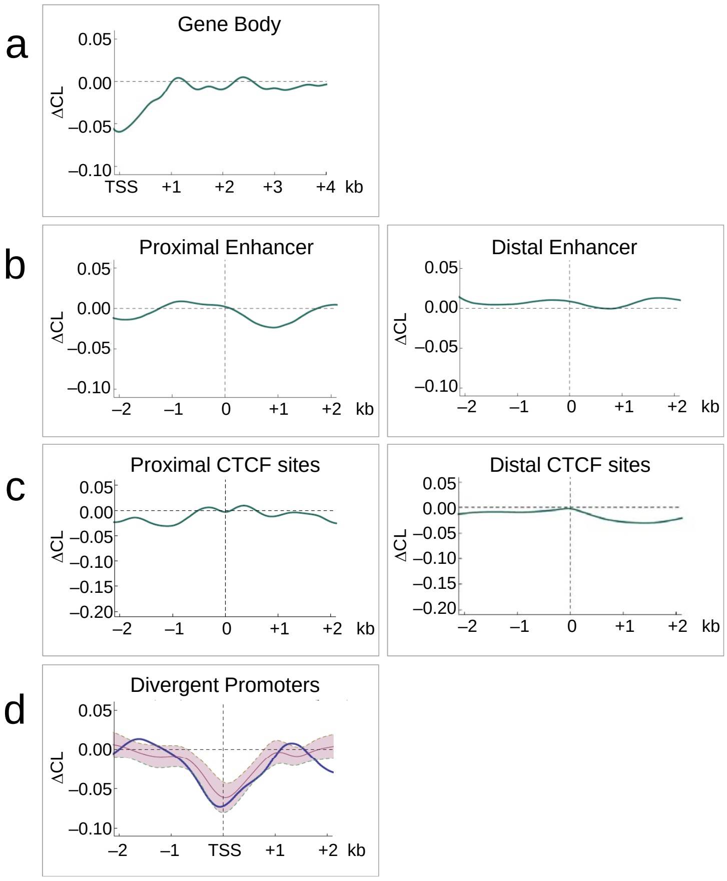
Supplementary Information

for: Transcription dependent dynamic supercoiling is a short range genomic force

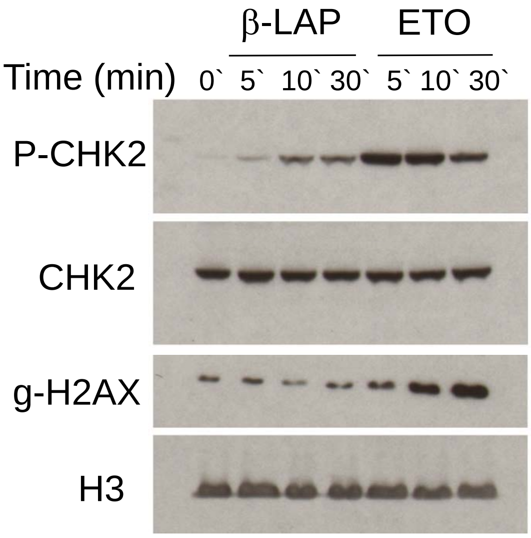
Fedor Kouzine, Ashutosh Gupta. Laura Branello, Damian Wojtowicz, Khadija Ben-Aissa, Juhong Liu, Teresa M. Przytycka and David Levens

- Supplementary Figure 1: ΔCL profiles at different regions of the genome.
- Supplementary Figure 2: Kinetics of γ-H2AX formation and CHK2 phosphorylation (P-CHK2) following β-Lapachone (β-LAP) or Etoposide (ETO) treatments for the indicated times.
- Supplementary Figure 3: ΔCL profiles in a 4 kb region centered on TSSs in presence or absence of campthothecin (CPT) or β-LAP
- Supplementary Table 1: List of transcribed Regions
- Supplementary Table 2: List of all detection primers used for qPCR
- Supplementary Note: Extracting supercoiling signals from noisy genomic data
  Ashutosh Gupta and David Levens

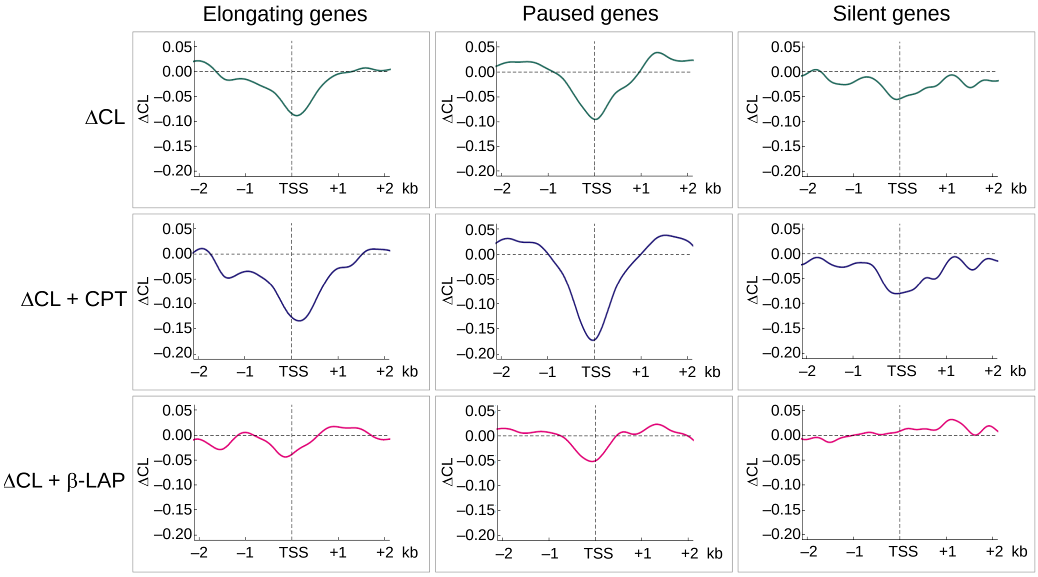
Supplementary Figures



Supplementary Figure 1: ΔCL profiles at different regions of the genome. (a) The average ΔCL profile for all genes starting from the TSS to 4 kb into the gene body (Suppl. Note 3.5). (b) The average ΔCL profile for all enhancers that are in or within ±3,000 bp of the gene body (proximal) or are more than ±3,000 bp away from the gene body (distal). Only enhancers with significant Pol II signal were considered (Suppl. Note 3.6). (c) The average ΔCL profile for all CTCF sites that are in or within ±3,000 bp of the gene body (proximal) or are more than ±3,000 bp away from the gene body (distal) (Suppl. Note 3.6). (d) The average ΔCL profile about the TSS for all divergent promoters that are separated by a minimum of 100 bp or a maximum of 4,000 bp. The shaded range shows the $\mu \pm \sigma$ region obtained by averaging 30 randomizations of an equivalent number of genes excluding divergent promoters (Suppl. Note 3.7).



Supplementary Figure 2: Kinetics of γ-H2AX formation and CHK2 phosphorilation (P-CHK2) following β-LAP or Etoposide (ETO) treatments for the indicated times. Equal loading is shown by histone H3 and total CHK2 detection.



Supplementary Figure 3: ΔCL profiles in a 4 kb region centered on TSSs in presence or absence of CPT or β-LAP. According to the pausing index (Suppl. Note 3.4) genes were grouped in 3 categories: elongating (left panel), paused (central panel) and silent (right panel).

Supplementary Tables

Table 1: List of transcribed regions

#	chr	Start Site	End Site	Accession #
1	19	59107802	59137444	59284
2	19	59107882	59138080	59284
3	20	33573306	33574658	343705
4	6	74135002	74136236	441161
5	20	33578212	33580862	140873
6	6	74129120	74130615	154288
7	6	74119507	74120674	340168
8	15	41772375	41778712	548596
9	X	153152125	153176632	1527
10	X	153101360	153114725	2652
11	15	41673536	41678892	548596
12	X	153062933	153077705	5956
13	20	33484240	33486662	554250
14	20	33484562	33489441	8200
15	X	153533444	153535036	30848
16	19	59927813	60070473	AF285439
17	X	153466704	153468263	653387
18	1	149603404	149611805	8991
19	20	33336947	33343639	128876
20	20	33652037	33656662	80307
21	1	149779404	149822683	7286
22	20	33609920	33651008	80307
23	11	5558682	5559690	340980
24	6	74161191	74183791	55510
25	6	73975313	74029640	80759
26	X	153556723	153632526	139716
27	X	153499058	153500716	246100
28	X	153717262	153904192	2157
29	11	4799191	4800220	119694
30	19	59739981	59748862	90011
31	11	4901179	4902145	79324
32	X	153177344	153211894	8277
33	X	153660161	153686957	4354
34	11	4826042	4827014	119692

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
35	11	5230996	5232587	3048
36	11	5036455	5037433	119678
37	6	74191313	74218720	115004
38	11	5109497	5110448	390054
39	11	5329313	5330252	390058
40	11	5024331	5025267	119679
41	11	5400006	5400960	390061
42	11	5177540	5178506	283111
43	19	59265068	59269173	441864
44	11	4923964	4924906	401666
45	11	4932577	4933519	401667
46	11	5522366	5523329	390067
47	22	31080871	31087147	10738
48	11	5492198	5494501	143630
49	22	31085892	31097063	10737
50	22	30875518	30885243	150297
51	11	4746784	4747723	256892
52	19	59187353	59207732	59285
53	19	59077278	59102713	5582
54	6	108593954	108616706	7101
55	9	131123115	131127005	414318
56	11	4859624	4860689	401665
57	6	41411504	41426593	9436
58	22	30769258	30836645	6523
59	15	41597132	41611110	4130
60	11	4965999	4970235	56547
61	22	30916425	30930718	10739
62	22	30944462	30981318	6527
63	X	152780580	152794505	3897
64	22	31526801	31589028	7078
65	20	33506563	33563216	11190
66	22	31238539	31732683	8224
67	22	31239399	31784329	8224
68	6	41829976	41834895	647014
69	5	131315195	131375214	23305
70	5	131424245	131426795	3562
71	5	131170738	131357870	23305
72	22	31140289	31183373	254240
				Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
73	5	131317500	131375553	23305
74	11	4892524	4893469	81282
75	11	4976788	4977736	119682
76	5	141953305	142045812	2246
77	X	153452672	153453380	286967
78	22	31087313	31107216	646618
79	5	131556201	131590834	8974
80	X	152891434	152901834	8269
81	11	4885175	4886114	119687
82	11	5466512	5467469	390066
83	22	31039083	31041792	646599
84	11	5431294	5432233	390064
85	X	152853916	152863426	5973
86	5	131905034	131907113	3567
87	11	131033416	131038060	399980
88	6	132309645	132314155	1490
89	11	5367204	5368185	390059
90	11	4781238	4782423	119695
91	13	112808105	112822346	2155
92	19	59158105	59177951	59283
93	X	153908257	153938385	65991
94	2	234316134	234317400	414061
95	21	32706622	32809568	59271
96	11	130745778	131710752	50863
97	21	32866419	32870062	55264
98	11	5203270	5204877	3043
99	11	5246158	5483410	3046
100	5	131466369	131511544	645029
101	11	5129236	5130175	23538
102	11	5714253	5716328	387748
103	21	33084854	33107868	56245
104	20	33720024	33750688	9054
105	5	132225179	132228124	2661
106	22	30845512	30846923	646580
107	11	5573934	5590217	117854
108	11	116196627	116199221	337
109	11	5210634	5212434	3045
110	9	130978873	131012683	389792

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
111	6	41714230	41729959	4188
112	11	5098469	5099384	390053
113	7	27134534	27136924	3201
114	1	149851285	149938183	81609
115	1	149851164	149933599	81609
116	X	152799320	152807619	643736
117	21	32870732	32879687	140290
118	18	59455481	59462470	6318
119	21	33066281	33093160	54067
120	11	5485107	5487744	50613
121	19	59705824	59713709	3904
122	21	33779706	33785650	54943
123	5	131612501	131658907	BC030525
124	19	59510164	59516221	353514
125	2	234209886	234343242	54575
126	18	59455932	59479430	AF428135
127	5	132111041	132118263	645121
128	1	149750499	149777792	57530
129	13	112349358	112386812	400165
130	9	130896893	130912904	1384
131	7	127020924	127029079	29999
132	5	131658043	131707798	6583
133	18	59473411	59480098	6317
134	11	5641363	5662869	85363
135	X	153943079	153952830	4515
136	2	234333657	234346684	54658
137	7	27106497	27108919	3199
138	X	153254304	153256200	AK125630
139	21	39699654	39739529	150082
140	11	64079673	64095575	55867
141	7	27151640	27153893	3203
142	7	27191681	27198951	646692
143	11	63934128	63944265	644541
144	6	41812427	41823099	5225
145	5	131621285	131637046	8572
146	7	27160814	27162821	3204
147	21	39739666	39809303	6450
148	X	122923269	123064027	10735

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
149	9	131138623	131140395	AK092192
150	X	153952903	154004543	79184
151	11	5574461	5622204	445372
152	21	32922943	33022148	8867
153	21	33782367	33785893	54943
154	2	234490781	234592905	79054
155	11	2118322	2126470	51214
156	5	56240856	56248767	133383
157	13	112670814	112800864	23263
158	2	234624084	234650515	6694
159	X	122922235	123063026	10735
160	7	125865894	126670548	2918
161	7	115952074	115988466	857
162	21	32869022	32870472	55264
163	7	27187653	27191355	3207
164	18	59528407	59541613	89778
165	7	27168581	27171674	3205
166	11	63973121	63975702	439914
167	7	117137940	117300797	83992
168	21	33936653	34183479	6453
169	7	116790511	116854779	136991
170	18	23784932	24011189	1000
171	7	116704517	116750579	7472
172	21	33320108	33323370	10215
173	22	30659507	30671336	25775
174	15	41652602	41769512	9677
175	18	59593623	59623592	8710
176	11	116165295	116167794	116519
177	7	113842511	114117391	93986
178	7	113842287	114117218	93986
179	7	27147520	27149812	3202
180	6	108722790	108950951	246269
181	21	34243099	34258130	400863
182	11	2273445	2279866	29125
183	7	116907252	117095951	1080
184	11	2106925	2109541	492304
185	7	89712444	89777638	79846
186	7	115926679	115935831	858

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
187	7	89678935	89704865	261729
188	7	89678993	89704927	261729
189	2	220016639	220039828	10290
190	7	89621624	89632077	26872
191	18	59705921	59722100	5055
192	19	59289813	59297806	126014
193	13	29674766	29779163	84056
194	2	219991342	219999705	1674
195	11	64114857	64126396	116085
196	7	27112333	27125739	3200
197	13	29680608	29779584	84056
198	2	220087135	220106998	55515
199	2	234351720	234406802	339766
200	21	33883516	33935936	9946
201	15	41906499	41946502	79968
202	11	2109739	2116400	AK074614
203	7	27121491	27129028	AK056230
204	5	132114415	132140966	23176
205	12	38905085	39051870	120892
206	19	59289744	59295960	126014
207	13	112825145	112851842	2159
208	7	27203023	27206221	3209
209	18	59733724	59753456	5273
210	2	220087295	220111738	55515
211	11	116205833	116208997	345
212	11	64130221	64247236	9379
213	6	41845891	41855608	10817
214	11	2110355	2116780	3481
215	X	122821728	122875503	331
216	7	27099136	27102119	3198
217	7	114349444	114446492	29969
218	11	1953071	1956250	AK126915
219	11	1972983	1975280	283120
220	14	98705376	98807575	64919
221	21	32687312	32688133	84996
222	10	55236344	55248144	387683
223	5	131733342	131759205	6584
224	2	220123699	220145134	23363

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
225	7	90729032	90731910	645794
226	11	116211678	116213548	335
227	20	33667222	33672379	6676
228	7	27176734	27180448	3206
229	X	153282772	153293621	1774
230	11	2279818	2296006	10077
231	16	48057	62591	64285
232	21	39479273	39607426	54014
233	21	39674139	39691496	7485
234	6	108469305	108502634	28962
235	11	2137586	2139015	3630
236	11	1817478	1819484	7136
237	19	59777070	59790833	11027
238	22	30650902	30652995	AK123899
239	5	132059221	132101163	11127
240	X	153365249	153368126	8266
241	X	153368841	153372189	8273
242	7	116099694	116225676	4233
243	21	33364442	33366596	116448
244	5	131437383	131439758	1437
245	2	220200526	220214936	6508
246	12	38629561	38786156	114134
247	19	59557944	59568280	3903
248	15	41815433	41825789	440278
249	11	64270605	64284763	5837
250	11	2141734	2149611	7054
251	5	132185910	132189901	134549
252	15	41612966	41669697	9677
253	11	1897511	1916512	7140
254	19	59064592	59071501	91663
255	16	142853	144504	3050
256	11	64348496	64368617	55561
257	19	59064504	59069685	91663
258	16	258310	265915	8786
259	21	33619083	33653999	3454
260	7	27248945	27252717	2128
261	2	220044627	220047344	AK098307
262	19	59668021	59676234	148170

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
263	16	265611	277210	64714
264	11	1808892	1815326	90019
265	19	59796924	59804352	11024
266	5	56146021	56227730	4214
267	X	153293070	153303259	6901
268	21	33872080	33882884	29980
269	16	155972	156767	445449
270	2	220145197	220148671	3623
271	5	56251187	56283697	166968
272	7	90731718	90736068	8321
273	2	234438819	234441829	151507
274	16	261827	265981	8786
275	19	59618416	59639892	57348
276	9	130883226	130892538	57171
277	7	116447578	116657391	7982
278	6	74007762	74076659	CR936715
279	16	162874	163708	3040
280	7	90063746	90674880	5218
281	5	132177177	132180377	134548
282	7	89870731	89883204	9069
283	11	2246303	2248758	430
284	9	130747629	130749833	22845
285	5	142130475	142586243	23092
286	5	142130155	142582945	23092
287	7	89813956	89858258	85865
288	16	166678	167520	3039
289	19	59355657	59368664	147798
290	19	59355714	59368756	147798
291	19	59434640	59452868	79168
292	7	90176647	90677840	5218
293	7	116380616	116657313	7982
294	5	131920528	132007498	10111
295	19	59446172	59452939	10990
296	19	59412608	59438414	11025
297	21	33726662	33774120	757
298	1	149641823	149698556	23126
299	X	153359307	153360790	8270
300	X	153387699	153397567	60343

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
301	11	116124098	116148914	84811
302	2	118393639	118491788	54520
303	13	112392643	112589470	23250
304	15	41884024	41904243	4236
305	X	152940457	153016323	4204
306	9	130913064	130951044	5524
307	2	234410765	234427885	55355
308	21	39469253	39477310	8624
309	22	30480068	30633001	9681
310	1	149521414	149531005	57592
311	9	130978768	130980347	389792
312	7	116289798	116346549	830
313	19	59386005	59389333	79042
314	9	130839073	130874172	84895
315	7	127007917	127012890	79571
316	2	118389618	118390940	54520
317	9	130749797	130809195	23511
318	19	59491665	59496050	11026
319	2	220116988	220123561	130612
320	22	31110991	31113822	646621
321	21	33028083	33066040	94104
322	19	59664787	59666706	94059
323	11	64250958	64269504	10235
324	21	33021613	33022627	644266
325	21	39607756	39608756	257357
326	X	153412799	153428663	2539
327	21	33560541	33591390	3588
328	21	32895965	32906784	56683
329	12	38904567	38905165	642606
330	16	372247	382955	645631
331	X	153310214	153318055	537
332	22	30670478	30683590	7533
333	16	277440	342465	8312
334	19	59351188	59355258	79165
335	20	33593191	33608819	51614
336	22	31201223	31224818	25793
337	16	170334	171178	3049
338	21	33798138	33836286	2618

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
339	20	33750740	33752294	140823
340	16	43016	47444	79622
341	6	108639409	108689156	8724
342	X	152823563	152825834	554
343	X	153339816	153355179	55558
344	5	132021763	132024700	3596
345	16	224801	258971	83986
346	22	31113568	31138235	51493
347	16	67017	75845	4350
348	22	30222260	30344534	9814
349	6	41856466	41865609	29964
350	5	132235912	132238286	116842
351	16	23876	26382	645582
352	1	149531036	149565348	5298
353	1	149437652	149488630	8394
354	1	149493820	149506560	5710
355	20	33330138	33336008	3692
356	1	149531231	149566511	5298
357	11	1730560	1741798	1509
358	19	59368920	59385478	79143
359	11	64313184	64327289	5871
360	16	415668	512482	9727
361	16	361859	371908	58986
362	X	153429255	153446455	8517
363	22	30165350	30215810	56478
364	7	126797588	126820003	168850
365	21	34197626	34210028	539
366	5	132037271	132046267	3565
367	X	122821265	122822820	643547
368	16	357396	360541	10573
369	16	387773	402487	26063
370	22	30402241	30438731	253143
371	16	356981	360226	10573
372	19	59866263	59873622	11006
373	2	220111921	220116682	79586
374	6	108298214	108386086	11231
375	19	59536503	59542233	23547
376	11	64418594	64441239	23130

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
377	9	130810133	130830400	56904
378	19	59412548	59418709	11025
379	2	118310049	118312244	389024
380	2	118288724	118306423	8886
381	11	1925113	1934408	6150
382	7	115637816	115686073	26136
383	6	41622141	41678100	116113
384	X	153325695	153334051	9130
385	X	153260980	153263075	2010
386	7	127015694	127018989	381
387	2	220071856	220079955	29926
388	11	116154485	116163949	8882
389	21	39420865	39421836	391282
390	22	30411027	30439831	253143
391	21	39636110	39642917	3150
392	22	30160554	30164552	AK127132
393	19	59470017	59476753	10288
394	20	33754944	33793607	9584
395	X	152848570	152853662	8260
396	19	59314702	59320534	AK128544
397	20	33677379	33705245	8904
398	15	41825881	41852096	2923
399	16	178970	219450	55692
400	2	220170839	220189418	114790
401	20	33700290	33716252	10137
402	11	64302486	64303493	644613
403	5	131774571	131825958	441108
404	21	33837219	33871682	6651
405	16	387192	390755	4833
406	16	36999	43625	51728
407	15	41874456	41879547	619189
408	11	64327563	64334764	4221
409	21	33524100	33558697	3455
410	11	1830883	1870068	4046
411	19	59333260	59351239	4849
412	8	128875987	129182678	5820
413	X	153644343	153659154	1736
414	X	152929150	152938536	3654

Continued on next page

Table 1 – continued from previous page

#	chr	Start Site	End Site	Accession #
415	15	41852089	41856794	80237
416	15	41879912	41882079	25764
417	18	59767573	59778624	284293
418	22	30345379	30388195	23761
419	7	27029881	27031053	402643
420	21	33697071	33731696	3460
421	16	4081	5847	375260
422	X	153318714	153325008	2664
423	X	152826026	152844908	393
424	15	41871843	41881362	25764
425	X	152866201	152883371	3054
426	6	41759693	41810776	7942
427	8	128816861	128821905	M13930
428	22	30344476	30356810	23761
429	16	25950526	25951759	647915
430	11	64376783	64402767	10938
431	5	132230255	132231276	27089
432	X	153230158	153256123	2316
433	18	59788242	59807588	5271
434	19	59297971	59302080	4696
435	1	149638664	149641036	5692
436	8	128817497	128822856	4609
437	X	153279911	153283874	6134
438	1	149579739	149586393	5993
439	11	5667630	5688668	10346
440	19	59652207	59665006	114823
441	11	64288653	64302817	7536
442	19	59396537	59403327	6203
443	5	131846678	131854333	3659
444	22	30765440	30765968	402057
445	6	74283958	74287475	1915

Table 2: List of all detection primers used for qPCR

Name	Gene	Forward Primer 5'>3'	Reverse Primer 5'>3'	Level Of Expression
A	CKMT1B	ATCCTCGCATCTTCACTTGG	ATGAGGCACGACTGGAAAAG	0-20%
B	CKMT1A	GCATTCATTCTCCTTGCTACC	GAGAGTAAAGGCGAGTGGTGTA	0-20%
C	GTAG2	CTGGGTTCGGCAGTATCAGT	CCTTTCCTGTGGATCTGACC	0-20%
D	TUFT1	TAAGGCAATGTGTCCCGC	GAAAGGCAGGCACCAAGG	0-20%
E	GDF5	GGATGGTCTCGATCTCCTGA	CATCATGTGGGAAATTGTGC	0-20%
F	ASCL2	CTCTGAGACCTCAGGGAACG	AGGCTGGCAGTAAACACTGG	60-80%
G	PFTK1	CAAAATAAGGCACCCTACATCTG	GAGTCCAGTTGTTTGAGCGG	60-80%
H	ARHGAP26	TGGCACAGTCTCAGCTCACT	CAGAGCGAGACTCCGTCTC	60-80%
I	MIER3	AGGAATGGGAGATGGAGACC	TTCTCTGCCCTGTCGATCTT	60-80%
J	HISPPD2A	CTTGATGCTCCCTTCCTTTG	GCACAAACTCTGCCTCTTCC	60-80%
K	PISD	CACATCTGTGGGAGCAACTG	CCGCTGGAATTGTATCCTGT	80-100%
L	IRF1	GGGAGGGTTTCAGTCCTAGC	CCATCACAGCAAACCATCAA	80-100%
M	UQCRQ	GCTGAGGAGAAGTGTGAGC	GGATGACGCCTTTGTCC	80-100%
N	MYC	GGACTCAGTCTGGGTGGAAGG	AAGGAGGAAAACGATGCCTAGA	80-100%
O	EEF1A1	CCTGCGAGTGTGTGTGTG	GCAAGTGTTGGGGTTAGGAA	80-100%
P	Intergenic	GCAGTTCAACCTACAAGCCAATAGAC	CACAAATTAGCGCATTGCCTGA	NA

Supplementary Note

For: Transcription dependent dynamic supercoiling is a short range genomic force
Fedor Kouzine, Ashutosh Gupta, Laura Branello, Damian Wojtowicz, Khadija Ben-Aissa, Juhong Liu, Teresa M. Przytycka and David Levens

Extracting supercoiling signals from noisy genomic data

Ashutosh Gupta and David Levens

1 Discussions

1.1 Reproducibility

Microarrays have been routinely used for the ChIP-chip experiments, where the enrichment of bound sequences is often 10–100 fold higher than the background. However, for the current series of experiments, namely psoralen intercalation, this is not the case. The maximum observed relative enrichment of psoralen photobinding under phisiological conditions is approximately two folds [1], as the free energy of intercalation of psoralen in negatively supercoiled DNA is much smaller than the corresponding binding energies of typical antibodies. There is also a finite, although smaller, free energy of intercalation in relaxed DNA. Psoralen binding sites are not focal, but are continuously distributed across the genome. As a result the unprocessed data have a very low signal-to-noise ratio (SNR)[1], and conventional methods and standards for mapping molecules bound to DNA are inadequate without modification.

Here we present a method developed to study such low energy / low specificity effects. This method is capable of extracting signal from low SNR data (as low as less than 15^{-2}), it is unsupervised and has been calibrated.[2] The underlying assumption is that the noise is of much higher frequency than the real signal and its uncorrelated to the real signal (which in this case is psoralen-binding[3]).

[1] See Suppl. Note 2.1 for definition.

[2] See Suppl. Note 1.2 for calibration details.

[3] Because we don't expect psoralen intercalation (and level of supercoiling) to change abruptly from one base-pair to next, while the microarray data does show high variation.

As an example, we define a hypothetical (low frequency) function and overlay increasing levels of white noise[4] (6 replicates).

The function was designed so that it has a low frequency signal (based on what we observed from our datasets) and distinctive features of different amplitudes (various peaks and valleys of different amplitudes). For this simulation, the chosen noise levels were in a range that was much wider than than the observed noise level from the experiment (see Suppl. Note 3.1 for more).

The noisy data is then smoothed using Fourier Convolution Smoothing [2], and plotted in figure on page 3 along with the raw data, and the original function.

We observe that as the noise level increases, the 6 replicates look increasingly different although they are all derived from the same starting function modified by same level of noise. This suggests that when noise levels are high, we cannot ask for reproducibility 'from individual experiments'.[5]

To achieve reproducibility/reliability we need to repeat the experiment several times.[6] The number of replicates required depends on the level of noise. If the noise levels are low one or two more experiments suffice. For higher noise levels, higher numbers of replicates are needed.

Lets say that we start with four replicates. These can be subdivided into four subsets of three replicates (by dropping one of them). Now if the average profiles of each subset are similar, then there are enough replicates to make a reliable inference from the data. If the averages are not comparable, that means more replicates are required. And so on.

This is the prescription for a generic case where the actual behavior is not known. For the simulation under discussion we have a direct benchmark for comparison, i.e. the original function which was corrupted with different levels of noise. The law of large numbers guarantees an accurate result.

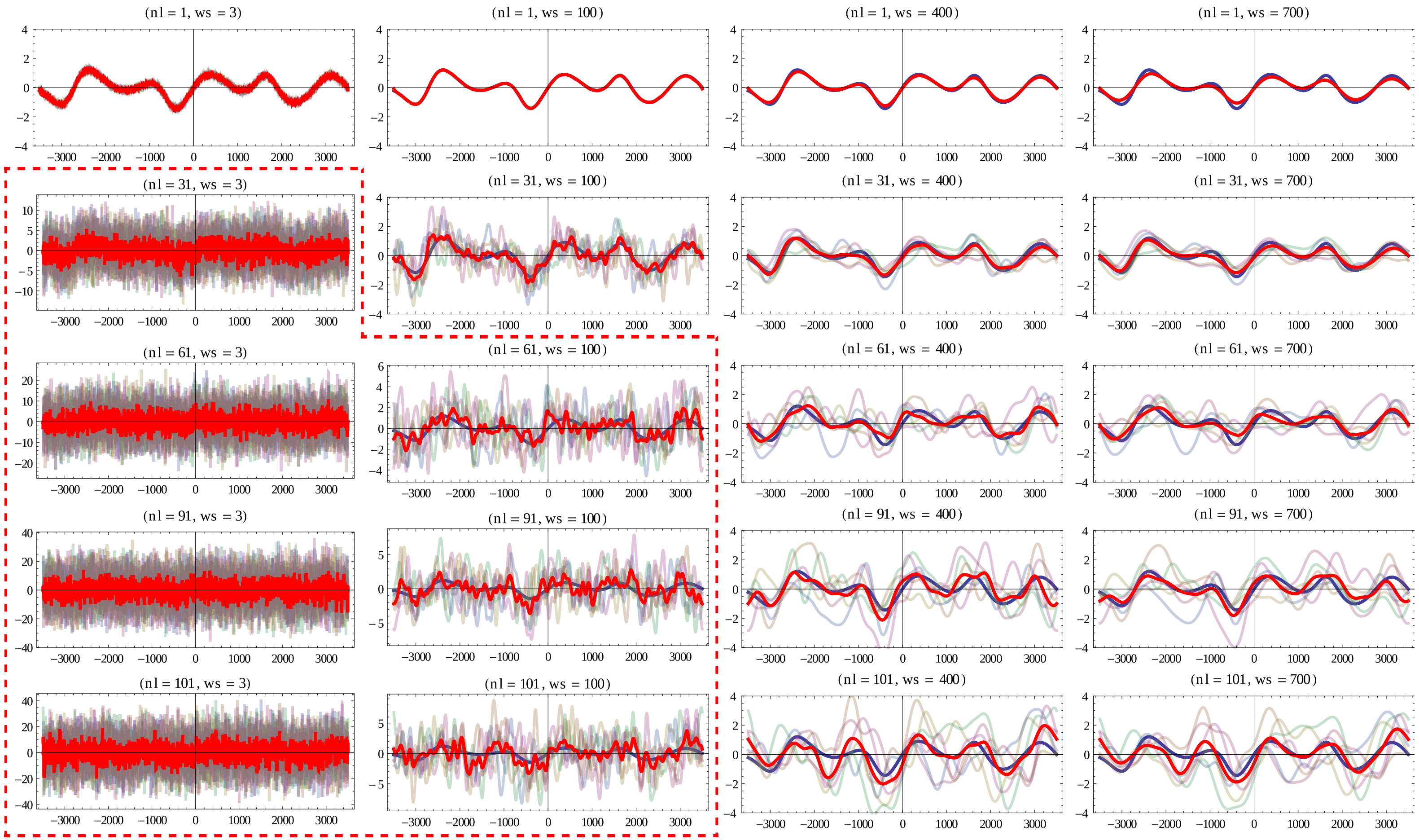
Figure on page 3 suggests that with the average of 6 replicates, we are able to qualitatively regenerate the original function for SNR as low as 15^{-2} (i.e. noise amplitude $\sim$15 times that of the signal amplitude).[7]

[4] Note that although white noise has a flat frequency spectrum (i.e. all frequencies are present) the net frequency component (power) for any given frequency is much smaller than the signal frequency.

[5] Reproducibility is a fundamental demand of any scientific experiment, and is key for its acceptability and validity. However, under certain stochastic conditions the system can have high degree of variability and exact reproducibility can't be achieved.

[6] Just like one will have to toss a coin several times to test whether its a fair coin or not, just one or two tosses wont be able to give a definitive answer.

[7] This is a conservative estimate, as we were able to recover good correlation for up to



Different panels show the same hypothetical response function (in Blue). Each panel has 6 replicates (in dimmed colors) with various noise levels $(nl)^a$ overlayed and smoothed with various window sizes (ws). The average of the replicates is shown in Red. Note that when the noise level is high, the replicates (of the response function with same noise level) behave very differently, but the original behavior is recovered upon averaging. The plots shown in dashed box are on different scales.

[a] See Suppl. Note 2.2 for a definition of noise level (nl).

If it is not possible to do enormously large number of replicates (due to say economic reasons), the average of all the replicates done is a better measure than the individual experiments.

It may seem that a large number of replicates might be needed, but that is not true. For high noise experiments like microarrays, even for our low free energy effect, 3–4 replicates are sufficient to achieve an adequate level of accuracy (with meta-analysis this number comes down to 2–3 experiments).

1.2 Calibration for SNR extraction from a given data

The method described in the previous section can be evolved to generate a calibration for estimation of signal-to-noise ratio (SNR) (or noise levels)[8] from a given data provided that the data meets the criterion described in the previous section.

To calibrate, we first define a characteristic function based on known features of data. Then we overlay different levels of white noise on this data, which are equivalent to different replicates. At low noise each replicate closely mimics the original function. But as the noise levels go up, the replicates are averaged in different combinations of increasing numbers until we get a close fit to the original profile (see figure on page 3).

Several thousand simulations were run for various noise levels[9] (ranging between 1 to 100) on unit signal amplitudes[10] with a mix of various small frequencies (which were chosen based on our experimental data). Each of these noisy dataset is then smoothed for various window sizes ranging from 400 to 700 (see Table below). The standard deviation of the differences between original noisy dataset and smooth datasets gives a metric for the preselected window sizes. By averaging a large number of entries, coefficient table below was generated.

This coefficient table is then used to predict the noise levels of any given dataset. This prediction algorithm was tested on several thousand simulated datasets[11] generated for various noise levels (ranging between 1 to 10^3) on various signal amplitudes (ranging between 10^{-4} to 10) with a mix of various small frequencies (which were much larger then experimental data).

about 50 times noise with only 6 replicates.

[8] See definition of noise level in Suppl. Note 2.2.

[9] See Suppl. Note 3.3 for the protocol used for simulating noisy data.

[10] Signal amplitude is defined as half of the difference between max and min values of all amplitudes.

[11] Each dataset is used alone, no replicates.

Table on the following page summarizes the prediction results.

Note that when we have some knowledge about the noise levels, we are able to successfully predict a much broader range, i.e. up to about noise level 40. However, when we have absolutely no knowledge about the noise level, we can still successfully predict the noise levels up to 23. Our meta-analysis data in Fig. 2 and 3 has a noise level of about 13, which is well within the successful prediction range.

The method presented here gives an unsupervised prediction of noise level. A supervised prediction (i.e. with more information about the data) will give better results, but the unsupervised method is sufficient for the present analysis.

This analysis can help predict the number of replicates needed, for a noisy experiment, up to a desired reproducibility-confidence-interval from just one experiment. A simulation on replicates shows that for noise levels at least up to 46, average of three replicates gives high enough noise reduction so that a fourth replicate doesnt add much improvement. This is a reconfirmation that for the purpose of this work 3 replicates are sufficient.

While generalizing this technique, the following facts must be kept in mind. The calibration (and smoothing) is a function of data size and density, frequency spectrum of the data, noise amplitude[12] and frequency etc. Although a complete analytical understanding of the calibration is beyond the scope of this paper, one can safely say that this method will work for very high noise levels for high frequency data also if the sampling frequency is sufficiently high.

[12] The dependence is only on the noise amplitude and not on the signal amplitude.

Calibrated correction coefficients for various window sizes.

Window Size	Coeff
400	3.46323
500	3.46300
600	3.46295
700	3.46295

Errors in prediction of noise for datasets with known or unknown noise levels.

Known Noise Level	Stdev (σ) of Prediction Errors	Unknown Noise Level	Stdev (σ) of Prediction Errors
1	0	1	0
2	0	2	0
3	1	3	0
4	1	4	1
5	1	5	1
6	1	6	1
7	1	7	1
8	1	8	1
9	2	9	1
10	2	10	1
11	2	11	1
12	2	12	2
13	2	13	2
14	3	14	2
15	2	15	2
16	2	16	3
17	3	17	3
18	2	18	5
19	3	19	4
20	3	20	4
21	3	21	6
22	3	22	7
23	3	23	16
24	4	24	257
25	4	25	1190
26	3		
27	4		
28	4		
29	6		
30	5		
31	5		
32	6		
33	6		
34	10		
35	8		
36	8		
37	9		
38	7		
39	10		
40	12		

2 Definitions

2.1 Signal-to-Noise Ratio

The signal-to-noise ratio is a commonly used term to describe the signal corruption by noise, and is defined as the ratio of signal power to the noise power, see Eq. 1, where A is the root mean square amplitude. For more details please see [3].

$$SNR = \frac{P_{signal}}{P_{noise}} = \left(\frac{A_{signal}}{A_{noise}}\right)^2 \tag{1}$$

2.2 Noise Level

The signal-to-noise ratio, as defined in the previous section, has it's origins in electrical engineering where it relates to the ratio of powers in signal and noise. For the convenience of remembering, and ease of intuitive understanding, we define a new term *noise level*. Eq. 2 defines the *noise level* in terms of the signal and noise amplitudes (a), which are given by the difference between max and min values of the amplitudes.

$$nl = 2\frac{a_{noise}}{a_{signal}} \simeq \frac{2}{\sqrt{SNR}} \tag{2}$$

Eq. 2 suggests that, a noise level of 10 would mean that the noise amplitude is 5 times larger than the signal amplitude.[13] In other words, one unit of signal is burried in 5 units of noise.

[13] See Suppl. Note 3.3 for how this definition is used to simulate noisy data.

2.3 Definition of Sets

Ratio	Short	Description	Equivalence
$\frac{XL}{nXL}$	$CL \rightarrow log_2(\frac{XL}{nXL})$	Relative enrichment of cross–linked DNA (or psoralen intercalation) in untreated (no drug treatment) Raji B cells	Psoralen binding due to a combined effects of sequence, inherent chromatin structure and transcriptionally generated dynamic supercoiling
$\frac{XL\,(DRB)}{nXL\,(DRB)}$	$CL(DRB) \rightarrow log_2(\frac{XL\,(DRB)}{nXL\,(DRB)})$	Relative enrichment of cross–linked DNA (or psoralen intercalation) in DRB treated cells	Psoralen binding mainly due to sequence and inherent chromatin structure (DRB would inhibit transcription, so no dynamic supercoiling)
	$\Delta CL \rightarrow CL(DRB) - CL$		Transcription generated dynamic DNA supercoiling (due to ongoing transcription)
$\frac{XL\,(CPT)}{nXL\,(CPT)}$	$CL(CPT) \rightarrow log_2(\frac{XL\,(CPT)}{nXL\,(CPT)})$	Relative enrichment of cross–linked DNA (or psoralen intercalation) in camptothecin treated cells	
	$\Delta CL(CPT) \rightarrow CL(DRB) - CL(CPT)$		Transcription generated dynamic DNA supercoiling in cells treated with CPT
$\frac{XL\,(\beta\ Lap)}{nXL\,(\beta\ Lap)}$	$CL(\beta\ Lap) \rightarrow log_2(\frac{XL\,(\beta\ Lap)}{nXL\,(\beta\ Lap)})$	Relative enrichment of cross–linked DNA (or psoralen intercalation) in β–lapachone treated cells	
	$\Delta CL(\beta\ Lap) \rightarrow CL(DRB) - CL(\beta\ Lap)$		Transcription generated dynamic DNA supercoiling in cells treated with β Lap

2.4 General Definitions

Meta Analysis

During meta-analysis we average multiple transcribed regions by aligning transcribed regions at the transcription start sites (±8,000 bp). For all our analysis we have averaged the raw data, and smoothed only the final average. The ratios are calculated for each individual probe of microarray.

Expression Levels

Expression levels were defined as the average of the scores (or signal) for all probes of an annotated gene body.[14] We had 3 replicates of the expression array hybridizations, and average of expression levels from these

[14] In other words the total score, normalized by the number of probes.

three experiments were used for further calculations. The expression level is calculated from raw data which was baseline shifted (no smoothing).

Expression Level Classes

Once the expression levels were defined, we classified data in several groups (decades, quintiles, quartiles, tertiles etc.). After looking at these different groups, it was apparent that at the level of resolution of our experiments, the data is best viewed in quintiles. For simplicity of explanation, transcribed regions were classified in three categories (based on the expression levels): Low (0–20%, 20–40%), medium (40–60%, 60–80%), high (80–100%).

2.5 Baseline Shifting

Since we expect ratios to be small,[15] we normalize the entire hybridization experiment so as to bring the overall baseline across the chromosome to zero. This is achieved simply by averaging the ratios of all probes across the chromosomes, and subtracting the average from all the probes.

We also used the same concept baseline shifting to remove the sequence dependent bias of psoralen for DNA intercalation.[16]

3 Analysis Methods

3.1 Data Analysis

Owing to the small free energy of intercalation of psoralen, the hybrization data was noisy, and had a very small signal to noise ratio.[17] The appearance of the raw data (for all regions) suggested that there was significant high frequency noise (i.e. large variations over short lengths along the DNA). Considering the magnitude of the bending and torsional persistence lengths for DNA $\sim$50–100 nm (about 150–300 bp) [4], variation in supercoiling occurring on a much shorter scale is unlikely unless accompanied by a dramatic structural transitions, almost certainly an infrequent

[15] Because psoralen has a small free energy corresponding to interaction in negatively supercoiled DNA. Moreover, it does have some affinity for intercalation in relaxed DNA as well. Also, see Suppl. Note 1.1.

[16] Also see Suppl. Note 3.2.

[17] See Suppl. Note 1.1.

phenomenon. Therefore the high frequency fluctuations were attributed to noise.

In order to suppress this noise, we used a technique called Fourier Convolution Smoothing (FCS) to smooth the data [2]. The benefit of FCS is that it dampens the high frequency noise much more than the low frequency noise. The technique uses moving window average as a reference,[18] as a result of which the local features are not lost during an unsupervised noise reduction.

Our microarrays are designed with each probe having 50 bp and a 12 bp overlap (i.e. 38 bp are unique between successive probes). So for any given region of genome or an individual transcribed region, we have a data density of 38 bp per data point (i.e. per probe). While doing the meta-analysis,[19] we align all the transcribed regions on the transcription start sites (TSS). Since the TSS are randomly distributed with respect to probes, for the meta-analysis the data density increases to 1.4 bp per data point. The meta-analysis presented in this study uses a window size of 400 data points (equivalent to 561 bp).

Based on the DNA properties, we improvised upon the previously described FCS technique to fit it for our data. The ENCODE data on Nimblegen microarrays was not continuous, so whenever we had a break of 600 bp or more (i.e. abt 15 probes), those data points were separated into distinct groups, and smoothed individually. Continuous regions with less than 400 probes were also dropped from individual transcribed regions.

Our Nimblegen ENCODE ($hg18$) microarrays had usable data for a total of 855 transcribed regions. Since many of these regions were overlapping, there was a possibility of over-representing a specific gene. In order to avoid this we identified clusters of transcripts/genes that were overlapping or had a TSS within 50 bp of each other; and used only the largest of "transcribed region" from each of these groups. This brought down the total number of transribed regions to 445 (with 415 unique genes). See the list of these transcribed regions in Table 1.

3.2 Sequence Dependent Background Correction

These 445 transcribed regions were sorted based on the expression levels[20] and segregated in various quantiles (decades, quintiles, quartiles,

[18] With a pre-decided window-size (ws), the only parameter used for smoothing.

[19] See Suppl. Note 2.4 for definition.

[20] See definition in Suppl. Note 2.4.

tertiles etc.). When meta-analysis[21] was performed for all 445 transcribed regions in these quantiled datasets, we observed a graded difference in baselines for each quantile.[22]

We wanted to understand this difference, and explain it. It is well known that psoralen has a sequence dependant bias for intercalation in DNA. So we sorted the transcribed regions based on the AT content within ±3,000 bp of TSS (instead of sorting them by expression). In the meta-analysis, it was very obvious that the AT-rich transcribed regions had a much higher psoralen intercalation, irrespective of expression level. So we have decided to do an AT content dependent baseline shift for different transcribed regions. To reduce systematic errors, these 445 transcribed regions were divided in 10 groups (each having about 44–45 transcribed regions). Now a correction term, for each of the decades, was calculated by averaging the raw ratios in the flanking regions of (-8,000, -2,000) bp and (2,000, 8,000) bp (about TSS) of the constituent transcribed regions.[23] The data for each of the constituent transcribed regions is then baseline shifted using this correction term to get the corrected data, which is used for further analysis.[24]

3.3 Addition of Noise Levels in Simulations

There are several ways one could add noise on a pure signal. For our simulations, we used the following protocol for noise addition: For a given dataset and noise level (say nl) we generate dataset of equal length such that each point is the product of nl and a (pseudo) random number in the range of $-\frac{1}{2}$ and $+\frac{1}{2}$.[25]

[21] See definition in Suppl. Note 2.4.

[22] The low expression quantiles had a higher baseline than the high expressing quantiles.

[23] If we had enough data points for all the transcribed regions, we could in principle do a baseline shift based on the flank psoralen profile of each individual gene, but due to lack of continuous data points, we have decided to use the the flanks: (-8,000, -2,000) bp and (2,000, 8,000) bp (about TSS).

[24] All the processing was done on raw data, and smoothing was applied only in final step to remove the high frequency noise.

[25] Another possibility could be to use a Gaussian distribution with mean, $\mu = 0$, and standard deviation, $\sigma = nl$.

3.4 Pausing index of RNA polymerase II

Pol II pausing (or stalling) index ([5], [6]) is a measure that reflects the dynamics of Pol II assembly and promoter clearance. It represents the ratio of Pol II read density around TSS (1 kb region centered at TSS) over the average read density in the gene body (starting 750 bp downstream of TSS). Genes were split into three groups as follows: paused – the pausing index above 6 and no detectable Pol II in gene body ($p = 0.05$), elongating the pausing index between 0.5 and 6, and detectable Pol II in TSS and gene body regions, and silent no detectable Pol II signal in TSS and gene body regions.

3.5 Gene body analysis

As shown in Table 1, we have a total of 445 genes. Suppl. Fig. 1a shows the average ΔCL profile for all 445 genes starting from the TSS to 4 kb into the gene body. Out of these 445 genes, 29 genes are shorter than 1 kb, 35 are in range 1-2 kb, 38 are in range 2-3 kb and 24 in 3-4 kb range. The remaining 319 genes are larger than 4 kb. For the analysis, only the data in the gene body was considered, i.e. data after the transcription termination site to 4 kb was dropped in Suppl. Fig. 1(a).

3.6 Enhancers and CTCF sites

The list of enhancers and CTCF sites in GM06990 B-lymphocyte cells were obtained from Heintzman et al. ([7]) study, where the detailed identification procedure is presented. All coordinates were converted from $hg17$ to $hg18$ human genome version using liftOver program (UCSC tools set). Only enhancers with significant Pol II signal were considered.

In order to understand the interaction of enhancers and CTCF sites with the dynamic supercoiling the lists were divided into two parts. The elements that are in or within ±3,000 bp of the gene body (proximal) or are more than ±3,000 bp away from the gene body (distal). Out of 463 enhancers (127 proximal, 336 distal) only 122 proximal and 111 distal enhancers had data near ENCODE regions. All of 729 CTCF sites (444 proximal, 285 distal) had some ENCODE data within ±5,000 bp of the sites. The results are plotted in Suppl. Fig. 1 (b) and (c). These lists were further studied by classification based on presence or absence of Pol II (data to be deposited online).

3.7 Analysis of divergent promoters

It has been previously shown (*in-vivo*) using a model system that divergent promoters generate a higher level of negative supercoiling in their shared upstream region [8]. We wanted to test the generality of this observation in our data.

It is known that Pol II footprint is about 40 bp and we have already seen that the effect of transcription generated dynamic supercoiling travels about 1,500 bp upstream of the transcribing Pol II. Therefore, we analyzed all the divergent promoters in our gene-list which were separated by more than 100 bp but less than 4,000 bp. There were a total of only 23 such promotor pairs in our ENCODE array. The average ΔCL profile about the TSS for all these divergent promoters is shown in Suppl. Fig. 1. The shaded range shows the $\mu \pm \sigma$ region obtained by averaging 30 randomizations of an equivalent number of genes excluding divergent promoters.

Considering our discussion in previous sections, one of the reasons for the absense of a statistically significant difference is that the number of promotor pairs is very small. Moreover, only 7 out of 23 promotor pairs were experessed in our experiments. Since mutual reinforcement of supercoiling would need simultaneous (or near simultaneous) firing of these promoters. So these data neither confirm nor refute the previous (*in-vivo*) observations of accumulation of negative supercoils in the shared upstream regions of the divergent promoters [8].

3.8 3D profiles

To generate the 3D profiles in Fig. 4(a) a moving window average with 20% genes was taken after sorting them based on the expression level. More specifically, the expression level sorted list of 445 genes was divided in successive overlapping groups of 89 genes (i.e. 20% of the 445 genes) giving a total of 357 such groups. These groups were individually averaged and smoothed (as described in Suppl. Note 3.1), and the resulting data were used to generate the 3D profiles in Fig. 4(a).

References

[1] R R Sinden, J O Carlson, and D E Pettijohn. Torsional tension in the dna double helix measured with trimethylpsoralen in living e. coli cells: analogous measurements in insect and human cells. *Cell*, 21(3):773–783, Oct 1980.

[2] M K Raghuraman, E A Winzeler, D Collingwood, S Hunt, L Wodicka, A Conway, D J Lockhart, R W Davis, B J Brewer, and W L Fangman. Replication dynamics of the yeast genome. *Science*, 294(5540):115–121, Oct 2001.

[3] Wikipedia. Signal-to-noise ratio - wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Signal-to-noise_ratio. [Online; accessed 31-March-2012].

[4] C Lavelle. Forces and torques in the nucleus: chromatin under mechanical constraints. *Biochem Cell Biol*, 87(1):307–322, Feb 2009.

[5] J Zeitlinger, A Stark, M Kellis, J W Hong, S Nechaev, K Adelman, M Levine, and R A Young. Rna polymerase stalling at developmental control genes in the drosophila melanogaster embryo. *Nat Genet*, 39(12):1512–1516, Dec 2007.

[6] G W Muse, D A Gilchrist, S Nechaev, R Shah, J S Parker, S F Grissom, J Zeitlinger, and K Adelman. Rna polymerase is poised for activation across the genome. *Nat Genet*, 39(12):1507–1511, Dec 2007.

[7] N D Heintzman, G C Hon, R D Hawkins, P Kheradpour, A Stark, L F Harp, Z Ye, L K Lee, R K Stuart, C W Ching, K A Ching, J E Antosiewicz-Bourget, H Liu, X Zhang, R D Green, V V Lobanenkov, R Stewart, J A Thomson, G E Crawford, M Kellis, and B Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, May 2009.

[8] F Kouzine, S Sanford, Z Elisha-Feil, and D Levens. The functional response of upstream dna to dynamic supercoiling in vivo. *Nat Struct Mol Biol*, 15(2):146–154, Feb 2008.