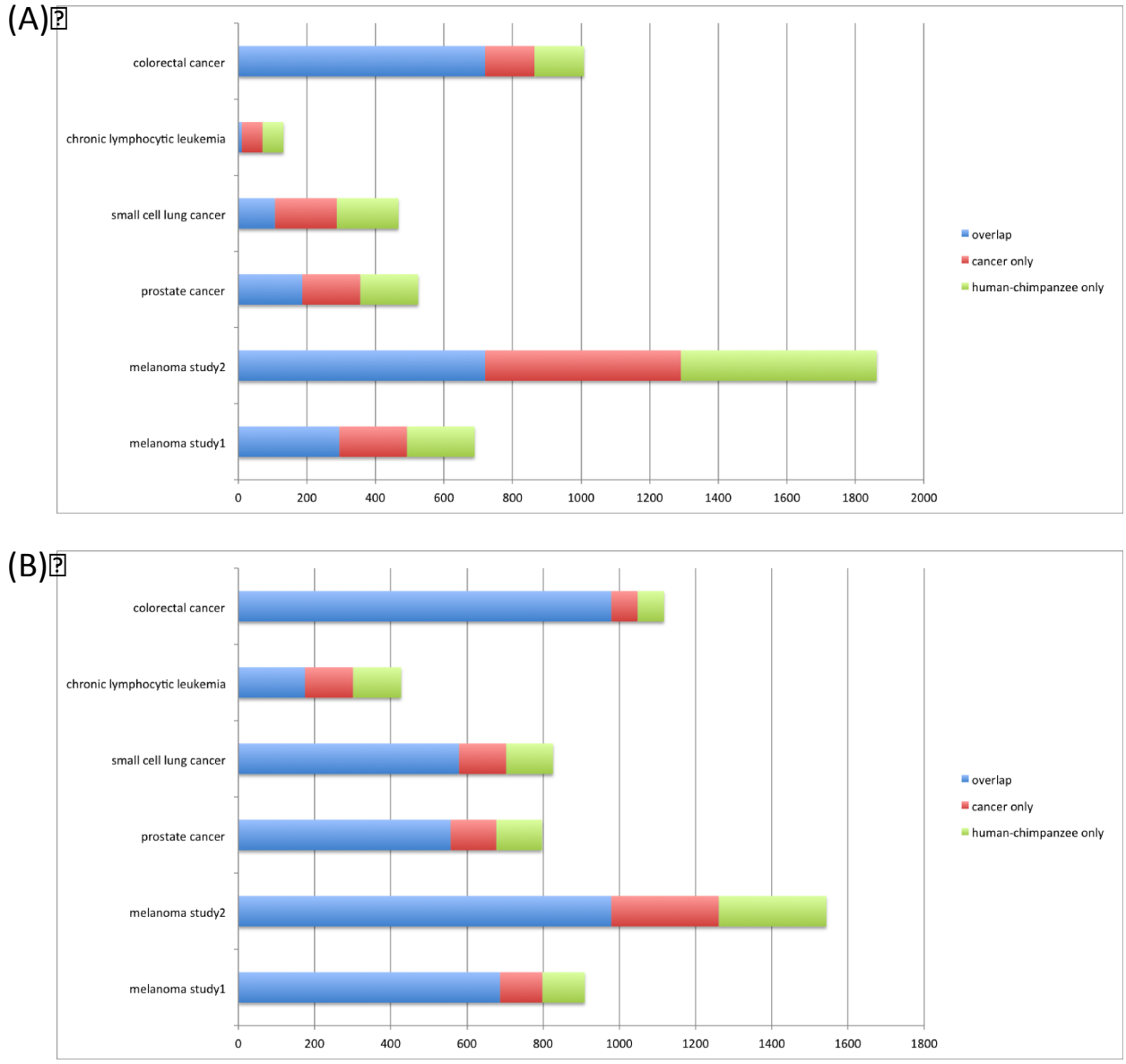
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. Statistical significance was evaluated using simple linear regression, and p-values were obtained. All p-values were less than 0.01 except chronic lymphocytic leukemia (D), probably due to the small number of mutations in this dataset. The green bar shows the genome-wide AIMF, the orange bar the AIMF in constant late DNA replication timing FIR, and the purple bar the AIMF in constant early DNA replication timing FIR. The blue dashed line, i.e. the fitted linear model, shows the positive association between the AIMF and the Hi-C counts. Due to the small number of mutations in the chronic lymphocytic leukemia genome, we only used Hi-C counts from 2 to 8 in panel D. The x-axes display the groups of regions stratified by the number of Hi-C interactions with constant late replication timing zones.

Supplementary Figure S43: AIMF in the 'transition-to-late' regions defined by different numbers of Hi-C interaction counts from the K562 cell line between regions inside and outside the 'constant late' DNA replication timing zones.



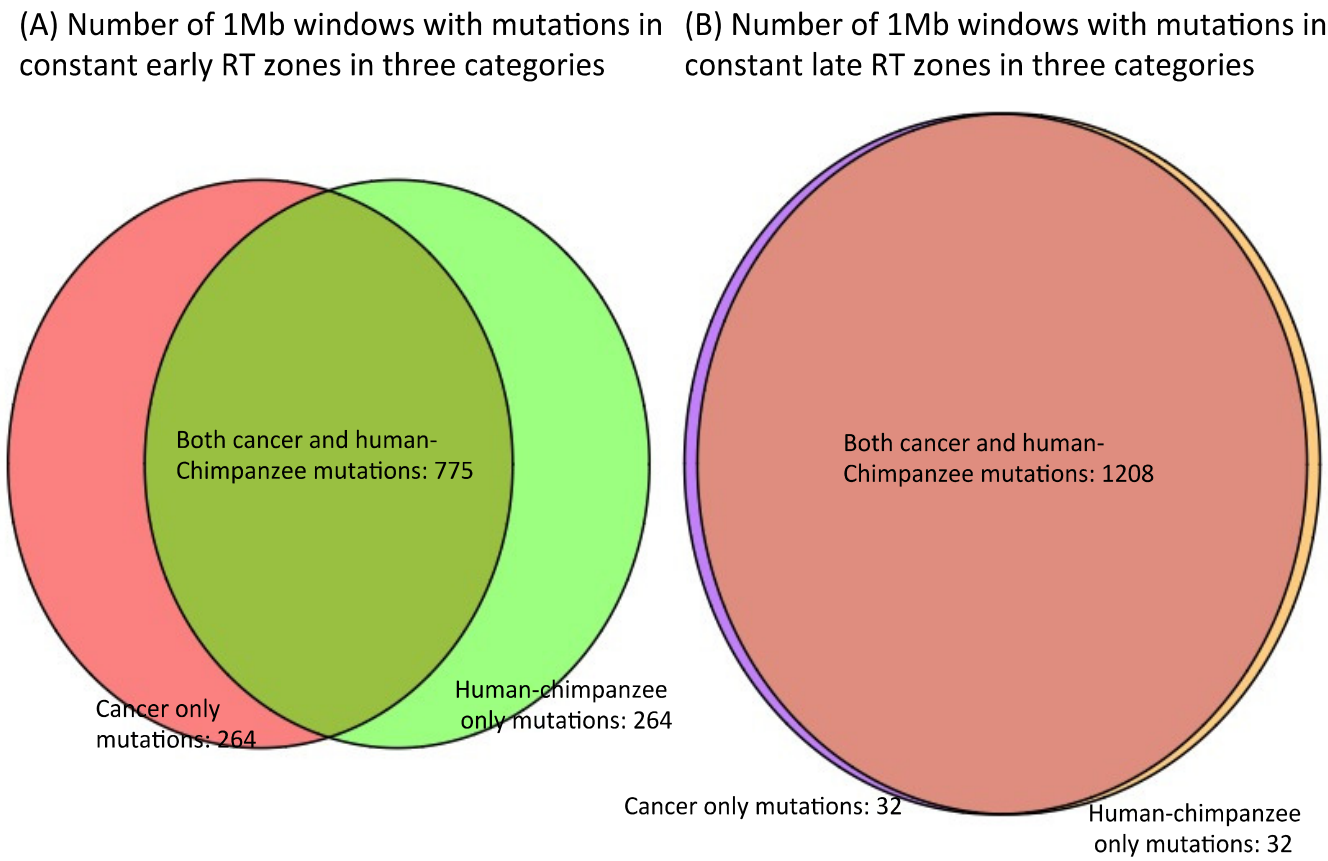
The panels show (A) melanoma samples in study 1[25], (B) melanoma samples in study 2[26], (C) prostate cancer samples[27], (D) small cell lung cancer samples[28], (E) chronic lymphocytic leukemia samples[29], (F) colorectal cancer samples[30], (G) Watson[31] and (H) HuRef genomes[32]. Statistical significance was evaluated using simple linear regression, and p-values were obtained. All p-values were less than 0.01 except chronic lymphocytic leukemia (D), probably due to the small number of mutations in this dataset. The green bar shows the genome-wide AIMF, the orange bar the AIMF in constant late DNA replication timing FIR, and the purple bar the AIMF in constant early DNA replication timing FIR. The blue dashed line, i.e. the fitted linear model, shows the positive association between the AIMF and the Hi-C counts. Due to the small number of mutations in the chronic lymphocytic leukemia genome, we only used Hi-C counts from 2 to 8 in panel D. The x-axes display the groups of regions stratified by the number of Hi-C interactions with constant late replication timing zones.

Supplementary Figure S44: Comparison of genomic regions containing cancer SNS versus human-chimpanzee SNS.



The frequencies of 1Mb non-overlapping windows were analyzed for three categories, defined based on whether they included both cancer and human-chimpanzee SNS, cancer SNS only, or human-chimpanzee SNS only. This analysis was performed for the melanoma, prostate cancer, small cell lung cancer, and chronic lymphocytic leukemia genomes. (A) Comparison of mutations in early DNA replication timing zones and (B) comparison of mutations in late DNA replication timing zones.

Supplementary Figure S45: Comparison of genomic regions containing all cancer SNS versus human chimpanzee SNS.



The number of 1Mb non-overlapping windows were analyzed for three categories, defined based on whether they included both SNS discovered in either of the four cancer types and human-chimpanzee SNS, any cancer SNS only, or human-chimpanzee SNS only. (A) Venn diagram of the three categories for windows with mutations in early DNA replication timing zones and (B) Venn diagram of the three categories for windows with mutations in late DNA replication timing zones.

SUPPLEMENTARY TABLES

Supplementary Table S1. Comparisons of mutation frequencies between cancer types with more than one sample, after adjusting for multiple hypothesis testing.

Adjusted p-value	Prostate cancer	Chronic lymphocytic leukemia	Colorectal cancer
Melanoma Study 2	5.21×10^{-7}	4.63×10^{-7}	7×10^{-6}
Prostate cancer		7.25×10^{-4}	8.04×10^{-16}
Chronic lymphocytic leukemia			2.59×10^{-15}

Supplementary Table S2. The mutation frequencies genome-wide, in gene regions, and in repeat elements across five cancer types. The differences between mutation frequencies in genes and promoters and the genome-wide mutation frequency were significant based on the chi-squared test, with p-values less than 0.0001 in all cancer types analyzed. The mutation frequencies in repeat elements in small cell lung cancer, colorectal cancer, and prostate cancer were also significantly higher than genome-wide mutation frequencies, with chi-squared p-values less than 0.0001 in all cases.

Datasets	Genome-wide mutation frequency (left: early, right: late)	Mutation frequencies in genes and promoters (left: early, right: late)	Mutation frequencies in repeat elements (left: early, right: late)
Melanoma Study 1[25]	6.83×10^{-6} 1.76×10^{-5}	5.91×10^{-6} 1.56×10^{-5}	5.86×10^{-6} 1.75×10^{-5}
Melanoma Study 2	1.97×10^{-5} 3.52×10^{-5}	1.93×10^{-5} 3.33×10^{-5}	1.97×10^{-5} 3.55×10^{-5}
Prostate cancer	1.20×10^{-6} 1.72×10^{-6}	1.13×10^{-6} 1.58×10^{-6}	1.55×10^{-6} 1.98×10^{-6}
Small cell lung cancer	4.73×10^{-6} 1.39×10^{-5}	4.09×10^{-6} 1.09×10^{-5}	5.74×10^{-6} 1.51×10^{-5}
Chronic lymphocytic leukemia	1.87×10^{-7} 5.71×10^{-7}	1.74×10^{-7} 4.86×10^{-7}	1.83×10^{-7} 5.48×10^{-7}
Colorectal cancer	3.34×10^{-6} 8.29×10^{-6}	3.07×10^{-6} 7.13×10^{-6}	4.00×10^{-6} 8.95×10^{-6}

Supplementary Table S3: The AIMF for four cancer types overlaid with *long intergenic noncoding RNAs* (lincRNAs) compared to the genome-wide AIMF using a one-sample binomial test.

	FIR Overlaid with Early RT			FIR Overlaid with Late RT		
	# of mutations	AIMF	p-value	# of mutations	AIMF	p-value
Melanoma Study 1	68	1.24×10^{-5}	0.07902	209	1.71×10^{-5}	0.3327
Melanoma Study 2	112	2.05×10^{-5}	0.5373	428	3.51×10^{-5}	0.1539
Prostate cancer	5	9.14×10^{-7}	0.6044	20	1.64×10^{-6}	0.7899
Small cell lung cancer	14	2.56×10^{-6}	0.04044	149	1.22×10^{-5}	0.2274
Chronic lymphocytic leukemia	1	1.83×10^{-7}	0.8246	9	7.38×10^{-7}	0.5345
Colorectal Cancer	16	2.93×10^{-6}	0.5981	94	7.71×10^{-6}	0.2

SUPPLEMENTARY METHODS

Differences of mutation frequencies among different cancer types

Differences of mutation frequencies among different cancer types: We used analysis of variance to assess the differences of mutation frequencies among the different cancer types (**Supplementary Table S1**) whose sample sets contained multiple samples (prostate and colorectal cancer as well as chronic lymphocytic leukemia and melanoma, see **Table 1**). The p-value for between-group variation was 2.04×10^{-8}.

Distributional difference in the adjusted intergenic mutation frequency (AIMF) in constant early and constant late DNA replication timing zones

Effects of genes and repeats: Since the genome-wide mutation frequency was larger than the AIMF in prostate cancer and small cell lung cancer, we examined whether the removed regions biased our results toward underestimating the mutation frequencies. Thus we calculated the mutation frequencies within genes and repeats separately. We found that the mutation frequencies in repeats in these two cancer types contributed to the higher genome-wide mutation frequencies as compared to the AIMF (**Supplementary Table S2**).

Effects of gene density: We explored whether the differences in the AIMF between early and late DNA replication timing zones could be explained (i.e., confounded) by the variations in gene density for which replication timing might be acting as a surrogate[10, 58, 59]. We segmented the genome into 1Mb non-overlapping bins, and then calculated the fraction of base pairs in each bin that belonged to any non-repetitive exons, using data from Refseq genes in hg18 from the UCSC genome browser[33]. The genome-wide median value for this fraction was 0.02. The bins were classified into "low gene density" and "high gene density" categories according to whether the bin contained a fraction of exonic bases that were above or below this genome-wide median value. Then the AIMF was computed for both low and high gene density categories. We observed that the AIMF was larger in constant late replication timing regions than in constant early replication timing regions (**Supplementary Figures S1** and **S2**) regardless of the gene density categories. Note that in some cases the mutation frequency approached 0, especially in chronic lymphocytic leukemia, since the number of mutations in this case was small. This pattern was consistent across all four analyzed cancer types. Note that there were no remarkable trends of the AIMF across the 23 chromosomes in either category. Although a single sample might not be representative of the mutation patterns in those cancer types, similar patterns were obtained across multiple cancer types and completely sequenced genomes (Mann-Whitney U-Test; p-value < 0.05). The two exceptions were the HuRef genome (p-value 0.1, i.e. marginally significant) and chronic lymphocytic leukemia (p-value = 0.3, potentially due to the small number of mutations, 18 on average). We also obtained consistent results across multiple samples in prostate cancer (7 samples) and chronic lymphocytic leukemia (4 samples), suggesting that the effect of gene density is be a major confounder in this case.

Effects of CpG islands: We next assessed whether the differences in the AIMF between early and late DNA replication timing could be explained by variations in the fraction of CpG islands for which replication timing might be acting as a surrogate[10, 58, 59]. We segmented the genome into 1Mb non-overlapping bins, and calculated the fraction of base pairs in each bin that belonged to any CpG island, using data from the UCSC genome browser[33]. The bins were classified into "low CpG island density" and "high CpG island density" categories according to whether the bin contained a fraction of bases residing in a CpG island that was above or below the genome-wide median value of 0.004. Then the AIMF was computed for both low high CpG island categories. We found that the AIMF was larger in constant late replication timing regions than in constant early replication timing regions (**Supplementary Figures S3** and **S4**) regardless of the two categories. Note that in some cases the mutation frequency approached 0, especially in chronic lymphocytic leukemia, since the number of mutations in this case was small. This pattern was consistent across all four analyzed cancer types and the two completely sequenced personal genomes. There were no remarkable trends of AIMF across the 23 chromosomes in either category.

Although a single sample might not be representative of the mutation patterns in those cancer types, similar patterns were obtained across multiple cancer types and completely sequenced human genomes (Mann-Whitney U-Test, p-value < 0.05), except for HuRef (p-value = 0.07) and chronic lymphocytic leukemia (p-value = 0.1, marginally significant) in the high CpG island category. We also obtained consistent results across multiple samples in prostate cancer (7 samples) and chronic lymphocytic leukemia (4 samples), suggesting that the effect of CpG islands should not be a major confounder in this case.

Effects of GC percentage: We then assessed if the differences in the AIMF between early and late DNA replication timing zones arose due to variations the GC percentage of genomic regions, which was previously shown to affect the mutation rate[10, 58, 59]. In order to examine the effects of variations of the regional GC percentage, we determined the fraction of GC bases for each 1Mb non-overlapping bins. We then binned the distribution of GC percentages into two categories: "low GC percent" and "high GC percent", according to the median GC percentage of 0.39. Then we determined the AIMF for both categories separately. We found that for both low and high GC percent categories, the AIMF was larger in genomic regions in constant late than constant early replication timing zones (**Supplementary Figures S5** and **S6**). There were no remarkable trends of the AIMF across the 23 chromosomes in either category. Although a single sample might not be representative of the mutation patterns in those cancer types, similar patterns were obtained across multiple cancer types and two completely sequenced personal genomes (Mann-Whitney U-Test, p-value < 0.05). We also obtained consistent results across multiple samples in prostate cancer (7 samples) and chronic lymphocytic leukemia (4 samples), suggesting that the effect of GC percentage should not be a major confounder in this case.

Effects of recombination rate: We then explored if the differences in the AIMF between early and late DNA replication timing zones arose due to variations between regions in the recombination rate, which was previously shown to affect the mutation rate[10, 58, 59]. In order to examine the effects of variations in the regional recombination rate, we obtained data of the recombination rate for each 1Mb non-overlapping bin from the UCSC genome browser[33]. We then binned the distribution of recombination rates into two categories: "low recombination rate" and "high recombination rate", according to the median recombination rate of 1.05. Then we determined the AIMF for both categories separately. We found that for both low and high recombination rate categories, the AIMF was larger in genomic regions in constant late replication timing than constant early replication timing zones (**Supplementary Figures S7** and **S8**). There were no remarkable trends of AIMF across the 23 chromosomes in either category. Although a single sample might not be representative of the mutation patterns in those cancer types, similar patterns were obtained across multiple cancer types and two completely sequenced personal genomes (Mann-Whitney U-Test, p-values < 0.05). We also obtained consistent results across multiple samples in prostate cancer (7 samples) and chronic lymphocytic leukemia (4 samples), suggesting that the effect of recombination rate should not be a major confounder in this case.

Effects of chromatin: We next examined whether the differences in the AIMF between early and late DNA replication timing zones arose due to the differences in chromatin composition. Giemsa-staining-based g-banding patterns, obtained from the USCS genome browser, were used as measurements for the chromatin status of 1Mb non-overlapping bins of hg18 genome[33,50]. We then stratified the genome into "G-Neg" and "G-Pos" categories and determined the AIMF for regions that resided in constant early and constant late replication timing zones for either category. We found that the AIMF was larger in constant late replication timing regions than in constant early replication timing regions, irrespective of the chromatin status (**Supplementary Figures S9** and **S10**). There were no remarkable trends of AIMF across the 23 chromosomes in both categories. Although a single sample might not be representative of the mutation patterns in those cancer types, similar patterns were obtained across multiple cancer types and two completely sequenced personal genomes (Mann-Whitney U-Test, p-value < 0.05). We also obtained consistent results across multiple samples in prostate cancer (7 samples) and chronic lymphocytic leukemia (4 samples), suggesting that the effect of chromatin should not be a major confounder in this case.

Effects of nuclear topology: Finally, we assessed whether the differences in the AIMF between early and late DNA replication timing zones arose due to the regions residing in different parts of the nucleus, i.e. in the nuclear core and periphery. We defined the nuclear periphery as those regions that contained lamina-associated regions, using data obtained from Guelen *et al*[11], and the nuclear core as the complement of these regions. We then overlaid the nuclear core and periphery data onto the Filtered Intergenic Regions (FIR) and thereby subdivided the regions into two categories: nuclear core and nuclear periphery. We then calculated the AIMF for both categories and found that the AIMF was higher in constant late than in constant early replication timing zones in both categories (**Supplementary Figures S11 and S12**). There were no remarkable trends of the AIMF across the 23 chromosomes in either category. Although a single sample might not be representative of the mutation patterns in those cancer types, similar patterns were obtained across multiple cancer types and the Watson genome (Mann-Whitney U-Test, p-value < 0.05). The p-value for the HuRef genome was 0.098 in the nuclear core and 0.18 in the nuclear periphery. We also obtained consistent results across multiple samples in prostate cancer (7 samples) and chronic lymphocytic leukemia (4 samples), suggesting that the effect of nuclear topology should not be a major confounder in this case. Since the dynamics of nuclear topology in different cell types could affect our results, we also used a bootstrap sampling approach to randomly draw 80% of the lamina associated domains 1,000 times, and repeated the analysis above. We obtained consistent results after controlling for the simulated data (**Supplementary Figure S13**).

Permutation analysis: We performed permutation analyses to investigate the statistical significance of our findings for different mutation frequencies in late vs. early replication timing regions. A schematic illustration of the permutation analysis is shown in **Supplementary Figure S14**. We chose a certain number of mutations in constant replication timing zones (i.e., constant early and constant late) that overlapped with FIR regions; we then randomly selected the same number of mutations from these regions. We repeated this random selection 1,000 times to calculate the simulated AIMF for constant late and constant earl' DNA replication timing zones. Thus the null distribution of AIMF in late or early DNA replication timing zones is given by the histograms emerging from the simulation analyses. Then, based on the ranking of our observed AIMF compared with its null distribution, we calculated the proportion of the simulated results that was more extreme than the observed results as the permutation p-value. We display the null distribution for AIMF in constant early and constant late DNA replication timing zones for melanoma in panel (C) and (D) of **Supplementary Figure S14**, respectively, with actual mutations in melanoma marked by a triangle. **Supplementary Figure S15** Panels (A) through (J) show the results for all genomes analyzed.

Other DNA replication timing datasets: We utilized the replication timing data of lymphoblastoid, myoblast, ESC, NPC, definitive endoderm and mesendoderm cell lines from the replication domain database[24]. The corresponding data are from Ryba et al.[35] and Pope et al.[36]. We then determined the AIMF based on the new replication timing data (**Supplementary Figure S16**). Compared to using the original replication timing data[23], our results were robust. All differences of mutation frequencies between early and late replication timing regions were significant based on the chi-squared test with p-values less than 0.01.

Effects of replication domains at the 1Mb scale: We then examined whether the different mutation frequencies in late versus early DNA replication timing zones could be affected by the relative proportion of late replication timing regions within 1Mb non-overlapping windows (**Supplementary Figure S17**). We calculated the proportion of late replication timing basepairs in non-overlapping 1Mb windows and stratified these windows into those that harbored a large proportion of late RT material and those that harbored a small proportion of late RT material. We then calculated the AIMF for early and late replication timing in the two strata. The result indicates that in the stratum with a large proportion of late RT material, the mutation frequencies were higher than in the stratum with a small proportion of late RT material (with chi-squared test p-values < 0.001 in six datasets), but the differences between specific early and late replication timing regions were not confounded, i.e. still remained robust (chi-squared test p-values < 0.001 in both groups in six datasets).

Effects of relative positions of DNA replication timing zones: We then examined whether the different mutation frequencies in late versus early DNA replication timing zones could be affected by the relative positioning of the mutation in each replication timing segment, i.e. whether it resided at the center or the border of each DNA replication timing region. We identified 40Kb regions in the middle and at the boundary of each constant late and constant early DNA replication timing zone and then calculated the AIMF for each of the four categories: early replication timing-center of the region, early replication timing-boundary of the region, late replication timing-center of the region, and late replication timing-boundary of the region. We observed consistent differences of mutation frequencies even after stratifying the data by the relative positioning within DNA replication timing zones (**Supplementary Figure S18**).

Stratification analyses by CTCF-binding sites, common fragile sites, and G-quadruplex secondary structures: We performed a set of stratification analyses using data on the genomic location of CTCF-binding sites[52], common fragile sites, and consensus sequences for G-quadruplex secondary structures[51]. The genome was divided into 1Mb non-overlapping windows and for each factor, its frequency of occurrence within each window was calculated. Using the median fraction as the cutoff, we classified the windows into two categories: "large proportion of factor x" and "small proportion of factor x" categories. We then calculated the AIMF for each category (**Supplementary Figure S19**).

Enrichment of AIMF in large intergenic noncoding RNA regions: We then aimed to investigate whether the AIMF depended on the presence of large intergenic noncoding RNAs (lincRNAs) in certain genomic regions. The catalogue of human lincRNAs was obtained from Cabili *et al*[43]. We overlaid the lincRNA regions with the FIR regions that had constant early versus constant late DNA replication timing. We found that 5,469,654 base pairs out of 79,233,321 constant early FIR base pairs were reported to be lincRNA regions, whereas 12,192,941 base pairs out of 169,496,567 constant late FIR base pairs were lincRNA regions. We observed that the AIMF in these two regions was not significantly different from the genome-wide AIMF; this finding was consistent across the four cancer types analyzed (**Supplementary Table S3**).

Differences in the mutation spectrum between constant late and constant early DNA replication timing zones

Effects of gene density: After observing that the mutation spectra (or mutation transversion signatures) differed between constant late and constant early DNA replication timing zones, we examined whether these differences were due to variations in gene density. We segmented the genome into 1Mb non-overlapping bins and calculated the fraction of base pairs that belonged to any non-repetitive exons, using data from Refseq genes in hg18 from the UCSC genome browser[33]. The genome-wide median value for this fraction was 0.02. The bins were classified into low and high gene density categories according to whether the bin contained a fraction of exonic bases that was above or below this genome-wide median value. For each category and each bin, we computed the proportions of the six types of single nucleotide substitution signatures, A→C: T→G, A→G: T→C, A→T: T→A, C→A: G→T, C→T: G→A, and C→G: G→C, in the FIR residing in constant late and constant early replication timing zones, respectively. We then examined the substitution patterns for both low and high gene density categories and found that for a given cancer type, some types of substitutions were consistent irrespective of the local gene density (**Supplementary Figures S20** and **S21**). Additionally, the differences in the mutation spectra were often similar across the four cancer types and the two completely sequenced personal genomes. Given the above analyses, we conclude that while gene density is associated with the chromatin environment[60], the mutation spectrum patterns observed are unlikely due to the effects of gene density alone.

Effects of CpG island density: We then examined whether the differences in the mutation spectra between constant early and constant late replication timing zones was due to variations in the fraction of CpG islands in corresponding genomic regions. We again segmented the genome into 1Mb non-overlapping bins and calculated the fraction of base pairs that belonged to CpG islands, using data obtained from the UCSC genome browser[33]. The

bins were classified into "low CpG island" and "high CpG island" categories according to whether the bin contained a number of bases belonging to a CpG island that was above or below the genome-wide median value. For each category and each bin, we computed the proportion of six types of single nucleotide substitution signatures, A→ C: T→ G, A→ G: T→ C, A→ T: T→ A, C→ A: G→ T, C→ T: G→ A, and C→ G: G→ C, in the FIR residing in constant late and constant early replication timing zones, respectively. We then examined the substitution patterns for both low high CpG island density categories and found that for a given cancer type, some types of substitutions were consistent irrespective of the local fraction of bases belonging to CpG islands (**Supplementary Figures S22** and **S23**). Additionally, the differences in the mutation spectra were often similar across the four cancer types and the two completely sequenced personal genomes. Given the above analyses, we concluded that the mutation spectra patterns are unlikely due to the effects of CpG island density alone.

Effects of GC percentage: We then examined whether the differences in the mutation spectra between constant early and constant late replication timing zones was due to variations of the GC percentage in the corresponding genomic regions. We segmented the genome into 1Mb non-overlapping bins and determined the GC percentage for each of these bins, using data obtained from the UCSC genome browser[33]. The bins were classified into "low GC percent" and "high GC percent" categories according to whether the bin contained a fraction of G and C nucleotides that were above or below the genome-wide median value of the GC percentage. For each category and each bin, we computed the proportions of six types of single nucleotide substitution signatures, A→ C: T→ G, A → G: T → C, A → T: T → A, C → A: G → T, C → T: G → A, and C → G: G → C, in the FIR residing in constant late and constant early replication timing zones, respectively. We then examined the substitution patterns for both low high GC percent categories and found that for a given cancer type, some types of substitutions were consistent irrespective the local GC percentage (**Supplementary Figures S24 and S25**). Additionally, the differences in the mutation spectra were often similar across the four cancer types and the two completely sequenced personal genomes. Given the above analyses, we concluded that the mutation spectrum patterns are unlikely due to the effects of GC percentage alone.

Effects of recombination rate: We then examined whether the differences in the mutation spectra between constant early and constant late replication timing zones was due to variations of the recombination rate in the corresponding genomic regions. We again segmented the genome into 1Mb non-overlapping bins, and determined the recombination rate for each of these bins, using data obtained from UCSC genome browser[33]. The bins were classified into "low recombination rate" and "high recombination rate" categories according to whether the bin had a recombination rate that was above or below the genome-wide median value. For each category and each bin, we computed the proportion of six types of single nucleotide substitution signatures, A→ C: T→ G, A→ G: T→ C, A→ T: T→ A, C→ A: G→ T, C→ T: G→ A, and C→ G: G→ C, in the FIR residing in constant late and constant early replication timing zones, respectively. We then examined the substitution patterns for both low and high recombination rate categories and found that for a given cancer type, some types of substitutions were consistent irrespective of the local recombination rate (**Supplementary Figure S26** and **S27**). Additionally, the differences in the mutation spectra were often similar across the four cancer types and the two completely sequenced personal genomes. Given the above analyses, we concluded that the mutation spectra patterns are unlikely due to the effects of recombination rate alone.

Effects of chromatin: We then examined whether the differences in the mutation spectra between constant early and constant late replication timing was due to differences in the chromatin composition of the corresponding genomic regions. We again obtained Giemsa-staining based g-banding patterns for each 1Mb non-overlapping bin of the hg18 genome from the UCSC genome browser[33]. We stratified the genome into "G-Neg" and "G-Pos" categories and for each category, we computed the proportions of six types of single nucleotide substitution signatures, A→ C: T→ G, A→ G: T→ C, A→ T: T→ A, C→ A: G→ T, C→ T: G→ A, and C→ G: G→ C, in the FIR residing in constant late and constant early replication timing zones, respectively. We then examined the substitution patterns for both G-Neg and G-Pos categories and found that for a given cancer type, some types of

substitutions were consistent irrespective the local chromatin status (**Supplementary Figures S28** and **S29**). Additionally, the differences in the mutation spectra were often similar across the four cancer types and the two completely sequenced personal genomes. Given the above analyses, we concluded that the mutation spectra patterns are unlikely due to the effects of chromatin composition alone.

Effects of nuclear topology: Finally, we examined whether the differences in the mutation spectra between constant early and constant late replication timing zones was due to differences in the nuclear topology, i.e. whether genomic regions resided within the nuclear core or periphery. We again defined the nuclear periphery as the genomic material containing lamina-associated regions, based on data obtained from Guelen *et al*[11], and the nuclear core as the remainder of the genome. We then overlaid the nuclear core and periphery data onto the FIR regions and thereby defined two categories: nuclear core and periphery. For each category, we computed the proportions of six types of single nucleotide substitution signatures, A→ C: T→ G, A→ G: T→ C, A→ T: T→ A, C→ A: G→ T, C→ T: G→ A, and C→ G: G→ C, in the FIR residing in constant late and constant early replication timing zones, respectively. We then examined the substitution patterns for both the nuclear core and nuclear periphery categories and found that for a given cancer type, some types of substitutions were consistent irrespective of the nuclear topology (**Supplementary Figure S30** and **S31**). Additionally, the differences in the mutation spectra were often similar across the four cancer types and the two completely sequenced personal genomes. Given the above analyses, we concluded that the mutation spectra patterns are unlikely due to the effects of nuclear topology alone.

Mutation signatures in genes and promoters: We then compared the mutation signatures within genes and promoters and within adjusted intergenic regions defined in our study (**Supplementary Figure S32**). The mutation frequencies in genes and promoters exhibited very similar patterns as compared to the adjusted intergenic mutations, with a correlation > 0.9 in all six cancer sample sets and two personal genomes.

The mutation signature patterns for all 47 samples surveyed in our analysis: To examine specifically which of the mutation signatures displayed the most common differences between early and late replication timing zones, we investigated all 25 melanoma samples, 7 prostate cancer samples, 9 colorectal cancer samples and 4 chronic lymphocytic leukemia samples. We found that A->T|T->A is the most common signature with a higher proportion in late replication timing zones (in 38/47 samples) (**Supplementary Figures S33-S40**).

The effects of higher-order chromatin structure on differences in mutation frequencies in early/late replication timing zones

We then aimed to investigate the effects of the higher-order genome organization onto the observed mutation patterns. We divided the genome into 100Kb non-overlapping windows and obtained Hi-C-based long-range interaction data from GM06990 (generated by HindIII and NCoI restriction cutting enzyme separately) and K562 cell lines from Lieberman-Aiden *et al*[12]. We excluded all interactions between two loci that were less than 20Kb apart from each other. We then stratified all pairs of windows according to the interaction counts (specifying the numbers between 2 and 10) between them and counted the windows linked with but outside of the constant late DNA replication timing zones for each such stratum. The regions that overlapped with the FIR were termed 'transition to late' regions for each interaction count between 2 – 10 (**Supplementary Figures S41-S43**).

Comparison between cancer and human-chimpanzee single nucleotide substitutions

We then aimed to investigate the differences between regions that harbored single nucleotide substitutions (SNS) in the cancer genomes and the regions that harbored SNS when comparing the human hg18 and chimpanzee panTro2[61] genomes. We utilized a similar approach as in Stamatoyannopoulos et al[18] with the axtToNet software. We obtained around $4x10^7$ substitutions in total across the autosomes[44]. The genome was again divided into non-overlapping 1Mb windows and the number of mutations of each type was counted. We then compared an equal

number of windows containing cancer SNS and windows containing human-chimpanzee SNS. We classified those windows into three categories depending on which types of mutations they contained: both cancer and human-chimpanzee SNS, cancer SNS only, and human-chimpanzee SNS only. The fractions of these three types of windows appeared to be different between early and late DNA replication timing zones, with a chi-squared p-value < 0.001 (**Supplementary Figure S44**). Next, after collapsing the windows with SNS in either of the four cancer types together, we found 1,039 such windows with SNS in early DNA replication timing zones. We then selected the same number of windows with the highest number of SNS. Out of these 1,039 windows, 775 overlapped in cancer and human-chimpanzee groups (**Supplementary Figure S45**). We then performed similar analyses in late DNA replication timing zones, and found that out of 1,240 windows, 1,208 overlapped in cancer and human-chimpanzee groups. Therefore, we concluded that, at the scale of 1Mb, most of the regions with human-chimpanzee SNS were also the regions with cancer SNS, suggesting some common mechanisms between evolutionary transversion and mutagenesis in cancers with no obvious differences in early and late DNA replication timing zones.

SUPPLEMENTARY REFERENCES

58. Tyekucheva, S. et al. Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol* **9**, R76 (2008).
59. Hellmann, I. et al. Why do human diversity levels vary at a megabase scale? *Genome Res* **15**, 1222-1231 (2005).
60. Surralles, J., Sebastian, S. & Natarajan, A.T. Chromosomes with high gene density are preferentially repaired in human cells. *Mutagenesis* **12**, 437-442 (1997).
61. Karolchik, D. et al. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51-54 (2003).