[Frontiers in Bioscience 10, 2576-2584, September 1, 2005]

HUMAN GENOME – FROM PIECES TO PATTERNS

Meena Kishore Sakharkar [1], Pandjassarame Kangueane [1], Bagavathi S. Perumal [1], Vincent Tak Kwong Chow [2], Eric Sorscher [3], Kishore R. Sakharkar [4] and Aubrey Hill [3]

[1] NCSV, MAE, Nanyang Technological University, Singapore 639798, [2] Department of Microbiology, National University of Singapore, Singapore 119260, [3]Department of Medicine, University of Alabama at Birmingham, 796 Birmingham, AL 35294-0005, USA, [4]BioInformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore 138671

TABLE OF CONTENTS

1. ABSTRACT

A profile of exon-intron lengths in genes shows a normal distribution. This observation suggests that different genes may have portions of their total exon and/or intron lengths in common. In order to explore the common exon-intron structural patterns that may arise due to common lengths across genes, we compared the exon-intron length patterns of annotated human genes. We discovered 1762278 conserved arrangements of exon-intron length across the otherwise unrelated and diverse genomic landscape. The existence of common exon-intron length patterns across unrelated genes suggests for their role of in gene assemblage and human genome design and architecture.

2. INTRODUCTION

The boundaries between exons and introns are blurring (1). Researchers have shown how dynamic the genome is and how the repetitive elements of so-called junk DNA are actually, as Makalowski puts it, "a genomic treasure" and "a source of 'ready-to-use-motifs' " increasing an organism's evolutionary flexibility (1). The structural and functional properties of DNA change as a function of its nucleotide composition. The human genome has been described as exon islands in a vast sea of introns (2). The size and prevalence of introns (about 25%) in more complex organisms suggests that introns could be important functional elements in large genomes (3). In order to understand the structure and evolution of genes and genomes, it is important to know the general statistical characteristics of the exon-intron structures. The first compilation of exon-intron structures in eukaryotic genes was published by Hawkins in 1988 (4). Since then many disparate reports have been presented and the use of patterns in exons and introns to understand gene structure is becoming increasingly ubiquitous. Also, the sequential arrangement of coding (exons) and non-coding (introns) regions is of particular interest from a biological viewpoint in revealing essential details necessary for understanding the assembly of the spliceosome and the splicing process in general. Recently, it was reported that different genes have portions of their total exon-intron sequential structure in common. The analysis reported more than 200 patterns of length 2 (length 2 implies a block of exon and an intron) or greater among the 72 human genes (5). The data is novel and needs further analysis at genome level. Exploration of such patterns at the genome level will confirm their factual nature and provide clues to their role in genome design and gene architecture. Due to the size and complexity of human genes, it has not been possible in the past to determine whether common exon-intron size patterns exist among different genes by visual inspection. Our goal in this exploration is to get a list of all such conserved patterns of length two or more and to compare the exon-intron patterns across the human genome landscape representing the length of the exons and introns as integers while ignoring the underlying nucleotide sequence (5). This allows us to study the distribution of exon-intron blocks in human genes to determine whether one can detect the reuse of exon-intron sequences and to use the frequency of such reuse to estimate how many ancestral exon-intron block sequences there might have been. This study is imperative for the theoretical study of the origin and evolution of genes and genomes. These findings could help improve gene structure prediction by computational methods by providing better understanding of factors that govern genome assemblage and design.

3. MATERIALS AND METHODS

The exon-intron structure of human genes was obtained from human genome data downloaded from NCBI build 34 and AceView (http://www.ncbi.nlm.nih.gov/AceView/) database (6). The length, in nucleotides, for each exon or intron was alternately entered as an element of an array representing the structure of that gene. The arrays were indexed such that the first element was designated exon (starting at '0'), the second element as intron, and so on. This scheme resulted in the exons residing in the even numbered elements of the array and the introns residing in the odd-numbered elements of the array. The array representing the exon-intron structure of the human FUCA1 and ADH5 gene is presented in Table 1. Pair wise comparisons of 27115 human genes (246632 exons and 218575 introns) were performed. This involved 7.35 X 10^9 gene pair comparisons. Each of the 'exon and the subsequent intron' or 'intron and subsequent exon' which we call a "block of length 2" from each gene was compared to other "blocks of length 2" from all annotated genes throughout the database, to identify pairs of genes that have at least one block of length 2 matched in length (99% similar in length). The lengths were extracted based on CDS feature as described earlier (7). Extensions of block lengths from 3 to >25 were computed and tabulated (Table 2). The identified pairs were further categorized into subgroups based on the start of the match with exon or intron, and whether the genes are on the sense or anti-sense strand (Table 3). Redundancy in blocks of length 2 to >25 was computed (at 99% of components length) to identify pairs of unique length patterns (Table 4). Each of the entries from genome data were matched with corresponding Aceview entry by using Gene name and the exon/intron lengths were identified in any of the splice variants in Aceview and the lengths were categorized as confirmed components. AceView is a database that offers an integrated view of the human genes as reconstructed by alignment of all publicly available mRNAs and ESTs on the genome sequence. These confirmed components were mapped back onto the initial pool of blocks of length 2 to >25 so as to identify the pairs that have Aceview validated block lengths (Table 4). A dataset on new patterns of blocks of length 2 from higher block lengths were generated and categorized (Table 5). Similar analysis was performed to identify unique blocks of length 3 and tabulated (Table 6).

4. RESULTS & DISCUSSION

4.1. From pieces to patterns by retroposition

The Human genome contains 246632 exons and 218575 introns from 27115 gene sequences (build 34 version 3.0). 1762278 patterns of "blocks of length 2" were identified across the genomic landscape that had at least 1 exon and 1 intron length in common (99% similar in length). Distributions of the total number of pairs based on the number of matched exon-intron lengths they contain show a wavelet distribution (Figure 1). The number of genes that share blocks of length 3 is 16380, this number decreases to 637 when the block size is 4 and then increases again to 1335 for 5 (Table 2). This pattern is observed in both the sense and the anti-sense strand (Table 3). Such alternate patterns suggest that odd number pairs are more frequent than even number pairs. This suggests that similar components are more preferred at the boundaries. For example, it is more frequent to find a block of length 5 with **exon**-intron-exon-intron-**exon** than to find a block of length 4 e.g. **exon**-intron-exon-**intron**. Also, pairs beginning with an exon and ending in exon were found to be more frequent than the ones beginning with introns and ending with introns (Table 3). Though, the dynamics of coding regions are expected to be more correlated with functional complexity and diversity than are the dynamics of non-coding regions, it suggests that the dynamics of coding regions are not independent of non-coding regions. However, the possible reasons for such occurrences are obscure.

Interestingly, genes that share blocks of length 4 or more lie on the same chromosome 90% of the time (Table3; Figure 2). The prevalence of similar blocks on same chromosome suggests for their evolution by duplication and divergence. These results are in accordance with previous findings and suggest that the dynamics of

Table 1. Gene comparison matrix for FUCA1 and ADH5

	Position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	Length (bp)	**E**	**I**	**E**	**I**	**E**	**I**	**E**	**I**	**E**	**I**	**E**	**I**	**E**	**I**	**E**
		374	2272	135	2219	138	3230	106	5237	201	5520	191	2475	100	218	141
E	12	0.032		0.089		0.087		0.113		0.060		0.063		0.120		0.085
I	3472		0.654		0.639		0.930		0.663		0.629		0.713		0.063	
E	102	0.273		0.756		0.739		0.962		0.507		0.534		0.980		0.723
I	2998		0.758		0.740		0.928		0.572		0.543		0.826		0.073	
E	142	0.380		0.951		0.972		0.746		0.706		0.743		0.704		0.993
I	522		0.230		0.235		0.162		0.100		0.095		0.211		0.418	
E	88	0.235		0.652		0.638		0.830		0.438		0.461		0.880		0.624
I	4441		0.512		0.500		0.727		0.848		0.805		0.557		0.049	
E	220	0.588		0.614		0.627		0.482		0.914		0.868		0.455		0.641
I	151		0.066		0.068		0.047		0.029		0.027		0.061		0.693	
E	261	0.698		0.517		0.529		0.406		0.770		0.732		0.383		0.540
I	1242		0.547		0.560		0.385		0.237		0.225		0.502		0.176	
E	136	0.364		0.993		0.986		0.779		0.677		0.712		0.735		0.965
I	2202		0.969		0.992		0.682		0.420		0.399		0.890		0.099	
E	139	0.372		0.971		0.993		0.763		0.692		0.728		0.719		0.986
I	131		0.058		0.059		0.041		0.025		0.024		0.053		0.601	
E	25	0.067		0.185		0.181		0.236		0.124		0.131		0.250		0.177

Row: FUCA1 (X-axis), Column: ADH5 (Y-Axis), Length is in nucleotides, E = Exon; I = intron

Table 2. Distribution of block lengths (2 to >25) showing # of matched pairs identified; # of pairs showing same gene names and # of pairs showing different gene names

Lengths of Blocks	Matches with different gene names	Matches with same gene names	Total # of matches	Matches with same gene name (%)
2	1737702	334	1738036	1.922e-2
3	15575	805	16380	4.9
4	358	279	637	44.0
5	845	490	1335	36.7
6	142	265	407	65.1
7	159	508	667	76.2
8	51	169	220	76.8
9	95	307	402	76.4
10	20	130	150	86.7
11	50	401	451	88.9
12	21	127	148	85.8
13	37	252	289	87.2
14	8	127	135	94.1
15	30	256	286	89.5
16	6	130	136	95.6
17	21	236	257	91.8
18	8	121	129	93.8
19	17	172	189	91.0
20	2	75	77	97.4
21	17	115	132	87.1
22	3	81	84	96.4
23	7	153	160	95.6
24	1	70	71	98.6
>=25	36	1464	1500	97.6

Note: Block lengths >25 are clustered together.

Table 3. Distribution of block lengths in the sense and anti-sense strand

Lengths of blocks	Blocks starting with exon				Blocks starting with intron				Total	# of matches on Same Chr.	Matches on same chromosome (%)
	Com-Com	Join-Join	Com-Join	Total	Com-Com	Join-Join	Com-Join	Total			
2	216172	219236	434200	869608	216700	218332	433396	868428	1738036	92212	5.3
3	3442	3682	5855	12979	830	914	1657	3401	16380	2499	15.3
4	108	112	79	299	128	134	76	338	637	509	79.9
5	327	804	156	1287	12	18	18	48	1335	1273	95.4
6	106	78	28	212	69	103	23	195	407	379	93.1
7	225	361	50	636	14	9	8	31	667	637	95.5
8	54	40	8	102	42	58	18	118	220	201	91.4
9	197	152	38	387	6	8	1	15	402	376	93.5
10	29	41	9	79	38	31	2	71	150	145	96.7
11	202	219	18	439	10	2	0	12	451	447	99.1
12	33	37	6	76	32	31	9	72	148	148	100
13	138	129	11	278	6	4	1	11	289	287	99.3
14	44	38	2	84	28	22	1	51	135	135	100
15	141	120	16	277	3	6	0	9	286	286	100
16	28	29	4	61	49	26	0	75	136	135	99.3
17	161	83	11	255	2	0	0	2	257	256	99.6
18	19	31	2	52	50	25	2	77	129	129	100
19	87	89	7	183	5	1	0	6	189	188	99.5
20	22	16	0	38	25	13	1	39	77	77	100
21	69	53	6	128	1	3	0	4	132	132	100
22	18	25	1	44	18	20	2	40	84	84	100
23	84	70	1	155	1	4	0	5	160	159	99.4
24	12	8	0	20	26	25	0	51	71	71	100
>=25	578	665	6	1249	156	93	2	251	1500	1499	99.9
Total	222296	226118	440514	888928	218251	219882	435217	873350	1762278		

Com = complementary strand, Join = sense strand, Note: Block lengths >25 are clustered together. Matched pairs on the same chromosome are also shown

Table 4. Distribution of block lengths, number of matched pairs identified, number of unique blocks.

Lengths of blocks	# of unique length patterns	Total # of matched pairs	# of verified pairs	# of verified pairs/total # of matched pairs (%)
2	206756	1738036	339574	19.5
3	10417	16380	2958	18.1
4	335	637	34	5.3
5	548	1335	57	4.3
6	232	407	14	3.4
7	361	667	51	7.6
8	155	220	18	8.2
9	255	402	38	9.5
10	109	150	8	5.3
11	213	451	92	20.4
12	98	148	10	6.8
13	169	289	23	8.0
14	91	135	11	8.1
15	156	286	17	5.9
16	87	136	13	9.6
17	131	257	25	9.7
18	81	129	12	9.3
19	127	189	45	23.8
20	50	77	15	19.5
21	98	132	29	22.0
22	50	84	13	15.5
23	78	160	41	25.6
24	40	71	11	15.5
>25	682	1500	354	23.6

Percentage of blocks verified by Aceview , Note: Block lengths >25 are clustered together.

Table 5. Distribution of block lengths, number of block lengths covered in blocks of length 2 and number of new blocks of length 2 identified in blocks of length 3 to >25

Lengths of blocks	# of unique patterns with Blocks of length 2 (A)	# of patterns that are covered in blocks of length 2 (B)	# of patterns that are not covered in blocks of length 2	B/A (%)	# of verified patterns (C)	C/B (%)
3	16373	16089	284	98.3	10176	63.2
4	851	790	61	92.8	380	48.1
5	1849	1619	230	87.6	639	39.5
6	998	902	96	90.4	400	44.3
7	1977	1761	216	89.1	837	47.5
8	988	912	76	92.3	472	51.8
9	1928	1733	195	89.9	838	48.4
10	937	851	86	90.8	496	58.3
11	1989	1814	175	91.2	1027	56.6
12	971	879	92	90.5	434	49.4
13	1924	1761	163	91.5	871	49.5
14	1144	1071	73	93.6	552	51.5
15	2103	1903	200	90.5	1013	53.2
16	1262	1165	97	92.3	695	59.7
17	2002	1800	202	89.9	989	54.9
18	1312	1212	100	92.4	834	68.8
19	2105	1927	178	91.5	1073	55.7
20	886	824	62	93.0	499	60.6
21	1785	1640	145	91.9	899	54.8
22	1032	969	63	93.9	608	62.7
23	1636	1506	130	92.1	856	56.8
24	918	871	47	94.9	652	74.9
>25	19720	18285	1435	92.7	9937	54.3
Total	**66690**	**62284**	**4406**	**93.4**	**35177**	**56.5**

Percentage of blocks verified by Aceview are also shown, Note: Block lengths >25 are clustered together.

coding regions are not independent of non-coding regions. The view is further validated by the observations on function of introns (8). The data implies on the role of segmental gene duplications in genome evolution (9) and also confirms that gene duplication on the same chromosome is prevalent than inter-chromosomal gene duplication (10).

Because of several confounding aspects such as domain shuffling, the existence of isoforms derived from alternative splicing, and annotation errors in the databases detecting duplicate genes in a genome is not a simple task. To test if length patterns of blocks could help identify duplicate genes we explored the minimum number of blocks required for two genes to have the same gene name. It is interesting to see that genes that share blocks of length 10 or more have same gene name >85% of the time (Table 2; Figure 3). These results suggest that exon lengths and intron lengths play a significant role in gene assembly and function. A distribution profile on the number of matched pairs (total # of blocks identified) on each chromosome identified chromosome 16 and chromosome 19 with maximum number of hits (Figure 4). Both of these chromosomes are reported as duplication rich with highest number of segmental duplications (11-12). Chromosome 19 is also notable for the prevalence of duplication structures such as tandemly clustered gene structures. These findings are thus complementary to previous reports and present a new method for identification and detection of duplicate genes. On average a human gene usually has 10 exons (13). The presence of exon–intron pairs (blocks of length 2) of almost identical length but of dissimilar base sequence in numerous locations throughout the human genome supports the idea that introns are the result of propagation by replicative transposons or retrotransposons as suggested by Roger *et al*. (14). If this is the case they should be expected

Table 6. Distribution of block lengths, number of block lengths covered in blocks of length 3 and number of new blocks of length 3 identified in blocks of length 4 to >25

Lengths of blocks	# of unique patterns with lengths of block 3 (A)	# of patterns that are covered in blocks of length 3 (B)	# of patterns that are not covered in blocks of length 3	B/A (%)	# of verified patterns (C)	C/B (%)
4	578	105	473	18.2	33	31.4
5	1414	192	1222	13.6	62	32.3
6	808	81	727	10.0	24	29.6
7	1665	176	1489	10.6	72	40.9
8	854	60	794	9.2	20	33.3
9	1702	144	1558	8.5	61	42.4
10	835	47	788	5.6	21	44.7
11	1800	130	1670	7.2	69	53.1
12	890	66	824	7.4	24	36.4
13	1774	123	1651	6.9	47	38.2
14	1061	75	986	7.1	35	46.7
15	1964	123	1841	6.3	46	37.4
16	1181	68	1113	5.8	36	52.9
17	1883	105	1778	5.6	55	52.4
18	1237	77	1160	6.2	43	55.8
19	1994	133	1861	6.7	57	42.9
20	846	44	802	5.2	24	54.5
21	1711	111	1600	6.5	60	54.1
22	986	51	935	5.2	31	60.8
23	1572	86	1486	5.5	40	46.5
24	880	49	831	5.6	34	69.4
25	19911	1197	18714	6.0	658	55.0
Total	47546	3243	44303	6.8	1227	49.3

Percentage of blocks verified by Aceview are also shown. Note: Block lengths >25 are clustered together.

to maintain similar lengths, while diverging in their base sequence. The existence of such patterns suggest on the repeated replication of mobile elements within the genome.

4.2. From pieces to patterns and gene assemblage

The modular view of evolution suggests for the assembly of new genes from copies of pieces of various older genes, rapidly building new functions from a novel collection of already reliable parts. By analyzing the structure of genes at the genome level, it should be possible to discern the ancient blocks within. To test this hypothesis we derived a unique set of blocks of length 2 present in different blocks lengths (length 3 to length >25). The number of unique patterns identified for a block of length 2 is 206756 and for a block of length 3 is 10417 (Table 4). This indicates that blocks of length 2 provide maximum number of pieces for gene assemblage and probable functional diversity. We further proceeded with identification of unique length patterns of blocks of length 2 and 3 among the higher block lengths 3 to >25 and 4 to >25, respectively (Table 5 and Table 6). 16373 blocks of length 2 were identified in blocks of length 3, however, 16089 of them (~98%) are previously covered in blocks of length 2. So the new patterns of blocks of length 2 identified in block length 3 are 284 (Table 5). A complete analyses from block lengths 3 to >25 reveals that there are 62284 unique blocks of length 2 as the initial set and there are only 4406 new patterns of blocks of length 2 in blocks of length 3 or more. This suggests that >90% of the patterns contributed by blocks of length three or more are already represented in the initial pool of block of length two. These numbers support the idea of reuse of pieces and gene assemblage from a pool of building blocks. Furthermore, more than 40% of these blocks of length 2 in higher block lengths are verified in Aceview. This confirms for an initial unique pool of smaller blocks or pieces (blocks of length 2) that are probably the building blocks for genes. These pieces could have perhaps been recruited in a combinatorial fashion to mix and match and give rise to new genes and build up the lengthy chains that make longer genes with novel functions. To rule out the possibilty of bigger blocks as initial building blocks, we repeated the analysis for blocks of length 3. It was interesting to see that only ~5-20% of the patterns contributed by blocks of length four or more are already represented in the initial pool of block of length three, thereby re-confirming our findings (Table 6). These results show that duplication has played a vital role in the evolution of new gene functions and evolution is a tinkerer that re-recruits pieces at its disposal rather than repeatedly starting from scratch.

5. CONCLUSION

The exon-intron length patterns of annotated human genes were examined and they revealed common structural patterns (1762278 conserved arrangements of exon-intron length) within the human genome. Exploration of the blocks of variant length shows interesting data highlighting their occurrence by retroposition and suggesting their possible role in gene assemblage. The results support the process of creating new combinations of exons by recruitment of pieces of exons/introns and creation of present-day gene architecture in the human genome. We conclude that the modular protein evolution by exon-shuffling and retroposition are the probable processes that have contributed significantly to human genome design.

6. CAVEATS

As this analysis is strictly based on CDS feature in genome data, it does not take into account the first exon and is biased towards internal coding exons of the gene. Nonetheless, this analysis hints at the possible role of non-coding DNA in genome architecture and design and provides a platform for understanding the human genome and issues in gene evolution. This stringent criterion of

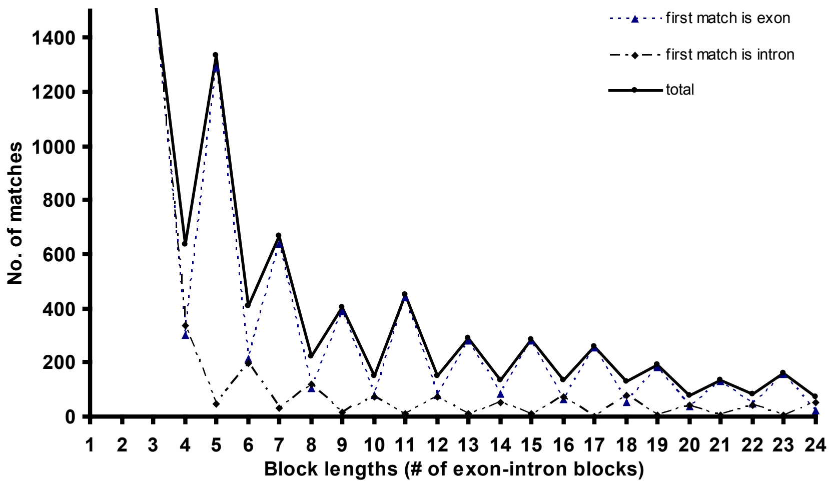

Figure 1. Distribution of block lengths (exon-intron and intron-exon). It is clear that odd length blocks are more than even length blocks and matched pairs with exons as start are more prevalent.

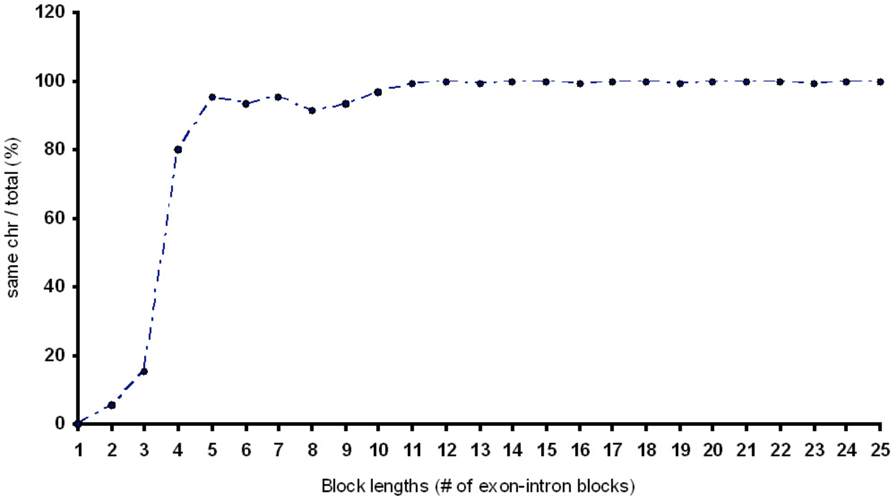

Figure 2. Percentage of block length matches on same chromosome.

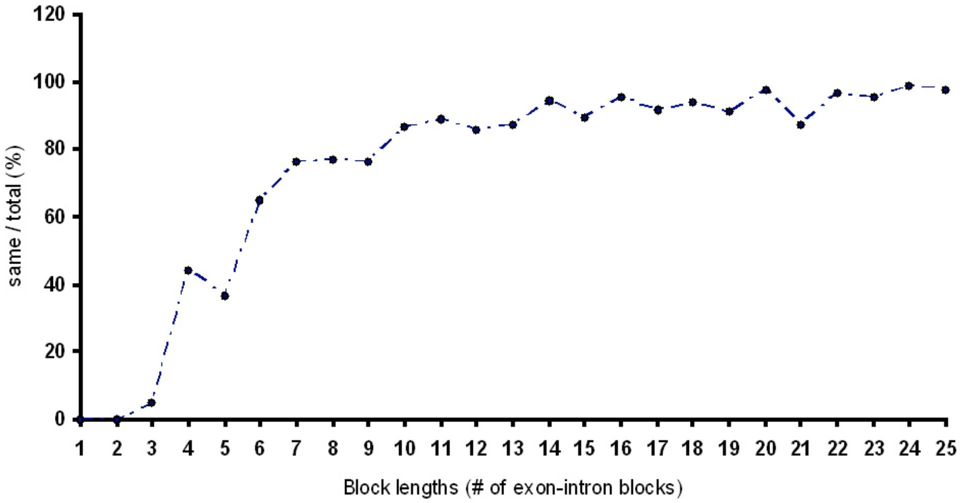

Figure 3. Percentage of block length matches with same gene name.

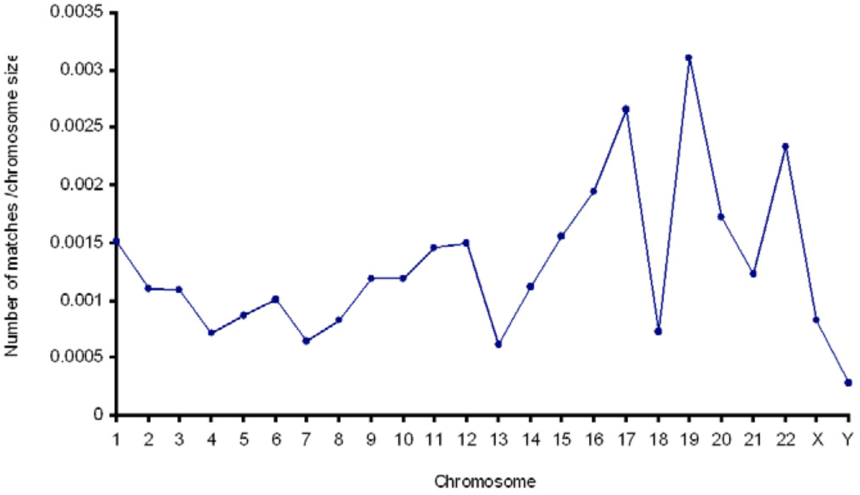

Figure 4. Percentage of block length matches on each chromosome.

99% length match excluded many interesting patterns but also ensured that the remaining ones were genuine and potentially biologically significant.

7. ACKNOWLEDGEMENTS

This manuscript is based on the project supported by A*STAR - BMRC grant #03/1/22/19/242.

8. REFERENCES

1. Makalowski W: Not Junk After All. *Science* 300, 1246-1247 (2003)

2. 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, 860-921 (2001)

3. Croft L, Schandorff S, Clark F, Burrage K, Arctander P, & Mattick JS: ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet*. 24, 340-341 (2000)

4. Hawkins JD: A survey on intron and exon lengths. *Nucleic Acids Res*. 16, 9893-9908 (1988)

5. Hill A, & Sorscher E: Common structural patterns in human genes. *Bioinformatics* 20, 1632-1635 (2004)

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

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

8. Mattick,J. & Gagen, M: The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNA's in the development of complex organisms. *Mol. Biol. Evol*. 18, 1611–1630 (2001)

9. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW & Eichler EE: Recent segmental duplications in the human genome. *Science*. 297,1003-1007 (2002)

10. Zhang L, Lu HH, Chung WY, Yang J & Li WH.: Patterns of segmental duplication in the human genome. *Mol Biol Evol.* 22, 135-141 (2005)

11. Martin J, Han C, Gordon LA, Terry A, Prabhakar S, She X, Xie G, Hellsten U, Chan YM, Altherr M, Couronne O, Aerts A, Bajorek E, Black S, Blumer H, Branscomb E, Brown NC, Bruno WJ, Buckingham JM, Callen DF, Campbell CS, Campbell ML, Campbell EW, Caoile C, Challacombe JF, Chasteen LA, Chertkov O, Chi HC, Christensen M, Clark LM, Cohn JD, Denys M, Detter JC, Dickson M, Dimitrijevic-Bussod M, Escobar J, Fawcett JJ, Flowers D, Fotopulos D, Glavina T, Gomez M, Gonzales E, Goodstein D, Goodwin LA, Grady DL, Grigoriev I, Groza M, Hammon N, Hawkins T, Haydu L, Hildebrand CE, Huang W, Israni S, Jett J, Jewett PB, Kadner K, Kimball H, Kobayashi A, Krawczyk MC, Leyba T, Longmire JL, Lopez F, Lou Y, Lowry S, Ludeman T, Manohar CF, Mark GA, McMurray KL, Meincke LJ, Morgan J, Moyzis RK, Mundt MO, Munk AC, Nandkeshwar RD, Pitluck S, Pollard M, Predki P, Parson-Quintana B, Ramirez L, Rash S, Retterer J, Ricke DO, Robinson DL, Rodriguez A, Salamov A, Saunders EH, Scott D, Shough T, Stallings RL, Stalvey M, Sutherland RD, Tapia R, Tesmer JG, Thayer N, Thompson LS, Tice H, Torney DC, Tran-Gyamfi M, Tsai M, Ulanovsky LE, Ustaszewska A, Vo N, White PS, Williams AL, Wills PL, Wu JR, Wu K, Yang J, Dejong P, Bruce D, Doggett NA, Deaven L, Schmutz J, Grimwood J, Richardson P, Rokhsar DS, Eichler EE, Gilna P, Lucas SM, Myers RM, Rubin EM & Pennacchio LA: The sequence and analysis of duplication-rich human chromosome 16. *Nature*. 432, 988-994 (2004)

12. Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, Hellsten U, Goodstein D, Couronne O, Tran-Gyamfi M, Aerts A, Altherr M, Ashworth L, Bajorek E, Black S, Branscomb E, Caenepeel S, Carrano A, Caoile C, Chan YM, Christensen M, Cleland CA, Copeland A, Dalin E, Dehal P, Denys M, Detter JC, Escobar J, Flowers D, Fotopulos D, Garcia C, Georgescu AM, Glavina T, Gomez M, Gonzales E, Groza M, Hammon N, Hawkins T, Haydu L, Ho I, Huang W, Israni S, Jett J, Kadner K, Kimball H, Kobayashi A, Larionov V, Leem SH, Lopez F, Lou Y, Lowry S, Malfatti S, Martinez D, McCready P, Medina C, Morgan J, Nelson K, Nolan M, Ovcharenko I, Pitluck S, Pollard M, Popkie AP, Predki P, Quan G, Ramirez L, Rash S, Retterer J, Rodriguez A, Rogers S, Salamov A, Salazar A, She X, Smith D, Slezak T, Solovyev V, Thayer N, Tice H, Tsai M, Ustaszewska A, Vo N, Wagner M, Wheeler J, Wu K, Xie G, Yang J, Dubchak I, Furey TS, DeJong P, Dickson M, Gordon D, Eichler EE, Pennacchio LA, Richardson P, Stubbs L, Rokhsar DS, Myers RM, Rubin EM & Lucas SM: The DNA sequence and biology of human chromosome 19. *Nature*. 428, 529-535 (2004)

13. Sakharkar MK, Chow VT & Kangueane P: Distributions of exons and introns in the human genome. *In Silico Biol*. 4, 387-393 (2004)

14. Roger,A.J., Keeling,P.J. & Doolittle,W.F: Introns, the broken transposons. *Soc. Gen. Physiol. Ser.*, 49, 27–37 (1994)

Key Words: Exon-Intron Length Patterns, Human Genome, Block Length, Duplication, Genome Design, Gene Architecture

Send correspondence to: Meena Kishore Sakharkar Ph.D., Nanyang Technological University, Singapore 639798; Tel: 65 6 790 5836; Fax: 65 6 774 4340; E-mail: mmeena@ntu.edu.sg

http://www.bioscience.org/current/vol10.htm