Evolutionary analyses of the human genome

Wen-Hsiung Li, Zhenglong Gu, Haidong Wang & Anton Nekrutenko

Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA

The completion of the human genome will greatly accelerate the development of a new branch of science—evolutionary genomics. We can now directly address important questions about the evolutionary history of human genes and their regulatory sequences. Computational analyses of the human genome will reveal the number of genes and repetitive elements, the extent of gene duplication and compositional heterogeneity in the human genome, and the extent of domain shuffling and domain sharing among proteins. Here we present some first glimpses of these features.

We have analysed the draft human genome sequence for data related to evolutionary genomics. Our investigations reveal new information about repetitive elements, domain sharing and conservation, and gene duplication in the human genome (for Methods, see Supplementary Information).

Numbers of repetitive elements

Analysing 76% of the human genome (using almost all available contigs, Table 1), we estimated that around 43% of the human genome is occupied by four major classes of interspersed repetitive element: (1) short interspersed elements (SINEs), (2) long interspersed elements (LINEs), (3) elements with long terminal repeats (LTR elements), and (4) DNA transposons. There are more than 4.3 million repetitive elements in the human genome, with *Alu* and LINE1 (L1) being the most frequent. These estimates largely agree with previous ones[1,2]. As many repetitive elements would have degenerated to the extent that they cannot be detected by the computer program RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker), more than 50% of the human genome would have come from insertion of repetitive elements.

Repetitive elements in proteins

Contrary to the belief that a repetitive element insertion into a gene is deleterious and unlikely to survive, many translated repetitive elements are found in proteins (Table 2). From the International Protein Index (ref. 3; http://www.ensemble.org/IPI), we derived a new database by eliminating 'isoforms' (due to alternative splicing). We set the expected (E) value $< 10^{-80}$ in BLASTing (using tBLASTN) the database by itself and deleted all but one copy of the genes whose chromosome locations overlapped by more than 50%. This procedure reduced the number of 'proteins' in the database from 45,112 to 43,195, of which 15,337 are 'known' proteins and 27,858 are predicted proteins (translated from gene predictions). Because of the stringent conditions used, the chance of misidentifying isoforms is negligible. The new database probably still contains some isoforms because the chromosomal locations of many sequences are unknown and their isoforms cannot be identified.

We then BLASTed each sequence in the new database against a recent release of RepBase (www.girinst.org). Predicted proteins on average contain many more matches to repetitive element fragments than 'known' proteins (Table 2), suggesting many false positives in gene prediction. This is not a serious problem for 'known' proteins, as they have been translated from genes cloned by traditional methods or from 'genes' that have a high similarity to known genes. Surprisingly, many 'known' proteins also contain (truncated) repetitive elements, especially L1 and Alu. A closer look suggests that repetitive elements were usually not inserted into the original open reading frames, but became part of a gene because of alternative splicing, which can sometimes extend or truncate the coding region. L1 has on average the highest E-scoring matches (Table 2), indicating that L1-mediated gene evolution may be common. In addition, there is evidence that transduction of 3′-flanking sequences (including exons) is common in L1 retrotransposition[4], so that L1 might have mediated many exon-shuffling events. Therefore, repetitive elements may have been significant in gene evolution and species differentiation.

To reduce the effect of false gene predictions, we deleted from the database 2,615 predicted proteins that had a significant hit ($E < 10^{-4}$) by a repetitive element and did not have a domain structure other than reverse transcriptase or transposase. The 'cleaned' database contains 15,337 'known' proteins and 25,243 predicted proteins (total, 40,580).

Domain sharing and conservation

A domain is a structural or functional unit in a protein. To investigate the frequency of domain sharing, where the same domain appears in different proteins, we obtained a collection of human, fruitfly, nematode and yeast proteins (15,312, 8,896, 9,254 and 3,136 polypeptides) containing at least one domain; we used the InterPro domain database. In each case of nested domains, only the shortest one was included in the final dataset. There are 1,865, 1,218, 1,183 and 973 domain types in human, fruitfly, nematode and yeast, respectively, and the proportions of mosaic proteins (containing

Table 1 Repetitive elements in the human genome

Type	No. found	Estimated no. in genome	Per cent of genome
SINE (all)	1,404,300	1,841,000	12.5
Alu	1,010,400	1,324,600	10.7
LINE (all)	1,045,800	1,371,100	18.9
L1	661,000	866,600	15.4
DNA	308,800	404,900	2.7
LTR	531,900	697,300	7.9
Other	7,300	9,600	0.1
Total	3,959,200	4,323,900	42.5

The numbers were obtained by using RepeatMasker to mask all contigs assigned to chromosomes. The sequence database used was the 17 July 2000 freeze. Sequencing gaps were removed. Total length of analysed sequences (2,440,850,649 bp) is ~76% of the human genome. Per cent of genome means the estimated proportion of the human genome occupied by the repetitive elements under consideration.

Table 2 Repetitive elements in 'known' and predicted proteins

Proteins	$E \le 10^{-50}$		$10^{-50} < E \le 10^{-10}$		$10^{-10} < E \le 10^{-4}$		$10^{-4} < E \le 10^{-2}$	
	n	Top hit	n	Top hit	n	Top hit	n	Top hit
'Real'	64	Line1	127	Alu	162	Alu	304	Line1
Predicted	870	Line1	1,519	Line1	1,232	Alu	997	Alu
All	1,034	Line1	1,646	Line1	1,394	Alu	1,301	Alu

n, number of cases. 'Known' proteins are translated from genes that have been cloned by traditional methods or have high similarities to known genes. Predicted proteins are translated from genes predicted from genomic sequences using computer programs.

more than one domain type) in the four taxa are 28%, 27%, 21% and 19%.

First, we consider the sharing of domain types (or domain combination), regardless of the order or the number of times a domain appears in a protein; for example, a protein with A-A-B-B-A contains only two domain types and has the combination AB. In our database, the largest number of domain types per protein is nine in human and fruitfly, and seven in nematode and yeast. The frequency of domain sharing is very high among human proteins (Table 3); for example, there are 88 cases where three proteins share two types of domain. There are also many human proteins that share more than one type of domain with *Drosophila*, (slightly less frequently) with *C. elegans*, and (much less frequently) with yeast proteins. But there are only three cases where a combination of more than three domain types is shared by human and yeast proteins and only two of these cases are shared by the four taxa. One of these two cases has a combination of seven domain types; it occurs once in human, nematode and yeast but twice in the fruitfly. It is a carbamoyl-phosphate synthase (EC 6.3.5.5) involved in the first three steps of *de novo* pyrimidine nucleotide biosynthesis (SwissProt accession nos P07259, Q18990, Q9VXD5, P27708).

We now consider the conservation of domain arrangements (the number and order of domains within a protein). There are 3,433, 1,702, 1,248 and 470 distinct arrangements of two or more domains in human, fruitfly, nematode and yeast proteins, respectively. Some proteins exhibit extensive domain repetition: in human, the largest number of domain types in a protein is nine, but the largest total number of domains in a protein is 130. Many human proteins have identical arrangements (Table 3). In the case of two domain types, many of the human arrangements are shared by fruitfly, (less frequently) by nematode, and (even less frequently) by yeast. The largest domain arrangement shared by all four taxa contains 11 domains, with only two domain types. It is ornithine decarboxylase, which catalyses a rate-limiting step in the biosynthesis of polyamines (SwissProt: P08432, Q94278, Q9V352, P11926). The shared arrangement that has the largest number of domain types (four) contains five domains; it is a sulphonylurea receptor (SURx) in fruitfly, which is a subunit of the ATP-sensitive potassium channel (SwissProt: P53049, Q9U6Z2, Q9V352).

Duplicate genes

Two genes that were derived from a gene duplication are said to be paralogous; two genes (in two species) are orthologous if they were derived from the same gene through speciation. Predicting whether two proteins are paralogous is relatively simple when their sequence identity (I) is high (>40% for long sequences) but becomes difficult when I is in the medium range (20–35%) or lower, especially for short sequences.

Rost[5] proposed an empirical formula for clustering proteins in a database (Table 4). Two proteins are assumed to be paralogous if the proportion (p) of identical residues over the L aligned amino-acid residues between the two proteins is higher than the cut-off point (p^I) defined by the formula. The cut-off point increases as L decreases because two unrelated short sequences may by chance have a high p value. A common practice in clustering proteins into groups is to use single linkage: if proteins A and B have a p higher than p^I and so do proteins B and C, then A, B and C are clustered in the same group, even if the p value for A and C does not meet the cut. Applying Rost's formula with $n = 5$ (n is a factor to raise the cut-off point) to the 'cleaned' protein database, we found that the largest group contained 15,121 members, which is more than one-third of the database and includes various proteins. Even for $n = 25$ the largest group still contained 4,519 members. Such large groups

Table 3 Domain sharing and order conservation within human and between human and other eukaryotes

Human versus	Domain sharing					Identical domain arrangements			
	No. of proteins sharing domains	No. of cases				Total no. of domains in a protein	No. of identical arrangements/no. of human proteins		
		No. of domain types					No of domain types*		
		1	2	3	>3		2	3	>3
Human	2	214	194	73	61	1	-	-	-
	3	147	88	25	18	2	141/556	-	-
	4	123	38	17	5	3	57/208	21/62	-
	5	67	17	5	3	4	53/168	18/63	4/10
	6	56	19	5	0	5	44/173	11/27	5/16
	>6	377	79	20	5	>5	150/605	66/172	34/78
Fly	1	143	129	32	23	1	-	-	-
	2	134	65	14	12	2	119/337	-	-
	3	97	47	11	5	3	35/98	10/18	-
	4	83	19	7	0	4	28/65	10/24	1/1
	5	51	9	2	2	5	25/74	8/17	5/13
	>5	359	65	14	2	>5	58/137	11/19	12/16
Worm	1	136	92	27	9	1	-	-	-
	2	124	56	11	12	2	89/307	-	-
	3	94	38	9	7	3	28/118	10/24	-
	4	84	17	5	2	4	16/39	6/20	0/0
	5	46	8	2	2	5	16/60	3/8	3/6
	>5	355	61	11	1	>5	43/118	8/16	9/13
Yeast	1	135	51	8	2	1	-	-	-
	2	91	27	5	0	2	51/199	-	-
	3	64	18	2	0	3	9/20	4/12	-
	4	58	5	0	0	4	4/7	3/3	0/0
	5	41	3	0	0	5	3/6	1/2	1/1
	>5	260	24	4	1	>5	36/16	1/3	0/0
Fly, worm and yeast	1	75	24	4	1	1	-	-	-
	2	78	16	3	0	2	26/145	-	-
	3	49	12	1	0	3	**4/10**	1/3	-
	4	48	3	0	0	4	3/5	0/0	0/0
	5	33	2	0	0	5	3/6	0/0	1/1
	>5	249	21	3	1	>5	7/18	0/0	0/0

*Number of unique domain arrangements/number of human proteins in which these arrangements are found. The second number is larger than the first because many proteins may share the same arrangement. For example, in the case 4/10 (bold numbers) there are four unique arrangements of three-domain proteins with two domain types (for example, the arrangement A-B-A has three domains but only two domain types: A and B) that have been conserved among human, fly, worm, and yeast; in human there are ten such proteins.

Table 4 Protein groups inferred from sequence similarities

Group size	$I' \geq 50\%$*	$I' \geq 40\%$†	$I' \geq 30\%$‡
	No. of groups	No. of groups	No. of groups
1	31,515	28,251	25,237
2	2,041	2,343	2,288
3–5	807	1,069	1,298
6–10	104	170	262
11–20	36	57	86
21–50	14	26	38
51	1		2
69		1	1
71			1
104	1		
122			1
124	1		
129	1		
132		1	
133		1	1
139	1		
221		1	
232		1	
265			1
292			1
331			1
358		1	
479			1

Total number of 'proteins': 40,580 (15,337 'known' proteins and 25,243 predicted proteins). All comparisons with $L \leq 20$ were excluded.
*$I' \geq 50\%$ for $L > 40$ and $I' \geq p^I$ for $L > 40$, where p^I is given by Frost's formula[4] $p^I = 0.01n + 4.8L^{(-0.32(1+\exp(-L/1,000)))}$. For $n = 0$, $p^I = 72\%$, 41%, 28% and 24% for $L = 20$, 50, 100 and 150, respectively.
†$I' \geq 40\%$ for $L > 70$ and $I' \geq p^I$ for $L > 70$.
‡ $I' \geq 30\%$ for $L > 150$ and $I' \geq p^I$ for $L > 150$.

occur probably because nonhomologous proteins may share the same domains (see above).

We propose to use $I' = I \times \mathrm{Min}(n_1/L_1, n_2/L_2)$, where I is the proportion of identical amino acids in the aligned region (including gaps) between the query (sequence 1) and target (sequence 2) sequences obtained by the alignment program FASTA, L_i is the length of sequence i, and n_i is the number of amino acids in the aligned region in sequence i. The factor $\mathrm{Min}(n_1/L_1, n_2/L_2)$, which means the smaller of n_1/L_1 and n_2/L_2, takes care of the situation where a high I value is obtained when a short protein shares one or more domains with a longer protein. Another difference between I' and p^I is that I' imposes a gap penalty in the aligned region. For short proteins, however, I' may become high by chance and so we impose $I' \geq p^I$ with $n = 5$.

Table 4 shows the protein groups inferred from our formula. $I' \geq 50\%$ corresponds to Dayhoff's definition of protein families. The largest group (139 members) contains the L1 reverse transcriptase (RT) and sequences with high I' values with L1 RT. This is surprising, but many 'known' and predicted proteins contain (truncated) L1 RT; note also that many L1 RTs may still be nearly intact in the human genome. The second largest group (129 members) contains 91 immunoglobulin heavy chains, 1 rheumatoid factor, 6 unnamed proteins and 31 predicted proteins; the third (124 members) contains 85 immunoglobulin light chains, 2 heavy chains, 1 microfibrillar protein, 2 unnamed proteins and 34 predicted proteins; the fourth (104 members) contains 38 zinc finger proteins, 6 unnamed proteins and 60 predicted proteins; and the fifth (51 members) contains 16 olfactory receptors and 35 predicted proteins. This criterion identifies 3,007 families, 2,041 of which are two-protein families. These should be taken as minimum estimates because many human genes remain unidentified. For $I' \geq 40\%$, the zinc finger group becomes the largest, and the L1 RT and olfactory receptor groups become the second and third largest. For $I' \geq 30\%$, the five largest groups are zinc finger proteins, olfactory receptors, immunoglobulins (both light and heavy chains), L1 RTs and keratins. For $I' \geq 25\%$, some of the largest groups become very heterogeneous, indicating that at this level of similarity it requires a more rigorous analysis to determine whether two proteins are related.

The $I' \geq 30\%$ criterion identifies 3,982 superfamilies (Table 4). Although some of the groupings may be false positives, this number may represent a minimum estimate because many human genes remain unidentified and because many of the proteins in the 'singleton' groups (25,237) may actually be related to each other. Taking the data at face value, the proportion of 'singleton' groups is 25,237/40,580 = 62% of the total 'proteins' in our 'cleaned' database. This may be an overestimate, but should be taken cautiously because many of the 'singletons' may be false positive and because the total number of human genes remains unknown.

Our analysis has provided some insights into the evolutionary genomics of the human genome. There are many repetitive elements in our genome (Table 1), and they may have been very important in the evolution of mammalian proteins (Table 2). Domain sharing is common among proteins, and many domain arrangements have been conserved (Table 3). But many challenges remain. For example, as the number of human genes is still unknown, it remains unclear how many human genes exist as single copies. Reliable annotation of the human genome and clean databases of human genes and proteins are required for a rigorous analysis. In addition, better tools are needed for analysis. Single linkage does not seem appropriate for clustering proteins. Finally, better methods are needed for deciding whether two proteins are homologous, especially for short proteins. □

1. Smit, A. F. A. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9,** 657–663 (1999).
2. Gu, Z., Wang, H., Nekrutenko, A. & Li, W.-H. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequences. *Gene* **259,** 81–88 (2000).
3. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
4. Goodier, J. L., Ostertag, E. M., Kazazian, H. H. Jr Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9,** 653–657 (2000).
5. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12,** 85–94 (1999).

Supplementary information is available from *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank R. Stevens for letting us use Argonne computers, E. Birney for help and NIH for research support.

Correspondence should be addressed to W.-H.L. (e-mail: whli@uchicago.edu).