Intergenic Distance in the Human Genome

We have analyzed intergenic distances and searched for the presence of bidirectional genes using the complete sequences and mapping information of human chromosomes 20, 21, and 22, which contain 2,122 known and predicted genes. Intergenic distances between genes with divergent transcripts were distributed in a biphasic manner with a strong peak of 25 kb and a weak peak of 0.3 kb between the divergent transcripts, suggesting that the genes might share a common promoter. The weak peak was not observed at the transcriptional ends of genes. Seventy-three percent (55/75 pairs of genes, from a total of 150 genes) of these divergent transcripts located within 1 kb of one another were CpG islands. Expression of the divergent transcript genes was not concordant in various human tissues, suggesting that they were independently regulated. Analyses of the frequency of occurrence of interspersed repeats in the intergenic sequences suggested that these repeats are strongly excluded from the regions of transcriptional starts. This exclusion might be responsible for the existence of these divergent transcripts.

In a recent report, Adachi and Lieber (2002) suggested that 20% of human genes were located within 1 kb of one another and that this frequency was higher in DNA repair genes. To evaluate this phenomenon in larger gene subsets, we have analyzed intergenic distances using the complete sequences of human chromosomes 20 (Deloukas et al. 2001), 21 (Hattori et al. 2000), and 22 (Dunham et al. 1999) in conjunction with gene mapping information from the GenBank database. We located the 5’ region of all known and predicted genes on these chromosomes and categorized them into four groups; genes with divergent transcripts, two genes sharing the same strand, genes overlapping with other genes, and genes located at the ends of contigs, which are the largest units of continuous sequences available or are adjacent to a sequence gap and cannot be accurately located (fig. 1a). The distributions of these categories were calculated, and the proportions are indicated in figure 1b, which also shows the distances between transcriptional start sites of divergent genes. Approximately half of the genes have divergent transcripts, whereas the other half are located on the same strand. This proportion is in accordance with the assumption that genes are located randomly in the genome so that for any given gene the direction of a gene located 5’ in the same direction or the opposite direction is 50%.

Lavia, Macleod, and Bird (1987) and Adachi and Lieber (2002) described an association between divergent transcripts and CpG islands; therefore the frequency of CpG islands in these divergent transcripts was also analyzed. Applying the CpG island searcher (Takai and Jones 2003) at the default parameters of %GC ≥ 55 ObsCpG/ExpCpG ≥ 0.65 and Length ≥ 500 bp, CpG islands were found in 73% of 55 of a total of 75 pairs of genes which were located within 1 kb of one another. This frequency of CpG islands in divergent transcripts appears, at first sight, to be significantly higher than a previous analysis (Takai and Jones 2002) (46%, 161 CpG islands in 350 genes). However, the value is as expected if CpG islands are not preferentially located in bidirectional promoters; each divergent transcript consists of two genes, and the expected frequency of CpG islands in the middle of bidirectional transcripts is calculated as 1 – (1 – 0.46) × (1 – 0.46) = 0.71, so that the observed frequency fits quite well with this expected value. Therefore, in contrast to the findings of Adachi and Lieber (2002), we do not find that CpG islands are preferentially associated with bidirectional promoters. Perhaps the reason for this discrepancy lies in the definition of a CpG island which is not clearly specified at the UCSC Web site. The frequency of overlapping genes on these chromosomes was also evaluated (fig. 1c). Overlapping sequences were seen in 182 genes, of which 48 shared 5’ regions, 35 were located on the same strand and downstream of others, and 41 were found within other genes. The genes in these three categories account for 124 “overlapping” genes in figure 1b. Additionally, 58 genes shared 3’ regions, and these were included in figure 1 according to the location of their 5’ ends rather than their 3’ ends.

To further investigate the underlying meaning of the values of intergenic distance, we applied the Kernel Density Estimation method (Simonoff 1996), which can be thought of as giving smoothed histograms with small Gaussian curves centered at each different value (fig. 1d–g). These plots, shown on logarithmic scales, show almost the same biphasic shape in chromosomes 20, 21, and 22—a major peak at 25 kb and a minor peak centered at 0.3 kb consisting of bidirectional transcripts within 1 kb of one another. These plots reveal that the distributions of intergenic distances of chromosomes 21 and 22 are not so different from each other as previously reported median values of intergenic distances of chromosomes 21 and 22 (Chen et al. 2002). In addition to these small chromosomes, a plot of chromosome 14, which has relatively low gene density as compared with chromosome 20 and 22, is indicated in figure 1g. This minor peak is not observed between the ends of genes (fig. 1h) or between the ends and starts of genes (fig. 1i), suggesting that the close proximity
might have been maintained for some functional reason. These bidirectional transcripts make up approximately 10% of genes; however, the actual first exons might have been missed in many genes, especially in predicted ones, and our estimate should therefore be considered a lower limit. On the other hand, Adachi and Lieber (2002) used expressed sequence tags (ESTs) to define the 5' regions of genes in many cases. Because of the way that ESTs are used to define genes, this could lead to an overestimation of the frequency of bidirectional transcripts.

The biphasic nature of the distribution of genes on human chromosomes 20–22 raises interesting questions regarding its significance. One possibility is that genes sharing promoters might be co-ordinately regulated in different tissues. However, this does not seem to be the case, because genes with divergent promoters within 1 kb of each other and genes that share 5' ends did not show a higher degree of concordant expression than other divergent transcript in the cancer genome anatomy project (CGAP) database (fig. 2). An alternative explanation...
might stem from the evolution of the human genome from a more compact genome. A large part of this expansion is thought to have been due to the radiation of repetitive elements (Smit 1996). Perhaps the bidirectional genes have been maintained because interference with the shared promoter, caused by insertion of a repetitive element, would result in the simultaneous disruption of the regulation of two genes.

We therefore assessed how often repetitive elements were present in the vicinity of start sites by analyzing the content in 100-bp windows at various distances from the point where transcription begins. Analysis of these windows in 578 genes which did not have divergent promoters within 10 kb of each other showed that repetitive elements were strongly excluded from the first 300 bp of such start sites (fig. 3a). Repetitive elements were also excluded, although less strongly, from the ends of transcripts (fig. 3b). Figure 3c shows that divergent promoters which had less than 1 kb between start sites were virtually devoid of repetitive elements. Because the average length of these promoters was 300 bp (fig. 1d), we also analyzed the content of repetitive elements in the 300-bp upstream regions of divergent transcripts separated by more than 10 kb (fig. 3c). The content of repetitive elements in the 150-bp upstream regions of divergent transcripts separated by more than 10 kb is also indicated in figure 3c. These data show that divergent promoters within 1 kb of each other are distinct in that they contain low levels of such repetitive elements. Perhaps these observations can explain the persistence of close divergent transcripts in the human genome after its presumed expansion from a more compact form. If promoters or essential sequences necessary for regulation of gene expression (Zhang 1998), like CpG islands, are inherently refractory to invasion by transposable elements (fig. 4), then it would follow that two overlapping refractory zones might become superimposed on each other, thus preventing their separation. This has happened in the majority of the genome.
different tissues. For each combination of tissue, expression was computed by dividing the number of ESTs or SAGE tags representing the gene, divided by the total number of ESTs or SAGE tags in all libraries with the given tissue. This ratio was then multiplied by 200,000, giving the number of ESTs or SAGE tags per 200,000. Then Pearson’s coefficient value was calculated for each combination of divergent or overlapping genes. If both genes have zero as the number of ESTs or SAGE tags in a tissue, that value was not used. This was done to avoid the introduction of bias of the coefficient by excessive data at origin. Genes with fewer than five informative expression data points of tissues were also excluded. As a control, pairs of genes located on different chromosomes were randomly selected from genes which makes divergent transcript or from genes that share 5’ ends.

For kernel density estimation (Simonoff 1996), we have used distribution function given by

\[
f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),
\]

where kernel function \( K \) is given by

\[
K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.
\]

As the bandwidth \( h \), we have used

\[
h_0 = \left[ \frac{R(K)}{\sigma^2 K(R(f))} \right]^{1/5} n^{-1/5} = 1.059\sigma n^{-1/5}.
\]

Repetitive elements were detected by the Repeat-Masker (University of Washington Genome Center, Seattle, http://ftp.genome.washington.edu/cgi-bin/RepeatMasker).

Acknowledgments

We thank T. Takai, G. Coetzee, D. Shibata, P. Laird, and M. Lieber for critical reading of the manuscript. D.T. and P.A.J. are supported by National Cancer Institute grants R01 CA82422 and R01 CA 83867.

Literature Cited


Materials and Methods

For analyses of chromosomes 20, 21, and 22, we obtained sequence and mapping information from the GenBank database. We used the contigs (build 30), NT_011387, NT_025215, NT_028392, NT_011362, NT_030871, NT_35608, NT_011333 (chromosome 20), NT_029490, NT_011512, NT_030187, NT_030188, NT_011515 (chromosome 21), NT_011516, NT_028395, NT_011519, NT_011520, NT_011521, NT_011522, NT_011523, NT_030872, NT_011525, NT_019197, NT_011526 (chromosome 22), NT_026437.10 (chromosome 14). The potential start position of a gene was based on GenBank mapping information, and possible multiple transcriptional start sites were not assessed. To obtain intergenic distances and status of direction, genes which were located in the 5’ end of contigs and genes which were included in other genes were excluded. The status of the 5’ end of genes was then evaluated.

For CpG island identification, the sequences between divergent promoters were extracted, or centered 500-bp sequences were extracted if distances between divergent promoters were less than 500 bp. We then applied the CpG islands searcher (http://www.uscnorris.com/cpgislands/) and its command line version (Takai and Jones 2003) with the criteria %GC ≥ 55 ObsCpG/ExpCpG ≥ 0.65 and Length ≥ 500 bp. The algorithm to search for CpG islands is described in our earlier report (Takai and Jones 2002). Other analyses were also done with programs coded by D.T.

For expression information, we used the “tissue” section of the cancer genome anatomy project Web site (http://cgap.nci.nih.gov). Correlation coefficient values were determined from the serial analysis of gene expression (SAGE) data and EST data for up to 50


Associate Editor, Peer Bork

Accepted October 09, 2003