A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome?

Arkarachai Fungtammasan,[1,2,3] Erin Walsh,[3,4] Francesca Chiaromonte,[3,5,7] Kristin A. Eckert,[3,6,7] and Kateryna D. Makova[2,3,7]

[1] *The Integrative Biosciences Graduate Program, Bioinformatics and Genomics Option, Pennsylvania State University, University Park, Pennsylvania 16802, USA;* [2] *Department of Biology, Pennsylvania State University, University Park, Pennsylvania, 16802, USA;* [3] *Center for Medical Genomics, Pennsylvania State University, University Park, Pennsylvania 16802, USA;* [4] *Cellular and Molecular Biology Graduate Program, Pennsylvania State University College of Medicine, Hershey, Pennsylvania 17033, USA;* [5] *Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, USA;* [6] *Department of Pathology, Jake Gittlen Cancer Research Foundation, Pennsylvania State University College of Medicine, Hershey, Pennsylvania 17033, USA*

Chromosomal common fragile sites (CFSs) are unstable genomic regions that break under replication stress and are involved in structural variation. They frequently are sites of chromosomal rearrangements in cancer and of viral integration. However, CFSs are undercharacterized at the molecular level and thus difficult to predict computationally. Newly available genome-wide profiling studies provide us with an unprecedented opportunity to associate CFSs with features of their local genomic contexts. Here, we contrasted the genomic landscape of cytogenetically defined aphidicolin-induced CFSs (aCFSs) to that of nonfragile sites, using multiple logistic regression. We also analyzed aCFS breakage frequencies as a function of their genomic landscape, using standard multiple regression. We show that local genomic features are effective predictors both of regions harboring aCFSs (explaining ~81% of the deviance in logistic regression models) and of aCFS breakage frequencies (explaining ~45% of the variance in standard regression models). In our optimal models (based on a combination of biological interpretability and high R-squared value), aCFSs are predominantly located in G-negative chromosomal bands and away from centromeres, are enriched in *Alu* repeats, and have high DNA flexibility. In alternative models, CpG island density, H3K4me1 coverage, and mononucleotide microsatellite coverage are significant predictors. Also, aCFSs have high fragility when colocated with evolutionarily conserved chromosomal breakpoints. Our models are predictive of the fragility of aCFSs mapped at a higher resolution. Importantly, the genomic features we identified here as significant predictors of fragility allow us to draw valuable inferences on the molecular mechanisms underlying aCFSs.

[Supplemental material is available for this article.]

Chromosomal fragile sites are loci that are prone to gaps or breaks within metaphase chromosomes. Common fragile sites (CFSs) are observed in all humans and constitute a component of normal chromosome structure (Durkin and Glover 2007; Freudenreich 2007). Such regions have been documented in many other mammalian species, including chimpanzee, gorilla, orangutan (Smeets and van de Klundert 1990), baboon (Soulie and De Grouchy 1981), cat (Stone et al. 1993), dog (Stone et al. 1991a,b), mouse (Elder and Robinson 1989), and rat (Robinson and Elder 1987). CFSs have an important role in chromosome instability; they are associated with sister chromatid exchange hotspots (Glover and Stein 1987), viral integration sites (Bester et al. 2006; Dall et al. 2008), and sites of deletion, amplification, and translocation in various cancers (Arlt et al. 2006; Durkin et al. 2008; Burrow et al. 2011). Recently, CFSs have been shown to be preferred sites of structural variation in stem cells (Hussein et al. 2011). Clearly, CFSs play an important role in genome dynamics and are medically relevant.

[7] **These authors contributed equally to this work.**
Corresponding authors: chiaro@stat.psu.edu; kae4@psu.edu; kdm16@psu.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.134395.111.

A subset of CFSs can be specifically induced by cellular treatment with aphidicolin (APH), a DNA polymerase inhibitor. Several models have been proposed to explain the underlying mechanisms responsible for preferential DNA strand breakage at APH-induced CFSs (hereafter called aCFSs) (Durkin and Glover 2007). Replication delay or inherent DNA replication difficulties are believed to underlie the initiation of aCFS expression (Arlt et al. 2006). Some aCFS regions undergo delayed or prolonged DNA elongation in S phase, and cells can enter G2 phase with only 50% of some aCFS loci completely replicated (Palakodeti et al. 2004; Pelliccia et al. 2008). DNA breakage within aCFSs is thought to be a consequence of failing to complete replication and/or resolving the arrested forks prior to the end of telophase and chromosome segregation (Chan et al. 2009). Specific DNA sequences, such as $[A/T]_n$ and $[AT/TA]_n$ repeats, and/or the formation of non-B DNA secondary structures within aCFSs can inhibit replicative DNA polymerases (Shah et al. 2010) and replication fork progression (Zhang and Freudenreich 2007). AT-rich high DNA flexibility regions have been described within some aCFSs, and may affect replication by acting as sinks for the superhelical density generated ahead of the replication fork, hindering efficient topoisomerase activity (Zlotorynski et al. 2003). More recently, a paucity of replication origins, inefficient origin firing, and failure to activate latent origins have all been suggested to play a role in

delayed replication at specific aCFSs (Palakodeti et al. 2010; Letessier et al. 2011; Ozeri-Galai et al. 2011).

Intrinsic DNA sequence features alone are sufficient to induce site-specific chromosomal breakage of aCFSs found at ectopic locations (Ragland et al. 2008). Unlike rare fragile sites that are associated with a single DNA element, many sequence motifs spread throughout the aCFS region may contribute to fragility (Ried et al. 2000; Durkin and Glover 2007; Ragland et al. 2008), making the characterization of aCFSs a computational challenge. Nevertheless, previous analyses of individual aCFS loci demonstrated that these regions may be enriched in several genomic features, such as *Alu* repeats (Tsantoulis et al. 2008), gene-containing regions (Helmrich et al. 2006), histone hypoacetylation (Jiang et al. 2009), highly AT-rich sequences, and high DNA flexibility sequences (Mishmar et al. 1998). Unfortunately, such studies have been mostly inconclusive at the genome-wide level. For example, a large fraction of the fragile site *FRA3B* is enriched in LINE1 elements, but these sequences are poorly represented in *FRA16D;* conversely, *Alu* repeats dominate *FRA16D* (Ried et al. 2000). Similarly, highly AT-rich sequence content was shown to be irrelevant (Helmrich et al. 2006) or even to have the opposite correlation (Tsantoulis et al. 2008) with fragility. The inconsistent results reported previously are best ascribed to two major factors. First, most studies considered only a few aCFSs, so the observed enrichment might not have reflected the global trend of genomic contexts surrounding such sites. Second, the studies analyzed the enrichment of single (and not multiple) genomic features at aCFSs, and compared these with control regions that differed from study to study (Mishmar et al. 1998; Helmrich et al. 2006; Ruiz-Herrera et al. 2006; Tsantoulis et al. 2008). Some of these studies previously used control regions that exhibited low-frequency breakage or were characterized as aCFSs in subsequent studies (Mrasek et al. 2010).

Traditionally, the organization of the human genome was interrogated using a variety of chromosomal banding techniques (Comings 1978). Differential banding patterns reflect variations in chromatin structure and base composition among chromosomal regions and have been correlated with various aspects of genome function (Craig and Bickmore 1993). For example, R bands have a higher gene and CpG island density than G bands, display higher levels of histone acetylation, and are enriched for SINE elements. S phase DNA replication is distinctly bimodal, and R bands correspond to early replicating regions, whereas G bands correspond to late replicating regions of the genome (Holmquist et al. 1982). Several types of fragile sites have been observed to lie more frequently in R bands or near the border of R and G bands (Yunis and Soreng 1984). It has also been suggested that aCFSs are located in G-band-like regions of R bands (Mishmar et al. 1999). T bands, a specific subset of R bands, have been shown to be sites of increased chromosomal rearrangement, both in cancer cells and during chromosomal evolution (Holmquist 1992). Interestingly, aCFSs may be depleted in evolutionarily homologous syntenic regions conserved between mammals and chicken (Ruiz-Herrera et al. 2006).

Current evidence suggests that aCFSs are caused by multiple interplaying genomic factors (Dillon et al. 2010). Fragility may result from several genomic properties characterizing a locus, instead of a single motif or genomic feature (Durkin and Glover 2007). The wealth of genome-wide profiling studies that are now available provides us with an unprecedented opportunity to investigate the underlying causes of chromosomal fragility. A model that considers multiple factors simultaneously is expected to be more biologically realistic and could illuminate how different genomic features interact to contribute to fragility. In addition, aCFSs vary in their breakage frequency, but previous studies have not incorporated this quantitation into their statistical models.

To advance our understanding of the relationship between aCFSs and genomic contexts, we built statistical models to explain the fragility of well-characterized autosomal aCFSs by considering their genomic landscape and contrasting them with nonfragile regions (NFRs) obtained from a genome-wide screening. We focused on CFSs induced by aphidicolin because they have been characterized genome-wide (Mrasek et al. 2010), are the most numerous CFSs, and CFSs induced by other agents might have different breakage mechanisms and characteristics. We used multiple logistic regression to predict the probability of a given region to be either an aCFS or an NFR and multiple linear regression to predict expected breakage frequency. Finally, to evaluate performance, we validated our models using mouse fragile sites.

Results

Defining aCFSs vs. NFRs

The cytogenetic locations of APH-induced CFSs from the genome-wide screening (Mrasek et al. 2010) were converted to genomic coordinates using the UCSC Genome Browser (Rhead et al. 2009). Among the known APH-induced CFSs, we focused on 76 well-characterized sites (Lukusa and Fryns 2008). These 76 aCFSs, which vary in size from 0.7 to 27.8 Mb, were originally defined by cytogenetic analyses (Lukusa and Fryns 2008; Mrasek et al. 2010). Notably, while some of the aCFS size variation may reflect actual biological differences, some might also be explained (especially in the upper size ranges) by the limited resolution of cytogenetic methods used to define the aCFSs (Durkin and Glover 2007). We did not use all 233 breakage regions identified by Mrasek and colleagues (2010) because their set included low-breakage-frequency sites (possibly representing experimental background) and rare fragile sites known to originate by a different molecular mechanism (Durkin and Glover 2007; Lukusa and Fryns 2008). From the 76 well-characterized aCFSs, initially we excluded sites located on sex chromosomes, because sex chromosomes possess very few aCFSs (three on chromosome X and none on chromosome Y) (Fig. 1) and are enriched in repetitive elements (Skaletsky et al. 2003; Ross et al. 2005), potentially biasing our analysis. A subsequent analysis including aCFSs on chromosome X led to similar results (see below). Our final data set, thus, consisted of 73 aCFSs that were distributed quite broadly across human autosomes (but none were located on chromosomes 19 and 21) and covered 490 Mb (or 17.00%) of the autosomal genome (Fig. 1; Supplemental Table S1). Visual inspection of their genome-wide distribution suggests that most autosomal aCFSs are located away from centromeres (Fig. 1).

To define NFRs against which to contrast aCFSs, the set of 233 known fragile regions identified in a genome-wide screen (Mrasek et al. 2010), fragile regions from other studies (Kuwano et al. 1988; Borgaonkar 1994), heterochromatin, and centromeric regions were removed from the human autosomal genome (see Methods for details). After such removal, we defined the leftover 117 fragments as NFRs, ranging in size from 1.4 to 32.9 Mb and covering 915 Mb (or 31.78%) of the autosomal genome (Supplemental Table S2).

Genomic features

To assess the location of aCFSs relative to the global organization of the genome, we initially used G banding, GC content,

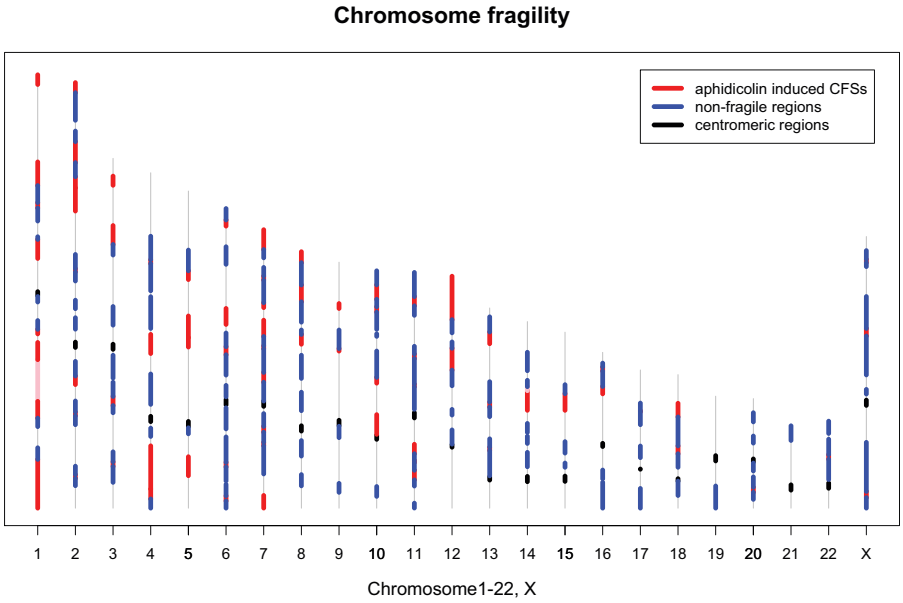

Figure 1. Locations of 76 APH-induced common fragile sites (aCFSs) and 124 nonfragile control regions (NFRs) used in this study. (Red) aCFSs (pink is used to differentiate among three fragile sites on chromosome 1). (Blue) NFRs. Gray regions were excluded from the analysis because they are either rare fragile sites or background breakage regions. (Black regions) Centromeres.

distance to the centromere, and distance to the telomere as potential predictors of fragility. The advent of whole-genome sequencing and other genome-profiling techniques has overlaid this global organization with more specific landscape features. Therefore, we considered a larger set of 54 genomic features, including those for which an association with aCFSs has been suggested in previous studies (Supplemental Table S3). This set included the above features associated with global genome organization, as well as with gene expression and chromatin organization (CpG islands, transcription start sites, H3K4me1 histone modification sites, nuclear lamina binding sites, and microRNA occurrence), DNA replication (replication timing and replication origin density), recombination and mutation (recombination rates and evolutionarily breakpoint regions), and DNA sequence/structure (direct and inverted repeats, triplex motifs, microsatellites, low complexity A/T rich regions, non-B DNA structures, DNA flexibility, and transposable elements). The DNA flexibility parameter (Twist) measures potential local variations in the DNA structure, expressed as fluctuations in the twist angle of two adjacent base pairs (Sarai et al. 1989; Mishmar et al. 1998; Travers 2004). All these predictors were measured in each aCFS and NFR and prescreened based on their pairwise associations with other predictors (a flowchart with our predictor screening pipeline is depicted in Fig. 2A). More specifically, we clustered predictors using their Spearman's rank correlation coefficients (Fox 2002) and selected one "representative" predictor from each tight cluster to ensure that pairwise correlations between any selected predictors did not exceed 0.7 (Fig. 3). In subsequent regression fits, this avoided strong multicollinearities, keeping variance inflation factors (VIFs), which measure linear association among the predictors in a regression, below five (Fox 2002). "Representative" features from each tight cluster were selected based on prior evidence of association with CFSs (from the literature; see Introduction). This prescreening produced a set of 19 predictors (Table 1; Fig. 3) that were used for subsequent model selection in our regressions (Fig. 2B). Genome-wide computational modeling is limited by the resolution/accuracy of the genome annotations that are available for the predictors. Some of this annotation, in turn, is limited by the experimental approaches used to derive the data (e.g., many "genome-wide" sequencing studies exclude repetitive sequences). Because all approaches have inherent error, we considered other genomic features (from Supplemental Table S3) as alternatives at various stages of our analyses, with the logic that true molecular predictors of aCFSs will likely overlap with several genomic features.

Contrasting genomic features between aCFSs and NFRs

In this analysis, we used logistic regression to contrast genomic features between aCFSs and NFRs. Well-established techniques for model selection applied to the 19 prescreened predictors (Table 1) led to a model with biological interpretability and high R-squared value—later called the *optimal logistic model* (importantly, some of the 19 prescreened predictors were subsequently replaced with alternative choices, i.e., predictors highly correlated with them, to produce alternative models—see below). The optimal logistic model captures 81.15% of the null deviance and comprises four highly significant genomic features (Table 2; *P*-values are given in the table). The strongest feature discriminating between aCFSs and NFRs was G band coverage (individual contribution ~88%). This was a negative predictor, indicating that aCFSs are positioned largely outside of G bands (e.g., in G-negative bands). Average twist value, distance to the centromere, and *Alu* repeat coverage were all significant positive predictors (individual contributions ~12%, ~3%, and ~4%, respectively). Distance to the centromere coverage lost its significance after Bonferroni correction for multiple testing. Thus, known aCFSs appear to be located distant from the centromere and preferably in a genomic landscape depleted in G bands, yet characterized by high DNA flexibility and enrichment in *Alu* repeats. Note that the existence of correlations among predictors (despite low VIFs) implies that the sum of the relative contributions of individual predictors does not necessarily add to 100% (the total deviance explained by the model).

We attempted to further investigate the genomic features associated with aCFSs by replacing predictors in the optimal model (Table 2) with predictors that were excluded during prescreening (see above; Fig. 3) but can provide alternative interpretations (Table 3; Supplemental Table S4). For instance, we replaced average Twist value with the coverage of low-complexity A/T-rich regions that have been suggested to associate with aCFS breakage (Dillon et al. 2010). Coverage of low complexity A/T-rich regions was a significant positive predictor ($P = 0.0017$) (model 1 in Table 3), and the resulting model had a higher pseudo R-squared as compared with the optimal model (86.67% vs. 81.15%). However, low complexity A/T-rich regions had high VIFs (>5), reflecting multicollinearity (Supplemental Table S4), and therefore, we consider this

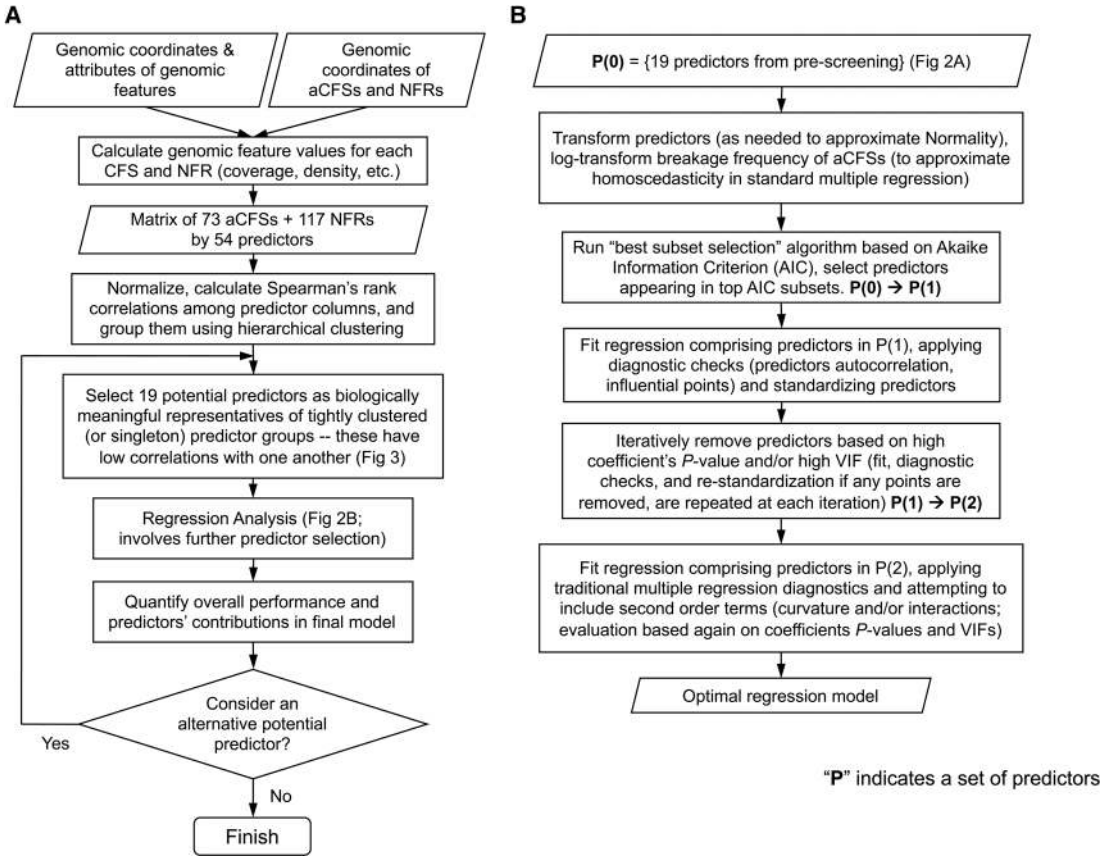

Figure 2. Statistical workflow. (*A*) Potential predictor selection (prescreening). (*B*) Regression analysis (box "Regression Analysis" in *A*).

model to be suboptimal. We also replaced Twist with negatively correlated predictors such as H3K4me1 site coverage (enriched in promoter regions) (Heintzman et al. 2009) or CpG island coverage (models 2 and 3 in Table 3, respectively). All of these predictors were significant ($P = 0.021$, $P = 0.0017$, respectively; for H3K4me1 site coverage, the significance was lost after Bonferroni correction for multiple testing). The resulting models had lower and higher pseudo R-squared (80.95% and 83.65%, respectively) as compared to that for the optimal model (81.15%).

Alu sequences are composed of two long A-rich stretches frequently containing mononucleotide microsatellites (Arcot et al. 1995; Kelkar et al. 2011). Mononucleotide A/T repeats cause pausing of replicative DNA polymerases in vitro (Shah et al. 2010) and, thus, may contribute to replication difficulties within aCFSs. Thus, we replaced *Alu* repeat coverage with mononucleotide microsatellite coverage in our modeling (model 5 in Table 3). Mononucleotide microsatellite coverage was a significant positive predictor ($P = 0.009$), resulting in a model with a lower fit than the optimal model (80.49% vs. 81.15%). We also replaced *Alu* coverage with either the coverage of mononucleotide microsatellites within or outside of *Alu*s (models 6 and 7 in Table 3). Interestingly, the model using mononucleotide microsatellites within *Alu* sequences had a lower fit than the model utilizing mononucleotides outside of *Alu* sequences (79.86% vs. 84.96%) and the optimal model (79.86% vs. 81.15%). Replacing *Alu* coverage with A/T-containing microsatellite coverage (i.e., genome-wide microsatellites with repeats consisting of exclusively A and/or T bases) did not yield a more predictive model (74.98% for model 8 in Table 3). However, note that, in all these models, *Alu* coverage and other predictors used to replace it were significant only prior to Bonferroni correction for multiple testing and were, in fact, the predictors with the second lowest explanatory power (among other predictors). Nevertheless, this analysis suggests that mononucleotide A/T-rich microsatellites, particularly when located within *Alu*s, may substantially contribute to fragility.

Despite evidence that some aCFSs replicate late during the cell cycle (Le Beau et al. 1998; Hellman et al. 2000; Durkin and Glover 2007), replication timing was not a significant predictor of fragility in any of our models. Three independent replication timing data sets were tested (Woodfine et al. 2004; Hansen et al. 2009; Ryba et al. 2010). These data sets were found to be highly correlated (all pairwise Pearson correlation coefficients were above 0.7, data not shown) and therefore appear to be robust. The utilization of replication origins has been recently hypothesized to be altered in aCFSs (Palakodeti et al. 2010; Letessier et al. 2011). However, origin density was not significant in our models, independent of the three replication origin data sets used in our analyses (Cadoret et al. 2008; Karnani et al. 2010; Chen et al. 2011).

Finally, we repeated our analysis including X chromosomes (here, all 76 aCFSs and 124 NFRs were used; chromosome Y was excluded as female cells were analyzed by Mrasek and colleagues [Mrasek et al. 2010]). The results (pseudo R-squared of 83.65%) (Supplemental Table S5) were similar to those obtained for autosomes only (Table 2).

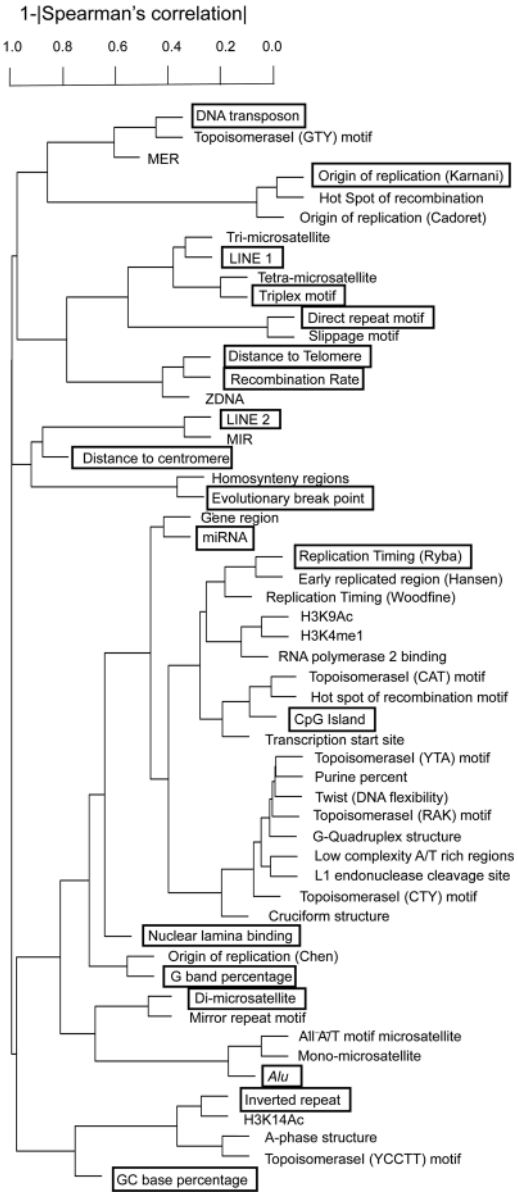

Figure 3. Hierarchical clustering of predictors using their Spearman correlation coefficients computed across all 73 aCFSs+117 NFRs. (*Y*-axis) 1–|correlation coefficient|. The lower predictors merge in the dendrogram, the higher their correlation. Predictors in black boxes were selected as potential predictors to run our regression analysis (which includes further predictor selection steps).

Effect of genomic features on the frequency of aCFS breakage

Across the genome, various aCFSs do not break with the same probability. Mrasek and colleagues (2010) evaluated breakage frequency of aCFSs within 25,000 peripheral blood lymphocyte-derived metaphase spreads and observed that only ~10% of aCFSs detected were fragile at a frequency >1%. We used these data to examine genomic features potentially affecting aCFS breakage frequency, focusing on the 73 APH-induced autosomal aCFSs, and performing a standard multiple regression analysis. Different variable selection approaches starting from the 19 prescreened predictors (Table 1) all led to a model (later called the *optimal standard model)* accounting for ~45% of variation in aCFS breakage frequency and containing four significant predictors (Table 4): distance to the centromere ($P = 0.0002$), G-band coverage ($P = 0.0267$; loses significance after Bonferroni correction for multiple testing), CpG island coverage ($P = 0.000679$), and evolutionary breakpoint region coverage ($P = 0.00000161$). Two of these features (distance to the centromere and G-band coverage) were also identified as discriminators (with the same sign) between aCFSs and NFRs in the optimal logistic regression model above (Table 2), and CpG island coverage was a significant predictor in an alternative logistic model (model 3 in Table 3). Most importantly, evolutionary breakpoint region coverage (individual contribution ~31.2%) emerged as a new positive predictor highly relevant to aCFS breakage frequency. Our optimal standard model suggests that the frequency of breakage of aCFSs appears to increase in a genomic landscape rich in evolutionary breakpoints, distant from the centromere, located in G-negative bands, and depleted in CpG islands.

We next added to or replaced the four significant predictors identified above (Table 4) with predictors excluded during prescreening (Supplemental Table S6). Transcription start site density was found to be a significant ($P = 0.0023$) negative predictor, and the model including it (model 1 in Supplemental Table S6) had explanatory power slightly higher than that for the optimal model (46.41% vs. 44.81%). Coverage of H3K4me1 histone modification sites, Twist, and coverage of low complexity A/T-rich regions (models 2, 3, and 4, respectively, in Supplemental Table S6) were not significant. All these alternative standard regression models had predictors with low VIFs (Supplemental Table S7).

The explanatory power (percentage of variance or deviance explained) for the breakage frequency model (44.81%) was lower than that obtained for the aCFS vs. NFR model (81.15%). There are two possible explanations. First, the breakage frequency data that we used came from one study (Mrasek et al. 2010). Therefore, some portion of the observed variation in breakage frequency across aCFSs may have actually been due to the stochastic nature of fragility in the APH-induced experimental assay. Second, even though the overall correlation in breakage frequency among individuals was high, some fragile sites exhibited a high degree of variability in breakage frequency between individuals (Supplemental Fig. S1).

Validation in cloned aCFSs

Because our models were derived and estimated using aCFSs initially defined by low-resolution cytogenetic methods, it was important to assess their predictive behavior for aCFSs mapped at higher resolution, i.e., with fluorescence probes around the breakage area (Ruiz-Herrera et al. 2006). We were able to find high-resolution genomic coordinates of 18 autosomal cloned aCFS either using BAC accession numbers (Ciullo et al. 2002; Fechter et al. 2007; Reshmi et al. 2007; Bosco et al. 2010; Pelliccia et al. 2010; Blumrich et al. 2011) or data collected by Ruiz-Herrera and colleagues (Ruiz-Herrera et al. 2006). Among these, 14 *(FRA2C, FRA2G, FRA2H, FRA3B, FRA4F, FRA6E, FRA6F, FRA7B, FRA7G, FRA7H, FRA9E, FRA11F, FRA11G,* and *FRA13A)* overlapped with aCFSs cytogenetically defined by Mrasek and colleagues (Mrasek et al. 2010), while high-resolution coordinates for the remaining four *(FRA1E, FRA7E, FRA7I,* and *FRA16D)* did not overlap with their cytogenetic coordinates (Supplemental Table S8). This was not completely unexpected, as both cytogenetic banding and fluorescent mapping methods have inherent technical limitations that contribute to variation among coordinates. Because of this uncertainty, these four aCFSs were excluded from further analysis.

Table 1. The 19 genomic features (after prescreening) used as potential predictors in regression analyses

Predictors	Type of data[a]	Data source	Previous studies[b]
Global genome organization			
G-bands	Coverage	(Furey and Haussler 2003)	(– [Mishmar et al. 1999])
GC content	Average	Genome-wide screen	(– [Mishmar et al. 1998]) (+ [Tsantoulis et al. 2008])
Distance to the telomere	Distance	Genome-wide screen	
Distance to the centromere	Distance	Genome-wide screen	
Gene expression/chromatin structure			
CpG islands	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
Nuclear lamina binding sites	Coverage	(Guelen et al. 2008)	
miRNA sites	Coverage	(Griffiths-Jones et al. 2007)	
DNA sequence/structure and replication			
Alu repeats	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	(+ [Tsantoulis et al. 2008])
LINE1 repeats	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
LINE2 repeats	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
DNA transposons	Coverage	(Karolchik et al. 2003) (Rhead et al. 2009)	
Dinucleotide microsatellites (>5 repeats)	Coverage	(Abajian [http://espressosoftware.com/sputnik/])	
Inverted repeats	Coverage	(Cer et al. 2010)	
Directed repeats	Coverage	(Cer et al. 2010)	
Triplex motif	Coverage	(Cer et al. 2010)	
Replication timing	Assigned value	(Ryba et al. 2010) (Weddington et al. 2008)	
Origin of replication	Density	(Karnani et al. 2010) (Chen et al. 2011) (Cadoret et al. 2008)	
Recombination and mutational pathways			
Recombination rate	Assigned value	(Myers et al. 2005)	
Evolutionary breakpoint regions	Coverage	(Larkin et al. 2009)	(0 [Ruiz-Herrera et al. 2006])

[a]Type of value for each predictor. (Coverage) Percentage of a particular fragile or nonfragile region that overlaps with a feature. (Assigned value) Value of the predictor for a particular interval. If there is no predictor interval that overlaps with the query interval, the query interval is marked as NA. If there is more than one predictor interval that overlaps with the query interval, the assigned value is calculated as the weighted average based on interval lengths. (Density) Counts of a feature normalized by the interval length. (Distance) Distance measured from the closest terminus of the region to either centromere or telomere.
[b]+ predictor is enriched in fragile sites (positive predictor); – predictor is enriched in nonfragile sites (negative predictor); 0 is nonsignificant.

We investigated genomic regions of the 14 cloned aCFSs defined by the intersection of their cytogenetically and clonally defined coordinates. Using the models described above (Tables 2, 4), we recalculated values for the significant predictors of the smaller genomic segments delineated by the intersection coordinates. Interestingly, in the logistic regression model, using genomic features defined at a higher resolution led to higher or equal expected probabilities of being fragile sites (as compared to using cytogenetic coordinates) for 10 of the 14 aCFSs tested *(FRA2G, FRA2H, FRA3B, FRA4F, FRA6E, FRA7B, FRA7H, FRA9E, FRA11F,* and *FRA13A)*. The remaining four aCFSs *(FRA2C, FRA6F, FRA7G,* and *FRA11G)* had slightly lower probabilities (Supplemental Table S9). Similarly, using the high-resolution genomic features, our standard regression model predicted an equal or higher than expected breakage frequency (as compared with using cytogenetic coordinates) for 11 out of 14 aCFSs (Supplemental Fig. S2). This analysis illustrates that our models capture important aspects of the molecular biology underlying aCFSs and are not a by-product of the lack of resolution of current genome-wide aCFS data.

Validation in mouse fragile sites

For additional validation, we tested 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). The use of the optimal (Table 2) and alternative (Table 3) logistic regression models resulted in 63% and 71%–79% correct predictions, respectively (Supplemental Table S10). Unfortunately, we cannot derive a false positive rate, as the lack of a genome-wide screen for mouse aCFSs precludes defining mouse NFRs. See Supplemental Note for details of the mouse fragile sites analyses.

Table 2. Optimal multiple logistic regression model contrasting autosomal aCFSs with NFRs

Predictor	Standardized coefficient	VIF[b]	*P*-value	Relative contribution
G-band coverage	−6.2290	2.518	2.09×10^{-6}	87.66
Twist (DNA flexibility)	2.8358	4.214	4.30×10^{-4}	12.49
Alu repeat coverage (*log*)	1.9282	2.475	1.08×10^{-2}	4.72
Distance to the centromere[a]	0.9052	1.204	3.30×10^{-2}	3.05
Pseudo R-squared				81.15

[a]Significance is lost after Bonferroni correction for multiple testing.
[b](VIF) Variance inflation factor.

Table 3. Alternative multiple logistic regression models contrasting autosomal aCFSs with NFRs

	(Positive/negative predictor) Relative contribution (*P*-value)							
Predictor	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
G-band coverage	(–) 88.30 (1.2×10^{-4})[c]	(–) 85.76 (3.1×10^{-6})	(–) 85.07 (2.8×10^{-5})	(–) 86.06 (9.1×10^{-7})	(–) 88.03 (2.0×10^{-6})	(–) 89.74 (8.9×10^{-7})	(–) 85.85 (1.3×10^{-4})	(–) 89.87 (1.2×10^{-7})
Twist (DNA flexibility index)					(+) 10.92 (7.5×10^{-4})	(+) 11.44 (5.1×10^{-4})	(–) 0.04 (0.778)[a]	(+) 12.77 (2.6×10^{-4})
Low complexity A/T-rich region coverage	(+) 9.93 (1.7×10^{-3})[c]							
H3K4me1 site coverage		(–) 3.30 (0.021)[b]						
CpG island coverage *(log)*			(–) 10.59 (1.7×10^{-3})					
Transcription start site density *(log)*				(–) 2.05 (0.072)[a]				
Distance to the centromere	(+) 4.08 (0.015)[b]	(+) 3.80 (0.018)[b]	(+) 3.24 (0.027)[b]	(+) 3.24 (0.021)[b]	(+) 4.02 (0.016)[b]	(+) 3.76 (0.016)[b]	(+) 1.19 (0.151)[a]	(+) 2.95 (0.035)[b]
Alu repeat coverage *(log)*	(+) 2.98 (0.041)[b]	(+) 2.97 (0.027)[b]	(+) 4.03 (0.016)[b]	(+) 2.01 (0.065)[a]				
Mononucleotide microsatellite coverage (*log*)					(+) 4.89 (9.8×10^{-3})			
Mononucleotide microsatellite coverage inside *Alus* (*log*)						(+) 4.17 (0.015)[b]		
Mononucleotide microsatellite coverage outside *Alus* (*log*)							(+) 3.83 (0.014)[b]	
Coverage of AT-containing microsatellites (*log*)								(+) 4.13 (0.016)[b]
Pseudo R-squared	86.67	80.95	83.65	79.64	80.49	79.86	84.96	74.98

[a]Not significant.
[b]Not significant after Bonferroni correction for multiple testing.
[c]Variance inflation factor (VIF) higher than 5.

Table 4. Optimal multiple standard regression model for breakage frequency of autosomal aCFSs

Predictor	Standardized coefficient	VIF[b]	*P*-value	Relative contribution
Evolutionary breakpoint region coverage *(log)*	0.5435	1.025	1.61×10^{-6}	31.20
Distance to the centromere	0.4174	1.095	2.10×10^{-4}	20.02
CpG island coverage *(log)*	−0.3889	1.152	6.79×10^{-4}	17.12
G band coverage[a]	−0.2385	1.077	2.67×10^{-2}	7.67
Multiple R-squared				44.81
Adjusted R-squared				41.25

[a]Significance is lost after applying Bonferroni correction for multiple testing.
[b](VIF) Variance inflation factor.

Discussion

In this study, we posed the following questions. First, which genomic features are enriched or depleted in APH-induced CFSs, and how much does each feature contribute to fragility? Second, what are the genomic features that aggravate fragility of aCFSs? Third, are models built based on the analysis of cytogenetically mapped aCFSs also relevant for predicting the available finely mapped aCFSs? We showed that our models predict aCFSs with high accuracy, explain a large portion of the observed variation in breakage frequency, and validate finely mapped aCFSs.

In the genome-wide aCFS vs. NFR logistic regression model (Table 2), G-band coverage is the dominant predictor, while average Twist value, distance to the centromere, and *Alu* repeat coverage have important but smaller roles. The same observations emerge when validating aCFSs mapped at a higher resolution, suggesting the robustness of our conclusions. Moreover, the model has a high success rate for predicting aCFSs in the mouse genome, despite the limited number of identified mouse fragile sites and unavailability of a negative control for mouse. That G banding is the strongest predictor may be due, in part, to the scale at which aCFS fragility information is available to us. G-banding and aCFSs (Mrasek et al. 2010) were both identified at the cytogenetic level, while most of the other genomic features we examined were identified at the primary sequence level. Our study, thus, depends on the accuracy with which the cytogenetic G-banding map was annotated onto the genome sequence (Furey and Haussler 2003). Nevertheless, the preferential location of aCFSs in G-negative bands has been described previously (Yunis and Soreng 1984; Hecht 1988).

Two predictors of genome-wide aCFS breakage frequency (Table 4) coincide with features identified in the aCFS vs. NFR comparison, namely distance to the centromere and G-band coverage. CpG island coverage is a significant predictor in the optimal standard regression model (Table 4) and in one of the alternative logistic regression models (model 3 in Table 3). However, evolutionary breakpoint region coverage is a significant predictor only in the breakage frequency model (Table 4), while *Alu* repeat coverage is significant (prior to Bonferroni correction for multiple testing) only in the logistic model (Table 2). Therefore, our results suggest that genomic features differentiating between aCFSs and NFRs and genomic features that affect fragility level of aCFSs are not necessarily the same.

Our genome-wide model demonstrates that G-banding is negatively correlated with aCFSs; in other words, most aCFSs are located within G-negative (R) bands, which are also the functional regions of the genome (Mishmar et al. 1999). The additional predictors identified in our study provide further insight into the chromosomal organization surrounding aCFSs and the mechanisms underlying aCFS expression, as discussed below.

Alu repeat coverage

Alu repeats have been documented previously to be enriched in aCFSs (Tsantoulis et al. 2008), and our optimal logistic model including *Alu* repeats has high predictive value (Table 2). Mechanistically, *Alu* repeats have been studied extensively for their effect on nonhomologous recombination, which might impact chromosome stability (Cordaux and Batzer 2009; Konkel and Batzer 2010). Most *alu* repeats contain mononucleotide microsatellites (Arcot et al. 1995; Kelkar et al. 2011) which are involved in replication slippage, unequal crossing over, secondary structure formation, and DNA polymerase inhibition (Bhargava and Fuentes 2009; Shah et al. 2010). Therefore, a role for mononucleotide microsatellites in genomic instability is well supported. We found mononucleotide microsatellite coverage to be a significant (prior to Bonferroni correction for multiple testing) positive predictor, although the model including it in place of *Alu* coverage had lower explanatory power (Table 3). Although the logistic regression model with mononucleotide microsatellites located within *Alu*s had a slightly lower overall fit than the model including *Alu* coverage, it is possible that mononucleotide microsatellites may be the factor that actually contributes to aCFS fragility, due to a high correlation in occurrence of *alu* repeats and mononucleotide microsatellites.

AT sequence content and DNA flexibility

Experimental studies have shown that long AT/TA palindromes and AT-rich sequences are associated with replication stalling and aCFS breakage (Dillon et al. 2010). G-negative bands are heterogeneous with regard to base composition and may contain AT-rich isochores (Costantini et al. 2006). We found Twist percentage (Sarai et al. 1989) to be a significant, positive predictor in the optimal logistic model contrasting aCFSs with NFRs (Table 2), which concurs with a previous study (Mishmar et al. 1998). Our results suggest that aCFSs tend to be located in G-negative banding regions that have AT-rich isochores with a high density of AT-rich repeats or a high A/T-base pair content (Mishmar et al. 1998,1999). Therefore, our results agree with the notion that aCFSs might be located in G-band-like regions of R bands (Mishmar et al. 1999).

CpG island coverage

We found that CpG island coverage is negatively associated with both breakage frequency and the probability of being an aCFS (Table 4 and model 3 in Table 3, respectively). Also, in some alternative logistic regression models, H3K4me1 site coverage was also a significant negative predictor, but loses its significance after Bonferroni correction for multiple testing (model 2 in Table 3). This finding is counterintuitive, given the dogma that G-negative bands display a higher density of genes, CpG islands, and histone acetylation, relative to G-positive bands (Craig and Bickmore 1993). One interpretation of our results is that aCFSs reside within G-negative chromosomal isochores that adopt a less open

chromatin structure, relative to NFRs. Because DNA repair mechanisms are known to be more efficient in open chromatin and active gene regions (Mellon et al. 1986), NFRs may be able to repair DNA more efficiently than aCFS regions.

Replication timing

Delayed replication has been considered to be a key molecular feature associated with aCFS expression (Arlt et al. 2006; Palumbo et al. 2010). However, we found G-negative banding to be the dominant predictor of fragility in our aCFS vs. NFR model, and initial studies examining G vs. R banding patterns and S-phase replication timing showed that G-negative bands are replicated early in S phase (Holmquist et al. 1982). Thus, despite the observation that some aCFSs remain incompletely replicated at the end of S phase (Palakodeti et al. 2004; Pelliccia et al. 2008), late replication of G bands per se is not a genome-wide predictor of fragility (see below). Instead, our modeling suggests that aCFSs may be sequences that experience replication delays but that lie within otherwise early replicating regions of the genome.

Late-replicating regions of the genome are known to harbor heterochromatin and centromeres, both highly repetitive sequences (Holmquist et al. 1982). Recently, genome-wide studies of replication timing have been performed, primarily examining replication dynamics within unique and low-complexity sequences of the genome. In these studies, repetitive DNA sequences are excluded from the sequencing analyses (Hansen et al. 2009), or arrays are used that are underrepresented for heterochromatic regions of the genome and do not contain centromeric DNA (Woodfine et al. 2004). These methodological limitations notwithstanding, we used the replication timing data from three studies (Woodfine et al. 2004; Hansen et al. 2009; Ryba et al. 2010) to determine whether it is a predictor of chromosomal fragility. However, we found neither replication timing nor content of early replicated regions to be a significant predictor in any of our models. Possibly, some aCFSs have average replication timing under normal physiological conditions but exhibit delayed or late replication under APH-induced stress. Replication timing can change with developmental stage and cell type (Hansen et al. 2009; Pope et al. 2010).

Density of replication origins

Recent studies of *FRA3B* suggest that differential utilization of replication origins contributes to fragility of aCFSs (Palakodeti et al. 2010; Letessier et al. 2011). However, we did not find the density of replication origins to be a significant genome-wide predictor in our models. We utilized origin-mapping data derived both computationally and experimentally (Cadoret et al. 2008; Karnani et al. 2010; Chen et al. 2011). Notably, the densities of computationally predicted vs. experimentally mapped replication origins were not highly correlated (Fig. 3), suggesting limitations in one or both of these approaches. The computational prediction of replication origins uses abrupt base skew changes to partially differentiate leading versus lagging DNA strand switches (Chen et al. 2011). However, a similar signal can be generated from a transcription start site. Experimental mapping of replication origins exhibits substantial variation among platforms utilized in the same laboratory (Karnani et al. 2010), and the set of origins identified by several platforms has a high false negative rate (low sensitivity). While aCFSs and NFRs appear to have a similar density of replication origins, they might differ in replication origin efficiency (Palakodeti et al. 2010; Letessier et al. 2011) or utilization (Gilbert 2010). In fact, a recent analysis suggested that the failure to activate origins in response to replication stress and fork stalling was involved in *FRA16C* instability (Ozeri-Galai et al. 2011).

Distance from the centromere

We observed a positive association between the presence of aCFSs and distance from the centromere, suggesting that chromosomal regions located farther away from the centromere have a higher probability of being aCFSs. Moreover, the farther away from the centromere, the higher is the breakage frequency of aCFSs. Several genomic contexts/features are known to vary along the length of a chromosome, creating a change in genomic landscape that affects the rates of various mutational events (Hardison et al. 2003; Kvikstad et al. 2007; Ananda et al. 2011). The best subset selection procedure we applied to select the optimal set of predictors in our regressions favors models with a small number of predictors. Therefore, in reality, distance from the centromere might be selected as an effective proxy capturing the effects on fragility of multiple genomic features as they vary along the length of the chromosome.

Enrichment of evolutionary breakpoint regions

Evolutionary breakpoint regions are genomic sites of intra- and inter-chromosomal breakages that were found to be frequently reused among ten amniote genomes analyzed (Larkin et al. 2009). From our study, this predictor is significant only in the breakage frequency model, where it is, in fact, the dominant predictor. Since the aCFS vs. NFR model includes a larger data set (73 aCFSs + 117 NFRs) than the breakage frequency model (73 aCFSs), the lack of predictive power of evolutionary breakpoints in the former model cannot be explained by sample size limitations. Our results suggest that evolutionary breakpoint regions are enriched specifically in highly fragile aCFSs.

Statistically, we cannot establish a causality direction, i.e., whether (1) evolutionary breakpoint regions make existing aCFSs more fragile, or (2) aCFSs are, indeed, hotspots of evolutionary breakpoints. Some human aCFSs were found to have orthologous aCFSs in other mammals, e.g., other primates (Smeets and van de Klundert 1990; Ruiz-Herrera et al. 2004), carnivores (Stone et al. 1991a,b, 1993), and mouse (Glover et al. 1998; Shiraishi et al. 2001; Krummel et al. 2002; Matsuyama et al. 2003; Rozier et al. 2004; Helmrich et al. 2006, 2007). Analysis of such loci is expected to shed light on the causative agents of fragility (CFSs vs. evolutionary breakpoints). A locus-specific analysis indicated that, for instance, both human and mouse orthologous CFSs are enriched in AT-repeats (Shiraishi et al. 2001). In addition, recent evidence suggests that evolutionary breakpoint regions are enriched for repeats that might alter chromatin conformation or recruit transposable elements and trigger genome instability (Farre et al. 2011). For a definite answer to this puzzling question, a genome-wide analysis of sequence features still conserved for such orthologous aCFSs is required. Nevertheless, our modeling suggests that some features of chromosomal regions that are conserved in their evolutionary fragility across species separated by hundreds of million of years (Larkin et al. 2009) are also associated with fragility of these regions in the human genome under conditions of replication stress. This implies similarity in the mechanisms of chromosomal fragility at micro- and macroevolutionary levels. Larkin and colleagues (2009) discovered that evolutionary breakpoint regions are enriched in structural variants, SNPs, genes, and

pseudogenes, and depleted in recombination hotspots and most conserved elements. The significance of evolutionary breakpoint regions in our modeling might capture a combination of some of these factors.

Summary

The ultimate goal of our computational analysis is to develop accurate and reliable models that can aid in the prediction of locations and fragility levels of aCFSs within individual human genomes. Using the models we describe here, we can currently predict the probability that a given chromosomal region is an aCFS and its corresponding breakage frequency, based on genomic context. We demonstrate that our models remain valid when we apply them to a handful of aCFSs that have been mapped using fine-scale fluorescence probe labeling. Our models did not identify several "expected" genomic features as being significant predictors of genome-wide chromosomal fragility. This does not eliminate these characteristics, which include replication timing and replication origin density, from the list of potential contributors to aCFS instability. Rather, our findings support the idea that, although aCFSs share characteristics that predict fragility globally (such as those found in our optimal models), other genome features might be contributors to only a unique subset of aCFSs. A full understanding of the mechanisms of aCFS instability will require further computational and experimental analyses. With advances in genome-wide sequencing technologies, we will soon be in a position to identify locations of chromosome breakage at the base pair level, allowing a more detailed analysis of aCFSs in individual genomes.

Methods

Mapping genomic locations of aCFSs and NFRs

The cytogenetically determined locations of all 73 autosomal aCFSs, as determined previously (Lukusa and Fryns 2008; Mrasek et al. 2010), were converted to human genomic coordinates (hg18) using the UCSC Genome Browser (Rhead et al. 2009). Breakage frequencies were obtained for three distinct individuals but presented high inter-individual concordance (correlations around 0.96–0.99); we, therefore, used the average breakage frequency across the three individuals for each aCFS.

The NFR set was constructed from regions that did not exhibit breakage after induction by aphidicolin in the genome-wide screen by Mrasek and colleagues (Mrasek et al. 2010) and were not indicated as fragile sites in other studies (Kuwano et al. 1988; Borgaonkar 1994). From these, we further excluded centromeric regions because they are enriched in minisatellites (Vergnaud and Denoeud 2000) and heterochromatin regions that do not have DNA sequence available. Sex chromosomes were also excluded from our initial analyses because of their high repetitive element content (Skaletsky et al. 2003; Ross et al. 2005). In all, we utilized 117 autosomal NFRs.

Calculating and prescreening predictors

Genomic features (Supplemental Table S3), as assigned to each aCFS and NFR, were downloaded from the UCSC Genome Browser (Rhead et al. 2009) or from the literature (see Supplemental Table S3 for references). The hg18 human genome annotations were used for most features, and those available only for other human genome assemblies were mapped to hg18 with the lift-over tool in Galaxy (Blankenberg et al. 2011). Most features were available in the corresponding data sets as genomic intervals and were intersected with aCFS or NFR cytogenetic coordinates to calculate coverage (percentage of overlap) for large-scale genomic features or density for small motifs. For replication timing, we used a weighted average value when several data intervals overlapped with an aCFS or an NFR. Features were measured as coverage (percentage), motif density, or average value across an interval, depending on their type (Table 1; Supplemental Table S3), and each was transformed to approximate a Gaussian distribution. The gaps in assembly were subtracted from each aCFS and NFR prior to calculating coverage, density, or assigned value.

In order to limit the number of, and correlations among, features used as potential predictors for our regression models, we performed a prescreening (Fig. 2A). We used hierarchical clustering based on pairwise Spearman's rank correlation coefficients (distance = 1–|coefficient|) to parse 54 features into tightly correlated groups (clusters) and selected 19 of them, each representing one such group and having correlations below 0.7 with one another. Notably, considering only the 73 aCFSs, or both the 73 aCFSs and the 117 NFRs, produced similar clustering patterns and led to selecting the same predictors. Such prescreening facilitates subsequent regression model building. It reduces computational time for best subset selection algorithms (see below; since computational time doubles for every additional predictor, excluding ~30 features reduced the computational burden by a trillion-fold). Moreover, it improves estimation through the model-building process, providing a higher "observations per predictor" ratio. This is especially important for logistic regression, where estimation is performed by numerical maximum likelihood and requires sufficiently large sample sizes to converge. Note that prescreening predictors by clustering mitigates, but does not eliminate, the risk of multicollinearity in our regressions (even though potential predictors are picked to have relatively low pairwise correlations, overall linear associations might still be high). We, therefore, still evaluate multicollinearity using variance inflation factors during model building (see below).

Regression analyses

Two types of regressions were used in our study—logistic and standard multiple linear regression. The former models a binary response (aCFS = "1" vs. NFR = "0") and the latter an approximately continuous response (breakage frequency of aCFSs). For both regressions, we (1) performed transformations on the 19 potential predictors to approximate Gaussian distributions; for the standard regression, also the response (breakage frequency) was transformed by natural logarithm to regularize its distribution and ensure homoscedasticity in the fits, (2) ran a best subset selection algorithm to select a smaller subset of predictors based on the Akaike information criterion, (3) checked this subset of predictors for autocorrelation using the partial autocorrelation function—no absolute partial autocorrelations above 0.15 were detected in any of the analyses, (4) identified and removed influential data points (outliers), based on Cook's distances computed with the model comprising this subset of predictors (we removed points with Cook's distance larger than 4/[sample size – number of predictors –1], which corresponded to ~0.02 and 0.05 for the logistic and standard regression, respectively), (5) further reduced the model iteratively eliminating predictors based on their coefficients' *P*-values and variance inflation factors; this led to models retaining only predictors significant or marginally significant after Bonferroni correction for multiple testing and with variance inflation factors below 5 (unless noted otherwise), and (6) considered quadratic models comprising square and pairwise product terms

obtained from these final sets of predictors—no quadratic terms were found significant in any of the analyses (Fox 2002). Common graphical diagnostics (e.g., residual plots) (Supplemental Fig. S3) were also employed to assess model performance throughout the process. The pipeline of our regression analysis is depicted in Figure 2B.

For both regressions, we evaluated the final models and individual contributions of the predictors retained in them based on explained deviance (logistic regression) or variance (standard regression). For the logistic regression, we calculated the pseudo R-squared of a model using $(D_o–D)/D_o$, where D_o is the null deviance and D is the residual deviance of the model. We calculated the relative contribution of each predictor to a model using $[(D_o–D) – (D_o–D_{(–p)})]/(D_o–D)$, where $D_{(–p)}$ is the deviance of the smaller model obtained removing the predictor of interest. For standard regression, we calculated the R-squared of a model using (TSS – SSE)/SSE, where TSS is the total sum of squares and SSE the residual sum of squares of the model. We calculated the relative contribution of each predictor to a model using the partial R-squared, which is defined as $(SSE_{(–p)} – SSE)/SSE_{(–p)}$, where $SSE_{(–p)}$ is the residual sum of squares of the smaller model obtained removing the predictor of interest.

All regression analyses were implemented in the R statistical package version 2.11.1 (R Development Core Team 2011). All tools developed for this project are freely available at the Galaxy (Blankenberg et al. 2011) website, http://main.g2.bx.psu.edu/. Tools "Logistic Regression" and "Partial R-squared" can be found under "Multiple Regression," and tool "Assigned Weighted Average Value of Genomic Feature" can be found under "Regional Variation." The "Standard Multiple Regression" tool used here was already available in Galaxy.

Cloned aCFSs

Eighteen autosomal aCFSs mapped by fluorescence probes (i.e., "cloned") (Ciullo et al. 2002; Ruiz-Herrera et al. 2006; Fechter et al. 2007; Reshmi et al. 2007; Bosco et al. 2010; Pelliccia et al. 2010; Blumrich et al. 2011) were studied. For analysis, we intersected their cloned and cytogenetically defined coordinates. For four CFSs, we found no intersections, although clonally defined coordinates were adjacent to the cytogenetically defined coordinates. The 14 intersected regions were used to validate our logistic and multiple regression models. We recalculated values for significant predictors using the same method we used with aCFSs defined at the cytogenetic level. Next, we compared breakage frequency and probability to be an aCFS for the intersected regions vs. their cytogenetically defined counterparts.

Mouse fragile sites and genomic contexts

Cytogenetic locations of 24 known APH-induced mouse fragile sites (Elder and Robinson 1989; Helmrich et al. 2006) were converted to mouse genomic coordinates (mm8). For 16 of these sites, breakage frequencies were estimated from 266 mouse cells (Elder and Robinson 1989). Genomic features for each mouse fragile site were calculated similarly as for human aCFSs. We used rodent B1s as an equivalent to *Alu* repeats in human as both evolved from the 7SL RNA gene (Yang et al. 2004). Mouse fragile sites were used as a test set to validate our multiple logistic regression for aCFSs prediction. We consider 0.5 to be the threshold for a positive call assuming equal probability to observe fragile and nonfragile regions. Breakage frequencies of mouse fragile sites were used to validate our multiple linear regression for breakage frequency.

Acknowledgements

We thank Benjamin Dickins for his comments on the manuscript and Yogeshwar Kelkar for his help at the early stages of this project. This study was supported by NIH grant R01-GM087472 to K.D.M. and K.A.E., and a Fulbright Scholarship and the 2011 ICS Summer Student Research Award, Pennsylvania State University, to A.F. This project was also funded, in part, under a grant from the Pennsylvania Department of Health using Tobacco CURE Funds (SAP# 4100042746). The Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

References

Ananda G, Chiaromonte F, Makova KD. 2011. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biol* **12:** R27. doi: 10.1186/gb-2011-12-3-r27.

Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA. 1995. Alu repeats: A source for the genesis of primate microsatellites. *Genomics* **29:** 136–144.

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

Bester AC, Schwartz M, Schmidt M, Garrigue A, Hacein-Bey-Abina S, Cavazzana-Calvo M, Ben-Porat N, Von Kalle C, Fischer A, Kerem B. 2006. Fragile sites are preferential targets for integrations of MLV vectors in gene therapy. *Gene Ther* **13:** 1057–1059.

Bhargava A, Fuentes FF. 2009. Mutational dynamics of microsatellites. *Mol Biotechnol* **44:** 250–266.

Blankenberg D, Coraor N, Von Kuster G, Taylor J, Nekrutenko A. 2011. Integrating diverse databases into a unified analysis framework: A Galaxy approach. *Database (Oxford)* **2011:** bar011. doi: 10.1093/database/bar011.

Blumrich A, Zapatka M, Brueckner LM, Zheglo D, Schwab M, Savelyeva L. 2011. The FRA2C common fragile site maps to the borders of MYCN amplicons in neuroblastoma and is associated with gross chromosomal rearrangements in different cancers. *Hum Mol Genet* **20:** 1488–1501.

Borgaonkar DS. 1994. *Chromosomal variation in man: A catalog of chromosomal variants and anomalies,* 7th ed. Wiley-Liss, New York.

Bosco N, Pelliccia F, Rocchi A. 2010. Characterization of FRA7B, a human common fragile site mapped at the 7p chromosome terminal region. *Cancer Genet Cytogenet* **202:** 47–52.

Burrow AA, Williams LE, Pierce LC, Wang Y-H. 2011. Over half of breakpoints in gene pairs involved in cancer-specific recurrent translocations are mapped to human chromosomal fragile sites. *BMC Genomics* **10:** 59. doi: 10.1186/1471-2164-10-59.

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

Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfovsky N, Luke BT, Bacolla A, Collins JR, Stephens RM. 2010. Non-B DB:A database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res* **39:** D383–D391.

Chan KL, Palmai-Pallag T, Ying S, Hickson ID. 2009. Replication stress induces sister-chromatid bridging at fragile site loci in mitosis. *Nat Cell Biol* **11:** 753–760.

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. *Mol Biol Evol* **28:** 2327–2337.

Ciullo M, Debily M-A, Rozier L, Autiero M, Billault A, Mayau V, El Marhomy S, Guardiola J, Bernheim A, Coullin P, 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.

Comings DE. 1978. Mechanisms of chromosome banding and implications for chromosome structure. *Annu Rev Genet* **12:** 25–46.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10:** 691–703.

Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. *Genome Res* **16:** 536–541.

Craig JM, Bickmore WA. 1993. Chromosome bands-flavours to savour. *Bioessays* **15:** 349–354.

Dall KL, Scarpini CG, Roberts I, Winder DM, Stanley MA, Muralidhar B, Herdman MT, Pett MR, Coleman N. 2008. Characterization of naturally

occurring HPV16 integration sites isolated from cervical keratinocytes under noncompetitive conditions. *Cancer Res* **68:** 8249–8259.
Dillon LW, Burrow AA, Wang Y-H. 2010. DNA instability at chromosomal fragile sites in cancer. *Curr Genomics* **11:** 326–337.
Durkin SG, Glover TW. 2007. Chromosome fragile sites. *AnnuRev Genet* **41:** 169–192.
Durkin SG, Ragland RL, Arlt MF, Mulle JG, Warren ST, Glover TW. 2008. Replication stress induces tumor-like microdeletions in FHIT/FRA3B. *Proc Natl Acad Sci* **105:** 246–251.
Elder FF, Robinson TJ. 1989. Rodent common fragile sites: Are they conserved? Evidence from mouse and rat. *Chromosoma* **97:** 459–464.
Farré M, Bosch M, López-Giráldez F, Ponsà M, Ruiz-Herrera A. 2011. Assessing the role of tandem repeats in shaping the genomic architecture of great apes. *PLoS ONE* **6:** e27239. doi: 10.1371/journal.pone.0027239.
Fechter A, Buettel I, Kuehnel E, Savelyeva L, Schwab M. 2007. Common fragile site FRA11G and rare fragile site FRA11B at 11q23.3 encompass distinct genomic regions. *Genes Chromosomes Cancer* **46:** 98–106.
Fox DJ. 2002. *An R and S-plus companion to applied regression,* 1st ed. Sage Publications, Inc., Thousand Oaks, CA.
Freudenreich CH. 2007. Chromosome fragility: Molecular mechanisms and cellular consequences. *Front Biosci* **12:** 4911–4924.
Furey TS, Haussler D. 2003. Integration of the cytogenetic map with the draft human genome sequence. *Hum Mol Genet* **12:** 1037–1044.
Gilbert DM. 2010. Evaluating genome-scale approaches to eukaryotic DNA replication. *Nat Rev Genet* **11:** 673–684.
Glover TW, Stein CK. 1987. Induction of sister chromatid exchanges at common fragile sites. *Am J Hum Genet* **41:** 882–890.
Glover TW, Hoge AW, Miller DE, Ascara-Wilke JE, Adam AN, Dagenais SL, Wilke CM, Dierick HA, Beer DG. 1998. The murine Fhit gene is highly similar to its human orthologue and maps to a common fragile site region. *Cancer Res* **58:** 3409–3414.
Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2007. miRBase: Tools for microRNA genomics. *Nucleic Acids Res* **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, Stamatoyannopoulos JA. 2009. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci* **107:** 139–144.
Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* **13:** 13–26.
Hecht F. 1988. Fragile sites, cancer chromosome breakpoints, and oncogenes all cluster in light G bands. *Cancer Genet Cytogenet* **31:** 17–24.
Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459:** 108–112.
Hellman A, Rahat A, Scherer SW, Darvasi A, Tsui L-C, Kerem B. 2000. Replication delay along FRA7H, a common fragile site on human chromosome 7, leads to chromosomal instability. *Mol Cell Biol* **20:** 4420–4427.
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.
Holmquist GP. 1992. Chromosome bands, their chromatin flavors, and their functional features. *Am J Hum Genet* **51:** 17–37.
Holmquist G, Gray M, Porter T, Jordan J. 1982. Characterization of Giemsa dark- and light-band DNA. *Cell* **31:** 121–129.
Hussein SM, Batada NN, Vuoristo S, Ching RW, Autio R, Narva E, Ng S, Sourour M, Hamalainen R, Olsson C, et al. 2011. Copy number variation and selection during reprogramming to pluripotency. *Nature* **471:** 58–62.
Jiang Y, Lucas I, Young DJ, Davis EM, Karrison T, Rest JS, Le Beau MM. 2009. Common fragile sites are characterized by histone hypoacetylation. *Hum Mol Genet* **18:** 4501–4512.
Karnani N, Taylor CM, Malhotra A, Dutta A. 2010. Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. *Mol Biol 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, Eckert KA, Chiaromonte F, Makova KD. 2011. A matter of life or death: How microsatellites emerge in and vanish from the human genome. *Genome Res* **21:** 2038–2048.
Konkel MK, Batzer MA. 2010. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* **20:** 211–221.
Krummel KA, Denison SR, Calhoun E, Phillips LA, Smith DI. 2002. The common fragile site FRA16D and its associated gene WWOX are highly conserved in the mouse at Fra8E1. *Genes Chromosomes Cancer* **34:** 154–167.
Kuwano A, Sugio Y, Murano I, Kajii T. 1988. Common fragile sites induced by folate deprivation, BrdU and aphidicolin: Their frequency and distribution in Japanese individuals. *Jpn J Hum Genet* 33: 355–364.
Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. 2007. A macaque's-eye view of human insertions and deletions: Differences in mechanisms. *PLoS ComputBiol* **3:** e176. doi: 10.1371/journal.pcbi.0030176.
Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res* **19:** 770–777.
Le Beau MM, Rassool FV, Neilly ME, Espinosa R, Glover TW, Smith DI, McKeithan TW. 1998. Replication of a common fragile site, FRA3B, occurs late in S phase and is delayed further upon induction: Implications for the mechanism of fragile site induction. *Hum Mol Genet* **7:** 755–761.
Letessier A, Millot GA, Koundrioukoff S, Lachages A-M, Vogt N, Hansen RS, Malfoy B, Brison O, Debatisse M. 2011. Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature* **470:** 120–123.
Lukusa T, Fryns JP. 2008. Human chromosome fragility. *Biochim Biophys Acta* **1779:** 3–16.
Matsuyama A, Shiraishi T, Trapasso F, Kuroki T, Alder H, Mori M, Huebner K, Croce CM. 2003. Fragile site orthologs FHIT/FRA3B and Fhit/Fra14A2: Evolutionarily conserved but highly recombinogenic. *Proc Natl Acad Sci* **100:** 14988–14993.
Mellon I, Bohr VA, Smith CA, Hanawalt PC. 1986. Preferential DNA repair of an active gene in human cells. *Proc Natl Acad Sci* **83:** 8878–8882.
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* **95:** 8141–8146.
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.
Mrasek K, Schoder C, Teichmann A-C, Behr K, Franze B, Wilhelm K, Blaurock N, Claussen U, Liehr T, Weise A. 2010. Global screening and extended nomenclature for 230 aphidicolin-inducible fragile sites, including 61 yet unreported ones. *Int J Oncol* **36:** 929–940.
Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310:** 321–324.
Ozeri-Galai E, Lebofsky R, Rahat A, Bester AC, Bensimon A, Kerem B. 2011. Failure of origin activation in response to fork stalling leads to chromosomal instability at fragile sites. *Mol Cell* **43:** 122–131.
Palakodeti A, Han Y, Jiang Y, Le Beau MM. 2004. The role of late/slow replication of the FRA16D in common fragile site induction. *Genes Chromosomes Cancer* **39:** 71–76.
Palakodeti A, Lucas I, Jiang Y, Young DJ, Fernald AA, Karrison T, Le Beau MM. 2010. Impaired replication dynamics at the FRA3B common fragile site. *Hum Mol Genet* **19:** 99–110.
Palumbo E, Matricardi L, Tosoni E, Bensimon A, Russo A. 2010. Replication dynamics at common fragile site FRA6E. *Chromosoma* **119:** 575–587.
Pelliccia F, Bosco N, Curatolo A, Rocchi A. 2008. Replication timing of two human common fragile sites: FRA1H and FRA2G. *Cytogenet Genome Res* **121:** 196–200.
Pelliccia F, Bosco N, Rocchi A. 2010. Breakages at common fragile sites set boundaries of amplified regions in two leukemia cell lines K562—Molecular characterization of FRA2H and localization of a new CFS FRA2S. *Cancer Lett* **299:** 37–44.
Pope BD, Hiratani I, Gilbert DM. 2010. Domain-wide regulation of DNA replication timing during mammalian development. *Chromosome Res* **18:** 127–136.
R Development Core Team. 2011. *R:A language and environment for statistical computing* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Ragland RL, Glynn MW, Arlt MF, Glover TW. 2008. Stably transfected common fragile site sequences exhibit instability at ectopic sites. *Genes Chromosomes Cancer* **47:** 860–872.

Reshmi SC, Huang X, Schoppy DW, Black RC, Saunders WS, Smith DI, Gollin SM. 2007. Relationship between FRA11F and 11q13 gene amplification in oral cancer. *Genes Chromosomes Cancer* **46:** 143–154.

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 Res* **38:** D613–D619.

Ried K, Finnis M, Hobson L, Mangelsdorf M, Dayan S, Nancarrow JK, Woollatt E, Kremmidiotis G, Gardner A, Venter D, et al. 2000. Common chromosomal fragile site FRA16D sequence: Identification of the FOR gene spanning FRA16D and homozygous deletions and translocation breakpoints in cancer cells. *Hum Mol Genet* **9:** 1651–1663.

Robinson TJ, Elder FF. 1987. Multiple common fragile sites are expressed in the genome of the laboratory rat. *Chromosoma* **96:** 45–49.

Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, et al. 2005. The DNA sequence of the human X chromosome. *Nature* **434:** 325–337.

Rozier L, El-Achkar E, Apiou F, Debatisse M. 2004. Characterization of a conserved aphidicolin-sensitive common fragile site at human 4q22 and mouse 6C1: Possible association with an inherited disease and cancer. *Oncogene* **23:** 6872–6880.

Ruiz-Herrera A, Garcia F, Frönicke L, Ponsà M, Egozcue J, Caldés MG, Stanyon R. 2004. Conservation of aphidicolin-induced fragile sites in Papionini (Primates) species and humans. *Chromosome Res* **12:** 683–690.

Ruiz-Herrera A, Castresana J, Robinson TJ. 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol* **7:** R115. doi: 10.1186/gb-2006-7-12-r115.

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

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

Shah SN, Opresko PL, Meng X, Lee MYWT, Eckert KA. 2010. DNA structure and the Werner protein modulate human DNA polymerase δ-dependent replication dynamics within the common fragile site FRA16D. *Nucleic Acids Res* **38:** 1149–1162.

Shiraishi T, Druck T, Mimori K, Flomenberg J, Berk L, Alder H, Miller W, Huebner K, Croce CM. 2001. Sequence conservation at human and mouse orthologous common fragile regions, FRA3B/FHIT and Fra14A2/ Fhit. *Proc Natl Acad Sci* **98:** 5722–5727.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825–837.

Smeets DF, van de Klundert FA. 1990. Common fragile sites in man and three closely related primate species. *Cytogenet Cell Genet* **53:** 8–14.

Soulie J, De Grouchy J. 1981. A cytogenetic survey of 110 baboons *(Papio cynocephalus)*. *Am JPhys Anthropol* **56:** 107–113.

Stone DM, Jacky PB, Prieur DJ. 1991a. Chromosomal fragile site expression in dogs: II. Expression in boxer dogs with mast cell tumors. *Am J Med Genet* **40:** 223–229.

Stone DM, Jacky PB, Hancock DD, Prieur DJ. 1991b. Chromosomal fragile site expression in dogs: I. Breed specific differences. *Am J Med Genet* **40:** 214–222.

Stone DM, Stephens KE, Doles J. 1993. Folate-sensitive and aphidicolin-inducible fragile sites are expressed in the genome of the domestic cat. *Cancer Genet Cytogenet* **65:** 130–134.

Travers AA. 2004. The structural basis of DNA flexibility. *Philos Transact A Math Phys EngSci* **362:** 1423–1438.

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:** 32563264.

Vergnaud G, Denoeud F. 2000. Minisatellites: Mutability and genome architecture. *Genome Res* **10:** 899–907.

Weddington N, Stuy A, Hiratani I, Ryba T, Yokochi T, Gilbert DM. 2008. ReplicationDomain: A visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics* **9:** 530. doi: 10.1186/1471-2105-9-530.

Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, Debernardi S, Mott R, Dunham I, 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, Hardison RC. 2004. Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res* **14:** 517–527.

Yunis JJ, Soreng AL. 1984. Constitutive fragile sites and cancer. *Science* **226:** 1199–1204.

Zhang H, Freudenreich CH. 2007. An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in *S. cerevisiae*. *Mol Cell* **27:** 367–379.

Zlotorynski E, Rahat A, Skaug J, Ben-Porat N, Ozeri E, Hershberg R, Levi A, Scherer SW, Margalit H, Kerem B. 2003. Molecular basis for expression of common and rare fragile sites. *Mol Cell Biol* **23:** 7143–7151.

Received November 4, 2011; accepted in revised form March 19, 2012.

***Genome Research* 22:** 993–1005 (2012)

Corrigendum: A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome?

Arkarachai Fungtammasan, Erin Walsh, Francesca Chiaromonte, Kristin A. Eckert, and Kateryna D. Makova

The authors found that the genomic coordinates for several fragile sites were accidentally truncated during conversion from cytogenetic bands. The errors shift the genomic coordinates at five out of a total of 76 aphidicolin-induced common fragile sites (aCFSs) and at 11 nonfragile regions. Seven nonfragile regions had to be removed because their coordinates now overlap with aCFSs, reducing the number of control regions from 131 to 124. These changes affected Figures 1 and 2, and Supplemental Tables S1 and S2. Upon reanalysis, the list of significant genomic features, their directional effects on fragility, and relative contributions did not change.

However, correcting of genomic coordinates resulted in the numerical changes of the regression models' attributes (coefficient, variance inflation factor, *P*-value, and relative contribution) throughout the manuscript. These changes affected Figure 3, Supplemental Figures S2 and S3, and Supplemental Tables S4–S7, S9, and S10. Additionally they include:

1. The pseudo R-squared of logistic regressions in Tables 2 and 3 are about 5% higher than previously published. This makes our models stronger. There are also small numerical changes in Table 4.
2. The logistic regression models that include CpG islands have R-squared values (Tables 2, 3) and correct prediction rates (Table S10) that are both higher than the corresponding values in the models including Twist.
3. The transcription start site density is no longer a significant predictor that can distinguish aCFSs from nonfragile regions (it was significant before Bonferroni correction); however, it is still a significant predictor in models explaining the level of fragility.

These corrections do not affect the conclusions of the article. The authors apologize for making this mistake and for any confusion this may have caused.

The article has already been corrected in both the PDF and full-text HTML files online.

doi: 10.1101/gr.214460.116

A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome?

Arkarachai Fungtammasan, Erin Walsh, Francesca Chiaromonte, et al.

Genome Res. 2012 22: 993-1005 originally published online March 28, 2012
Access the most recent version at doi:10.1101/gr.134395.111