Human Molecular Genetics, 2003, Vol. 12, No. 18 **2249–2254**
DOI: 10.1093/hmg/ddg246

Functional analysis of human promoter polymorphisms

Bastiaan Hoogendoorn*, Sharon L. Coleman, Carol A. Guy, Kaye Smith, Tim Bowen, Paul R. Buckland and Michael C. O'Donovan

Department of Psychological Medicine, University of Wales College of Medicine, Heath Park, Cardiff CF14 4XN, UK

Received November 22, 2002; Revised and Accepted March 6, 2003

The potential importance of gene regulation in disease susceptibility and other inherited phenotypes has been underlined by the observation that the human genome contains fewer protein coding genes than expected. Promoter sequences are potential sources of polymorphism affecting gene expression, although to date there are no large-scale systematic studies that have determined how frequently such variants occur. We have used denaturing high performance liquid chromatography to screen the first 500 bp of the 5′ flanking region of 170 opportunistically selected genes identified from the Eukaryotic Promoter Database (EPD) for common polymorphisms. Using a screening set of 16 chromosomes, single-nucleotide polymorphisms were found in ~35% of genes. It was attempted to clone each of these promoters into a T-vector constructed from the reporter gene vector pGL3. The relative ability of each promoter haplotype to promote transcription of the luciferase gene was tested in each of three human cell lines (HEK293, JEG and TE671) using a co-transfected SEAP-CMV plasmid as a control. The findings suggest that around a third of promoter variants may alter gene expression to a functionally relevant extent.

INTRODUCTION

Based upon the unexpectedly low number of protein encoding genes (1) and, more recently, the types of variants that have been shown to be responsible for some quantitative traits in simpler organisms (2,3), it has been proposed that the genetic causes for susceptibility to complex diseases may reflect a different spectrum of sequence variants to the missense and nonsense mutations that dominate simpler genetic disorders. Amongst this spectrum, polymorphisms that alter gene expression are suspected of playing a prominent role. If this is correct, the implication is that a considerable proportion of human genes show inter-individual variation in gene expression and that this is substantially attributable to *cis*-acting genetic mechanisms.

Because our knowledge of regulatory elements in the human genome is not comprehensive, and because functional polymorphisms in regulatory elements cannot yet be reliably distinguished from those with no effect simply by their sequence context, the data from the large-scale SNP discovery programmes instituted by public and private initiatives (4,5) cannot be used to test this hypothesis directly. However, indirect studies suggest that sequence variants do affect the expression of a substantial number of genes in cultured human cell lines (6), a finding supported by our own unpublished data from RNA extracted from human brain.

Promoters are involved in initiating transcription and are therefore among the many important *cis*-acting elements that regulate gene expression that might harbour functionally relevant polymorphisms. However, they differ from most regulatory elements in that their locations are fixed relative to the positions of their respective genes, and therefore promoter sequences are an ideal choice for undertaking large-scale analysis and functional annotation. However, because direct laborious experimental analysis is required to determine functionality, studies thus far have investigated a small number (typically 1) of promoters, generally selected because they were already known or suspected of involvement in a specific disease. Consequently, we do not even have approximate unbiased estimates of the frequency with which functional sequence variants occur in promoters. In this study, we have started to address this. Our study is systematic in that we targeted all the experimentally proven human promoters with >250 bases of sequence available as of August 1999. We screened the promoters for sequence variation, cloned the polymorphic promoter sequences into a reporter gene vector, and tested haplotype pairs for each promoter for intrinsic differences in their ability to drive transcription in a reporter gene assay. The results indicate that a surprisingly high proportion of promoter variants, approximately a third in this study, may modify gene expression by 50% or more.

*To whom correspondence should be addressed. Tel: +44 2920742535; Fax: +44 2920746554; Email: hoogendoornb@cf.ac.uk

RESULTS

To ensure we examined true promoters, we drew sequences exclusively from the Eukaryotic Promoter Database (7,8), which curates experimentally characterized transcription start points. In August 1999, EPD contained 233 unique human promoters with at least 250 bases of available sequence. Of these, we were able to amplify and screen 170 promoters representing 73 995 bases of proximal promoter sequence. As EPD still only contains a total of 281 human promoter sequences (October 2002), our completed analysis represents 60% of all the curated experimentally defined human promoters.

Denaturing high performance liquid chromatography (DHPLC) (9) analysis followed by sequencing revealed a total of 120 sequence variants which comprised 109 transitions and transversions, six single base insertion/deletions and five multiple base insertion/deletions. The ratio of transitions to transversions was 1.6, similar to the expected ratio of 1.4 (10). All detected variants are listed in Appendix 1 which is available at our web site (www.uwcm.ac.uk/study/medicine/psychological_medicine/pub_data/EPD.htm). The average nucleotide diversity in 5′ flanking sequence was 4.9×10^{-4} which is similar to previous reports of that in untranslated exons and the intronic sequence immediately adjacent to exons (11).

As expected, the GC content of the fragments we were able to amplify was lower (mean = 51%) than that of those we were unable to amplify (mean = 62%). It is possible that this bias in the GC content of amplified promoters might result in an underestimate of diversity because the mutation rate increases with high GC content (12,13). However, the degree of underestimation is likely to be marginal as the effect of GC content appears to plateau above 50% (13).

The polymorphisms were distributed across 59 promoters (Appendix 1). We attempted to clone all 59 but were unable to clone 18 promoters because of low cloning efficiency and, despite the use of proofreading enzyme and bacterial hosts designed to minimize plasmid re-arrangement, spontaneous occurrence of single base changes or re-arrangements in the insert occurred. The 41 cloned promoters represent 91 haplotypes. Full details of all cloned haplotype sequences are given in Appendix 2 of our web site (mentioned above). All cloned haplotypes were tested for their ability to drive transcription of the luciferase reporter gene in three human cell lines, HEK293t, JEG-3 and TE671, using a 96-well cell culture format (14). The cell lines are clearly not exhaustive but were chosen to represent a variety of human tissues types and represent a balance between choosing cell lines with different tissue origins and robustness of the assay conditions. Several other cell lines were tested but found that either low transformation efficiencies or poor adherence to 96-well plates made them unsuitable for high-throughput assays.

It has previously been shown that a 10-fold increase in reporter gene activity over the basic promoter-less vector provides a conservative definition of promoter activity (14). However, this applies to predicted rather than proven promoters. As all sequences in this study fall into the latter group this criterion was considered to be unnecessary. However, to avoid 'floor effects', comparisons were restricted to those promoters with at least one haplotype showing at least 2-fold promoter activity over the pGL3-basic control. Twenty-one promoters did not show functional differences by our criteria: RNU4C, MMP1, IGF2, DAF, IFIT1, PRL, MT1B, WT1, H4FG, TNNI1, KRT1 (MUT), APOC2, APOE, ORM1, BCKDHA, PGC, AMY1A, CYP21A2, CA3, H3FL, NPPA. A further three promoters did not show activity greater than that of the pGL3-basic plasmid in any of the three cell lines: IFNA13, PROC, FGB. The constructs and reporter gene activities of the remaining 17 promoters with haplotypes displaying either confirmed or suggestive differences in functional activity are shown in Table 1. Full details of all reporter assays are given in Appendix 2 of our web site. All promoter constructs were co-transfected with an internal control expression vector (CMV-SPAP) to ensure transfection efficiencies were similar for all test samples.

Approximate allele frequencies were estimated by sequencing the relevant PCR fragments in pooled DNA samples constructed from 180 anonymous white blood donors born in the UK (Table 2). Although estimates using this method are extremely crude, 70% of the single-nucleotide polymorphisms (SNPs) had variants that could be clearly detected in pools, suggesting frequencies of >0.1%.

DISCUSSION

The main goal of this study was to provide an estimate of the frequency with which functional sequence variation occurs in proximal promoter sequences. In the absence of a consensus on functionally important differences in promoter activity, we chose to classify differences between haplotype pairs as significant where they met each of the following criteria. First, across eight replicates of each haplotype, the reporter gene activity of one haplotype was at least 1.5 times that of its comparator. Second, the difference was significant at $P \leq 0.05$. Third, the findings replicated at $P \leq 0.05$ in an assay using eight replicates of independent clones. A relative difference of 1.5 was chosen for two reasons. First, if this is reflected *in vivo*, homozygotes for the high activity allele could be considered to carry an extra copy of the gene relative to homozygotes for the low activity allele. However, this is a working definition, and it is likely that smaller changes in some genes may have biological relevance (as opposed to simply altering expression) while larger changes in others may not. The second reason was pragmatic. In pilot experiments (data not shown), we have shown that changes of this magnitude are greater than the random errors intrinsic to the assay, and are highly reproducible in independent replication studies using fresh clones. Of the 38 promoters yielding functional data, 17 had at least one haplotype that drove expression 1.5 times greater than another haplotype from the same promoter and in which that difference was statistically significant. Of these, 13 were replicated at a level that was statistically significant, while the other four showed a trend in the same direction as the original assay, but which was not significant at $P \leq 0.05$. The results confirm the high reproducibility we expected to achieve using the criteria above. Thus, according to our full criteria, including that of replication, 34% of the promoters tested showed significant differences between pairs of haplotypes in at least one cell line. From the perspective of haplotype pairs rather than promoters, our data allow a total of 68 comparisons between pairs of haplotypes from the same promoter. Nineteen (28%)

Table 1. List of genes tested in reporter gene assay yielding functional differences. Reporter gene activity corrected by SPAP is given relative to the promoterless pGL3 basic vector. Relative allele activity is expressed as a percentage of the highest expressing haplotype. Where the difference is significant according to our full criteria, 95% confidence intervals for the difference are given in brackets

Gene symbol	Haplotype	Activity relative to control			Relative allele activity			Haplotype
		HEK293T	JEG-3	TE671	HEK293T	JEG-3	TE671	
NOS2A	A	13	1	5	94	N/A	100	G −278, C +38
	B	14	1.1	3	100	N/A	61 (39, 83)**	A −278, G +38
MMP3	A	11	1	3	100	N/A	100	C −377
	B	10	1	1.1	91	N/A	33 (0, 77)*	G −377
GAS	A	46	7.5	6.7	66 (39, 94)*	53 (45, 61)***	56 (39, 93)***	G −327, G −289, G −99
	B	69	14.2	12	100	100	100	A −327, A −289, T −99
TRD@	A	2.4	<1	1.3	100	N/A	61 (48, 76)***	G −210
	B	2.2	1.3	2.2	86	N/A	100	A −210
SST	A	115	1.8	4	100	53 (36, 70)***	100	T −318, T −275, C −249, del −152
	B	74	3.4	4	64	100	90	G −318, A −275, T −249, T ins −152
FSHB	A	6.7	1.2	1.3	86	46 (38, 55)***	58 (46, 71)***	A −428, T −212
	B	7.8	2.6	2.3	100	100	100	T −428, G −212
CEACAM6	A	114	76	37	100	100	100	G −379, A −210
	B	81	41	33	70	54 (45, 63)***	90	A −379, G −210
TNP1	A	80	<1	5.7	52 (44, 61)***	N/A	100	C −363, G −256
	B	154	<1	5.4	100	N/A	96	T −363, A −256
IVL	A	35	3.5	2.5	100	100	100	A −182, G −150
	B	29	2.7	1.5	83	76	60 (41, 80)**	G −182, A −150
APOA2	A	21	1.1	2	66 (51, 82)***	N/A	89	C −110
	B	31	1.9	2.2	100	N/A	100	T −110
ALB	A	15	3.1	4.6	74	100	87	A −358, T −315
	B	20	1.2	5.3	100	39 (29, 48)***	100	G −358, A −315
GHRH	A	12	1.7	2.3	100	N/A	100	T −57
	B	2	1	1.8	21 (1, 42)***	N/A	76	C −57
HLA-DRA	A	38	5	3	87	66 (T)	80	A −361, T −353, G −261, G −232, T −225, T −197
	B	44	7	4	100	100	100	G −361, G −353, C −261, C −232, C −225, C −197
NEFL	A	114	6.2	9.7	95	54 (T)	69	A −273
	B	121	11	14	100	100	100	G −273
SLC9A1	A	411	684	108	66	95	62 (T)	G −156
	B	618	723	174	100	100	100	T −156
RNU3	A	2722	496	101	100	100	100	T −186, G −12, T +22
	B	2087	398	84	77	80	83	T −186, G −12, G +22
	C	2720	379	85	100	70	84	T −186, A −12, T +22
	D	2583	338	80	95	68	79	C −186, G −12, T +22
	E	2383	319	84	87	64 (T)	83	C −186, A −12, G +22
	F	1864	310	90	68	63 (T)	89	C −186, G −12, G +22
PRM2	A	28	1.5	21	100	N/A	100	A −289, T −286, G −271, C −221, G −126
	B	3.9	1.4	11	14 (0, 45)***	N/A	51 (14, 87)*	A −289, T −286, G −271, T −221, G −126
	C	1.6	1	16	6 (0, 37)***	N/A	76	G −289, T −286, C −271, C −221, A −126
	D	1.3	1	13	5 (0, 37)***	N/A	60 (38, 82)**	G −289, C −286, G −271, C −221, A −126
	E	1.7	1	11	6 (0, 38)***	N/A	51 (19, 80)**	G −289, C −286, G −271, C −221, G −126
PRM2: in HEK B haplotype = 100, C haplotype 42 (26, 53)***, D haplotype 36 (15, 51)***, E haplotype 42 (26, 58)***								

NA, levels of reporter gene activity not high enough for analysis. Statistical significance for all assays that showed significant differences upon replication: where $*P < 0.05$; $**P < 0.005$, and $***P < 0.000$. 'T' denotes differences that met initial statistical and functional criteria for significance but which only yielded a trend upon replication. For the PRM2 genes, the values given in the table refer to comparison with the A haplotype. The values for comparison with the B haplotype are given as the last row of the table. None of the other pairs of haplotypes yielded significant differences.

showed significant replicated differences between pairs as defined above.

Three of the promoters did not yield activity greater than double that of the pGL-basic in any of the three cell lines. The activity of other promoters varied widely between cell lines (Table 1). This in part is likely to represent quantitative differences in the transformation efficiencies of different cell lines. However, given that there were only five haplotype-pair comparisons representing only three promoters for which significant changes occurred in more than one cell line

Table 2. List of SNPs from which functional haplotypes were comprised. The frequency and base sequence of the minor alleles are given as estimated by sequencing a pool of 180 blood donors of northern European descent

Gene	SNP	Minor allele frequency
NOS2A	G/A −278	0.5/G
NOS2A	G/C +38	0.5/G
MMP3	G/C −377	0.5/G
GAS	A/G −327	0.2/G
GAS	A/G −289	ND/G
GAS	T/G −99	0.1/G
TRD@	G/A −210	0.2/A
SST	G/T −318	ND/T
SST	A/T −275	ND/T
SST	T ins/del −152	ND
SST	T/C −249	0.1/G
FSHB	G/T	0.2/T
FSHB	T/A −428	N/A
CEACAM6	G/A −379	0.5/G
CEACAM6	G/A −210	0.4/A
TNP1	C/T −363	0.4/T
TNP1	G/A −256	0.5/G
PRM2	C/T −221	0.1/T
PRM2	G/A −289	ND/A
PRM2	T/C −286	0.3/C
PRM2	G/C −271	0.3/C
PRM2	G/A −126	0.1/A
IVL	G/A −182	ND/A
IVL	G/A −150	0.1/A
APOA2	C/T −110	ND/C
ALB	G/A −358	0.4/A
ALB	T/A −315	0.4/A
GHRH	C/T −57	ND/T

ND, below the detectable level.

(Table 1), it is likely that there are also qualitative differences in the responses of the cell lines to individual promoter haplotypes. Although one can question the specific choice of cell lines used, this observation supports the importance of using multiple cell lines in this analysis.

Ideally, it would be possible to determine likely functionality using bio-informatics rather than 'wet laboratory' procedures. We explored this possibility by attempting *post hoc* to distinguish between haplotypes that resulted in a change in function and those that did not. We initially sought to determine if the former were more likely to change a putative transcription factor (TF) binding site as predicted by the programme Consite (http://forkhead.cgr.ki.se/cgi-bin/consite). We examined 48 haplotype pairs comprising 18 pairs in which functional differences were observed and 30 pairs in which they were not. Haplotypes where findings were found at the trend level (Table 1) or which contained more than three variants were excluded. A higher proportion of haplotypes (13/18) showing functional differences between pairs contained one or more variants that altered the sequence of a predicted TF site than did variants with no demonstrated function (17/30) but this was not significantly different ($\chi^2 = 0.47$, $P = 0.49$) and only represents a very modest enrichment of functional haplotypes by using this procedure. Promoter haplotypes with a polymorphic site within a predicted TF site were also more likely to display significant differences in more than one cell line (4/4) than those showing significant differences in only one cell line (9/14) but, again, this was not significant (Fisher's exact test, $P = 0.27$). We also undertook a more restrictive analysis of the sequences, using Consite to search for conserved TF binding sites in the human and orthologous mouse sequences. None of the variants we report as functional were located in a conserved TF binding site using the default settings. Although we were unable to distinguish between functional and non-functional variation by bio-informatics approaches, we note that the first approach may lack specificity, and the second sensitivity. Thus, a literature search revealed that variants in involucrin (IVL) (15), gastrin (GAS) (16) and solute carrier family 9, isoform 1 (SLCA1) (17) are within experimentally determined TF binding sites that were not identified by human–mouse comparative analyses. We also used both CONPRO (18) and MatInspector (19) at a range of levels of stringency. While each of the three programmes predicted somewhat different TF binding site profiles for a given fragment, the results were similar in that none of the programmes could significantly differentiate between functional and non-functional haplotypes (data not shown).

Our sample size for mutation analysis was determined with the common disease, common variant hypothesis in mind, (20,21). Regardless of whether or not that hypothesis is correct (22,23), clearly our observation that around 35% of proximal promoters are polymorphic is an underestimate because, while screening 16 chromosomes is expected to detect most of the common variants in our population, detection is not exhaustive (24). We also found that a high proportion (~35%) of promoters that are polymorphic and that we were able to clone showed significant replicated differences between haplotype pairs. We are not in a position to note if our inability to clone sequences introduced a bias in this estimate.

There are a number of limitations to our study. First, our estimate of the degree of functional variation in promoters is almost certainly an underestimate as we have only examined the proximal promoters, and we have not measured the effects of polymorphisms under dynamic states, for example in response to hormonal challenges, during development, nor exhaustively in cell lines from all tissues. Moreover, although it was not the objective of this study, it is worth stressing that our analysis does not address the effects of polymorphism in the many, often distal, regulatory elements outside of the proximal promoter. A second caveat is that the degree to which the magnitude of the mRNA changes we have observed will be reflected by mRNA change *in vivo* is unknown as regulation of expression of genes in living tissue in their natural genomic contexts is likely to be more sophisticated than it is in a reporter gene assay. Similarly, as a result of translational and post-translational regulatory process it is possible that changes in mRNA may not result in changes in protein abundance or activity.

Estimates of the effect of *cis*-acting variants on mRNA expression *in vivo* can be obtained by measuring relative allelic expression (6,25). This method has the advantage of indirectly summing the effect within individuals of heterozygosity at all regulatory loci, known and unknown. Conversely, this advantage limits application of the method for testing specific haplotypes of single regulatory elements or of well-circumscribed regions. In principle the method could be applied if one had tissue samples from multiple organs at

multiple stages of development from individuals who are heterozygous for the specific circumscribed haplotype to be tested who are also known to be homozygous for variation at all other regulatory sites. It is therefore simply not practical for us to use this method to test the specific promoter variants we have identified, but it is of interest that the indirect methods concur with our own analysis in that a high proportion of genes in human cell lines (6) and brain tissue (our own unpublished data) display evidence for polymorphisms affecting gene expression.

From the 170 promoters we were able to screen, we determined that around 35% were polymorphic. Of those we were able to clone, about 35% had at least one haplotype that met our criteria for displaying functional differences. Thus, we estimate that 0.35×0.35 or around 10% of promoters have functionally relevant sequence variants in their promoters. Of the 27 variants comprising the functional haplotypes, ~70% (19/27) had minor allele frequencies estimated greater than 0.1, and ~50% (14/27) had minor frequencies greater than 0.2 (Table 2). Thirty per cent of the variants in functional haplotypes were rare, and possibly private to those individuals in whom they were detected. Because the power to detect rare haplotypes is low (our power to detect variants with a frequency of 1% is 0.14), there are clearly many more functional haplotypes of low frequency than we have observed. Like the common variants, the cumulative effects of a large number of rare functional variants can result in common phenotypes (22). Thus, we conclude that our data lend empirical credibility to the hypothesis that polymorphism within promoters may be a common source of phenotypic variation and possibly susceptibility to common disease, particularly since our estimate of the frequency of functional variation is likely to be rather conservative.

MATERIALS AND METHODS

Promoter selection and identification of sequence variation

Promoters from all 233 sequences in EPD (10,11) that contained at least 250 bases of 5′ flanking sequence were extracted from EPD. Primers were designed using Primer 3 (26). The 3′ primer for each target promoter was designed to include sequence corresponding to the transcription start site in each amplicon, but not include any coding sequence in order to avoid changing the open reading frame of the reporter gene or making a target-reporter fusion protein. Where possible we included no more than 50 bp of 5′-UTR and avoided inclusion of any untranslated ATG sequences. For promoter regions without obvious TATA-boxes, we endeavoured to include at least 35 bp of the 5′-UTR as these may contain downstream promoter elements (27).

Mutation screening was performed using DHPLC, utilizing sensitive procedures based upon DHPLCMelt (http://insertion.stanford.edu/melt.html) we have published elsewhere (28). We screened DNA from eight unrelated anonymized UK residents of northern European descent. DNA was amplified by PCR using QiaTaq (Qiagen). Fragments yielding chromatograms indicating heteroduplex formation were sequenced using Big Dye Terminator chemistry (Applied Biosystems).

Cloning and reporter gene assay

This is described in detail elsewhere (16). Briefly, DNA from heterozygous individuals was amplified using Expand High Fidelity DNA polymerase (Roche) to minimize mis-incorporation of nucleotides. PCR products were ligated into a pGL3 vector (Promega) that we modified for T/A cloning and cloned into SURE 2 supercompetent cells (Stratagene). Plasmid DNA was purified using Qiagen chemistry and proprietary procedures (Qiagen), and was sequenced using Big Dye Terminator chemistry (Applied Biosystems) in both directions using RV3 (ctagcaaaataggctgtccc) and GL2 (ctttatgtttttggcgtcttcca) to confirm that the haplotypes present in genomic DNA were faithfully represented. Each cloned haplotype represents a naturally occurring haplotype found in one or more chromosomes.

The ability of each sequence to promote transcription of the luciferase gene was tested transiently in human cell lines HEK293t (human embryo kidney, a gift from GlaxoSmithKline), TE671 (human medulloblastoma) and JEG-3 (human choriocarcinoma). The latter two lines were obtained from the European Collection of Cell Cultures (ECACC). The cell lines have all been successfully used by other researchers in similar transient transfection assays and represent varied tissues types. TE671 was thought to be of medulloblastoma origin by ECACC, but is also now known to be identical to the human rhabdomyosarcoma RD cell line (ECACC no. 85111502). All plasmid DNA was quantitated fluorimetrically using Pico Green (Molecular Probes) and a TD-700 (Turner Designs) fluorimeter. Cell lines were transfected with plasmid using lipofectamine (Gibco) in 96-well format (eight replicates per clone for each cell line), and cultured according to ECACC specifications at 37°C with 5% CO_2. Cells were seeded into black, clear bottomed 96-well luminometric plates (Perkin Elmer) at ~60% of confluency the day prior to transfection. To control for transfection efficiency, cells were co-transfected with CMV-SPAP (a gift from GlaxoSmithKline). Cell lines were transfected overnight in serum-free medium which was replaced with complete heat inactivated medium (PAA Laboratories) and incubated for a further 24 h. SPAP activity was measured in the culture medium after transferring to a second 96-well plate using a phospha-light kit (Tropix) according to the manufacturer's instructions. Luciferase activity in the remaining cells was measured in the original plate using a Luc Screen assaying kit (Tropix). Both plates were read on a TR717 luminometer for 1–10 s per well.

Promoter activity was normalized by dividing luciferase activity by SPAP activity. In representative assays the coefficient of variance for 10 data sets averaged for HEKs: 0.24 for luciferase measurements alone, 0.25 for SPAP measurements alone, and 0.14 for luciferase corrected with SPAP; for TEs the values were 0.38, 0.34 and 0.20, respectively, and for JEGs 0.26, 0.28 and 0.18. Pairs of haplotypes were compared by Student's *t*-test or, where the data were not normally distributed, by Mann–Whitney test using Minitab version 13 (Minitab Inc.). Where significant differences were observed between haplotypes, the analysis was

replicated using fresh clone preparation to discount the possibility of differences being attributable to random variation in plasmid preparation.

In unpublished work we have found rare spontaneous changes in the sequence of the reporter gene. In each case, the reporter gene activity dropped below our definition of promoter activity for this study (×2 basic). To ensure that the present data were not the result of disruption of the reporter gene, where activities of pairs of clones showed differences in same direction (even where they were not significant) across all the cell lines in which they were expressed, we sequenced the entire luciferase gene. Thus, of the promoters showing a significant difference in pairs (Table 1) we sequenced seven of 13. In no cases was a change evident. One clone, however, was discarded prior to detailed sequencing of the luciferase gene because of a mutation in the 5′-UTR of the luciferase gene within the region that was routinely sequenced along with the insert in all clones.

Transcription factor binding site analysis

Putative transcription factor binding sites were identified by the programme Consite (http://forkhead.cgr.ki.se/cgi-bin/consite). Each sequence was screened at high levels of stringency for human transcription factors with a transcription factor score of 90% (the default is 80%). For the analyses based upon cross-species conservation, we aligned each human promoter sequence with 10 kb of mouse sequence that is immediately 5′ to the relevant orthologous reference mRNA sequence in GenBank. We used the 'all vertebrate' transcription factor setting, combined with the default settings for conservation and transcription factor scores. We also applied more relaxed parameters of 70% for conservation and 80% for transcription factor scores where the default criteria were more stringent.

ACKNOWLEDGEMENTS AND DECLARATION

This work was funded by grants from the MRC (UK). We have no competing financial interests in the data presented.

REFERENCES

1. Peltonen, L. and McKusick, V.A. (2001) Genomics and medicine. Dissecting human disease in the postgenomic era. *Science*, **291**, 1224–1229.
2. MacKay, T. (2002) Quantitative trait loci in *Drosophila*. *Nat. Rev. Genet.*, **2**, 11–20.
3. Toma, D.P., White, K.P., Hirsch, J. and Greenspan, R.J. (2002) Identification of genes involved in *Drosophila melanogaster* geotaxis, a complex behavioral trait. *Nat. Genet.*, **31**, 349–353.
4. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
5. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
6. Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. and Kinzler, K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.
7. Praz, V., Périer, R.C., Bonnard, C. and Bucher, P. (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucl. Acids Res.*, **30**, 322–324.
8. Périer, R.C., Praz, V., Junier, T., Bonnard, C. and Bucher, P. (2000) The Eukaryotic Promoter Database (EPD). *Nucl. Acids Res.*, **28**, 302–303.
9. Oefner, P.J. and Underhill, P.A. (1998) DNA mutation detection using denaturing high-performance liquid chromatography (DHPLC). *Curr. Protoc. Hum. Genet.*, **19** (suppl.), 7.10.1–7.10.12.
10. Collins, D.W. and Jukes, T.H. (1994) Rates of transition and transversion in coding sequences since the human–rodent divergence. *Genomics*, **20**, 386–396.
11. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, **22**, 231–238.
12. Cooper, D.N. and Youssoufian, H. (1988) The CpG dinucleotide and human genetic disease. *Hum. Genet.*, **78**, 151–155.
13. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
14. Coleman, S.L., Hoogendoorn, B., Guy, C., Smith, S.K., O'Donovan, M.C. and Buckland, P.R. (2002) A streamlined approach to functional analysis of promoter polymorphisms. *BioTechniques*, **33**, 412–418.
15. Kubo, E., Fatma, N., Sharma, P., Shinohara, T., Chylack, L.T. Jr., Akagi, Y. and Singh, D.P. (2002) Transactivation of involucrin, a marker of differentiation in keratinocytes, by lens epithelium-derived growth factor (LEDGF). *J. Mol. Biol.*, **320**, 1053–1063.
16. Tillotson, L.G. (1999) A novel zinc finger gene, encodes proteins that bind to the CACC element of the gastrin promoter. *J. Biol. Chem.*, **274**, 8123–8128.
17. Yang, W., Wang, H. and Fliegel, L. (1996) Regulation of Na^+/H^+ exchanger gene expression. Role of a novel poly(dA.dT) element in regulation of the NHE1 promoter. *J. Biol. Chem.*, **271**, 20444–20449.
18. Liu, R. and States, D.J. (2002) Consensus promoter identification in human genome utilizing expressed sequence markers and gene modeling. *Genome Res.*, **12**, 462–469.
19. Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector—New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.*, **23**, 4878–4884.
20. Lander, E.S. (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.
21. Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet.*, **17**, 502–510.
22. Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–137.
23. Wright, A.F. and Hastie, N.D. (2001) Complex genetic diseases: controversy over the Croesus code. *Genome Biol.*, **2**, 2007.1–2007.8.
24. Kruglyak, L. and Nickerson, D.A. (2001) Variation is the spice of life. *Nat. Genet.*, **27**, 234–236.
25. Cowles, C.R., Hirschhorn, J.N., Altshuler, D. and Lander, E.S. (2002) Detection of regulatory variation in mouse genes. *Nat. Genet.*, **32**, 432–437.
26. Rozen, S. and Skaletsky, H.J. (1998) Primer3. Code available at www.genome.wi.mit.edu/genome_software/other/primer3.html
27. Burke, T.W. and Kadonaga, J.T. (1997) The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev.*, **11**, 3020–3031.
28. Jones, A.C., Austin, J., Hansen, N., Hoogendoorn, B., Oefner, P.J., Cheadle, J.P. and O'Donovan, M.C. (1999) Optimal temperature selection for mutation detection by denaturing HPLC and comparison to single-stranded conformation polymorphism and heteroduplex analysis. *Clin. Chem.*, **45**, 1133–1140.