doi:10.1016/j.jmb.2005.02.072 J. Mol. Biol. (2005) 349, 27–45

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Integrated Pseudogene Annotation for Human Chromosome 22: Evidence for Transcription

Deyou Zheng[1], Zhaolei Zhang[2], Paul M. Harrison[3], John Karro[1] Nick Carriero[4] and Mark Gerstein[1,4]*

[1]*Department of Molecular Biophysics and Biochemistry Yale University, 266 Whitney Avenue, New Haven, CT 06520 USA*

[2]*Banting and Best Department of Medical Research, 112 College Street, University of Toronto Toronto, Ont., Canada M5G 1L6*

[3]*Department of Biology, McGill University, 1205 Dr Penfield Avenue, Montreal, Que., Canada H3A 1B1*

[4]*Department of Computer Science, Yale University, 51 Prospect Street, New Haven CT 06520, USA*

**Corresponding author*

Pseudogenes are inheritable genetic elements formally defined by two properties: their similarity to functioning genes and their presumed lack of activity. However, their precise characterization, particularly with respect to the latter quality, has proven elusive. An opportunity to explore this issue arises from the recent emergence of tiling-microarray data showing that intergenic regions (containing pseudogenes) are transcribed to a great degree. Here we focus on the transcriptional activity of pseudogenes on human chromosome 22. First, we integrated several sets of annotation to define a unified list of 525 pseudogenes on the chromosome. To characterize these further, we developed a comprehensive list of genomic features based on conservation in related organisms, expression evidence, and the presence of upstream regulatory sites. Of the 525 unified pseudogenes we could confidently classify 154 as processed and 49 as duplicated. Using data from tiling microarrays, especially from recent high-resolution oligonucleotide arrays, we found some evidence that up to a fifth of the 525 pseudogenes are potentially transcribed. Expressed sequence tags (EST) comparison further validated a number of these, and overall we found 17 pseudogenes with strong support for transcription. In particular, one of the pseudogenes with both EST and microarray evidence for transcription turned out to be a duplicated pseudogene in the cat eye syndrome critical region. Although we could not identify a meaningful number of transcription factor-binding sites (based on chromatin immunoprecipitation-chip data) near pseudogenes, we did find that ~12% of the pseudogenes had upstream CpG islands. Finally, analysis of corresponding syntenic regions in the mouse, rat and chimp genomes indicates, as previously suggested, that pseudogenes are less conserved than genes, but more preserved than the intergenic background (all notation is available from http://www.pseudogene.org).

Keywords: chromosome 22; pseudogene; transcription; microarray; CESCR

Introduction

The completion of DNA sequencing is only the first step in understanding the human genome; with its completion we are faced with the challenge of deciphering the genetic components of the genome.[1–3] Investigators have already achieved a great deal, identifying many human genes using genome-scale experimental and computational approaches, but a great deal of work is still required to identify accurately the structures of all human genes. Coding exons of genes, however, cover only ~1.2% of the euchromatic genome while the untranslated regions of gene transcripts occupy ~0.7% of the euchromatic genome.[1,3] For a large section of the human genome, especially intergenic regions, our knowledge is sparse. Limited explorations of these often overlooked genomic regions

Present address: J. Karro, Department of Biology, Pennsylvania State University, 206 Wartik, PSU, University Park, PA 16802, USA.

Abbreviations used: TAR, transcriptionally active regions; EST, expressed sequence tags; ncRNA, non-coding RNA; CESCR, cat eye syndrome critical region; ChIP, chromatin immunoprecipitation.

E-mail address of the corresponding author: mark.gerstein@yale.edu

have shed light on their biological importance by revealing some highly conserved elements,[4–6] transcriptionally active regions (TAR)[7–10] and non-coding RNA genes (ncRNA).[11–13]

Pseudogene definition and type

One type of genetic element often found in these poorly understood regions is pseudogenes. While the structure and property of genes are generally clear, they have not been well addressed for pseudogenes. For a genomic sequence to be classified as a pseudogene based on conventional definition, it should be "non-functional" and display sequence similarity to a functional gene.[14–16] Sequence similarity can be computed rather easily, though there is no standard threshold value to be used as "close enough". The meaning of non-functional, however, is largely open for interpretation. It is usually regarded as either failure of transcription or translation, or production of a defective protein. Sometimes it is also viewed as lack of evolutionary selection pressure, since if a sequence is a gene, it will experience functional constraints and be under purifying selection according to the theory of neutral evolution.[17] In practice, such different interpretations often result in different operational definitions for pseudogene annotation (see below).

Pseudogenes are generated by two processes: direct gene DNA duplication or retrotransposition (i.e. the insertion of a reverse transcription product of a mature mRNA from a functional gene). A pseudogene resulting from the former is often referred to as a duplicated (or non-processed) pseudogene, and from the latter as a processed pseudogene (or retro-pseudogene). In general, the two distinct processes result in different sequence features, which can in turn be used to distinguish them. These features include, for example, the absence of introns, the presence of flanking direct repeats, and a 3′-polyadenylation tract for processed pseudogenes.

Pseudogene transcription

As mentioned above, it is conventionally thought that a pseudogene cannot generate a functional product. This often results from either the lack of an appropriate transcriptional promoter, the lack of critical RNA processing signals or the accumulation of various disruptions (such as insertions, deletions, premature stop codons, or frameshifts) in its putative protein coding region.[14,15,18] However, about 25 pseudogenes have been shown in the literature to be transcribed, some of which have even been shown to carry out biochemical functions.[18] Transcripts of several human pseudogenes have been described, including those from the interferon pseudogene[19] and the DNA topoisomerase I pseudogene.[20] Pseudogene transcription has also been reported in fly, mouse, cow, chimp and other organisms.[18] While these discoveries could simply be regarded as anecdotal cases and ignored, it is possible that they reflect our limited knowledge of pseudogenes or intergenic regions in general.

Furthermore, two transcribed pseudogenes have been shown to carry out biochemical function. Pseudogene *makorin1-p1* was demonstrated to regulate the mRNA stability of its parent gene, *makorin1*.[21] The synthesis of neutral nitric oxide synthase (nNOS) was suppressed by its homologous pseudogene in the neurons of mollusk *Lymnaea stagnalis*.[22] These reports indicate that the conventional view of pseudogenes being "dead" or "defunct" needs to be re-evaluated. However, the extent of pseudogene transcription is not clear and is one of the main questions that we set out to address.

Previous studies of human pseudogenes

In the past, pseudogenes were often discovered as by-products of studying individual genes or gene families. These studies have demonstrated many important aspects of pseudogene annotation. Because of their close sequence similarity to genes, pseudogenes can be mistakenly annotated as genes and can generate artifacts in molecular biological experiments caused by cross-hybridization. Moreover, they are "molecular fossils" and valuable for studying molecular evolution.[16,18] To further appreciate their importance and complexity, several groups have recently surveyed and characterized pseudogenes on genome scales.[16,23–27] Due to the lack of a consistent standard for pseudogene annotation, these studies were conducted using various approaches and with different focuses. Through exhaustive sequence comparisons with known human proteins, Zhang *et al*. reported that there were about 20,000 putative pseudogenes (~8000 processed) in the human genome.[26] Non-functionality of pseudogenes in this study was manifested as various disablements in the putative coding regions. Using basically the same approach but with more restrictive criteria, Ohshima and colleagues identified ~3600 human processed pseudogenes.[25] In another study using the lack of evolutionary selection pressure for inferring non-functionality, Torrents *et al*. showed that the human genome contains 19,724 pseudogenic regions.[24] These studies show that pseudogenes are nearly as prevalent as protein coding genes in the human genome, and therefore are an important component of our genome.

Chromosome 22 was one of the first chromosomes completely sequenced and is often used for pilot genomic scale studies and therefore many functional genomics data sets have been accumulated for this chromosome, such as microarray data identifying transcriptionally active regions (TARs)[7,8] and chromatin immunoprecipitation (ChIP)-chip data mapping transcription factor binding sites.[28,29] Since the release of its DNA sequence, this chromosome has been carefully examined for

accurate gene annotation, with 546 protein-coding genes identified†.[30,31] The annotation process also identified 234 pseudogenes, of which 168 were processed and 66 duplicated.[31] In an earlier independent study, 112 processed and 123 non-processed pseudogenes were discovered for chromosome 22, with a 5% false-positive rate.[23] This chromosome was also included in the previously cited works that searched for pseudogenes in the entire human genome.[24–26] In summary, these studies reported that there were 200–300 pseudogenes on chromosome 22, about half of which appeared to have been derived from retrotransposition.

With its careful annotation and extensive transcription relevant data, chromosome 22 becomes the ideal choice for examining pseudogene transcription. Here, we first integrated several annotations to generate a comprehensive list of 525 chromosome 22 pseudogenes with detailed descriptions of their features. We then examined our pseudogenes for potential transcriptional activity, considering a variety of expression evidence including microarray expression data, expressed sequence tags (ESTs), potential promoters, transcription factor binding sites and sequence conservation. Our analyses showed that up to a fifth of them could be transcribed with various degrees of evidence. To our knowledge, this is the first study of this nature. Our results suggest that transcribed pseudogenes might represent a new type of transcribed ncRNA and more studies are needed to comprehend their roles in genome organization, expression and evolution.

Results

Construct a list of putative pseudogenes on chromosome 22

Identify putative pseudogene regions using updated genome sequence and annotation

We first used a homology-based approach to look for putative pseudogene sequences as described.[23,26] Briefly, we set up a TBLASTN[32] search for chromosome 22 regions similar to human proteins. For the BLAST comparison we used the chromosome 22 sequence from build 34, proteins from both the ENSEMBL (version 19.34a) and the December 2003 release of UniProt database. BLAST hits overlapping with annotated exons were removed; mutual overlapping hits were consolidated. In the end, 509 regions on chromosome 22 were identified as potential pseudogenes (designated as DZ pseudogenes) (Figure 1).

† All our annotation is available from http://www.pseudogene.org

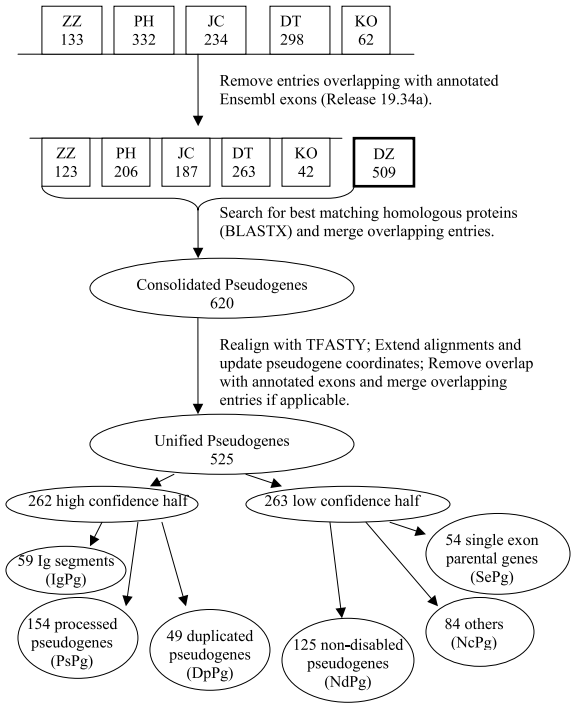

Figure 1. The data sources and process used to generate a comprehensive list of putative pseudogenes for chromosome 22. Pseudogenes from previous reports are labeled as PH,[23] JC,[31] KO,[25] DT,[24] ZZ,[26] and those from a process of this study as DZ.

Map pseudogenes from previous studies

Since the initial release of the human genome several research groups have developed computational pipelines to identify human pseudogenes. We retrieved the DNA sequences of pseudogenes on chromosome 22 as defined by Harrison *et al.*,[23] Collins *et al.*,[31] Ohshima *et al.*,[25] Torrents *et al.*[24] and Zhang *et al.*[26] These sequences were mapped to the human genome build 34 using BLAST[32] with a manual examination performed to resolve ambiguities. In a few cases, the sequences of previously identified pseudogenes were not found in the new build of the human genome. In the end, 332, 234, 62, 298 and 133 sequences from the above five studies were successfully mapped to build 34 of chromosome 22 (designated as PH, JC, KO, DT and ZZ pseudogenes, respectively) (Figure 1). Some of these remapped pseudogenes overlapped with newly annotated genes and thus were removed. As expected, pseudogenes derived from earlier drafts of human genome (PH and KO) showed more substantial overlaps with current gene annotations than those derived from the recent genome assembly (Figure 1).

Merge pseudogenes from different studies

We then merged all pseudogenes from the six sources described above to generate a total of 620 pseudogene candidates (Figure 1). These nucleotide

sequences were then aligned to their homologous proteins using the FASTA software package.[33,34] This step also updated the chromosome coordinates for each pseudogene to match the new alignment. As a result of the realigning and coordinate updating, several pseudogene candidates overlapped with others or with exons. They were consolidated with other pseudogenes or removed from subsequent analysis. The majority of sequences removed in this process were assigned using purely hypothetical proteins (e.g. translations of cDNA fragments) in the original studies. In the end, our integration process yielded 525 genomic sequences, referred to here as unified pseudogenes.

Chromosomal distribution of pseudogenes

Since the six sets of pseudogenes were determined differently, each represented a set in the pool of diverse pseudogenes on chromosome 22 with some overlap between sets (Figure 2(a) and (b)). The number of "core" pseudogenes identified by all studies is small, with many missed in at least one study. The inconsistency is a consequence of different thresholds for detecting pseudogenes' similarity to genes and different strategies for inferring "non-functionality" of pseudogenes (see Discussion for more details). In addition, as shown in Figure 2(a), five regions near the pericentromere show higher densities of pseudogenes than the rest of the chromosome. Three of these regions (located at 14.4–15.0 M, 15.6–16.0 M, and 20.7–21.6 M) appear to be specific to human lineage as they do not have mouse synteny. These regions are also coincidental with the highly duplicated pericentromeric regions.[35] In contrast, the chromosome distribution of functional genes is very different. Although both genes and pseudogenes are spread along the entire chromosome, genes are mainly found in genomic regions with mouse and rat synteny (data not shown).

Create feature lists for the 525 unified pseudogenes

We next generated a list of features to characterize these 525 unified pseudogenes. Table 1 lists these features, which can be separated into two classes: inherent and external. Inherent features (top section in Table 1) are directly associated with the pipeline process for assigning pseudogenes. They are either essential for describing a pseudogene or easily computed from any process for identifying pseudogenes. Most of these features, such as chromosome location, sequence identities and disablements (i.e. frameshifts or premature stop codons in the putative coding regions), are derived from either DNA or protein sequences. All these features are standard attributes of pseudogenes stored in our database†. These features are stored in a format compatible with the gene feature format (GFF) used to describe human gene annotation and therefore convenient for parsing and exchanging. External features (bottom of Table 1) describe sequence conservation, transcriptional evidence, and transcription factor binding, and are somewhat more specific to this study. For instance, the EST match is included as a means for identifying potentially expressed pseudogenes. Some of these features are also useful for characterizing functional genes and therefore provide a means for comparisons between genes and pseudogenes. Two examples are shown in Table 1; the full list of 525 putative pseudogenes with their features can be downloaded from the www†.

† http://www.pseudogene.org

Analysis of 525 putative pseudogenes

Classification of our putative pseudogenes

We then separated our 525 unified pseudogenes into two parts: (I) 262 that were easy to classify with high confidence; and (II) the rest (263) (Figure 1). The 262 (I) pseudogenes include (Figure 1):

IgPg, 59 immunoglobulin gene segments;

PsPg, 154 processed pseudogenes; of these, 124 were defined using these criteria: (1) harboring disablement(s) in the middle of the sequence alignment with the parental protein; (2) no "intron" (a gap >60 nt) in the pseudogene; (3) parental gene having more than one exon; and (4) the alignment covering $>70\%$ of the parental protein. The other 30 were classified as processed after a manual examination for traces of retrotransposition.

DpPg, 49 duplicated pseudogenes. These were identified manually by comparing the protein sequence alignments with the corresponding parental genes' exon-intron structures.

The 263 pseudogenes in (II) lack the clear characteristics used to identify the above IgPg, PsPg and DpPg pseudogenes. They are separated into three distinct groups:

NdPg, 125 non-disabled pseudogenes that are potential pseudogenes but without disablements (also referred to as pre-pseudogenes[31]);

SePg, 54 single-exon pseudogenes whose parental genes have only one exon (thus not easily classified as duplicated or processed pseudogenes); and

NcPg, 84 others that did not admit of any clear classification. Note: (1) this group also contains some sequences with low complexity; (2) two of the NcPg could be real genes as they have been separately incorporated into two genes in a new annotation release

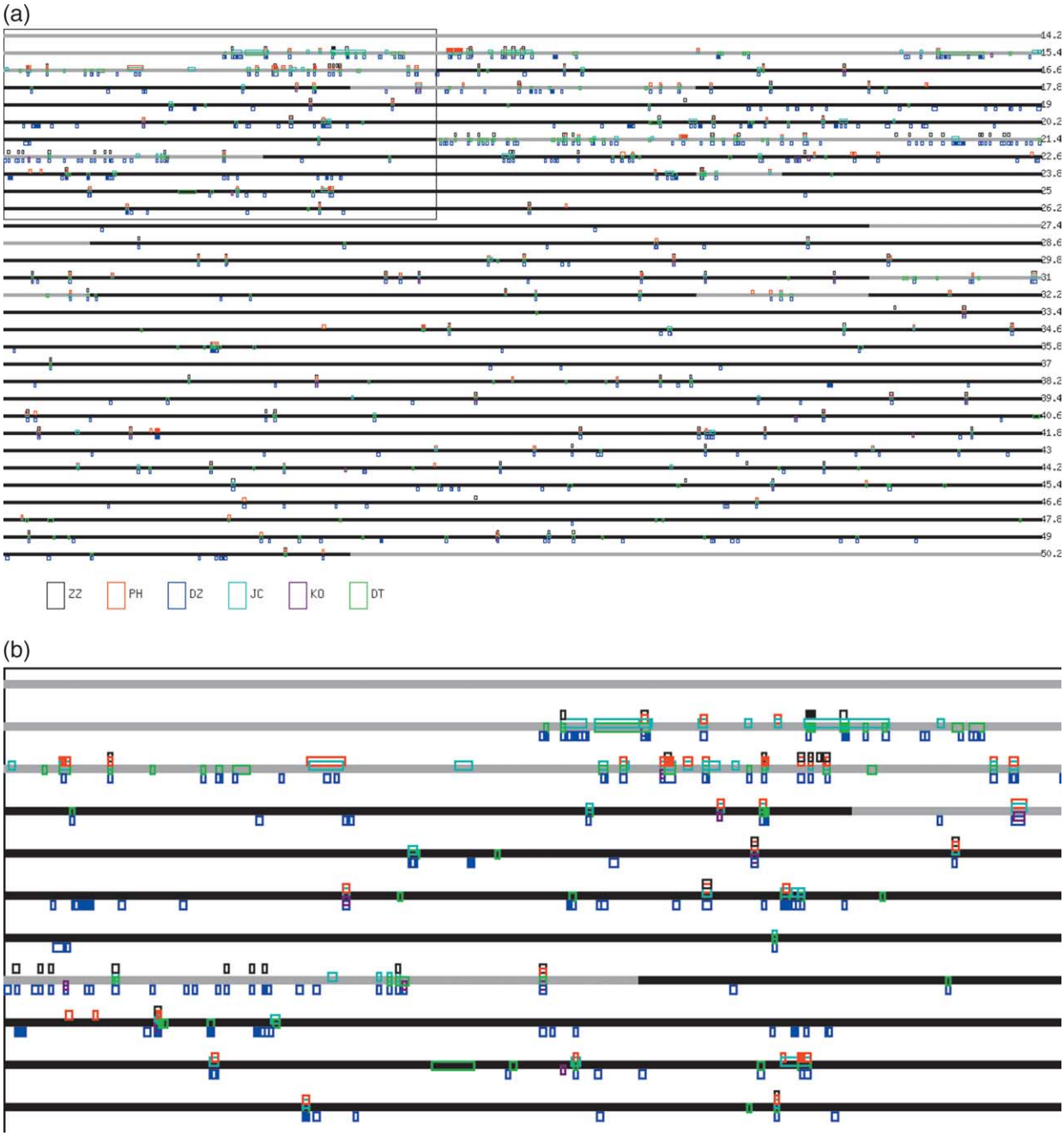

Figure 2. Chromosomal map of pseudogenes. (a) Pseudogenes identified in different studies. The chromosome is sketched in grey from centromere to telomere, with coordinates labeled on the right. The mouse syntenic regions are in black. Shown on the map are pseudogenes from different studies (colored as indicated and labeled as for Figure 1), excluding those overlapping with exons. The strand information is not plotted, and data from different studies are shifted slightly and not plotted in an exact scale for display purposes. (b) Higher-resolution view of the box area in (a).

(ENSEMBL v25.34e), with the regions containing disablements annotated as introns. For consistency, we included them in all subsequent analyses, since all our data refer to the ENSEMBL v19.34a. They are likely to be excluded in the future when we update our annotation using a new build of the human genome and its associated new annotation.

The above classification is somewhat arbitrary and should not be interpreted as providing a precise ratio of pseudogenes in different classes. It is worth mentioning that 20% of the 466 non-immunoglobulin pseudogenes were in the three aforementioned regions with high pseudogene densities, but without mouse synteny (data not shown).

We believe that these 525 pseudogene sequences form a comprehensive list of potential pseudogenes for chromosome 22; they are the union of the pseudogenes found by multiple approaches

Table 1. List of pseudogene features investigated in this study

Feature list	Description	Examples		Applicable to genes
A. *Identifier*				
Database Acc ID	Unique identifier	116 (HSFYpg)	1025	N
Source	ZZ/PH/JC/DT/KO/DZ[a]	JC	DZ	Y
Pgene_start, Pgene_end	Pseudogene start and end coordinate on chr22	15682995, 15684657	29374829, 29375355	Y
Strand	+/−	+	−	Y
B. *Alignment statistics*				
Parental ProteinID	Parental protein ID (the best matching protein)	ENSP00000303,599	ENSP00000270,634	N
Query_start, Query_end	Alignment start and end on parental protein	1, 401	1, 175	N
E-value, AA_ident, DNA_ident	Alignment *e*-value, amino acid and nucleotide identities	2.1e-68, 0.74, 0.89	3.3e-63, 0.81, 0.89	N
Completeness	Alignment coverage of the parental protein	1	0.86	N
Query_len	Parental protein length	401	203	N
Query_chr	Parental gene's chromosome	Y	19	N
Query_exon	No. exons in parental gene	2	8	N
Query_distance	Distance to the parental gene if it is also on chr22	NA	NA	N
C. *Sequence features*				
Poly(A)	0/1/2/3[b]	0	3	N
Disable	0/1 (0, No; 1, Yes)	1	1	N
Ig_fragment	0/1 (0, No; 1, Yes)	0	0	N
GC_pgene	Pseudogene GC%	0.39	0.58	Y
GC_query	Parental gene GC%	0.35	0.57	N
K_a/K_s	Ratio	1	1	Y
Pgene_intron	No. introns in pgene	1	0	Y
Disable_middle	No. disablements in the middle	11	6	N
Disable_edge	No. disablements at the N or C termini	0	0	N
D. *Intersection with transcription features*				
ESTid	GenBank accession ID	BX104099	NA	Y
EST_overlap	No. overlapping base-pairs between a pgene and these features[b]	402	0	Y
CpG island		0	0	Y
Rinn_positive		0	0	Y
NASA_Tar		0	0	Y
Nfkb_site		0	0	Y
CREB_site		0	0	Y
Mouse_preserve	0/1/2/3[b]	0	1	Y
Rat_preserve	0/1/2/3[b]	0	1	Y
Chimp_preserve	0/1/2/3[b]	3	3	Y

[a] Pseudogenes taken from previous reports are labeled PH,[23] JC,[31] KO,[25] DT,[24] ZZ;[26] those identified in this study are labeled DZ.
[b] See Methods for definitions.

and have been reconciled with a recent genome build (34). Future sequence changes for chromosome 22 are likely to be small, since chromosome 22 has been well characterized and annotated.[30,31] Therefore, our pseudogenes and their associated features are a solid and valuable resource for any future studies that aim to characterize chromosome 22 pseudogenes or intergenic regions.

Summary of selected pseudogene characteristics

Many characteristics of human pseudogenes have been described in previous analyses of both chromosome 22 and the whole human genome.[23,24,26] The pseudogenes in our unified list display many of the same properties. For example, we also found that the most abundant pseudogenes are derived from ribosomal protein genes. Table 2 summarizes some important features for our chromosome 22 pseudogenes. We selected NdPg, PsPg and DpPg for this Table because their characteristics are less ambiguous than those of the other classes. One-third of the processed pseudogenes still bear polyadenine (poly(A)) tails, as defined by a candidate polyadenylation signal and a region of elevated polyadenine content (>60% A) in the 1000 nt sequence at the 3′ downstream to the pseudogene.[23] These are likely recent processed pseudogenes. Of all pseudogenes excluding IgPg, about a quarter (23–24%) were derived from genes on chromosome 22 (109 *versus* 24 from chromosome 2, the second largest source), but genes from all human chromosomes contributed.

Table 2. Summaries of selected features for NdPg, DpPg and PsPg pseudogenes

	Non-disabled pseudogenes (NdPg)	Duplicated pseudogenes (DpPg)	Processed pseudogenes (PsPg)
Total number	125	49	154
With poly(A) tail	27	7[a]	50
Parental gene on chr22	50	9	32
$K_a/K_s \geqq 0.5$	74	40	109
≧40% amino acid identity	122	47	143
≧70% completeness	38	12	137
On regions with mouse synteny	87	30	110
With homologs in mouse syntenic regions	33	7	6
On regions with rat synteny	85	29	104
With homologs in rat syntenic regions	27	6	13
On regions with chimp synteny	125	49	154
With homologs in chimp syntenic regions	70	32	85
CpG island	0	8	28
Intersect TAR by PCR array	4	5	36
Intersect loose TAR by NASA array	12	5	16

[a] Seven DpPg had poly(A) tails, but four of them were labeled as class 3, i.e. having a detectable poly(A) tail but no polyadenylation signal.[23] The other three were false-positives of the method used to identify poly(A).

48% (52/109) of those from chromosome 22 are within 1 Mb of their parental genes. The percentages of DpPg and PsPg originating from chromosome 22 are similar (∼20%) (Table 2).

Distribution of K_a/K_s ratio and sequence identity

Estimation of synonymous (K_s) and non-synonymous (K_a) substitution rates often provides an important quantitative measure that distinguishes functional genes from non-functional genes (or pseudogenes).[24,26] In theory, K_a/K_s ratios should be ∼1 for pseudogenes and ≪1 for most genes. However, due to the limitation of how K_a/K_s is calculated, the benchmark value for most pseudogenes is 0.5–0.8[24,26] while the K_a/K_s ratio is <0.2 for most genes.[24] As shown in Figure 3(a), the K_a/K_s ratio of chromosome 22 pseudogenes follows the distribution characteristic of non-functional genes. That NdPg behave somewhat differently indicates that a few of them may be, in fact, components of real genes yet to be annotated (or pseudogenes arising very recently). However, the NdPg do exhibit a distribution of sequence identity with parental proteins similar to that for processed and duplicated pseudogenes (Figure 3(b)). Figure 3(c) plots the protein sequence identity of processed pseudogenes with and without poly(A) tail. It indicates that more recent processed pseudogenes, whose poly(A) tails have not decayed, show a higher level of sequence identity with the original genes.

Pseudogene preservation in mouse, rat and chimp genomes

As indicated above, it is generally known that pseudogenes are less conserved than genes because most of them evolve neutrally without functional constraints. Our previous study of single nucleotide polymorphisms (SNPs) also showed that SNP density was higher in pseudogenes than genes and that the ratio of non-synonymous to synonymous change was also higher for pseudogenes.[36] Here, we compare the conservation and preservation of human genes and pseudogenes in mouse, rat and chimp genomes by focusing only on syntenic regions. We use the term "preservation" instead of conservation to be consistent with the conventional view of no evolutionary pressure on pseudogenes. Our data revealed that 38% of NdPg and 23% of DpPg pseudogenes in chromosome 22 regions with mouse synteny actually had homologous sequences in the corresponding mouse syntenic regions (Table 2). Here, we refer to these sequences, which are similar to our pseudogenes and located within syntenic regions, as "putative mouse syntenic pseudogenes". The number of PsPg pseudogenes with putative mouse syntenic pseudogenes was significantly lower (5%) (Table 2). That PsPg were less preserved is consistent with that reported as a possible burst of PsPg in the human genome after the divergence of human and mouse lineages.[25,27] About half of the putative mouse syntenic pseudogenes contained obvious disruptions in their hypothetical coding regions. Similar results were obtained for rat, although a larger percentage of PsPg pseudogenes appear to have rat homologs. As one would expect, a significant number of pseudogenes were preserved in chimp chromosome 23 (the syntenic chromosome of human chromosome 22) (Table 2). In many cases, the positions of disablements were also preserved. By way of comparison, 89% of chromosome 22 genes were on regions with mouse and 87% were on regions with rat synteny; 82% and 77% of those had mouse or rat homologs, respectively; 88% of the genes had chimp homologs. To establish a scale, only 8% of randomly selected DNA sequences in mouse syntenic regions appear to have homologous sequences in the corresponding mouse syntenic regions. The equivalent figure for rat is 10%. Overall, pseudogenes are less conserved than genes, but are preserved better than background

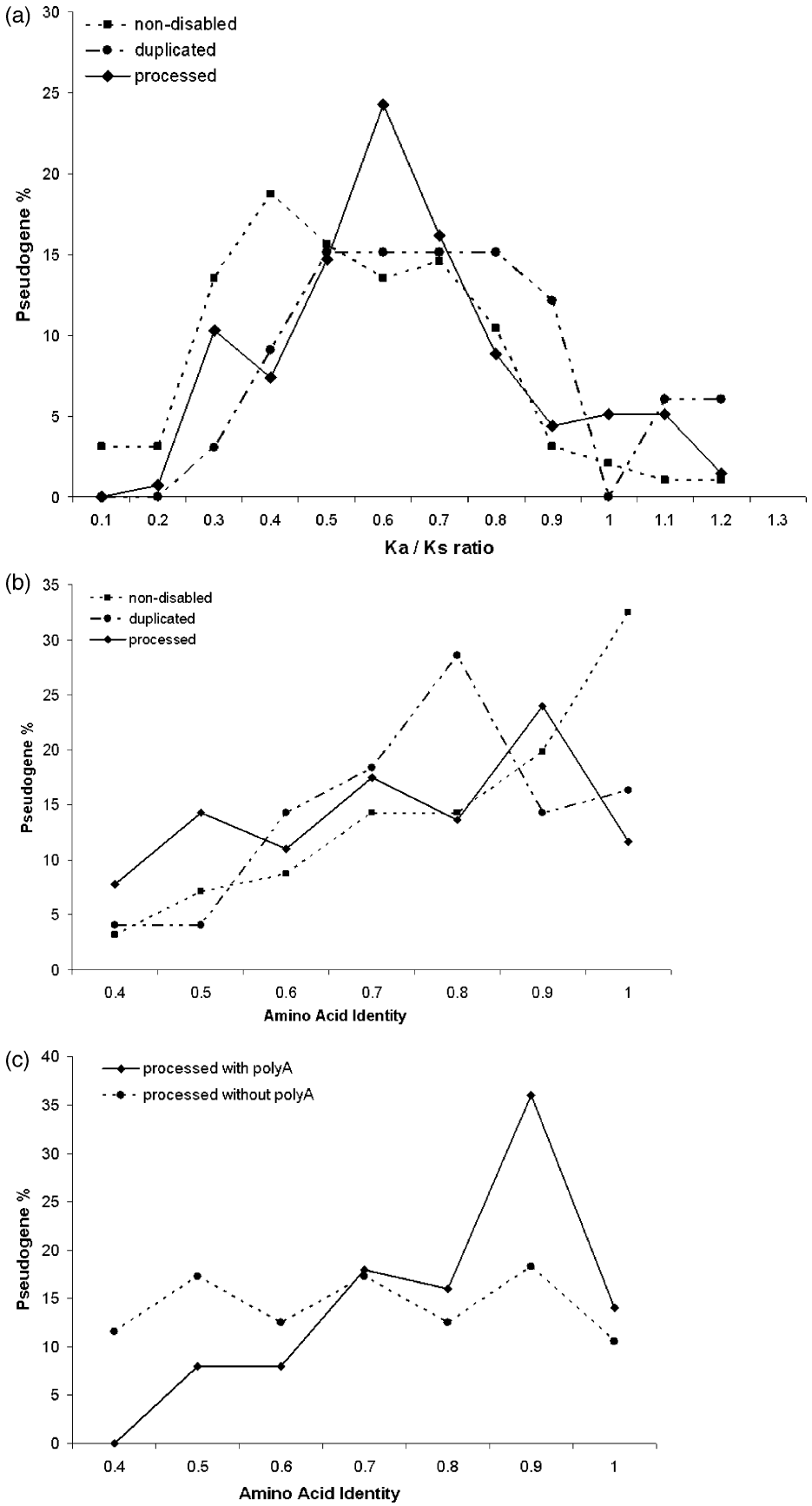

Figure 3. Distributions of K_a/K_s ratio and amino acid sequence identity of non-disabled, duplicated and processed pseudogenes. (a) The K_a/K_s ratio. (b) Amino acid identity with the parental genes. (c) Amino acid identity with the parental genes of a processed pseudogene with and without identifiable poly(A) tails.

genetic sequences. These analyses suggest that some pseudogenes could have experienced selection pressure and thus might be functional according to the theory of neutral evolution.[17]

Integration of pseudogene annotation with microarray expression data

There have been several examples of transcribed pseudogenes in the literature. In addition, various microarray technologies have shown a high degree of transcription in intergenic regions, where pseudogenes exist.[7,8,10] For example, using a high-density oligonucleotide array produced at NASA (referred to as the NASA experiment here), Bertone *et al*. identified a total of 13,899 transcription units, ranging in size from 209 nt to 3438 nt, but only a third of which corresponded to previously annotated exons.[10] Follow-up investigation will elaborate how transcription of these intergenic regions is related to cellular function. Nevertheless, these observations together suggest transcripts from pseudogenes could be part of the complex human transcriptome. Here, we used recent functional genomics data to address whether pseudogenes are "dead" or "alive" (i.e. transcriptionally active). With several types of expression data derived from EST comparisons and microarray assays (both expression and ChIP-chip), we screened for potentially transcribed pseudogenes on chromosome 22. Figure 4 gives an overview of our results. It shows a section of chromosome 22 with various transcription data mapped onto it. Tables 2 and 3 summarize our results, which go into detail below.

Intersection with the NASA array data

The NASA experiment contained a series of high-density oligonucleotide (36 nt) tiling microarrays representing both sense and antisense strands of the entire non-repetitive sequence of the human genome.[10] Due to the complexity of the data, two strategies were implemented for identifying transcribed genes and TARs. We refer to the originally reported TARs as "strict TARs" (Table 3A), consisting of at least five consecutive probes exhibiting fluorescence intensities in the top 90th intensity percentile.[10] We found ten pseudogenes that intersected with these TARs. However, these criteria were very strict and suitable only for highly transcribed regions. An intensity threshold at the 75th percentile is more suitable for identifying weakly transcriptionally active regions ("loose TARs"), where transcribed pseudogenes likely exist. We generated 1378 loose TARs (*versus* 112 strict TARs) and intersected them with our pseudogenes. This analysis identified 45 potentially transcribed pseudogenes (Table 3(A)), 13 of which had EST matches and seven overlapped with TARs from the PCR array (see below). In addition, analysis of the NASA array using a sign test identified an additional 24 transcribed pseudogenes ($P<0.05$) besides the above 45 (Table 3(B)). All together, the high-resolution NASA array data indicated that up to 18% of chromosome 22 pseudogenes could potentially be transcribed. After those pseudogenes with $\geq$95% nucleotide sequence identity with their parental genes were removed in order with to avoid possible cross-hybridization artifacts (see Methods), this analysis suggested that up to 12% of chromosome 22 pseudogenes were likely transcribed (Table 3).

Intersection with the Affymetix data

Transcriptional activity of chromosome 22 was probed independently using a distinct set of oligonucleotide arrays (referred to as the Affymetrix array).[7] Eleven cell lines were used in the study. Those authors also found that many intergenic regions were transcriptionally active. Using a sign test ($P<0.05$), we found that up to 19% of our pseudogenes were transcribed in an individual cell line (Table 3B) and 79 were transcriptionally active in at least six of 11 cell lines. Excluding those that may arise from cross-hybridization, we still found up to 16% pseudogenes potentially transcribed (Table 3B). There were nine pseudogenes that were transcriptionally active in all 11 cell lines, eight of which contained obvious disablements in their hypothetical coding regions. Moreover, three of these nine were also identified as transcribed in the analysis of the NASA array data, and one of the three had an EST match (see below).

Pseudogenes adjacent to transcription factor binding site

We next investigated the upstream regions of pseudogenes for sequence elements that are possibly involved in transcription regulation. A survey of the upstream 2 kb regions found 62 pseudogenes near CpG islands, of which eight also had EST matches (Table 2). In addition to the presence of a promoter signal (i.e. CpG island), another interesting question is whether a pseudogene is in the vicinity of a transcription factor binding site(s). Using the PCR tiling array and ChIP-chip, an unbiased mapping of NF-κB binding along human chromosome 22 identified 209 unique sites.[28] These sites were distributed along the entire chromosome in both coding and non-coding regions.[28] We found that 13% of chromosome 22 genes had NF-κB sites in the range from 1 kb upstream to the 3′ end, but less than 2% (8/466) of pseudogenes did. However, the figure for genes dropped significantly after we restricted our search to only within the upstream regions (Table 3A). This result indicates that many NF-κB sites are actually located within genes rather than in their promoter regions. A quick comparison of the numbers in Table 3A for genes and exons supports this interpretation as well, as many more exons have upstream NF-κB binding sites than genes. Similarly, using data derived from a ChIP-chip study for mapping cyclic AMP-responsive elements

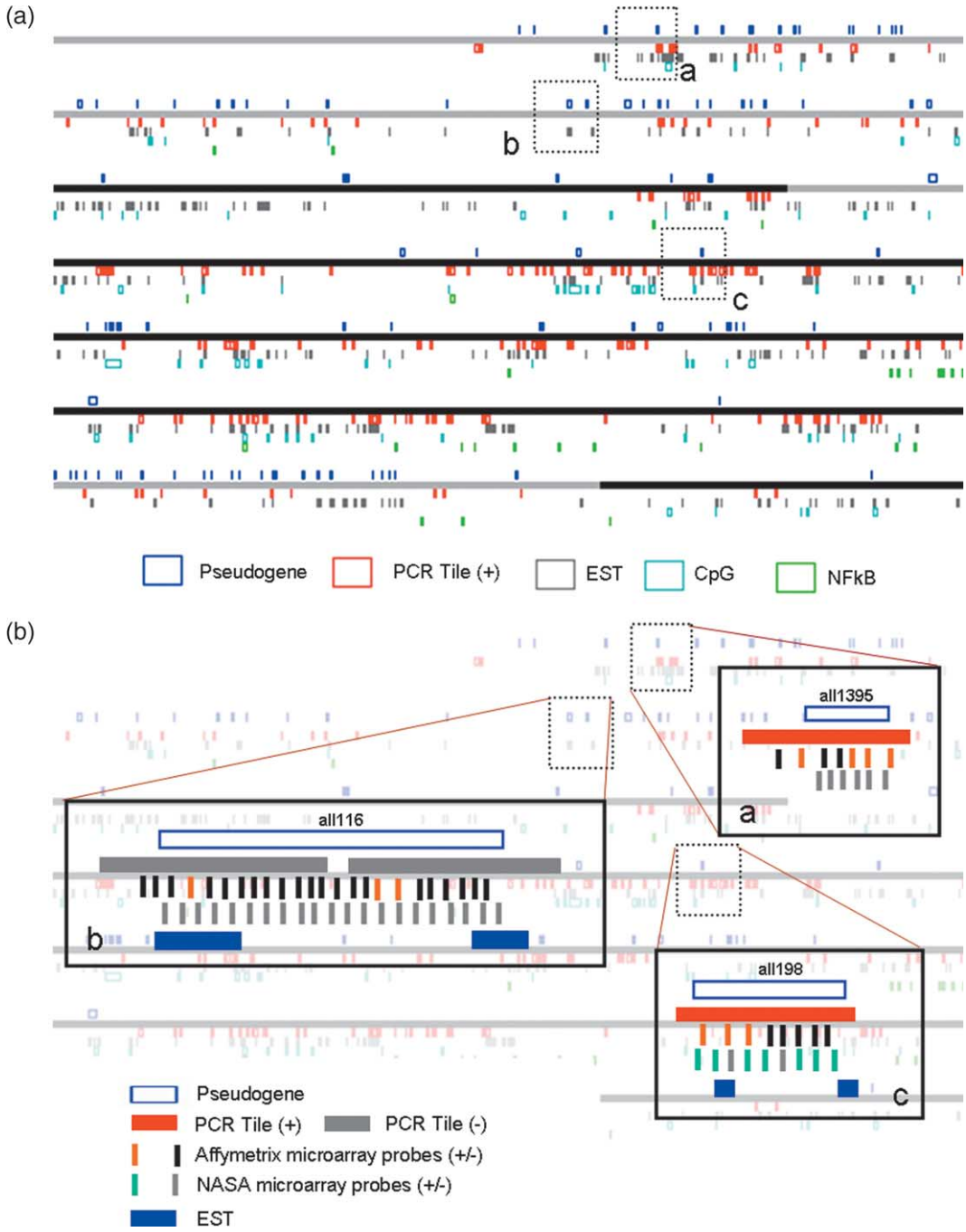

Figure 4 (a) and (b) (*legend opposite*)

(CREs),[29] we found that ∼2% of chromosome 22 pseudogenes were near CREs. Considering these findings together with the numbers derived from simulated fragments (Table 3A), it appears that none of the chromosome 22 pseudogenes has an NF-κB binding site or a cyclic AMP-responsive element near its immediate 5′ end.

Pseudogenes with confident EST matches

The integrative analyses of CpG islands, transcription factor binding sites and DNA microarray data present a global picture of genomic transcriptional activity. However, they do not provide conclusive evidence for transcription of a particular chromosome region. On the other hand, a unique EST match usually means a genomic region is indeed a TAR. Moreover, a pooled set of sequences in the EST database is a collection of ESTs from a variety of tissues, providing potentially a broader coverage of the human transcriptome. We believe that tissue or cell-line specific expression is a reason that we did not find a significant number of pseudogenes transcribed in all the microarray assays. The same problem is observed in cross-comparison of gene expression in different

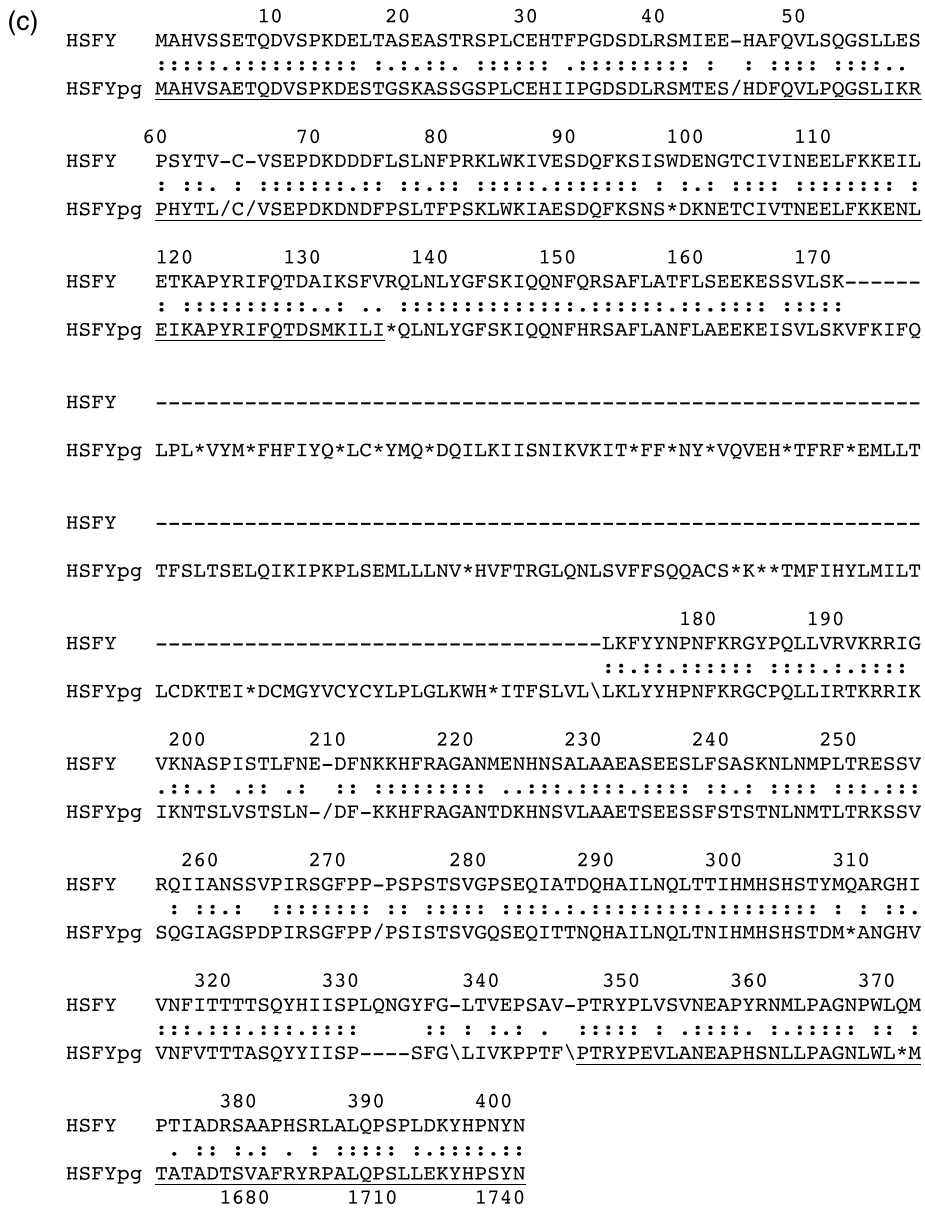

Figure 4. Map of a pseudogene with transcription evidence. (a) The chromosomal regions shown in Figure 2(b) are plotted with evidence of transcription from the PCR tiling array, ESTs and regulatory elements. The data from the NASA and Affymetrix arrays are not shown to reduce clutter. (b) High-resolution view of the box areas a, b and c in (a). Three transcribed pseudogenes are shown with a variety of transcription evidence. One in box a is supported by evidence from both the PCR tiling array and the Affymetrix array data; b, mainly from ESTs but also some Affymetirx data; and c, from all three microarray data set and ESTs. (c) Alignment of the pseudogene in box b with its parent gene. At the top is the functional protein HSFY (Ensembl ID, ENSP00000303599) from a gene on chromosome Y with two exons. At the bottom is its duplicated pseudogene located at 15,682,995–15,684,682 on chromosome 22, regions with EST match of BX104099 (N-) and AI214704 (C-) underlined. These two ESTs have a higher level of sequence identity relative to this pseudogene than to the HSFY gene.

microarrays. Therefore, to identify transcribed pseudogenes, we turned our focus to those with EST matches. In our initial screening described above, we used the pre-identified EST matching genomic regions, downloaded from the UCSC browser, and found 42 pseudogenes (Table 3A) overlapping these regions (73, if strand information was not considered). These sequences require further analyses in order to rule out false-positives, since the ESTs might be products of functional genes. We compared the 42 ESTs against the transcripts in ENSEMBL release 19.34a and the whole human genomic DNA sequence (see Methods). In the end there were 17 pseudogenes whose matching ESTs were significantly better aligned to the pseudogenic regions than to other locations in the genome or to any annotated transcripts (Table 4). Five of these were also among the transcriptional candidates from the analyses of oligonucleotide array data. These analyses suggest that there are at least 17 expressed pseudogenes (∼4% of non-Ig pseudogenes) on

Table 3. Intersection of pseudogenes with indicators of transcription activity

A. *Comparative analysis of pseudogenes, genes and exons with transcription evidence*

	Pseudogenes[a]		Chr22 genes		Chr22 exons	
	Pseudogenes	Random fragments	Genes	Random fragments	Exons	Random fragments
Total number	525 [466]	525	528	528	5265	5265
Intersect with PCR tiles (+/−) on Rinn *et al.* microarray	370 [315]	251	483	463	4599	2329
Intersect with PCR tiles (+)	67 [63][b]	34	310	259	754	312
of the above	(18 [20])	(14)	(64)	(56)	(16)	(13)
Intersect with ≥5 oligo probes (+/−) on NASA microarray[c]	293 [248]	195	463	451	921	599
Intersect with strict TARs (+)	10 [10]	1	57	32	56	3
of the above	(3 [4])	(<1)	(12)	(7)	(6)	(<1)
Intersect with loose TARs (+)	45 [45][d]	7	271	156	333	56
of the above	(15 [18])	(4)	(59)	(35)	(36)	(9)
EST match	47 [42]	16	397	192	1020	185
	(9 [9])	(3)	(75)	(36)	(19)	(4)
CpG island (−2 kb∼50 bp)	62 [62]	22	232	20	804	232
	(12 [13])	(4)	(44)	(4)	(15)	(4)
Intersect with NF-κB binding site (−2 kb∼50 bp)	6 [5]	7	10	5	107	60
	(1 [1])	(1)	(2)	(1)	(2)	(1)
Intersect with CREB binding site (−2 kb∼50 bp)	11 [8]	6	5	7	82	63
	(2 [2])	(1)	(<1)	(1)	(2)	(1)

B. *Number of transcribed pseudogenes identified by oligonucleotide microarrays (sign test* P*-value <0.05)*

	NASA array (liver)	Cell lines studied by Affymetrix array										
		A-375	CCRF-CEM	COLO 205	FHS 738Lu	HepG2	Jurkat	NCCIT	OVC-AR-3	PC-3	SK-N-AS	U-87 MG
All pseudogenes (Pg)	71	90	85	87	88	84	77	94	85	84	80	65
Non-Ig Pg	69	86	81	86	85	78	72	89	84	81	77	65
Non-Ig Pg, <95% i.d. to its parent gene	55	72	64	69	68	62	59	75	67	66	62	50

[a] The numbers in square brackets exclude Ig segments; the numbers in parentheses are percentages.
[b] Seventeen of them could be due to cross-hybridization.
[c] Only pseudogenes, genes and exons with at least five probes were counted because five consecutive positive probes were used to define a TAR.
[d] Ten of them could be due to cross-hybridization.

chromosome 22. This figure represents a lower-bound estimation, since sequences transcribed at a low level are under-represented in the EST collection. Two of these had a K_a/K_s ratio <0.2, indicating they are possibly real genes mis-annotated or very recent pseudogenes.

Summary of evidence relating to transcription

Our study of the NASA and Affymetrix microarray data indicated that up to 16% of pseudogenes could be transcriptionally active at some level. Independently, ∼4% of pseudogenes were found to be transcribed from our analysis of EST data. Taken together, we observed that five of the chromosome 22 pseudogenes (∼1%) had very strong evidence for transcription coming from both ESTs and all oligonucleotide microarray data, using very conservative thresholds for cross-hybridization. Conversely, in the broadest union we find that up to 19% (87/466) of the pseudogenes had some support from either ESTs or at least one of the oligonucleotide array experiments. Our best guess for the level of transcription of pseudogenes falls between these extremes. A somewhat conservative but reasonable estimate would be 17 (∼4%), which represents the number of pseudogenes with EST evidence and some array support. Figure 4(b) shows three examples of transcribed pseudogenes with a variety of evidence.

It is useful to compare our pseudogene results with some forms of random expectation. For instance, as described above, we found that 42 pseudogenes (excluding IgPg) were in genomic regions with EST matches (Table 3A) and 45 intersected TARs uncovered by the NASA array.[10] Comparing these numbers with simulated data suggests that there are more pseudogenes with EST matches ($P<0.0001$) or overlapping with the NASA array TARs ($P<0.0001$) than would be expected by chance (Table 3(A)). As a comparison, the percentages of genes (and exons) with EST matches or intersecting with the NASA array TARs are significantly higher than those of pseudogenes ($P<0.0001$) (Table 3(A)). However, the TARs identified by the PCR array appear evenly distributed on chromosome 22, with no significant enrichment in gene, exon or pseudogene regions, even though we found 67 pseudogenes intersected with these TARs (Table 3A). Finally, our analyses did not suggest that duplicated pseudogenes were more likely to be transcribed than processed

Table 4. Potentially transcribed pseudogenes with confident EST match

ID	Chr start	Chr end	Matching EST ID	Sequence identity with EST (overlap size) (%)		
				Pseudogene region	Other best match genomic location	Best match coding sequence
3	14,502,720	14,503,765	AI859005	97.1 (413)	95.0 (263)	95.1 (203)
116 (HSFYpg)	15,682,995	15,684,657	BX104099	98.0 (455)	90.0 (411)	89.6 (402)
			AI214704	100 (279)	84.1 (277)	84.3 (267)
138	15,902,765	15,902,890	AA328631	100 (57)	–	93.1 (58)
149	16,272298	16,273,429	BU584025	97.0 (133)	93.3 (105)	93.3 (105)
248	17,600,520	17,600,816	BF365681	100 (168)	83.7 (159)	84.3 (159)
271	20,710,099	21710,401	BU934404	100 (331)	90.7 (311)	93.0 (227)
319	20,893,875	20894,210	BU584342	95.0 (220)	–	–
357	21,111,062	21,111,355	BU584572	100 (207)	94.7 (207)	94.7 (207)
371	21,285,944	21,286,209	AL040506	99.8 (429)	81.8 (308)	87.5 (104)
423	21,547,845	21,548,117	AW407149	98.5 (262)	89.0 (181)	88.5 (174)
1152	21,589,319	21,589,636	AW393782	99.0 (381)	93.8 (370)	93.3 (372)
1118	22,410,547	22,411,878	BM146160	100 (259)	–	82.4 (159)
			BX095903	100 (350)	–	82.4 (159)
559	30,989,922	30,993,648	M78983	99.1 (331)	92.3 (91)	90.0 (331)
959	34,565,127	34,566,265	AI422197	99.6 (469)	92.4 (172)	93.1 (466)
			BE090973	99.2 (491)	94.1 (169)	92.3 (465)
618	36,444,591	36,445,172	AA314050	97.7 (480)	–	–
			AA628732	96.5 (366)	–	–
678	42,812,141	42,812,371	H55200	100 (102)	–	86.7 (158)
760	49,102,149	49,102,618	AA397693	100 (123)	–	96.8 (123)

pseudogenes (Table 2). This is somewhat counter-intuitive, since duplicated genes presumably have a higher chance to preserve a transcriptional promoter than processed pseudogenes.

An example of a transcribed pseudogene

Figure 4 shows a transcribed pseudogene. This pseudogene is a duplicate of a gene on chromosome Y, a heat-shock transcriptional factor (HSFY, a 410 residue protein of ENSEMBL ID ENSP00000303599). This pseudogene (HSFYpg, on the + strand of chromosome 22 with coordinates from 15,682,995 to 15,684,657) still preserves the intact two-exon structure of its functional counterpart, with 74% amino acid sequence identity (Table 1). It lies in the cat eye syndrome critical region (CESCR) on chromosome 22q11.2.[37] This disease is a rare developmental disorder characterized by a variety of congenital defects and the presence of three or four copies of a segment of 22q11.2. Its location outside a mouse/rat syntenic region suggests that the duplication likely arose after the split of human and rodent lineages. This locus has also been investigated in a previous attempt to discover candidate genes associating with cat eye syndrome, but without any conclusion.[38] There are two ESTs matching the 5′ and 3′ ends of HSFYpg (Figure 4). Both ESTs were from a testis cDNA library. The first, BX104099 (GenBank accession ID, 455 bp), is matched to chromosome 22 at 15,682,980–15,683,431 with 98% identical nucleotide residues. The second (AI214704, 281 bp) is aligned to 15,684,780–15,684,502 with 100% sequence identity (Table 4). In comparison, the best matches for BX104099 outside this pseudogene region are a 411 bp genomic sequence on chromosome Y (90% identity; also predicted as a pseudogene in the ENSEMBL database, ENSG00000183974) and the coding sequence (CDS) of protein ENSP00000303599 (89% identity over 395 bp). The best matches for AI214704 are a 277 bp genomic region on chromosome Y (84% identity) and the transcript of ENSP00000303599 (84% identity over 267 bp). Considering the significantly different sequence identities, alignment sizes (Table 4) and *e*-values (data not shown), the two ESTs are most likely associated with the HSFYpg. In addition, Northern blots analysis of a tissue sample from testis also confirmed that this locus is transcribed.[38] These data together indicate that HSFYpg is transcriptionally acitve. However, we could not find a CpG island within the 10 kb upstream region of HSFYpg. Nor did HSFYpg match a TAR from expression microarrays. As shown in Figure 4, this pseudogene contains several disruptions in the potential coding region and is therefore unlikely to produce a functional protein.

Discussion

We have carried out a study to develop a comprehensive catalogue of pseudogenes on chromosome 22 and to assess their transcriptional activity. The results show that there are about 500 pseudogene candidates on chromosome 22, and that up to 19% of them could be potentially transcribed. Our data and results should prove useful for future investigations of the function of pseudogenes. In particular, we hope that our study will stimulate interest in analyzing experimentally the role of pseudogenes in genome organization, expression and evolution. Our work also shows that

more studies are needed for reliable identification of pseudogenes and their transcription status.

Challenges in pseudogene identification

Like gene annotation, annotating pseudogenes is a challenging, but essential, task if we are to fully understand the human genome. Lacking evolutionary selection pressure, pseudogenes can accumulate various mutations (e.g. insertions and deletions) that make them extremely difficult to recognize. The concept of a pseudogene is generally understood, but it can be interpreted in many ways that lead to distinct operational definitions in practice for inferring pseudogene's non-functionality and their sequence similarity to functioning genes.[16,24,26] Such differences result in various algorithms and parameters (e.g. thresholds of *e*-value and sequence identity) that yield partially overlapping sets of pseudogenes. As shown in Figure 2, different approaches indeed yielded different lists of pseudogenes. Unfortunately, it is currently not meaningful to evaluate different strategies because there is not a gold standard set of pseudogenes and no experimental method is suitable for validating putative pseudogenes. It has also been pointed out that a homology-based strategy can detect most pseudogenes, but sometimes fails to reveal "full pseudogene structures".[31] It is, however, not clear what components a pseudogene may have. Can a pseudogene include intron(s), a promoter, a 5′ untranslated region (UTR) or a 3′ UTR? We observed that pseudogenes from the Sanger group (using more information than homology)[31] were, on average, at least twice as long as those identified by others (ZZ, 605 bp; PH, 659 bp; JC, 2224 bp; DT, 1102 bp; KO, 957 bp; DZ, 528 bp). The average length of DZ pseudogenes is shorter than that reported in previous studies because we did not follow the common practice of filtering out putative pseudogenes of lengths less than 70% of their parental genes. Therefore, many of our pseudogenes may be referred to as pseudogene fragments by others and not reported.

Future improvements for pseudogene identification

In the process of identifying DZ pseudogenes, we kept nearly all genomic regions that have recognizable sequence similarities to known proteins but are located outside of repetitive sequences (as marked by RepeatMasker) and annotated exons. Whereas a few of these sequences could eventually turn out to be components of real genes, most of them are likely pseudogenes or pseudogene fragments. Since the pseudogene definition is opened for interpretation, a probabilistic model-based approach (similar to approaches used in gene prediction) might better distinguish pseudogenes from other genetic elements because uncertainty can be built into such a model directly. This model should capture and integrate sequence homology, evolutionary pressure, sequence features associated with the mechanisms of pseudogene generation, and other statistical parameters that distinguish pseudogenes from real genes. Although identification of pseudogenes can be viewed as just a by-product of gene annotation, it probably needs its own model to catch characteristics unique to pseudogenes. We are currently developing such a model-based strategy. Toward this end, we found that a significantly large fraction of human processed pseudogenes from the work of Zhang *et al*.[26] were flanked with repetitive sequences (within 500 bp in the 5′ or 3′ end, data not shown), which are partial evidence of retrotransposition.

Explanation of our pseudogene classification

In this study we separate 525 unified pseudogenes into several groups. It must be emphasized that our classification cannot be interpreted to yield a precise ratio of pseudogenes between different categories. Some pseudogenes in the NdPg group may prove to be real genes (or part of them) or duplicated pseudogenes. Nevertheless, about half of candidates in our NdPg group, for which K_a/K_s values could be calculated, showed an elevated K_a/K_s ratio (>0.5). Our classification of processed pseudogenes (PsPg) is conceptual, based on the fact that a pseudogene is most likely a result of retrotransposition event if it does not contain an intron but its parent gene does.

Pseudogene preservation in mouse, rat and chimp genomes

The 525 putative pseudogenes identified here provide the most comprehensive list of pseudogenes for chromosome 22. This is likely to be a near-final pseudogene annotation, since any future sequence change for chromosome 22 should be very small. Most properties of these 525 pseudogenes are consistent with what have been described in the literature and thus have not been repeated here, e.g. the most prominent pseudogenes were derived from ribosomal protein genes or immunoglobulin genes. However, the preservation of these pseudogenes on the rat and mouse genomes appears to be weaker than that reported for the whole human genome.[27] This difference is attributed largely to recent gene duplications in the human lineage resulting in pseudogene dense regions without mouse/rat synteny near the chromosome 22 pericentromere (Figure 2). In fact, about 25% of our duplicated pseudogenes are in these regions. Our data therefore are chromosome 22-specific and not typical for pseudogene preservation in the human genome.

Pseudogenes transcription

Conventionally, pseudogenes are thought to be non-transcribed and non-functional. This view is probably true for most pseudogenes, but more than

25 examples of transcribed pseudogenes have been reported so far.[18] This list is expected to grow with new analyses using more functional genomics data. Therefore, the classical view of "dead" pseudogenes needs revision. That is a motivation to screen for transcribed pseudogenes using several kinds of expression data in this study. Although we found 50 or so pseudogene candidates with each type of transcriptional evidence, i.e. EST, CpG islands or individual microarray expression, we did not find one meeting all criteria. Such a result is not totally unexpected. Genomic sequences that are transcribed at a very low level are under-represented in the EST database but can be identified by microarrays. In addition, here we pooled all sequences in the EST database, and thus our set of ESTs represents a large collection of expressed sequences from many tissues and various conditions. A microarray experiment, however, is usually performed in one physiological condition for some cell lines or tissues, so it will identify only a subset of transcriptome. Therefore, we do not expect all EST-supported cases of transcribed pseudogenes to be validated by microarray data, and *vice versa*. Neither will we necessarily expect a complete overlap between transcripts identified from different microarray data sets. We plan to scale up our analysis with more functional genomics data to the entire human genome to get a broader picture of human pseudogene transcription.

We also need to keep in mind that transcription and its regulation of pseudogenes could be quite different from those of protein-coding genes. It has been suggested that pseudogenes might represent a reservoir of diverse "extra parts" that can only be resurrected during a cell "emergency" such as environmental stress.[39] Pseudogenes may have different promoter structures or may be transcribed only under special conditions. Therefore, a strategy different from what is currently used for studying gene expression might be needed for investigating pseudogene transcription. Nevertheless, with careful analyses, we still found 17 potentially expressed pseudogenes on chromosome 22 with EST evidence. Although EST data are sometimes very noisy and not always reliable, the general conclusion that an appreciable percentage of pseudogenes might be transcribed is overwhelmingly supported by all the types of expression evidence that were examined. Finally, it should be mentioned that an earlier study using EST sequences to screen processed pseudogenes also suggested that 2–3% of human processed pseudogenes might be expressed.[40] That approach, however, differs from ours, in that we took extra care to distinguish the EST of a pseudogene from the EST of its parental gene and to avoid ambiguous matches to other genomic regions.

Pseudogene function

Pseudogene study is still in its initial stage. Most investigations have been focused on the application of pseudogenes in studying molecular evolution. For example, pseudogenes have been used as a means for studying gene birth and death.[3] Since they are generally considered as dead and junk DNA, their molecular and cellular roles have largely been ignored even though many studies have reported individual cases of functional pseudogenes.[15,18] Our analyses of transcription and preservation not only provide evidence that some pseudogenes are transcriptionally active, but also suggest that a good fraction of them could be functional. Recently, a significant number of intergenic sequences have been found to be transcribed and to function as RNA (referred to as ncRNA) rather than protein. It has been shown that pseudogenes could regulate the stability or translation of their parent genes' mRNA using their RNA transcripts.[21,22] However, it remains to be studied whether some pseudogenes represent one type of ncRNA genes or an entirely new kind of functional element. Although our study does not directly answer such questions, pseudogenes found to be well preserved and transcriptionally active will be good candidates for subsequent investigation of pseudogene function. Finally, it has to be mentioned that a sequence can realize its function in many different ways, such as encoding a protein, regulating gene expression, or existing as a reservoir for new genes. Right now, it is too early to speculate on all the possible roles of pseudogenes.

Putative mechanism for the function of the HSFY pseudogene

We discovered a potentially expressed duplicated pseudogene (HSFYpg) in the CESCR on chromosome 22q11.2.[37] The complete genomic sequence of the HSFY gene can also be aligned to this pseudogene region with $>80\%$ sequence identity. There are two ESTs matching the 5′ and 3′ portions of this HSFYpg pseudogene (Figure 4). The nucleotide sequence identities suggest that these two ESTs are more likely products of the HSFYpg pseudogene than products of its parent gene. However, we could not locate a CpG island within the 10 kb upstream region of HSFYpg. Neither did this pseudogene overlap with a transcriptionally active fragment identified from the NASA array or other microarrays. This could be explained by the testis origin of those two ESTs and the possibility that the HSFYpg pseudogene has an unusual promoter structure. As shown in Figure 4, the HSFY pseudogene contains many disruptions in its putative coding region and is therefore unlikely to generate a functional protein. If the HSFY pseudogene is indeed transcribed, it will be interesting to see whether it plays any role in cat eye syndrome. It is possible that HSFYpg carries out its function as an RNA transcript using a mechanism similar to that shown for pseudogene *makorin1*-p1[21] or the nNOS pseudogene.[22]

Methods

Assign, merge and update putative pseudogenes on human chromosome 22

The human genome DNA sequence (NCBI build 34) and annotation (release 19.34a) were downloaded from ENSEMBL†.[41,42] This release contained 29,802 proteins. It also identified 528 genes on chromosome 22. We retrieved the DNA sequences of pseudogenes on chromosome 22 as determined by Harrison *et al.*,[23] Collins *et al.*,[31] Ohshima *et al.*,[25] Torrents *et al.*[24] and Zhang *et al.*[26] These sequences were mapped to the human genome build 34 using BLAST[32] with human examination. From these sets, 332, 234, 62, 298 and 133 sequences, respectively, were collected and mapped to chromosome 22 (Figure 1). A subsequent process then removed any sequence overlapping an annotated exon by at least 30 nucleotides.

Separately, a six-frame TBLASTN search was set up to search for DNA sequences on chromosome 22 (with repeats masked), similar to proteins in a set containing the human proteins in ENSEMBL (annotation release 19.34a) and proteins in UniProt (released December 2003, comprised of Swiss-Prot and TrEMBL release 42 but excluding purely hypothetical proteins translated from cDNAs). The resulting BLAST hits were merged to reduce mutual overlap, and then those intersecting (>30 nt) with exons were discarded. Details and parameters of this process have been described.[26,27] In the end, we kept 509 DNA sequences (designated as DZ pseudogenes) from this pipeline process for the analyses that follow.

Sequences from the five previous pseudogene sources and the 509 DZ pseudogenes were merged using their chromosome coordinates. We used the BLASTX program to search for the closest human protein homolog (defined as the smallest *e*-value) for each individual sequence. Those without a match (threshold *e*-value 1e-5) were discarded. This merging and updating process produced a total of 620 consolidated pseudogenes (Figure 1). Each sequence was then realigned with its homologous protein using the TFASTY program (with default parameters) of the FASTA package.[33,34] The realigning process was repeated, if necessary, to extend for an optimal alignment with maximal length. Next, the 620 sequences were subject to a final round of clean-up using their new extended chromosome coordinates from FASTA; mutual overlapping sequences were again consolidated and the one with the best FASTA alignment score was selected; any extended fragment overlapping with an exon (due to alignment extension) was removed. Details of this process have been described.[26] In the end, 525 DNA sequences were kept as putative pseudogenes and designated here as unified pseudogenes.

Create feature lists for the 525 putative pseudogenes

We created a list of features to describe each individual unified pseudogene (Table 1). The features include chromosome coordinates, strand, the parental protein ID, start and end positions on the query protein, *e*-value, amino acid identity to its parental protein, and the coverage of the DNA fragment on the parental protein. These values were derived from FASTA alignments. We considered an intron present in a putative pseudogene if there was a gap (on the parental protein) longer than 20 residues within the alignment.[26] We also counted the number of stop codes and frameshifts within the alignment region. The count was used to infer whether a pseudogene contained any disablements (≥ 1) in its putative coding sequence. Such disablements in the ten amino acid residues at the N/C terminus of the alignment were defined as disablements at the edge; otherwise, as in the middle.

Using the protein alignments as guidance, coding sequences of the parental genes (taken from ENSEMBL) were aligned to the putative pseudogenes. These DNA alignments were used to compute DNA sequence identity and GC content. Gaps were omitted when calculating sequence identities. The presence and classification of poly(A) tails were determined as described.[23,26] We also calculated the ratio between the non-synonymous *versus* synonymous rates of substitution (K_a/K_s) for an individual pseudogene using the YN00 program within the PAML evolutionary package.[43,44] Additionally, there are many immunoglobulin gene segments in chromosome 22.[23,31] We flagged 59 DNA fragments as immunoglobulin gene segments (IgPg) based on their homologies to immunoglobulin genes. In addition, we recorded for each pseudogene sequence the chromosome and the number of exons of its parental gene. If a parental gene was also on chromosome 22, we calculated the distance between the pseudogene and its parent.

Intersection of pseudogenes with transcription evidence

Several functional genomic data resources were used to screen for potentially transcribed pseudogenes, including CpG islands, EST sequences, microarray expression data and transcription factor binding sites. It has been shown that 40–60% of human genes have distinctive CpG islands at their 5′ end.[45] Since CpG islands are usually associated with transcriptional promoters, it is interesting to study how many pseudogenes are near CpG islands. We downloaded genomic regions (in the form of chromosome coordinates) with CpG islands from the UCSC browser.‡ There were 688 predicted CpG islands in the downloaded data. From the UCSC browser we also obtained a list of genomic regions with matching ESTs. From the list we removed those entries which either contained a gap in the EST alignment or for which the size of the alignment was less than 90% of the EST length. These criteria left us with 19,205 chromosome 22 regions with reliable EST matches.

Microarray expression data from three platforms were used in our study. Rinn *et al.* recently constructed a tiling PCR microarray for probing transcriptional activity.[8] DNA nucleotides on the array represent almost all unique sequences of chromosome 22. Additionally, two distinct types of oligonucleotide arrays were constructed for identifying novel transcripts.[7,10] The main differences among these three arrays are the size of nucleotides (i.e. probes) and the probe spacing; as a result they generated maps of chromosome transcriptional activity with different resolutions. Despite these differences, all three studies reported that there were as many transcriptions in unannotated regions as in the annotated regions that harbor genes. We retrieved all these data and mapped the corresponding nucleotide probes and transcriptionally active regions (TARs) to build 34 of the human genome, excluding those that could be mapped to more than one region in chromosome 22 with 100% sequence identity.

† http://www.ensembl.org

‡ http://genome.ucsc.edu.[46]

The PCR array was also used for screening transcription factor binding sites on chromosome 22 by the ChIP-chip technique. We used these data as well.

In order to identify potentially transcribed pseudogenes, we screened for pseudogenes that were near CpG islands, located at genomic regions with EST matches, or at TARs. In these screens, an intersecting size larger than 200 base-pairs, or longer than 0.75 of the shorter sequences being compared (if both sequence lengths were less than 200 nt) was used to define positive overlapping. Strand information was considered in analysis of EST data. Additionally, in the examination of upstream regions for CpG islands or transcription factor binding sites, we looked only at the region from −2 kb to +50 bp of a pseudogene, since it has been shown that the average distance of CpG islands is ∼1 kb upstream for chromosome 22 genes.[45] For comparisons, the above analyses identifying transcription evidence were also carried out for 528 genes and 5265 (unique) exons on chromosome 22. Furthermore, for background statistics we generated 525, 528 and 5265 fragments with the same size distributions as the pseudogenes, genes and exons, respectively, distributed them randomly on chromosome 22, and intersected them with transcription evidence. This process was repeated 100 times. The average numbers of random fragments intersecting transcription evidence are shown in Table 3A.

During the analysis of data from the NASA array, we first studied the intersections of our pseudogenes with the 112 TARs ("strict" TARs) identified by the authors.[10] We then reduced the intensity threshold from 90% to 75% in order to identify genomic regions that are weakly transcribed. With this modified parameter and the same requirement of at least five consecutive positive probes, 1378 TARs were identified. These TARs were called loose TARs and subsequently used to screen for transcriptional activity.

In addition, a sign test[10] was used to identify potentially transcribed pseudogenes in the analyses of expression data from the NASA array and the Affymetrix array. For the NASA array, each probe was assigned a value of 1 if its fluorescence intensity was greater than the median intensity of all probes on the array, and 0 otherwise. For the Affymetrix array, a positive probe identified by the original authors was assigned a value of 1. Pseudogenes with P-values <0.05 in the sign test were regarded as demonstrating positive hybridization.

Identification of false positives due to cross-hybridization or mis-assignment of ESTs

One particular concern when studying pseudogene expression using microarray data is cross-hybridization. This is especially problematic if the size of the probes is relatively large, such as those used in the PCR array; the DNA fragments spotted on the Rinn *et al*. microarray were PCR products of size from 300 bp to 1.4 kb.[8] We cannot tell precisely the genomic origin of a positive probe in a microarray when several similar genomic regions exist as candidates. Here, we made a conservative assumption that a transcription unit did not belong to a pseudogene if we cannot rule out a gene as its potential source. Following this assumption we filtered out and did not use TARs that could be aligned to more than one genomic region (with 100% identity). Nevertheless, we were still concerned with assigning a TAR from a functional gene to a pseudogene by mistake due to cross-hybridization. We adopted the following procedures to identify artifacts potentially caused by cross-hybridization in microarray experiments.

(1) *For the PCR array*. In our study of the Rinn *et al*. microarray data, we found 67 pseudogenes overlapping with TARs (Table 3A). To limit possible false-positives due to cross-hybridization, we searched the predicted coding sequences (CDSs) of the whole human genome using the PCR fragments (probes) of the positive microarray spots. A BLAST hit of 150 bp with ≥95% identity was interpreted as an indication that the original probe detected the transcription of a CDS and therefore our assignment of a "transcribed" pseudogene based on this probe was a false-positive. We found that 15 probes met this criterion of likely cross-hybridization, accounting for 17 falsely assigned transcribed pseudogenes (two of the 15 probes each spanned two pseudogenes). Therefore, 46 non-Ig pseudogenes are potentially transcriptionally active from our analysis of the expression data from PCR tiling array (Table 3A).

(2) *For the oligonucleotide arrays*. Because the data from the NASA array and the Affymetrix array had better resolution, we simply excluded pseudogenes that had a nucleotide sequence identity with their parental genes of ≥95% when counting transcriptionally active pseudogenes without cross- hybridization (Table 3). This criterion shows that ten of the 45 transcribed pseudogenes identified using the loose TARs from the NASA microarray could be results of cross-hybridization (Table 3A). Overall, approximately 20% of the pseudogene transcription identified by the NASA and Affymetrix tiling microarrays may conservatively be due to cross-hybridization at some degree (Table 3). If we take a very aggressive approach and change the parameter of sequence identity from 95% to 85%, about a half of our transcribed pseudogenes will be excluded. However, the transcribed pseudogene *makorin1-p1* is actually more than 85% identical with the *makorin1* gene.[21] It has to be stressed that identifying microarray signals from cross- hybridization is a very challenging task. Here, we used the similar criteria for inferring cross-hybridization as described.[10]

(3) *For the EST data*. This is not really an issue about cross-hybridization, but mis-identification of transcribed pseudogenes using ESTs generated from functional genes. In order to avoid such a mis-classification we developed a strategy to distinguish ESTs of functional genes from those of pseudogenes. We compared those ESTs that were matched to pseudogenic regions against the whole genome and the ENSEMBL transcripts for sequence(s) of >80% identity. We chose and subsequently studied an EST only if the BLAST e-value for its alignment to a pseudogene was at least a factor of 10^{10} smaller than the e-values of its alignment to either a CDS or other genomic region. The sizes and sequence identities of the alignments between such an EST and its three types of matching sequences were compared (Table 4) in order to determine if the EST best matched a pseudogene among the three candidates. The goal of this process is to identify ESTs that are genuine products of pseudogenes.

Preservation in mouse, rat and chimp syntenic regions

We investigated the preservation of each of our 525

putative pseudogenes in the syntenic regions of the mouse, rat and chimp genomes. We obtained the mouse, rat and chimp synteny maps of chromosome 22 from the ENSEMBL website. The data segregated chromosome 22 into nine mouse and ten rat syntenic regions. Chromosome 23 of the chimp genome is the syntenic chromosome of human chromosome 22. With these maps as a guide, we used the parental gene from which a pseudogene was derived as the query to search for homologous DNA sequences within its syntenic regions using TBLASTN (*e*-value <1e-5). A BLAST hit was further realigned and extended with the TFASTY program as described above for the process of pseudogene identification. In the end, we defined four categories for describing the preservation of a pseudogene: (0) if a pseudogene was located in a chromosome region without mouse/rat synteny; (1) if it was in a syntenic region but without identifiable mouse, rat or chimp homolog; (2) if it had a mouse, rat or chimp homolog but no disablement detected in the corresponding protein-pseudogene sequence alignment; and (3) if disablement(s) was observed.

Acknowledgements

The authors thank Paul Bertone, Joel Rozowsky, Thomas Royce and Olof Emanuelsson for useful discussions. M.G. acknowledges financial support from the NIH (P50 HG02357-01).

References

1. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al*. (2001). The sequence of the human genome. *Science*, **291**, 1304–1351.
3. International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
4. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
5. Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M. & Frazer, K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.
6. Pennacchio, L. A. & Rubin, E. M. (2001). Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2**, 100–109.
7. Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. & Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**, 916–919.
8. Rinn, J. L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N. M., Hartman, S. *et al*. (2003). The transcriptional activity of human chromosome 22. *Genes Dev.* **17**, 529–540.
9. Semon, M. & Duret, L. (2004). Evidence that functional transcription units cover at least half of the human genome. *Trends Genet.* **20**, 229–232.
10. Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X. *et al*. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
11. Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2**, 919–929.
12. Scherer, S. W., Cheung, J., MacDonald, J. R., Osborne, L. R., Nakabayashi, K., Herbrick, J. A. *et al*. (2003). Human chromosome 7: DNA sequence and biology. *Science*, **300**, 767–772.
13. Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D. *et al*. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
14. Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.* **19**, 253–272.
15. Mighell, A. J., Smith, N. R., Robinson, P. A. & Markham, A. F. (2000). Vertebrate pseudogenes. *FEBS Letters*, **468**, 109–114.
16. Zhang, Z. & Gerstein, M. (2004). Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.* **14**, 328–335.
17. Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.
18. Balakirev, E. S. & Ayala, F. J. (2003). Pseudogenes: are they "junk" or functional DNA? *Annu. Rev. Genet.* **37**, 123–151.
19. Goeddel, D. V., Leung, D. W., Dull, T. J., Gross, M., Lawn, R. M., McCandliss, R. *et al*. (1981). The structure of eight distinct cloned human leukocyte interferon cDNAs. *Nature*, **290**, 20–26.
20. Zhou, B. S., Beidler, D. R. & Cheng, Y. C. (1992). Identification of antisense RNA transcripts from a human DNA topoisomerase I pseudogene. *Cancer Res.* **52**, 4280–4285.
21. Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S. *et al*. (2003). An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*, **423**, 91–96.
22. Korneev, S. A., Park, J. H. & 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.
23. Harrison, P. M., Hegyi, H., Balasubramanian, S., Luscombe, N. M., Bertone, P., Echols, N. *et al*. (2002). Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**, 272–280.
24. Torrents, D., Suyama, M., Zdobnov, E. & Bork, P. (2003). A genome-wide survey of human pseudogenes. *Genome Res.* **13**, 2559–2567.
25. Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y. & Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol.* **4**, R74.
26. Zhang, Z., Harrison, P. M., Liu, Y. & 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.
27. Zhang, Z., Carriero, N. & Gerstein, M. (2004). Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.* **20**, 62–67.
28. Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T. E., Luscombe, N. M. *et al*. (2003).

Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc. Natl Acad. Sci. USA*, **100**, 12247–12252.
29. Euskirchen, G., Royce, T. E., Bertone, P., Martone, R., Rinn, J. L., Nelson, F. K. *et al*. (2004). CREB binds to multiple loci on human chromosome 22. *Mol. Cell. Biol.* **24**, 3804–3814.
30. Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E. *et al*. (1999). The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
31. Collins, J. E., Goward, M. E., Cole, C. G., Smink, L. J., Huckle, E. J., Knowles, S. *et al*. (2003). Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13**, 27–36.
32. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
33. Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
34. Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. (1997). Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.
35. Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S. *et al*. (2002). Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.
36. Balasubramanian, S., Harrison, P., Hegyi, H., Bertone, P., Luscombe, N., Echols, N. *et al*. (2002). SNPs on human chromosomes 21 and 22-analysis in terms of protein features and pseudogenes. *Pharmacogenomics*, **3**, 393–402.
37. Schinzel, A., Schmid, W., Fraccaro, M., Tiepolo, L., Zuffardi, O., Opitz, J. M. *et al*. (1981). The "cat eye syndrome": dicentric small marker chromosome probably derived from a no.22 (tetrasomy 22pter to q11) associated with a characteristic phenotype. Report of 11 patients and delineation of the clinical picture. *Hum. Genet.* **57**, 148–158.
38. Footz, T. K., Brinkman-Mills, P., Banting, G. S., Maier, S. A., Riazi, M. A., Bridgland, L. *et al*. (2001). Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere. *Genome Res.* **11**, 1053–1070.
39. Harrison, P. M. & Gerstein, M. (2002). Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318**, 1155–1174.
40. Yano, Y., Saito, R., Yoshida, N., Yoshiki, A., Wynshaw-Boris, A., Tomita, M. & Hirotsune, S. (2004). A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *J. Mol. Med.* **82**, 414–422.
41. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L. *et al*. (2002). The Ensembl genome database project. *Nucl. Acids Res.* **30**, 38–41.
42. Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y. & Clarke, L. (2004). An overview of Ensembl. *Genome Res.* **14**, 925–928.
43. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556.
44. Yang, Z. & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43.
45. Scherf, M., Klingenhoff, A., Frech, K., Quandt, K., Schneider, R., Grote, K. *et al*. (2001). First pass annotation of promoters on human chromosome 22. *Genome Res.* **11**, 333–340.
46. Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D. & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucl. Acids Res.* **32**, D493–D496.

Edited by M. Levitt

(Received 20 October 2004; received in revised form 16 February 2005; accepted 23 February 2005)