[Frontiers in Bioscience 10, 1382-1395 May 1, 2005]

COMPUTATIONAL PREDICTION OF SEG (SINGLE EXON GENE) FUNCTION IN HUMANS

Meena K. Sakharkar [1], Vincent T. K. Chow [2], Kingshuk Ghosh [2], Iti Chaturvedi [2], Pern Chern Lee [1], Sundara Perumal Bagavathi [1], Paul Shapshak [3], Subramanian Subbiah [4] and Pandjassarame Kangueane [1]

[1] School of Mechanical and Production Engineering, Nanyang Center for Supercomputing and Visualization, Nanyang Technological University, Singapore 639798, [2] Department of Microbiology, National University of Singapore, Singapore, [3] Department of Psychiatry and Behavioral Sciences, University of Miami Medical School, USA, [4] Department of Applied Physics, Stanford University, USA

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Materials and methods
- *3.1. Human SEG dataset*
- *3.2. Functional assignment*
- *3.3. PROSITE assignment*
- *3.4. PFAM assignment*
- *3.5. SUPERFAMILY assignment*
- *3.6. HK-SEG*
- *3.7. SEG cDNA*
- *3.8. Retrotransposed SEG*
- *3.9. SEG in GenBank description*

4. Results and discussion
- *4.1. Biologically meaningful signatures in SEG proteins*
- *4.2. Protein families in SEG proteins*
- *4.3. Known structural folds in SEG proteins*
- *4.4. HK-SEG*
- *4.5. SEG cDNA*
- *4.6. Retrotransposed SEG*
- *4.7. SEG characteristics in GenBank description*

5. Conclusions
6. Acknowledgements
7. References

1. ABSTRACT

Human genes are often interrupted by non-coding, intragenic sequences called introns. Hence, the gene sequence is divided into exons (coding segments) and introns (non-coding segments). Consequently, a majority of them are multi exon genes (MEG). However, a considerable amount of single exon genes (SEG) are present in the human genome (approximately 12%). This amount is sizeable and it is important to probe their molecular function and cellular role. Hence, we performed a genome wide functional assignment to 3750 SEG sequences using PFAM (protein family database), PROSITE (database of biologically meaningful signatures or motifs) and SUPERFAMILY (a library covering all proteins of known 3 dimensional structure). PFAM assigned 13% SEG to trans-membrane receptor genes of the G-protein coupled receptor (GPCR) family and showed that a majority of SEG proteins have DNA binding function. PROSITE identified 336 unique motif types in them and this accounts for 25% of all known patterns, with a majority having PHOSPHORYLATION and ACETYLATION signals. SUPERFAMILY assigned 33% SEG to the membrane *all alpha* (proteins containing alpha helix structural elements according to SCOP (structural classification of proteins) definition). Functional assignment of SEG proteins at multiple levels (sequence signals, sequence families, 3D structures) using PFAM, PROSITE and SUPERFAMILY is envisioned to suggest their selective and predominant molecular function in cellular systems. Their function as DNA binding, phosphorylating, acetylating and house-keeping agents is intriguing. The analysis also showed evidence of SEG expression and retro-transposition. However, this information is inadequate to draw concerted conclusion on the prevalent role played by these proteins in cellular biology. A complete understanding of SEG function will help to explore their role in cellular environment. The derived datasets from these analyses are available at http://sege.ntu.edu.sg/wester/intronless/human/

2. INTRODUCTION

Human genes are frequently interrupted by non-coding sequences called introns (1). Hence, they are often intron bearing and the gene structure is made of multiple exons (2-3). However, the human genome contains many single exon genes (SEG) that are not interrupted by introns (4-6). The CELERA (a genome company) human genome team identified 901 SEG with 298 instances of single-exon

gene to multi-exon gene (MEG) correspondence (7). The SEG to MEG correspondence discovered by the CELERA team provide insights to their possible origin by retrotransposition (8), which occurs by homologous recombination between the genomic copy of a gene and an intronless cDNA (9). The current update of the human genome contains about 12% SEG and this fraction is larger than an expected 5% (5-6). The presence of a substantial amount of SEG is interesting and their cellular role is puzzling. Hence, it is important to systematically document SEG molecular function to deduce their predominant role in cellular environment. However, molecular functions are known for only a handful of human SEG such as, D1 (dopamine) receptors (10), melanocortin 4-receptor (11), 5HT1D (serotonergic) receptors (12) and AR (β2. adrenergic receptor) (13). These molecules have G-protein coupled receptor (GPCR) function. The C14orf4 SEG is found to have house-keeping (HK) function (14). An analysis based on a dataset of GPCR sequences extracted from GenBank reported their prevalent occurrence as SEG (15). Nonetheless, the characterization of human SEG is limited and a comprehensive functional assignment of all SEG using specific biochemical, gene expression and gene knockout analyses are expensive, laborious and often inconclusive. Therefore, it is important to perform functional assignments to SEG proteins using alternative procedures that are driven by computational predictions. This will enable us to design roadmaps to study their collective role in cellular systems. We assigned predicted functions to SEG sequences using PFAM, PROSITE and SUPERFAMILY assignments. Here, we discuss SEG molecular function to understand their prime cellular role.

3. MATERIALS AND METHODS

3.1. Human SEG dataset

Human SEG sequences were obtained from the Genome SEGE database (5) created using a procedure (4) described in Figure 1. This procedure utilized CDS annotation in the FEATURE (GenBank formatted record) for the identification and extraction of SEG sequences from the human genome (5). The CDS annotation in the FEATURE contains several patterns (complete (direct or complimentary) or partial (direct or complimentary)) for representing SEG and these patterns are summarized in Table 1. Thus, we obtained 3750 SEG nucleotide sequences from the human genome. The human genome file does not contain protein translations. A protein translation file called 'protein.fa' (file containing all protein sequences in the human genome) containing protein sequences was downloaded. The 'protein.fa' file contains protein translation for 3656 SEG sequences. These SEG protein sequences formed the dataset for the current analysis and this dataset is thereafter, referred as Genome-SEG. Similarly, MEG protein sequences [24275] were obtained from the human genome for analysis using a procedure described elsewhere (2-3).

3.2. Functional assignment

The 3656 SEG sequences were subject to functional assignments (Figure 2) using three complementary computational procedures, such as [1] PFAM (16), [2] PROSITE (17), [3] SUPERFAMILY (18-19). The HK gene set (20) and GenBank descriptions from SEGE (4) were also used for additional inference. We used a wide variety of assignment tools to achieve best possible specificity and sequence coverage by detecting all known biologically meaningful signatures, protein families and structural folds in a concerted manner (Figures 3a, 3c, 3e). These results were then compared with a parallel analysis performed with the human MEG protein sequences (Figure 3b, 3d, 3f).

3.3. PROSITE assignment

PROSITE (release 18.32) is a database of protein motifs and it contains 1275 documented entries describing 1374 different patterns, rules and profiles (17). The human protein SEG data was subject to PROSITE motif identification and the analyses revealed 19218 motifs with 336 distinct types of signatures. This number accounts for nearly 25% of known patterns. The distribution is given in Figure 3a and the top 10 patterns are given in Table 2. The distribution is further compared to a corresponding distribution for MEG proteins (Figure 3b).

3.4. PFAM assignment

PFAM is a collection of hidden Markov models (HMM) and multiple sequence alignments that have been developed for the identification of functional regions within proteins (16). PFAM (version 14.0) contains alignments and models for 7459 protein families. PFAM models were used to assign molecular functions to the 3656 SEG protein sequences, using the HMMER^{TM} package version 2.3.2 (Oct 2003). The HMM model assigns protein family for each SEG protein. The PFAM distribution among human SEG proteins is shown in Figure 3c and the top 10 assignments are given in Table 4. The distribution is further compared to a corresponding distribution for MEG proteins (Figure 3d).

3.5. SUPERFAMILY assignment

SUPERFAMILY (release 1.63) is a library of HMM models based on SCOP and contains 1232 known structural super-families (18-19). The HMM based assignment tool provides structural assignments to protein sequences at the super-family level using known structural information. SUPERFAMILY was used to assign structural super-families and domains for human SEG sequences. This exercise produced 1702 assignments to 1659 sequences with 199 unique SCOP super-families. The top 10 assignments are given in Table 5 and the distribution is shown in Figure 3e. The distribution is further compared to a corresponding distribution for MEG proteins (Figure 3f).

3.6. HK-SEG

By definition, human housekeeping (HK) genes are constitutively expressed to maintain cellular function (20). A recent investigation, documented a list of HK genes in human (21). Therefore, we compared the human SEG list with the human HK list. The comparison identified 18 SEG with house-keeping function (Table 6). These genes are referred to thereafter as HK-SEG.

The 'CDS' line in the FEATURES should contain a continuous span of bases indicated by the number of the first and last bases in the range separated by two periods (e.g. 23..78).

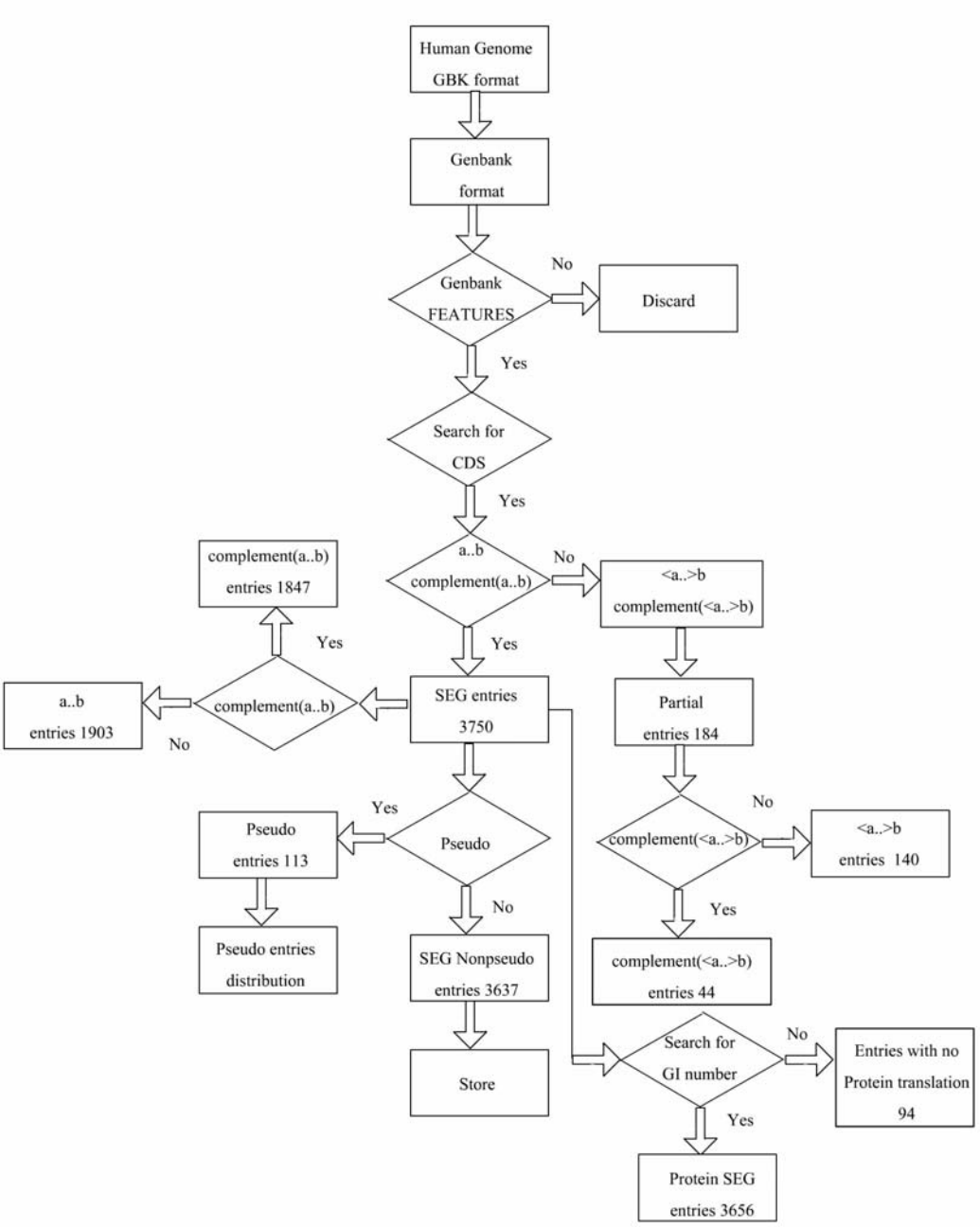

Figure 1. This flowchart describes the generation of a human SEG dataset. This procedure utilized CDS annotation in the FEATURE (GenBank formatted record) for the identification and extraction of SEG sequences from the human genome. The CDS annotation in the FEATURE contains several patterns (complete {direct or complimentary} or partial {direct or complimentary}) for representing SEG and these patterns are summarized in Table 1.

If symbols '<' or '>' are indicated at the end points of the range, the entry is partial because the range is beyond specified base number in such cases. When operators such as 'complement (location)' are used in the 'CDS' line, the feature is read as complementary to the location indicated and therefore the complementary strands are read from 5′ to 3′.

3.7. SEG cDNA

The human full length cDNA sequences were downloaded from the NIH mammalian gene collection [MGC] (22). The MGC program is a multi-institutional effort to identify and sequence cDNA clones containing a full-length open reading frame (FL-ORF). To date, the MGC has produced over 413 cDNA libraries derived from human tissue and sequenced and verified the complete FL-ORFs for a non-redundant set of 12,330 human genes. Therefore, we compared the human SEG nucleotide sequences with the full length cDNA sequences from MGC

Table 1. Different CDS (coding sequence) patterns used for SEG annotation in GenBank FEATURE format is shown in this table

Entries	Nature	CDS Patterns for SEG	# entries	Total entries
Complete	direct	a..b	1903	3750
	complement	complement(a..b)	1847	
Partial	direct	<a..>b	140	184
	complement	complement(<a..>b)	44	

Table 2. The top 10 PROSITE motifs in human SEG proteins are given

S.No.	Pattern/profile	Accession number	Motifs #	Consensus pattern	Function	Description	Remarks
1	PKC_PHOSPHO_SITE	PS00005	3312	[ST]-x-[RK]	signaling mechanism	Protein kinase C phosphorylation site	Signaling protein
2	MYRISTYL	PS00008	3157	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}	Protein myristylation; signal transduction/protein–protein interaction	N-myristoylation site	Mediator of protein and membrane trafficking
3	CK2_PHOSPHO_SITE	PS00006	3014	[ST]-x(2)-[DE]	Phosphorylation and regulation of proteins	Casein kinase II phosphorylation site	Multi-functional protein
4	ASN_GLYCOSYLATI-ON	PS00001	2041	N-{P}-[ST]-{P}	glycosylation of proteins	N-glycosylation site	Post -translational modification of proteins
5	CAMP_PHOSPHO_SITE	PS00004	1313	[RK](2)-x-[ST]	phosphorylation of proteins	cAMP- and cGMP-dependent protein kinase phosphorylation site	Post -translational modification of proteins
6	AMIDATION	PS00009	923	x-G-[RK]-[RK]	amidation of peptides	Amidation site	Protein modification
7	G_PROTEIN_RECEP_F1_1/2	PS50262	863	[GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)-[LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)-[LIVM]	signal transduction	G-protein coupled receptors family 1 profile	Signaling protein
8	TYR_PHOSPHO_SITE	PS00007	717	[RK]-x(2,3)-[DE]-x(2,3)-Y	phosphorylation of proteins	Tyrosine kinase phosphorylation site	Post translational modification of proteins
9	SULFATION	PS00003	525	(Glu or Asp) within two residues of tyrosine at -1; three acidic residues from -5 to +5; three hydrophobic residues from -5 to +5	regulation of protein-protein interactions	Tyrosine sulfation site	Post translational modification of proteins
10	NUCLEAR	PS00015	348	Two adjacent basic amino acids (Arg or Lys);a spacer region of any 10 residues; at least three basic residues (Arg or Lys)after the spacer region	nuclear translocation of proteins	Bipartite nuclear targeting sequence	Nuclear protein

Table 3. PROSITE motifs present exclusively in SEG proteins but absent in MEG proteins are given

S.No.	Pattern/profile	Description	Pattern	#Sequences
1	HISTONE_H2B	Histone H2B signature	[KR]-E-[LIVM]-[EQ]-T-x(2)-[KR]-x-[LIVM](2)-x-[PAG]-[DE]-L-x-[KR]-H-A- [LIVM]-[STA]-E-G	18
2	INVOLUCRIN	Involucrin signature	<M-S-[QH]-Q-x-T-[LV]-P-V-T-[LV]	1
3	ZF_BED	Zinc finger BED-type profile	-	6
4	ZN2_CY6_FUNGAL_1	Fungal Zn(2)-Cys(6) binuclear cluster domain signature	[GASTPV]-C-x(2)-C-[RKHSTACW]-x(2)-[RKHQ]-x(2)-C-x(5,12)-C-x(2)-C-x(6,8)- C	1
5	IF3	Initiation factor 3 signature	[KR]-[LIVM](2)-[DN]-[FY]-[GSTN]-[KR]-[LIVMFYS]-x-[FY]-[DEQTAHI]-x(2)- [KRQ]	1
6	CEREAL_TRYP_AMYL_INH	Cereal trypsin/alpha-amylase inhibitors family signature	C-x(4)-[SAGDV]-x(4)-[SPAL]-[LF]-x(2)-C-[RH]-x-[LIVMFYA](2)-x(3,4)-C	1
7	CRF	Corticotropin-releasing factor family signature	[PQA]-x-[LIVM]-S-[LIVM]-x(2)-[PST]-[LIVMF]-x-[LIVM]-L-R-x(2)-[LIVMW]	2
8	ASP_PROT_RETROV	Aspartyl protease, retroviral-type family profile	-	1
9	PRION_1	Prion protein signature 1	A-G-A-A-A-A-G-A-V-V-G-G-L-G-G-Y	2
10	PRION_2	Prion protein signature 2	E-x-[ED]-x-K-[LIVM](2)-x-[KR]-[LIVM](2)-x-[QE]-M-C-x(2)-Q-Y	2
11	THIONIN	Plant thionins signature	C-C-x(5)-R-x(2)-[FY]-x(2)-C	1
12	GRAM_POS_ANCHORING	Gram-positive cocci surface proteins LPxTG motif profile	-	1
13	HIPIP	High potential iron-sulfur proteins signature	C-x(6,9)-[LIVM]-x(3)-G-[YW]-C-x(2)-[FYW]	1
14	INTERFERON_A_B_D	Interferon alpha, beta and delta family signature	[FYH]-[FY]-x-[GNRC]-[LIVM]-x(2)-[FY]-L-x(7)-[CY]-A-W	17

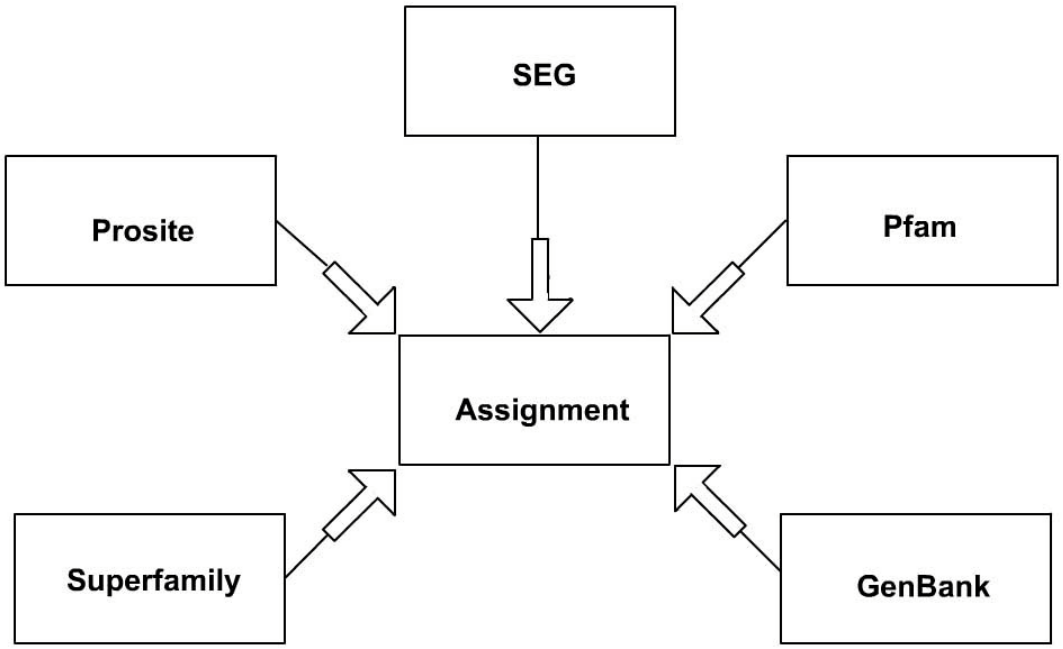

Figure 2. This illustration summarizes the different functional assignment models used in this analysis. The SEG sequences were subject to functional assignments using computational procedures, such as PFAM (protein family), PROSITE (database of biologically meaningful signatures or motifs), SUPERFAMILY (a library covering all proteins of known 3 dimensional structure) and GenBank descriptions from SEGE. We used a wide variety of assignment tools to achieve best possible specificity and sequence coverage by detecting all known biologically meaningful signatures, protein families and structural folds in a concerted manner.

Figure 3. The distribution of different PROSITE signatures (3a and 3b), PFAM families (3c and 3d) and SUPERFAMILY folds (3e and 3f) to SEG (3a, 3c, 3e) and MEG (3b, 3d, 3f) proteins is shown in this diagram. PROSITE: The distribution of most common motifs among SEG proteins is shown in Figure 3a. The distribution suggests that the protein kinase C (PKC) phospho site motif ([ST]-x-[RK]) constitutes the largest class (17%), followed by (16%) myristyl (G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}) and (16%) casein kinase II (CK2) ([ST]-x(2)-[DE]) phospho site motifs. This implies that human SEG proteins are rich in phosphorylation and acetylation signals. The result is compared with a distribution for MEG proteins, which showed that PKC, CK2 and myristyl motifs are predominant among them (Figure 3b). Thus, the distributions of the most prominent motif signals are similar in human SEG and MEG proteins. PFAM: The distribution of different protein families in SEG proteins is shown in Figure 3c. The data show that 13% SEG were assigned to 7-transmembrane receptors, 3% each to zinc finger - C2H2 type, leucine rich repeat and formin homology region. However, comparison of this distribution with that of MEG proteins suggests differences between SEG and MEG protein families (Figure 3d). Nonetheless, MEG proteins have a predominance of C2H2 type zinc fingers, protein kinase domain and immunoglobin domain. Moreover, the 7-transmembrane receptors family predominant in SEG is subdominant in MEG. SUPERFAMILY: The distribution of different structural family assignments to SEG is shown in Figure 3e. The results show that, (1) membrane-all-alpha (33%), (2) C2H2 and C2HC zinc fingers (6%), and (3) histone-fold (5%) with 568, 104, and 92 assignments, respectively are predominant among SEG proteins. Comparison with the MEG proteins show a different distribution with P loop containing nucleotide triphosphate hydrolases (6%), protein kinase - PK (5%) and C2H2 and C2HC zinc fingers (4%). This indicates that the SEG and MEG proteins have distinctly different structural folds. ABBREVIATIONS: TK - tyrosine kinase phosphorylation; PK - protein kinase; CK - casein kinase; GPCR - G protein coupled receptors; TS - tyrosine sulphation.

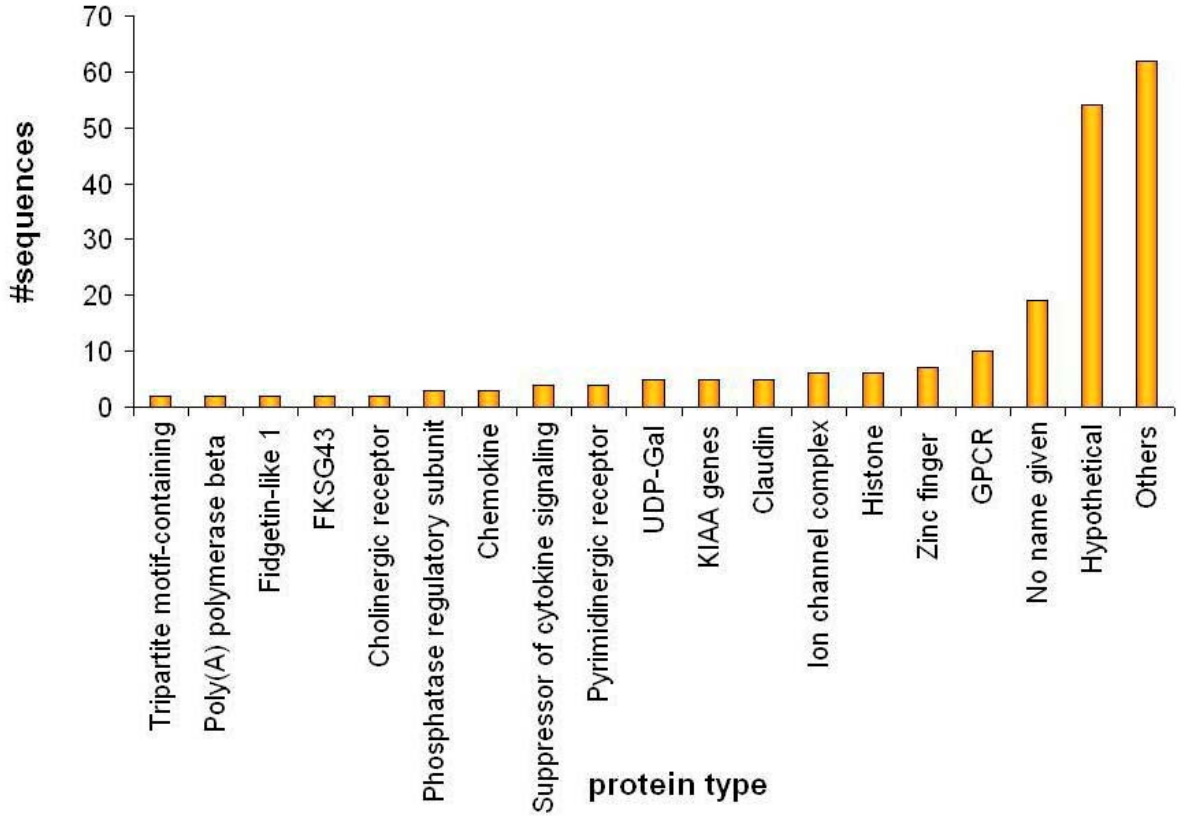

Figure 4. The characteristics of expressed SEG are shown in this figure. This distribution is obtained by comparing SEG with a dataset of full length cDNA sequences from MGC (Mammalian Gene Collection). It shows that a majority of expressed SEG are histones, GPCR, zinc finger, ion channel complexes. These SEG are expressed and it is thus, likely to capture their mRNA by RT-PCR (reverse transcriptase – polymerase chain reaction).

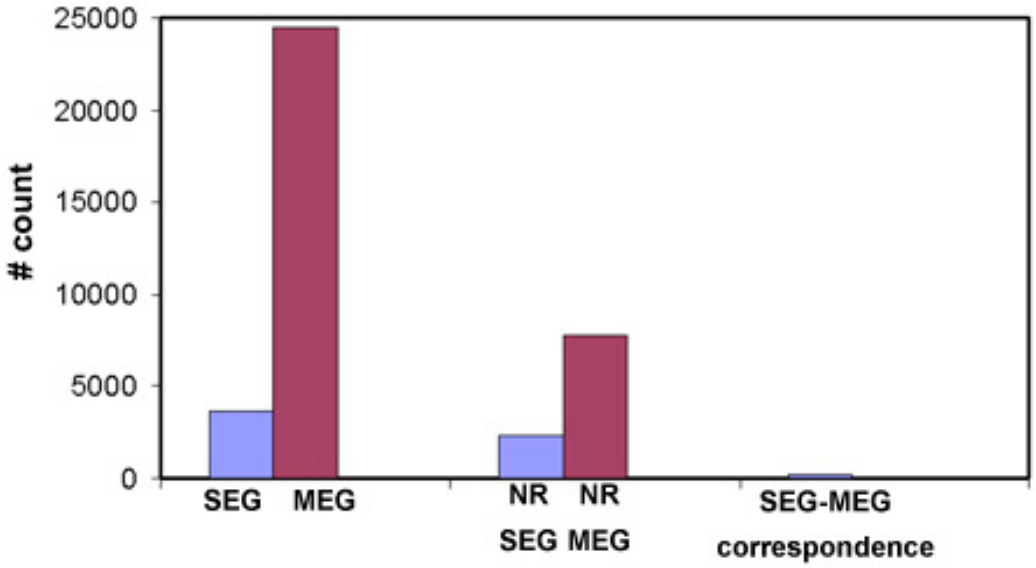

Figure 5. This diagram illustrates SEG-MEG correspondences in the human genome. This is obtained by creating datasets of nr-SEG (non-redundant SEG) and nr-MEG (non-redundant MEG) at 40% (40% cutoff was chosen because homologous sequences below 40% still share similar structural fold) using the sequence purging program CD-HIT (23). Subsequent clustering of nr-SEG and nr-MEG datasets using CD-HIT at 40% sequence identity produced 232 instances of SEG-MEG correspondences (Figure 5) and these genes are referred as retroposed SEG.

Table 4. The top 10 PFAM families in human SEG proteins are given

S.No.	Protein family	PFAM accession	No. of sequences	Location	Function	Features	Remark
1	7 transmembrane receptor (rhodopsin family)	PF00001	491	membrane	rhodopsin-like receptor activity	7 transmembrane (TM) helices	GPCR
2	Zinc finger, C2H2 type	PF00096	180	nucleus	zinc ion binding	two conserved cysteines and histidines co-ordinate a zinc ion	DNA binding protein
3	Leucine rich repeat	PF00560	82	diverse locations	diverse functions	short sequence motifs of β-α unit	DNA binding protein
4	Formin homology region 1	PF06346	69	cytoplasmic or nuclear proteins	cytoskeletal organisers	two sequence domains- the low-complexity, proline-rich FH1 and the carboxy-terminal FH2	Profilin binding protein
5	Core Histone H2A/H2B/ H3/H4	PF00125	63	nucleus	DNA compaction; Gene regulation	histone fold motif	DNA binding protein
6	Cadherin domain	PF00028	56	membrane	calcium ion binding	five cadherin domains - EC1-EC5	Trans-membrane glycoprotein
7	Keratin	PF01500	38	cellular	hair fiber formation	cysteine-rich proteins	High sulfur proteins
8	MAGE family	PF01454	36	cellular	tumor specific antigen	MAGE conserved domain	Melanoma-associated antigen
9	Mammalian taste receptor protein	PF05296	34	cellular	taste receptor	Predicted 7 transmembrane segments	GPCR
10	Helix-loop-helix DNA domain binding	PF00010	34	nucleus	developmental responses via transcriptional regulation	conserved bHLH domain	DNA binding protein

using BLASTN (nucleotide sequence comparison program) at a high E (expect) value cut-off ≤ 10^{-50} over at least 90% of the query sequence length. This exercise identified 203 unique SEG matching known cDNA and these genes are thereafter, referred as SEG cDNA. This number accounts for 5% of SEG in the human genome. The characteristics of SEG cDNA is shown in Figure 4.

3.8. Retrotransposed SEG

The proposed hypothesis in the origin of human SEG is retrotransposition (8). This occurs by homologous recombination between the genomic copy of a gene and an intronless cDNA (9). Therefore, our interest is to identify SEG-MEG correspondences in the human genome. This is done by creating datasets of nr-SEG (non-redundant SEG) and nr-MEG (non-redundant MEG) at 40% (40% cutoff was chosen because homologous sequences below 40% still share similar structural fold) using the sequence purging program CD-HIT (23). Subsequent clustering of nr-SEG and nr-MEG datasets using CD-HIT at 40% sequence identity produced a list of SEG-MEG correspondences (Figure 5) and these genes are referred to thereafter as retrotransposed SEG.

3.9. SEG in GenBank description

GenBank is an annotated collection of all publicly available DNA sequences from a wide variety of sources. Each GenBank entry includes a concise description of the sequence, and a table of features that identify coding regions and other sites of biological significance, such as transcription units, sites of mutations or modifications, and repeats. We used GenBank (release 138) to derive SEGE (4) consisting of 65628 human nuclear encoded SEG. This dataset is largely redundant due to multiple submissions from different authors. Therefore, it is important to remove all redundant sequences in the dataset. This is achieved by purging 65628 sequences at 40% sequence identity cut-off using CD-HIT (23). The purging produced 14279 non-redundant SEG sequences and this gene set is thereafter, referred as GenBank-SEG. We then clustered 3656 Genome-SEG with GenBank-SEG. This exercise produced 1496 clusters consisting of 2663 SEG with at least one SEG from Genome-SEG and one from GenBank-SEG (Figure 6). We assume that GenBank-SEG entries contain a concise description of the sequence with its biological significance obtained by experimental investigation. The characteristic feature of SEG in these clusters is shown in Figure 7.

4. RESULTS AND DISCUSSION

The human genome contains about 28000 genes. They are predominantly intron-bearing (roughly about 88% of human genes) and are frequently MEG (2-3). However, a considerable number of SEG (roughly about 12%) is

Table 5. The top 10 SUPERFAMILY folds in human SEG proteins are given

S. No	Superfamily	Superfamily Accession	No. of sequences	Class	Fold	Function	Features	Remark
1	Membrane receptor all-alpha	56869	568	Membrane and cell surface proteins and peptides	Membrane all-alpha	-	-	Cell-signaling
2	Zinc finger, C2H2 and C2HC	57667	104	Small proteins	C2H2 and C2HC zinc fingers	Zn ion binding	two conserved cysteines and histidines co-ordinate a zinc ion	DNA binding Protein
3	Histone-fold	47113	92	All alpha proteins	Histone-fold	DNA compaction; Gene regulation	histone fold motif	DNA binding protein
4	P-loop containing nucleotide trisphosphate hydrolases	52540	56	Alpha and beta proteins (α/β)	P-loop containing nucleotide triphosphate hydrolases	substrate binding, catalysis and regulation of activity	Conserved glycine-rich region with a phosphate-binding loop (P-loop)	ATP- and GTP-binding proteins
5	Cadherin	49313	56	All beta proteins	Immunoglobulin-like beta-sandwich	calcium ion binding	five cadherin domains - EC1-EC5	DNA binding protein
6	Homeodomain-like	46689	31	All alpha proteins	DNA/RNA-binding 3-helical bundle	transcription regulator	helix-turn-helix (HTH) motif	DNA binding protein
7	RING finger domain, C3HC4	57850	29	Small proteins	RING finger domain, C3HC4	Zn ion binding, ubiquitin-protein ligase activity	cysteine rich RING finger domain	protein-protein interactions
8	Helix-loop-helix DNA-binding domain	47459	29	All alpha proteins	Helix-loop-helix DNA-binding domain	developmental responses via transcriptional regulation	conserved bHLH domain	DNA binding protein
9	L domain-like	52058	29	Alpha and beta proteins (α/β)	Leucine-rich repeat	diverse functions	less regular structure consisting of variable repeats	protein-protein interactions
10	Winged helix DNA-binding domain	46785	28	All alpha proteins	DNA/RNA-binding 3-helical bundle	diverse functions	compact α/β structure, a winged helix motif	DNA binding protein

present in the human genome (5). A similar fraction is present in many mammalian genomes (6). Therefore, it is of interest to investigate their molecular function to deduce possible relationships in their cellular roles. It is also equally important to study the functional prevalence between intronless and intron-bearing human genes. Thus, we systematically analyzed the complete set of human SEG proteins to decipher their concerted molecular function and cellular role. It has already been shown that some human SEG functions as dopamine receptors (10) and adrenergic receptors (13), besides their involvement in HK function (14). Therefore, prediction of SEG function is an important task, considering their significant role in cellular environment. This is done at multiple levels of investigation using PFAM, PROSITE and SUPERFAMILY.

4.1. Biologically meaningful signatures in SEG proteins

The PROSITE database consists of a large collection of biologically meaningful motifs that are described as representative signatures (17). Our analysis identified a total of 19218 motifs consisting of 336 unique types in 3656 SEG proteins. This accounts for nearly 25% of known motifs suggesting that about one-fourth of biologically meaningful signatures are distributed among human SEG proteins. This implies that a wide diversity of biologically important sites is present in SEG proteins, indicating their putative involvement in cellular biology. The top 10 motifs occurring in these proteins are given in Table 2. The distribution of different motifs among SEG proteins is shown in Figure 3a. The distribution suggests that PKC (protein kinase C) phospho site motif ([ST]-x-[RK]) constitutes the largest class (17%), followed by (16%) myristyl (G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}) and (16%) CK2 (casein kinase II) ([ST]-x(2)-[DE]) phospho site motifs. This implies that human SEG proteins are rich in phosphorylation and acetylation signals. The

Table 6. HK-SEG is given in this table. By definition, human housekeeping genes (HK) are constitutively expressed to maintain cellular function

S. No	Genome ID	Gene name	Product name	HK ID	Description
1	1926_NT_011362	GNAS	neuroendocrine secretory protein 55	NM_000516	Homo sapiens GNAS complex locus (GNAS), transcript variant 1,
2	1003_NT_026437	RPL36AL	ribosomal protein L36a-like protein	NM_001001	Homo sapiens ribosomal protein L36a-like (RPL36AL),
3	2811_NT_007592	AIF1	allograft inflammatory factor 1 isoform 2	NM_001623	Homo sapiens allograft inflammatory factor 1 (AIF1), transcript variant 3,
4	2311_NT_022517	CCBP2	chemokine binding protein 2	NM_001296	Homo sapiens chemokine binding protein 2 (CCBP2),
5	1304_NT_037887	SSTR5	somatostatin receptor 5	NM_001053	Homo sapiens somatostatin receptor 5 (SSTR5),
6	1475_NT_010859	MC2R	melanocortin 2 receptor	NM_000529	Homo sapiens melanocortin 2 receptor (adrenocorticotropic hormone) (MC2R),
7	3474_NT_023935	GAS1	growth arrest-specific 1	NM_002048	Homo sapiens growth arrest-specific 1 (GAS1),
8	1935_NT_011387	CENPB	centromere protein B	NM_001810	Homo sapiens centromere protein B, 80kDa (CENPB),
9	2689_NT_077451	MGAT1	mannosyl (alpha-1,3-)-glycoproteinbeta-1,2-N-acetylglucosaminyltransferase	NM_002406	Homo sapiens mannosyl (alpha-1,3-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase (MGAT1),
10	607_NT_033899	FEZ1	zygin 1 isoform 2	NM_005103	Homo sapiens fasciculation and elongation protein zeta 1 (zygin I) (FEZ1), transcript variant 1,
11	431_NT_079542	KIAA0514	KIAA0514	NM_014696	Homo sapiens KIAA0514 gene product (KIAA0514),
12	777_NT_009755	UBC	ubiquitin C	NM_021009	Homo sapiens ubiquitin C (UBC),
13	1689_NT_011295	JUND	jun D proto-oncogene	NM_005354	Homo sapiens jun D proto-oncogene (JUND),
14	1158_NT_010194	ISLR	immunoglobulin superfamily containingleucine-rich repeat	NM_005545	Homo sapiens immunoglobulin superfamily containing leucine-rich repeat (ISLR),
15	1159_NT_010194	ISLR	immunoglobulin superfamily containingleucine-rich repeat	NM_005545	Homo sapiens immunoglobulin superfamily containing leucine-rich repeat (ISLR),
16	1362_NT_010718	UBB	ubiquitin B precursor	NM_018955	Homo sapiens ubiquitin B (UBB),
17	1493_NT_025004	SDCCAG33	serologically defined colon cancer antigen 33	NM_005786	Homo sapiens serologically defined colon cancer antigen 33 (SDCCAG33),
18	1401_NT_010783	HIS1	HMBA-inducible	NM_006460	Homo sapiens HMBA-inducible (HIS1),

result is compared with a distribution for MEG proteins in human, which showed that PKC, CK2 and myristyl motifs are predominant in them (Figure 3b). Thus, the distributions of prominent biological signals are similar in SEG and MEG proteins. Although, the proportion of SEG and MEG in the human genome is significantly different, the distribution of 'weak biological signals' is almost similar. The analysis also identified 14 motifs that are exclusively present in SEG but absent in MEG, and this observation requires further investigation (Table 3). However, PROSITE did not distinctly distinguish SEG and MEG proteins in human. This may be due to "high sensitivity" and "low specificity" offered by short PROSITE signals. Nonetheless, this analysis is valuable in understanding the biological significance of SEG proteins.

4.2. Protein families in SEG proteins

The PFAM database contains a comprehensive set of protein domain families (16). Our results show that about 76% of SEG sequences have a PFAM assignment. The top 10 PFAM families assigned to SEG proteins are given in Table 4. The distribution of different protein families in SEG proteins is shown in Figure 3c. Results show that 13% SEG were assigned to 7-transmembrane receptors, 3% to zinc finger - C2H2 type, 3% to LEUCINE rich repeat and 3% to FORMIN homology region. Table 4 shows that a majority of SEG proteins have DNA binding and GPCR protein family assignment. However, comparison of this distribution, with that of MEG proteins suggests differences between them (Figure 3d). This shows that MEG proteins have a predominance of C2H2 type zinc fingers, protein kinase domain and immunoglobin domain. Moreover, the 7-transmembrane receptor predominant in SEG is subdominant in MEG. This analysis suggests that PFAM is able to differentiate between SEG and MEG proteins. Furthermore, the prevalent assignment of PFAM

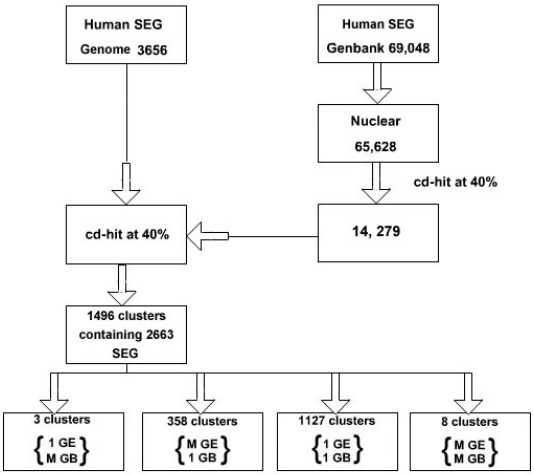

Figure 6. An illustration summarizing the annotation process for SEG proteins using GenBank description is shown in this figure. GenBank is an annotated collection of all publicly available DNA sequences from a wide variety of sources. Each GenBank entry includes a concise description of the sequence, and a table of features that identify coding regions and other sites of biological significance, such as transcription units, sites of mutations or modifications, and repeats. We used GenBank (release 138) to derive SEGE (4) consisting of 65,628 human nuclear encoded SEG. This dataset is largely redundant due to multiple submissions from different authors. Therefore, it is important to remove all redundant sequences in the dataset. This is achieved by purging 65628 sequences at 40% sequence identity cut-off using CD-HIT (23). The purging produced 14,279 non-redundant SEG sequences and this gene set is thereafter, referred as GenBank-SEG. We then clustered 3656 Genome-SEG with GenBank-SEG. This exercise produced 1496 clusters consisting of 2663 SEG with at least one SEG from Genome-SEG and one from GenBank-SEG. GE = genome entry, GB = GenBank entry, M = multiple

7-transmembrane receptors, C2H2 type zinc fingers, LEUCINE rich repeat and FORMIN homology region is valuable.

4.3. Known structural folds in SEG proteins

The SUPERFAMILY is a library of HMM models derived using SCOP defined structural folds (18-19). The top 10 structural families assigned to SEG proteins are given in Table 5 and a distribution of different structural family assignments is shown in Figure 3e. Results show that, (1) membrane-all-alpha (33%), (2) C2H2 and C2HC zinc fingers (6%), and (3) histone-fold (5%) have 568, 104, and 92 assignments, respectively. Comparison with MEG proteins (Figure 3f) shows a different distribution. In MEG proteins, the P loop containing nucleotide triphosphate hydrolases (6%), protein kinase (5%) and C2H2 and C2HC zinc fingers (4%) are prevalent. This analysis suggests that SEG and MEG proteins have distinctly different, yet prevalent structural folds.

4.4. Housekeeping intronless genes

The human HK genes encode for proteins, essential for the maintenance of cell function and are constitutively expressed (20). Eisenberg and colleagues documented a list of 575 HK genes in human that are found to play a key role in the maintenance of cellular function (21). Comparison of SEG gene list with HK gene list identified 18 SEG proteins performing HK functions (Table 6). It has been shown that HK genes are compact. This is true, if the gene expression levels are high and constantly expressed in all cells, which makes them even more valuable to transcribe (21). Therefore, it is obvious that some HK genes are SEG and are constitutively expressed. Nevertheless, a complete list of all HK genes in human is unavailable. SEG are intronless. Intronless genes, not requiring post-transcriptional splicing, might be transcribed efficiently and with potentially greater abundance and rate of protein expression. It is of interest to determine the effect of SEG on mRNA abundance. Genes without introns are not liable to differential and aberrant splicing, which might result in higher transcriptional fidelity.

4.5. Expressed SEG

The origin of SEG is explained by the mechanism of retrotransposition, which occurs by homologous recombination between the genomic copy of a gene and an intronless cDNA (9). Retrotransposed SEG is proposed to exist as pseudo-genes with lost molecular function. Alternatively, they would have gradually evolved with novel molecular function. Therefore, it is our interest to identify SEG that are expressed and captured using reverse transcriptase reaction in RT-PCR experiments. The comparison of SEG with a dataset of full length cDNA sequences from MGC identified 203 unique SEG that are expressed. These are expressed and it is thus captured by RT-PCR. Analysis found that a majority of expressed SEG are histones, GPCR, zinc finger and ion channel complexes (Figure 4). It should be noted that the description of 203 expressed SEG is not representative and a thorough analysis is required to map these data points in a relational system for subsequent interpretation and conclusion.

4.6. Retrotransposed SEG

Retrotransposition of SEG occurs by homologous recombination between the genomic copy of a gene and an intronless cDNA (9). The intronless cDNA originates from an intron-bearing ancestral gene. Therefore, it is our interest to identify a list of SEG having MEG correspondences in the genome. This exercise identified about 232 instances of SEG–MEG correspondence (Figure 5). This is consistent with a similar figure reported by the CELERA team (7). Slight variations in the numbers are due to the different procedures used in each of these analyses. The correspondence provides insights to their possible origin by retroposition, which occurs by homologous recombination between the genomic copy of a gene and an intronless cDNA (12). Retrotransposition of processed mRNA transcripts into the genome results in functional genes, called intronless paralogs, or inactivated genes (pseudogenes). A paralog refers to a gene that appears in more than one copy in a given organism as a result of a duplication event.

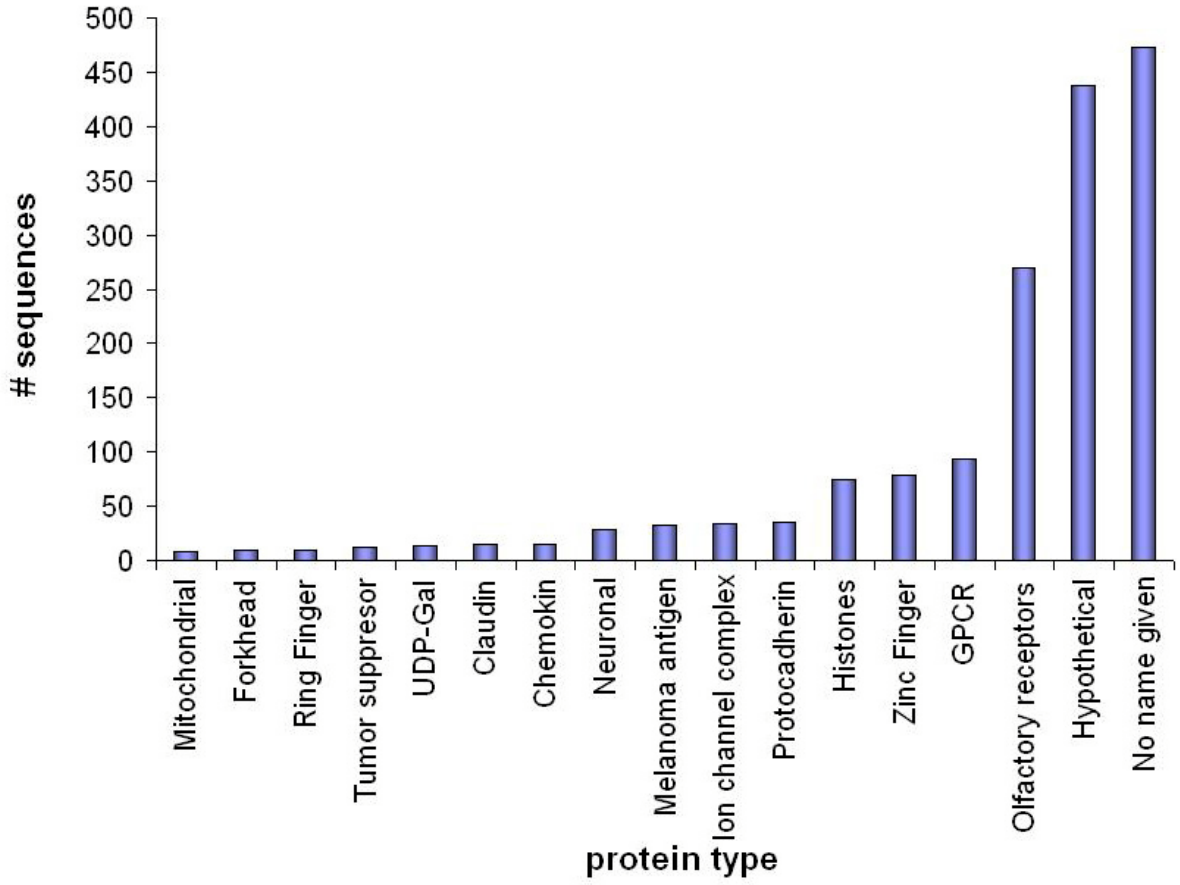

Figure 7. The characteristics of expressed SEG are shown in this figure. Here, we show that human SEG function as olfactory receptors, GPCRs, zinc fingers, histones, protocadherin, ion channel complexes, etc. (Figure 7). We assume that GenBank-SEG entries contain a concise description of the sequence with its biological significance obtained from experimental investigation.

4.7. SEG Characteristics in GenBank description

The concise description of SEG function in GenBank entries is an excellent resource to annotate human SEG in a comprehensive manner. Hence, we compared human SEG from SEGE (derived from GenBank) and Genome SEGE (human genome data) to assign function using GenBank description. An illustration summarizing the annotation process is shown in Figure 6. This exercise grouped 2663 SEG into 1496 clusters, such that each cluster contains at least one SEG from Genome-SEGE and one from SEGE. GenBank annotation of human SEG proteins suggests that a majority of them are olfactory receptors, GPCRs, zinc fingers, histones, protocadherin and ion-channel complexes (Figure 7). This data point is valuable for further experimentation, but not representative of the complete SEG set in human. Therefore, it is necessary to use this information to design experiments (devise hybridization probes and anti-sense RNA probes) to deduce functions for a large number of SEG proteins. A complete functional understanding of all SEG proteins is important to infer their cellular role.

5. CONCLUSIONS

The challenge in bioinformation knowledge discovery is to establish the concerted role played by related group of genes. A comprehensive understanding of their role is essential to compare and contrast the functional features of intronless and intron-bearing human genes. Here, we discussed the complex role of human SEG in cellular environment using computational assignments and this enabled us to predict molecular function for more than half the SEG proteins in humans. This exercise helped us to identify the predominant role played by SEG as DNA binding, phosphorylating, acetylating, housekeeping and also that most SEG are expressed. We also found that many expressed SEG are histones, GPCR, zinc fingers and ion-channel complexes. Despite the fact that the cellular role of reverse transcriptase is not completely understood, several studies have demonstrated that this enzyme is responsible for generating retro-genes which significantly alter genomic activities (24-25). This analysis show evidence of retrotransposed SEG and support the hypothesis of SEG origin by retrotransposition.

6. ACKNOWLEDGEMENTS

This study was supported by Singapore A*STAR-BMRC research grants #01/1/21/19/191 and # 03/1/22/19/242. M.K.S and P.K thanks Tan Tin Wee, Dmitri A Petrov, Sandro de Souza, Manyuan Long and

Liew Kim Meow for their initial support and help in this project.

7. REFERENCES

1. Sakharkar, M. K., V. T. K. Chow and P. Kangueane: Distributions of exons and introns in the human genome. *In Silico Biol* 4, 0032 (2004)

2. Sakharkar, M., M. Long, T. W. Tan and S. J. de Souza: ExInt: an Exon/Intron database. *Nucleic Acids Res* 28, 1, 191-192 (2000)

3. Sakharkar, M., F. Passetti, J. E. de Souza, M. Long, and S. J. de Souza: ExInt: an Exon Intron Database. *Nucleic Acids Res* 30, 1, 191-194 (2002)

4. Sakharkar, M. K, P. Kangueane, D. A. Petrov, A. S. Kolaskar and S. Subbiah: SEGE: A database on 'intron less/single exonic' genes from eukaryotes. *Bioinformatics* 18, 9, 1266-1267 (2002)

5. Sakharkar, M. K and P. Kangueane: Genome SEGE - A database for 'intronless' genes in eukaryotic genomes. *BMC Bioinformatics* 5, 67, 1-5 (2004)

6. Sakharkar, M. K., V. T. K. Chow, I. Chaturvedi, V. S. Mathura, P. Shapshak and P. Kangueane: A report on single exon genes (SEG) in eukaryotes. *Front Biosci* 9, 3, 3262-3267 (2004)

7. Venter, C. J, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C Evangelista, AE Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart B, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M .J. Campbell, K.V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu: The sequence of the human genome. *Science* 291, 5507, 1304-1351 (2001)

8. Brosius, J: Many G-protein coupled receptors are encoded by retro-genes. *Trends Genet* 15, 8, 304-305 (1999)

9. Fink, G. R: Pseudogenes in yeast? *Cell* 49, 1, 5-6 (1987)

10. Sunahara, R. K., H. B. Niznik, D. M. Weiner, T. M. Stormann, M. R. Brann, J. L. Kennedy, J. E. Gelernter, R. Rozmahel, Y. L. Yang, Y. Israel: Human dopamine D1 receptor encoded by an intronless gene on chromosome 5. *Nature* 347, 6288, 80-83 (1990)

11. Brocke, K. S., G. Neu-Yilik, N. H. Gehring, M. W. Hentze, and A. E. Kulozik: The human intronless melanocortin 4-receptor gene is NMD insensitive. *Hum Mol Genet* 11, 3, 331-335 (2002)

12. Demchyshyn, L., R. K. Sunahara, K. Miller, M. Teitler, B. J. Hoffman, J. L. Kennedy, P. Seeman, H. H. Van Tol and H. B. Niznik: A human serotonin 1D receptor variant (5HT1D beta) encoded by an intronless gene on chromosome 6. *Proc Natl Acad Sci* 89, 12, 5522-5526 (1992)

13. Kobilka, B. K., T. Frielle, H. G. Dohlman, M. A. Bolanowski, R. A. Dixon, P. Keller, M. G. Caron and R. J. Lefkowitz: Delineation of the intronless nature of the genes for the human and hamster beta 2-adrenergic receptor and their putative promoter regions. *J Biol Chem* 262, 15, 7321-7327 (1987)

14. Rampazzo, A., F. Pivotto, G. Occhi, N. Tiso, S. Bortoluzzi, L. Rowen, L. Hood, A. Nava, and G. A. Danieli: Characterization of C14orf4, a novel intronless human gene containing a polyglutamine repeat, mapped to

the ARVD1 critical region. *Biochem Biophys Res Commun* 278, 3, 766-774 (2000)

15. Gentles, A. J. and S. Karlin: Why are human G-protein-coupled receptors predominantly intronless? *Trends Genet* 15, 2, 47-49 (1999)

16. Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats and S. R. Eddy: The Pfam protein families database. *Nucleic Acids Res* 32, 1, D138-D141 (2004)

17. Hulo, N., C. J. Sigrist, V. Le Saux, P. S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher and A. Bairoch: Recent improvements to the PROSITE database. *Nucleic Acids Res* 32, 1, D134-D137 (2004)

18. Madera, M., C. Vogel, S.K. Kummerfeld, C. Chothia and J. Gough: The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res* 32, 1, D235-239 (2004)

19. Gough, J. and C. Chothia: SUPERFAMILY - HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* 30, 1, 268-272 (2002)

20. Butte, A. J., V. J. Dzau and S. B. Glueck: Further defining housekeeping, or "maintenance," genes. Focus on "A compendium of gene expression in normal human tissues". *Physiol Genomics* 7, 2, 95-96 (2001)

21. Eisenberg, E. and E. Y. Levanon: Human housekeeping genes are compact. *Trends Genet* 19, 7, 362-365 (2003)

22. Strausberg, R. L., E. A. Feingold, R. D. Klausner and F. S. Collins: The Mammalian Gene Collection. *Science* 286, 5439, 455-457 (1999)

23. Li, W., L. Jaroszewski and A. Godzik: Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17, 3, 282-283 (2001)

24. Brosius, J. and H. Tiedge: Reverse transcriptase: mediator of genomic plasticity. *Virus Genes* 11, 2-3, 163-179 (1995)

25. Brosius, J: The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* 118, 2-3, 99-116 (2003)

Key Words: SEG, Intronless, MEG, Intron bearing, Retroposition, Reverse transcriptase activity, cDNA, mRNA, Evolution, Intron function, Intron loss, Intron, Exon, Pseudo genes, RNA world

Send correspondence to: Dr Pandjassarame Kangueane, School of Mechanical and Production Engineering, 50 Nanyang Avenue, Nanyang Technological University, Singapore 639 798, Tel: 65-6-790-5836, Fax: 65-6-774-4340, E-mail: mpandjassarame@ntu.edu.sg