JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 5, 2010
© Mary Ann Liebert, Inc.
Pp. 755–765
DOI: 10.1089/cmb.2009.0027

A New Approach for the Identification of Processed Pseudogenes

*IVAN MOLINERIS,[1] *GABRIELE SALES,[1] FEDERICO BIANCHI,[2]
FERDINANDO DI CUNTO,[2] and MICHELE CASELLE[1]

ABSTRACT

Processed pseudogenes are DNA sequences generated through reverse transcription (RT) and retrotransposition of mature mRNAs. These sequences are usually considered junk DNA, since in most cases they lack a suitable promoter and are no longer transcribed. Nonetheless, due to their origin, they represent a valuable source of information on the transcriptome, which becomes particularly interesting for organisms lacking large EST collections. Here, we describe REtrotransposed Gene EXPlorer (REGEXP), a new method for the systematic identification of retrotransposition events that, unlike existing approaches, does not rely on *a priori* knowledge of mRNA sequences. Using our pipeline, we were able to identify 2288 processed pseudogenes in the human genome, showing a good overlap with the ENSEMBL, VEGA, and pseudogene.org datasets. These pseudogenes could be traced back to 987 genes, mostly corresponding to already known genes. In many cases, we recovered the signature of additional exons, likely due to alternative splicing. Interestingly, some of our predictions did not match previously known or predicted genes, and we were able to validate most of them by RT–polymerase chain reaction (PCR). Similar results were obtained with the mouse genome. Our data show that the REGEXP method is capable of identifying processed pseudogenes and to predict most of the corresponding genes with high specificity. Therefore, it may represent a valuable integration to the current genome annotation pipelines.

Key words: algorithms, alignment, computational molecular biology, distance geometry, gene networks, percolation theory, sequence analysis.

1. INTRODUCTION

Pseudogenes are currently defined as nonfunctional copies of genes, originating either from segmental genome duplications or from retrotransposition events. Nevertheless, it has been recently recognized that pseudogenes may play a crucial role in various stages of gene regulation and in particular in fine-tuning the expression of their parent genes (Watanabe et al., 2008; Tam et al., 2008; Korneev et al., 1999; Weil et al., 1997; Zhou et al., 1992; Hirotsune et al., 2003). Developing precise knowledge of the content of pseudogenes is thus required to fully understand the structure and functions of a genome (Zheng et al., 2007).

*These two authors contributed equally to the work.

[1]Theoretical Physics Department and [2]Molecular Biotechnology Center, Universit di Torino, Torino, Italy.

The identification of pseudogenes is also very challenging from a computational point of view, and it is perfectly suited for bioinfomatic methods. Several pseudogene databases exist in the literature (Torrents et al., 2003; Suyama et al., 2006; Pavlicek et al., 2006; Yao et al., 2006; Shemesh et al., 2006; Karro et al., 2007; Ortutay and Vihinen, 2008; Khelifi et al., 2005; Ohshima et al., 2003). Although these methods are based on different computational pipelines, they all share the need for substantial information on the transcriptome and/or proteome of the organism, in addition to the genomic sequence.

These tools perform rather well with genomes for which a large amount of ESTs and/or protein information exist, but these tools must rely on phylogenetic conservation in other cases and are expected to perform poorly on fast-evolving or non-canonical genes.

Among the different pseudogenes, a particularly interesting class is processed pseudogenes (PPGs), where copies of cellular RNAs typically contain poly(A) and lack introns, which were reverse-transcribed and inserted into the genome by L1/LINE1 retrotransposons (Esnault et al., 2000).

PPGs exist in most of the higher eukaryotes, although their number can vary by orders of magnitude. Indeed, whereas thousands of PPGs are present in the mouse and human genomes, the *Caenorhabditis elegans* genome contains only 208 processed pseudogenes (Harrison et al., 2001), the chicken genome contains at most 51 PPGs (Hillier et al., 2004), and the *Drosophila melanogaster* genome contains at most 34 PPGs (Harrison et al., 2003).

While the rate of the emergence of PPGs in mammalians is about 1–2% per gene per million years, this seems to have drastically decreased in the hominid lineage (Sakai et al., 2007). For comparison, the rate of gene duplication in the human genome is about 0.9% per gene per million years.

Since they are derived from a mature mRNA product, PPGs lack the upstream promoters of normal genes; thus, they are usually "dead on arrival," becoming non-functional pseudogenes. Nevertheless, according to a recent study (Sakai et al., 2007), about 1% of PPGs in human shows evidence of transcriptional activity. Moreover, compared to their parent sequence, PPGs are often truncated at their 5′ end, probably as a result of the relatively nonprocessive mechanism that creates them (Pavlícek et al., 2002).

Since PPGs may derive from normal protein-coding mRNAs, alternatively spliced mRNAs (Karro et al., 2007), non-protein-coding RNAs (Jurka et al., 1988) and antisense transcripts (Ejima and Yang, 2003), they may represent a rich sample of the transcriptome (Pavlicek et al., 2006), although they cannot be expected to completely cover it.

In this article, we propose REtrotransposed Gene EXPlorer (REGEXP), a new approach for the identification of gene-PPG pairs based on the DNA sequence only, with no need for additional information on EST or proteins. Using this approach, we have identified 2288 PPGs in human and 2063 in mouse, corresponding to 987 human and 709 mouse parent genes. Interestingly, although most of the parent genes were already known or supported by EST tracks, in a few cases they were completely new predictions, not supported by any type of evidence in the UCSC or ENSEMBL databases. Importantly, we were able to experimentally validate some of these predictions. We conclude that, even though our method is perfectly suited for organisms for which only the sequence is known, it can lead to the identification of new genes even in extensively annotated genomes.

2. METHODS

2.1. The alignment database

We start from the full set of local alignments found by comparing the repeat masked sequence of the human genome (build 36) with itself; we compute these alignments with the Megablast software (Zhang et al., 2000).

To avoid excessive memory occupation, we split the chromosome sequences into smaller fragments and compare them all. We perform a sequence split when we find a repeat masked region longer than 1000 base pairs (usually a LINE); thus, we don't need to postprocess the alignments to merge overlapping fragments. We are confident that no alignment containing a masked region of 1000 bps or more can exist since its score would be under any reasonable statistical cutoff. The alignment database contains about 12 million high scoring pairs (HSPs; pairs of regions sharing high sequence similarity) longer than 30 bps.

We label each HSP a using two aligned regions r_{a1} and r_{a2}, which are identified by their starting and ending points in absolute chromosomal coordinates. This induces a natural definition of distance between HSPs $a = (r_{a1}, r_{a2})$ and $b = (r_{b1}, r_{b2})$ as the length of the smallest segment joining two endpoints, i.e.,

$d(a,b) = \min(d(r_{a1}, r_{b1}), d(r_{a1}, r_{b2}), d(r_{a2}, r_{b1}), d(r_{a2}, r_{b2}))$, where $d(r_{ai},\ r_{bj})$ is the euclidean distance between two points, and $d(a,\ b) = \infty$ if they are on different chromosomes.

2.2. Location clusters

Since a processed pseudogene is the union of the exons of the original gene, one would expect to find it in the alignment database looking for clusters of nearby HSPs. On one side of the alignment (the "pseudogene side"), we expect multiple HSPs very close to each other (ideally, if no insertion occurred after the retrotransposition event they should be contiguous); on the other side (the "gene side"), they will be near, but separated by gaps corresponding to the introns that are missing from the pseudogene. Even if we allow for the presence of mutations in one or both the sequences, the scenario remains similar. Some of the original HSPs may now have a lower score, some may as well have disappeared, but the picture still consists of a number of HSPs clumped one next to the other. To extract these HSP clusters (which we shall denote in the following as "location clusters") from the alignment database, we developed the following clustering procedure. Each HSP can be represented as a segment in the bidimensional plane spanned by the two sequences (in a way that closely resembles dot-plots); we cluster together two consecutive alignments/segments if the distance between the two segments is lower then a certain threshold (we chose 22Kbps because only 5% of known human introns are longer than that) along both directions. If at least three of these segments are concatenated together, we consider the resulting group a location cluster.

As a result of this definition, each location cluster can be considered as the bidimensional bounding box of a set of at least three nearby segments, and any two location clusters are separated both horizontally and vertically by more than 22Kbps.

These location clusters are the starting point of our analysis. The remaining part of the computational pipeline is devoted to refine them and to filter out those that do not conform to certain requirements. We consider each location cluster surviving the entire filtering process as a candidate gene-pseudogene pair.

2.3. Corruption gaps

In some cases, processed pseudogenes may have accumulated so many mutations that only a small portion of the original duplicated region can be retrieved using a standard alignment algorithm. Typically, this lack of homology with the original sequence shows up as a series of gaps in the alignment cluster: we call them "corruption gaps." Our goal is to separate these gaps from those due to intron splicing.

To identify corruption gaps, we use the HSPs as anchors (each HSP can have itself small gaps, as a consequence of standard alignment algorithms, but we ignore them during this filtering phase).

As mentioned above, each alignment can be represented as a segment on the cartesian plane having as x and y axes the two genomic regions. Similarly to what happens in dot-plot graphs, these segments lie on lines with angular coefficent exactly ± 1 if there are no gaps in the HSP (the sign of the angular coefficient depends on the strand of the alignment). Given that we use a scoring system penalizing gaps, the angular coefficient of segments representing an HSP is always near ± 1.

We join two HSPs, represented by segments a and b, with a new segment c (that we define a "corruption gap") if the distance $d(a,\ b)$ is smaller than 3000 bps and if the angular coefficent of c is 45 ± 5 degrees.

We chose the values of these parameters considering some exemplar cases; the final results are only slightly influenced by such values.

We call a set of high scoring pairs joined by corruption gaps a "diagonal": its projections on the two axes define two regions that are a candidate exon or pseudoexon (homologous of an exon in a pseudogene).

2.4. Splicing gaps

We expect to find another class of gaps in the alignment clusters: those deriving from the splicing of introns in the processed pseudogenes. These are of great importance for our identification process since they allow us to distinguish the original gene from its retrotransposed copy.

Introns in the mRNA of a gene are expected to be spliced before the retrotransposition event, so we expect to see candidate pseudoexons that are close together while the corresponding candidate exons are separated by gaps that we call "splicing gaps."

A splicing gap is found by looking at the geometrical distribution of diagonals: if the segment joining two diagonals has a projection on one of the two axes that is less than σ bps in length, while on the other axis the projection is larger than β bps, then we add this segment to the location cluster as a splicing gap.

We set the threshold β looking at the intron length distribution and choosing a value such that only the 5% of all introns are smaller than β bps; i.e., we expect to lose only 5% of true introns because of this cutoff. In the human case, the threshold turns out to be $\beta = 74$. The parameter σ accounts for the fuzziness of diagonals that may not be precise at the extremes; for this parameter, we use a value of 15.

We can project a splicing gap on both axes of the cartesian plane: we consider the longest projection as a candidate intron.

Another reason for which the identification of the splicing gaps is of crucial importance is that it allows us to separate the "true" processed pseudogenes from alignments (and possibly unprocessed pseudogenes) deriving from segmental duplications of the genome. To this end, we discard all location clusters without splicing gaps; to further reduce the number of false positives, we actually require the presence of at least three splicing gaps in each location cluster to continue its processing along the pipeline (in fact, only 4% of the human genes contain only one intron).

In some cases, it may happen that splicing gaps are found on both sides of a location cluster, for instance due to large repeat insertions on the pseudogene side. To avoid misclassification, we eliminate these location clusters from our dataset (669 out of 22123 location clusters with splicing gaps).

For all the remaining location clusters, we can unambiguously recognize which of the two axes holds a candidate exon (we call that side b) or a candidate pseudoexon (side s). The segments associated to the splicing gaps (which have projections only on the b side) denote our putative introns.

2.5. Trimming

Once we have identified the two sides (gene and pseudogene) of the location cluster, we can perform a further refinement of our candidate. Indeed, it often happens that the central alignment core, the signal of a retrotransposition event, is flanked by spurious alignments having no relation with the gene-pseudogene pair. We may eliminate them, imposing the constraint that the pseudoexons on the pseudogene side should be "close enough" to each other.

To implement this constraint, we evaluate the median μ of the gaps g_i between consecutive pseudoexons and the median s of their square variance defined as

$$s = \text{median}_i\left\{(\mu - g_i)^2\right\}$$

We then recursively remove alignments at the extremes of location clusters if the gap they open on the pseudogene side is larger than $\mu + 2\sqrt{s}$.

2.6. Analysis of the repeat content of candidate introns

A possible source of misclassification in our analysis is the presence in a duplicated genomic region of one or more transposons inserted after the duplication. These inserted sequences could be erroneously interpreted as spliced introns by the pipeline described above, thus leading to a wrong classification of the location cluster.

To avoid this problem, we look at the transposon content of all the candidate introns and discard those sequence composed for more than 90% by transposons. We then discard all the location clusters with less than two surviving introns.

Out of the initial 1588810 location clusters, only 2288 survived all the steps of the above pipeline; they represent our predictions.

2.7. Retrieval of external datasets

We obtained the lists of previously annotated genes from ENSEMBL (ENS, 2007) release 40 (August 2006), VEGA (Ashurst et al., 2005) release 40 (August 2006), and UCSC releases hg18 and mm8 (downloaded in September 2006). We obtained the lists of VEGA PPGs filtering the full VEGA gene dataset for the biotype "processed pseudogene." We also downloaded the full pseudogene set provided as the pseudogene.org pipeline output (Karro et al., 2007) in October 2007, and we later extracted all the processed pseudogenes linked to a valid ENSEMBL gene ID.

2.8. Identification of pseudogene families

A relevant number of location clusters overlap with some other location clusters. This happens in two cases: either when a single gene produced many pseudogenes, or when a single processed pseudogene shares high sequence similarity with more than one gene belonging to the same family. In the first case, we can define pseudogene families and associate them with a single original gene; in this way, we classify 2288 total pseudogenes in 987 families (see Supplementary File 7 in online Supplementary Material at www.liebertonline.com). In the second case, we report all the putative genes associated with the pseudogene and do not perform any further analysis. One or more of the candidate genes associated with a single PPG could be unprocessed pseudogenes; in principle, one could distinguish them from the gene which originated the retrotransposition event looking in detail at the alignments. Suppose that a single gene gives rise to both a processed and an unprocessed pseudogene: if the pseudogenes are free from selective pressure and therefore mutate randomly, the mutation events are independent and one could expect to find a better sequence homology between the PPG and the gene than between the PPG and the unprocessed pseudogene.

2.9. Experimental validation of the new candidate genes

The amplification primers were designed on two consecutive exons on the gene side of our predictions. To ensure their specificity, all the sequences differed from the corresponding pseudogene sequences at least on their 3′ end nucleotide. The sequences of primers are reported in Supplementary File 1 (see online Supplementary Material at www.liebertonline.com). Human testis cDNA was commercially obtained (Clontech). The amplification was performed in 50μl of 1×Go Taq Flexi Buffer (Promega), containing 0.2μM of each primer, 0.2mM of each dNTP, 1.5mM $MgCl_2$, 1.25μ GoTaq DNA Polymerase (Promega), 10% DMSO, and 5ng human testis cDNA (Clontech). Samples were amplified by 25 cycles of 95°C 1 min, 50°C 1 min, 72°C 1 min, followed by a final extension step of 72°C for 5 min. B-Actin primers were used as positive control.

3. RESULTS

3.1. Construction of the pseudogene database

We based our work on the idea that a gene-PPG pair can be recognized, in a set of pairwise paralogous alignments, as a cluster of HSPs that are nearby on one side of the alignment (the exons of the retrotransposed gene), but are separated by unaligned sequences on the other side (the introns of the retrotransposed gene; Fig. 1).

As mentioned above, the major interest of this strategy with respect to the existing ones (Torrents et al., 2003; Suyama et al., 2006; Pavlicek et al., 2006; Yao et al., 2006; Shemesh et al., 2006; Karro et al., 2007; Ortutay and Vihinen, 2008; Khelifi et al., 2005; Ohshima et al., 2003) is that it does not rely on known protein sequences, thus allowing us to identify previously unknown genes. The main problem of our approach is to discriminate the events due to retrotransposition from those caused by other causes and especially from sequence duplications followed by insertions. Therefore, we devised a pipeline capable of identifying the gaps most likely due to splicing events and to predict the structure of the original gene on the basis of this information. Moreover, since a single gene can give rise to many processed pseudogenes, we included a step to recognize these cases and to associate them to the unique original gene. To reduce the false positive rate and to increase the proportion of complete predictions, we require at least three splicing gaps to accept a candidate. It is well known that PPGs are synthetized from the 3′ end of the original mRNA and that very often the resulting cDNA is truncated before it reaches the 5′ end. However, it was recently observed that the length distribution of retrotransposed LINEs has a bimodal behavior, with a portion of complete LINEs 20 times larger than expected on the basis of a simple random truncation model (Pavlícek et al., 2002). By requiring at least three introns, we expected to reduce the risk of an incomplete annotation of the original gene. Using this parameter, we found 2288 gene-PPG pairs in the human genome, corresponding to 987 parent genes. Out of these, 965 genes had at least one exon annotated in ENSEMBL (both coding [948] and noncoding [17] genes are considered), and 943 had at least one overlapping UCSC known gene or RefSeq; in seven cases, we found neither. Among the sequences overlapping with ENSEMBL genes, there are only nine pairs (A, B) in which the predicted genes A and B are both associated to the same ENSEMBL gene; we did not observe any triplets with this behavior.

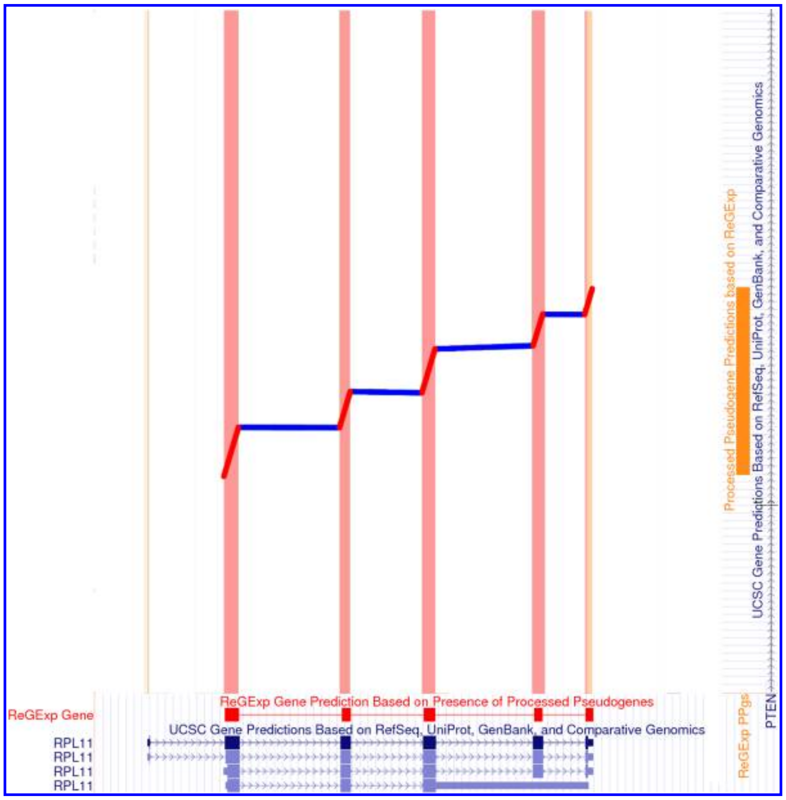

FIG. 1. Graphical representation of one entry of our dataset, corresponding to a gene-PPG pair in the human genome. The graph is similar to a dot-plot. On the horizontal axis, we put the region where we identified the gene and its annotation from the UCSC genome browser; on the vertical axis, the region corresponding to the pseudogene. Each alignment between the two regions is represented as a red segment in the central panel, while blue segments are the splicing signatures recovered by our pipeline. Finally, the background is colored in vertical stripes mirroring the exons.

Interestingly, virtually all the PPGs originated from parent genes annotated in ENSEMBL as coding genes retained a portion of their original coding sequence (Fig. 2). More precisely, nearly 30% have a complete coding sequence, and about 7% retained also the 5′ UTR end. To directly address the effect of the three-introns threshold, we performed the same analysis on a less stringent version of our database, in which we required only two splicing gaps to give a prediction. As expected, this database contains a strongly increased number (1694) of parent genes, but a large fraction of the new predictions does not overlap with known coding sequences (Fig. 2). Altogether, these observations confirm that, with the three-introns threshold, we may obtain more reliable results.

3.2. Validation of the REGEXP pipeline

To obtain an independent validation of our approach, we compared our results with the pseudogene.org dataset (Karro et al., 2007) and with processed pseudogenes reported in the human section of the Vertebrate Genome Annotation (VEGA) database, a central repository of high-quality, manually curated annotations (Ashurst et al., 2005). In both databases, the starting point of the annotation pipeline is the list of known proteins, which allows them to keep the threshold on the alignment score lower than in our case.

Globally, more than 81% of our entries were confirmed by at least one of these two databases (Fig. 3A). Comparison of our 2288 candidates with the 7816 processed pseudogenes reported in the pseudogene.org dataset revealed an overlap of 1640 PPGs, which corresponds to around 72% of our database (Table 1).

In the case of VEGA, the global intersection is not very informative, because when we perfomed the analysis the annotation of pseudogenes had been completed only for the chromosomes 1, 9, 10, 13, 20, 21, and X (Havana-Helpdesk, 2007). However, when limiting the comparison only to those seven chromosomes, the number of predicted pseudogenes supported by VEGA is about 95%; we also observed that our predictions overlap with VEGA remarkably better than with pseudogene.org (Fig. 3B).

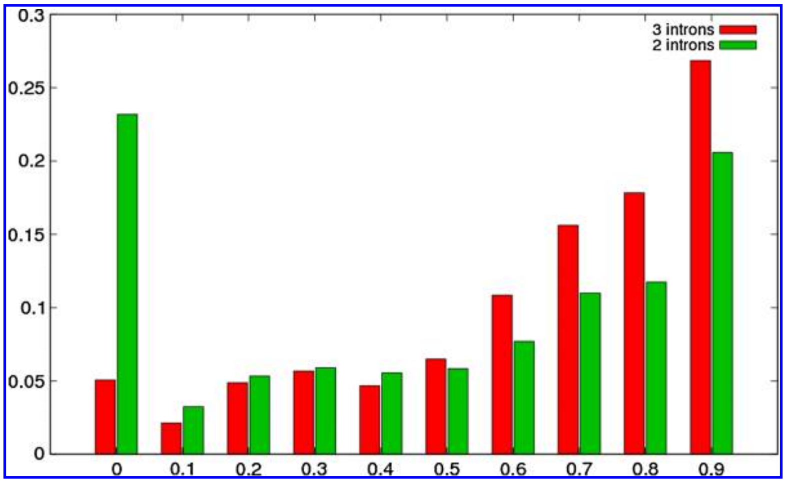

FIG. 2. Estimate of the fraction of truncated predictions in our database. On the horizontal axis, we report the fraction of gene coding regions (according to the ENSEMBL annotation) covered by our gene predictions; on the vertical axis, the percentage of gene predictions. Green bars refer to predictions with at least two introns, red ones to predictions with at least three introns.

To address whether relaxing the very stringent three-introns threshold could improve the sensitivity of the method without compromising its specificity, we made the same comparison using the version of our database requiring only two splicing gaps. Interestingly, we found 1288 candidates PPGs supported by overlapping entries in the VEGA dataset, 2593 in pseudogene.org, and 916 in both (Table 1). However, at the same time, the overall overlap with VEGA for the annotated chromosomes was much worse than with the previous database (Fig. 3B,C). Due to this reduced specificity, we reported the list of these retrotransposed genes in Supplementary File 2, (see online Supplementary Material at www.liebertonline.com) but we did not use them in the following steps of our analysis.

Altogether, these results indicate that, although our method is not very sensitive, it is remarkably specific and therefore could be used to reliably predict pseudogenes in non-annotated genomes.

3.3. Functional characterization of the parent genes

It is interesting to look at the functional characterization of the parent genes that originated the PPGs contained in our database. To address this point, we studied the GO annotation of the genes in our database and weighted them with the size of the corresponding pseudogene families. We found that most of the entries in our database correspond to protein coding genes (Table 2). Moreover, we observed a clear overrepresentation of sequences derived from ribosomal protein genes and, apparently, no other particular bias in the GO annotations (see Supplementary File 4 in online Supplementary Material at www.liebertonline.com) in good agreement with what already observed by Yao et al. (2006). As already noted in Yao et al. (2006), this strong preference denotes a special affinity of the retrotransposition machinery for genes involved in the ribosome complex.

Table 1. Comparison with VEGA and Pseudogene.org Datasets

	3 introns		*2 introns*	
	all chrs	*vega chrs*	*all chrs*	*vega chrs*
regexp	2288	686	4254	1329
vega	4012	3195	4012	3195
pseudogene.org	7816	2362	7816	2362
regexp and vega	850	653	1288	998
regexp and pseudogene.org	1640	484	2593	792
vega and pseudogene.org	2275	1749	2275	1749
regexp and vega and pseudogene.org	630	476	916	694

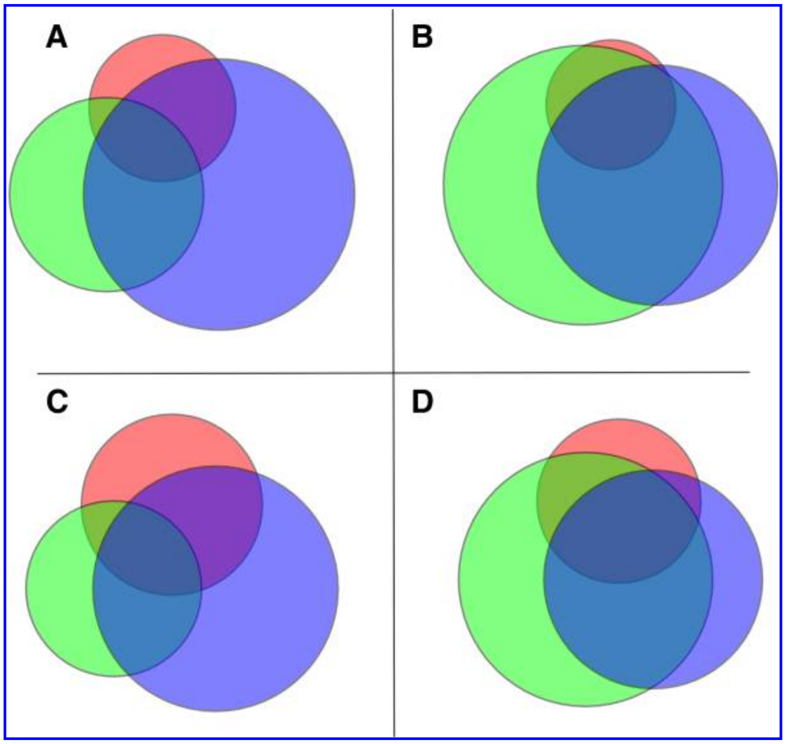

FIG. 3. Venn diagrams showing the intersections among our dataset (red), ENSEMBL VEGA (green), and pseudogene.org pipeline dataset (blue). Regexp predictions with at least three introns on all chromosomes (**A**) and only on those chromosomes completely annotated by VEGA (**B**). Regexp predictions with at least two introns on all chromosomes (**C**) and only on those chromosomes completely annotated by VEGA (**D**). For the associated numerical values, see Table 1.

3.4. Identification of new putative alternative splicing events

In several cases, we find instances of previously unknown, alternatively spliced transcripts. However, due to the peculiar features of our pipeline, out of the splicing variants that we observe, we can only be confident of those associated to additional exons. In particular, among the 965 transcripts that we could associate to ENSEMBL entries, we find 57 instances of additional exons.

It is interesting to compare this result with the analogous one reported in Shemesh et al. (2006), in which a similar analysis was perfomed starting from the entries of the pseudogene.org database. In Shemesh et al. (2006), the authors found 30 cases of additional exons, out of which 22 can be associated to transcripts contained in the ENSEMBL database. Interestingly, only three of these 22 alternative exons were in common with our predictions. Among the 19 alternative transcripts found by Shemesh et al. (2006) but missed by our algorithm, 13 were not present in our database from the very beginning, while the remaining six were discarded in the pipeline due to our filters on the transposon content, and the minimum and maximum length of the putative introns. This comparison may give an idea of the number of false negatives that we have, due to the very stringent constraints that we imposed in our analysis. On the other hand, it is

Table 2. Summary of Results of Our Analysis

	Human	*Mouse*
Total number of genes	987	709
Supported by UCSC known genes	928	649
Supported by RefSeq	922	655
Supported by ENSEMBL or VEGA genes	965	668
Supported by ENSEMBL or VEGA coding genes	948	661
New predictions	7	29

interesting to note that 54 out of 57 instances of additional exons that we find were missed in Shemesh et al. (2006). In 34 cases, this was due to the fact that the genes were absent in the pseudogene.org database from the very beginning, while the remaining 20 cases were effectively missed by the pipeline of Shemesh et al. (2006). Finally, out of the additional exons that we found, about 50% were supported by EST tracks, while the others are completely new. These results further illustrate that our approach is able to efficiently complement the existing genome annotation pipelines.

3.5. Identification and validation of putative new genes

One of the most interesting aspects of our method is that, in principle, it should be able to reveal the existence of functional genes independently from homology with previously identified cDNAs, even when they correspond to completely species-specific sequences. Among our predictions, 22 human sequences (2%) did not correspond to known genes in the ENSEMBL database. Nevertheless, for most of them, we could find EST tracks covering the majority of the predicted exons. However, seven putative genes identified by our method did not correspond to available ESTs and were not predicted by other gene-finding programs. In order to test these predictions, we performed a direct experimental validation. We reasoned that, if these sequences were produced by functional genes that are still active in the modern genomes, the corresponding mRNAs should be expressed at least in the germ line. To obtain direct support for this hypothesis, we designed specific PCR primers matching the nucleotides of two different exons of the gene side sequence and used them to perform reverse transcription–polymerase chain reaction (RT-PCR) on human testis cDNA. Remarkably, in four cases, we recovered amplification products of the expected molecular weight (Fig. 4), which were further confirmed by direct sequencing. Interestingly, all the predictions that were not confirmed by RT-PCR corresponded to putative genes located within introns of annotated genes (Table 3).

Blastx analysis of the validated candidates revealed that most of them show significant homology with other protein-coding genes. In particular, candidate 836759 may encode a protein very similar to a portion of the membrane protein LRRC37B (ENSG00000185158), a member of the "SLIT-like" family of genes. Moreover, candidates 1043493 and 338893 may encode for proteins similar to the putative proteins encoded by LOC255649 (which display strong similarity to rodents' Oocyte secreted protein 1) and LOC686205, respectively. The only new validated candidate gene that did not show any significant homology was 931840. Interestingly, the mRNA sequence that we reconstructed for this gene does not contain a significant open reading frame, thus suggesting that it may correspond to a new human-specific non-coding gene.

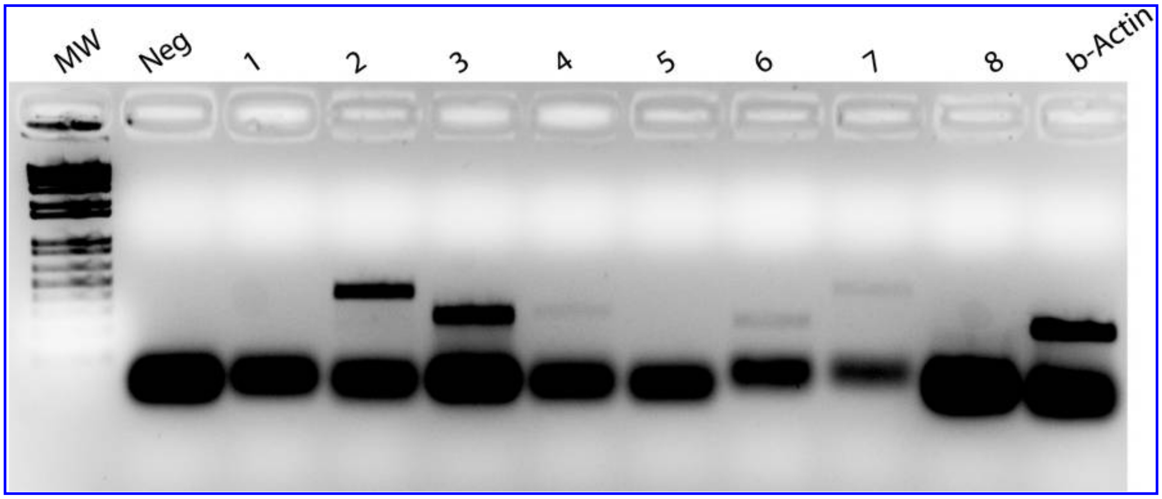

FIG. 4. Experimental validation of the predictions corresponding to novel genes. Reverse transcription–polymerase chain reaction (RT-PCR) amplification of human testis cDNA, with primers specific for predictions 316640 (1), 338893 (2), 836759 (3), 931840 (4), 961127 (5), 1043493 (6), 128365 (7), and 268231 (8). The major band in lanes 2, 3, 4, 6, and 7 shows an amplification product corresponding to the expected molecular weight. Candidate 338893, considered as a new prediction in the first release of the program, has shown a positive matching with EST sequences in the present version. We included it in the experiment as a further positive control. b-Actin, beta-actin positive control; Neg, no cDNA negative control amplified with primers of candidate 338893; MW, molecular weight standard.

Table 3. Summary of the Putative New Genes that We Tested in Our RT-PCR Experiment

ID	*PCR*	Δ*bp*	*Description*
836759	Positive	50k	New prediction, similar to SLIT-like and LR37B-HUMAN
1043493	Positive	8k	New prediction
268231	Negative	Intron	Genome duplication (plus transposon insertion)
316640	Negative	Intron	Most probably a set of 3’UTR exons
931840	Positive	5k	New prediction
961127	Negative	Intron	Genome duplication (plus transposon insertion)
338893	Positive	12k	New prediction

In the first column, we report the identifier in the REGEXP database; in the second column, the result of the PCR test; in the third column, the distance Δ*bp* with respect to the nearest annotated gene or the tag “intron” if the putative new gene is located inside the intron of an already known transcript. In the last column, we list some non-trivial features of the putative new genes.

4. CONCLUSION

The complete annotation of genomes requires not only the identification of genes, but also of pseudogenes. Although pseudogenes are commonly referred to as nonfunctional copies of working genes, it has been recently recognized that they may play important functions in gene regulation and, in particular, in fine-tuning the expression of the genes from which they are derived. A particularly interesting class of pseudogenes is given by processed pseudogenes (PPGs), copies of cellular RNAs typically containing poly(A) and lacking introns, which were reverse-transcribed and inserted into the genome. All the methods so far developed to identify PPGs are based on the use of known mRNAs and protein sequences as input data for suitable alignment programs. Therefore, they could be expected to have a lower sensitivity on genomes lacking extensive transcriptome annotation or on PPGs derived from non-canonical genes. In this article, we have proposed REGEXP, a new approach for the identification of gene-PPG pairs based only on the analysis of the genomic sequence, which does not require *a priori* knowledge of the transcriptome. We showed that specificity of REGEXP is comparable to that reached in the VEGA annotation database by manual curation of PPGs. Finally, using REGEXP, we were able to identify and to experimentally validate a few previously unknown genes, even in the highly annotated human genome. Therefore, we conclude that REGEXP could represent a significant addition to the current genome annotation pipelines.

ACKNOWLEDGMENTS

We warmly thank Davide Corà, Elena Grassi, Luigi Grassi, Loredana Martignetti, Chiara Peyron, Paolo Provero, and Angela Re for useful suggestions and discussions. This work was partially supported by the Fund for Investments of Basic Research (FIRB) from the Italian Ministry of the University and Scientific Research (no. RBNE03B8KK-006).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Ashurst, J.L., Chen, C.K., Gilbert, J.G.R., et al. 2005. The vertebrate genome annotation (VEGA) database. *Nucleic Acids Res.* 33, 459–465.

Ejima, Y., and Yang, L. 2003. Trans-mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. *Human Mol. Genet.* 12, 1321–1328.

ENS. 2007. Ensembl.

Esnault, C., Maestre, J., and Heidmann, T. 2000. Human line retrotransposons generate processed pseudogenes. *Nat. Genet.* 24, 363–367.

Harrison, P.M., Echols, N., and Gerstein, M.B. 2001. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res.* 29, 818–830.

Harrison, P.M., Milburn, D., Zhang, Z., et al. 2003. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res.* 31, 1033–1037.

Havana-Helpdesk. 2007. Private communication.

Hillier, L., Miller, W., Birney, E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695–716.

Hirotsune, S., Yoshida, N., Chen, A., et al. 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 423, 91–96.

Jurka, J., Smith, T.F., and Labuda, D. 1988. Small cytoplasmic ro RNA pseudogene and an alu repeat in the human alpha-1 globin gene. *Nucleic Acids Res.* 16, 766.

Karro, J.E., Yan, Y., Zheng, D., et al. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.* 35, D55–D60.

Khelifi, A., Adel, K., Duret, L., et al. 2005. Hoppsigen: a database of human and mouse processed pseudogenes. *Nucleic Acids Res.* 33, D59–D66.

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.

Ohshima, K., Hattori, M., Yada, T., et al. 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.

Ortutay, C., and Vihinen, M. 2008. Pseudogenequest—service for identification of different pseudogene types in the human genome. *BMC Bioinform.* 9, 299.

Pavlicek, A., Gentles, A., Paces, J., et al. 2006. Retroposition of processed pseudogenes: the impact of RNA stability and translational control. *Trends Genet.* 22, 69–73.

Pavlícek, A., Paces, J., Zíka, R., et al. 2002. Length distribution of long interspersed nucleotide elements (lines) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. *Gene* 300, 189–194.

Sakai, H., Koyanagi, K., Imanishi, T., et al. 2007. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. *Gene* 389, 196–203.

Shemesh, R., Novik, A., Edelheit, S., et al. 2006. Genomic fossils as a snapshot of the human transcriptome. *Proc. Natl. Acad. Sci. USA*, 103, 1364–1369.

Suyama, M., Harrington, E., Bork, P., et al. 2006. Identification and analysis of genes and pseudogenes within duplicated regions in the human and mouse genomes. *PLoS Comput. Biol.* 2, e76.

Tam, O., Aravin, A., Stein, P., et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453, 534–538.

Torrents, D., Suyama, M., Zdobnov, E., et al. 2003. A genome-wide survey of human pseudogenes. *Genome Res.* 13, 2559–2567.

Watanabe, T., Totoki, Y., Toyoda, A., et al. 2008. Endogenous sirnas from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453, 539–543.

Weil, D., Power, M.A., Webb, G.C., et al. 1997. Antisense transcription of a murine fgfr-3 psuedogene during fetal developement. *Gene* 187, 115–122.

Yao, A., Charlab, R., and Li, P. 2006. Systematic identification of pseudogenes through whole genome expression evidence profiling. *Nucleic Acids Res.* 34, 4477.

Zhang, Z., Schwartz, S., Wagner, L., et al. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 7, 203–214.

Zheng, D., Frankish, A., Baertsch, R., et al. 2007. Pseudogenes in the encode regions: consensus annotation, analysis of transcription, and evolution. *Genome Res.* 17, 839–851.

Zhou, B.S., Beidler, D.R., and Cheng, Y.C. 1992. Identification of antisense RNA transcripts from a human DNA topoisomerase in pseudogene. *Cancer Res.* 52, 4280–4285.

Address correspondence to:
Dr. Michele Caselle
Theoretical Physics Department
Universit di Torino
INFN
Via Pietro Giuria 1
Torino, Italy

E-mail: caselle@to.infn.it