Expressed Structurally Stable Inverted Duplicates in Mammalian Genomes as Functional Noncoding Elements

Zhen-Xia Chen[1,2,3,*], Brian Oliver[2], Yong E. Zhang[4,5], Ge Gao[3], and Manyuan Long[6,*]

[1]College of Life Sciences and Technology, Huazhong Agricultural University, Wuhan, P.R. China

[2]National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland

[3]Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing, P.R. China

[4]Key Laboratory of Zoological Systematics and Evolution and State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, P.R. China

[5]University of Chinese Academy of Sciences, Beijing, P.R. China

[6]Department of Ecology and Evolution, University of Chicago

*Corresponding authors: E-mails: zhen-xia.chen@mail.hzau.edu.cn;mlong@uchicago.edu.

Accepted: March 10, 2017

Abstract

Inverted duplicates are a type of repetitive DNA motifs consist of two copies of reverse complementary sequences separated by a spacer sequence. They can lead to genome instability and many may have no function, but some functional small RNAs are processed from hairpins transcribed from these elements. It is not clear whether the pervasive numbers of such elements in genomes, especially those of mammals, is the result of high generation rates of neutral or slightly deleterious duplication events or positive selection for functionality. To test the functionality of intergenic inverted duplicates without known functions, we used mirror duplicates, a type of repetitive DNA motifs with few reported functions and little potential to form hairpins when transcribed, as a nonfunctional control. We identified large numbers of inverted duplicates within intergenic regions of human and mouse genomes, as well as 19 other vertebrate genomes. Structure characterization of these inverted duplicates revealed higher proportion to form stable hairpins compared with converted mirror duplicates, suggesting that inverted duplicates may produce hairpin RNAs. Expression profiling across tissues demonstrated that 7.8% of human and 5.7% of mouse inverted duplicates were expressed even under strict criteria. We found that expressed inverted duplicates were more likely to be structurally stable than both unexpressed inverted duplicates and expressed converted mirror duplicates. By dating inverted duplicates in the vertebrate phylogenetic tree, we observed higher conservation of inverted duplicates than mirror duplicates. These observations support the notion that expressed inverted duplicates may be functional through forming hairpin RNAs.

Key words: inverted repeats, *Homo sapiens*, *Mus musculus*; noncoding RNA, mirror repeats, hairpin RNA.

Introduction

Over 98% of the human genome is outside protein-coding exons (Bernstein et al. 2012), and two-thirds of the human genome may be repetitive or repeat-derived (de Koning et al. 2011). Despite the prevalence of these sequences, our understanding of what role, if any, these sequences might play is rudimentary (Doolittle 2013), which makes the functional analysis of repetitive elements an important problem in genomics.

Inverted duplicates (we use "inverted duplicates" instead of the commonly used term "inverted repeats" to indicate that they consist of two copies of sequences) are a specific category of repetitive elements. They consist of two reverse complementary sequences (termed as arms) and a spacer ($\geq$ 0 nt in length) between the arms. They have been observed in the genomes of bacteria (Lillo et al. 2002), yeast (Strawbridge et al. 2010), flies (Chen et al. 2011), mouse (Bollag and Liskay 1988), and human (Warburton et al. 2004).

Multiple types of noncoding genes have been discovered (Khalil et al. 2009; Ruby et al. 2007) and if inverted duplicates are transcribed, they would be noncoding gene candidates. Transcription of inverted duplicates results in the formation of an RNA that is capable of forming dsRNA, or a hairpin, via base-pairing. The formation of secondary structures in

noncoding RNAs is essential for known functional inverted duplicates (Castel and Martienssen 2013). The secondary structures of RNAs are determined by both Watson-Crick (A-U and C-G) and wobble (G-U) base-pairings, and are usually assessed by thermodynamic stability; specifically, the predicted minimum free energy (MFE) structure (Hofacker 2003; McCaskill 1990; Zuker and Stiegler 1981). These hairpins could subsequently be a substrate of Dicer for production of miRNAs, siRNAs, or other small functional RNAs. There are examples of miRNAs and endo-siRNAs, processed from hairpin RNAs transcribed from inverted duplicates, that function to suppress genes through silencing, mRNA degradation, or translational repression mechanisms (Castel and Martienssen 2013). Additionally, some RNAs (e.g., rRNA, tRNA, snRNA, snoRNA, and lncRNA) use hairpin structures as scaffolding for complex assembly and/or catalytic activity (Azad 1979; Dick and Schamel 1995; Jady et al. 2012; Piekna-Przybylska et al. 2007; Rinn and Chang 2012). However, it is unclear if inverted duplicates play any general role in biology and evolution. Indeed, there is evidence that many inverted duplicates are deleterious, as they lead to genomic instability by promoting chromosome rearrangements including those resulting in acentric and dicentric chromosomes (Darmon et al. 2010; Lobachev et al. 2000; Lu et al. 2015; Mizuno et al. 2009; Tanaka and Yao 2009). For example, inverted duplicates of *Alu* (the major class of repetitive DNA of human genome [Schmid 1996]), stimulate intra and interchromosomal recombination when transformed into yeast, and appear to be under purging selection in the human genome (Lobachev et al. 2000). In the classic mutation-selection balance theory (Li 1997), the large numbers of inverted duplicates may be rapidly generated and then subsequently subjected to negative selection that purges them from the genome, but should any of these inverted duplicates become functional they could be retained.

Mirror duplicates (we use "mirror duplicates" instead of the commonly used term "mirror repeats" to indicate that they consist of two copies of sequences) are a realistic nonfunctional control essential for determining the functionality of inverted duplicates. By definition, mirror duplicates consist of two copies of reverse sequences separated by a spacer sequence, and thus simulated inverted duplicates can be produced by converting one arm of mirror duplicates to complementary sequence. Biologically, mirror duplicates, as well as inverted duplicates, are a type of repetitive DNA motifs that may induce transcriptional repression and genomic instability, including deletions, point mutations and translocations, in mammals (Bacolla et al. 2004, 2006; Liu et al. 2012; Sundararajan and Freudenreich 2011; Wang et al. 2008; Wang and Vasquez 2006). Moreover, few mirror duplicates are reported to be functional (Bacolla et al. 2004, 2006; Liu et al. 2012; Sundararajan and Freudenreich 2011; Wang et al. 2008; Wang and Vasquez 2006). Technically, mirror duplicates have similar definition as inverted duplicates, and can be identified by the same program with the same sensitivity (Warburton et al. 2004).

In the current study, we used mirror duplicates as a nonfunctional control to evaluate the functionality and related evolvability of intergenic inverted duplicates in human and mouse in four angles, including background distribution, folding probability, expression, and evolutionary age distribution. This study may facilitate our understanding on the roles of repetitive regions in complex mammalian genomes in their biology and evolution.

Materials and Methods

Genomes

Genome sequences of human (GRCh38/hg38), chimpanzee (panTro4), orangutan (ponAbe2), rhesus macaque (rheMac3), marmoset (calJac3), mouse (GRCm38/mm10), rat (rn6), guinea pig (cavPor3), rabbit (oryCun2), dog (canFam3), cow (bosTau8), armadillo (dasNov3), African elephant (loxAfr3), tenrec (echTel2), opossum (monDom5), platypus (ornAna1), chicken (galGal4), lizard (anoCar2), frog (xenTro3), fugu (fr3), and zebrafish (danRer10) were downloaded from the UCSC Genome Browser (Lindblad-Toh et al. 2011; Karolchik et al. 2014). Simulated chromosomes in human and mouse were generated by randomly shuffling all nucleotides of real chromosomes with the command shuffleseq in the European Molecular Biology Open Software Suite (EMBOSS, v6.3.1) (Rice et al. 2000). Chromosomes X and 7, which constitute 10% of both genomes, were simulated for 600 times.

Detection of Inverted and Mirror Duplicates

We used Inverted Repeats Finder (IRF, version 3.05; Warburton et al. 2004) on the human and mouse genomes and simulated chromosomes with randomly shuffled sequences to identify two types of sequences flanking a central axis of symmetry. The first were inverted duplicates, which is a sequence followed by its reverse complement. The second were control mirror duplicates, which is a sequence followed by its reverse. IRF searches for exact reverse-complement matches (for inverted duplicates) or reverse matches (for mirror duplicates) of 4–7 nt tuples between nonoverlapping segments of a sequence. The matched short tuple pairs centered in the same or nearly the same sites were clustered together for the calling of inverted duplicates and mirror duplicates. If the alignment of a matched tuple cluster exceeded a predefined minimum alignment score (Minscores), then it would be reported as an inverted duplicate or mirror duplicate. We assigned weights for each match (+2), mismatch (−3) and insertion/deletion (−5) to calculate scores for inverted duplicates and mirror duplicates (using the -mr command). The five parameters proportion of matches (PM) and indels (PI), minimum alignment score (Minscores), maximum arm length (MaxLength), and maximum spacer length

(MaxLoop) influenced the number of identified inverted duplicates and mirror duplicates by IRF. We tuned parameters to maximize the number of human inverted duplicates returned with moderate consideration on the expense of machine time. First, to determine PM and PI, we set loose parameters as follows: Minscores (20), MaxLength (500,000), and MaxLoop (10,000), and varied parameters PM (80, 85, and 90) and PI (0 and 5). We identified the most inverted duplicates when PM = 80 and PI = 5, and thus examined other three parameters with determined PM (80) and PI (5). We found only 197 (0.003%) human inverted duplicates had arms > 10,000 nt, and thus determine MaxLength to 10,000. We found the majority of inverted duplicates had spacer length ≤1,000 nt and the number of identified inverted duplicates decreased slowly with the increase of spacer length when spacer length was > 1,000 nt. Furthermore, considering inverted duplicates with spacer length > 1,000 nt were less likely to form hairpin structures, we thus determined MaxLoop to be 1,000. We found the number of identified inverted duplicates dropped rapidly with increasing IRF score, and thus determined MinScore to be 20.

Secondary Structure of Inverted Duplicates

The most stable secondary structure of a sequence is the minimum free energy (MFE) structure of the sequence. We predicted the MFE structures of RNAs encoded by inverted duplicates with the *RNAfold* program (v2.1.1, in the Vienna RNA Package)(Zuker and Stiegler 1981; McCaskill 1990; Hofacker 2003). Briefly, we used a Perl script to extract the sequence of each inverted duplicate (including two arms and a spacer) and folded the sequence. Then, we filtered out the MFE structures with low stability (MFE > −15 kcal/mol, as applied in Lu et al. [2008]), or those that could not form hairpins (parings between arms < 6, or stem length < 21 bp). We required a stable hairpin MFE structure have stem length ≥ 21 bp, because the length of mature miRNAs and siRNAs, which are embedded in the stems, are ~22 nt (Ruby et al. 2007) and ~21 nt (Zamore et al. 2000), respectively.

As a control, we also predicted the MFE structures of RNAs encoded by converted mirror duplicates. We converted the right arms of mirror duplicates to complementary sequences and followed the process above to predict the secondary structures.

Classification of Inverted Duplicates

We classified inverted duplicates into genic and intergenic inverted duplicates based on the General Transfer Format (GTF) format gene annotation of human (GENCODE v24) and mouse (GENCODE vM9) downloaded from GENCODE (http://www.gencodegenes.org; last accessed March 25, 2017). The gene annotation includes both automatic annotation generated by Ensembl and manual annotation curated by HAVANA group at the Welcome Trust Sanger Institute for all annotated transcripts from protein-coding, pseudogenes and noncoding genes. We used bedtools command *intersectBed* (Quinlan and Hall 2010) to find inverted duplicates having at least 1 nt overlap with a gene, including all the annotated protein-coding, pseudogenes and noncoding genes, and classified them into genic inverted duplicates. Other inverted duplicates were classified into intergenic inverted duplicates. To avoid expression signals from nearby genes, we excluded inverted duplicates whose spacers were within 1 kb of any gene in expression analysis.

We also classified inverted duplicates into repetitive and nonrepetitive inverted duplicates based on the repeat annotations created by RepeatMasker (http://www.repeatmasker.org; last accessed March 25, 2017) from UCSC. RepeatMasker identified low complexity regions and repetitive regions, including long Interspersed Elements (LINE), Short Interspersed Elements (SINE), simple repeat, LTR and other repeats, in DNA sequences. We referred inverted duplicates that have at least 1 nt overlap with any repeat as repetitive inverted duplicates and others as nonrepetitive inverted duplicates.

We further separated inverted duplicates into five groups based on their arm length to test whether inverted duplicates of different arm length might have different patterns. The 0, 20, 40, 60, 80, and 100 quantiles of arm length for human inverted duplicates were 10, 14, 17, 22, 36, and 10,000 nt. We thus classified inverted duplicates, as well as mirror duplicates, in human and mouse into the five groups: SS (10–13 nt), S (14–16 nt), M (17–21 nt), L (22–35 nt), and LL (36–10,000 nt).

Dating Inverted Duplicates on the Phylogenetic Tree of Vertebrates

We modified a previous pipeline (Zhang et al. 2010a) to date inverted duplicates on the vertebrate phylogenetic tree based on reciprocal syntenic alignments between genomes. We performed whole-genome pairwise alignments between human/mouse and the 19 species (chimpanzee, orangutan, rhesus macaque, marmoset, rat, guinea pig, rabbit, dog, cow, armadillo, African elephant, tenrec, opossum, platypus, chicken, lizard, frog, fugu, zebrafish, as well as between human and mouse, by LASTZ (Harris 2007). Following the UCSC Genome Browser manual (Meyer et al. 2013), we generated reciprocal best LiftOver chain files from the alignments and aligned human and mouse inverted duplicates to the other species. We defined an inverted duplicate as present in a species if it could be aligned to the species using this method no matter whether it was still an inverted duplicate in the species. Based on the presence and absence of inverted duplicates in the genomes, we dated human and mouse inverted duplicates to the branches on the phylogenetic tree, and used the mid-point of the branches as their origination time (Zhang et al. 2010a).

Expression Profiling

RNA-seq reads from 20 human tissues (trachea, thyroid, thymus, testis, spleen, prostate, placenta, ovary, muscle,

lung, liver, kidney, intestine, heart, esophagus, colon, cervix, brain, bladder, and adipose; Clark et al. 2015) and 22 mouse tissues (thymus, testis, subcutaneous adipose, stomach, spleen, small intestine, placenta, ovary, mammary gland, lung, liver, large intestine, kidney, heart, genital adipose, frontal lobe, duodenum, cortex, colon, cerebellum, bladder, and adrenal; Lin et al. 2014) were download from NCBI's Sequence Read Archive (SRA) (accession numbers: SRP047192, SRP012040). We aligned SRA sequencing data, as well as raw reads in FASTQ format, by HISAT2 (v2.0.4) (Kim et al. 2015) with provided index for the reference genomes (-x) and annotations (-known-splicesite-infile).

To estimate the expression level of inverted duplicates, we quantified discrete counts of reads overlapping the spacer and flanking 1 kb of each inverted duplicate by *coveragebed* in BedTools v2.25.0 (Quinlan and Hall 2010), normalized the read counts by R package DESeq2 v1.6.3 (Love et al. 2014). For each inverted duplicate, maximum normalized read counts among 20 human tissues and 22 mouse tissues were calculated as expression of the inverted duplicate for human and mouse.

To define expression status for intergenic inverted duplicates with high confidence, we used a stringent cutoff derived from the expression level of genic inverted duplicates. The median expression of all the genic inverted duplicates were taken as the expression cutoff for intergenic inverted duplicates. Intergenic inverted duplicates with normalized read count in any tissue higher than the cutoff were termed as expressed inverted duplicates, and others were unexpressed.

Results

Abundance of Inverted Duplicates in Human and Mouse Genomes

We carried out a genome-wide search for inverted duplicates (defined as two copies of 10–10,000 nt reverse complementary arm sequences separated by a 0–1,000 nt spacer; fig. 1*A*) in 21 vertebrate genomes, including human, mouse, chimpanzee, orangutan, rhesus macaque, marmoset, rat, guinea pig, rabbit, dog, cow, armadillo, African elephant, tenrec, opossum, platypus, chicken, lizard, frog, fugu, and zebrafish (Lindblad-Toh et al. 2011). Abundant inverted duplicates were observed in the genomes, ranging from 377,552 in fugu to 5,087,596 in opossum (see supplementary table S1, Supplemental Material online). Specifically, we identified 4,315,271 inverted duplicates in human (19% of the genome, included 1,891,032 intergenic inverted duplicates) and 3,174,675 in mouse (14% of the genome, included 1,752,433 intergenic inverted duplicates). To understand the mechanisms underlying such abundance, we focused on human and mouse because they had high-quality reference sequences (Dietrich et al. 1996; Mueller et al. 2013; Olivier et al. 2001) and annotations (Church et al. 2009; Ross et al. 2005), as well as rich high-throughput sequencing data for functional genomic analysis.

To test whether the abundance of inverted duplicates was common for other comparable elements, like mirror duplicates (defined as inverted duplicates, but the two copies are in reverse rather than reverse complementary) and direct duplicates (defined as inverted duplicates, but the two copies are the same), we used mirror duplicates (fig. 1*B*) as a close comparative system with the inverted duplicates of interest. We did not use direct duplicates because they could not be identified through the same process as inverted duplicates with similar sensitivity, whereas mirror duplicates could. In total, we identified 6,547,895 mirror duplicates in human (19% of the genome, included 2,936,365 intergenic mirror duplicates) and 7,760,250 in mouse (16% of the genome, included 4,511,292 intergenic mirror duplicates), demonstrating even higher abundance than inverted duplicates (fig. 1*C*, $P < 0.001$). The higher abundance of mirror duplicates than inverted duplicates was also observed in 17 of the other 19 vertebrate genomes, except for platypus and frog (see supplementary table S1, Supplemental Material online). Meanwhile, inverted and mirror duplicates were similarly abundant in randomly shuffled chromosomes (fig. 1*C*, Wilcoxon signed rank test, $P > 0.05$), ruling out the possibility that the higher abundance of mirror duplicates was due to bias introduced by detection methodologies. Therefore, the abundance of inverted duplicates might be common for other duplicated elements. The abundance of inverted duplicates could be resulted from rapid gain through mutational mechanisms and/or maintenance of functional elements under selection. We then focused on intergenic inverted duplicates without known functions to explore the mechanisms. Intergenic inverted duplicates were referred to as inverted duplicates in the following text useless otherwise specified.

Repetitive regions may facilitate the growth of duplicates through mutational mechanisms (Bailey et al. 2003; Fiston-Lavier et al. 2007) by repeated insertion of transposable elements (Lobachev et al. 2000) and DNA replication errors (Brewer et al. 2011). To estimate the contribution of mutational mechanisms to the abundance of inverted duplicates, we classified the duplicates into those in repetitive and nonrepetitive regions of the genomes. We found that inverted duplicates and mirror duplicates were enriched 2- to 8-fold in repetitive regions (inverted duplicate density: 1.71 per kb in human and 1.67 per kb in mouse; mirror duplicate density: 2.99 per kb in human and 5.19 per kb in mouse) relative to nonrepetitive regions (inverted duplicate density: 0.82 per kb in human and 0.62 per kb in mouse; mirror duplicate density: 0.83 per kb in human and 0.67 per kb in mouse), supporting the idea that repetitive regions facilitate the origination of inverted and mirror duplicates. Interestingly, although the counts of mirror duplicates were 75–210% more than inverted duplicates for repetitive regions, they were only 1–8% more for nonrepetitive regions. Compared with mirror

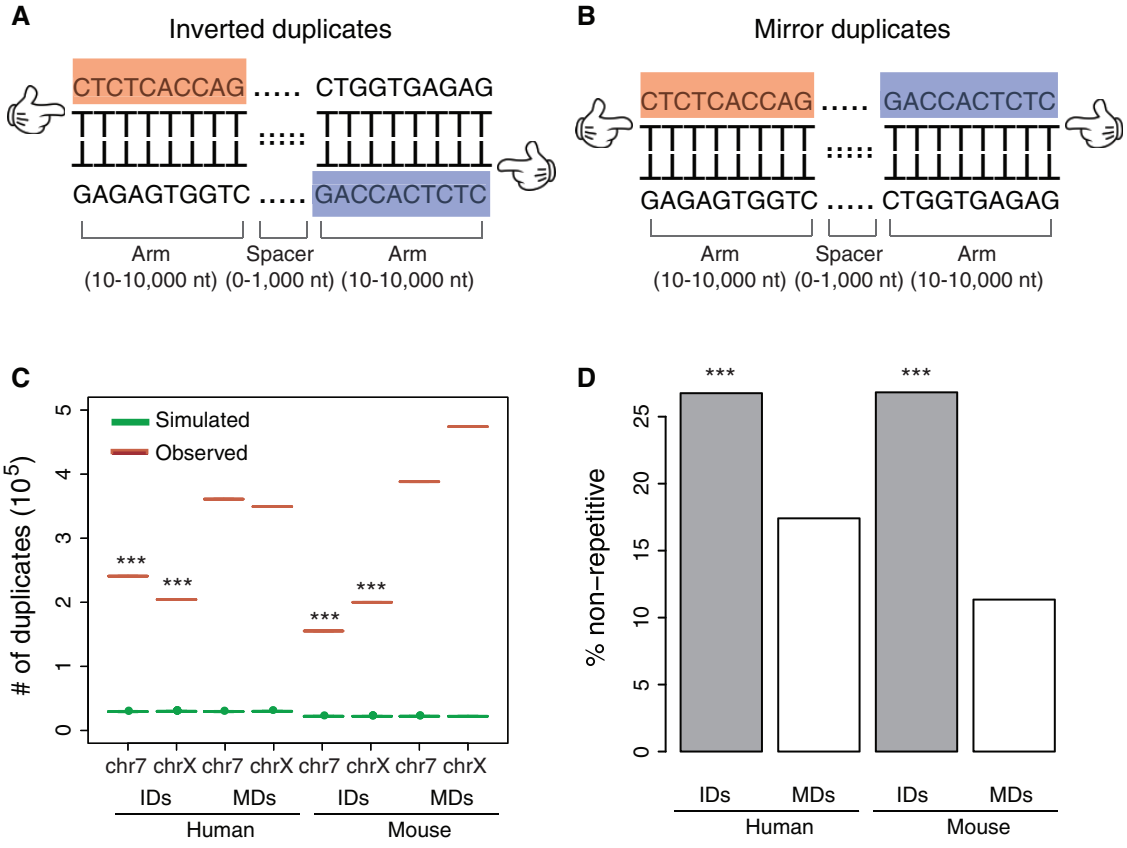

Fig. 1.—Abundance of inverted duplicates and mirror duplicates. (*A* and *B*) Schematic representation of inverted duplicates (*A*) and mirror duplicates (*B*). The orange and blue blocks have the same sequences in opposite directions. The hand symbols indicate the direction of the sequences. (*C*) Box and whisker plot for amounts of inverted duplicates and mirror duplicates identified in real chromosomes (red) and 600 randomly shuffled chromosomes (green). The boxes of simulated duplicates span from the first quantile to the third quantile (the interquartile range or IQR) for the 600 simulations, and the whiskers show the minimum and maximum. (*D*) Barplot for percentage of nonrepetitive inverted duplicates (gray) and mirror duplicates (white). chr7: chromosome 7; chrX: chromosome X; IDs: inverted duplicates; MDs: mirror duplicates. ***FET, $P < 0.001$ (compared with corresponding mirror duplicates).

duplicates, thus, inverted duplicates were enriched in nonrepetitive regions, where they might form at lower frequency under mutational mechanisms (fig. 1*D*, Fisher's exact test, FET, $P < 0.001$).

Higher Potential for Inverted Duplicates to Form Hairpins

Some inverted duplicates may play a functional role through the formation of hairpin structures when transcribed (Castel and Martienssen 2013; Tao et al. 2007a, 2007b). To determine the potential for inverted duplicates to form hairpins, we first examined arm and spacer length, as inverted duplicates with longer arms and shorter spacers should form more stable hairpin structures (Zuker and Stiegler 1981). Although we found human inverted duplicates were longer than mirror duplicates (fig. 2*A*), mouse inverted duplicates were shorter (fig. 2*A*). Moreover, the spacers of inverted duplicates were longer in both human and mouse (fig. 2*A*). We further separated inverted duplicates into five groups (SS, S, M, L, and LL, from shortest to longest) based on their arm length, and test whether each group showed similar patterns. We found inverted duplicates in all the groups had longer spacer than mirror duplicates. The LL group of inverted duplicates had the longest spacers among all the groups in both human and mouse genomes, whereas LL group of mirror duplicates had the smallest spacers (fig. 2*C*–*F*).

To directly assay folding ability of inverted duplicates, we predicted the secondary structures of RNAs transcribed from inverted duplicates and converted mirror duplicates (inverted duplicates computationally generated from mirror duplicates by converting their right arms to complementary sequences) using Minimum Free Energy (MFE) structure and pairing probability. We used structures with at least six Watson-Crick (A-U or C-G) or wobble base pairs (U-G) between the arms ≥ 21 nt in length and MFE ≤ -15 kcal/mol as a cutoff for defining a hairpin encoding element (termed as "structurally stable", see

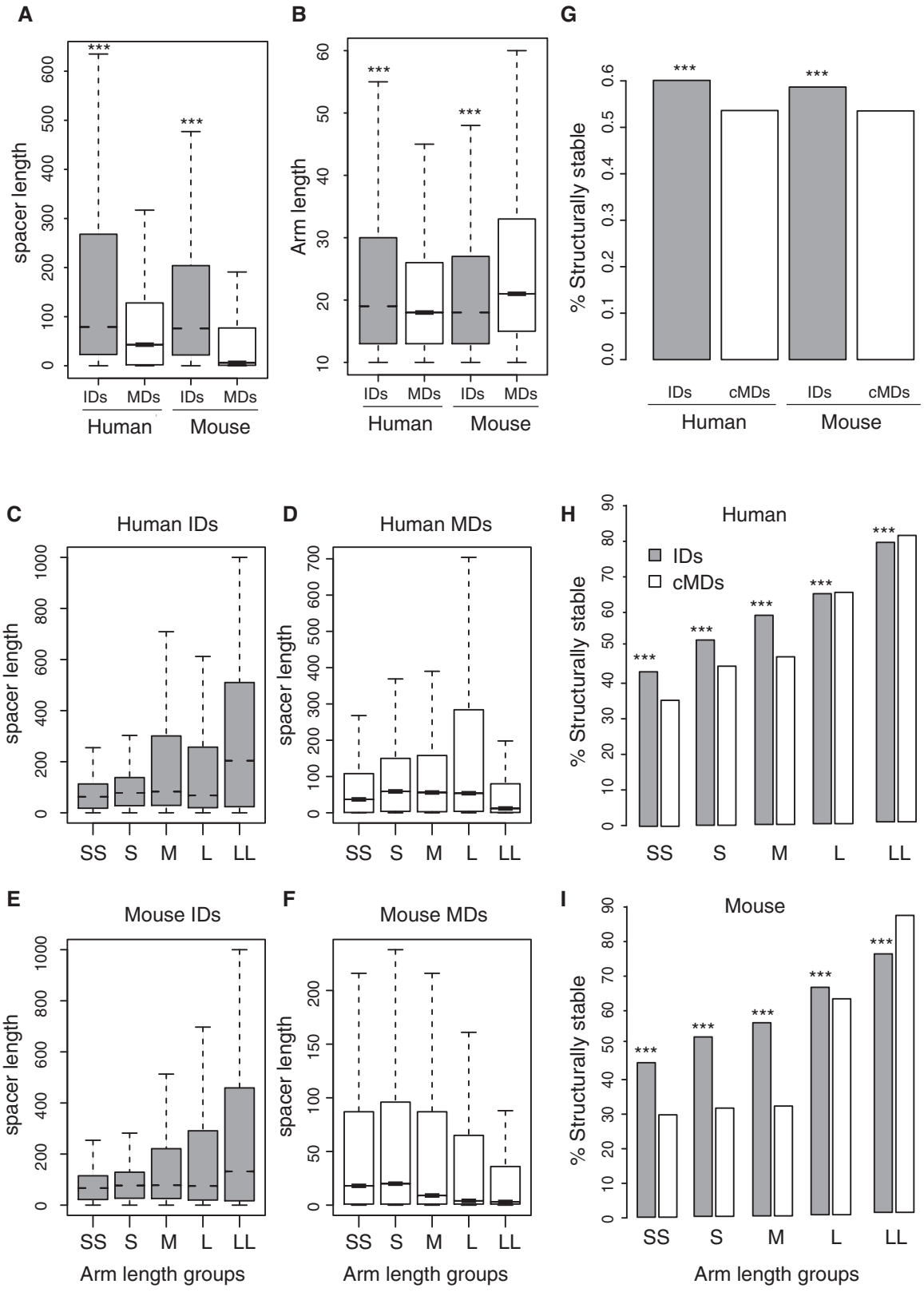

Fig. 2.—Structures of inverted duplicates. (*A* and *B*) Box and whisker plot for arm length (*A*) and spacer length (*B*) of inverted duplicates (gray) and mirror duplicates (white). (*C*–*F*) Box and whisker plot for spacer length of inverted duplicates (*C*, *E*) and mirror duplicates (*D*, *F*) in human (*C*, *D*) and mouse (*E*, *F*). (*G*–*I*)

Materials and Methods). We found 1,196,909 (63%) human inverted duplicates and 1,073,564 (61%) mouse inverted duplicates were structurally stable, whereas 1,655,063 (56%) human converted mirror duplicates and 2,515,846 (56%) mouse converted mirror duplicates were structurally stable (fig. 2*G*). Although converted mirror duplicates also had reverse complementary arms as inverted duplicates, lower percentage of them were structurally stable in both human and mouse than inverted duplicates (fig. 2*G*). Distinctively, we found lower percentage of structurally stable inverted duplicates in the L and LL groups in human and LL group in mouse than converted mirror duplicates, whereas higher percentage of structurally stable inverted duplicates in other groups (fig. 2*H* and *J*).

Enrichment of Expressed Structurally Stable Inverted Duplicates

Inverted duplicates can function as hairpin RNAs only if they are expressed. We took advantage of previously published RNA-seq data from 20 human tissues (Clark et al. 2015) and 22 mouse tissues (Gan et al. 2013; Lin et al. 2014) to identify expressed inverted duplicates. We used stringent criteria to define expressed inverted duplicates. The maximum normalized read counts of an inverted duplicate across all the tissues were taken as the expression level of the inverted duplicates. Inverted duplicates with normalized read counts in any tissue higher than the median expression level of all genic inverted duplicates were termed as expressed inverted duplicates. In total, 147,567 (7.8%) human inverted duplicates and 99,078 (5.7%) mouse inverted duplicates were expressed, whereas 216,067 (7.4%) human mirror duplicates and 248,345 (5.5%) mouse mirror duplicates were expressed, demonstrating the statistically significant, albeit small, enrichment of expressed inverted duplicates (FET, $P < 0.001$), which motivated further examination.

Furthermore, we tested the cooccurrence between hairpin structure and expression. If structure and expression were correlated, then more expressed inverted duplicates should be structurally stable and more structurally stable inverted duplicates should be expressed. Interestingly, we found the proportion of structurally stable inverted duplicates in expressed inverted duplicates (63.9% in human and 65.4% in mouse) were higher than that in unexpressed inverted duplicates (63.2% in human and 61.0% in mouse) (fig. 3*A* and *B*, FET, $P < 0.001$) and the proportion of expressed inverted duplicates in structurally stable inverted duplicates (7.9% in human and 6.0% in mouse) were also higher than that in unstable inverted duplicates (7.7% in human and 5.1% in mouse) (fig. 3*C* and *D*, FET, $P < 0.001$). Meanwhile, both the proportion of structurally stable inverted duplicates in expressed inverted duplicates and of expressed inverted duplicates in structurally stable inverted duplicates were higher than that for mirror duplicates (56.6% and 7.4% in human, 56.7% and 5.6% in mouse) (fig. 3*E*–*H*, FET, $P < 0.001$). Besides, we found the enrichment of expressed inverted duplicates in structurally stable inverted duplicates and of structurally stable inverted duplicates in expressed inverted duplicates existed in all the five groups classified by arm length (fig. 3*I*–*L*).

Higher Portion of Old Inverted Duplicates than That for Mirror Duplicates

To measure evolutionary age distribution of inverted duplicates, we dated all intergenic inverted duplicates based on presence and absence in the phylogenetic tree of vertebrates (fig. 4*A* and *B*). Nearly all the intergenic inverted duplicates, including 1,858,342 human inverted duplicates (98%, branch 5–12) and 1,735,811 mouse inverted duplicates (99%, branch 5–11), were originated after the split of eutherian mammals and marsupial (fig. 4*C*). Among the eutherian-specific inverted duplicates, 1,487,223 human inverted duplicates were primate-specific (79%, branch 8–12) (see supplementary table S2, Supplemental Material online) and 1,626,075 mouse inverted duplicates were rodent-specific (93%, branch 8–11) (see supplementary table S3, Supplemental Material online, fig. 4*C*). We referred these inverted duplicates as young inverted duplicates and others as old inverted duplicates. Considering there are only 9% primate-specific and 15% rodent-specific protein-coding genes, and 19% primate-specific and 21% rodent-specific miRNAs in human and mouse (Zhang et al. 2010b) using the same criteria, the portion of order-specific inverted duplicates was much higher than known functional elements. Although it might suggest rapid evolution of inverted duplicates, it could also be partly explained by the bias of age dating method (Moyers and Zhang 2015). The method depended on the detection of orthologs via sequence homology, and the age of DNA elements would be underestimated especially when they were short and evolved fast as most inverted duplicates (Moyers and Zhang 2015). Such method bias would be avoided in the comparison between inverted duplicates and mirror duplicates, because they were similar in length and were dated with the same method. We found the portion of old inverted duplicates was higher than that for mirror duplicates (fig. 4*D*), demonstrating slower evolution of inverted duplicates than mirror duplicates.

Fig. 2.—Continued

Barplot for percentage of structurally stable inverted duplicates (gray) and converted mirror duplicates (white). The five groups SS, S, M, L, LL were classified by arm length from shortest to longest. The boxes represent IQR, and the whiskers indicate 1.5 × IQR above and below the boxes. IDs: inverted duplicates; MDs: mirror duplicates; cMDs: converted mirror duplicates. ***FET, $P < 0.001$ (compared with corresponding converted mirror duplicates).

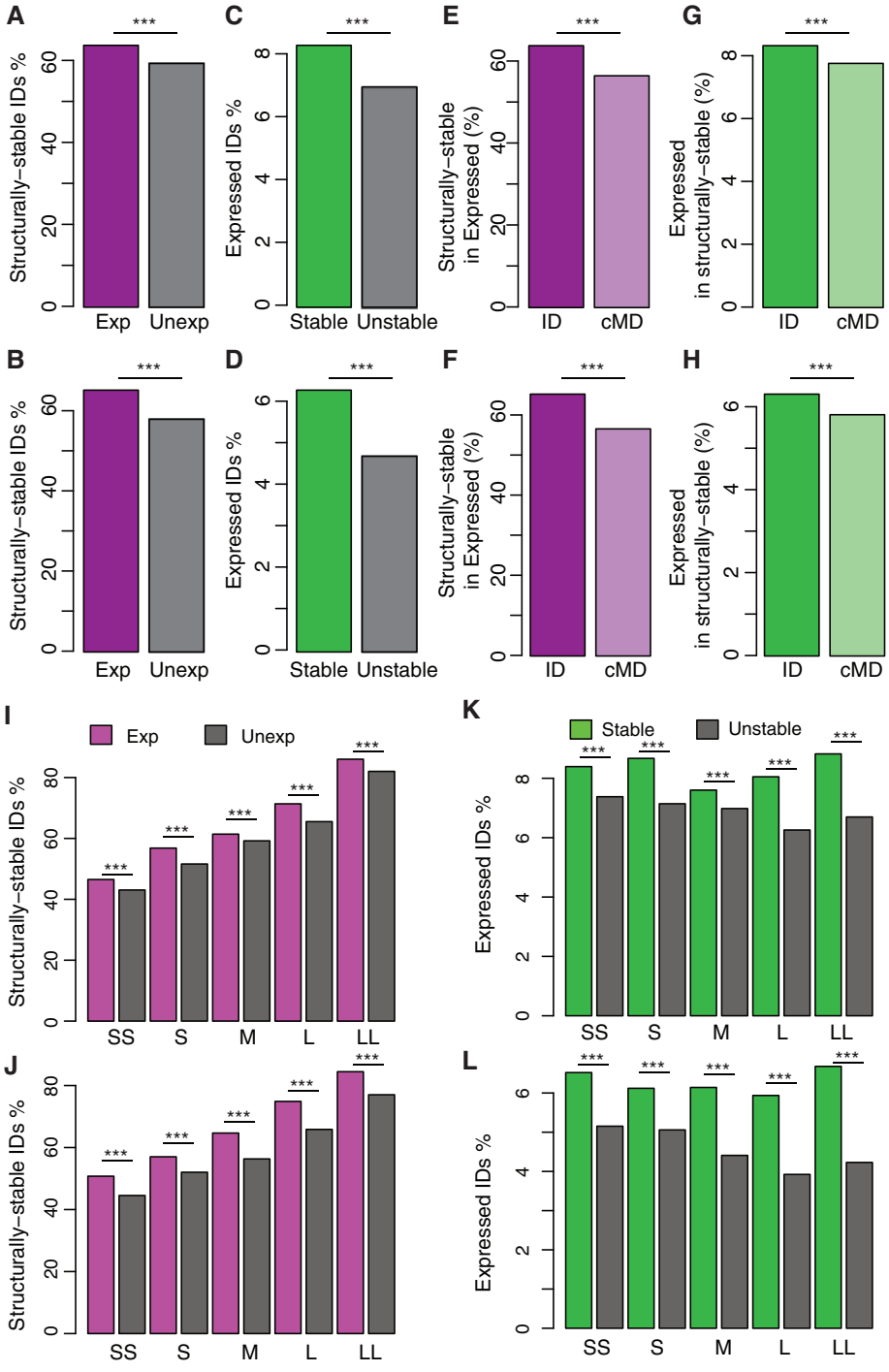

Fig. 3.—Expression of inverted duplicates. (*A* and *B*) Percentage of structurally stable inverted duplicates in expressed (purple) and unexpressed inverted duplicates (gray) in human (*A*) and mouse (*B*). (*C* and *D*) Percentage of expressed inverted duplicates in structurally stable (green) and unstable inverted duplicates (gray) in human (*C*) and mouse (*D*). (*E* and *F*) Percentage of structurally stable inverted duplicates in expressed inverted duplicates (dark) and converted mirror duplicates (light) in human (*E*) and mouse (*F*). (*G* and *H*) Percentage of expressed inverted duplicates in structurally stable inverted duplicates (dark) and converted mirror duplicates (light) in human (*G*) and mouse (*H*). (*I* and *J*) Percentage of structurally stable inverted duplicates in expressed (purple) and unexpressed (gray) inverted duplicates in human (*I*) and mouse (*J*). (*K* and *L*) Percentage of expressed inverted duplicates in structurally stable (green) and unexpressed (gray) inverted duplicates in human (*K*) and mouse (*L*). The five groups SS, S, M, L, and LL were classified by arm length from shortest to longest. IDs: inverted duplicates; cMDs: converted mirror duplicates; Exp: expressed; Unexp: unexpressed; Stable: structurally stable. ***FET, $P < 0.001$.

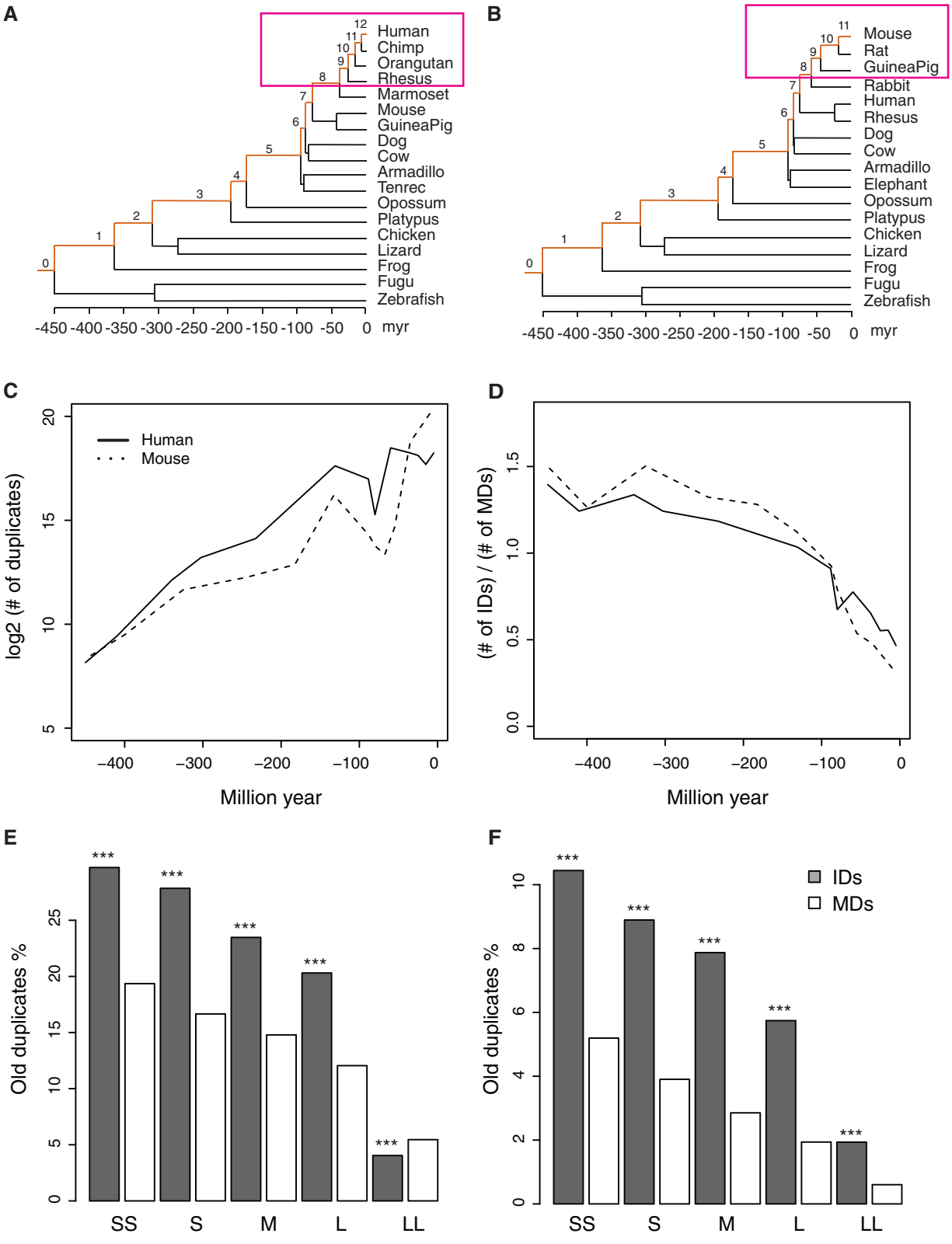

Fig. 4.—Conservation of inverted duplicates. (*A*) Human inverted duplicates originated in each branch of phylogenic tree of vertebrates; (*B*) Mouse inverted duplicates originated in each branch. The phylogenic tree of vertebrates was constructed as described in (Zhang et al. 2010b). Branches leading to human and mouse are marked with orange line. Branch names are shown at each branch. Primate-specific and rodent-specific inverted duplicates are

The slower evolution of inverted duplicates than mirror duplicates may be resulted from either slower gain and/or slower loss of inverted duplicates compared with mirror duplicates. We found younger inverted duplicates had lower abundance than mirror duplicates (fig. 4*D*), suggesting slower gain of inverted duplicates. The relative abundance of inverted duplicates increased with age (fig. 4*D*), demonstrating slower loss (or higher conservation in other words) of inverted duplicates than mirror duplicates. Furthermore, the percentage of old inverted duplicates was greater than that of old mirror duplicates in all the five groups classified by arm length, except for the LL group in human (fig. 4*E* and *F*).

Discussion

We identified large number of inverted duplicates dispersed in the intergenic regions of human and mouse genomes. To find out mechanisms underlying the abundance of inverted duplicates, we analyzed their density distribution, structures, expression and phylogenetic distribution. In this discussion, we discuss about the rationality of our methods in determining the potential role of noncoding elements. We also discuss about the potential roles of inverted duplicates as a basis to understand the biological property of already known gigantic numbers of noncoding RNAs in large transcribed intergenic regions in metazoan genomes.

Controls for Inverted Duplicates

Repetitive regions are abundant in mammalian genomes (de Koning et al. 2011). It's not clear whether their abundance, structures, expression patterns and evolutionary age distribution are due to background evolutionary forces on neutral or slightly deleterious mutations or selection for favorable elements. Therefore, a nonfunctional control is essential for separating positive selection from background evolutionary forces, and determining the functionality of inverted duplicates. In this study, we used mirror duplicates as a major nonfunctional control for inverted duplicates. We identified similar number of inverted and mirror duplicates in randomly shuffled chromosomes and thus excluded the possibility of introducing differences to their comparison from identification methods. Also, we observed that mirror duplicates, compared with inverted duplicates, had higher abundance in repetitive regions and less conservation. These observations support that mirror duplicates are at least less likely to be functional than inverted duplicates, and thus can be used as nonfunctional control for inverted duplicates.

Besides mirror duplicates, we also used inverted duplicates in simulated chromosomes and converted mirror duplicates as controls to determine the function of inverted duplicates based on different hypothesis. A simple null hypothesis is that inverted duplicates were originated randomly in the genomes. By comparing with inverted duplicates in simulated randomly shuffle chromosomes, we can tell inverted duplicates are not originated randomly because they are enriched 7- to 9-fold than those in simulated chromosomes. To explore the hypothesis that inverted duplicates function as hairpins, we compared them with converted mirror duplicates. Converted mirror duplicates have reverse complementary arms as inverted duplicates. By comparing the secondary structures between inverted duplicates and converted mirror duplicates, we can tell that inverted duplicates tend to form hairpins because more inverted duplicates are structurally stable than converted mirror duplicates. The simulated inverted duplicates and converted mirror duplicates fit well with our null hypothesis, and thus can be used as complementary controls to mirror duplicates for inverted duplicates

Roles of Intergenic Inverted Duplicates

Our observations revealed dual effects, advantageous effects via hairpin RNAs (Castel and Martienssen 2013) and deleterious effects by forming hairpins (Lobachev et al. 2000), of inverted duplicates.

Clues for the functionality of inverted duplicates via hairpin RNAs were found in four angles: In density distribution, inverted duplicates showed enrichment relative to mirror duplicates in nonrepetitive regions; In structures, the potential of inverted duplicates to form hairpins was higher than that of converted mirror duplicates; In expression, expressed inverted duplicates were enriched in structurally stable inverted duplicates than in unstable inverted duplicates; In evolutionary age distribution, inverted duplicates showed higher portion of old ones than mirror duplicates. Furthermore, given that previously reported vast majority of genomic regions in human and *Drosophila* are transcribed (Li et al. 2009; Pennisi 2007), we propose that at least part of the identified large numbers of intergenic inverted duplicates are likely to be located in the vast transcribed genomic regions and some may even function as noncoding RNAs. In fact, 147,567 (7.8%) human inverted duplicates and 99,078 (5.7%) mouse inverted duplicates have been identified to be parts in transcripts from intergenic DNAs even with strict criteria. This also supports the potential role of inverted duplicates as noncoding RNAs. To experimentally test the function of inverted duplicates may be a challenge but not

Fig. 4.—Continued

marked in pink boxes. (*C*) The number of inverted duplicates originated in each branch in human (solid line) and mouse (dotted line). (*D*) Ratio of inverted duplicate number to mirror duplicate number in human (solid line) and mouse (dotted line). (*E* and *F*) Percentage of old inverted duplicates (gray) and mirror duplicates (white). The five groups SS, S, M, L, and LL were classified by arm length from shortest to longest. IDs: inverted duplicates; MDs: mirror duplicates. ***FET, $P < 0.001$ (compared with mirror duplicates in corresponding group).

impossible, given the genome editing techniques that have been developed and are being improved (Chen et al. 2015; Platt et al. 2014; Xue et al. 2014).

Meanwhile, the lower abundance of inverted duplicates than mirror duplicates, especially in repetitive regions and for young duplicates, suggests that inverted duplicates may subject to stronger background selection against their deleterious effects than mirror duplicates. It is in consistent with the deleterious effects of large inverted duplicates (Darmon et al. 2010; Lobachev et al. 2000; Lu et al. 2015; Mizuno et al. 2009; Tanaka and Yao 2009).

We can roughly weigh the two effects from the five groups of inverted duplicates classified by arm length. Generally, all the groups showed similar patterns supporting the function of inverted duplicates via hairpin RNAs. However, the L group of inverted duplicates in human and LL group in both human and mouse showed lower potential to form hairpin structures than mirror duplicates. Besides, the LL group in human showed lower portion of old ones than that for mirror duplicates. These observations suggest that deleterious effects of large inverted duplicates can overweigh advantageous effects when they form large hairpin structures. Therefore, large hairpin structures are disfavored even for functional inverted duplicates (Darmon et al. 2010; Lobachev et al. 2000; Lu et al. 2015; Mizuno et al. 2009; Tanaka and Yao 2009). Our inference is supported by previous study in human on 166 large (arm length >8 kb) inverted duplicates, which may function in male germ-line gene expression and/or maintaining sequence integrity by gene conversion (Warburton et al. 2004). Despite the large arm size, 159 (96%) of the inverted duplicates have spacer >1 kb and thus are not likely to form hairpin structures (Warburton et al. 2004), supporting that large hairpin structures are disfavored even for functional inverted duplicates.

To summarize, we have examined the intergenic inverted duplicates in human and mouse in four angles, and found that at least some of them are likely to be functional. Furthermore, the expression and hairpin structure could be important to the function of inverted duplicates. Meanwhile, large hairpin structures are disfavored even for functional inverted duplicates.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Author Contributions

Z.C., G.G., Y.Z., B.O., and M.L. conceived and designed the analysis. Z.C. performed the analysis. Z.C., B.O., Y.Z., and M.L. wrote the paper. All authors contributed comments and critiques.

Acknowledgments

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov). Z.C. was supported by Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Program No.2016RC011); B.O. was supported in part by the Intramural Research Programs of the National Institutes of Health, NIDDK (DK015600-18); M.L. was supported in part by the National Institutes of Health, R01GM116113-01A1 and National Science Foundation, 1051826 and 1026200. The authors declare no conflicts of interest. Certain commercial equipment, instruments, or materials are identified in this document. Such identification does not imply recommendation or endorsement by NIH, nor does it imply that the products identified are necessarily the best available for the purpose.

Literature Cited

Azad AA. 1979. Intermolecular base-paired interaction between complementary sequences present near the 3′ ends of 5S rRNA and 18S (16S) rRNA might be involved in the reversible association of ribosomal subunits. Nucleic Acids Res. 7:1913–1929.

Bacolla A, et al. 2004. Breakpoints of gross deletions coincide with non-B DNA conformations. Proc Natl Acad Sci U S A. 101:14162–14167.

Bacolla A, et al. 2006. Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. Nucleic Acids Res. 34:2663–2675.

Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 73:823–834.

Bernstein BE, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74.

Bollag RJ, Liskay RM. 1988. Conservative intrachromosomal recombination between inverted repeats in mouse cells: association between reciprocal exchange and gene conversion. Genetics 119:161–169.

Brewer BJ, Payen C, Raghuraman MK, Dunham MJ. 2011. Origin-dependent inverted-repeat amplification: a replication-based model for generating palindromic amplicons. PLoS Genet. 7:e1002016.

Castel SE, Martienssen RA. 2013. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nat Rev Genet. 14(2):100–112.

Chen S, et al. 2015. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell 160:1246–1260.

Chen ZX, et al. 2011. Deficiency of X-linked inverted duplicates with male-biased expression and the underlying evolutionary mechanisms in the Drosophila genome. Mol Biol Evol. 28:2823–2832.

Church DM, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 7:e1000112.

Clark MB, et al. 2015. Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. Nat Methods 12:339–342.

Darmon E, et al. 2010. *E. coli* SbcCD and RecA control chromosomal rearrangement induced by an interrupted palindrome. Mol Cell. 39(1):59–70.

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7(12):e1002384.

Dick TP, Schamel WA. 1995. Molecular evolution of transfer RNA from two precursor hairpins: implications for the origin of protein synthesis. J Mol Evol. 41:1–9.

Dietrich WF, et al. 1996. A comprehensive genetic map of the mouse genome. Nature 380:149–152.

Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci U S A. 110:5294–5300.

Fiston-Lavier AS, Anxolabehere D, Quesneville H. 2007. A model of segmental duplication formation in *Drosophila melanogaster*. Genome Res. 17:1458–1470.
Gan H, et al. 2013. Integrative proteomic and transcriptomic analyses reveal multiple post-transcriptional regulatory mechanisms of mouse spermatogenesis. Mol Cell Proteomics 12(5):1144–1157.
Harris R. 2007. Improved pairwise alignment of genomic DNA. University Park, PA: The Pennsylvania State University.
Hofacker IL. 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31:3429–3431.
Jady BE, Ketele A, Kiss T. 2012. Human intron-encoded Alu RNAs are processed and packaged into Wdr79-associated nucleoplasmic box H/ACA RNPs. Genes Dev. 26:1897–1910.
Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 42:D764–D770.
Khalil AM, et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 106:11667–11672.
Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360.
Li W-H. 1997. Molecular evolution. Sunderland, MA: Sinauer Associates.
Li Z, et al. 2009. Detection of intergenic non-coding RNAs expressed in the main developmental stages in *Drosophila melanogaster*. Nucleic Acids Res. 37:4308–4314.
Lillo F, Basile S, Mantegna RN. 2002. Comparative genomics study of inverted repeats in bacteria. Bioinformatics 18:971–979.
Lin S, et al. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 111:17224–17229.
Lindblad-Toh K, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482.
Liu G, et al. 2012. Replication fork stalling and checkpoint activation by a PKD1 locus mirror repeat polypurine-polypyrimidine (Pu-Py) tract. J Biol Chem. 287:33412–33423.
Lobachev KS, et al. 2000. Inverted Alu repeats unstable in yeast are excluded from the human genome. EMBO J. 19:3822–3830.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. 15(12):550.
Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu CI. 2008. The birth and death of microRNA genes in Drosophila. Nat Genet. 40:351–355.
Lu S, et al. 2015. Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. Cell Rep. doi: 10.1016/j.celrep.2015.02.039.
McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29:1105–1119.
Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B. et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 41:D64–D69.
Mizuno K, Lambert S, Baldacci G, Murray JM, Carr AM. 2009. Nearby inverted repeats fuse to generate acentric and dicentric palindromic chromosomes by a replication template exchange mechanism. Genes Dev. 23(24):2876–2886.
Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. Mol Biol Evol. 32:258–267.
Mueller JL, et al. 2013. Independent specialization of the human and mouse X chromosomes for the male germ line. Nat Genet. 45:1083–1087.
Olivier M, et al. 2001. A high-resolution radiation hybrid map of the human genome draft sequence. Science 291:1298–1302.
Pennisi E. 2007. Genomics. DNA study forces rethink of what it means to be a gene. Science 316:1556–1557.
Piekna-Przybylska D, Liu B, Fournier MJ. 2007. The U1 snRNA hairpin II as a RNA affinity tag for selecting snoRNP complexes. Methods Enzymol. 425:317–353.
Platt RJ, et al. 2014. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell 159:440–455.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16(6):276–277.
Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 81:145–166.
Ross MT, et al. 2005. The DNA sequence of the human X chromosome. Nature 434:325–337.
Ruby JG, et al. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome Res. 17(12):1850–1864.
Schmid CW. 1996. Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. Prog Nucleic Acid Res Mol Biol. 53:283–319.
Strawbridge EM, Benson G, Gelfand Y, Benham CJ. 2010. The distribution of inverted repeat sequences in the *Saccharomyces cerevisiae* genome. Curr Genet. 56:321–340.
Sundararajan R, Freudenreich CH. 2011. Expanded CAG/CTG repeat DNA induces a checkpoint response that impacts cell proliferation in *Saccharomyces cerevisiae*. PLoS Genet. 7:e1001339.
Tanaka H, Yao MC. 2009. Palindromic gene amplification—an evolutionarily conserved role for DNA inverted repeats in the genome. Nat Rev Cancer. 9(3):216–224.
Tao Y, et al. 2007a. A sex-ratio meiotic drive system in *Drosophila simulans*. II: an X-linked distorter. PLoS Biol. 5(11):e293.
Tao Y, Masly JP, Araripe L, Ke Y, Hartl DL. 2007b. A sex-ratio meiotic drive system in *Drosophila simulans*. I: an autosomal suppressor. PLoS Biol. 5(11):e292.
Wang G, Carbajal S, Vijg J, DiGiovanni J, Vasquez KM. 2008. DNA structure-induced genomic instability *in vivo*. J Natl Cancer Inst. 100:1815–1817.
Wang G, Vasquez KM. 2006. Non-B DNA structure-induced genetic instability. Mutat Res. 598:103–119.
Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res. 14:1861–1869.
Xue W, et al. 2014. CRISPR-mediated direct mutation of cancer genes in the mouse liver. Nature 514:380–384.
Zamore PD, Tuschl T, Sharp PA, Bartel DP. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell 101:25–33.
Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010a. Age-dependent chromosomal distribution of male-biased genes in Drosophila. In. Genome Res. 20:1526–1533.
Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. 2010b. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol. 8:
Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 9:133–148.

Associate editor: Balazs Papp