Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Theoretical Biology 240 (2006) 200–208

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

Genome-wide analysis of mammalian DNA segment fusion/fission

Zhihua Zhang[a,1], Hong Sun[a,1], Yong Zhang[a], Yi Zhao[b], Baochen Shi[a], Shiwei Sun[b], Hongchao Lu[b], Dongbo Bu[b], Lunjiang Ling[b], Runsheng Chen[a,b,*]

[a] *Institute of Biophysics Academia sinica, Chinese Academy of Sciences, Beijing, 100101 China*
[b] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080 China*

Received 3 January 2005; received in revised form 9 September 2005; accepted 12 September 2005
Available online 14 November 2005

Abstract

As a powerful tool for gene function prediction, gene fusion has been widely studied in prokaryotes and certain groups of eukaryotes, but it has been little applied in studies of mammalian genomes. With the first fully sequenced mammalian genomes (human, mouse, rat) now available, we defined and collected a set of fusion/fission event-linked segments (FFLS) based on structured organized genomic alignment. The statistics of the sequence features highlighted the FFLSs against their random context. We found that there are three groups of FFLSs with different component pairs (i.e. gene–gene, gene–noncoding and noncoding–noncoding) in all three mammalian genomes. The proteins encoded by the components of FFLSs in the first group shown a strong tendency to interact with each other. The segmental components in the last two groups which did not contain any protein-coding genes, were found not only to be transcribed to some level, but also more conserved than the random background. Thus, these segments are possibly carrying certain biologically functional elements. We propose that FFLS may be a potential tool for prediction and analysis of function and functional interaction of genetic elements, including both genes and noncoding elements, in mammalian genomes. The full list of the FFLSs in the genomes of the three mammals is available as supporting information at doi:10.1016/j.jtbi.2005.09.016.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: Fusion; Fission; Mammalian comparative genomic; Non-protein coding DNA segment

1. Introduction

Completion of the genomic sequencing is not the end of understanding the human genome, as we are now facing the challenge of deciphering the various genetic components of the genome (Consortium, 2004a; Lander et al., 2001). Since Marcotte et al. (1999) and Enright et al. (1999) introduced fusion analysis to study protein–protein interaction (PPI) and predict protein function, the method has played an increasingly important role in genome studies. Yanai et al. (2001), followed by Enright and Ouzounis (2001), applied the method on 30 complete genomes to obtain indications of the functions of uncharacterized genes. Yanai et al. (2002) also studied the evolution of gene fusion by phylogenetic analysis of domains involved in such events.

However, little genome-wide gene fusion/fission analysis has been done in mammals. This is partly due to the highly complicated structure of mammalian genome. As much as 15% of human genes are duplicated, with segment duplications covering 5.2% of the genome (Bailey et al., 2002; Li et al., 2001). This high degree of duplication in addition to other genomic rearrangements makes it very difficult to detect the orthologous genes from paralogous genes and pseudogenes. Recent studies have nevertheless demonstrated that gene order in eukaryotes, especially in mammals is not random (Hurst et al., 2004). Experimental evidence suggests that coordinately expressed or, functionally related genes are likely to cluster on the chromosomes (Hurst et al., 2004). To some degree, a fusion/fission gene

Abbreviations: FFLS, fusion/fission event-linked segment; TLSC, top level segmental chain; PPI, protein–protein interaction.

*Corresponding author at: 100101, Bioinfomatics Laboratory, Institute of Biophysics Academia sinica, Datun Road 15#, Beijing, PR China. Tel.: 086 10 64888546; fax: 086 10 64877837.

E-mail addresses: biozhang2001@yahoo.com.cn (Z. Zhang), shpolor@yahoo.com.cn (H. Sun), crs@sun5.ibp.ac.cn (R. Chen).

[1] Joint first authors.

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2005.09.016

may be represented as an overlapping gene. A number of instances of overlapping genes in eukaryotes have been reported (Bachman et al., 1999; Williams and Fried, 1986), and Veeramachaneni et al. (2004) recently produced an genome-wide distribution of overlapping genes in human and mouse. However, all previous studies have focused on coding genes, which cover only 1–2% of the euchromatic genome (Consortium, 2004a; Lander et al., 2001). Several studies have shown that many intergenic regions may also be transcribed (Balakirev and Ayala, 2003; Bertone et al., 2004), and pseudogenes, previously regarded as functionless, have been shown to carry out certain functions in mammals (Balakirev and Ayala, 2003). To investigate and understand the non-coding region of genome is a considerable challenge, and analysis of fusion/fission events may give the scientist an additional handle on those challenges. To this end, a well-defined segmental orthology between mammalian DNA sequences is required, which should be organized local with alignments which maintain a synteny with gaps that are acceptable in both query and target sequences. W.J. Kent et al. (Kent et al., 2003) introduced a program called AXTCHAIN to do so. AXTCHAIN detects pairwise nucleotide alignments and chains them into longer spans of gapped alignments. The chained alignments are further organized into a hierarchic structure called a "net". Chain and net let us define orthologous segments between mammalian genomes more clearly than in the past (Kent et al., 2003). In this work, by using the top level chain as well defined orthologous segments pairs, we define an event called "segments fusion/fission linking" between highly conserved genomic segments in the three complete mammalian genomes (human, mouse, and rat). We show that (i) there is a set of genomic segments that is possibly related to fusion/fission events in the three mammalian genomes, and that stands out from the sequence background with respect to several statistical criteria; (ii) for the FFLS where the two components contain protein-coding genes, we have shown that the two encoded proteins tend to interact, and; (iii) conservation and transcriptional features of components with non-protein coding segments suggested that they may carry biologically functional elements.

2. Materials and methods

2.1. Functional elements

The list of overlapping gene was copied from the website (http://posnania.cbio.psu.edu/research/overlapping_genes.html) (Veeramachaneni et al., 2004). We obtained the chromosome coordinates of CpG islands, known gene exons, pseudo genes (Zhang et al., 2003b), KEGG pathway data (Kanehisa et al., 2004) and the regions matching with EST from the UCSC Genome Browser (University of California, Santa Cruz; Kent et al., 2002). High-density oligonucleotide tiling microarrays data were used for screening transcribed regions (Bertone et al., 2004; Cheng et al., 2005).

PPI: There are not high throughput PPI data available in mammalians yet. We downloaded experimental proven interaction data from BIND (Bader et al., 2003), DIP (Salwinski et al., 2004), MIPS (Pagel et al., 2005), MINT (Zanzoni et al., 2002) and IntAct (Hermjakob et al., 2004). A recent computational prediction of PPIs in human (Lehner and Fraser, 2004) and two-hybrid screening to map Smad signaling PPIs (Colland et al., 2004) have also been included in this analysis.

2.2. The genomic alignments among the mammalian genomes

We downloaded the three assembled and annotated mammal genomic drafts, human (NCBI Build 35, May 2004) (Lander et al., 2001), mouse (NCBI Build 33, May 2004) (Consortium, 2002) and rat (Baylor HGSC v. 3.1, June 2003) (Consortium, 2004b) from the GoldenPath (ftp://genome.cse.ucsc.edu/goldenPath/). "Net" and chained alignments among the three mammalian genomes were also downloaded from the GoldenPath. In a pair-wise genomic comparison, for example human vs. rat, we say the human genome is the target genome if it is set as out-group, and the rat genome will then be a query.

2.3. Assigning segmental orthology

Instead of analysing protein-coding genes only, we retrieved segments irrespective of protein-coding sequence content. For that purpose, we used chained and netted BLASTZ alignments from human, mouse and rat genome comparisons (Kent et al., 2003) as the segment pool. A chain is a series of associated BLASTZ matches ("blocks") satisfying the condition that the order of blocks is the same in both species. As there may exist gaps in both chained alignments, the chain data has been further organized into nets. The segments organized at the top level of the net are more likely to be orthologous, since these are the regions covered by the highest-scoring chains, and not just selected by simple sequence similarity. In AXTCHAIN, transposon repeats and simple repeats of period 12 or less, recognizable by REPEATMASKER (Smit, 1999) and TANDEM REPEAT FINDER (Benson, 1999), were masked out from the original genome sequences. However, not all repetitive elements are well-recognized by these programs, especially in rodent genomes (Consortium, 2002, 2004b), and in addition, there are huge numbers of duplicated segments in mammalians genomes, which makes the job of defining orthologous segments very difficult. In this work, we defined segmental chains at the top level of the nets as orthologous segments. These top-level segmental chains (TLSC) are the basic units in this study. A TLSC contains two segmental chains, one at the query genome (qSC) and the other at the target genome (tSC). In other words, two chained BLASTZ aligned segments were regarded as orthologous if and only if the tSC is found at the top level of the net in the target genome. This definition of orthologous segments takes into account both sequence similarity and the order of blocks within the chain (Kent et al., 2003).

2.4. Identifying fusion/fission links

As pointed out above, we considered not only protein-coding regions, but also intronic and intergenic non-coding regions in the three mammalian genomes. Let *A* and *B* be two tandem TLSCs in the target genome containing the chained segments tA and qA, tB and qB, respectively (Fig. 1). We regard the two tandem TLSCs, *A* and *B*, as potentially "fusion/fission linked" if they satisfy the following three criteria: (i) the two target-chained segments tA and tB should be separated by a gap not exceeding 10 bp (ii) the two query chained segments qA and qB should not overlap, and qA and qB should be sufficiently separated (i.e. they are either separated by no less than 1000 bp, or located in different "syntenic" block (annotated by NETFILTER) (Kent et al., 2003) or on different chromosomes); (iii) the total number of unsequenced bases plus the total number of bases masked as repeats by the two softwares mentioned above, should not exceed 50% of the total segment length of tA, qA, tB and qB. The thresholds of 10 and 1000 bp were based on the distribution of gap sizes in the chain-covered genomes (see supplementary material). TLSCs with spans longer than 100 kb or shorter than 50 bp were also excluded from our FFLS detection procedure, because the longer chains were either large primary synteny units or large scale inversions (Kent et al., 2003). There are also substantial amounts of interspersed and tandem segmental duplications in the mammalian genomes (Bailey et al., 2002; Consortium, 2002, 2004b; Lander et al., 2001). To handle this issue, we limited our study to the top-level chains. Although there are much more chains located at the lower levels, the short chains at the lower levels are often processed pseudogenes (Kent et al., 2003). At this early stage of genome-wide analysis of segmental fusion/fission linkage, we therefore decided to ignore the lower level chains. The tandem TLSCs that passed all the above criteria were collected as "fusion/fission event-linked segments (FFLS)".

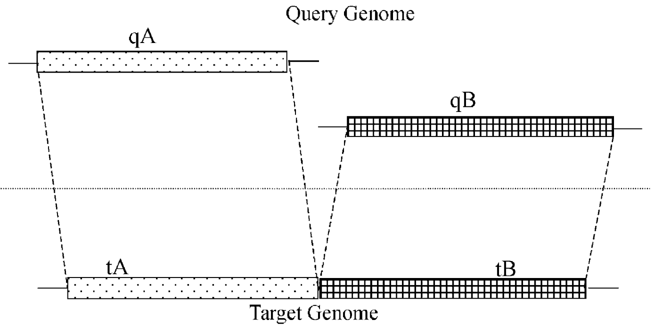

Fig. 1. Visualization of a fusion/fission event-linked segment (FFLS). The chained blocks are represented as bars. To indicate that the two query chains are not necessarily located on the same chromosomes, they have not been aligned in the same line.

2.5. Random sampling

As a negative control for our results, we used randomly sampled chains. We did not sample the whole length chains, because all the chains (the chained blocks) were separated by blocks at different levels. Instead, we randomly selected successively chained blocks across the whole net of genome alignments. The number of selected chains was equal to the number of FFLSs detected in each chromosome. To evaluate the statistical significance of the features of the detected FFLSs, each analysis was repeated 1000 times with independent, randomly sampled data sets. The fraction of times in which the random sample set average scored higher (or lower) than the average of the FFLS provided the basis for the statistical significance. All the analyses on the random data sets were normalized by the average size difference between the detected FFLSs and the random data sets (see Results).

3. Results

Using organized BLASTZ alignments among the human, mouse and rat genomes (Net; Kent et al., 2003), we defined and collected a set of FFLSs. Here, we show the FFLSs are statistically and biologically meaningful segments for functional analysis of mammalian genomes.

3.1. Overview of the FFLSs in the three mammals

For all six possible comparisons between the three genomes, we detected thousands of "FFLS" (Table 1). There were far more FFLSs detected in comparisons between human and the rodents than between rat and mouse. This may reflect the longer divergence time from the common ancestor of the human and rodents than the divergence time between the two rodents (Consortium, 2004b; Thomas et al., 2003). When we looked into the distribution of the FFLSs between chromosomes of all the three mammalians, we saw that some chromosomes were either enriched or depleted in FFLSs compared to the rest of chromosomes (enrichment on Chr19 and Chr22 in human, Chr17 in mouse and Chr4, Chr6 and Chr7 in rat; depletion on Chr9 and Chr20 in human, Chr6, Chr14 and Chr15 in mouse and Chr5 and Chr9 in rat). To test whether this variation indicated a clustered organization of FFLSs among chromosomes, the averages of the nearest neighbor distances of the FFLSs on each chromosome were compared to an equivalent number of distances between randomly selected nearest neighbor positions. The FFLSs and random positions whose nearest neighbor distances were larger than the expected were excluded from the analysis. Average nearest neighbor distances of FFLSs

Table 1
Basic statistics for the FFLSs

Target genome	Human		Mouse		Rat	
Query genome	Mouse	Rat	Human	Rat	Human	Mouse
chr1	68	62	40	9	272	6
chr2	135	184	34	6	32	7
chr3	34	44	23	16	35	13
chr4	19	15	6	15	217	8
chr5	18	18	29	21	9	0
chr6	39	42	147	1	157	0
chr7	82	100	122	9	163	10
chr8	64	68	15	22	52	0
chr9	2	5	49	4	4	8
chr10	56	62	21	30	11	10
chr11	44	45	15	19	12	0
chr12	64	96	68	6	29	2
chr13	24	20	38	8	63	20
chr14	103	101	30	2	16	0
chr15	55	38	2	0	49	2
chr16	86	79	15	11	25	0
chr17	31	35	118	53	28	13
chr18	27	36	18	21	12	9
chr19	278	311	16	11	44	1
chr20	2	11	N	N	24	3
chr21	24	18	N	N	N	N
chr22	59	55	N	N	N	N
chrX	24	19	29	87	17	15
Total FFLS	1338	1464	835	351	1271	127
Mean distance	1.75E+06	2.94E+06	1.04E+06	1.97E+07	752882.7	2.91E+06
Median distance	84 233	82 643	22 792	149 142	45 523	16 125
Unique query gene	163	38	170	3	292	9

The number of detected FFLSs in the human, mouse and rat genomes are shown for each chromosome. The rows "Mean distance" and "Median distance" show the distribution of distances between the query pair-members of the FFLSs. The mean and median distances were calculated only for those FFLSs for which query pair-members are located on the same chromosome. The "Unique query gene" row shows the number of FFLSs whose two query components belong to the same gene.

were significantly lower than those of random groups on nearly all chromosomes. In human, except for chr9 ($p = 0.701$, 0.128, queried by mouse and rat, respectively) and chr20 ($p = 0.580$, 0.169), FFLS were clustered on all the other 22 chromosomes with p values less than 0.001. In mouse, the exceptions were chr1 ($p < 0.001$, $= 1$ queried by human and rat, respectively), chr2 ($p < 0.001$, $= 1$) chr9 ($p = 0.913$, 0.419) and chr19 ($p = 1$, 1), and a similar clustering was observed in rat genome, despite fewer detected FFLS (Table 1).

Apart from clustering on chromosomes, the FFLSs were highlighted from the random background by several other features (Table 2). Both the average GC-content and the percentage of FFLS containing CpG island are higher than that in the random background. Human FFLSs contain more CpG island than those in rodents. The ratios of protein-coding to non-protein-coding regions in FFLSs were also higher than in the random background. The FFLSs detected between the rodents contained fewer protein-coding regions than detected between the human and the rodent genomes. This difference may possibly be explained by the fact that the drafts of rodent genomes are still unfinished, alternatively, it might also be influenced by the shorter distance of the rodents to their common ancestor.

We found three types of FFLSs: (i) both component chains containing protein-coding genes; (ii) one of the component chains containing protein-coding genes, and (iii) none of component chains containing protein-coding genes. The first group may represent gene fusion (or fission) events. As to the non-protein-coding regions of the second and the third FFLS groups, we guess it could be either fusion (or fission) events involving non-protein-coding genes, or caused by the incomplete annotation of the mammalian genomes.

3.2. Functional linkages between the components of the FFLSs

Marcotte et al. (1999) and Enright et al. (1999) have already shown that two proteins are more likely to interact if their homologues in another organism have been fused into one protein; however, as a method for predicting PPIs, this approach has been reported to yield a high proportion of false positives (82%) (Snel et al., 2000). We checked if this association to protein interactions also holds in our

Table 2
Sequence features of the FFLSs

	Human		Mouse		Rat	
	Mouse	Rat	Human	Rat	Human	Mouse
G + C content[a]	0.4287 ($p = 0.314$)	0.4274 ($p = 0.327$)	0.4146 ($p = 0.413$)	0.3998 ($p = 0.515$)	0.4147 ($p = 0.310$)	0.4089 ($p = 0.325$)
G + C content[b]	0.4242 ($p = 0.323$)	0.4272 ($p = 0.301$)	0.4144 ($p = 0.367$)	0.4022 ($p = 0.391$)	0.4129 ($p = 0.512$)	0.4029 ($p = 0.418$)
CpG island[a]	0.0896 ($p<0.001$)	0.0770 ($p<0.001$)	0.0133 ($p = 0.075$)	0.0158 ($p = 0.077$)	0.0180 ($p = 0.002$)	0.0460 ($p = 0.003$)
CpG island[b]	0.090 ($p<0.001$)	0.0939 ($p<0.001$)	0.0026 ($p = 0.003$)	0.020 ($p = 0.060$)	0.0184 ($p = 0.012$)	0.051 ($p<0.001$)
EST[a]	0.8585 ($p<0.001$)	0.8540 ($p = 0.009$)	0.9049 ($p<0.001$)	0.7586 ($p<0.001$)	0.3032 ($p<0.001$)	0.2931 ($p<0.001$)
EST[b]	0.8074 ($p = 0.003$)	0.7993 ($p<0.001$)	0.8839 ($p<0.001$)	0.7142 ($p<0.001$)	0.2911 ($p<0.001$)	0.2769 ($p = 0.002$)
Coding/non-coding	0.3370 ($p<0.001$)	0.3385 ($p<0.001$)	0.2176 ($p<0.001$)	0.1837 ($p = 0.24$)	0.028 ($p = 0.026$)	0

The statistical significances are show in the parentheses under each value. The first line displays the target genomes. The "CpG island" rows show the percentage of FFLSs containing a CpG island. The "EST" rows show the percentage of FFLSs overlapped with a EST. The "coding/non-coding" rows show the ratio of protein-coding exons to non-protein-coding (intergenic) region contained in the FFLSs.

[a] All FFLS.

[b] Non-protein-coding FFLS.

detections. A data set of PPIs was integrated from several databases (see Methods) to investigate, whether the two proteins encoded by two components either in target or query genome tend to interact. In the human genome, we found 109 such FFLSs when queried by both mouse and rat, similarly, 164 and 30 FFLSs with interacting proteins were found in the query genomes of mouse and rat, respectively. When we used the opposite querying direction, fewer such FFLSs were found in rodents (29 and 4 cases in mouse and rat when used as target genomes, and 82 cases in the human when used to query to mouse genome). Unsurprisingly, very few such FFLSs with interacting proteins were found among FFLSs obtained from comparisons between rodents, since only few FFLS were detected in these comparisons (Table 1). This number of interacting FFLS proteins is significantly higher than in randomly selected segments (Fig. 2). We used the annotated chromosomal coordinates of a gene to indicate whether the gene is contained in an FFLSs. However, because the chain does not always maintain the exon structure of a gene, we say a gene is associated to a chain if 10% or more of its total exon length overlaps with the blocks of the chain. Testing some alternative thresholds other than 10% (e.g. 30%, 70% and 90%) also gave a statistically higher number of interacting proteins. However, as currently there are no available genome-wide high-throughput data on PPIs in mammals, the results may also be due to bias in the PPI data.

Gene fusions can be also represented as overlapping genes. With the same threshold of gene region matching as above (10%), we used a recent genome-wide survey of overlapping mammalian genes (Veeramachaneni et al., 2004) to test how many such genes could be recovered by our method. In the human genome, of 542 genes with overlapping exons, 83 and 109 cases were consistent with FFLSs obtained by querying the mouse and rat genomes, respectively. In the mouse genome, out of 455 genes with overlapping exons, 29 and 2 cases were recovered by querying human and rat, respectively (Fig. 2). Interestingly, according to our PPI data set, all overlapping FFLS genes encode the proteins with experimentally proven mutual interactions, whereas among non-FFLS overlapping genes, only a fraction also encode mutually interacting proteins. Not all overlapping genes were detected with our method, lack of detection being possibly caused by the differences in alignment strategies. The overlapping genes were identified among well annotated proteins, whereas we have used a net of organized genome alignment (Kent et al., 2003), which was more focused on genomic structure than on protein structure. As conservation of protein-coding regions is different from non-protein-coding regions (Zhang et al., 2003a), it is not surprising that the two alignments methods do not produce identical results.

In Marcotte et al.'s pioneer work (Marcotte et al., 1999) on gene fusion analysis, the authors proposed that fusion may greatly reduce the entropy of dissociation of two components. By the same logic, the entropy of cascade reactions in pathways should also be reduced if genes encoding cascade protein are fused. However, when checked against the KEGG pathway annotation (Kanehisa

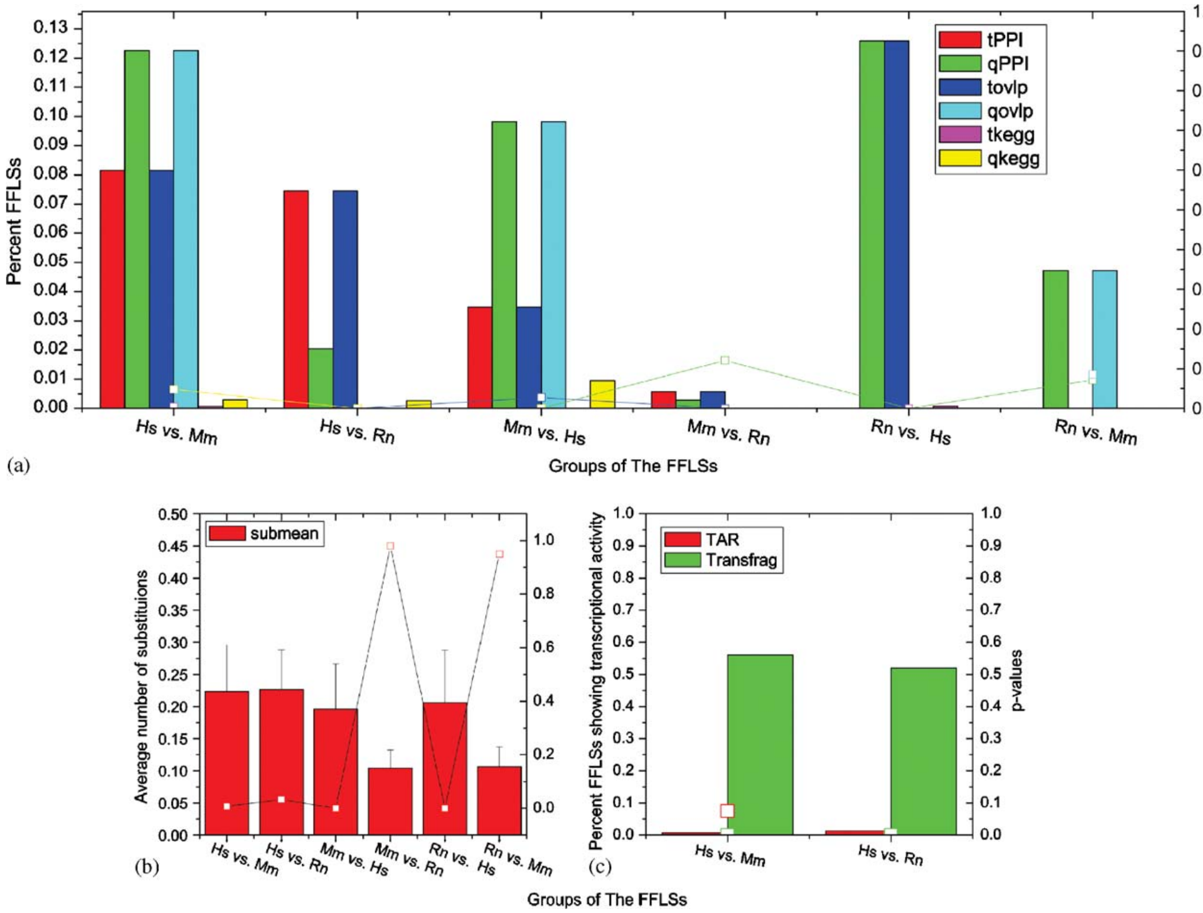

Fig. 2. FFLSs putative biological functions. (a) Functional linkage between components of the FFLSs. The percentage of FFLSs showing either of the three types of functional linkage; protein–protein interaction (PPI), gene overlap (ovlp) and common KEGG pathway (kegg), are shown as bars. Components in target and query genome are distinguished by a prefix "t" or "q" in the legend. The *p* values are shown as squares. (b) Conservational, and (c) transcriptional features of the FFLSs. Tiling array data is so far only available for the human genome.

et al., 2004), only a small fraction of the gene products of components in the target or query genomes were found to act within the same pathway. Although the fraction of FFLSs sharing pathways is statistically higher than that of the random background (Fig. 2), it is still too low to be considered as potential tool for of pathway studies.

3.3. Evidence relating to the putative functions of non-protein-coding FFLSs components

Among our detected FFLSs, many contain components that have not yet been annotated with any protein-coding genes. Whereas this may be due to incomplete genome annotation, another possibility is that the non-protein-coding components harbor other functional elements, such as non-coding RNA. It is reasonable to assume, that if the FFLS components contain functional elements, these should either be transcribed more often or be more conserved than the random context, or both.

We first investigated the non-protein-coding FFLS by intersecting with the sequence elements that are possibly involved in transcription regulation. CpG islands are usually associated with transcriptional promoters, and as shown in Table 2, there are significantly more CpG island in FFLSs than in the random background. The tendency to contain CpG islands is stronger in human FFLSs than in FFLSs from intra-rodents comparisons. We next intersected the non-protein-coding FFLSs with transcriptional data from the recently released high density oligonucleotide arrays (TAR; Bertone et al., 2004; Transfrag; Cheng et al., 2005). The TAR experiment contained a series of high-density oligonucleotide (36 nt) tilling microarrays representing both sense and antisense strands of the entire non-repetitive sequence of the human genome. These arrays identified a large number of transcription units of which only a third corresponded to previously annotated exons. As shown in Fig. 2, 5 and 11 FFLSs in human were found to overlap with TAR data (3 and 7 were non-protein-coding FFLSs). Although this observation is also significantly higher than for the random background, the numbers are still very low. One reason for this could be that the criteria for accepting a microarray signal as evidence of transcription in the TAR data was very strict, and suitable only for detecting highly transcribed regions (Bertone et al., 2004). Transcription levels of possible functional elements in non-protein-coding FFLSs may be

Table 3
Expressed FFLS

		Total	Exonic	Intronic	Intergenic
H17 vs. M5	Number of FFLS in the 10 chromosomes	348	86	78	184
	Number of FFLS intersected with Transfrag	110	40	7	63
	Percentage of total length covered by Transfrag	0.689	0.672	0.684	0.701
	p value	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
H17 vs. R3	Number of FFLS in the 10 chromosomes	448	132	90	226
	Number of FFLS intersected with Transfrag	139	53	14	72
	Percentage of the FFLS length covered by Transfrag	0.649	0.621	0.578	0.683
	p value	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$

The table shows FFLS intersecting with Transfrag data from the at human genome. The Transfrag data cover 10 human chromosomes (Chr6, Chr7, Chr13, Chr14, Chr19, Chr20, Chr21, Chr22, ChrX, and ChrY or approximately 30% of the human genome.) profiled by tiling microarrays. The percentage of the FFLS length covered by Transfrag is higher than for the random sampling data set. The p value row gives the statistical significances of the percentage of total length covered by Transfrag.

low or tissue specific, and we therefore checked it further against the Transfrag tiling array data (Cheng et al., 2005). Transfrag is a denser oligonucleotide (5 nt) tiling microarrays data set from 10 human chromosomes in eight difference cell lines. These results indicated that more than 50% of the non-protein-coding FFLSs in these 10 human chromosomes were potentially transcribed (Table 3).

The integrative analysis of CpG islands and DNA microarray data presented a global picture of genomic transcriptional activity. However, only a unique EST match can verify and represent conclusive evidence for transcription from particular FFLSs. We therefore compared the pre-identified EST matching genomic regions (UCSC Genome Browser; Kent et al., 2002), with our FFLSs data set. Regions identified by ESTs intersected close to 80% of the total non-protein-coding FFLSs length in human (Table 2). This is a higher percentage than what was identified by Transfrag, and may be contaminated by false-positives EST mapping; however, have much higher than the percentages of average coverage of random background segments (72%). Our data and analyses indicate that at least one-third of the non-protein-coding components of FFLSs, may be transcriptionally active at some level.

To perform a conservation analysis on the FFLSs, we joined, for each FFLS, the qSCs in the two componential TLSCs and aligned it to the two joined tSCs. Then, using the general time-reversible Markov model of base substitution, REV(Yang, 1994) implemented in the PAML program, we estimated the average number of substitutions per site in each pair of the two compared mammalian genomes. As shown in Fig. 2, we obtained 0.22 ($SD = 0.07$) and 0.23 ($SD = 0.06$) substitutions per site in human when compared with mouse and rat, respectively. The substitutions rate of FFLSs is somehow lower when rodents are used to query the human genome (0.19, $SD = 0.07$ and 0.20, $SD = 0.08$ substitutions per site for mouse and rat, respectively). Predictably, the substitutions rate in rodents were much lower (0.1 substitutions per site), since the time from divergence from the common ancestor of the rodents, is much shorter than that from the common ancestor with human. These substitution rates are much lower than the about 0.45 substitution per neutral site reported previously (Consortium, 2004b; Cooper et al., 2004; Hardison et al., 2003). However, the substitution rate for our random sampling data sets, (e.g. 0.26 substitutions per site in humans) is also lower than neutral site, since they were sampled from the chained BLASTZ alignments. The substitution rate of the human FFLSs is, nevertheless, significantly lower than the random background (Fig. 2). A similarly lower substitution rate has also been suggested in previous studies on overlapping genes (Lipman, 1997; Miyata and Yasunaga, 1978).

In summary, we have defined and collected a set of chained segments, which we have called FFLSs, and which may be associated with fusion/fission events in the evolutionary history of the three recently sequenced mammals. The sequence statistics of the FFLSs highlighted them from the random background. Three types of FFLSs were observed. FFLS with protein-coding components show a strong tendency towards their encoded proteins being interaction partners, whereas FFLS with non-protein-coding components shown strong evidence of being transcribed at some level. The latter are also more conserved than the random background, indicating that the non-protein-coding components may carry certain biological functions. A full list of the FFLSs in the genomes of three mammals is available as supporting information at http://bioinfo.ibp.ac.cn/network/paper/suppl/SupplementList.htm.

4. Discussion

In the original definition of the gene fusion, the authors aligned genes according to several different criteria (Enright et al., 1999; Marcotte et al., 1999). This strategy is not applicable to mammalian genome research because of several confounding factors. Mammalian genomes have many segmental duplications, repetitive elements are not always recognized by available software, and many gene

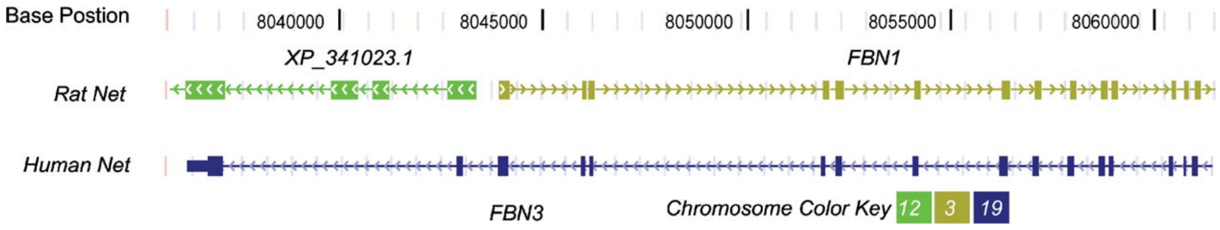

Fig. 3. An example of FFLS at human chromosome 19. "Base Position" row gives the coordinates of human chromosome 19 in the region at 8040–8060 kb. "Human Net" row shows top-level structure of human chromosome 19 at the region, and "Rat Net" row shows the top-level structure of orthology regions in rat genome. The colored rectangles represent the top-level blocks in the corresponding chromosomes. The sequence in the region were aligned by BLASTZ and further organized into a net structure. The rectangles aligned in the same column indicate the blocks were orthologs. Genes in the regions are labeled with its name.

families have expanded by duplication or contracted by deletions. However, with the recent progress of genomics and proteomics, we have obtained more knowledge on the structure and evolution of the mammalian genomes (Bailey et al., 2004; Bourque et al., 2004; Cooper et al., 2004; Kent et al., 2003; Murphy et al., 2001; Samonte and Eichler, 2002; Wienberg, 2004), which make it possible to handle these problems more efficiently. The growing amount of available PPI data in mammals (Bader et al., 2003; Hermjakob et al., 2004; Pagel et al., 2005; Salwinski et al., 2004; Zanzoni et al., 2002) strengthens the case for potential fusion/fission, as fusion/fission events have previously proven to be a powerful tool for PPI prediction (Enright et al., 1999; Marcotte et al., 1999). New kinds of DNA microarrays which can screen genomic region for transcriptional activity (Bertone et al., 2004; Cheng et al., 2005) have also emerged. Against the backdrop of those new advances, we have defined and collected a set of chained genome segments called FFLS, which are potentially related to fusion/fission events in the three sequenced mammals since the divergence from their common ancestor. The definition of FFLS was based on the organized genome alignment, AXTNET (Kent et al., 2003), and only chains at the top level were accepted as orthologous. In addition to masking simple repeats and tandem repeats by REPEATMASKER (Smit, 1999) and TANDEM REPEAT FINDER (Benson, 1999), we filtered out very short chains from further analysis. By comparing to randomly sampled data sets, we found several statistically significant characteristics the FFLS that have been further verified by components functional linkage analysis. For the protein-coding components, we showed that the proteins encoded by the components have strong tendency to interact with each other. For the non-protein-coding components, we showed that a considerable fraction is likely to be transcribed to some level, and they are more conserved than other chained alignment, both indicating that they may harbor some biologically functional elements. As to the gene fusion/fission events, our results showed similar significant linkage of PPI components as it did in previous works (Enright et al., 1999; Marcotte et al., 1999).

An example of FFLS is found in human chromosome 19. A segment at 8040–8060 kb, was shown to have homology to two segments in rat chromosome 12, at 1415–1439 kb, and chromosome 3, at 112 518–112 537 kb (Fig. 3). The segment in human contains 16 exons of the *fibrillins-3* (*FBN3*), which were represented as the top level blocks in the human–rat comparison net. The corresponding homologous segment in rat chromosome 12 that is annotated contains the transcript XP_341023.1. The homology segment in rat chromosome 3 is constructed by 14 top-level blocks, which were annotated as 14 exons a of *fibrillins-1* (*FBN1*). This FFLS and its orthologs have recently been suggested involved in a fission events (Corson et al., 2004).

At this early stage of genome-wide segmental fusion/fission analysis, the present study cannot obtain full support from the annotation of mammalian genomes. This is particularly true for rat (Consortium, 2004b). However, given the rapid accumulation of knowledge on the mammalian genomic structure, FFLSs is a potential tool for prediction of gene function and the PPIs.

Acknowledgments

We thank Dr. G.Skogerbø for helpful comments on the manuscript. We also thank the comments and suggestions from the two anonymous reviewers. This work was supported by the National High Technology Development Program (973) of China under Grant no. 2002CB713805 and no. 2003CB715907, Chinese Academy of Sciences Grant no. KSCX2-2-07 and no. KJCX 1-08.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2005.09.016.

References

Bachman, N.J., Wu, W., Schmidt, T.R., Grossman, L.I., Lomax, M.I., 1999. The 5′ region of the COX4 gene contains a novel overlapping gene, NOC4. Mamm. Genome. 10, 506–512.

Bader, G.D., Betel, D., Hogue, C.W., 2003. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 31, 248–250.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E., 2002. Recent segmental duplications in the human genome. Science 297, 1003–1007.

Bailey, J., Baertsch, R., Kent, W., Haussler, D., Eichler, E., 2004. Hotspots of mammalian chromosomal evolution. Genome. Biol. 5, R23.

Balakirev, E.S., Ayala, F.J., 2003. Pseudogenes: are they "junk" or functional DNA? Annu. Rev. Genet. 37, 123–151.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al., 2004. Global identification of human transcribed sequences with genome tiling arrays. Science 306, 2242–2246.

Bourque, G., Pevzner, P.A., Tesler, G., 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. Genome Res. 14, 507–516.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al., 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308, 1149–1154.

Colland, F., Jacq, X., Trouplin, V., Mougin, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P., Gauthier, J.-M., 2004. Functional proteomics mapping of a human signaling pathway. Genome Res. 14, 1324–1332.

Consortium, M.G.S., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Consortium, I.H.G.S., 2004a. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

Consortium, R.G.S.P., 2004b. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493–521.

Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S., Sidow, A., 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. Genome Res. 14, 539–548.

Corson, G.M., Charbonneau, N.L., Keene, D.R., Sakai, L.Y., 2004. Differential expression of fibrillin-3 adds to microfibril variety in human and avian, but not rodent, connective tissues. Genomics 83, 461–472.

Enright, A.J., Ouzounis, C.A., 2001. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol. 2, Research0034.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A., 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402, 86–90.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al., 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome. Res. 13, 13–26.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., et al., 2004. IntAct: an open source molecular interaction database. Nucl. Acids Res. 32, D452–D455.

Hurst, L.D., Pal, C., Lercher, M.J., 2004. The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. 5, 299–310.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M., 2004. The KEGG resource for deciphering the genome. Nucl. Acids Res. 32, D277–D280.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D., 2002. The human genome browser at UCSC. Genome. Res. 12, 996–1006.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., Haussler, D., 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl Acad. Sci. USA 100, 11484–11489.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Lehner, B., Fraser, A.G., 2004. A first-draft human protein-interaction map. Genome Biol. 5, R63.

Li, W.H., Gu, Z., Wang, H., Nekrutenko, A., 2001. Evolutionary analyses of the human genome. Nature 409, 847–849.

Lipman, D.J., 1997. Making (anti)sense of non-coding sequence conservation. Nucl. Acids Res. 25, 3580–3583.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D., 1999. Detecting protein function and protein–protein interactions from genome sequences. Science 285, 751–753.

Miyata, T., Yasunaga, T., 1978. Evolution of overlapping genes. Nature 272, 532–535.

Murphy, W.J., Stanyon, R., O'Brien, S.J., 2001. Evolution of mammalian genome organization inferred from comparative gene mapping. Genome Biol. 2 REVIEWS0005.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H.W., et al., 2005. The MIPS mammalian protein–protein interaction database. Bioinformatics 21, 832–834.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D., 2004. The database of interacting proteins: 2004 update. Nucl. Acids Res. 32, D449–D451.

Samonte, R.V., Eichler, E.E., 2002. Segmental duplications and the evolution of the primate genome. Nat. Rev. Genet. 3, 65–72.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.

Snel, B., Bork, P., Huynen, M., 2000. Genome evolution. Gene fusion versus gene fission. Trends Genet. 16, 9–11.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al., 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424, 788–793.

Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R., Makalowska, I., 2004. Mammalian overlapping genes: the comparative perspective. Genome Res. 14, 280–286.

Wienberg, J., 2004. The evolution of eutherian chromosomes. Curr. Opin. Genet. Dev. 14, 657–666.

Williams, T., Fried, M., 1986. A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3′ ends. Nature 322, 275–279.

Yanai, I., Derti, A., DeLisi, C., 2001. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc. Natl Acad. Sci. USA 98, 7940–7945.

Yanai, I., Wolf, Y.I., Koonin, E.V., 2002. Evolution of gene fusions: horizontal transfer versus independent events. Genome Biol. 3 research0024.

Yang, Z., 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39, 105–111.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G., 2002. MINT: a Molecular INTeraction database. FEBS Lett. 513, 135–140.

Zhang, L., Pavlovic, V., Cantor, C.R., Kasif, S., 2003a. Human-mouse gene identification by comparative evidence integration and evolutionary analysis. Genome Res. 13, 1190–1202.

Zhang, Z., Harrison, P.M., Liu, Y., Gerstein, M., 2003b. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 13, 2541–2558.