Pseudogenes in the ENCODE Regions: Consensus Annotation, Analysis of Transcription and Evolution

Deyou Zheng [1, *], Adam Frankish [2], Robert Baertsch [3], Philipp Kapranov [4], Alexandre Reymond [5, 6], Siew Woh Choo [7], Yontao Lu [3], France Denoeud [8], Stylianos E Antonarakis [6], Michael Snyder [9], Yijun Ruan[7], Chia-Lin Wei [7], Thomas R. Gingeras [4], Roderic Guigo [8, 10], Jennifer Harrow [2], and Mark B. Gerstein [1, 11, 12, *]

Running title: pseudogenes in the ENCODE regions

Addresses:
[1] Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA.
[2] Wellcome Trust Sanger Institute; Wellcome Trust Genome Campus; Hinxton, Cambridgeshire, CB10 1HH, UK.
[3] Department of Biomolecular Engineering; University of California, Santa Cruz; 1156 High Street, Santa Cruz, CA 95064, USA.
[4] Affymetrix, Inc.; Santa Clara, CA, 92024, USA.
[5] Center for Integrative Genomics; University of Lausanne; Genopode building; 1015 Lausanne, Switzerland.
[6] Department of Genetic Medicine and Development; University of Geneva Medical School; 1 rue Michel-Servet, 1211 Geneva, Switzerland.
[7] Genome Institute of Singapore; 60 Biopolis Street; Singapore 138672, Singapore.
[8] Grup de Recerca en Informática Biomèdica; Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra. Passeig Marítim de la Barceloneta, 37-49, 08003, Barcelona, Catalonia, Spain.
[9] Molecular, Cellular & Developmental Biology Department; Yale University; New Haven, CT 06520, USA.
[10] Center for Genomic Regulation; Passeig Marítim de la Barceloneta, 37-49, 08003, Barcelona, Catalonia, Spain.
[11] Department of Computer Science, Yale University, 51 Prospect Street, New Haven, CT 06520, USA.
[12] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA.

[*]Corresponding author:
Mark Gerstein
Tel: (203) 432 6105; Fax: (360) 838 7861; Email: Mark.Gerstein@yale.edu

Or Deyou Zheng, deyou.zheng@yale.edu

ABSTRACT

Arising from either retrotransposition or genomic duplication of functional genes, pseudogenes are "genomic fossils" valuable for exploring the dynamics and evolution of genes and genomes. Pseudogene identification is an important problem in computational genomics, and is also critical for obtaining an accurate picture of a genome's structure and function. However, no consensus computational scheme for defining and detecting pseudogenes has been developed thus far. As part of the ENCyclopedia Of DNA Elements (ENCODE) project, we have compared several distinct pseudogene annotation strategies and found that different approaches and parameters often resulted in rather distinct sets of pseudogenes. We subsequently developed a consensus approach for annotating pseudogenes (derived from protein coding genes) in the ENCODE regions, resulting in 201 pseudogenes, two-thirds of which originated from retrotransposition. A survey of orthologs for these pseudogenes in 28 vertebrate genomes showed that a significant fraction (~80%) of the processed pseudogenes is primate specific sequences, highlighting the increasing retrotransposition activity in primates. Analysis of sequence conservation and variation also demonstrated that most pseudogenes evolve neutrally, and processed pseudogenes appear to have lost their coding potential immediately or soon after their emergence. In order to explore the functional implication of pseudogene prevalence, we have extensively examined the transcriptional activity of the ENCODE pseudogenes. We performed systematic series of pseudogene-specific RACE analyses. These, together with complementary evidence derived from tiling microarrays and high throughput sequencing, demonstrated that at least a fifth of the 201 pseudogenes are transcribed in one or more cell lines or tissues.

Key Words: pseudogene, ENCODE, transcription, conservation, evolution

INTRODUCTION

The goal of the ENCyclopedia Of DNA Elements (ENCODE) project is to produce a comprehensive catalog of structural and functional components encoded in the human genome (Consortium 2004). In its pilot phase, ~30 megabases (Mb) (~1%) of the human genome were chosen as representative targets. Most of the functional components (e.g., genes and regulatory elements) are essentially determined by high-throughput experimental technologies with the assistance of computational analyses (Consortium 2004); however, one component whose identification depends almost exclusively on computational analysis is pseudogenes.

Pseudogenes are usually defined as defunct copies of genes that have lost their potential as DNA templates for functional products (Balakirev and Ayala 2003; Harrison et al. 2002; Mighell et al. 2000; Vanin 1985; Zhang and Gerstein 2004; Zhang et al. 2003; Zheng et al. 2005). As only pseudogenes derived from protein coding genes will be characterized here, the term pseudogene in this study applies to genomic sequences that cannot encode a functional protein product. Pseudogenes are often separated into two classes: processed pseudogenes, which have been retrotransposed back into a genome via an RNA intermediate, and non-processed pseudogenes, which are genomic remains of duplicated genes or residues of dead genes. These two classes of pseudogenes exhibit very distinct features: processed pseudogenes lack introns, possess relics of a poly(A) tail, and are often flanked by target-site duplications (Balakirev and Ayala 2003; Brosius 1991; Jurka 1997; Long et al. 2003; Mighell et al. 2000; Schmitz et al. 2004). It has to be mentioned that retrotransposition sometimes generates new genes that are often called retroposed genes (or processed genes) (Brosius 1991; Long et al. 2003).

The common assumption is that pseudogenes are non-functional and thus evolve neutrally. As such, they are frequently considered as "genomic fossils" and often used for calibrating parameters of various models in molecular evolution, such as estimates of neutral mutation rates (Bustamante et al. 2002; Gojobori et al. 1982; Gu and Li 1995; Li et al. 1981; Li et al. 1984; Ota and Nei 1995; Zhang and Gerstein 2003). However, a few pseudogenes have been indicated to have potential biological roles (Balakirev and Ayala 2003; Korneev et al. 1999; Mighell et al. 2000; Ota and Nei 1995). Whether these are anecdotal cases or pseudogenes do play cellular roles is still a matter of debate at this point, simply because not enough studies have been conducted with pseudogenes as the primary subjects. To be clear, in this study the non-functionality of a pseudogene is strictly interpreted as a sequence's lacking protein coding potential, regardless of whether it can produce a (functional or non-functional) RNA transcript.

The prevalence of pseudogenes in mammalian genomes (Balakirev and Ayala 2003; Mighell et al. 2000; Zhang et al. 2003) has been problematic for gene annotation (van Baren and Brent 2006) and can introduce artifacts to molecular experiments targeted at functional genes (Hurteau and Spivack 2002; Kenmochi et al. 1998; Ruud et al. 1999; Smith et al. 2001). The correct identification of pseudogenes, therefore, is critical for obtaining a comprehensive and accurate catalog of structural and functional elements of the human genome. Several computational algorithms have been described previously

for annotating human pseudogenes (Bischof et al. 2006; Coin and Durbin 2004; Harrison et al. 2002; Khelifi et al. 2005; Ohshima et al. 2003; Torrents et al. 2003; van Baren and Brent 2006; Zhang et al. 2006; Zhang et al. 2003). Although these methods often present similar estimates for the number of pseudogenes in the human genome, they can produce rather distinct pseudogene sets (Khelifi et al. 2005; Zhang and Gerstein 2004; Zheng et al. 2005). In order to obtain an accurate list of pseudogenes in the ENCODE regions, we have compared several methods and subsequently developed a uniform computational framework for annotating pseudogenes, which can be applied to the human and other mammalian genomes. Furthermore, the final list of pseudogenes is good benchmark data for developing and improving methods of pseudogene annotation.

To characterize the ENCODE pseudogenes in detail, we have subsequently synthesized data from a number of the ENCODE research groups (ref to the main ENCODE paper). We examined the transcriptional activity of pseudogenes using data from the ENCODE genes and transcripts group, and the transcription regulation group. In addition, rapid amplification of cDNA ends (RACE) analyses coupled with tiling microarrays (Kapranov et al. 2005) were carried out in this study with pseudogene loci as the specific targets. These empirical transcriptional data from multiple techniques together revealed a complex picture of pseudogene transcription: low in abundance and specific in tissues or cells.

Using data generated by the ENCODE multi-species sequence analysis group and variation group, we have begun to explore several fundamental concepts concerning the evolution and preservation of pseudogenes. Specifically, with orthologous genomic sequences from 28 mammalian or vertebrate species, we have characterized in detail the sequence decay and preservation of pseudogenes, in comparison to both their surrounding genomic materials and protein coding genes.

RESULTS

Strategies and Results of Pseudogene Annotation in the ENCODE Regions

As a sub-group within the ENCODE project, our first goal is to obtain an accurate list of pseudogenes in order to facilitate the creation of a comprehensive catalog of structural and functional elements in the ENCODE regions (Consortium 2004). This is realized in a consortium fashion and executed in two stages. We began with an examination of five methods for detecting pseudogenes. These methods, which have been developed independently, are:
1) *GIS-PET method,* from the Genome Institute of Singapore;
2) *HAVANA method* of manual pseudogene annotation, by the Human And Vertebrate Analysis aNd Annotation team (HAVANA) at the Wellcome Trust Sanger Institute as part of the GENCODE collaboration (Harrow et al. 2006);
3) *PseudoPipe* (Zhang et al. 2006; Zheng and Gerstein 2006), from the pseudogene research group at Yale University;
4) *pseudoFinder*, from the University of California Santa Cruz (UCSC); and
5) *retroFinder*, also from UCSC but focused specifically on processed pseudogenes.

Details of these computational methods are described in Methods and Supplement. In summary, all five methods detected pseudogenes by their sequence similarity to at least one entry in a collection of query sequences representing known human genes (referred to as the parent genes). The major differences are in (i) queries (either proteins or nucleotides) used to search for pseudogenes and (ii) strategies (including parameters) used to assess a sequence's coding potential and to distinguish pseudogenes into types of processed or non-processed.

The initial sets of pseudogenes annotated by individual methods for the ENCODE regions contained similar numbers (except GIS) of pseudogenes: 56 by GIS, 165 by HAVANA, 167 by PseudoPipe, 172 by pseudoFinder, and 163 by retroFinder; but, the annotated pseudogene sets were rather different. A simple union of these five sets yielded 252 non-overlapping pseudogenes, of which only 45 (17.9%) were identified by all methods while 69 (27.3%) were method specific (Figure 1). Setting aside the GIS data (see supplement for reasons), we found that 87 (34.5%) pseudogenes were agreed upon by the remaining four methods. Furthermore, pair-wise comparisons showed that the overlaps between two lists ranged from 62.2% to 80%, with the two protein-based methods exhibiting the best agreement: 132 of 165 (80%) HAVANA pseudogenes were also discovered by the Yale method.

The results above show clearly that none of the individual methods initially applied to the study provided a completely authoritative description of the pseudogenes in the ENCODE regions. After careful comparisons and investigations, it was determined that the most critical factor contributing to the discrepancies among the pseudogene sets was the nature of the queries (i.e., the parent genes/proteins used for detecting pseudogenes) rather than uncertainty of pseudogene assignment. In most cases, when a pseudogene was missed by one or more methods, careful manual inspection identified the same problem: the parent gene or the coding sequence (CDS) assigned to it was dubious or simply not used by other approaches (see supplement).

In order to minimize such consequences, as the second stage of our annotation we have developed a consensus procedure that involves intense manual curation to obtain an accurate and reliable list of pseudogenes. Such a procedure also provides a uniform definition and computational scheme for consolidating lists of pseudogenes from different sources. Our current approach is based on known proteins in the UniProt database (Bairoch et al. 2005), i.e., we only considered pseudogenes with support from reliable parent protein coding sequences. Classification of processed and non-processed pseudogenes was based on retention of parent gene structure, evidence of a retrotransposition, and preservation of flanking genomic sequence. By this procedure and starting from the 252 non-redundant pseudogenes annotated in the first stage, we identified a consensus set of 201 pseudogenes, 77 of which were non-processed and 124 processed. This pseudogene annotation is available at http://www.pseudogene.org/ENCODE/ and http://genome.ucsc.edu/ENCODE/. (Under the UCSC browser, a special track named "ENCODE Pseudogene Predictions" was built to present both our final consensus annotation and the initial annotations from the individual methods). It is important to point out that each of the five methods except for

GIS-PET contributed new pseudogenes to the final consensus set. All subsequent analyses described below were done on these 201 consensus pseudogenes.

Characterization of the ENCODE Pseudogenes

The genomic distribution of pseudogenes is similar overall to that of functional coding genes: i.e., gene-rich ENCODE regions usually have more pseudogenes than gene-poor regions (Figure 2). In addition, different gene families seem to have contributed very different numbers of pseudogenes. The two dominant families were ribosomal protein genes and olfactory receptor (OR) genes, which accounted for 37 (18.5%, all processed) and 29 (14.5%, all non-processed) of the 201 pseudogenes, respectively. Additionally, ~10% of the pseudogenes were from genes involved in immune response. Contributions from other gene families were relatively small (<5 pseudogenes per family). Notably, the overrepresentation of OR pseudogenes simply reflects the inclusion of a single region (ENm009) in the ENCODE pilot project that contains a large cluster of coding OR genes and OR pseudogenes (Glusman et al. 2001) and does not, therefore, represent the statistics for the entire human genome.

Most pseudogenes are decayed gene copies and have accumulated nonsense or frameshift mutations that would usually disrupt an open reading frame (ORF). The ENCODE processed and non-processed pseudogenes share mean sequence identities of 67.6% (± 14%) and 61.8% (± 18%) with their parent proteins in alignment coverage of 82.4% (± 26%) and 69.4% (± 33%), respectively. In addition, 83.2% of processed and 79% of non-processed pseudogenes display disablements (defined as nonsense or frameshift mutations) in their putative ORFs, with average disablements of 6.2 per processed pseudogene and 2.4 per non-processed pseudogene. Overall, such disablements were located uniformly across the hypothetical coding regions of pseudogenes. The differences in sequence identity and disablements between processed and non-processed pseudogenes are significant ($p < 0.001$, Wilcoxon rank-sum test), appearing to suggest that the sequences giving rise to processed pseudogenes lose coding potential more quickly than those for non-processed pseudogenes. It needs to be pointed out that disablements can sometimes escape detection due to the limitation of available sequence alignment tools (Zheng and Gerstein 2006). Therefore, they should not be used as the exclusive criterion for distinguishing pseudogenes from genes.

Pseudogene Transcription

Using pre-existing data, several recent surveys have indicated that pseudogene transcription could contribute to the complexity of the human transcriptome (Harrison et al. 2005; Shemesh et al. 2006; Strichman-Almashanu et al. 2003; Vinckenbosch et al. 2006; Yano et al. 2004; Zheng et al. 2005). In order to obtain direct evidence of pseudogene transcription, we have systematically interrogated the transcription of 160 pseudogenes (49 non-processed and 111 processed) with locus specific RACE/microarray analysis (Kapranov et al. 2005) using poly A+ RNA from 12 tissues. In 51 cases (26 non-processed and 25 processed pseudogenes) we were able to design pseudogene locus-specific 5' RACE primers, which typically had five or more

mismatched base pairs when compared to the parent genes while matching the pseudogenes perfectly. For the remainder it was not possible to design such primers. To take this into account, a careful examination of the transcriptionally active regions (termed RACEfrags, Danoeud et al., companion paper) identified by hybridizing RACE products onto tiling microarrays was performed in the subsequent data analysis. Specifically, we assigned a RACEfrag to a pseudogene only if it was uniquely mapped to this pseudogene locus (see Methods). The resulting data supported transcription for 14 (eight processed and six non-processed) of the 160 pseudogenes loci, nine of which from RACE experiments where pseudogene specific primers were used. Interestingly, nine of these 14 pseudogenes were found to be transcriptionally active (and five exclusively) in testes. This unusual pseudogene expression in testes may have biological implication, and this observation is in accordance with previous reports (Kleene et al. 1998; Marques et al. 2005; Reymond et al. 2002) and especially a recent finding that transcription of human retrocopies mainly (and/or initially) occurs in testes (Vinckenbosch et al. 2006). The final number of 14 seems a conservative estimate since we decided not to assign a (ambiguous) RACEfrag to a pseudogene if it could be mapped to both the pseudogene and another locus.

In addition to this pseudogene targeted RACE analysis, we have also intersected our pseudogenes with various empirical transcription data obtained by the ENCODE genes and transcripts group (ref to ENCOE main paper), including transfrags, 5' specific Cap Analysis Gene Expression (CAGE) tags, and Paired-End 5' and 3' diTags (PET). These analyses suggested that a large number of pseudogenes were potentially transcribed (Table 1). A survey of known mRNA/ESTs in public databases also identified 21 transcribed ENCODE pseudogenes. Figure 3 shows one example of pseudogene transcription, and data for all our individual pseudogenes are available in the UCSC browser (which can be accessed through a table in supplement).

We believe that the data obtained by RACE experiments or by sequencing analyses (CAGE, PET, EST, and mRNA) provide unambiguous evidence for pseudogene transcription. Altogether, these data indicate that 38 (19% of 201, 20 non-processed and 18 processed) pseudogenes are the sources of novel RNA transcripts. This may well represent a low-bound estimate and does not include the ambiguous and possibly inconclusive cases supported only by transfrags. We should emphasize that most cases of pseudogene transcription were only detected in one or a few experiments (manifested by small overlaps between data from different evidence, Table 1) and thus the example in Figure 3 is not typical. This indicates that pseudogene transcription is quite tissue specific, as RACEfrags, CAGE, PET, and transfrags were obtained from different cell lines or tissues (see Materials and Methods). On the other hand, such a pattern of tissue (or cell line)-specific transcription was a common characteristic of novel non-coding transcripts (Cheng et al. 2005).

We have subsequently examined the ENCODE pseudogenes for potential cryptic promoters. A comparison with high quality regulatory elements discovered by integrative analyses of ~130 chromatin immunoprecipitation (ChIP)-chip experiments (ref to a companion paper by Trinklein et al.) showed that 19 pseudogenes (three non-

processed and 16 processed) likely contained transcriptional regulation sites in their "promoter" regions (-2 kb ~ +200 bp). Five of these were among the 38 pseudogenes exhibiting transcription evidence, but the association of regulatory elements with transcription was not statistically significant ($p = 0.58$, χ^2-test).

Pseudogene Preservation

Pseudogenes are usually considered the evolutionary end point of genomic material whose ultimate fate is to be removed from a genome. Nevertheless, millions of years of evolution has left the human genome with thousands of pseudogenes (Torrents et al. 2003; Zhang et al. 2003). Within the ENCODE project, the MSA group has identified and sequenced the orthologous regions of the individual ENCODE target regions in 20 to 28 vertebrate (mostly mammalian) species (see Methods for the list). Several algorithms such as TBA (Threaded Blockset Aligner) (Blanchette et al. 2004) have also been applied to construct multi-species sequence alignments across the entire ENCODE regions (ref to ENCODE main paper and MSA companion paper by Margulies et al.). With these data, it is possible to survey the preservation of sequences corresponding to the human pseudogenes in other species to get a glimpse of the evolutionary process leading to the human lineage.

For each of our 201 pseudogenes, the aligned block containing this pseudogene was extracted from the multi-species sequence alignments constructed by the MSA group, and this excerpt was defined as the orthologous region for this pseudogene. A sequence relative (i.e., ortholog) of a human pseudogene was considered to be present (i.e., "preserved") in a species if at least 50 nucleotides from that species were found in the aligned block. Data in Figure 4 shows that as a species' divergence from humans increases, fewer orthologs of (current) human non-processed pseudogenes are preserved, suggesting that the majority of duplication events giving rise to these genomic materials occurred a long time ago. This pattern slightly deviates from that of protein coding genes, as expected. However, the trend for processed pseudogenes is dramatically different; preservation decreases very sharply before reaching a near plateau (Figure 4). The turning point appears to be between the New World monkeys and strepsirrhines, about 40 to 63 million years ago (MYA) (Goodman 1999; Goodman et al. 1998) or later. There is no significant difference between the introns (i.e., *pseudointrons*) and exons (i.e., *pseudoexons*) of pseudogenes (see supplementary figS1). As the ortholog assignment for distantly related species can be tricky, we have used the MSA data from other alignment programs, MAVID (Bray and Pachter 2004) and MLAGAN (Brudno et al. 2003), and obtained similar results (shown in Figure 4 for processed pseudogenes only). These results demonstrate that most (~80%) human processed pseudogenes arise from sequences specific to the primate lineage and are in good agreement with previous data estimated with molecular clocks using pseudogenes and SINE (short interspersed elements) repeats (Ohshima et al. 2003).

The overall sequence decay rate of pseudogenes is very similar to that of neutrally evolving DNA. The nucleotide sequence identity between human pseudogenes and their orthologs indicates apparently that the majority of pseudogenes experience no

evolutionary constraints, as their sequence decay pattern is not much different from that derived from four-fold degenerative sites, at least within the lineage of mammals (Figure 5A). We subsequently analyzed these 201 pseudogenes and the corresponding MSA data using the program phastOdds (Siepel et al. 2005), which computes the log odds ratio of the probability that a sequence fragment fits a model of "constrained" versus "neutral" evolution. The result supports that the evolution of pseudogenes as a group is better described by the neutral model, but it suggests that a few pseudogenes (mostly non-processed ones) may have experienced evolutionary constraints in certain periods of their evolution (most likely as genes) (Figure 5B).

The evolutionary constraint of a genomic sequence can also be evaluated in the context of its local genomic environment. As known and shown in Figure 6, the nucleotide sequence identity in CDS of genes is significantly higher than their adjacent 5' and 3' genomic sequences (human-mouse, human-dog; such a pattern is not obvious when very closely related species like human-chimp are considered). Pseudogenes, however, do not display such a clear profile of sequence constraints. In fact, constraints on processed pseudogenes are not much different from those on their surrounding genomic sequences. The profile for non-processed pseudogenes is rather intricate. On one hand, the data from the human-mouse comparison indicate that some of these pseudogenes may have evolved with constraints (Figure 6). On the other hand, the data from human-chimp and human-macaque comparisons suggest that non-processed pseudogenes may have speeded up their evolution recently. This is probably due to increasing mutation rate that can be attributed to the higher GC-content (51.5%) in these non-processed pseudogenes versus their adjacent sequences (43.4%) and processed pseudogenes (46.1%), suggesting that such sequences only became pseudogenes recently and were genes for much of their histories. Notably, about one half of our non-processed pseudogenes were derived from olfactory receptor genes and genes involved in immune response, which have been suggested to be under positive selection (Consortium 2005; Gilad et al. 2005; Lander et al. 2001; Lindblad-Toh et al. 2005).

In summary, as a group of genomic components, pseudogenes appear to evolve neutrally with few candidates exhibiting evolutionary constraints as measured by cross-species sequence preservation and phastOdds ratios. The "constraints" could be either a direct result of functional constraints or simply a consequence of recent pseudogenization. It has to be pointed out that our results may be complicated by the challenge in identifying orthologous sequences in species very divergent from human (ref to MSA companion paper by Margulies et al. and another by King et al.) and thus reflect alignment artifacts. On the other hand, our conclusion is independently supported by analyses of SNP (single nucleotide polymorphism) density and non-synonymous versus synonymous substitution (Ka/Ks) ratios (Figure 7), which showed that SNP density and Ka/Ks ratios of pseudogenes were overall significantly higher than those of genes ($p<0.01$), but outliers nonetheless existed.

Pinpointing the Timeline of Pseudogenization

With the MSA data we have attempted to track the history of individual pseudogene sequence and discover when the sequence appeared and lost its protein coding ability (i.e., pseudogenize). In this analysis, the orthologous sequences of each ENCODE pseudogene were retrieved from MSA data and then compared to the pseudogene's parent protein sequence using the alignment programs GeneWise (Birney et al. 2004) or FASTA (Pearson et al. 1997). The resulting alignments were then examined for nonsense or frameshift mutations. These analyses showed that disablements of a human processed pseudogene were often observed in their orthologous sequences as well (see supplementary Table S1 and Figure 8), further supporting the hypothesis that these sequences were dead on arrival or became a pseudogene soon after emergence. However, the scenario for non-processed pseudogenes is more complicated. Even in species like chimp, baboon and macaque that are very close to human, the pseudogenization of orthologous sequences is not always consistent with what one might expect from phylogeny (Figure 8). For instance, a non-processed pseudogene (ID: AC087380.14) located in region ENm009 appears to have originated from duplication of a functional gene with an olfactomedin-like domain. A disruption in its ORF is observed in the orthologous sequences of human, baboon, macaque, and many other species but not chimp, marmoset, or galago. This suggests that pseudogenization is most likely a random process in which disablements accumulate gradually and randomly once evolutionary constraint on a sequence relaxes. As a result, for recently pseudogenized sequences, we see disablements occurring in various species randomly. It has to be emphasized again, however, that a precise interpretation of our data should account for the quality of sequencing for each species and the reliability of ortholog assignments, which can be problematic for species very distantly related to humans. Also, gene conversion would add further complication to the final species pattern of disablements.

DISCUSSION

Comparison of Different Pseudogene Annotation Methods

In this study five methods of pseudogene annotations were extensively examined and compared. All methods first defined a set of pseudogene candidates based on their sequence similarity to a parent gene or protein. Empirical evidence or heuristic algorithms were then used to distinguish pseudogenes from gene-like candidates that may have protein coding potential. We found that the quality of the datasets for annotated human genes (or their translated proteins) is the most critical factor leading to inconsistent (likely false) annotation of pseudogenes for two main reasons: firstly, it is vital to be able to distinguish a locus as being either coding or pseudogenic and secondly spurious translations have contributed a significant pollution effect to current protein databases (see supplement for further discussion). This clearly indicates that gene and pseudogene annotation are intertwined and dynamics processes that need to be improved coordinately. In addition, we found that processed pseudogenes are more easily identifiable than non-processed pseudogenes, as the former constituted a large part of the common pseudogenes identified by multiple methods.

Our final consensus approach is based on a collection of well annotated protein sequences. It provides a relatively straightforward way of defining pseudogene boundaries. Although this approach is presented here as a way to integrate pseudogene annotation from different sources, it is by no means restricted to such a usage. It can be easily modified for *de novo* pseudogene identification and therefore is applicable to the entire human genome and other mammalian genomes. The strategy can be largely implemented through computational programs, but we believe that much manual intervention is necessary for achieving a high-quality annotation, as manual curation allows very detailed investigation, bringing numerous sources of evidence external to the initial prediction to bear -- e.g., literature reports, mRNA, and examination of parent genes. Manual curation is highly specific (i.e., very few manually curated pseudogenes were rejected from the final consensus set), capable of unraveling complex cases that proved problematic to all the automated methods (e.g. the mitochondrial pseudogenes AC006326.2, .3, .4 and .5 in ENm001 (Figure 9)), and is the most effective method of discriminating processed and non-processed pseudogenes. Furthermore, the HAVANA group also produced high quality annotation for all coding and transcript loci in the ENCODE regions, in addition to pseudogenes (Harrow et al. 2006). The simultaneous annotation of genes and pseudogenes has the advantage of allowing accurate assignment of a locus as coding or not, which is essential in interpreting regional context, e.g., identifying coding and pseudogene members of the KIR and LILR gene families in ENm007, a task that proved problematic for all computational methods (Guigó et al. 2006; Harrow et al. 2006).

Pseudogene Activity and Functional Implications

Using pre-existing transcriptional data, several studies have shown that a good fraction (>5%) of the human pseudogenes were potentially transcribed (Frith et al. 2006; Harrison et al. 2005; Yano et al. 2004; Zheng et al. 2005). Our RACE analysis, which was directly targeted at pseudogene loci, provided experimental evidence that up to 10% of the ENCODE pseudogenes are transcribed in at least one of the 12 human tissues. Moreover, a survey of additional transcription data generated by the ENCODE project increases the estimate of the proportion of pseudogenes that are transcribed to nearly 20%. Comparison of our pseudogenes with putative promoters discovered by ChIP-chip experiments suggested that some transcribed pseudogenes might possess their own promoters. On the other hand, careful examination found a few cases where pseudogene transcription could have been initiated from the promoters of neighboring genes (e.g., a leukocyte immunoglobulin-like receptor pseudogene at ENm007: 476942-477651) or LINE elements (e.g., a RBPMS processed pseudogene at ENr223: 134009-134631). Such a “co-option” mechanism of pseudogene transcription has been suggested previously (Harrison et al. 2005) and has been experimentally demonstrated for retroposed genes (Bradley et al. 2004; Vinckenbosch et al. 2006). Certainly, recent non-processed pseudogenes can be transcriptionally active if the function of their promoters has not been lost entirely.

Although transcription of a pseudogene is not sufficient to indicate whether it has a meaningful biological function, our data showed that pseudogene transcription often

occurred at a low level and with a pattern of tissue or cell line specificity. These are similar to the transcriptional characteristics that have been observed for anti-sense RNA (Dahary et al. 2005; Katayama et al. 2005) and many intronic and intergenic transcripts whose biochemical functions are yet to be unraveled (Bertone et al. 2004; Cheng et al. 2005; Johnson et al. 2005; Willingham and Gingeras 2006). It would not, therefore, be surprising if pseudogenes proved to be one source of novel, functional non-coding RNAs.

We have also investigated the possibility that the ENCODE pseudogenes harbored known ncRNA genes (such as miRNA), but we found no such evidence; however, some non-processed pseudogenes were found to contain pseudogenes of ncRNA genes (data not shown).

Pseudogene Preservation

The prevalence of pseudogenes in mammalian genomes is itself of considerable interest. It is generally believed that this prevalence relates to increasing retrotransposition activity mediated by LINE (long interspersed elements) or other transposed elements (Brosius 1991; Esnault et al. 2000; Long et al. 2003; Maestre et al. 1995; Marques et al. 2005; Pavlicek et al. 2006; Wheelan et al. 2005). Our first multi-species survey of orthologous sequences for human pseudogenes supports this belief, showing that ~80% of the human processed pseudogenes arise from retroposed sequences specific to primate lineage. This is in accordance with previous studies suggesting that a burst of retrotransposition events occurred in ancestral primates about 40-50 MYA (Ohshima et al. 2003; Zhang et al. 2003). Many human retroposed genes also emerged from these events (Marques et al. 2005). Interestingly, the lack of mouse orthologs was used by two research groups as a criterion for assigning human processed pseudogenes (Torrents et al. 2003; van Baren and Brent 2006).

As either measured by sequence preservation or assessment of sequence constraints (either by phastOdds or Ka/Ks ratios), our study indicated that a small number of pseudogenes might have been under evolutionary constraints. Non-processed pseudogenes constitute the majority of such candidates. Subsequent detailed examination of evolutionary histories indicated that these are likely recent pseudogenes, deriving from sequences that have spent part of their histories as genes during evolution. In any case, our results strongly support the hypothesis that the sequences for processed pseudogenes are likely dead on arrival or at least lose their protein coding ability much sooner than those leading to human non-processed pseudogenes after their appearances during genome evolution.

Our analyses were based on MSA alignment data, and the possibility exists that our conclusions could be limited by the difficulty in identifying orthologous sequences in species very divergent from human (ref to MSA companion paper by Margulies et al. and another by King et al.). For example, the chicken or fish sequences aligned to a human non-processed pseudogene may not be orthologous but paralogous sequences from elsewhere in the genome. Therefore, our estimate of primate specific sequences (for both processed and non-processed human pseudogenes) is probably in a lower bound. It is

worth mentioning that our analyses with alignment data from a local aligner (TBA) and two global aligners (MAVID and MLAGAN) produced essentially the same results (data not shown), suggesting that our overall conclusions were not subject to the biases of the alignment algorithms. Furthermore, independent support of our results also came from the ENCODE variation group, whose analyses showed that the ENCODE pseudogenes had less nucleotide variation than ancient repeats.

Finally, our study found that the transcribed pseudogenes did not show significantly different evolutionary constraints compared to those not transcribed as measured by Ka/Ks, SNP density (Figure 7) or sequence similarity with respect to their parental genes (see supplement). A simple and intuitive inference of these data will thus hypothesize that pseudogene transcription is biological "noise" resulting from stochastic cellular transcription. However, these results do not exclude the possibility that some transcribed pseudogenes play biological roles, since it has been found that many experimentally determined functional elements (e.g., promoters) are not significantly conserved either (ref to the ENCOE main paper). On the other hand, in accordance with our finding, several recent studies have showed that conserved and transcribed pseudogenes are generally exceptional (tens out of thousands of human pseudogenes), but such pseudogenes could be good candidates with biochemical functions (Harrison et al. 2005; Svensson et al. 2006; Zheng et al. 2005).

Scaling Pseudogene Annotation to the Entire Human Genome

Using semi-automated analyses, we have defined 201 pseudogenes for 1% of the human genome. Interestingly, even with all the caveats of automated computational pipeline, this number agrees remarkably well with the ~20,000 pseudogenes identified for the whole genome using automated computational pipelines (Torrents et al. 2003; Zhang et al. 2003). However, the population of ENCODE pseudogenes is not necessarily a good representation for the entire genome simply because the regions were specially selected and included some unusually dense clusters of non-processed pseudogenes. If we only consider randomly picked targets in the ENCODE regions, there are 59 processed and 15 non-processed pseudogenes. This would extrapolate to approximately 10,000 pseudogenes in the human genome and thus put us in disagreement with previous reports. One factor contributing to this discrepancy is pseudogene fragments, short pieces of DNA related to protein coding genes and excluded from current analysis. In the future, we will expand our annotation to accommodate such fragments and other pseudogene sequences that have escaped detection currently.

MATERIALS and METHODS

Pseudogene Annotation

Five computational methods were used for identifying pseudogenes in the ENCODE regions. These methods use either protein or nucleotide sequences as queries (referred to as parents) to look for genomic sequences similar to human genes but unlikely to code for a protein product. Details of the computational algorithms and implementations have

been presented previously (Harrow et al. 2006; Zheng and Gerstein 2006) or can be found in Supplementary Materials.

Consensus approach for unifying pseudogene annotation We next developed a consensus approach accommodating the major feature in each of the individual methods. We first compared pseudogenes from the five methods with genes annotated by GENCODE annotation group (Harrow et al. 2006) and removed pseudogenes that occupied the same genomic position as a coding gene -- note: this happened as the pseudogene annotations were carried out independently of GENCODE gene annotation. This is a quite reasonable step as gene annotation should supersede pseudogene annotation when ambiguity arises, because the former can be tested with biochemical assays, but the latter is more difficult to establish experimentally. The October 2005 release of GENCODE annotation was used. We then made a union of the remaining pseudogenes to eliminate redundancy. A protein from UniProt (Bairoch et al. 2005) was assigned as the parent protein for each pseudogene in the union and pseudogenes without a recognizable parent protein were discarded. A sequence alignment was subsequently constructed between a pseudogene and its parent protein. This alignment was used to define the genomic boundary of a pseudogene and to distinguish processed from non-processed pseudogenes. In the end, all pseudogenes were examined manually by the VEGA/HAVANA annotation team to remove dubious pseudogenes and resolve ambiguous classification. Essentially, the final pseudogenes are genomic loci that cannot produce a protein coding transcript with the following features: (i) containing frameshifts or premature stop codons, or (ii) truncated fragments of the parent genes without such disablements and unlikely to be part of another gene structure (due to lacking evidence of transcription), or (iii) significant disruption in structure due to rearrangement compared to the parent sequences, or (iv) expert advice suggesting that even minor changes in the CDS would abolish function (e.g., in the cases of OR pseudogenes). The separation of processed and non-processed pseudogenes followed the general strategy of HAVANA Method (see supplement).

Pseudogene Transcription

The degree of pseudogene transcription was assessed with evidence from multiple sources. Most of the data were obtained from the ENCODE gene and transcript group (ref to ENCODE main paper). These included transcribed regions (transfrags) identified by tiling microarray covered non-repetitive sequences within the ENCODE regions using RNA samples from 11 cell lines or conditions, 5'-specific Cap Analysis Gene Expression tags from 15 tissues (Shiraki et al. 2003), and Paired-End 5' and 3' diTags from HCT116 and MCF7 cells (Ng et al. 2005). We also used mRNA/ESTs in public databases as a source of expression evidence. When comparing pseudogenes with transfrags, we would only assign transcription evidence to a pseudogene if at least one of its "exons" overlapping >50 nt of a transfrag. In the analysis of expression tags a pseudogene was considered to be transcriptionally active if there was a CAGE tag on the same strand near its 5' end, or if a pair of ditags spanned this pseudogene. In both cases we only considered tags (5' or 5'/3') that were <100 bp from the ends of a pseudogene. Spliced

ESTs or mRNAs were assigned to a pseudogene locus only if they were mapped to this region much better (or uniquely) than any other genomic locations of the human genome.

We have also chosen 160 (49 non-processed and 111 processed) of our pseudogenes randomly to test for expression by the use of locus specific RACE/microarray analysis (Kapranov et al. 2005). Poly A+ RNA from 12 tissues (brain, colon, heart, kidney, liver, lung, muscle, placenta, small intestine, spleen, stomach and testis) were extracted and used as substrates for these studies. Primers specific to pseudogenes or with 0~3 mismatches with their parent genes were used for the RACE experiments. The RACE products were pooled to four groups and then hybridized to ENCODE tiling microarrays. Genomic fragments corresponding to RACE products were identified and called RACEfrags, as described previously (Kapranov et al. 2005). Non-specific RACEfrags (i.e., present in more than one of the four pools) were discarded. In the meantime, we also scanned all RACEfrags against the entire human genome and kept the "unique" ones, which contained at least one stretch (>25 nt) of nucleotide sequence that did not share >85% sequence identity with a sequence in other genomic location. We considered a pseudogene to be transcribed if such a unique RACEfrag(s) was detected from the location of RACE primer up to -5 kb upstream of a pseudogene.

Pseudogene Conservation and Evolutionary History

The preservation of the ENCODE human pseudogenes was assessed using data derived from multi-species sequence alignment constructed by the ENCODE-MSA group (ref to the main ENCODE paper and MSA companion paper by Margulies et al.). The alignment data were obtained from this site (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/encode/alignments/SEP-2005/), and the MSA alignments were used to infer ortholog assignment for each of our pseudogenes. The alignment block containing a pseudogene was designated as orthologous regions for this pseudogene. A pseudogene (or its exon) was considered as "preserved" in a species if >50 bp and 20% of this pseudogene was aligned to its orthologous sequence from that species. We then computed pair-wise sequence identity from the alignment, excluding gaps. Data for other genomic features (e.g., exons, introns and CDS) were calculated in a similar fashion and available as the Supplement. For assessing sequence disablements, we aligned a pseudogene or its orthologous sequences to the parent genes using the programs GeneWise (Birney et al. 2004) (for non-processed pseudogenes) or FASTA (Pearson et al. 1997) (for processed pseudogenes). In all analyses disablements were defined as premature stop codons (i.e., nonsense) or frameshift mutations present in the alignment.

SNP data were obtained from the UCSC browser (genome.ucsc.edu) and Ka/Ks ratios were analyzed by the software package PAML (Yang 1997). Indels were not included in this study.

Supplementary Materials

Supplementary Materials are available on Genome Research website and additional figures, tables and data are available at
http://www.pseudogene.org/ENCODE/supplement/

Acknowledgements

The authors would like to thank the ENCODE Project Consortium for making their data publicly available, the MSA group for providing the multi-species sequence alignment data, and David Haussler, Elliott H. Margulies, Adam Siepel, and Zhaolei Zhang for valuable discussions and comments. This work has been funded by National Human Genome Research Institute (NHGRI)/National Institutes of Health (NIH) grants to the ENCODE project, especially to the following ENCODE subgroups: GENCODE [# U01HG03150], Yale [# U01HG03156], and Affymetrix, Inc [# U01HG03147]. Portions of this study have also been funded in part with federal funds from the National Cancer Institute (NCI) and NIH under Contract No. N01-CO-12400 (T.R.G.), NCI [# NO1-CO-12400/22XS013A] (to R. B.), by the Swiss National Science Foundation and the Child Care Foundation (S.E.A and A.R.), and Spanish Ministry of Education and Science (R.G.).

Table 1. Numbers of ENCODE Consensus Pseudogenes with Transcriptional Evidence.

	Transfrags	CAGE	DiTag	RACEfrags	mRNA/EST
Transfrags	105 *	8	2	5	14
CAGE		8	1	0	1
DiTag			2	0	0
RACEfrags				14	5
mRNA/EST					21

* About 50% of the transfrags intersecting pseudogenes could be mapped to multiple locations in the human genome. As a result, cross-hybridization might be the source of transcription evidence for one half of these pseudogenes.

Figure Legends

Figure 1. Comparison of results from five methods of pseudogene identification. (A) Pseudogenes annotated by a method were binned into groups based on the number of methods that recognized them as pseudogenes. In this scheme method-specific pseudogenes were labeled as (found by) "1" method. (B) A four-way comparison of pseudogenes identified by HAVANA, PseudoPipe, retroFinder, and pseudoFinder. Note: one pseudogene could overlap more than one pseudogene from other method(s).

Figure 2. The distribution of genes and the final 201 consensus pseudogenes within 44 ENCODE regions. Both genes and pseudogenes were concentrated in the manually picked regions (001-014).

Figure 3. A pseudogene with multiple evidence of transcription. This is a processed pseudogene identified by all five methods (in pink color). The evidence of transcription includes RACEfrags, EST, GIS-PET, Riken CAGE, and transfrags (Affy RNA or Yale TARs). Near its 5' end there is a putative promoter region ("ENCODE_ChIP", top) derived from many ChIP-chip experiments targeted at DNA elements regulating transcription.

Figure 4. Preservation of human genomic components in other species. The number of human pseudogenes (or genes) with orthologous sequences in individual species was computed and then plotted (by normalization with the total number in human) against each species. Only exons (or *pseudoexons*) were used in these analyses; NPS and PS represent non-processed and processed pseudogenes, respectively. Data were derived from sequence alignment constructed by the program TBA except PS-mavid, which was by MAVID. Note that species with sequences available for ENm001 region only are omitted in this figure. A more comprehensive plot (of Fig 4 and also Fig 5A) with data for introns and other genomic data can be found in the Supplement (Fig S1, S2). The data for non-mammalian species (right of the vertical line) should be taken with more caution because ortholog assignments for these species are likely more difficult.

Figure 5. ENCODE pseudogenes overall exhibit a characteristic pattern of neutral evolution. (A) The orthologous sequences of each human genomic component (e.g., pseudogene) were retrieved from MSA data, and pair-wise nucleotide sequence identity was calculated. Shown here are the means for each type of components (data labeled as Figure 4). A line representing neutral evolution is also shown using data derived from four-fold degenerate sites. (B) A score based on the log-likelihood of observing a genomic fragment under a model of constrained versus neutral evolution was computed for individual exons of genes or pseudogenes using the phastOdds program (Siepel et al. 2005). These scores were then normalized by exon length and plotted here as a histogram. A value near zero or negative indicates that the evolution of a sequence can be described better by a neutral model.

Figure 6. Comparison of sequence conservation for genes and pseudogenes in the context of adjacent genomic sequences. The orthologous sequences in chimp, macaque, mouse, and dog were retrieved from the MSA data for protein "coding" regions (CDS) of genes and pseudogenes. Their regions were divided into 10 blocks, and pair-wise nucleotide sequence identities were calculated for each block. Data shown here are the means for all genes, processed (PS) or non-processed (NPS) pseudogenes. For comparison, 500 bp upstream and downstream sequences of CDSs were also analyzed. The p-values of t-test for the differences between genes and pseudogenes (for all four species) and between NPS and PS (in chimp and macaque) are <0.01.

Figure 7. Comparison of Ka/Ks ratio and SNP density for genes and pseudogenes. Only the CDS of a gene or pseudogene was used for analyses of Ka/Ks ratio and SNP density (number of SNP per 300 nucleotides). The Ka/Ks ratio was derived from the sequences between baboon and human. Data for transcribed pseudogenes are circled, and they are not statistically significant from the rest.

Figure 8. Detection and disabled pattern of pseudogene orthologs. For each pseudogene its orthologous sequences were retrieved and compared to the parent protein sequence. Respectively, boxes and circles represent whether a pseudogene ortholog is detected or not in a species. A cross (X) means that the hypothetical CDS is disabled. Data for non-mammalian species are not shown. The five pseudogenes shown here are (from A to E), CTA-440B3.1-001 (ENm004, PS), RP11-374F3.2-001 (ENr111, PS), RP11-98F14.4-001 (ENr132, PS), AC087380.17-001 (ENm009, NPS), and AC087380.14-001 (ENm009, NPS).

Figure 9. Complexity in pseudogene annotation – insertion of one pseudogene into another. A set of "nested" pseudogenes (in green) was found in the ENm001 region with protein homology (shown in blue) supporting the annotation. This arrangement appears to have been generated through the insertion of a heterogeneous nuclear ribonucleoprotein A1 (HNRPA1) processed pseudogene (1) into the genome on the negative strand. This was followed by a second insertion event in which a transcript originating from the mitochondrial genome was transposed into the HNRPA1 pseudogene sequence. Gene order and orientation suggest that this mitochondria-derived sequence has undergone further rearrangement, including deletions, to leave an NADH dehydrogenase 2 (MTND2) pseudogene (2a) and an NADH dehydrogenase 4 (MTND4) pseudogene (2b) on the positive strand and a cytochrome B (CYTB) pseudogene (2c) on the negative strand. A view of the protein alignment for the 5' end of the HNRPA1 pseudogene (in yellow) is shown with an in-frame stop codon (indicated by *) and a shift from frame +2 to +3 (highlighted by the red box) clearly visible.

References

Bairoch, A., R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33:** D154-159.

Balakirev, E.S. and F.J. Ayala. 2003. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* **37:** 123-151.

Bertone, P., V. Stolc, T.E. Royce, J.S. Rozowsky, A.E. Urban, X. Zhu, J.L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242-2246.

Birney, E., M. Clamp, and R. Durbin. 2004. GeneWise and Genomewise. *Genome Res* **14:** 988-995.

Bischof, J.M., A.P. Chiang, T.E. Scheetz, E.M. Stone, T.L. Casavant, V.C. Sheffield, and T.A. Braun. 2006. Genome-wide identification of pseudogenes capable of disease-causing gene conversion. *Hum Mutat* **27:** 545-552.

Blanchette, M., W.J. Kent, C. Riemer, L. Elnitski, A.F. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, D. Haussler, and W. Miller. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14:** 708-715.

Bradley, J., A. Baltus, H. Skaletsky, M. Royce-Tolland, K. Dewar, and D.C. Page. 2004. An X-to-autosome retrogene is required for spermatogenesis in mice. *Nat Genet* **36:** 872-876.

Bray, N. and L. Pachter. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* **14:** 693-699.

Brosius, J. 1991. Retroposons--seeds of evolution. *Science* **251:** 753.

Brudno, M., C.B. Do, G.M. Cooper, M.F. Kim, E. Davydov, E.D. Green, A. Sidow, and S. Batzoglou. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13:** 721-731.

Bustamante, C.D., R. Nielsen, and D.L. Hartl. 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Mol Biol Evol* **19:** 110-117.

Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D.K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D.S. Gerhard, and T.R. Gingeras. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149-1154.

Coin, L. and R. Durbin. 2004. Improved techniques for the identification of pseudogenes. *Bioinformatics* **20 Suppl 1:** I94-I100.

Consortium, C.S.a.A. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69-87.

Consortium, E.P. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636-640.

Dahary, D., O. Elroy-Stein, and R. Sorek. 2005. Naturally occurring antisense: transcriptional leakage or real overlap? *Genome Res* **15:** 364-368.

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

Frith, M.C., L.G. Wilming, A. Forrest, H. Kawaji, S.L. Tan, C. Wahlestedt, V.B. Bajic, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, T.L. Bailey, and L. Huminiecki. 2006. Pseudo-messenger RNA: phantoms of the transcriptome. *PLoS Genet* **2:** e23.

Gilad, Y., O. Man, and G. Glusman. 2005. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* **15:** 224-230.

Glusman, G., I. Yanai, I. Rubin, and D. Lancet. 2001. The complete human olfactory subgenome. *Genome Res* **11:** 685-702.

Gojobori, T., K. Ishii, and M. Nei. 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J Mol Evol* **18:** 414-423.

Goodman, M. 1999. The genomic record of Humankind's evolutionary roots. *Am J Hum Genet* **64:** 31-39.

Goodman, M., C.A. Porter, J. Czelusniak, S.L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C.P. Groves. 1998. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* **9:** 585-598.

Gu, X. and W.H. Li. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J Mol Evol* **40:** 464-473.

Guigó, R., P. Flicek, J.F. Abril, A. Reymond, J. Lagarde, F. Denoeud, S.E. Antonarakis, M. Ashburner, V.B. Bajic, E. Birney, R. Castelo, E. Eyras, T.R. Gingeras, P. Good, J. Harrow, S. Lewis, T. Hubbard, and M.G. Reese. 2006. EGASP: the human ENCODE genome annotation assessment project. *Genome Biol* **7:** S2.

Harrison, P.M., H. Hegyi, S. Balasubramanian, N.M. Luscombe, P. Bertone, N. Echols, T. Johnson, and M. Gerstein. 2002. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res* **12:** 272-280.

Harrison, P.M., D. Zheng, Z. Zhang, N. Carriero, and M. Gerstein. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res* **33:** 2374-2383.

Harrow, J., F. Denoeud, A. Frankish, A. Reymond, C. C.K., J. Chrast, J. Lagarde, J.G.R. Gilbert, R. Storey, D. Swarbreck., C. Ucla, T. Hubbard, S.E. Antonarakis, and R. Guigó. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol* **7:** S4.

Hurteau, G.J. and S.D. Spivack. 2002. mRNA-specific reverse transcription-polymerase chain reaction from human tissue extracts. *Anal Biochem* **307:** 304-315.

Johnson, J.M., S. Edwards, D. Shoemaker, and E.E. Schadt. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* **21:** 93-102.

Jurka, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* **94:** 1872-1877.

Kapranov, P., J. Drenkow, J. Cheng, J. Long, G. Helt, S. Dike, and T.R. Gingeras. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* **15:** 987-997.

Katayama, S., Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C.C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K.C. Pang, J. Hallinan, J. Mattick, D.A. Hume, L. Lipovich, S. Batalov, P.G. Engstrom, Y. Mizuno, M.A. Faghihi, A. Sandelin, A.M. Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard, and C. Wahlestedt. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309:** 1564-1566.

Kenmochi, N., T. Kawaguchi, S. Rozen, E. Davis, N. Goodman, T.J. Hudson, T. Tanaka, and D.C. Page. 1998. A map of 75 human ribosomal protein genes. *Genome Res* **8:** 509-523.

Khelifi, A., L. Duret, and D. Mouchiroud. 2005. HOPPSIGEN: a database of human and mouse processed pseudogenes. *Nucleic Acids Res* **33:** D59-66.

Kleene, K.C., E. Mulligan, D. Steiger, K. Donohue, and M.A. Mastrangelo. 1998. The mouse gene encoding the testis-specific isoform of Poly(A) binding protein (Pabp2) is an expressed retroposon: intimations that gene expression in spermatogenic cells facilitates the creation of new genes. *J Mol Evol* **47:** 275-281.

Korneev, S.A., J.H. Park, and M. O'Shea. 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.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczky R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L. Naylor R.S. Kucherlapati D.L. Nelson G.M. Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saurin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H. Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglou E. Birney P. Bork

D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M. Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kaspryzk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G. Slater A.F. Smit E. Stupka J. Szustakowski D. Thierry-Mieg J. Thierry-Mieg L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos M.J. Morgan P. de Jong J.J. Catanese K. Osoegawa H. Shizuya S. Choi and Y.J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Li, W.H., T. Gojobori, and M. Nei. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature* **292:** 237-239.

Li, W.H., C.I. Wu, and C.C. Luo. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol* **21:** 58-71.

Lindblad-Toh, K. C.M. Wade T.S. Mikkelsen E.K. Karlsson D.B. Jaffe M. Kamal M. Clamp J.L. Chang E.J. Kulbokas, 3rd M.C. Zody E. Mauceli X. Xie M. Breen R.K. Wayne E.A. Ostrander C.P. Ponting F. Galibert D.R. Smith P.J. DeJong E. Kirkness P. Alvarez T. Biagi W. Brockman J. Butler C.W. Chin A. Cook J. Cuff M.J. Daly D. DeCaprio S. Gnerre M. Grabherr M. Kellis M. Kleber C. Bardeleben L. Goodstadt A. Heger C. Hitte L. Kim K.P. Koepfli H.G. Parker J.P. Pollinger S.M. Searle N.B. Sutter R. Thomas C. Webber J. Baldwin A. Abebe A. Abouelleil L. Aftuck M. Ait-Zahra T. Aldredge N. Allen P. An S. Anderson C. Antoine H. Arachchi A. Aslam L. Ayotte P. Bachantsang A. Barry T. Bayul M. Benamara A. Berlin D. Bessette B. Blitshteyn T. Bloom J. Blye L. Boguslavskiy C. Bonnet B. Boukhgalter A. Brown P. Cahill N. Calixte J. Camarata Y. Cheshatsang J. Chu M. Citroen A. Collymore P. Cooke T. Dawoe R. Daza K. Decktor S. DeGray N. Dhargay K. Dooley P. Dorje K. Dorjee L. Dorris N. Duffey A. Dupes O. Egbiremolen R. Elong J. Falk A. Farina S. Faro D. Ferguson P. Ferreira S. Fisher M. FitzGerald K. Foley C. Foley A. Franke D. Friedrich D. Gage M. Garber G. Gearin G. Giannoukos T. Goode A. Goyette J. Graham E. Grandbois K. Gyaltsen N. Hafez D. Hagopian B. Hagos J. Hall C. Healy R. Hegarty T. Honan A. Horn N. Houde L. Hughes L. Hunnicutt M. Husby B. Jester C. Jones A. Kamat B. Kanga C. Kells D. Khazanovich A.C. Kieu P. Kisner M. Kumar K. Lance T. Landers M. Lara W. Lee J.P. Leger N. Lennon L. Leuper S. LeVine J. Liu X. Liu Y. Lokyitsang T. Lokyitsang A. Lui J. Macdonald J. Major R. Marabella K. Maru C. Matthews S. McDonough T. Mehta J. Meldrim A. Melnikov L. Meneus A. Mihalev T. Mihova K. Miller R. Mittelman V. Mlenga L. Mulrain G. Munson A. Navidi J. Naylor T. Nguyen N. Nguyen C. Nguyen R. Nicol N. Norbu C. Norbu N. Novod T. Nyima P. Olandt B. O'Neill K. O'Neill S. Osman L. Oyono C. Patti D. Perrin P. Phunkhang F. Pierre M. Priest A. Rachupka S. Raghuraman R. Rameau V. Ray C. Raymond F. Rege C. Rise J. Rogers P. Rogov J. Sahalie S. Settipalli T. Sharpe T. Shea M. Sheehan N. Sherpa J. Shi D. Shih J. Sloan C. Smith T. Sparrow J. Stalker N. Stange-Thomann S. Stavropoulos C. Stone S. Stone S. Sykes P. Tchuinga P. Tenzing S. Tesfaye D. Thoulutsang Y.

Thoulutsang K. Topham I. Topping T. Tsamla H. Vassiliev V. Venkataraman A. Vo T. Wangchuk T. Wangdi M. Weiand J. Wilkinson A. Wilson S. Yadav S. Yang X. Yang G. Young Q. Yu J. Zainoun L. Zembek A. Zimmer and E.S. Lander. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803-819.

Long, M., E. Betran, K. Thornton, and W. Wang. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4:** 865-875.

Maestre, J., T. Tchenio, O. Dhellin, and T. Heidmann. 1995. mRNA retroposition in human cells: processed pseudogene formation. *Embo J* **14:** 6333-6338.

Marques, A.C., I. Dupanloup, N. Vinckenbosch, A. Reymond, and H. Kaessmann. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* **3:** e357.

Mighell, A.J., N.R. Smith, P.A. Robinson, and A.F. Markham. 2000. Vertebrate pseudogenes. *FEBS Lett* **468:** 109-114.

Ng, P., C.L. Wei, W.K. Sung, K.P. Chiu, L. Lipovich, C.C. Ang, S. Gupta, A. Shahab, A. Ridwan, C.H. Wong, E.T. Liu, and Y. Ruan. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* **2:** 105-111.

Ohshima, K., M. Hattori, T. Yada, T. Gojobori, Y. Sakaki, and N. Okada. 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.

Ota, T. and M. Nei. 1995. Evolution of immunoglobulin VH pseudogenes in chickens. *Mol Biol Evol* **12:** 94-102.

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

Pearson, W.R., T. Wood, Z. Zhang, and W. Miller. 1997. Comparison of DNA sequences with protein sequences. *Genomics* **46:** 24-36.

Reymond, A., V. Marigo, M.B. Yaylaoglu, A. Leoni, C. Ucla, N. Scamuffa, C. Caccioppoli, E.T. Dermitzakis, R. Lyle, S. Banfi, G. Eichele, S.E. Antonarakis, and A. Ballabio. 2002. Human chromosome 21 gene expression atlas in the mouse. *Nature* **420:** 582-586.

Ruud, P., O. Fodstad, and E. Hovig. 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.

Schmitz, J., G. Churakov, H. Zischler, and J. Brosius. 2004. A novel class of mammalian-specific tailless retropseudogenes. *Genome Res* **14:** 1911-1915.

Shemesh, R., A. Novik, S. Edelheit, and R. Sorek. 2006. Genomic fossils as a snapshot of the human transcriptome. *Proc Natl Acad Sci U S A* **103:** 1364-1369.

Shiraki, T., S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* **100:** 15776-15781.

Siepel, A., G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller, and D. Haussler. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034-1050.

Smith, R.D., C.W. Ogden, and M.A. Penny. 2001. Exclusive amplification of cDNA template (EXACT) RT-PCR to avoid amplifying contaminating genomic pseudogenes. *Biotechniques* **31:** 776-778, 780, 782.

Strichman-Almashanu, L.Z., M. Bustin, and D. Landsman. 2003. Retroposed copies of the HMG genes: a window to genome dynamics. *Genome Res* **13:** 800-812.

Svensson, O., L. Arvestad, and J. Lagergren. 2006. Genome-Wide Survey for Biologically Functional Pseudogenes. *PLoS Comput Biol* **2:** e46.

Torrents, D., M. Suyama, E. Zdobnov, and P. Bork. 2003. A genome-wide survey of human pseudogenes. *Genome Res* **13:** 2559-2567.

van Baren, M.J. and M.R. Brent. 2006. Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res* **16:** 678-685.

Vanin, E.F. 1985. Processed pseudogenes: characteristics and evolution. *Annu Rev Genet* **19:** 253-272.

Vinckenbosch, N., I. Dupanloup, and H. Kaessmann. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A* **103:** 3220-3225.

Wheelan, S.J., Y. Aizawa, J.S. Han, and J.D. Boeke. 2005. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res* **15:** 1073-1078.

Willingham, A.T. and T.R. Gingeras. 2006. TUF love for "junk" DNA. *Cell* **125:** 1215-1220.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555-556.

Yano, Y., R. Saito, N. Yoshida, A. Yoshiki, A. Wynshaw-Boris, M. Tomita, and S. Hirotsune. 2004. A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *J Mol Med* **82:** 414-422.

Zhang, Z., N. Carriero, D. Zheng, J. Karro, P.M. Harrison, and M. Gerstein. 2006. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*.

Zhang, Z. and M. Gerstein. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* **31:** 5338-5348.

Zhang, Z. and M. Gerstein. 2004. Large-scale analysis of pseudogenes in the human genome. *Curr Opin Genet Dev* **14:** 328-335.

Zhang, Z., P.M. Harrison, Y. Liu, and M. Gerstein. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13:** 2541-2558.

Zheng, D. and M. Gerstein. 2006. A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biol* **7:** S13.

Zheng, D., Z. Zhang, P.M. Harrison, J. Karro, N. Carriero, and M. Gerstein. 2005. Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *J Mol Biol* **349:** 27-45.

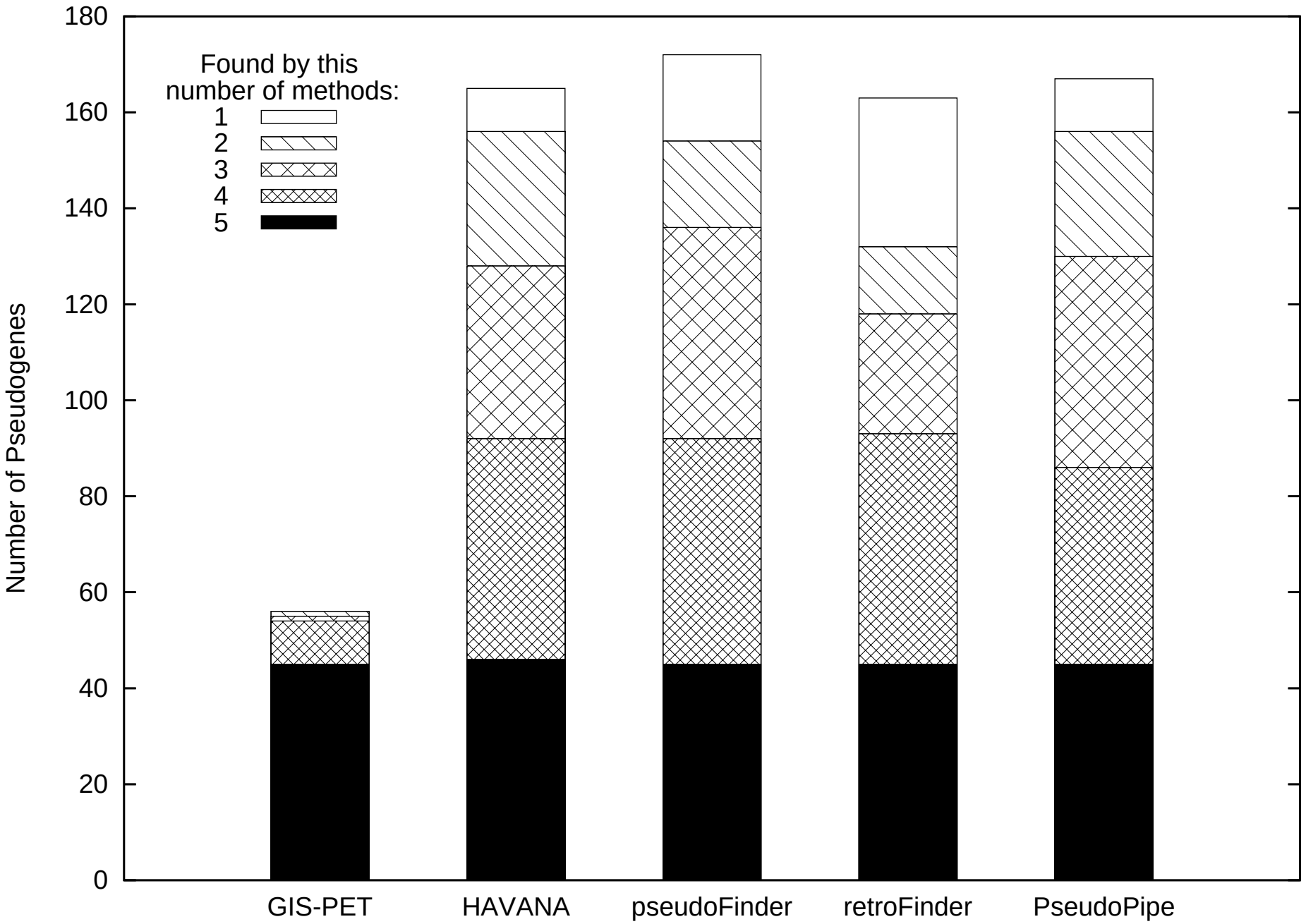
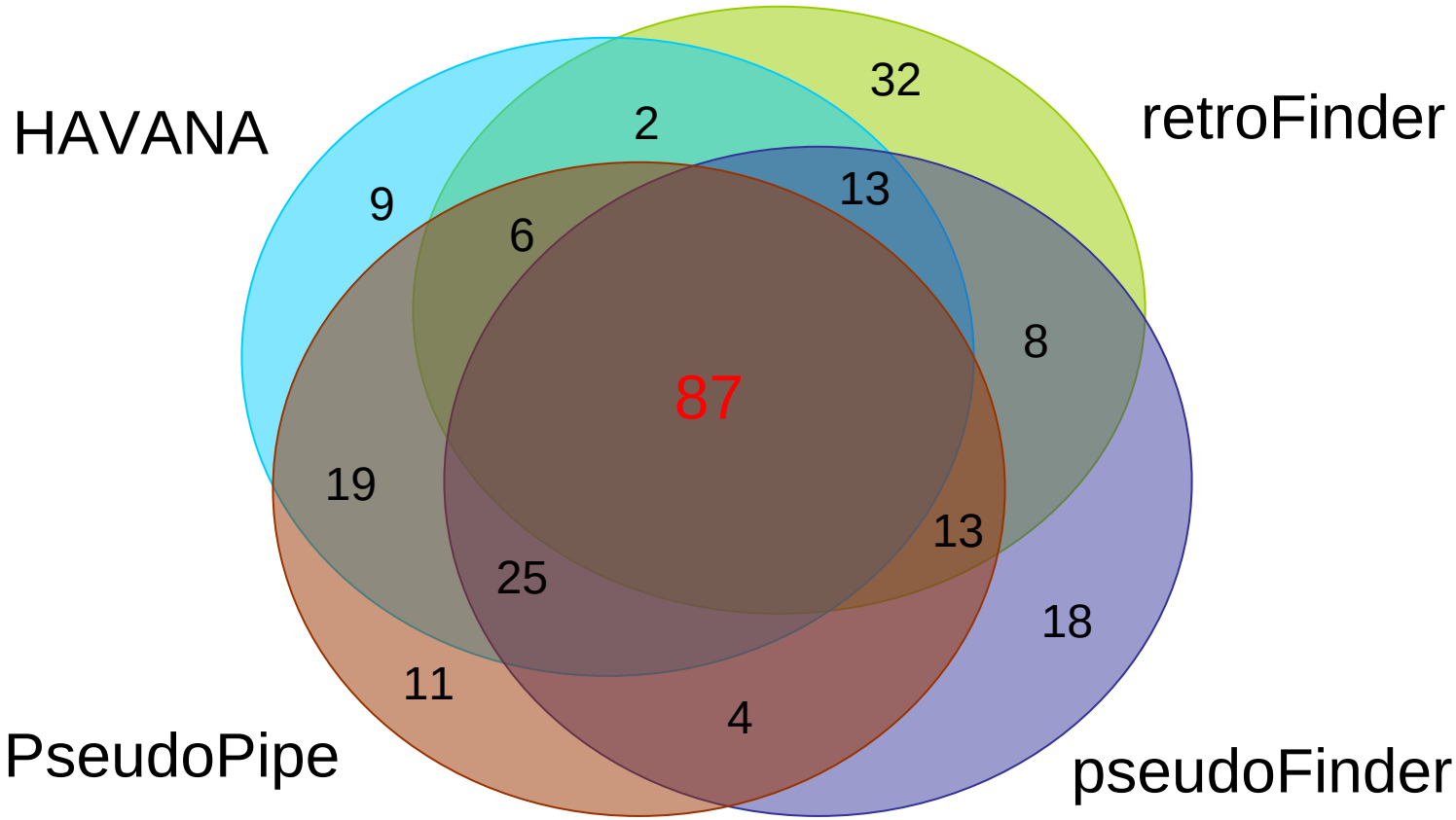
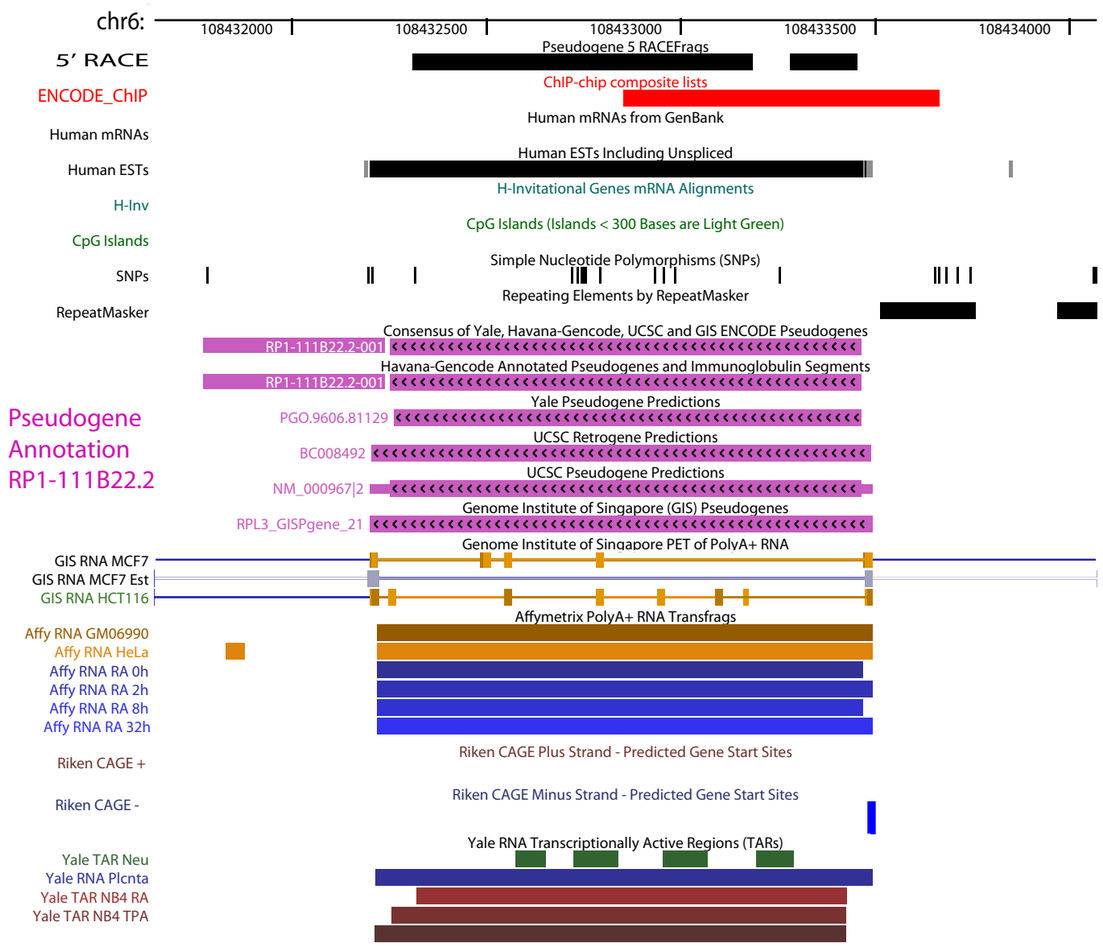
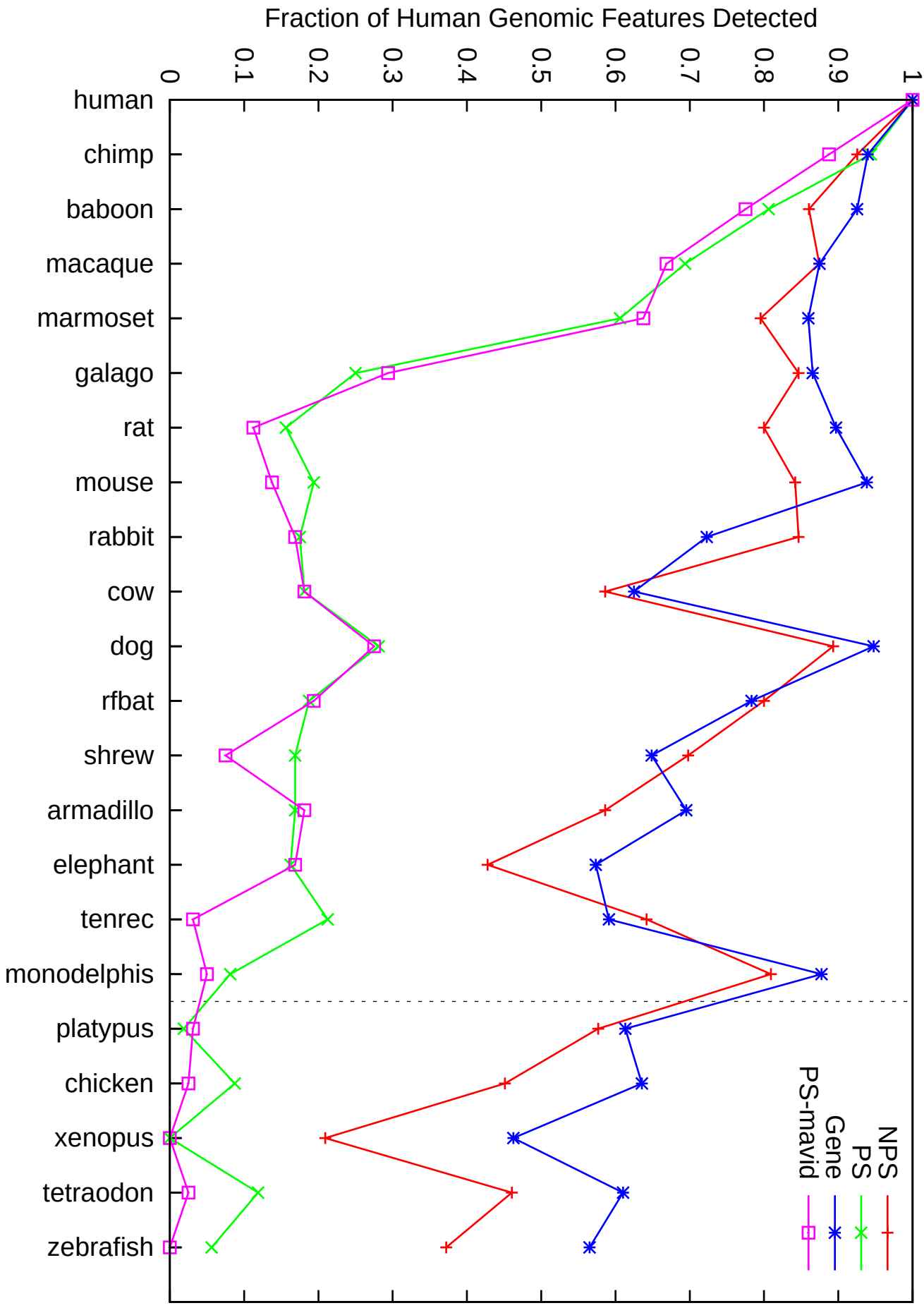

Found by this number of methods:
1
2
3
4
5
Number of Pseudogenes
0
20
40
60
80
100
120
140
160
180
GIS-PET
HAVANA
pseudoFinder
retroFinder
PseudoPipe

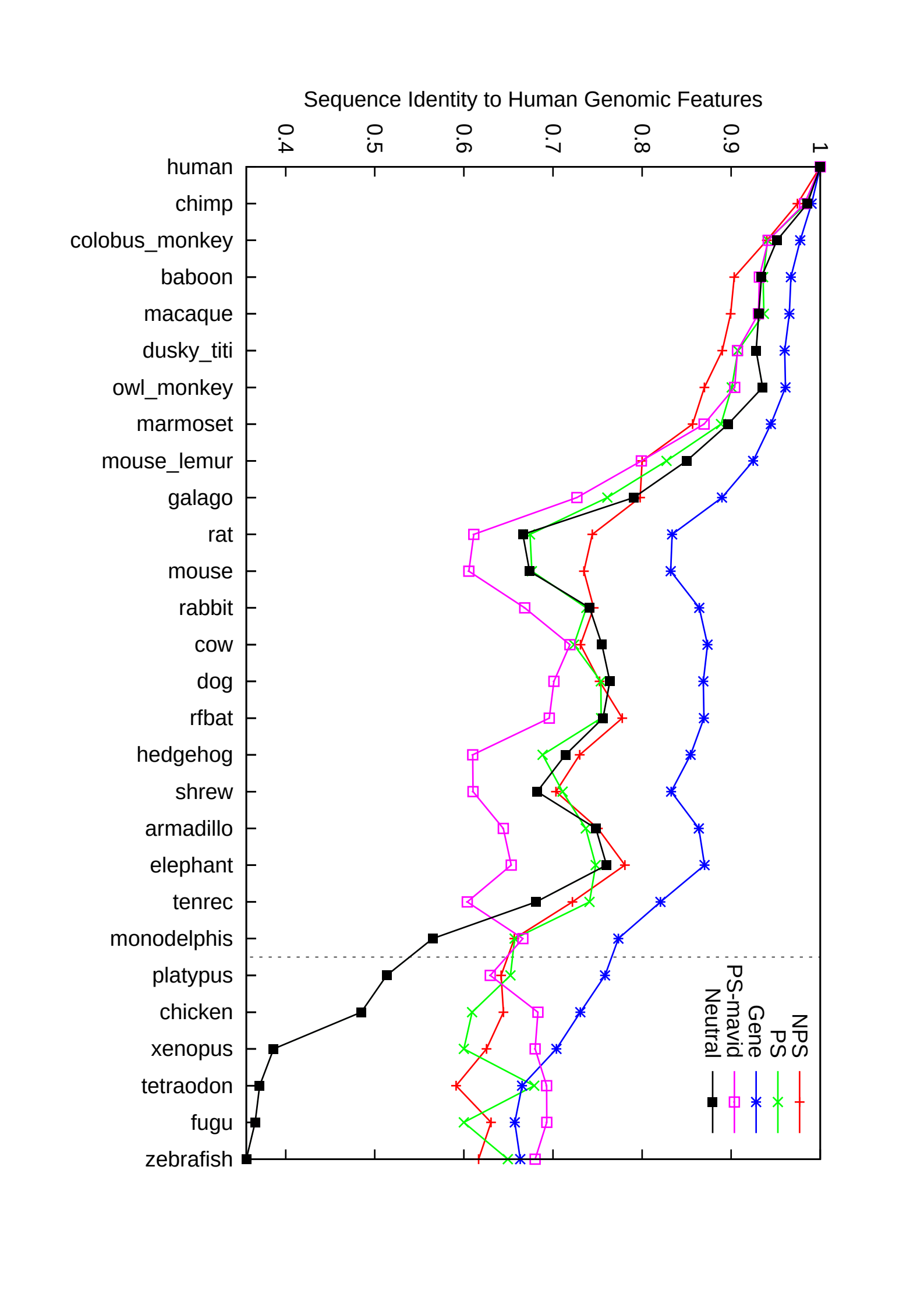
Fig1B

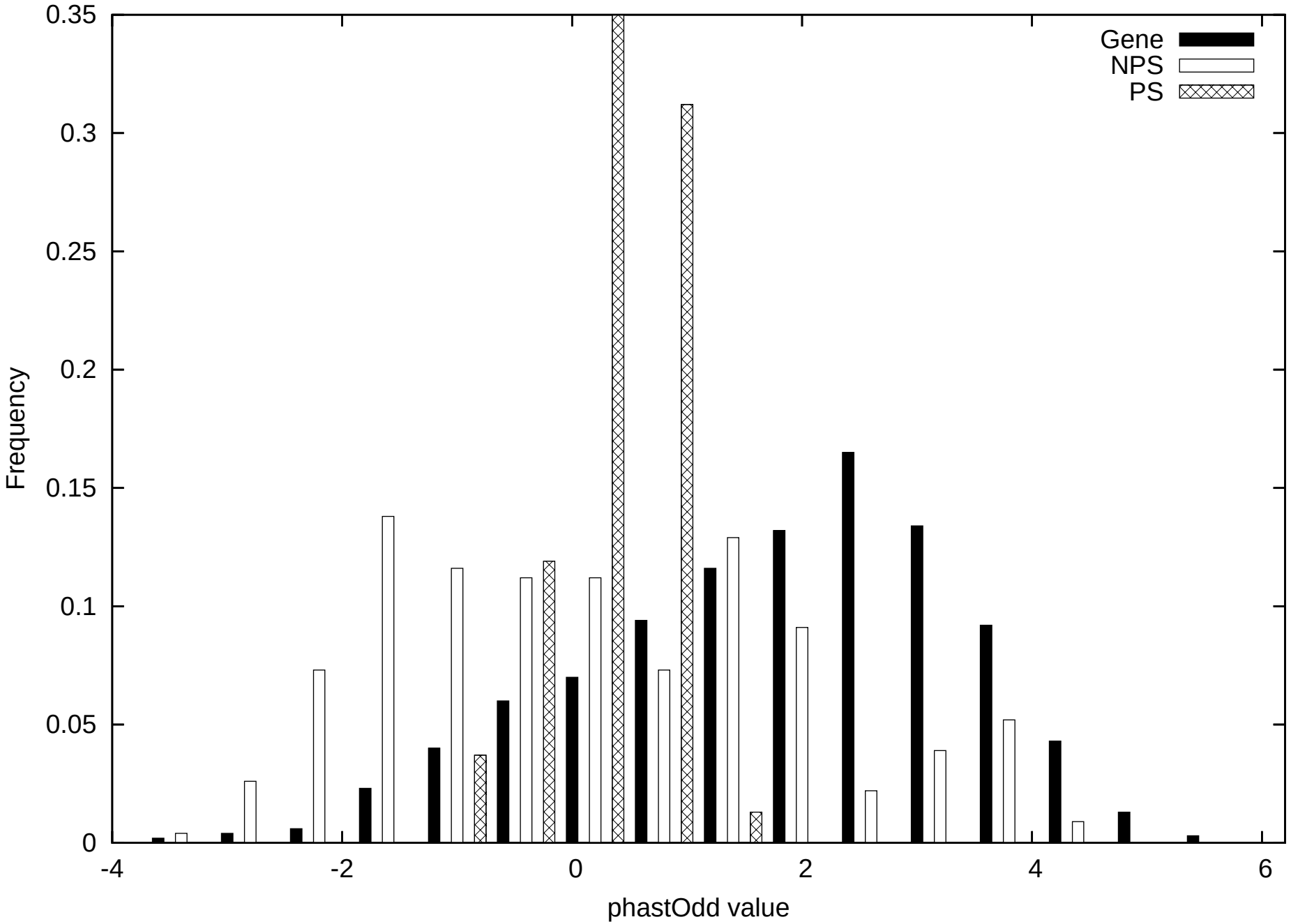

HAVANA
retroFinder
PseudoPipe
pseudoFinder
9
2
32
6
13
8
87
19
13
25
18
11
4

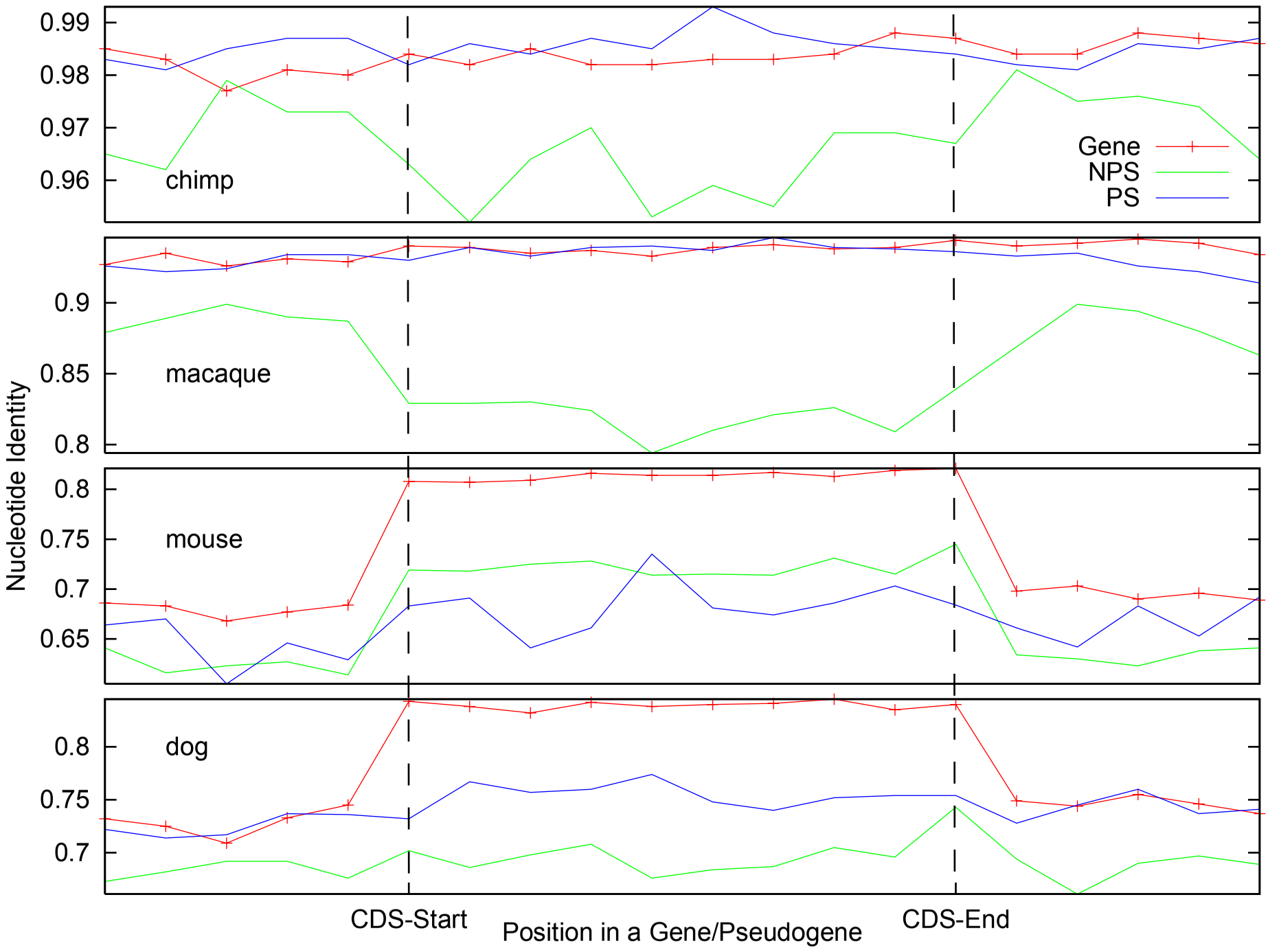
Number of Genes/Pseudogenes
0
5
10
15
20
25
30
35
40
45
50
ENCODE Regions
001
002
003
004
005
006
007
008
009
010
011
012
013
014
111
112
113
114
121
122
123
131
132
133
211
212
213
221
222
223
231
232
233
311
312
313
321
322
323
324
331
332
333
334
Manual Picks
Random Picks
Gene
Non-Processed Pseudogene
Processed Pseudogene

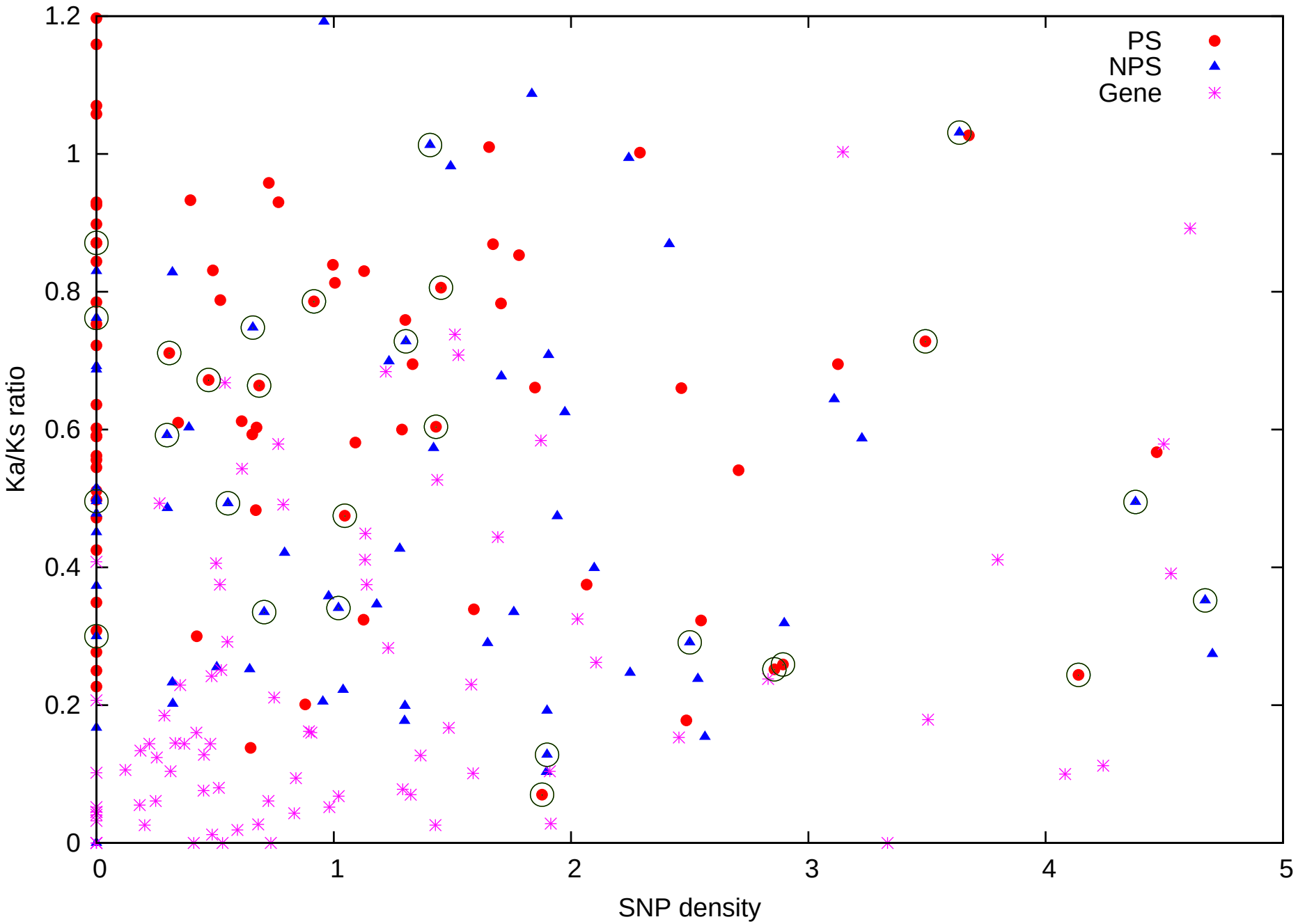

chr6:
108432000
108432500
108433000
108433500
108434000
5′ RACE
Pseudogene 5 RACEFrags
ENCODE_ChIP
ChIP-chip composite lists
Human mRNAs
Human mRNAs from GenBank
Human ESTs
Human ESTs Including Unspliced
H-Inv
H-Invitational Genes mRNA Alignments
CpG Islands
CpG Islands (Islands < 300 Bases are Light Green)
SNPs
Simple Nucleotide Polymorphisms (SNPs)
RepeatMasker
Repeating Elements by RepeatMasker
Consensus of Yale, Havana-Gencode, UCSC and GIS ENCODE Pseudogenes
RP1-111B22.2-001
Havana-Gencode Annotated Pseudogenes and Immunoglobulin Segments
RP1-111B22.2-001
Pseudogene Annotation RP1-111B22.2
Yale Pseudogene Predictions
PGO.9606.81129
UCSC Retrogene Predictions
BC008492
UCSC Pseudogene Predictions
NM_000967|2
Genome Institute of Singapore (GIS) Pseudogenes
RPL3_GISPgene_21
Genome Institute of Singapore PET of PolyA+ RNA
GIS RNA MCF7
GIS RNA MCF7 Est
GIS RNA HCT116
Affymetrix PolyA+ RNA Transfrags
Affy RNA GM06990
Affy RNA HeLa
Affy RNA RA 0h
Affy RNA RA 2h
Affy RNA RA 8h
Affy RNA RA 32h
Riken CAGE +
Riken CAGE Plus Strand - Predicted Gene Start Sites
Riken CAGE -
Riken CAGE Minus Strand - Predicted Gene Start Sites
Yale RNA Transcriptionally Active Regions (TARs)
Yale TAR Neu
Yale RNA Plcnta
Yale TAR NB4 RA
Yale TAR NB4 TPA


Fraction of Human Genomic Features Detected
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
human
chimp
baboon
macaque
marmoset
galago
rat
mouse
rabbit
cow
dog
rfbat
shrew
armadillo
elephant
tenrec
monodelphis
platypus
chicken
xenopus
tetraodon
zebrafish
NPS
PS
Gene
PS-mavid


Sequence Identity to Human Genomic Features
0.4
0.5
0.6
0.7
0.8
0.9
1
human
chimp
colobus_monkey
baboon
macaque
dusky_titi
owl_monkey
marmoset
mouse_lemur
galago
rat
mouse
rabbit
cow
dog
rfbat
hedgehog
shrew
armadillo
elephant
tenrec
monodelphis
platypus
chicken
xenopus
tetraodon
fugu
zebrafish
NPS
PS
Gene
PS-mavid
Neutral


Gene
NPS
PS
Frequency
phastOdd value
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-2
0
2
4
6


0.99
0.98
0.97
0.96
chimp
Gene
NPS
PS
0.9
0.85
0.8
macaque
0.8
0.75
0.7
0.65
mouse
0.8
0.75
0.7
dog
Nucleotide Identity
CDS-Start
CDS-End
Position in a Gene/Pseudogene


PS
NPS
Gene
Ka/Ks ratio
SNP density
0
0.2
0.4
0.6
0.8
1
1.2
0
1
2
3
4
5

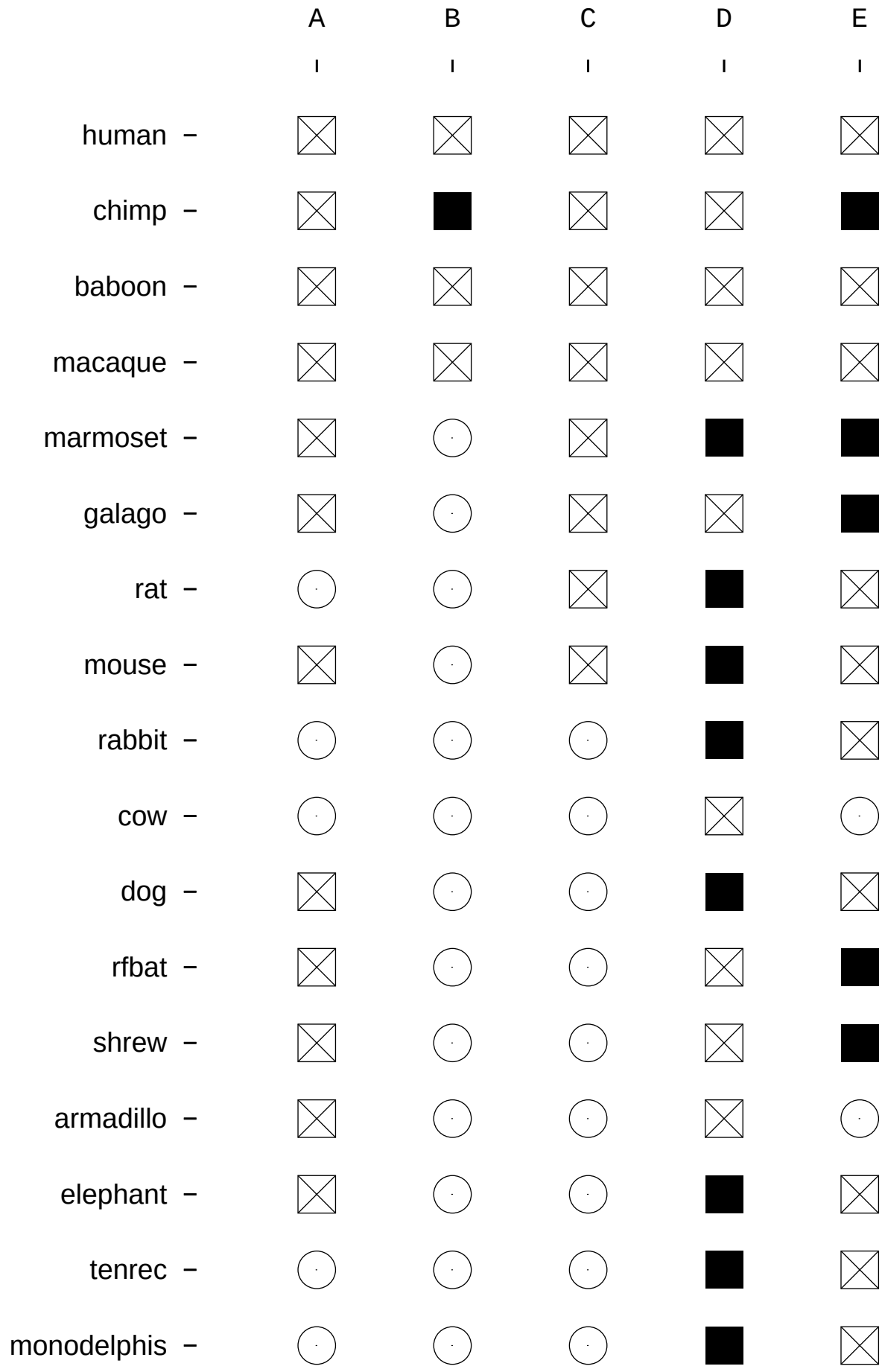
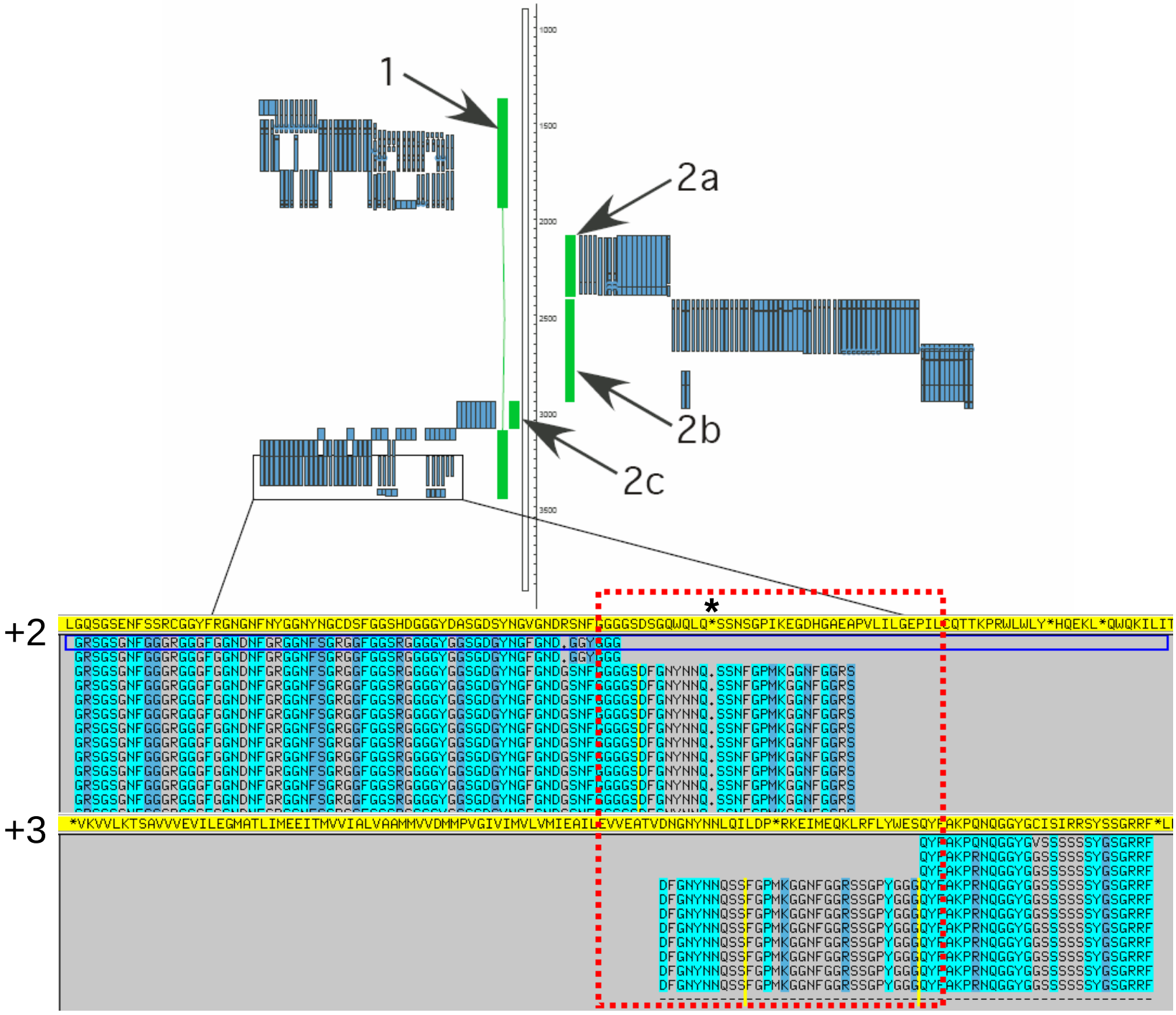
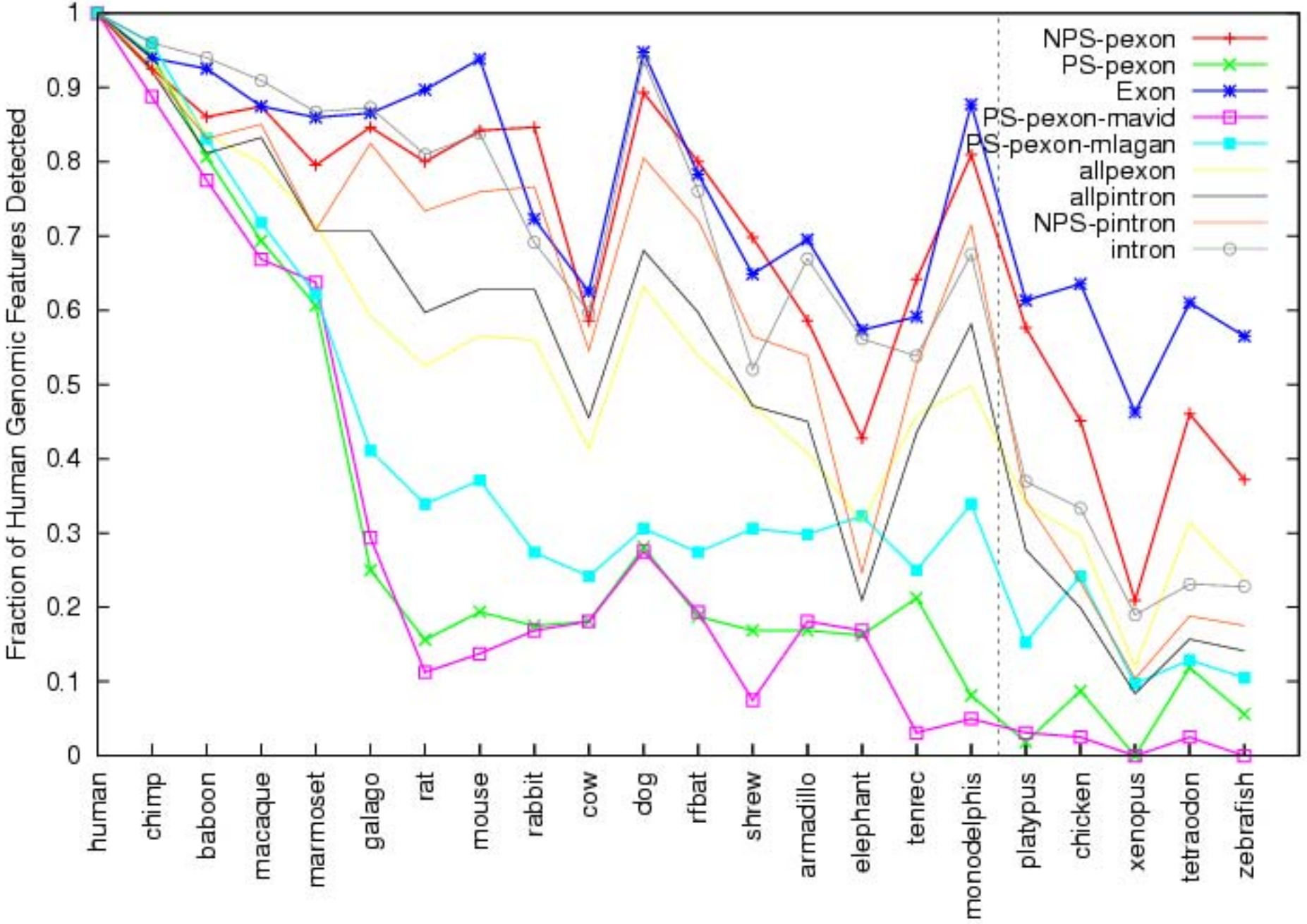
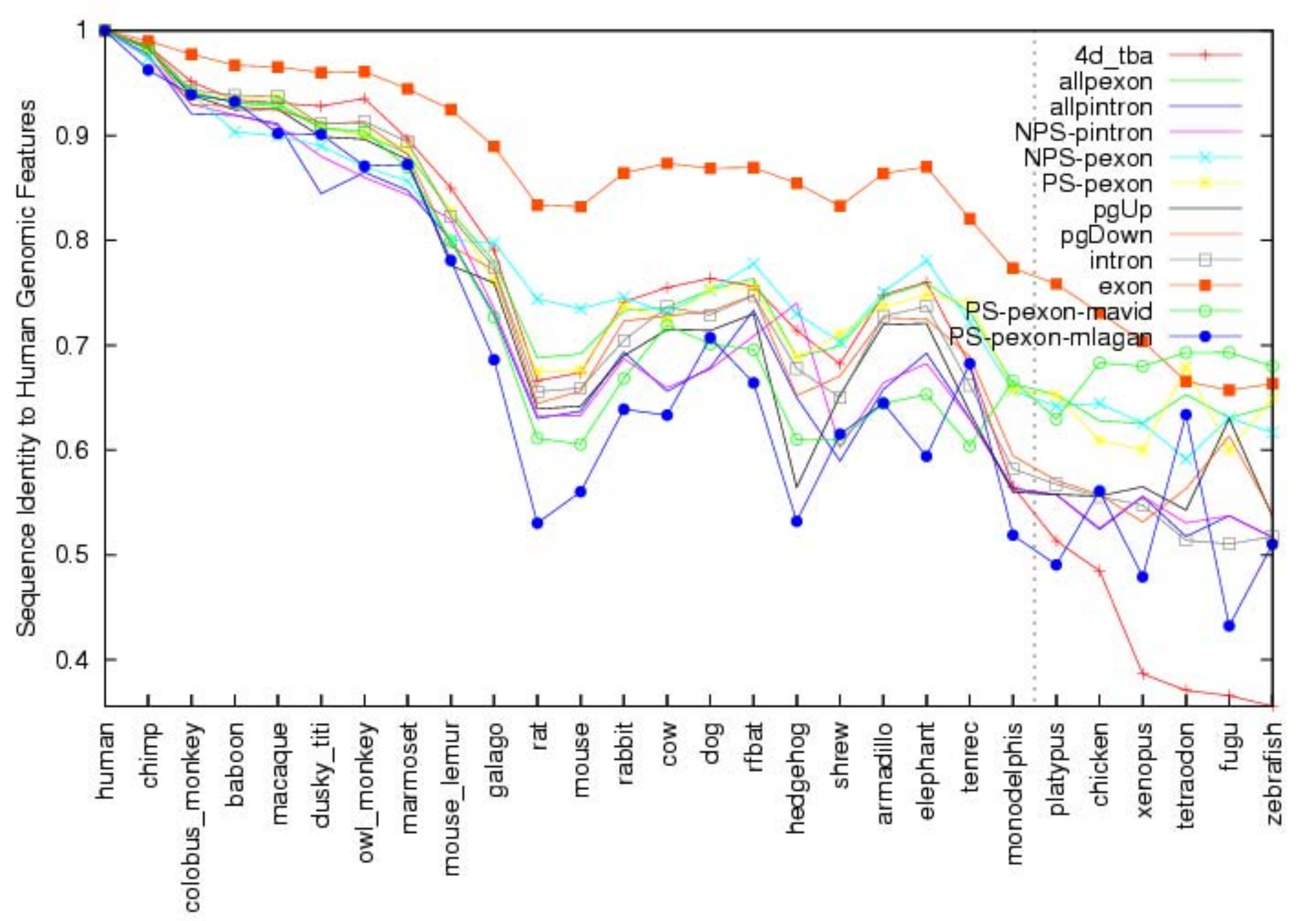

A
B
C
D
E
human
chimp
baboon
macaque
marmoset
galago
rat
mouse
rabbit
cow
dog
rfbat
shrew
armadillo
elephant
tenrec
monodelphis

Fig9




Fraction of Human Genomic Features Detected
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
NPS-pexon
PS-pexon
Exon
PS-pexon-mavid
PS-pexon-mlagan
allpexon
allpintron
NPS-pintron
intron
human
chimp
baboon
macaque
marmoset
galago
rat
mouse
rabbit
cow
dog
rfbat
shrew
armadillo
elephant
tenrec
monodelphis
platypus
chicken
xenopus
tetraodon
zebrafish


Sequence Identity to Human Genomic Features
1
0.9
0.8
0.7
0.6
0.5
0.4
4d_tba
allpexon
allpintron
NPS-pintron
NPS-pexon
PS-pexon
pgUp
pgDown
intron
exon
PS-pexon-mavid
PS-pexon-mlagan
human
chimp
colobus_monkey
baboon
macaque
dusky_titi
owl_monkey
marmoset
mouse_lemur
galago
rat
mouse
rabbit
cow
dog
rfbat
hedgehog
shrew
armadillo
elephant
tenrec
monodelphis
platypus
chicken
xenopus
tetraodon
fugu
zebrafish


chr20:
33572000
33572500
33573000
33573500
33574000
33574500
33575000
Encyclopedia of DNA Elements (ENCODE) Regions
ENCODE Pseudogene Predictions - All ENCODE Regions
Consensus Pseudogenes
Havana-Gencode Pseudogenes
Yale Pseudogenes
UCSC Retrogenes
UCSC Pseudogenes
GIS Pseudogenes

Pseudogenes in the ENCODE Regions: Consensus Annotation, Analysis of Transcription and Evolution

Deyou Zheng [1], Adam Frankish [2], Robert Baertsch [3], Philipp Kapranov [4], Alexandre Reymond [5, 6], Siew Woh Choo [7], Yontao Lu [3], France Denoeud [8], Stylianos E Antonarakis [6], Michael Snyder [9], Yijun Ruan[7], Chia-Lin Wei [7], Thomas R. Gingeras [4], Roderic Guigo [8, 10], Jennifer Harrow [2], and Mark B. Gerstein [1, 11, 12, *]

Running title: pseudogenes in the ENCODE regions

Addresses:
[1] Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA.
[2] Wellcome Trust Sanger Institute; Wellcome Trust Genome Campus; Hinxton, Cambridgeshire, CB10 1HH, UK.
[3] Department of Biomolecular Engineering; University of California, Santa Cruz; 1156 High Street, Santa Cruz, CA 95064, USA.
[4] Affymetrix, Inc.; Santa Clara, CA, 92024, USA.
[5] Center for Integrative Genomics; University of Lausanne; Genopode building; 1015 Lausanne, Switzerland.
[6] Department of Genetic Medicine and Development; University of Geneva Medical School; 1 rue Michel-Servet, 1211 Geneva, Switzerland.
[7] Genome Institute of Singapore; 60 Biopolis Street; Singapore 138672, Singapore.
[8] Grup de Recerca en Informática Biomèdica; Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra. Passeig Marítim de la Barceloneta, 37-49, 08003, Barcelona, Catalonia, Spain.
[9] Molecular, Cellular & Developmental Biology Department; Yale University; New Haven, CT 06520, USA.
[10] Center for Genomic Regulation; Passeig Marítim de la Barceloneta, 37-49, 08003, Barcelona, Catalonia, Spain.
[11] Department of Computer Science, Yale University, 51 Prospect Street, New Haven, CT 06520, USA.
[12] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA.

Supplementary Materials

Description of the Five Individual Methods of Pseudogene Identification

Method I – PseudoPipe from Yale This is a computational pipeline specifically designed for annotating pseudogenes. The detailed algorithms and parameters have been described previously (Zhang et al. 2006; Zheng and Gerstein 2006).

Method II – HAVANA manual annotation This is a semi-automated method. It first sends all genomic sequences to an automated analysis pipeline for similarity searches and *ab initio* gene predictions. The searches are run on a computer farm and stored in an ENSEMBL MySQL database using the ENSEMBL analysis pipeline system (Searle et al. 2004). A pseudogene is annotated where the total length of the protein homology to the genomic sequence is >20% of the length of the parent protein or >100 aa in length, whichever is least (Harrow et al. 2006). For all annotated pseudogenes an active homologous gene (the parent) can be identified elsewhere in the genome. Where an open but truncated ORF is present other evidence is used (for example, a 3' genomic polyA tract) to allow classification as a pseudogene. Where a parent gene has only a single coding exon (e.g., olfactory receptors), a small 5' or 3' truncation to the CDS at the pseudogene locus (compared to other family members) is sufficient to confirm pseudogene status where the truncation is predicted to significantly affect secondary structure by the literature and/or expert community. Processed and non-processed pseudogenes are distinguished on the basis of structure and genomic context. Processed pseudogenes, which arise via retrotransposition, lose the intron-exon structure of the parent gene, often have an A-rich tract indicative of the insertion site at their 3' end and are flanked by a different genomic sequence to the parent gene. Non-processed pseudogenes, which arise via gene duplication, share both the intron-exon structure and flanking genomic sequence with the parent gene.

Method III – retroFinder by UCSC This method is specific for genes or pseudogenes (referred together as retrogenes) originating from retrotranspositions and does not attempt to identify non-processed pseudogenes that are created by a different evolutionary process. The method starts with an alignment of all human mRNAs to the genome using BLASTZ (Schwartz et al. 2003) and uses a set of biological features to assign a score representing the likelihood that a retrotransposition event has occurred at each locus. These features include 1) the number of introns removed from the retrogene, 2) breaks in synteny with mouse and dog relative to the size of the retrogene, 3) the number of exons in the parent gene, 4) the number of splice sites (or alignment breaks) in the same relative position in the mRNA, 5) coverage of repetitive elements, 6) percent identity and 7) coverage of the alignment and 8) length of the polyA tail inserted in the genome after the pseudogene. We define a weighted linear combination of the features trained on a known set of HAVANA pseudogenes. A heuristic threshold was determined based on those known pseudogenes and then used for selecting processed pseudogenes, (and thus retrogenes are not included here).

Method IV -- pseudoFinder by UCSC Like other homology-based pseudogene finding methods, this method first identified homologues of a given set of human reference genes by the HomoMap (homologous mapping method) method, which first maps the DNA sequence of a gene via Human Blastz Self Alignment results and then chains mapped segments (Kent et al. 2003; Schwartz et al. 2003). Each homologue was then compared with its reference gene to collect a set of features, such as sequence identity, Ka/Ks ratio, splicing site score, and number of premature stop codons. Instead of removing homologues overlapping reference genes, they were labeled as negative samples while homologues overlapping known pseudogenes were labeled as positive samples. These labeled samples were used to train Support Vector Machines (SVMs) to learn how to separate positive samples from negative samples. After that, the trained SVMs were used to identify pseudo homologues from all homologues. To get the final set of pseudogenes, a heuristic method removed weak pseudo homologues, which had few pseudogene-like features and supports from both gene and mRNA evidence, and added weak functional homologues, which had multiple pseudogene-like features and no support from gene or mRNA evidence.

Method V – PET based method by GIS The mRNA transcripts in HCT116 and MCF7 cells have been determined by paired-end diTag sequencing (Ng et al. 2005). These transcripts in turn are a good resource for identifying processed pseudogenes. To this end, PETs mapped to multiple locations were used to identify pseudogene locations. The genomic coordinates of the multiple mapped PETs were clustered into PET-based gene families based on the sequence homologies. A representative member (e.g., shortest genomic sequence) was selected from each family to search the whole genome using the program BLAT (Kent 2002) to identify putative pseudogenes and those without introns were classified as processed pseudogenes. As this method was targeted at one special subset of pseudogenes whose parent genes were transcribed in HCT116 and MCF7 cell lines, it detected many fewer pseudogenes than the previous four methods did.

Number of Pseudogenes in ENCODE Regions

Three ENCODE regions (ENr213, ENr312 and ENr321) out of 44 do not appear to contain any pseudogenes, and 31 regions have only processed pseudogenes. On the other hand, three regions, ENr112, ENr311 and ENr313, do not contain coding genes but have four, one, and one pseudogene, respectively. The apparent correlation of gene and pseudogene distribution is interesting as one might expect pseudogenes, at least the processed type, to be dispersed randomly in the human genome. Although local GC content could have some influence on the insertion of pseudogenes (Zhang et al. 2003), better accessibility of chromatin in gene-rich regions may be a more important deterministic factor.

Comparison between Transcribed and not Transcribed Pseudogenes

The mean sequence identity between transcribed pseudogenes and their parent genes is 71.4% for non-processed and 74.9% for processed pseudogenes, and the mean number of

disablements is 2.1 and 5.6. These figures are not significantly different from the corresponding statistics based on all ENCODE pseudogenes ($p>0.05$).

Current Annotation is Protein-based and has not Included Pseudogenes without a Known Protein Homolog

An important early decision in setting the frame of reference for this project was how best to define the boundaries of a pseudogene. The choice, to define a pseudogene on the basis of homology to the protein or the DNA/mRNA of the parent gene, was reflected in the different methods used by the various predictors. After discussion it was determined that protein homology would be used. Therefore, pseudogenes without a known protein homolog were not pursued in this study. Primary efforts were put into the accurate detection of pseudogenes with protein coding parent genes. The identification of mutations (e.g., frameshifts and nonsense mutations that would disable the CDS and confirm the pseudogenization of the element identified) is facilitated by a protein–centered approach. In addition, using protein evidence to define the pseudogene provided the opportunity for an additional quality control step in the annotation by allowing the validity of the parent gene and its CDS to be assessed.

Comparison of Different Pseudogene Annotation Methods

Although all five methods were designed with the intention of obtaining a complete collection of pseudogenes, they resulted in quite distinct lists of pseudogenes (Figure 1). Since each of them uses an independent operational definition of pseudogenes and adopts a different set of computational schemes and parameters, it is difficult to evaluate them by directly intersecting their results. Furthermore, without a "gold standard" pseudogene set defining the "truth", it is not very meaningful to estimate which method performs best. Here we discuss the pros and cons of each method in details.

HAVANA pseudogene annotation is protein based and has a very large manual component. It differs from the other four methods as it is not pseudogene-specific. Instead, pseudogenes are annotated as part of a region-by-region approach to identify all coding, non-coding and pseudogene loci. HAVANA manual annotation was used to make decisions about the validity of all other pseudogene predictions as it allows very detailed investigation, bringing numerous sources of evidence external to the initial prediction to bear -- e.g., literature reports, mRNA and examination of parent genes. Manual curation is highly specific (i.e., very few manually curated pseudogenes were rejected from the final consensus set), capable of unraveling complex cases that proved problematic to all the automated methods [e.g. the mitochondrial pseudogenes AC006326.2, .3, .4 and .5 in ENm001 (Figure 9)], and the most effective method of discriminating processed and non-processed pseudogenes. However, the initial HAVANA pseudogene set was smaller than the final consensus set, due mostly to a failure in the annotation of pseudogenes supported by shorter and weaker protein homologies that were not visible in the annotation interface. This is most likely to have been a function of the set-up of the initial genomic sequence analysis pipeline for proteins

that are calibrated for the detection of coding genes rather than specifically for the detection of pseudogenes.

The Yale pseudogene annotation pipeline, PseudoPipe, is also protein based and succeeded in identifying many pseudogenes supported by short and weak homologies. This was not unexpected as the pipeline is optimized for the detection of pseudogenes and pseudogene fragments. As a method without much manual intervention, the Yale pipeline depends on the reliability of the gene set that it uses to filter its predictions and the quality of the proteins set on which its predictions are based. From analysis of several pseudogenes that were annotated by this pipeline but not by HAVANA method, it became apparent that there were many dubious sequences in the protein sets used by the Yale pipeline. Such spurious proteins were almost all the result of automated CDS predictions from mRNA submitted to GenBank/EMBL. These mRNAs represented either dubious gene structures (e.g., single-exon genes with no cross species support), heavily repeat-masked single-exon genes with consequently dubious CDSs, genuine gene structures assigned an inappropriate CDS (e.g., in 3' UTR of a coding gene) or presenting a likely target for nonsense mediated mRNA decay. These problems are difficult to identify computationally and can only be resolved by manual intervention.

Unlike the previous methods retroFinder uses mRNA from coding genes to make pseudogene predictions. Since our final consensus approach is protein based, some pseudogenes from this method were excluded from the consensus list if they were identified solely by their sequence similarities to non-coding parts of an mRNA, such as 3' UTR. This contributes to the reduced specificity of retroFinder in the context of this study. This is also a function of both the methodology and the directional (3'-5') and often incomplete insertion of retrotransposed sequence that can result in some degree of 5' truncation and can lead to the complete loss of the CDS in the pseudogenic sequence. However, the corollary of this is the ability of this method to identify short protein matches at the C-terminus where the alignment of the 3'UTR sequence supports the alignment over the CDS region. The method is able to more accurately classify processed pseudogenes that are missed by other methods due to the number of orthogonal features used in the classifier. Although it uses mRNA alignments, this method also suffers from the contamination of the protein databases via mRNAs that are assigned unconvincing CDSs.

The pseudoFinder aligns the genomic sequence of the parent coding gene to identify pseudogenes. By doing this, this method is able to take the advantage of other existing genomic annotations related to the parental gene and pseudogene by incorporating them as features, such as number of exons, conservation of splicing sites, mRNA evidence, and insertion of repeat elements. These features help improve the method's sensitivity and accuracy. Furthermore, the set of prospective parent genes must be specified which provides the opportunity for users to reduce database contamination issues and the machine learning approach applied in the method gives a chance to correct misannotations in the parental gene set as well. Like retroFinder, this method can identify processed pseudogenes with short or no protein match but long 3' UTR sequence support. However, this method cannot find pseudogenes from gene families that are lost in human

but still functioning in other species and its accuracy depends on the availability and accuracy of other annotations. The majority of pseudogenes not identified by any other method were based on very weak but relatively long alignments.

The GIS-PET method is based on transcripts, using changes in the distance between ditags, which indicate the 5' and 3' ends of the parent coding gene. Because only transcripts from two cell lines were used for generating ditags, the set of genes with available ditags and used for pseudogene identification is much smaller than the datasets used by the other methods, and this is very likely to explain the smaller pseudogene set identified by this method. However, this method does appear to be highly specific, with all pseudogenes identified included in the final consensus set. Furthermore, the pipeline can be easily modified to identify potentially expressed pseudogene loci.

Additional supplementary data (including FigS1-3 and TableS1) are available at http://www.pseudogene.org/ENCODE/supplement/

Reference

Harrow, J., F. Denoeud, A. Frankish, A. Reymond, C. C.K., J. Chrast, J. Lagarde, J.G.R. Gilbert, R. Storey, D. Swarbreck., C. Ucla, T. Hubbard, S.E. Antonarakis, and R. Guigó. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol* **7:** S4.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Kent, W.J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100:** 11484-11489.

Ng, P., C.L. Wei, W.K. Sung, K.P. Chiu, L. Lipovich, C.C. Ang, S. Gupta, A. Shahab, A. Ridwan, C.H. Wong, E.T. Liu, and Y. Ruan. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* **2:** 105-111.

Schwartz, S., W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, and W. Miller. 2003. Human-mouse alignments with BLASTZ. *Genome Res* **13:** 103-107.

Searle, S.M., J. Gilbert, V. Iyer, and M. Clamp. 2004. The otter annotation system. *Genome Res* **14:** 963-970.

Zhang, Z., N. Carriero, D. Zheng, J. Karro, P.M. Harrison, and M. Gerstein. 2006. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*.

Zhang, Z., P.M. Harrison, Y. Liu, and M. Gerstein. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13:** 2541-2558.

Zheng, D. and M. Gerstein. 2006. A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biol* **7:** S13.