FEBS Letters 580 (2006) 1472–1478

Functional and evolutionary analyses on expressed intronless genes in the mouse genome

Kishore Ramaji Sakharkar[a,1], Meena Kishore Sakharkar[b,1], Cymbeline T. Culiat[c], Vincent T.K. Chow[d], Shazib Pervaiz[a,e,*]

[a] National University Medical Institutes, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[b] Nanyang Centre for Supercomputing and Visualisation (MAE, MD), Affiliated staff BioInformatics Research Centre (SCE), Nanyang Technological University, Singapore
[c] Mammalian Genetics and Genomics Group Life Sciences, Oak Ridge National Laboratory Oak Ridge, TN 37831-6445, USA
[d] Department of Microbiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[e] Department of Physiology, Oncology Research Institute, NUS Graduate School for Integrative Sciences and Engineering, Yong Loo Lin School of Medicine, National University of Singapore, 2 Medical Drive, MD9, #02-05 Singapore 117597, Singapore

Received 9 December 2005; revised 18 January 2006; accepted 23 January 2006

Available online 31 January 2006

Edited by Takashi Gojobori

Abstract **Using computational approaches we have identified 2017 expressed intronless genes in the mouse genome. Evolutionary analysis reveals that 56 intronless genes are conserved among the three domains of life – bacteria, archea and eukaryotes. These highly conserved intronless genes were found to be involved in essential housekeeping functions. About 80% of expressed mouse intronless genes have orthologs in eukaryotic genomes only, and thus are specific to eukaryotic organisms. 608 of these genes have intronless human orthologs and 302 of these orthologs have a match in OMIM database. Investigation into these mouse genes will be important in generating mouse models for understanding human diseases.**

Keywords: Intronless; Single exonic genes; Housekeeping; Ortholog; Mouse models; Human disease

1. Introduction

The mouse genome has the distinction of being the mammalian genome runner-up, with a genome that is 14% smaller than the human genome [1]. Mouse has emerged as the most popular model system for studies on mammalian genetics and development for several reasons, including the fact that the genomes of mouse and other mammals such as humans are highly conserved. Most human genes have counterparts in the mouse genome [2]. The recent completion of the mouse genome draft sequence led to the surprising result that approximately 40% of the human genome's 3 billion base pairs could be aligned to the mouse genome at the nucleotide level [2]. Virtually all (99%) of the protein-coding genes in humans align with homologs in the mouse, and over 80% are clear 1:1 orthologs [3]. In most cases, the intron–exon structures are highly conserved [2,4]. This extensive conservation in protein-coding regions may be expected, because humans and mice share many metabolic pathways [3]. Due to the ease in manipulation and design of experimental studies in mouse, mouse has emerged as a model system for studying genetic basis of diseases.

Although the gene structures of most eukaryotic genes are characterized by having multiple exons separated by introns, intronless genes constitute a significant part of eukaryotic genomes [5]. Several isolated reports on intronless genes in human and mouse are available [6–8]. Takeda et al. reported 178 intronless non-chemosensory G-protein coupled receptors in humans [9]. Olfactory receptor genes are the largest superfamily of receptors in mouse and a substantial proportion of them have been reported to be intronless [10]. Recently, we generated a database (Genome SEGE – http://sege.ntu.edu.sg/wester/SEG) on intronless genes from nine completely sequenced eukaryotic genomes [11]. The presence of intronless genes in mouse provides an opportunity for comparative genomics and evolutionary studies across mouse and other groups of organisms, e.g., bacteria, archea and other eukaryotes. This comparative genomic analysis is relevant to the understanding of the evolution of gene architecture and genomes and in gene identification and validation of predicted gene structures. Furthermore, it is essential to the annotation of intronless genes in eukaryotes as the gene structure prediction algorithms employed in the analysis of eukaryotic genomes are biased towards accurately predicting intron-containing genes.

A gene could be annotated as single exon (i.e., intronless) CDS based on three main reasons: (a) the gene is truly 'intronless' and functional; (b) the gene is of retroposition origin (c) false positive predictions by gene finding algorithms. Despite the "handicap" that retrogenes rarely "meet" promoter elements at the locus of integration (resulting in the plethora of dead retropseudogenes), retroposition offers the opportunity to recombine the retrogene with a different promoter at the new gene locus [12], leading to a temporary or permanent "death on arrival" of gene transcription [13]. Thus, to perform an analysis on functional characteristics and annotations of intronless genes it is essential to remove non-functional intronless pseudogenes and get a "clean" list of protein-coding intronless genes. In this report, we propose a methodology

*Corresponding author. Fax: +65 67735461.
E-mail address: phssp@nus.edu.sg (S. Pervaiz).

[1] Both authors have contributed equally to the manuscript and thus should be treated as joint first authors.

0014-5793/$32.00
doi:10.1016/j.febslet.2006.01.070

to remove intronless genes that are putative pseudogenes based on a procedure by Harrison et al. and use full-length cDNA sequences to confirm the gene's structure, expression and function before performing analysis [14,15]. This approach makes the list of identified intronless genes more reliable and accurate. It also circumvents the greatest challenges in using EST databases to understand gene structure and expression. ProtFun provides insights into major functional categories of intronless genes in the mouse genome [16]. The extracted and validated list of intronless genes was also compared against the major taxonomic classes in GenBank to gain insights into their evolutionary conservation across the various lineages. 608 mouse intronless genes had intronless human orthologs at 80% sequence identity and 302 of the orthologs had a match in OMIM. The results shed light on the underlying evolutionary and genetic mechanisms that shaped the mouse genome and will be of importance in developing mouse models for human intronless genes involved in pathological states.

2. Materials and methods

Step 1: Dataset creation. GenBank format files for *Mus musculus* (mouse) (Build 32) genome were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/M_musculus/) (27th June, 2005), (National Centre for Biotechnology Information) to create a dataset on "intronless" genes based on the CDS FEATURE table annotation [11]. We identified 5258 putative intronless genes in *M. musculus* and identified 5085 non-pseudogenic intronless genes by using the methodology proposed by Harrison et al. (dataset-1) [14]. Here, we eliminated all identifiable processed pseudogenes by scanning for the polyadenylation signal (AATAAA) and polyadenylation tail. In this procedure, we define a sequence to represent a pseudogene if it contains a polyadenylation tail (>15A) within 1000 nucleotides from the stop codon with a preceding polyadenylation signal. 5076 of the 5085 intronless genes were found to code for proteins and CD-HIT was performed at 80% against 21 849 mouse multi-exonic protein-coding genes [17]. We identified 133 intronless genes in mouse genome with intron-containing paralogs (at an identity of $\geqslant 80\%$ along 90% of the sequence length).

Step 2: Identification of expressed intronless genes. The mouse full-length cDNA sequences were downloaded from the MGC (mammalian gene collection) at ftp://ftp1.nci.nih.gov/pub/MGC/. The MGC contains full-length open reading frame for mammalian genomes (including mouse) [10]. Therefore, we compared the mouse intronless genes against the MGC using BLASTN at a low *E*-value cutoff of 10^{-50} [18]. We identified 2150 intronless genes that had a hit in MGC. 133 of these were from the set that had a multi-exonic paralog identified in step 1. These were removed and 2017 genes were identified as expressed intronless genes without multi-exonic paralogs. This gene list represents the curated/validated dataset of expressed intronless genes and was subject to further comparative genomics analysis. It is noteworthy that >85% of the sequences showed more than 90% identity along 90% of the length of the gene under comparison.

Step 3: Expressed intronless genes shared across kingdoms of life. Similarities of all the validated gene entries (2017) to genes from the three major domains of life (archea, bacteria, eukaryotes) were determined by using Blast-link (Blink). Blink provides pre-computed sequence alignments for BLAST hits against a given sequence. Blink displays the number of hits for the query sequence to major taxonomic groups at the protein level (archea, bacteria, eukaryotes). A note on the presence or absence of hits to the specified taxonomic group(s) was performed and tabulated (data available online).

Step 4: Eukaryotic homologs. Homologene is a system for automated detection of homologs among the annotated genes of 18 completely sequenced eukaryotic genomes, including mouse. Homologene data was downloaded from NCBI (Build 41) and 970 clusters containing mouse intronless genes were identified. The homologous genes in humans, rats, invertebrates, plants, fungi were identified and plotted.

Step 5: Protein function classification. In order to gain insights into the protein functions of the mouse intronless genes ProtFun classification was employed. The ProtFun produces ab initio predictions of protein function directly from sequence. Prediction for each of the 2017 intronless proteins using ProtFun was performed [16].

Step 6: Human/mouse orthologs. In order to compile a list of mouse intronless genes that have intronless counterparts in humans and are involved in genetic disorders in humans, we performed a CD-HIT (at 80%) between intronless mouse proteins and human proteomes. Approximately 47% (942/2017) of mouse intronless genes have an ortholog in the human genome and 608 of these orthologs are intronless in human and 302 of these have a hit in OMIM.

All the data generated is available at: http://www.cellfate.org/mouse/intronless/.

3. Results and discussion

We initially identified 5085 non-pseudogenic intronless genes in the mouse genome. A total of 2017 genes (~40%) were later confirmed to be expressed and intronless in gene architecture by matching against full-length cDNA data using Blast (*e*-value = 10^{-50}) and not having intron-containing paralogs. Although the MGC database does not currently represent the full set of transcripts from the mouse genome, it is clear that there is a significant proportion of functional and expressed intronless genes (~8%) in the mouse genome. We performed functional and evolutionary characterization on these mouse intronless genes. The results of our observations are summarized below.

3.1. Distribution of mouse intronless genes

Chromosomal distributions of mouse intronless genes and the percentage of chromosomes represented by these genes are shown in Figs. 1 and 2. Generally, the genes appear to be randomly distributed on the mouse chromosomes. However, several clusters (clusters contain more than 3 genes) of intronless genes in certain chromosomes are evident in dataset-1, e.g., olfactory receptor, Tas2r receptor, small proline-rich protein, histone-2, interferon alpha, vomeronasal-1-receptor, chemokine receptors, keratin-associated protein and protocadherin beta (http://www.cellfate.org/mouse/intronless/). It appears that chromosomes 2 and 7 have relatively higher numbers of intronless genes, an observation further validated by the higher percentage of genome represented in intronless genes on these chromosomes. Chromosomes 2 and 7 also show maximum number of clusters for olfactory receptors, the largest superfamily of intronless receptors in mouse. The distribution of olfactory receptor gene clusters is available online. It should be noted that not all the genes in the clusters could be mapped to MGC data. However, literature survey confirmed the presence of these clusters on the respective chromosomes and their expansion by gene duplication (data not shown).

3.2. Taxonomic distribution, conservation and functional assignment

In order to explore the conservation of intronless genes across major kingdoms (Archea, Bacteria and Eukaryotes), a comparison of intronless genes against GenBank pre-computed data via *Blink* was performed. Each entry was clustered based on its homology to archeal (A), bacterial (B) or eukaryotic (E) kingdoms. This homology clustering provides important information on evolution of intronless genes (Fig. 3A).

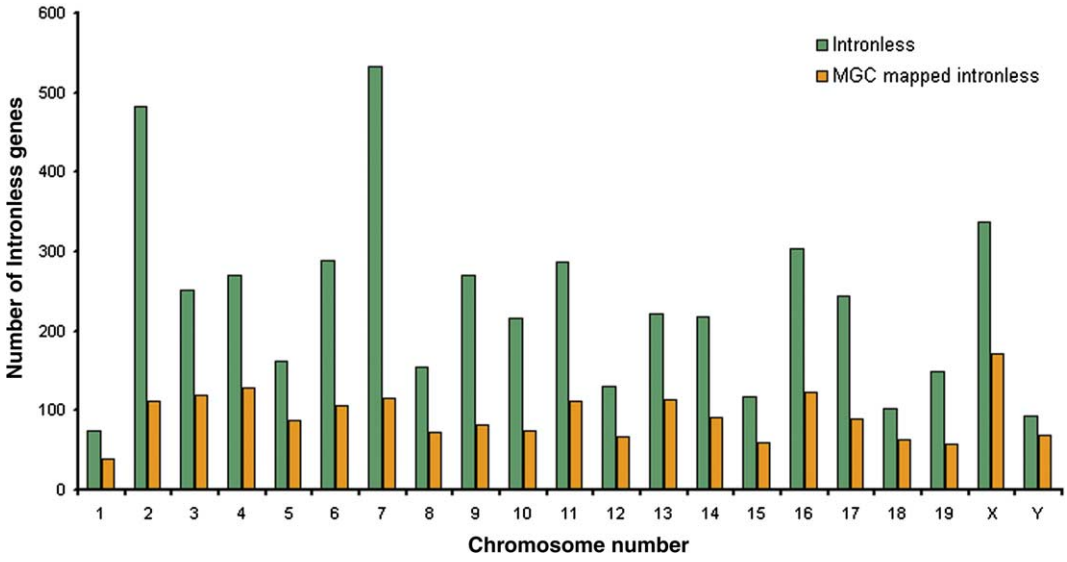

Fig. 1. Chromosome wise distribution of number of putative intronless genes and number of intronless genes that had a match in MGC.

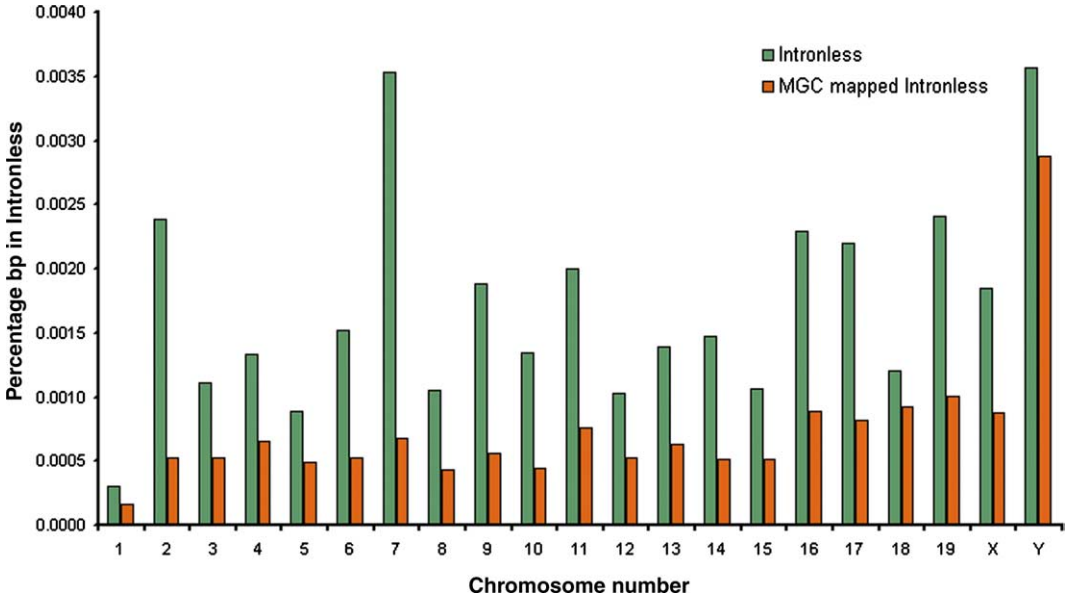

Fig. 2. Chromosome wise distribution of percentage of bp in putative intronless genes and percentage of bp in intronless that had a match in MGC.

It is interesting to note that out of the 2017 expressed intronless genes of the mouse genome, 85 are conserved across all the three kingdoms (ABE), while 56 of these are non-pseudogenic and expressed at least in mouse (~3%). These proteins are involved in basic cellular processes of life and thus could be termed as slowly evolving proteins (e.g., ribosomal proteins). Consistent with this idea, Wilson and colleagues proposed that essential genes evolve more slowly than non-essential genes [19]. Furthermore, Jordan et al. suggested that functionally important proteins are more evolutionarily conserved than less vital proteins [20]. Our data thus suggest that these 85(56) proteins are probably essential for survival for all kingdoms of life (ABE) (Fig. 3A). These proteins are well represented in the categories: protein translation; energy metabolism, amino acid biosynthesis, transport and binding; central intermediary metabolism, purines and pyramidines. It is interesting to see that 406 (225) intronless genes of mouse are shared with bacteria and eukaryotes (BE) (~18%). This category of proteins is well represented in the following functional classes: translation, energy metabolism, cell envelope, amino acid biosynthesis, regulatory functions and transport and binding. 60 (49) proteins share homology with archea and eukaryotes (AE) (~2%). The later group is well represented in the following functional categories: translation, energy metabolism, amino acid biosynthesis, growth factors and structural proteins. A large portion of expressed mouse intronless genes, 4036 (1616) (~80%), have homologs in eukaryotic genomes only (E) (Fig. 3a). The products of these genes are specific to eukaryotic organisms and include: histones, G-protein coupled receptors, channel proteins, protocadherins, etc. None of the

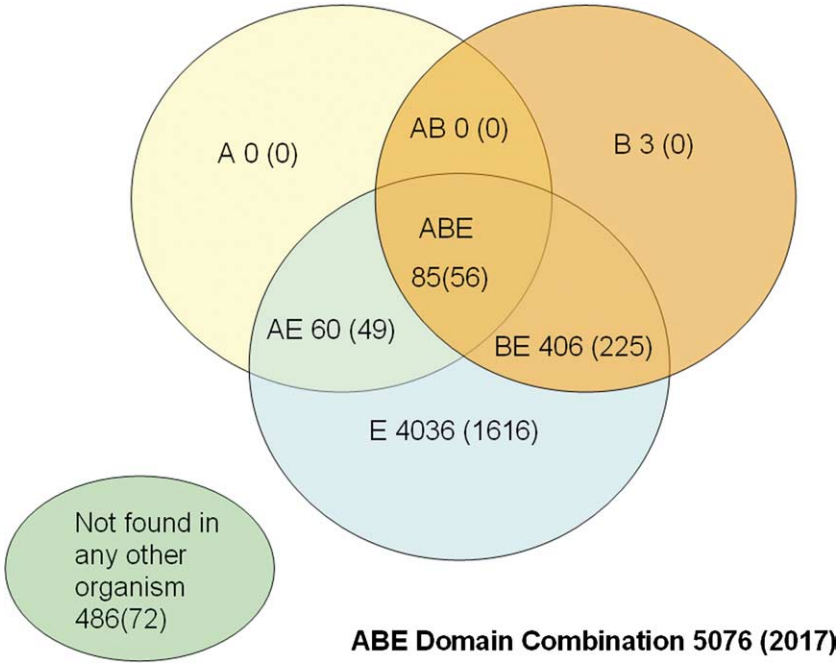

Fig. 3A. Venn diagram showing the classification of 5076 putative and 2017 expressed *M. musculus* intronless genes into three domains of life based on homology. Value in parenthesis indicates the number of genes sharing similarity. A, archea; B, bacteria and E, eukaryote.

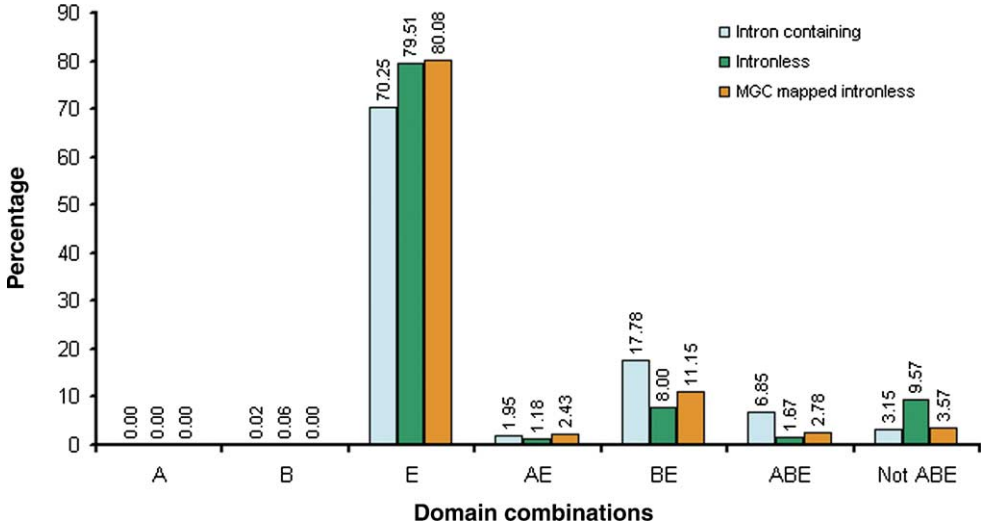

Fig. 3B. Distribution of putative intron-containing, putative intronless and MGC mapped intronless proteins into domain combinations A, archea; B, bacteria and E, eukaryote.

proteins from the confirmed expressed intronless genes were found to be homologous to bacteria (B), archea (A) or archea and bacteria (AB). 72 of the MGC mapped gene products do not share homology to any protein in NCBI under *Blink*. These proteins are mouse-specific. Most of these are hypothetical and have no function assigned as of today. These could be probable candidates for further explorations into mouse genome (Fig. 3A). We refer to them as ''ORFans''.

Analogous analysis on the intron-containing predicted gene complement showed similar homology ratio profiles across archea, bacteria and eukaryotes (Fig. 3B). Interestingly, the homology profiles for mouse genes are similar to the data on human homology profiles [21].

ProtFun assignment showed proteins that are growth factors, or involved in transcription, translation, immune response, transport, and structural proteins to be shared between bacteria and eukaryotes (BE); growth factors, structural proteins, immune response, and those involved in transcription, to be shared between archea and eukaryotes (AE) and growth factors, transport proteins, immune response, translation, transcription and structural proteins to be shared between all the three kingdoms of life (ABE) (Figs. 4 and 5).

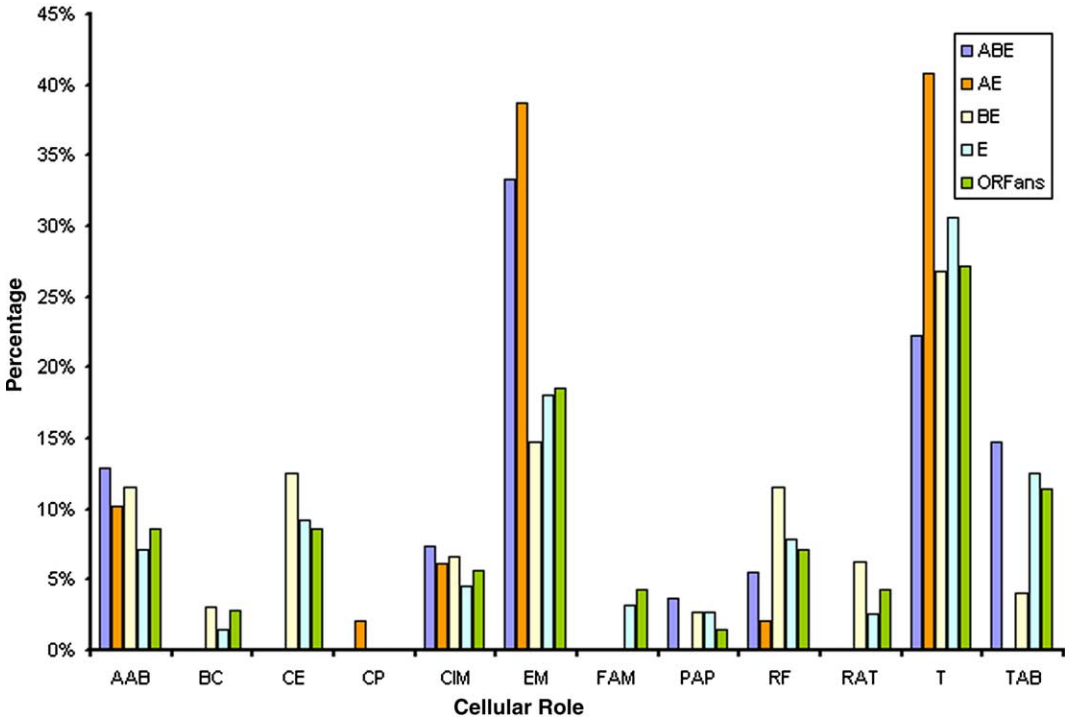

Fig. 4. Distribution of 12 protein cellular role categories across various domain combinations. Amino acid biosynthesis (AAB); biosynthesis of cofactors (BC); cell envelope (CE); cellular processes (CP); central intermediary metabolism (CIM); energy metabolism (EM); fatty acid metabolism (FAM); purines and pyrimidines (PAP); regulatory functions (RF); replication and transcription (RAT); translation (T); transport and binding (TAB).

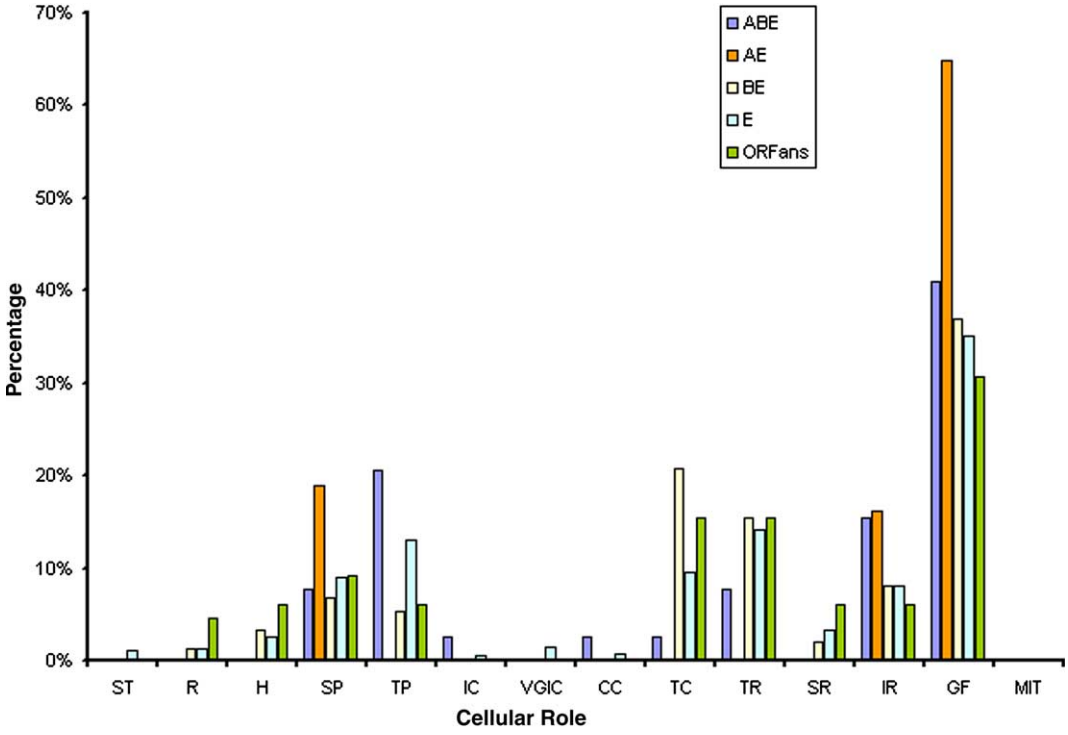

Fig. 5. Distribution of 13 Gene Ontology categories across various domain combinations. Signal transducer (ST); receptor (R); hormone (H); structural protein (SP); transporter (TP); ion channel (IC); voltage-gated ion channel (VGIC); cation channel (CC); transcription (TC); transcription regulation (TR); stress response (SR); immune response (IR); growth factor (GF); metal ion transport (MIT).

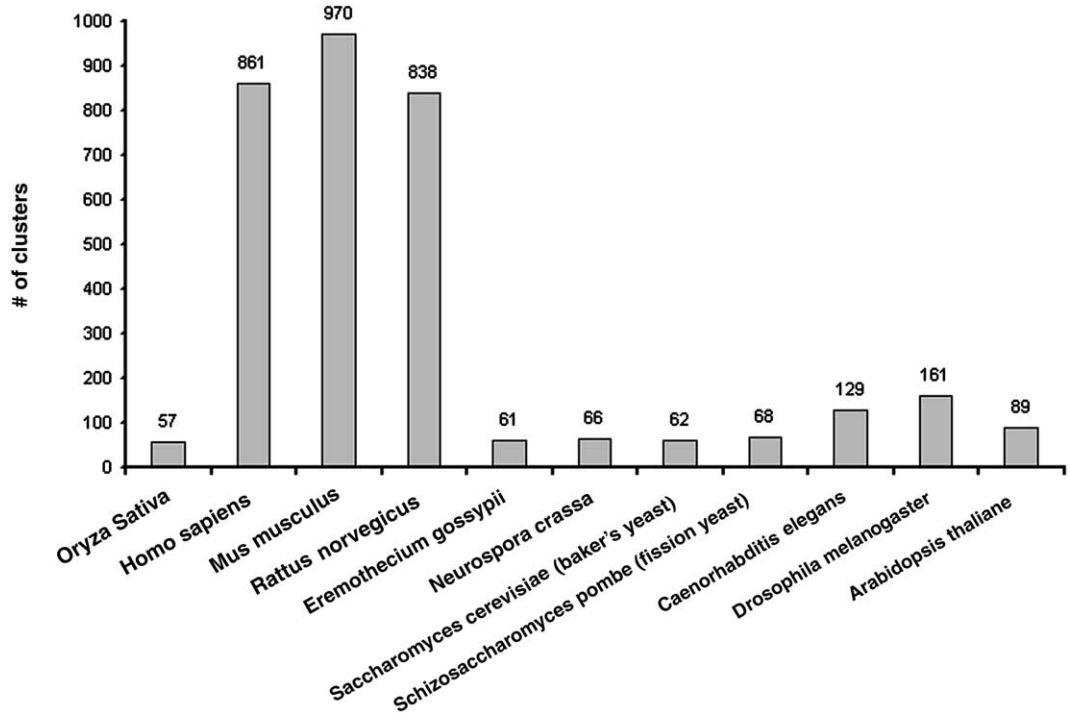

Fig. 6. Homologs of intronless genes from *M. musculus* with other eukaryotes in Homologene.

These data reinforce the notion that intronless genes are indeed excellent candidates for genes with housekeeping functions as they do not undergo alternative splicing that is reported as a major cause of many diseases [22].

These data thus provide clues towards understanding the evolutionary and organizational relationships among kingdoms. To gain further insight into the kinds of domains present in these proteins, we searched the proteins involved in fusion against the conserved domain database (CDD) [23]. This exercise revealed the presence of several important domains (homeodomain, serine/threonine protein kinases, actins, DNA binding domains, ubiquitin) in intronless gene products.

3.3. Homology search against eukaryotic genomes in Homologene

To further explore on the homology of the 2017 intronless genes of *M. musculus* with respect to the model eukaryotic genomes, we performed a match against Homologene database. It is interesting to note that 970 clusters incorporated intronless genes from the mouse genome. A distribution of the homologs within the clusters with respect to other genomes was plotted. As expected *M. musculus* shares maximum numbers of intronless homologs with *Homo sapiens* (861) and *Rattus norvegicus* (838). It is interesting to note that higher numbers of intronless genes are shared with invertebrate genomes (*Drosophila melanogaster* and *Caenorhabditis elegans*) than are shared with plant genomes (*Arabidopsis thaliana* and *Oryza sativa*) (161, 129 as against 89,57). The four fungal genomes display homology to <100 intronless genes. This is expected as plant and animal kingdom divergence precedes the evolution of metazoans. Since Homologene does not include all the eukaryotic genomes, it is possible that the remaining intronless genes are homologous to other genomes not present in the Homologene database. Nonetheless, the results hint at the dynamics of eukaryotic genome evolution and the differential selection of intronless genes across crown eukaryotic genomes represented in Homologene (Fig. 6).

3.4. Human/mouse intronless disease orthologs

Tissue-specific genes, genes related to asthma, obesity, osteoporosis and other diseases specific to vertebrates, cannot be addressed using lower organisms. The mouse has been exploited as a genetic system for more than 100 years and provides a rich resource of genetic mutations and inbred strains for biomedical research. Mice and humans diverged from a common ancestor about 65 million years ago, yet many salient aspects of mammalian physiology have not diverged significantly in these lineages during this time. The anatomical and physiological parallels between the two species are reflected in the comparable numbers of genes and similar gene architectures. For these reasons, and because mouse genetics offers many technological advances that are not available in other mammalian organisms, many have argued that the mouse will continue to play a key role in modeling studies of human genes and human diseases.

In order to estimate the number of mouse intronless genes that have orthologs in humans, we performed a match of mouse intronless proteins against human proteomes. It was interesting to see that 942 of the 2017 (~50%) mouse genes had orthologs in human proteome. Further analysis revealed that about 608 of these had intronless gene architectures in human. It is noteworthy that about 302 of the human intronless orthologs revealed matches in the OMIM database. These genes may be excellent targets for the creation of precise mouse models for studying human disease by targeted mutagenesis (data available online).

4. Conclusion

In this report evolutionary and functional classifications were performed for mouse intronless genes. Our data clearly suggest that there is a substantial proportion of mouse expressed intronless genes, and that these genes are involved in essential housekeeping biological functions. A considerable proportion of these genes reveal significant conservation across human and mouse genomes and these may facilitate understanding the pattern and extent to which these genomes have common genes and organization. These conserved genes may also serve as candidates for the construction of knock-out mouse models for human intronless genes to facilitate the understanding of related human diseases.

Acknowledgments: The authors acknowledge the help by Dr. Lars Juhl Jensen, EMBL, Germany, in performing protein function assignment by ProtFun. M.K.S. wishes to acknowledge funding support from A*STAR-BMRC, Singapore (Grant No. 03/1/22/19/242).

References

[1] Sands, A.T. (2003) The master mammal. Nat. Biotechnol. 21, 31–32.
[2] Waterston, R.H. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.
[3] Hardison, R.C. (2003) Comparative genomics. PLoS Biol. 1 (2), e58.
[4] Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L. and Dubchak, I. (2003) Strategies and tools for whole-genome alignments. Genome Res. 13, 73–80.
[5] Sakharkar, M.K., Chow, V.T., Chaturvedi, I., Mathura, V.S., Shapshak, P. and Kangueane, P. (2004) A report on single exon genes (SEG) in eukaryotes. Front Biosci. 9, 3262–3267.
[6] Sunahara, R.K., Niznik, H.B., Weiner, D.M., Stormann, T.M., Brann, M.R., Kennedy, J.L., Gelernter, J.E., Rozmahel, R., Yang, Y.L. and Israel, Y. (1999) Human dopamine D1 receptor encoded by an intronless gene on chromosome 5. Nature 347, 80–83.
[7] Gentles, A.J. and Karlin, S. (1999) Why are human G-protein coupled receptors predominantly intronless? Trends Genet. 15, 47–49.
[8] Rampazzo, A., Pivotto, F., Occhi, G., Tiso, N., Bortoluzzi, S., Rowen, L., Hood, L., Nava, A. and Danieli, G.A. (2000) Characterization of C14orf4, a novel intronless human gene containing a polyglutamine repeat, mapped to the ARVD1 critical region. Biochem. Biophys. Res. Commun. 278, 766–774.
[9] Takeda, S., Kadowaki, S., Haga, T., Takaesu, H. and Mitaku, S. (2002) Identification of G protein-coupled receptor genes from the human genome sequence. FEBS Lett. 520, 97–101.
[10] Zhang, X. and Firestein, S. (2002) The olfactory receptor gene superfamily of mouse. Nature 2, 124–133.
[11] Sakharkar, M.K. and Kangueane, P. (2004) Genome SEGE: a database for intronless genes in eukaryotic genomes. BMC Bioinformatics 5, 67.
[12] Brosius, J. (2005) Echoes from the past – are we still in an RNP world? Cytogenet. Genome Res. 110, 8–24.
[13] Brosius, J. (1991) Retroposons – seeds of evolution. Science 251, 753.
[14] Harrison, P.M., Hegyi, H., Balasubramanian, S., Luscombe, N.M., Bertone, P., Echols, N., Johnson, T. and Gerstein, M. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res. 12, 272–280.
[15] Gerhard, D.S., Wagner, L., Feingold, E.A., Shenmen, C.M., Grouse, L.H., Schuler, G., Klein, S.L., Old, S., Rasooly, R., Good, P., Guyer, M., Peck, A.M., Derge, J.G., Lipman, D. and Collins, F.S.The MGC Project Team (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res. 14, 2121–2127.
[16] Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H.H., Rapacki, K., Workman, C., Andersen, C.A., Knudsen, S., Krogh, A., Valencia, A. and Brunak, S. (2002) Prediction of human protein function from post-translational modifications and localization features. J. Mol. Biol. 319, 1257–1265.
[17] Li, W., Jaroszewski, L. and Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17, 282–283.
[18] Altschul, S.F., Madden, T.L., Schaer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
[19] Wilson, A.C., Carlson, S.S. and White, T.J. (1977) Biochemical evolution. Annu. Rev. Biochem. 46, 573–639.
[20] Jordan, I.K., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12, 962–968.
[21] Lander, E.S. et al. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860–921.
[22] Venables, J.P. (2004) Aberrant and alternative splicing in cancer. Cancer Res. 64, 7647–7654, Review.
[23] Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Marchler, G.H., Mullokandov, M., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Yamashita, R.A., Yin, J.J., Zhang, D. and Bryant, S.H. (2005) CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res. 33, D192–D196.