Large-scale methylation analysis of human genomic DNA reveals tissue-specific differences between the methylation profiles of genes and pseudogenes

Christoph Grunau[1,+], Winfried Hindermann[2] and André Rosenthal[1,3,§]

[1]Department of Genome Analysis, Institute for Molecular Biotechnology, Beutenbergstrasse 11, D-07745 Jena, Germany, [2]Friedrich-Schiller-University, Department of Pathology, Ziegelmühlenweg 1, D-07743 Jena, Germany and [3]Department of Biology and Pharmacy, Friedrich-Schiller-University, D-07743 Jena, Germany

Received 27 June 2000; Revised and Accepted 8 September 2000

Cytosine in CpG dinucleotides is frequently found to be methylated in the DNA of higher eukaryotes and differential methylation has been proposed to be a key element in the organization of gene expression in man. To address this question systematically, we used bisulfite genomic sequencing to study the methylation patterns of three X-linked genes and one autosomal pseudogene in two adult individuals and across nine different tissues. Two of the genes, *SLC6A8* and *MSSK1*, are tissue-specifically expressed. *CDM* is expressed ubiquitously. The pseudogene, ψ*SLC6A8*, is exclusively expressed in the testis. The promoter regions of the *SLC6A8*, *MSSK1* and *CDM* genes were found to be essentially unmethylated in all tissues, regardless of their relative expression level. In contrast, the pseudogene ψ*SLC6A8* shows high methylation of the CpG islands in all somatic tissues but complete demethylation in testis. Methylation profiles in different tissues are similar in shape but not identical. The data for the two investigated individuals suggest that methylation profiles of individual genes are tissue specific. Taken together, our findings support a model in which the bodies of the genes are predominantly methylated and thus insulated from the interaction with DNA-binding proteins. Only unmethylated promoter regions are accessible for binding and interaction. Based on this model we propose to use DNA methylation studies in conjunction with large-scale sequencing approaches as a tool for the prediction of *cis*-acting genomic regions, for the identification of cryptic and potentially active CpG islands and for the preliminary distinction of genes and pseudogenes.

INTRODUCTION

Considerable effort is underway to decipher the human genome. However, once the sequence is determined we will still know very little about how the genetic information is organized, how the expression of genes is regulated and how the tremendous complexity of the human body is developed and controlled. DNA methylation is one of the features that organizes and controls the expression of genetic information (reviewed in ref. 1). Methylation of cytosines in certain regions of genes can inhibit gene expression and artificial demethylation of genes has been shown to result in reactivation (2). Differentially methylated regions are key elements in the transcriptional regulation of genes whose expression state depends on the sex of the parent from whom they were inherited (imprinted genes) (reviewed in ref. 3). Likewise, the initiation and/or maintenance of the inactive X chromosome in female eutherians was found to depend on methylation (reviewed in ref. 4). Tumor tissue in mammalia is characterized by a genome-wide demethylation and a local hypermethylation of tumor-suppressor genes (reviewed in ref. 5). Recent work has shown that gene silencing in methylated regions is accomplished through the interaction of methylcytosine-binding proteins with other structural compounds of the chromatin (reviewed in ref. 6). This interaction makes the DNA probably inaccessible to transcription factors through histone deacetylation and chromatin structure changes.

A number of hypotheses about the origin and function of DNA methylation have been proposed. Methylation was suggested to serve as a host defense mechanism that silences most of the parasitic sequence elements (reviewed in ref. 7) or as a molecular instrument to lock developmental changes and to determine or to reflect tissue-specific gene expression (reviewed in ref. 8). Methylation was also proposed to contribute to the reduction of background expression and to result into the division of complex genomes into regulatory and non-regulatory regions (9). Whereas a number of experiments gave evidence for a correlation between the expression state and methylation strength, others showed that demethylation does not necessarily result in transcriptional activation and

[+]Present address: Institute for Human Genetics, CNRS UPR 1142, Laboratoire de Séquences Répétées et Centromères Humains, Rue de la Cardonille, 34396 Montpellier, France

[§]To whom correspondence should be addressed at: metaGen GmbH, Ihnestrasse 63, D-14195 Berlin-Dahlem, Germany. Tel: +49 30 8413 1673; Fax: +49 30 8413 1674; Email: andrex1x@aol.com

that the promoter region of inactive genes can be unmethylated (reviewed in refs 10,11). Thus, (de)methylation is apparently no general mechanism to determine tissue-specific gene expression. Since most investigations used isoschizomeric digest with restriction-sensitive endonucleases it remains, however, open whether local differences between expressing and non-expressing tissues exist that could not have been resolved by this method. It has been demonstrated that different tissues show regular differences in 5-methylcytosine content (12). So far it is not clear whether these differences are also reflected in the degree of methylation of single genes or if they are contributed exclusively by repetitive sequences. If indeed differences exist between the methylation profiles of genes in different tissues, it will be interesting to investigate how faithfully these differences are conserved among distinct individuals.

The genome contains a large number of pseudogenes which were created by duplication events or by the reintegration of reverse transcribed mRNA. Many of these pseudogenes possess promoter regions identical or very similar to those of the active ancestral genes. However, the pseudogenes are not transcribed in wild-type somatic cells. To study the methylation status of a pseudogene compared with that of the gene we have included the gene–pseudogene pair *SLC6A8*–ψ*SLC6A8* into our study.

Our knowledge about DNA methylation has profited from the investigation of model genes, but general conclusions are also hampered to a certain degree by the restriction to these genes. DNA sequences of many organisms are now readily available through the large-scale genomic sequencing approaches. Given this information, the bisulfite genomic sequencing technique (13) will deliver methylation data from any piece of DNA. To elucidate the tissue-specific methylation patterns to single cell and single base resolution by bisulfite genomic sequencing we have chosen a housekeeping gene, two tissue-specific genes and a pseudogene: the X-linked human *CDM* gene, the muscle-specific serine kinase gene *MSSK1*, the creatine transporter gene *SLC6A8* and the autosomal pseudogene ψ*SLC6A8*, respectively. These genes were studied in the regions joining the predicted CpG islands to the body of the gene. *SLC6A8*, *CDM* and *MSSK1* are located in the human telomere-near band Xq28 in a region spanning ~100 kb of GenBank accession no. U52111. The studied X-chromosomal region (26.5 kb) is duplicated on chromosome 16p11.1 (14). This duplication contains the investigated pseudogene of *SLC6A8* (ψ*SLC6A8*) and the five last exons of *CDM* (14).

SLC6A8 (solute carrier family 6, member 8) (15) encodes for a Na^+-coupled transporter that transfers creatine into the cell. So far no transcription factors have been identified controlling the expression of this gene. The prediction of a TATA box at position 936 and the existence of a CpG island from position 568 to 2507 of GenBank entry U52111 suggest regulatory functions in this region (16). Northern blot analysis indicates a strong expression of the gene in skeletal muscle, kidney, testis, colon, heart, brain, small intestine and prostate. No transcripts were found in liver and pancreas (17) where creatine is synthesized.

The gene structure of ψ*SLC6A8* in the duplicated region on chromosome 16 resembles entirely the ancestral region on Xq28. A number of insertions/deletions and base exchanges have led to an overall similarity of 94.6% but the putative coding sequence is still 97.1% identical to the X-chromosomal sequence (14). Changes were, however, sufficient to create a premature stop codon in the putative exon 4 (14). The gene is not expressed in any tissue except testis (18). This tissue specificity led to the hypothesis that it might serve to compensate for the inactivation of the X-chromosomal copy during certain stages of spermatogenesis (18).

MSSK1 (muscle-specific serine kinase 1) is exclusively expressed in skeletal and heart muscle. Alternative splicing of the first two exons has been proposed (19). Comparative analysis with the ortholog mouse gene revealed a conserved TATA box 71 bp upstream of the putative translation start at position 94 126 (19). No other transcriptional control factors have been disclosed.

The function of *CDM* has not yet been elucidated. The deduced amino acid sequence suggests the presence of membrane-associated segments and a weak similarity with the rod-like tail portion of heavy chain myosins from different species (20). The gene is ubiquitously expressed (20). A large CpG island exists from position 36 398 to 39 303 but no transcription factor binding sites have been assigned yet.

We present here a comprehensive description of the DNA methylation in the promoter-near regions of four human genes: the X-linked *CDM* gene, the muscle-specific serine kinase gene *MSSK1*, the human creatine transporter gene *SLC6A8* and the autosomal pseudogene ψ*SLC6A8*. Our data indicate that the methylation profiles of the genes are tissue specific. The level of expression and the degree of methylation appear not to be correlated. Whereas the promoter regions of all the genes are essentially unmethylated, the pseudogene ψ*SLC6A8* was found to be highly methylated in somatic tissue, but not in testis. The testis is the only tissue in which ψ*SLC6A8* is expressed. From the nature of the pseudogene we propose that it is not a testis-specific isoform but transcribed as a result of transient demethylation during spermatogenesis.

RESULTS

Methylation profile of *CDM*

The methylation patterns of the *CDM* gene was determined for a region spanning the first two exons and the boundary of the predicted CpG island. DNA was extracted from eight tissues from patient A (heart, brain, prostate, lung, skeletal muscle, pancreas, kidney and liver). The entire region consists of 1627 bp (positions 35 806–37 432 of U52111). The methylation profile was reconstructed using four overlapping PCR fragments spanning 470–523 bp. The methylation profile of *CDM* is represented in Figure 1. Hardly any of the densely packed CpGs in the CpG island region are methylated. A clear-cut increase in methylation occurs downstream of the predicted CpG island at the first CpG after the second exon (position 36 167 of U52111). Only in brain does methylation start earlier (position 36 361 of U52111). The first methylated cytosine is usually also the strongest or second strongest methylated cytosine of the region. A second preferentially methylated site exists in position 143 of the investigated region (position 35 948 of U52111). The methylation profiles of all tissues are similar in this regard; however, the degree of methylation at individual CpG sites varies among the tissues. The strongest overall methylation sites are found in heart and brain and the

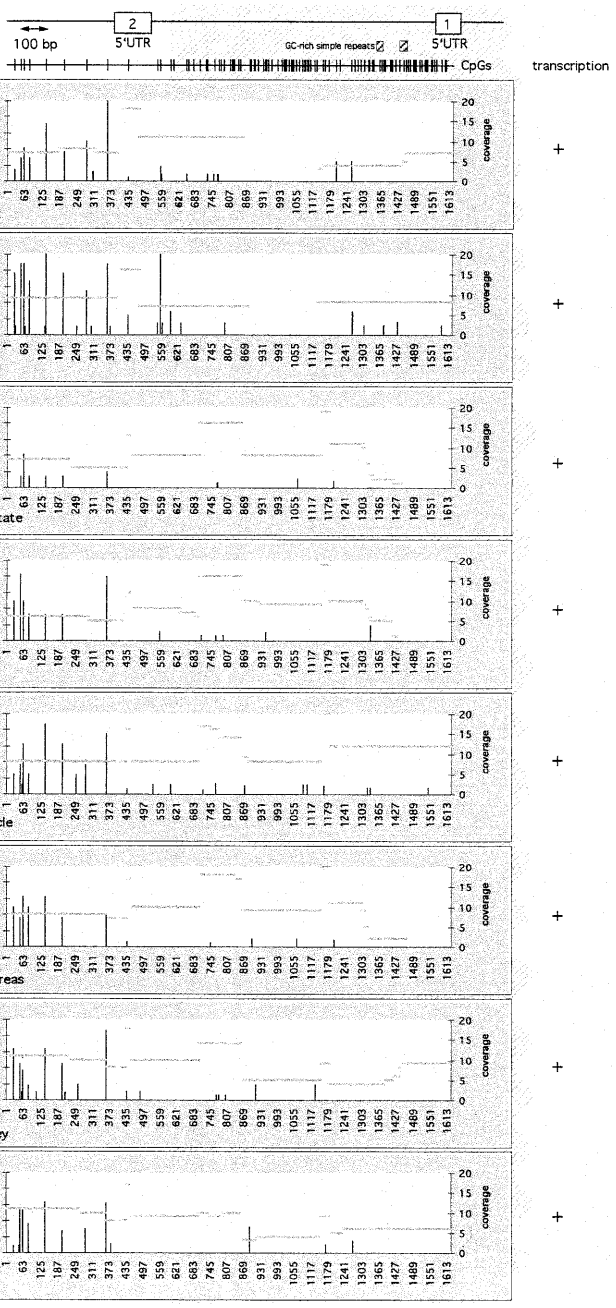

Figure 1. Methylation profiles of the *CDM* gene in eight tissues of patient A. At the top, the *CDM* gene is schematically represented and corresponds to the *x*-axes in the panels below. Boxes indicate exons with numbers in order of transcription. On the *x*-axes of the panels the genomic sequence of *CDM* is given in base pairs, with position 1 corresponding to nucleotide 35 806 of the sequence of GenBank accession no. U52111. The vertical bars represent the percentage methylation at individual cytosine positions and correspond to the left *y*-axes. The horizontal lines indicate the total number of clones that were analyzed for methylation at each position and correspond to the right *y*-axes. At the top of the first panel, the distribution of CpG pairs is represented by vertical lines. On the right of the panels showing the methylation profile of the tissue, the corresponding relative strength of transcription is indicated. The shaded box in the background depicts a predicted CpG island (71% G+C, observed/expected CpG = 0.75). Shaded boxes in the foreground of the CpG island panel indicate repeats.

weakest methylation sites are situated in the prostate. Except for prostate, the degree of methylation of individual CpG sites in the body of the gene usually exceeds 40%. The distribution of methylation patterns is mosaic-like, i.e. individual clones show a pattern of methylation signals which is similar but not identical to the patterns of other clones derived from the same PCR (detailed maps are available under http://genome.imb-jena.de/PublicationSupplements). Despite the high G+C content of the region (average 64.4%, locally >75%) and the dense occurrence of CpGs (107 in 1627 bp) only 4.5% of the cytosines in CpG pairs were found to be methylated. In addition, 0.6% of the cytosines in a sequence context other than CpG showed a methylation signal. *CDM* is expressed in all investigated tissues (20). Given the use of autopsy material in our study, it would have been impossible to obtain reliable expression measurements for the investigated genes. Instead, expression levels of the genes were taken from the literature.

Methylation profile of *MSSK1*

For the determination of the methylation patterns in *MSSK1* the same DNA samples were used for *MSSK1* as for *CDM*. The investigated region included the first four exons of the gene and a putative TATA box. It had a length of 1153 bp (positions 93 995–95 147 of U52111) and was covered by three overlapping PCR fragments. The entire region contains 58 CpG pairs. Lowest methylation was found in the predicted CpG island (Fig. 2). In heart, brain and prostate, the small CpG island of the gene is free of methylation. In lung, muscle, pancreas, kidney and liver the island is partly methylated. In all these tissues except muscle, two distinct fractions of cells exist: a small fraction with dense methylation and a larger fraction with weak methylation. Only a small number of clones contribute to the observed methylation signal: 1 in 13 in the lung, 3 in 20 in the pancreas, 2 (identical) in 13 in the kidney and 7 in 28 in the liver. The distribution downstream of position 400 (position 94 394 of U52111), gives place to irregular mosaic-like methylation patterns as noted for *CDM* (detailed maps available under http://genome.imb-jena.de/PublicationSupplements). The region of low methylation extends from the CpG island into the third exon. In contrast to *CDM*, the methylation level increase gradually. The shape of the methylation profile across the region is similar in all tissues. Yet, in the expressing tissues of muscle and heart, the methylation reaches its maximum more towards the 3′ end compared with most non-expressing tissues. This tendency is confirmed when the difference between the average methylation in non-expressing and expressing tissues is calculated and plotted (data not shown). Between position 250 and 700 (positions 94 244–94 694 of U52111), DNA from heart and muscle but also from the brain is on average less methylated than DNA from the other tissues. Further downstream the methylation is equilibrated. The non-expressing brain also shows this late onset of DNA methylation; therefore, this characteristic is not exclusive to expressing tissues. In 347 investigated PCR clones, 30.4% of the CpGs and 0.92% of the cytosines in non-CpG pairs showed a methylation signal. *MSSK1* is exclusively expressed in heart and muscle. Our data indicate no correlation between the level of methylation and expression in individual tissues.

Methylation profile of the X-linked *SLC6A8*

As for the other genes, a set of tissue-specific DNA from patient A was used. Additionally, DNA was extracted from the

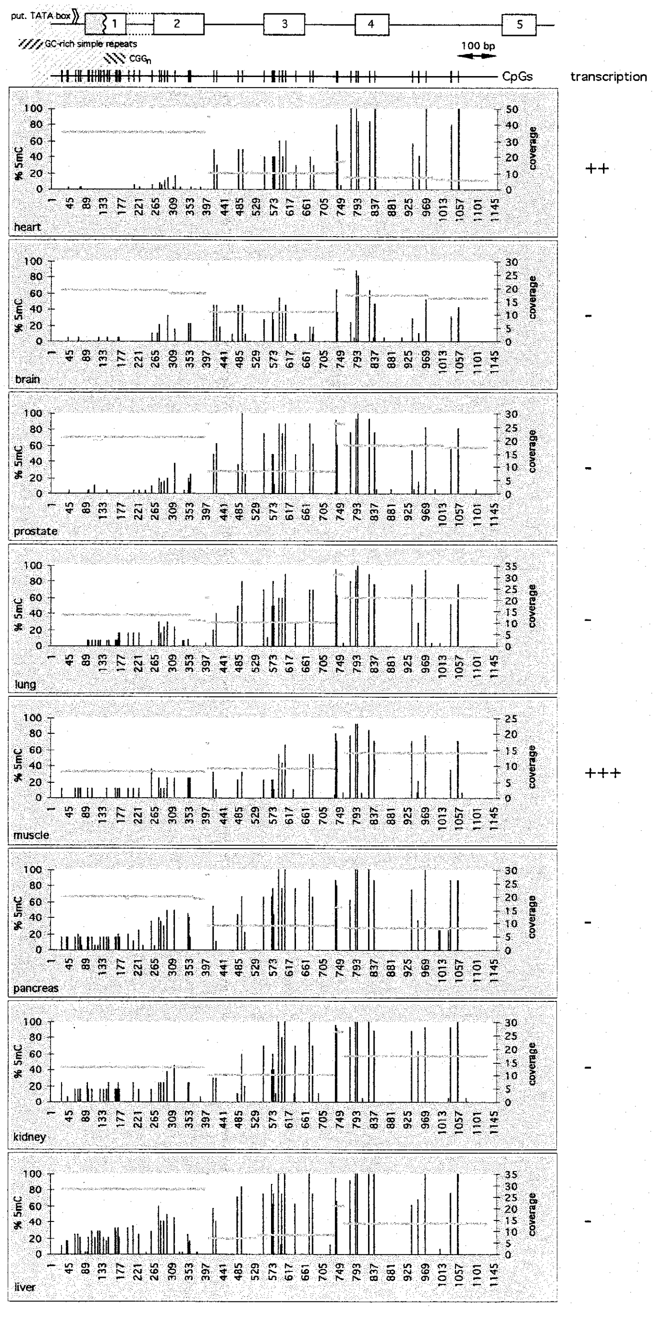

Figure 2. Methylation profiles of the muscle-specific serine kinase gene (*MSSK1*) in eight tissues of patient A. *MSSK1* is schematically represented. Boxes indicate exons in order of transcription. Alternative splicing of the first two exons is represented by a dotted line. The transcription start of the first exon is putative and the unconfirmed part is shaded in grey. On the *x*-axes of the panels the genomic sequence of *MSSK1* is given in base pairs, with position 1 corresponding to nucleotide 93 995 of the sequence of GenBank accession no. U42111. The structure of the graph is identical to that in Figure 1. The shaded box in the background indicates a predicted CpG island (73% G+C, observed/expected CpG = 0.75). Alternatively shaded boxes in the foreground of the CpG island panel stand for GC-rich simple repeat and (CGG)$_n$ repeats.

testis of patient J. Methylation was studied from 473 bp upstream of the 3′ end of the first exon to the 3′ end of the predicted CpG island (Fig. 3). The investigated region had a length of 1195 bp (positions 1404–2598 of U52111) and was covered by three overlapping PCR fragments and contains 137 CpG pairs. Similar to the genes described above, hardly any methylation was found in the CpG island of the X-chromosomal copy of *SLC6A8*. However, in this gene the methylation increases already 100–300 bp upstream of the 3′ end of the predicted CpG island. Methylation is extremely low in prostate and virtually non-existent in testis (Fig. 4). However, comparison of the methylation density plot for prostate with the plots for other tissues (available under http://genome.imb-jena.de/PublicationSupplements) indicated that the methylation increased also in prostate at around position 700 (position 2103 in U52111). It is not likely that the methylation density will drop suddenly. Thus, the extremely low methylation signal in the 3′ part is certainly due to the low number of investigated PCR products here. The true methylation profile is probably similar to that of the brain or the heart. In general, *MSSK1* methylation grows gradually and reaches, in *SLC6A8*, its maximum around position 1050 of the investigated region (position 2453 of U52111). This position is consistent with the end of the predicted CpG island. No pronounced correlation between the degree and the onset of methylation and expression level can be found. As noted with *MSSK1*, a disposition for undermethylation in expressing tissues is noted when the average methylation of non-expressing and expressing tissues are compared (data not shown). Between position 400 and 1100 (positions 1803 and 2503 of U52111) methylation is on average higher in non-expressing tissues. However, this general tendency cannot be used to characterize individual tissues: for example muscle and liver show almost identical methylation profiles, whereas the gene is highly expressed in muscle and completely repressed in liver. In both brain and heart at the 3′ end of the investigated region, methylation signals are exclusively found in two of the nine PCR clones. In contrast, all other tissues show a scattered and mosaic-like distribution of methylation.

To study whether methylation profiles are conserved among individuals, identical sets of tissues were investigated in patients A and B. An arbitrary chosen set of data is shown in Figure 5. The analysis was restricted to the informative, i.e. differentially methylated, region from 780 to 1195 bp of the investigated region (positions 2183–2598 of U52111). The methylation profiles of the same tissues from the two different people show very small differences compared with the differences between the tissues of the same person. These findings suggests that methylation profiles of genes are tissue-specific and that the tissue specificity is constant between individuals.

In 359 investigated PCR clones, 7.2% of the CpGs and 0.54% of the cytosines in other pairs showed a methylation signal.

Methylation profile of the autosomal ψ*SLC6A8*

The secondary copy of *SLC6A8* on chromosome 16 (ψ*SLC6A8*) was found to be nearly completely methylated in all investigated specimens, i.e. eight different tissue samples from individual A (Fig. 6) and DNA from the white cerebral matter of seven further individuals (C–I) (Fig. 6; other data not shown here, for details see http://genome.imb-jena.de/PublicationSupplements). The methylation density is very high and lies between 30 and 90% in a 100 bp window. One remark-

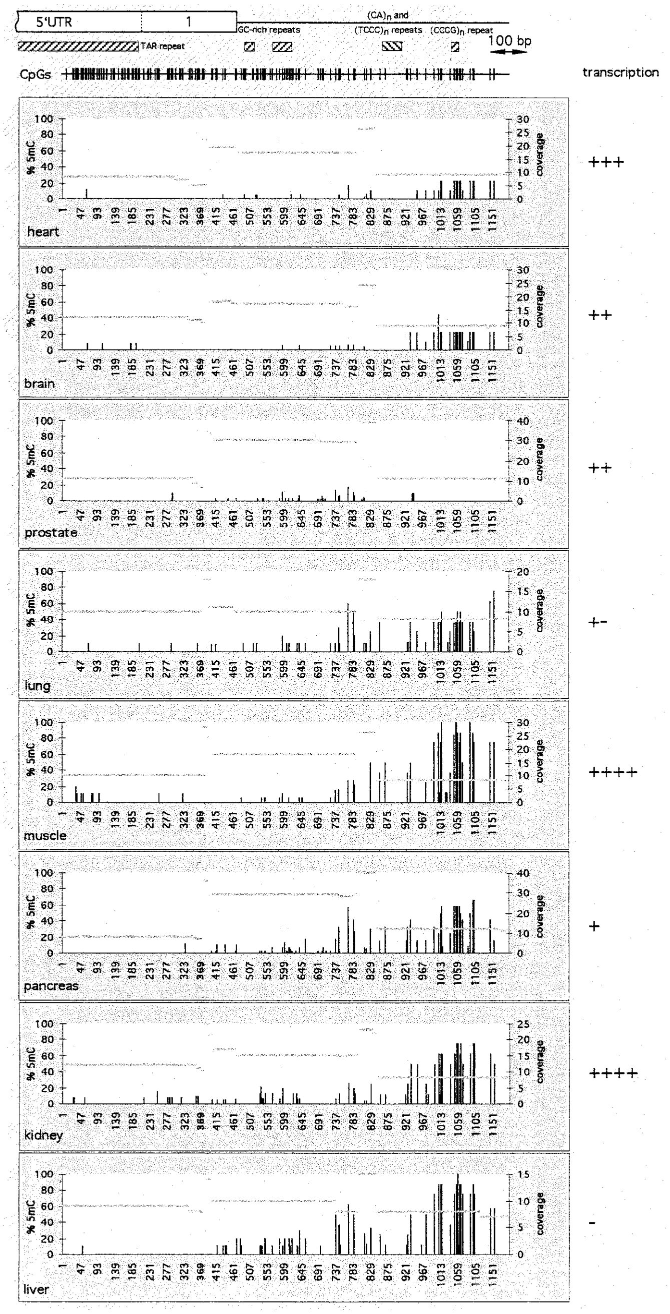

Figure 3. Methylation profiles of the creatine transporter gene (*SLC6A8*) on the X chromosome in eight tissues of patient A. *SLC6A8* is represented as in Figures 1 and 2. The box indicates the first exon. On the *x*-axes of the panels, the genomic sequence of *SLC6A8* is given in base pairs, with the position 1 corresponding to nucleotide 1404 of the sequence of GenBank accession no. U52111. The structure of the figure follows that of Figures 1 and 2. The shaded box reflects a predicted CpG island (78% G+C, observed/expected CpG = 0.88). The differently shaded box between the CpG island panel and the gene scheme indicates different repeats.

able exception is the complete absence of methylation in the testis (Fig. 4).

Despite the high degree of overall methylation and very similar methylation profiles in all tissues and patients, detailed examination of the 15 data sets (a total of 333 subcloned PCR products) revealed that only 7 of the 114 CpGs are ≥90% methylated in all tissue samples. These conserved sites are located at positions 189, 193, 202, 229, 235, 253 and 286 of the studied fragment. Their occurrence coincides with the 3′ end of a predicted TAR repeat within the first exon of ψ*SLC6A8* and a local minimum of the G+C content (Fig. 6). Whether this site is of functional importance or simply more accessible to the methyltransferase or guiding factors remains to be elucidated. The computer program TSSW identified a promoter in this region, but its significance is low. The differences between methylation profiles of different tissues of an individual person and of the brain of different individuals are comparable. The autosomal copy has lost 16.7% of methylatable CpGs (Fig. 4). Their number dropped from 137 in the X-linked copy to 114 in the fragment on chromosome 16. Fifteen of the 23 CpGs were lost as the result of a C→T or G→A exchange. The number of other base pairs remained comparably constant. As for the other genes, methylation in non-CpG pairs is low (1.2%) compared with CpGs (63.6%). Mosaic-like methylation is also observed in this case. ψ*SLC6A8* is the only investigated gene that shows a clear correlation between methylation and repression (in all studied somatic tissues) and demethylation and expression (in testis).

Sequence context of methylated cytosines

Information theory has been used successfully to characterize sequence patterns (21) and was proposed as a general improvement to consensus sequences (21). We adapted the algorithm for the search for conserved methylation sites. A weak periodicity of information content was found in all tissues for the X-linked genes *MSSK1*, *CDM* and *SLC6A8*. However, this regularity was lost when the average for all genes was calculated. For the methylation patterns of the autosomal ψ*SLC6A8*, no consensus other than the regular occurrence of guanosine after a methylated cytosine was found (data not shown). Therefore, in the investigated region no motives other than CpG exist that are preferentially methylated. Methylation in other CpN pairs is rare. The frequencies for all 5mCpN are listed in Table 1. Taken together, methylation in non-CpG pairs occurs in 0.81 ± 0.27% of the Cp(ATC). This value lies within the range of the systematic errors of PCR and sequencing reaction (see Materials and Methods). Therefore, our data can neither exclude nor confirm the existence of methylation in non-CpG pairs.

To address the question whether repetitive sequences are primary targets for methylation, the four investigated genomic regions were analyzed *in silico* for the existence of repetitive sequences. In all cases, only simple repeats and repeats of low complexity were identified. For *CDM*, two short GC-rich repeats were found at positions 1337–1357 and 1437–1464 of the investigated region. For the *MSSK1* fragment, a simple $(CGG)_n$ repeat was identified at around position 130–180. For both genes, these short repeats do not show a higher level of methylation than the non-repetitive sequences in their vicinity (Figs 1 and 2). In the case of *SLC6A8* and ψ*SLC6A8*, a number of repeats were found including C-rich and GC-rich repeats of low complexity and simple $(CA)_n$, $(TCCCC)_n$ and $(CCCG)_n$ repeats. However, with the exception of the $(CCCG)_n$ repeat at position 1049–1073 of the investigated fragment, none of the repeats match with a local increase of the methylation level. The $(CCCG)_n$ repeat is located in a region where a majority of CpG pairs are methylated. However, this region also matches

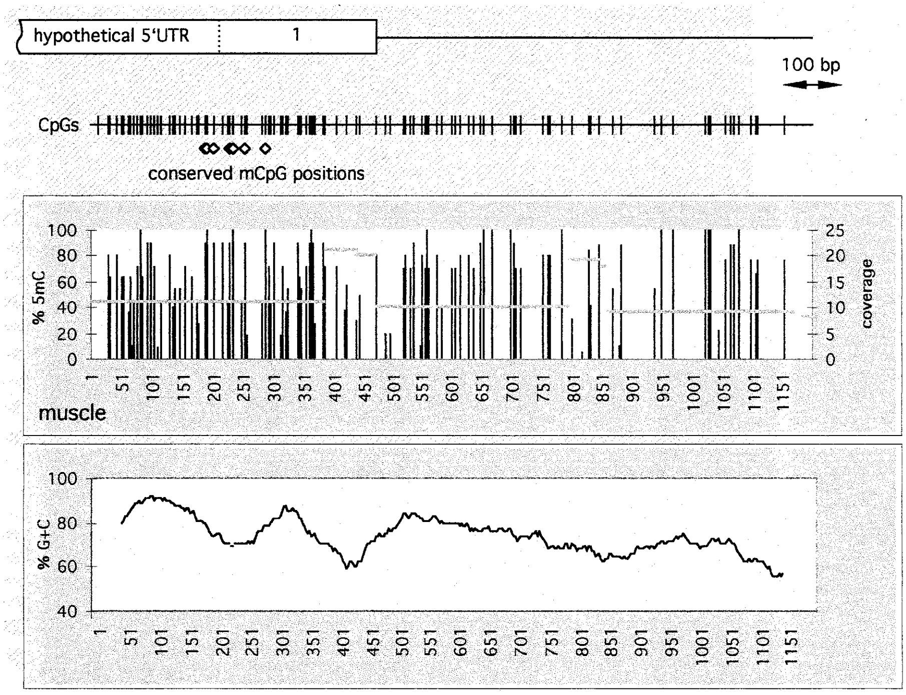

Figure 4. DNA methylation profiles for the creatine transporter pseudogene (Ψ*SLC6A8*) on chromosome 16 and the creatine transporter gene (*SLC6A8*) on the X chromosome in the testis of patient J. An *SLC6A8* consensus gene is shown on the top and corresponds to the *x*-axes in the panels below. The box indicates the first exon. On the *x*-axes of the panels the genomic sequence of *SLC6A8* is given in base pairs, with the position 1 corresponding to nucleotide 4904 of the sequence of GenBank accession no. U41302 (chromosome 16) and to nucleotide 1404 of the sequence of GenBank accession no. U52111 sequence (X chromosome). The vertical bars represent the percentage methylation at individual cytosine positions and correspond to the left *y*-axes. At the top of each panel, the distribution of CpG pairs is represented by vertical lines. Both gene and pseudogene are expressed in testis. In both gene and pseudogene, almost no methylation is found.

the end of the predicted CpG island (Fig. 3). Thus, further data are needed to decide if the co-location of high 5mC levels and the $(CCCG)_n$ repeat is a coincident in this case or if it occurs regularly.

All methylation data were deposited at the DNA-methylation database (http://www.methdb.de) under the sequence IDs 21–24 (*CDM*), 25–27 (*MSSK1*), 28–31 (*SLC6A8*) and 32–34 (ψ*SLC6A8*).

DISCUSSION

Methylation of the X-linked genes *CDM*, *MSSK1* and *SLC6A8* is not expression specific

Tissue-specific, i.e. cell type-specific, methylation has been discussed for a long time. The hypothesis is tempting that gene expression in different cell types is either established or locked by differential methylation during the early embryonic development (22,23). Correlations between changes in expression and methylation were found in several cases (24). However, serious doubts have been expressed about whether methylation serves as a general mechanism that determines tissue-specific gene expression (25). It is experimentally well established that in mammals methylation in the promoter region of a gene inhibits its transcription (26), yet undermethylation in the promoter region *in vivo* does not result necessarily in gene expression (see catalog in refs 10,11). Accordingly, methylation was proposed to serve to organize the complex genome of vertebrates (27) and to neutralize potentially dangerous DNA elements (reviewed in refs 7,28). In fact, most (de)methylation events *in vivo* are associated with transcriptional activation or repression being part of developmental changes. Well-described examples—besides early embryogenesis and tumor development (reviewed in refs 29,30, respectively)—include the demethylation of the vitellogenin gene during the transition of immature chickens to egg-laying hens (reviewed in ref. 31), methylation changes in the human globin gene switching during embryogenesis (reviewed in ref. 32, and citations therein) and (de)methylation processes in B- and T-cell maturation (33,34). These global reorganization processes might require remodulation of the distribution of methylated cytosines. Our data are in good agreement with the hypothesis that methylation reduces the complexity of the genome and that undermethylation serves as a tag for regions with regulatory function. All three investigated genes *SLC6A8*, *MSSK1* and *CDM* show undermethylation in the putative promoter-near region regardless of their state of activity. A similar behavior was recently described using bisulfite genomic sequencing for the tissue-specifically expressed α-actin gene (35). In contrast, two further studies utilized bisulfite sequencing to analyze methylation in the tissue-specific genes tyrosine hydroxylase (36) and galectin-1 (37) genes and found correlation between tissue-specific expression and methylation. However, only three different somatic tissues have been examined in each study. The reported differences in methylation might therefore reflect a coincidence rather than a causal relationship. Figure 5 illustrates that small arbitrarily chosen sample sizes can simulate correlation.

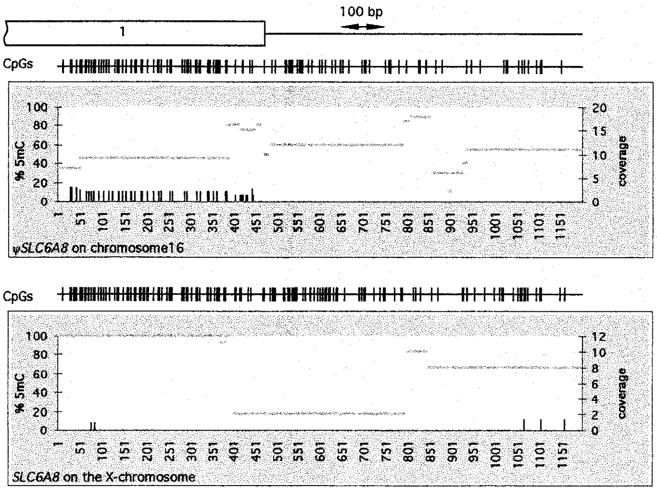

Figure 5. Methylation profiles for the creatine transporter gene (*SLC6A8*) on the X chromosome in four tissues of patients A and B. Profiles for patient A are given in the left column of panels and those for patient B are given in the right column. Each row of panels represents one tissue. At the top, the distribution of CpG pairs is represented by vertical lines and corresponds to the *x*-axes in the panels below. On the *x*-axes of the panels the genomic sequence of *SLC6A8* is given in base pairs, with the first position (marked 805) corresponding to position 805 in Figure 3 and nucleotide 2208 of the sequence of GenBank accession no. U52111. The vertical bars represent the percentage methylation at individual cytosine positions and correspond to the left *y*-axes. The horizontal lines indicate the total number of clones that were analyzed for methylation at each position and correspond to the right *y*-axes. On the left, the corresponding relative strength of expression in the individual tissues is indicated. Except for the brain, the methylation profiles in identical tissues of different individuals resemble each other.

The weak tendency to undermethylation in expressing tissues for both *SLC6A8* and *MSSK1* might be explained—as proposed earlier (27)—by the permanent presence of transcription factors that reduce the accessibility of DNA to the methyltransferase. However, it has to be emphasized that this result does not allow conclusions to be drawn from the methylation state to the expression state of genes in individual tissue. Whereas undermethylation cannot serve as a signal to predict gene expression, it can be of use to confirm putative promoter regions or to predict possible transcription start sites. In particular where *in silico* analysis has left open the question whether further exons exist upstream of a predicted first exon, it will be of benefit to study the methylation density at this site. Promoter prediction programs still require preselection of short regions of genomic sequence for a meaningful analysis (38). We propose to use the local demethylation as a sign to identify these regions. Methylation analysis could also allow to distinguish between active and silent CpG islands.

In all investigated tissues mosaic-like methylation patterns were found. The methylation density distributions along DNA molecules of singular cells of a tissue were similar; however, the exact methylation patterns were different. A preliminary parsimony analysis gave no indication for the clonal inheritance of different patterns (data not shown).

Different relative methylation levels were found for each gene in all investigated tissues. For instance, in the brain *CDM* is strongly methylated, but *MSSK1* shows moderate and *SLC6A8* weak methylation. Thus, the general degree of methylation in a tissue is not reflected directly in the degree of methylation of individual genes.

Methylation of the autosomal ψ*SLC6A8*

In contrast to the X-linked genes, ψ*SLC6A8* was found to be highly methylated in all tissues analyzed except testis. The testis is also the only tissue that shows transcription from ψ*SLC6A8* (18). This observation has led to the hypothesis that it might serve as a spare copy to rescue the transiently inactivated X-linked gene during spermatogenesis (18). In mammals, early condensation of the X chromosome during the prophase of the first meiotic division in spermatogenesis has been known for a long time (39). The expression of the X-linked gene *PGK-1* declines precisely from the pachytene stage (40) indicating the onset of the inactivation at this stage. In parallel—after *de novo* methylation in the leptotene or zygotene stage of meiosis I (reviewed in refs 3,41)—global demethylation takes place (12,42–45) including CpG islands and Alu repeats (46,47). In contrast, at least some non-CpG island genes remain methylated (48). A point that has been speculated about for a long time is the ectopic expression of many autosomal genes in the testis. Genes expressed during spermatogenesis can be classified into four groups: (i) genes that are necessary for sperm development, such as protamines, small basic proteins, histone (Hlt) and enzymes necessary for penetration into the egg; (ii) intronless copies of genes such as *MYCL2* (49), *PGK-2* (50), *PDHA-2* (51) and, for example, pseudogenes of the human GTP-binding protein α subunit (Unigene Hs.138204), human creatine transporter gene *SLC6A8* and human DNA damage repair and recombination protein, *RAD52* (Unigene Hs.73046); (iii) proto-oncogenes (52); and (iv) transgenes (see below). It has been speculated that the autosomal copies of phosphoglycerate kinase *PGK-2* (human 19p13.3) and pyruvate dehydrogenase E1 α subunit *PDHA-2* (human 4q22–23) serve to rescue the function of their X-chromosomal ancestors *PGK-1* (human Xq13.3) and *PDHA-1* (human Xp22.1), respectively, during the transient X-inactivation in pachytene sperm. Based on this hypothesis a similar mechanism was proposed for the *SLC6A8* gene and pseudogene (18).

However, certain findings throw doubt on the hypothesis. (i) It is supposed that *PDHA* translocated from an autosome (where it is still situated in marsupials) (53) to the X chromosome. This is a disadvantageous position if the hypothesis is true, that expression of this gene is also required when the X chromosome is inactivated during spermatogenesis. Supposing that the most likely hypothesis is the one that requires the fewest assumptions, it would be more plausible to assume that the X-chromosomal site provides no disadvantage. (ii) If both the autosomal and the X-chromosomal copies serve the same purpose, then they should be quite similar. However, it has been shown (51) that the evolutionary distance between mouse and man is larger for the both autosomal genes [number of nucleotide substitutions per 100 non-degenerated sites = 7.2 (*PGK-2*) and 13.9 (*PDHA-2*)] than for their X-chromosomal counterparts (0.9–1.0 for both *PGK-1* and *PDHA-1*). (iii) *SLC6A8* and ψ*SLC6A8* show an expression pattern similar to *PDHA-1/2* and *PGK-1/2*. As outlined above, the autosomal copy is active only in the testis whereas the X-chromosomal gene is expressed in many somatic tissues except liver and

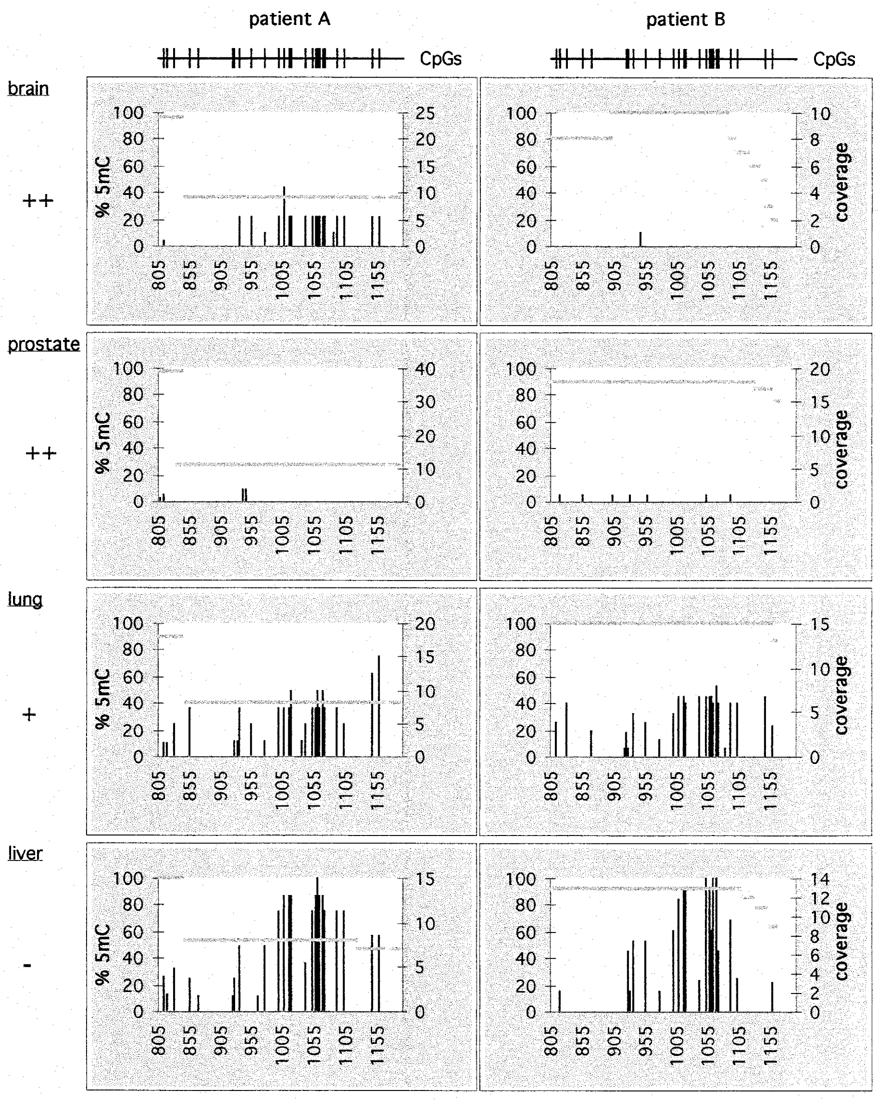

Figure 6. Methylation profile of muscle tissue of patient B for the creatine transporter pseudogene (ϕ*SLC6A8*) on chromosome 16. The tissue is arbitrarily chosen as an example. Methylation profiles are similar in other tissues and other patients, except testis. At the top, ϕ*SLC6A8* is schematically represented and corresponds to the *x*-axes in the panels below. The box indicates the first exon. On the *x*-axes of the panels the genomic sequence of ϕ*SLC6A8* is given in base pairs, with position 1 corresponding to nucleotide 4904 of the sequence of GenBank accession no. U41302. The structure of the figure is identical to that of Figures 1–3. Above the first panel, the distribution of CpG pairs is represented by vertical lines. The rhombi below mark the position of seven CpG sites found to be ≥90% methylated in all tissues (for details see text). The shaded box in the background reflects a predicted CpG island (78% G+C, observed/expected CpG = 0.82). The bottom panel shows the distribution of the G+C density calculated in a 100 bp window shifted in 3 bp steps.

pancreas. However, ψ*SLC6A8* possesses a premature stop codon, leading either to a short testis-specific isoform or a functionless protein stump. The occurrence of the testis-specific isoform would be restricted to primates (14), which seems unlikely.

In at least 13 cases it has been demonstrated that expression of transgenes was restricted to the testis. In all five cases in which methylation was investigated, expression was correlated with hypomethylation (54–58).

In somatic cells, *PGK-2* (46), *PDHA-2* (59) and (as presented in this paper) the autosomal ψ*SLC6A8* are methylated and silent, whereas they are active and demethylated in the male reproductive cells. Taken together, we suggest that the testis-specific transcription of pseudogenes is a side-product of the transient demethylation during spermatogenesis. The global methylation changes during germ cell generation are probably too important to the cell to be abandoned. They might occur to allow recombination which is, at least in somatic cells, suppressed by methylation (60,61) or are necessary to remove epigenetic defects as proposed by Holliday (62) or both. Against gene expression that leads to deleterious effects in germ cells the organism could be protected by apoptosis of these cells. During sperm cell maturation ~50% of the sperms are subject to this cell death (63). Further investigations are necessary to clarify the role of demethylation of pseudogenes during spermatogenesis.

Table 1. Occurence of 5mC in different CpN pairs in % [5mCpN/(5mCpN + CpN) × 100]

	% occurrence of 5mC			
	SLC6A8 (359 clones)	*CDM* (259 clones)	*MSSK1* (347 clones)	ψ*SLC6A8* (333 clones)
CpG	7.19	4.49	30.37	63.56
CpA	0.08	0.18	0.42	0.36
CpT	0.19	0.18	0.21	0.29
CpC	0.27	0.28	0.29	0.49

The number of investigated clones for each gene is given in parentheses.

Table 2. Positions of the first exons of the investigated genes

Gene	First exon (bp)	GenBank accession no.
SLC6A8	1042–1876	U52111
CDM	Complement (37 373–37 464)	U52111
MSSK1	94 082–94 184	U52111
ySLC6A8	6909–7040	U41302

Both copies of *SLC6A8* share an identity of 97%. This makes it unlikely, but not impossible, that the signal for the methylation of the autosomal copy is contained in the sequence itself. The deletion of two potential Sp1-binding sites in ψ*SLC6A8* could be of interest in this regard.

Dense methylation has been reported for many repetitive sequences (64) including the centromeric and pericentromeric regions (65). Independently, methylation was reported to spread into adjacent regions (66). Thus, the methylation of ψ*SLC6A8* could be the consequence of its centromere-near location. A number of pseudogenes arisen from duplications or translocations are located in pericentromeric regions (67) and our data support the idea that they are silenced there by methylation.

Alternatively, methylation of ψ*SLC6A8* could be the result of an active decision-making process. The silencing of invading homologs together with their endogenous counterparts by methylation has been described in many cases for plants and lower eukaryotes (homology-dependent silencing) (68) and a similar phenomenon was shown to exist in mammals (57). A *CAT* reporter gene ligated to the *PGK-2* core promoter was shown to undergo a similar tissue-, stage- and cell type-specific demethylation in the 5′ portion of the *CAT* coding sequence to that of *PGK-2* itself, suggesting that the signal for methylation is encoded in the promoter region (69). The transcription of the cardiotrophin-1 gene that is located in the same band 16p11.1–2 as *SLC6A8*, indicates that the region itself is not completely inactive (70).

Are methylation profiles tissue specific?

It is interesting to note that, whereas the differences of methylation profiles between tissues of one individual can be remarkably high, profiles are very similar in identical tissue types from different individuals. This was found for *SLC6A8* in a set of tissues from two people (Fig. 6) and for ψ*SLC6A8* in brain tissue from eight people (data not shown; details available at http://genome.imb-jena.de/PublicationSupplements). This argues for a regulated process, and in fact regular differences between the global 5-methylcytosine content of different tissues have been described previously (12). Although our data are not comprehensive in this regard, they suggest that methylation profiles and tissue types are closely related. Given the number of individuals investigated, a statistical analysis of the data will not deliver unambiguous results. Further work is needed to provide a solid data basis and to prove this hypothesis.

The potential use of methylation studies in large-scale sequencing approaches

The data presented here are in good agreement with the idea that methylation is a method of reducing the complexity of large genomes: *cis*-acting elements are accessible to *trans*-acting factors whereas other DNA regions are shielded from this interaction by methylation. According to this model, methylation studies can be of great benefit in large-scale sequencing projects. They can help in identifying pseudogenes, in distinguishing active and inactive CpG islands, in guiding gene prediction and promoter identification and in recognizing the real 5′ ends of genes. Although our findings are entirely sufficient to propose this approach only the analysis of a large number of sequences from different origins will prove it. We have therefore established a public database for DNA methylation (http://www.methdb.de) and attempt to collect all available data in this resource. Given the current interest in this research field, the information necessary to prove or to disprove our hypothesis will soon be available.

MATERIALS AND METHODS

Throughout this paper, positions are indicated with respect to GenBank accession no. U52111 from 22 March 1996 and accession no. U41302 from 24 November 1995. Since the entries of the GenBank database are continuously updated, the boundaries of the first exons of the genes studied are additionally listed in Table 2. In case of future modifications of the GenBank sequence data, this will guide the adjustment of the positions that are referred to in this study.

Tissue material and DNA preparation

Tissue samples were obtained from autopsy material of male adults. All 10 patients died between the ages of 60 and 88 years. None of the individuals suffered from a malignant tumor. Cause of death, age and severe disorders confirmed by the autopsy are listed in Table 3. For the purpose of the experiments described in this paper, the individuals were arbitrarily given the letters A–J. Autopsies were carried out between 24 and 72 h after death. Prior to autopsy, the bodies were kept between 4 and 10°C. Homogeneous tissue material was carefully dissected, stored immediately at –20°C and was subsequently transferred to –80°C for further use. DNA from eight different tissues [white cerebral matter (brain), heart, kidney, liver, prostate, pancreas, lung, skeletal muscle] was taken from

Table 3. List of patients from whom tissue material was derived for this study

Patient	Cause of death	Age (years)	Severe disorders
A	Myocardial infarction	72	Coronary heart disease, arteriosclerosis
B	Myocardial infarction	79	Coronary heart disease, arteriosclerosis, diabetes
C	Pneumonia	67	Arteriosclerosis, diabetes
D	Myocardial infarction	72	Coronary heart disease, arteriosclerosis
E	Ventricular fibrillation	65	Bronchitis
F	Septical shock	71	Arteriosclerosis, diabetes
G	Myocardial infarction	67	Coronary heart disease, arteriosclerosis
H	Pulmonary embolism	60	Arteriosclerosis, hypertrophy of the prostate
I	Myocardial infarction	88	Diabetes
J	Myocardial infarction	82	Coronary heart disease, arteriosclerosis

patients A and B. Testis DNA was obtained from patient J and brain tissue was extracted from patients C–J.

A sample of phenotypically homogeneous tissue material (200–500 mg) was quickly washed in sterile water and incubated in 1 ml of buffer (20 mM Tris pH 8.0, 1 mM EDTA, 100 mM NaCl, 0.5% SDS) and 0.3 mg of protease K (Boehringer, Mannheim, Germany) under shaking for 16 h at 55°C. DNA was separated by two rounds of phenol–chloroform extraction and traces of phenol were removed with chloroform. Genomic DNA was precipitated with sodium acetate/isopropanol, washed with 70% ethanol, dissolved in 10 mM Tris pH 8.0 and stored at –20°C (71).

Bisulfite treatment, PCR amplification and sequencing

Bisulfite genomic sequencing was carried out as described by Frommer *et al.* (13). Briefly, the DNA was denatured in 111 µl of 0.3 M NaOH (in the presence of 10 µg of yeast mRNA) at 42°C for 20 min. Freshly prepared sodium bisulfite solution [1.2 ml, 541 g/l (ACS grade reagent; Sigma, St Louis, MO)] in 10 mM hydroquinone pH 5.0) were added directly to the denatured DNA. The reaction was overlaid with 200 µl of mineral oil (IR grade; Sigma) and incubated at 55°C for 4 h in the dark. The DNA was desalted using the Wizard DNA purification kit (Promega, Madison, WI) and eluted in 105 µl of 1 mM Tris–HCl pH 8.5. One hundred microliters of this DNA solution was desulfonated in 0.3 M NaOH at 37°C for 30 min. For neutralization, ammonium acetate was added to a final concentration of 3 M and the DNA was precipitated with ethanol using 1 µg of yeast mRNA as carrier. The DNA was dissolved in 100 µl of 10 mM Tris–HCl pH 8.5 and stored at –20°C until further use. The bisulfite-treated DNA could be used for at least 6 months without loss of PCR yield.

A sample (2.5 µl) of the bisulfite-treated DNA was used as template for the PCR amplification of overlapping fragments. The amplification was carried out in two consecutive PCR reactions with nested primers sets. The primer design followed the guidelines by Clark and Frommer (72). PCR primers, annealing temperatures and the length of the PCR products are listed in Table 4. For each PCR at least 500 pg but usually 50 ng of genomic template DNA was used, not considering the loss during the bisulfite treatment. A total of 50 pmol of each primer, 2.5 U *Taq* polymerase (Qiagen, Hilden, Germany), the standard buffer supplied with the enzyme and 250 µM dNTPs were incubated in a total volume of 50 µl. After the first PCR, 2.5 µl of the first PCR mixture were used as template for the second PCR. Incubation times and temperatures were 94°C for 2 min followed by 5 cycles (94°C for 1 min, annealing temperature for 2 min, 72°C for 3 min), 25 cycles (94°C for 0.5 min, annealing temperature for 2 min, 72°C for 1.5 min) and 72°C for 10 min. Annealing temperatures are listed in Table 4. The PCR products were separated on a 1% TAE agarose gel and purified using the QIAquick gel extraction kit (Qiagen). To ensure maximum cloning efficiency, an aliquot of the PCR product was incubated with 2.5 U *Taq* polymerase (Qiagen) and 200 µM dATP in PCR buffer for 10 min at 72°C and subsequently ligated into the pGEM-T-easy vector (Promega). Sequencing of the subcloned PCR products was performed with the ABI Prism BigDye or Dye Terminator cycle sequencing kits (Perkin Elmer, Foster City, CA). The sensitivity of this method has been found previously to be 99.50 ± 0.45% for unmethylated cytosines and 100.00 ± 0.00% for 5-methylcytosines (data not shown). The total error rate of the PCR and the sequencing reaction was 0.7% (estimated from the occurrence of non-cytosine base exchanges in a subset of 1298 clones).

Analysis of DNA methylation

The sequences of the PCR products were aligned to the genomic sequence using the GAP4 computer package (73), manually edited and further analyzed. Cytosines in the genomic sequence that were converted to uracils and subsequently PCR amplified as thymines correspond to unmethylated cytosines, whereas cytosines in the PCR products indicate the presence of 5-methylcytosines. PCR products from *SLC6A8* of X-chromosomal and autosomal origin were identified on the basis of diagnostic non-cytosine base exchanges. The alignments were used (i) to calculate the average methylation level for each cytosine position; (ii) to visualize the methylation state of individual CpGs (red circles represent 5mCpG and blue circles unmethylated CpG); and (iii) to determine the methylation density for each clone. The latter was calculated as the ratio of methylated cytosines in CpG pairs to total number of CpG pairs in a window of 100 bp shifted in 1 bp steps over the sequence.

The G+C density was determined using the program 'window' of the GCG version 8 package (Genetics Computer Group, Madison, WI). The chosen window size was 100 bp and the step size was 3 bp.

In order to identify conserved sequence patterns around methylated cytosines, the frequencies of each nucleotide for positions up to 300 bp upstream and downstream of this site were determined. The information content R and the relative entropy H' of each position (21) was calculated and plotted using the MethTools software package (74).

Table 4. Primers and annealing temperatures for the PCR on bisulfite-converted DNA

PCR product	First PCR round			Second PCR round			Length (bp)
	Annealing at 5×/25×[a] (°C)	Primer 1[b]	Primer 2[b]	Annealing at 5×/ 25×[a] (°C)	Primer 3[b]	Primer 4[b]	
SLC6A8-1	50/50	aTaTaTatgagattTttTaggTtTaTtt	AcacaatAtAccCttAtAtAcccAac	50/50	TtaagtgTttGgtggaTtgTttTtgaTtg	TtcttAtaAcacaAAtaAAAAaaAc	465
SLC6A8-2	54/54	tTaagtgTttGgtggaTtgTttTtgaTtg	Aaacctttactctaaacctctatttcc	50/50	TtagatggaTttTatTatgt	AcacaatAtAccCttAtAtAcccAac	462
SLC6A8-3	45/50	TTagatggaTttTatTatgt	TcacaAtcccctActAAtAAAatAAc	50/50	GtTtTTgGgaggtaaggagTTTtgg	AAAcctttActctAAacctctAtttcc	416
SLC6A8-3×	As for SLC6A8-3			65	GaggtaaggagTTTtggTtgTTTTTa	CtAAacctctAtttcccacccatcacc	397
CDM-1	50/50	gaTaagagtTaTTaaatTagTaaTaa	CctcttAAtAccctAAcactActc	40/50	GaTaagagtTaTTaaatTagTaaTaa	AAtctAcaAtAAactAcaAttAcc	475
CDM-2	48/50	gaTaagagtTaTTaaatTagTaaTaa	TatAAAAccttAaAAtActtAAt	50/50	GtagagaagTaaTaTaaTaaag	CctcttAAtAccctAAcactActc	470
CDM-3	50/50	gTagagaagTaaTaTaaTaaag	ActActAtAAAaAaAttctAttAc	50/50	GtagggTTtTttggTTagTag	tatAAAAccttAaAAtActtAAt	523
CDM-4	50/50	GTagggTTtTttggTTagTag	ActActAtAAAaAaAttctAttAc	40/50	AttTTtagagggTaggatt	ActActAtAAAaAaAttctAttAc	494
MSSK1-1	48/50	gTttggTTtaaggaTTaggttgTTaag	AaacccttAtaAActAtAAccc	42/50	GagTagTtgggaggTtattta	tAaAAccCcatAtccccaaAAcc	423
MSSK1-2	50/50	gagTagTtgggaggTtattta	CcaaAAAAccacaAatAAttccac-cctAcc	50/50	GaagaTTTTaaagaTtaTtgTaagg	aAacccttAtaAActAtAAccc	410
MSSK1-3	50/50	gaagaTTTTaaagaTtaTtgTaagg	CtaActtAttactAcctccac	50/50	Ttgggtttgtttggggttattttg	ccaaAAAAccacaAatAAttccaccctAcc	445

[a] Two-stage PCR: annealing temperature of the first 5 cycles followed by that for the final 25 cycles.
[b] Upper case letters, sites that result from de-amination of unmethylated cytosines in the template DNA.

Search for promoter regions and transcription factor binding sites was performed with the computer programs SignalScan (75) and TSSW (V.V. Solovyev, A.A. Salaman and C.B. Lawrence, Department of Cell Biology, Baylor College of Medicine, Houston, TX). CpG islands were predicted with X-Grail (version 1.3c; Informatics Group, Oak Ridge National Laboratory, Oak Ridge, TN) and the program CpG-islands-finder (76). Repetitive DNA was identified with the computer programs RepeatMasker (A. Smit, University of Washington Genome Center, Washington DC) and Censor (77).

ACKNOWLEDGEMENTS

C.G. is grateful to S. Clark for the introduction into the bisulfite sequencing technique, to D. Stötzer, S. Fabisch and H. Ludewig for technical assistance and M. Platzer for helpful discussions of the manuscript. The authors are indebted to J. Buard and R.D. Brown for carefully reading the manuscript. C.G. was funded by the EC BIOMED 2 program (contract no. BMH4-CT96-0338).

REFERENCES

1. Jost, J.P. and Saluz, H.P. (eds) (1993) *DNA Methylation: Molecular Biology and Biological Significance*. Birkhäuser Verlag, Basel.
2. Chen, B., Kung, H.F. and Bates, R.R. (1976) Effects of methylation of the beta-galactosidase genome upon *in vitro* synthesis of beta-galactosidase. *Chem. Biol. Interact.*, **14**, 101–111.
3. Sasaki, H., Allen, N.D. and Surani, M.A. (1993) DNA methylation and genomic inprinting in mammals. In Jost, J.P., and Saluz, H.P. (eds), *DNA: Molecular Biology and Biological Significance*. Birkhäuser Verlag, Basel, pp. 469–477.
4. Goto, T. and Monk, M. (1998) Regulation of X-chromosome inactivation in development in mice and humans. *Microbiol. Mol. Biol. Rev.*, **62**, 362–378.
5. Schulz, W.A. (1998) DNA methylation in urological malignancies. *Int. J. Oncol.*, **13**, 151–167.
6. Razin, A. (1998) CpG methylation, chromatin structure and gene silencing-a three-way connection. *EMBO J.*, **17**, 4905–4908.
7. Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.*, **13**, 335–340.
8. Cedar, H. (1988) DNA methylation and gene activity. *Cell*, **53**, 3–4.
9. Bird, A.P. (1995) Gene number, noise reduction and biological complexity. *Trends Genet.*, **11**, 94–100.
10. Yeivin, A. and Razin, A. (1993) Gene methylation patterns and expression. In Jost, J.P. and Saluz, H.P. (eds), *DNA Methylation: Molecular Biology and Biological Significance*. Birkhäuser Verlag, Basel, pp. 521–568.
11. Yisraeli, J. and Szyf, M. (1984) Gene methylation patterns and expression. In Razin, A., Cedar, H. and Riggs, A.D. (eds), *DNA Methylation: Biochemistry and Biological Significance*. Springer Verlag, New York, NY, pp. 353–378.
12. Gama-Sosa, M.A., Midgett, R.M., Slagel, V.A., Githens, S., Kuo, K.C., Gehrke, C.W. and Ehrlich, M. (1983) Tissue-specific differences in DNA methylation in various mammals. *Biochim. Biophys. Acta*, **740**, 212–219.
13. Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L. and Paul, C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl Acad. Sci. USA*, **89**, 1827–1831.
14. Eichler, E.E., Lu, F., Shen, Y., Antonacci, R., Jurecic, V., Doggett, N.A., Moyzis, R.K., Baldini, A., Gibbs, R.A. and Nelson, D.L. (1996) Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.*, **5**, 899–912.
15. Gregor, P., Nash, S.R., Caron, M.G., Seldin, M.F. and Warren, S.T. (1995) Assignment of the creatine transporter gene (SLC6A8) to human chromosome Xq28 telomeric to G6PD. *Genomics*, **25**, 332–333.
16. Sandoval, N., Bauer, D., Brenner, V., Coy, J.F., Drescher, B., Kioschis, P., Korn, B., Nyakatura, G., Poustka, A., Reichwald, K. *et al.* (1996) The genomic organization of a human creatine transporter (CRTR) gene located in Xq28. *Genomics*, **35**, 383–385.
17. Nash, S.R., Giros, B., Kingsmore, S.F., Rochelle, J.M., Suter, S.T., Gregor, P., Seldin, M.F. and Caron, M.G. (1994) Cloning, pharmacological characterization, and genomic localization of the human creatine transporter. *Receptors Channels*, **2**, 165–174.
18. Iyer, G.S., Krahe, R., Goodwin, L.A., Doggett, N.A., Siciliano, M.J., Funanage, V.L. and Proujansky, R. (1996) Identification of a testis-expressed creatine transporter gene at 16p11.2 and confirmation of the X-linked locus to Xq28. *Genomics*, **34**, 143–146.
19. Brenner, V. (1998) *Von der Sequenz zur Funktion: Genomanalyse einer 102 KB-Region des humanen X-Chromosoms*. PhD thesis, Friedrich-Schiller-University, Jena, Germany.
20. Mosser, J., Sarde, C.O., Vicaire, S., Yates, J.R. and Mandel, J.L. (1994) A new human gene (DXS1357E) with ubiquitous expression, located in Xq28 adjacent to the adrenoleukodystrophy gene. *Genomics*, **22**, 469–471.
21. Schneider, T.D. and Sephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
22. Holliday, R. and Pugh, J.E. (1975) DNA modification mechanisms and gene activity during development. *Science*, **187**, 226–232.

23. Riggs, A.D. (1975) X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.*, **14**, 9–25.
24. Razin, A. and Cedar, H. (1991) DNA methylation and gene expression. *Microbiol. Rev.*, **55**, 451–458.
25. Bestor, T.H. (1998) The host defence function of genomic methylation patterns. *Novartis Found. Symp.*, **214**, 187–195.
26. Hsieh, C.L. (1997) Stability of patch methylation and its impact in regions of transcriptional initiation and elongation. *Mol. Cell. Biol.*, **17**, 5897–5904.
27. Bird, A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
28. Bird, A.P. (1993) Functions for DNA methylation in vertebrates. *Cold Spring Harb. Symp. Quant. Biol.*, **58**, 281–285.
29. Jaenisch, R. (1997) DNA methylation and imprinting: why bother? *Trends Genet.*, **13**, 323–329.
30. Spruck, C.H., Rideout, W.M. and Jones, P.A. (1993) DNA methylation and cancer. *EXS*, **64**, 487–509.
31. Jost, J.P. and Saluz, H.P. (1993) Steroid hormone dependent changes in DNA methylation and its significance for the activation or silencing of specific genes. In Jost, J.P. and Saluz, H.P. (eds), *DNA Methylation: Molecular Biology and Biological Significance*. Birkhäuser Verlag, Basel, pp. 423–451.
32. Mavilio, F., Giampaolo, A., Care, A., Migliaccio, G., Calandrini, M., Russo, G., Pagliardi, G.L., Mastroberardino, G., Marinucci, M. and Peschle, C. (1983) Molecular mechanisms of human hemoglobin switching: selective undermethylation and expression of globin genes in embryonic, fetal, and adult erythroblasts. *Proc. Natl Acad. Sci. USA*, **80**, 6907–6911.
33. Lichtenstein, M., Keini, G., Cedar, H. and Bergman, Y. (1994) B cell-specific demethylation: a novel role for the intronic kappa chain enhancer sequence. *Cell*, **76**, 913–923.
34. Fitzpatrick, D.R., Shirley, K.M., McDonald, L.E., Bielefeldt-Ohmann, H., Kay, G.F. and Kelso, A. (1998) Distinct methylation of the interferon gamma (IFN-gamma) and interleukin 3 (IL-3) genes in newly activated primary CD8(+) T lymphocytes: regional IFN-gamma promoter demethylation and mRNA expression are heritable in CD44(high)CD8(+) T Cells. *J. Exp. Med.*, **188**, 103–117.
35. Warnecke, P.M. and Clark, S.J. (1999) DNA methylation profile of the mouse skeletal alpha-actin promoter during development and differentiation. *Mol. Cell. Biol.*, **19**, 164–172.
36. Okuse, K., Matsuoka, I. and Kurihara, K. (1997) Tissue specific methylation occurs in the essential promoter element of the tyrosine hydroxylase gene. *Mol. Brain Res. J.*, **46**, 197–207.
37. Salvatore, P., Benvenuto, G., Caporaso, M., Bruni, C.B. and Chiariotti, L. (1998) High resolution methylation analysis of the galectin-1 gene promoter region in expressing and non-expressing tissues. *FEBS Lett.*, **421**, 152–158.
38. Werner, T. (1999) Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome*, **10**, 168–175.
39. Lifschytz, E. and Lindsley, D.L. (1972) The role of X-chromosome inactivation during spermatogenesis. *Proc. Natl Acad. Sci. USA*, **69**, 182–186.
40. McCarrey, J.R., Berg, W.M., Paragioudakis, S.J., Zhang, P.L., Dilworth, D.D., Arnold, B.L. and Rossi, J.J. (1992) Differential transcription of Pgk genes during spermatogenesis in the mouse. *Dev. Biol.*, **154**, 160–168.
41. Surani, A. (1998) Imprinting and the initiation of gene silencing in the germ line. *Cell*, **93**, 309–312.
42. Rocamora, N. and Mezquita, C. (1984) Hypomethylation of DNA in meiotic and psotmeiotic rooster testis cells. *FEBS Lett.*, **177**, 81–84.
43. Rocamora, N. and Mezquita, C. (1989) Chicken spermatogenesis is accompanied by a genomic-wide loss of DNA methylation. *FEBS Lett.*, **247**, 415–418.
44. Vanyushin, B.F., Mazin, A.L., Vasilyev, V.K. and Belozersky, A.N. (1973) The content of 5-methylcytosine in animal DNA: the species and tissue specificity. *Biochim. Biophys. Acta*, **299**, 397–403.
45. Feinstein, S.I., Racanielle, V.R., Ehrlich, M., Gehrke, C., Miller, D.A. and Miller, O.J. (1985) Pattern of undermethylation of the major satellite DNA of mouse sperm. *Nucleic Acids Res.*, **13**, 3969–3978.
46. Ariel, M., McCarrey, J. and Cedar, H. (1991) Methylation patterns of testis-specific genes. *Proc. Natl Acad. Sci. USA*, **88**, 2317–2321.
47. Rubin, C.M., Van de Voort, C.A., Teplitz, R.L. and Schmid, C.W. (1994) Alu repeated DNAs are differentially methylated in primate germ cells. *Nucleic Acids Res.*, **22**, 5121–5127.
48. Rahe, B., Erickson, R.P. and Quinto, M. (1983) Methylation of unique sequence DNA during spermatogenesis in mice. *Nucleic Acids Res.*, **11**, 7947–7959.
49. Robertson, N.G., Pomponio, R.J., Mutter, G.L. and Morton, C.C. (1991) Testis-specific expression of the human MYCL2 gene. *Nucleic Acids Res.*, **19**, 3129–3137.
50. Van de Berg, J.L., Cooper, D.W. and Close, P.J. (1976) Testis specific phosphoglycerate kinase B in mouse. *J. Exp. Zool.*, **198**, 231–240.
51. Fitzgerald, J., Hutchison, W.M. and Dahl, H.H. (1992) Isolation and characterisation of the mouse pyruvate dehydrogenase E1 alpha genes. *Biochim. Biophys. Acta*, **1131**, 83–90.
52. Kumar, G., Patel, D. and Naz, R.K. (1993) c-MYC mRNA is present in human sperm cells. *Cell Mol. Biol. Res.*, **39**, 111–117.
53. Fitzgerald, J., Wilcox, S.A., Graves, J.A. and Dahl, H.H. (1993) A eutherian X-linked gene, PDHA1, is autosomal in marsupials: a model for the evolution of a second, testis-specific variant in eutherian mammals. *Genomics*, **18**, 636–642.
54. Salehi-Ashtiani, K., Widrow, R.J., Markert, C.L. and Goldberg, E. (1993) Testis-specific expression of a metallothionein I-driven transgene correlates with undermethylation of the locus in testicular DNA. *Proc. Natl Acad. Sci. USA*, **90**, 8886–8890.
55. Nagashima, H., Imai, M. and Iwakura, Y. (1993) Aberrant tissue specific expression of the transgene in transgenic mice that carry the hepatitis B virus genome defective in the X gene. *Arch. Virol.*, **132**, 381–397.
56. Dupressoir, A. and Heidmann, T. (1996) Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice. *Mol. Cell. Biol.*, **16**, 4495–4503.
57. Mehtali, M., LeMeur, M. and Lathe, R. (1990) The methylation-free status of a housekeeping transgene is lost at high copy number. *Gene*, **91**, 179–184.
58. Goto, T., Christians, E. and Monk, M. (1998) Expression of an Xist promoter-luciferase construct during spermatogenesis and in preimplantation embryos: regulation by DNA methylation. *Mol. Reprod. Dev.*, **49**, 356–367.
59. Iannello, R.C., Young, J., Sumarsono, S., Tymms, M.J., Dahl, H.H., Gould, J., Hedger, M. and Kola, I. (1997) Regulation of Pdha-2 expression is mediated by proximal promoter sequences and CpG methylation. *Mol. Cell. Biol.*, **17**, 612–619.
60. Chen, R.Z., Petterson, U., Beard, C., Jackson-Grusby, L. and Jaenisch, R. (1998) DNA hypomethylation leads to elevated mutation rates. *Nature*, **395**, 89–93.
61. Hsieh, C.L. and Lieber, M.R. (1992) CpG methylated minichromosomes become inaccessible for V(D)J recombination after undergoing replication. *EMBO J.*, **11**, 315–325.
62. Holliday, R. (1984) The biological significance of meiosis. *Symp. Soc. Exp. Biol.*, **38**, 381–394.
63. Blanco, R.J. and Martinez, G.C. (1996) Spontaneous germ cell death in the testis of the adult rat takes the form of apoptosis: re-evaluation of cell types that exhibit the ability to die during spermatogenesis. *Cell Prolif.*, **29**, 13–31.
64. Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A. and Gehrke, C. (1982) Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res.*, **10**, 2709–2721.
65. Schnedl, W., Erlanger, B.F. and Miller, O.J. (1976) 5-methylcytosine in heterochromatic regions of chromosomes in Bovidae. *Hum. Genet.*, **31**, 21–26.
66. Orend, G., Kuhlmann, I. and Doerfler, W. (1991) Spreading of DNA methylation across integrated foreign (adenovirus type 12) genomes in mammalian cells. *J. Virol.*, **65**, 4301–4308.
67. Ruault, M., Trichet, V., Gimenez, S., Boyle, S., Gardiner, K., Rolland, M., Roizes, G. and De Sario, A. (1999) Juxta-centromeric region of human chromosome 21 is enriched for pseudogenes and gene fragments. *Gene*, **239**, 55–64.
68. Matzke, A.J., Neuhuber, F., Park, Y.D., Ambros, P.F. and Matzke, M.A. (1994) Homology-dependent gene silencing in transgenic plants: epistatic silencing loci contain multiple copies of methylated transgenes. *Mol. Gen. Genet.*, **244**, 219–229. [Erratum (1995) *Mol. Gen. Genet.*, **247**, 264.]
69. Zhang, L.P., Stroud, J.C., Walter, C.A., Adrian, G.S. and McCarrey, J.R. (1998) A gene-specific promoter in transgenic mice directs testis-specific demethylation prior to transcriptional activation *in vivo*. *Biol. Reprod.*, **59**, 284–292.
70. Pennica, D., Swanson, T.A., Shaw, K.J., Kuang, W.J., Gray, C.L., Beatty, B.G. and Wood, W.I. (1996) Human cardiotrophin-1: protein and gene structure, biological and binding activities, and chromosomal localization. *Cytokine*, **8**, 183–189.
71. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Labortatory Manual*, 2nd edn.Cold Spring Harbor Laboratory Press, New York, NY.
72. Clark, S.J. and Frommer, M. (1997) Bisulphite genomic sequencing of methylated cytosines. In Taylor, G.R. (ed.), *Laboratory Methods for the Detection of Mutations and Polymorphisms in DNA*. CRC Press, Boca Raton, pp. 151–161.

73. Dear, S. and Staden, R. (1991) A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.*, **19**, 3907–3911.
74. Grunau, C., Schattevoy, R., Mache, N. and Rosenthal, A. (2000) MethTools—a toolbox to visualize and analyze DNA methylation data. *Nucleic Acids Res.*, **28**, 1053–1058.
75. Prestridge, D.S. (1991) SIGNAL SCAN: a computer program that scans DNA sequences for eucaryotic transcriptional elements. *CABIOS*, **7**, 203–206.
76. Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
77. Jurka, J., Klonowski, P., Dagman, V. and Pelton, P. (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.*, **20**, 119–121.