Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 11995–11999, December 1993
Genetics

Number of CpG islands and genes in human and mouse

FRANCISCO ANTEQUERA AND ADRIAN BIRD

Institute of Cell and Molecular Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JR, Scotland

Communicated by Stanley M. Gartler, September 17, 1993

ABSTRACT **Estimation of gene number in mammals is difficult due to the high proportion of noncoding DNA within the nucleus. In this study, we provide a direct measurement of the number of genes in human and mouse. We have taken advantage of the fact that many mammalian genes are associated with CpG islands whose distinctive properties allow their physical separation from bulk DNA. Our results suggest that there are ≈45,000 CpG islands per haploid genome in humans and 37,000 in the mouse. Sequence comparison confirms that about 20% of the human CpG islands are absent from the homologous mouse genes. Analysis of a selection of genes suggests that both human and mouse are losing CpG islands over evolutionary time due to *de novo* methylation in the germ line followed by CpG loss through mutation. This process appears to be more rapid in rodents. Combining the number of CpG islands with the proportion of island-associated genes, we estimate that the total number of genes per haploid genome is ≈80,000 in both organisms.**

The total number of genes in organisms like *Escherichia coli* and *Saccharomyces cerevisiae* can be approximately estimated by extrapolation from regions of their genomes that have already been sequenced, because most of their DNA contains coding information (1, 2). In contrast, genes in mammals are made of exons dispersed along very variable lengths of DNA, and the distance between genes is often larger than the genes themselves. As a consequence, a mammalian DNA sequence does not immediately provide information about the number and position of the genes that it contains. The most often quoted figure is 50,000–100,000 genes per haploid genome, although the derivation of this figure is unspecified. The only systematic experimental approach to the measurement of gene number in mammals has relied on analysis of the complexity of the mRNA populations (reviewed in ref. 3). Reassociation experiments suggested the existence of 30,000–40,000 genes. However, difficulties in obtaining full representation of low-abundance RNAs from all cell types make it likely that this is an underestimate.

Ideally, one would like to physically separate gene from non-gene DNA, thus providing a direct estimate independent of gene expression. Such a strategy requires the identification of genes within the structure of a DNA molecule for which there is no functional information. CpG islands make this approach possible because they can be used as gene markers in the mammalian genome (see refs. 4–6 for reviews). CpG islands constitute a distinctive fraction of the genome because, unlike bulk DNA, they are nonmethylated and contain the dinucleotide CpG at its expected frequency. Another notable property of CpG islands is that their G+C content is significantly higher than that of non-island DNA. This facilitates their identification even in cloned DNA, where the native methylation pattern has been erased (7).

In this study, we have exploited these properties to separate CpG islands from the rest of the genome and determine their absolute number in human and mouse. A previous approach to the number of CpG islands was carried out in mouse by quantitation of end-labeled restriction fragments generated upon digestion of total genomic DNA with the methyl-sensitive restriction endonuclease *Hpa* II (8). An approximate figure of 30,000 CpG islands per haploid genome was suggested. Our results show significant differences between mouse and human that are relevant to our understanding of the origin and maintenance of CpG islands.

Because not all genes have CpG islands, the total number of genes cannot be deduced directly from their number. We have taken the study further by establishing the proportion of genes that are CpG island-associated. Combining the number of CpG islands per genome and the percentage of CpG island-associated genes, we obtain a direct estimate of the total number of genes in human and mouse.

MATERIALS AND METHODS

Cell Culture Conditions. The human lymphoblastoid PES cell line (9) was grown in RPMI 1640 medium containing 10% tryptose phosphate broth and supplemented with 10% fetal bovine serum. The mouse embryo stem-cell line EFC-1 (10) was grown in Glasgow minimal essential medium supplemented with 10% fetal bovine serum, 0.1 mM 2-mercaptoethanol, and 1 mM nonessential amino acids.

Metabolic Labeling of Cells and Isolation of Nuclei. Exponentially growing unsynchronized cells were incubated with [*methyl*-^{3}H]thymidine (1 μCi/ml; specific activity, 5.0 Ci/mmol; 1 Ci = 37 GBq) for 12 hr, followed by a further 12 hr with a second aliquot of label. Cells were washed with ice-cold phosphate-buffered saline (PBS) and suspended in lysis buffer 1 (10 mM Tris·HCl, pH 7.5/10 mM NaCl/2 mM $MgCl_2$) or in lysis buffer 2 (25 mM KCl, 20 mM Hepes, pH 7.8/0.15 mM spermine/0.5 mM spermidine/1 mM EDTA/0.5 mM EGTA/0.23 M sucrose) and lysed by addition of Nonidet P-40 (0.05%). Nuclei were purified by two rounds of centrifugation of the cell lysates through a 5-ml 1.2 M sucrose cushion for 3 min at 5000 rpm in a Beckman JS13.1 swing-out rotor. The nuclear pellet was resuspended in lysis buffer without Nonidet P-40 and lysed by incubation in 10 mM EDTA/0.5% SDS with proteinase K (200 μg/ml) at 37°C for 2–3 hr. DNA was extracted with phenol and chloroform, ethanol-precipitated and resuspended in 10 mM Tris·HCl/1 mM EDTA, pH 8.

Sucrose Gradients. After digestion of purified DNA with *Hpa* II, the restriction fragments were loaded on a seven-step sucrose gradient (5%–20% in steps of 2.5%) made up in 10 mM Tris·HCl, pH 7.4/50 mM NaCl/1 mM EDTA. Gradients were centrifuged at 20°C for 3 hr at 50,000 rpm in a Beckman SW-60 swing-out rotor. Seven fractions of 550 μl were collected from each gradient; 20 μl of each fraction was set aside for end-labeling, and the rest was precipitated with an equal volume of 10% (wt/vol) trichloroacetic acid. Precipitates were collected on glass fiber filters and the incorporated radioactivity was measured by scintillation counting.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

RESULTS

Number of CpG Islands in Human and Mouse. A human lymphoblastoid cell line (PES) and a mouse pluripotent embryonic stem cell line (EFC-1) were chosen for analysis. Both lines were derived recently and have normal male diploid karyotypes (9, 10). They show patterns of genomic DNA methylation indistinguishable from samples derived from human and mouse tissues. In particular, we found no evidence of *de novo* methylation of CpG islands compared with tissue DNA, either at tissue-specific genes (11) or when the total CpG island fraction was measured by end-labeling (data not shown). Thus we considered them to be representative of the animal tissue. We chose not to use primary human cells, due to reported variability of methylation levels which could potentially interfere with the analysis (12).

To quantitate the percentage of the total genome represented by CpG islands, cells were first labeled with [^{3}H]thymidine. DNA was purified from isolated nuclei and digested to completion with the methyl-sensitive endonuclease *Hpa* II (recognition sequence, CCGG). CpG islands were predominantly cut to fragments smaller than 500 bp due to the high density of nonmethylated *Hpa* II sites (8, 13). Bulk genomic DNA, on the other hand, gave large fragments, as *Hpa* II sites are comparatively infrequent and mostly methylated. The fragments were fractionated in a sucrose gradient and the radioactivity in each fraction was determined. Fractions 1 and 2 contained fragments smaller than 70 bp and 500 bp, respectively (Fig. 1). The average size of CpG-island fragments was lower in human cells than in mouse cells, due to the significantly higher G+C content of islands in the human genome (see below). The radioactivity in fractions 1 and 2, relative to the total in all seven fractions served as a basis for estimating the total number of CpG islands (Table 1).

The accuracy of the estimates was improved by the following corrections (Table 1). First, as some contamination with genomic fragments derived from non-island bulk DNA was expected, the true proportion of fragments in fractions 1 and 2 derived from CpG islands was determined by cloning at random a number of fragments from these fractions and determining the percentage that were CpG island-like by

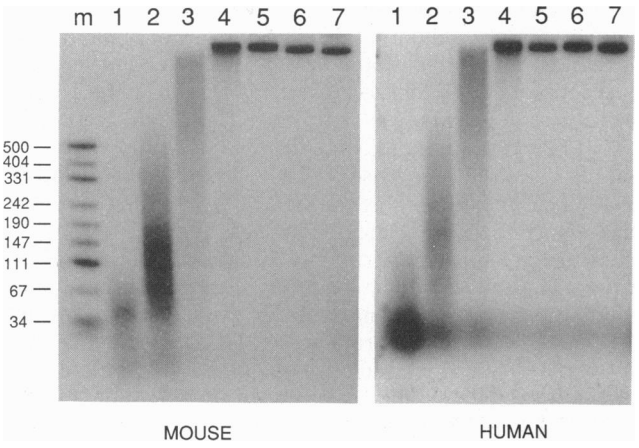

FIG. 1. Fractionation of human and mouse *Hpa* II-digested DNA after end-labeling with ^{32}P. Uniformly [^{3}H]thymidine-labeled DNA was digested to completion with *Hpa* II and the restriction fragments were fractionated in a sucrose gradient. To monitor fractionation, a small amount of DNA from each of the seven fractions was end-labeled with [α-^{32}P]dCTP and the fragments were separated in a high-resolution 2.5% NuSieve (FMC)/1% agarose composite gel. Lanes 1–7 represent gradient fractions from top to bottom; lane m, size markers in base pairs. The smaller average size of human CpG-island fragments compared with those of mouse (fractions 1 and 2) is due to their greater G+C content, resulting in an increased frequency of *Hpa* II (CCGG) sites (see text).

Table 1. Quantitation of the number of CpG islands in human and mouse

Species	Exp.	% ^{3}H in fr. 1 + 2*	% of total genome†	No. of CpG islands‡	No. of genes§
Human	1	1.26	1.72	51,600	92,307
	2	1.01	1.33	39,900	71,377
	3	1.10	1.47	44,100	78,890
Mean				45,200	80,858
Mouse	1	0.77	1.30	39,000	83,155
	2	0.69	1.15	34,500	73,560
Mean				36,750	78,357

*Percentage of the total ^{3}H cpm present in gradient fraction (Fr.) 1 and 2 (Fig. 1). The DNA in every fraction of the sucrose gradient was precipitated with trichloroacetic acid onto glass fiber filters and the radioactivity was measured. The experiments were repeated twice in human and once in mouse, using two different methods to prepare nuclei (see *Materials and Methods*). Results for each experiment are shown.

†Percentage of the total genome represented by CpG islands was calculated by multiplying the percentages in column 1 by three correcting factors: (*i*) contamination with non-CpG-island fragments (0.714 and 0.791 in human and mouse, respectively); (*ii*) fraction of the CpG islands released upon *Hpa* II digestion (1.15 and 1.42); (*iii*) differential base composition between CpG island and bulk DNA (1.88 and 1.68). From the resulting product, we subtracted the contribution of the rRNA genomic clusters [0.22% and 0.15% of the genome in human and mouse, respectively (14); see text].

‡Absolute number of CpG islands per haploid genome was calculated by multiplying the fraction of the total genome represented by the CpG islands by 3×10^9 (the genome size in base pairs) and dividing by 10^3 [the average size of CpG islands in both organisms is 1 kb (refs. 8 and 15 and data not shown)].

§The total gene number was calculated by dividing the number of CpG islands by the fraction of genes in the database that are CpG island-associated. This percentage was estimated by selecting genomic sequences from the GenBank/EMBL database (see text). Altogether, 152 human and 81 mouse genes were screened. Of these, 55.9% and 46.9% had CpG islands, respectively.

DNA sequencing. The criteria were a G+C content of at least 50% (bulk genome content is 40%) and an observed/expected CpG ratio above 0.6 (the ratio for bulk DNA in mouse and human is 0.25). Altogether, we sequenced 28 human and 25 mouse clones. Analysis of the sequences showed that 71.4% (human) and 79.1% (mouse) of the fragments were derived from CpG islands (data not shown).

A second correction was needed because some CpG islands are not quantitatively cut by *Hpa* II to fragments smaller than 500 bp; some sites are occasionally more than 500 bp apart, and some extend across the edge of the island into non-island DNA. The fraction of an average CpG island that was released into fractions 1 and 2 upon *Hpa* II digestion was estimated by analysis of 46 CpG island sequences in the GenBank/EMBL sequence database. The analysis showed that 87% (human) and 70% (mouse) of an average CpG island is cut by *Hpa* II to fragments smaller than 500 bp. The difference between human and mouse is due to base composition differences between their islands (see below).

For the third correction, [^{3}H]thymidine incorporation was adjusted to take account of the average base composition of CpG islands in human and mouse. Based on the same 46 sequences, the G+C contents are 67.1% and 64.3%, respectively, whereas bulk DNA is 40% in both species. Thus, thymidine is underrepresented in CpG islands relative to non-CpG island DNA by a factor of 1.88 in human and 1.68 in mouse.

The fourth correction was required to subtract sequences derived from 18S and 28S rRNA genes (rDNA) from the island fraction. Like CpG islands, rDNA has a high G+C content and is mainly nonmethylated (16, 17). The proportion

of rDNA fragments in fractions 1 and 2 was estimated by hybridizing the complete rDNA repeat unit to a panel of cloned fragments from these fractions. The results for both mouse and human indicated that about 10% of the clones were rDNA (ref. 8 and data not shown), which is the proportion expected if all genomic rDNA was in these low molecular weight fractions. A more recent estimate, based on hybridization to a library of 375 CpG island-like sequences, showed that 37 (or 9.9%) were derived from rDNA (S. Cross, personal communication). We therefore assumed that the entire rDNA cluster is present in fractions 1 and 2 and have corrected accordingly (see legend to Table 1).

Applying these four corrections and considering 1 kb as the average length of CpG islands (8, 15), we deduced that the approximate number of CpG islands per haploid genome is 45,000 in human and 37,000 in mouse (Table 1).

Number of Genes in Human and Mouse. We next used the GenBank/EMBL sequence database (release 33) to determine the proportion of genes that are associated with CpG islands. Only complete transcription units whose entire sequence, including 5′ and 3′ flanks, was present in the database were analyzed. On this basis, 152 human and 81 mouse sequences were selected. Of these, 55.9% in human and 46.9% in mouse had CpG islands (data not shown). An independent survey of 362 genes transcribed by RNA polymerase II gave a proportion of island-associated genes in human of 57.0% (15), which is very close to our estimate. Considering that there are approximately 45,000 and 37,000 CpG islands in the human and mouse genomes and that 55.9% and 46.9% of all the genes are associated with islands, respectively, we estimate the total number of genes to be around 80,000 in both organisms (Table 1). Thus, even though the number of islands is different, the total number of genes appears to be similar. Our estimate depends on the assumption that each CpG island identifies a gene. While it is possible that some are not associated with genes (18), the number is likely to be small, as gene mapping studies across long chromosome regions have shown that novel CpG islands are usually associated with transcripts (19, 20). In a few cases genes with two associated CpG islands have been reported (5). Of the 124 human and mouse genes associated with 5′ CpG islands in our survey, only 3 had another at the 3′ end of the gene. It is not known whether these should be considered 3′ CpG islands or whether they are associated with the 5′ end of another gene immediately downstream. Occasionally, a CpG island includes divergent promoters for two genes (21, 22). We have assumed that this configuration is the exception rather than the rule.

Loss of CpG Islands in the Mouse Genome. The database survey showed a lower proportion of CpG island-associated genes in the mouse genome than in the human genome (16% less). This agrees with our biochemical finding that mice have about 19% fewer CpG islands (Table 1). Moreover, a comparison of the genes in our selection that had been sequenced in both species showed that out of 23 genes, 16 had a CpG island in human and 13 in mouse, a difference once again of 19%. We conclude that approximately 1 in 5 of the genes that have a CpG island in human have no CpG island in mouse.

Is the lower number of CpG islands in the mouse due to creation of new islands in human or to the loss of ancestral islands in the mouse lineage? To address this, we compared the sequences of the three genes from our selection that had CpG islands in human but not in mouse (α-globin, ζ-globin, and skeletal α-actin). This comparison showed that CpG dinucleotides in the human gene had most often been replaced in the mouse by TpG or its complement, CpA (Fig. 2). Since methylated CpG dinucleotides are known to mutate at an accelerated rate to TpG (23–25), the most likely explanation is that these three genes all had CpG islands in a common ancestor, but some time after the divergence of mouse and human lineages in evolution, the islands became *de novo* methylated in the mouse germ line, leading to progressive CpG loss. In line with this idea, the mouse α-globin gene is methylated in germ cells, unlike its human homolog (26, 27). Creation of new CpG islands in the human lineage would require a tendency for TpG/CpA sequences to mutate to CpG. Since no such tendency has been observed, loss by the mouse genome is a more plausible scenario than gain by the human genome.

DISCUSSION

CpG Islands and Genes. Our results show that the total number of islands in the mouse genome is about 20% less that in humans. Because a similar difference is seen in the proportion of genes associated with CpG islands, the estimated total number of genes is the same in both organisms. The method we have used is direct in the sense that it exploits the association of many genes to CpG islands, a characteristic that is independent of their level of expression. We estimate that there are ≈80,000 genes per haploid genome in both organisms, which refines and provides experimental support for the current guesses of 50,000–100,000 genes.

All housekeeping genes tested so far are associated with CpG islands. However, not all islands are associated with housekeeping genes. Well-studied examples of tissue-restricted genes which are associated with CpG islands are the human α-globin genes and the human and mouse MyoD1 and Thy-1 genes. Based on an extensive analysis of DNA sequences, Larsen *et al.* (15) have found that all the human housekeeping (HK) genes and 40% of tissue-restricted (TR) genes are associated with CpG islands. Applying this percentage to our estimate of ≈45,000 islands per human haploid genome means that HK + (0.4)TR = 45,000. In addition, since the total number of genes is around 80,000, then HK + TR = 80,000. Solving these equations, we estimate that there are ≈22,000 housekeeping genes and 58,000 tissue-restricted per human haploid genome. This number agrees well with the data obtained from RNA reassociation analysis (reviewed in ref. 3), which suggest that the number of tissue-restricted genes is 2–3 times the number of ubiquitously expressed genes.

The above calculation implies that only half of all CpG islands in the genome (around 22,000) are associated with housekeeping genes, the other half being at tissue-restricted genes. Support for this estimate comes from the observation that two mouse cell lines in culture have methylated, and therefore most probably inactivated, about half of all their CpG-island genes (11). In each case, the same subset of islands had been methylated, and analysis of specific genes showed that these were consistently tissue-restricted genes. This is the expected result, as tissue-restricted genes are probably dispensable in culture. The implication is that half of CpG island genes in the mouse genome can be regarded as essential housekeeping genes, thereby agreeing with the calculation above.

Evolution of CpG Islands. Loss of CpG islands may not be restricted to the mouse genome, as a similar process has been seen in the case of the human α-globin pseudogene (27). The difference in total number of islands between human and mouse may therefore result from a higher rate of loss in mouse. The results suggest a hypothesis for the sequence of CpG-island evolution, based on the idea that CpG islands arose once at the dawn of vertebrate evolution. According to this hypothesis, genes of an invertebrate ancestor were embedded in entirely nonmethylated DNA (as are all current invertebrate genes that we know of; see ref. 4). With the appearance of the vertebrates, DNA methylation spread through the genome, but promoters, which would be suppressed by dense methylation, were spared. The mechanisms

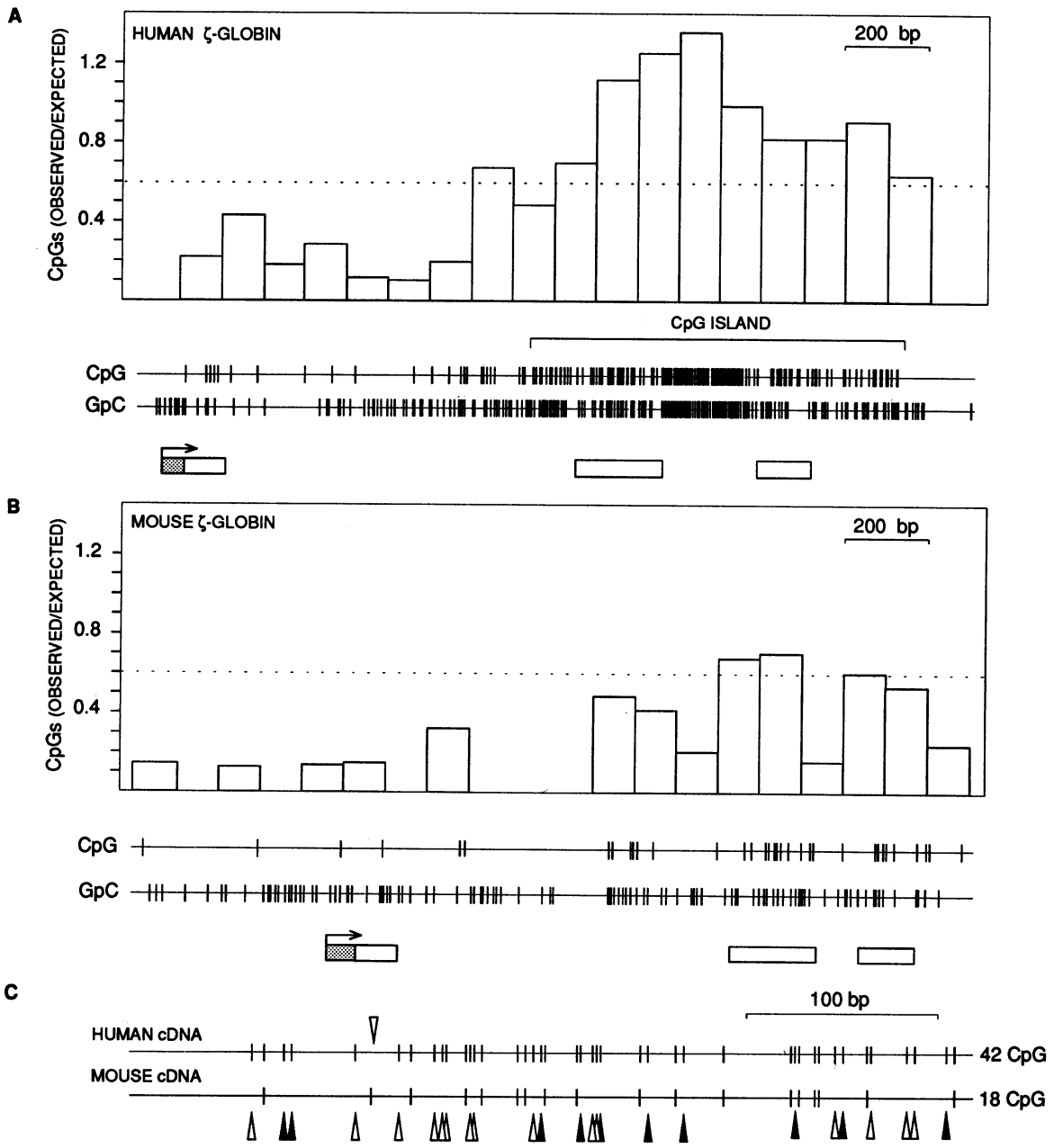

FIG. 2. The CpG island at the human ζ-globin gene has been lost at the corresponding locus in the mouse. (*A*) The histogram represents the observed/expected ratio of CpG dinucleotides across the 2.0-kb genomic region containing the human ζ-globin gene. Map plots below the histogram compare the densities of CpG and GpC over the same region. The CpG island is the region where the CpG observed/expected ratio exceeds 0.6 (dotted line on histogram) and the densities of CpG and GpC are similar (map plots). Outside the island, the CpG frequency falls to the bulk genome value of ≈0.25. Map plots show that in these regions CpG dinucleotides are underrepresented relative to GpC dinucleotides. Boxes represent the three exons of the gene. Arrow indicates the origin of transcription; hatched box shows the 5′ untranslated sequence. (*B*) The mouse ζ-globin gene represented as in *A*. Despite the overall conservation of the gene organization, the CpG island spanning the last two exons in the human gene is absent in the mouse. The observed/expected frequency of CpG dinucleotides and their underrepresentation relative to GpC dinucleotides are typical of methylated bulk DNA. Only coding sequences, spanning <200 bp, have an observed/expected ratio of about 0.6. Maintenance of some CpG dinucleotides across coding regions is most likely due to selection against changes that affect protein function. (*C*) Direct comparison of the protein coding sequences of the human and mouse ζ-globin genes. The two cDNAs are equal in length and their nucleotide sequences are 81.3% identical. Intron sequences showed no significant sequence similarity and were therefore excluded from the comparison. Of the 42 CpG dinucleotides in human, 17 are conserved in the mouse sequence. Black arrowheads indicate CpG dinucleotides present in the human sequence that have been replaced by TpG or CpA dinucleotides in the mouse. White arrowheads represent CpG dinucleotides that have been replaced by other dinucleotides (seven different dinucleotides altogether). Mutation from CpG to TpG/CpA accounts for 40% of all CpGs in the human coding sequence that are absent in the mouse. Only one CpG is present exclusively in the mouse cDNA (white arrowhead pointing downward). It has been replaced by CpC at the equivalent position in the human gene. We also compared the coding sequences of two other genes (α-globin and skeletal α-actin) that have CpG islands in human but not in mouse (see text). Substitution by TpG/CpA in the mouse homologous gene occurred at 60% and 72.5% of human CpG positions, respectively (data not shown).

of spread and immunity to methylation are unknown. Loss of methylated CpG dinucleotides by mutation and retention of CpG dinucleotides in the nonmethylated regions (coupled in some cases with increased G+C content) gave rise to today's obvious CpG islands. With time, an increasing proportion of genes have succumbed to methylation, and therefore loss, of

their CpG islands, without losing their capacity for expression. The genes that have accomplished this transition appear to be exclusively those with tissue-restricted expression patterns, as all known housekeeping genes are still associated with islands. It is noteworthy that the three genes in our survey associated with a CpG island in human but not in mouse are tissue-restricted genes.

Density and Distribution of CpG Islands in the Genome. CpG islands have been mapped across several contiguous regions of human DNA. Although the mapped regions represent a minute fraction of the total genome, it is possible to make an approximate estimate of the total number of islands that they predict, for comparison with our estimate. A region of 680 kb in the class III region of the human major histocompatibility complex on chromosome 6p21.3 contains 20 islands, which represents an average of one every 34 kb (28). The Huntington disease region in chromosome 4p16.3 showed 26 islands in ≈740 kb, which gives an average of 1 every ≈28 kb (20, 29, 30). An analysis of the *ERCC1* locus of human chromosome 19q13.3 has detected 6 islands in 106 kb of DNA, representing an average of about 1 in 18 kb (19). Another 6 islands have been mapped in a 180-kb region in human chromosome 10q11.2 at a region tightly linked to the gene responsible for multiple endocrine neoplasia type 2A. The density in this case is 1 every 30 kb (31). Finally, 9 islands have been detected in a contig spanning 850 kb in the Wilms tumor locus on chromosome 11p13, representing a density of one every 94 kb (32).

Although the regions concerned were sometimes described as unusually rich in CpG islands, the density is in fact comparable between them. When the data for the five loci are pooled, the average is one island every 36 kb. It has been reported previously that CpG islands are concentrated in the G-light bands of metaphase chromosomes and are relatively rare in G-dark bands (33, 34). If these bands represent about half of the human genome, and the island density is one per 36 kb, the predicted total number of CpG islands in G-light bands is 42,000 per haploid genome, which accounts for almost all the genomic total of 45,000 (Table 1). The data strongly support the idea that most or all CpG islands are confined to G-light bands. Accordingly, the assignment of genes to a G-light or G-dark band has been found to correlate well with the presence or absence of an associated CpG island (34).

We thank Donald Macleod, Eric Selker, Cathy Abbott, Richard Meehan, and Sally Cross for criticism of the manuscript. This work was supported by the Imperial Cancer Research Fund and the Wellcome Trust. F.A. is an Imperial Cancer Research Fund postdoctoral fellow.

1. Daniels, D., Plunket, G., III, Burland, V. & Blattner, F. (1992) *Science* **257**, 771–778.
2. Oliver, S., van der Aart, Q. J. M., Agostini-Carbone, M. L., Aigle, M., Alberghina, L., *et al.* (1992) *Nature (London)* **357**, 38–46.
3. Lewin, B. (1990) *Genes IV* (Oxford Univ. Press, Oxford, U.K.), pp. 466–481.
4. Bird, A. (1987) *Trends Genet.* **3**, 342–347.
5. Gardiner-Garden, M. & Frommer, M. (1987) *J. Mol. Biol.* **196**, 261–282.
6. Antequera, F. & Bird, A. (1993) in *DNA Methylation: Molecular Biology and Biological Significance*, eds. Jost, J. P. & Saluz, H. P. (Birkhauser, Basel), pp. 169–185.
7. Bickmore, W. & Bird, A. (1992) *Methods Enzymol.* **216**, 224–244.
8. Bird, A., Taggart, M., Frommer, M., Miller, O. & Macleod, D. (1985) *Cell* **40**, 91–99.
9. Cooke, H. & Smith, B. (1986) *Cold Spring Harbor Symp. Quant. Biol.* **51**, 213–219.
10. Nichols, J., Evans, E. & Smith, A. (1990) *Development (Cambridge, U.K.)* **110**, 1341–1348.
11. Antequera, F., Boyes, J. & Bird, A. (1990) *Cell* **62**, 503–514.
12. Schmookler-Reis, R. & Goldstein, L. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 3949–3953.
13. Cooper, D., Taggart, M. & Bird, A. (1983) *Nucleic Acids Res.* **11**, 647–658.
14. Long, E. & Dawid, I. (1980) *Annu. Rev. Biochem.* **49**, 727–764.
15. Larsen, F., Gundersen, G., Lopez, R. & Prydz, H. (1992) *Genomics* **13**, 1095–1107.
16. Bird, A. & Taggart, M. (1980) *Nucleic Acids Res.* **8**, 1485–1497.
17. Bird, A., Taggart, M. & Gehring, C. (1981) *J. Mol. Biol.* **152**, 1–17.
18. Fischel-Ghodsian, N., Nicholls, R. & Higgs, D. (1987) *Nucleic Acids Res.* **15**, 6197–6207.
19. Martín-Gallardo, A., McCombie, W., Gocayne, J., FitzGerald, M., Wallace, S., Lee, B., Lamerdin, J., Trapp, S., Kelley, J., Liu, L., Dubnick, M., Johnston-Dow, L., Kerlavage, A., de Jong, P., Carrano, A., Fields, C. & Venter, J. (1992) *Nat. Genet.* **1**, 34–39.
20. McCombie, W., Martin-Gallardo, A., Gocayne, J., FitzGerald, M., Dubnick, M., Kelley, J., Castilla, L., Liu, L., Wallace, S., Trapp, S., Tagle, D., Whaley, W., Cheng, S., Gusella, J., Frischauf, A., Poustka, A., Lehrach, H., Collins, F., Kerlavage, A., Fields, C. & Venter, J. (1992) *Nat. Genet.* **1**, 348–353.
21. Lavia, P., Macleod, D. & Bird, A. (1987) *EMBO J.* **6**, 2773–2779.
22. Colombo, P., Yon, J., Garson, K. & Fried, M. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 6358–6362.
23. Cooper, D. & Krawczak, M. (1989) *Hum. Genet.* **83**, 181–188.
24. Sved, J. & Bird, A. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4692–4696.
25. Jones, P., Rideout, W., Shen, J., Spruck, C. & Tsai, Y. (1992) *Bioassays* **14**, 33–36.
26. Boyes, J. (1991) Ph.D. thesis (Univ. of Vienna, Vienna, Austria).
27. Bird, A., Taggart, M., Nichols, D. & Higgs, D. (1987) *EMBO J.* **6**, 999–1004.
28. Milner, C. & Campbell, R. (1992) *Bioessays* **14**, 565–571.
29. Weber, B., Collins, C., Kowbel, D., Riess, O. & Hayden, M. (1991) *Genomics* **11**, 1113–1124.
30. Carlock, L., Wisniewski, D., Lorincz, M., Pandrangi, A. & Vo, T. (1992) *Genomics* **13**, 1108–1118.
31. Brooks-Wilson, A., Smailus, D. & Goodfellow, P. (1992) *Genomics* **13**, 339–343.
32. Bonetta, L., Kuehn, S., Huang, A., Law, D., Kalikin, L., Koi, M., Reeve, A., Brownstein, B., Yeger, H., Williams, B. & Feinberg, A. (1990) *Science* **250**, 994–997.
33. Bickmore, W. & Sumner, A. (1989) *Trends Genet.* **5**, 144–148.
34. Holmquist, G. (1987) *J. Mol. Evol.* **28**, 469–486.