CBS_ESAD+features_extended.tsv

	OR	CI_95	p_value	adj.OR	adj.CI_95	adj.p_value
Binary_Genic1	0.71617	0.6452, 0.79483	0	0.75201	0.66922, 0.84493	0.0002
Rep_Time	0.64379	0.60892, 0.68031	0	0.53928	0.50272, 0.57801	0
XPD	0.77617	0.73594, 0.81836	0	0.88363	0.82624, 0.94484	0.00218
CTCF	1.76313	1.66711, 1.86589	0	2.43157	2.20184, 2.6895	0
DHS	1.53712	1.45527, 1.62438	0	1.25726	1.14328, 1.38284	0.00057
POLR2A	0.90977	0.86355, 0.95829	0.00435	0.67035	0.61659, 0.72809	0

Started with a total of 31252 rows, with 2965 of them having mutations.

But 90/31252 has Rep_Time = ".", so after removing them, we have a total of 31162 rows left, and that with only 2960 mutations.

Then we log1p (i.e., log(x + 1)) the DHS column because it is skewed.
* log is actually better but it will result in negative infinity values.

Then we do z-score normalization.

So recall we have 2960 mutation and 28202 non-mutation,
we will take 10% of non-mutation —> 2820 non-mutation to do the 100 times.

CBS_MELA+features_extended.tsv

	OR	CI_95	p_value	adj.OR	adj.CI_95	adj.p_value
Binary_Genic1	0.64311	0.59688, 0.69286	0	0.67268	0.62318, 0.72606	0
Rep_Time	0.8295	0.79851, 0.86152	0	0.76229	0.7305, 0.79529	0
XPD	0.99138	0.9553, 1.02879	0.55477	1.03293	0.99163, 1.07595	0.18974
CTCF	1.30378	1.25556, 1.35405	0	1.40083	1.33504, 1.47017	0
DHS	1.18783	1.14429, 1.23315	0	1.04035	0.99442, 1.08843	0.16317
POLR2A	0.98961	0.95361, 1.02696	0.50473	0.92354	0.88769, 0.96078	0.00144

Started with a total of 31252 rows, with 5913 of them having mutations.

But 90/31252 has Rep_Time = ".", so after removing them, we have a total of 31162 rows left, and that with only 5904 mutations.

Then 125/31162 has XPD = ".", so after removing them, we only have a total of 31037 rows left, and that with only 5882 mutations.

Then we log1p (i.e., log(x + 1)) the DHS column because it is skewed.
* log is actually better but it will result in negative infinity values.

Then we do z-score normalization.

So recall we have 5882 mutation and 25155 non-mutation,
we will take 23% of non-mutation —> 5785.65 → 5786 non-mutation to do the 100 times.

CBS_XPD+features_extended.tsv

	OR	CI_95	p_value	adj.OR	adj.CI_95	adj.p_value
Binary_Genic1	0.71816	0.6125, 0.8418	0.00076	0.79307	0.66654, 0.94359	0.03049
Rep_Time	1.06794	0.98675, 1.15587	0.1899	0.8625	0.7827, 0.94982	0.01492
XPD	1.14853	1.06096, 1.24371	0.00484	0.93586	0.83958, 1.04285	0.30351
CTCF	2.4606	2.21184, 2.74861	0	2.22163	1.96242, 2.52617	0
DHS	1.69392	1.55656, 1.8457	0	1.37049	1.20186, 1.56409	0.00014
POLR2A	0.86444	0.79733, 0.9363	0.00488	0.77175	0.70348, 0.84544	0.00001

Started with a total of 31252 rows, with 1288 of them having mutations.

But 90/31252 has Rep_Time = ".", so after removing them, we have a total of 31162 rows left, and that with only 1283 mutations.

Then 125/31162 has XPD = ".", so after removing them, we only have a total of 31037 rows left, and that with only 1281 mutations.

Then we log1p (i.e., log(x + 1)) the DHS column because it is skewed.
* log is actually better but it will result in negative infinity values.

Then we do z-score normalization.

So recall we have 1281 mutation and 29756 non-mutation,
we will take 4% of non-mutation —> 1190.24 → 1191 non-mutation to do the 100 times.