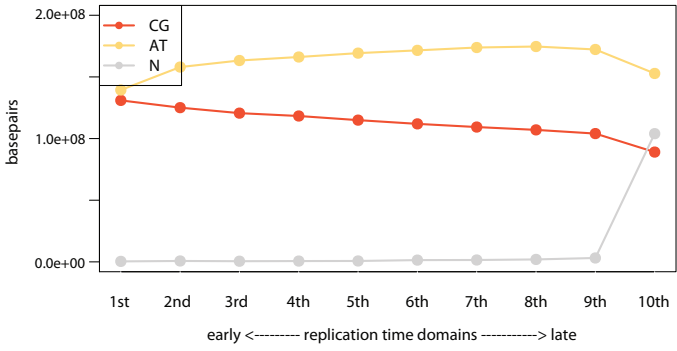
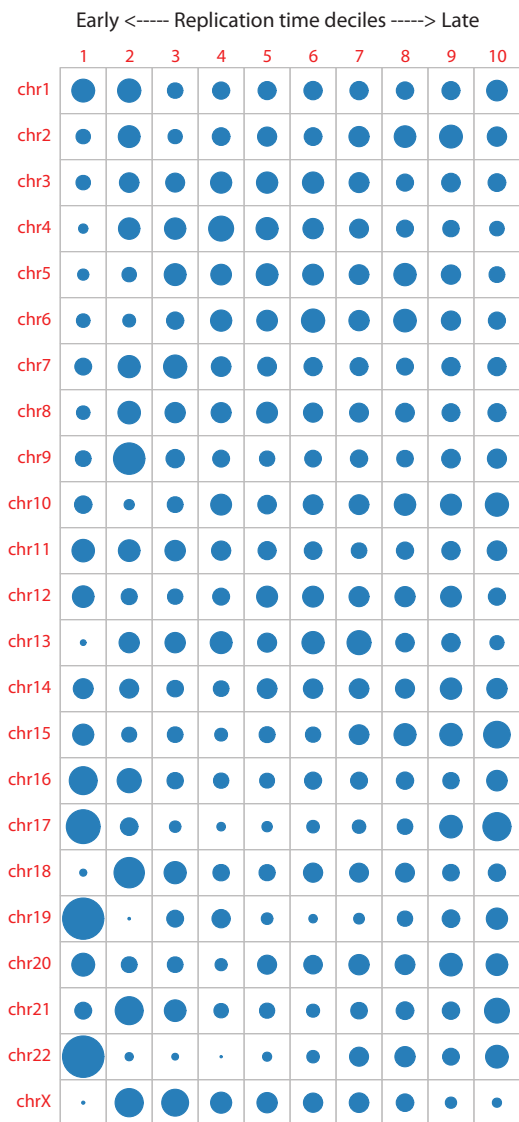
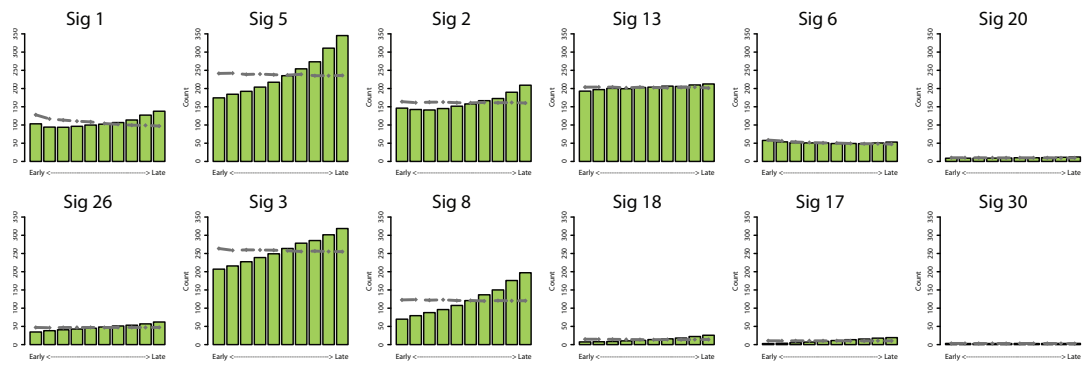


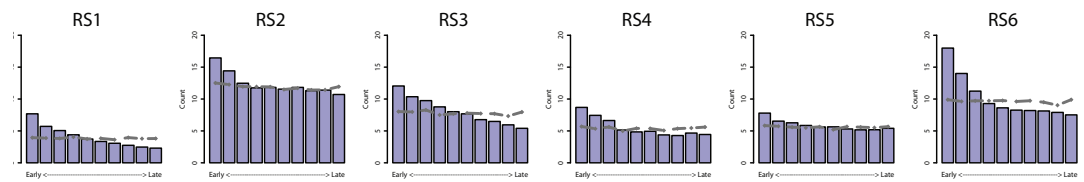
A

**B**

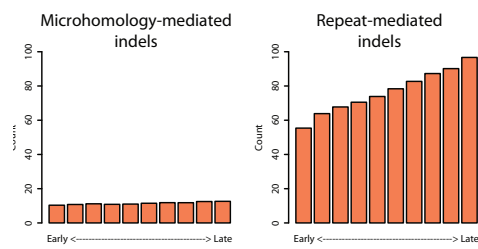
C



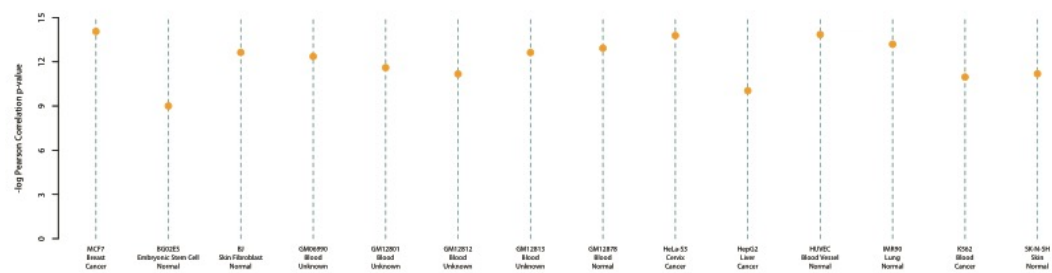
D

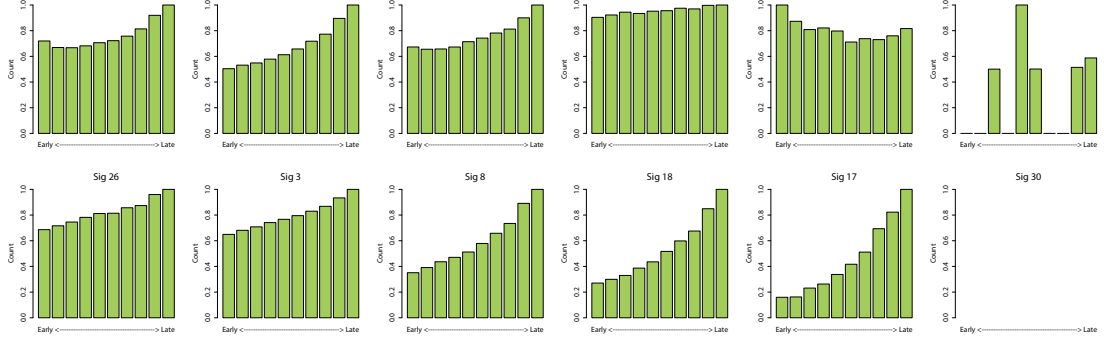
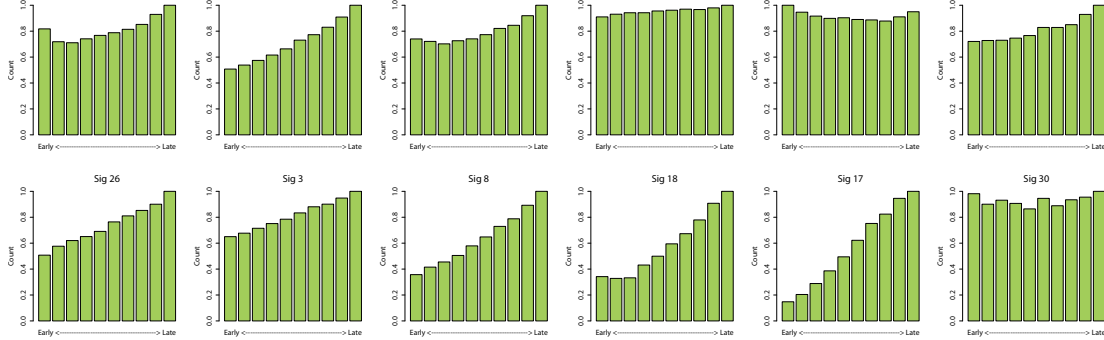
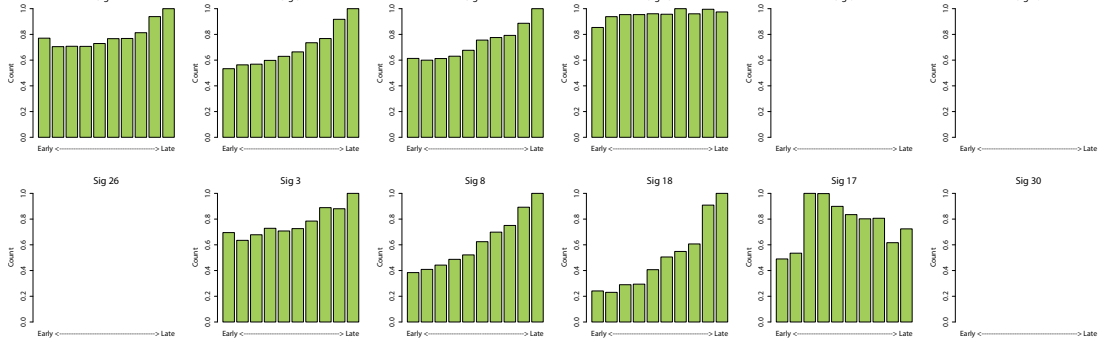
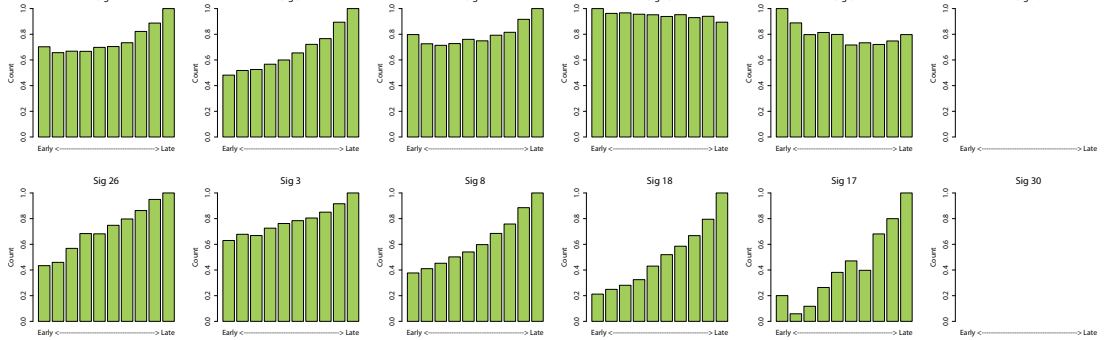


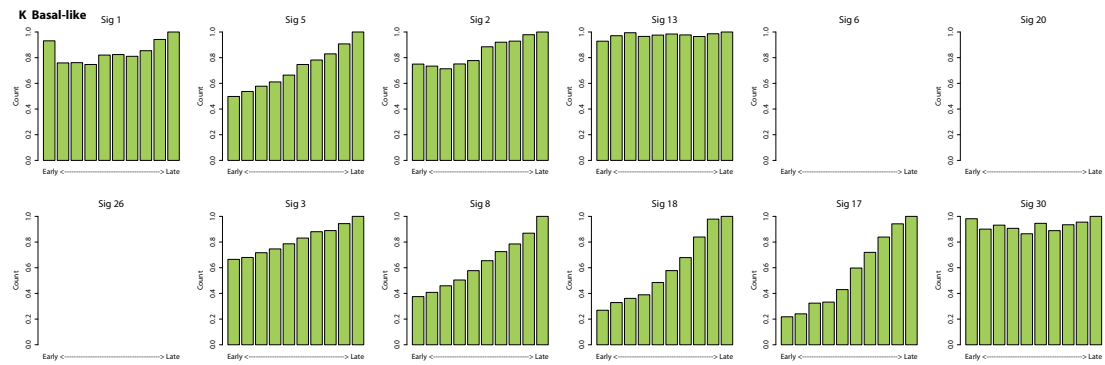
E



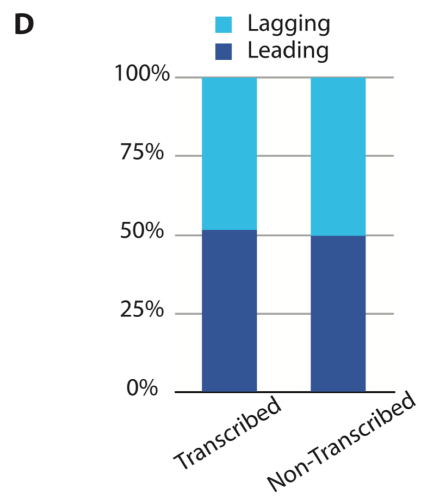
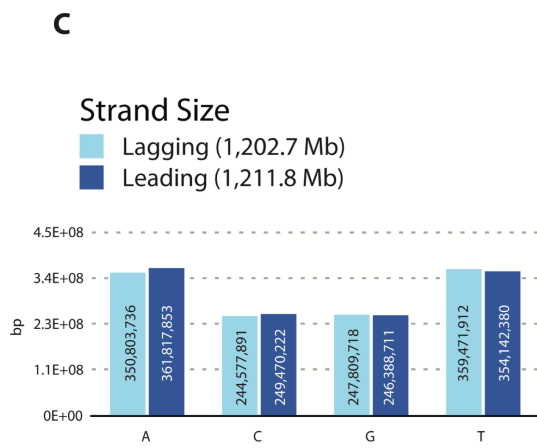
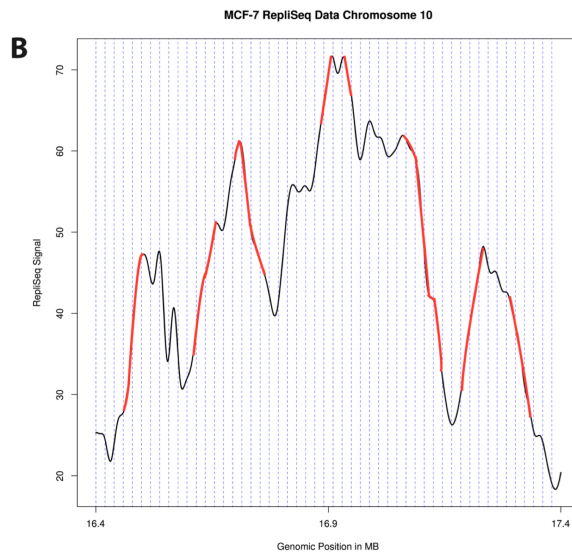
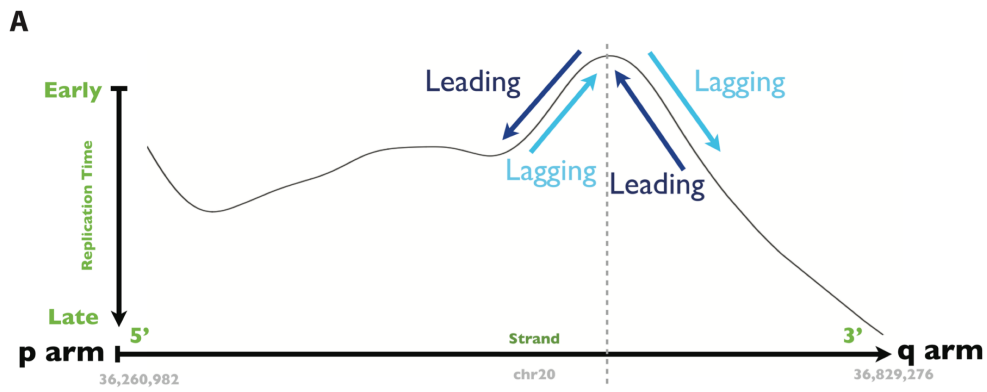
F



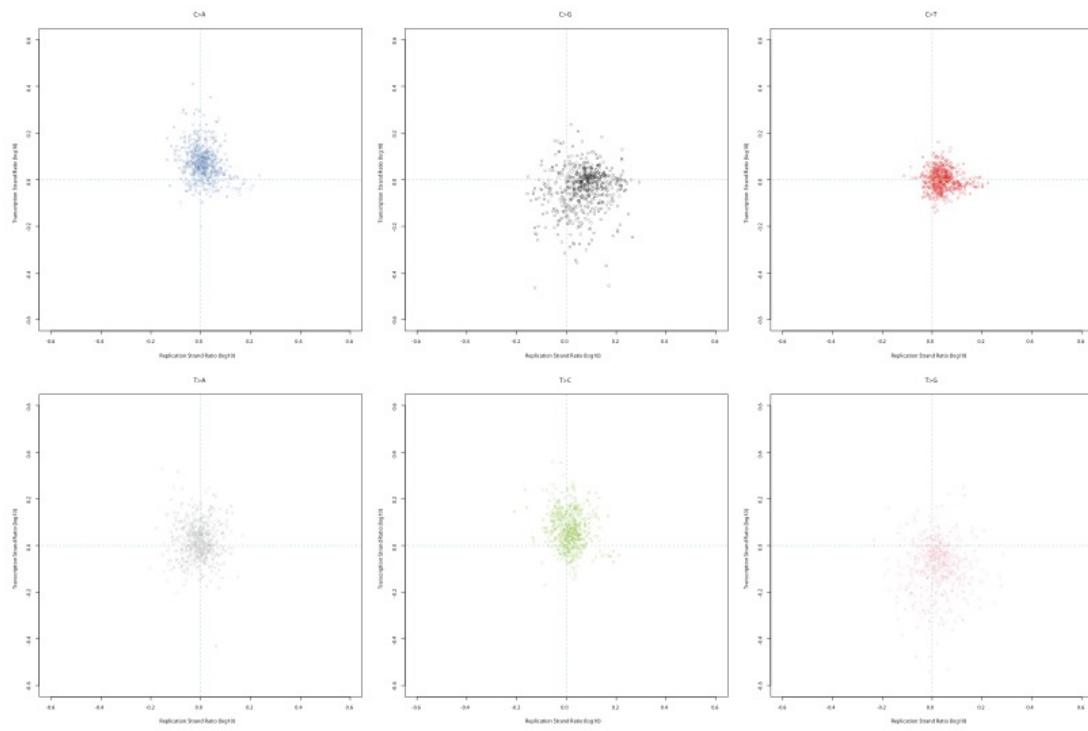
G ER POSITIVE**H ER NEGATIVE****I Luminal A****J Luminal B**



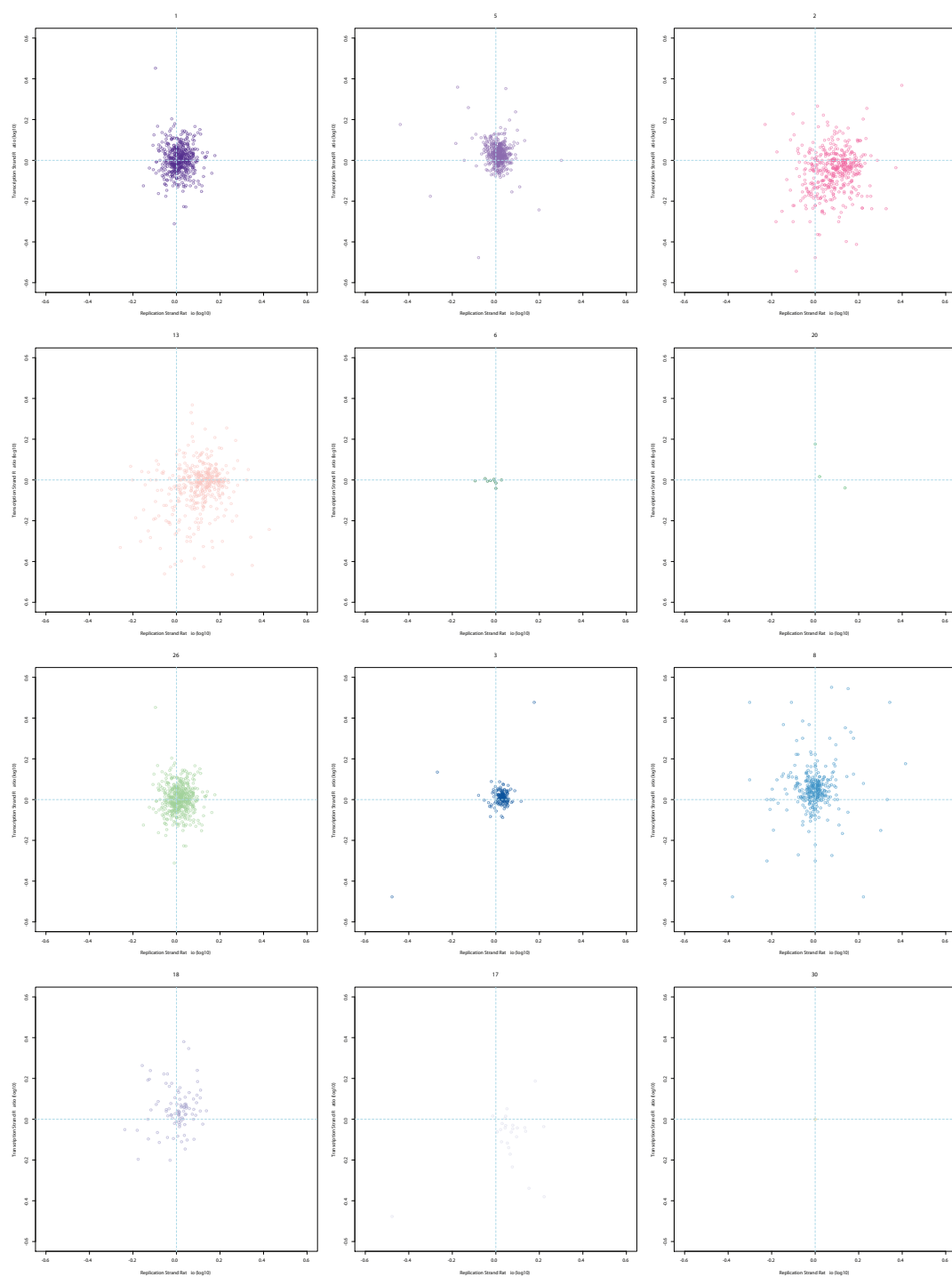
Supplementary Fig. 1 | Characterization of replication time domains. A, Distribution of CG/AT and N across the replication time deciles extracted from MCF-7 Repli-seq data. **B,** Overview of distribution of replication time domains across chromosomes. The size of each circle is proportional to the percentage of DNA in the chromosome (row) that is present in a specific decile (column). Absolute counts across the replication time deciles of **C,** mutations **D,** rearrangements and **E,** indels by signature. The dashed-diamond lines show the trend expected from the randomization experiments. **F,** Exploring correlations between replication time and somatic mutations across fourteen cell lines. Y-axis shows the $-\log$ of the p -value obtained from a Pearson's correlation test that estimates the relationship between aggregated base substitutions and replication timing domains for each cell line. **G-K,** Replication time analysis for the twelve base substitution signatures, where samples were grouped together by using ER status (ER positive 366 samples and ER negative 194 samples), and gene expression profiles LumA (101 samples), LumB (75 samples) and Basal (74 samples) respectively [see ¹⁵ for further details on complete cohort].

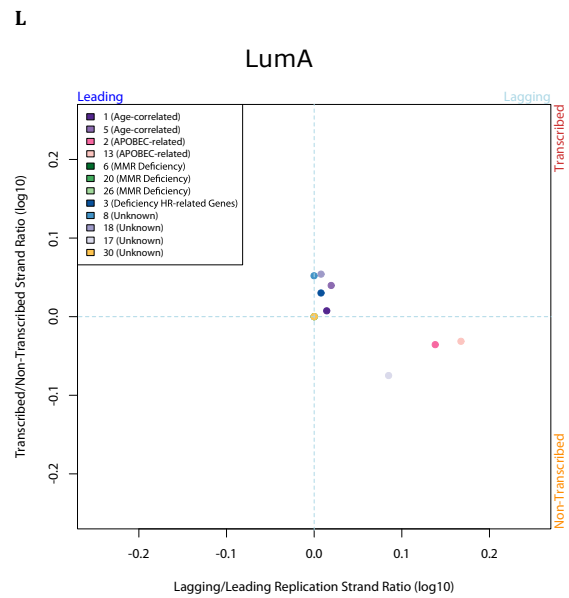
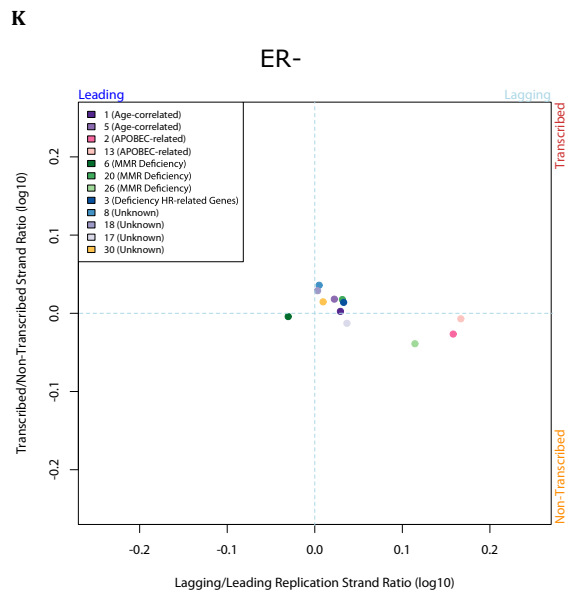
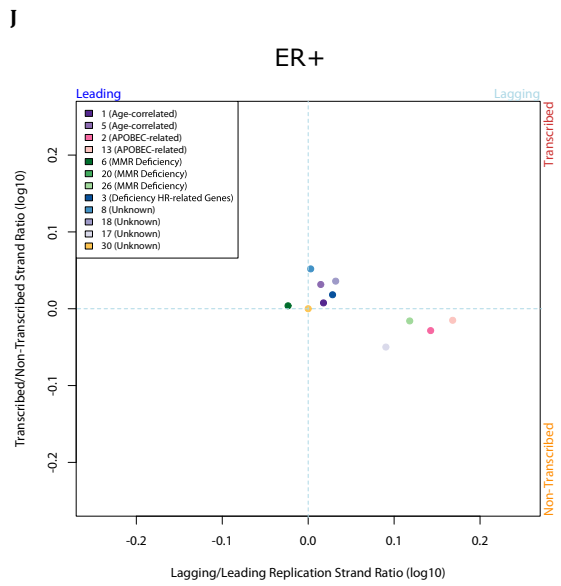
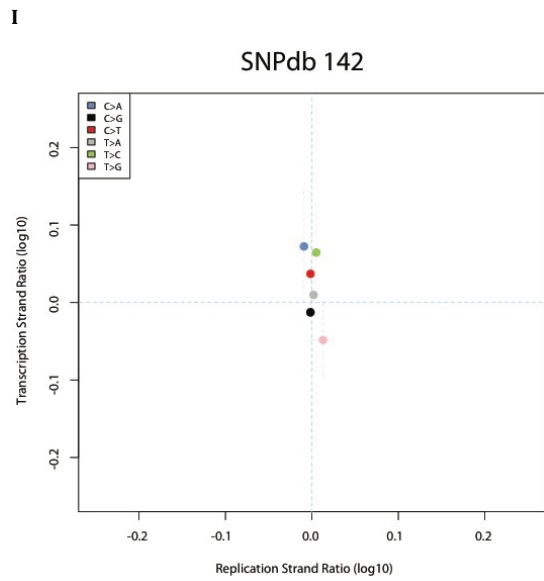
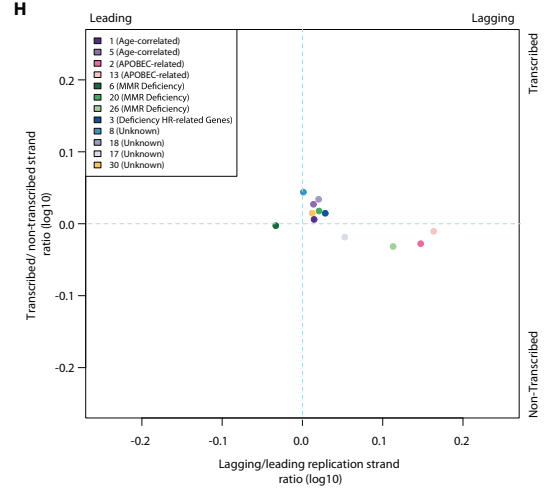
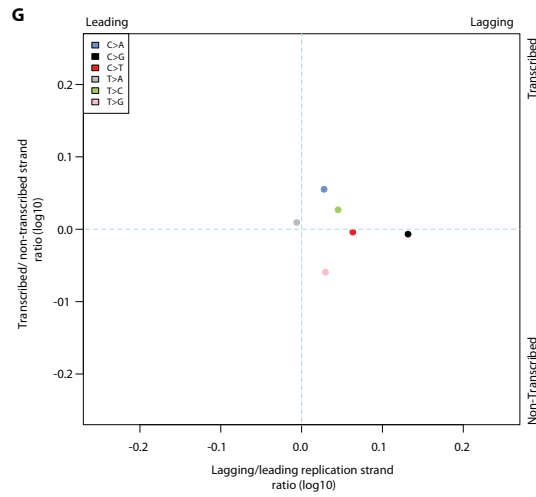


E

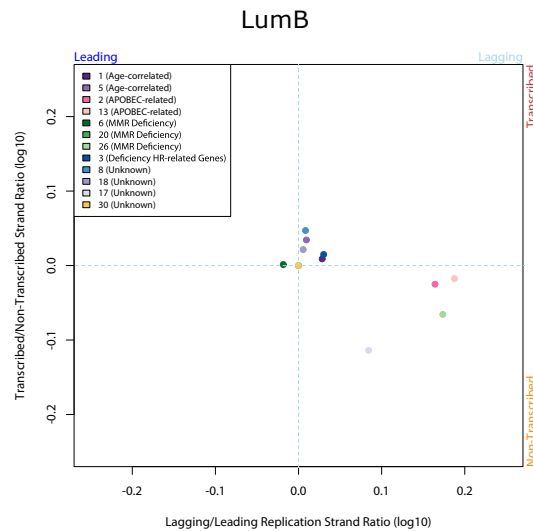


F

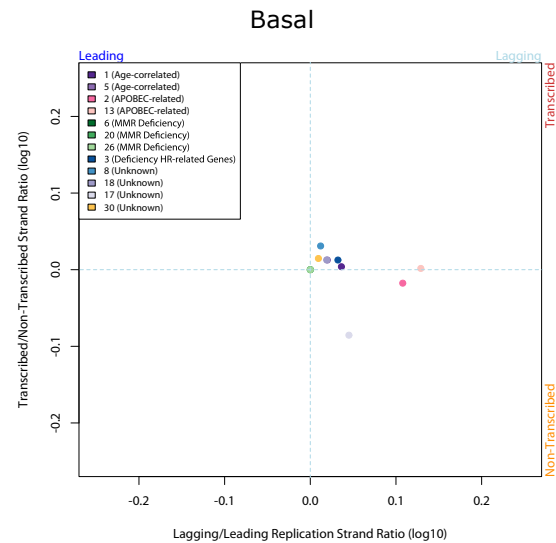




M

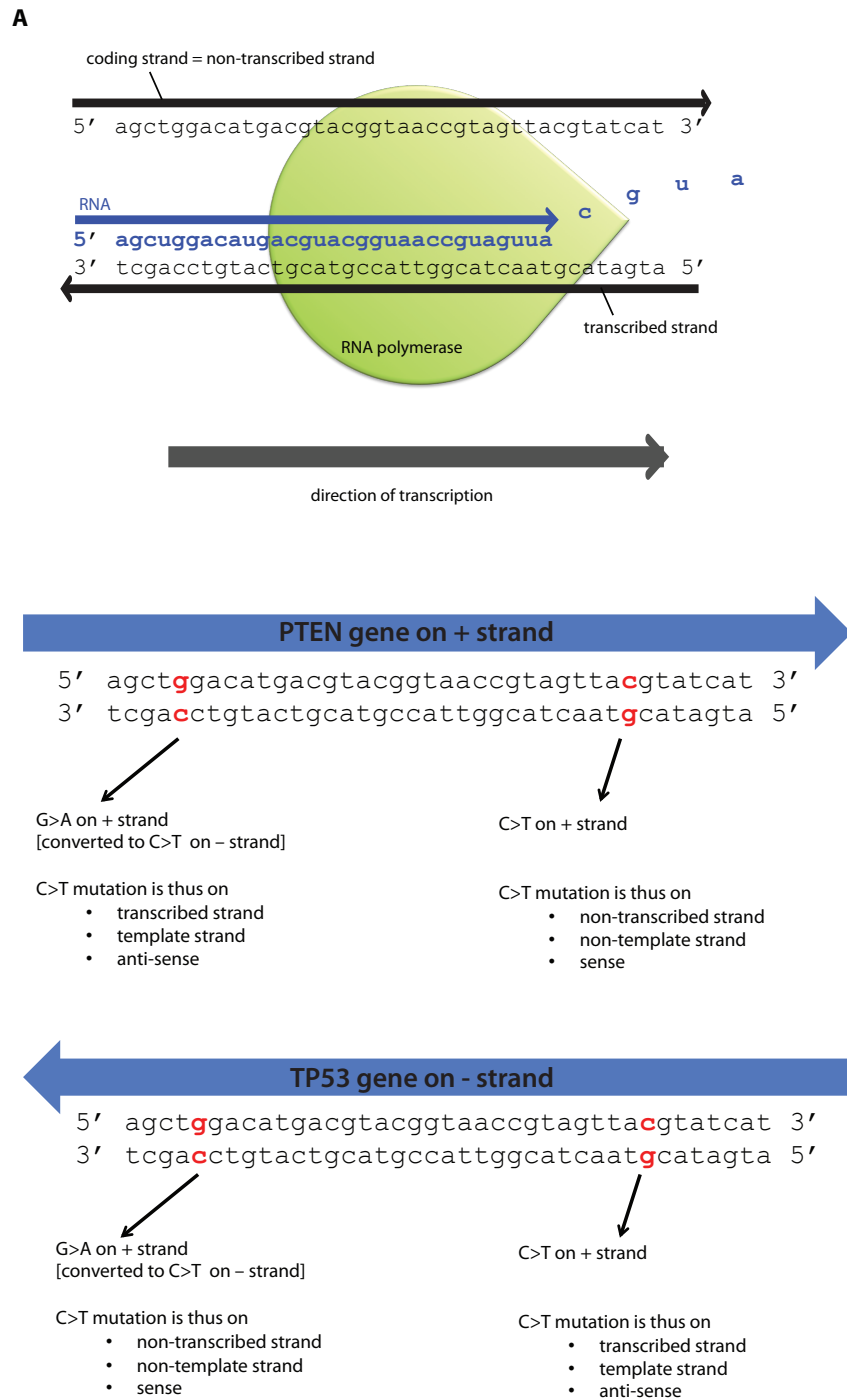


N

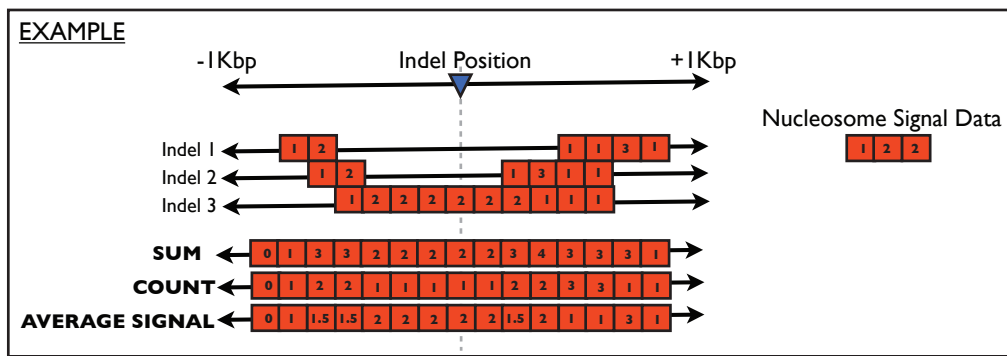
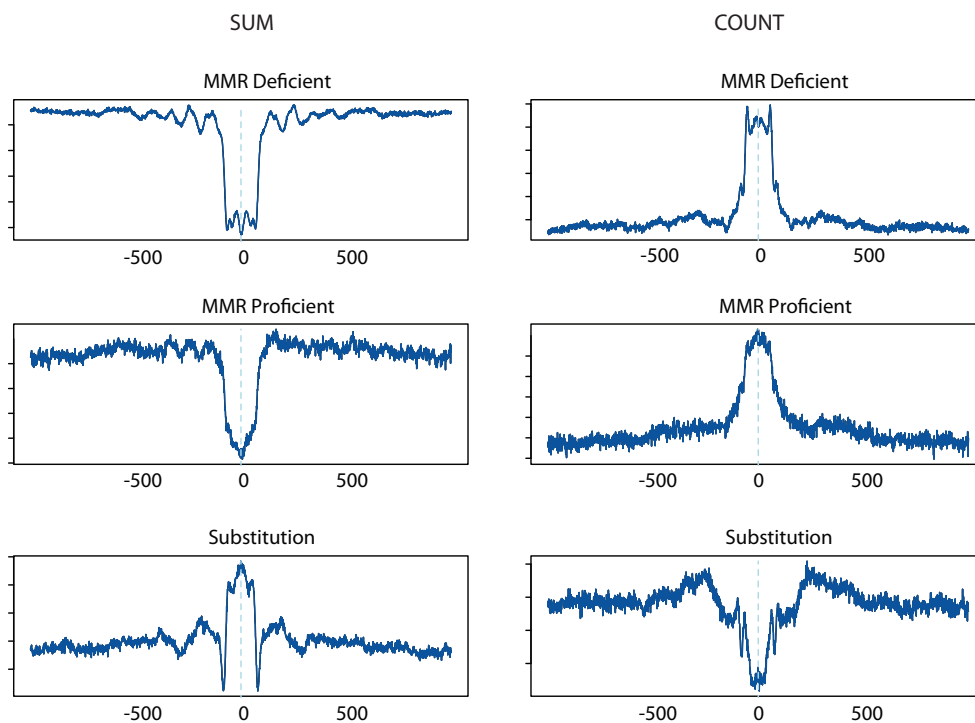


Supplementary Fig. 2 | Characterization of replication strands. A, Overview of the approach used for the identification of the replication strands based on Repli-seq data signal. The figure shows an example of a processed Repli-seq signal, where the local maxima is highlighted with a vertical dashed grey line. This is the point that is replicated before other local regions and acts as a replication initiation point. Peaks in these processed signal data are therefore replicated earlier than troughs. Because of the directionality of DNA synthesis, replicative strands can be inferred. From the local maxima, the direction of the replication fork moving leftwards means that the 3' to 5' DNA strand to the left of the local maxima acts as a template for 5' to 3' DNA synthesis of nascent strands, thus is the template for the leading strand. Conversely, for the replication fork moving rightwards (right of the local maxima), the 5' to 3' DNA strand is the template for the lagging strand. **B,** Smoothened profile of replication time signal obtained from MCF-7 RepliSeq data. The x-axis shows the genomic position of the region in Mb (in this example, chr10:16,437,499-16,437,499), and the y-axis reports the observed replication time signal. The dashed vertical lines highlight distance intervals of 25 Kb. Between the earliest and latest domains, regions with several mini-peaks/valleys are noted and are excluded from reference coordinates. Only long stretches of consistent early-to-late or late-to-early transition zones are used in defining replication strands (highlighted in red). **C-**

D, Size, base content, and distribution of transcribed and non-transcribed strands for the p to q-leading and p to q-lagging replication strands. **E**, Distribution of replication and transcription strand ratios (in log10 scale) for the six pyrimidine mutation classes (each dot represents a sample). **F**, Distribution of replication and transcription strand ratios (in log10 scale) for the twelve signatures (each dot represents a sample). **G**, Distribution of the replication and transcription strand ratios (in log10 scale) of the six pyrimidine mutation classes **H**, of the twelve base substitution signatures **I**, and of variations from dbSNP human build 142. **J-N**. Replication and transcription strand bias for the twelve base substitution signatures, where samples were grouped together by using ER status (ER positive 366 samples and ER negative 194 samples) **J**, and **K**, respectively, and gene expression profiles LumA (101 samples), LumB (75 samples) and Basal (74 samples) **L**, **M** and **N**, respectively [see ¹⁵ for further details on complete cohort].



Supplementary Fig. 3 | Characterization of transcriptional strands. A, The nucleotide sequence of transcribed RNA is identical to the sense/non-template/non-transcribed strand, except that U replaces T, and is complementary to that of the anti-sense/template/transcribed strand. **B** Example of mutation presented in pyrimidine context and assigned to transcriptional strands

A**B**

Supplementary Fig. 4 | Relationships between base substitution signatures and nucleosome occupancy. **A**, Overview of the approach used to compute the relationship between mutations (within individual signatures) and nucleosome occupancy. For all mutations in any given signature, the signal obtained from the MNase experiments (indicating the putative presence of a nucleosome) was summed and divided by the total number of mutations relevant to that signature, for a window of 2kb around each mutation (Supplementary Fig. 4A provides a diagram and example) providing an average signal (y-axis of Figure 3). Thus, the mutations that account for each signature are at the centermost point of the

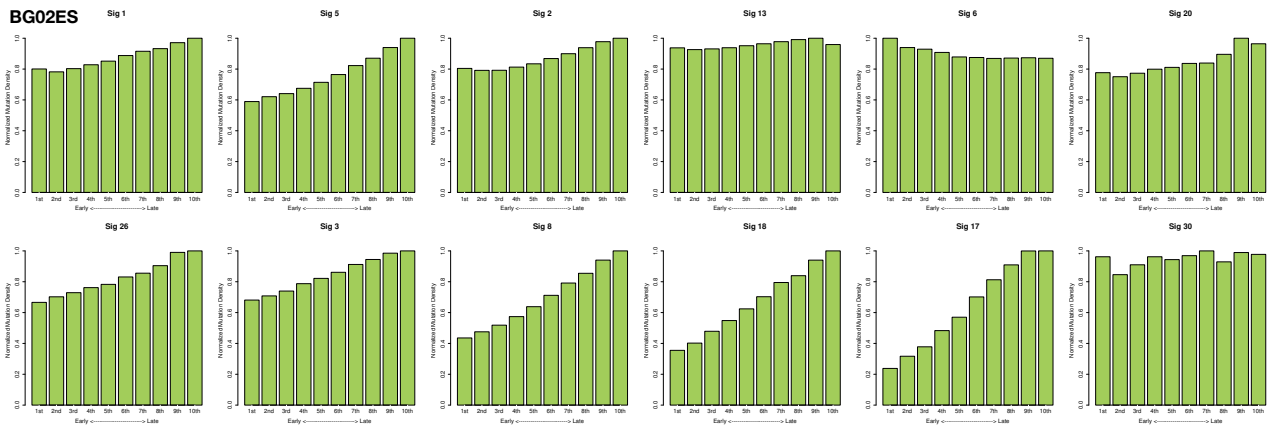
windows in Figure 3. **B**, Intermediary steps in obtaining the averaged signal. Left panels show the sum of the signal across all the somatic variants for a given signature, the right panels show the total number or count of somatic variants contributing to the total signal.

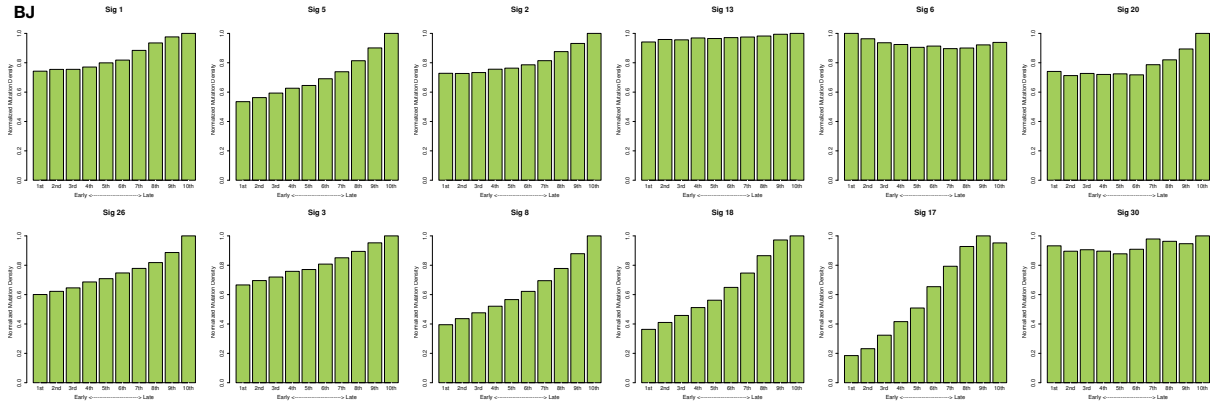
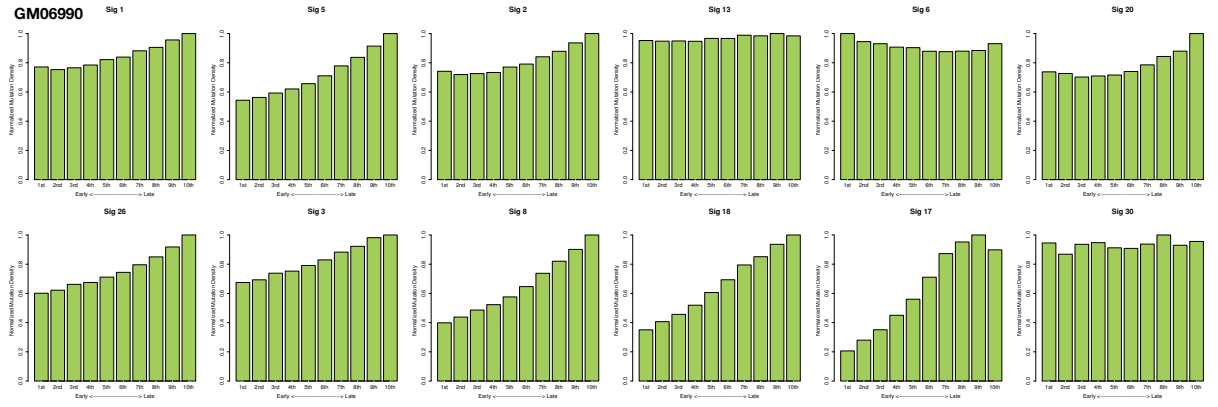
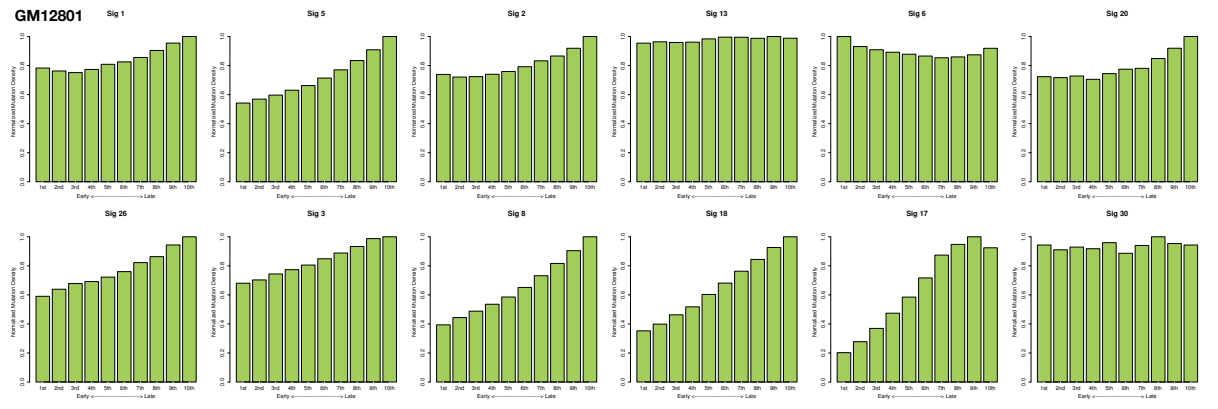
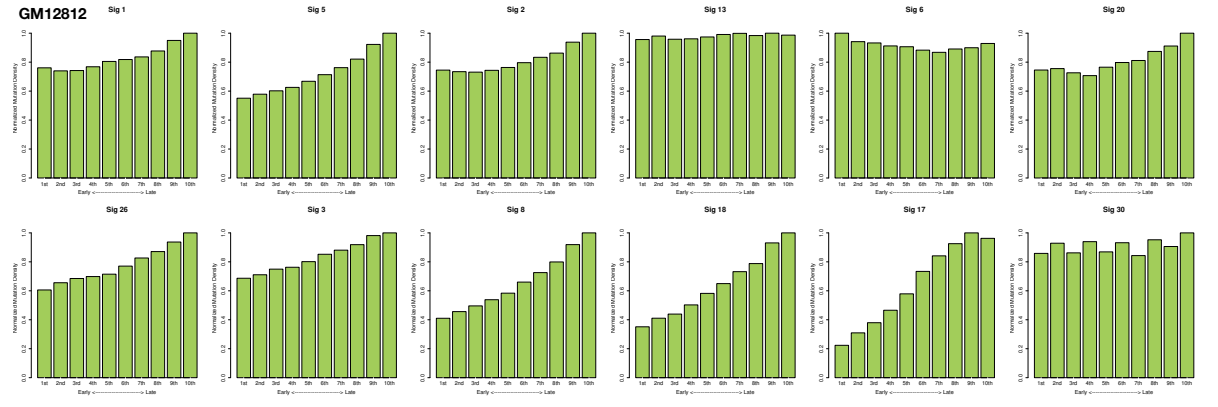
A

Early <---- Replication time deciles ----> Late

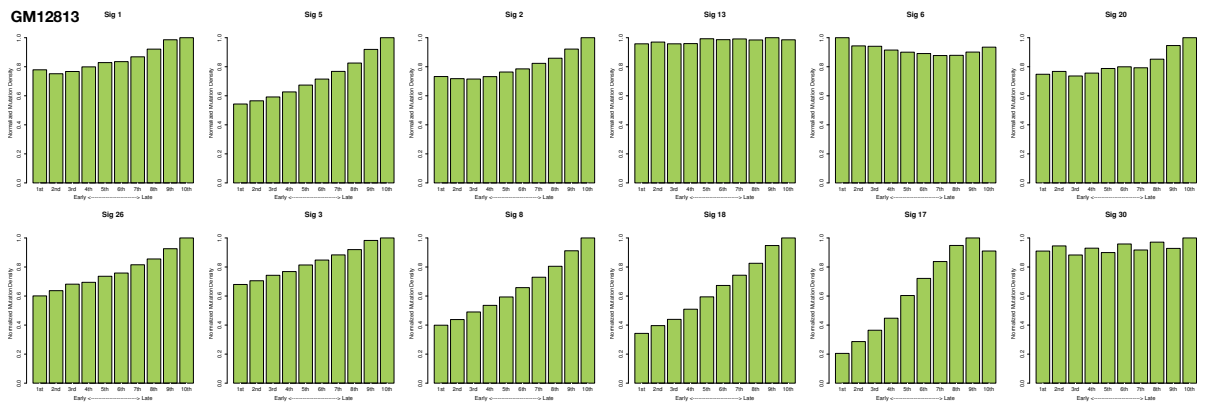
	1	2	3	4	5	6	7	8	9	10
BG02ES	51.7	25.3	17.5	15.9	14.7	15.3	15.6	18.2	21.4	70.1
BJ	56.7	27.5	19.8	17.4	17.2	16.8	17.9	20.3	28.2	52.6
GM06990	58.6	29.9	23.1	19.4	18.6	16.9	19.9	20.3	25.9	81.2
GM12801	60.9	31.1	23.3	19	17.9	16.3	19.2	18.4	24	81.9
GM12812	53.9	26.8	19.8	18.4	18.2	16.3	18.9	18.8	25.2	80.9
GM12813	58.1	30.3	22.9	20.5	19.5	17.3	19	20.5	25.8	83
GM12878	59.4	29.1	23.1	19.7	18.7	16.8	17.9	20.3	27.9	82.5
HeLa-S3	64.1	32.9	23.2	18.5	17.8	16.6	18.9	22.3	28	82.9
HepG2	64	34.7	26.2	22.2	21.1	20.8	19.9	21.4	26.4	77.6
HUVEC	60.1	30.9	19.3	17.9	16.2	16.3	17.8	20.2	27.2	85
IMR90	56.3	25	18.8	17.1	16	16.2	16.7	19.9	27	79.8
K562	64.6	34.2	24.4	21.2	20.2	18.3	19.1	19.4	24	80.1
SK-N-SH	69	34.8	23.6	18.8	17.4	16.4	15.7	17.5	22.8	74.8

BG02ES

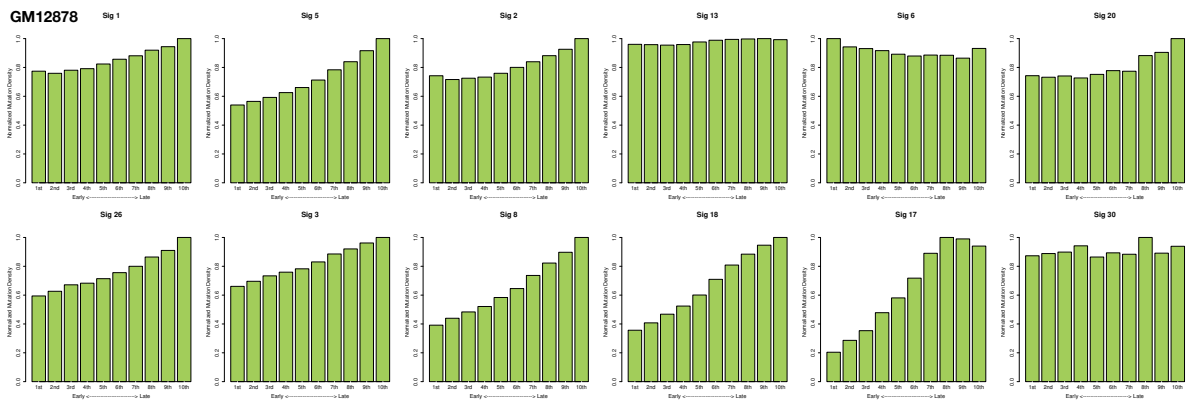


BJ**GM06990****GM12801****GM12812**

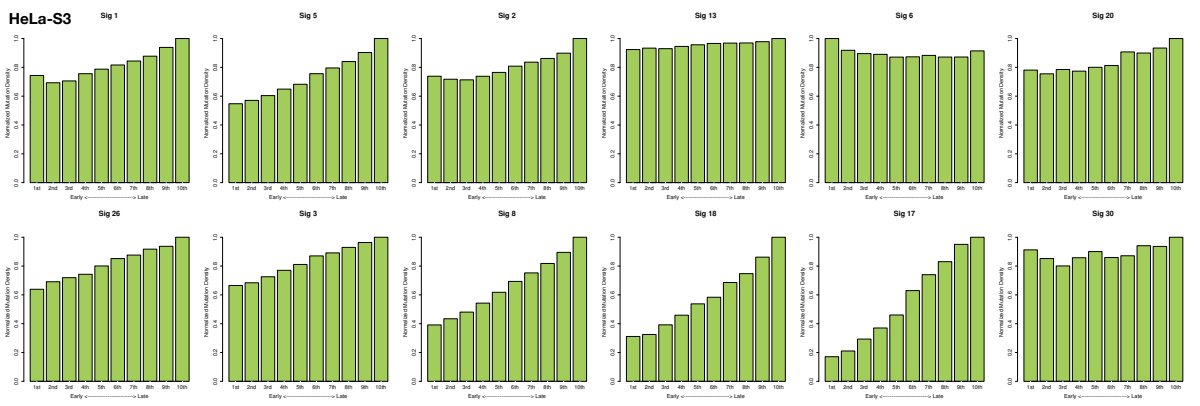
GM12813



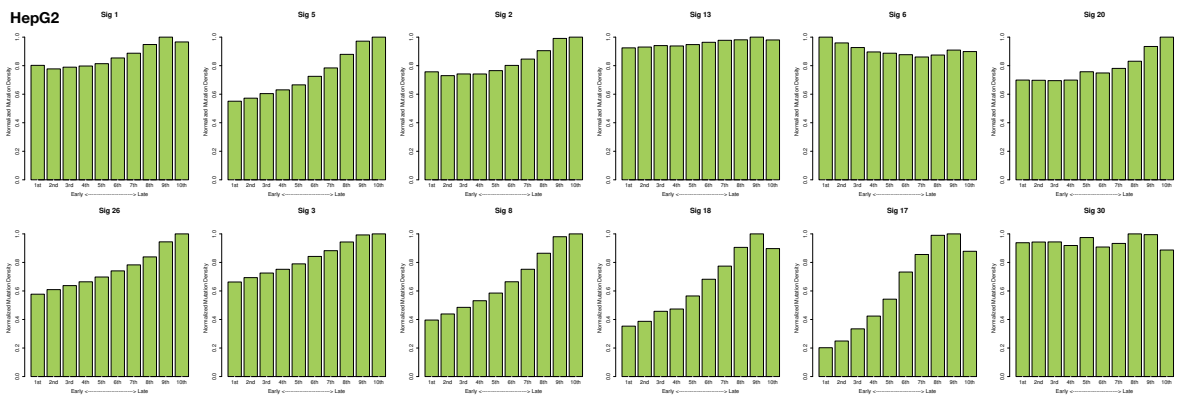
GM12878



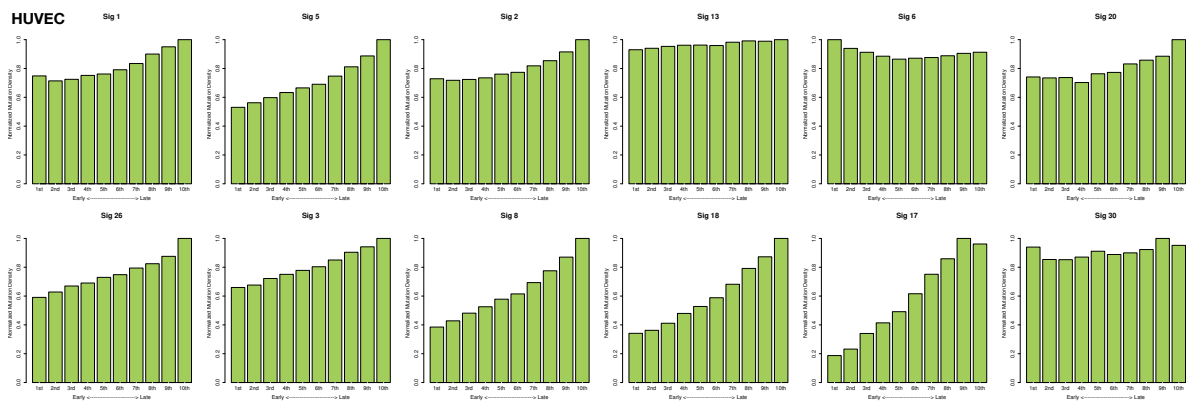
HeLa-S3



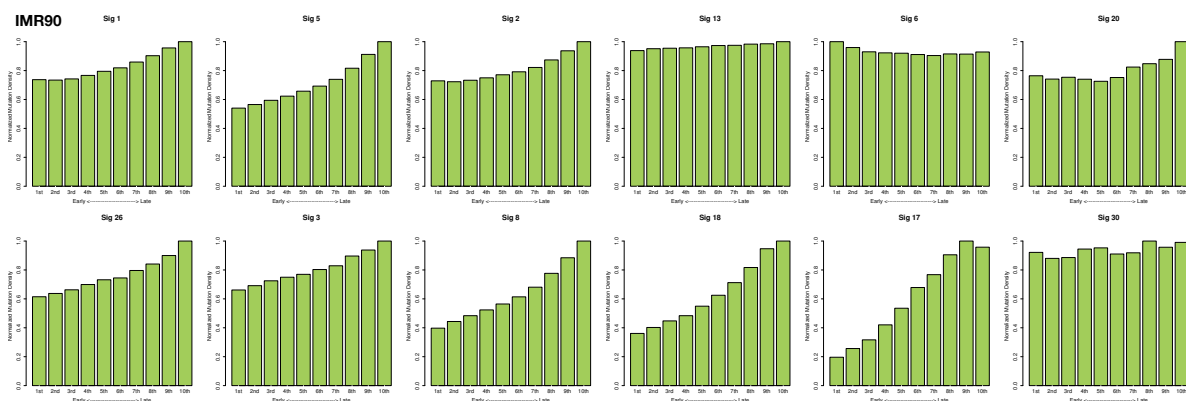
HepG2



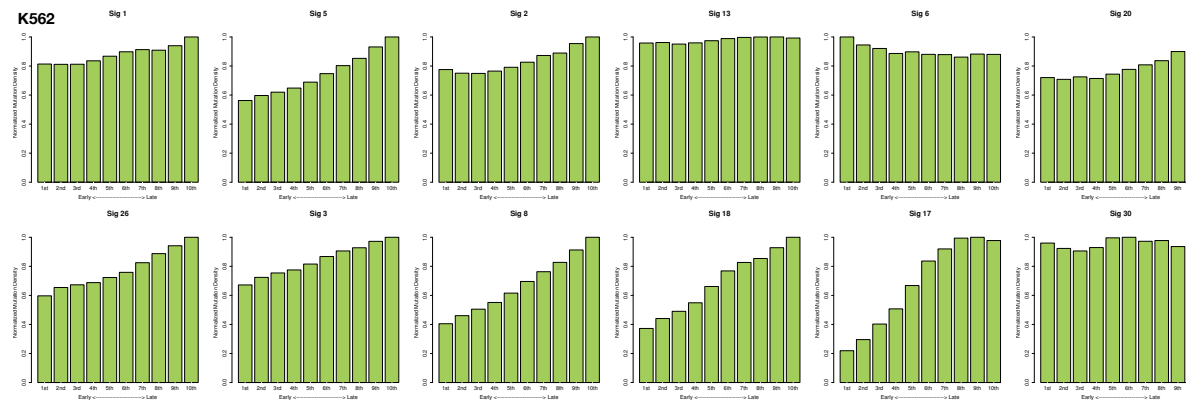
HUVEC



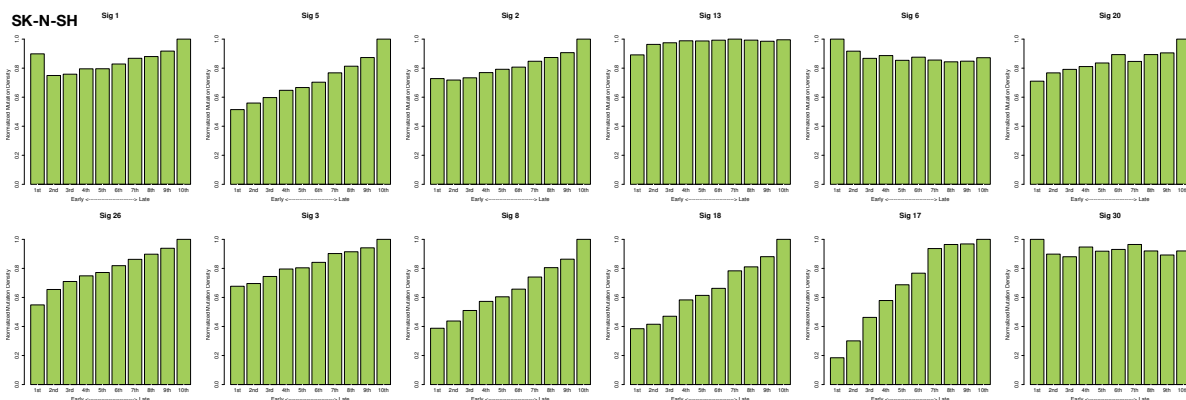
IMR90

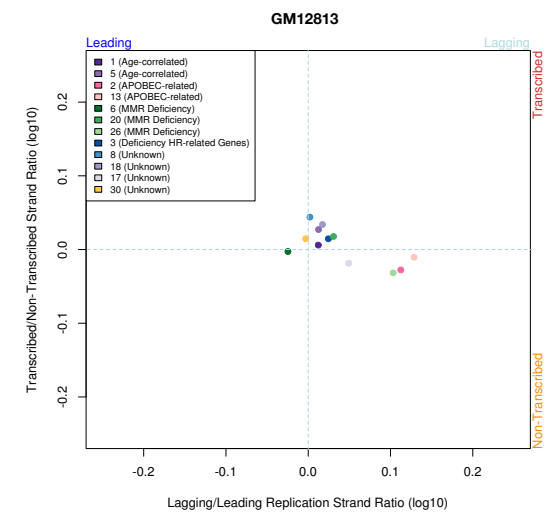
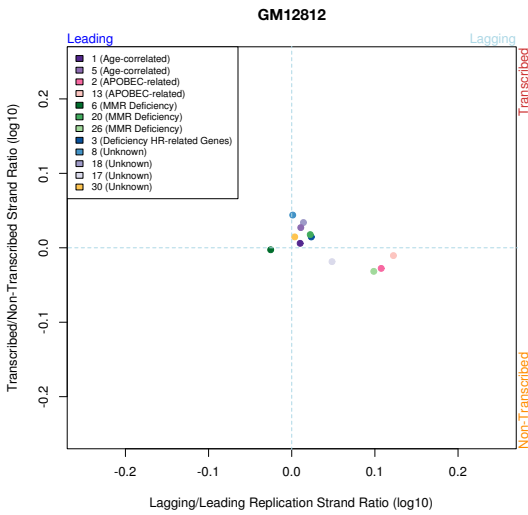
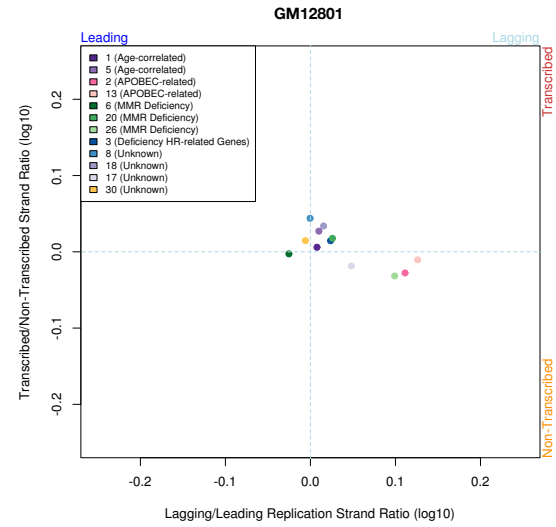
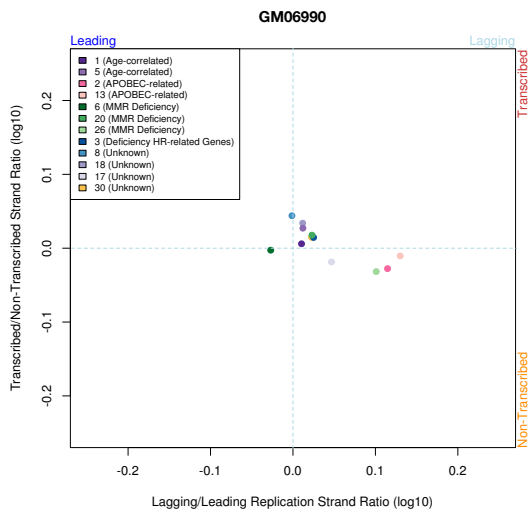
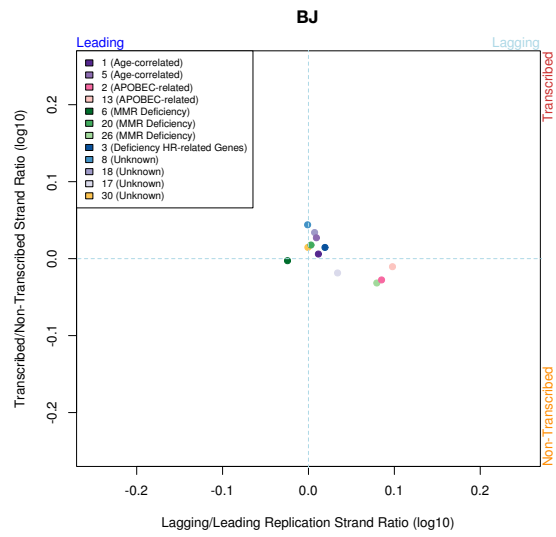
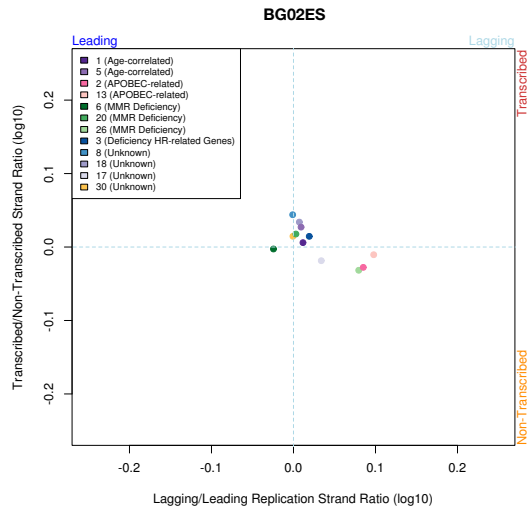


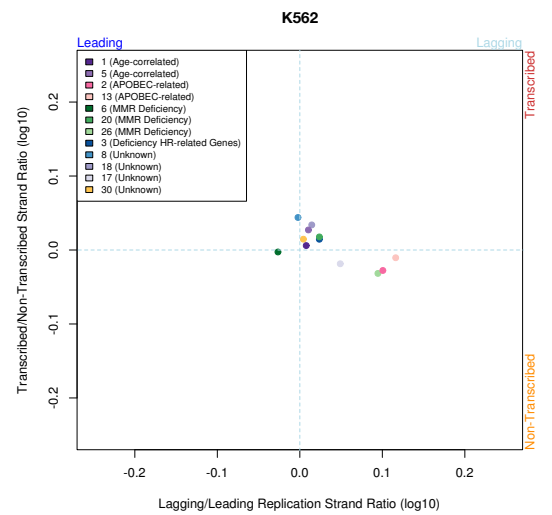
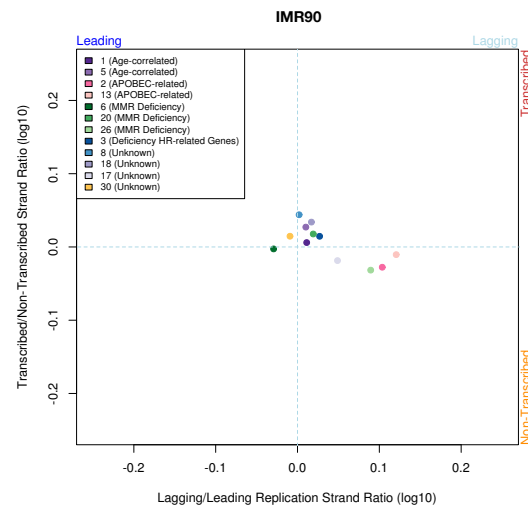
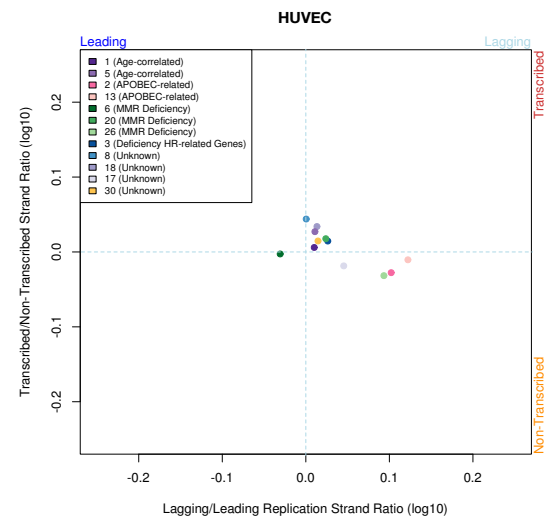
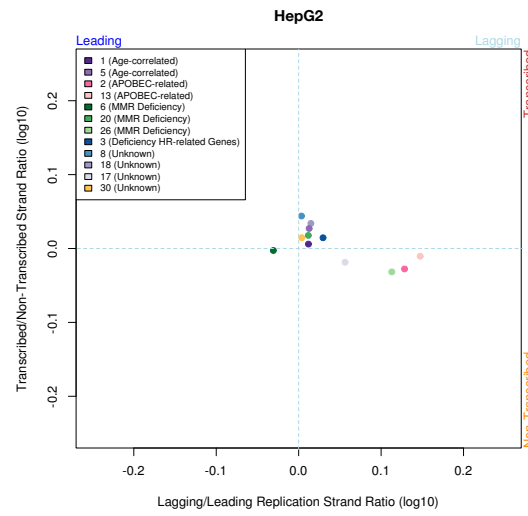
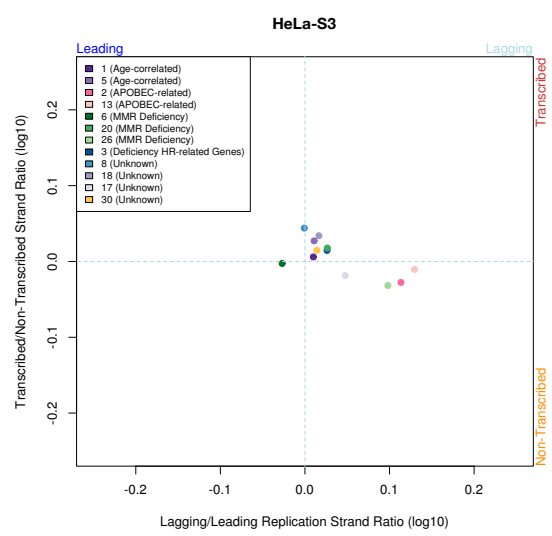
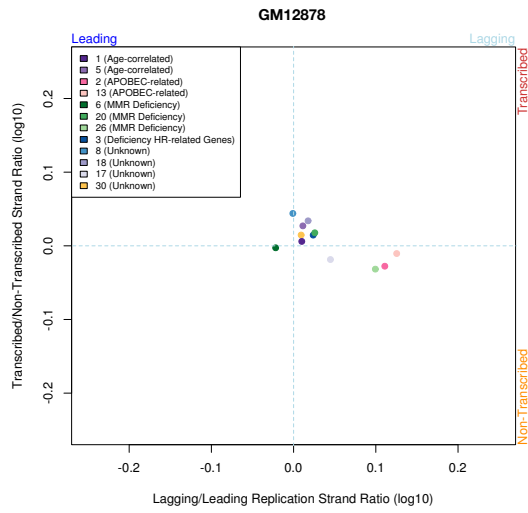
K562

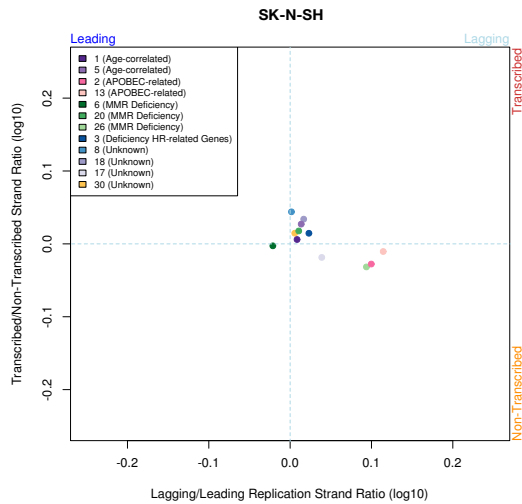


SK-N-SH

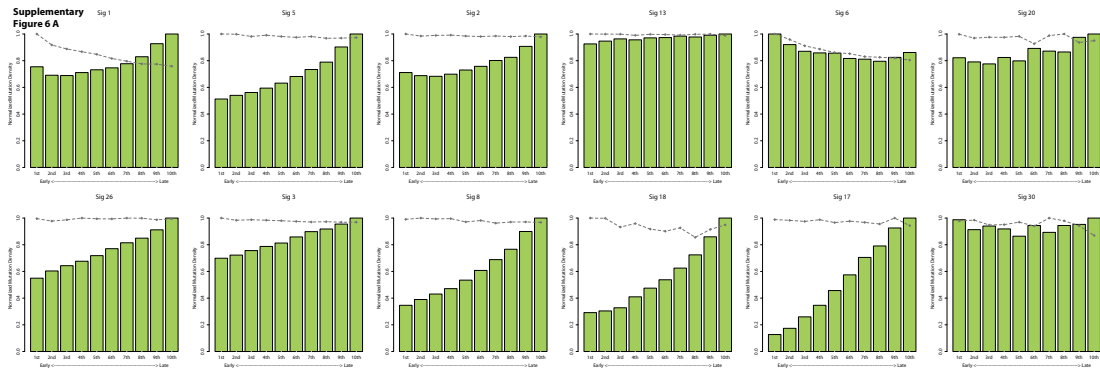






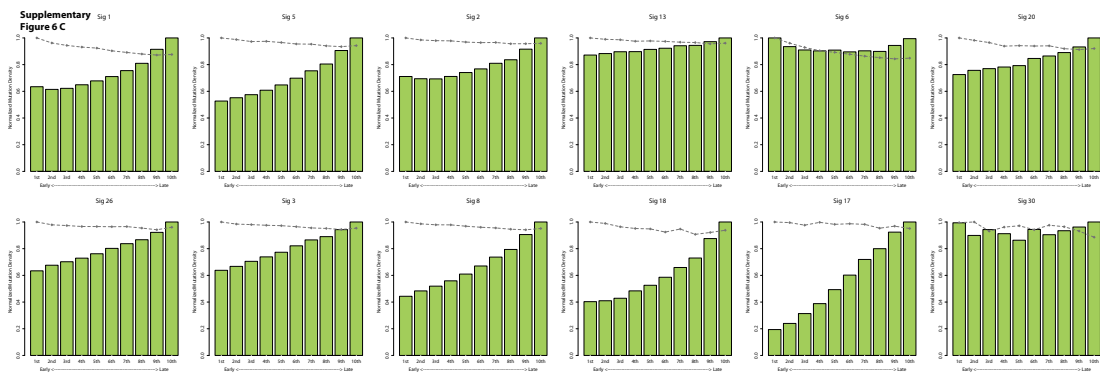
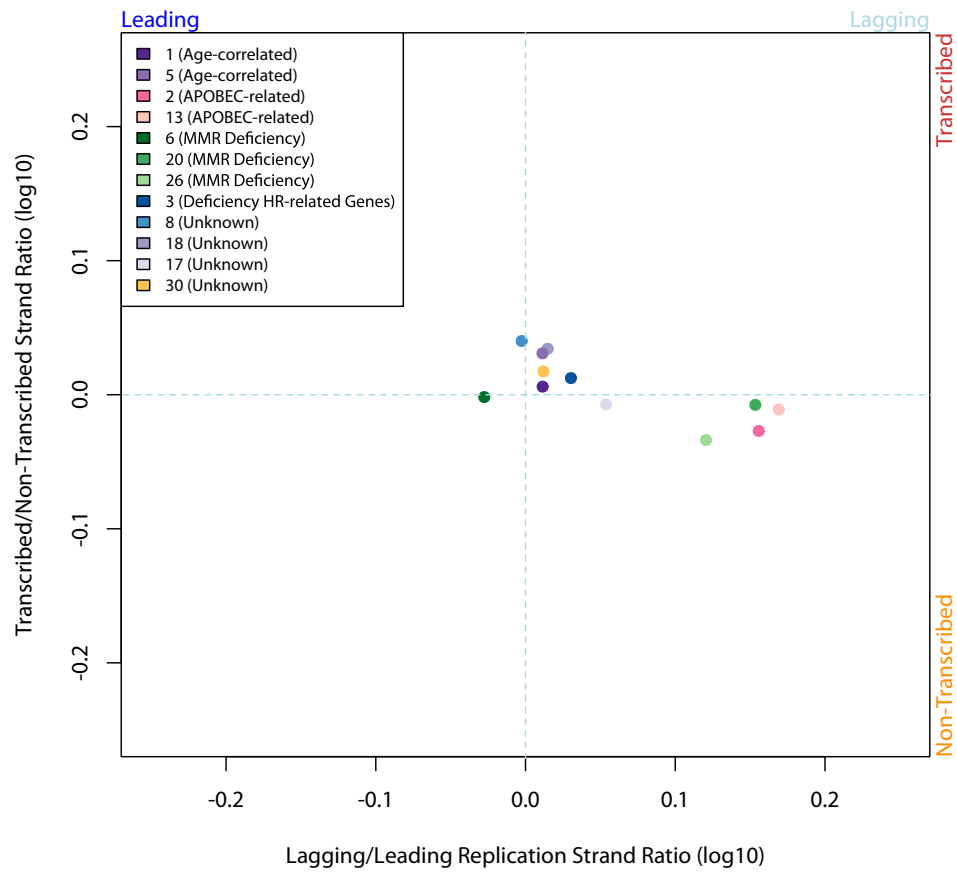


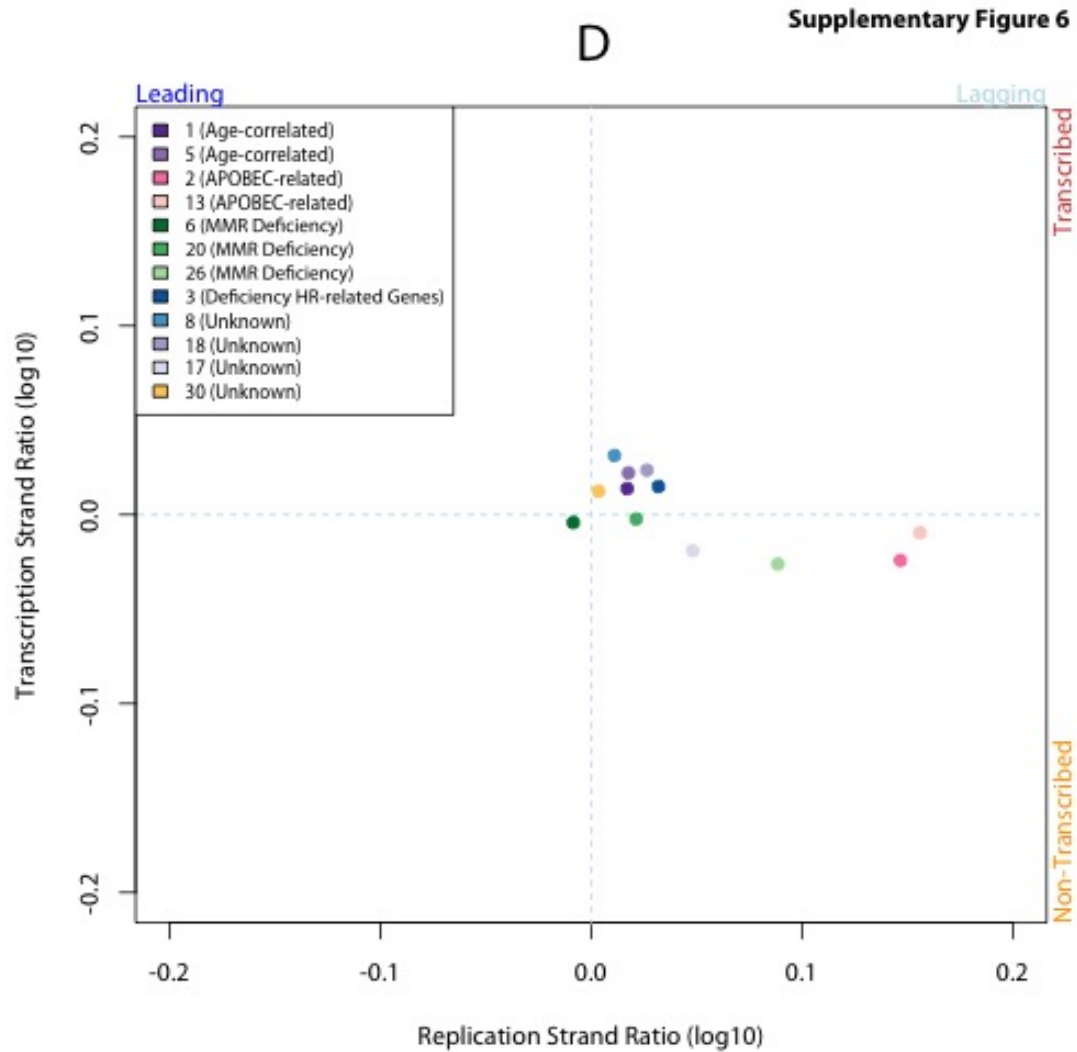
Supplementary Fig. 5 | Comparison between MCF7 and publicly available datasets on ENCODE. A, Details of the replication time domain shared between MCF-7 and other cell lines. The figure reports the percentage of genome that MCF-7 shares with the other thirteen cell lines (rows) in each computed decile (columns). Deciles are numbered from 1 (earliest decile) to 10 (latest decile). A large proportion of the earliest (average 59.8%) and latest (average 77.9%) replication domains are shared between all thirteen cell lines and MCF7. Subsequent pages demonstrate results of analysis of replication time and replication and transcriptional strand biases for the remaining thirteen cell lines. Note the striking similarity between cell lines, particularly for the signatures with the strongest biological signals.



B

Supplementary Figure 6





Supplementary Fig. 6 | Replication time, and replication vs transcription strand analyses performed by using ‘maximum likelihood with probability threshold’ and ‘belief propagation’ approaches. A-B, Replication time and replication vs transcription strand analyses performed by using maximum likelihood with probability threshold (signatures with a posteriori probability lower or equal to 0.5 were filtered out). **C-D,** Replication time and replication vs transcription strand analyses performed by using belief propagation. Comparison of these plots with the ones reported in the main manuscript (Figure 1 and Figure 2) shows that the main results were consistent across the different methodologies.

Processive Group Length	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)	Number of Groups	Mut Density (per Mb)
2	21,792	41	62,208	42	48,104	33.8	59,777	31.8	16,233	70.3	2,201	164.1	14,468	434.0	80,249	7.7	2,314	11.4	2,412	4.6	2,324	34.9	1,064	16.6
3	2,774	22	9,463	1.7	1,615	2.3	12,015	26.7	3,485	5,035	20.9	1,067	39.4	15,010	21.2	15,010	2,996	6.6	3.4	3.6	31.4	23.7	7.0	8.5
4	433	1.7	1.615	1.7	3,485	25.4	22.5	1,067	39.4	2.79	44.2	95.0	12.9	15,010	21.2	15,010	2,996	6.6	3.4	3.6	31.4	23.7	7.0	8.5
5	76	1.3	3.20	1.4	1,019	22.5	1,067	39.4	2.79	44.2	95.0	12.9	15,010	21.2	15,010	2,996	6.6	3.4	3.6	31.4	23.7	7.0	8.5	
6	16	1.9	65	1.3	398	24.5	1,831	19.8	32.8	3	17.3	96	9.6	141	5.2	81	9.7	0	0	55	22.3	16	16.6	
7	5	1.7	17	7.0	134	22.7	346	18.6	27	30.8	0	0	0	0	0	0	9.3	0	0	22	22.7	2	5.9	
8	1	0.6	2	18.2	65	28.3	159	18.3	14	31.9	0	0	0	0	0	0	0	0	0	13	17.8	0	0.0	
9	0	0.0	2	0.5	27	33.2	86	16.4	5	23.4	0	0	0	0	0	0	0	0	0	1	43.0	0	0.0	
10	0	0.0	1	0.4	11	79.4	48	17.6	3	46.2	0	0	0	0	0	0	0	0	0	1	43.0	0	0.0	
11	0	0.0	0	0.0	4	31.4	18	16.2	1	27.1	0	0	0	0	0	0	0	0	0	0	0	0	0.0	
12	0	0.0	0	0.0	4	31.4	18	16.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	
13	0	0.0	0	0.0	1	397.0	10	17.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	
14	0	0.0	0	0.0	0	0.0	5	18.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	
15	0	0.0	0	0.0	0	0.0	5	12.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	
16	0	0.0	0	0.0	0	0.0	5	12.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	
17	0	0.0	0	0.0	0	0.0	2	17.7	0	0	0	0	0	0	0	0	0	0	0	0	1	51.5	0	
18	0	0.0	0	0.0	0	0.0	0	0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	
19	0	0.0	0	0.0	0	0.0	2	17.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	

Supplementary Table 2 | Details of strand-coordinate mutagenesis of mutation signatures. For each mutation signature (columns) and processive group length (rows) the table reports (i) the number of groups (ii) the median of the mutation density per Mb across the groups. The mutation density for each group was computed as the number of mutations contained in the group, that is, the processive group length, divided by the size of the group in Mb.

Processive Group Length	Signature 2			Signature 13		
	Lagging	Leading	log10(ratio)	Lagging	Leading	log10(ratio)
1	147,346	115,683	0.105	181,853	140,591	0.112
2	50,935	32,053	0.201	63,734	39,180	0.211
3	19,259	9,919	0.288	25,980	12,564	0.315
4	7,310	3,450	0.326	11,385	4,899	0.366
5	2,856	1,254	0.357	5,083	1,957	0.414
6	1,107	531	0.319	2,566	878	0.465
7	476	210	0.354	1,313	535	0.389
8	244	140	0.240	754	158	0.677
9	134	64	0.317	460	125	0.563
10	39	41	-0.021	223	87	0.406
11	22	0	1.362	205	59	0.536
12	28	8	0.508	87	9	0.944
13	13	0	1.146	55	10	0.707
14	0	0	0.000	50	6	0.862
15	0	0	0.000	47	13	0.535
16	16	0	1.230	60	20	0.463
17	0	0	0.000	16	1	0.929
>3	12,245	5,698	0.332	22,304	8,757	0.406
>4	4,935	2,248	0.341	10,919	3,858	0.452
>5	2,079	994	0.320	5,836	1,901	0.487
>6	972	463	0.322	3,270	1,023	0.505
>7	496	253	0.292	1,957	488	0.603

Processive Group Length	Signature 2			Signature 13		
	Transcribed	Non-Transcribed	log10(ratio)	Lagging	Leading	log10(ratio)
1	66,455	70,795	-0.027	84,148	85,904	-0.009
2	19,654	21,090	-0.031	26,432	26,850	-0.007
3	6,191	6,760	-0.038	8,826	9,366	-0.026
4	2,431	2,517	-0.015	3,834	3,810	0.003
5	803	757	0.026	1,533	1,642	-0.030
6	421	401	0.021	733	869	-0.074
7	119	154	-0.111	345	383	-0.045
8	135	97	0.142	144	184	-0.106
9	44	19	0.352	91	134	-0.167
10	19	31	-0.204	41	59	-0.155
11	0	0	0.000	65	45	0.157
12	30	18	0.213	12	36	-0.454
13	0	0	0.000	4	9	-0.301
14	0	0	0.000	0	0	0.000
15	0	0	0.000	7	8	-0.051
16	0	0	0.000	0	0	0.000
17	0	0	0.000	0	0	0.000
>3	4,002	3,994	0.001	6,809	7,179	-0.023
>4	1,571	1,477	0.027	2,975	3,369	-0.054
>5	768	720	0.028	1,442	1,727	-0.078
>6	347	319	0.037	709	858	-0.083
>7	228	165	0.140	364	475	-0.116

Supplementary Table 3 | Replication and transcriptional strand bias associated with APOBEC-related signatures processivity. The tables report the replication strand (**A**) and the transcription strand (**B**) counts and ratios for the different length of processive groups associated with the APOBEC-related signature 2 and 13. The last five rows report the aggregated counts and ratios (e.g. >3 reports the number of mutations and the respective ratios across all groups composed of more than 3 mutations). From the table, it is clear that the replication strand bias of Signature 13 increases with the length of the processive group. Unlike replication strand, the transcriptional strand bias observed for these signatures does not seem related to the length of the processive group.

Signature	MCF7	BG025	BI	GM06990	GM12801	GM12812	GM12813	GM12878	HeLa-S3	HepG2	HUVEC	IMR90	K562	SK-N-SH
1	0.0145	0.0116	0.0101	0.0104	0.0077	0.0101	0.0122	0.0099	0.0099	0.0119	0.0101	0.0112	0.0077	0.0085
5	0.0137	0.0092	0.0109	0.0120	0.0099	0.0109	0.0125	0.0113	0.0108	0.0127	0.0109	0.0102	0.0104	0.0136
2	0.1473	0.0851	0.0994	0.1147	0.1113	0.1076	0.1125	0.1109	0.1134	0.1284	0.1022	0.1036	0.1006	0.0997
13	0.1634	0.0978	0.1181	0.1301	0.1261	0.1223	0.1286	0.1254	0.1295	0.1474	0.1222	0.1206	0.1162	0.1144
6	-0.0332	-0.0244	-0.0301	-0.0270	-0.0253	-0.0253	-0.0247	-0.0218	-0.0270	-0.0307	-0.0307	-0.0292	-0.0265	-0.0213
20	0.0206	0.0029	0.0124	0.0230	0.0259	0.0223	0.0307	0.0258	0.0265	0.0117	0.0239	0.0192	0.0237	0.0106
26	0.1129	0.0796	0.0876	0.1010	0.0991	0.0987	0.1032	0.0994	0.0980	0.1129	0.0934	0.0894	0.0946	0.0938
3	0.0284	0.0192	0.0257	0.0250	0.0237	0.0235	0.0245	0.0238	0.0261	0.0295	0.0263	0.0270	0.0236	0.0230
8	0.0011	-0.0010	0.0031	-0.0013	-0.0004	0.0012	0.0020	-0.0009	-0.0008	0.0034	0.0004	0.0020	-0.0023	0.0016
18	0.0202	0.0071	0.0106	0.0118	0.0155	0.0141	0.0173	0.0176	0.0165	0.0148	0.0133	0.0170	0.0144	0.0167
17	0.0527	0.0339	0.0439	0.0467	0.0481	0.0484	0.0491	0.0448	0.0475	0.0563	0.0454	0.0489	0.0490	0.0389
30	0.0121	-0.0007	-0.0070	0.0225	-0.0059	0.0036	-0.0031	0.0092	0.0141	0.0038	0.0145	-0.0092	0.0043	0.0057

Supplementary Table 4 | Results of replication strand bias analysis across 14 ENCODE cell lines Replication strand ratio (log10) computed for each signature (row) for 14 different cell lines (column). Negative and positive values are associated with an asymmetry towards the p-to-q leading and p-to-q lagging strand respectively. Results show that the replication strand biases demonstrated by specific signatures are consistent across all the cell lines. Note that MCF-7 exhibits the strongest asymmetry for APOBEC-related signatures 2 (ratio = 0.1473) and 13 (ratio = 0.1634).