2374–2383 *Nucleic Acids Research, 2005, Vol. 33, No. 8*
doi:10.1093/nar/gki531

Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability

Paul M. Harrison*, Deyou Zheng[1], Zhaolei Zhang[3], Nicholas Carriero[2] and Mark Gerstein[1,2]

Department of Biology, McGill University, Stewart Biology Building, 1205 Dr. Penfield Avenue, Montreal, Quebec, Canada H3A 1B1, [1]Department of Molecular Biophysics and Biochemistry, [2]Department of Computer Science, Yale University, New Haven, CT, USA and [3]Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada

Received January 19, 2005; Revised March 14, 2005; Accepted April 4, 2005

ABSTRACT

Pseudogenes, in the case of protein-coding genes, are gene copies that have lost the ability to code for a protein; they are typically identified through annotation of disabled, decayed or incomplete protein-coding sequences. Processed pseudogenes (*PΨg*s) are made through mRNA retrotransposition. There is overwhelming genomic evidence for thousands of human *PΨg*s and also dozens of human processed genes that comprise complete retrotransposed copies of other genes. Here, we survey for an intermediate entity, the transcribed processed pseudogene (*TPΨg*), which is disabled but nonetheless transcribed. *TPΨg*s may affect expression of paralogous genes, as observed in the case of the mouse *makorin1-p1 TPΨg*. To elucidate their role, we identified human *TPΨg*s by mapping expressed sequences onto *PΨg*s and, reciprocally, extracting *TPΨg*s from known mRNAs. We consider only those *PΨg*s that are homologous to either non-mammalian eukaryotic proteins or protein domains of known structure, and require detection of identical coding-sequence disablements in both the expressed and genomic sequences. Oligonucleotide microarray data provide further expression verification. Overall, we find 166–233 *TPΨg*s (~4–6% of *PΨg*s). Proteins/transcripts with the highest numbers of homologous *TPΨg*s generally have many homologous *PΨg*s and are abundantly expressed. *TPΨg*s are significantly over-represented near both the 5′ and 3′ ends of genes; this suggests that *TPΨg*s can be formed through gene–promoter co-option, or intrusion into untranslated regions. However, roughly half of the *TPΨg*s are located away from genes in the intergenic DNA and thus may be co-opting cryptic promoters of undesignated origin. Furthermore, *TPΨg*s are unlike other *PΨg*s and processed genes in the following ways: (i) they do not show a significant tendency to either deposit on or originate from the X chromosome; (ii) only 5% of human *TPΨg*s have potential orthologs in mouse. This latter finding indicates that the vast majority of *TPΨg*s is lineage specific. This is likely linked to well-documented extensive lineage-specific SINE/LINE activity. The list of *TPΨg*s is available at: http://www.biology.mcgill.ca/faculty/harrison/tppg/bppg.tov (or) http:pseudogene.org.

INTRODUCTION

The search for novel functional elements in the human genome is imperative and ongoing (1–3). Pseudogenes (gene copies that have lost their protein-coding ability) are a form of sequence of potential functional utility (4). Substantial progress has been made in the annotation of pseudogenes (5–11). There may be twice as many pseudogenes (derived from protein-coding genes) in the human genome as protein-coding genes (6–10).

Pseudogenes (derived from protein-coding genes) are typically 'diagnosed' through searching for the 'symptoms' of a lack of protein-coding ability. These symptoms include: frame disablement (from premature stop codons and frameshifts), coding sequence decay (typically detectable through examination of non-synonymous and synonymous substitution rates) or incompleteness (either from sequence truncation or from the loss of essential signals for transcription, splicing and translation) (6–10). Processed pseudogenes (*PΨg*s) are made

*To whom correspondence should be addressed. Tel: +1 514 398 6420; Fax: +1 514 398 5069; Email: paul.harrison@mcgill.ca

through retrotransposition of mRNAs. There is ubiquitous genomic evidence for thousands of *PΨg*s in mammals (5–10). Similarly, dozens of processed genes (i.e. genes made by retrotransposition of the complete sequence of other genes) have arisen in both the mouse and human genomes (12,13). This mass gene retrotransposition may arise, at least in part, as a by-product of long interspersed element (LINE) retrotransposition (14). Retrotransposition is clearly an active process in mammalian gene evolution (15). Here, we search for an intermediate type of retrotransposed gene sequence: the transcribed processed pseudogene (shortened as *TPΨg*), which is a *PΨg* that is disabled but nonetheless transcribed.

Historically, there have been several isolated reports of transcribed pseudogenes, of either the duplicated or the processed form (16–21). Two recent studies have demonstrated that such transcribed pseudogenes can regulate transcription of homologous protein-coding genes. Transcription of a pseudogene in *Lymnea stagnalis*, that is homologous to the nitric oxide synthase gene, decreases the expression levels for the gene through formation of a RNA duplex; this is thought to arise via a reverse-complement sequence found at the 5′ end of the pseudogene transcript (20). In a second example, transcription of the *makorin1-p1 TPΨg* in mouse was required for the stability of the mRNA from a homologous gene *makorin1* (21). This regulation was deduced to arise from an element in the 5′ areas of both the gene and the pseudogene (21).

In addition to helping to elucidate such regulatory roles, annotation of *TPΨg*s will further add to our understanding of the dynamics of gene evolution through retrotransposition (15). Also, it is crucial to annotate *TPΨg*s correctly as a part of the ongoing process of correct cDNA/expressed sequence tag (EST) mapping during genome annotation, and for more accurate interpretation of microarray expression data (22,23). Here, we have performed a data-mining expedition for human *TPΨg*s using a rigorous method that applies stringent filters to avoid data pollution. *TPΨg*s have a markedly distinct distribution in the genome when compared with other *PΨg*s and processed genes. A key result is that *TPΨg*s are significantly likely to insert near the 5′ and 3′ ends of genes, implying that *TPΨg*s can be generated by co-option of promoter elements or by intrusion into untranslated regions (UTRs) as 'molecular passengers'. Also, we find that the vast majority of *TPΨg*s are human-lineage specific compared with mouse.

Definitions and terms

An mRNA can be reverse transcribed and re-integrated into the genomic DNA, possibly as a by-product of LINE-1 retrotransposition (14). The parent gene of the mRNA need not be on the same chromosome as the retrotransposed copy. Such a retrotransposed mRNA has three possible fates in the present-day genome: (i) formation of a non-transcribed *PΨg*, (ii) formation of a *TPΨg* or (iii) formation of a processed gene (or part of a gene).

A *PΨg* can be defined as any disrupted, decayed or incomplete copy of a gene that has arisen through such retrotransposition. In the process of evolution, *PΨg*s accumulate disablements (frameshifts and premature stop codons) in their apparent coding sequences. Procedures to annotate *PΨg*s using disablement detection have been described previously (4,5,7), and serve as the basis for the present analysis.

Operationally, a *TPΨg* is defined as a *PΨg* for which an expressed sequence is mappable across any of its coding-sequence disablements, i.e. the disablement occurs in both the expressed sequence and the genomic sequence (see Methods for details).

A processed gene is any undisrupted retrotransposed copy of a gene that also has low K_a/K_s values indicative of selection pressure on coding ability (see Methods for details).

Each of our *TPΨg*s has ≥ 1 disablement verified by alignment of the expressed sequences to genomic DNA, in a region of the *TPΨg* that maps to a known structural protein domain, or to a protein sequence that is conserved in non-mammalian eukaryotes. This three-level verification procedure (genome: transcript:protein) is termed triple alignment. Each verified disablement has an estimated probability of being the result of a sequencing error of $\leq 10^{-6}$, since the error rate for the genomic sequence build is $\leq 10^{-4}$ (24) and the error rate for cDNAs/ESTs is $\leq 10^{-2}$ (25,26). We made a subset of *TPΨg*s, termed the C set, which has further evidence of lack of coding ability. These have: (i) no continuous segment of sequence that can code for a protein domain (as defined in Methods); (ii) high K_a/K_s values (≥ 0.5).

As it is possible that a fraction of the *TPΨg*s that map to introns arise from intron retention in cDNAs or ESTs in the source expressed sequence data, we analyzed all of the data both including and excluding the 67 *TPΨg*s that map to introns (see Table 3 and below). Our results are unaffected by such potential contamination, as explained below.

METHODS

Detection of *TPΨg*s

(i) Mapping expressed sequence data onto existing PΨgs annotations. *PΨg*s were annotated previously using a method based on the detection of disabled protein homology in genomic DNA (4,5,7). We mapped >6200 of these onto human genome build 34 (from http://www.ensembl.org), through detection of 100% nucleotide sequence matches, removing overlap with coding exons. For each *PΨg*, the genomic sequence was extracted, both with and without a 6000 nt extension added on to either end to allow for homology matching to 'pseudo-UTR' regions. (These sets of genomic DNA are named *genPΨg* and $genP\Psi g_{+/-6000}$.) Three sources of expressed sequences (Refseq mRNAs, Unigene consensuses, and ESTs from dbEST) were downloaded from http://www.ncbi.nih.gov. They were mapped onto *genPΨg* and $genP\Psi g_{+/-6000}$, using BLASTN with low-complexity masking (*E*-value $\leq 10^{-10}$, minimum match length 100 nt) (27,28). From the resulting significant matches, those that align with $\geq 95\%$ identity were used to generate a second BLASTN search against *genPΨg* and $genP\Psi g_{+/-6000}$, but this time without low-complexity masking, to insure correct sequence identity. Matches to both *genPΨg* and $genP\Psi g_{+/-6000}$ with $\geq 99\%$ identity over >0.998 of the length of the expressed sequence were then extracted. These expressed sequence matches were filtered to insure that they match more significantly to the *PΨg* than to any homologous gene. The matching expressed sequences were then

re-aligned to the *PΨg* sequence using FASTY (29), to check that ⩾1 disablement (frameshift or premature stop codon) in the *PΨg* occurs in both the genomic sequence and expressed sequence. Each disablement verified in this way has an estimated probability of being the result of a sequencing error of $\leqslant 1 \times 10^{-6}$; this is because the genomic sequence error rate is $\leqslant 1 \times 10^{-4}$, and the cDNA/EST sequencing error rate is $\leqslant 1 \times 10^{-2}$.

(ii) Extraction of PΨgs that are in Refseq mRNAs. All human Refseq entries corresponding to known mRNAs (total = 20 741) were compiled from data downloaded from the NCBI website (http://www.ncbi.nih.gov). These were compared with all known, non-fragmentary human proteins in the SWISSPROT database (30), using a modification of the disabled protein homology-based procedure developed previously for *PΨg* annotation (4,5,7,31–33). To insure that all of the candidate *TPΨg*s in Refseq mRNAs map to a single continuous piece of genomic DNA, we extracted the appropriate mRNA subsequences and mapped them to the human genome using BLASTN. Those segments that matched over their complete length exactly were retained. The resulting *TPΨg* data were then filtered along with those generated in (i), as detailed in (iii) below.

(iii) Filtering the (transcribed) PΨg data. We applied a set of filters to insure that we were compiling a bona fide list of *TPΨg*s. All *TPΨg* data sets were filtered as follows:

(a) *Removal of homologies to purely hypothetical proteins or fragmentary proteins: TPΨg*s based only on homology to predicted reading frames or reading-frame fragments were removed through BLASTP comparisons (*E*-value $\leqslant 10^{-4}$) against a library of hypothetical or fragmentary proteins from SWISSPROT (30). These are removed because their disablements may be erroneous (which is inappropriate for the method employed here). Also, they may be inaccurately dated (values for K_s, K_a, etc., may be incorrect).

(b) *Verification that the disablements are in conserved parts of a known protein sequence or domain:* We verified that the disablements examined are in known conserved parts of sequences, as detailed below. This list of filters has an 'if-else-if-else-if' structure:
   (1) First, we assigned protein structural domains to the *TPΨg*s, by comparing them with the ASTRALSCOP 95% identity set of protein domains (34), using BLASTP (27) (*E*-value $\leqslant 10^{-4}$). The total assigned *TPΨg* subsequence was determined (from the most N-terminal residue that was assigned to a domain, to the most C-terminal). This assigned subsequence was considered disabled, if a frameshift or stop codon occurred >10 residues in from either terminus. This accounts for ∼54% of *TPΨg*s.
   (2) Otherwise, secondly, *TPΨg*s not meeting criterion (1) were checked manually for occurrence of disablements in conserved domains using the InterPro (http://www.ebi.ac.uk/interpro) and CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) domain annotation tools.
   (3) Otherwise, thirdly, *TPΨg*s not meeting criterion (2) (<10% of the sequences) were checked for disablement within a part of the sequence that is conserved in other mammals, and in ⩾1 non-mammalian species, using BLASTP (*E*-value $\leqslant 10^{-4}$).

(c) *Removal of candidates with small introns:* Putative *PΨg*s, *TPΨg*s, and the expressed sequences that match them, were filtered for small intron sequences. A library of introns of <1000 nt was made from genes on human genome build 34. TBLASTN (27) was used to annotate any significant matches to these introns of >0.80 of their length (*E*-value $\leqslant 10^{-4}$). Any such matches could either be from introns in an aberrant cDNA or be a previously disregarded small intron in the genome. Some additional examples of *TPΨg*s that map to introns may arise from intron retention in cDNAs or ESTs; however, the main points of the analysis reported in this paper are unaffected by such potential contamination, as explained below.

(d) *Removal of possible duplications of single-exon genes and large exons*: We wished to insure that there were no single-exon gene duplications in our *PΨg* and *TPΨg* sets. To do this, all *PΨg*s and *TPΨg*s were compared using BLASTP (*E*-value $\leqslant 10^{-4}$), to the set of proteins for build 34 (removing those whose genes overlap putative *PΨg*s) (27). All *PΨg*s and *TPΨg*s that had closest-matching homologies to single-exon proteins were removed. Furthermore, we insured that they aligned to their closest-matching homologous human proteins around at least one 'exon seam', i.e. a position in a protein sequence that corresponds to an intron–exon boundary. This exon seam filter insures that the pseudogenes considered are processed, and is particularly useful for removing homologies to genes with large exons (e.g. some zinc-finger-containing proteins).

(e) *Filtering for processed genes:* All *TPΨg*s were filtered for overlap with annotated processed genes (resulting in the removal of only one putative *TPΨg*).

After applying these rigorous filters, we had 3418 *PΨg*s (both transcribed and non-transcribed), and 233 *TPΨg*s, 218 from mapping expressed sequences to *PΨg*s and 15 from *PΨg* extraction from Refseq mRNAs. Almost half (97/233, 42%) of the *TPΨg* set represent 100% exact matches of expressed sequences to *PΨg*s. Restricting analysis to just these matches does not affect any of the main trends and results reported here.

Making an obviously decayed C set of *TPΨg* sequences

We derived a 'core set' of *TPΨg*s that have further evidence of coding-sequence decay. These are dubbed the C set (totaling 177/233, 76%). This set is the union of the following two subsets: (i) *TPΨg*s without continuous segment of sequence that can code for a protein domain (106/233 *TPΨg*s, 45%) or (ii) *TPΨg*s with high K_a/K_s values (>0.50) indicative of lack of coding ability (127/233 *TPΨg*s, 54%).

(i) Lack of protein domain coding ability. We parsed each *TPΨg* into subsequences according to the positions of its disablements. If all subsequences could be labeled as 'unlikely to code for a protein domain', then the *TPΨg* was included in the C set. This resulted in inclusion of 106 *TPΨg*s in the C set. We labeled a subsequence as 'unlikely to code for a protein domain' if:

(a) Its length was ≤32 residues. The vast majority (95%) of non-cysteine-rich protein domains in the ASTRALSCOP 40% identity set have sequence lengths >32 residues (34). Cysteine-rich domains (which are likely disulfide-bridged or metal-chelating) are defined as having cysteine concentration <0.077/residue, a value suggested by a bimodality in cysteine concentration, in surveys of cysteine and cystine occurrence in proteins (35,36). Condition (a) was not applied to any fragments that were adjudged cysteine-rich.
(b) It contained a disrupted SCOP domain, as defined in part (iii)(b)(1) above. Such fragments are likely not to constitute a large enough fragment; the reasoning behind this criterion is that evolution has defined and refined the integrity of a body of recurrent folding units (protein domains) (34), and we can therefore use their disruption to evaluate whether a piece of sequence is no longer protein-coding.

(ii) K_a/K_s analysis. We calculated the K_a/K_s values for whole *TPΨg*s, using the Yang and Nielsen method in PAML (37), using the present-day gene sequence to compare against the pseudogene, as described previously (7). Also, similarly, we calculated K_a/K_s values for subsequences of *TPΨg*s (≥50 residues) derived by parsing at disablement positions. This parsing allows for the possibility that some of the pseudogene subsequences have coding ability, while others do not, i.e. we can test for a coding ability 'imbalance'. From these K_a/K_s calculations, we found that only ∼4% of both *PΨg*s and *TPΨg*s have two adjacent regions where one is <0.25 (potentially coding) and the other >0.5 (potentially non-coding), indicating that such imbalance is rare. From consulting independent analysis of populations of human genes and *PΨg*s (6), we ascertained that for a threshold value of $K_a/K_s \geq 0.5$, >95% of sequences are predicted to be *PΨg*s and not genes. We use this as the expectation for the distribution of *PΨg*s in general. Calculation of K_a/K_s values for gene/pseudogene pairs errs on the side of under-estimation of coding-sequence decay (7).

Conservation of *TPΨg* in mouse

For each human *TPΨg*, we searched against potentially orthologous mouse *TPΨg*s. These 'mo*TPΨg*s' were derived by mapping expressed sequences (Refseq mRNAs, Unigene consensus sequences and ESTs) for mouse onto a previously derived set of mouse *PΨg*s (8), in a similar manner to the human mappings (see above). These were pooled with any existing mo*TPΨg* annotations, and a small number of mouse genes that might be potentially misannotated mo*TPΨg*s. A potentially orthologous mo*TPΨg* was required to match ≥0.5 of the length of the human *TPΨg* (for BLASTP matches, *E*-value ≤ 10^{-4}), and to share the same closest-matching human protein with any potential human *TPΨg* homologs. We did not require that the retrotranspositions be in syntenic positions, since orthologous gene retrotranspositions are not necessarily syntenic (38).

Processed genes

We mapped an independently derived list of processed genes (13) to human genome build 34. In addition to the criteria in (13), we required K_a/K_s values <0.25, and coverage of ≥0.95 of the parent gene's length. Any examples that overlap the *TPΨg* data set of annotations were removed; vice versa, any *TPΨg*s that have $K_a/K_s < 0.25$ and cover ≥0.95 of their parent gene were deleted from the *TPΨg*s list. Our definitions give two distinct sets of processed genes and *TPΨg*s; naturally, we miss some sequences that cannot be classified as either a *TPΨg* or a processed gene.

RESULTS AND DISCUSSION

Number of *TPΨg*s

In total, we found 233 human *TPΨg*s (Table 1). These *TPΨg*s form a subset of 3418 previous *PΨg* annotations that were mapped to build 34 of the human genome (7). These *PΨg*s were filtered in the same way as the *TPΨg*s (from a starting total of ∼6200), to remove predicted reading frames, retained introns and potential duplications of single-exon genes or large exons. Using these data, we can estimate that ∼6% (218/3418) of *PΨg*s are *TPΨg*s. An additional 15 *TPΨg*s were derived from a reciprocal process of searching for *PΨg*s in known Refseq mRNAs, followed by subsequent mapping to the genome.

A small fraction of the *TPΨg*s (8%) corresponds to known Refseq mRNAs (Table 1). About a third are supported by Unigene consensus sequences, with a large fraction (71%) matching individual ESTs [of this last group, about a quarter (∼23%) are supported by a Refseq mRNA or a Unigene consensus; Table 1]. We sought additional expression verification

Table 1. Summary of numbers of *TPΨg*s

Set or subset of *TPΨg*s	Total number	Total number (without those mapped to introns)
Mappings to existing pseudogene annotations	218	154
Pseudogene extraction from Refseq mRNAs	15	12
Total *TPΨg*s	233	166
Expressed sequence support		
*TPΨg*s that are supported by Refseq mRNAs	18 (8%)	16 (10%)
*TPΨg*s that are supported by Unigene consensus sequences	74 (32%)	50 (30%)
*TPΨg*s that are supported by dbEST expressed sequence tags	167 (72%)	111 (67%)
*TPΨg*s that are supported by dbEST expressed sequence tags and by either a Refseq mRNA or a Unigene consensus	38 (16%)	25 (16%)
*TPΨg*s that are additionally supported by oligonucleotide microarray data	75 (32%)	53 (32%)
Further evidence of decay		
*TPΨg*s that have no continuous segment likely to code for a protein domain	106 (45%)	70 (42%)
*TPΨg*s that have $K_a/K_s \geq 0.5$	127 (54%)	88 (53%)
C set (*TPΨg*s that have no continuous segment likely to code for a protein domain or $K_a/K_s \geq 0.5$)	177 (76%)	123 (74%)

from a series of high-density oligonucleotide microarrays, composed of ~52 million 36mers (23). These microarrays were applied to probe the transcriptionally active regions of the human genome, in a strand-sensitive way. Using the same data and statistical method (i.e. a sign test) for scoring the genes' transcriptional activity (22,23), we found that 75/233 (32%) *TPΨg*s were transcriptionally active in liver ($P < 0.05$) (Table 1). In comparison, 64% of genes from RefSeq mRNAs, 57% of Ensembl annotated genes (39), and 35% of genes predicted with the program GENSCAN (40), were found to be transcribed in liver.

The C set of more obviously decayed *TPΨg*s comprises 76% of the total population; 45% of *TPΨg*s having no continuous segment likely to code for a protein domain, and 54% of *TPΨg*s having $K_a/K_s \geq 0.5$ (Table 1). Obvious degradation of the coding sequences is demonstrated for this set from analysis of protein-domain mapping and K_a/K_s (see Methods). Additionally, other factors (not examined in the present analysis) are expected to cause lack of coding ability in *TPΨg*s or arise as further consequences. It is likely that *TPΨg*s will not have appropriate start codon context (41), therefore leading to little or no efficient translation initiation. Also, those *TPΨg*s that are inserted into 3′-UTRs of mRNAs will be unlikely to become protein-coding through being downstream of a clearly defined coding sequence (although it is conceivable that they may be translatable in the 5′-UTR). Furthermore, a consequence of any frameshift in a sequence is the likelihood of an additional 20 residues or so of non-coding DNA, added onto the end of the sequence truncation (on average, in randomly picked, conceptually translated intergenic DNA, a stop codon will appear ~20 residues downstream of any starting point); such additional sequence may lead to aggregation or misfolding in the cell.

The proportions of *TPΨg*s break down in a similar fashion to that just described above for the total data set, when the 67 examples that map to introns are removed (Table 1).

Closest matching human proteins for *TPΨg*s

*TPΨg*s were grouped according to their closest-matching human protein (Table 2). Each table entry represents a single 'parent gene'. The total counts are also shown for the *TPΨg*s that do not map to introns (in square brackets, Table 2). There are 4 human proteins that have ≥4 homologous *TPΨg*s. The highest number of *TPΨg*s (5) occur for cyclophilin A, which is required for *cis*-peptide isomerization (42). All of these proteins arise from highly expressed mRNAs. They also occur in the top 20 proteins when apportioning all *PΨg*s, in the same way (7).

Table 2. Human proteins with four or more homologous *TPΨg*s

Number[a]	Name of human protein[b]
5 [4]	Peptidyl-prolyl *cis-trans* isomerase A (Cyclophilin A) [P62937]
4 [3]	Prohibitin [P35232]
4 [3]	40S ribosomal protein S12 [P25398]
4 [3]	Actin, cytoplasmic 2 (Gamma-actin) [P63261]
	Glyceraldehyde-3-phosphate dehydrogenase [P04406, P00354]

[a]The totals in square brackets are for when those mapping to introns are removed.
[b]The Swissprot accession numbers are given in square brackets.

TPΨg position relative to genes and the implications for their expression mechanisms

A number of mechanisms for *TPΨg* expression are plausible. First, *TPΨg*s may co-opt nearby promoter elements of protein-coding genes. Secondly, they may intrude into the UTRs of another mRNA, as a sort of 'molecular passenger'. Thirdly, they may make use of cryptic promoter elements in the intergenic DNA; such promoter elements may have originated from transposable elements, or from genomic duplication of genic promoter regions, or sporadically (*de novo*).

Such mechanisms for *TPΨg* expression may have a bearing on their overall positional distribution in the genome relative to genes. To investigate this, we classified the *TPΨg*s into those that: (i) overlap existing coding-sequence exons; (ii) appear inserted in introns; (iii) are inserted in a 3000 or 10 000 nt region 5′ to annotated genes; (iv) are inserted in a 3000 or 10 000 nt region 3′ to annotated genes.

Table 3 summarizes these data. A minor proportion (8%) of *TPΨg*s entail gene coding-sequence annotations, i.e. they are erroneously annotated reading frames. There are 67 *TPΨg*s that map to introns (Table 3); it is unclear how many of these may arise from intron retention in cDNAs or ESTs. Expectations based on random insertion in the genome were calculated for classes (ii) to (iv). We focus on (iii) and (iv) in particular.

*TPΨg*s are significantly more likely than random ($P < 0.01$, chi-squared tests) to be inserted in the regions 5′ and 3′ of annotated genes; this effect is most obvious in the 3000 nt regions 5′ and 3′ to genes, but is still significant up to 10 000 nt in either direction (Table 3). Similar results are observed for the C set of more obviously decayed *TPΨg*s. The enrichment of *TPΨg*s observed in the 5′ and 3′ areas of genes can be seen as a simple logical consequence of randomly inserted *PΨg*s having an increased probability of being transcribed, and is clear support for either co-option of genic promoter elements, or insertion into UTRs as molecular passengers, leading to *TPΨg* expression. This result is also unaffected by possible contamination from intron retention in cDNAs/ESTs, as, in general, *PΨg*s are significantly under-represented in introns (Table 3); if one assumed that, in the extreme, all of the *TPΨg* mappings near the 5′ and 3′ ends of genes were actually mappings to introns, then this would make their over-representation even more significant. The general dearth of *PΨg*s in introns may be a reflection of an overall genomic tendency for a lack of retroelement insertion in introns (43).

Roughly half of the *TPΨg*s are located away from genes (>10 000 nt 5′ and 3′ to genes, and overlapping neither an exon nor an intron; Table 3). These thus may be co-opting cryptic promoters of unknown origin in the intergenic DNA, such as those derivable from transposable elements.

In summary, the distribution of *TPΨg*s in the vicinity of genes is significantly different from that observable for other non-transcribed *PΨg*s (that have no transcription evidence), and for processed genes in the following ways (Table 3):

(i) *TPΨg*s are significantly over-represented in the 10 000 nt 5′ and 3′ to genes, whereas other *PΨg*s and processed genes are not;
(ii) Other *PΨg*s are significantly over-represented in intergenic DNA and significantly under-represented in introns, and processed genes are significantly under-represented in introns; *TPΨg*s now show such trends for introns or

Table 3. Position of *TPΨg*s, other *PΨg*s and processed genes relative to annotated genes

Categories of sequence grouped by position relative to genes	Type of sequence					
	*TPΨg*s		Other *PΨg*s		Processed genes	
	Observed number[a]	Expected number[b]	Observed number[a]	Expected number[b]	Observed number[a]	Expected number[b]
Sequences that overlap gene annotations	18 (8%)	—	—	—	—	—
Sequences mapped to introns of annotated genes	*67 (28%)*	*79.7*	*693 (22%)*	*1100.0*¶¶	*3 (5%)*	*21.2*¶¶
Sequences <3000 nt 5′ of start codon of annotated genes	20 (9%)	6.8**	78 (0.7%)	93.6	5 (8%)	1.9
Sequences <10 000 nt 5′ of start codon of annotated genes	36 (15%)	22.3*	278 (9%)	307.8	7 (11%)	5.9
Sequences <3000 nt 3′ of translation stop of annotated genes	22 (9%)	6.7**	55 (1.7%)	92.3¶¶	0 (0%)	1.8
Sequences <10 000 nt 3′ of translation stop of annotated genes	42 (18%)	22.2**	241 (7%)	306.4¶¶	9 (14%)	5.8
Sequences that are in intergenic DNA[c]	109 (47%)	129.8	2109	1371.7**	43	31.4

[a]These categories are not additive, as they are not mutually exclusive, i.e. some *TPΨg* may be within 10 000 nt of the 5′ end of one gene, and be in the intron of another gene or within 10 000 of the 3′ end of a third gene.

[b]Expected values are calculated assuming random insertion in the whole genome (without the genomic DNA for annotated genes). For significant over-representation, ** indicates $P < 0.001$, and * indicates $P < 0.01$ for a chi-squared test (1 degree of freedom) using Yates correction (similarly, ¶¶ is used for significant under-representation for $P < 0.01$).

[c]Intergenic DNA is defined as all of the genomic DNA that does not comprise exons, introns or the regions of genes within 10 000 nt of the translation stop and start of gene coding sequences.

intergenic DNA. In addition, there is a dearth of *PΨg*s 3′ to genes (Table 3). The reasons for this are unclear; there may be a compositional effect, similar to the relationship between genomic G+C content and ribosomal-protein *PΨg*s insertion, observed previously (44).

We examined the distribution of 20 *TPΨg*s that are directly mappable onto known Refseq mRNAs. Thirteen of these overlap an erroneously predicted open reading frame, and two are already annotated as transcribed pseudogenes. None of the five remaining *TPΨg*s are inserted in the 5′-UTR of a messenger RNA. One explanation for this absence in 5′-UTRs is that a *TPΨg* would introduce upstream ORFs that interfere with translation initiation (41). The five *TPΨg*s inserted in the 3′-UTRs of mRNAs are all in the forward direction (i.e. they are all on the same DNA strand as the annotated coding sequence). An example of this is discussed below. In addition, we checked the list of *TPΨg*s 3′ to annotated genes and within 3000 nt of the end of the coding sequence (Table 3), for additional examples of this 'passenger' phenomenon, through manual examination of cDNAs or ESTs for the 5′ genes, but could find no further examples of cDNAs with polyadenylation signals to define the end of the mRNA. Such analysis is complicated by the fact that, in some cases, it may not be possible to distinguish between the original polyadenylation signal of the gene, and an inserted polyadenylation signal arising from the *TPΨg*.

Distribution on chromosomes

Analysis of the distribution of processed genes in the human and mouse genome has indicated that the X chromosome is a marked outlier, both for processed gene deposition onto the X chromosome and origination from X (13). A similar outlier preference was observed for *PΨg* deposition onto the X chromosome (but not origination from X) (13). These phenomena may be due to selection pressures to compensate for X-chromosome inactivation during spermatogenesis, in combination with some unaccounted-for mutational biases (13,38).

To compare with this previous analysis, we examined the distribution of *TPΨg* 'parent genes' on each chromosome, and also the distribution of the number of *TPΨg*s per chromosome (Figure 1A and B). Figure 1A indicates the data for origination of *TPΨg*s, and Figure 1B shows the trend for deposition of *TPΨg*s onto each chromosome. In each case (origination and deposition), the X chromosome is not an outlier. This may indicate that, in general, *TPΨg* formation is deleterious, unlike processed gene and non-transcribed *PΨg* formation, which are arguably, by comparison, beneficial and selectively neutral, respectively. Interestingly, there is some outlier behavior for *TPΨg* origination from chromosome 12. The same result is obtained, if the 67 *TPΨg*s that map to introns are removed.

Search for potential orthologs in mouse

We investigated mouse/human cross-species conservation of *TPΨg*s, as an indicator of human-lineage specificity. The 233 human *TPΨg*s were compared against a set of 215 putative mouse *TPΨg*s (mo*TPΨg*s) (see Methods for details). We found that 5% (11/233) have potential orthologous *TPΨg*s. Four of these are for the metabolic enzyme, glyceraldehyde-3-phosphate dehydrogenase, which is ubiquitously and highly expressed, giving this sequence the status of a notable 'parent gene' for *TPΨgs* (see also Table 2). If the human and mouse *TPΨg*s are not restricted to having the same closest-matching human gene homolog, 28/237 (12%) have potential orthologs.

These results suggest that a minor fraction of *TPΨg*s could be used in conserved functional roles in mammals. However, given that ~40% of human *PΨg*s are conserved in the mouse genome (8), these results imply that *TPΨg*s are significantly under-conserved between human and mouse ($P < 0.001$ using binomial statistics) compared with *PΨg*s in general, and also compared with processed genes (13), which are at most ~20% lineage-specific. The vast majority of *TPΨg*s are thus human lineage-specific compared with mouse; indeed, both *Alu*s (which are primate-specific) and *PΨg*s can be made as by-products of LINE retrotransposition (14), and have similar overall age profiles in the genome (8). These results are also evidence for a general evolutionary selection pressure to delete *TPΨg*s. This may be because they form a source of transcriptional interference for adjacent genes or homologous genes. However, one must stress that, in the future, increased cDNA coverage for both the mouse and human

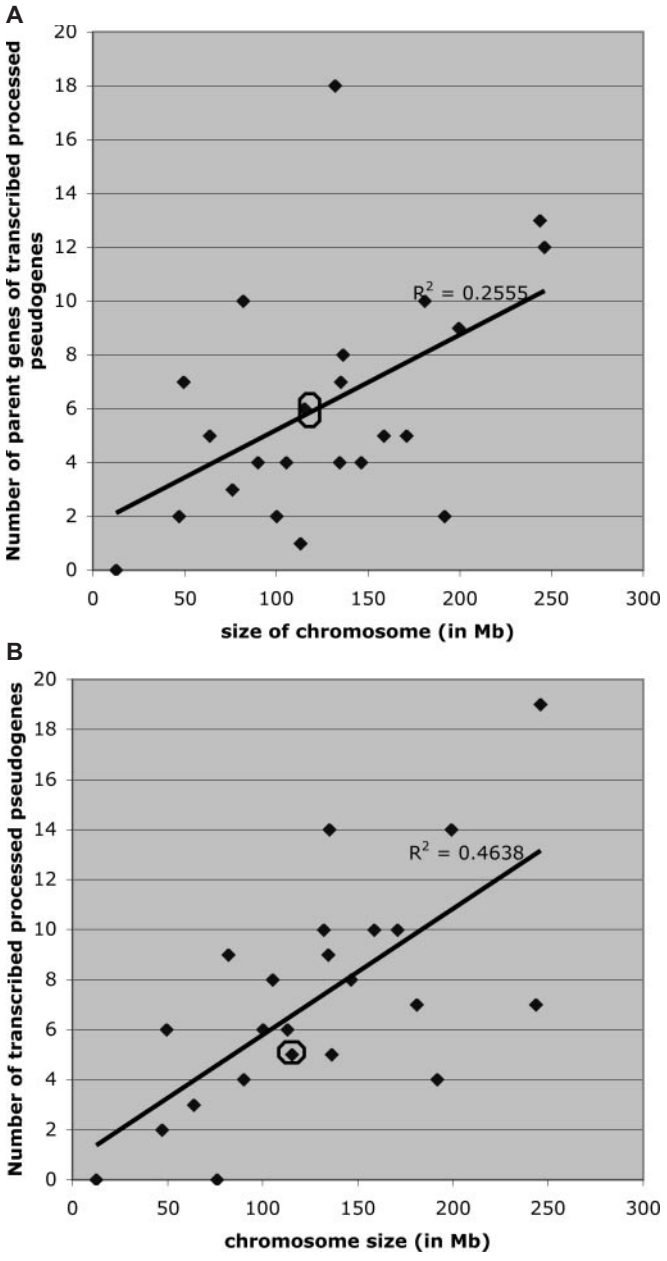

Figure 1. Origination and deposition of *TPΨg*s for different chromosomes. (**A**) Origination of *TPΨg*s: this plot shows the number of parent genes of *TPΨg*s in a chromosome versus the chromosome size (in Mb). (**B**) Deposition of *TPΨg*s: this shows the number of *TPΨg*s per chromosome versus chromosome size (in Mb). Only retrotranspositions from one chromosome to another are considered in each plot. The X chromosome is ringed. Note that for each plot we have corrected for the probability of X and Y chromosome inclusion in gametes [i.e. the size of X is multiplied by 0.75 and Y by 0.25; for comparison see figure 1 in (13)].

genomes may modify these statistics somewhat. Such a lack of saturation in current databases of expressed sequences can be demonstrated using some simple sampling analysis. Sampling of *TPΨg*-matching expressed sequences from random fractional subsets of the total expressed sequence database used in the present analysis (i.e. ESTs + Unigene consensuses + Refseq mRNAs), indicates that we are not near finding all of the *TPΨg*s in the human genome (or, at least, those discoverable through mapping of expressed sequences). (This sampling analysis is presented in Supplementary Figure 1.)

Examples of *TPΨg*s

A *TPΨg* derived from the prohibitin gene is shown in Figure 2A. A prohibitin *TPΨg* is inserted into the 3′-UTR of a Zn-finger-containing protein. Prohibitin is highly and ubiquitously expressed, and is involved in inhibition of DNA synthesis; its mRNA contains a putative functional RNA element in its own 3′ UTR (45). It is beyond the scope of this present study to ascertain whether this RNA element, in this *TPΨg*, is intact, as it has not yet been characterized extensively by mutational and biophysical analysis. This *TPΨg* is one of four that derived from the prohibitin gene (Table 2).

The second example is derived from the precursor sequence of mitochondrial 2-amino-3-ketobutyrate coenzyme A. The crystal structure of the *Escherichia coli* homolog of this enzyme is known (PDB code *1fc4*a). We have indicated how the 'triple alignment' of genomic sequence, EST and known protein-domain sequence overlap (Figure 2B). The protein chain is divided into colored segments, with each disablement defining a segment boundary. One can clearly see that the triple alignment covers two disablements in the *TPΨ*g.

CONCLUSIONS

Diverse efforts to map novel elements of potential functional utility in our genome are ongoing (1–3). In the spirit of such endeavors, we have derived a rigorous procedure for annotating a specific novel type of element of potential functional utility, the *TPΨg*. Applying this method to the human genome, we discovered 166–233 *TPΨg*s, which represent ∼4–6% of all *PΨg*s (the lower total arises from setting aside any examples that map to introns). One should point out that we might have missed some *TPΨg*s; e.g. those without extensive homology to a coding sequence (i.e. those consisting largely of UTR homologies), or *TPΨg*s formed from single-exon and large-exon genes, or *TPΨg*s that are transcribed in a low-level beyond detectability through EST/cDNA sequencing.

*TPΨg*s are significantly more likely in regions close to the 5′ and 3′ ends of genes, compared with both a random insertion model for them throughout the genome, and compared with the distribution observed in general for *PΨg*s. Furthermore, if one assumes that these 5′ and 3′ regions are actually introns, the significance of the increased 5′ and 3′ density of *TPΨg*s improves. (This indicates that the increased 5′ and 3′ density is not an artifact of intron retention in cDNA/EST libraries.) This increased density provides evidence that *TPΨg*s may be expressed through co-option of genic promoter elements or through insertion into UTRs as 'molecular passengers'. Specific detailed evidence was found for molecular passengers in the 3′-UTRs of known mRNAs; an example of this derived from the prohibitin gene was illustrated (Figure 2A). *TPΨg*s could thus also have a role as intermediates in protein-coding sequence evolution. A reasonable hypothesis that can be further investigated is that, *TPΨg*s may represent a source of evolutionary protein novelty, either as 'molecular passengers',

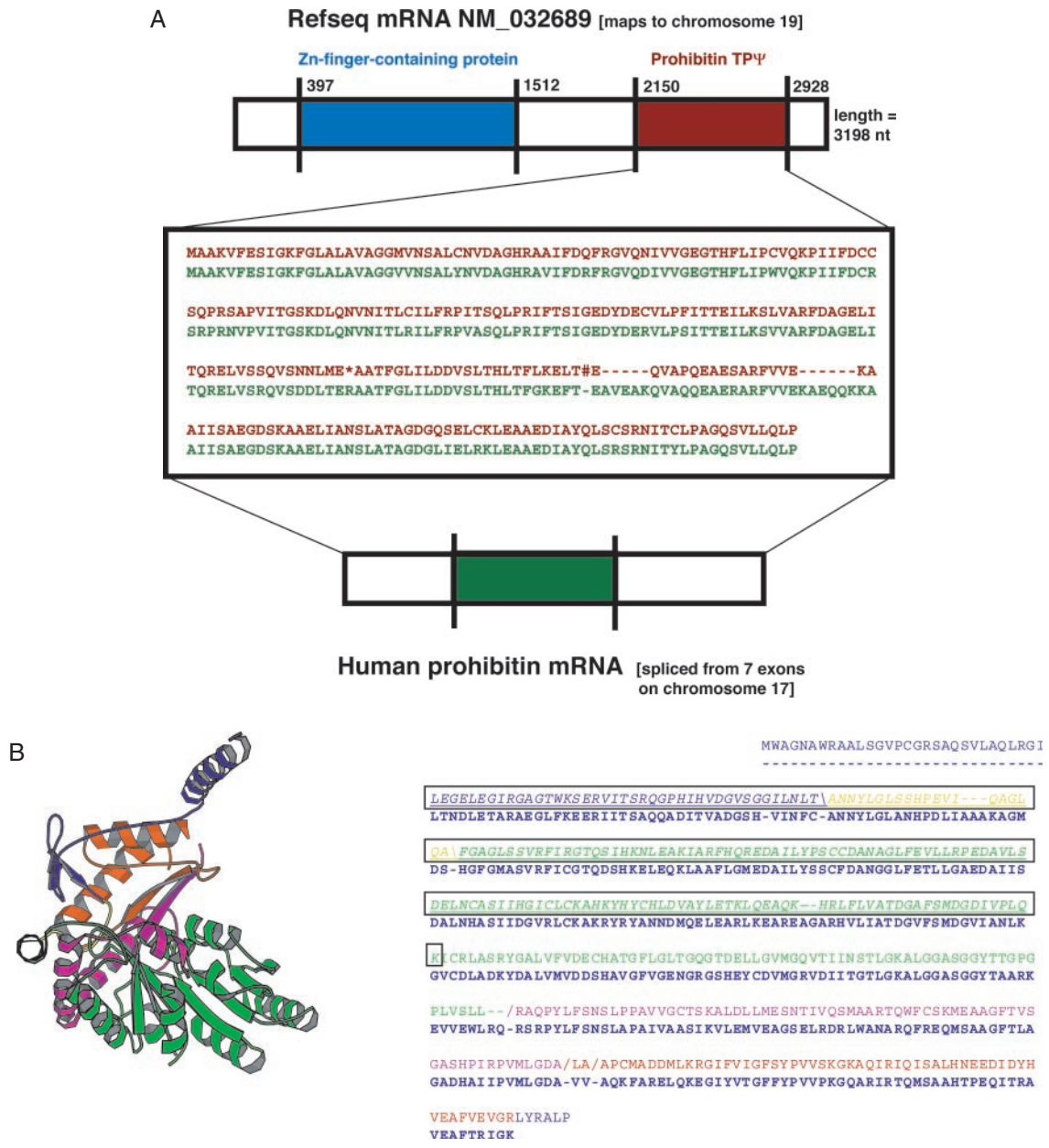

Figure 2. Examples of *TPΨg*s. (**A**) This is a *TPΨg* derived from the human prohibitin gene. The prohibitin gene contains both a protein-coding region and an RNA in its 3′-UTR (45), but only the segment of the *TPΨg* corresponding to the protein-coding sequence is shown. In the center is an alignment of the *TPΨg* (in red) with prohibitin protein (in green). The graphic above it shows the position of the *TPΨg* (red segment) in the 3′-UTR of an mRNA that codes for a Zn-finger-containing protein (blue segment). **(B)** An example of a *TPΨg* that maps to a known globular protein domain. The *TPΨg* derives from the mRNA for the precursor sequence of mitochondrial 2-amino-3-ketobutyrate coenzyme A. The domain is from the closest-matching protein structure (from *E.coli*, PDB code *1fc4*a). In the Molscript (54) picture, the protein chain trace color changes at the position of each disablement. The alignment of the *E.coli* domain sequence and the human *TPΨg* sequence is shown. The part of the sequence that maps to an EST (gi | 6138420) is boxed and italicized.

or as part of alternative splicings (46), through being temporarily released from coding-sequence selection pressures (31,47–50). Use of additional sequence segments may underlie the influence of the *[PSI+]* prion on phenotypic variability in budding yeast (31,51); analogs of this phenomenon are possible in mammals.

Two examples of regulation by transcribed pseudogenes of homologous genic transcripts have been observed (20,21). Transcriptional analysis showed that the stability of the *makorin1* mRNA in mouse relies upon the expression of its homologous *makorin1-p1 TPΨg*, through the action of an element at the 5′ end of the *makorin1-p1* sequence. However, *makorin1-p1* only seems to be conserved in one line of *Mus*, and has not been found in the rat genome (52). In a second example, transcription of a pseudogene in *Lymnea stagnalis*, that is homologous to the nitric oxide synthase gene, decreases expression levels for the gene; this is thought to arise via a reverse-complement sequence found at the 5′ end of the pseudogene transcript (20). Alternatively, *TPΨg*s near genes or in UTRs may also exert a controlling/interfering influence

on the genes' transcription and translation, through upstream ORF formation, or the action of other undiscovered elements. Such *TPΨg*s could exert such effects through co-option as alternative splicings, as has been observed for *Alu*s (53). Also, it is possible that some *TPΨg*s produce a short peptide that does not misfold or aggregate in the cell, but is still targeted and serves an alternative function as a truncated peptide. Certainly, *TPΨg*s represent a source of transcriptional 'noise', which may have implications for selection pressures on transcription levels, and the degree of variation on which such pressures can act.

Our survey provides evidence for the existence in the human genome of a small population of *TPΨg*s, which are an intermediate class of retrosequence derived from genes, since they have expression evidence (like genes), but also have evidence of lack of coding ability (like other pseudogenes). The distribution of *TPΨg*s near the 5′ and 3′ ends of genes indicates that *TPΨg*s can co-opt genic promoters or intrude into UTRs; furthermore, this is a robust observation that verifies our expression-data mappings. One must also point out, however, that about half of the *TPΨg*s are located away from genes in intergenic DNA (Table 3), and thus may be co-opting cryptic promoters of undesignated origin. Also, *TPΨg*s differ from other *PΨg*s (without transcription evidence) and from processed genes in terms of their distribution per chromosome, and their projected conservation in mouse. Our analysis indicates that, unlike processed genes and other *PΨg*s, the vast majority (∼95%) of *TPΨg*s are human lineage-specific. In combination, the chromosomal distribution and mouse conservation for *TPΨg*s suggests that there is some general evolutionary pressure to delete *TPΨg*s from the genome. One should point out that the cDNA coverage of both genomes is far from complete (as illustrated here, with some simple sampling analysis), so that the analysis of conservation in mouse should be regarded as tentative.

This *TPΨg* analysis has important implications for genome annotation. It is still common practice to assume that an mRNA contains one undisrupted open reading frame; however, it is clear that one should routinely check for *TPΨg*s in the manner described here. Also, this *TPΨg* annotation is useful for improved interpretation of microarray expression data (22,23). The list of *TPΨg*s is available at: http://www.biology.mcgill.ca/faculty/harrison/tppg/tppg.tar (or) http://pseudogene.org.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

Thanks to T. Bureau, N. Juretic and D. Hoen (McGill U.) for discussions. This work was supported in part by a Discovery Grant from the National Science and Engineering Council of Canada to P.M.H., and by National Institutes of Health grant # P50 HG02357-01 to M.G. Funding to pay the Open Access publication charges for this article was provided by McGill University.

Conflict of interest statement. None declared.

REFERENCES

1. Dermitzakis,E.T., Reymond,A., Scamuffa,N., Ucla,C., Kirkness,E., Rossier,C. and Antonarakis,S.E. (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*, **302**, 1033–1035.
2. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
3. Consortium,E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
4. Harrison,P. and Gerstein,M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.*, **318**, 1155–1174.
5. Harrison,P.M., Hegyi,H., Balasubramanian,S., Luscombe,N.M., Bertone,P., Echols,N., Johnson,T. and Gerstein,M. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.*, **12**, 272–280.
6. Torrents,D., Suyama,M., Zdobnov,E. and Bork,P. (2003) A genome-wide survey of human pseudogenes. *Genome Res.*, **13**, 2559–2567.
7. Zhang,Z., Harrison,P., Liu,Y. and Gerstein,M. (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.*, **13**, 2541–2558.
8. Zhang,Z., Carriero,N. and Gerstein,M. (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.*, **20**, 62–67.
9. Consortium,M.G.S. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
10. Consortium,R.G.S.P. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
11. Chen,C., Gentles,A.J., Jurka,J. and Karlin,S. (2002) Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 2930–2935.
12. 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.
13. Emerson,J.J., Kaessmann,H., Betran,E. and Long,M. (2004) Extensive gene traffic on the mammalian X chromosom. *Science*, **303**, 537–540.
14. Esnault,C., Maestre,J. and Heidmann,T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.*, **24**, 363–367.
15. Brosius,J. (1999) Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica*, **107**, 209–238.
16. Fujii,G.H., Morimoto,A.M., Berson,A.E. and Bolen,J.B. (1999) Transcriptional analysis of the PTEN/MMAC1 pseudogene, psiPTEN. *Oncogene*, **18**, 1765–1769.
17. Bristow,J., Gitelman,S.E., Tee,M.K., Staels,B. and Miller,W.L. (1993) Abundant adrenal-specific transcription of the human P450c21A 'pseudogene'. *J. Biol. Chem.*, **268**, 12919–12924.
18. Zhou,B.S., Beidler,D.R. and Cheng,Y.C. (1992) Identification of antisense RNA transcripts from a human DNA topoisomerase I pseudogene. *Cancer Res.*, **52**, 4280–4285.
19. Olsen,M.A. and Schechter,L.E. (1999) Cloning, mRNA localization and evolutionary conservation of a human 5-HT7 receptor pseudogene. *Gene*, **227**, 63–69.
20. Korneev,S.A., Park,J.H. and O'Shea,M. (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J. Neurosci.*, **19**, 7711–7720.
21. Hirotsune,S., Yoshida,N., Chen,A., Garrett,L., Sugiyama,F., Takahashi,S., Yagami,K., Wynshaw-Boris,A. and Yoshiki,A. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*, **423**, 91–96.
22. Rinn,J.L., Euskirchen,G., Bertone,P., Martone,R., Luscombe,N.M., Hartman,S., Harrison,P.M., Nelson,F.K., Miller,P., Gerstein,M. *et al.* (2003) The transcriptional activity of human Chromosome 22. *Genes Dev.*, **17**, 529–540.
23. Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M., Weissman,S. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
24. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature*, **409**, 860–921.

25. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
26. Wheeler,D., Church,D., Edgar,R., Federhen,S., Helmberg,W., Madden,T., Pontius,J., Schuler,G., Schriml,L., Sequeira,E. *et al.* (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D32–D40.
27. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
28. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
29. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
30. Apweiler,R., Bairoch,A., Wu,C., Barker,W., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
31. Harrison,P.M., Kumar,A., Lan,N., Echols,N., Snyder,M. and Gerstein,M. (2002) A small reservoir of disabled ORFs in the sequenced yeast genome and its implications for the dynamics of proteome evolution. *J. Mol. Biol.*, **316**, 409–419.
32. Harrison,P.M., Carriero,N., Liu,Y. and Gerstein,M. (2003) A 'polyORFomic' analysis of prokaryote genomes using disabled-homology filtering reveals conserved but undiscovered short ORFs. *J. Mol. Biol.*, **333**, 885–892.
33. Liu,Y., Harrison,P.M., Kunin,V. and Gerstein,M. (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.*, **5**, R64.
34. Chandonia,J., Hon,G., Walker,N., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
35. Harrison,P.M. and Sternberg,M.J.E. (1994) Analysis and classification of disulphide connectivity in proteins. The entropic effect of cross-linkage. *J. Mol. Biol.*, **244**, 448–463.
36. Harrison,P.M. and Sternberg,M.J.E. (1996) The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *J. Mol. Biol.*, **264**, 603–623.
37. Yang,Z. and Nielsen,R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol*, **17**, 32–43.
38. Bradley,J., Baltus,A., Skaletsky,H., Royce-Tolland,M., Dewar,K. and Page,D. (2004) An X-to-autosome retrogene is required for spermatogenesis in mice. *Nature Genet.*, **36**, 872–876.
39. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.
40. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
41. Lu,P.D., Harding,H.P. and Ron,D. (2004) Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response. *J. Cell. Biol.*, **167**, 27–33.
42. Haendler,B. and Hofer,E. (1990) Characterization of the human cyclophilin gene and of related processed pseudogenes. *Eur. J. Biochem.*, **190**, 477–482.
43. Semon,M. and Duret,L. (2004) Evidence that functional transcription units cover at least half of the human genome. *Trends Genet.*, **20**, 229–232.
44. Zhang,Z., Harrison,P. and Gerstein,M. (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.*, **12**, 1466–14482.
45. Manjeshwar,S., Branam,D.E., Lerner,M.R., Brackett,D.J. and Jupe,E.R. (2003) Tumor suppression by the prohibitin gene 3′untranslated region RNA in human breast cancer. *Cancer Res.*, **63**, 5251–5256.
46. Modrek,B. and Lee,C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, **34**, 177–180.
47. Letunic,I., Copley,R.R. and Bork,P. (2002) Common exon duplication in animals and its role in alternative splicing. *Hum. Mol. Genet.*, **11**, 1561–1567.
48. Kondrashov,F.A. and Koonin,E.V. (2003) Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet.*, **19**, 115–119.
49. Trabesinger-Ruef,N., Jermann,T., Zankel,T., Durrant,B., Frank,G. and Benner,S.A. (1996) Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function? *FEBS Lett*, **382**, 319–322.
50. Balakirev,E.S. and Ayala,F.J. (2003) Pseudogenes: Are they 'junk' or functional DNA? *Annu. Rev. Genet.*, **37**, 123–151.
51. True,H.L., Berlin,I. and Lindquist,S.L. (2004) Epigenetic regulation of translation reveals hidden genetic variation to produce complex traits. *Nature*, **431**, 184–187.
52. Podlaha,O. and Zhang,J. (2003) Non-neutral evolution of the transcribed pseudogene Makorin1-p1 in mice. *Mol. Biol. Evol.*, **21**, 2202–2209.
53. Dagan,T., Sorek,R., Ast,G. and Graur,D. (2004) AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res.*, **32**, D489–D492.
54. Kraulis,P.J. (1991) Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, **24**, 946–950.