SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
EVOLUTION
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Received
17 June 2014

Accepted
28 August 2014

Published
29 September 2014

Correspondence and requests for materials should be addressed to Z.L. (zhlu@seu.edu.cn)

High order intra-strand partial symmetry increases with organismal complexity in animal evolution

Shengqin Wang[1], Jing Tu[1], Zhongwei Jia[3] & Zuhong Lu[1,2]

[1]State Key Lab of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096, China, [2]Department of Biomedical Engineering, College of Engineering, Peking University, Beijing, 100781, China, [3]National Institute of Drug Dependence, Peking University, Beijing 100191, China.

For sufficiently long genomic sequence, the frequency of any short nucleotide fragment on one strand is approximately equal to the frequency of its reverse complement on the same strand. Despite being studied over two decades, the precise mechanism involved has not yet been made clear. In this study, we calculated the high order intra-strand partial symmetry (IPS) for 14 animal species by using a fixed sliding window method to scan each genome sequence. The study showed that the IPS was positive associated with organismal complexity measured by the number of distinct cell types. The results indicated that the IPS might be resulted from the increasing of functional non-coding DNAs, and plays an important role in the evolution process of complex body plans.

It is well known that the frequency of single nucleotide is very similar to the frequency of its reverse complement (%A=%T and %C=%G) within single-stranded DNA sequence in most complete genome sequences. This kind of intra-strand compositional symmetry phenomenon is also called the Chargaff second parity rule, which was discovered in the late 1960s[1]. As growing of deposited genome sequences, such symmetry is commonly found in a sequence which is longer than 50 Kb[2].

Interestingly, the extending researches of Chargaff's Second Parity Rule reveal that the symmetry is very strongly supported by di-, tri or higher order N-mer oligonucleotides for sufficiently long genomic sequences[2,3]. A large number of published papers introduce the intra-strand symmetry in genomic era, but the issue about the origin of this kind of universal phenomenon is still controversial, such as point mutation[4], local recombination rates[5], inversion and inverted transpositions[6,7], etc. Stem-loop structure is also one proposed factor for the intra-strand symmetry[8,9], though the contribution is limited[10]. Moreover, it is an intrinsic property in the genome evolution established by a cumulative effects of a number of mechanisms for multiple orders and length scales[11,12].

It is noted that, the fragments in some genome sequences are found not to be obedient the intra-strand symmetry[13,14]. Besides, different nucleotide composition are found to exist in the two replicating strands of most genomes, even mammalians, which affects the intra-strand symmetry in local region[15,16]. Particularly, the intra-strand symmetry for higher order nucleotides will drop sharply as the increasing of order N[5]. It is known that, the organism complexity measured by the number of different cell types has increased greatly during the course of evolution. In order to know if there is relationship between intra-strand partial symmetry (IPS) of genome and organismal complexity, we investigated the IPS values on high order N-mer nucleotides within available animal organisms via fixed sliding window approach in this study. Increased data of full complete genome sequences provides us an opportunity for global properties analysis of this kind of symmetry. At last, we found that the high order IPS is significantly correlated with organismal complexity, and the increasing of functional non-coding DNAs might be one of the reasons.

Results

Here, the measurement of different high order intra-strand symmetry was performed out, and the ability to get the difference of IPS values is maximized by using 9-mer oligonucleotides (N=9) (Figure 1). For one sequence, the IPS values of N-mer nucleotides ranging from 0 to 1 usually drop sharply as the increasing of order N[11]. In order to maximize the ability to detect the difference of IPS values, we calculated the standard deviation of IPS values under a different order of oligonucleotides to get the optimal parameter (Figure 1).

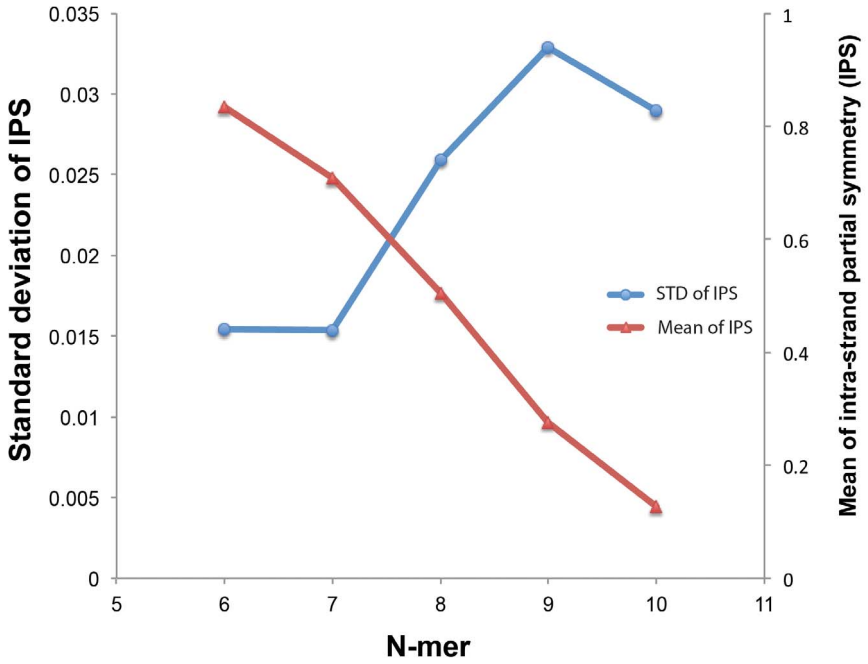

Figure 1 | Plot of standard deviation of the IPS values under different N-mer nucleotides. 9-mer was selected to maximize the divergence among organisms in this study.

Fourteen organisms with their organismal complexity greater than 64 were selected by three filter steps, in which the number of remained fragments is greater than 10. Although there are many sequenced organism genome with low organismal complexity, they have quite a lot of degenerated nucleotides in continuous 50 Kb sliding windows. These degenerated nucleotides could artificially increase the degree of symmetry when allowed fuzzy match. Therefore, we simply discarded all of the fragments containing degenerated nucleotides.

We built a simple linear regression model to find out the relationship between the IPS value and organismal complexity. The average IPS values increase with organismal complexity in a given GC range, since the GC content of the genomic fragment might influence the IPS value. Figure 2 gives relationship between Average IPS and organismal complexity in 0.01 interval from 0.4 to 0.51, as well as their combined range of 0.4 ~ 0.6. The results show that the average IPS values are in the range from 0.21 to 0.29, and all of the correlation effects with organismal complexity are strong (Figure 2). We also calculated other IPS values of N-mer oligonucleotides, and the 7-, 8-, and 10-mer also give significant correlation between the IPS value and organismal complexity except the 6-mer (Supplementary Figure S1). One of the reasons we use the fragments with GC content range from 0.4 ~ 0.6 is that these sequences usually present functional elements, for instance more than 75% of UCSC human genes with GC content within the range from 0.4 to 0.6.

To detect the functional role of such kind of symmetry, average 9-mer IPS values were compared with the density of UCSC genes among chromosomes in the well-annotated human genome. Interestingly, we found that the average IPS values ranging from 0.269 to 0.289 are different among chromosomes, and significantly correlated with the chromosome gene density (Figure 3). High IPS could help producing compact structures, which might prevent the genome degradation by folding stably secondary structure. It is also noted that, there are a lot of functional non-coding regions existed in UCSC genes, such as UTRs, introns etc. Compared with coding DNA, the functional non-coding regions should have more evolutional pressure.

Non-coding DNA is different from coding DNA and does not suffer by the codon choices[17]. It should promote symmetry better and has more chance to fold as secondary structure. The secondary structure in non-coding region, such as the intron region, can be selected for optimizing gene expression of pre-mRNA by increasing the folding free energy[18,19]. Recent research also shows folding into secondary structures to exhibit longer half-lives at the 3′ end are the major determinant of mRNA stability[20]. In order to check if the emergence of functional non-coding DNA increases with the degree of symmetry, we downloaded coding and functional non-coding regions with FASTA format from UCSC genome browser based on "human" Refseq annotation[21], and compared the difference of the IPS values between them. Here, we removed sequences containing degenerate bases. Continuous stretches of 50 Mb nucleotide sequences were made from non-redundant subset sequences by removing gene names, and then we quantified 9-mer symmetries as the increasing of contiguous sequences in each dataset. In order to ignore the order effect of pooled genes, each data was randomly shuffled and calculated by 10 times. Our result shown that the functional non-coding regions approach higher symmetry than coding regions at 9-mer nucleotides as the size of the pool increases (Figure 4), which also give us the clue that the emergence of functional non-coding sequence may give the reason to the increasing of the symmetry degree in animal evolution.

Discussion

The significance of correlation between IPS and organismal complexity suggests functional implication for such symmetry, perhaps in promoting structural evolution of functional DNAs (Figure 2, Supplementary Figure S1). It is known that, the genome changes both in size and structure on long-term evolution[22]. The various changes in the genome can be interpreted by occurred duplication, insertion, transposable elements, recombination, mutation, et al. These variants correlating with phenotype will go through accumulation of mutations unfettered by the restraining selective forces before being fixed in the genome. In other words, the functional genome structures can be generated from the mechanism of random genetic drift and selection pressure during the evolutional process. For example, as the hypothesis of "Kissing" model, the stem-loop structure contributes to the initiation of meiotic recombination[9], so the

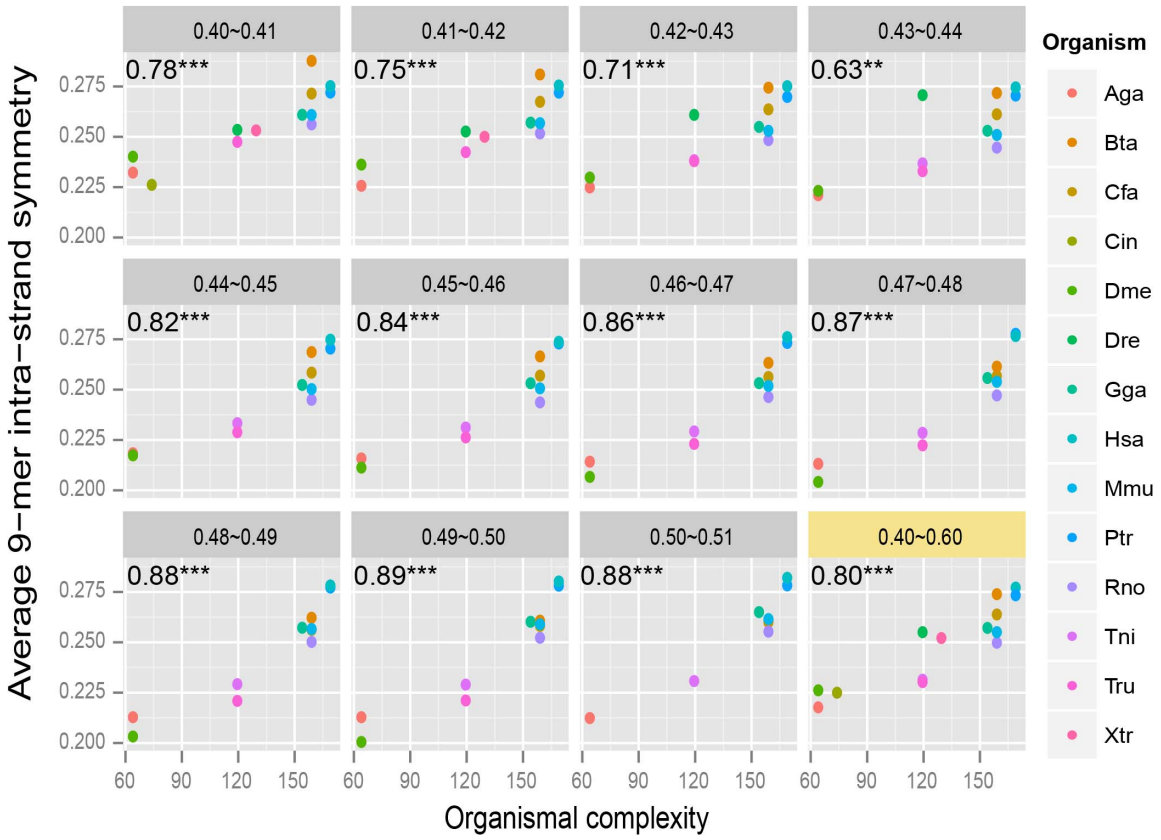

Figure 2 | The average value of 9-mer order intra-strand symmetry is positively correlated with organismal complexity (measured by number of different cell types). Scatterplot was generated using the ggplot2 R package[33]. The top of each scatterplot shows the GC content of fragments. The R-squared and significance are marked at the top-right corner of scatter plots (*P* value: **P*<0.05, ***P*<0.01, ****P*<0.001). Fourteen organisms were used in this study: Aga (Anopheles gambiae), Bta (Bos taurus), Cfa (Canis familiaris), Ciona intestinalis (Cin), Dme (Drosophila melanogaster), Dre (Danio rerio), Gga (Gallus gallus), has (Homo sapiens), Mmu (Mus musculus), Ptr (Pan troglodytes), Rno (Rattus norvegicus), Tni (Tetraodon nigroviridis), Tru (Takifugu rubripes), and Xtr (Xenopus tropicalis).

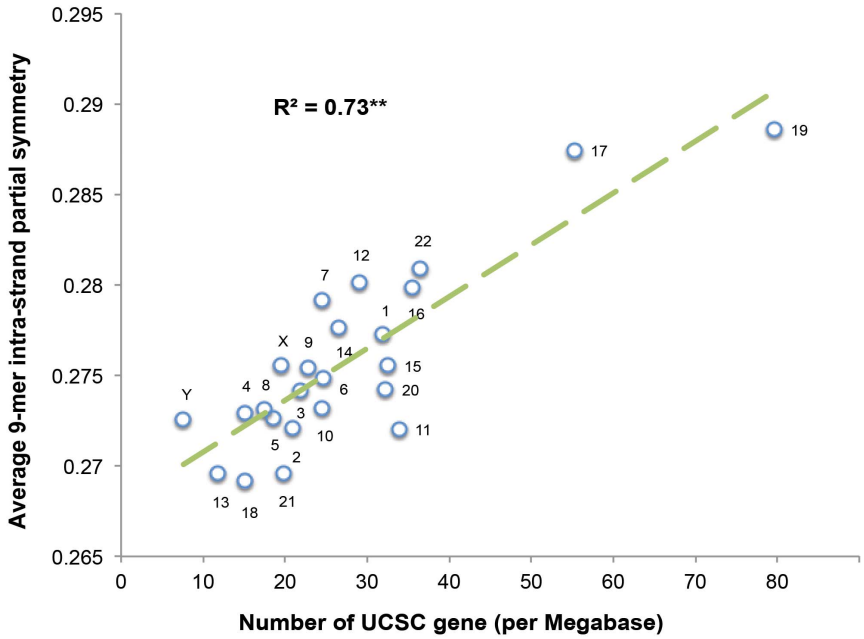

Figure 3 | Positive correlation between high order intra-strand symmetry (9-mer) and density of UCSC genes in human chromosomes. The GC content of the fragments is in the range from 0.4 to 0.6.

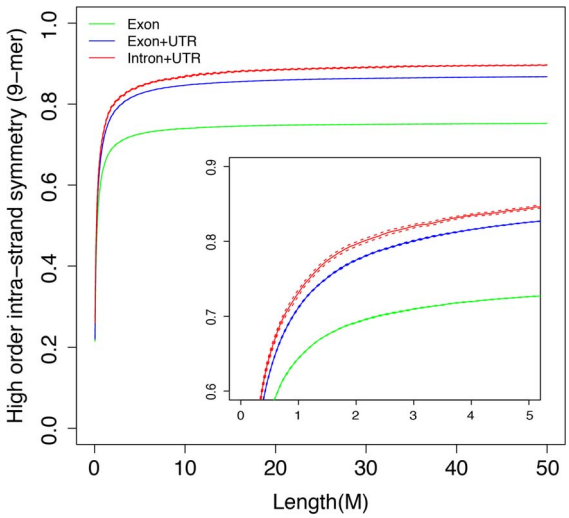

Figure 4 | Plot shows the mean of high order intra-strand partial symmetry (9-mer) for each pooled datasets. The mean of each data were calculated by subsampling each data 10 times, and the strand error of IPS values for each dataset are shown by the dashed lines.

variation increasing the symmetry will tend to be fixed on long-term evolution. As shown in Figure 2, the average IPS values between Homo sapiens and Pan troglodytes is quite similar in each GC content, suggesting the contribution of different nucleotide composition between the two replicating strands to the IPS should be limited. One of the important contributions of increasing symmetry is the enrichment of homogenous feature in sequence, which should be made advantage of the productivity of primary secondary and tertiary structure in both DNA and RNA.

During the course of evolution, the organismal complexity of an organism is not well related with its total number of genes based on the G-value paradox[23]. The regulatory potential of coding DNA is not sufficient to give reasons for the evolution of complex organisms. In contrast to the limited diversity of proteins in phylogeny, the functional non-coding DNA, which has much greater chemical versatility to interact with other molecules easily, can be given the new role to dramatically improve the complicated regulatory framework in complex organism[24,25]. The organismal complexity can be attribute by the expansion of functional non-coding genome sequences, at least for the first approximation[26,27]. For example, the length of 3′ and 5′ untranslatable regions has expanded with the evolution of complex organisms, particularly in animals, suggesting an increase in cis-acting regulatory sequences that control translation and mRNA stability[28,29].

Many clues show the emergence of the functional non-coding DNA can increase the organismal complexity. Our results show the IPS is well correlated with the increasing of functional non-coding DNA, providing clues to the evolution of organismal complexity (Figure 3, 4). It has also been proved that the upstream region of coding DNA can promote symmetry better than coding region[11]. Since functional non-coding DNA does not suffer from codon choices, the variations in functional non-coding DNA could be kept more easily to increase the secondary structure and get higher gene expression level. Based on the stable secondary structure, the functional DNA could not only fold as regular secondary structure to prevent degraded, but also construct a huge amount of complex conformations for further biological functions. This process can be driven by the genome duplication events[30].

To our knowledge, it is the first time for us to report that the high order IPS is strongly correlated with organismal complexity with animal evolution. More importantly, we proved that the functional non-coding DNA shows higher IPS value in comparison with coding DNA. The genomes of complex organisms usually have abundant of functional non-coding DNAs, which coordinate with the conclusion that high complex organisms have high IPS values. During the organism evolution process, organisms are producing new cell type and expand post-transcriptional regulation level. Therefore, the increasing of functional non-coding DNA can be attributed to the increasing of compositional symmetry, which might be a possible event to explain its increasing of the regulation of gene transcription and post-transcription in complex organisms.

Methods

There are 14 animals used in this study, including *six Mammal* genomes: Homo sapiens (Version: hg19), Pan troglodytes (panTro4), Mus musculus (mm10), Rattus norvegicus (rn5), Bos Taurus (bosTau7), Canis familiaris (canFam3), *fiveVertebrate* genomes: Gallus gallus (galGal4), Xenopus tropicalis (xenTro3), Takifugu rubripes (fr3), Tetraodon nigroviridis (tetNig2), Danio rerio (danRer7),*two Insect genomes*: Drosophila melanogaster (dm3), Anopheles gambiae (anoGam1),and *one Deuterostome* genome: Ciona intestinalis (ci2). For each organism, we extracted complete genome sequences from the UCSC Genome Browser[21]. The organismal complexity information measured by the number of distinct cell types was obtained from the result of a recent research[31]. In addition, we used Perl-based bioinformatics scripts and R language to process the statistical analysis.

In order to ignore the effect of sequence length, we employed a sliding window approach to divide each downloaded chromosome sequence into non-overlapping 50 Kb fragments from the first nucleotide of each sequence. The length was commonly found to obey the Chargaff second parity rule[2]. For each fragment, the high order IPS was calculated using the measurement defined by previously developed algorithm[11], $IPS_N = 1\text{-}(\Sigma_i |f_i - f_i'|)/(\Sigma_i |f_i + f_i'|)$, where f_i is the frequency of the i-th N-mer oligonucleotides in one fragment, and f_i' is the frequency of its reverse complement in the same fragment. In palindromic nucleotides, such as "GAATTC", which contains the same order of oligonucleotides with its reverse complement, we simply split them into two same parts, where f_i is equal to f_i'. Finally, the IPS of N-mer oligonucleotides for each fragment can be determined respectively.

To decrease the impact of other factors, such as the local recombination[5] and dramatically highly repetitive DNA in the Y chromosome sequence[32], we performed three filter steps. Firstly, we discarded fragments containing degenerate bases, which comes to be more similar between f_i and f_i'. Secondly, low complex sequence tends to have intra-strand symmetry, so we discarded fragments with extreme GC content (<0.4 or >0.6) to decrease putative positive. Thirdly, we performed the interquartile range (IQR) to filter outliers. It is described as IQR = Q3 - Q1, where Q1 and Q3 are the first quartile or 25th percentile and the third quartile or 75th percentile, respectively. Any score below the Lower fence (Q1 - 1.5 * IQR) or above the Upper fence (Q3 + 1.5 * IQR) can be considered as an outlier. At last, we only kept organisms with count of fragments greater than 10 within defined GC content of all chromosomes after removing the outliers.

1. Rudner, R., Karkas, J. D. & Chargaff, E. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proc. Natl. Acad. Sci. U.S.A.* **60**, 921–922 (1968).
2. Prabhu, V. V. Symmetry observations in long nucleotide sequences. *Nucleic Acids Res.* **21**, 2797–2800 (1993).
3. Qi, D. & Cuticchia, A. J. Compositional symmetries in complete genomes. *Bioinformatics* **17**, 557–559 (2001).
4. Lobry, J. R. & Lobry, C. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol. Biol. Evol.* **16**, 719–723 (1999).
5. Chen, L. & Zhao, H. Negative correlation between compositional symmetries and local recombination rates. *Bioinformatics* **21**, 3951–3958 (2005).
6. Albrecht-Buehler, G. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 17828–17833 (2006).
7. Albrecht-Buehler, G. Inversions and inverted transpositions as the basis for an almost universal 'format' of genome sequences. *Genomics* **90**, 297–305 (2007).
8. Forsdyke, D. R. & Mortimer, J. R. Chargaff's legacy. *Gene* **261**, 127–137 (2000).
9. Forsdyke, D. R. A stem-loop 'kissing' model for the initiation of recombination and the origin of introns. *Mol. Biol. Evol.* **12**, 949–958 (1995).
10. Zhang, S. H. & Huang, Y. Z. Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA. *Bioinformatics* **26**, 478–485 (2010).
11. Baisnée, P.-F., Hampson, S. & Baldi, P. Why are complementary DNA strands symmetric? *Bioinformatics* **18**, 1021–1033 (2002).
12. Rapoport, A. E. & Trifonov, E. N. Compensatory nature of Chargaff's second parity rule. *J. Biomol. Struct. Dyn.* (2012) doi:10.1080/07391102.2012.736757.
13. Powdel, B. R. *et al.* A Study in Entire Chromosomes of Violations of the Intra-strand Parity of Complementary Nucleotides (Chargaff's Second Parity Rule). *DNA Research* **16**, 325–343 (2009).
14. Nikolaou, C. & Almirantis, Y. Deviations from Chargaff's second parity rule in organellar DNA. *Gene* **381**, 34–41 (2006).

15. Guo, F.-B. Replicating strand asymmetry in bacterial and eukaryotic genomes. *Curr. Genomics* **13**, 2–3 (2012).
16. Touchon, M. *et al*. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9836–9841 (2005).
17. Lorente-Galdos, B. *et al.* Accelerated exon evolution within primate segmental duplications. *Genome Biol.* **14**, R9 (2013).
18. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475–479 (2013).
19. Trotta, E. Selection on codon bias in yeast: a transcriptional hypothesis. *Nucleic Acids Res.* **41**, 9382–9395 (2013).
20. Geisberg, J. V., Moqtaderi, Z., Fan, X., Ozsolak, F. & Struhl, K. Global Analysis of mRNA Isoform Half-Lives Reveals Stabilizing and Destabilizing Elements in Yeast. *Cell* **156**, 812–824 (2014).
21. Meyer, L. R. *et al*. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* **41**, D64–9 (2013).
22. Petrov, D. A. Evolution of genome size: new approaches to an old problem. *Trends Genet.* **17**, 23–28 (2001).
23. Hahn, M. W. & Wray, G. A. The g-value paradox. *Evol. Dev.* **4**, 73–75 (2002).
24. Mattick, J. Video Q&A: Non-coding RNAs and eukaryotic evolution - a personal view. *BMC Biol.* **8**, 67 (2010).
25. Liu, G., Mattick, J. S. & Taft, R. J. A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle* **12**, 2061–2072 (2013).
26. Prasanth, K. V. & Spector, D. L. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes & Development* **21**, 11–42 (2007).
27. Taft, R. J., Pheasant, M. & Mattick, J. S. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29**, 288–299 (2007).
28. Chen, C. Y., Chen, S. T., Juan, H. F. & Huang, H. C. Lengthening of 3'UTR increases with morphological complexity in animal evolution. *Bioinformatics* **28**, 3178–3181 (2012).
29. Vinogradov, A. E. & Anatskaya, O. V. Organismal complexity, cell differentiation and gene expression: human over mouse. *Nucleic Acids Res.* **35**, 6350–6356 (2007).
30. Spring, J. Genome duplication strikes back. *Nat Genet* **31**, 128–129 (2002).
31. Vogel, C. & Chothia, C. Protein family expansions and biological complexity. *PLoS Comput Biol* **2**, e48 (2006).
32. Hoskins, R. A. *et al*. Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science* **316**, 1625–1628 (2007).
33. Ginestet, C. ggplot2: Elegant Graphics for Data Analysis. *Journal of the Royal Statistical Society: Series A (…* (2011).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61227803); and the Natural Science Foundation of Jiangsu Province of China (BK2012331). We thank Chunpeng He for helpful advice.

Author contributions

Z.L. and S.W. conceived of the study and wrote the main manuscript text. S.W. and J.T. performed the data analysis. Z.L., Z.J. and S.W. advised and revised the manuscript text. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Wang, S., Tu, J., Jia, Z. & Lu, Z. High order intra-strand partial symmetry increases with organismal complexity in animal evolution. *Sci. Rep.* **4**, 6400; DOI:10.1038/srep06400 (2014).