Published online 31 August 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 16* ***4477–4485***
doi:10.1093/nar/gkl591

Systematic identification of pseudogenes through whole genome expression evidence profiling

Alison Yao, Rosane Charlab[1] and Peter Li[1,*]

Celera Genomics, 45 West Gude Dr, Rockville, MD 20850, USA and [1]Applied Biosystems Inc, 45 West Gude Dr, Rockville, MD 20850, USA

Received October 30, 2005; Revised July 28, 2006; Accepted July 31, 2006

ABSTRACT

The identification of pseudogenes is an integral and significant part of the genome annotation because of their abundance and their impact on the experimental analysis of functional genes. Most of the computational annotation systems are not optimized for systematic pseudogene recognition, often annotating pseudogenes as functional genes, and users then propagate these errors to subsequent analyses and interpretations. In order to validate gene annotations and to identify pseudogenes that are potentially mis-annotated, we developed a novel approach based on whole genome profiling of existing transcript and protein sequences. This method has two important features: (i) equally detects both processed and non-processed pseudogenes and (ii) can identify transcribed pseudogenes. Applying this method to the human Ensembl gene predictions, we discovered that 2011 (9% of total) Ensembl genes in the categories of known and novel might be pseudogenes based on expression evidence. Of these, 1200 genes are found to have no existing evidence of transcription, and 811 genes are found with transcription evidence but contain significant translation disruption. Approximately 40% of the 2011 identified pseudogenes presented a multi-exon structure, representing non-processed pseudogenes. We have demonstrated the power of whole genome profiling of expression sequences to improve the accuracy of gene annotations.

INTRODUCTION

Pseudogenes are defined as non-functional genomic sequences derived from functional genes. The loss of function is generally viewed as either a failure of transcription or translation, or production of a defective protein (1,2). Most pseudogenes are thought to be transcriptionally silent, but transcribed pseudogenes have been experimentally identified (3–10). Over the past several years, substantial efforts have been devoted to genome-scale identification and characterization of pseudogenes (11–15). However, none of these methods is optimized for the detection of non-processed pseudogenes that have retained the original exon/intron structures. Of these, the transcribed types have been known to be especially problematic for gene annotation (16,17). Because of the similarity to functional paralogs, pseudogenes are often mis-incorporated into gene collections (18–21) introducing errors that propagate downstream to many subsequent analyses. For example, a challenge of designing targeted expression assays is to avoid cross-reacting paralogs, thus by knowing which of the paralogs are reacting and non-reacting pseudogenes, it will simplify the task of assay design and result analysis.

Among the causes for mis-incorporation are (i) multiple gene instantiations supported by one piece of evidence aligned to multiple places on the genome and (ii) the use of poor quality or fragmented evidence (22). Moreover, most gene prediction algorithms were designed or trained for detecting functional genes, but in practice, will make pseudogenes into viable gene models by adjusting splicing patterns (23). Accurate pseudogene annotation may ultimately rely on manual curation, which is a labor-intensive and time-consuming process.

Expression evidence, such as mRNA, expressed sequence tags (ESTs) and protein sequences, could be potentially mapped to many gene loci including its own locus, paralogs and potential pseudogenes. In the process of gene annotation, we have observed that these alignments usually have different degrees of match statistics, such as identity, coverage and splicing status (see Figure 1), but the 'best hit' of any given evidence is always associated with the originating locus. This would then serve as confirming evidence for gene expression of that locus. In addition, we define a pseudogene without confirming transcriptional products as non-transcribed pseudogene and one with transcriptional products, but without translational products, as transcribed pseudogene. We developed a novel bioinformatics method

*To whom correspondence should be addressed. Tel: +1 240 453 3154; Fax: +1 240 453 3587; Email: Peter.Li@appliedbiosystems.com
Present address:
Alison Yao, National Institute of Allergy and Infectious Diseases, NIH 6610 Rockledge Dr, Bethesda, MD 20892, USA

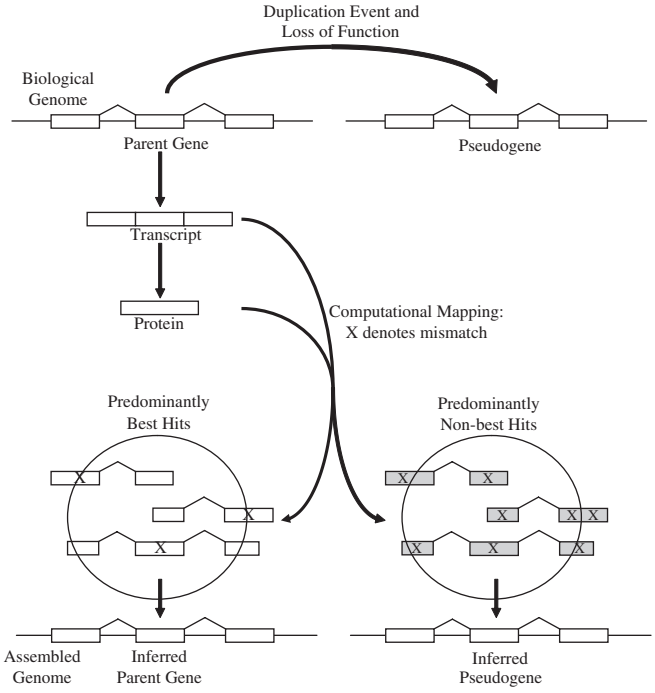

Figure 1. The conceptual relationships between parental gene, pseudogene and expression evidence. More mismatches (X) would be seen in the non-best hits of evidence aligned to the assembled genome.

to systematically identify and validate pseudogenes by carefully profiling expression evidence over the whole genome. In this study, we applied the method on the Ensembl gene annotation and the results are presented.

MATERIALS AND METHODS

Sequence data

We obtained the human genome assembly Build 35 from NCBI (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.35.1/). The 25 assembled chromosome sequences and 85 NT sequences were used as a reference genome. The January 2005 release of the human mRNA and EST sequences were extracted from the NCBI repositories: 22 887 sequences from a curated subset of the RefSeq experimental mRNA database (24), 195 073 sequences from GenBank mRNA database (25) and 6 020 341 sequences parsed from the dbEST database (26). Human SWISS-PROT protein sequences, a subset of protein sequences with high quality annotation, were obtained from the SWISS-PROT database in October 2004 (27) (http://www.ebi.ac.uk/swissprot/) and 11 777 sequences were obtained.

Gene sets

The Ensembl gene (22) release 27.35a.1 was extracted from Ensembl database (http://www.ensembl.org/Homo_sapiens/). There are 22 216 functional genes corresponding to 33 860 transcripts in the categories of known and novel, and 1978 pseudogenes in this release. A Vega pseudogene set on chromosomes 9 and 10, which was manually annotated and curated by the international vertebrate genome annotation (VEGA) project (28), was provided by the Vega project leader Dr Ashurst through website http://vega.sanger.ac.uk/. A total of 1031 pseudogenes were obtained from Vega.

Transcript-based sequence alignments

ESTmapper (29) was used to align EST, cDNA, Ensembl and Vega transcripts onto human Build 35 genome. ESTmapper uses a hash-index of 20mers in the genome to quickly locate areas of the genome likely to contain the query, then invokes the core of the sim4 algorithm (30) to produce a spliced alignment between the query and each genomic region selected. Different cut-offs were applied to retain alignments from mRNA and EST sequences due to their different qualities. The retained alignments for all alignments must contain 50% or more of the original sequence. For RefSeq and GenBank mRNA sequences, alignments that have at least 70% of sequence identity were retained; whereas, this sequence identity cut-off was raised to 95% for EST sequences. These less stringent thresholds used for full-length cDNA sequences allow additional alignments on locations other than its source gene. The average number of alignment per sequence was 1.61, 19.87 and 1.38 for RefSeq, GenBank mRNA and EST, respectively. The overall retention rate was 99.93, 95.03 and 84.55%, respectively.

ESTmapper was used to map Ensembl and Vega transcript sequences onto human Build 35 genome because the exon structure information was not provided in the original gene FASTA files downloaded from the websites, but was needed in this study to establish relationships between evidence and genes. When there were multiple alignments for a single gene, the alignment whose coordinates from ESTmapper that overlap with the ones reported in the original files was used. Because the genome used by Vega has incorporated 34 additional clones, the chromosomal locations of Vega pseudogenes do not always correspond to the genome used in this study (Dr Ashurst, personal communication). As a result, only 667 out of 1031 Vega pseudogenes unambiguously overlap their reported coordinates with ESTmapper coordinates. For Ensembl genes, 22 131 out of 22 216 functional genes and 1976 out of 1978 pseudogenes were mapped onto the Build 35 genome with consistent locations.

Protein sequence alignments

Protein sequences from SWISS-PROT were mapped onto human Build 35 genome using a combined method of TBLASTN (31) and GeneWise (32). Sequences were initially searched against the genome using TBLASTN with an expectation score of $<1 \times 10^{-10}$. TBLASTN generated one or more alignments for each protein-coding exon, which identified the approximate genomic locations of the putative exons. The aligned genomic sequences were extracted with additional 100 bp sequence on each side and joined together in the order of the original protein sequence (see Figure 2). If multiple sets of alignments were generated, as seen in gene clusters, then each member of the set would have its own extracted sequence defined by the protein sequence. Then GeneWise was run on this extracted sequence to produce the final protein alignments and report frameshifts and in-frame stop codons when detected. This two-step process

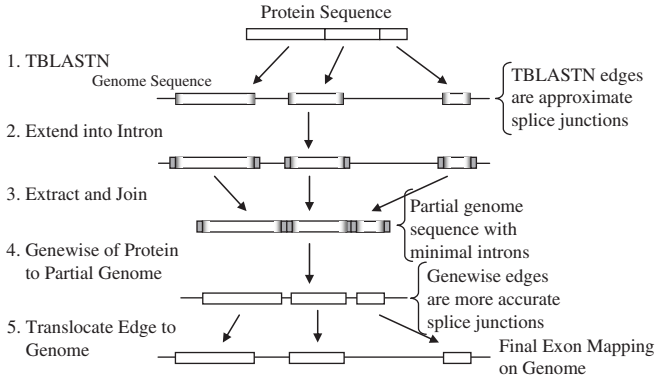

Figure 2. Composite mapping of protein sequence. Mapping of protein sequence to genome by linking segments of TBLASTN results prior to Genewise alignment.

greatly reduced genomic search space for GeneWise while not losing any structure information. The overall mapping rate of the SWISS-PROT sequences was 95.03% with an average of 5.73 alignments per sequence. The mapping statistics are summarized in Table 1.

Expression evidence profiles

We constructed transcript (mRNA, EST) or protein expression profiles in three steps. Step one identifies a best hit for every sequence aligned on the genome. The minimum requirement for a best hit is ⩾98% identity with the genomic sequence on ⩾90% of the original sequence coverage. By definition, a unique hit that passes the initial quality thresholds is always a best hit. Additional rules were applied to further narrow down the selection to one single alignment, in the order of: identity percentage, splicing status (spliced preferred over non-spliced) and coverage percentage. In rare cases, there is more than one alignment showing the same match statistics, then we would consider all of them as best hits. In addition, there are 240 GenBank mRNAs and 6 RefSeq mRNAs already annotated and marked as pseudogenes. All the alignments from these known pseudogenes are considered as non-best hits. For protein sequences, we first computed a frameshift rate for each GeneWise alignment. The frameshift rate is defined as the sum of frameshifts divided by the sum of the match length. The best hit was picked with the lowest frameshift rate. If more than one hit has the same lowest frameshift rate, then the same rules used for transcripts are adapted here in order to determine a single best hit. The percentage of unique best hits among all the best hits was 98.79, 98.59 and 98.27 and 97.76 for mapped RefSeqs, GenBank mRNAs, ESTs and SWISS-PROT proteins, respectively.

Step two establishes association relationships between evidence and genes through a collocation algorithm. This algorithm matches the two alignments by their genomic locations. If an evidence alignment overlaps with the exon regions of a gene on the same strand, this evidence is considered to collocate with the gene. This association does not require the two sequences to have sequence similarities or share the same splice junction.

Step three collects supporting evidence to construct expression evidence profiles. For a transcript expression profile, all the collocated alignments from EST, GenBank mRNA and RefSeq mRNA sequences are pooled together. Similarly, a protein expression profile is constructed for every gene that has support from SWISS-PROT protein sequences. The number of structure disruptions, i.e. frameshifts, presented in the alignments of supporting evidence is calculated in each profile.

Pseudogene identification

Pseudogenes were identified through the following three processes. First, genes with transcript profiles that do not contain any best hits, regardless of protein profile, are defined as non-transcribed pseudogenes. Second, genes with best hits from their transcript profiles, but the corresponding protein profiles lack best hit and contains frameshifts, are defined as transcribed pseudogenes. Third, genes without any transcript profile, but with a protein profile that lack best hit and contains frameshifts, are also considered as non-transcribed pseudogenes since no detectable transcript evidence strongly suggests that they have never been transcribed. Any remaining genes are not considered as pseudogenes due to lack of information.

Calibration of parameters

We applied the described process to identify pseudogenes in Celera gene annotation from human and mouse. Celera human annotation was based on human genome assembly release R27 and mouse annotation was based on mouse genome assembly release WGA3. Both gene sets were manually annotated by expert annotators and pseudogenes were carefully curated and flagged. Using these gene sets as reference, we varied and tested the criteria and thresholds used in selecting best hits and identifying pseudogenes, followed by thorough manual curation on the results. We found that the thresholds used in this study were consistently accurate for capturing annotated pseudogenes in both species with minimum false negatives and false positives.

RESULTS

Completeness of evidence

An important consideration for our approach is the completeness of transcript evidence to ensure coverage for the functional genes. If the evidence is incomplete, then some genes will not have primary cDNA coverage and could be potentially misclassified as pseudogenes. We addressed this problem by looking at successively increasing samples of existing ESTs and their impact on our pseudogene analysis. The full set of ESTs overlaps 20 536 functional and 1159 pseudogenes from Ensembl. We randomly selected five subsets of EST sequences with the number of sequences being 1, 2, 3, 4 and 5 million per subset. The number of genes supported by each subset increased from 18 757 for 1 million to 20 524 for 5 million, while the number of pseudogenes increased from 739 to 1158, respectively (see Table 2). Using criteria of either no best hit in the supporting evidence, or a complete lack of supporting ESTs, the number of

Table 1. Summary of gene and evidence mapping

Gene and evidence	Total sequences	Criteria	Mapped sequences	Percentage	Alignments per sequence
Ensembl functional genes	22 216	Location verification	22 131	99.62%	NA
Ensembl pseudogenes	1978	Location verification	1976	99.90%	NA
Vega chromosome 9 and 10 pseudogenes	1031	Location verification	667	64.69%	NA
Refseq	22 887	70% identity, 50% length	22 871	99.93%	1.61
GenBank mRNA	195 073	70% identity, 50% length	185 385	95.03%	19.87
EST	6 020 341	95% identity, 50% length	5 089 981	84.55%	1.38
Swiss-Prot	11 777	TBLASTN $<1 \times 10^{-10}$	11 192	95.03%	5.73

Table 2. Summary of EST profiling of Ensembl genes and pseudogenes

# of EST mapped	Overlapping Ensembl genes	Overlapping Ensembl pseudogenes	Non-transcribed Ensembl genes	Non-transcribed Ensembl pseudogenes	Total non-transcribed pseudogenes
1 million	18 757	739	2942	912	3854
2 million	19 780	911	1874	802	2676
3 million	20 161	1007	1388	725	2113
4 million	20 387	1090	1104	662	1766
5 million	20 524	1158	918	613	1531
Total	20 536	1159			

potential non-transcribed pseudogenes identified per subset decreased from 3854 to 1531. Plotting the potential pseudogenes, in Figure 3, we showed that this number decreases as the number of evidence increases, but it flattens out when the number of ESTs reaches 4 million. From the whole genome EST mapping, we also observed a limited number of genome regions covered by EST but not Ensembl. This is consistent with a recent study (33) aimed at assessing human protein-coding genes through longitudinal database surveys, which suggests that the human gene count has shown a static number at ~28 000. The EST coverage for human genes is approaching saturation and new sequence submissions are predominantly extending known genes or sampling new splice variants.

Validation of accuracy and sensitivity

To evaluate the performance of our process, we tested its discriminative power on two sets of pre-defined pseudogenes: the manually annotated Vega pseudogenes from chromosomes 9 and 10 and the Ensembl pseudogene predictions.

Vega pseudogene validation

From the 667 Vega pseudogenes consistently located on the NCBI Build 35 genome, transcript expression profiles were constructable for 469 genes. In these 469 genes: 266 were not supported by any best hits and, therefore, qualified as non-transcribed pseudogenes; 150 were collocated with protein evidence from SWISS-PROT, but the protein profiles suggested 92 of them are not translated by our definition. In the remaining 198 genes without transcript evidence, we were able to construct protein profiles for 120, of which 93 met the criteria for pseudogenes. In total, 451 (266 + 92 + 93) out of the 536 genes that were supported by any expression evidence are confirmed as pseudogenes. The overall validation rate for mapped Vega pseudogene is 76.57%. These numbers are summarized in Table 3. We excluded from our analysis the 78 genes for which we could not construct either expression profiles. After manual review, we established that these pseudogenes were annotated based on protein evidence not included in our snapshot of the SWISS-PROT database.

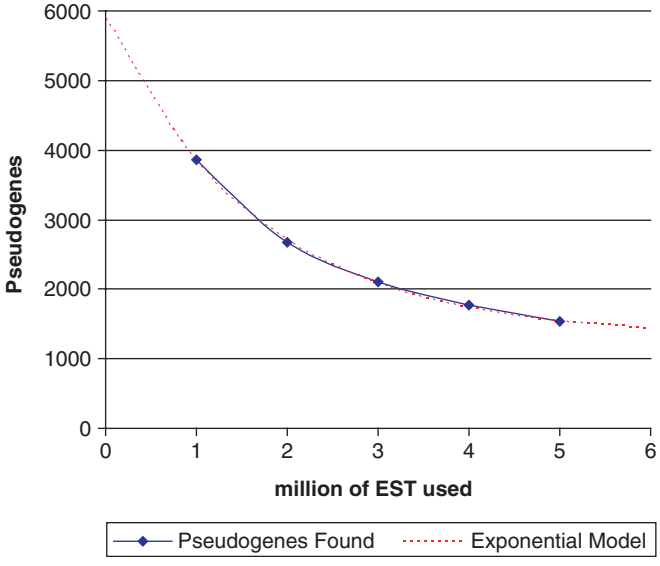

Figure 3. Potential pseudogenes as a function of EST evidence used. Total potential pseudogenes without best hits or lack of any EST hits plotted against the amount of EST evidence used.

The Vega gene set contains both processed and non-processed pseudogenes. For the purpose of verifying the sensitivity of our method towards the detection of non-processed pseudogenes, we classified genes that lack introns as processed pseudogenes, while the rest as the non-processed, although some of the multiple-exon forms may actually represent partially processed pseudogenes (1). Of the 667 pseudogenes from the original dataset, 153 contain intronic sequences representing 23% non-processed pseudogenes. Of the 451 genes confirmed as pseudogenes by our method, 85 are non-processed, which represents 19% of the validated

Table 3. Summary statistics for the validation of Vega and Ensembl pseudogenes

Categories	Vega #	% of respective category	% of total	Ensembl #	% of respective category	% of total
Original	1031			1978		
Aligned with consistent locations	667			1974		
cDNA support	469			1611		
No cDNA best hits (non-transcribed pseudogene)	266	100.00	45.16	977	100.00	52.95
With cDNA best hits	203			642		
Swiss-Prot protein support	150			380		
No best hits & presence of frameshifts (transcribed pseudogene)	92	61.33	15.62	311	81.84	16.86
No cDNA support	198			634		
Swiss-Prot protein support	120			234		
No best hits and presence of frameshifts (non-transcribed pseudogene)	93	77.50	15.79	212	90.60	11.49
No support from cDNA & Swiss-Prot	78			129		
Non-transcribed pseudogenes	78	NA	NA	129	NA	NA
Total genes supported by evidence	589			1845		
Total pseudogenes validated	451	NA	76.57	1500	NA	81.30

pseudogenes, a percentage consistent with the overall proportion of non-processed pseudogenes (23%) in the original dataset. This indicates that our approach can detect both processed and non-processed pseudogenes without bias.

Ensembl pseudogene validation

We applied the same validation procedure to the processed pseudogenes from Ensembl prediction. The results are shown in Table 3. We obtained consistent mapping of 1974 out of 1978 Ensembl pseudogenes. The transcript (EST and mRNA) expression profiles for 1611 pseudogenes were constructed. Of these, 977 genes were considered as non-transcribed pseudogenes due to lack of best hits. For the remaining 634 genes potentially active at transcription, we could only construct protein expression profiles for 380 genes based on SWISS-PROT sequences. Of these, 311 were identified as transcribed pseudogenes. There are an additional 234 genes supported only by protein evidence. Of these, 212 genes passed the pseudogene thresholds. In total 1500 (977 + 311 + 212) are non-evidential at either the transcription or translation level. The overall validation rate is 81.30% (1500/1845).

Validation statistics

We achieved a validation rate of 76.57 and 81.30% on Vega and Ensembl pseudogenes, respectively. However, this rate was calculated without considering the fact that the protein sequences used for profiling in this study were just a subset of the dataset used by Vega and Ensembl annotation process. As a result, some of the genes do not show any protein evidence in our analysis. If our protein data were expanded to include sequences in TrEMBL or NCBI NRAA, the validation rate would be higher since more genes could be subjected to evaluation. However, these protein datasets would also generate more false positives due to the inclusion of computational predicted translations. There were 53 and 262 genes in Vega and Ensembl, respectively, which were supported by best hits of transcript evidence, but not covered by any protein from SWISS-PROT. If these gene counts were factored out, the validation rate would be increased to 84 and 94% for Vega and Ensembl, respectively. Such sensitivity provides confidence in our method to identify true pseudogenes.

Our analysis depends on an accurate alignment of evidence sequences, which determines the assignment of best hits. There is a potential for error in a small number of ESTmapper and GeneWise alignments, typically less than 1 out of 10^5 bases mapped due to algorithmic biases (Peter Li, unpublished data). In addition, true variations between mRNA sequences and the finished human genome such as polymorphisms can lead to frame shifts (34) and incorrect assignment of best hits. These allele-specific pseudogenes should still be considered as pseudogenes (35,36) and do not invalidate the premise behind our approach.

We estimate 10–20% of the genes are potentially false negatives (pseudogenes wrongly inferred as genes) by our method, because we cannot exclude the possibility that the original protein evidence might be partial or might represent defective proteins. Although they qualify as best hits, they do not represent real expression products. Such low quality sequences directly affect the outcomes of gene annotation. Extra effort should be paid to identify and exclude them before being used as evidence. On the other hand, the fact that these genes (potentially false negatives by our method) present transcriptional activities, and possibly translational activities or intact structure without disruptions may cast a doubt on their pseudogene status previously assigned by their respective sources. To estimate the rate of false positives (genes wrongly inferred as pseudogenes), we examined the asymptotic limit of ESTs to identify pseudogenes. Under the assumption that as we enlarge the mapped EST set, more candidate pseudogenes will overlap a best hit EST and thus convert to a transcribed gene. From Figure 3, the decrease of pseudogenes asymptotically approaches to 1310 using an exponential probability model of 'pseudogene failures', this gives us an upper bound of false positive rate at 15%. An independent estimate of 10% came from manual curation of the pseudogenes identified by this pipeline on the Celera human gene set. In the opinion of the

expert annotators, the major sources of false positives were: (i) error in the assignment of best hits mostly due to the discrepancies between the evidence sequence and the genome sequence and (ii) evidence that is not included in the pipeline, e.g. sequences from NRAA or TrEMBL. The feedback from manual curation had been incorporated into the algorithm to calibrate the parameters to minimize false positives.

Identification of pseudogenes from ensembl genes

Ensembl human annotation release 27.35a.1 contains 22 216 genes in the categories of known and novel as functional genes. We retained 22 131 genes after mapping to B35 and verifying location consistency with their original coordinates. Transcript expression profiles can be constructed for 21 655 genes. Of these, 1018 were considered as non-transcribed pseudogenes due to lack of best hits in their profiles. An additional 412 genes only have protein evidence. After checking properties of: (i) lack of best hits and (ii) presence of frameshifts, 182 genes were identified as non-transcribed pseudogenes. We then examined the group of genes with best hits from mRNA or EST evidence. There are 20 637 genes in this category. Of these, 13 803 genes had protein profiles established. From these, 811 genes were not supported by any best hit from protein and contained frameshifts. We consider these 811 genes as transcribed pseudogenes since we have expression evidence at transcription, but no evidence of intact translation. However, a genome assembly without sequence or assembly errors may be required before this interpretation can be made with more certainty. This group of pseudogenes are currently considered as non-functional at translation, but with recognition that they are potentially functional and may have to be re-evaluated in the future.

We concluded that 2011 (1018 + 182 + 811) Ensembl predicted genes have been identified as pseudogenes based on expression evidence profiling (The gene list is available in the Supplementary Data). Of these, 1200 genes are non-transcriptional and 811 genes are non-translational. This number represents 9.1% (2011/22 067) of the total Ensembl genes. Checking the exon structure, the number of pseudogenes with more than one and two exons is 1341 (66.68%) and 815 (40.53%), respectively, suggesting at least 40% of them are non- or partially processed pseudogenes.

Evaluation of pseudogenes identified from ensembl genes

To evaluate the accuracy of our methods, we conducted in-depth analysis of the pseudogenes identified above to evaluate whether their functional parental genes exist and to understand the underlying mechanisms for pseudogene origination by comparing exon structures between each pair of pseudogene and their respective parent. A subset of 526 genes from this group was selected because they are supported by evidence from RefSeq mRNA sequences. The genomic alignments of these RefSeq sequences were used to track and identify parental genes. To qualify for a parental gene, we require that the transcript of this gene must be supported by the best hits of the RefSeq sequences and covers at least 50% of the transcript length of the respective pseudogene. Based on the above criteria, 288 genes had putative parental genes identified. These parental genes were Ensembl genes in the categories of known or novel and located somewhere else in the genome. Of the 288 pseudogenes, 239 were assigned to a unique parental gene while 49 were associated to more than one paralogous parent. Due to a very high degree of sequence identity among those paralogs, it is impossible to distinguish which one is the real parent of the respective pseudogene.

We then compared the exon structure between each pair of genes, and divided the gene pairs into four different groups. Group A contains the gene pairs in which the pseudogene is single-exon and the parental gene is multi-exon, representative of the 'processed pseudogenes from retrotransposition' model. Group B contains the gene pairs that both genes lack intronic sequences. It is difficult to determine which mechanism was responsible for the formation of these pseudogenes if simply judged by the exon structure. Group C contains the gene pairs in which both genes are spliced and the parental gene has an equal or a greater number of exons than the pseudogene. In Group D, the pseudogenes contain a single putative intron that does not exist in the respective parental genes. Table 4 summarized the detailed statistics for the four groups described above. Based on the exon structure and the result of comparison between pseudogenes and parental genes, it is most likely that the 112 genes in the group C are non-processed pseudogenes. This number represents 38.89% of the 288 pseudogenes with putative parental genes identified. The percentage is consistent with the calculation from the previous section based on the total identified pseudogenes in which the fraction of non-processed pseudogenes is 40.53% when considering genes with more than two exons as the non-processed.

Group D represents some interesting gene pairs that were investigated further. We found that, in most of the cases, the putative intron present in the pseudogene appears to be the result of the insertion of transposons or other DNA sequence after retrotransposition. However, we also discovered cases where the intron was artificially created by the prediction algorithm. For instance, gene ENSG00000188712 was collocated with RefSeq NM_001004484 and several other protein sequences from SWISS-PROT. All of the protein sequences contain a frameshift that resides within the region where the intron locates. This artificial intron bypasses the defect so that the gene does not appear to be disrupted

Table 4. Comparison of exon structures between Ensembl pseudogenes and their parental genes from the 288 identified pairs

Categories	Pairs of genes Total	Relationship of pseudo to parent	
		One-to-one	One-to-many
Group A: pseudogene single-exon, parent multi-exon	109 (37.85%)	102	7
Group B: both single-exon	42 (14.58%)	27	15
Group C: both multi-exon	112 (38.89%)	96	16
Group D: pseudogene 2-exon, parent single-exon	25 (8.68%)	14	11
Total	288	239	49

Table 5. Functional classification of pseudogenes

Name	Number
Olfactory receptor	53
Keratin type I/II	24
Immunoglobulin	21
Peptidyl-prolyl *cis-trans* isomerase	17
Heterogeneous nuclear ribonucleoprotein	16
HMGI/II (high mobility group)	14
Dynein heavy chain	12
Nucleophosmin	12
40S ribosomal S2	11
Elongation factor 1 alpha	10

in structure. This result demonstrates that our approach is able to capture this type of 'disrupted' processed pseudogenes (by natural or artificial means) that are otherwise missed by methods that rely on the absence of introns.

Functional classification of pseudogenes

We grouped the 2011 pseudogenes using Ensembl protein family classification (http://www.ensembl.org/). Table 5 summarizes the top 10 functional classes. These 10 classes represent multigene families and many of them are highly expressed. All of them have been previously identified as having a large number of pseudogenes in human (11,12). Other than 308 (15.3%) unclassified pseudogenes, the most frequent pseudogenes come from ribosomal protein genes, for which we found 137 (6.8%) copies of the combined set of all types of ribosomal proteins. A key glycolytic enzyme involved in energy production, Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), has been reported to have more than 400 processed GAPDH pseudogenes in mouse (20). In human, however, results from previous studies were not consistent. One study (12) identified 78 processed pseudogenes while the other (11) did not find any. Through our method, we identified four GAPDH pseudogenes (ENSG00000163410, ENSG00000188796, ENSG00000188885 and ENSG00000183299). All of these four genes are partial compared with the functional counterpart. They are in the category of novel and have no orthologs from mouse or rat according to Ensembl. The first two genes have a single-exon and the latter two have four exons each, but with suspiciously small intron sizes (2–5 bp) for the majority of the introns. Another gene family, Cytochrome P450 (*CYP*), is relatively pseudogene-rich with 58 known pseudogenes in human, but these pseudogenes are sometimes hard to identify due to their almost intact structures (37). In fact, only one *CYP* pseudogene was identified among ~12 000 annotated pseudogenes (http://bioinfo.mbb.yale.edu/genome/pseudogene/human-all/index.html). We identified eight copies, and six of them are full-length or nearly full-length pseudogenes with multi-exon (exon number ranges from 5 to 12) structures. The remaining two genes, ENSG00000198461 and ENSG00000130612, have only 1 and 2 exons with a length of 132 and 309 bp, respectively, representing the detritus exons type of pseudogenes (37). In addition, we identified ENSG00000184235 as a transcribed protein tyrosine phosphatase (PTP) pseudogene in agreement with a recent study on PTPs (18).

DISCUSSION

We presented a novel method for the systematic identification and validation of pseudogenes from a given set of annotation. Because the annotated gene sets are derived from evidence, our process reevaluates the relationship of the genes and their expression evidence, such as ESTs, mRNAs and proteins, and assesses their functionality through a detailed profiling of supporting evidence in a whole genome-scale. This global view of properly assigning evidence to the gene overcomes the typical shortcomings of gene annotation from the local pattern of supporting transcription and translation evidence. Consequently, we can infer whether a gene is functional or disabled, and at which level during the process of gene expression with a mechanism that is universally applicable to all types of pseudogenes regardless of their individual structure and sequence features, and the synonymous and non-synonymous nucleotide substitution rates. To our knowledge, this is the first report that uses whole genome expression evidence to systematically identify pseudogenes through a computational approach. This method was designed as a post-processing step following computational gene annotation to identify potential incorrect annotations and facilitate the subsequent efforts with manual curation. Because the annotated genes form the foundation for the subsequent experimental design and computational work on a genome, it is critical that high quality annotation was established from the outset.

The strength of our method depends on the relative completeness of transcriptional and translational evidence. For species with deep cDNA and protein coverage, such as human and mouse, this method would be appropriate to validate gene annotations. However, for species with limited evidence, it may not be appropriate. In addition, the inherent error rate of transcriptome sequencing is an important factor in the interpretation of transcript alignments on the genome: the best alignment of a given transcript may not be identifiable if there are too many errors. This will become more critical in the future when researchers apply the next generation of sequencing technology to the transcriptome (38) whose the initial error rate might be higher than the intrinsic polymorphism or mutation rate. While this is resolvable by constructing consensus genome sequence from multiple copies, it raises the risk of collapsing transcripts from different paralogous genes in transcriptome sequencing.

The method of whole genome expression sequence profiling contains two important features. First, it is equally powerful for the identification of processed and non-processed pseudogenes as demonstrated by validating manually curated Vega pseudogenes. The latter type of pseudogenes has been thought to be a major source of annotation errors (16,33). This is proved to be true from our results since 40% of the 2011 pseudogenes identified from Ensembl genes are non-processed pseudogenes with multi-exon gene structures. Second, our method is able to identify at which level the loss of gene expression most likely occurs. This ability allows the detection of transcriptionally active pseudogenes, which represent a challenge to the annotation process because of the existence of both homology and expression evidence. Additionally, undetected transcribed pseudogenes may lead to misinterpretation of experimental results from both gene

expression and genotyping assays and from microarrays intended for functional genes. We discovered through the whole genome expression profiling that 14% (92/667) and 16% (311/1974) of the pseudogenes from Vega and Ensembl, respectively, have supporting expression evidence from cDNA and EST sequences. These numbers are well in line with data previously reported. This significant percentage of transcribed pseudogenes deserves more attention from assay and array developers because it makes the primer and probe selection more difficult in order to ensure that they precisely amplify the expected gene product. Correctly characterizing all the pseudogenes allows accurate design of locus-specific assays and microarrays for functional genes, many of which are clinically important (39,40). Further studies on transcribed pseudogenes will add to our understanding of their potential roles as non-coding RNA genes or other new types of functional elements.

SUPPLEMENTARY DATA

Supplemenetary Data are available at NAR online.

ACKNOWLEDGEMENTS

The authors thank Richard Mural, Weiniu Gan and Gennady Merkulov for helpful suggestions and discussions. The authors would also like to thank Eugene Spier for his initial request of a cleaner gene sets for assay design that inspired this study. Funding to pay the Open Access publication charges for this article was provided by Applied Biosystems, Inc.

Conflict of interest statement. None declared.

REFERENCES

1. Vanin,E.F. (1985) Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.*, **19**, 253–272.
2. Mighell,A.J., Smith,N.R., Robinson,P.A. and Markham,A.F. (2000) Vertebrate pseudogenes. *FEBS Lett.*, **468**, 109–114.
3. Guo,N., Mogues,T., Weremowicz,S., Morton,C.C. and Sastry,K.N. (1998) The human ortholog of rhesus mannose-binding protein-A gene is an expressed pseudogene that localizes to chromosome 10. *Mamm. Genome*, **9**, 246–249.
4. Balakirev,E.S. and Ayala,F.J. (2003) Pseudogenes: are they 'junk' or functional DNA? *Annu. Rev. Genet.*, **37**, 123–151.
5. Boger,E.T., Sellers,J.R. and Friedman,T.B. (2001) Human myosin XVBP is a transcribed pseudogene. *J. Muscle Res. Cell. Motil.*, **22**, 477–483.
6. Edgar,A.J. (2002) The human L-threonine 3-dehydrogenase gene is an expressed pseudogene. *BMC Genet.*, **3**, 18–31.
7. Hirotsune,S., Yoshida,N., Chen,A., Garrett,L., Sugiyama,F., Takahashi,S., Yagami,K., Wynshaw-Boris,A. and Yoshik,A. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*, **423**, 91–96.
8. 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.
9. Yousef,G.M., Borgono,C.A. and Diamandis,E.P. (2004) Cloning of a kallikrein pseudogene. *Clin. Biochem.*, **37**, 961–967.
10. Berger,I.R., Buschbeck,M., Bange,J. and Ullrich,A. (2005) Identification of a transcriptionally active hVH-5 pseudogene on 10q22.2. *Cancer Genet. Cytogenet.*, **159**, 155–159.
11. Torrents,D., Suyama,M., Zdobnov,E. and Bork,P. (2003) A genome-wide survey of human pseudogenes. *Genome Res.*, **13**, 2559–2567.
12. Zhang,Z., Harrison,P.M., Liu,Y. and Gerstein,M. (2003) Millions of years of evolution reserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.*, **13**, 2541–2558.
13. Zhang,Z. and Gerstein,M. (2004) Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.*, **14**, 328–335.
14. Khelifi,A., Duret,L. and Mouchiroud,D. (2005) HOPPSIGEN: a database of human and mouse processed pseudogenes. *Nucleic Acids Res.*, **33**, D59–D66.
15. Ohshima,K., Hattori,M., Yada,T., Gojobori,T., Sakaki,Y. and 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.
16. Harrison,P.M., Hegyi,H., Balasubramanian,S., Luscombe,N.M., Bertone,P., Echols,N., Johnson,T. and Gerstein,M. (2002) Molecular fossils in the humen genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.*, **12**, 272–280.
17. Yeh,R.F., Lim,L.P. and Burge,C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.*, **11**, 803–816.
18. Andersen,J.N., Del Vecchio,R.L., Kannan,N., Gergel,J., Neuwald,A.F. and Tonks,N.K. (2005) Computational analysis of protein tyrosine phosphatases: practical guide to bioinformatics and data resources. *Methods*, **35**, 90–114.
19. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
20. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
21. Mounsey,A., Bauer,P. and Hope,I.A. (2002) Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. *Genome Res.*, **12**, 770–775.
22. Curwen,V., Eyras,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M.J. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.
23. Nelson,D.R. (2004) 'Frankenstein genes', or the Mad Magazine version of the human pseudogenome. *Hum. Genomics*, **1**, 310–316.
24. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
25. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, 23–26.
26. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST–database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
27. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
28. Ashurst,J.L., Chen,C.K., Gilbert,J.G.R., Jekosch,K., Keenan,S., Meidl,P., Searle,S.M., Stalker,J., Storey,R., Trevanion,S. *et al.* (2005) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **33**, D459–D465.
29. Florea,L., Di Francesco,V., Miller,J., Turner,R., Yao,A., Harris,M., Walenz,B., Mobarry,C., Merkulov,G., Charlab,R. *et al.* (2005) Gene and alternative splicing annotation with AIR. *Genome Res.*, **15**, 54–66.
30. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
31. 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.
32. Birney,E., Clamp,M. and Durbin,R. (2004) Genewise and genomewise. *Genome Res.*, **14**, 942–950.
33. Southan,C. (2004) Has the yo-yo stopped? An assessment of human protein-coding gene number. *Proteomics*, **4**, 1712–1726.
34. Furey,T.S., Diekhans,M., Lu,Y., Graves,T.A., Oddy,L., Randall-Maher,J., Hillier,L.W., Wilson,R.K. and Haussler,D. (2004) Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing. *Genome Res.*, **14**, 2034–2040.
35. Pai,H.V., Kommaddi,R.P., Chinta,S.J., Mori,T., Boyd,M.R. and Ravindranath,V. (2004) A frameshift mutation and alternate splicing in human brain generate a functional form of the pseudogene cytochrome P4502D7 that demethylates codeine to morphine. *J. Biol. Chem.*, **279**, 27383–27389.

36. Hollyoake,M., Campbell,R.D. and Aguado,B. (2005) NKp30 (NCR3) is a pseudogene in 12 inbred and wild mouse strains, but an expressed gene in *Mus caroli*. *Mol. Biol. Evol*., **22**, 1661–1672.
37. Nelson,D.R., Zeldin,D.C., Hoffman,S.M.G., Maltais,L.J., Wain,H.M. and Nebert,D.W. (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics*, **14**, 1–18.
38. Metzker,M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res*., **15**, 1767–1776.
39. Ruud,P., Fodstad,O. and Hovig,E. (1999) Identification of a novel cytokeratin 19 pseudogene that may interfere with reverse transcriptase-polymerase chain reaction assays used to detect micrometastatic tumor cells. *Int. J. Cancer*, **80**, 119–125.
40. Harper,L.V., Hilton,A.C. and Jones,A.F. (2003) RT–PCR for the pseudogene-free amplification of the glyceraldehyde-3-phosphate dehydrogenase gene (gapd). *Mol. Cell. Probes*, **17**, 261–265.