Domain-wide regulation of gene expression in the human genome

Hinco J. Gierman,[1] Mireille H.G. Indemans,[1,4] Jan Koster,[1,4] Sandra Goetze,[2] Jurgen Seppen,[3] Dirk Geerts,[1] Roel van Driel,[2] and Rogier Versteeg[1,5]

[1]*Department of Human Genetics, Academic Medical Centre, University of Amsterdam, 1100 DE Amsterdam, The Netherlands;* [2]*Swammerdam Institute for Life Sciences, University of Amsterdam, 1100 DE Amsterdam, The Netherlands;* [3]*AMC Liver Centre, 1105 BK Amsterdam, The Netherlands*

Transcription factor complexes bind to regulatory sequences of genes, providing a system of individual expression regulation. Targets of distinct transcription factors usually map throughout the genome, without clustering. Nevertheless, highly and weakly expressed genes do cluster in separate chromosomal domains with an average size of 80–90 genes. We therefore asked whether, besides transcription factors, an additional level of gene expression regulation exists that acts on chromosomal domains. Here we show that identical green fluorescent protein (GFP) reporter constructs integrated at 90 different chromosomal positions obtain expression levels that correspond to the activity of the domains of integration. These domains are up to 80 genes long and can exert an eightfold effect on the expression levels of integrated genes. 3D-FISH shows that active domains of integration have a more open chromatin structure than integration domains with weak activity. These results reveal a novel domain-wide regulatory mechanism that, together with transcription factors, exerts a dual control over gene transcription.

[Supplemental material is available at www.genome.org. The microarray data from this study have been submitted to GEO under accession no. GSE6629. Sequences of all integration sites were submitted to GenBank under accession nos. EF214748–EF214837.]

A few groups of adjacent genes in mammalian genomes have been found to exhibit coregulated expression, and examples of such domain-wide control include the Hox clusters (Gould 1997), X chromosome inactivation (Plath et al. 2002), and position effect variegation (PEV) exerted by heterochromatin on adjacent regions (Weiler and Wakimoto 1995). However, it is assumed that the vast majority of human genes are individually regulated by transcription factor complexes.

The Human Transcriptome Map integrated high-throughput expression data measured by SAGE (serial analysis of gene expression) with the human genome sequence, which revealed that the genome consists of many domains of highly and weakly expressed genes (Caron et al. 2001). The highly expressed domains (called ridges) are gene dense, GC rich, and SINE repeat rich, and the genes have short introns, whereas the weakly expressed domains (called anti-ridges) show the opposite characteristics (Lercher et al. 2003; Versteeg et al. 2003). Ridges were also described in the mouse genome (Mijalski et al. 2005) and were found to be relatively conserved compared to the human genome (Singer et al. 2005).

The highly expressed genes in ridges are generally broadly expressed throughout different tissue types (Lercher et al. 2002). However, not all genes in ridges are highly expressed in each tissue type, and tissue-specific regulation of gene expression also occurs in ridges (Versteeg et al. 2003). Adjacent genes in ridges can therefore have very different expression levels. This is in line with genome-wide analyses where expression of individual genes was not found to correlate over distances of more than two genes (Semon and Duret 2006).

Ridges were recently found to be enriched for open chromatin fibers (Gilbert et al. 2004) and active promoters (Kim et al. 2005). Nonetheless, it is not known whether the differential expression in ridges and anti-ridges is due to individual gene regulation, or if domain-wide mechanisms exert an additional effect (for reviews, see Hurst et al. 2004; Sproul et al. 2005). Here we present data that for the first time show that active domains in the genome contribute substantially to the expression of their embedded genes.

Results

Construction and sequencing of clone collection

To ascertain whether chromosomal domains can influence the activity of embedded genes, we studied the expression level of the same reporter gene integrated at many different positions in ridges, anti-ridges, and domains displaying intermediate gene expression. We infected human embryonic kidney cells (HEK293) with a lentiviral construct harboring the GFP gene driven by the ubiquitously expressed human phosphoglycerate kinase (PGK) promoter (Dull et al. 1998; Zufferey et al. 1998). Cells were transduced at a low multiplicity of infection (MOI = 0.03) to favor single integrations. Individual GFP-positive cells were isolated by fluorescence-activated cell sorting (FACS), and equal numbers of clones with low, medium, and high GFP expression were selected for further expansion. More than 100 clones were cultured and analyzed by Southern blotting to select for single integrations (>90% of clones). The integration sites of the viral constructs were PCR-amplified, sequenced, and mapped onto the genome (see Methods and Supplemental Protocol S1). Insertion sites were unequivocally determined in 90 clones that had unique integrations in 21 chromosomes (Fig. 1; Supplemental Table S1).

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-mail r.versteeg@amc.uva.nl; fax 31-20-6918626.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6276007. Freely available online through the *Genome Research* Open Access option.

17:1286–1295 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

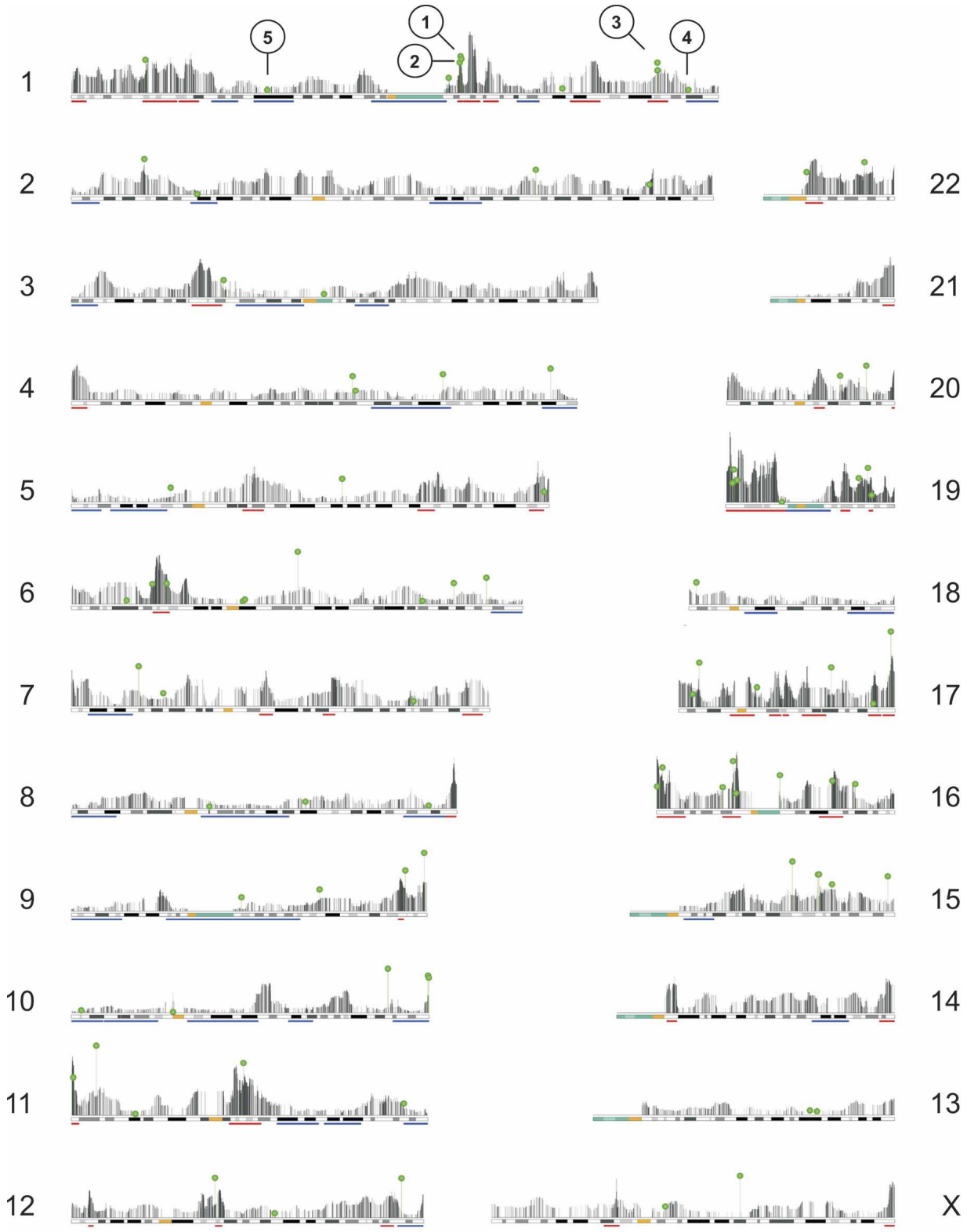

Figure 1. Physically mapped transcriptome profiles of all chromosomes, showing the integration sites and expression levels of all GFP constructs. Giemsa banding is illustrated below each transcriptome map: (yellow) centromere; (green) heterochromatic region; (red) ridges and (blue) anti-ridges are indicated by bars below the Giemsa banding. Black vertical bars represent genes (n = 20,382); their height indicates domain activity for a window of 49 genes (median expression of the surrounding 49 genes in 133 pooled SAGE libraries). Green lollipops indicate integration sites of all GFP constructs (n = 90). The height of each lollipop corresponds to the expression level of the integrated GFP construct. The numbered clones on Chromosome 1 were used for 3D-FISH analysis (see Fig. 6).

FACS analysis of all clones showed a broad range of GFP expression levels. Expression of GFP mRNA of 10 representative clones was analyzed by Northern blotting. The mRNA levels were quantified by PhosphorImaging and normalized to GAPDH levels, which revealed a linear correlation with the levels of GFP fluorescence (Pearson $R^2 = 0.98$, $P = 10^{-8}$; Supplemental Fig. S1), and thereby validated the use of fluorescence as a measure of transcriptional activity of the GFP gene. Analysis of seven representative clones showed that levels of GFP fluorescence were constant over an extended culturing period (Supplemental Fig. S2).

GFP expression in ridges is higher than in anti-ridges

The set of clones enabled us to analyze whether the domain of integration influenced the GFP expression levels. Figure 1 shows the position and GFP expression of all the 90 clones in the expression profiles of the Human Transcriptome Map. We analyzed whether integration in ridges confers a higher expression level to integrated constructs than integration in anti-ridges. We identified 22 clones with integrations in ridges and 14 clones with integrations in anti-ridges. Most clones with ridge integrations displayed high fluorescence, whereas most anti-ridge clones had a low to intermediate fluorescence (Figs. 1, 2A). The average GFP expression of the ridge clones was 4.0-fold higher than of the anti-ridge clones, indicating a strong effect of ridges and anti-ridges on GFP transcription (Fig. 2A; $P = 7.6 \times 10^{-3}$, unpaired *t*-test; for this and all other analyses, GFP fluorescence and moving median values were $\log_2$ transformed to obtain a normal distribution). The average expression of endogenous genes in ridges and anti-ridges differs by a factor of 3.9 (Fig. 2B). The observation that an identical gene integrated in ridges or anti-ridges acquires the relative expression level of the domain of integration suggests a strong regulatory effect of the domain of integration. As the ratio of GFP expression in ridges and anti-ridges (4.0) is comparable to the ratio of endogenous gene expression in ridges and anti-ridges (3.9), a similar domain effect seems to act on the endogenous genes in ridges and anti-ridges.

The different ridges in the genome vary with regard to their median expression level, and hence they may also differ with respect to the effect they have on integrated GFP constructs. To estimate the maximal impact of domains on embedded genes, we compared clones with GFP constructs integrated in the most active and most inactive domains. Domain activity was defined as previously described (Versteeg et al. 2003). In short, the median expression level of the 49 genes surrounding an integration site is determined for each clone (see Methods). The 10 clones with integrations in the most active domains have an 8.4 times higher average GFP expression than the 10 clones with integrations in the least active domains (Fig. 2C, $P = 6.5 \times 10^{-5}$, unpaired *t*-test). These findings show that identical transgenes integrated in different chromosomal regions acquire expression levels that strongly correlate with the expression levels of the domains of integration.

GFP expression correlates with activity of domains up to 80 genes long

The effect of the chromosomal domain on the expression level of the integrated GFP constructs could either result from local effects of genes adjacent to the integration site, or from a mechanism that acts on the domain as a whole. A local effect exerted by nearby active promoters and enhancers would predict a high correlation between expression of GFP and neighboring genes, while a domain-wide effect predicts a high correlation of GFP with the expression of the domain as a whole rather than with the neighboring genes. To analyze both possibilities, we calculated the median gene expression level for window sizes from 1 to 201 genes around each of the integration sites. Gene expression data were obtained by combining 133 different SAGE libraries of various tissue types (see Methods). This median expression level showed a strong correlation with GFP expression, being the highest for window sizes of roughly 19–79 genes around the integration sites (Fig. 3, average $R = 0.50$, $P < 10^{-6}$) . The correlation between GFP and the immediate neighboring genes was much weaker (window size 1, $R = 0.29$, $P = 0.007$) (Fig. 3), but the value of this observation might be limited because of the fewer expression values included in these smaller window sizes, which could result in a higher variance and a lower correlation (see below). To further test the significance of the observed positive correlations, we performed a Monte Carlo simulation in which GFP expression values were randomly distributed among the clones. The correlation between domain activity and GFP value was calculated for 1 million permutations per window size (see also Supplemental Fig. S3). Values of $R > 0.45$ were observed with a frequency of $<2 \times 10^{-5}$ for any individual window size. Therefore, the observation of correlations of $R > 0.45$ for all window sizes of 19–79 within the actual data is highly significant and suggests that the inserted GFP gene acquires an expression level related to domains of up to 80 genes surrounding the integration site. This length agrees well with the average size of

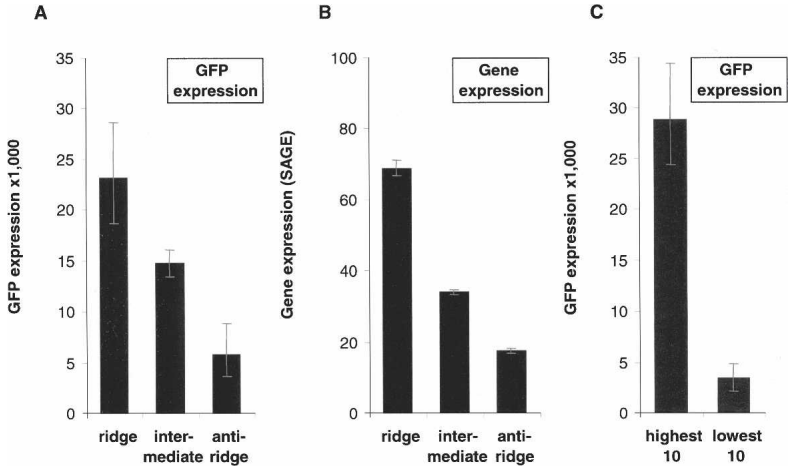

Figure 2. Expression of GFP constructs and genes in different chromosomal domains. (*A*) Average GFP expression of all clones with integrations in ridges ($n = 22$), intermediate domains (i.e., neither ridge nor anti-ridge; $n = 54$), or anti-ridges ($n = 14$). (*B*) Average expression (based on 133 pooled SAGE libraries) of all human genes embedded in ridges ($n = 4250$), intermediate domains ($n = 13{,}226$), or anti-ridges ($n = 2906$). (*C*) Average GFP expression of the clones harboring integration sites with the highest ($n = 10$) and lowest ($n = 10$) domain activities (defined as the median expression of the surrounding 49 genes in 133 pooled SAGE libraries). Error bars represent standard error of the mean.

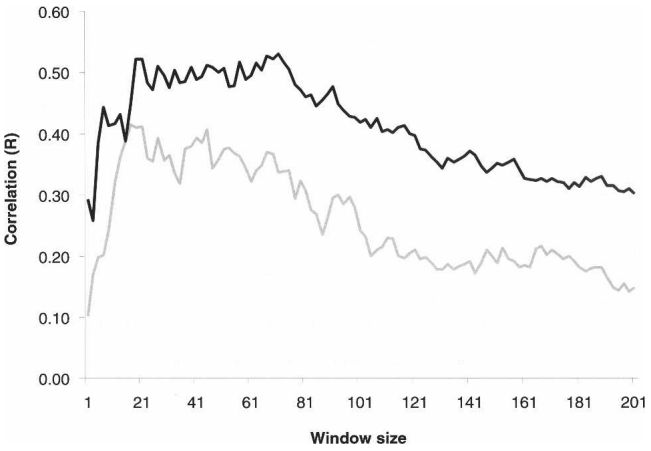

Figure 3. Correlation between GFP expression and domain activity for window sizes of 1–201 genes. The Pearson correlation coefficient (*R*, *Y*-axis) was calculated for GFP expression and domain activities of the integration sites of all 90 clones for window sizes increasing from 1 to 201 genes (*X*-axis). This was done using "all-tissue" domain activity data represented by 133 SAGE libraries from different human tissues (black line) and cell-line-specific HEK293 activity measured by Affymetrix microarrays (gray line).

ridges and anti-ridges of 94 and 81 genes, respectively. As ridges are more gene-dense than anti-ridges, these values correspond to an average length of 6.2 Mb for ridges and 17.1 Mb for anti-ridges. The various ridges and anti-ridges in the human genome are, however, highly variable in length, and these numbers represent average values only.

Domain effect on GFP expression is stronger than effect of neighbor genes

The transcriptome profiles based on SAGE libraries from different tissues are not necessarily representative for HEK293 cells; thus we compared our GFP data with transcriptome profiles specific for HEK293. Even a large SAGE library is not powerful enough to generate a reliable transcriptome profile of an individual cell line; thus we used Affymetrix U133 Plus 2.0 microarrays to generate a HEK293 expression profile. The expression values were related to the same transcriptional units (TUs) on the genome as used for SAGE (see Methods). However, the SAGE data were obtained for 20,382 TUs, whereas the Affymetrix data covered only 16,841 (83%) of the TUs and were hence not as powerful in this respect. Nevertheless, we obtained transcriptome maps for all chromosomes of HEK293 with profiles comparable to the SAGE-based all-tissue map (data not shown). These maps enabled us to determine whether GFP expression in the clones was correlated with HEK293-specific expression data. Figure 3 demonstrates that the all-tissue activity and the HEK293-specific activity showed very similar patterns of correlation between GFP levels and domain-wide expression, with the HEK293 data again giving maximum correlations for the same domain sizes of roughly 19–79 genes and an average correlation of $R = 0.36$ ($P < 4.9 \times 10^{-3}$). Of note, also the HEK293-specific analysis showed a much weaker correlation between expression of GFP and the closest neighboring gene ($R = 0.10$, $P = 0.39$). This was also true for neighboring genes located either parallel or antiparallel to the GFP insert (data not shown). We performed Monte Carlo simulations as described above, using the HEK293 expression data. The confidence intervals are very similar to those calculated on the SAGE data and confirm the significance of the observed correlations (values of $R > 0.35$ were observed with a frequency of $<1 \times 10^{-3}$ for any individual window size).

The breakdown of the correlation between GFP expression and domain activity at lower window sizes could suggest that the effect of the domain at large is much stronger than the effect of immediate neighboring genes. However, also here, it should be considered that the smaller window sizes include less expression data, which might result in a higher variance and consequently a drop in correlation. We therefore specifically analyzed whether neighboring genes affect GFP expression in HEK293. We first calculated the correlation between GFP levels and the average expression of the two immediate neighboring genes, then of GFP expression and the next two neighbors, and so on (Supplemental Fig. S4). This analysis is unbiased, as equal amounts of expression data are used for each calculation. The analysis showed that up to a distance of roughly 20 genes, there is a low positive correlation (average $R = 0.1$). Although this correlation is not significant for most individual points, it suggests a weak effect of neighboring genes on GFP expression levels.

To investigate the relative contributions to GFP expression by neighbor genes and by the domains at large, we made use of the earlier observation that not all highly expressed genes of the genome cluster in ridges. In fact, two-thirds of the highly expressed genes are found outside ridges. Moreover, not all genes in ridges are highly expressed. Ridges also include weakly and non-expressed genes. This enabled us to independently assess the contribution of neighboring genes and of expression domains on GFP expression. We first split the set of GFP clones in two equally large groups, according to the average expression level of the two neighboring genes. The group of clones with high neighbor gene expression (Hi-N) and the group with low neighbor gene expression (Lo-N) showed only a slightly different average GFP expression (Fig. 4A). We subsequently analyzed the relation to domain activity in each of the two groups. The Hi-N group was split in two equal groups, according to the median activity of the 49 surrounding genes (window size 49). Now we observed a strong relation to the average GFP expression level: domains with high activity had a 2.1-fold higher GFP expression than domains of low activity (Fig. 4B). We also split the group of Lo-N in two halves, according to domain activity. Also here, a 2.3-fold higher GFP expression was found in the highly active domains, compared to the weakly active domains. This analysis shows that the domain of integration has a strong influence on the GFP expression level, while the effect of neighboring genes is limited. Controls for the distribution of domain activity and neighbor gene activity over the analyzed groups validate this conclusion (see legends to Fig. 4C–F). We have repeated this analysis using only the expression of the closest neighbor gene as well as separating the Hi-N and Lo-N groups according to integration in ridge, intermediate, and anti-ridge domains. Each time we observed very similar results leading to the same conclusion (data not shown).

The conclusion that the expression of genes in ridges and anti-ridges is strongly influenced by an effect of the domain at large would make two predictions: Firstly, GFP expression should correlate with domain activity throughout the entire domain, including parts of the domain distant from the GFP integration site. Secondly, the correlation should break down at the border of

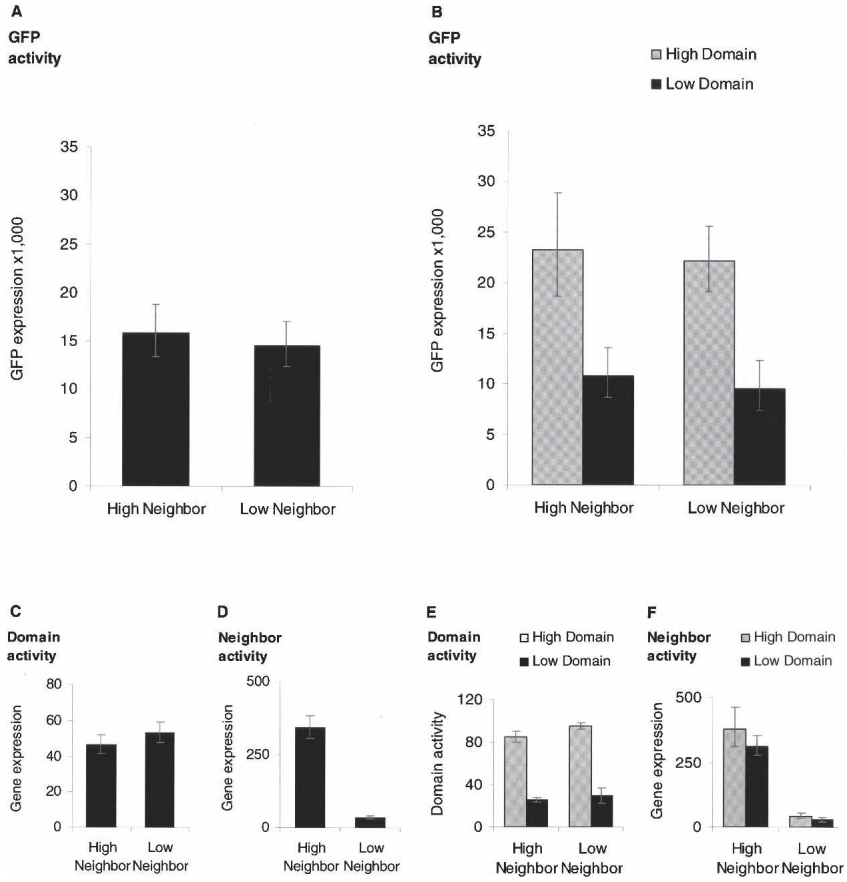

Figure 4. The effect of neighbor genes on GFP expression. (*A*) Average GFP expression of all clones divided into two equally sized groups with either high (Hi-N) or low (Lo-N) neighbor gene activity. The 1.1-fold difference in GFP expression is not significant ($P = 0.71$). (*B*) Average GFP expression for both groups of Hi-N and Lo-N clones, each divided into two equally sized groups with either high (gray bars) or low (black bars) domain activity (window size 49). GFP expression differs 2.1-fold between both domain types for the Hi-N group ($P = 0.02$) and 2.3-fold between both domain types for the Lo-N group ($P = 0.008$). There is no significant difference in GFP expression between the Hi-N and Lo-N groups of the same domain type ($P > 0.71$). (*C*,*E*) Average domain activity per group of clones. Domain activity does not differ significantly between the Hi-N and Lo-N clones in *C*, or between the Hi-N and Lo-N clones of the same domain types in *E*. (*D*,*F*) Average neighbor gene activity per group of clones. The difference in neighbor gene activity is 9.9-fold between the Hi-N and Lo-N clones in *D*, but not significant between the different domain types of the Hi-N and Lo-N clones in *F* ($P > 0.21$). Neighbor gene activity was calculated as the average expression of the two immediate neighboring genes as measured by Affymetrix arrays. Domain activity was defined as the median expression of the surrounding 49 genes in 133 pooled SAGE libraries. Out of 90 clones, six have a pair of neighboring genes without a probe set and were therefore excluded from the analysis. Clone numbers in each group are thus: Hi-N ($n = 42$) and Lo-N ($n = 42$) for the analysis in *A*, *C*, and *D* and $n = 21$ for all four groups in *B*, *E*, and *F*. All *P*-values were calculated with an unpaired *t*-test. Error bars represent standard error of the mean.

each domain. To test these predictions, we aligned all clones with a GFP construct in ridges or anti-ridges ($n = 36$) on their GFP integration site. As the GFP insertions divide each domain in two unequal parts, we oriented the domains such that the larger fragments were on the same side. We calculated the correlation between GFP expression of each clone and domain activity (using a window size of 21 genes) at various positions within and outside the domain. Figure 5 shows a plot of the correlation between GFP expression and domain activity at distances of 0, 25%, 50%, 75%, and 100% from the integration site to the domain ends. Outside the domains, we chose fixed positions of 11, 31, and 51 genes from the domain border. The correlation remains high throughout the domain but completely breaks down at the domain boundaries. Taken together, these results show that ridges and anti-ridges exert a domain-wide effect on GFP expression and form functional domains within the human genome.

Chromosome structure corresponds to GFP expression

Several recent studies have examined differences in chromatin condensation and nuclear position of chromosomal domains. Among others, the group of Cremer has shown that gene-dense domains are positioned toward the nuclear interior (Bolzer et al. 2005). Bickmore and coworkers found that gene-dense domains throughout the genome possess open chromatin fibers (Gilbert et al. 2004), and they postulated that this domain-wide feature facilitates transcription (Sproul et al. 2005).

Goetze et al. (2007) analyzed the three-dimensional (3D) properties of specific ridge and anti-ridge domains in different cells and observed that ridge domains are less condensed and located more interiorly in the nucleus compared to anti-ridges, independent of cell type. To consolidate these structural studies with our functional analysis of chromosomal domains, we examined the 3D structure of the domains of integration. This was done using 3D-FISH in five clones with GFP insertions in chromosome 1 (labeled 1–5 in Fig. 1). Two of the clones had integrations in anti-ridges in chromosomal bands 1p34 and 1q43 and showed relatively low GFP expression. In the other three clones, constructs integrated in ridges at 1q21 (2 clones) and 1q42 exhibited high GFP expression (Fig. 6A). Nuclei of each clone were fixed to preserve the 3D structure and hybridized with fluorescently labeled BACs. Each clone was hybridized with 11 pooled BACs covering a domain of 2.2 Mb surrounding the specific integration site. 3D images of the integration domains were reconstructed from two-dimensional (2D) images obtained by confocal laser microscopy. The three integration domains of the clones with high GFP expression had significantly larger ($P < 9 \times 10^{-4}$, unpaired *t*-test) diameters than the integration domains of the two clones with low GFP expression, which suggests a more open chromatin structure. The domains with high GFP expression also had a more interior nuclear position ($P < 5 \times 10^{-10}$, unpaired *t*-test) (Fig. 6B–E), compared to the domains with low GFP expression (Supplemental Tables S2, S3). These findings show that the previously observed 3D characteristics of ridges and anti-ridges correspond to the functional activity of these domains that is described here.

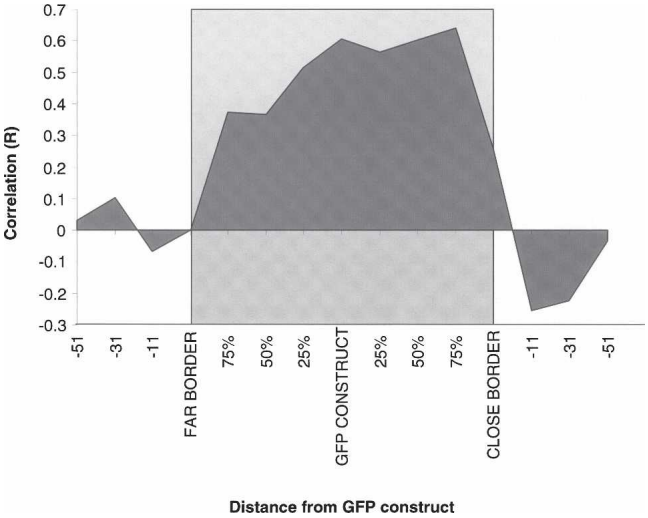

Figure 5. Correlation between GFP expression and domain activity for various relative positions within and outside the domain. All clones with a GFP construct in a ridge or anti-ridge ($n = 36$) were aligned on the position of their GFP construct. For all clones, the border with the largest physical distance to the GFP integration site was determined (far border) and used to orient them alike. Using a window of 21 genes, the domain activity was determined at various relative positions within the domain at 25%, 50%, 75%, and 100% (i.e., the border) of the physical distance between the GFP construct and each border. The domain is represented by a gray box in between both borders. Using the same window size, correlation was calculated outside the domain at a distance of 11, 31, and 51 genes. Correlation is significant for all positions within the domain ($P < 0.028$), but not for any position on the border or outside the domain.

Discussion

Our results show that identical transgenes integrated in different chromosomal regions acquire expression levels that strongly correlate with the expression levels of the domains of integration. These chromosomal domains can exert a general activating or attenuating influence on embedded genes. Immediate neighboring genes also influence GFP expression, but this effect is more limited. The effect of the domains on the level of expression of inserted genes is considerable, and it is plausible that the endogenous genes in these domains are influenced in a similar manner. We have previously reported that expression of genes in ridges is not uniformly high (Versteeg et al. 2003). In that study, we observed that some ridge genes displayed a tissue-specific expression pattern, because they could be silent in one tissue and highly expressed in other tissues. Ridges are defined not by high expression of all embedded genes, but by a high median expression of the domain as a whole. We discerned this high median expression in all studied tissues, although different genes made contributions in different tissues. The dynamic regulation of the expression of individual genes in ridges together with our finding of domain-wide control of expression suggest the existence of a dual mechanism of gene regulation: Transcription factors determine whether a gene will be expressed and also establish a basic level of transcription. In addition, there is a substantial effect of the domain in which genes are positioned, which potentiates the ultimate expression levels. Transcription factors controlling the PGK promoter probably determine a basal level of GFP expression, which can be modified up to eightfold by properties of the whole domain of integration. Such a dual mechanism would considerably augment the dynamic range of transcription factors: the same transcription factor could induce substantial expression of a target gene located in a ridge and low expression of a target gene situated in an anti-ridge. Clearly, that type of mechanism would bear on the evolutionary dynamics of gene repositioning in genomes. Comparison of conservation of gene position in expression clusters in mouse and man are in line with this idea (Singer et al. 2005).

The domain-wide regulation could be based on an activating or a suppressive mechanism, or a combination of both. Activation as well as silencing of genes is often accompanied by changes in the histone code and/or DNA methylation, which can also trigger alterations in chromatin condensation. It is tempting to speculate that histone codes also play a role in the domain-wide regulation of gene expression that we describe here. Histone modifications can spread over considerable genomic distances and have both been associated with silencing and activating mechanisms. In PEV and X chromosome inactivation, the long-range silencing of genes is accompanied by the spreading of trimethylated lysine 9 on histone 3 (H3K9me3) (Heard et al. 2001). This silencing mechanism is mediated by the Swi6/HP1 proteins and Clr4/Su(var)3–9 histone methyltransferases (Bannister et al. 2001; Lachner et al. 2001; Nakayama et al. 2001; Noma et al. 2001). Two recent publications have identified large regions of down-regulated genes in colon cancer (Frigola et al. 2006) and bladder tumors (Stransky et al. 2006). In both cases, silencing was accompanied by trimethylated H3K9.

Also, active marks can spread and influence expression in large chromosomal domains. In *Drosophila*, the dosage compensation complex mediates the spreading of an active histone mark along the X-chromosome by acetylation of H3K16 (Kelley et al. 1999). Genome-wide analyses also detected increased H3K9 and H3K14 acetylation (Roh et al. 2005) and H3K4 methylation (Bernstein et al. 2005) in transcriptionally active regions of the human genome, but these marks were mainly restricted to promoters and regulatory elements of genes and were thus concluded not to represent domain-wide modifications. Interestingly, Finnegan et al. (2004) observed that the expression of a transgene in *Arabidopsis* increased upon activation of the insertion domain, suggesting that spreading of an activating effect can occur. The ability to perform genome-wide analyses of a multitude of histone modifications will enable a further search for domain-wide marks (Barski et al. 2007). Our present results demonstrate that domain-wide regulation of gene expression is a general principle of the human genome, rather than a phenomenon restricted to a few specific loci.

Methods

Constructs

We obtained the lentiviral construct pRRL-PGK-GFPsin-18 and the packaging plasmids pMDLg/pRRE, pMD.G(VSV-G), and RSV-REV, as a kind gift of D. Trono and R. Zufferey (University of Geneva, Geneva, Switzerland) (Dull et al. 1998; Zufferey et al. 1998). The 3′ long terminal repeat (LTR) has largely been deleted, abrogating the enhancer activity of the virus LTR (Zufferey et al. 1998). From this plasmid, we constructed pRRL-FLL by cloning two LoxP sites (flanking the PGK-GFP cassette) and a Flp-In Recombination Target (FRT) site (directly upstream of the first LoxP site) into the pRRL-PGK-GFPsin-18 vector (see Supplemental Table S4 for DNA oligonucleotides used for construction of pRRL-FLL). Using a three-point ligation, two double-stranded (an-

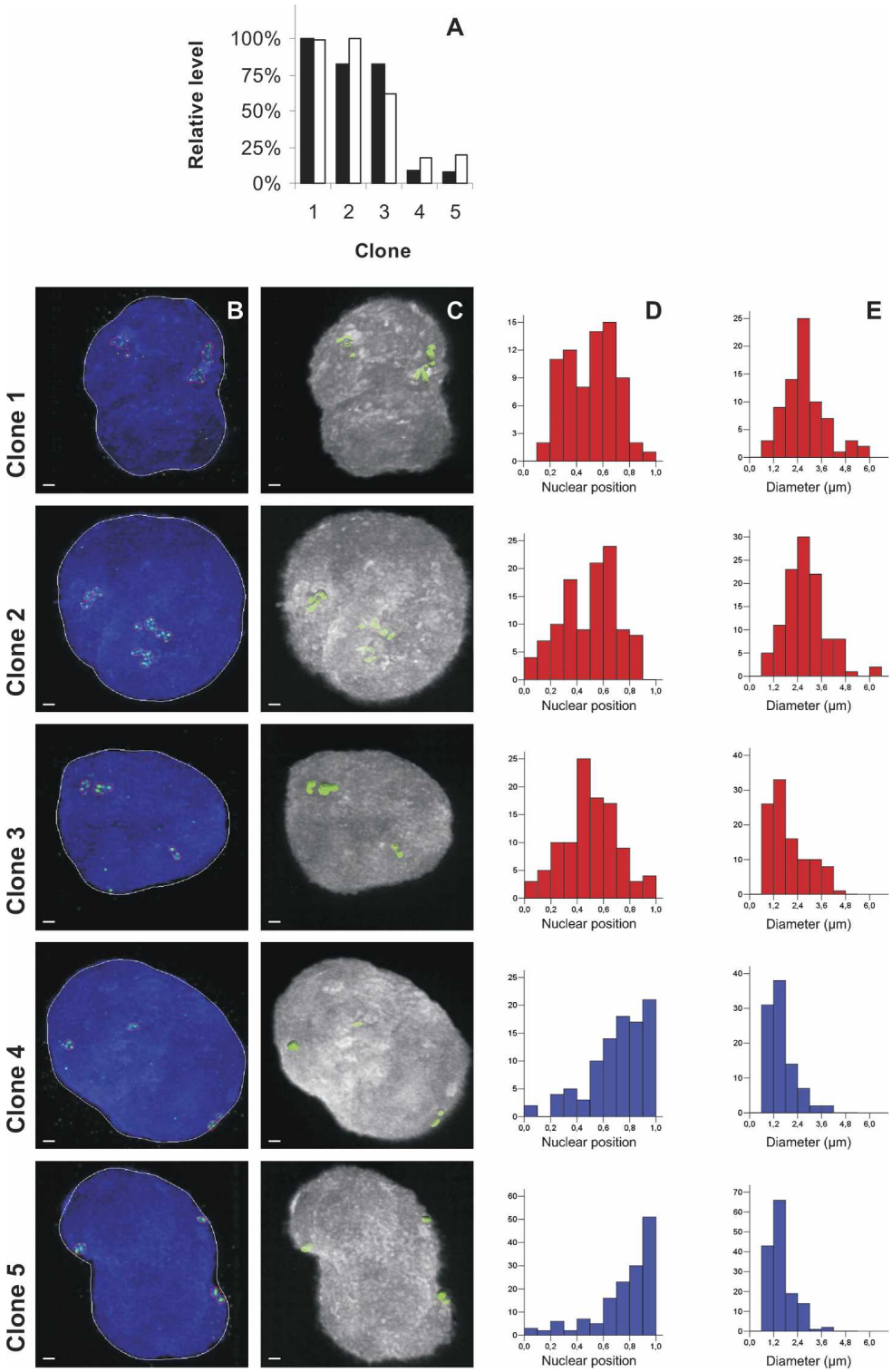

Figure 6. Three-dimensional FISH analysis of domains in five clones harboring GFP integrations on Chromosome 1. (*A*) Relative levels of GFP expression (black bars) and domain activity of the integration sites (shaded bars) in five clones. Highest values are set at 100% and domain activity is defined as the median expression of the surrounding 49 genes (133 pooled SAGE libraries). (*B*,*C*) FISH analysis of 2.2-Mb regions surrounding the integrated GFP construct of five clones (rows) illustrates the 3D structure of each domain. Representative 3D FISH images of each domain are shown as projection (*B*) and after volume rendering (*C*). Per clone, 30–60 nuclei were analyzed. (Green) The transgene integration domain; (blue) nuclei were counterstained in DAPI. Scale bars, 1 μm. Red lines in *B* represent contours of the hybridized areas. (*D*) Histograms showing the distribution of integration domains (count per signal, *Y*-axis) of each of the clones [(red) ridge clones; (blue) anti-ridge clones] with respect to their squared nuclear positions (*D*) or their diameter (*E*). In *D* the *X*-axis shows the relative position of the domain, ranging from 0 (center) to 1 (nuclear periphery). In *E* the *X*-axis depicts the largest diameter in 3D of a domain. The histograms and images show that the ridge domains (clones 1–3) localize more to the interior and are less condensed than the anti-ridge domains (clones 4 and 5).

nealed) oligonucleotides containing an FRT site (FRTfw and FRTrev) and a LoxP site (LOX1fw with LOX2fw), were cloned into pRRL-PGK-GFPsin-18 after digestion with XhoI and ClaI (Roche). Subsequently, one double-stranded (annealed) oligonucleotide (Lox2fw and Lox2rev) containing a second LoxP site was cloned into the SalI (Roche) digested vector. Ligation mix was re-digested with SalI, to select for successfully ligated plasmids (insert disrupts the SalI site).

Cell culture, lentiviral transduction, and FACS analysis

HEK293 cells and 293T cells were cultured in DMEM (Invitrogen) containing 10% fetal calf serum. For production of lentivirus, 293T cells were calcium-transfected with the lentiviral construct pRRL-FLL and packaging plasmids. Titer was determined by FACS analysis using a FACSCalibur (BD Biosciences) of counted HEK293 cells transduced with different dilutions of virus. One week after transduction, clones were single-cell-sorted on a FACSVantage SE or FACSAria (BD Biosciences). Cells displaying fluorescence (~3%) were gated for low, medium, and high fluorescence. From each gate, equal amounts of 96-well plates were seeded with single cells. Clones were passaged to 6-well plates, trypsinized, kept on ice, and analyzed for fluorescence (FITC channel) on a LSRII FACS (BD Biosciences), which was calibrated with EGFP Calibration Beads (BD Biosciences).

RNA isolation and Northern blot analysis

Total RNA was isolated using TRIzol (Invitrogen) and purified with RNeasy (QIAGEN). For Northern blot analysis, samples were separated on a 1% agarose gel containing 6.7% formaldehyde and transferred to Hybond-N membranes (Amersham Biosciences), which were hybridized with radioactively labeled probes. For the GAPDH probe, HEK293 cDNA was made using HEK293 total RNA and a Superscript II RT-PCR Kit (Invitrogen). PCR on the HEK293 cDNA was performed with GAPDHfw (GGGCT GCTTTTAACTCTG) and GAPDHrev (AGGCTGTTGTCATA CTTCTC) primers. The PCR product was checked by sequencing. The GFP probe was made using PCR on the pRRL-FLL construct with the GFPfw and GFPrev primers. A STORM 860 PhosphorImager (Amersham Biosciences) was used to quantify signal intensities.

DNA isolation and Southern blot analysis

Genomic DNA was isolated from HEK293 cells with the Wizard SV Genomic DNA Purification System (Promega) and digested overnight with PstI or BamHI (Roche). Digests were separated on 0.8% agarose gels, transferred to Hybond-N+ membranes (Amersham Biosciences), and hybridized with radioactively labeled probes for detection of the DNA fragment containing the lentiviral construct. Probes were generated by PCR on pRRL-FLL. For detection of the 3′-fragment, a GFP probe was generated with GFPfw (GACGTAAACGGCCACAAGTT) and GFPrev (GAACTC CAGCAGGACCATGT) primers. For detection of the 5′-fragment, a probe against the lentiviral backbone was made with HIVfw (GAGAGAGATGGGTGCGAGAG) and HIVrev (GATGCCCCA GACTGTGAGTT) primers.

PCR amplification and sequencing of integration sites

Integration sites were PCR-amplified using a splinkerette-based PCR approach (Devon et al. 1995; Mikkers et al. 2002). In short, chromosomal DNA (isolated as described above) was digested and ligated to annealed adapter oligonucleotides (splinkerettes), and the fragment containing the lentivirus was amplified by nested PCR using primers complementary to the splinkerette and the LTRs and subsequently sequenced. For more details, see Supplemental Protocol S1.

Mapping of sequences

We developed Perl scripts to analyze sequences. In short, LTR and splinkerette sequences were removed from the ends of all sequences, leaving the genomic sequences. To prevent using low-quality sequences, sequences containing more than 10 ambiguous base calls or more than 600 bp were cut off at either the tenth ambiguous base or base 600. Sequences were mapped against the human genome (UCSC build HG15) using the BLAT algorithm (http://genome.ucsc.edu/). Matches were excluded if (1) the aligned genomic DNA fragment was <45 bp, or (2) the sequence identity was <95%, or (3) there was no unique best hit within the BLAT ranking. On average, aligned sequences were 270 bp long and had a sequence identity of 99.3%. The first nucleotide in the sequence outside of the viral LTR represents the exact position of the viral integration site. As a control, 29 clones were sequenced from both the 5′- and 3′-LTR, which yielded the same genomic position in all cases.

Microarray analysis

RNA was isolated from two duplicate cultures of untransduced HEK293 cells that were both analyzed on microarrays. RNA sample integrity was checked on an Agilent 2100 Bioanalyzer (Agilent Technologies). Four micrograms of total RNA was used for cRNA synthesis and fragmented. Labeling was performed with One-Cycle cDNA Synthesis Kit (Affymetrix). Sample quality was checked on a Bioanalyzer prior and after fragmentation. Ten micrograms of labeled cRNA was hybridized to Affymetrix Human Genome U133 Plus 2.0 arrays according to the manufacturer's protocol (Affymetrix) at the Microarray Department of the University of Amsterdam (MAD, Amsterdam, The Netherlands). Arrays were scanned with a GeneChip Scanner 3000 (Affymetrix).

Microarray data analysis and generation of transcriptome maps

Expression data (.cel files) were normalized using the MAS5 algorithm using GCOS software (Affymetrix). Data from the duplicate arraying experiments were averaged and further analyzed. We developed a Perl script that linked Affymetrix probesets to transcriptional units (TUs), which have been described previously (Versteeg et al. 2003), in the following manner. All consensus and exemplar sequences used for Affymetrix U133 plus 2.0 probe set generation were mapped to the human genome sequence (UCSC build HG15) using the BLAT algorithm (http://genome.ucsc.edu/). We then checked for overlap of every mapped sequence with the exons of TUs. If the orientation of TU and Affymetrix sequence was in agreement or if the TU was unoriented (1273 TUs), the probe set associated with the Affymetrix sequence was linked to the TU. All probe sets linked to more than three TUs were removed from the data set. This resulted in 16,841 (of 20,382) TUs linked to one or more probe sets. Expression values of TUs not coupled to any probe set were treated as missing values (but do determine window size). For each TU, the average signal for all linked probe sets was calculated. Transcriptome maps were generated by calculating a moving median over 49 TUs (MM49). In short, the median is calculated over the 24 TUs situated upstream and downstream of the TU in question and the TU itself. This is done separately for each TU on every chromosome, providing a set of moving medians. The Human Transcriptome Map, based on 133 SAGE libraries from different tissue types, was also mapped to UCSC genome build HG15 and

generated in a similar manner as described previously (Versteeg et al. 2003).

Data analysis and statistics

Fluorescence values obtained by FACS analysis were corrected for background by subtracting fluorescence values measured for normal HEK293 cells. For all analyses, background-corrected fluorescence values and expression data (SAGE and Affymetrix) were $\log_2$-transformed in order to obtain data sets with a normal distribution. All calculations and statistical tests were done using SPSS 12.0.1 (SPSS Inc.), Perl scripts, and Excel 2002 (Microsoft Corporation). All correlations were calculated using the Pearson correlation coefficient (two-tailed significance). Significance of all differences was calculated with an unpaired *t*-test (assuming unequal variance, two-tailed significance). The fold difference (F_d) between groups of clones was calculated by subtracting the mean $\log_2$-transformed GFP values from each other, which yields the $\log_2$ of the fold difference: $^2\log(F_d) = \Sigma[^2\log(GFP_A)]/n_A - \Sigma[^2\log(GFP_B)]/n_B$. Median values over window sizes were calculated in the following manner: for window size 1, the ($\log_2$-transformed) expression value of the TU closest to the integrations site was used. For window size 3, the median was calculated over the nearest TU and the two adjacent TUs, and so on. Monte Carlo analysis was performed with a Perl script that distributed the actual GFP expression values randomly over the clones (i.e., one permutation). Subsequently the Pearson correlation with the actual domain activities was calculated, using either expression data of 133 SAGE libraries or HEK293 microarray data. This was done for 1 million permutations per window size. The significance of *R*-values of $R \geq 0.45$ was calculated by taking the sum of all frequencies of $R \geq 0.45$ for any given window size.

Confocal laser-scanning microscopy and 3D-FISH

All experiments were performed in duplicate. For each experiment, 30–60 nuclei were 3D-imaged. Twelve-bit images were recorded using an LSM 510 confocal laser-scanning microscope (Carl Zeiss Inc.) equipped with a 63 × /1.4 NA Apochromat objective. We used an Ar-ion laser at 364 nm and an Ar laser at 488 nm to excite DAPI and FITC fluorochromes, respectively. Fluorescence was detected with a 385–470-nm bandpass filter for DAPI and a 505–530-nm bandpass filter for FITC. Images were scanned as 512 × 512 × 90 to 512 × 512 × 140 voxel images with a sampling rate of 50 × 50 × 100 nm (*x*, *y*, *z*). Per nucleus, 150 images with 100-nm height increments (in *z*-direction) were combined for the construction of 3D images. See Supplemental Protocol S2 for more details.

3D-FISH analysis

The radial nuclear position p_n [$p_n = (r_0/r_n)^2$] of a chromosomal domain was calculated as the distance between the center of gravity (COG) of a domain and the center of the nucleus (r_0), divided by the length of a line from the nuclear center to the nuclear envelope through the COG of the domain (r_n). This (relative) value is then squared to correct for the higher probability of a domain to be located more toward the nuclear periphery. Diameter is the largest longitudinal section of the 3D domain.

Acknowledgments

We thank D. Trono and R. Zufferey (University of Geneva) for providing the lentivirus vectors and the Sanger Institute for supplying BACs. We also thank B. Hooijbrink, D. Markusic, M.W. Tanck, P. van Sluis, R. Volckmann, H.Y. Man, and N. Ponne of the AMC, W. de Leeuw, O. Giromus, J. Mateos-Langerak, the Centre of Advanced Microscopy and the Microarray Department of the University of Amsterdam, A. Uren (Netherlands Cancer Institute), and Scientific Volume Imaging BV for advice and support. We thank the reviewers for valuable comments. This research was supported by grants from the European Commission for the FP6 3D-Genome project (contract LSHG-CT-2003-503441), the Stichting Kindergeneeskundig Kankeronderzoek (SKK), and the BioRange program of the Netherlands Bioinformatics Centre (NBIC).

References

Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C., and Kouzarides, T. 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410:** 120–124.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., Gingeras, T.R., et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120:** 169–181.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Muller, S., Eils, R., Cremer, C., Speicher, M.R., et al. 2005. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3:** e157. doi: 10.1371/journal.pbio.0030157.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Devon, R.S., Porteous, D.J., and Brookes, A.J. 1995. Splinkerettes improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Res.* **23:** 1644–1645.

Dull, T., Zufferey, R., Kelly, M., Mandel, R.J., Nguyen, M., Trono, D., and Naldini, L. 1998. A third-generation lentivirus vector with a conditional packaging system. *J. Virol.* **72:** 8463–8471.

Finnegan, E., Sheldon, C., Jardinaud, F., Peacock, W., and Dennis, E. 2004. A cluster of *Arabidopsis* genes with a coordinate response to an environmental stimulus. *Curr. Biol.* **14:** 911–916.

Frigola, J., Song, J., Stirzaker, C., Hinshelwood, R.A., Peinado, M.A., and Clark, S.J. 2006. Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nat. Genet.* **38:** 540–549.

Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N.P., and Bickmore, W.A. 2004. Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers. *Cell* **118:** 555–566.

Goetze, S., Mateos-Langerak, J., Gierman, H.J., de Leeuw, W., Giromus, O., Indemans, M.H., Koster, J., Ondrej, V., Versteeg, R., and van Driel, R. 2007. The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. *Mol. Cell. Biol.* **27:** 4475–4487.

Gould, A. 1997. Functions of mammalian Polycomb group and trithorax group related genes. *Curr. Opin. Genet. Dev.* **7:** 488–494.

Heard, E., Rougeulle, C., Arnaud, D., Avner, P., Allis, C.D., and Spector, D.L. 2001. Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. *Cell* **107:** 727–738.

Hurst, L.D., Pal, C., and Lercher, M.J. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5:** 299–310.

Kelley, R.L., Meller, V.H., Gordadze, P.R., Roman, G., Davis, R.L., and Kuroda, M.I. 1999. Epigenetic spreading of the *Drosophila* dosage compensation complex from roX RNA genes into flanking chromatin. *Cell* **98:** 513–522.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C.X., Singer, M.A., Richmond, T.A., Wu, Y.N., Green, R.D., and Ren, B. 2005. A high-resolution map of active promoters in the human genome. *Nature* **436:** 876–880.

Lachner, M., O'Carroll, N., Rea, S., Mechtler, K., and Jenuwein, T. 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410:** 116–120.

Lercher, M.J., Urrutia, A.O., and Hurst, L.D. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31:** 180–183.

Lercher, M.J., Urrutia, A.O., Pavlicek, A., and Hurst, L.D. 2003. A unification of mosaic structures in the human genome. *Hum. Mol.*

Genet. **12:** 2411–2415.
Mijalski, T., Harder, A., Halder, T., Kersten, M., Horsch, M., Strom, T.M., Liebscher, H.V., Lottspeich, F., de Angelis, M.H., and Beckers, J. 2005. Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues. *Proc. Natl. Acad. Sci.* **102:** 8621–8626.
Mikkers, H., Allen, J., Knipscheer, P., Romeyn, L., Hart, A., Vink, E., and Berns, A. 2002. High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. *Nat. Genet.* **32:** 153–159.
Nakayama, J., Rice, J.C., Strahl, B.D., Allis, C.D., and Grewal, S.I.S. 2001. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292:** 110–113.
Noma, K., Allis, C.D., and Grewal, S.I.S. 2001. Transitions in distinct histone H3 methylation patterns at the heterochromatin domain boundaries. *Science* **293:** 1150–1155.
Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A., and Panning, B. 2002. Xist RNA and the mechanism of X chromosome inactivation. *Annu. Rev. Genet.* **36:** 233–278.
Roh, T.Y., Cuddapah, S., and Zhao, K. 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes & Dev.* **19:** 542–552.
Semon, M. and Duret, L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol. Biol. Evol.* **23:** 1715–1723.
Singer, G.A., Lloyd, A.T., Huminiecki, L.B., and Wolfe, K.H. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol. Biol. Evol.* **22:** 767–775.
Sproul, D., Gilbert, N., and Bickmore, W.A. 2005. The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.* **6:** 775–781.
Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., de Medina, S.G.D., Segraves, R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., et al. 2006. Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet.* **38:** 1386–1396.
Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J., and van Kampen, A.H. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13:** 1998–2004.
Weiler, K.S. and Wakimoto, B.T. 1995. Heterochromatin and gene expression in *Drosophila*. *Annu. Rev. Genet.* **29:** 577–605.
Zufferey, R., Dull, T., Mandel, R.J., Bukovsky, A., Quiroz, D., Naldini, L., and Trono, D. 1998. Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery. *J. Virol.* **72:** 9873–9880.

Received January 11, 2007; accepted in revised form June 22, 2007.