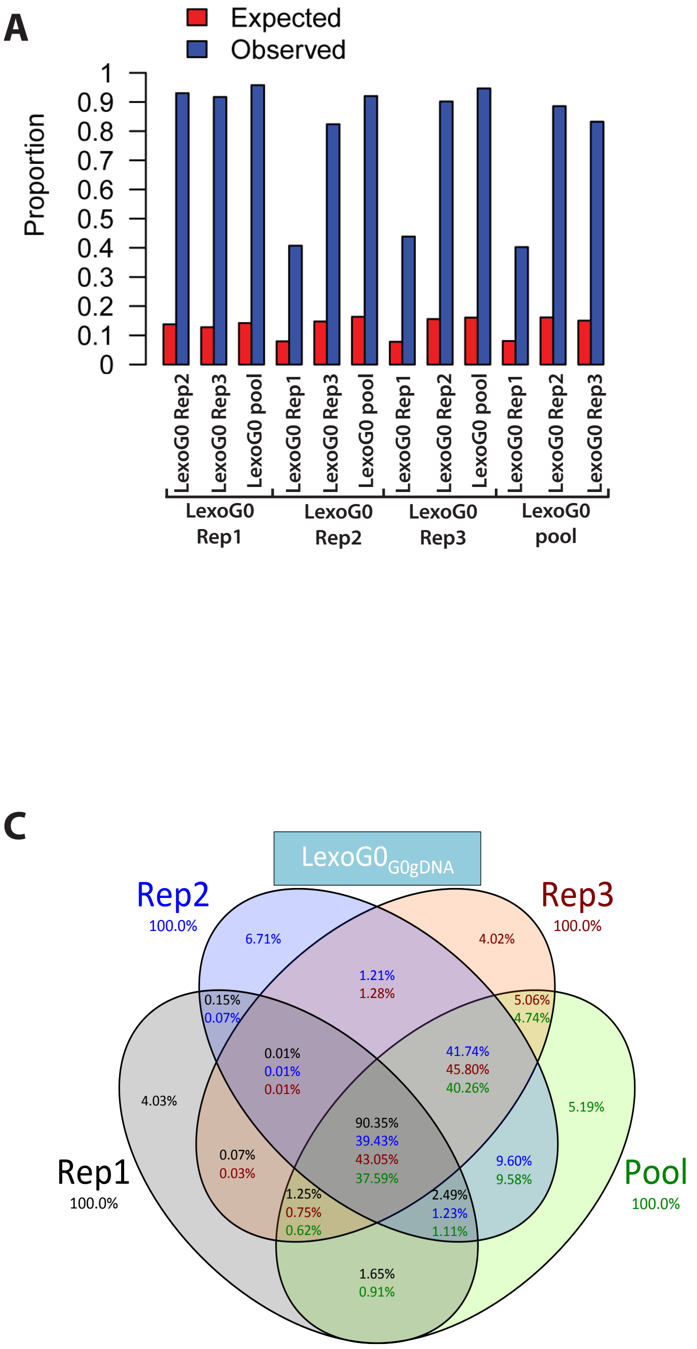
Figure S1.



A
Expected
Observed
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
LexoG0 Rep2
LexoG0 Rep3
LexoG0 pool
LexoG0 Rep1
LexoG0 Rep3
LexoG0 pool
LexoG0 Rep1
LexoG0 Rep2
LexoG0 pool
LexoG0 Rep1
LexoG0 Rep2
LexoG0 Rep3
LexoG0 Rep1
LexoG0 Rep2
LexoG0 Rep3
LexoG0 pool
B
Expected
Observed
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
NS-seq Rep2
NS-seq Rep3
NS-seq pool
NS-seq Rep1
NS-seq Rep3
NS-seq pool
NS-seq Rep1
NS-seq Rep2
NS-seq pool
NS-seq Rep1
NS-seq Rep2
NS-seq Rep3
NS-seq Rep1
NS-seq Rep2
NS-seq Rep3
NS-seq pool
C
LexoG0G0gDNA
Rep2
100.0%
Rep3
100.0%
Rep1
100.0%
Pool
100.0%
6.71%
4.02%
1.21%
1.28%
0.15%
0.07%
5.06%
4.74%
4.03%
0.01%
0.01%
0.01%
41.74%
45.80%
40.26%
5.19%
90.35%
39.43%
43.05%
37.59%
0.07%
0.03%
9.60%
9.58%
1.25%
0.75%
0.62%
2.49%
1.23%
1.11%
1.65%
0.91%
D
NSG0gDNA
Rep2
100.0%
Rep3
100.0%
Rep1
100.0%
Pool
100.0%
2.69%
9.74%
0.01%
0.01%
0.18%
0.17%
32.80%
18.31%
20.66%
0%
0%
0%
51.70%
30.93%
16.82%
16.02%
16.75%
17.39%
18.34%
9.72%
0.01%
0.01%
7.97%
5.14%
8.24%
8.17%
4.26%
24.36%
20.07%
11.61%
29.8%
18.12%

Figure S2.

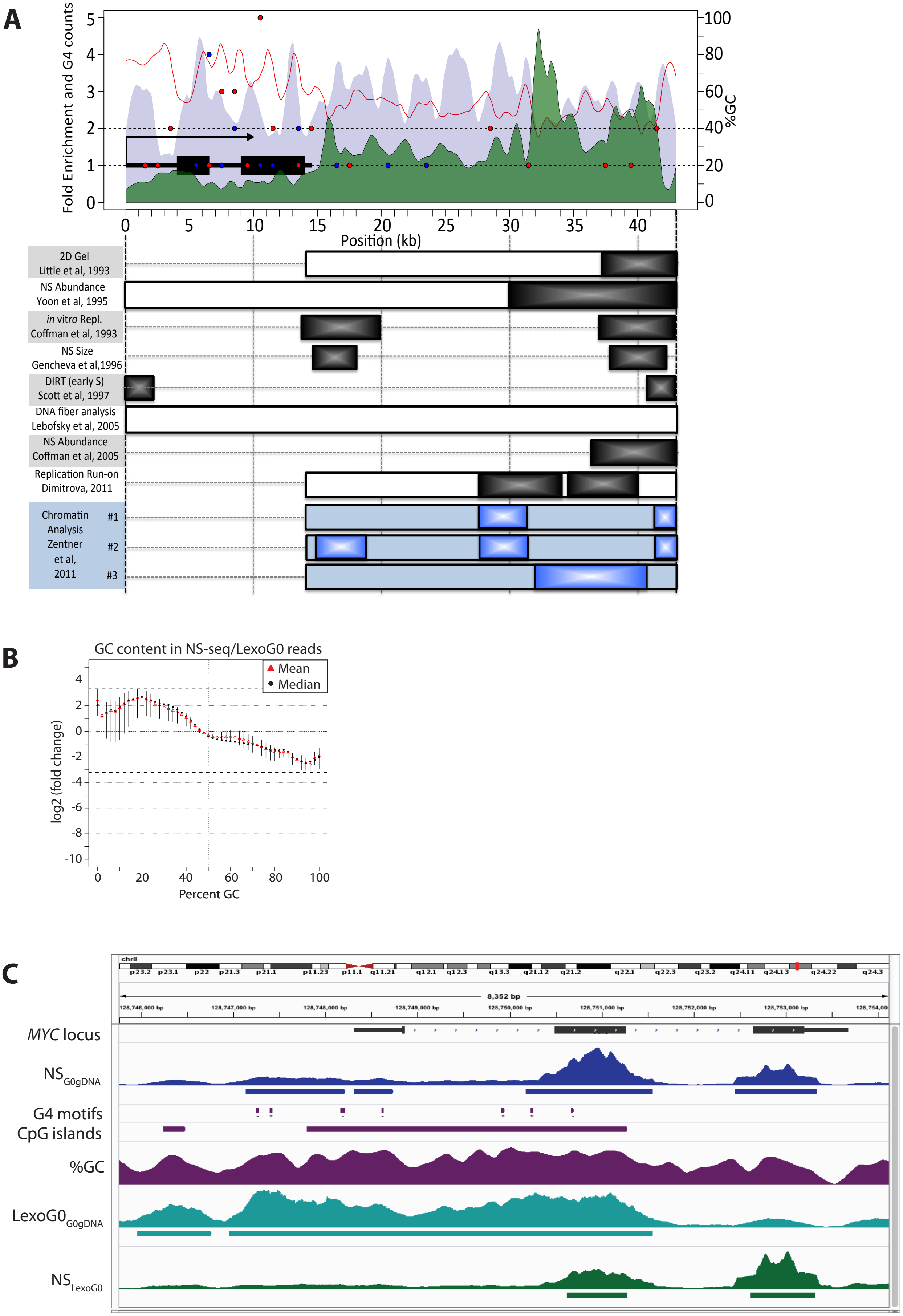

A
Fold Enrichment and G4 counts
0
1
2
3
4
5
100
80
60
40
20
0
%GC
0
5
10
15
20
25
30
35
40
Position (kb)
2D Gel
Little et al, 1993
NS Abundance
Yoon et al, 1995
in vitro Repl.
Coffman et al, 1993
NS Size
Gencheva et al,1996
DIRT (early S)
Scott et al, 1997
DNA fiber analysis
Lebofsky et al, 2005
NS Abundance
Coffman et al, 2005
Replication Run-on
Dimitrova, 2011
Chromatin
Analysis
Zentner
et al,
2011
#1
#2
#3
B
GC content in NS-seq/LexoG0 reads
Mean
Median
log2 (fold change)
4
2
0
-2
-4
-6
-8
-10
0
20
40
60
80
100
Percent GC
C
chr8
p23.2
p23.1
p22
p21.3
p21.1
p11.23
p11.1
q11.21
q12.1
q12.3
q13.3
q21.12
q21.2
q22.1
q22.3
q23.2
q24.11
q24.13
q24.22
q24.3
8,352 bp
128,746,000 bp
128,747,000 bp
128,748,000 bp
128,749,000 bp
128,750,000 bp
128,751,000 bp
128,752,000 bp
128,753,000 bp
128,754,00
MYC locus
NS_{G0gDNA}
G4 motifs
CpG islands
%GC
$LexoG0_{G0gDNA}$
NS_{LexoG0}

Figure S3.

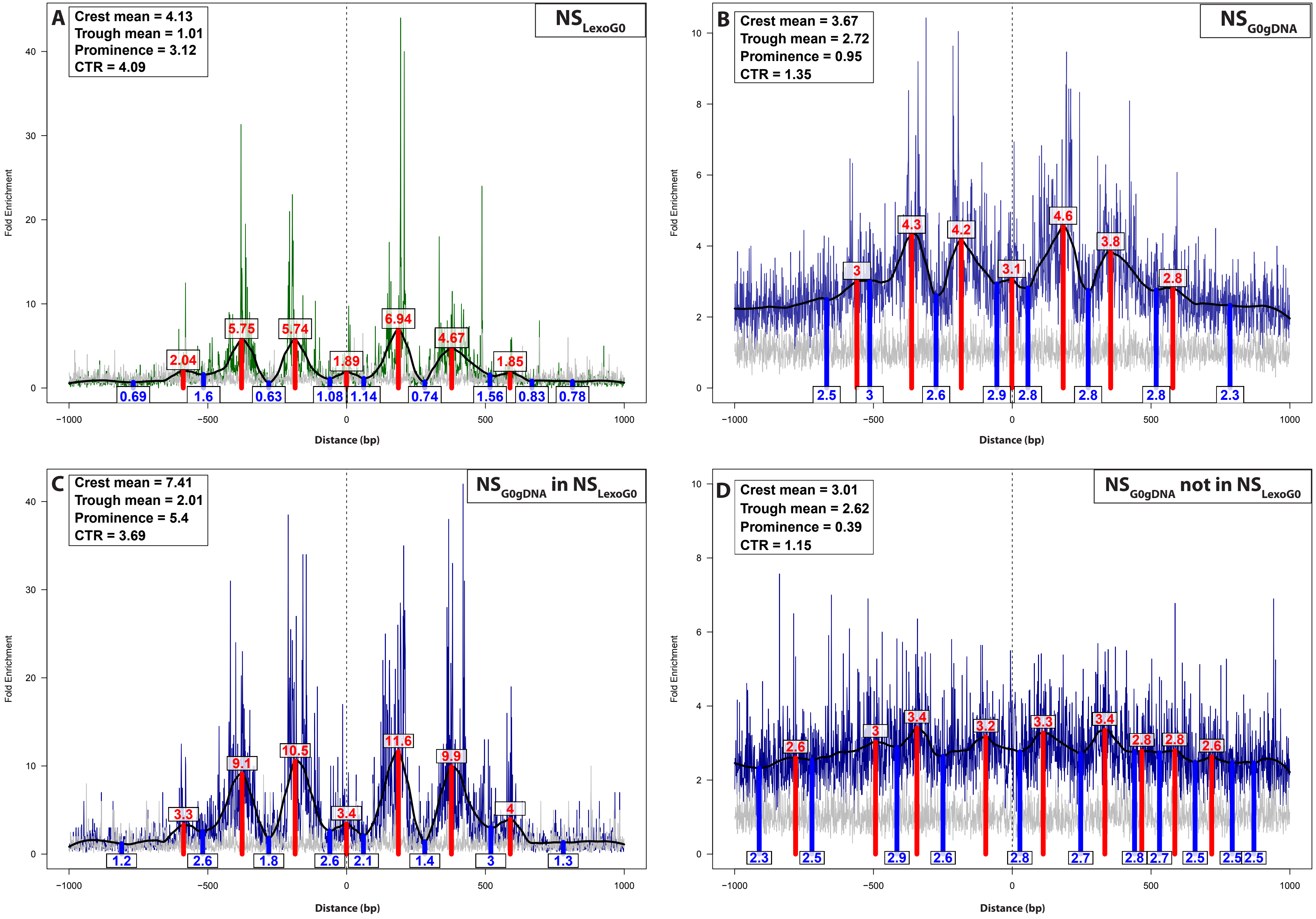

A
Crest mean = 4.13
Trough mean = 1.01
Prominence = 3.12
CTR = 4.09
NS_{LexoG0}
Fold Enrichment
0
10
20
30
40
2.04
5.75
5.74
1.89
6.94
4.67
1.85
0.69
1.6
0.63
1.08
1.14
0.74
1.56
0.83
0.78
−1000
−500
0
500
1000
Distance (bp)
B
Crest mean = 3.67
Trough mean = 2.72
Prominence = 0.95
CTR = 1.35
NS_{G0gDNA}
Fold Enrichment
0
2
4
6
8
10
3
4.3
4.2
3.1
4.6
3.8
2.8
2.5
3
2.6
2.9
2.8
2.8
2.8
2.3
−1000
−500
0
500
1000
Distance (bp)
C
Crest mean = 7.41
Trough mean = 2.01
Prominence = 5.4
CTR = 3.69
NS_{G0gDNA} in NS_{LexoG0}
Fold Enrichment
0
10
20
30
40
3.3
9.1
10.5
3.4
11.6
9.9
4
1.2
2.6
1.8
2.6
2.1
1.4
3
1.3
−1000
−500
0
500
1000
Distance (bp)
D
Crest mean = 3.01
Trough mean = 2.62
Prominence = 0.39
CTR = 1.15
NS_{G0gDNA} not in NS_{LexoG0}
Fold Enrichment
0
2
4
6
8
10
2.6
3
3.4
3.2
3.3
3.4
2.8
2.8
2.6
2.3
2.5
2.9
2.6
2.8
2.7
2.8
2.7
2.5
2.5
2.5
−1000
−500
0
500
1000
Distance (bp)

Figure S4.

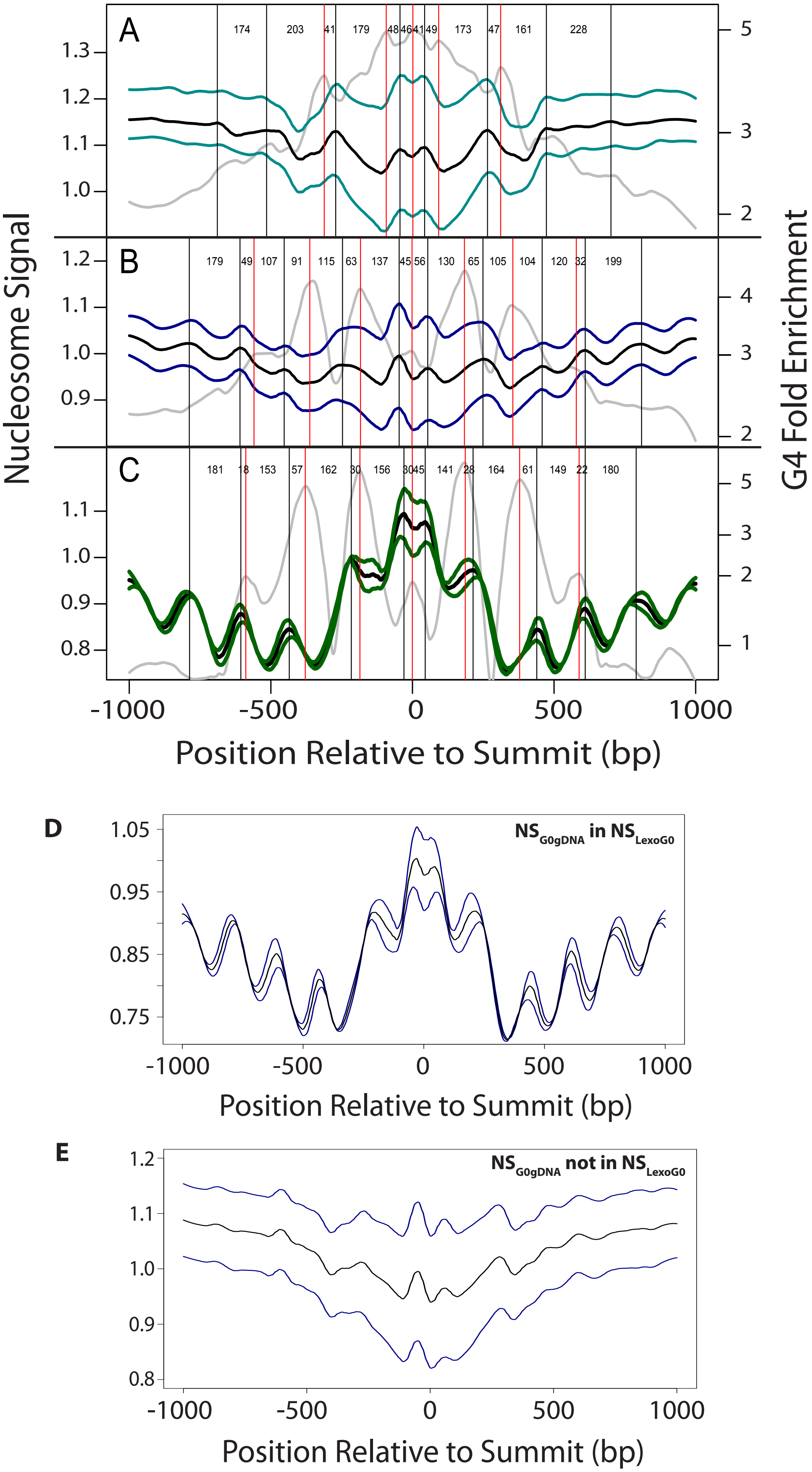

A
174
203
41
179
48
46
41
49
173
47
161
228
1.3
1.2
1.1
1.0
5
3
2
B
179
49
107
91
115
63
137
45
56
130
65
105
104
120
32
199
1.2
1.1
1.0
0.9
4
3
2
C
181
18
153
57
162
30
156
30
45
141
28
164
61
149
22
180
1.1
1.0
0.9
0.8
5
3
2
1
Nucleosome Signal
G4 Fold Enrichment
-1000
-500
0
500
1000
Position Relative to Summit (bp)
D
NS$_{G0gDNA}$ in NS$_{LexoG0}$
1.05
0.95
0.85
0.75
-1000
-500
0
500
1000
Position Relative to Summit (bp)
E
NS$_{G0gDNA}$ not in NS$_{LexoG0}$
1.2
1.1
1.0
0.9
0.8
-1000
-500
0
500
1000
Position Relative to Summit (bp)

Figure S5.

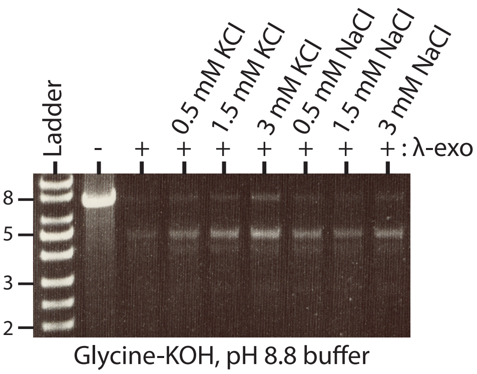

Ladder
0.5 mM KCl
1.5 mM KCl
3 mM KCl
0.5 mM NaCl
1.5 mM NaCl
3 mM NaCl
-
+
+
+
+
+
+
+ : λ-exo
8
5
3
2
Glycine-KOH, pH 8.8 buffer

Table S1: Read Mapping and Peak Calling Statistics

Mapping Statistics				
Sample	Read Length	Total # Reads	Mappable Reads	% Mappable
LexoG0 Rep1	50	162102273	115150124	71.1
LexoG0 Rep2	50	196524435	174625028	88.9
LexoG0 Rep3	50	174860103	153657189	87.9
LexoG0 pool	50	533486811	443432341	83.1
NS-seq Rep1	50	128247879	57397008	44.7
NS-seq Rep2	50	139320824	89354586	64.1
NS-seq Rep3	50	136227515	96762425	71
NS-seq pool	50	403796218	243514019	60.3
G0gDNA	50	193565007	181911420	94

Number of Peaks		
Sample	Background Control	# of Peaks Called
LexoG0$_{G0gDNA}$ Rep1	G0gDNA	110704
LexoG0$_{G0gDNA}$ Rep2	G0gDNA	194025
LexoG0$_{G0gDNA}$ Rep3	G0gDNA	183622
LexoG0$_{G0gDNA}$ pool	G0gDNA	196851
NS$_{G0gDNA}$ Rep1	G0gDNA	100594
NS$_{G0gDNA}$ Rep2	G0gDNA	95030
NS$_{G0gDNA}$ Rep3	G0gDNA	87013
NS$_{G0gDNA}$ pool	G0gDNA	162098
NS$_{LexoG0}$ pool	LexoG0	66831

Table S2: Fold Enrichment Correlation

Fold Enrichment Correlation of Replicates		
Sample1	**Sample2**	**FE**
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ Rep2	0.789778
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ Rep3	0.781041
LexoG0$_{G0gDNA}$ Rep2	LexoG0$_{G0gDNA}$ Rep3	0.950533
LexoG0$_{G0gDNA}$ Pool	LexoG0$_{G0gDNA}$ Rep1	0.84553
LexoG0$_{G0gDNA}$ Pool	LexoG0$_{G0gDNA}$ Rep2	0.975653
LexoG0$_{G0gDNA}$ Pool	LexoG0$_{G0gDNA}$ Rep3	0.971312
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ Rep2	0.675875
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ Rep3	0.571822
NS$_{G0gDNA}$ Rep2	NS$_{G0gDNA}$ Rep3	0.806056
NS$_{G0gDNA}$ Pool	NS$_{G0gDNA}$ Rep1	0.783419
NS$_{G0gDNA}$ Pool	NS$_{G0gDNA}$ Rep2	0.927668
NS$_{G0gDNA}$ Pool	NS$_{G0gDNA}$ Rep3	0.924786

NSG0gDNA and LexoG0gDNA Correlation		
Sample1	**Sample2**	**FE**
NS$_{G0gDNA}$ Pool	LexoG0$_{G0gDNA}$ Pool	0.557402
NS$_{G0gDNA}$ Pool	LexoG0$_{G0gDNA}$ Rep1	0.682616
NS$_{G0gDNA}$ Pool	LexoG0$_{G0gDNA}$ Rep2	0.511703
NS$_{G0gDNA}$ Pool	LexoG0$_{G0gDNA}$ Rep3	0.507415
LexoG0$_{G0gDNA}$ Pool	NS$_{G0gDNA}$ Pool	0.557402
LexoG0$_{G0gDNA}$ Pool	NS$_{G0gDNA}$ Rep1	0.714478
LexoG0$_{G0gDNA}$ Pool	NS$_{G0gDNA}$ Rep2	0.569375
LexoG0$_{G0gDNA}$ Pool	NS$_{G0gDNA}$ Rep3	0.325473

NS$_{LexoG0}$ Correlation		
Sample1	**Sample2**	**FE**
NS$_{LexoG0}$ Pool	LexoG0$_{G0gDNA}$ Pool	0.171399
NS$_{LexoG0}$ Pool	LexoG0$_{G0gDNA}$ Rep1	0.327537
NS$_{LexoG0}$ Pool	LexoG0$_{G0gDNA}$ Rep2	0.133805
NS$_{LexoG0}$ Pool	LexoG0$_{G0gDNA}$ Rep3	0.13193
NS$_{LexoG0}$ Pool	NS$_{G0gDNA}$ Pool	0.781847
NS$_{LexoG0}$ Pool	NS$_{G0gDNA}$ Rep1	0.422503
NS$_{LexoG0}$ Pool	NS$_{G0gDNA}$ Rep2	0.703584
NS$_{LexoG0}$ Pool	NS$_{G0gDNA}$ Rep3	0.849206

Table S3: LexoG0$_{G0gDNA}$ overlap analysis

fileA	fileB	expNum	obsNum	expProportion	obsProportion	pVal*	obsToExpRatio
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ Rep1	7383.991861	110704	0.066700317	1	0	14.99243256
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ Rep2	15100.44786	102960	0.136403814	0.930047695	0	6.818340817
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ Rep3	14015.89049	101496	0.126606902	0.91682324	0	7.241494936
LexoG0$_{G0gDNA}$ Rep1	LexoG0$_{G0gDNA}$ pool	15586.88905	105998	0.140797885	0.957490244	0	6.800459005
LexoG0$_{G0gDNA}$ Rep2	LexoG0$_{G0gDNA}$ Rep1	15100.83794	79055	0.077829341	0.407447494	0	5.235139953
LexoG0$_{G0gDNA}$ Rep2	LexoG0$_{G0gDNA}$ Rep2	30250.35358	194025	0.155909566	1	0	6.413974616
LexoG0$_{G0gDNA}$ Rep2	LexoG0$_{G0gDNA}$ Rep3	28146.5769	159841	0.145066754	0.823816518	0	5.67887884
LexoG0$_{G0gDNA}$ Rep2	LexoG0$_{G0gDNA}$ pool	31158.04845	178511	0.160587803	0.920041232	0	5.729209912
LexoG0$_{G0gDNA}$ Rep3	LexoG0$_{G0gDNA}$ Rep1	14016.20383	80497	0.076331833	0.438384289	0	5.743138511
LexoG0$_{G0gDNA}$ Rep3	LexoG0$_{G0gDNA}$ Rep2	28146.47907	165512	0.153284895	0.901373474	0	5.880380263
LexoG0$_{G0gDNA}$ Rep3	LexoG0$_{G0gDNA}$ Rep3	26181.34223	183622	0.142582818	1	0	7.013467774
LexoG0$_{G0gDNA}$ Rep3	LexoG0$_{G0gDNA}$ pool	28998.48508	173809	0.157924895	0.946558691	0	5.993726897
LexoG0$_{G0gDNA}$ pool	LexoG0$_{G0gDNA}$ Rep1	15587.34068	79185	0.079183447	0.402258561	0	5.080083999
LexoG0$_{G0gDNA}$ pool	LexoG0$_{G0gDNA}$ Rep2	31158.14638	174283	0.158282896	0.885354913	0	5.593497055
LexoG0$_{G0gDNA}$ pool	LexoG0$_{G0gDNA}$ Rep3	28998.67702	163796	0.147312826	0.832081117	0	5.648395612
LexoG0$_{G0gDNA}$ pool	LexoG0$_{G0gDNA}$ pool	32085.86815	196851	0.162995708	1	0	6.135130864

* pVal = 0 corresponds to a p value < 10e-323

Table S4: NS$_{G0gDNA}$ overlap analysis

fileA	fileB	expNum	obsNum	expProportion	obsProportion	pVal*	obsToExpRatio
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ Rep1	4765.535408	100594	0.047373953	1	0	21.10864602
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ Rep2	6336.469076	41533	0.062990527	0.412877508	0	6.554596811
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ Rep3	6876.176742	25147	0.068355734	0.249985089	0	3.657119493
NS$_{G0gDNA}$ Rep1	NS$_{G0gDNA}$ pool	10775.47803	79622	0.107118496	0.791518381	0	7.389184939
NS$_{G0gDNA}$ Rep2	NS$_{G0gDNA}$ Rep1	6336.781606	35764	0.066681907	0.376344312	0	5.643874481
NS$_{G0gDNA}$ Rep2	NS$_{G0gDNA}$ Rep2	7719.422757	95030	0.08123143	1	0	12.31050598
NS$_{G0gDNA}$ Rep2	NS$_{G0gDNA}$ Rep3	8083.091148	65670	0.085058309	0.691044933	0	8.124367126
NS$_{G0gDNA}$ Rep2	NS$_{G0gDNA}$ pool	13136.28301	92302	0.138233011	0.971293276	0	7.026492952
NS$_{G0gDNA}$ Rep3	NS$_{G0gDNA}$ Rep1	6876.732809	23076	0.079031097	0.265201751	0	3.355663313
NS$_{G0gDNA}$ Rep3	NS$_{G0gDNA}$ Rep2	8083.346126	42878	0.092898143	0.492776941	0	5.30448645
NS$_{G0gDNA}$ Rep3	NS$_{G0gDNA}$ Rep3	8330.722321	87013	0.095741123	1	0	10.44483259
NS$_{G0gDNA}$ Rep3	NS$_{G0gDNA}$ pool	13759.67614	78512	0.158133568	0.902301955	0	5.705948252
NS$_{G0gDNA}$ pool	NS$_{G0gDNA}$ Rep1	10776.0039	70852	0.066478327	0.43709361	0	6.574979058
NS$_{G0gDNA}$ pool	NS$_{G0gDNA}$ Rep2	13136.27618	70182	0.081039101	0.432960308	0	5.342609963
NS$_{G0gDNA}$ pool	NS$_{G0gDNA}$ Rep3	13759.23495	79603	0.084882201	0.49107947	0	5.78542341
NS$_{G0gDNA}$ pool	NS$_{G0gDNA}$ pool	22354.11578	162098	0.137904945	1	0	7.251371585

* pVal = 0 corresponds to a p value < 10e-323

Table S5: Overlap Analysis

file A	file B	expNum	obsNum	expProportion	obsProportion	pVal	obsToExpRatio
LexoG0$_{\text{G0gDNA}}$	LexoG0$_{\text{G0gDNA}}$	32085.86815	196851	0.162995708	1	0	6.135130864
LexoG0$_{\text{G0gDNA}}$	NS$_{\text{G0gDNA}}$	26783.96267	62230	0.136062111	0.316127426	0	2.323405269
LexoG0$_{\text{G0gDNA}}$	NS$_{\text{LexoG0}}$	12895.44162	12513	0.065508642	0.063565844	0.999762207	0.970342883
LexoG0$_{\text{G0gDNA}}$	G4	29934.79818	72554	0.152068306	0.368573185	0	2.423734396
LexoG0$_{\text{G0gDNA}}$	CpG	3830.944083	23865	0.019461136	0.121233827	0	6.229534935

file A	file B	expNum	obsNum	expProportion	obsProportion	pVal	obsToExpRatio
NS$_{\text{G0gDNA}}$	LexoG0$_{\text{G0gDNA}}$	26784.04091	76265	0.16523363	0.470486989	0	2.847404552
NS$_{\text{G0gDNA}}$	NS$_{\text{G0gDNA}}$	22354.11578	162098	0.137904945	1	0	7.251371585
NS$_{\text{G0gDNA}}$	NS$_{\text{LexoG0}}$	10741.98268	60434	0.066268447	0.372823847	0	5.625963271
NS$_{\text{G0gDNA}}$	G4	25311.32205	56600	0.156148269	0.349171489	0	2.236153445
NS$_{\text{G0gDNA}}$	CpG	3207.148297	13405	0.019785243	0.082696887	0	4.17972565

file A	file B	expNum	obsNum	expProportion	obsProportion	pVal	obsToExpRatio
NS$_{\text{LexoG0}}$	LexoG0$_{\text{G0gDNA}}$	12895.94605	13492	0.192963536	0.20188236	3.0316E-09	1.046220258
NS$_{\text{LexoG0}}$	NS$_{\text{G0gDNA}}$	10742.37149	62357	0.16073935	0.933055019	0	5.804770398
NS$_{\text{LexoG0}}$	NS$_{\text{LexoG0}}$	5057.979259	66831	0.07568313	1	0	13.21298419
NS$_{\text{LexoG0}}$	G4	13814.15162	23718	0.206702752	0.354895183	0	1.71693497
NS$_{\text{LexoG0}}$	CpG	1590.659661	39	0.023801225	0.000583562	1	0.02451813

file A	file B	expNum	obsNum	expProportion	obsProportion	pVal	obsToExpRatio
G4	LexoG0$_{\text{G0gDNA}}$	29931.68791	174050	0.083394456	0.484931056	0	5.814907617
G4	NS$_{\text{G0gDNA}}$	25308.61823	93270	0.070513846	0.259865094	0	3.685305897
G4	NS$_{\text{LexoG0}}$	13812.17603	24383	0.038482925	0.067934926	0	1.765326474
CpG	LexoG0$_{\text{G0gDNA}}$	3830.800926	26189	0.134366921	0.918589968	0	6.83642938
CpG	NS$_{\text{G0gDNA}}$	3207.019083	12600	0.112487516	0.441950193	0	3.928882141
CpG	NS$_{\text{LexoG0}}$	1590.538004	33	0.055788776	0.001157489	1	0.020747697

Table S6: G4 and CpG island density correlations

A. Peak density vs. feature density

	G4 (100 kb bins)		CpG Islands (1 Mb bins)	
q < 0.001 peak sets	**Pearson**	**Spearman**	**Pearson**	**Spearman**
LexoG0$_{G0gDNA}$ pool	0.704	0.704	0.646	0.746
NS$_{G0gDNA}$ pool	0.692	0.363	0.802	0.490
NS$_{LexoG0}$ pool	-0.248	-0.260	-0.364	-0.472

B. Average Fold Enrichment Signal vs. G4 density

	G4 (100 kb bins)	
Fold Enrichment Signal	**Pearson**	**Spearman**
LexoG0$_{G0gDNA}$ pool	0.862	0.776
NS$_{G0gDNA}$ pool	0.692	0.564
NS$_{LexoG0}$ pool	-0.124	0.004

Table S7: Peak summit windows and G4 overlap

Summits +/- 1kb	G4s	expNum	obsNum	expProportion	obsProportion	pVal	obsToExpRatio
LexoG0$_{\text{G0gDNA}}$	G4 centers	49865.5276	90810	0.2533161	0.4613134	0	1.821098
NS$_{\text{G0gDNA}}$	G4 centers	41062.03318	70772	0.2533161	0.4366001	0	1.723539
NS$_{\text{LexoG0}}$	G4 centers	16929.36828	23248	0.2533161	0.3478625	0	1.373235

G4s	Summits +/- 1kb	expNum	obsNum	expProportion	obsProportion	pVal	obsToExpRatio
G4 centers	LexoG0$_{\text{G0gDNA}}$	49856.47853	156537	0.1389081	0.436137	0	3.139752
G4 centers	NS$_{\text{G0gDNA}}$	41054.57748	115280	0.1143846	0.3211885	0	2.807969
G4 centers	NS$_{\text{LexoG0}}$	16926.2996	26481	0.04715937	0.07378029	0	1.564489

	LexoG0$_{\text{G0gDNA}}$	NS$_{\text{G0gDNA}}$	NS$_{\text{LexoG0}}$
numSummits +/- 1kb that overlap G4s	90810	70772	23248
% of all summits +/- 1kb that overlap >= 1 G4	46.13134	43.66001	34.78625
Of those that overlap >= 1 G4, % that overlaps 1 G4:	56.77128	60.97044	91.63369
Of those that overlap >= 1 G4, % that overlaps 2 G4s:	23.2067	18.25157	5.785444
Of those that overlap >= 1 G4, % that overlaps 3 G4s:	9.767647	9.399197	0.8817963
Of those that overlap >= 1 G4, % that overlaps 4 G4s:	4.740667	5.269033	0.5204749
Of those that overlap >= 1 G4, % that overlaps 5 G4s:	2.469992	2.797717	0.4129387
Of those that overlap >= 1 G4, % that overlaps 6 G4s:	1.342363	1.514723	0.2623882
Of those that overlap >= 1 G4, % that overlaps >= 7 G4s:	1.701351	1.79732	0.5032679
On average, of those that overlap >= 1 G4, overlaps this many G4s:	1.865907	1.853035	1.163068

Table S8: Nucleosome and G4 correlations

A

Experiment	A	B	Pearson	Spearman
LexoG0$_{\text{G0gDNA}}$	mean nucleosome smoothed	G4	-0.84944833	-0.9062846
LexoG0$_{\text{G0gDNA}}$	K562 smoothed nucleosome	G4	-0.95448538	-0.96137874
LexoG0$_{\text{G0gDNA}}$	GM12878 smoothed nucleosome	G4	-0.00048484	-0.20675595
LexoG0$_{\text{G0gDNA}}$	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.1891064	0.3402856
LexoG0$_{\text{G0gDNA}}$	K562 smoothed nucleosome	mean nucleosome smoothed	0.9499651	0.976812
LexoG0$_{\text{G0gDNA}}$	GM12878 smoothed nucleosome	mean nucleosome smoothed	0.4863646	0.5113621
LexoG0$_{\text{G0gDNA}}$	K562 raw nucleosome	GM12878 raw nucleosome	0.2135404	0.366179
NS$_{\text{G0gDNA}}$	mean nucleosome smoothed	G4	-0.80415049	-0.87139479
NS$_{\text{G0gDNA}}$	K562 smoothed nucleosome	G4	-0.79586782	-0.86944983
NS$_{\text{G0gDNA}}$	GM12878 smoothed nucleosome	G4	-0.42899402	-0.48135622
NS$_{\text{G0gDNA}}$	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.308657032	0.355864991
NS$_{\text{G0gDNA}}$	K562 smoothed nucleosome	mean nucleosome smoothed	0.909581044	0.912055128
NS$_{\text{G0gDNA}}$	GM12878 smoothed nucleosome	mean nucleosome smoothed	0.675986399	0.680323131
NS$_{\text{G0gDNA}}$	K562 raw nucleosome	GM12878 raw nucleosome	0.340282529	0.413766276
NS$_{\text{LexoG0}}$	mean nucleosome smoothed	G4	-0.00690628	-0.04702974
NS$_{\text{LexoG0}}$	K562 smoothed nucleosome	G4	-0.08227209	-0.08133883
NS$_{\text{LexoG0}}$	GM12878 smoothed nucleosome	G4	0.047354732	-0.03519783
NS$_{\text{LexoG0}}$	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.950578073	0.968156115
NS$_{\text{LexoG0}}$	K562 smoothed nucleosome	mean nucleosome smoothed	0.983128609	0.9888404
NS$_{\text{LexoG0}}$	GM12878 smoothed nucleosome	mean nucleosome smoothed	0.991333183	0.993413977
NS$_{\text{LexoG0}}$	K562 raw nucleosome	GM12878 raw nucleosome	0.946049507	0.961888765

B

Experiment	A	B	Pearson	Spearman
NS$_{\text{G0gDNA}}$ (in NS$_{\text{LexoG0}}$)	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.9105249	0.926113
NS$_{\text{G0gDNA}}$ (in NS$_{\text{LexoG0}}$)	K562 raw nucleosome	GM12878 raw nucleosome	0.9051878	0.9230251
NS$_{\text{G0gDNA}}$ (not in NS$_{\text{LexoG0}}$)	K562 smoothed nucleosome	GM12878 smoothed nucleosome	0.901602	0.938893
NS$_{\text{G0gDNA}}$ (not in NS$_{\text{LexoG0}}$)	K562 raw nucleosome	GM12878 raw nucleosome	0.903426	0.933372

C

Sample	Sum of deviations^2 from mean over each position
Before Deconvolution	
LexoG0$_{\text{G0gDNA}}$	28.86581618
NS$_{\text{G0gDNA}}$	17.95427157
NS$_{\text{LexoG0}}$	1.43678066
After Deconvolution of NS$_{\text{G0gDNA}}$	
NS$_{\text{G0gDNA}}$ (peaks in NS$_{\text{LexoG0}}$)	1.4291
NS$_{\text{G0gDNA}}$ (peaks NOT in NS$_{\text{LexoG0}}$)	28.85379

Table S9: rDNA read mapping statistics

Sample Name	Total num raw Reads	num Mappable Reads to hg19+rDNA	num reads mapped to rDNA repeat in context of hg19	num reads (mapq >= 2) mapped to rDNA repeat in context of hg19	number non-redundant* reads with mapq >= 2
LexoG0 Rep1	162102273	116086307	1901110	1615177	1203114
LexoG0 Rep2	196524435	176323067	3291863	2767654	1893121
LexoG0 Rep3	174860103	155299989	3097505	2588199	1781096
LexoG0 Pool	533486811	447709363	8290478	6971030	4877331
NS-seq Rep1	128247879	57840335	848305	725977	605176
NS-seq Rep2	139320824	89873195	954705	809750	702167
NS-seq Rep3	136227515	97416920	1209455	1028929	882839
NS-seq pool	403796218	245130450	3012465	2564656	2190182
G0 gDNA	193565007	181619332	696286	614554	610713

SUPPLEMENTARY MATERIALS

SUPPLEMENTARY FIGURES

Figure S1

(A) and (B): Barplot visualizations of the observed proportion of peak overlaps compared to the expected proportion of peak overlaps between replicates with each other and between replicates and the peak set resulting from pooling all reads. The proportions are of the peak set written in horizontal words that brackets three other peak sets. For example, the first three pairs of expected and observed proportions are of Rep1 that overlap the sets labeled under each expected and observed pair of bars (Rep2, Rep3, and Pool). This figure is related to Tables S3 and S4 where the expected and observed proportion values can be found along with p-values and other information.

(A) LexoG0$_{\text{G0gDNA}}$ peaks were called as described in Supplementary Methods. We identified 110,704 peaks, 194,025 peaks and 183,622 peaks in LexoG0 Reps1-3, respectively, and 196,851 peaks in the LexoG0 pooled data set. We observed significantly higher overlap of the peaks in each replicate with those in the LexoG0$_{\text{G0gDNA}}$ peak set from pooled reads (95.7%, 92.0% and 94.7%, respectively) than would be expected at random ($p<10^{-323}$). These results strongly support the conclusion that we were able to reproducibly identify peaks derived from λ-exonuclease (λ-exo) digested non-replicating DNA genome-wide. Moreover, the peak set from the pooled LexoG0 reads is representative of all the biological replicates, and most analyses were performed using this set of peaks from pooled reads.

(B) NS$_{\text{G0gDNA}}$ peaks were called as described in the Supplementary Methods. We identified 100,594 peaks, 95,030 peaks and 87,013 peaks in NS-seq Rep1-3, respectively, and 162,098 peaks from the NS-seq pooled read data set. The replicates all had significantly higher overlap than expected at random (also see Table S4). All of the replicates showed significant overlap ($p<10^{-323}$) with the NS_{G0gDNA} peak set called from pooled reads (79.1%, 97.1% and 90.2%, respectively). These results suggest that we were able to reproducibly detect peaks enriched by λ-exo digestion of replicating DNA and that the NS_{G0gDNA} peak set from pooled reads is representative of the individual replicates, so most analyses were performed with peaks resulting from the pooled set of reads.

(C) and (D): Venn diagrams of the number of overlaps between (C) all $LexoG0_{G0gDNA}$ peak sets (Reps 1-3 and set from pooled reads) and (D) all NS_{G0gDNA} peak sets (Reps 1-3 and set from pooled reads).

For both (C) and (D), black text is replicate 1, blue text is replicate 2, brick red text is replicate 3, and green text is the set of peaks from pooled reads. Text and ellipsis color correspond for a given set. The four-way Venn diagram graphic was downloaded from http://www.math.cornell.edu/~numb3rs/lipa/imgs/venn4.png. Venn diagram values were obtained with a custom Python script that employed pybedtools (Dale et al, 2011) and were used to annotate the four-way Venn diagram graphic. Note that the area (size) of each section in the four-way Venn diagram does not correspond to the values within it. For both $LexoG0_{G0gDNA}$ and NS_{G0gDNA}, most of the mass (sum of percentages) of each replicate is within the pooled data set (green ellipsis), showing that it represents each well. Summing all percentages of a given color gives 100% (i.e. the full dataset represented by that color).

Figure S2

(A) Integrative look at origin activity and chromatin marks in human rDNA repeats: At the top, NS-seq fold-enrichment signal when using G0gDNA (light blue-grey) and LexoG0 (green) as the control is shown the same way as in Figure 5B. Fold enrichment values are on the left Y-axis. The blue and red circles represent G4 counts in 1 kb bins on the positive and negative strand, respectively. G4 counts are also on the left Y-axis. The %GC signal is shown as a red line and is measured on the right Y-axis. The rRNA gene is depicted inside the plot with an arrow representing the transcription start site and direction of transcription. Note that rDNA repeats are typically tandem repeats and that the positions from 30-43 kb represent the upstream region, including the promoter, for the next rDNA repeat. Below are bars representing sites where previous studies found rDNA replication initiation activity. In general, the white bar (black outline) represents the entire area where initiation was detected, while the black bars represent sites of most frequent initiation activity. Although two studies detect initiation events everywhere across the rDNA repeat, most studies find replication activity restricted, or most frequent, in the intergenic spacer (also known as the 'non-transcribed spacer', NTS). Moreover, the most initiation activity across all studies appears to be between positions 30-43 kb and/or around positions 14-20 kb depending on the study. These areas are also the highest areas in the fold enrichment signal from our data, which suggest there are 3 preferred areas for initiation (near 15.5-16.5 kb, 31.5-34 kb, and 38-41.5 kb). Below the replication initiation bars are bars that represent three groups (#1, #2, and #3) of chromatin marks from a recent study (Zentner et al, 2011). Group #1 represents H3K4me1, H3K4me2, H3K4me3, and H3K9ac. Group #2 represents H3K27ac only. Group #3 represents both H3K27me3 and H4K20me1. Similar to the representation of initiation activity above, the light blue bars (black outline) represent the area where a given chromatin mark is detected and the blue bars represent where the given mark was most enriched. H3K27ac marks the initiation zone that seems to occur near 15-20 kb and a pair of marks, H3K27me3 and H4K20me1, seem to coincide with most of the initiation zone between 30-43 kb. All relevant rDNA references are listed in the figure.

(B) Comparison of GC content in NS-seq vs. LexoG0 reads: The log2(fold change) of the distribution of GC content in the three replicates of NS-seq reads relative to the pooled LexoG0 reads (i.e. log2(NS-seq/LexoG0)). This figure is supplementary to Figures 2A and 2B in the paper, is purposefully plotted on the same Y-axis scale (for direct comparison), and shows directly that NS-seq reads are enriched in AT-rich reads and depleted in GC-rich reads relative to LexoG0 reads. Over each GC% the minimum to maximum (line segment), median (black dot), and mean (red triangle) values for the NS-seq replicates (relative to the pooled LexoG0 reads) are shown.

(C) *MYC* locus: The $\text{NS}_{\text{G0gDNA}}$, $\text{LexoG0}_{\text{G0gDNA}}$ and $\text{NS}_{\text{LexoG0}}$ peak sets are illustrated at the *MYC* locus where origin activity in the promoter region and in exon two has been well characterized in HeLa cells (Tao et al. 2000). $\text{NS}_{\text{G0gDNA}}$ identified three peaks across this locus (blue), overlapping (i) the first MYC exon and upstream promoter region, (ii) the second MYC exon, and (iii) the last MYC exon. In purple are The locations of CpG islands, G4 motifs and GC content across the locus are indicated in purple. The $\text{LexoG0}_{\text{G0gDNA}}$ peaks (cyan) overlap the CpG islands and G4 motifs as well as the first two $\text{NS}_{\text{G0gDNA}}$ peaks. $\text{NS}_{\text{LexoG0}}$ (green) lacks the upstream peak that overlaps with CpG islands, G4 motifs, and a $\text{LexoG0}_{\text{G0gDNA}}$ peak, but contains the second exon peak that also overlaps these features. The third exon peak, which does not overlap these features, was preserved as expected. The first exon peak had the weakest fold-enrichment in $\text{NS}_{\text{G0gDNA}}$, suggesting that the preferred initiation sites in MCF7 cells are over the second and third exons in contrast to HeLa cells (Tao et al. 2000). Given that the first exon peak is absent in $\text{NS}_{\text{LexoG0}}$, there may be some loss of sensitivity to weakly enriched origins in strongly λ-exo-biased regions. However, the presence of the second exon peak in $\text{NS}_{\text{LexoG0}}$ demonstrates the advantage of controlling NS-seq with LexoG0 over the alternative procedure of discarding all $\text{NS}_{\text{G0gDNA}}$ peaks that overlap $\text{LexoG0}_{\text{G0gDNA}}$ peaks.

Figure S3

Each panel shows the G4 enrichment signal around the specified set of summits (not strand-oriented) and, for each, measures crest heights (red vertical bars and numbers), trough heights (blue vertical bars and numbers), calculates the crest and trough means from each set of heights, and from the means computes “prominence” ($crest_{mean}$-$trough_{mean}$) and the crest-to-trough ratio (CTR = $crest_{mean}$/$trough_{mean}$), which is a measure of how phased the signal is around crests relative to troughs. Height, prominence, and CTR measurements were performed for **(A)** all NS_{LexoG0} summits, **(B)** all NS_{G0gDNA} summits **(C)** NS_{G0gDNA} summits represented in NS_{LexoG0}, and **(D)** NS_{G0gDNA} summits not represented in NS_{LexoG0}. NS_{G0gDNA} summits were considered to be represented in NS_{LexoG0} if they mapped inside of an NS_{LexoG0} summit window, where a summit window is a summit +/- 1 kb. Partitioning the NS_{G0gDNA} summits this way decomposes the relatively dampened wave-like NS_{G0gDNA} G4 enrichment signal (compared to NS_{LexoG0}) into a more prominent and phased signal **(C)** and a roughly uniform signal **(D)**.

Figure S4

(A-C): This figure is similar to Figure 6 in the paper, but shows both the crest positions of nucleosomes (vertical black lines) and of G4 enrichments (vertical red lines) with distances between adjacent crests (regardless of crest-type) for the three subsets of peak summits within 1 kb of G4s: **(A)** $LexoG0_{G0gDNA}$ (cyan), **(B)** NS_{G0gDNA} (blue), **(C)** NS_{LexoG0} (green). This allows easy comparison of where G4 enrichment crests are with respect to nucleosome signal crests. In all cases they are offset from each other. The colored lines show the mean nucleosome signal over each position around summits for K562 and GM12878 cell lines while the central black lines show the mean of the two cell line signals. The light grey line shows the log transformed G4 enrichment profile. The right Y-axis shows G4 enrichment values (not log transformed), which are not uniformly spaced since they are mapped onto a log scale. Also see Table S7, which is related to this figure. The distances from left to right in bp are: $LexoG0_{G0gDNA}$

= 174, 203, 41, 179, 48, 46, 41, 49, 173, 47, 161, 228; NS_{G0gDNA} = 179, 49, 107, 91, 115, 63, 137, 45, 56, 130, 65, 105, 104, 32, 199; NS_{LexoG0} = 181, 18, 153, 57, 162, 30, 156, 30, 45, 141, 28, 164, 61, 149, 22, 180.

(D-E): The nucleosomal signal around **(D)** NS_{G0gDNA} summits that are represented in NS_{LexoG0} and around **(E)** NS_{G0gDNA} summits that are not represented in NS_{LexoG0}. For **D-E**, NS_{G0gDNA} summits were considered to be represented in NS_{LexoG0} if they overlapped a NS_{LexoG0} summit window (summit +/- 1 kb). Partitioning the NS_{G0gDNA} summits this way decomposes the NS_{G0gDNA} nucleosomal signal that has less consistent positioning compared to NS_{LexoG0} into a stronger, more consistent wave-like signal **(D)** and a less wave-like, less consistent signal **(E)**. Also see Table S8.

Figure S5

Increasing the concentration of potassium present in the plasmid experiments (glycine-KOH buffer, pH 8.8) by titrating KCl resulted in stronger bands signifying that more G4 structures were stabilized, thwarting λ-exo digestion. Increasing the concentration of sodium ions present in the glycine-KOH buffer (pH 8.8) by titrating NaCl did not result in stronger bands signifying that Na+ ions did not contribute to the stabilization of more G4 structures. It is likely that sodium ions did not contribute much to G4 stability compared to K^+ ions because G4 folding and unfolding kinetics differ in the presence of each (Shim at al, 2009). G4s fold fast and unfold slowly in the presence of K^+ while both folding and unfolding are fast in the presence of Na^+. In addition, the melting temperature for G4s stabilized by Na^+ ions is much lower than that for G4s stabilized by K^+ ions (Kankia and Marky, 2001). At 37˚C (the temperature of λ-exo digestion in our experiments and those of others), most G4s are not likely stabilized by Na^+, but they are very likely stabilized in the presence of K+.

SUPPLEMENTARY TABLES

Table S1

Shows numbers of reads obtained from Illumina HiSeq 2000, numbers of reads that were mapped to hg19, and numbers of peaks in the peak sets discussed in the paper.

Table S2

Shows Pearson's product-moment correlation coefficients (Pearson's *r*) of genome-wide fold enrichment (FE) signals (wigCorrelate). When comparing replicates to each other, it is a measure of reproducibility. When comparing a replicate to the fold enrichment signal resulting from the pooled reads, it is a measure of how well the pooled data represent the replicate.

LexoG0$_{\mathbf{G0gDNA}}$ replicates and pooled set:

The genome-wide fold enrichment signals from the three replicates were highly correlated with each other showing Pearson's *r* ranging from 0.78 to 0.95. All three replicates were highly correlated with the fold enrichment signal obtained from pooled reads as well (Pearson's *r* = 0.84, 0.97 and 0.97, for Rep1-3 respectively). As all replicate fold enrichment signals were highly correlated (and peak sets significantly overlapped; Table S3) and since the pooled data set was a balanced representation of each (determined by FE signal correlation and peak overlaps; Table S3), the peaks resulting from the pooled data set were used for most analyses and the pooled set of LexoG0 reads was used as the LexoG0 control for NS$_{\text{LexoG0}}$.

NS$_{\mathbf{G0gDNA}}$ replicates and pooled set:

The fold enrichment signals of NS$_{\text{G0gDNA}}$ replicates were highly correlated with each other displaying Pearson's *r* ranging from 0.67 to 0.81. Additionally, each of the replicates was highly correlated with peaks called from the pooled set of reads (Pearson's *r* = 0.78, 0.93 and 0.92, for Rep1-3 respectively). As all replicates were highly correlated and reproducible and since the pooled data set was representative of each (both also determined by peak overlap; Table S4), the peaks resulting from the pooled data set were used for most analyses and the pooled set of NS-seq reads was used as the NS treatment file for NS_{LexoG0} MACS2 peak calling.

Tables S3 and S4

Overlap statistics between replicate peak sets and the peak set resulting from pooled reads for NS_{G0gDNA} and $LexoG0_{G0gDNA}$. The observed and expected proportions are visualized in Figure S2. The observed and expected number of overlaps is the observed and expected number of peaks in fileA that overlap peaks in fileB. Note that in the tables, a p-value of 0 simply means it was so small that R considered it 0, which occurs around 2.5e-324. Thus, elsewhere when discussing p-values, if it was 0 in R, we say $p < 10^{-323}$. The expected proportion and p-values were obtained through the binomial model described in the Supplementary Methods. Significant overlap between replicates is a measure of reproducibility while significant overlap between replicates and the peak set resulting from pooled reads is a measure of how well the pooled set represents the replicate.

Table S5

Overlap analysis of peak sets (resulting from pooled read sets) with each other and with other features (CpG islands and G4s). The observed number of overlaps is the number of peaks in file A that overlap peaks in file B. When p-value is 0, interpret it as $p < 10^{-323}$.

Table S6

Correlations of peaks or average fold enrichment with given feature in 100 kb or 1 Mb bins. Both Pearson's *r* and Spearman's rank-order correlation are given.

Table S7

Since the G4 enrichment signal around NS_{G0gDNA} and NS_{LexoG0} was phased, with inter-crest distances reminiscent of nucleosome spacing, it suggested that there was a relationship between G4s and nucleosomes. Thus, nucleosomal signal was assayed around the subset of summits that were proximal to G4s, defined as summits that have $\geq$ 1 G4 motif within 1 kb in either direction. A summit window is defined here as a summit +/- 1 kb. This table provides the statistics on how many summit windows overlap G4s and vice versa. Moreover, for summit windows that overlap $\geq$ 1 G4, how many overlap 1 G4, 2 G4s, 3 G4s (etc) is shown. The G4 enrichment signal that summarizes all NS_{LexoG0} summits is highly prominent and phased around the summit position (Figure S3A). Nonetheless, most (91.6%) of the 34.7% of NS_{LexoG0} summits that are proximal to G4s have just a single G4 nearby, which is typically 3' to the summit when strand information is considered (Figure 3F) and is typically spaced 1-3 nucleosomal distance units (185-210 bp) away from the NS summits (Figure 3C).

Table S8

Nucleosome signal was plotted around the subset of peak summits that were proximal to G4s (had G4s within 1 kb; Table S7). The consistency of nucleosomal positioning relative to these peak summits was tested by correlation as well as how much variation there was between the 2 cell line signals. Note that the raw nucleosome signal for a given cell line is the fold enrichment of the mean nucleosome score over each relative position from the summit (for all specified summits) divided by the "genomic mean score at random" over each position (from shuffling the specified peak summits). The smoothed nucleosomal signal for a given cell line results from lightly loess smoothing the raw signal defined above to round out jagged edges. The “mean nucleosome smoothed” signal results from taking the overall mean from the 2 cell line raw nucleosomal signal means (defined above) at each position and lightly loess smoothing it to round out jagged edges. **(A)** Correlations between cell line signals and for cell line signals with mean signals. The correlations between G4 and nucleosome signals are also provided. **(B)** Pearson's *r* and Spearman's rank-order correlation between the nucleosome signal from each cell line after partitioning the NS_{G0GDNA} summits into NS_{G0GDNA} summits that overlap with NS_{LexoG0} summit windows and those that do not overlap with NS_{LexoG0} summit windows. **(C)** The amount of variation between the two cell line signals was measured by taking the sum of squared deviations of the "smoothed cell line signals" (defined above) from the "mean nucleosome smoothed signal" (defined above) over each position.

Table S9

Numbers of Illumina HiSeq 2000 reads for each sample that mapped to hg19+rDNA, a modified hg19 genome that contained a copy of the 43 kb rDNA repeat (http://www.ncbi.nlm.nih.gov/nuccore/555853?report=fasta) as an additional "chromosome", and how many reads mapped to the rDNA repeat itself.

SUPPLEMENTARY METHODS

λ-exonuclease (λ-exo) digestion of plasmid DNA. pFRT.myc6xERE is a 7180 bp plasmid that contains a 2.4 kb genomic fragment from the promoter region of the *MYC* gene (Malott and Leffak, 1999) with a region shown to be unnecessary for origin activity (Δ11; Liu et al. 2003) replaced by a 6x estrogen response element (6xERE) cassette. This construct contains the NHE III_1 element of the *MYC* promoter. The purine-rich strand of this element has been shown to form a G4 structure (Pu27; reviewed by Brooks and Hurley 2010). Plasmid DNA was linearized with BglII (New England Biolabs (NEB)), purified using Ampure beads (Beckman Coulter) and labeled at the 3' end using terminal transferase (New England Biolabs) and α^{32}P-dCTP (Perkin Elmer) under conditions that add 3-6 nucleotides (~1:400 ratio of 3' ends to α^{32}P-CTP; 37° C for 1 hr.). Labeled fragments were purified over a Sephadex G-50 column (Sigma) and 200 ng was digested overnight (16-18 hours) with λ-exo in the buffer indicated; ten units of a custom, high concentration preparation of λ-exo from Fermentas (20 units/μl) were used per reaction (enzyme:DNA = 50 units/μg). Four buffer conditions were used: 67 mM glycine-KOH pH 8.8 and pH 9.4 or 67 mM glycine-NaOH pH 8.8 and pH 9.4; all with 2.5 mM $MgCl_2$ and 50 μg/ml bovine serum albumin. The linearized plasmid was made single stranded, as necessary, by boiling for 5 minutes and transferring directly to ice. Reaction products were run out on 0.8% agarose, dried on a gel dryer and exposed to a phosphoimaging plate. In unlabeled plasmid experiments, about 700 ng of single stranded plasmid DNA and 20 units of λ-exo were used per reaction (enzyme:DNA = 28.6 units/μg). The digestion products were run on 0.8% agarose and stained with ethidium bromide. G4 deletion mutants of pFRT.myc6xERE were generated with the Q5 Site-directed Mutagenesis Kit (New England Biolabs) following the manufacturer's directions. Pu27 was deleted and replaced with a HindIII restriction site using the following primers: oMycG4Pu27for, 5'-CTTATAAGCGCCCCTCCCGGG-3'; oMycG4Pu27rev, 5'-

CTTGAGGAGACTCAGCCGGGC-3’). Pu30 was deleted and replaced with a BamHI restriction site (oMycG4Pu30for, 5’-TCCGTACAGACTGGCAGAGAG-3’; oMycG4Pu30rev 5’-TCCACACGGAGTTCCCAATTTC-3’).

Predicting G4s in the plasmid sequence. The QGRS mapper (Kikin et al. 2006; http://bioinformatics.ramapo.edu/QGRS) was used with default parameters to predict another G4 sequence (Pu30) in the pFRT.myc6xERE sequence. QGRS was used for this analysis as it offers the advantage of providing “G scores” for each G4 candidate, with higher scores belonging to candidates that are more likely to actually form G4s.

λ-exonuclease (λ-exo) digestion and sequencing of non-replicating DNA (LexoG0). Three biological replicates were performed. For each, genomic DNA was purified from serum starved cells (9.6% S-phase) with 15 ml of DNAzol (Invitrogen) following the manufacturer’s directions and resuspended in DNA hydration buffer (Qiagen). 150 μg of DNA was sonicated to a size range of 200 bp to 10 kb in a Biorupter Standard (Diagenode) and purified with Agencourt Ampure XP beads (Beckman Coulter). In order to investigate the genome-wide, nascent strand independent λ-exo biases in the genomic DNA it is important to avoid enriching the small amount of contaminating S-phase DNA that may be present. Fragmentation of the DNA by sonication breaks any long, RNA-primed nascent strands associated with replication forks into smaller fragments, ensuring that short RNA protected fragments (if present) are distributed throughout the genome rather than only near origins, thus preventing origin sequences from being accidentally enriched. The fragmented DNA was made single stranded by boiling for 10 minutes, and transferring to ice. The 5’ ends were phosphorylated with 50 units of T4 Polynucleotide Kinase (T4 PNK; New England Biolabs) for 1 hour at 37° C. The reaction was stopped by incubating 15 minutes at 75° C. Following phenol:chloroform extraction and ethanol precipitation, the phosphorylated fragments were digested with 100 units of λ-exo (Fermentas) in glycine-KOH pH 9.4 buffer in a total volume of 100 µl. The reaction was stopped by incubating 15 minutes at 75° C before the samples were phenol:chloroform extracted and ethanol precipitated. λ-exo digested fragments were electrophoresed on a 1.5% UltraPure LMP agarose (Invitrogen) gel. Fragments in the range of 500-1500 nt were then purified by melting at 65° C for 10 min before sequential extraction with phenol, phenol-chloroform, and chloroform, followed by resuspension in 10 µl elution buffer (Qiagen). The concentration was determined by NanoDrop (Thermo Scientific); the starting DNA samples were depleted ~1000-fold. The purified single stranded fragments were made double stranded with random hexamers and Klenow (New England Biolabs), then sonicated to a size of 100-600 bp. Illumina libraries were prepared using the NEBNext kit (New England Biolabs) following the manufacturer's directions. 200-500 bp library fragments were size selected on 2% NuSieve agarose (Lonza) and were gel purified (Qiagen). Libraries were sequenced on the Illumina HiSeq 2000 platform.

Sequencing of undigested non-replicating genomic DNA (G0gDNA). For the G0gDNA control, undigested genomic DNA from serum starved MCF7 cells (6.8% S-phase) was sonicated to a size range of 100-600 bp, and Illumina libraries were prepared using the NEBNext kit (New England Biolabs) following the manufacturer's directions. 200-500 bp library fragments were size selected on 2% NuSieve agarose (Lonza) and were gel purified (Qiagen). Libraries were sequenced on the Illumina HiSeq 2000 platform.

λ-exonuclease (λ-exo) digestion and sequencing of replicating DNA: Nascent-strand sequencing (NS-seq). Three biological replicates were performed. For each, genomic DNA was purified from asynchronously growing MCF7 cells (35-40% S-phase) with 15 ml of DNAzol (Invitrogen) following the manufacturer's directions and resuspended in DNA hydration buffer (Qiagen). Nascent strands were prepared by adapting the protocol developed for replication initiation point mapping (Gerbi and Bielinsky, 1997) for NS-seq. DNA was handled gently to prevent breakage of long RNA-primed nascent DNA throughout the entire preparation in order to keep short RNA-primed nascent DNA close and specific to origins (in contrast to LexoG0 where purposeful fragmentation was performed). Replicative Intermediate (RI) DNA was enriched from 150 µg of genomic DNA by BND-cellulose chromatography (Sigma). Typically, 40 to 50 µg (~25-30%) of starting material was recovered. The RI DNA was made single stranded by boiling for 10 minutes and transferring to ice. The 5' ends were phosphorylated with 50 units of T4 PNK for 1 hour at 37° C and the reaction was stopped by incubating 15 minutes at 75° C. The fragments were digested with 100 units of λ-exo in glycine-KOH pH 8.8 buffer in a total volume of 100 µl. The enzyme:DNA ratio was kept low (2-2.5 units/µg DNA) to preserve the nascent strands because λ-exo can lose specificity at high enzyme:DNA ratios and digest the RNA primer at the 5' end of DNA (Yang and Li 2013). λ-exo digested fragments were electrophoresed on a 1.5% UltraPure LMP agarose gel and fragments in the range of 500-1500 nt were purified and resuspended in Qiagen elution buffer. The concentration was determined by Nanodrop; 21-96 ng of DNA was recovered for the replicates reported here, representing a ~500-2500 fold depletion of the starting DNA. Nascent strand enrichment was determined by qPCR at the *MYC* locus using the following primers:

control locus:

oMyc RT set 1-2 fwd, 5'-TTGCCAATTGCCTCTGGTTGAGAC-3';

oMyc RT set 1-2 rev, 5'-GACTTTGCTGTTTGCTGTCAGGCT-3';

test primers:

oMyc RT set 16-2 fwd, 5'- TGAACCAGAGTTTCATCTGCGACC-3';

oMyc RT set 16-2 rev, 5'- AGAAGCCGCTCCACATACAGTCCT-3'.

Sequencing libraries were made from nascent strand preparations where the *MYC* origin was >60-fold enriched. Single stranded nascent strands were made double stranded with random hexamers and Klenow and sonicated to a size of 100-600 bp. Illumina libraries were prepared using the NEBNext kit following the manufacturer's directions. 200-500 bp library fragments were size selected on 2% NuSieve agarose, gel purified and sequenced on the Illumina HiSeq 2000.

Although other studies used higher enzyme:DNA ratios, we kept the ratio lower to preserve RNA-primed DNA (Yang and Li 2013). Nonetheless, that there was only 21-96 ng. of enriched DNA at the end of the preparations (up to 2500-fold depletion of the starting DNA) and that the *MYC* origin was enriched >60-fold in each replicate indicates that the amount of λ-exo was sufficient.

Mapping and manipulating reads. Fasta files of human genome build hg19 were downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz; The Genome Sequencing Consortium 2001; Kent et al. 2002; Karolchik et al. 2004; Kent et al. 2010). A Bowtie 2 index was made with 'bowtie2-build –f hg19.fa hg19' (Langmead and Salzberg 2012). Illumina reads in fastq format were mapped to the hg19 Bowtie 2 index, using the parameters "--very-sensitive -N 1". The SAM format output of Bowtie 2 was piped into SAMtools (Li et al. 2009) to retain only reads that mapped to hg19 and converted to BAM format with "samtools view -F 4 -bS". "samtools sort" was used to sort the BAM files. In cases where BAM files of reads from replicates needed to be merged, "samtools merge" was used. See Table S1 for hg19 mapping statistics.

GC content in mappable reads. GC content in mappable reads was obtained with a custom Python script (https://github.com/JohnUrban/LexoNSseq2015) that collects this information from SAM files. Only mappable reads with 50 unambiguous bases were used (no N content) for calculating GC content of reads. Briefly, the Python script counted the number of G and C bases in each 50 bp read and reported how many reads had each GC count from 0-50 (i.e. a histogram of numberGC vs. numberReads). This histogram information was brought into R (R Core Team, 2013) where GC counts (0-50) were turned into percents (0-100) by 100*GCcount/50 (where 50 is the read length) and the number of reads with each GC count in a given dataset was normalized by the total number of reads summed over all GC counts in that dataset (i.e. percentGC vs. proportionOfReads): numberReadsWithGCcount/totalNumberReads. The normalized distributions of GC content in LexoG0 (Figure 2A) and NS-seq (Figure 2B) reads for each replicate were plotted in R as the log2(fold change) compared to the normalized distribution of GC content in G0 genomic DNA reads -- i.e. log2(NS-seq/gDNA) and log2(LexoG0/gDNA). This was also done for NS-seq reads relative to LexoG0 reads (Figure S2).

FRiT scores. For FRiT scores (fraction of reads in telomeres), mappable reads in BAM files were converted back to fastq files with SamToFastq.jar from Picard Tools (http://picard.sourceforge.net). The fastq files of mappable reads were re-mapped (using the same Bowtie 2 parameters as above) to a model telomere sequence composed of 1000 human telomere repeats (TTAGGG). The number of telomere-mappable reads was then divided by the total number of input mappable reads for normalization and multiplied by one million to get the number of hits per million reads (i.e. the FRiT score).

G4-CPMR and G4-Start-Site-CPMR. G4 motifs in hg19 mappable reads were identified and counted (for G4-CPMR where CPMR is counts per million reads) using a Python script modified from Dario Beraldi’s quadparser.py (http://bioinformatics-misc.googlecode.com/svn-history/r16/trunk/quadparser.py and https://github.com/JohnUrban/LexoNSseq2015) to analyze only the forward strand of reads in fastq files. For both G4-CPMR and G4-start-site-CPMR scores, we only considered the original read sequences (forward strands), not their reverse complements, as the read sequences represent 5’ ends of fragments that λ-exo may have encountered (whereas reverse complements represent 3’ ends of fragments). This is accomplished by searching only for ‘([gG]{3,}\w{1,7}){3,}[gG]{3,}' (the Python regular expression for $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$) and not for ‘([cC]{3,}\w{1,7}){3,}[cC]{3,}' ($C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}N_{1-7}C_{3+}$), which identifies G4s on the opposite strand. Hg19 mappable reads were converted back to fastq format with SamToFastq.jar from Picard Tools (http://picard.sourceforge.net) with the specification to return the original forward strand sequences for all reads. The number of G4 motifs in each fastq file of mappable reads was counted with the Python script (G4 counts), then divided by the total number of input mappable reads (for the given sample) to normalize and multiplied by one million to get the G4-CPMRs. The Python script was also used to keep track of which position each G4 motif started on in order to get G4 start site counts over each position of the reads (which represent the 5’ ends of fragments). To get G4-start-site-CPMRs, the start site count for each position was divided by the total number of input mappable reads to normalize and multiplied by one million. Note that when the G4-start-site-CPMR is summed up over all positions, it is equal to the G4-CPMR.

rDNA locus profiling. For profiling signals over the rDNA repeat, the fastq files of raw reads from the HiSeq2000 were mapped with Bowtie 2 (Langmead and Salzberg 2012) using the same parameters as above to a modified version of hg19, referred to here as hg19+rDNA, that contained a copy of the 43 kb rDNA repeat (http://www.ncbi.nlm.nih.gov/nuccore/555853?report=fasta) as an additional “chromosome”. Only mappable reads were retained in BAM format by piping the Bowtie 2 output into “samtools view -F 4 –bS -” (Li et al. 2009). Mapping reads to the rDNA repeat in the context of hg19 was done to ensure that reads that would map elsewhere in the genome with higher alignment scores were not forced to map to the rDNA, as was performed in a recent paper studying the chromatin landscape of the rDNA repeat (Zentner et al, 2011). Reads that mapped to the rDNA repeat with higher alignment scores than elsewhere in hg19 were extracted with SAMtools specifying '-q 2' and the name of the rDNA chromosome. Since the human genome contains >400 copies of the rDNA repeat, there was very high read depth coverage over each bp. To reduce possible spurious effects of PCR biases, "macs2 filterdup" was used with the 'auto' option on the extracted rDNA reads, which allowed the binomial distribution to determine how many reads can pileup at the same position on the same strand given the length of the repeat (~43 kb) and number of reads mapped to it with a p-value of 0.00001. The BED format results of macs2 filterdup were piped into BEDTools "sortBed" (Quinlan and Hall 2010) to sort and then into BEDTools bedToBam to convert back into BAM format. It is noteworthy that the rDNA fold enrichment results (in Figure 5B) were extremely robust and similar with and without any read filtering steps. The depth over each bp of the 43 kb rDNA repeat was obtained using BEDTools "genomeCoverageBed" with "-d" set and was normalized by the number of reads (in millions) that mapped to hg19+rDNA for the given sample to give the signal per million mapped reads (SPMR) over the rDNA locus for that sample. The depth files containing SPMR information were taken into R (R Core Team, 2013) and plotted. Fold enrichments were taken over each position and fold enrichment trends were obtained by loess smoothing the fold enrichment signal (span=0.05). G4 motifs were mapped strand-specifically across the rDNA locus using our customized quadparser Python script. The position information was taken into R, and strand-specific G4 counts were taken in 1 kb bins across the locus for visualization with the FE plots. For %GC signal across the rDNA repeat, BEDTools "makewindows" was used with "-w 5 -s 1" to create 5 bp sliding windows (incremented by 1 bp) across the rDNA locus. BEDTools "nucBed" was used to obtain the %GC in each window and this score was assigned to the middle bp in the 5 bp window. This raw %GC signal was brought into R and loess smoothed (span=0.05) before plotting with the fold enrichment signals.

Genome-wide Peak Calling. Genomic regions that were significantly enriched over a background control (called "peaks") were identified with MACS2 (Zhang et al. 2008). To avoid calling low complexity peaks (e.g. regions with only one or a few positions with numerous reads), before peak calling each replicate of mapped reads was further filtered for redundant reads that mapped to the same location on the same strand (potential PCR artifacts) by keeping only one read per position with 'macs2 filterdup'. Each replicate of mappable reads was filtered individually before pooling instead of filtering the pooled set to avoid eliminating reads that independently align to the same position in separate replicates, which should be treated as true positive alignments in the pooled set. For peak calling, 'macs2 callpeak' was used with '--nomodel', which turns off the ChIP-seq specific model builder, '--keep-dup all' since redundant read filtering was already performed as a pre-processing step, and '--extsize=350', which MACS2 uses as an estimate of the average Illumina library fragment size and for smoothing. Peak set names below are in treatment$_{\text{control}}$ format consistent with the nomenclature in the paper. LexoG0$_{\text{G0gDNA}}$ and NS$_{\text{G0gDNA}}$ peaks were called relative to the undigested G0gDNA reads to control for amplicons or deletions present in the MCF7 genome and to control for any biases introduced during library construction and sequencing. Thus, LexoG0 or NS-seq was set as the treatment (-t) and G0gDNA as the control (-c). NS$_{\text{LexoG0}}$ peaks were called with NS-seq set as the treatment (-t) and LexoG0 as the control (-c) to control for nascent strand independent biases of λ-exo, such as its %GC and G4 biases, in addition to any amplicons or deletions and biases introduced in the sequencing process. All peaks were called with '--downsample' to use equivalent numbers of reads between the treatment and control and to avoid the assumption of linearity introduced in downscaling. The local windows used to estimate local biases in the controls ('dynamic lambda') while scanning the genome for peaks were 5000 (--slocal) and

50000 (--llocal). These window sizes were chosen to cover the local region around a source of nascent strands (or G4-protected fragments), which we size selected up to 1500 bp, and to cover a region spanning the typical width of replication initiation zones. For all peak calling, we set a high stringency cutoff of $q < 0.001$ corresponding to a false discovery rate (FDR) of 0.1%. Since MCF7 is a female cell line, chrY data were excluded from all subsequent analyses. chrM (mitochondrial chromosome) was also removed from consideration. The output from MACS2 contains both the peak regions and peak summits (the bp of highest coverage inside a given peak region), which were each used for various analyses. It should be noted that the LexoG0 samples were designed first and foremost to characterize λ-exo biases in non-replicating cells (with undigested gDNA as the control). LexoG0 data were subsequently used to control these biases in NS-seq. It is possible that the LexoG0 control does not control for biases introduced by BND, if any BND biases exist and if they remain after the λ-exo digestion step. However, since the BND step only reduces the input DNA ~3-fold and the λ-exo-digestion step reduces the input up to 2500-fold, it is likely that BND biases are lost and overwritten by the strong λ-exo biases characterized in this paper. An alternative approach to LexoG0 could be to pass non-replicating DNA through BND before λ-exo digestion. However, BND enrichment of non-replicating DNA recovers a much smaller amount of DNA leading to larger relative enrichments of select sites in the genome specific to non-replicating gDNA. Since this would create BND biases not in proportion to those in replicating DNA, it raises additional BND issues rather than alleviating the potential BND bias issue and therefore this alternative is not necessarily an improvement upon LexoG0. LexoG0 side steps these issues by improving upon the standard undigested G0gDNA control, which only corrects for copy number and biases introduced in library construction and sequencing. In contrast, controlling with λ-exo-digested G0gDNA (LexoG0) corrects for nascent strand independent λ-exo biases while also controlling for copy number and biases introduced during library construction and sequencing.

Shuffling peaks/features and computing %GC of peak sequences. For analyses where peaks, peak summits, or other genomic features (e.g. G4 motifs) required shuffling throughout the genome, shuffleBed from BEDTools (Quinlan and Hall 2010) was employed with the constraints that the peaks stay on the chromosome they start out on (-chrom), do not overlap after the shuffle (-noOverlapping), and were not shuffled into hg19 gap regions nor onto chromosomes Y and M (-excl). The shuffled features were piped into sortBed to sort before being written to file. Hg19 gap locations were obtained from the UCSC Table Browser (Kent et al. 2002; Karolchik et al. 2004; Kent et al. 2010). The '.genome' file needed for this and some other BEDTools analyses was made with UCSC Kent Utilities Tool 'faSize -detailed' (http://hgdownload.cse.ucsc.edu/admin/exe/); Kent et al. 2002; Karolchik et al. 2004; Kent et al. 2010) on our copy of hg19.fa. For analyses interrogating the %GC in peaks and shuffled peaks (Figure 2C), "nucBed" from BEDTools was used to obtain the %GC information for each feature in a BED file and those results were brought into R for visualization. In R, a histogram of the %GC scores (which range from 0-100) for a given BED file was made with breaks=seq(0,100,0.5). The resulting bin counts were then loess smoothed (span=0.075) over the bin midpoints before plotting to lightly smooth out jagged edges.

Overlap analyses. For overlap analyses, 'intersectBed' from BEDTools was used (Quinlan and Hall 2010). To obtain the number of features in BED file A that overlapped features in BEDfile B, '-u' was set, file A was set to '-a', file B set to '-b', and the output was piped into 'wc – l'. Any feature in A that overlapped at least 1 feature in B by at least 1 bp was counted. To test for significance, we used a binomial model that conservatively estimates the upper-tailed p-value obtained if one did permutation tests into infinity. Briefly, the number of distinct positions that a feature from A can be shuffled onto in the genome is estimated as:

$$|\text{Total positions}| = G - C^*(\mu_A - 1)$$

Where G is the size of the mappable genome, which for hg19 is 2.835679040e9, C is the number of contiguous sequence components (i.e. regions separated by gaps, of which there are 257 in hg19 when considering only chr 1-22 and chrX), and μ_A is the mean interval size of features in file A. The number of distinct "successful positions" a feature in A can be shuffled to (where success indicates overlap with a feature in B), the probability of success, the probability of seeing x overlaps, and the upper tailed p value were estimated as:

|Successful positions| = **min(**(μ_A+μ_B-1)*|B|**,** |Total positions|**)**

Probability of success = **p** = |Successful positions| **/** |Total positions|

$$P(X = x) = \binom{|A|}{x} p^x * (1-p)^{n-x}$$

$$Pvalue = \sum_{x=|obs|}^{|A|} P(X = x)$$

Where μ_A and μ_B are the mean interval sizes of features in file A and B respectively, |A| and |B| are the number of features in file A and B respectively, and |obs| is the observed number of overlaps of features in A with features in B. The expected number of overlaps, |exp|, is obtained by p*|A|. This estimate of the number of successful positions results in a conservative p-value estimate because it assumes that all peaks in B are capable of forming disjoint sets of successful positions. In other words, it assumes that a successful position as determined by an arbitrary feature b_i is not also a successful position as determined by another arbitrary feature, b_k. Often this assumption is true. In cases when it is not true, the probability of success (p) and, therefore, |exp| are both overestimated, which is conservative with respect to |obs|.

Features and feature densities across genome. Feature density correlation analysis casts a wide net to see if the density of feature A in a genomic neighborhood is able to predict the density of feature B in that genomic neighborhood. When comparing feature set A to feature set B (eg. NS-seq peaks and CpGs), we chose to make bin sizes big enough such that feature counts in the bins for both A and B have a dynamic range and are not mostly zero counts. If one feature is numerous and the other is not, small bin sizes would result in mostly zeros for the rare feature and a range of counts for the other. This means the detectable correlation, if any, will be low at that level of resolution (bin size) due to the prevalence of zero counts for the rare feature. Using larger bin sizes allows both features to have a dynamic range of counts and allows the possibility to detect higher correlations if they exist, despite the lower resolution. Generally our analyses used 100 kb bins (for example, when comparing of NS-seq peak counts with G4 motif counts; Figure 3 B-D) as was used for many similar analyses in a previously published NS-seq paper (Besnard et al 2012), but it was more appropriate to use 1 Mb bins to explore correlations of peaks with CpG islands. There are relatively very few CpG islands compared to the size of the genome. When 100 kb bins are used, only ~40% of the bins have CpG islands in them and 60% have zero counts. In contrast, ~88% of the 1 Mb bins contain CpG islands and 100% contain NS-seq peaks, both with a dynamic range of counts. Thus, 1 Mb is an appropriate bin size for this particular analysis of peaks and CpG islands, despite the lower resolution. How close the CpG islands and NS-seq peaks (or other pairs of features) are to each other is the subject of other analyses such as direct overlap and proximity distributions.

To obtain feature (peaks, G4 motifs, CpG islands, etc) densities, defined as counts in 100 kb or 1 Mb bins, first BEDTools 'make windows' was used to partition hg19 into 100kb or 1Mb bins (Quinlan and Hall 2010). To eliminate noise from the analysis, the following bins were discarded: any bin smaller than the specified size, any bin that overlapped a gap, and any bin on chrM or chrY. To get the feature counts inside each retained bin, BEDTools "coverageBed" was used with the feature BED file as '-a', the genomic windows as '-b', and '-counts' set. Each resulting bedGraph file was sorted with sortBed to ensure that the counts in the same bins for different features were all in the same order and brought into R where they were subject to both Pearson and Spearman correlation tests (using cor()) and, in some cases, scatter plotted (peak sets vs G4 motifs in Figure 4D). For predicted G4 motif densities, G4 motifs were predicted with our Python implementation of quadparser (searching for G_3-$N_{1\text{-}7}$-G_3-$N_{1\text{-}7}$-G_3-$N_{1\text{-}7}$-G_3 and C_3-$N_{1\text{-}7}$-C_3-$N_{1\text{-}7}$-C_3-$N_{1\text{-}7}$-C_3 to predict G4s on both strands). We also downloaded the predicted G4 motifs from the Non-B DataBase (Cer et al, 2013, http://nonb.abcc.ncifcrf.gov/apps/Query-GFF/feature/) to compare to our set and found that it was identical. RefSeq genes and CpG island locations were downloaded from the UCSC Table Browser (Kent et al. 2002; Karolchik et al. 2004; Kent et al. 2010). All peaks, density signals, fold enrichment signals, and –log10(p) signals across the genome or genomic stretches were visualized in the Integrative Genomics Viewer (IGV; Robinson JT et al, 2011; Thorvaldsdóttir et al, 2012). For example, Figure 4 B-C shows G4 density and peak density in 100 kb bins across chromosomes 3 and 6, respectively.

Profiling G4s within 1 kb around peak summits. G4 positions were defined as the center position of each predicted G4 motif. The peak summits were identified by MACS2 (Zhang et al. 2008) as the bp of highest coverage inside each peak. "slopBed" from BEDTools (Quinlan and Hall 2010) was used to extend the peak summits equal lengths (e.g. 1kb or 2kb) in each direction. The slopBed output was piped into "intersectBed -wb -a G4centers.bed -b -". The '-wb' flag instructs BEDTools to return the pair of entries that overlapped. Here that means that both the G4 center that overlapped a windowed peak summit and the windowed peak summit that was overlapped are returned on the same line. The windowed peak summits in the paired-entry BEDTools output were then converted back to single bp summit positions such that the paired information contained a peak summit and a G4 center within the window size. The resulting file was loaded into R for further analysis. In R, the start sites of the G4 centers were subtracted from the start sites of their corresponding peak summits. This returns G4 center distances from the peak summit between -1*windowSize to windowSize, with 0 representing the peak summit position. When not considering what strand the G4 is on: if a G4 center start site is to the right of the peak summit, then subtraction results in a positive distance between 1 and windowSize; if the G4 center is to the left of the peak summit, then subtraction results in a negative distance between -1*windowSize and -1; if the G4 center start site is the same position as the peak summit start site, it returns 0. To incorporate information about which strand the G4 was on such that any G4 5' to the peak summit produces a negative distance and anything 3' to the peak summit produces a positive distance: distances for G4 motifs on the positive strand need no further correction, but the distances for G4 motifs on the negative strand need to be multiplied by -1. Thus, for G4 motifs that occur on the negative strand of the genome sequence: those that are to the right of the peak summit incur a negative distance; those to the left incur a positive distance; those that share the summit position remain as a distance of 0. The distances of G4 centers to peak summits were then counted and plotted. To test what G4 motif centers around peak summits would look like at random, the G4 motif locations were shuffled with shuffleBed (using parameters established above) and the same process was applied to the shuffled G4 motif centers. It was then possible to calculate the fold enrichment of the G4 counts near peak summits over the random distribution at each position (Test/Control1). The fold enrichment signal was loess smoothed to more clearly show the trend (span=0.1). As an additional control for the calculation of fold enrichment with the random distribution, G4 motif locations were shuffled with shuffleBed a second time for “control #2”. Both randomized controls were then used together to calculate the fold enrichment at random (Control2/Control1), which is centered around 1-fold so long as the procedure works correctly. The crest to crest distances of the wave-like G4 enrichment signal around peak summits, were calculated by first using a custom R script to objectively identify crests with a standardized definition. Specifically, using the loess smoothed G4 count data around the peak summits, we required a G4 enrichment crest to have a higher smoothed count than the counts of at least 55 bp to each side and for the crest count to have a fold enrichment >= 1.5 over the count in that position when shuffled at random. A range of other window size values gives the same results for NS_{G0gDNA} and NS_{LexoG0}. With crest positions identified, distances between crests could then be calculated. All plotting was done in R.

Prominence, CTR, and decomposition of the G4 enrichment signal around NS_{G0gDNA}.

Trough positions in the G4 enrichment signal around summits were identified in each G4 enrichment signal in a similar fashion to how crests were identified (above). Troughs for G4 enrichment signal around all NS_{LexoG0} summits, all NS_{G0gDNA} summits, and the subset of NS_{G0gDNA} summits that overlapped NS_{LexoG0} summit windows were identified by requiring that the smoothed count in a trough position be lower than the counts of at least 55 bp to each side, but lower than $\geq$ 144 surrounding positions total, and have a fold enrichment of $\leq$ 3. The G4 Fold Enrichment scores over crest and trough positions were collected and the means for each ($crest_{mean}$, $trough_{mean}$) were computed. Prominence of the crests, qualitatively defined as the amount that crests jut out above troughs was quantified by: $crest_{mean}$ - $trough_{mean}$. The phasing of the G4 enrichment at the crests, qualitatively defined as how concentrated the signal is at crests (relative to troughs) was quantified as the crest-to-trough ratio (CTR): $crest_{mean}/trough_{mean}$. The NS_{G0gDNA} summits were partitioned into two subsets: one subset containing summits that overlapped (were inside of) NS_{LexoG0} summit windows (summit +/- 1 kb) and the other subset containing summits that did not overlap NS_{LexoG0} summit windows. These subsets are described as NS_{G0gDNA} summits represented in NS_{LexoG0} and NS_{G0gDNA} summits not represented in NS_{LexoG0}, respectively. The subsets were then treated individually as described in the section titled, “Profiling G4s within 1 kb around peak summits”. Decomposition of the G4 enrichment signal around NS_{G0gDNA} summits into a stronger wave-like component (for those represented in

NS_{LexoG0}) and a roughly uniform component (for those not represented in NS_{LexoG0}) was the result of analyzing the partition this way. All plotting was done in R.

Profiling nuclesome signal around peak summits. Since the spacing of the crests of the waves of G4 enrichment around our peak summits was suggestive of nucleosome spacing, we also looked at the nuclesome signal around those summits for which G4s were nearby (within 1 kb to either side, see Table S7). The available nucleosome data at UCSC (Kent et al. 2002; Karolchik et al. 2004; Kent et al. 2010) was downloaded (K562 and GM12878 cells; Kundaje et al, 2012). Peak summits were extended 1 kb to each side to produce 2001 bp summit windows (same as for the G4 analysis above). For each summit window, the nucleosome signal over each individual bp of the 2001 bp was obtained. Then the mean over each individual relative position (-1000 to 1000) around all the summits was calculated. Genome positions for which nucleosome signal was not available, represented as ".", were treated as missing data. In other words, means over each position were calculated only from the sum of available scores divided by the number of available scores (in contrast to treating all missing data as 0, which is an invalid assumption). The same procedure was done after shuffling the peaks (shuffleBed) to obtain the genome-wide mean nucleosome scores over each position expected at random. In R, for each raw cell line signal (see Table S8), the ratio of the two means (the mean score for the test sample, μ_{test}, and the mean score for the shuffled sample, $\mu_{shuffle}$) at each position, j, in the 2001 bp window was plotted ($\mu_{test,j}/\mu_{shuffle,j}$). The cell line signals were lightly loess smoothed (span=0.075) for plotting (colored lines in Figures 6 and S4; smoothed cell line nucleosome signal in Table S8). For the mean signal between the 2 cell lines, the mean between the two raw cell line signals at each position was taken, $((\mu_{test,j,K562}/\mu_{shuffle,j,K562}) + (\mu_{test,j,GM12878}/\mu_{shuffle,j,GM12878}))/2$, before light loess smoothing (span=0.075) (black lines in Figures 6 and S4). The crests in the wave-like mean nucleosome signal between the two cell lines were identified similar to how crests were identified in the G4 signal around summits. Specifically, we required crest positions of the mean nucleosome signal between cell lines to have higher scores than $\geq$ 50 bp to each side, but at least higher than 130 surrounding positions total, and to have minimum height difference (or greater) between the potential crest position and the lowest point within the left or right window in order to ignore positions that are arbitrarily higher than surrounding area. The subset of NS$_{\mathrm{G0gDNA}}$ summits with $\geq$ 1 G4 within 1 kb, were further partitioned, as in the G4 analysis above, to two subsets with one containing all NS$_{\mathrm{G0gDNA}}$ summits that are represented in N$_{\mathrm{SLexoG0}}$ and the other containing all NS$_{\mathrm{G0gDNA}}$ summits <u>not</u> represented in NS$_{\mathrm{LexoG0}}$. The raw and smoothed cell line nucleosome signals as well as the raw and smoothed mean nucleosome signal between cell lines around these two subsets of NS$_{\mathrm{G0gDNA}}$ summits were computed as described above. The decomposition of the nucleosome signal around NS$_{\mathrm{G0gDNA}}$ summits into a stronger wave-like component resembling the nucleosome signal around NS$_{\mathrm{LexoG0}}$ summits and a less wave-like component resembling the nucleosome signal around LexoG0$_{\mathrm{G0gDNA}}$ summits was the result of this partitioning process. All plotting, correlations, and “divergence” calculations were done in R.

SUPPLEMENTARY REFERENCES

Brooks TA, Hurley LH. 2010. Targeting MYC expression through G-quadruplexes. *Genes Cancer* **1**: 641-649.

Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GH, Volfovsky N, Yi M, Luke BT, et al. 2013. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* **41:** 94-100.

Coffman FD, Georgoff I, Fresa KL, Sylvester J, Gonzalez I, Cohen S. 1993. In vitro replication of plasmids containing human ribosomal gene sequences: origin localization and dependence on an aprotinin-binding cytosolic protein. *Exp Cell Res* **209**: 123–32.

Coffman FD, He M, Diaz M-L, Cohen S. 2005. DNA replication initiates at different sites in early and late S phase within human ribosomal RNA genes. *Cell Cycle* **4**: 1223–6.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**: 3423-24.

Dimitrova DS. 2011. DNA replication initiation patterns and spatial dynamics of the human ribosomal RNA gene loci. *J Cell Sci* **124**: 2743–52.

Gencheva M, Anachkova B, Russev G. 1996. Mapping the sites of initiation of DNA replication in rat and human rRNA genes. *J Biol Chem* **271**: 2608–14.

Gerbi SA, Bielinsky AK. 1997. Replication initiation point mapping. *Methods* **13**: 271–28.

Kankia BI, Marky LA. 2001. Folding of the thrombin aptamer into a G-quadruplex with Sr(2+): stability, heat, and hydration. *J Am Chem Soc* **123:** 10799-10804.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32:** D493-6.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996-1006.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed data sets. *Bioinformatics* **26**: 2204-2207.

Kikin O, D'Antonio L, Bagga PS. 2006. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequence. *Nucleic Acids Res* **34** (Web Server issue): W676-W682.

Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A. 2012. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* **22**: 1735-47.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9:** 357-9.

Lebofsky R, Bensimon A. 2005. DNA replication origin plasticity and perturbed fork progression in human inverted repeats. *Mol Cell Biol* **25**: 6789–97.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Little RD, Platt TH, Schildkraut CL. 1993. Initiation and termination of DNA replication in human rRNA genes. *Mol Cell Biol* **13**: 6600–13.

Liu G, Malott M and Leffak M (2003) Multiple functional elements comprise a mammalian chromosomal replicator. *Mol Cell Biol* **23**: 1832-1842.

Malott M, Leffak M. 1999. Activity of the c-myc replicator at an ectopic chromosomal location. *Mol Cell Biol* **19**: 5685-5695. Erratum in: *Mol Cell Biol* **19**: 8694 (1999).

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-82.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Robinson, JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov, JP. 2011. Integrative Genomics Viewer. *Nature Biotechnology* **29:** 24–26

Scott RS, Truong KY, Vos JM. 1997. Replication initiation and elongation fork rates within a differentially expressed human multicopy locus in early S phase. *Nucleic Acids Res* **25**: 4505–12.

Shim JW, Tan Q. Gu LQ. 2009. Single-molecule detection of folding and unfolding of the G-quadruplex aptamer in a nanopore nanocavity. *Nucleic Acids Res* **37**: 972-982.

Tao L, Dong Z, Leffak M, Zannis-Hadjopoulos M, Price G. 2000 Major DNA replication initiation sites in the c-myc locus in human cells. *J Cell Biochem* **78**: 442-457.

The Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. 2001. *Nature*. **409** (6822): 860-921.

Thorvaldsdóttir H, Robinson JT and Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. **14**: 178-192.

Yang W, Li X 2013. Next-generation sequencing of Okazaki fragments extracted from *Saccharomyces cerevisiae*. *FEBS Lett* **587**: 2441-2447.

Yoon Y, Sanchez JA, Brun C, Huberman JA. 1995. Mapping of replication initiation sites in human ribosomal DNA by nascent-strand abundance analysis. *Mol Cell Biol* **15**: 2482–9.

Zentner GE, Saiakhova A, Manaenkov P, Adams MD, Scacheri PC. 2011. Integrative genomic analysis of human ribosomal DNA. *Nucleic Acids Res*. **39**: 4949-60.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.