SUPPLEMENTAL NOTE

Mouse fragile sites analysis

We evaluated the ability of our models derived from human data to predict 24 known APH-induced mouse fragile sites (Elder and Robinson 1989; Helmrich et al. 2006), including 16 such sites for which breakage frequencies were evaluated from 266 mouse cells (Elder and Robinson 1989). Mouse aCFSs were mapped onto the mouse genome (mm8), and the genomic features identified as significant predictors in our human-based models (i.e. G-band coverage, distance to centromere, CpG islands, Twist, low complexity A/T rich region coverage, mononucleotide microsatellite coverage, and *Alu* coverage) were computed. Instead of *Alu* element coverage, for the mouse genome we considered "young" SINE element content (all SINEs excluding MIRs) and B1 content. B1s in mouse and *Alu*s in human are both derived from the 7SL RNA gene and they both contain mononucleotide microsatellites (Yang et al. 2004). B1s constitute ~38% of all mouse SINEs (Yang et al. 2004). Our optimal (Table 2) and alternative (Table 3) logistic regression models can predict 79% and 71-79% correct predictions, respectively (Table S10). However, because mouse fragile sites have not been screened genome-wide, we cannot define mouse NFRs, and therefore, cannot estimate a false positive rate.

Our human-based breakage frequency standard regression model (Table 4) did not show a high predictive power for mouse APH-induced fragile sites breakage frequency (data not shown). The lower predictive power of this model compared with our logistic model might come from the lower pseudo R- squared of our breakage frequency model. Alternatively, it is possible that genomic landscapes affecting breakage frequency are not conserved between human and mouse, or that a larger data set is required to capture global pattern of APH-induced fragile sites in mouse. This issue will be important to re-investigate once the genome-wide screening of CFSs in other mammalian species becomes available.

SUPPLEMENTAL TABLES

Table S1. List of aCFSs and genomic locations (hg18) in BED format. Gaps in assembly were removed to ensure accuracy in coverage and density calculations. Therefore, some aCFSs are separated into several intervals.

FRA	Chromosome	Start	Stop
1A	chr1	0	167280
	chr1	217280	257582
	chr1	307582	461231
	chr1	511231	2624080
	chr1	2674080	3835128
	chr1	3895128	5335220
	chr1	5385220	9200000
1B	chr1	51300000	60900000
1D	chr1	84700000	94500000
1E	chr1	99400000	102000000
1F	chr1	142436048	142562525
	chr1	142612525	142807140
	chr1	142857140	142935838
	chr1	142985838	143113101
	chr1	143163101	143333770
	chr1	143383770	143422081
	chr1	143522081	144544475
	chr1	144594475	144876007
	chr1	144926007	146492662
	chr1	146542662	146727982
	chr1	146777982	146950771
	chr1	147000771	147221084
	chr1	147271084	147726269
	chr1	147776269	153300000
1G	chr1	171200000	174300000
1I	chr1	241700000	246974833
	chr1	247024833	247199719
1K	chr1	184000000	197500000
1L	chr1	60900000	84700000
2C	chr2	17000000	19100000
2D	chr2	52700000	54800000
2E	chr2	70500000	75400000
2F	chr2	134800000	136600000
2G	chr2	169500000	182700000
2H	chr2	182700000	189100000
2I	chr2	197100000	209100000
2J	chr2	237000000	239420971

	chr2	239453971	239466915
	chr2	239496915	240449069
	chr2	240474069	242751149
3A	chr3	23800000	26400000
3B	chr3	58500000	63700000
3C	chr3	184200000	189400000
3D	chr3	150400000	161200000
4A	chr4	5200000	8783740
	chr4	8933740	10900000
4C	chr4	151000000	155100000
4D	chr4	10900000	31421166
	chr4	31492166	32468752
	chr4	32527752	35500000
4F	chr4	88200000	99100000
5C	chr5	130400000	135400000
5D	chr5	91900000	97300000
5E	chr5	18500000	29300000
5F	chr5	102800000	109600000
6B	chr6	4100000	7000000
6C	chr6	23500000	26100000
6E	chr6	160900000	164400000
6F	chr6	104800000	113900000
6G	chr6	87500000	92100000
7B	chr7	34000	327567
	chr7	477567	7200000
7C	chr7	35600000	37500000
7D	chr7	43300000	46600000
7E	chr7	90900000	92600000
7F	chr7	104400000	107200000
7G	chr7	114400000	117200000
7H	chr7	130100000	132400000
7I	chr7	147500000	153901567
	chr7	154001567	154737899
	chr7	154817899	158821424
7J	chr7	71800000	74353660
	chr7	74603660	77400000
8B	chr8	93500000	99100000
8C	chr8	117700000	127300000
8D	chr8	140000000	144006276
	chr8	144106276	145396296
	chr8	145403396	146274826
9B	chr9	113900000	116700000
9D	chr9	89600000	91000000
10D	chr10	71300000	74600000

10E	chr10	111800000	114900000
10F	chr10	119100000	125859462
	chr10	125909462	127400000
10G	chr10	42100000	45746970
	chr10	45896970	46849175
	chr10	46999175	47262482
	chr10	47412482	47575713
	chr10	47725713	48715542
	chr10	48865542	50807416
	chr10	50857416	51068851
	chr10	51118851	53300000
11C	chr11	16100000	21600000
11D	chr11	26000000	27200000
11E	chr11	31000000	36400000
11F	chr11	85300000	87366026
	chr11	87378026	87900000
11G	chr11	115400000	120700000
11H	chr11	69200000	69437111
	chr11	69454899	70700000
12B	chr12	78700000	91200000
12E	chr12	107500000	107857874
	chr12	107914874	121006020
	chr12	121156020	131272065
	chr12	131317065	132289534
13A	chr13	32900000	34700000
13C	chr13	57600000	60500000
13D	chr13	93800000	98100000
14B	chr14	55800000	67000000
14C	chr14	67000000	69300000
15A	chr15	55800000	65300000
16C	chr16	65200000	69400000
16D	chr16	78200000	80500000
17B	chr17	54900000	55600000
18A	chr18	31000000	35500000
18B	chr18	52000000	59800000
20B	chr20	9000000	11900000
22B	chr22	27900000	30500000
XB	chrX	6000000	9500000
XC	chrX	98200000	102500000
XD	chrX	140100000	141900000

Table S2. List of NFRs and genomic locations (hg18) in BED format. Gaps in assembly were removed to ensure accuracy in coverage and density calculations. Therefore, some NFRs are broken into several intervals.

NFR	Chromosome	Start	Stop
1	chr1	9200000	12975585
	chr1	13025585	13142499
	chr1	13192499	13429749
	chr1	13479749	16998245
	chr1	17048245	29750669
	chr1	29800669	34400000
2	chr1	46500000	51300000
3	chr1	102000000	103636494
	chr1	103686494	103843838
	chr1	103893838	107000000
4	chr1	117600000	120498679
	chr1	120548679	120700000
5	chr1	153300000	154800000
6	chr1	163800000	171200000
7	chr1	174300000	184000000
8	chr2	7000000	16171403
	chr2	16221403	17000000
9	chr2	19100000	21012571
	chr2	21037571	23900000
10	chr2	47600000	52700000
11	chr2	54800000	61100000
12	chr2	75400000	83700000
13	chr2	102100000	108600000
14	chr2	113800000	118600000
15	chr2	129600000	134800000
16	chr2	136600000	144700000
17	chr2	189100000	197100000
18	chr2	209100000	215100000
19	chr2	221300000	233711985
	chr2	233731985	237000000
20	chr3	8700000	11500000
21	chr3	14700000	23800000
22	chr3	26400000	32100000
23	chr3	54400000	58500000
24	chr3	63700000	66115833
	chr3	66375833	71800000
25	chr3	74200000	87200000

26	chr3	99800000	101500000
27	chr3	104400000	115000000
28	chr3	144400000	150400000
29	chr4	0	1413146
	chr4	1464146	3883456
	chr4	3963456	5200000
30	chr4	40900000	45600000
31	chr4	59200000	59429326
	chr4	59479326	69275441
	chr4	69375441	71536854
	chr4	71711854	75641303
	chr4	75671303	76500000
32	chr4	102500000	120600000
33	chr4	124000000	151000000
34	chr5	8200000	18500000
35	chr5	42400000	45800000
36	chr5	97300000	97589886
	chr5	97612886	102800000
37	chr5	135400000	138812257
	chr5	138817257	147200000
38	chr6	5000	4100000
39	chr6	7000000	9199728
	chr6	9249728	13500000
40	chr6	15500000	23500000
41	chr6	26100000	40600000
42	chr6	45200000	57200000
43	chr6	63500000	70000000
44	chr6	75900000	87500000
45	chr6	92100000	95737264
	chr6	95937264	99900000
46	chr6	139100000	149100000
47	chr6	164400000	167766922
	chr6	167784922	170021897
	chr6	170171897	170896992
48	chr7	7200000	13300000
49	chr7	19500000	35600000
50	chr7	37500000	43300000
51	chr7	46600000	48167949
	chr7	48207949	50338125
	chr7	50378125	53900000
52	chr7	61100000	61314455
	chr7	61364455	61554592
	chr7	61604592	71800000
53	chr7	77400000	90900000

54	chr7	92600000	104400000
55	chr7	117200000	130100000
56	chr7	132400000	137300000
57	chr7	142800000	147500000
58	chr8	6200000	7462059
	chr8	7562059	12099352
	chr8	12199352	19100000
59	chr8	29700000	29732668
	chr8	29798768	38500000
60	chr8	55600000	66100000
61	chr8	74000000	86069241
	chr8	86193341	86763703
	chr8	86851003	87200000
62	chr8	99100000	101600000
63	chr8	106100000	117700000
64	chr8	127300000	140000000
65	chr9	9000000	14100000
66	chr9	40200000	40223029
	chr9	40273029	40415834
	chr9	40465834	40930341
	chr9	40980341	41133214
	chr9	41183214	41355793
	chr9	41405793	42603951
	chr9	42653951	43203694
	chr9	43253694	43886565
	chr9	43936565	44616642
	chr9	44666642	44848289
	chr9	44898289	45190199
	chr9	45240199	45705517
	chr9	45755517	46106426
	chr9	46156426	46351035
	chr9	46401035	46700000
67	chr9	91000000	91533236
	chr9	91583236	91668616
	chr9	91718616	101600000
68	chr10	6700000	12300000
69	chr10	74600000	81231464
	chr10	81241464	89600000
70	chr10	98000000	99400000
71	chr10	102000000	111800000
72	chr10	114900000	119100000
73	chr10	127400000	128606059
	chr10	128656059	133271394

	chr10	133281394	133527517
	chr10	133577517	135374737
74	chr11	50000	1152759
	chr11	1169335	10700000
75	chr11	12600000	16100000
76	chr11	21600000	26000000
77	chr11	27200000	31000000
78	chr11	56400000	68846377
	chr11	68848982	69200000
79	chr11	70700000	85300000
80	chr11	87900000	92300000
81	chr11	96700000	101600000
82	chr11	110000000	115400000
83	chr11	120700000	134452384
84	chr12	5300000	7059293
	chr12	7132293	14800000
85	chr12	36500000	44600000
86	chr12	53100000	56300000
87	chr12	69800000	74791940
	chr12	75041940	78700000
88	chr12	91200000	94800000
89	chr12	100000000	107500000
90	chr13	18400000	32900000
91	chr13	34700000	39500000
92	chr13	52200000	57600000
93	chr13	60500000	72100000
94	chr13	98100000	109100000
95	chr14	23600000	31800000
96	chr14	36900000	41000000
97	chr14	43200000	48300000
98	chr14	69300000	72900000
99	chr14	78400000	106360585
	chr15	23300000	24687593
	chr15	24731593	25169606
100	chr15	25270606	25700000
101	chr15	31400000	37900000
102	chr15	65300000	70400000
	chr15	96300000	96345672
103	chr15	96367672	100338915
	chr16	0	8576922
104	chr16	8594422	14700000
105	chr16	16700000	21700000
106	chr16	69400000	78200000
107	chr16	80500000	82700000

	chr17	0	296854
108	chr17	343376	11200000
	chr17	28800000	31699961
	chr17	31799961	33429230
109	chr17	33529230	35400000
110	chr17	47600000	54900000
111	chr17	55600000	59900000
112	chr18	7200000	15400000
113	chr18	23300000	31000000
	chr18	35500000	50313134
114	chr18	50360134	52000000
	chr19	11000	7297004
	chr19	7302004	8593198
115	chr19	8598198	12600000
116	chr20	5000000	9000000
117	chr20	11900000	17800000
118	chr20	21200000	25700000
119	chr20	37100000	41100000
120	chr20	49200000	54400000
	chr21	38600000	41877429
	chr21	41878628	43505733
121	chr21	43507092	46944323
	chr22	16300000	18889431
122	chr22	18939431	27900000
123	chr22	30500000	39300000
	chr22	42600000	43057958
	chr22	43107958	47356150
	chr22	47366250	48750436
	chr22	48767136	49087576
	chr22	49089176	49107103
124	chr22	49126803	49591432
	chrX	0	34821
	chrX	84821	171384
	chrX	201384	967557
	chrX	1017557	1054113
	chrX	1104113	1184234
	chrX	1274234	2028238
125	chrX	2128238	6000000
	chrX	9500000	37008177
126	chrX	37033177	47300000
127	chrX	65000000	67700000
	chrX	76000000	76570348
128	chrX	76590348	98200000
129	chrX	102500000	113403924

	chrX	113473924	115596318
	chrX	115616318	120700000
130	chrX	137800000	140100000
	chrX	141900000	143335010
131	chrX	143365010	146900000

Table S3. A list of 54 genomic predictors considered in the study

Predictors	Type of data[a]	Data source	Previous studies[b]
Global Genome Organization			
G-banding	Coverage	(Furey and Haussler 2003)	[-, (Mishmar et al. 1999)]
Distance to the telomere	Distance	Genome-wide screen	
Distance to the centromere	Distance	Genome-wide screen	
Purine percent	Average	Genome-wide screen	
GC content (% base G+% base C)	Average	Genome-wide screen	[-, (Mishmar et al. 1999)] [+, (Tsantoulis et al. 2008)]
Gene Expression/Chromatin Structure			
Gene region	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	[+, (Helmrich et al. 2006)]
CpG island	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
Transcription start site	Density	(Karolchik et al. 2003) (Rhead et al. 2009)	
Nuclear lamina binding sites	Coverage	(Guelen et al. 2008)	
H3K4me1 sites	Coverage	(Barski et al. 2007)	
H3K9Ac sites	Coverage	(Wang et al. 2008)	[-,(Jiang et al. 2009)]
H3K14Ac sites	Coverage	(Wang et al. 2008)	[-, (Jiang et al. 2009)]
RNA polymerase II binding sites	Coverage	(Barski et al. 2007)	
miRNA sites	Coverage	(Griffiths-Jones et al. 2007)	

DNA Sequence Elements			
Alu repeats	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	[+, (Tsantoulis et al. 2008)]
MIR (type of SINE) repeats	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
LINE1 repeats	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
LINE2 repeats	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
L1 endonuclease cleavage sites	Density	Genome-wide screen	
DNA transposons	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
MER (type of DNA transposon) repeats	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
Twist value calculated from sequence	Average	Genome-wide screen	[+, (Mishmar et al. 1998)] [0, (Helmrich et al. 2006)] [-, (Tsantoulis et al. 2008)]
Low complexity A/T rich regions	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
Mononucleotide microsatellites (>9 repeats)	Coverage	(Abajian) (Kelkar et al. 2010)	
Dinucleotide microsatellites (>5 repeats)	Coverage	(Abajian) (Kelkar et al. 2010)	
Trinucleotide microsatellites (>4 repeats)	Coverage	(Abajian) (Kelkar et al. 2010)	
Tetranucleotide microsatellites (>4 repeats)	Coverage	(Abajian) (Kelkar et al. 2010)	
All A/T motif microsatellites	Coverage	(Abajian)	
G-Quadruplex forming repeats	Coverage	(Cer et al. 2010)	
Z-DNA motifs	Coverage	(Cer et al. 2010)	

Inverted repeats	Coverage	(Cer et al. 2010)	
Cruciform motifs	Coverage	(Cer et al. 2010)	
Directed repeats	Coverage	(Cer et al. 2010)	
Slipped motifs	Coverage	(Cer et al. 2010)	
Mirror repeats	Coverage	(Cer et al. 2010)	
Triplex motifs	Coverage	(Cer et al. 2010)	
A-Phased repeats	Coverage	(Cer et al. 2010)	
DNA Replication			
Replication timing (Woodfine)	Assigned value	(Woodfine et al. 2004)	
Replication timing (Ryba)	Assigned value	(Ryba et al. 2010) (Weddington et al. 2008)	
Early replicated regions in lymphocyte cell line (Hansen)	Coverage	(Hansen et al. 2009)	
Origins of replication computationally predicted from base skew (Chen)	Density	(Chen et al. 2011)	
Origins of replication from purified origin-centered nascent strand using union set of lambda exonuclease digestion and anti-bromodeoxyuridine immunoprecipitation assays (Karnani)	Density	(Karnani et al. 2010)	
Origins of replication from ENCODE data	Density	(Cadoret et al. 2008)	
Topoisomerase I motif (CAT)	Density	Genome-wide screen	[+/-, (Arlt and Glover 2010)]

Topoisomerase I motif (CTY)	Density	Genome-wide screen	[+/-, (Arlt and Glover 2010)]
Topoisomerase I motif (GTY)	Density	Genome-wide screen	[+/-, (Arlt and Glover 2010)]
Topoisomerase I motif (RAK)	Density	Genome-wide screen	[+/-, (Arlt and Glover 2010)]
Topoisomerase I motif (YCCTT)	Density	Genome-wide screen	[+/-, (Arlt and Glover 2010)]
Topoisomerase I motif (YTA)	Density	Genome-wide screen	[+/-, (Arlt and Glover 2010)]
Recombination and Mutational Pathways			
Sex-averaged recombination rates	Assigned value	(Myers et al. 2005)	
Hotspots of recombination	Coverage	(Myers et al. 2005)	
Hotspot of recombination motif (CCNCCNTNNCCNC)(Myers et al. 2008)	Density	Genome-wide screen	
Evolutionary breakpoint regions	Coverage	(Larkin et al. 2009)	[0, (Ruiz-Herrera et al. 2006)]
Homologous synteny blocks	Coverage	(Larkin et al. 2009)	

[a] Type of value for each predictor. Coverage = the percentage of particular fragile or non-fragile region that overlaps with predictor. Assigned value = the value of a predictor in an interval. If there is no predictor interval that overlaps with query interval, query interval is marked as "NA". If there is more than one predictor interval overlapping with query interval, assigned value will be calculated as a weighted average based on proportions of overlap with each predictor interval. Average = the average value of that predictor value within that query interval. Density = the count of a feature normalized by the length of the interval. Distance = the distance measured from the region terminus closest to the telomere or centromere.

[b] Prediction column composed of a symbol and a literature reference in square brackets. Symbols: '+' -- predictor is enriched in fragile sites (positive predictor); '-' -- predictor is enriched in non-fragile sites (negative predictor); '0' -- non-significant; '+/-' -- unknown.

Table S4. Standardized coefficients and VIFs (variance inflation factors) in alternative multiple logistic regression models contrasting autosomal aCFSs with NFRs. Numbers in cells are standardized coefficients. Sign of coefficient indicates the relationship between a predictor and fragility. VIFs are in parentheses.

Predictor	Standardized coefficient (with VIF[a] in parentheses)							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
G-band coverage	- 6.78 (4.24)	- 4.63 (1.44)	- 4.75 (1.45)	- 4.36 (1.23)	- 5.28 (2.72)	- 6.05 (3.93)	- 5.48 (3.86)	- 5.59 (3.97)
Twist (DNA flexibility index)					2.17 (3.43)	2.08 (3.51)	1.24 (3.29)	1.69 (3.10)
Low complexity A/T-rich region coverage	3.04 (5.44)[b]							
H3K4me1 site coverage		- 1.05 (2.08)						
CpG island coverage *(log)*			- 1.65 (2.15)					
Transcription start site density *(log)*				- 1.38 (2.35)				
Distance to the centromere	1.48 (1.72)	0.79 (1.24)	0.80 (1.12)	0.68 (1.12)	0.94 (1.32)	1.29 (1.95)	0.94 (1.34)	1.07 (1.58)
Alu repeat coverage *(log)*	1.17 (2.40)	0.68 (1.75)	0.94 (1.77)	0.90 (2.19)				
Mononucleotide microsatellite coverage (*log*)					1.20 (1.77)			
Mononucleotide microsatellite coverage inside *Alu*s (*log*)						1.29 (1.92)		
Mononucleotide microsatellite coverage outside *Alu*s (*log*)							1.10 (1.85)	
Coverage of AT-containing microsatellites (*log*)								1.09 (1.56)

[a] VIF (variance inflation factor)

[b] VIF(variance inflation factor) higher than 5

Table S5. Multiple logistic regression model comparing aCFSs with NFRs using autosomes and X chromosomes

Predictor	Standardized coefficient	VIF[b]	P-value	Relative contribution
G-band coverage	- 5.4208	1.835	1.20×10^{-7}	88.72
Twist (DNA flexibility)	2.0559	3.207	2.74×10^{-3}	7.04
Distance to the centromere	0.9185	1.268	1.18×10^{-2}	3.89
Alu repeat coverage *(log)*[a]	1.3688	2.348	3.67×10^{-2}	2.81
Pseudo multiple R-squared				77.16

[a] Not significant after Bonferroni correction for multiple testing

[b] VIF (variance inflation factor)

Table S6. Alternative breakage frequency multiple regressions using various combinations of genomic predictors. Numbers in cells are relative contributions (partial R-squared). Predictors' *P*-values are in parentheses. Red color indicates positive predictor. Blue color indicates negative predictor.

Predictor	Relative Contribution (*P*-value)			
	Model 1	Model 2	Model 3	Model 4
Evolutionary breakpoint region coverage	26.45 (8.5e-6)	25.22 (1.3e-5)	29.60 (2.0e-6)	31.31 (1.3e-6)
Distance to the centromere	21.92 (6.5e-5)	21.30 (7.5e-5)	24.51 (2.1e-5)	29.38 (3.1e-6)
CpG islands coverage *(log)*				14.01 (2.1e-3)
Transcription start site density *(log)*	13.37 (2.3e-3)			
H3K4me1 site coverage		2.42 (0.206)[a]		
Twist			4.23 (0.095)[a]	
Coverage of low complexity A/T rich regions				2.55 (0.20)[a]
Multiple R-squared	43.37	36.20	41.28	48.20
Adjusted R-squared	40.75	33.30	38.57	44.92

[a] Not significant

Table S7. Standardized coefficients and VIFs (variance inflation factors) **in alternative breakage frequency multiple regression models.** Numbers in cells are standardized coefficients. Sign of coefficient indicates correlation with the level of fragility. VIFs are given in parentheses.

Predictor	Standardized coefficient (with VIF[a] in parentheses)			
	Model 1	Model 2	Model 3	Model 4
Evolutionary breakpoint region coverage	0.50 (1.01)	0.51 (1.01)	0.55 (1.01)	0.54 (1.02)
Distance to the centromere	0.44 (1.01)	0.47 (1.04)	0.49 (1.04)	0.53 (1.06)
CpG islands coverage *(log)*				- 0.44 (1.89)
Transcription start site density *(log)*	- 0.33 (1.00)			
H3K4me1 site coverage		- 0.14 (1.04)		
Twist			0.18 (1.03)	
Coverage of low complexity A/T rich regions				- 0.17 (1.85)

[a] VIF (variance inflation factor)

Table S8. List of 12 autosomal cloned aCFSs and their clonally and cytogenetically defined coordinates (hg18).

FRA	Cytogenetic level			Cloned level			Overlap
	Chromosome	Start	Stop	Chromosome	Start	Stop	
1E	chr1	99400000	102000000	chr1	97660567	98160567	No
2C	chr2	17000000	19100000	chr2	14852340	15598790	Yes
2G	chr2	169500000	182700000	chr2	169182739	170182739	Yes
2H	chr2	182700000	189100000	chr2	182954118	191779297	Yes
3B	chr3	58500000	63700000	chr3	59600000	63800000	Yes
4F	chr4	88200000	99100000	chr4	90421845	97461845	Yes
6E	chr6	160900000	164400000	chr6	159949579	165849579	Yes
6F	chr6	104800000	113900000	chr6	111500000	112600000	Yes
7B	chr7	34000	327567	chr7	34000	327567	Yes
	chr7	477567	7200000	chr7	477567	12358621	
7E	chr7	90900000	92600000	chr7	79193285	84693285	No
7G	chr7	114400000	117200000	chr7	110393285	116893285	Yes
7H	chr7	130100000	132400000	chr7	130053285	130493285	Yes
7I	chr7	147500000	153901567	chr7	143672652	145959490	
	chr7	154001567	154737899				No
	chr7	154817899	158821424				
9E	chr9	113900000	116700000	chr9	107960266	117460267	Yes
11F	chr11	85300000	87366026	chr11	83931848	87366026	Yes
	chr11	87378026	87900000	chr11	87378026	91537002	
11G	chr11	115400000	120700000	chr11	113119646	117662914	Yes
13A	chr13	32900000	34700000	chr13	34400000	35200000	Yes
16D	chr16	78200000	80500000	chr16	76700000	77500000	No

Gaps in assembly were removed to ensure accuracy in coverage and density calculations. Therefore, some aCFSs are separated into several intervals.

Table S9. Predicted probability of 18 aCFSs to be fragile sites, as evaluated using their cytogenetically defined, clonally defined or intersected (between cytogenetically and clonally defined) coordinates.

aCFS	Predicted probability ratio to be aCFS/NFR	
	Cytogenetic level	Intersected coordinates
FRA 1E	0.99759	No overlap
FRA 2C	0.99639	0.99443
FRA 2G	0.98989	0.99953
FRA 2H	0.00459	0.00481
FRA 3B	0.98803	0.99110
FRA 4F	0.00418	0.01161
FRA 6E	0.00114	0.00114
FRA 6F	0.99836	0.99774
FRA 7B	0.22044	0.22044
FRA 7E	0.99766	No overlap
FRA 7G	0.99886	0.99884
FRA 7H	0.99470	0.99909
FRA 7I	0.89350	No overlap
FRA 9E	0.99417	0.99417
FRA 11F	0.99601	0.99601
FRA 11G	0.97247	0.96104
FRA 13A	0.99408	0.99919
FRA 16D	0.99034	No overlap

Table S10. Validation of multiple logistic regression models in mouse fragile sites.

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
G-band coverage	-	-	-	-	-	-	-	-
Distance to the centromere	+	+	+	+	+	+	+	+
CpG island coverage *(log)*		-		-				-
Twist	+		+					
Low complexity A/T-rich region coverage						+	+	
Mononucleotide microsatellite coverage		+	+			+		
B1 coverage *(log)*	+			+	+		+	
Coverage of newly evolved SINEs								+
Number of fragile sites predicted correctly (out of 24)	19	18	17	19	19	19	19	19
Percentage of sites predicted correctly	79	75	71	79	79	79	79	79

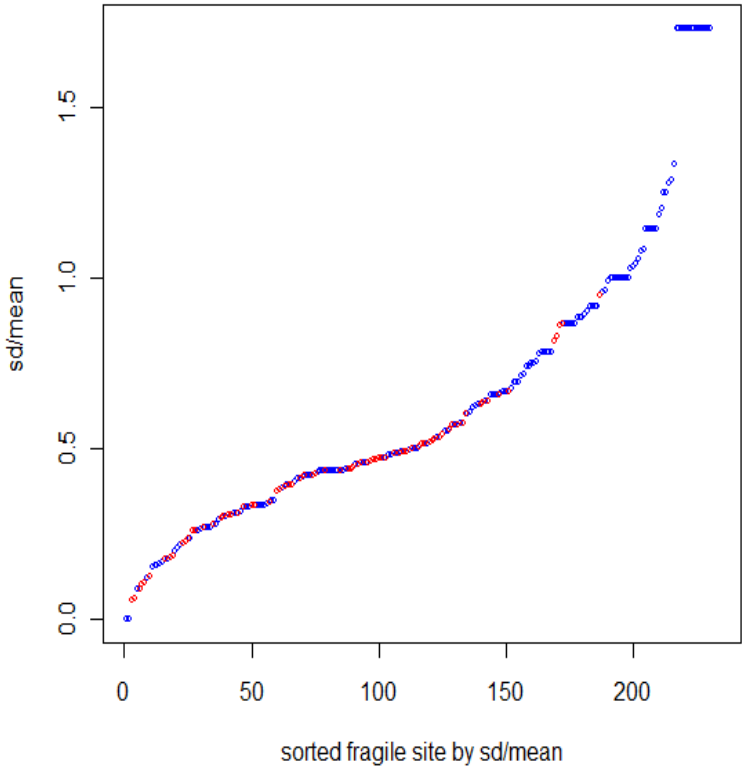
"+"and "-" indicate positive and negative role of predictor of fragility, respectively

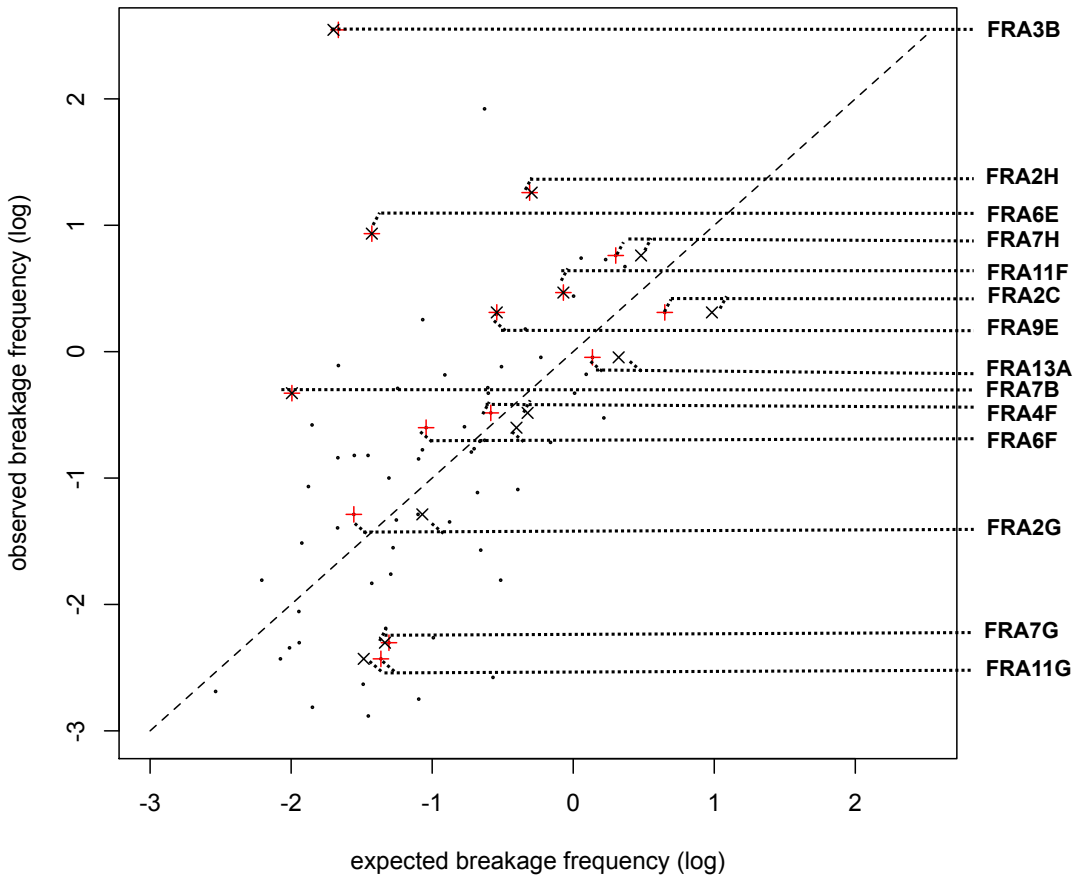
SUPPLEMENTAL FIGURE LEGENDS

Figure S1. 233 genomic regions in Mrasek et al. (2010) ranked by coefficient of variation of breakage frequency among three individuals. The 76 APH-induced CFSs are in red, rare fragile regions and previously unreported fragile regions are in blue.

Figure S2. Scatter plot of expected (according to the optimal multiple regression model) and observed breakage frequency of all 73 autosomal aCFSs mapped at cytogenetic level and super-imposed to 14 aCFSs mapped clonally. Dots are all 73 aCFSs mapped at cytogenetic level. Black crosses are 14 aCFSs modeled according to intersection of their cytogenetic and clonally defined coordinates. Red pluses are the same 14 aCFSs modeled according to their cytogenetic coordinates.

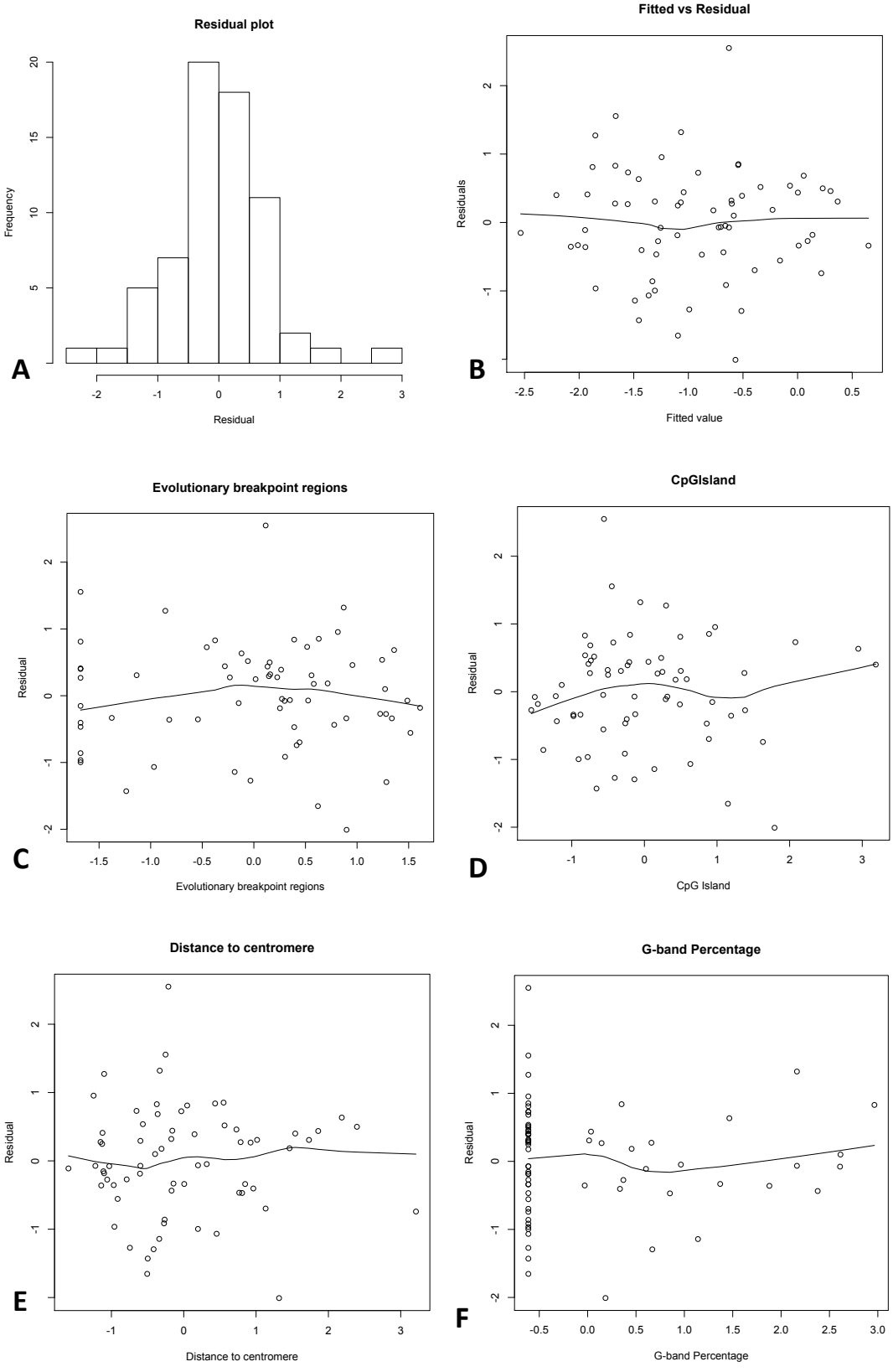
Figure S3. Regression diagnostics. (A) Histogram of residuals has normal distribution. (B) Plot between fitted value and residuals. Variances are constant. Plots between evolutionary breakpoint regions coverage (C), CpG islands coverage (D), distance to the centromere (E), and G-banding (F), and their respective residuals. Each predictor has a linear relationship with its residuals. Local polynomial regression fitting lines were drawn to show tendencies of the data. Sixty-six percent of data were used to draw local fitting at each data point.


sd/mean
0.0
0.5
1.0
1.5
0
50
100
150
200
sorted fragile site by sd/mean



X intersection of cytogenetically and clonally defined aCFSs

\+ cytogenetically defined aCFSs


Residual plot
Frequency
Residual
A
Fitted vs Residual
Residuals
Fitted value
B
Evolutionary breakpoint regions
Residual
Evolutionary breakpoint regions
C
CpGIsland
Residual
CpG Island
D
Distance to centromere
Residual
Distance to centromere
E
G-band Percentage
Residual
G-band Percentage
F

REFERENCES FOR THE SUPPLEMENTAL MATERIAL`

Abajian C. Sputnik - DNA microsatellite repeat search utility. http://espressosoftware.com/sputnik/ (Accessed July 12, 2011).

Arlt MF, and Glover TW. 2010. Inhibition of topoisomerase I prevents chromosome breakage at common fragile sites. *DNA Repair (Amst.)* **9**: 678–689.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, and Zhao K. 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**: 823–837.

Cadoret J-C, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, and Prioleau M-N. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proceedings of the National Academy of Sciences* **105**: 15837 –15842.

Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfovsky N, Luke BT, Bacolla A, Collins JR, and Stephens RM. 2010. Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Research* **39**: D383–91.

Chen C-L, Duquenne L, Audit B, Guilbaud G, Rappailles A, Baker A, Huvet M, d' Aubenton-Carafa Y, Hyrien O, Arneodo A, et al. 2011. Replication-Associated Mutational Asymmetry in the Human Genome. *Molecular Biology and Evolution* **28**: 2327 –2337.

Elder FF, and Robinson TJ. 1989. Rodent common fragile sites: are they conserved? Evidence from mouse and rat. *Chromosoma* **97**: 459–464.

Furey TS, and Haussler D. 2003. Integration of the cytogenetic map with the draft human genome sequence. *Human Molecular Genetics* **12**: 1037 –1044.

Griffiths-Jones S, Saini HK, van Dongen S, and Enright AJ. 2007. miRBase: tools for microRNA genomics. *Nucleic Acids Research* **36**: D154–D158.

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**: 948–951.

Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, and Stamatoyannopoulos JA. 2009. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences* **107**: 139–144.

Helmrich A, Stout-Weider K, Hermann K, Schrock E, and Heiden T. 2006. Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. *Genome Res* **16**: 1222–1230.

Jiang Y, Lucas I, Young DJ, Davis EM, Karrison T, Rest JS, and Le Beau MM. 2009. Common fragile sites are characterized by histone hypoacetylation. *Hum Mol Genet* **18**: 4501–4512.

Karnani N, Taylor CM, Malhotra A, and Dutta A. 2010. Genomic Study of Replication Initiation in Human Chromosomes Reveals the Influence of Transcription Regulation and Chromatin Structure on Origin Selection. *Molecular Biology of the Cell* **21**: 393 –404.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**: 51–54.

Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, and Makova KD. 2010. What Is a Microsatellite: A Computational and Experimental Definition Based upon Repeat Mutational Behavior at A/T and GT/AC Repeats. *Genome Biology and Evolution* **2**: 620–35.

Larkin DM, Pape G, Donthu R, Auvil L, Welge M, and Lewin HA. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Research* **19**: 770–777.

Mishmar D, Mandel-Gutfreund Y, Margalit H, Rahat A, and Kerem B. 1999. Common fragile sites: G-band characteristics within an R-band. *Am J Hum Genet* **64**: 908–910.

Mishmar D, Rahat A, Scherer SW, Nyakatura G, Hinzmann B, Kohwi Y, Mandel-Gutfroind Y, Lee JR, Drescher B, Sas DE, et al. 1998. Molecular characterization of a common fragile site (FRA7H) on human chromosome 7 by the cloning of a simian virus 40 integration site. *Proc. Natl. Acad. Sci. U.S.A* **95**: 8141–8146.

Myers S, Bottolo L, Freeman C, McVean G, and Donnelly P. 2005. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science* **310**: 321 – 324.

Myers S, Freeman C, Auton A, Donnelly P, and McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* **40**: 1124–1129.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, et al. 2009. The UCSC Genome Browser database: update 2010. *Nucleic Acids Research* **38**: D613–D619.

Ruiz-Herrera A, Castresana J, and Robinson TJ. 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol* **7**: R115–R115.

Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, and Gilbert DM. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**: 761–770.

Tsantoulis PK, Kotsinas A, Sfikakis PP, Evangelou K, Sideridou M, Levy B, Mo L, Kittas C, Wu X-R, Papavassiliou AG, et al. 2008. Oncogene-induced replication stress preferentially targets common fragile sites in preneoplastic lesions. A genome-wide study. *Oncogene* **27**: 3256–3264.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh T-Y, Peng W, Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40**: 897–903.

Weddington N, Stuy A, Hiratani I, Ryba T, Yokochi T, and Gilbert DM. 2008. ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics* **9**: 530.

Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, Debernardi S, Mott R, Dunham I, and Carter NP. 2004. Replication timing of the human genome. *Hum. Mol. Genet* **13**: 191–202.

Yang S, Smit AF, Schwartz S, Chiaromonte F, Roskin KM, Haussler D, Miller W, and Hardison RC. 2004. Patterns of Insertions and Their Covariation With Substitutions in the Rat, Mouse, and Human Genomes. *Genome Research* **14**: 517 –527.