DNA methylation profiling in X;autosome translocations supports a role for L1 repeats in the spread of X chromosome inactivation

Neeta Bala Tannan1, Manisha Brahmachary1, Paras Garg1, Christelle Borel1,†, Randah Alnefaie2, Corey T. Watson1, N. Simon Thomas3,4 and Andrew J. Sharp1,*

1Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, USA, 2Program in Bioinformatics, Graduate Center for Professional Studies, Polytechnic Institute of New York University, New York, USA, 3Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Salisbury, UK and 4Division of Human Genetics, University of Southampton, Southampton, UK

Received May 30, 2013; Revised September 20, 2013; Accepted October 29, 2013

X chromosome inactivation (XCI) is an epigenetic mechanism that silences the majority of genes on one X chromosome in females. Previous studies have suggested that the spread of XCI might be facilitated in part by common repeats such as long interspersed nuclear elements (LINEs). However, owing to the unusual sequence content of the X and the nonrandom distribution of genes that escape XCI, it has been unclear whether the correlation between repeat elements and XCI is a functional one. To test the hypothesis that the spread of XCI shows sequence specificity, we have analyzed the pattern of XCI in autosomal chromatin by performing DNA methylation profiling in six unbalanced X;autosome translocations. Using promoter hypermethylation as an epigenetic signature of XCI, we have determined the inactivation status of 1050 autosomal genes after translocation onto an inactive derivative X. By performing a comparative sequence analysis of autosomal genes that are either subject to or escape the X inactivation signal, we identified a number of common repetitive elements, including L1 and L2 LINEs, and DNA motifs that are significantly enriched around inactive autosomal genes. We show that these same motifs predominantly map to L1P repeat elements, are significantly enriched on the X chromosome versus the autosomes and also occur at higher densities around X-linked genes that are subject to X inactivation compared with those that escape X inactivation. These results are consistent with a potential causal relationship between DNA sequence features such as L1s and the spread of XCI, lending strong support to Mary Lyon’s ‘repeat hypothesis’.

INTRODUCTION

X chromosome inactivation (XCI) is a mechanism that causes the transcriptional silencing of most genes on one X chromosome in female mammals, allowing for dosage compensation of the X chromosome between males and females (1). The X inactivation center at Xq13.2 mediates inactivation in cis (2) and contains the XIST gene, which produces a non-coding RNA that specifically coats the inactive X (Xi) (3). Genes that are inactivated on the Xi are characterized by epigenetic modifications including a number of different histone modifications and methylation of CpG islands (4). However, ~20% of the genes are expressed from both the active X (Xa) and the Xi and thus escape the X inactivation process (5). Genes that escape X inactivation show a markedly different epigenetic profile (4,6,7).

One of the first hypotheses to explain how the X inactivation signal is able to spread over the entire Xi suggested the presence of booster elements or way stations, which act to strengthen the propagation of the X inactivation signal (8). Based on their enrichment on the X chromosome, long interspersed nuclear
elements (LINEs), a class of mammalian-specific retrotransposon, have been implicated as possible candidates for these booster elements (9). Enrichment of LINEs has specifically been found in the region of the X IC, and regions containing genes escaping XCI have significantly lower LINE content than on the rest of the X (10). Furthermore, there is evidence suggesting selectivity for LINEs in regions of the X chromosome containing inactive genes (11) and that these LINEs participate in the formation of a silent nuclear compartment induced by Xist RNA (12). More recent sequence analyses have identified differential content of a variety of repeat elements associated with the inactivation status of X-linked genes. These include increased content of L1s, L2s, mammalian interspersed repeat elements (MIRs) and ERV long terminal repeats (LTRs) with inactivated genes and increased Alu density associated with genes escaping XCI (13).

Furthermore, motifs that are present within L1 LINEs have been found to occur at significantly different frequencies between genes that are subject to and those that escape XCI on the X chromosome (14). The density of certain types of tandem repeat motif has also been shown to correlate with XCI, with [GATA]_n repeats associated with genes escaping XCI, and [AT]_n, [AC]_n, and [AG]_n, showing significant enrichment on the X versus the autosomes (15). However, because the X chromosome has a unique evolutionary history (16), markedly different repeat content compared with the autosomes (10) and a nonrandom distribution of genes escaping XCI (5), these confounding factors make it unclear whether the distribution of these repetitive elements is specifically related to the XCI process.

Unbalanced X;autosome translocations provide a unique system for the study of the spread of XCI. The ability of the XCI signal to spread into neighboring autosomal chromatin and silence autosomal genes in cis in X;autosome translocation carriers has been demonstrated in both mice and humans (17–22). Exactly as occurs on the X chromosome, it has been shown that many autosomal genes attached in cis to the Xi become inactivated, whereas others escape XCI. Although complicated owing to the potential effects of selection that can remove cells with a non-favorable spread of XCI, multiple lines of evidence show that the propagation of the XCI signal is generally less efficient in autosomal chromatin compared with the X. These observations have led to the suggestion that autosomal DNA might therefore be deficient in certain ‘booster elements’ found on the X chromosome that are required for normal spreading of XCI (9,23–26).

Given the unusual sequence and repeat properties of the X chromosome (27) that represent a strong confounder in any attempts to associate sequence features with the spread of XCI, another advantage of X;autosome translocations is that they allow a relatively unbiased system for such studies. Motivated by evidence from analyses of the X chromosome suggesting a possible role for repeat elements in the spread of XCI, in this study we have performed quantitative genome-wide DNA methylation profiling in six unbalanced X;autosome translocations to test the hypothesis that the spread of XCI is sequence dependent. Using promoter methylation levels as a proxy to measure the spread of XCI in each X;autosome translocation, we have determined the X inactivation status of 1050 autosomal genes translocated onto the Xi, providing the first detailed assessment of the spread of XCI over entire chromosome segments. Subsequent sequence analysis comparing the prevalence of common repeats and DNA motifs in translocated autosomal genes that are subject to versus those that escape the spread of XCI shows that the density of many types of common repeat is significantly correlated with the spread of XCI into autosomal chromatin, recapitulating similar observations made on the X chromosome. We also identify motifs enriched around autosomal genes silenced by the spread of XCI and show that these motifs map predominantly to L1 elements, are enriched on the X chromosome compared with the autosomes and occur at significantly higher density around X-linked genes that are silenced on the Xi compared with those that escape XCI. Finally, we show that motifs previously identified as enriched around X-linked genes that are subject to XCI also occur at significantly increased density around autosomal genes that are silenced by the spread of XCI. These data confirm the sequence-specific nature of the spread of XCI, supporting a role for LINEs in the XCI process.

RESULTS

Increased DNA methylation correlates with the spread of XCI in X;autosome translocations

We used the Illumina 450k bead array platform to perform quantitative profiling of DNA methylation at ~482 000 CpGs distributed across the genome. In agreement with previous studies (20), we observed large and highly significant increases in DNA methylation specifically on the translocated autosomal segment in each X;autosome translocation carrier, showing that as on the X chromosome, the spread of XCI into autosomal DNA results in wide-scale epigenetic changes (Figs 1 and 2). Comparison of translocation breakpoints as localized by array CGH showed complete agreement with the most proximal increases in methylation observed on the trisomic autosomal segment in each case. However, the extent of change in DNA methylation levels varied among the different cases analyzed. In the X;10 translocation in Cases 3a/3b, significant increases in DNA methylation spread across most of the 39.5 Mb of translocated 10q chromatin, although the magnitude of these increases showed a strong inverse relationship with increasing distance from the translocation breakpoint (Fig. 1). In contrast, in Case 2 in which ~50 Mb of 7q chromatin is translocated onto the Xi, the extent of spread of methylation changes was mostly limited to the first ~16 Mb of the translocated 7q segment. Marked discontinuity in the spread of XCI was also evident in this case, with a distal cluster of highly methylated genes in 7q35 separated from these more proximal methylated loci by an intervening region of ~15 Mb in size in which almost no changes in methylation occur (Fig. 2).

Previous studies have shown that on the X chromosome, the increase in DNA methylation at gene promoters is directly correlated with the probability of silencing by XCI (4). We also observed that in the majority of X;autosome translocation cases, increased methylation of translocated autosomal genes occurred specifically at promoter regions, with 2392 of 3844 (62.2%) hypermethylated probes located within ±2 kb of the transcription start site (TSS) of RefSeq genes. Accordingly, in order to classify autosomal genes as either subject to or escaping the spread of XCI, we determined the increase in the level of DNA methylation within the promoter region on each translocated autosomal segment as a proxy to measure the spread of
XCI in each translocation case. Using a 1-kb sliding window, we identified clusters of multiple probes that showed significantly increased methylation levels on the translocated chromosome as a signature of silencing by a spread of XCI. This approach showed extremely high specificity for identifying hypermethylated regions on the translocated segments of autosomes. For example, in Case 3a who carries a t(X;10), 1222 windows on the 39.5-Mb translocated segment of 10q23.3-qter were classified as hypermethylated, whereas none of the 13 700 probes on the remaining 96-Mb disomic portion of chromosome 10 exceeded our thresholds (≥3 probes within a 1-kb window, each with $P$-value difference ≥0.1 and $P$ ≤ 0.01, Fig. 1).

Similar results were seen in all other cases studied. In total, 373 autosomal genes were classified as hypermethylated and scored as silenced by the spread of XCI, and 607 were classified as normally methylated and scored as escaping the spread of XCI (Table 1, Supplementary Material, Table S1). These genes were subsequently used as the basis to search for sequences that showed a significant association with XCI.

We observed high concordance for genes scored as active and inactive between relatives carrying the same X;autosome translocation. In Cases 3a and 3b, of the total of 130 genes that were classified as inactive on the translocated segment of chromosome 10 in either of the two individuals, 100 (77%) were scored concordantly between the two. Similarly, 106 of the 125 (85%) genes classified as active in either of the two individuals were scored concordantly. Inspection of the raw data showed that in most cases, genes that were scored discordantly between Case 3a and 3b in fact showed very similar methylation patterns in both individuals. However, at most of these discordant sites, the methylation changes lay close to our statistical thresholds, and thus owing to small fluctuations in the data, the change in $\beta$-values did not meet our thresholds for hypermethylation in one of the two individuals.

Expression-based studies of the spread of XCI in Cases 2, 3a and 4 have been reported previously (19,20). Of 19 genes tested by either allele-specific RT-PCR or methylation-sensitive restriction digestion of promoter CpG islands, 16/19 (84%) of genes were scored concordantly between these previous analyses and in the present study, suggesting that the use of array-based methylation profiling to characterize the spread if XCI is largely accurate (Supplementary Material, Table S2).
three genes that were scored discordantly, two that had previously been shown not to be expressed from the translocated autosomal segment showed hypermethylation of sites outside their promoter region that resulted in these genes not being classified as inactive by our promoter-centric criteria. *ABLIM1* showed hypermethylation of an intronic CpG island 50 kb downstream

![Figure 2. Variable and discontinuous spreading of XCI in five X;autosome translocation cases. Marked variation in the pattern of methylation changes can be seen in different chromosome regions among the five cases. For example, in Case 5 (d) apart from a single gene, the region from \(~82~\text{to}~87~\text{Mb}\) appears almost epigenetically unchanged compared with controls, indicating that most of this segment escapes the spread of XCI. In contrast, the entire 10-Mb region distal to this shows multiple sites of significantly increased methylation, indicating that many genes within this region are silenced by the XCI signal. Each plot shows the \(-\log_{10} P\)-value from the Weisberg \(t\)-test comparing \(\beta\)-values for each probe in the translocation carrier with those in the other six or seven individuals studied who carry a normal chromosome complement for that same region. Boxed regions on each chromosome ideogram and gray-shaded regions on each plot indicate the trisomic autosomal segment that is translocated onto the inactive X.](https://academic.oup.com/hmg/article-abstract/23/5/1224/634637)
Table 1. Description of eight carriers of six unbalanced X;autosome translocations assayed

<table>
<thead>
<tr>
<th>Case</th>
<th>Karyotype</th>
<th>X inactivation ratio</th>
<th>Phenotype</th>
<th>Translocated trisomic autosomal segment</th>
<th>Autosomal breakpoint, Mb from pter</th>
<th>Approximate size of translocated autosomal segment</th>
<th>Number of RefSeq genes on translocated autosomal segment</th>
<th>Number of translocated autosomal RefSeq genes assayed (% of total)</th>
<th>Number of genes classified as active (% of genes assayed)</th>
<th>Number of genes classified as inactive (% of genes assayed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46,X,der(X)t(X;3)(q27.3;p26.2)</td>
<td>100</td>
<td>POF, SA</td>
<td>3pter-p26.2</td>
<td>3.56–3.68</td>
<td>3.6 Mb</td>
<td>7</td>
<td>4 (57%)</td>
<td>1 (25%)</td>
<td>3 (75%)</td>
</tr>
<tr>
<td>2</td>
<td>46,X,der(X)t(X;7)(q27;q31)</td>
<td>100</td>
<td>DD, PA</td>
<td>7q31.1-qter</td>
<td>108.64–108.66</td>
<td>50.2 Mb</td>
<td>362</td>
<td>214 (59%)</td>
<td>164 (77%)</td>
<td>50 (23%)</td>
</tr>
<tr>
<td>3a</td>
<td>46,X,der(X)t(X;10)(q26.3;q23.3)</td>
<td>100</td>
<td>PA, niece of case 3b</td>
<td>10q23.3-qter</td>
<td>95.80–95.82</td>
<td>39.5 Mb</td>
<td>353</td>
<td>254 (72%)</td>
<td>143 (56%)</td>
<td>111 (44%)</td>
</tr>
<tr>
<td>3b</td>
<td>46,X,der(X)t(X;10)(q26.3;q23.3)</td>
<td>100</td>
<td>POF, SA, aunt of case 3a</td>
<td>10q23.3-qter</td>
<td>95.80–95.82</td>
<td>39.5 Mb</td>
<td>353</td>
<td>254 (72%)</td>
<td>143 (56%)</td>
<td>111 (44%)</td>
</tr>
<tr>
<td>4</td>
<td>46,X,der(X)t(X;11)(q26;p12)</td>
<td>100</td>
<td>DD, menstruated at 21 years Short stature</td>
<td>11pter-p12</td>
<td>41.94–41.97</td>
<td>41.95 Mb</td>
<td>421</td>
<td>245 (58%)</td>
<td>116 (47%)</td>
<td>129 (53%)</td>
</tr>
<tr>
<td>5</td>
<td>46,X,der(X)t(X;14)(q13;q24) [27]/45,X[3]</td>
<td>100</td>
<td>DD, menstruated at 21 years Short stature</td>
<td>14q24.3-qter</td>
<td>74.39–74.79</td>
<td>31.8 Mb</td>
<td>335</td>
<td>294 (88%)</td>
<td>206 (70%)</td>
<td>88 (30%)</td>
</tr>
<tr>
<td>6a</td>
<td>46,X,der(X)t(X;20)(q26.3;q13)</td>
<td>&gt;95</td>
<td>DD, menstruated at 16 years, sibling of case 6b Subtelomeric</td>
<td>20q13.33-qter</td>
<td>&lt;1 Mb</td>
<td>50</td>
<td>39 (78%)</td>
<td>35 (90%)</td>
<td>4 (10%)</td>
<td></td>
</tr>
<tr>
<td>6b</td>
<td>46,X,der(X)t(X;20)(q26.3;q13)</td>
<td>&gt;95</td>
<td>Menstruated at 16 years, sibling of case 6a Subtelomeric</td>
<td>20q13.33-qter</td>
<td>&lt;1 Mb</td>
<td>50</td>
<td>39 (78%)</td>
<td>35 (90%)</td>
<td>4 (10%)</td>
<td></td>
</tr>
</tbody>
</table>

The approximate percentage of cells in which the der(X) is inactivate. Only genes with ≥ 3 probes within 2 kb of their TSS that had a mean β-value of < 0.5 in controls were included in the analysis. POF, premature ovarian failure; PA, primary amenorrhea; SA, secondary amenorrhea; DD, developmental delay.
of the TSS, whereas LMO2 showed hypermethylation of an intergenic CpG island ~25 kb distal to the 3’ end of the gene in the X;11 translocation carrier.

Comparison of our methylation data with these previous expression analyses also supports the idea that even subtle changes in methylation that we observed on translocated autosomes are likely to be of functional significance. For example, expression of CNTNAP2 was previously shown by allele-specific studies to be reduced by ~70% on the translocated segment of chromosome 7. While this gene was scored as inactive by our criteria, the epigenetic changes observed in this case were subtle, with only 5 of the 17 probes in this gene’s promoter showing β-value increases of >0.1 compared with controls, and only one showing an increase of >0.2.

Repeat content of translocated autosomal genes shows significant correlation with susceptibility to XCI

Given previous studies of the X chromosome that have suggested a link between XCI and certain classes of common repeat (9,10,13–15), we analyzed the density of repeats associated with 373 autosomal genes classified as silenced and 607 classified as escaping the spread of XCI. Analyzing regions within ±50 kb of these genes, we observed several different repeat types that showed significant relative enrichments in either the inactive or active set of autosomal genes. Using repeat classifications annotated by RepeatMasker, we observed significantly higher densities of eight different repeat families (L1, L2, CR1, ERVL, MaelR, MIR, MER1 and MER2) around inactive autosomal genes, whereas two different repeat types (Alu and simple repeats) were significantly enriched around active autosomal genes (Fig. 3 and Supplementary Material, Table S3).

In addition to this combined analysis in which we considered all 980 translocated autosomal genes, we also performed individual analysis of repeat content for each translocated autosomal segment in a chromosome-specific manner (excluding cases 1 and 6, which both have only a small number of translocated autosomal genes). This analysis showed that L1 and L2 are the only two classes of repeat that are significantly enriched around inactive translocated genes in all of the four X;autosome translocations examined. In contrast, other repeat types either show no significant enrichments in some of the translocation cases or significant trends in the opposite direction to that observed in the combined analysis (Supplementary Material, Fig. S1).

Given a previous report that [GATA]n repeats are associated with genes escaping XCI (15), we tested the prevalence of [GATA]n repeats (including [ATAG]n, [TAGA]n, and [AGAT]n repeats and their reverse complements) annotated by Tandem Repeats Finder in hg18 (http://genome.ucsc.edu/) within ±50 kb of translocated autosomal genes, but observed no significant difference in density between active and inactive genes (data not shown).

We also observed a highly significant tendency for inactive autosomal genes to occur preferentially in Giemsa (G)-dark versus G-light bands (Supplementary Material, Fig. S2). The mean Giemsa-staining score of inactive genes was 32.9 compared with 10.4 for autosomal genes escaping XCI (P = 8.6 × 10<sup>-22</sup>, Wilcoxon Rank Sum test).

Using published RNAseq data from 58 CEU HapMap B-cell lines (28), we compared the relative expression levels in control individuals of active and inactive translocated autosomal genes. This analysis showed that, in control individuals, autosomal genes that escape the spread of X inactivation show a mean of 1.5-fold higher expression compared with autosomal genes that are silenced by the spread of X inactivation (P = 0.006, Wilcoxon Rank Sum test). However, considering all genes in the genome, there is a significant negative correlation between gene expression levels and G-band staining intensity (r = -0.116, P = 2 × 10<sup>-7</sup>, Spearman rank correlation), with genes in G-dark bands showing lower expression levels on average than those in G-light bands (Supplementary Material, Fig. S3). As translocated autosomal genes silenced by X inactivation occur more frequently in G-dark bands, this difference in gene expression levels between active and inactive translocated autosomal genes would be expected simply due to this fact alone.

We also observed that gene density in regions containing active translocated autosomal genes is on average higher than that in regions containing inactive translocated autosomal genes (mean density in active regions is one gene per 53.5 kb, mean density in inactive regions is one gene per 90.4 kb). As with the analysis of gene expression levels, this observation mirrors the strong genome-wide tendency for higher gene densities to occur in G-light bands in general.

Motifs associated with inactive autosomal genes are enriched on the X chromosome, map predominantly to recently active L1 elements and occur at higher densities around inactive X-linked genes

We used CisFinder (29) to identify 52 consensus motifs ranging from 9 to 21 bp in length that were significantly enriched (1% false discovery rate (FDR)) within ±50 kb of inactive autosomal genes compared with those that remain active (Supplementary Material, Fig. S4). Fourteen of these 52 motifs were comprised solely of AT repeats and therefore likely represent false positive associations (30) and were thus excluded from further analysis. In order to investigate the properties of the remaining 38 motifs that were enriched around inactive autosomal genes, we mapped their locations genome-wide and observed that 31 of these motifs showed a significantly increased density (P < 0.05, permutation test) on the X chromosome compared with the autosomes (mean enrichment on chrX, 1.61-fold, range 0.95–2.22) (Fig. 4A, Supplementary Material, Table S4). We studied the distribution of these motif positions on the X chromosome and observed that, as a group, these motifs tend to occur at significantly higher densities within ±50 kb 242 Refseq genes that were previously identified as subject to XCI compared with 52 Refseq genes that escape XCI (mean enrichment within ±50 kb of Refseq genes that are subject to XCI, 1.6-fold, range 0.5–3) (5) (Fig. 4A, Supplementary Material, Table S4).

We then compared the positions of these motifs on the X chromosome with Repeatmasker annotations and found that for 31 of 38 of these motifs, the most common overlap was with L1 elements (Fig. 4B). Six of the remaining seven motifs were predominantly located in non-repetitive sequence, whereas one motif mapped mostly to MIRs (Supplementary Material, Table S4). Considering the positions of all 38 motifs on the X chromosome, we compared the fraction of matches located within the various different sub-families of L1s to that expected under a random distribution based on their relative prevalence on
Significantly different repeat densities between active and inactive translocated autosomal genes. Each plots show the fraction of base pairs within ± 50 kb of each gene (transcription start to transcription end, extended 50 kb both 3' and 5') of 607 active and 373 inactive translocated autosomal genes that overlap common repeats annotated by RepeatMasker. Each plot shows fold enrichment of repeats for inactive genes versus active genes (based on median), and the Bonferroni-corrected P-value from a Wilcoxon Rank Sum test. A list of all repeat types is shown in Supplementary Material, Table S3.

Figure 3.
the X chromosome. We observed a strong enrichment above the null for these motifs to occur within L1PA and L1HS repeats, corresponding to L1s that were active specifically in primates and recent human history (Fig. 4C).

We also used CisFinder to search for motifs enriched around inactive autosomal genes after first masking common repeats but only identified motifs composed exclusively of low-complexity AT sequences, likely representing nonspecific enrichments.

**Measurement of relative enrichments of sub-regions of repeats between active and inactive translocated autosomal regions**

As motif-finding approaches could be biased against finding sequences within evolutionarily older elements as a result of their increased nucleotide divergence, we applied an alternate approach to study the relative representation of sub-regions of...
the consensus sequences of repetitive elements. By using the positional information of each repeat fragment within the consensus sequence of that repeat family, we searched for regions within each repeat consensus sequence that were differentially represented within ± 50 kb of active and inactive translocated autosomal genes (Fig. 4, Supplementary Material, Fig. S5). Although most repeats tested showed nominally significant differences in certain sub-parts of their consensus, the most significant difference observed among all repeats tested occurs at position 7733 in the 3′ end of L1 consensus sequence ($P = 6.9 \times 10^{-13}$), with this region occurring at 1.25-fold higher frequency around active autosomal genes compared with inactive genes.

**Motifs previously associated with inactive X chromosome genes are enriched around autosomal genes silenced by the spread of X inactivation**

A previous study performed a sequence analysis of genes in Xp22 of known XCI status, identifying motifs that are enriched around X-linked genes that are either subject to or escape XCI (13). Taking advantage of this dataset, we measured the occurrence of 105 12mers that showed significant association ($P < 0.05$) with the XCI status of X-linked genes in sequences within ± 50 kb of the set of 980 translocated autosomal genes. As a group, we observed that motifs that are enriched around genes subject to XCI in Xp22 occurred at significantly higher density around autosomal genes that were silenced by the spread of inactivation compared with autosomal genes that remain active (mean of 1.54-fold enrichment around inactive autosomal genes, range 0.52–3.25, $P = 0.00059$, Wilcoxon Rank Sum test) (Supplementary Material, Fig. S6, and Table S5).

**DISCUSSION**

Several previous sequence analyses of the X chromosome have shown correlations between a number of different types of common repeat element and the location of genes that either are subject to or escape the XCI process (10,13–15). However, the human X chromosome has both a highly nonrandom distribution of genes escaping XCI (5) and a markedly different repeat content compared with the autosomes (10). As a result, it has been unclear whether the distribution of these repetitive elements is specifically linked with XCI or instead might be a consequence of the unusual evolutionary history of the X, which was produced by successive additions of autosomal chromatin onto distal Xp (16). Furthermore, it is also known that retrotransposons can show an integration bias that depends on both the transcriptional activity and epigenetic state of a chromosomal region (31–33). Thus, because X-linked genes that are silenced by XCI have spent one-third of their evolutionary history in a repressed transcriptional state and with a correspondingly different set of epigenetic marks compared with genes that escape XCI, it is possible that different types of transposon would preferentially integrate into certain regions dependent on their XCI status. As a result, the observed correlation of certain repeats with XCI status might be a consequence, rather than a cause, of XCI. Our investigation of the spreading of XCI into autosomal DNA provides a unique system that is free from these confounding factors, allowing a relatively unbiased assessment of sequence features that are associated with the spread of XCI. We hypothesized that if the spread of XCI is influenced by sequence features such as L1 density, in cases of X;autosome translocation, these same sequences should occur at increased frequency around autosomal genes that are silenced by the spreading of X inactivation compared with those that remain active.

Using DNA methylation profiling in six different X;autosome translocations, we have determined how the spread of XCI influences the epigenetic state of 1050 autosomal genes when translocated onto an Xi. Consistent with our hypothesis, sequence analysis of 980 of these autosomal genes that showed consistent methylation patterns revealed that their differential susceptibility to XCI is correlated with the local density of several classes of common repeat. Remarkably, we observed that many of the same repeat types that occur at significantly different densities between active and inactive autosomal genes have been previously identified as showing correlations with the expression status of X-linked genes on the inactive X chromosome. For example, Wang et al. (13) reported that increased densities of L1, L2, ERV1 and MIR repeats are associated with X-linked genes that are subject to XCI. Our analysis of spreading of X inactivation in autosomal DNA also identifies these same repeats as enriched around inactive autosomal genes. Increased densities of Alu repeats have been associated with X-linked genes that escape XCI (13), and we similarly observed a relative enrichment for Alu elements around autosomal genes that escape the spread of XCI.

Our comparative motif analysis provides further evidence in support of a role for L1s in the spread of XCI. Utilizing two complementary approaches, we first performed a de novo motif analysis to identify motifs enriched around autosomal genes silenced by the spreading of XCI. Mapping the positions of these motifs genome-wide, we found that this set of motifs associated with inactive autosomal genes shows distinct properties that are consistent with a role in the XCI process: (i) they are enriched on the X chromosome compared with autosomes, (ii) they are significantly enriched around X chromosome genes subject to XCI compared with those that escape XCI and (iii) they map predominantly to L1s, with an over-representation in evolutionarily younger L1P and L1HS sub-families. Taking the reverse approach, we then analyzed a set of motifs previously associated with the XCI status of genes on the X chromosome (13). Again, consistent with the hypothesis that the spread of X inactivation shows sequence specificity, we found that motifs associated with inactive genes on the X chromosome also occur at significantly higher densities around translocated autosomal genes that silenced by the spreading of X inactivation.

It should be noted, however, that our motif analysis may be inherently biased toward finding enrichments of specific sequences that are located within evolutionarily younger repeats, such as L1P and L1HS elements, simply because these recent repeat insertions have had less time to accumulate divergent mutations and thus have higher similarity to each other compared with more ancient repeat types, such as L2s. We attempted to control for this by complementing our motif analysis using a second approach that searched for sub-regions of repeat consensus sequences that were associated with inactivation status, independent of their exact sequence (Supplementary Material, Fig. S5). The most significant difference observed among all repeats tested in this analysis occurs in the 3′ end of L1 consensus sequence, with this
region observed significantly more frequently around active translocated autosomal genes. This mirrors similar observations made on the X chromosome, which showed that motifs associated with inactive X-linked genes are also relatively depleted in the 3’ UTR of L1s (14).

Our observations indicate that the spread of XCI into autosomal DNA is associated with local repeat density in a way that largely recapitulates similar observations made on the X chromosome, supporting a role of local repeat content in the spreading of X inactivation. Our results are therefore consistent with the Lyon ‘repeat hypothesis’ (9) and indicate a potential functional role for L1 elements in the XCI process. Our results are also similar to those obtained in studies of transgenic mouse cell lines in which Xist transgenes were inserted into different autosomal locations. Observations in this system suggest that genes residing in LINE-rich chromosomal regions show much more efficient silencing than those in LINE-poor domains and that these repeat elements likely aid in heterochromatin formation within the nuclear compartment defined by Xist RNA (12). While ChIP-seq studies of chromatin marks associated with the inactive X during the onset of XCI have reported no significant increase of H3K27me3 over repetitive sequences such as LINES, such conclusions are inherently limited owing to the use of read mapping parameters in these studies that only considered uniquely mappable reads (34).

Interestingly, studies of regions of the genome containing genes that are imprinted or undergo random mono-allelic expression have also identified differential repeat content in these regions compared with bi-allelically expressed genes in a way that mirrors our observations of XCI. This includes an increased density of evolutionarily recent L1 insertions (35) and a reduction in local SINE content in association with mono-allelic expression. Given the parallels between these different modes of epigenetic silencing, it is tempting to speculate that there might therefore be shared mechanisms underlying these phenomena.

An alternative hypothesis that could potentially explain the association of differential repeat content with local susceptibility to XCI is that these repeats are not themselves the ‘booster elements’ that functionally propagate the spread of XCI (8), but instead are nonspecific markers of chromosomal regions that have inherently different properties, including their probability of epigenetic silencing by the spread of XCI. For example, it is well known that chromosomal banding patterns produced by Giemsa-staining correspond to a number of underlying differences, with G-light bands being more GC-rich, having a higher gene density and transcriptional activity, reduced chromatin compaction and earlier replication, and containing a higher density of Alu repeats and lower LINE density than G-dark bands (38,39). Many of these properties of G-light bands are traditionally associated with euchromatic regions, whereas many features of G-dark bands such as higher chromosome compaction, later replication and lower transcriptional activity are more reminiscent of heterochromatin. Indeed, we observed that autosomal genes silenced by XCI show a strong preference to be located in G-dark bands. A simple hypothesis therefore that unifies our observations of differential content of common repeats, DNA motifs, expression levels and chromosomal banding patterns is that the XCI signal is more likely to result in stable inactivation of a gene that lies in an environment that already possesses features associated with silent chromatin, which by nature will tend to be G-dark bands that are LINE/AT-rich, and Alu poor. Conversely, genes located in more euchromatic G-light bands that are inherently more Alu and GC-rich will have an increased tendency to escape the spread of XCI. Further studies of additional X-autosome translocations will be necessary to confirm this hypothesis.

It should be noted that the XCI patterns we observe in adult patients likely result from the combined influence of the initial spreading and subsequent maintenance of XCI into autosomal chromatin, and also the subsequent effects of post-XCI selective pressures that might remove cells that have a selective or growth disadvantage because of the presence of functionally trisomic dosage-sensitive autosomal genes that escaped the spread of XCI. Without performing studies of XCI in the developing embryo when such selection would likely occur, we are unable to distinguish between these possibilities, but it is highly plausible that such selective effects could have a strong influence on our observations. However, we do note that studies of the behavior of mouse X-autosome translocations during early development suggest specifically that it is the primary spreading of the XCI signal that is disrupted in such cases (26). All of the studies we performed utilized DNA derived from peripheral blood, and thus we can exclude epigenetic artifacts induced by immortalization and cell culture (40).

As our study included two pairs of relatives who carried the same translocation chromosome, a simple prediction of the hypothesis that the spread of XCI is sequence dependent is that there should be concordance in the pattern of autosomal gene silencing between relatives. Within the limits of our methodology, we indeed observed very similar patterns of methylation across >300 translocated genes that were common between related individuals. Previous studies have also suggested that the location of the X chromosome breakpoint might influence the extent of spread of XCI into autosomal chromatin (23). However, in our six cases, we observed no clear relationship between these parameters. Finally, as array CGH also localized the breakpoints on the X chromosome in each translocation case, we utilized our methylation data to look for possible disturbances of normal XCI of X-linked genes near the breakpoints that might result from position effects. However, we observed no significant changes in methylation of X-linked genes near the translocation breakpoint in any of the eight cases analyzed (data not shown).

In conclusion, our large-scale epigenetic analysis of six X-autosome translocations identifies L1 repeats as the most strongly associated sequence feature with the extent of spread of the XCI signal. Although L1s are undoubtedly not the sole determinant of the susceptibility of a locus to silencing by XCI, our findings strongly support the notion that, as proposed by Mary Lyon (9), L1s act as ‘booster elements’ (8) that aid in propagating the X inactivation signal in cis.

MATERIALS AND METHODS

Study population

We studied eight individuals who each carried a derivative X chromosome resulting from an unbalanced X-autosome translocation. Two of the chromosomes were each analyzed in two
relatives carrying the same inherited X;autosome translocation, and thus our cohort comprises six independent derivative X chromosomes, summarized in Table 1. Although all X chromosome breakpoints occur within Xq, they are not clustered and all six translocations have different partner autosomes. Each carrier showed completely skewed X inactivation by methylation-sensitive analysis of the androgen receptor locus, with preferential silencing of the derivative chromosome in >95% of cells. Array comparative genomic hybridization using either custom 1 Mb BAC array or Agilent 44k or 180k oligonucleotide arrays had previously been used to map the translocation breakpoints. Full details are reported in (41). Analyses of the spread of XCI using a variety of methods in three of those individuals (Cases 2, 3a and 4) have been reported previously (19,20).

DNA methylation profiling
Genomic DNA was extracted from peripheral blood samples from each individual and extracted by salt/ethanol precipitation. Genome-wide measurement of CpG methylation levels was performed by hybridization of bisulfite-converted DNA from each case to Infinium HumanMethylation450 BeadChips, according to the manufacturer’s protocol (Illumina Inc., San Diego, CA, USA). All eight samples were processed as a single batch and hybridized on the same slide. Array data were processed using the Methylation Module of GenomeStudio v1.8 software using default parameters. Probes with a detection P-value of >0.01 were removed (mean n = 505 per sample). Owing to the variable size of the deleted X chromosome material, variable size and origin of the trisomic autosomal material, and the variable spread of XCI in each case, all of which result in significant changes to the global distribution of methylation values, no inter-array normalization was performed. However, all samples showed raw β-value profiles that had almost identical means (mean β-value of all arrays 0.4996, range 0.4966–0.5023), suggesting no significant technical bias between samples.

Classification of hypermethylated translocated autosomal genes
As the translocations we studied involved movement of different segments of autosomal chromatin onto the inactive X, to identify methylation changes resulting from the spread of XCI into attached autosomal DNA, we compared methylation levels among the eight individuals in our study population. To identify translocated autosomal genes with significantly elevated methylation, we compared probe β-values in each translocation carrier against those obtained in the other six or seven individuals studied who did not have a translocation of that same autosome. Thus, the normal disomic chromosomes present in the other cases of X;autosome translocation were considered as controls for each individual.

As signals from single probes are potentially unreliable because of the possibility of SNPs underlying either the probe binding site or at the CpG being assayed, we adopted a sliding window approach that only considered clusters of multiple concordant probes as robust signals. As many sites show constitutively high levels of methylation in normal individuals and thus by definition cannot undergo large increases in methylation, in order to avoid inappropriately classifying these loci as active (i.e. showing no significant increase in methylation), we first excluded any probe that showed a mean β-value of ≥0.5 in controls. We then used the Weisberg t-test for outliers to assign P-values to each probe by comparing the β-value on the translocated chromosome with β-values obtained in the controls. The difference in β-value between the translocation carrier and the mean of the controls was also calculated for each probe. To robustly identify clusters comprising multiple significant sites, we applied a 1-kb sliding window based on each probe start coordinate, excluding any windows that contained ≤2 probes. Loci were classified as hypermethylated when ≥3 probes within a 1-kb window each had P-value of ≤0.01 and β-value difference of ≥0.1. The remaining windows were classified as normally methylated. A total of 1528 autosomal RefSeq genes are located on the six independent X;autosome translocations, of which 1050 were assayed (69%). The majority of autosomal genes that were not assayed showed high promoter methylation levels in controls (mean β-value > 0.5), presumably corresponding to genes with tissue-specific expression patterns that are not transcribed in blood.

We downloaded RefSeq genes annotations for the hg18 assembly and intersected these with the probe positions (http://genefacilities.ucsc.edu/). Based on the known promoter hypermethylation of X-linked genes silenced by XCI, we classified any Refseq genes that had one or more hypermethylated loci within ±2 kb of their TSS as inactive. Genes with no hypermethylated probes within ±2 kb of their TSS were scored as active. A full list of all probes mapping to the translocated autosomal segments, with β-values, P-values and their classification status based on sliding window analysis is shown in Supplementary Material, Table S6. Genes that had <3 probes within ±2 kb of their TSS were removed (total n = 102 across all six translocated regions). For subsequent sequence analysis, where two relatives carried the same translocation (Cases 3a/3b and 6a/6b), only genes that were scored concordantly between the two individuals were utilized (total n = 980 across all six translocated regions, Supplementary Material, Table S1).

Sequence analysis and expression levels of translocated autosomal genes
We measured the content of common repeat elements around 373 inactive and 607 active translocated autosomal genes. Data on Refseq genes, Giemsa bands and common repeats annotated by Repeatmasker in the hg18 assembly were downloaded from the UCSC genome browser (http://genome.ucsc.edu/). Genes were extended ±50 kb both 3′ and 5′, active and inactive set each merged down to form a non-redundant set, and then any active regions that overlapped inactive regions were removed. Intersections were performed between repeats and genes using Galaxy (https://main.g2b2.psu.edu/). In order to avoid potential false positive enrichments associated with low frequency repeat elements, repeat types that comprised <0.1% of sequence in either the active or inactive set were excluded. For enrichment analysis of repeat sub-regions within ±50 kb of active and inactive genes, the position of each repeat fragment within the
consensus sequence of that repeat family was taken from the RepeatMasker annotation, and the total number of occurrences of each nucleotide position of the consensus calculated using BEDTools v2.17 (42). This analysis was performed for eight of the ten repeat families that were found at significantly different density between active and inactive regions. We excluded CR1 repeats, which are too rare to give meaningful data, and simple repeats, which represent a diverse collection of sequences and for which there is no single consensus sequence. P-values per base were calculated using a chi-square test of observed versus expected values, based on the null hypothesis of equal representation at each position of the repeat consensus sequence between active and inactive regions. For generating the plots shown in Supplementary Material, Figure S5, counts within active regions were normalized against those in inactive regions based on the total count of all bases observed to allow easy visual comparison.

For association testing between gene inactivation and Giemsa staining, we used the G-band staining intensity score (a value ranging from 0 to 100) in the UCSC Cytoband track and overlapped these scores with each translocated autosomal gene.

To determine steady-state mRNA expression levels, we used published RNAseq data from 58 CEU HapMap B-cell lines (28). Expression levels per gene were calculated based on the median RPKM value in this population based on all annotated isoforms of a gene. Gene expression levels within G-bands were calculated by intersection with the UCSC Cytoband track.

Gene densities within translocated autosomal regions were calculated by extending the coordinates of the 373 inactive and 607 active translocated autosomal genes defined earlier by 50 kb both 3′ and 5′, and merging these down to a non-redundant set of ‘active’ and ‘inactive’ intervals.

De novo motif analysis associated with inactive autosomal genes

We used CisFinder (http://lgsun.grc.nia.nih.gov/CisFinder/) (29) to perform de novo identification of motifs that were significantly enriched around inactive autosomal genes, as follows. We extended the coordinates of the 373 inactive and 607 active translocated autosomal genes we defined earlier by 50 kb both 3′ and 5′ and merged these down to a non-redundant set of ‘active’ and ‘inactive’ intervals. Regions in the active set that overlapped the inactive set were then removed, generating a ‘inactive’ test set of 139 intervals covering a total of 54.3 Mb, and a ‘active’ control set of 120 intervals covering a total of 40.6 Mb that were used as an input for motif analysis. We applied default settings, considering repeats, with minimum threshold of 1.2-fold enrichment. Individual motifs were clustered by CisFinder based on sequence similarity and a 1% FDR applied to define motifs that were significantly enriched around inactivated autosomal genes.

Investigating the genomic distribution of motifs identified as enriched around inactive translocated autosomal genes

Prior to conducting further analysis of the distribution of consensus motifs identified by CisFinder as significantly enriched around inactive translocated autosomal genes, we first excluded 14 of the 52 motifs that were composed predominantly of AT repeats and mapped the locations of each of the remaining 38 motifs genome-wide using the fuzznuc function in EMBOSS (version 6.5.7) (43). We then calculated the relative density of each motif on the X chromosome compared with the autosomes, applying a permutation test (1000 replicates) to calculate significance versus random expectation. We determined the fraction of occurrences of each motif on the X chromosome that overlapped common repeats by at least 1 bp as defined by RepeatMasker.

Motifs were then tested for significant association with genes on the X chromosome that are silenced by XCI compared with those that escape XCI. We used results of a previous study of XCI that defined the XCI status of 624 X-linked ESTs using a panel nine somatic cell hybrids containing a single inactive X chromosome (5). However, when mapping these ESTs to hg18, we found that a significant number did not match well with current gene annotations. We therefore decided to be conservative in our approach, including only those loci that we could map unambiguously to current RefSeq annotations (n = 328). Of these, genes that were expressed in ≥8 of the 9 hybrids tested (n = 52) were considered as escaping XCI, whereas genes that were expressed in ≤1 of the 9 hybrids tested (n = 242) were considered as subject to XCI. Using the same strategy as in our motif analysis of translocated autosomal genes, we extended each of these 294 X-linked genes of known XCI status +50 kb 3′ and 5′, and then merged these down to form a non-redundant set of active and inactive intervals. The relative density of each motif identified as enriched around in-active translocated autosomal genes was measured in these two X chromosome sets.

Testing of motifs significantly associated with XCI status on the X chromosome with translocated autosomal genes

Previous work has identified a set of 12mer motifs that show significant enrichment around genes that are either subject to or escape XCI within Xp22 (14). We used the wordcount function of EMBOSS (version 6.5.7) (42) to measure the occurrence of each of these 12mer motifs that showed significantly biased distributions in Xp22 (P < 0.05) in the total set of non-redundant intervals located within ±50 kb of autosomal genes that are either subject to or escape the spreading of XCI.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS

The authors thank Patricia Jacobs. Array data have been deposited in the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE50837.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by National Institutes of Health grants RDA033660, RHG006696, HD073731 and Alzheimer’s Association grant 2012ALZNIRG69983 to A.J.S.
REFERENCES


