Human Molecular Genetics, 2005, Vol. 14, No. 3 ***421–427***
doi:10.1093/hmg/ddi038
Advance Access published on December 8, 2004

Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance

Marie Sémon*, Dominique Mouchiroud and Laurent Duret

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558 Université Claude Bernard Lyon 1, 16 rue Raphaël Dubois, 69622 Villeurbanne Cedex, France

Received July 2, 2004; Revised and Accepted December 3, 2004

Mammalian chromosomes are characterized by large-scale variations of DNA base composition (the so-called isochores). In contradiction with previous studies, Lercher *et al.* (*Hum. Mol. Genet.*, 12, 2411, 2003) recently reported a strong correlation between gene expression breadth and GC-content, suggesting that there might be a selective pressure favoring the concentration of housekeeping genes in GC-rich isochores. We reassessed this issue by examining in human and mouse the correlation between gene expression and GC-content, using different measures of gene expression (EST, SAGE and microarray) and different measures of GC-content. We show that correlations between GC-content and expression are very weak, and may vary according to the method used to measure expression. Such weak correlations have a very low predictive value. The strong correlations reported by Lercher *et al.* (2003) are because of the fact that they measured variables over neighboring genes windows. We show here that using gene windows artificially enhances the correlation. The assertion that the expression of a given gene depends on the GC-content of the region where it is located is therefore not supported by the data.

INTRODUCTION

The analysis of mammalian chromosome sequences revealed complex genomic landscapes: some regions of the genome are very gene-rich, whereas some other large regions are devoid of genes (1). These variations in gene density are linked to large-scale variations in DNA base composition (the so-called isochores): gene density and introns length vary 5–20-fold between GC-poor and GC-rich isochores (1–3). This isochore organization is also correlated with many other important genomic features: replication timing (4), recombination (5), methylation pattern (6) and distribution of transposable elements (1,7).

The question of the functional significance of these peculiar chromosomal landscapes is, however, still highly debated: do they reflect an adaptation or are they simply a by-product of neutral evolutionary processes (8–11)? In other words, it is not yet known whether this isochore organization has any significant impact on the phenotype.

To address this important issue, several recent studies in human and mouse have analyzed the relationship between the GC-content of isochores and the expression patterns of the genes they contain. Surprisingly, these studies gave conflicting results (Table 1). Several papers reported very weak correlations, either negative (12–14) or positive (15–17), between the GC-content and gene expression. In contrast, Lercher *et al.* (19) found strong positive correlations, suggesting that there might be some selective advantage to concentrate housekeeping genes on transcriptionally competent, GC-rich, chromosomal domains.

The discrepancy among the studies conducted on the relation between GC-content and expression might be due to the methods used to measure expression (EST, SAGE or DNA microarray), the expression parameter considered (expression level or tissue breadth of expression), differences in the measure of GC-content (in introns, third codon positions or intergenic regions) or differences in the tissues and gene data sets analyzed. Moreover, these studies differ in the way correlations were computed: in Lercher *et al.* (19), correlations were assessed after averaging all variables over 15 neighboring genes, whereas in other studies, correlations were computed using individual genes.

To try to understand the discrepancy between the different studies, we compared on a same gene data set the correlations

*To whom correspondence should be addressed. Tel: +33 4724480000; Fax: +33 472431388; Email: semon@biomserv.univ-lyon1.fr

Table 1. Review of the correlations between GC-content and expression published in the literature

Species	Measure of GC-content	Expression		Data set		Correlation			References
		Method	Parameter	No. genes	No. tissues	Sign	R^2 (%)	P-value	
Human	Coding region	EST	Breadth	2399	19	−	1.7	$<10^{-4}$	(12)
	Third codon positions		Breadth	1396	22	−	0.8	0.0008	(13)
	Third codon positions		Log(peak)	1396	22	−	0.4	0.03	
	Genomic (20 kb)	SAGE	Breadth	11549	14	+	2.8	$<10^{-5}$	(21)
	Genomic (20 kb)		Log(peak)	8170	14	+	3.4	$<10^{-6}$	
	Intergenic (1–10 kb)		Mean	6430	22	+	0.4	$<10^{-4}$	(16)
	Third codon positions	Microarray	Log(mean)	6078	32	+	4.0	$<10^{-6}$	(15)
	Coding region		Log(mean)	6078	32	+	4.0	$<10^{-6}$	
	5′-UTR		Log(mean)	6078	32	+	2.9	$<10^{-6}$	
	3′-UTR		Log(mean)	6078	32	+	3.2	$<10^{-6}$	
	Intron		Log(mean)	6078	32	+	6.3	$<10^{-6}$	
	Intergenic (1–10 kb)		Log(mean)	6078	32	+	5.8	$<10^{-6}$	
	Intron[a]	SAGE	Mean(breadth)	542	19	+	24.0	$<10^{-5}$	(19)
	Intron[a]		Mean(log(peak))	542	19	+	5.0	$<10^{-5}$	
	Intron		Mean(breadth)	8 classes[b]	19	+	89.0	$<10^{-5}$	
	Intron		Mean(log(peak))	8 classes[b]	19	+	83.0	$<10^{-5}$	
	Genomic[c]		Median/window	510 windows[c]	NA	+	NA	$<10^{-211}$	(17)
Mouse	Third codon positions	Microarray	Log(mean)	7708	45	+	1.2	$<10^{-6}$	(15)
	Coding region		Log(mean)	7708	45	+	1.0	$<10^{-6}$	
	5′-UTR		Log(mean)	7708	45	+	0.6	$<10^{-6}$	
	3′-UTR		Log(mean)	7708	45	+	0.8	$<10^{-6}$	

There is a large variability in the values and in the signs of the correlations. The different analyses were based on different measures of GC-content, different methods to detect gene expression (SAGE, EST and microarray) and different parameters of expression (breadth, number of tissues where genes are expressed; mean, average level of expression for expressed genes; peak, maximum level of expression). The sign and R^2-value (%) of correlations are given. No. genes, number of genes in the data set. No. tissues, number of tissues in the expression data.
[a]Mean intronic GC-content over windows of 15 genes.
[b]Correlation assessed after splitting the data set into eight classes of GC-content.
[c]Windows of 49 transcription units.

between GC-content and gene expression obtained with different experimental methods, different estimators of GC-content and different scales of measure (gene by gene or by genomic regions). These analyses were performed both in human and in mouse.

We show that in both species, whatever the method used to measure expression or base composition, the correlations between gene expression and GC-content are very weak. We also show that the analyses performed on sets of neighboring genes are not appropriate, as they lead to overestimation of the real relationship between gene expression and GC-content. Given the weakness of the correlations and the noisiness of present gene expression data, one should be extremely cautious when trying to interpret the biological significance of the relationship between gene expression and GC-content.

RESULTS

We analyzed 6242 human genes for which patterns of expression in 11 different tissues could be estimated using three independent experimental methods (EST, SAGE and DNA microarray). We considered different expression parameters: expression breadth (the number of tissues where expression is detected), mean expression level (the average level of expression for expressed genes in the 11 tissues) and peak expression (the maximum level of expression in the 11 tissues). We measured the GC-content in introns (GCi) and at the third position of codons (GC3).

To assess for possible biases in the sampling of genes or tissues for which we had expression data from the three methods (EST, SAGE and DNA microarray), we also measured correlations on sets of genes for which we had (i) SAGE data but no microarray data (6523 genes), (ii) EST data only (19 988 genes), and on sets of tissues for which we had (i) microarray data but no SAGE data (14 tissues), (ii) EST data only (18 tissues). These analyses did not reveal any significant difference with the common data set (data not shown). Hence, we will mainly present results obtained with the set of 6242 human genes and 11 tissues for which we had expression data from the three methods.

We also assessed the correlation between GC-content and gene expression in the mouse genome, using the three measures of expression (EST, SAGE and DNA microarray). As very few tissues were available for SAGE data, it was not possible to build a common gene data set for the three methods. We therefore studied three data sets corresponding to genes for which we had EST data (26 749 genes, 45 tissues), SAGE data (6906 genes, 11 tissues) and DNA microarray data (5297 genes, 45 tissues).

Correlations measured on individual genes

Table 2 gives the correlations computed on individual genes between GC-content and different measures of expression. All these correlations are in agreement with previous results (Table 1). For each method (EST, SAGE and microarray), the different parameters of expression (breadth, peak or

Table 2. Correlation between GC-content and expression, for different measures of genes expression and for different estimators of base composition in human and mouse

Species	Measure of GC-content	Expression Method	Parameter	Data set No. genes	No. tissues	Correlation Sign	R^2 (%)	*P*-value
Human	GCi	SAGE	Breadth	5977	11	+	1.60	$<10^{-16}$
			Peak			+	0.07	NS
			Mean			+	0.02	NS
		EST	Breadth	5977	11	+	0.02	NS
			Peak			−	0.08	NS
			Mean			−	0.07	NS
		Microarray	Breadth	5977	11	+	4.10	$<10^{-16}$
			Peak			+	0.81	10^{-12}
			Mean			+	1.80	$<10^{-16}$
	GC3	SAGE	Breadth	6246	11	+	0.03	NS
			Peak			+	0.10	NS
			Mean			+	0.02	NS
		EST	Breadth	6246	11	−	0.26	$<10^{-16}$
			Peak			−	0.02	NS
			Mean			−	0.04	NS
		Microarray	Breadth	6246	11	+	3.06	$<10^{-16}$
			Peak			+	2.13	$<10^{-16}$
			Mean			+	2.20	$<10^{-16}$
Mouse	GCi	SAGE	Breadth	6355	11	+	1.13	10^{-15}
			Peak			+	0.12	NS
			Mean			+	0.23	10^{-4}
		EST	Breadth	24127	45	+	0.91	$<10^{-16}$
			Peak			+	0.01	NS
			Mean			+	0.00	NS
		Microarray	Breadth	4832	45	+	1.28	10^{-15}
			Peak			+	0.09	NS
			Mean			+	0.62	NS
	GC3	SAGE	Breadth	6906	11	+	0.19	10^{-4}
			Peak			+	0.32	10^{-6}
			Mean			+	0.55	10^{-10}
		EST	Breadth	26749	45	+	0.05	10^{-4}
			Peak			+	0.01	NS
			Mean			+	0.03	NS
		Microarray	Breadth	5297	45	+	0.54	10^{-8}
			Peak			+	0.03	NS
			Mean			+	0.41	10^{-6}

The human sample consists of genes and tissues for which expression data are available for SAGE, EST and microarray. The three mouse data sets correspond, respectively, to the genes and tissues available for EST, SAGE and microarray data. GCi, GC-content in introns; GC3, GC-content at third codon positions. Expression parameters: breadth, number of tissues where genes are expressed; mean, average level of expression for expressed genes; peak, maximum level of expression. The sign and R^2-value of correlations are given, No. genes, number of genes in the data set, No. tissues, number of tissues in the expression data. NS = *P*-value non-significant after Bonferroni correction.

mean) gave similar results: when correlations are significant, they always are in the same direction. Correlations are generally stronger with the breadth than with the peak or mean expression levels. In human, EST data indicate a weak negative correlation between expression breadth and GC3 ($R^2 = 0.3\%$), but no significant correlation with GCi; on the contrary, SAGE and microarray data revealed a weak positive correlation between expression breadth and GC-content ($R^2 = 1.6$–4.1%), and correlations are stronger with GCi than with GC3. In mouse, the three measures of expression are positively correlated with GC-content, but again correlations are very weak ($R^2 = 0.9$–1.3% for expression breadth versus GCi). Thus, with the exception of human ESTs, all the measures of expression indicate a weak positive correlation between expression breadth and gene GC-content.

How does one explain the contradictory results obtained in human with ESTs. Is it simply due to an artifact in EST data? It is clear that gene expression data are noisy. The measures of expression breadth obtained by the three methods are only weakly correlated (SAGE/microarray $R^2 = 25\%$, SAGE/EST $R^2 = 27\%$, EST/microarray $R^2 = 16\%$ on the common data set of 6246 human genes in 11 tissues). It is not possible to determine which one of the three measures of expression is the most reliable. In principle, quantitative estimation of expression obtained with SAGE or DNA microarrays should be more reliable than those obtained with EST data. Indeed, the first goal of EST projects was to identify new genes (and not to measure expression), and hence EST data often derive from cDNA libraries that have been normalized, to decrease the number of cDNA clones deriving from abundant transcripts. EST data are therefore expected to underestimate the level of expression of highly expressed genes. Conversely, this process of normalization allows the detection of rare transcripts, and hence should improve the measure of tissue distribution

breadth (i.e. the number of tissues where genes are expressed). Hence, although ESTs are clearly not appropriate to measure expression level, there is a priori no reason why this method should be less reliable than SAGE or microarray to measure the breadth of expression.

To assess the sensitivity of the three methods, we selected in RefSeq (18) 1493 human genes, supported by experimental evidence (i.e. for which a manually curated mRNA was available) and that are complete in their 3′ end (i.e. with a polyA tail and a canonic polyadenylation signal <50 bp of the 3′ end). The proportion of RefSeq mRNAs that are not detected to be transcribed in any of the 11 studied tissues is higher for microarray than for SAGE and EST (30, 7 and 7%, respectively), which suggests that microarray is less sensitive than both the other methods.

To assess the consistency of the different methods, we compared in human and mouse orthologous genes, the measures of expression breadth obtained by EST and microarray (NB: this analysis could not be performed for SAGE because there are presently too few tissues for which data are available in both human and mouse). EST-based estimates of expression breadth are highly correlated between orthologs ($R^2 = 50\%$ on a data set of 10 950 orthologous genes and 17 tissues in common between human and mouse). Surprisingly, microarray estimates are less correlated ($R^2 = 11\%$ on 2485 orthologous genes and 18 tissues). The restriction of the data sets to the 2485 genes and the 11 tissues in common between microarray and EST gives similar results. This suggests that for the measure of expression breadth, microarray data might be more noisy than ESTs. It is therefore not clear whether the negative correlation between GC3 and expression breadth that we observed with ESTs in human is due to an artifact of the EST approach or the fact that for some genes, expression breadth might be better estimated by ESTs than by other methods.

Whatever the answer to this question, it is important to stress that in reality the discrepancy between the measures (EST versus SAGE or microarray) is not strong, as all methods agree on the fact that correlations are very small ($R^2 = 0.02–4.1\%$). Thus, the only safe conclusion that can be drawn from these analyses is that the GC-content of genes is a very poor predictor of their expression breadth.

Correlation measured on sets of genes grouped according to their GC-content

To analyze the relationship between GC-content and expression, Lercher *et al.* (19) classified genes according to their GC-content, into eight categories of 5% width. For each category, they computed the average GC-content and the average expression breadth (SAGE). With these averages, they observed a strikingly strong linear correlation between GC-content and expression breadth ($R^2 = 89\%$). As shown in Figure 1, microarray data give similar results: after having grouped genes into GC-content categories, one can observe a strong positive correlation ($R^2 = 85\%$) between the average GC-content and the average expression breadth.

The grouping of genes into GC-content categories is a useful way to visualize the trend of the relationship between GC-content and expression. However, we would like to stress that this approach cannot be used to quantify this relationship: when computed on individual genes, the correlation coefficient represents the percentage of the total variance that is explained by the variable. But after the grouping, the correlation coefficient represents the percentage of the inter-category variance that is explained by the variable.

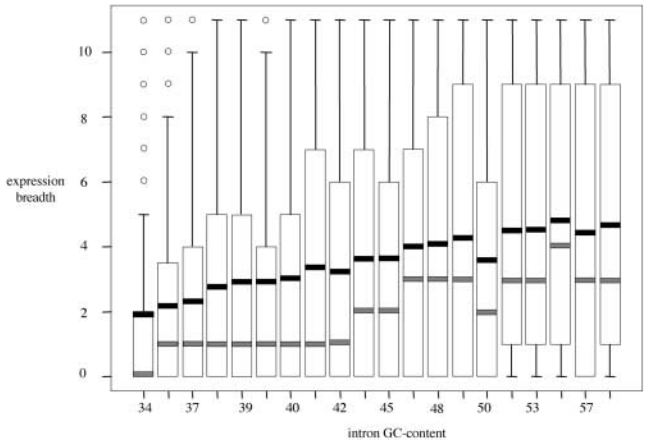

Figure 1. Relationship between intron GC-content and expression breadth in human. Expression breadth was measured with microarray data. Genes were grouped into 20 equal-sized categories according to their intronic GC-content. For each category, the distribution of expression breadth is represented by a box plot: the lower and higher sides of the boxes represent, respectively, the 25 and 75% quantile. The median of expression for each class is drawn in thin lines and the mean in thick lines. The inter-category correlation between GC-content and expression (mean, thick lines) is high. However, there is a huge intra-category variance (see the boxes size).

To illustrate this effect, we performed a simulation: we considered two variables (X and Y) linearly correlated, with a correlation coefficient of 5% (i.e. X explains 5% of the variability of Y). We randomly generated a sample of 5000 points, according to this linear model. We then grouped the points into categories according to the value of X, computed the average of X and Y for each category and then computed the correlation between these averages. As can be seen in Table 3, the correlation coefficient increases steadily as the size of groups increases. Thus, the grouping of points is misleading because it suggests that there is a strong relationship, whereas in reality it is impossible to predict the value of Y knowing the value of X of a given point.

Correlation measured on groups of neighbor genes

Lercher *et al.* (19) also analyzed the relationship between regional variations of GC-content and expression breadth. For this purpose, they assessed these correlations after averaging all variables over 15 neighboring genes (whatever the physical distance between these genes). They found a strong positive correlation ($R^2 = 24\%$) that raised steadily up to 50% when increasing the size of windows from 15 to 100 neighboring genes (19). Thus, by analyzing regional variations of GC-content and gene expression, they were able to detect this strong correlation that is hardly visible when analyzing individual genes.

Table 3. Simulations to assess the impact of grouping data on the measure of correlation coefficient between two linearly correlated variables

No. points per category	No. categories in data set	R^2 (%)
1	5000	5
5	1000	21
10	500	35
15	333	43
20	250	53
50	100	72
100	50	83
500	10	95
1000	5	97

Two linearly correlated variables (X and Y, with a correlation coefficient of 5%) were randomly generated ($N = 5000$ genes). Then, points were grouped into categories according to the value of X, and correlations were computed between X and Y, averaged within each category. Correlation coefficients are indicated for different levels of grouping (i.e. number of points per category).

How does one explain that the correlation between expression breadth and GC-content is much stronger when measured on sets of neighboring genes than on individual genes. Two hypothesis can be proposed. The first possible explanation comes from the fact that genes are not randomly distributed along mammalian genomes: it has been shown recently that tissue-specific and broadly expressed genes tend to cluster in different regions (17,20,21). These regional variations of gene expression are correlated with GC-content and gene density (17) (i.e. with isochores). However, these regional variations of gene expression are partly independent of the isochore structure: the clustering of housekeeping genes is significantly stronger in the human genome than in randomized genomes of identical isochore structure (21). Thus, it is possible that by analyzing groups of neighbor genes, some effects due to regional variations of gene expression were better captured by Lercher *et al.* (19). A second possible explanation comes from the fact that, in mammals, neighbor genes tend to have similar GC-contents (because of the isochore structure of mammalian genomes). Thus, measuring average GC-content and expression breadth in sets of neighboring genes may have the same consequence as the grouping of genes with similar GC-content: the grouping of genes with similar GC-content results in a decrease of the variance in expression, and hence to an increase in the existing correlation (as mentioned previously).

To distinguish between these hypotheses, we first assessed the correlations between GC-content and expression breadth after averaging both variables over neighboring genes (19). As shown in Table 4, the correlations increase steadily with window size (i.e. the number of genes per window), up to $R^2 = 52\%$ for SAGE data and $R^2 = 72\%$ for microarray data for a window of 100 genes (which represents in average a genomic fragment of 50 Mb). We then re-assessed the correlations after having permutated genes in the genome, keeping the isochore structure unchanged. More precisely, we classified the genes according to their intronic GC-content into 20 categories of equal size, and permutated genes within each of these categories. Correlations between mean expression breadth per window and mean GC-content per window were then computed. As shown in Table 3, after permutations, we still observed strong correlation between GC-content and expression breadth measured over 'neighboring' genes: up to $R^2 = 29\%$ for SAGE data and $R^2 = 58\%$ for microarray data for a window of 100 genes. These results indicate that the strong correlations reported by Lercher *et al.* (19) between average regional expression breadth and GC-content, are mainly a consequence of the fact that neighbor genes have similar GC-content.

DISCUSSION

In agreement with previous reports (Table 1), we observed that both in mouse and in human, the different measures of gene expression generally show a positive correlation between the GC-content of genes and their breadth of expression. However, these correlations are very weak: in the entire human data set, the percentage of the variance of gene expression breadth explained by the correlation with GC3 or GCi (R^2-values) are, respectively, 0.09 and 1.21% for SAGE data (12 205 genes, 18 tissues) and 1.72 and 3.33% for microarray data (6197 genes, 25 tissues). The relationship between GCi and expression breadth is hardly visible (Table 2). In mouse, the correlations are even weaker (Table 2).

In contradiction with these results, Lercher *et al.* (19) reported strong correlations between expression breadth and GC-content, which led them to predict 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 to conclude that there is probably a selective pressure favoring the concentration of housekeeping genes in GC-rich regions. The discrepancy with our results is because of the fact that Lercher *et al.* (19) computed their correlations not on individual genes but on groups of genes. We would like to stress that this grouping of genes is strongly misleading because it suggests that there exists a strong relationship between the expression breadth of genes and their GC-content, whereas in reality the relationship is very weak. Indeed, the correct interpretation of the strong correlations obtained with groups of genes is that if the average GC-content of a large set of genes is known, then it is possible to predict the average expression breadth. However, in contradiction with the conclusion of Lercher *et al.* (19), it is impossible to predict the expression of any particular individual gene in this set.

This work illustrates the problem of over-interpretation of statistical tests that is becoming recurrent in genomics. Thanks to the very large amount of data presently available, it is possible to detect extremely weak correlations that are significantly different from zero. However, what is the real usefulness of correlations that have such low predictive values? Correlation is not causality, and such weak correlations may reflect indirect relationships with some unknown variables. Moreover, as illustrated by the conflicting results obtained with human ESTs, they are very sensitive to possible methodological artifacts. In conclusion, although these correlations are statistically significant, it is difficult to assess their real biological significance.

Table 4. Correlation between intron GC-content and gene expression breadth, computed on windows of neighboring genes: impact of window size

Window			Real data R^2% (*P*-value)		Permutated genome R^2%[d]	
No. genes[a]	Size[b] (Mb)	No. windows[c]	SAGE	Microarray	SAGE	Microarray
1	NA	5977	1 % (10^{-15})	4 % (10^{-16})	NA	NA
5	2.0 (1.1)	1185	10 % (10^{-16})	20 % (10^{-16})	4	13
10	4.4 (3.2)	586	18 % (10^{-16})	32 % (10^{-16})	7	21
15	6.9 (5.4)	387	23 % (10^{-16})	39 % (10^{-16})	9	28
20	9.3 (7.8)	290	27 % (10^{-16})	47 % (10^{-16})	11	31
50	24.0 (23.5)	108	44 % (10^{-16})	62 % (10^{-16})	22	49
100	49.6 (47.3)	48	52 % (10^{-9})	72 % (10^{-14})	29	58

[a]Number of genes per window.
[b]Mean (median) of window size in Mb.
[c]Number of windows in the data set. R^2 (%) and *P*-values of correlations reassessed after averaging variables (expression breadth and intronic GC-content) on neighboring gene windows.
[d]Data set obtained after randomly permuting genes of similar intron GC-content. R^2 (%) and *P*-values of correlations reassessed after averaging variables (expression breadth and intronic GC-content) on the new 'neighboring' gene windows.

MATERIALS AND METHODS

Gene selection

We selected all human and mouse manually curated mRNAs from the RefSeq database (18), for which the expression could be computed from SAGE, EST and microarray data. We mapped them on the human genome [Ensembl, release 16.33, August 2003 (22)] or mouse genome (Ensembl, release 18.30, November 2003) using Ensembl links between CDS and RefSeq mRNAs. CDS for which total intron length was >1000 bp were retained to compute intronic GC-content. Orthologous gene pairs were found using the Hovergen database (23).

SAGE data

We performed the association between RefSeq mRNAs and SAGE data by determining the tags corresponding to each mRNAs. In total, 1% of the mRNA sequences lack the site *Nla* III (190 mRNAs out of 19 025 for human mRNAs), and were removed from the data set. The tag (10 pb upstream of the most 3′ *Nla* III site) was extracted from the other sequences. In some cases, one tag may match to more than one Refseq mRNA. We looked at the genomic location of these mRNAs to determine whether they correspond to alternative transcripts of a same gene or to different genes. In the latter case, genes were removed from the data set. We finally retained 13 435 human and 8951 mouse Refseq mRNAs that are non-redundant and unambiguously located on the human genome.

SAGE experiment results, called 'libraries', were obtained on the SAGE Genie website [ftp://cgap.ncbi.nih.gov/SAGE/Download (24)] for human data and on Gene Expression Omnibus site [http://www.ncbi.nlm.nih.gov/pub/geo/ (25)] for mouse data. Each of them contains a list of tags that corresponds to a sample of the transcriptome in a given tissue at a given developmental time. We retained 141 libraries for the human data set (41 for mouse) containing more than 20 000 tags and not corresponding to tumoral tissues. The libraries were then grouped into 19 tissues types (11 for mouse). After adding all counts for libraries representing the same tissue type, we converted absolute tag counts to relative tag counts (c.p.m., count per million).

EST

We selected from GenBank (release 133, December 2002) 4 906 743 ESTs from human tissues and 3 660 463 ESTs from mouse tissues. cDNA libraries from cell culture, tumors, pooled organs or unidentified tissues were excluded. To limit stochastic variations in expression measures, we only retained cDNA libraries that had been sampled with at least 10 000 ESTs. We retained 44 non-tumoral tissues for human and mouse data sets. CDS were then compared with the EST data set by using MEGABLAST (26). MEGABLAST alignments showing at least 95% identity over 100 nucleotides or more were counted as a sequence match. This criterion was chosen to be low enough to allow the detection of most ESTs despite sequencing error, but stringent enough to distinguish in most cases different members of highly conserved gene families. Normalization of the absolute tag count was done as for SAGE data.

Microarray

Oligonucleotide microarray data were extracted from the Gene Expression Atlas [http://expression.gnf.org (27)] that contains 25 human non-tumoral tissues and 45 mouse non-tumoral tissues. The sample replicates corresponding to the same tissue were averaged. The signals of probes corresponding to the same gene were averaged. In total, 7735 different human mRNAs and 5297 mouse mRNAs are represented into the resulting data set. As recommended by the authors (27), genes whose expression level exceeded 200 arbitrary units were noted as expressed.

Final data sets

For human data sets, 11 tissues are common to the three methods (blood, brain, heart, kidney, liver, lung, ovary,

pancreas, placenta, prostate and uterus) and expression could be evaluated for 6246 RefSeq mRNAs. For each of these genes, we calculated expression breadth (number of tissues with positive expression), expression mean (average level of expression for expressed genes) and peak rate (maximum level of expression), using each of the three methods.

For mouse data sets, very few tissues common to the three methods were available, and we maintained one separate data set for each method. Expression could be evaluated for 26 749 mRNAs and 45 tissues with EST data, 6906 mRNAs and 11 tissues with SAGE data and 5297 mRNAs and 45 tissues with microarray data. The statistical analyses were done using R (28).

ACKNOWLEDGEMENTS

We thank Vincent Navratil for providing us the orthologs data sets, Laurent Gueguen, Anne Beatrice Dufour and Eric Tannier for their help in statistics.

REFERENCES

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C. and Bernardi, G. (1991) The distribution of genes in the human genome. *Gene*, **100**, 181–187.
3. Duret, L., Mouchiroud, D. and Gautier, C. (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.*, **40**, 308–317.
4. Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y. and Ikemura, T. (2002) Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Hum. Mol. Genet.*, **11**, 13–21.
5. 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.
6. Jabbari, K., Rayko, E. and Bernardi, G. (2003) The major shifts of human duplicated genes. *Gene*, **317**, 203–208.
7. Smit, A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.*, **9**, 657–663.
8. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–958.
9. Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**, 3–17.
10. Galtier, N., Piganeau, G., Mouchiroud, D. and Duret, L. (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, **159**, 907–911.
11. Eyre-Walker, A. and Hurst, L.D. (2001) The evolution of isochores. *Nat. Rev. Genet.*, **2**, 549–555.
12. Goncalves, I., Duret, L. and Mouchiroud, D. (2000) Nature and structure of human genes that generate retropseudogenes. *Genome Res.*, **10**, 672–678.
13. Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.*, **12**, 640–649.
14. Ponger, L., Duret, L. and Mouchiroud, D. (2001) Determinants of CpG islands: expression in early embryo and isochore structure. *Genome. Res.*, **11**, 1854–1860.
15. Vinogradov, A.E. (2003) Isochores and tissue-specificity. *Nucleic Acids Res.*, **31**, 5212–5220.
16. Urrutia, A.O. and Hurst, L.D. (2003) The signature of selection mediated by expression on human genes. *Genome Res.*, **13**, 2260–2264.
17. 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.
18. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2003) NCBI reference sequence project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
19. 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.
20. 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.
21. 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.
22. Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928
23. Duret, L., Mouchiroud, D. and Gouy, M. (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.*, **22**, 2360–2365.
24. Liang, P. (2002) SAGE Genie: a suite with panoramic view of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11547–11548.
25. Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
26. Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
27. Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
28. Ihaka, R. (1996) R: a language for data analysis and graphics. *J. Comp. Graph. Genet.*, **16**, 418–420.