life

MDPI

Communication

Protein-Coding Genes in Euarchontoglires with Pseudogene Homologs in Humans

Lev I. Rubanov [1], Oleg A. Zverkov [1], Gregory A. Shilovsky [1,2,3], Alexandr V. Seliverstov [1] and Vassily A. Lyubetsky [1,*]

1 Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow 127051, Russia; rubanov@iitp.ru (L.I.R.); zverkov@iitp.ru (O.A.Z.); gregory_sh@list.ru (G.A.S.); slvstv@iitp.ru (A.V.S.)
2 Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow 119234, Russia
3 Faculty of Biology, Lomonosov Moscow State University, Moscow 119192, Russia
* Correspondence: lyubetsk@iitp.ru

Received: 14 July 2020; Accepted: 8 September 2020; Published: 10 September 2020

Abstract: An original bioinformatics technique is developed to identify the protein-coding genes in rodents, lagomorphs and nonhuman primates that are pseudogenized in humans. The method is based on per-gene verification of local synteny, similarity of exon-intronic structures and orthology in a set of genomes. It is applicable to any genome set, even with the number of genomes exceeding 100, and efficiently implemented using fast computer software. Only 50 evolutionary recent human pseudogenes were predicted. Their functional homologs in model species are often associated with the immune system or digestion and mainly express in the testes. According to current evidence, knockout of most of these genes leads to an abnormal phenotype. Some genes were pseudogenized or lost independently in human and nonhuman hominoids.

Keywords: recently pseudogenized genes; human pseudogenes; efficient software; independent pseudogenization in hominoids; Euarchontoglires group

1. Introduction

Pseudogenes were for long relegated to the "junk" portion of DNA, but nowadays they attract an increasing attention for bearing important biological functions [1,2], e.g., in gene expression regulation and development of human diseases. Association of pseudogenes with human disorders has been approached in many works. Among numerous examples are interfaces in cancer [3–6], type 2 diabetes [7], pulmonary fibrosis, adrenal hyperplasia, chronic pancreatitis, AIDS and others [1]. With many pseudogenes already known from human and model species [8], research continues towards an effective computer technique for large-scale prediction of pseudogenes with common protein-coding homologs in a wide set of species [9].

The challenge is to detect pseudogenes in a species of interest (in our case, humans) or a set of species based on ancestry patterns of protein-coding homologs in a query species set (here, the Euarchontoglires group).

2. Materials and Methods

We searched for the protein-coding genes in 3 reference species, namely, the mouse (*Mus musculus*), rat (*Rattus norvegicus*) and rabbit (*Oryctolagus cuniculus*), which are present in at least 4 of 5 nonhuman hominoids, at least 2 of 12 Old World monkeys and pseudogenized in humans. Figure 1 shows the phylogenetic tree indicating the lineages that were queried to discover such genes.

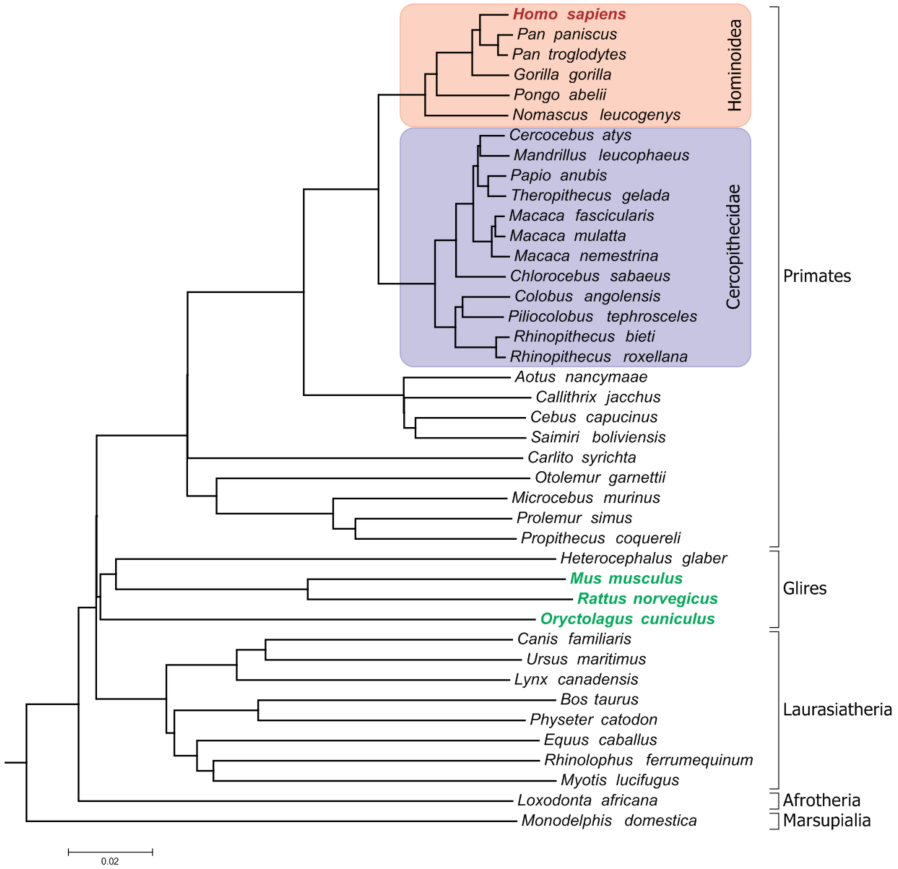

Figure 1. The phylogenetic tree of the species involved in the study. The reference species are shown in green. The lineages that were queried to discover the gene-loss/pseudogenization events are highlighted.

The solution was realized by verifying all the protein-coding genes *X* in the reference species against genes *Y*, including pseudogenes, in a given species set (here, 18 hominoids and cercopithecoids, including humans). The verification includes concurrent verifications of synteny in the neighborhoods of *X* and *Y* and gene orthology; pseudogene *Y* is verified for the same local synteny, homology and conservation of the ancestral exon-intronic structure. Such *X* and *Y* genes are referred to as consistent. The method allows for a reasonable parameter choice. For example, the neighborhood size is defined by a typical length of the topologically associated domain [10]. For nonhuman primates, the predicted consistency was verified in the neighborhood of 2 Mb under the constraining presence of at least two pairs of protein-coding one-to-one orthologs different from each other and the candidate gene; the neighborhood was 5 Mb and at least 4 (occasionally 3) such pairs were required for humans. Such orthologs are defined as witnesses to the candidate genes. The main stages of the method are efficiently implemented using fast computer software [11] that allows joint processing of over 100 complete genomes in a reasonable time. This program is an essential enhancement of our previous software once used in [12,13]. Genomic data were retrieved from Ensembl v99 [14], and data on the organ-specific expression of genes from Expression Atlas [15].

The results reported in this short communication were obtained with the following procedure (Figure 2). Initially, the program lossgainRSL was used to infer list *A* of the protein-coding genes in each reference species, which exist in the overwhelming majority of the nonhuman hominoids and many cercopithecoids but are lost or pseudogenized in human. The presence or absence of a gene is determined from orthology and local synteny: as stated above, at least two witnesses are required in the gene's neighborhood. The list is then pruned: for each gene *X*, its neighborhood of at most 1 Mb is verified on the genomic alignment with the human genome. The neighborhood is verified for the presence of at least four witness genes having one-to-one orthologs in the counterpart human region.

The gene *X*-aligned region in the human is then considered. Three cases are examined: (1) exonic gene *X* sequences hitting the noncoding regions in humans; (2) human regions are protein-coding but non-orthologous to gene *X*; and (3) human regions contain a pseudogene. Pseudogene types (processed, unprocessed, unitary, etc.) may be analyzed or chosen on a setting basis. In this work, all types are treated equally. In Case (1), gene *X* is assumed as lost in humans and removed from list *A*. In Case (2), the gene's organ-specific expression and exon-intronic structural patterns are compared between the reference species and human. If similarity is detected between any of the patterns, non-orthology is attributed to a low exonic similarity; the gene is not assumed as lost and also omitted from list *A*. In Case (3), a human pseudogene rather strongly aligns only to some exons of gene *X*, thus suggesting its pseudogenization; such genes are kept in list *A* and build a list of the reference species genes that are pseudogenized in human. Naturally implied parameters of this selection were estimated empirically.

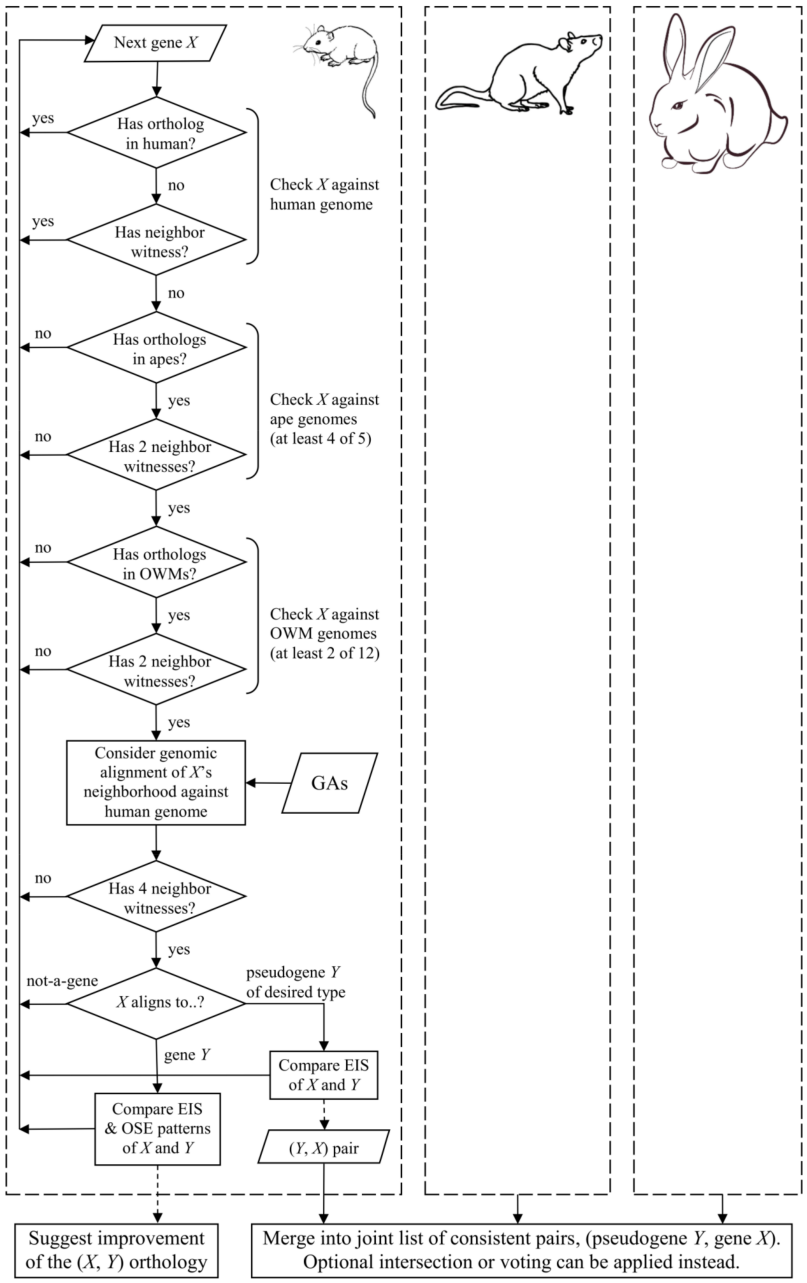

Figure 2. Schematic of the proposed method as applied to species and parameters involved in this study. OWM, Old World monkeys; GA, genomic alignment; EIS, exon-intronic structure; OSE, organ-specific expression. Dashed arrows signify side effects.

Pseudogene lists generated with this method are united across the three reference species. The lists can instead intersect or vote by reference species. With many reference species, voting is superior. Prior inference of list *A* is a key step for reducing the dimension of the reference species gene entries from tens of thousands to several hundreds, which speeds up the downstream analyses by two orders of magnitude and enables varying the computationally heavy parameters.

3. Results

3.1. Consistent Pseudogene Gene Pairs Identified

Our technique isolated only 50 pseudogenes in humans: 42 consistent with mouse genes, 42 of slightly various content with the rat and 38 with the rabbit (Table 1). An expanded version of the table is provided in Table S1, where the human pseudogenes are described in Columns A–H, their coding consistencies in the mouse in Columns I–R, in the rat in Columns S–AB and in the rabbit in Columns AC–AK. Columns Q and R contain the species numbers of the nonhuman hominoids and Old World monkeys, respectively—found to have a mouse ortholog in the current row under the imposed conditions (including number of witnesses). The numbers for the rat and rabbit are provided in Columns AA–AB and AJ–AK, analogously.

Table 1. Human pseudogenes and their consistencies in the mouse, rat and rabbit.

Human Pseudogene	Consistent Mouse Gene(s)	Consistent Rat Gene(s)	Consistent Rabbit Gene
A2MP1	*A2ml1*	*A2ml1*	ENSOCUG00000000313
ABCC13			ENSOCUG00000012802
AC063977.5	*4931406B18Rik*	*LOC690483*	
ADAM20P1	*Gm4787*		ENSOCUG00000008563
ADAM5	*Adam5*	*Adam5*	ENSOCUG00000005475
AHSA2P	*Ahsa2*	*Ahsa2*	ENSOCUG00000017817
AL160191.3		*Adam4*	
AL589987.1		*RGD1560171*	ENSOCUG00000012140
ALOX12P2	*Alox12e*	*Alox12e*	
C4BPAP1	*Zp3r*	*Zp3r*	ENSOCUG00000023993
CCDC162P	*Ccdc162*		ENSOCUG00000006035
CCDC92B	*Ccdc92b*	*Ccdc92b*	ENSOCUG00000027623
CLCA3P	*Clca3a1, Clca3a2, Clca3b*	*Clca5, Clca4l*	ENSOCUG00000003548
CMAHP	*Cmah*	*Cmahp*	ENSOCUG00000004430
CRYGFP			ENSOCUG00000010967
CST13P	*Cst13*	*Cst13*	ENSOCUG00000026263
CTF2P	*Ctf2*	*Ctf2*	ENSOCUG00000027802
CYMP	*Cym*	*Cym*	ENSOCUG00000026023
CYP2G1P	*Cyp2g1*	*Cyp2g1*	ENSOCUG00000005745
CYP2G2P	*Cyp2g1*	*Cyp2g1*	ENSOCUG00000005745
DPY19L2P1	*Dpy19l2*	*Dpy19l2*	ENSOCUG00000009890
FBXL21P	*Fbxl21*	*Fbxl21*	
FER1L4	*Fer1l4*	*Fer1l4*	ENSOCUG00000012872
FMO6P	*Fmo6*	*Fmo6*	ENSOCUG00000005169
GUCY1B2	*Gucy1b2*	*Gucy1b2*	
H1-9	*Hils1*	*Hils1*	ENSOCUG00000012304
H2BU2P	*H2bu2*	*Hist3h2ba*	
HTR5BP	*Htr5b*	*Htr5b*	ENSOCUG00000016884
KLRA1P	*Klra2*	*Ly49i7*	ENSOCUG00000007301
KRT43P			ENSOCUG00000011173
LINC00643			ENSOCUG00000026292
METTL21EP	*Mettl21e*	*Mettl21cl1*	ENSOCUG00000012425
OFCC1		*AABR07027339.1*	ENSOCUG00000016635
OR10AA1P	*Olfr433*		ENSOCUG00000036381

Table 1. *Cont.*

Human Pseudogene	Consistent Mouse Gene(s)	Consistent Rat Gene(s)	Consistent Rabbit Gene
OR13C6P	*Olfr159*	*Olr836*	
OR2S1P	*Olfr155*	*Olr840*	
PCDHB17P	*Pcdhb14*	*Pcdhb14*	
PRORSD1P	*Prorsd1*	*Prorsd1*	ENSOCUG00000011673
PRSS40A		*Prss40*	
PRSS40B	*Prss40*		ENSOCUG00000006394
PRSS46P	*Prss46*	*Prss46*	ENSOCUG00000039130
SKINT1L	*Skint1*	*Skint1*	
SLC22A20P	*Slc22a20*	*Slc22a20*	ENSOCUG00000003132
TAAR4P	*Taar4*	*Taar4*	ENSOCUG00000024372
TDH	*Tdh*	*Tdh*	ENSOCUG00000011294
TMED11P	*Tmed11*	*Tmed11*	
TMEM198B	*Tmem198b*	*Tmem198b*	ENSOCUG00000027671
TMEM30CP	*Tmem30c*	*Tmem30c*	ENSOCUG00000027470
UOX	*Uox*	*Uox*	ENSOCUG00000027397
ZNF271P	*Zfp35*	*Zfp35*	ENSOCUG00000029705

Reference species were found to possess 43 (mouse), 42 (rat) and 37 (rabbit) protein-coding genes pseudogenized in humans and some other hominoids. Among those, 10 genes in the mouse (*Zp3r, Prss40, Prss46, Tmem30c, Dpy19l2, Adam5, Gm4787, Hils1, 4931406B18Rik,* and *Cst13*) and 13 in the rat (*Clca4l, Zp3r, Prss40, Prss46, Tmem30c, Taar4, Dpy19l2, Adam5, Adam4, Hils1, Cyp2g1, LOC690483*, and *Cst13*) are expressed almost exclusively in the testes. Another 9 murine and 15 rat genes are highly expressed in the testes, but also active in other organs (*H2bu2, Prorsd1, Ahsa2, Htr5b, Pcdhb14, Ccdc162, Tdh, Slc22a20, Zfp35, Clca5*, *Cym, Hist3h2ba, Prorsd1, Ahsa2, Tmed11, Fbxl21, Pcdhb14, Tdh, Olr836, Slc22a20, Tmem198b, Mettl21cl1, Zfp35* and *RGD1560171,* respectively).

Among the human pseudogenes in Table S1, only one is annotated as processed, whilst the others represent transcribed unitary pseudogenes (21), transcribed unprocessed pseudogenes (17), unprocessed pseudogenes (8) and unitary pseudogenes (3). Accordingly, a total of 50 pseudogenes are predicted, with one of them listed three times in Table S1.

Many of the identified genes are associated with the immune system, such is the *Zfp35* gene, whose knockout in mice results in an abnormal T-helper 2 cell differentiation, increased airway responsiveness, increased circulating interleukin-13 level, increased circulating interleukin-4 level, increased circulating interleukin-5 level and increased eosinophil cell number [16]. This gene is expressed mainly in the testes, but also maintains high levels in the thymus, spleen and brain. In humans it is represented by the unitary transcribed pseudogene *ZNF271P*. Some detected protein-coding genes are involved in digestion, e.g., murine chymosin-encoding *Cym*.

Available evidence shows that knockout of many identified human pseudogene homologs in model species leads to severe disorders. For example, mice with a human-like *Cmah* deficiency have hyperactive macrophages, T cells, B cells, etc. [17]. Knockout of the heat shock protein ATPase 2 activator gene *Ahsa2* leads to abnormal cornea morphology and decreased total retina thickness [18], whilst urate oxidase-encoding *Uox* knockout causes abnormal kidney morphology and uremia [19]. Knocking out gene *Tmem198b* of the pituitary gland transmembrane protein 198b provokes reduced sensorimotor gating [20]. In contrast, two mouse genes, *4931406B18Rik* and *Htr5b*, exhibit no abnormal phenotype in ablation [21].

Of interest is the identified F-box and leucine-rich repeating protein 21-encoding gene *Fbxl21*. In mice and rats it is highly expressed within the suprachiasmatic nuclei, the site of the master clock, where it displays marked circadian oscillations apparently driven by members of the PAR-bZIP family [22]. Its knockout is associated with a shortened circadian behavioral period through ubiquitination and stabilization of cryptochromes [23], limb grasping and decreased grip strength [18]. In humans, it is consistent with the pseudogene *FBXL21P* transcribed in the brain, kidneys and

prostate, and the pseudogene ENSGGOG00000014124 in the gorilla. In the gibbon, chimp, bonobo and orangutan its coding homolog is preserved in the same syntenic context. Pseudogenization of this circadian rhythm regulator may presumably be associated with increased longevity, as well as emerging neoteny in humans [24]. Gui et al. [25] studied the impact of SNPs in *FBXL21* on the success of a kidney transplantation.

The identified *Ctf2, Cyp2g1, Fmo6, Olfr155, Olfr159, Olfr433* and *Taar4* genes are highly expressed in the vomeronasal organ or olfactory epithelium, which come in concordance with its reduction and degraded olfaction in humans [24].

Homologs of the rat gene *RGD1560171* survived in all five of the non-human hominoids studied and many short-living mammals except mice (pseudogene *Gm715*). This protein-coding gene is pseudogenized or lost in species with a relatively high longevity, including humans (*AL589987.1*), naked mole rats and elephants. The gene *Ofcc1* in the mouse, *AABR07027339.1* in the rat and ENSOCUG00000016635 in the rabbit are consistent with the human pseudogene *OFCC1* and belong to same family with a gene potentially causal for orofacial cleft in humans [26]. However, ablation of this gene had no effects in head development in mice.

Each human pseudogene is consistent with exactly one gene in mouse, except for *CLCA3P*, consistent with three syntenically linked paralogs in mice and two in rats. Each human pseudogene is consistent with exactly one protein-coding gene in the rabbit.

The human pseudogene homologs of 16 murine protein-coding genes are presumably functional in the five non-human hominoids, whilst other murine genes survived in only four species. The same ratio is observed with other reference species. Human pseudogenes presumably retain functionality in many Old World monkeys and other placental mammals (Tables S2–S4).

3.2. Consistent Genes of the Reference Species and Their Counterparts in Other Species

Predictions in nonhuman primates and other placental mammals against the mouse, rat and rabbit are given in Tables S2–S4, respectively. The species column contains gene IDs if a gene is present in the species. Mouse genes were found on average in 4.4 nonhuman hominoids (17 genes per 5 species) and 9.6 Old World monkeys; rat genes—in 4.4 nonhuman hominoids (17 genes per 5 species) and 9.4 Old World monkeys; and rabbit genes—in 4.4 nonhuman hominoids (15 genes per 5 species) and 9.5 Old World monkeys.

The mouse was found to have 43 consistencies with 42 human pseudogenes. Consistency is "one-to-one", with the exception of the three mouse genes consistent with the same pseudogene *CLCA3P* and the mouse gene *Cyp2g1*, consistent with two neighboring human pseudogenes, *CYP2G1P* and *CYP2G2P*. Consistency is supported by local alignment of corresponding genes accounting for the exon-intronic structure and genomic alignment with at least 4 pairs of orthologous witnesses in a neighborhood of a specified size (3 pairs in the sole case), with the pair numbers normally being much greater. Table S5 describes the witnesses; pseudogenes are shadowed green and the syntenic blocks separated by horizontal lines. The nearest protein-coding genes flanking the candidate genes, where possible, were chosen as witnesses within the specified neighborhoods.

The rat was predicted to have 42 genes consistent with 42 human pseudogenes. Consistency is "one-to-one", with the exception of two genes in the rat consistent with the same human pseudogene *CLCA3P* and the rat gene *Cyp2g1*, consistent with two syntenic human pseudogenes, *CYP2G1P* and *CYP2G2P*. The exceptions thus mimic those in the mouse. The consistency is supported by the local alignment of the corresponding genes, accounting for the exon-intronic structure and genomic alignment with at least 4 pairs of orthologous witnesses in a neighborhood of a specified size (3 pairs in two cases), with the pair numbers normally being much greater. Table S6 describes the witnesses; pseudogenes are shadowed green and the syntenic blocks separated by horizontal lines.

Rabbit was found to have 37 genes consistent with 38 human pseudogenes. Consistency is "one-to-one", with the exception of the rabbit gene ENSOCUG00000005745, consistent with same syntenic human pseudogenes, *CYP2G1P* and *CYP2G2P*. Consistency is supported by the local alignment

of the corresponding genes accounting for the exon-intronic structure and genomic alignment with at least 4 pairs of orthologous witnesses in a neighborhood of a specified size (3 pairs in the sole case), with the pair numbers normally being much greater. Table S7 describes the witnesses; pseudogenes are shadowed green and the syntenic blocks separated by horizontal lines.

In most cases, the detected pseudogenes are confined strictly within a syntenic region, with the following 10 outliers found at the boundary: *A2MP1, AL589987.1, CST13P, CYP2G1P, CYP2G2P, H2BU2P, KLRA1P, METTL21EP, OFCC1* and *TMED11P*; see Tables S5–S7.

3.3. Human Pseudogenes Independently Pseudogenized or Lost in Exactly One Nonhuman Hominoid

In contrast to humans, a few pseudogenes are known from nonhuman primates: 71% (human), 2% (gorilla), 3% (bonobo), 2% (chimpanzee), 5% (orangutan) and 62% (mouse) of the total protein-coding genes. The lack of some inferred genes in exactly one nonhuman hominoid may indicate their evolution into pseudogenes that escaped detection.

It follows from Tables S1–S4 that the pseudogenes found in humans presumably all retain functionality in the common chimpanzee and bonobo—all except *CCDC92B, H1-9* and *SKINT1L*; in the gibbon—all except *ADAM20P1, AL160191.3, CST13P* and *TDH*; in gorilla—all except *A2MP1, ADAM5, CYP2G1P, CYP2G2P, GUCY1B2, FBXL21P* and *LINC00643*; and in the orangutan—all except for 17 genes. These 31 cases of human pseudogenes that lost function in other hominoids are provided in Table S8 (shadowed green in the first column) along with the relevant witnesses. Such gene groups are separated by horizontal lines. Pseudogenes are consistent with the reference species genes specified in Table S1 that lack in exactly one nonhuman hominoid (bonobo, gorilla, gibbon or orangutan). Each group, starting from Column I, contains consistent nonhuman hominoid genes. For every pseudogene (except two), the genomic alignment contains a region without a consistent coding gene (marked "no gene", except for three cases in gorilla and three in orangutan). In two cases in the gorilla, the pseudogenes are consistent with known pseudogenes, and in one case—to a coding gene with no orthology in the reference species, thus introducing conflict in the consistency with the reference and human. In the orangutan, one instance lacks genomic alignment (shadowed blue), while in the other two cases a pseudogene is consistent with three coding genes and a coding gene with a pseudogene (shadowed green), respectively. In the rest of the cases, 3 in bonobo, 4 in gorilla and gibbon, 14 in orangutan, a pseudogene-consistent coding gene is not inferred in the nonhuman hominoids, although the sequence comparison sometimes suggests its possible pseudogenization. The description fields contain the gene attributes.

4. Discussion

More than a half (28 of 52) of the consistencies in Table S1 span the mouse, rat and rabbit, with the corresponding protein-coding genes being one-to-one orthologs with at least two orthologous witness pairs (normally more) within the 2 Mb neighborhood. The two exceptions are the already discussed pseudogene *CLCA3P*, consistent with three mouse genes, *Clca3a1*, *Clca3a2* and *Clca3b*; two rat genes, *Clca5* and *Clca4l*; and one rabbit gene, ENSOCUG00000003548. Here, the genes *Clca4l* and *Clca3b* are one-to-one orthologs; *Clca5*, *Clca3a1* and *Clca3a2*—one-to-many orthologs; and the rabbit gene is a one-to-many ortholog to the above listed three mouse genes and two rat genes. The second exception, killer cell lectin-like receptor A1 pseudogene *KLRA1P*, is consistent with the murine *Klra2*. However, its ortholog *Klra2* in the rat has no reliable genomic alignment with the *KLRA1P* neighborhood, which might be attributed to the available rat genome assembly. This pseudogene is therefore consistent with the immunoreceptor gene *Ly49i7* in the rat, non-orthologous to the murine *Klra2*. Pseudogene *KLRA1P* is also consistent with the gene ENSOCUG00000007301 in the rabbit, orthologous to 11 murine genes and 23 rat genes, including the aforementioned *Klra2* and *Ly49i7*.

Consider cases when a human pseudogene is consistent with the coding homologs in a subset of reference species; such pseudogenes are marked yellow in Table S1. The following 4 pseudogenes have a single consistency in the rabbit, with no orthology in the mouse and rat: *ABCC13, CRYGFP,*

KRT43P and *LINC00643*. Conversely, the following 8 pseudogenes have consistencies in the mouse and rat, with no orthology in the rabbit: *AC063977.5, ALOX12P2, FBXL21P, GUCY1B2, OR13C6P, OR2S1P, PCDHB17P* and *SKINT1L*. Pseudogene *OR2S1P* is consistent with the non-orthologous murine *Olfr155* and rat *Olr840* genes, both encoding olfactory receptor proteins, however. The rabbit has neither an *OR2S1P* consistency nor orthologs of *Olfr155* and *Olr840* in the rat and mouse, respectively.

The same is true with pseudogene *H2BU2P*: its consistencies in the mouse and rat are one-to-many orthologs of the gene ENSOCUG00000033562 in the rabbit. However, the rabbit gene is not consistent with the pseudogene *H2BU2P* but is a one-to-one ortholog of the active *H2BU1* gene in humans. A special case is pseudogene *TMED11P*, consistent with *Tmed11* in the mouse and rat. These are orthologs of ENSOCUG00000021044 in the rabbit that aligns to *TMED11P* but within the short, less than 70 Kb scaffold GL019412 in the available assembly, which allows no reliable consistency prediction. Recall that murine *Fbxl21* is expressed in many parts of the brain (listed in lowering order): pituitary gland, medulla oblongata, arcuate nucleus of hypothalamus, dorsal raphe nucleus, preoptic area, corpora quadrigemina, corpus striatum, diencephalon, cerebellum, spinal cord and hippocampal formation; its highest expression in the rat is also observed in the brain.

Let us describe cases where a pseudogene has no coding consistencies in the mouse or rat. Pseudogenes *PRSS40B* and *PRSS40A* neighbor in human Chromosome 2, the first being consistent with *Prss40* in the mouse and the second in the rat, both orthologs encoding serine protease 40 but hitting different pseudogenes in the genomic alignment. Pseudogene *PRSS40B* is consistent with the gene ENSOCUG00000006394 in the rabbit. A similar case is observed with pseudogenes *AL160191.3* and *ADAM20P1* co-located in human Chromosome 14 at about a 200 Kb distance, the first being consistent with the rat *Adam4* and the second to the murine *Gm4787*. These genes are one-to-many orthologs verified with the same human witnesses and both expressed mainly in the testes in these rodents. Pseudogene *ADAM20P1* in the rabbit is consistent with the gene ENSOCUG00000008563.

Pseudogene CCDC162P is consistent with the murine *Ccdc162* and rabbit ENSOCUG00000006035, with no orthologs in the rat. Alignment of this pseudogene's neighborhood to the rat genome does not support consistency to any coding gene in the rat, although the flanking genes are orthologous (relative to the pseudogene spot in humans and a putative gene spot in rats). Variance between the available mouse and rat genomes can hence be illustrated. Pseudogene *AL589987.1* is analogously consistent with *RGD1560171* in the rat and ENSOCUG00000012140 in the rabbit, with no murine orthology. The same is true with pseudogene *OFCC1*, consistent with *AABR07027339.1* in the rat and ENSOCUG00000016635 in the rabbit, with no murine orthology. Pseudogene *OR10AA1P* is consistent with the murine *Olfr433* and ENSOCUG00000036381 in the rabbit, with no orthologs in the rat.

Specially consider the following two rows in Table S1. Pseudogene *DPY19L2P1* finds consistency with *Dpy19l2* in the rat within one genome alignment locus verified with witnesses from Table S6. In another alignment spot, however, this rat gene has an orthologous hit to coding gene *DPY19L2* in human Chromosome 12. Unlike with the pseudogene in Chromosome 7, this ortholog is not verified with any witness in Chromosome 12. Therefore, Table S1 shows consistency to a pseudogene and not a protein-coding gene. The same is true for *Dpy19l2* in the mouse and *DPY19L2* in the rabbit.

Pseudogene *A2MP1* is located at the boundary of a syntenic locus (Tables S5 and S6), with witnesses found within a 200 Kb neighborhood in the mouse and rat, and within 6.5 Mb in humans, thus exceeding the value of 5 Mb suitable in all other cases. Worth noticing also is that the genomic alignment of the rat alpha-2-macroglobulin-like 1 gene *A2ml1* aligns the non-orthologous ovostatin 2-encoding human gene *AC024940.7* in the alternative region CHR_HSCHR12_4_CTG2 of Chromosome 12, but not pseudogene *A2MP1* co-located with *AC024940.7*. Genome alignment was likely affected in this case by close presence of the rat pregnancy-zone protein-encoding gene *Pzp* that is also highly affine to this pseudogene. A similar pattern is observed in the mouse and rabbit. Table S1 shows the human pseudogene *A2MP1*'s consistencies with the genes in all three reference species.

5. Conclusions

A method is proposed for the prediction of the protein-coding genes in a large set of species that are pseudogenized or lost in a species of interest. A fast computer implementation is freely available. The method allows processing of over 100 complete genomes in a reasonable time. It was applied to predict 50 pseudogenes in humans with functional homologs in rodents, lagomorphs and primates, associated with the immune system, digestion or expressed predominantly in the testes. These genes may represent important targets in studies of human longevity and neoteny, while their co-regulation network may be involved in the development of airway responsiveness and other autoimmune disorders.

Supplementary Materials: The following are available online at http://www.mdpi.com/2075-1729/10/9/192/s1, Table S1: Human pseudogenes and their consistencies in mouse, rat and rabbit. Table S2: Selected mouse genes and their presence in other mammals. Table S3: Selected rat genes and their presence in other mammals. Table S4: Selected rabbit genes and their presence in other mammals. Table S5: Syntenic neighborhoods of human pseudogenes and their consistent functional genes in mouse. Table S6: Syntenic neighborhoods of human pseudogenes and their consistent functional genes in rat. Table S7: Syntenic neighborhoods of human pseudogenes and their consistent functional genes in rabbit. Table S8: Human pseudogenes independently pseudogenized or lost in exactly one nonhuman hominoid.

Author Contributions: Conceptualization, A.V.S. and V.A.L.; methodology, V.A.L. and A.V.S.; software, L.I.R.; validation, O.A.Z., G.A.S., A.V.S., L.I.R., and V.A.L.; formal analysis, A.V.S. and L.I.R.; investigation, L.I.R., O.A.Z., A.V.S., and V.A.L.; resources, A.V.S. and L.I.R.; data curation, L.I.R.; writing—original draft preparation, A.V.S.; writing—review and editing, L.I.R. and V.A.L.; supervision, V.A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Russian Foundation for Basic Research, grant number 18-29-13037.

Acknowledgments: Computations were performed at the Joint Supercomputer Center of the Russian Academy of Sciences (JSCC RAS).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Cheetham, S.W.; Faulkner, G.J.; Dinger, M.E. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat. Rev. Genet.* **2020**, *21*, 191–201. [CrossRef] [PubMed]
2. Kovalenko, T.F.; Patrushev, L.I. Pseudogenes as functionally significant elements of the genome. *Biochemistry (Mosc)* **2018**, *83*, 1332–1349. [CrossRef] [PubMed]
3. Han, L.; Yuan, Y.; Zheng, S.; Yang, Y.; Li, J.; Edgerton, M.E.; Diao, L.; Xu, Y.; Verhaak, R.G.W.; Liang, H. The pan-cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat. Commun.* **2014**, *5*. [CrossRef] [PubMed]
4. Kalyana-Sundaram, S.; Kumar-Sinha, C.; Shankar, S.; Robinson, D.R.; Wu, Y.M.; Cao, X.; Asangani, I.A.; Kothari, V.; Presner, J.R.; Lonigro, R.J.; et al. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **2012**, *149*, 1622–1634. [CrossRef]
5. Karreth, F.A.; Reschke, M.; Ruocco, A.; Ng, C.; Chapuy, B.; Leopold, V.; Sjoberg, M.; Keane, T.M.; Verma, A.; Ala, U.; et al. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. *Cell* **2015**, *161*, 319–332. [CrossRef]
6. Poliseno, L.; Salmena, L.; Zhang, J.; Carver, B.; Haveman, W.J.; Pandolfi, P.P. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **2010**, *465*, 1033–1038. [CrossRef]
7. Chiefari, E.; Iiritano, S.; Paonessa, F.; Pera, I.L.; Arcidiacono, B.; Filocamo, M.; Foti, D.; Liebhaber, S.A.; Brunetti, A. Pseudogene-mediated posttranscriptional silencing of HMGA1 can result in insulin resistance and type 2 diabetes. *Nat. Commun.* **2010**, *1*, 40. [CrossRef]
8. Zhang, Z.D.; Frankish, A.; Hunt, T.; Harrow, J.; Gerstein, M. Identification and analysis of unitary pseudogenes: Historic and contemporary gene losses in humans and other primates. *Genome Biol.* **2010**, *11*, R26. [CrossRef]
9. Sharma, V.; Hecker, N.; Roscito, J.G.; Foerster, L.; Langer, B.E.; Hiller, M. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat. Commun.* **2018**, *9*, 1215. [CrossRef]

10. Razin, S.V.; Gavrilov, A.A. Structural-functional domains of the eukaryotic genome. *Biochemistry (Mosc)* **2018**, *83*, 302–312. [CrossRef]
11. LossgainRSL. A Program for Prediction of Gene Losses and Gains between Several Groups of Species. Available online: https://figshare.com/articles/software/lossgainRSL_a_program_for_prediction_of_gene_losses_and_gains_between_several_groups_of_species/9173243 (accessed on 15 July 2020). [CrossRef]
12. Korotkova, D.D.; Lyubetsky, V.A.; Ivanova, A.S.; Rubanov, L.I.; Seliverstov, A.V.; Zverkov, O.A.; Martynova, N.Y.; Nesterenko, A.M.; Tereshina, M.B.; Peshkin, L.; et al. Bioinformatics screening of genes specific for well-regenerating vertebrates reveals c-answer, a regulator of brain development and regeneration. *Cell Rep.* **2019**, *29*, 1027–1040. [CrossRef] [PubMed]
13. Rubanov, L.I.; Zaraisky, A.G.; Shilovsky, G.A.; Seliverstov, A.V.; Zverkov, O.A.; Lyubetsky, V.A. Screening for mouse genes lost in mammals with long lifespans. *BioData Min.* **2019**, *12*, 20. [CrossRef] [PubMed]
14. Yates, A.D.; Achuthan, P.; Akanni, W.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R.; et al. Ensembl 2020. *Nucleic Acids Res.* **2020**, *48*, D682–D688. [CrossRef] [PubMed]
15. Papatheodorou, I.; Fonseca, N.A.; Keays, M.; Tang, Y.A.; Barrera, E.; Bazant, W.; Burke, M.; Füllgrabe, A.; Fuentes, A.M.; George, N.; et al. Expression Atlas: Gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* **2018**, *46*, D246–D251. [CrossRef] [PubMed]
16. Kitajima, M.; Iwamura, C.; Miki-Hosokawa, T.; Shinoda, K.; Endo, Y.; Watanabe, Y.; Shinnakasu, R.; Hosokawa, H.; Hashimoto, K.; Motohashi, S.; et al. Enhanced Th2 cell differentiation and allergen-induced airway inflammation in Zfp35-deficient mice. *J. Immunol.* **2009**, *183*, 5388–5396. [CrossRef]
17. Kawanishi, K.; Dhar, C.; Do, R.; Varki, N.; Gordts, P.L.S.M.; Varki, A. Human species-specific loss of CMP-N-acetylneuraminic acid hydroxylase enhances atherosclerosis via intrinsic and extrinsic mechanisms. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 16036–16045. [CrossRef]
18. Dickinson, M.E.; Flenniken, A.M.; Ji, X.; Teboul, L.; Wong, M.D.; White, J.K.; Meehan, T.F.; Weninger, W.J.; Westerberg, H.; Adissu, H.; et al. High-throughput discovery of novel developmental phenotypes. *Nature* **2016**, *537*, 508–514. [CrossRef]
19. Wu, X.; Wakamiya, M.; Vaishnav, S.; Geske, R.; Montgomery, C., Jr.; Jones, P.; Bradley, A.; Caskey, C.T. Hyperuricemia and urate nephropathy in urate oxidase-deficient mice. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 742–746. [CrossRef]
20. Tang, T.; Li, L.; Tang, J.; Li, Y.; Lin, W.Y.; Martin, F.; Grant, D.; Solloway, M.; Parker, L.; Ye, W.; et al. A mouse knockout library for secreted and transmembrane proteins. *Nat. Biotechnol.* **2010**, *28*, 749–755. [CrossRef]
21. Miyata, H.; Castaneda, J.M.; Fujihara, Y.; Yu, Z.; Archambeault, D.R.; Isotani, A.; Kiyozumi, D.; Kriseman, M.L.; Mashiko, D.; Matsumura, T.; et al. Genome engineering uncovers 54 evolutionarily conserved and testis-enriched genes that are not required for male fertility in mice. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 7704–7710. [CrossRef]
22. Dardente, H.; Mendoza, J.; Fustin, J.M.; Challet, E.; Hazlerigg, D.G. Implication of the F-Box protein FBXL21 in circadian pacemaker function in mammals. *PLoS ONE* **2008**, *3*, e3530. [CrossRef] [PubMed]
23. Hirano, A.; Yumimoto, K.; Tsunematsu, R.; Matsumoto, M.; Oyama, M.; Kozuka-Hata, H.; Nakagawa, T.; Lanjakornsiripan, D.; Nakayama, K.I.; Fukada, Y. FBXL21 regulates oscillation of the circadian clock through ubiquitination and stabilization of cryptochromes. *Cell* **2013**, *152*, 1106–1118. [CrossRef] [PubMed]
24. Skulachev, V.P.; Holtze, S.; Vyssokikh, M.Y.; Bakeeva, L.E.; Skulachev, M.V.; Markov, A.V.; Hildebrandt, T.B.; Sadovnichii, V.A. Neoteny, prolongation of youth: From naked mole rats to "naked apes" (humans). *Physiol. Rev.* **2017**, *97*, 699–720. [CrossRef]
25. Gui, Z.; Li, W.; Fei, S.; Guo, M.; Chen, H.; Sun, L.; Han, Z.; Tao, J.; Ju, X.; Yang, H.; et al. Single nucleotide polymorphisms of ubiquitin-related genes were associated with allograft fibrosis of renal transplant fibrosis. *Ann. Transplant.* **2019**, *24*, 553–568. [CrossRef]
26. Ohnishi, T.; Yamada, K.; Watanabe, A.; Ohba, H.; Sakaguchi, T.; Honma, Y.; Iwayama, Y.; Toyota, T.; Maekawa, M.; Watanabe, K.; et al. Ablation of Mrds1/Ofcc1 induces hyper-γ-glutamyl transpeptidasemia without abnormal head development and schizophrenia-relevant behaviors in mice. *PLoS ONE* **2011**, *6*, e29499. [CrossRef] [PubMed]