Oncogene (2008) 27, 3256–3264

www.nature.com/onc

ORIGINAL ARTICLE

Oncogene-induced replication stress preferentially targets common fragile sites in preneoplastic lesions. A genome-wide study

PK Tsantoulis[1], A Kotsinas[1], PP Sfikakis[2], K Evangelou[1], M Sideridou[1], B Levy[3], L Mo[4], C Kittas[1], X-R Wu[4], AG Papavassiliou[5] and VG Gorgoulis[1]

[1]*Department of Histology and Embryology, Molecular Carcinogenesis Group, School of Medicine, University of Athens, Athens, Greece;* [2]*First Department of Propedeutic and Internal Medicine, School of Medicine, University of Athens, Athens, Greece;* [3]*Clinical Cytogenetics Laboratory, College of Physicians and Surgeons of Columbia University, New York, USA;* [4]*Department of Urology, New York University School of Medicine, New York, USA and* [5]*Department of Biochemistry, School of Medicine, University of Athens, Athens, Greece*

Common fragile sites (CFSs) are regions of the genome prone to breakage by replication inhibitors (extrinsic replication stress). Recently, we and others observed that oncogene-induced replication stress (RS) induces DNA damage from the earliest stages of cancer. Our aim was to perform a genome-wide analysis in precancerous and cancerous experimental models to examine whether allelic imbalance occurs within CFSs. Subsequently, CFSs sequence characteristics were assessed. We used a growth-factor-induced human skin hyperplasia and a *H-ras*-induced mouse hyperplastic urothelium as preneoplastic models, along with an inducible U2OS-*CDT1*$^{\text{Tet-ON}}$ cancer cell line model, all bearing established oncogene-induced RS stimuli. Human DNA was analysed with Affymetrix SNP microarrays, while mouse DNA was analysed with Nimblegen array CGH. We studied 56 aphidicolin-type CFSs and 1914 regions of control, nonfragile DNA. Our theoretical *in silico* analysis spanned 2.16 billion nonoverlapping bases on human chromosomes 1–22. Our results provide direct experimental evidence indicating that genomic alterations were more common within CFSs in epidermal and urothelial preneoplastic lesions as well as in cancer. CFSs were on average less flexible than nonfragile regions, contained more guanine–cytosine (GC) and Alu sequences. Importantly, regions with loss-of-heterozygosity were also less flexible and had a higher Alu percentage.
Oncogene (2008) **27,** 3256–3264; doi:10.1038/sj.onc.1210989; published online 17 December 2007

Keywords: fragile sites; DNA damage; cancer; nucleic acid repetitive sequences; preneoplastic lesions; replication stress

Correspondence: Dr VG Gorgoulis, Antaiou 53 Street, Lamprini, Ano Patissia, Athens GR-11146, Greece.
E-mail: histoclub@ath.forthnet.gr and
Professor AG Papavassiliou, Department of Biochemistry, School of Medicine, University of Athens, Mikras Asias 75, Goudi, Athens GR-11527, Greece.
E-mail: papavas@med.uoa.gr
Received 7 September 2007; revised 30 October 2007; accepted 14 November 2007; published online 17 December 2007

Introduction

Genomic instability is a prominent feature of most, if not all, cancer types. It is most frequently manifested as chromosomal instability, including gross chromosomal alterations and changes in ploidy. In a smaller subset of cancer types, particularly hereditary nonpolyposis colon cancer, genomic instability is observed at the nucleotide level with nucleotide insertions, deletions and substitutions due to mismatch repair gene defects (Lengauer *et al*., 1998). Through the alteration of key genes, genomic instability enables the acquisition of malignant traits and fuels cancer progression (Hanahan and Weinberg, 2000).

We and others (Bartkova *et al*., 2005; Gorgoulis *et al*., 2005) have reported the presence of genomic instability in preneoplastic lesions. Specifically, we detected double strand breaks and an associated DNA damage response in lesions pathologically defined as hyperplastic or dysplastic, which are the earliest pathological stages in cancer development. The sequence of these events was triggered by oncogenic stimuli, and was collectively defined as oncogene-induced replication stress (RS) (Bartkova *et al*., 2005, 2006; Gorgoulis *et al*., 2005; DiMicco *et al*., 2006). We were intrigued by the frequent and preferential involvement of specific regions of the genome known as common fragile sites (CFSs). In particular, the CFSs *FRA3B* (*3p14.2*), *FRA9E* (*9q32*) and *FRA11C* (*11p15.1*) were commonly affected (Bartkova *et al*., 2005; Gorgoulis *et al*., 2005).

CFSs are regions of the genome prone to breakage under conditions of RS (Glover, 2006). This definition is based on the effects of replication inhibitors, typically aphidicolin, under cell culture conditions. Although extremely helpful, this methodological procedure does not take into account intrinsic stressogenic stimuli, such as oncogene-induced RS, that could recapitulate drug-induced RS acting as a putative source for genomic aberrations. In addition, even though we can assume that a considerable overlap between the traditionally termed CFSs and oncogene-induced RS vulnerable regions should exist, this does not have to be identical and the degree of their relation remains to be determined.

CFSs are frequently affected in cancer (Arlt *et al*., 2006; and references therein), but their study as a whole class, concurrently in the entire genome, within precancerous context, has not been carried out so far. Such an approach would help clarify whether their involvement in cancer is a late by-product of extensive cellular deregulation or an early phenomenon with putative significance in cancer development (Glover, 2006).

The human genome is heterogeneous even in a very large scale (Lander *et al*., 2001) and contains regions with varying sequence characteristics. Furthermore, the structural properties of the DNA molecule depend on the underlying nucleotide sequence (Sarai *et al*., 1989; Olson *et al*., 1998; Packer *et al*., 2000). For example, adenine–thymine pairs are generally considered more flexible than guanine–cytosine pairs. Clearly, it is reasonable to assume that the vulnerability of fragile sites (FS) can be attributed, at least in part, to their nucleotide sequence. Rare FS, which are of limited interest in cancer studies, are known to contain di- or tri-nucleotide repeats that can cause spontaneous breaks in affected individuals. The sequence characteristics of CFSs are much more complex. It appears that some CFSs contain flexibility peaks (Mishmar *et al*., 1998, 1999; Mimori *et al*., 1999; Zlotorynski *et al*., 2003), adenine–thymine runs (Zlotorynski *et al*., 2003) or an increased percentage of repetitive DNA (Morelli *et al*., 2002). Despite the usefulness of these sporadic observations, which generally concern the subset of the molecularly defined CFSs (Helmrich *et al*., 2007), a broad survey of the sequence characteristics of CFSs is missing.

Our current work attempts to address the open question of whether CFSs, as a general class, are frequently affected from the earliest stages of carcinogenesis. We hope that this will serve as a starting point for the determination of areas, which are vulnerable to oncogene-induced RS. In addition, for the first time since the completion of the human genome sequence, we strive to detect in a comprehensive, statistically rigorous and objective manner potentially interesting sequence characteristics of CFSs.

Results

In vivo *and* in vitro *analyses*

FS are preferentially targeted in preneoplastic lesions. Initially a growth-factor-induced human skin hyperplasia (GF-hSHy) model derived from human skin implanted on immunodeficient mice, and treated weekly with injections of adenoviral-vectors bearing growth-factor genes was investigated (Berking *et al*., 2004; Gorgoulis *et al*., 2005). An array single nucleotide polymorphism (SNP) analysis showed that loss-of-heterozygosity (LOH) was more common in FS (13.5% of SNP markers) than in non-FS (9.4%, logistic regression, $P=0.04$) (Figure 1a, Table 1). It should be noted that there was extensive LOH in chromosomes 6, 9, 21 and 22, leading us to hypothesize that the genotype might have been affected by the inclusion of adjacent dysplastic areas, known to emerge during the progression of this model (Berking *et al*., 2004).. The heterogeneity of malignant tumours has been studied before (Hanahan and Weinberg, 2000; and references therein), but little is known about precancerous lesions. Our GF-hSHy sample was found to be genetically heterogeneous, as shown in Figures 1b and c. Two different areas of the same GF-hSHy sample were evaluated for the markers *D3S1289* and *D3S1300* that map close to *FRA3B*. The *D3S1300* marker was homozygous and therefore noninformative, while the *D3S1289* marker showed distinct patterns in each of the two areas. At the same time, the normal sample was clearly homogeneous. Therefore, the region containing *FRA3B* is differentially affected in distant cell populations of the same GF-hSHy, implying that even a very early lesion can be genetically heterogeneous.

Similarly, an array CGH analysis of DNA from a *H-ras*-induced mouse hyperplastic urothelium (*H-ras*-mHyU), expressing the active *H-ras* gene fused with the promoter of the mouse uroplakin-II (*UPK2*) gene (Wu, 2005), showed predominant involvement of FS. We have listed all segments in Table 2, together with an estimate of FS sensitivity from Helmrich *et al*. (2006). In such an early lesion it is hard to predict what subset of the cell population, possibly very small, is really affected with these alterations. As demonstrated above, it is possible that different areas of a hyperplastic lesion may have different genomic alterations, making their detection rather challenging. It is even harder to predict which of these alterations will prove to be significant during the progression of the lesion. However, it is clear that the majority (14 of 16, 88%) of probable copy number alterations are found in areas of medium or high CFS sensitivity (Table 2).

FS sensitivity persists in cancer. Next, we examined the human U2OS osteosarcoma cell line transfected with an inducibly controlled h*CDT1*, a replication licensing factor known to cause RS via re-replication (Vaziri *et al*., 2003).

The SNP microarray analysis of the U2OS cell line before and after the induction of the RS-inducing gene h*CDT1* demonstrates the preferential involvement of CFSs in cancer (Table 3). Out of 22 alterations, 7 were located in pericentromeric bands and 6 of the remaining 15 (40%) alterations corresponded to known CFSs. Two more FS are indicated in parentheses because they did not meet our study criteria (*FRA19B* is a rare FS and *FRA17B* is located at *17q23.1*, adjacent to *17q23.2*).

Whole-genome in silico *analysis*

Our data set was, on average, highly representative of the human genome in comparison with the initial sequence analysis of the human genome project (Lander *et al*., 2001). Mean guanine–cytosine (GC) content was 40.80 versus 41.00%, the average percentage of interspersed repeats was 45.42 versus 44.83%, the average percentage of LINE-elements was 20.37 versus 20.42%

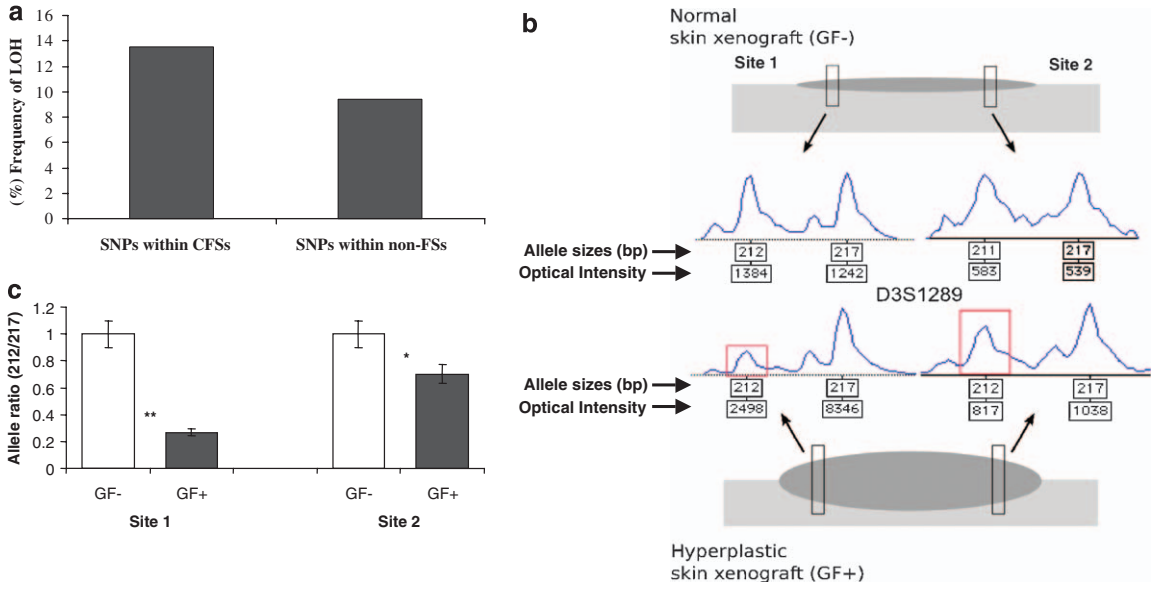

Figure 1 Alterations within single nucleotide polymorphism (SNPs) located in common fragile sites (CFSs) and non-FSs (**a**), and clonal evolution in different areas of the growth-factor-induced human skin hyperplasia (GF-hSHy) model (**b** and **c**), respectively. (**a**) Bar-chart showing the significantly increased loss-of-heterozygosity (LOH) frequency at SNPs located in CFSs in comparison to LOH at SNPs within non-FSs in the GF-hSHy model ($P=0.04$). (**b**) Representative outputs from automated sequence analysis show allele sizes (212 and 217 base pairs) and corresponding optical densities (in arbitrary units) for each allele, respectively. Note the heterogeneous allelic imbalance (AI) in the different sites ('peaks' marked in red boxes) of the growth-induced sample, as compared to corresponding normal counterparts (noninduced sample). (**c**) Histogram depicting mean (±s.d.) of allele ratios (212 or 217 bp alleles), in sites 1 and 2 respectively, of the GF− (white) and GF+ (gray) treated hSHy xenografts. (*$P<0.05$; **$P<0.001$).

and the average percentage of SINE-elements was 13.56 versus 13.10%. Small differences may be explained by the fact that we used a newer version of RepeatMasker. We calculated the average flexibility of our data (2.16 Gb), including fragile and nonfragile regions, to be 10.786 (95% confidence interval, 10.776–10.796).

The investigated FS were defined as follows. From a total of 113 FS recorded in NCBI Gene until May 2006, we discarded 28 as rare FS, 12 as nonaphidicolin type CFSs and 17 FS mapped to bands longer than 10 Mb or mapped without accuracy (see Materials and methods). Our final data set of 56 aphidicolin CFSs with a total length of 245 Mb (Supplementary Table 1), provided good coverage of chromosomes 1–22 (approximately 9%). CFSs in the mouse genome were obtained from a recent study (Helmrich *et al*., 2006).

Fragile sites are inflexible, GC and Alu rich. FS were less flexible than non-FS (10.742 versus 10.787, Wilcoxon's test, $P=0.045$). This may be attributed to the increased presence of guanine and cytosine (41.69 versus 40.79% for non-FS, Wilcoxon's test, $P=0.030$). The percentage of LINE1 sequences seemed to be lower in FS (15.67%) than in non-FS (16.82%), but the difference was not statistically significant (Wilcoxon's test, $P=0.055$). However, the percentage of Alu elements was higher in FS (11.66 versus 10.65%, Wilcoxon's test, $P=0.009$). The results are summarized in Figure 2, where dashed lines represent the estimated genome mean. Regions rich in Alu elements were also less flexible (Figure 3, Kendall's $\tau=-0.539$, $P<0.001$) and rich in GC (Kendall's $\tau=0.516$, $P<0.001$) across our whole dataset, leading us to conclude that these characteristics generally coexist.

Regions with loss-of-heterozygosity are inflexible and Alu rich. Using the location of SNP markers on the Affymetrix 10 k version 2 array we calculated for every marker the sequence properties of a region spanning 50 kb in the 5′ and 50 kb in the 3′ direction. Markers with LOH were located in less flexible 100 kb segments (mean flexibility 10.819 versus 10.848, Wilcoxon's test $P=0.049$) with a higher Alu content (9.61 versus 8.00%, Wilcoxon's test $P=0.010$). The GC content was similar around markers with and without LOH (40.2 and 39.9%, Wilcoxon's test $P=0.138$). It should be noted that marker position was nonrandom to optimize array performance, which is why the numbers quoted above differ from the genome averages given in the preceding section. Evidently, even when considering short genomic regions (100 kb), the sequence properties that were found in bands containing FS are seen to be predictive of vulnerability to cancer-related, oncogene-induced RS.

Smaller chromosomes are also inflexible, GC and Alu rich. Interestingly, smaller chromosomes appear to share common characteristics with CFSs. As shown in Figure 4, chromosome number was inversely associated with mean flexibility (Kendall's $\tau=-0.429$, $P=0.005$). In addition, the average GC content increases in smaller chromosomes (Kendall's $\tau=0.146$, $P<0.001$) together with Alu content (Kendall's $\tau=0.138$, $P<0.001$).

Table 1 Common fragile sites (CFSs) status in the growth factor-induced human skin hyperplasia (GF-hSHy) model as deduced from informative single nucleotide polymorphism (SNPs) within their region

Fragile site	*Region*	*LOH*
FRA1B	*1p32*	No
FRA1C	*1p31.2*	—
FRA1D	*1p22*	Yes
FRA1E	*1p21.2*	No
FRA1G	*1q25.1*	No
FRA1I	*1q44*	Yes
FRA2C	*2p24.2*	No
FRA2D	*2p16.2*	No
FRA2E	*2p13*	No
FRA2F	*2q21.3*	No
FRA2H	*2q32.1*	No
FRA2J	*2q37.3*	—
FRA3A	*3p24.2*	—
FRA3B	*3p14.2*	Yes
FRA3C	*3q27*	No
FRA4A	*4p16.1*	No
FRA4C	*4q31.1*	No
FRA5C	*5q31.1*	Yes
FRA5D	*5q15*	—
FRA6B	*6p25.1*	Yes
FRA6C	*6p22.2*	Yes
FRA6E	*6q26*	Yes
FRA6F	*6q21*	Yes
FRA6G	*6q15*	Yes
FRA7B	*7p22*	—
FRA7C	*7p14.2*	—
FRA7D	*7p13*	Yes
FRA7E	*7q21.2*	No
FRA7F	*7q22*	No
FRA7G	*7q31.2*	No
FRA7H	*7q32.3*	No
FRA8B	*8q22.1*	No
FRA8C	*8q24.1*	—
FRA8D	*8q24.3*	No
FRA9D	*9q22.1*	Yes
FRA9E	*9q32*	—
FRA10D	*10q22.1*	No
FRA10E	*10q25.2*	No
FRA10F	*10q26.1*	—
FRA11C	*11p15.1*	No
FRA11D	*11p14.2*	No
FRA11E	*11p13*	No
FRA11F	*11q14.2*	No
FRA11G	*11q23.3*	No
FRA13A	*13q13.2*	No
FRA13C	*13q21.2*	No
FRA13D	*13q32*	No
FRA14C	*14q24.1*	No
FRA15A	*15q22*	No
FRA16C	*16q22.1*	No
FRA16D	*16q23.2*	—
FRA17B	*17q23.1*	—
FRA18A	*18q12.2*	No
FRA18B	*18q21.3*	—
FRA20B	*20p12.2*	No
FRA22B	*22q12.2*	Yes

Abbreviation: LOH, loss of heterozygosity. —, noninformative SNPs.

Table 2 Areas harbouring copy number alterations in hyperplastic mouse urothelial (*H-ras-mHyU*) samples

	From	*To*	*Probes*	*Fragile*
Sample A				
17qE1.1	61186086	61692038	80	No
17qD	59230751	59880571	102	Medium
13qD2.3	120572095	120612152	8	Medium
2qA3	27772860	27784226	3	Medium
Sample B				
15qB3.3	46074936	49967834	553	High
3qF3	110733779	114477017	493	Medium
14qE4	119669213	119818839	25	Medium
16qC3	78999581	84427605	815	Medium
16qC1.3	59782980	61410366	237	No
16qC2	66970386	69558114	370	Medium
14qE4	118910165	118996478	16	High
10qD2	116535429	116610129	14	Medium
14qE5	120595066	120612339	4	Medium
4qC6	103644013	103661071	4	Medium
4qC1	68574700	68597909	5	Medium
5qB3	40721995	40733380	3	Medium

Fragile site sensitivity was based on Helmrich *et al*. (2006). The total number of probes over the whole genome for every sample and every replicate was approximately 385 000, as explained in the Materials and methods section. The number of probes per segment in the fourth column is associated not only with the length of the segment (a median inter-probe distance of 6 kb), but also with the statistical certainty with which it was selected. The details of the circular binary segmentation algorithm are beyond the scope of this article and are fully described in Olshen *et al*. (2004).

Table 3 Alterations in the U2OS-*hCDT1*$^{Tet\text{-}ON}$ cell line at 30 and 60 days of forced h*CDT1* expression

Region	*Lesion*		*Days*		*Fragile*	*Centro*
			30	*60*		
1q21.1	37 kb	Gain	Yes	Yes		
2p23.2	83 kb	Gain	Yes	Yes		
3p11.1	230 kb	Gain	No	Yes		Yes
3q13.11	166 kb	Gain	No	Yes		
5p15.33	136 kb	Gain	No	Yes		
7p22.3	161 kb	Gain	Yes	Yes	*FRA7B*	
7p22.3	223 kb	Gain	No	Yes	*FRA7B*	
7q21.12	110 kb	Gain	No	Yes	*FRA7E*	
7q22.1	72 kb	Gain	Yes	Yes	*FRA7F*	
9p11.1	202 kb	Gain	Yes	Yes		Yes
9p11.2	1960 kb	Gain	Yes	Yes		Yes
9p11.2	28 kb	Gain	Yes	Yes		Yes
9p11.2	459 kb	Gain	Yes	Yes		Yes
9p12	1400 kb	Del	No	Yes		
10q11.22	31 kb	Gain	Yes	No	*FRA10G*	Yes
12p13.33	175 kb	Gain	Yes	No		
13q34	175 kb	Gain	No	Yes		
14q23.1	73 kb	Gain	Yes	Yes	*FRA14B*	
14q23.1	497 kb	Gain	No	Yes	*FRA14B*	
16p11.2	476 kb	Del	No	Yes		Yes
17q23.2	162 kb	Gain	Yes	Yes	*(FRA17B)*	
19p13.33	165 kb	Gain	Yes	Yes	*(FRA19B)*	

Aphidicolin-type common fragile sites and centromeric regions are listed.

Discussion

Our study provides direct experimental evidence supporting our preliminary hypothesis (Bartkova *et al*., 2005; Gorgoulis *et al*., 2005) that CFSs, as a general class, are more frequently affected than non-FS during the earliest stages of carcinogenesis. This is most likely a consequence of structural vulnerability to oncogene-driven RS (Bartkova *et al*., 2005, 2006; Gorgoulis *et al*., 2005; DiMicco *et al*., 2006).

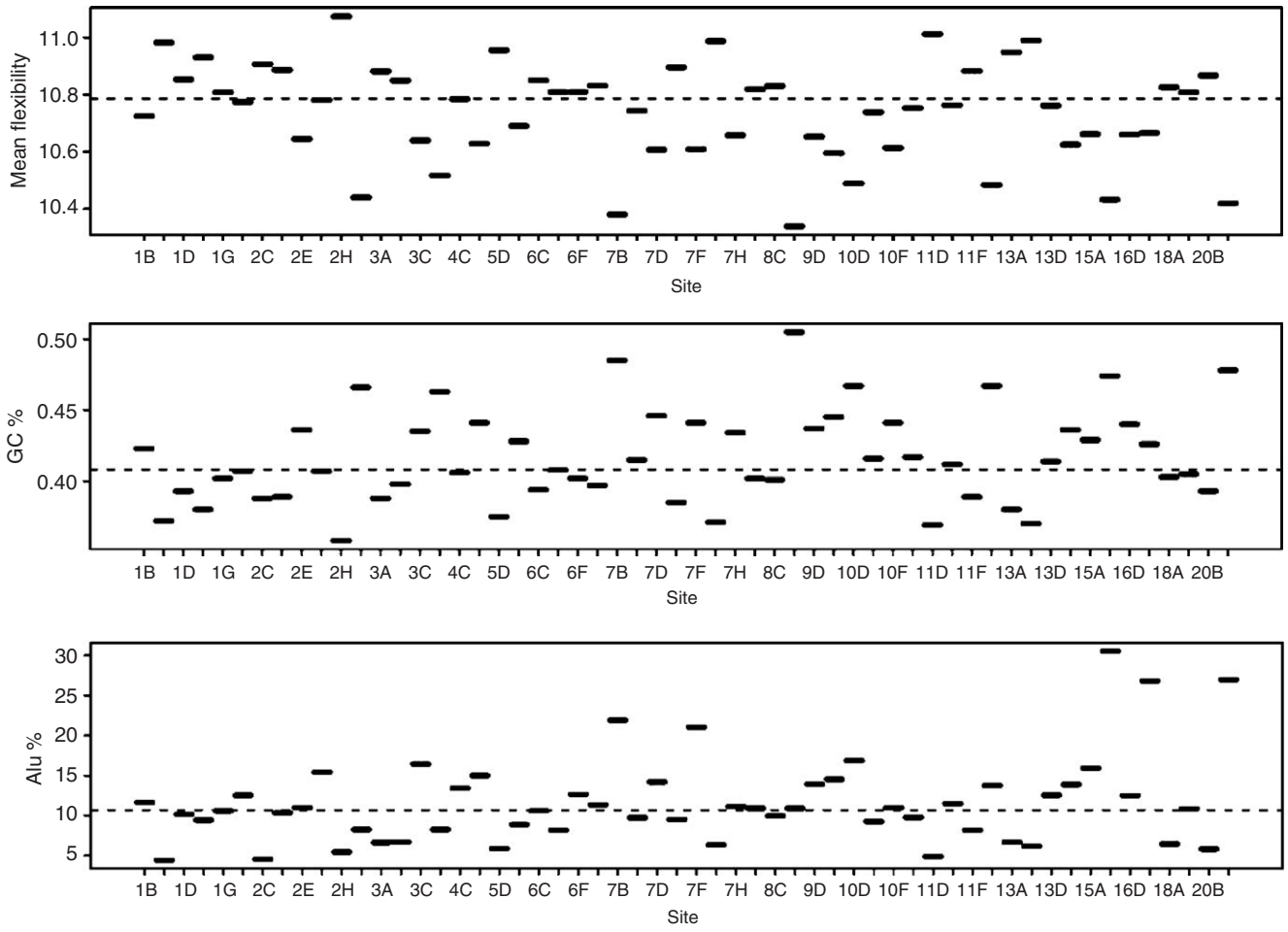

Figure 2 Summary of fragile site (FS) sequence characteristics. Horizontal lines show the mean value for each one of the 56 aphidicolin common fragile sites (CFSs) listed in Supplementary Table 1.

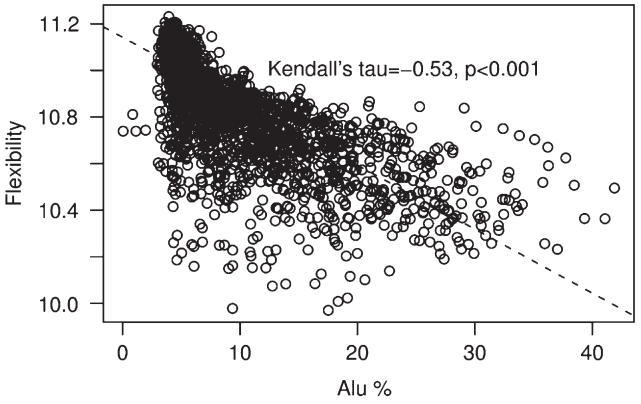

Figure 3 Flexibility is inversely related with Alu content.

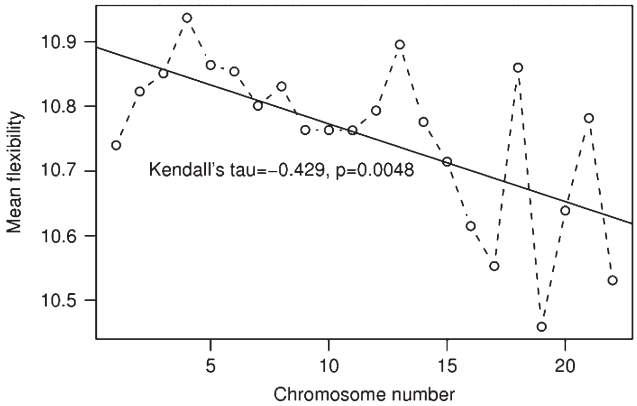

Figure 4 Small chromosomes are on average less flexible than large chromosomes.

These results raise a new question regarding the role of CFSs in the early stages of carcinogenesis. It is tempting to speculate that critical cancer-related genes may reside close to CFSs, but with a few notable exceptions, like *FHIT* (Zanesi *et al*., 2001) and *WWOX* (Aqeilan *et al*., 2007), it would be unrealistic and evolutionary unlikely that all CFSs harbour important genes. CFS lesions may have far-reaching consequences

if they can initiate a telomeric breakage-fusion-breakage cycle (Ciullo *et al*., 2002; Murnane, 2006). In such a scenario, a vicious cycle of progressively worsening chromosome instability could be created. On the other hand, it is interesting to note that their early involvement may serve to amplify the cellular response to a potential genomic threat, as an early sensor of excessive RS. In that case, breaks at these sites could function as an 'alarm system' inducing a prompt protective response under normal conditions, when the DNA damage response mechanism is intact.

Our whole-genome analysis of CFS sequence properties has a number of novel and potentially important implications. The unexpected observation that CFSs are on average rich in GC and Alu and relatively inflexible raises several interesting questions. However, it should be stressed that this does not necessarily invalidate previous reports. There are considerable differences between CFSs (Figure 2) and even within a given site there can be important local trends (for example, AT-runs) or other relatively short features that cannot be visualized in a large-scale study. For example, *FRA2H* is seen to be more flexible than average, relatively AT rich and Alu poor. On the other hand, site *FRA7B* has a low mean flexibility, almost 50% GC content and more than 20% Alu content, far greater than the genome average. Such differences will have to be reconciled through a more detailed mapping of FS, possibly with the aid of high-resolution microarray studies, and with the exploration of new sequence-dependent properties along with chromatin structure, not covered here. For example, origin of replication sites are spread throughout the genome in locations that have not yet been satisfactorily mapped (Toledo *et al*., 2000) or characterized. It appears, though, that they are densely packed in areas of high flexibility (Toledo *et al*., 2000). Notably, in a very recent study focusing on FS on chromosome 7, similar findings with regard to DNA-helix flexibility were drawn (Helmrich *et al*., 2007).

Despite the differences between sites, our results are coherent. It is well known that Alu elements tend to favour GC- and gene-rich regions (Lander *et al*., 2001). It has also been proposed that Alu element mobilization may mediate recombination events (Batzer and Deininger, 2002; and references therein), which are known to be relatively frequent within the CFSs (Glover and Stein, 1988; Hirsch, 1991) and the shorter chromosome arms (Figure 16 in Lander *et al*., 2001). Furthermore, Alu elements are known to contain a central adenine-rich region and a polyA tail (Batzer and Deininger, 2002), which could explain the detection of flexibility peaks and AT runs in some CFSs (Mishmar *et al*., 1998; Mimori *et al*., 1999), even in an otherwise AT-poor/GC-rich inflexible region. Most importantly though, the FS characteristics, that is, low mean-flexibility and high Alu content, were also found in the regions surrounding SNP markers with LOH in our GF-hSHy xenograft sample.

At this point, it is important to highlight the limitations of our approach, which should serve as directions for future research. First of all, the mapping of FS was not as precise as the methods we used for genome analysis. This is an unavoidable consequence of the fact that many CFSs have been approximately mapped in the past with less accurate methods. We hope that more high-resolution data will gradually become available, especially using whole-genome methods like microarray CGH, so that a similar analysis can be repeated. In addition, we have to acknowledge the fact that many of the known FS were submitted by authors who used different cell lines and experimental setups. A certain degree of biological heterogeneity within our list of CFSs (Supplementary Table 1) must be anticipated and is also evident by the variation between sites in Figure 2. Therefore, the applicability of our results to various organisms and cancer types will have to be further verified and expanded for each case separately. Nevertheless, research has shown that FS are remarkably conserved between species (Helmrich *et al*., 2006) and many have been repeatedly found affected in different cancer types (Glover, 2006; and references therein). Furthermore, our results were consistent across three different experimental systems, strengthening our conviction that they are representative of the general behaviour of FS.

In conclusion, due to the variable presence of potentially influential sequence features throughout the genome, it is important to realize that the concept of fragility probably represents a continuum, where some regions are more fragile than others. In that sense, the division in 'fragile' and 'nonfragile' regions could be substituted in the future with a more accurate estimate of vulnerability to stress based on specific sequence properties. Such an approach will allow a better understanding of the early events that precede the emergence of the malignant phenotype.

Materials and methods

Human and mouse precancerous lesions

Growth-factor-induced human skin hyperplasia model Human genomic DNA was extracted from a model of hyperplasia in which human skin xenografts were grown under the influence of exogenous growth factors (Berking *et al*., 2004; Gorgoulis *et al*., 2005). Briefly, human skin was implanted on immunodeficient mice and treated weekly with injections of adenoviral vectors bearing growth-factor genes (*bFGF*, *SCGF*, endothelin-3). Treated xenografts rapidly evolved from normal to skin hyperplasia (Figure 5a) and finally to dysplasia, while untreated xenografts retained their normal appearance and served as controls.

H-ras-*induced mouse hyperplastic urothelium* Mouse genomic DNA was extracted from the hyperplastic urothelium of two *H-ras*-transgenic mice (Wu, 2005; Mo *et al*., 2007). Briefly, fertilized eggs were microinjected with active *H-ras* gene fused with the promoter of the mouse uroplakin-II (*UPK2*) gene. The expression of mutant *H-ras* was confined to the urothelium, as verified by reverse transcription (RT)–PCR. This urothelium had the pathological appearance of simple hyperplasia, with 6–7 layers of normally oriented and well-differentiated urothelial cells (Figure 5b). Paired tail-DNAs from the same mice were used as controls. Of note, the role of

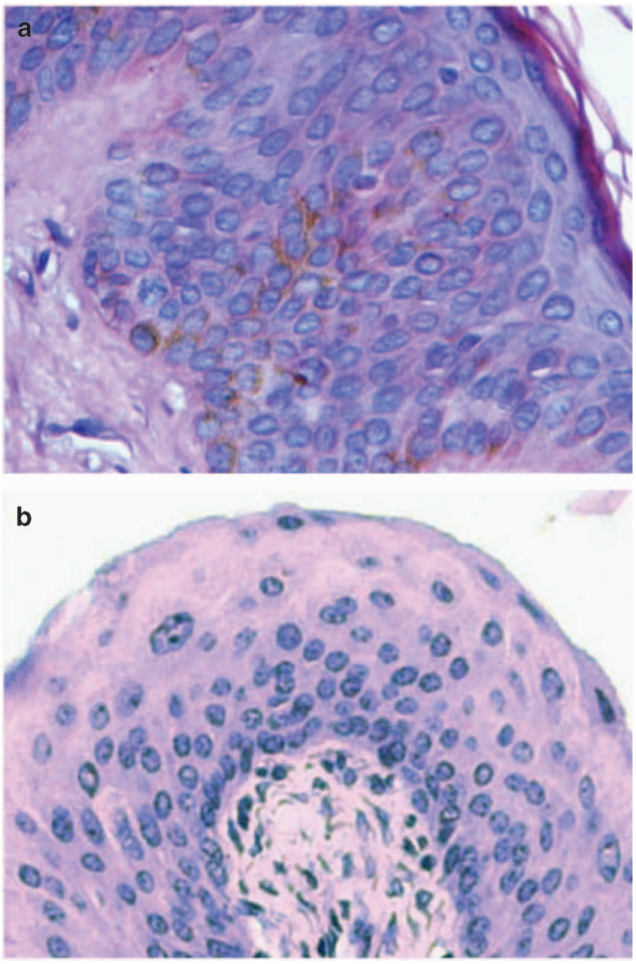

Figure 5 Human skin xenograft (**a**) and mouse urothelium sections (**b**), hematoxylin-eosin, 400 ×.

H-ras as an oncogenic RS factor has been recently established (DiMicco *et al*., 2006).

Both these models have a predictable progression and are free of uncontrollable influences. In contrast with human hyperplastic and dysplastic lesions, which usually develop over the course of several years and have a variable prognosis, these models evolve quickly and can be readily observed and obtained at precise time points.

Cancer cell model

The U2OS osteosarcoma cell line was stably transfected with the *CDT1* gene, producing the U2OS-*CDT1*$^{\text{Tet-ON}}$ cell line. This cell line, derived from the femoral osteosarcoma of a young patient, has an intact p53, a hypo-phosphorylated pRb and a deleted p16. Therefore, U2OS-*CDT1*$^{\text{Tet-ON}}$ serves as a model of cancer cells proliferating under conditions of experimentally induced RS. Briefly, U2OS cells were transfected with the pTRE2hyg-h*CDT1* vector using the Lipofectamine 2000 kit. The induction and expression of h*CDT1* with tetracycline was verified with RT–PCR and immunoblotting. Noninduced cells were used as controls.

Microarray analysis

Human genomic DNA from GF-hSHy xenografts was analysed with the Affymetrix 10 k version 2 SNP array, with a median distance of 113 kb between consecutive markers. Human genomic DNA from the U2OS-*CDT1*$^{\text{Tet-ON}}$ cell line was analysed with Affymetrix 100 k SNP array (manuscript in preparation), with a median intermarker distance of 8.5 kb. In both cases, the results were read and analysed with the Affymetrix GType 4.0 software.

Mouse genomic DNA was subjected to array CGH analysis with the Nimblegen MM8_WG_CGH array due to the lack of an equivalent Affymetrix SNP array for the mouse genome. This array contains 385 k probes with a median distance of approximately 6 kb. Two replicates were run for each sample pair and the mean normalized log 2 ratio was used for the subsequent segmentation analysis. Segmentation analysis was performed inside the R environment for statistics with the DNAcopy (Olshen *et al*., 2004) method (circular binary segmentation) using the default settings.

Allelic imbalance analysis

Allelic imbalance analysis (AI) included established microsatellite markers for chromosome *3p* (*D3S1300*, *D3S1289*), known to span the *FRA3B* region (Wistuba *et al*., 2000). In a paired normal and GF-hSHy xenograft, previously characterized for the presence of AI (Gorgoulis *et al*., 2005) at *FRA3B* with the above markers, sections were obtained from two distal parts, respectively, as shown in Figure 1. Genomic DNA extraction, PCR and automated sequence analysis (Applied Biosystems, Warrington, UK) were performed as previously described (Gorgoulis *et al*., 2005). Scoring and calculated differences in allele ratios between control and growth-factor-treated xenografts were repeated twice.

Whole-genome flexibility and sequence analysis

We used the assembled chromosome sequences of the human genome project release 36.1 obtained as FASTA text files and the corresponding annotation (MapView) for all analyses. Our study encompasses human chromosomes 1–22 and mouse chromosomes 1–19.

We obtained the list of all FS recorded in NCBI Gene until May 2006 and mapped them to their respective cytogenetic bands using the ideogram file of the human genome project annotation. For example, *FRA3B* was mapped to chromosome 3, bases 58500000–63700000, that is, band *3p14.2*. Rare FS and those responding to agents other than aphidicolin were removed (Figure 6) to have a relatively homogeneous set. We also removed CFSs, not localized with enough precision, like *FRA4D* (*4p15*, length of approximately 25 Mb) so that the breakpoints could be contained with accuracy. For that purpose, and taking also into consideration the data in the literature showing that CFSs cover large genomic regions up to, and sometimes more, than 10 Mb (Smith *et al*., 2007), which compares favourably with the resolution of traditional cytogenetic methods that were used in many previous studies of fragile sites, we set a 10 Mb limit. The same mapping procedure was applied to all included CFSs, even if a more detailed molecular definition exists in the bibliography to ensure a consistent and objective approach.

Thus, from a total of 113 FS we discarded 28 rare FS, 12 nonaphidicolin type CFSs and 17 FS mapped to bands longer than 10 Mb to arrive at our final data set of 56 aphidicolin CFSs with a total length of 245 Mb (Supplementary Table 1), providing good coverage of chromosomes 1–22 (approximately 9%). The expression of CFSs in the mouse genome was taken from a recent study (Helmrich *et al*., 2006).

The set of 'normal' sites was defined by removing all fragile and centromeric bands (Supplementary Table 1) from chromosomes 1–22. The resulting set contained 1914 regions

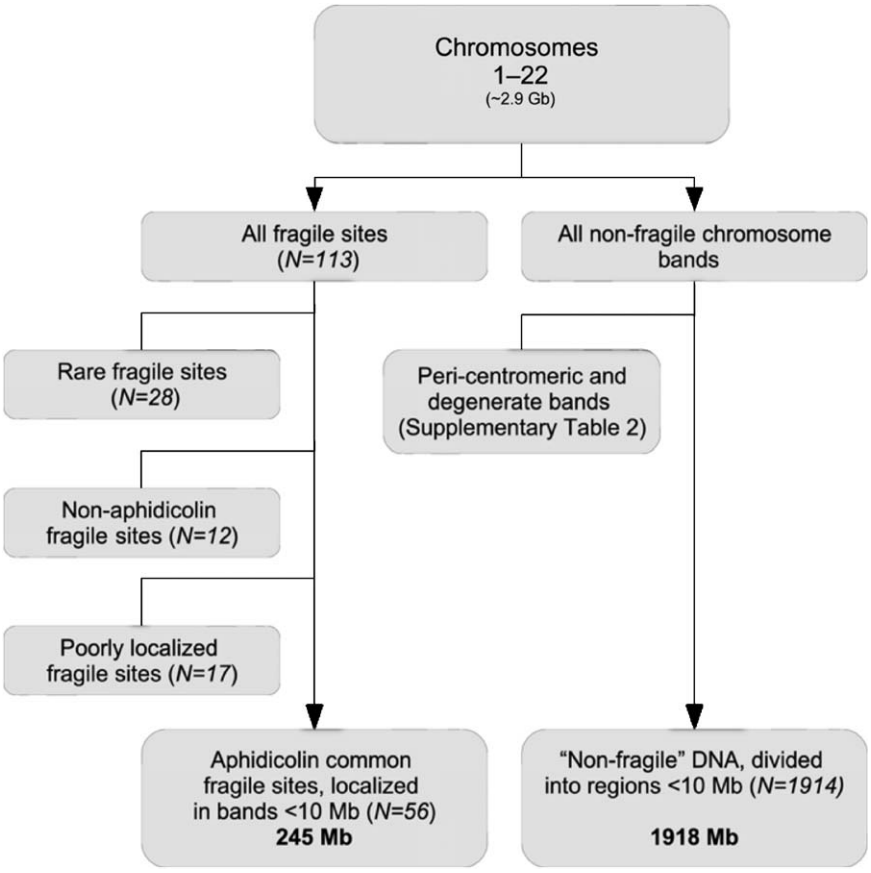

Figure 6 Dividing chromosomes 1–22 into fragile and nonfragile (reference) sets.

with a total 1918 Mb length. Overall, we analysed 2.16 billion bases or approximately 70% of the sequence of human chromosomes 1–22.

Flexibility is usually studied with FlexStab software (Mishmar *et al*., 1998). We implemented a similar algorithm using identical nucleotide-pair flexibility values (Sarai *et al*., 1989) to automate the whole-genome scanning process. Briefly, a region is divided into consecutive, potentially overlapping windows of user-defined length, as in the original FlexStab program. The flexibility mean is calculated over each window and the window is classified as 'high' or 'low' flexibility in comparison with the genome mean and s.d. For each region, the mean and s.d. of window flexibilities is recorded along with the percentage of highly flexible and nonflexible windows and the percentage of guanine and cytosine. For example, using a window length of 200 bases and a window shift of 1 base, region *3p14.2* had 39.8% GC, 10.849 mean window flexibility and a 0.544 s.d. The C-source code and accompanying manual are available upon request.

Repetitive DNA was detected with RepeatMasker version 3.1.3 using RepBase version January 2006 (http://www.repeatmasker.org), in 'sensitive' mode. We recorded the overall percentage of total interspersed repeats, LINE (LINE1 and LINE2), SINE and ALU for each region.

Statistics

Statistical analysis was carried out with the R language (Ihaka and Gentleman, 1996) using the appropriate parametric and nonparametric tests. Results were considered significant when $P<0.05$.

Abbreviations

AI, allelic imbalance; CFSs, common fragile sites; CGH, comparative genomic hybridization; FS, fragile sites; GF-hSHy, growth-factor-induced human skin hyperplasia; *H-ras*-mHyU, *H-ras*-induced mouse hyperplastic urothelium; RS, replication stress; SNP, single nucleotide polymorphism.

Acknowledgements

We thank Professor Meenhard Herlyn for providing the human skin xenograft material. This work was co-funded by European Social Fund and National Resources-(EPEAEK-II) PYTHAGORAS II (SARG 7952) and National Kapodistrian University of Athens (SARG 56/90/6599, 70/4/9174).

Conflicts of Interest

None.

References

Aqeilan RI, Trapasso F, Hussain S, Costinean S, Marshall D, Pekarsky Y *et al.* (2007). Targeted deletion of Wwox reveals a tumor suppressor function. *Proc Natl Acad Sci USA* **104**: 3949–3954.

Arlt MF, Durkin SG, Ragland RL, Glover TW. (2006). Common fragile sites as targets for chromosome rearrangements. *DNA Repair (Amst)* **5**: 1126–1135.

Bartkova J, Hoesí Z, Koed K, Krämer A, Tort F, Zieger K *et al*. (2005). DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature* **434**: 864–870.

Bartkova J, Rezaei N, Liontos M, Karakaidos P, Kletsas D, Issaeva N *et al*. (2006). Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. *Nature* **444**: 633–637.

Batzer MA, Deininger PL. (2002). Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370–379.

Berking C, Takemoto R, Satyamoorthy K, Shirakawa T, Eskandarpour M, Hansson J *et al*. (2004). Induction of melanoma phenotypes in human skin by growth factors and ultraviolet B. *Cancer Res* **64**: 807–811.

Ciullo M, Debily MA, Rozier L, Autiero M, Billault A, Mayau V *et al*. (2002). Initiation of the breakage-fusion-bridge mechanism through common fragile site activation in human breast cancer cells: the model of PIP gene duplication from a break at FRA7I. *Hum Mol Genet* **11**: 2887–2894.

DiMicco R, Fumagalli M, Cicalese A, Piccinin S, Gasparini P, Luise C *et al*. (2006). Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. *Nature* **444**: 638–642.

Glover TW. (2006). Common fragile sites. *Cancer Lett* **232**: 4–12.

Glover TW, Stein CK. (1988). Chromosome breakage and recombination at fragile sites. *Am J Hum Genet* **43**: 265–273.

Gorgoulis VG, Vassiliou LVF, Karakaidos P, Zacharatos P, Kotsinas A, Liloglou T *et al*. (2005). Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature* **434**: 907–913.

Hanahan D, Weinberg RA. (2000). The hallmarks of cancer. *Cell* **100**: 57–70.

Helmrich A, Stout-Weider K, Hermann K, Schrock E, 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.

Helmrich A, Stout-Weider K, Matthaei A, Hermann K, Heiden T, Schrock E. (2007). Identification of the human/mouse syntenic common fragile site FRA7K/Fra12C1-Relation of FRA7K and other human common fragile sites on chromosome 7 to evolutionary breakpoints. *Int J Cancer* **120**: 48–54.

Hirsch B. (1991). Sister chromatid exchanges are preferentially induced at expressed and nonexpressed common fragile sites. *Hum Genet* **87**: 302–306.

Ihaka R, Gentleman R. (1996). R: a language for data analysis and graphics. *J Comp Graph Stat* **5**: 299–314.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lengauer C, Kinzler KW, Vogelstein B. (1998). Genetic instabilities in human cancers. *Nature* **396**: 643–649.

Mimori K, Druck T, Inoue H, Alder H, Berk L, Mori M *et al*. (1999). Cancer-specific chromosome alterations in the constitutive fragile region FRA3B. *Proc Natl Acad Sci USA* **96**: 7456–7461.

Mishmar D, Mandel-Gutfreund Y, Margalit H, Rahat A, 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 S, Nyakatura G, Hinzmann B, Kohwi Y *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 USA* **95**: 8141–8146.

Mo L, Zheng X, Huang HY, Shapiro E, Lepor H, Cordon-Cardo C *et al*. (2007). Hyperactivation of Ha-ras oncogene, but not Ink4a/Arf deficiency, triggers bladder tumorigenesis. *J Clin Invest* **117**: 314–325.

Morelli C, Karayianni E, Magnanini C, Mungall AJ, Thorland E, Negrini M *et al*. (2002). Cloning and characterization of the common fragile site FRA6F harboring a replicative senescence gene and frequently deleted in human tumors. *Oncogene* **21**: 7266–7276.

Murnane JP. (2006). Telomeres and chromosome instability. *DNA Repair* **5**: 1082–1092.

Olshen AB, Venkatraman ES, Lucito R, Wigler M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**: 557–572.

Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci USA* **95**: 11163–11168.

Packer MJ, Dauncey MP, Hunter CA. (2000). Sequence-dependent DNA structure: dinucleotide conformational maps. *J Mol Biol* **295**: 71–83.

Sarai A, Mazur J, Nussinov R, Jernigan RL. (1989). Sequence dependence of DNA conformational flexibility. *Biochemistry* **28**: 7842–7849.

Smith DI, McAvoy S, Zhu Y, Perez DS. (2007). Large common fragile site genes and cancer. *Semin Cancer Biol* **17**: 31–41.

Toledo F, Coquelle A, Svetlova E, Debatisse M. (2000). Enhanced flexibility and aphidicolin-induced DNA breaks near mammalian replication origins: implications for replicon mapping and chromosome fragility. *Nucleic Acids Res* **28**: 4805–4813.

Vaziri C, Saxena S, Jeon Y, Lee C, Murata K, Machida Y *et al*. (2003). A p53-dependent checkpoint pathway prevents rereplication. *Mol Cell* **11**: 997–1008.

Wistuba II, Behrens C, Virmani AK, Mele G, Milchgrub S, Girard L *et al*. (2000). High resolution chromosome 3p allelotyping of human lung cancer and preneoplastic/preinvasive bronchial epithelium reveals multiple, discontinuous sites of 3p allele loss and three regions of frequent breakpoints. *Cancer Res* **60**: 1949–1960.

Wu XR. (2005). Urothelial tumorigenesis: a tale of divergent pathways. *Nat Rev Cancer* **5**: 713–725.

Zanesi N, Fidanza V, Fong LY, Mancini R, Druck T, Valtieri M *et al*. (2001). The tumor spectrum in FHIT-deficient mice. *Proc Natl Acad Sci USA* **98**: 10250–10255.

Zlotorynski E, Rahat A, Skaug J, Ben-Porat N, Ozeri E, Hershberg R *et al*. (2003). Molecular basis for expression of common and rare fragile sites. *Mol Cell Biol* **23**: 7143–7151.

Supplementary Information accompanies the paper on the Oncogene website (http://www.nature.com/onc).