Human Molecular Genetics, 2003, Vol. 12, No. 19 **2411–2415**
DOI: 10.1093/hmg/ddg251

A unification of mosaic structures in the human genome

Martin J. Lercher[1,*], Araxi O. Urrutia[1], Adam Pavlíček[2] and Laurence D. Hurst[1]

[1]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK and [2]Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, 16637 Prague, Czech Republic

Received February 17, 2003; Revised and Accepted July 21, 2003

The human genome is a mosaic structure on many levels: there exist cytogenetic bands, GC composition bands (isochores) and clusters of broadly expressed genes. How might these inter-relate? It has been proposed that to optimize gene regulation, housekeeping genes should concentrate on transcriptionally competent chromosomal domains. Prior evidence suggests that regions of high GC and R bands are associated with such domains. Here we report that broadly expressed genes cluster in regions of high GC, and in R and lightest Giemsa bands. This is not only a confirmation of the adaptive hypothesis, but is also the first direct systematic evidence of a general interdependence of expression patterns with base composition and chromosome structure.

INTRODUCTION

What determines gene order in the human genome? Genes are not randomly distributed along chromosomes. We have recently shown that they are arranged according to their breadth of expression: broadly expressed genes tend to cluster (1), although factors that account for this clustering remain unknown. In prokaryotes, genes related to a particular function are clustered in operon structures and their expression is co-regulated. While in eukaryotes co-regulatory gene units have been observed, in some cases, as in the case of HOX genes, there is no evidence for these to be a common case.

Unlike prokaryotes and other invertebrates, mammalian genomes show great variability in their base composition (2). Several hypotheses have been proposed to explain this pattern. Some authors favouring a selectionist explanation argue that high contents of G + C in some regions of the genome help to preserve chromatin structure in thermo-regulated organisms (2). Theories of mutational processes to explain base compositional differences have also been proposed (3). Nevertheless, the reason for the heterogeneity in base composition is still a matter of debate. How, if at all, do the compositional mosaic structure of the genome and the gene expression patterns interact?

If selectively neutral processes determine both the mosaic structure of chromosomes and the clustering of broadly expressed genes, then we expect no relationship between regional composition and functional properties of the genes such as their expression patterns. On the other hand, regions differing in their base composition may be differently suitable for transcription. If local chromatin characteristics affect access to the transcription machinery (4–6), then we expect genes expressed in many cell types to be concentrated in transcriptionally competent regions, even when gene density effects are corrected for.

It is well known that chromosomal regions of high GC exhibit higher gene densities (7). These regions also contain a higher density of CpG islands (8). Because it has been reported that housekeeping genes—in contrast to tissue-specific genes—are always associated with CpG islands (9), this has led to the widely accepted notion that housekeeping genes are preferentially located in regions of high GC (2). However, a detailed analysis found that the association between CpG islands and the expression patterns of genes is more complex: 10% of housekeeping genes are not associated with CpG islands, while this fraction varies for tissue-specific genes between GC-poor and GC-rich regions (10). Furthermore, the latter study concluded that housekeeping genes are slightly more prevalent in GC poor regions, once gene density has been accounted for. Thus, this systematic study (as well as two others from the same group) (11,12) contradicts widely held beliefs on the association between expression breadth and regional nucleotide composition. However, these reports measured expression breadth from expressed sequence tag (EST) data, and GC content from coding sequences; both are not ideal measures. Thus, the question of how housekeeping genes are distributed in relation to tissue specific genes in the human genome is currently not fully resolved.

*To whom correspondence should be addressed. Tel: +44 1225385902; Fax: +44 1225386779; Email: m.j.lercher@bath.ac.uk

RESULTS

Our aim is to evaluate whether such a relation between regional base composition and gene expression exists. Until recently it was not possible to systematically address this question due to the lack of reliable quantitative expression data necessary to discriminate expression rate from expression breadth. Serial Analysis of Gene Expression (SAGE) technology (13) allows quantitative identification of genes expressed in a particular tissue. To examine whether gene order in the genome is related to base composition variation, we compared expression patterns of over 10 000 autosomal human genes across 19 normal tissues with the GC content of their introns. It has recently been shown that under some experimental conditions, SAGE libraries may tend to over-represent GC rich sequences (14). As this could bias our results, all analyses are based on a curated dataset, which excludes libraries that showed a bias towards GC rich sequences (see Materials and Methods).

There appear to be two types of models that predict a correlation between local chromatin characteristics and expression pattern. The first type assumes that chromatin remodelling acts like a switch, either allowing or preventing the transcription of genes. This would predict a correlation of GC and banding pattern with expression breadth (the number of tissues where a gene is expressed), but not with measures of expression rate. The second type of model assumes that chromatin remodelling dominantly affects the rate of transcription, e.g. by ensuring that highly expressed genes (be they tissue-specific or broadly expressed) are in open chromatin. This model would predict an association of chromatin characteristics with peak expression rate, but not necessarily with expression breadth. To distinguish between these two models, we report results for both of these measures (1): breadth of expression and peak rate of expression. We also performed corresponding analyses for other measures of expression rate (mean across all tissues, mean across tissues with positive expression, standard deviation over mean across all tissues), although we are not aware of a model that would predict a direct effect on these measures. All measures of expression rate are highly correlated, and all results are in qualitative agreement with those presented here for the peak rate (data not shown).

Analysis of expression breadth and local nucleotide composition (GC) reveals a highly significant correlation ($r^2 = 0.24$, $P < 10^{-5}$; for an example see Fig. 1). A similar although weaker pattern appears when comparing GC content and the logarithm of the expression rate ($r^2 = 0.05$, $P < 10^{-5}$). To account for the great degree of variability in expression patterns at a one-gene resolution, these correlations were assessed after averaging all variables over 15 neighbouring genes. Furthermore, after sorting individual genes according to their surrounding DNA composition into GC categories of 5% width, mean expression breadth and log(rate) both have a strikingly strong linear relationship with base composition ($r^2 = 0.89$, $P < 0.0005$; $r^2 = 0.83$, $P < 0.005$, respectively; Fig. 2). We previously reported a limited although significant correlation between expression patterns and base composition on a one-gene basis (1). Correlation coefficients rise as the number of genes per window is increased (Fig. 3; all correlations are highly significant, $P < 0.0005$). Thus, while much of the variation in expression breadth and rate is based in the properties of individual genes, a large fraction of the *long-scale* variability (up to almost 50%, see Fig. 3) is predicted by a related variation in GC composition. This strongly supports the notion that isochores are real and may have some functional importance.

Our earlier analyses showed that clustering of genes was related to expression breadth and that the previously described clustering of highly expressed genes (15) is a by-product of the dependence of rate on breadth (1). Accordingly we found that the correlation of log(rate) with GC content fades out when we look at residuals from the breadth correlation. In contrast, when examining the residuals from breadth on log(rate), the correlation with GC remains unchanged (Table 1). These results provide evidence for a strong relationship between breadth of expression of a gene and the base composition at the genomic region where it is situated.

In contrast to the above results, some previous analyses have reported a small *negative* correlation between local GC content and the breadth of expression estimated from expressed sequence tag (EST) data (10–12). To reconfirm that our results are not an artefact of the SAGE method, we therefore repeated our analysis using the breadth of expression obtained from the ESTs contained in the UniGene database (16). In qualitative agreement with the SAGE analysis in Figure 3, we found a highly significant positive correlation between intron GC and EST breadth of expression, which increased with the number of neighbouring genes averaged (Supplementary Material Figure A). The discrepancy between our results and previous studies appears to be caused mainly by the previous studies examining individual genes rather than regional averages. Another contribution to this difference may stem from the use of (total or third site) coding sequence GC instead of intron GC; coding region and intron GC appear to measure different genomic properties. However, we found qualitatively similar results for intergenic GC, intron GC excluding repetitive sequence and transcript GC (data not shown). It has been suggested that a discrepancy between SAGE and EST results might be due to a differential decay of SAGE tags with different GC (12). This appears not to be relevant: there is hardly any correlation between SAGE tag GC and expression breadth in our curated data set ($r^2 = 0.0001$).

GC content has been associated with CpG density. Given that housekeeping genes tend to be located near CpG islands (10,17), the concentration of housekeeping genes was expected to be higher in GC rich regions (2). This suggests a possible explanation for our findings, i.e. the correlation between expression breadth and GC content might simply reflect the higher CpG density rather than GC content *per se*. However, we found very similar results when correlating expression breadth with intron GC excluding CpG islands ($r^2 = 0.79$ for 5% bins of GC). Thus, CpG island preference alone fails to explain the concentration of housekeeping genes in GC rich regions. From the above we might presume that isochores are, to a very large extent, regions of comparable breadth of expression.

The mammalian genome is also heterogeneous in its structure. Giemsa staining of metaphasic chromosomes reveals a banding pattern. The Giemsa bands are related to chromatin compaction and distribution of chromosomes inside the

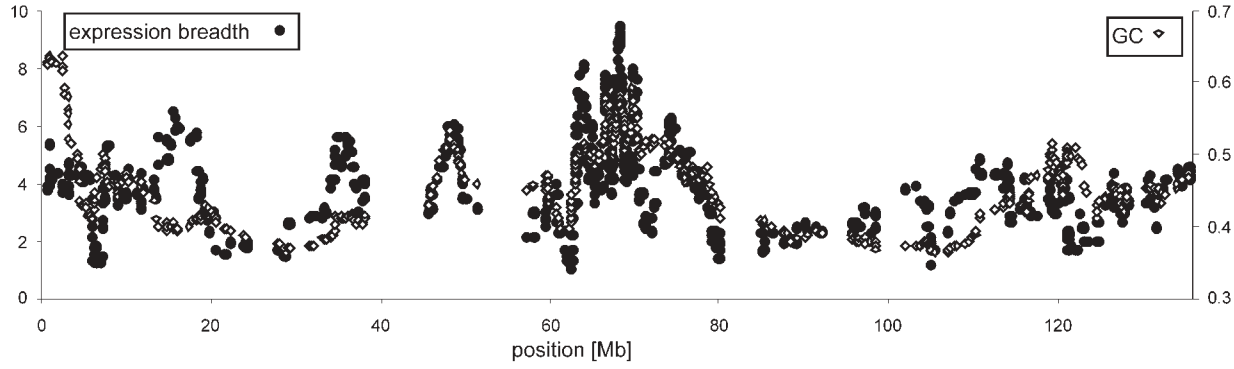

Figure 1. Expression breadth (black dots) and intron GC (grey diamonds) for genes on chromosome 11. Each point represents the average of GC content /breadth for 15 neighbouring genes.

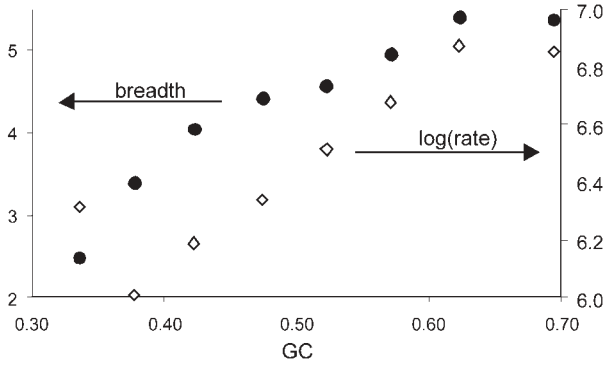

Figure 2. Expression breadth and log(rate) averaged over contiguous intron GC windows of 5% width. The correlation coefficients give $r^2 = 0.89$ (breadth) and 0.83 (rate), respectively.

Table 1. Correlations for rate and breadth with base composition; 15-gene averages

	r	r^2	P
Breadth versus log(rate)	0.43	0.19	$<10^{-5}$
Breadth versus GC	0.49	0.24	$<10^{-5}$
Residuals of breadth = $a + b \times$ log(rate) versus GC	0.43	0.19	$<10^{-5}$
Log(rate) versus GC	0.23	0.05	$<10^{-5}$
Residuals of log(rate) = $a + b \times$ breadth versus GC	0.02	0.0005	0.58

nucleus, where darker and more compacted regions tend to occupy the nuclear periphery (4). Moreover, band types have been correlated with base composition: GC-poorest DNA segments are preferentially located on the most intensely staining G bands, while a subset of the R bands contains the GC-richest isochores (18,19). Therefore we asked whether clustering of housekeeping genes in GC-rich regions relates to these chromosome bands. Indeed, we found that broadly expressed genes are preferentially located in the lightest staining G and R bands (Fig. 4), which contain the most GC-rich segments. Overall 81% of housekeeping genes (expressed in 13 or more tissues) are in one of these two bands. Gene density is generally higher in these two bands (19,20); nonetheless, controlling for gene density we still find enrichment of broadly expressed genes in the R and lightest staining G bands (747 genes compared with 687 expected; $P = 0.023$ from χ^2 test).

The observed mean expression breadth decreases much steeper from R- to dark G-bands than predictions derived from either total band GC or from the intron GC of the genes under study (Fig. 4). This suggests that at least part of the correlation between banding patterns and expression breadth is independent of GC. Consequently, examining the regression residuals of expression breadth versus intron GC for individual genes, we find that genes are not randomly distributed across cytogenetic bands (ANOVA; $P = 0.038$ from F-test). Thus, broadly expressed genes show independent preferences for regions of high GC as well as for the R and lightest staining G bands.

DISCUSSION

Our results provide the first direct systematic evidence of a general relationship between expression patterns and chromatin structures and base composition. This however leaves unresolved the issue of the evolution of isochores. Might GC content evolve as a by-product? Or is it necessary that regions of broad expression have a high GC content, i.e. is the GC content itself under selection? Assuming that housekeeping genes tend to concentrate in regions of open chromatin in order to facilitate transcription (4), our data could be consistent with two models that explain the higher GC content in DNA segments containing housekeeping genes. In the first model, GC content is selectively driven since GC-rich DNA tends to be open and taken to the centre of the nucleus. Alternatively, high GC content could, via biased gene conversion, be a by-product of open chromatin being more prone to recombination.

Both models are consistent with the correlation between recombination rates and base composition (21–25). In the former model this would be a side consequence of the fact that open chromatin is GC rich and open chromatin may be prone to recombination. In the latter model, the GC content is

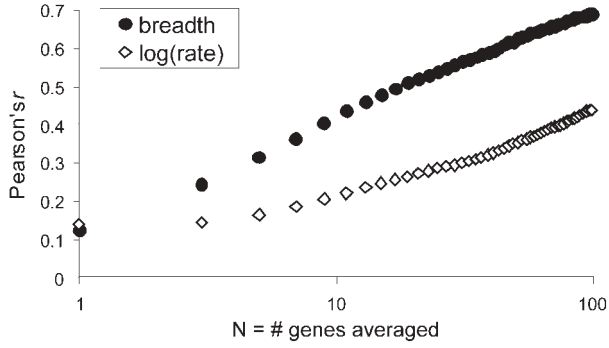

Figure 3. Pearson's r for the correlation between intron GC and expression breadth and rate, for sliding averages over N neighbouring genes.

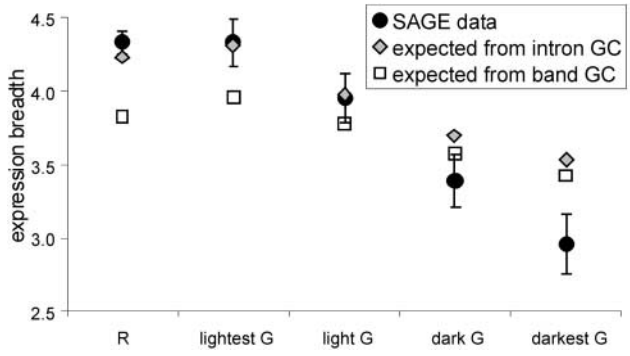

Figure 4. Mean expression breadth of genes in differently staining cytogenetic bands and predictions from intron GC and from total band GC. Error bars show standard errors of the means.

caused by recombination. Therefore both models are also consistent with a correlation between breadth and recombination. Indeed we find such pattern, although the correlation is extremely weak, possibly due to the low resolution of the data available ($r^2 = 0.0034$, $P < 10^{-4}$ for 15-gene averages; recombination data from 25). However, we can imagine a discriminating prediction. Under the second model, all genes expressed exclusively in germ cells just prior to chiasmata formation are prone to recombination and hence to high GC content, while the former predicts that, as such genes are tissue specific, they need not be GC rich. When SAGE libraries for these cell types become available, the test could be performed.

How might the association between expression patterns and local chromatin characteristics shown above be tested experimentally? The above model predicts that when genes are inserted into a non-native chromosomal environment together with their promoter regions, their expression pattern should depend on local GC content and cytogenetic banding pattern. It is indeed well known that randomly inserted transgenes are often not transcribed. In agreement with the competent chromatin model, transgene expression—at least in the case of globin genes—can be rescued with locus control region elements that modify chromatin structure (26). By a systematic examination of the local chromatin characteristics and the expression pattern for a large number of randomly located transgene insertions, the predictions of our model can thus in principle be tested. Unfortunately, currently available data is not of adequately high resolution to address this issue (F. Grosveld, personal communication), and we have to leave this test for future work.

In summary, our results are consistent with gene location being an adaptive property related to regional base composition and chromosome structure (2), where selective pressures favour the concentration of housekeeping genes in genomic regions with particular structural properties, most probably to facilitate access to transcription machinery (4). In accord with this picture, it has been shown that actively transcribed chromatin is predominantly located within the nuclear interior comprising early replicating R bands, which contain the GC richest and gene richest domains (27). The null model, in which genes in the genome are randomly assorted with respect to their expression, is no longer tenable.

MATERIALS AND METHODS

The Serial Analysis of Gene Expression (13) (SAGE) data was obtained from SAGEmap (28) (ftp://ncbi.nlm.nih.gov/pub/sage). The dataset was curated to avoid possible GC biases in SAGE libraries following the approach of Margulies *et al.* (14); we removed 14 libraries with mean tag GC > 0.5. The resulting SAGE tag/tissue data set was based on 40 libraries representing 19 tissues. Tag counts were converted to relative values (cpm, counts per million) after joining all libraries representing the same tissue type. If tags were found only once in one tissue type, we discarded the observation as a likely sequencing error. This data was cross-linked to the mRNA sequences in RefSeq (ftp://ncbi.nlm.nih.gov/refseq), by extracting the 3′-most *Nla*III SAGE tag for each mRNA. If the same tag occurred more than once in RefSeq, all corresponding genes were excluded. To be conservative, the gene set was further restricted to those sequences who's tag was also reported by NCBI as reliable for the corresponding UniGene cluster (16) (UniGene build #155, ftp://ncbi.nih.gov/pub/sage/map/Hs/NlIII/SAGEmap_tag_ug-rel.zip). For the remaining genes, we calculated breadth of expression as the number of tissues with positive expression. For genes expressed in at least one tissue, we also calculated the peak rate of expression (maximum cpm across tissues). As with all forms of expression assay, the SAGE data employed here will inevitably miss some genes expressed at low levels. However, this is not likely to unduly bias our results: as we have demonstrated earlier (1), controlling for rate of expression hardly affects regional variation in expression breadth.

Of the genes with valid expression information, 10774 could be located unambiguously on the June 2002 UCSC genome assembly (29) (ftp://genome-archive.cse.ucsc.edu). Gene position was defined as the midpoint between 5′ and 3′ ends of the transcribed sequence.

For each gene, we extracted the coding sequence from the RefSeq mRNA. We also extracted transcripts (containing both exon and intron sequences, and including information on repetitive DNA) from the genomic data at the UCSC web site. Owing to sequencing errors, mistakes in the assembly, or mis-annotations, intron sequences may be wrongly identified from this kind of data. To ensure proper identification, we compared the coding part of the corresponding exons against the RefSeq sequences. Genes were excluded if we found a length difference or if an internal stop codon occurred in the genomic

coding sequence. Nucleotide composition was measured as the guanine and cytosine (GC) fraction. Intron GC was calculated for 8128 genes with total intron length >100 bp. For 7986 genes with total intron length >500 bp, we also calculated intron GC excluding CpG islands. CpG islands were defined as regions of at least 200 bp, with mean GC > 0.5, and CpG observed/CpG expected >0.6 (10).

Recombination data (25) and cytogenetic band positions (based on FISH data) (30) were also obtained from the UCSC web site. Band positions are imprecise by up to several 100 kb or even more. When including only genes at least 1 Mb away from start and end of their cytogenetic band, results are qualitatively unchanged (data not shown).

To reconfirm that the observed patterns are not due to any remaining bias of the SAGE data, we also examined the correlation between nucleotide composition and local breadth of expression obtained from expressed sequence tag (EST) data. Each UniGene group not only contains the RefSeq mRNA sequence, but also all ESTs believed to map to the same gene. We used these to cross-link genes to 622 EST libraries constructed from normal tissue samples, each containing at least 50 ESTs. This resulted in a data set of 8763 genes, each known to be expressed in at least one out of 73 normal tissues (16 prenatal and 57 postnatal). We calculated breadth of expression as the number of tissues with positive expression information.

For all correlations, r is Pearson's coefficient. Significance levels were estimated from 10 000 random pairings of the raw data value pairs: $P = (1 + \text{number of random pairings with smaller or equal } r^2)/(1 + \text{number of random pairings})$. Correlations and regressions for expression rate were calculated after taking the logarithm.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We thank Laurent Duret and Frank Grosveld for interesting discussions. This work was supported by CONACyT and ORS (A.O.U.), BBSRC (L.D.H.), and The Wellcome Trust (M.J.L.).

REFERENCES

1. 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.
2. Bernardi, G. (1993) The vertebrate genome: isochores and evolution. *Mol. Biol. Evol.*, **10**, 186–204.
3. Gu, X. and Li, W.H. (1994) A model for the correlation of mutation-rate with GC content and the origin of GC-rich isochores. *J. Mol. Evol.*, **38**, 468–475.
4. Cremer, T. and Cremer, C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292–301.
5. Mahy, N.L., Perry, P.E. and Bickmore, W.A. (2002) Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *J. Cell Biol.*, **159**, 753–763.
6. Williams, R.R. (2003) Transcription and the territory: the ins and outs of gene positioning. *Trends Genet.*, **19**, 298–302.
7. IHGSC (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
8. Cross, S.H., Clark, V.H., Simmen, M.W., Bickmore, W.A., Maroon, H., Langford, C.F., Carter, N.P. and Bird, A.P. (2000) CpG island libraries from human Chromosomes 18 and 22: landmarks for novel genes. *Mamm. Gen.*, **11**, 373–383.
9. Antequera, F. and Bird, A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA*, **90**, 11995–11999.
10. Ponger, L., Duret, L. and Mouchiroud, D. (2001) Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.*, **11**, 1854–1860.
11. Goncalves, I., Duret, L. and Mouchiroud, D. (2000) Nature and structure of human genes that generate retropseudogenes. *Genome Res.*, **10**, 672–678.
12. Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.*, **12**, 640–649.
13. Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–488.
14. Margulies, E.H., Kardia, S.L. and Innis, J.W. (2001) Identification and prevention of a GC content bias in SAGE libraries. *Nucl. Acids Res.*, **29**, e60.
15. 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.
16. Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E. *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540–546.
17. Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
18. Saccone, S., Desario, A., Wiegant, J., Raap, A.K., Dellavalle, G. and Bernardi, G. (1993) Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl Acad. Sci. USA*, **90**, 11929–11933.
19. Federico, C., Andreozzi, L., Saccone, S. and Bernardi, G. (2000) Gene density in the Giemsa bands of human chromosomes. *Chromosome Res.*, **8**, 737–746.
20. Saccone, S., Federico, C., Solovei, I., Croquette, M.F., Della Valle, G. and Bernardi, G. (1999) Identification of the gene-richest bands in human prometaphase chromosomes. *Chromosome Res.*, **7**, 379–386.
21. Bernardi, G. (1989) The Isochore organization of the human genome. *A. Rev. Genet.*, **23**, 637–661.
22. Ikemura, T. and Wada, K.-N. (1991) Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromsome numbers; relation between nucleotide sequence data and cytogenetic data. *Nucl. Acids Res.*, **16**, 4333–4339.
23. Holmquist, G.P. (1992) Chromosome bands, their chromatin flavors and their functional features. *Am. J. Hum. Genet.*, **51**, 17–37.
24. Fullerton, S.M., Carvalho, A.B. and Clark, A.G. (2001) Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.*, **18**, 1139–1142.
25. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–247.
26. Grosveld, F., van Assendelft, G.B., Greaves, D.R. and Kollias, G. (1987) Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell*, **51**, 975–985.
27. Sadoni, N., Langer, S., Fauth, C., Bernardi, G., Cremer, T., Turner, B.M. and Zink, D. (1999) Nuclear organization of mammalian genomes. Polar chromosome territories build up functionally distinct higher order compartments. *J. Cell Biol.*, **146**, 1211–1226.
28. Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J. and Altschul, S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
29. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
30. Cheung, V.G., Nowak, N., Jang, W., Kirsch, I.R., Zhao, S., Chen, X.N., Furey, T.S., Kim, U.J., Kuo, W.L., Olivier, M. *et al.* (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature*, **409**, 953–958.