[Frontiers in Bioscience 9, 3262-3267, September 1, 2004]

A REPORT ON SINGLE EXON GENES (SEG) IN EUKARYOTES

Meena Kishore Sakharkar [1], Vincent Tak Kwong Chow [2], Iti Chaturvedi [2], Venkatarajan Subramanian Mathura [3], Paul Shapshak [4], Pandjassarame Kangueane [1]

[1] School of Mechanical and Production Engineering, NCSV, Nanyang Technological University, Singapore, [2] Department of Microbiology, National University of Singapore, Singapore, [3] Roskamp Institute, 2040 Whitfield Ave, Sarasota FL 34243, USA, [4] Department of Psychiatry and Behavior Science, University of Miami Medical School, Miami, FL 33136

TABLE OF CONTENTS

1. ABSTRACT

Single exon genes (SEG) are archetypical of prokaryotes. Hence, their presence in intron-rich, multi-cellular eukaryotic genomes is perplexing. Consequently, a study on SEG origin and evolution is important. Towards this goal, we took the first initiative of identifying and counting SEG in nine completely sequenced eukaryotic organisms - four of which are unicellular (*E. cuniculi, S. cerevisiae, S. pombe, P. falciparum*) and five of which are multi-cellular (*C. elegans, A. thaliana, D. melanogaster, M. musculus, H. sapiens*). This exercise enabled us to compare their proportion in unicellular and multi-cellular genomes. The comparison suggests that the SEG fraction decreases with gene count (r = -0.80) and increases with gene density (r = 0.88) in these genomes. We also examined the distribution patterns of their protein lengths in different genomes.

2. INTRODUCTION

Many eukaryotic genes contain 'non-coding', intra-genic segments called introns within the coding sequence (1-2). Hence, eukaryotic genes often have interrupted gene structures with multiple exons. However, prokaryotic genes lack introns and are characterized by un-interrupted single exon gene (SEG) structures. Therefore, prokaryotic genes are typically SEG. Although, multiple exon genes (MEG) are characteristic of many vertebrate eukaryotic genomes, several SEG have also been identified in them (3-7). Recently, specialized databases (SEGE and Genome SEGE) have been constructed for SEG from GenBank (6) and genome data (7). The presence of SEG in eukaryotic genomes is intriguing and their presence in intron-rich eukaryotic genomes is perplexing (3-7). Hence, it is important to study the origin and evolution of SEG in eukaryotes, more importantly in higher eukaryotes. To study their evolution on a genomic scale is resource intensive, information demanding and extremely complex. We undertook the first initiative of identifying and counting SEG for nine eukaryotic genomes. Here, we compare the proportional selection of SEG in different genomes and discuss the distribution patterns of their protein lengths.

3. MATERIALS AND METHODS

The current analysis is performed for nine completely sequenced eukaryotic organisms - four of which are unicellular (*E. cuniculi, S. cerevisiae, S. pombe* and *P. falciparum*) and five of which are multi-cellular (*C. elegans, A. thaliana, D. melanogaster, M. musculus* and *H. sapiens*). Data is obtained from Genome SEGE for this analysis (7). Table 1 lists SEG count and fraction in each genome. Figure 3 shows SEG protein lengths distribution in nine genomes.

4. RESULTS AND DISCUSSION

4.1. Fraction of SEG in genomes

Table 1 shows SEG count in nine eukaryotic genomes. The differences reflect inherent variations in different genome architectures and evolutionary divergences. Although, this trend is not surprising, the actual estimates are interesting in the sense that their proportions in some genomes are distinctly greater than others. We note from Table 1 that unicellular (45%-98%) and multi-cellular (3%-20%) genomes are distinguished by SEG proportion in them. Generally, the SEG fraction is greater in unicellular than multi-cellular genomes. This implies that unicellular genomes with very short generation times have larger fraction, while multi-cellular genomes with long generation times have smaller fraction. The Pearson correlation co-efficient (r) between SEG count and genome size is 0.2. This is much weaker than the Pearson correlation co-efficient between total gene count and genome size (r = 0.61). The r value between SEG count and gene count is 0.3. However, the r value between SEG

Table 1. An estimate of SEG count[1] in different eukaryotic genomes

Genomes	Single Exon Genes (SEG)			SEG Fraction (%)	Genome Size (Mb)	Total Genes	Reference
	Total	Pseudo	SEG[(a)]			Gene count (current update)	
Encephalitozoon cuniculi	1,981	0	1,981	97.7	2.9	2,028	(8)
Saccharomyces cerevisiae	5,551	60	5,491	92.5	12.068	6,004	(9)
Plasmodium falciparum	2,471	991	1,480	44.6	23	5,544	(10)
Schizosacharomyces pombe	2,585	17	2,468	49.6	13.8	5,213	(11)
Caenorhabditis elegans	654	3	651	2.7	97	24,607	(12)
Arabidopsis thaliana	5,920	84	5,836	20.1	125	29,483	(13)
Drosophila melanogaster	2,049	29	2,020	18.0	180	11,357	(14)
Mus musculus	4,218	105	4,113	15.8	2,500	26,771	(15, 16)
Homo sapiens	3,408	103	3,305	12.3	2900	27,675	(5, 17)

[1] The SEG fraction is defined as the percentage ratio of SEG count and gene count. Pseudo = processed pseudo genes that are SEG. SEG[(a)] = SEG count after eliminating processed pseudo genes.

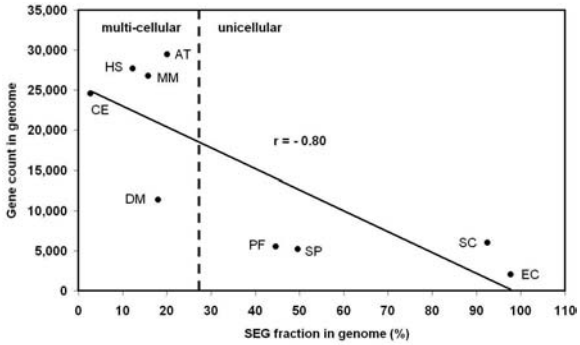

Figure 1. Relationship between SEG fraction and gene count is given. *EC = E. cuniculi, SC = S. cerevisiae, SP = S. pombe, PF = P. falciparum, CE = C. elegans, AT = A. thaliana, DM = D. melanogaster, MM = M. musculus, HS = H. sapiens*

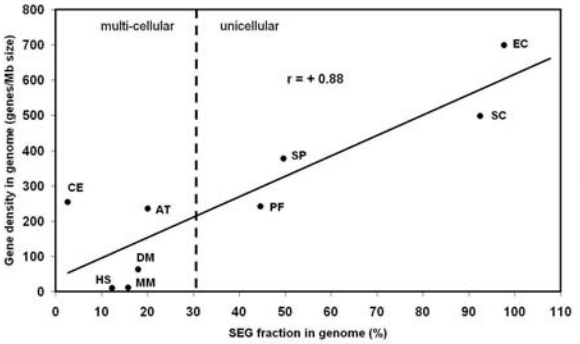

Figure 2. Relationship between SEG fraction and gene density is given. *EC = E. cuniculi, SC = S. cerevisiae, SP = S. pombe, PF = P. falciparum, CE = C. elegans, AT = A. thaliana, DM = D. melanogaster, MM = M. musculus, HS = H. sapiens*

fraction and genome size is -0.45. This suggests that SEG fraction decreases with genome size. Interestingly, the r value between SEG fraction and gene count is -0.80 (Figure 1). Thus, SEG fraction strongly decreases with total gene count in these genomes. In other words, genomes with high gene count contain low SEG fraction. We also found that the r value between SEG fraction and gene density (total gene count / Mb genome size) is 0.88. This relationship is strong and SEG fraction increases linearly with increase in gene density in these genomes (Figure 2). These patterns are very interesting and subsequent analysis is required to gain further insight into their selection and genome design. However, the bits and pieces of derived information have to be bridged together to signify the trend between SEG fraction and genome content. We hope to compare and contrast estimates from different genomes of distant phylogeny.

4.2. Mechanism of SEG origin

Table 1 shows that multi-cellular genomes contain about 12-20% SEG. This is not true for *C. elegans* and it contains only 2.7% SEG. The latest update (October, 2003) of the human genome contains 3,408 SEG sequences (about 12 % of total genes). These estimates are relatively large and their mere existence in many intron-rich genomes demands further investigations. It has been suggested that a significant fraction of human SEG have been generated by retro-position (18). Therefore, the presence of SEG can be explained by the mechanism of retro-position. This occurs by homologous recombination between the genomic copy of a gene and an intronless cDNA (19). The later is produced by reverse transcription of the corresponding mRNA, a mechanism that produces SEG genes in eukaryotes. In an independent experiment by sequence comparison, we found that about 20% (366) of unique SEG (purged at 40% sequence identity) show (MEG) correspondence with at least 40% sequence identity (data not shown). This strongly supports the hypothesis that human SEG arose by retroposition.

The human genome team suggested that a very small fraction of total human genes (<1%) is exclusively homologous to bacterial genes (17). Therefore, we compared human SEG with 430,011 prokaryotic protein sequences derived from 135 prokaryotic genomes. About 99% of human SEG lack homology with prokaryotic sequences. This suggests that human SEG did not evolve by

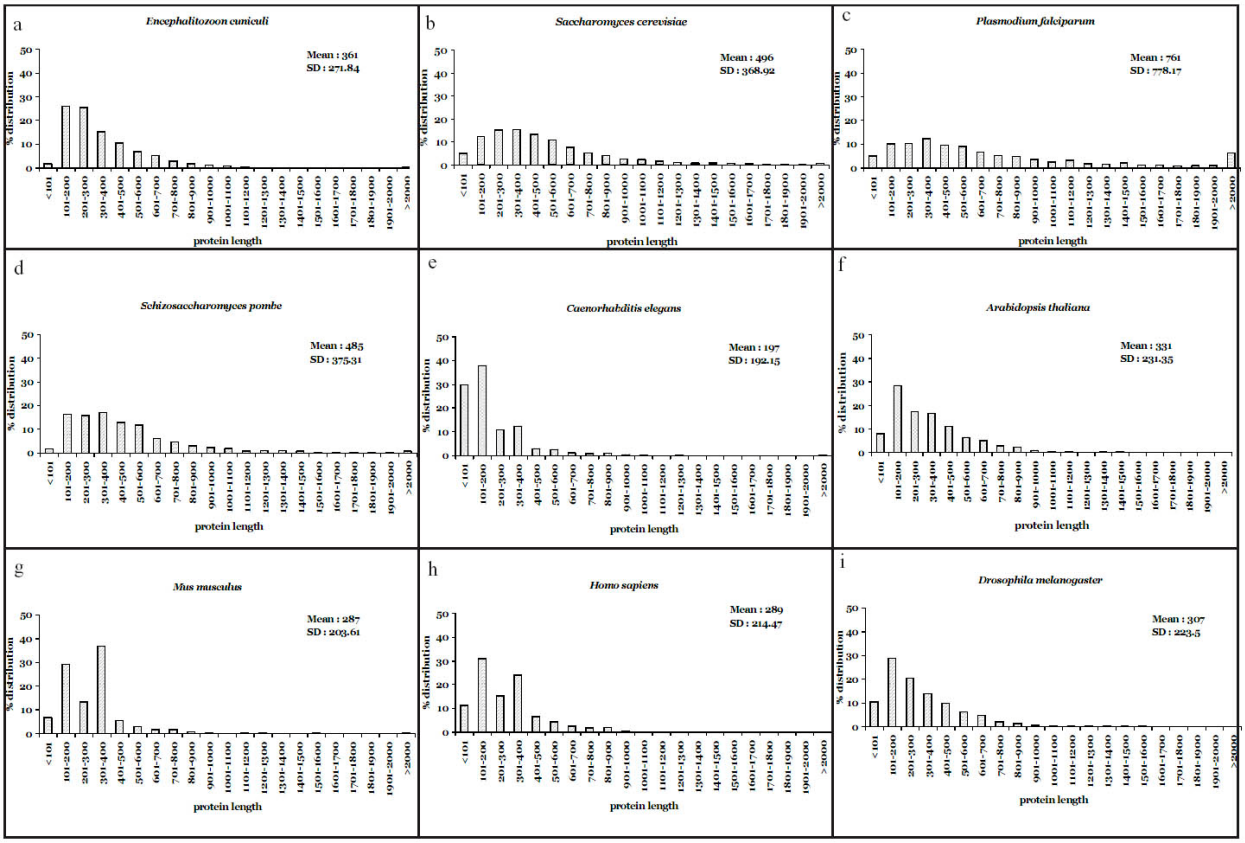

Figure 3. Distribution of SEG protein length in nine genomes is shown. The mean length and standard deviation (SD) is given for each genome.

gene transfer from bacteria to human. Nonetheless, the absence of homology between human SEG and prokaryotic proteins supports the hypothesis that SEG probably arose by retro-position. Additional data on paralogous SEG may provide further evidence towards the possible mechanism of their origin by retro-position.

4.3. Distribution of SEG protein length in genomes

SEG proportion varies between genomes (Table 1) and their fraction is related to gene count (Figure 1) and gene density (Figure 2). Therefore, the SEG fraction is related to genome content. We thus probed into the distribution patterns of SEG protein length in nine genomes (Figure 3). Figure 3 shows that the protein length distribution patterns are not identical. The mean and standard deviation of SEG protein length in genome facilitate comparison among them and suggest some degree of similarity. Multi-cellular genomes (*C. elegans, A. thaliana, D. melanogaster, M. musculus, H. sapiens*) show a mean length of about 200-330 residues and unicellular genomes (*E. cuniculi, S. cerevisiae, S. pombe, P. falciparum*) show a mean length of about 350-750 residues (Figure 3). Thus, SEG fraction and mean length can distinguish between unicellular and multi-cellular genomes. The standard deviation about the mean is found to increase with mean length in these genomes (r = 0.97). In general, unicellular genomes contain longer SEG than multi-cellular genomes (Figure 3). The r value between mean SEG length and SEG fraction in genomes is 0.47. We also found that the r value between mean SEG length and gene count is -0.64. However, mean lengths show poor correlation to SEG count (r = -0.04) and gene density (r = 0.28). Thus, mean SEG length is more related to gene count and SEG fraction than to SEG count and gene density.

5. CONCLUSION

The biological role of SEG in the genomes of higher organism is not completely understood. Here, we show that different eukaryotic genomes have varying SEG fraction and a sizeable portion of them is found in many intron-rich multi-cellular genomes. This report provides an overview of SEG count, fraction, protein lengths distribution patterns and mean SEG length in nine eukaryotic genomes and shows their relationship to genome size, gene count and gene density. It is also interesting to note that a large proportion of SEG are associated with unicellular organisms with very short generation times, while a small proportion of SEG is common in relatively complex multi-cellular organisms with long generation times. We hope that these estimates will help to probe into the biological role of SEG and their contribution in genome design.

6. ACKNOWLEDGEMENT

This research is supported by A*STAR (BMRC research) grant #01/1/21/19/191. The infrastructure support provided by NCSV, NTU and the utility support provided by NUS is duly acknowledged. We wish to express our sincere appreciation to all members of the research group for many discussions on the subject of this article. PS is supported by NIH grants DA014533 and GM056529.

7. REFERENCES

1. Gilbert W: Why genes in pieces? *Nature* 271(5645), 501 (1978)

2. Sakharkar M, F. Passetti, J.E. de Souza, M. Long, S.J. de 2. Souza: ExInt: an Exon Intron Database. *Nucleic Acids Res.* 30(1), 191-194 (2002)

3. Brosius J: Genomes were forged by massive bombardments with retro-elements and retro-sequences. *Genetica* 107(1-3), 209-238 (1999)

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

5. 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, JD 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 & X. Zhu: The sequence of the human genome. *Science* 291(5507), 1304-1351 (2001)

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

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

8. Katinka M.D, S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, F. Delbac, H. El Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach & C.P. Vivares: Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414(6862), 450-453 (2001)

9. Goffeau A, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin & S.G. Oliver: Life with 6000 genes. *Science* 274(5287), 563-567 (1996)

10. Gardner M.J, N. Hall, E. Fung, O. White, M. Berriman, R.W. Hyman, J.M., A. Carlton Pain, K.E. Nelson, S. Bowman, I.T. Paulsen, K. James, J.A. Eisen, K. Rutherford, S.L. Salzberg, A. Craig, S. Kyes, M.S. Chan, V. Nene, S.J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M.W. Mather, A.B. Vaidya, D.M. Martin, A.H. Fairlamb, M.J. Fraunholz, D.S. Roos, S.A. Ralph, G.I. McFadden, L.M. Cummings, G.M. Subramanian, C. Mungall, J.C. Venter, D.J. Carucci, S.L. Hoffman, C. Newbold, R.W. Davis, C.M. Fraser & B. Barrell: Genome sequence of human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906), 498-511 (2002)

11. Wood V, R. Gwilliam, M.A. Rajandream, M. Lyne, R. Lyne, A. Stewart, J. Sgouros, N. Peat, J. Hayles, S. Baker,

D. Basham, S. Bowman, K. Brooks, D. Brown, S. Brown, T. Chillingworth, C. Churcher, M. Collins, R. Connor, A. Cronin, P. Davis, T. Feltwell, A. Fraser, S. Gentles, A. Goble, N. Hamlin, D. Harris, J. Hidalgo, G. Hodgson, S. Holroyd, T. Hornsby, S. Howarth, E.J. Huckle, S. Hunt, K. Jagels, K. James, L. Jones, M. Jones, S. Leather, S. McDonald, J. McLean, P. Mooney, S. Moule, K. Mungall, L. Murphy, D. Niblett, C. Odell, K. Oliver, S. O'Neil, D. Pearson, M.A. Quail, E. Rabbinowitsch, K. Rutherford, S. Rutter, D. Saunders, K. Seeger, S. Sharp, J. Skelton, M. Simmonds, R. Squares, S. Squares, K. Stevens, K. Taylor, R.G. Taylor, A. Tivey, S. Walsh, T. Warren, S. Whitehead, J. Woodward, G. Volckaert, R. Aert, J. Robben, B. Grymonprez, I. Weltjens, E. Vanstreels, M. Rieger, M. Schafer, S. Muller-Auer, C. Gabel, M. Fuchs, A. Dusterhoft, C. Fritzc, E. Holzer, D. Moestl, H. Hilbert, K. Borzym, I. Langer, A. Beck, H. Lehrach, R. Reinhardt, T.M. Pohl, P. Eger, W. Zimmermann, H. Wedler, R. Wambutt, B. Purnelle, A. Goffeau, E. Cadieu, S. Dreano, S. Gloux, V. Lelaure, S. Mottier, F. Galibert, S.J. Aves, Z. Xiang, C. Hunt, K. Moore, S.M. Hurst, M. Lucas, M. Rochet, C. Gaillardin, V.A. Tallada, A. Garzon, G. Thode, R.R. Daga, L. Cruzado, J. Jimenez, M. Sanchez, F. del Rey, 11. J. Benito, A. Dominguez, J.L. Revuelta, S. Moreno, J. Armstrong, S.L. Forsburg, L. Cerutti, T. Lowe, W.R. McCombie, I. Paulsen, J. Potashkin, G.V. Shpakovski, D. Ussery, B.G. Barrell, P. Nurse & L. Cerrutti: The genome sequence of *Schizosaccharomyces pombe*. *Nature* 415(6874), 871-880 (2002)

12. The *Caenorhabditis elegans* sequencing consortium, Genome Sequence of the Nematode *Caenorhabditis elegans*. *Science* 282(5396), 2012-2018 (1998)

13. The *Arabidopsis* genome initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814), 796-815 (2000)

14. Adams M.D, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, S.E. Lewis, S. Richards, M. Ashburner, S.N. Henderson, G.G. Sutton, J.R. Wortman, M.D. Yandell, Q. Zhang, L.X. Chen, R.C. Brandon, Y.H. Rogers, R.G. Blazej, M. Champe, B.D. Pfeiffer, K.H. Wan, C. Doyle, E.G. Baxter, G. Helt, C.R. Nelson, G.L. Gabor, J.F. Abril, A. Agbayani, H.J. An, C. Andrews-Pfannkoch, D. Baldwin, R.M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E.M. Beasley, K.Y. Beeson, P.V. Benos, B.P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M.R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K.C. Burtis, D.A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J.M. Cherry, S. Cawley, C. Dahlke, L.B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A.D. Mays, I. Dew, S.M. Dietz, K. Dodson, L.E. Doup, M. Downes, S. Dugan-Rocha, B.C. Dunkov, P. Dunn, K.J. Durbin, C.C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A.E. Gabrielian, N.S. Garg, W.M. Gelbart, K. Glasser, A. Glodek, F. Gong, J.H. Gorrell, Z. Gu, P. Guan, M. Harris, N.L. Harris, D. Harvey, T.J. Heiman, J.R. Hernandez, J. Houck, D. Hostin, K.A. Houston, T.J. Howland, M.H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G.H. Karpen, Z. Ke, J.A. Kennison, K.A. Ketchum, B.E. Kimmel, C.D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A.A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T.C. McIntosh, M.P. McLeod, D. McPherson, G. Merkulov, N.V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S.M. Mount, M. Moy, B. Murphy, L. Murphy, D.M. Muzny, D.L. Nelson, D.R. Nelson, K.A. Nelson, K. Nixon, D.R. Nusskern, J.M. Pacleb, M. Palazzolo, G.S. Pittman, S. Pan, J. Pollard, V. Puri, M.G. Reese, K. Reinert, K. Remington, R.D. Saunders, F. Scheeler, H. Shen, B.C. Shue, I. Siden-Kiamos, M. Simpson, M.P. Skupski, T. Smith, E. Spier, A.C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A.H. Wang, X. Wang, Z.Y. Wang, D.A. Wassarman, G.M. Weinstock, J. Weissenbach, S.M. Williams, T. Woodage, K.C. Worley, D. Wu, S. Yang, Q.A. Yao, J. Ye, R.F. Yeh, J.S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X.H. Zheng, F.N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H.O. Smith, R.A. Gibbs, E.W. Myers, G.M. Rubin & JC Venter: The genome sequence of *Drosophila melanogaster*. *Science* 287(5461), 2185-2195 (2000)

15. Waterston R.H, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S.E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M.R. Brent, D.G. Brown, S.D. Brown, C. Bult, J. Burton, J. Butler, R.D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A.T. Chinwalla, D.M. Church, M. Clamp, C. Clee, F.S. Collins, L.L. Cook, R.R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K.D. Delehaunty, J. Deri, E.T. Dermitzakis, C. Dewey, N.J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D.M. Dunn, S.R. Eddy, L. Elnitski, R.D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G.A. Fewell, P. Flicek, K. Foley, W.N. Frankel, L.A. Fulton, R.S. Fulton, T.S. Furey, D. Gage, R.A. Gibbs, G .Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T.A. Graves, E.D. Green, S. Gregory, R. Guigo, M. Guyer, R.C. Hardison, D. Haussler, Y. Hayashizaki, L.W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D.B. Jaffe, L.S. Johnson, M. Jones, T.A. Jones, A. Joy, M. Kamal, E.K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W.J. Kent, A. Kirby, D.L. Kolbe, I. Korf, R.S. Kucherlapati, E.J. Kulbokas, D. Kulp, T. Landers, J.P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D.R. Maglott, E.R. Mardis, L. Matthews, E. Mauceli, J.H. Mayer, M. McCarthy, W.R. McCombie, S. McLaren, K. McLay, J.D. McPherson, J. Meldrim, B. Meredith, J.P. Mesirov, W. Miller, T.L. Miner, E. Mongin, K.T. Montgomery, M. Morgan, R. Mott, J.C. Mullikin, D.M. Muzny, W.E. Nash, J.O. Nelson, M.N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M.J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K.H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C.S. Pohl, A. Poliakov, T.C. Ponce, C.P. Ponting, S. Potter, M. Quail, A. Reymond, B.A. Roe, K.M. Roskin, E.M. Rubin, A.G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M.S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J.B. Singer, G. Slater, A. Smit, D.R.

Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J.P. Vinson, A.C. Von Niederhausern, C.M. Wade, M. Wall, R.J. Weber, R.B. Weiss, M.C. Wendl, A.P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R.K. Wilson, E. Winter, K.C. Worley, D. Wyman, S. Yang, S.P. Yang, E.M. Zdobnov, M.C. Zody & E.S. Lander: Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915), 520-562 (2002)

16. Gregory S.G, M. Sekhon, J. Schein, S. Zhao, K. Osoegawa, C.E. Scott, R.S. Evans, P.W. Burridge, T.V. Cox, C.A. Fox, R.D. Hutton, I.R. Mullenger, K.J. Phillips, J. Smith, J. Stalker, G.J. Threadgold, E. Birney, K. Wylie, A. Chinwalla, J. Wallis, L. Hillier, J. Carter, T. Gaige, S. Jaeger, C. Kremitzki, D. Layman, J. Maas, R. McGrane, K. Mead, R. Walker, S. Jones, M. Smith, J. Asano, I. Bosdet, S. Chan, S. Chittaranjan, R. Chiu, C. Fjell, D. Fuhrmann, N. Girn, C. Gray, R. Guin, L. Hsiao, M. Krzywinski, R. Kutsche, S.S. Lee, C. Mathewson, C. McLeavy, S. Messervier, S. Ness, P. Pandoh, A.L. Prabhu, P. Saeedi, D. Smailus, L. Spence, J. Stott, S. Taylor, W. Terpstra, M. Tsai, J. Vardy, N. Wye, G. Yang, S. Shatsman, B. Ayodeji, K. Geer, G. Tsegaye, A. Shvartsbeyn, E. Gebregeorgis, M. Krol, D. Russell, L. Overton, J.A. Malek, M. Holmes, M. Heaney, J. Shetty, T. Feldblyum, W.C. Nierman, J.J. Catanese, T. Hubbard, R.H. Waterston, J. Rogers, P.J. de Jong, C.M. Fraser, M. Marra, J.D. McPherson & D.R. Bentley: A physical map of the mouse genome. *Nature* 418(6899), 743-750 (2002)

17. Lander E.S, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. 18. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, J. Szustakowki, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi & Y.J. Chen: Initial sequencing and analysis of the human genome. *Nature* 409(6822), 860-921 (2001)

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

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

Key Words: SEG, eukaryotic genomes, protein length distribution, evolution

Send correspondence to: Pandjassarame Kangueane Ph.D, Nanyang Technological University, Singapore - 639798, Tel: +65 6790 5836, Fax: +65 6774 4340, E-mail: mpandjassarame@ntu.edu.sg