SCIENTIFIC REPORTS

Gene direction in living organisms

Xiu-Qing Li[1] & Donglei Du[2]

[1]Molecular Genetics Laboratory, Potato Research Centre, Agriculture and Agri-Food Canada, 850 Lincoln Road, Fredericton, NB, E3B 4Z7, Canada, [2]Quantitative Methods Research Group, Faculty of Business Administration, University of New Brunswick, 7 Macaulay Lane, Fredericton, NB, E3B 5A3, Canada.

SUBJECT AREAS:
MOLECULAR EVOLUTION
BIOINFORMATICS
GENOMICS
MOLECULAR BIOLOGY

Received
27 July 2012

Accepted
24 September 2012

Published
21 December 2012

Correspondence and requests for materials should be addressed to X.-Q.L. (xiu-qing.li@agr.gc.ca; lixiuqing2008@gmail.com)

Gene direction, which is important for function, has not been subjected to statistical testing for randomness and for the degree of evolutionary changes. We analyzed 747 sequenced species and 2,061 genomes/chromosomes and detected clear differences in gene direction between kingdoms. All the archaeans, bacteria, and protozoa analyzed have genes characterized mainly by same-direction neighbors (i.e., in head-to-foot or foot-to-head order), with up to 391 genes in tandem in protozoan *Leishmania infantum*. Fungi and photosynthetic protists have genes characterized by opposite-direction neighbors, except chromosome VII of *Ashbya gossypii*, a progenitor fungus. The gene direction analysis suggests that the same-direction dominance originated from the last common ancestor of these living organisms, then was strengthened in protozoa, but weakened or lost in fungi, photosynthetic protists and some plants/animals, giving chromosomes/genomes with gene opposite-direction dominance (i.e., towards the random use of both DNA strands).

Neighborhood conservation of gene arrangement was found in various bacteria[1] and eukaryotic organisms[2–8] from studying specific species or a group of genes. Gene direction is important for gene arrangement and function. Since random arrangement of a large number of genes along the chromosomes can theoretically generate a multiplicity of gene direction orders, a statistical test of gene direction randomness is required. To the best of our knowledge, however, there are no literature reports on algorithms suitable for testing gene direction randomness, likely because of the lack of a readily available algorithms for testing whether a series of two numbers or two letters (e.g., 1 for forward, 2 for backward) is random. Research is needed to develop a statistical algorithm to test gene direction randomness and to analyze many genomes for general information on gene direction distribution.

Genes with similar function or coordinated expression seem to be clustered in sequenced genomes[3]. Furthermore, the order of transcriptionally and functionally linked genes was found to be conserved in some eukaryotes, in a study using various analysis methods, including protein sequence BLAST searches, gene ontology assignments, and phylogenetic tree reconstruction[4]. It has been proposed that the range for which DNA neighborhood optimizes biochemical interactions might therefore be defined by DNA topology[1]. Recently, the notion that expression neighborhoods are a feature of eukaryotic genome organization necessary for correct gene expression was publically challenged because a targeted separation of one well-defined gene expression neighborhood in the Drosophila genome did not significantly alter gene expression[9]. Since gene direction order is an important aspect of gene order and architecture, an analysis of the gene direction in a large number of genomes may provide insights into whether gene neighborhoods are random, or likely the result of selection and inheritance.

In this study, we developed a statistical approach to test the significance of gene direction order. Since an intergenic region can have four possible configurations, that is, FF, BB, FB, and BF, where F denotes forward gene direction and B backward gene direction (i.e., on the complementary strand), the probability of occurrence of these four types of intergenic regions should be approximately equal if gene order on the annotated DNA sequence of a chromosome is random. The chi-square test approach can test the randomness of these four configurations. We tested the randomness of the direction of annotated genes on chromosomes (GenBank full version files; see **Tables S1–S7** for sequence ID list) of all or nearly all complete and annotated genomes of bacteria, archaeans, protists, fungi, plants, and animals available in NCBI GenBank (http://www.ncbi.nlm.nih.gov/), and present the findings below.

Results

Gene direction was not statistically random in any of the 63 archaean (**Supplementary Table S1**), 631 bacterial (**Supplementary Table S2**), 9 protist-protozal species (**Supplementary Table S3**) and a total of 1,127 genomes

Table 1 | Summary of gene direction arrangement, inferred from gene interval distribution, at the species and chromosomal (chr) levels

Kingdom	No. of species	Species with both chr types[a]	Species with only random gene direction chrs[b]	Total genomes or chrs[c]	Chrs (FF+BB) > (FB+BF) (%)	Chrs (FB+BF) ⩾ (FF+BB) (%)
Archaea	63	0	0	78	100	0
Bacteria	631	0	0	898	100	0
Protista: protoza	9	0	0	151	85.33	14.67
Protista: Chlorophyta	2	2	0	38	0	100
Plantae	5	3	0	53	86.79	13.21
Fungi	21	9	4	181	17.68	82.32
Animalia	16	13	0	662	52.11	47.89
Overall	747	27	4	2,061	63.13	36.87

[a]:The number of species that have both random and non-random gene-direction chromosomes.
[b]:All chromosomes are random in gene direction based on ChiTest with $P<0.01$.
[c]:Complete chromosomes are included in the analysis, but for the scaffolds and contigs, we include here only the ones that are annotated and are larger than 0.5 Mb.

analyzed (**Table 1**). Archaea and bacteria have only non-random gene direction chromosomes; while a majority of the fungi, chlorophyta protists, plants, and animal species have both random and nonrandom gene direction chromosomes (**Table 1**).

All of the analyzed genomes of archaea, bacteria, protista and protozoa have a greater number of same-direction gene pairs than opposite-direction pairs; in other words, they all have neighbors mainly characterized by the same direction. The same/opposite gene direction ratios of chromosomes are approximately 2.74, 2.00, and 46.20, on average, among the bacteria, archaeans, and protozoa, respectively (**Table 2**). In the protozoa species, 75% of the intervals have genes in the same direction, either on the forward strand or complementary strand; whereas, interestingly, majority of the genes in the chlorophyta and fungi species are in the opposite direction (**Table 2**).

The largest string of same-direction genes, consisting of 391 genes, was found on the complementary strand of *Leishmania infantum* chromosome 31 (NC_009415); there were only two genes in the forward direction (**Supplementary Table S3**). The second largest string, comprising 371 genes, was found on the forward strand on chromosome 26 (NC_007267) of *Leishmania major* strain Friedlin (**Supplementary Table S3**).

An extreme case of opposite-direction genes was found in *Ostreococcus lucimarinus* CCE9901 and *Micromonas* sp. RCC299, two species of Chlorophyta (protists), an early-diverging photosynthetic class within the green plant lineage (**Supplementary Table S4**). The majority (68%) of gene pairs in these two species exhibit opposite direction, either FB or BF (**Table 1**). Each of the 21 *O. lucimarinus* chromosomes and the 17 *Micromonas* sp. RCC299 chromosomes had fewer same-direction gene pairs than opposite-direction pairs. This stands in contrast with the situation for *Leishmania* species.

Four species of fungi (*Debaryomyces hansenii*, *Encephalitozoon cuniculi*, *Saccharomyces cerevisiae*, and *Encephalitozoon intestinalis*) have only chromosomes with randomly distributed gene direction (**Supplementary Table S5**). Another 17 species have fewer same-direction gene pairs than opposite-direction pairs; this is also the case for the average of these 21 fungal species (**Table 1**).

All the fungal chromosomes, except the chromosome VII of *Ashbya gossypii* (ATCC 10895; NC_005788.4), have more opposite gene neighbors (**Supplementary Table S5**). It is interesting that *Ashbya gossypii* has the smallest eukaryote genome and was used as a tool for mapping the ancient *Saccharomyces cerevisiae* genome[10]. Since fungi are less primitive than bacteria, which have same direction dominance of genes, this may indicate that fungi are further along in the progression towards opposite-direction dominance.

In plants, Arabidopsis (*Arabidopsis thaliana*), rice (*Oryza sativa* ssp. Japonica), poplar (*Populus trichocarpa*), and sorghum (*Sorghum bicolor*) have significantly more same-direction gene pairs than opposite-direction ones. However, a diploid yellow-flowered alfalfa (*Medicago truncatula*) was found to have fewer same-direction genes than opposite-direction ones with a (FF+BB)/(FB+BF) ratio, also called the same/opposite ratio, of 0.98 (**Supplementary Table S6**). Overall, for these plants, the (FF+BB)/(FB+BF) ratio per chromosome is 1.15, which means there are more same-direction than opposite-direction genes (**Table 2**).

For animal species, on average, there are statistically more same-direction gene pairs than opposite-direction pairs, but the difference is quite slim with a (FF+BB)/(FB+BF) ratio of 1.07 (**Table 2**). Among the animal genomes analyzed, the genomes of *Caenorhabditis elegans* (nematode) and *Drosophila melanogaster* (fruit fly) have been completely sequenced and annotated. Each of the five *C. elegans* chromosomes has a greater proportion of same-direction gene pairs and the same/opposite ratio is 1.15 on average (**Table 2**). In *D. melanogaster*, chromosome 2R gave a similar result in terms of same/opposite direction, but all the other chromosomes showed significantly fewer same-direction than opposite-direction genes (**Supplementary Table S7**).

Table 2 | Gene direction arrangement, inferred from gene interval distribution, in different kingdoms

Kingdom	Intergenic regions (n)	Same direction gene pairs (FF+BB)[a] (n)	Opposite gene pairs FB+BF (n)	Save direction (%)	Opposite direction (%)	Same/opposite gene ratio per chr[b]
Archaea	185,703	121,854	63,849	66	34	2
Bacteria	5,458,008	3,892,006	1,566,002	71	29	2.74
Protista: protoza	64,368	48,236	16,132	75	25	46.2
Protista: Chlorophyta	17,616	5,553	12,063	32	68	0.52
Plantae	163,598	86,086	77,512	53	47	1.15
Fungi	124,849	55,809	69,040	45	55	0.83
Animalia	208,512	106,349	102,163	51	49	1.07
Overall	6,222,654	4,315,893	1,906,761	69	31	7.79

[a]:F is assigned for the forward gene direction, B for the backward gene direction, and FF, BB, FB, and BF represent the types of intergenic regions. Same/opposite gene ratio: mean of individual chromosome's (FF+BB)/(FB+BF) ratios. [b]:The mean of ratios of same/opposite gene direction per chromosome (See the Supplementary information file for details).

The kingdoms showed clear-cut differences in terms of the same/opposite ratio (FF+BB)/(FB+BF) on chromosomes. More same direction gene pairs than opposite direction ones at the chromosomal level occurred in all 78 archaean chromosomes (genomes), 898 bacterial chromosomes (genomes), in 17.68% of the fungal chromosomes, 86.79% of the plant chromosomes, 85.33% of protozoan chromosomes, and 52.11% of the animal chromosomes (**Table 1**). None of the 38 protista-Chlorophyta chromosomes showed this dominance (**Table 1**).

Overall, 99% of the species (741 out of 747) have at least one chromosome on which gene direction is not random (**Table 1**). However, it is worth noting that some species (i.e., 4 fungi) are characterized by random order of gene direction at the annotated sequences at the chromosome level in their genomes (**Table 1**). In some species, such as alfalfa (*Medicago truncatula*) (**Supplementary Table S6**), zebra finch (*Taeniopygia guttata*), chimpanzee (*Pan troglodytes*), and humans (**Tables S7**), most chromosomes exhibit random gene direction in terms of gene pair configurations.

Discussion

In this study, we examined gene direction randomness at the whole chromosome level; therefore, we cannot rule out that regional non-random islands exist on the random gene-direction chromosomes. Similarly, chromosomes with non-random gene direction can be expected to have regions with random gene direction.

Most plants and some animals have more same-direction gene pairs than opposite-direction gene pairs in their genomes. In view of the fact that some lower kingdoms such as the Fungi and the Protista (protozoa) have already lost same-direction gene dominance, the maintenance of the statistical dominance of same-direction genes in these lower and higher organisms (i.e., fungi, protozoa, plants and some animals) must be attributed to functional advantages. This may correspond to the evolutionary conservation of non-randomness of gene neighborhoods which was reported previously[6]. The data suggest that the tendency in animals is towards randomness of gene direction at the chromosome level. There must be an unknown mechanism in these animal genomes to ensure animal fitness after members of previously defined gene blocks get physically split off. This might explain the observation that neighborhood continuity is not required for correct testis gene expression in *Drosophila*[9].

The same gene direction dominance with non-randomness likely originated from the last common ancestor of living organisms analyzed in this study. This hypothesis is supported by the non-random, same-direction dominance of genes found in the most primitive species (63 archaean species, 631 bacterial species), an evolutionary middle level species (*Ashbya gossypii*, a progenitor fungus), and most higher organisms (plants and some animals). This non-randomness was likely strengthened in archaea, bacteria, and some protist species notably in *Mycoplasma suis* (same/opposite = 6.95) and *Leishmania infantum* (391 genes in tandem), but weakened in many others species, including fungi, chlorophytes, and some plants and animals.

Extra-attention should be given to the interpretation of statistical randomness and its biological meaning. This is because the statistical test is based on the annotated DNA sequence, which is in format of a single strand (from 5′ to 3′) in GenBank, but its information of gene location and direction represents both DNA strands. The gene direction annotation on the DNA sequence is similar to combining all the signs from both sides (parallel but with opposite-traffic flows) of a highway. The same direction dominance of genes in this study is in the same meaning of "non-randomness of gene direction" in the literature. The opposite direction dominance detected in this study is non-random statistically on the annotated genome sequences but is totally opposite to the conventional meaning of non-random in gene direction; it is actually equivalent to nearly the extreme case of conventional randomness. Our interpretation is that the opposite direction dominance is likely created by nearly random use of both DNA strands.

The model we propose here can be used to explain how gene direction evolved from same-direction dominance to opposite-direction dominance in some species, such as Chlorophyta and fungal species. Same-direction dominance was likely needed in earliest life forms to maximize the use of the limited DNA/RNA sequences. As genome size increased, some species and some gene regions developed random gene direction because the opportunity for gene mutation and diversity could complement species' functional needs. The nearly equal use or random use of both strands created new advantages for certain species, and therefore selection for opposite gene direction dominance occurred in some species, allowing both strands to have an approximately equal distribution of genes. The chromosomes that have the annotated sequences with statistically random gene directions is likely at the interim stages on the way toward the opposite direction dominance. Unknown *in trans* mechanisms must exist which ensure that functionally related genes work together effectively in these species after the same direction dominance is lost. Such mechanisms play a greater role in animals than in plants. Although neighborhood continuity is clearly needed to a certain degree, we predict that there will be a trend toward less same-direction dominance and greater opposite-direction dominance in higher organisms, particularly animals, in the future.

In brief, the results of this analysis of the completely sequenced genomes suggest the following: The same gene direction dominance is likely derived from the common ancestor of these living organisms. This dominance is further strengthened in archaeans, bacteria and some protozoa-protists, but weakened in fungi, Chlorophyta, and some plants and animals, likely owing to the increase in genome size and the opportunity to use both strands. Gene direction experienced a V-shape evolution. One branch is from moderately non-random to extremely non-random. In this branch, genes mainly located on one DNA strand. The other branch is from moderately non-random to mainly random. In this second branch, gene locations evolved from one DNA strand to nearly random between both strands. There is an evolutionary shift in gene direction, from predominantly same-direction to opposite-direction or approximately equivalent number of same- and opposite-directions during evolution of more complex species. Functional neighborhood continuity will likely be conserved to a certain degree, but the future trend is likely to be toward increasing opposite-gene direction dominance and decreasing same-direction dominance in most animals. This study expands current knowledge of the genomes of living organisms, and may increase understanding of gene regulation in existing species as well as provide useful insights for designing synthetic genomes.

Methods

Most genomes were downloaded from the NCBI GenBank FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/). Plant genomes were individually searched and downloaded from http://www.ncbi.nlm.nih.gov/nuccore/and the genome browser website http://www.ncbi.nlm.nih.gov/genome/browse/. For the protist Micromonas sp. RCC299 genome, there were two series of IDs, only the series named with CP were used in the final analysis because the other series started by NC_ were unpublished versions and were identical in the analysis output to the CP series. For archaea, bacteria, fungi and protists, only the completed genomes were used. Plant genomes were analyzed if they had complete genomes or pseudomolecules available. The genomes of humans and the majority of the widely studied animals including rat, mouse, chimpanzee, monkey, and dog were not complete but had large scaffolds. Therefore, scaffolds of animal chromosomes larger than 0.5 Mb were also analyzed as long as they had clear indication of chromosome number and the sequences were unique. The GenBank files (GBK, GB, or GBS) of the chromosomes/scaffolds were used if they had clear annotation of gene and coding region locations. The gene direction and location of the chromosomes were counted. The types of gene intervals (i.e., intergenic regions) were determined by direct neighboring genes and classified as FF, BB, FB, and BF, where F is for forward and B is for backward or complement strand. A chi-square test was employed to test whether the four types of intervals were random and to test whether the same direction gene pairs (FF and BB) vs. opposite direction gene pairs (FB and BF) were statistically equal. The counting also included the total number of species analyzed, the species with only random gene direction

chromosomes, the species having both random and nonrandom gene direction chromosomes, the same/opposite ratio of genes on the chromosomes, and the percentage of chromosomes that had more same- than opposite-direction genes.

1. Junier, I., Hérisson, J. & Képès, F. Genomic organization of evolutionarily correlated genes in bacteria: limits and strategies. *J. Mol. Biol.* **419**, 369–86 (2012).
2. Downing, T. *et al*. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res.* **21**, 2143–2156 (2011).
3. Hurst, L. D., Pál, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310 (2004).
4. López, M. D., Guerra, J. J. M. & Samuelsson, T. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS ONE* **5** (2010).
5. McDonagh, P. D., Myler, P. J. & Stuart, K. The unusual gene organization of *Leishmania major* chromosome 1 may reflect novel transcription processes. *Nucleic Acids Res.* **28**, 2800–2803 (2000).
6. Michalak, P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* **91**, 243–248 (2008).
7. Ivens, A. C. *et al*. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**, 436–442 (2005).
8. Peacock, C. S. *et al*. Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nat. Genet.* **39**, 839–847 (2007).
9. Meadows, L. A., Chan, Y. S., Roote, J. & Russell, S. Neighbourhood continuity is not required for correct testis gene expression in Drosophila. *PLoS Biol.* **8**, e1000552 (2010).
10. Dietrich, F. S. *et al*. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**, 304–307 (2004).

Acknowledgements

This research was supported by XQL's AAFC research funding.

Author contributions

XQL conceived, designed, and executed the experiments, analyzed the data, and wrote the manuscript. DL wrote and tested the PERL program. All authors reviewed, edited and finalized 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: Li, X.Q. & Du, D. Gene direction in living organisms. *Sci. Rep.* **2**, 982; DOI:10.1038/srep00982 (2012).