Sun et al. BMC Evolutionary Biology 2010, 10:49
http://www.biomedcentral.com/1471-2148/10/49

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Gene duplication in the genome of parasitic *Giardia lamblia*

Jun Sun[1,2†], Huifeng Jiang[1,3†], Roberto Flores[3], Jianfan Wen[1*]

Abstract

Background: *Giardia* are a group of widespread intestinal protozoan parasites in a number of vertebrates. Much evidence from *G. lamblia* indicated they might be the most primitive extant eukaryotes. When and how such a group of the earliest branching unicellular eukaryotes developed the ability to successfully parasitize the latest branching higher eukaryotes (vertebrates) is an intriguing question. Gene duplication has long been thought to be the most common mechanism in the production of primary resources for the origin of evolutionary novelties. In order to parse the evolutionary trajectory of *Giardia* parasitic lifestyle, here we carried out a genome-wide analysis about gene duplication patterns in *G. lamblia*.

Results: Although genomic comparison showed that in *G. lamblia* the contents of many fundamental biologic pathways are simplified and the whole genome is very compact, in our study 40% of its genes were identified as duplicated genes. Evolutionary distance analyses of these duplicated genes indicated two rounds of large scale duplication events had occurred in *G. lamblia* genome. Functional annotation of them further showed that the majority of recent duplicated genes are VSPs (Variant-specific Surface Proteins), which are essential for the successful parasitic life of *Giardia* in hosts. Based on evolutionary comparison with their hosts, it was found that the rapid expansion of VSPs in *G. lamblia* is consistent with the evolutionary radiation of placental mammals.

Conclusions: Based on the genome-wide analysis of duplicated genes in *G. lamblia*, we found that gene duplication was essential for the origin and evolution of *Giardia* parasitic lifestyle. The recent expansion of VSPs uniquely occurring in *G. lamblia* is consistent with the increment of its hosts. Therefore we proposed a hypothesis that the increment of *Giradia* hosts might be the driving force for the rapid expansion of VSPs.

Background

Giardia are a group of flagellated unicellular protists which are the most common infective parasites of a number of vertebrates. For example, *G. lamblia* is a common human parasite. In the United States, about 20,000 cases of giardisis are reported each year [1]. Aside from being a prevalent pathogen, in the last two decades *G. lamblia* has caught a lot of attentions, as being the most primitive eukaryotes [2]. Phylogenetic and cellular evidence indicate that this organism might branch away from the ancestor of extant eukaryotes around the endosymbiotic origin of mitochondria in eukaryotes [2-5]. Therefore before the emergence of multicellular animals, *G. lamblia* may have survived freely in the world for several hundred million years [6]. It suggested that later on it developed the ability to successfully parasitize vertebrates, as it is now recognized as one of the most prevalent intestinal parasites in a variety of vertebrates from amphibians to mammals [7]. An intriguing question is how this ancient eukaryote became an obligate parasite in the later multicellular animals. The draft genome sequence of *G. lamblia* provides us an opportunity to uncover what genomic features resulted in its parasitic lifestyle [8]. Comprehension of how the parasitic ability developed would not only be of evolutionary biological significance, but also shed light on the mechanism of giardisis.

Genetic novelties emerge in organisms by creation of new genes through three major mechanisms: de novo creation, lateral gene transfer and gene duplication [9,10]. The Origin of new genes de novo in *G. lamblia*

* Correspondence: wenjf@mail.kiz.ac.cn
† Contributed equally
[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences (CAS), Kunming, Yunnan 650223, PR China

is impossible to detect because of the deficiency of close relatives in the lineage. Lateral gene transfer (LGT), which is a predominant force of acquisition of new genes in many microorganisms [11], may play an important role for the adaptive evolution of *G. lamblia* in animal intestines because half of the 15 LGT genes identified are associated with its surveillance in an anaerobic environment [12]. However, gene duplication, which has long been thought to be the primary mechanism in producing resources for the origin of evolutionary novelties, has not yet been thoroughly studied in *G. lamblia*. The most obvious contribution of gene duplication to organisms is that it provides genetic material to generate neo-function or sub-function while maintaining the original function of duplicated genes [9,13]. Moreover, the generation of duplicated genes can increase genetic robustness within cellular networks [14]. The dynamic evolution of duplicated genes inflects adaptive evolution of organisms under varying environments [15,16]. Therefore there is no doubt that gene duplication is extremely pervasive, conducting function in almost all organisms from prokaryotes to eukaryotes [9,13].

Previously, many studies on fungi, plants and animals have shown that gene duplication contributes novelties for their adaptive evolution [17-21]. In order to investigate the impact of gene duplication on the parasitic lifestyle of *G. lamblia*, we surveyed and depicted the evolutionary relationships of all the duplicated genes in its genome. Our results showed that two rounds of large scale duplication events took place in the evolutionary process of *G. lamblia*. Furthermore, most of the recent duplicated genes in the second round duplication events are VSP genes, which are essential for the parasitic properties of *G. lamblia* that utilizes them to evade the host's immune response [22]. Largely expanded VSP genes are helpful to parasitize a variety of hosts, because they allow *G. lamblia* to enact a more complex regulation of VSPs in different hosts and at different times [23-25].

Results

Identification of Gene duplication events in *G. lamblia*

Gene duplication and subsequent divergence of the duplicated copies provide opportunities to generate neo-function for adaptive evolution of organisms in varying environments. Identification of duplicated genes is valuable to understand the adaptive evolution of organisms. In order to identify all putative gene duplication events in *G. lamblia*, a very ancient organism which have survived in the world for several billion years [6], a global survey of protein similarities was conducted using the BLASTP program with loose parameters (E-value < 10^{-4}) [26]. After all-against-all alignments for the entirety of proteins in *G. lamblia*, we detected 2,403 duplication genes which cover about 40% of the total proteins in *G. lamblia*. Duplicated genes usually can be classified into tandem, segmental and dispersed duplicates. According to the location of duplicated genes in the assembled contigs, we found only 23 tandem duplicated genes whose original gene and duplicated copy are tandemly located on the same contig. Additionally eight genes were involved in segmental duplication events, which resulted in two or more duplicated genes located in the same contig (Additional file 1: The list of 2,403 duplicated genes). These results imply that the majority of duplication events happened dispersedly.

In order to further analyze the evolutionary scenario of these duplication events, subsequently, a cluster analysis was done based on sequence divergence of the duplicated genes. Briefly, we constructed a matrix with 2,403 rows and columns where each represented a duplicated gene. The amino acid similarities for each gene pair in the row and column were then extracted from the BLASTP results. Based on this matrix, a hierarchical clustering was conducted by the program AGNES (agglomerative hierarchical clustering algorithms) where proteins are clustered closer if their protein sequences have higher similarity (Materials and Methods) [27]. Interestingly, as shown in figure 1, roughly 30% of the duplicated genes were classified into two large duplicated groups: 235 genes in Group I and 500 genes in Group II. The rest of the 1,668 genes formed many small gene groups. A control dataset which contains these 1,668 genes (Group III) was also constructed, in order to see if the evolutionary pattern of the two large groups is different from these small groups. Comparison of average similarities among genes in the three groups showed that the average similarity for genes from Group I (35.26 ± 0.09) is significantly higher than those from Group II (31.32 ± 0.02) and III (29.03 ± 0.18). This implies perhaps the majority of duplication events in Group I took place more recently than those in Groups II and III did.

Two rounds of large scale duplication events happened in *G. lamblia*

In order to further clarify the evolutionary order of the duplication events in each group, we defined the best hit for each duplicated gene as its direct parental gene (Materials and Methods). Due to a large divergence in sequence, it is impossible to detect the parental genes for some ancient duplicated genes. Fortunately, we identified 1,907 pairs of parent-daughter relationships among all duplicated genes in *G. lamblia*, but the rest of the 496 duplicated genes lacked detectable parental genes. To gauge the evolutionary distances for each parent-daughter pair, we used non-synonymous distance

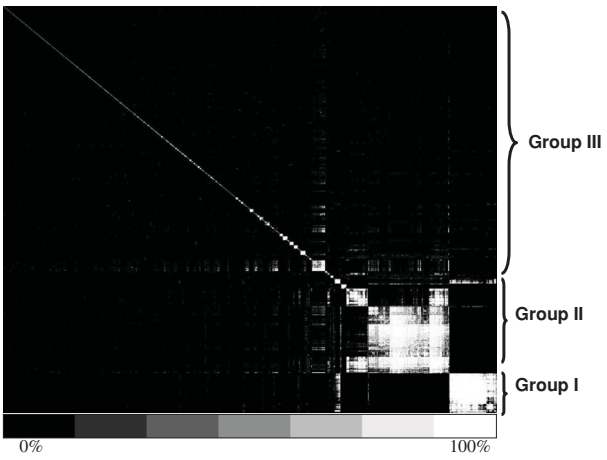

Figure 1 Cluster analysis of duplicated genes. The similarities of duplicated genes were denoted with different colors. As the bottom bar showed, the black means the similarity value is 0% and the white means the similarity value is 100%. Group I and Group II were defined based on the cluster results. As a control for comparing analysis, the rest proteins were put together as Group III.

(dN) and synonymous distance (dS) [28,29] (See Additional file 1 for the list of dN and dS value for all duplicated pairs). As shown in figure 2A, most genes in Group I were created very recently compared with the genes in Group II and III. For example 56% of the genes in Group I have dS smaller than 1, whereas the genes with such a low dS in Group II and III are only 11% and 17%, respectively. Thus, more than half of the duplicated genes in Group I were originated recently. Since the larger dS values for most duplicated genes in Group II and Group III might result from reverse mutations at synonymous sites, the evolutionary distances for a large number of duplicated genes in Group II and III cann't only be based on dS. As a result, we turned to dN.

Although positive selection would accelerate non-synonymous mutation rate in a short period, in the long term, especially for duplicated genes in Group II and III with such a large divergence, dN is possible to use as an index of evolutionary time. Therefore we used dN to characterize the duplication events within the three groups. Unexpectedly, the distribution of dN values for the totality of duplicated genes clearly shows that there were two rounds of gene duplication events in the genome of *G. lamblia* (Figure 2B): the first round of duplication events that were mainly enriched in Groups II and III happened earlier, while the second round of events focusing on Group I occurred very recently. Larger dN values in Groups II and III maybe the result of functional relaxation of the genes within both groups. However functional relaxation of genes in Groups II and III is not enough to explain why the dN values in both groups are higher than those in Group I, due to no significant difference between their dN/dS ratio (*P*-value of t-test is 0.5 between dN/dS from Group I and II and 0.15 between Group I and III). Based on the distribution of dN values, we arbitrarily defined two types of duplicated genes in each group: recent duplicated gene (RDG) with dN < 0.5 and ancient duplicated gene (ADG) with dN > 0.5. Thus 68.5% of the genes in Group I belongs to RDG. This proportion is much higher than those in Group II (13.4%) and III (16.5%). Therefore no matter what methods we utilize, duplicated genes in Group I seem much younger than those in Groups II and III.

Recently duplicated genes in *G. lamblia* are significantly biased towards VSP genes

The current adaptation of an organism relies on recent genomic contents. We are interested in testing if the recently duplicated genes in Group I are related to the parasitic life in *G. lamblia*. In order to do this, we first annotated functional domains for duplicated genes in each group using the Pfam database (Materials and methods). For the two large duplicated gene groups, 86% of the duplicated genes in Group I were annotated as VSP function, while in Group II 72% of the genes were annotated as Ank function and 27% as Pkinase (See Additional file 1 for functional domain annotation). As we mentioned above, more recent duplication genes are rich in Group I than those in Group II and III. By counting RDG and ADG in each group, we found that in Group II more than 80% of Ank genes and Pkinase belong to ADG, while 74% VSP genes belong to RDG in Group I.

Secondly, in order to study the functional distribution of duplicated genes in detail we clustered all proteins from *G. lamblia* into different gene families by TribeMCL [30]. We checked the proportion of RDG and ADG in each gene family to see if functional bias occurred between RDG and ADG. As shown in figure 3, functional bias between RDG and ADG is obvious in most of the gene families (here only families with more than 4 members were listed, Additional file 2 listed the ratio of RDG and ADG in each gene family). Many gene families with important functions had been duplicated long before such as: the Ank domain which possibly plays roles for localization in cells, the Pkinase which may be functional in signal transduction and protein degradation. Interestingly, two types of motor proteins (Kinesin and Dynein) were also significantly enriched in the ADG. It may be possible that such an enrichment of motor proteins were related to the earlier adaptation of *G. lamblia*, which moved in water by its flagella [31,32]. In addition, six out of the nine gene families enriched in RDG were annotated as hypothesis proteins and the

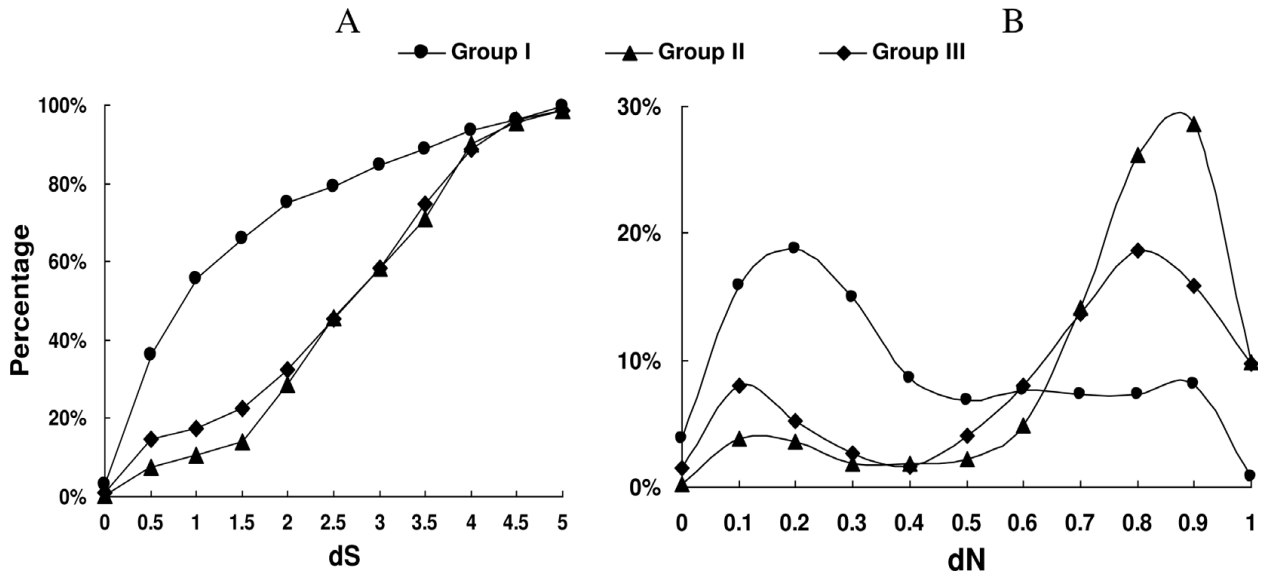

Figure 2 Distribution of evolutionary distances for duplicated gene. A) Cumulative distribution of dS for duplicated genes in Group I, II and III. B) dN distribution of duplicated genes in *G. lamblia*. X axis means values for evolutionary distances (dS and dN) and Y axis displayed the percentage of duplicated genes.

additional three families were Ank, Copine and VSP. Although Ank function is also enriched in RDG, there are only four members in the family, as well as in Copine. It was noticed that six gene families were annotated as hypothesis proteins. Further functional studies of these families would provide more valuable insights into the recently adaptive evolution of *G. liamblia*. Most of the VSPs were created very recently, which is consentient with our results above. Many studies have shown that VSPs play a profound role in antigenic variation of *G. lamblia*, which expresses only one of the VSP genes at a particular time and spontaneously switches to varying VSP genes every 6-13 generations [23]. Therefore numerous duplications of VSPs are indispensable for *G. lamblia* to be able to infect a variety of hosts.

The rapid expansion of VSPs in *G. lamblia* may be consistent with the evolutionary radiation of placental mammals

Evolutionary arms-race is an important driving force for the adaptive evolution [33-35]. The most important thing for a successful parasite in this race is to develop a mechanism to allow antigenic variation to escape arrest from the host immune system. VSPs are an essential gene family in *G. lamblia* to carry out antigenic variation. Therefore, we asked if the rapid expansion of VSPs in *G. lamblia* is associated with the evolution of its hosts. Comparison of the evolution rate of VSPs in *G. lamblia* with the divergence of its hosts may provide us some valuable insights. As the host range of *Giardia* extends from amphibians to mammals, we identified orthologous relationships for all proteins from human to mouse, platypus and fish by InParanoid [36]. We constructed a phylogenetic tree based on the VSP homologs (Materia and Methods. Additional file 3: the phylogenetic tree). Based on amino acid similarities, we found that about 60% of VSPs have higher amino acid similarities than the VSP homologs in human and platypus, while 78% of VSPs have higher amino acid similarities between VSP homologs in human and fish.

Many recent duplicated genes underwent positive selection, which would accelerate the evolutionary rate of proteins [37]. dS is presumably considered to be neutral during evolutionary process [37,38]. Thus, we compared the dS values of VSPs with the synonymous substitution rates of orthologous genes from fish to human. As shown in figure 4, about half of the VSPs have smaller dS values than the average dS values of orthologs between human and mouse, and approximately 80% of such duplicated genes have a dS value smaller than the average dS between human and platypus. Although the evolutionary rates in unicellular organisms is more rapid than multi-cellular organisms, for the reason that unicellular organisms usually have shorter generation time, it would be conservative to conclude that VSPs were rapidly expanded in *G. lamblia* after the period in which platypus separated from the ancestor of human and mouse. This time scale is almost consistent with the evolutionary radiation of placental mammals [39-41]. Therefore, it is available to propose a hypothesis that the increment of *Giardia* hosts is the driving force for the rapid expansion of VSPs.

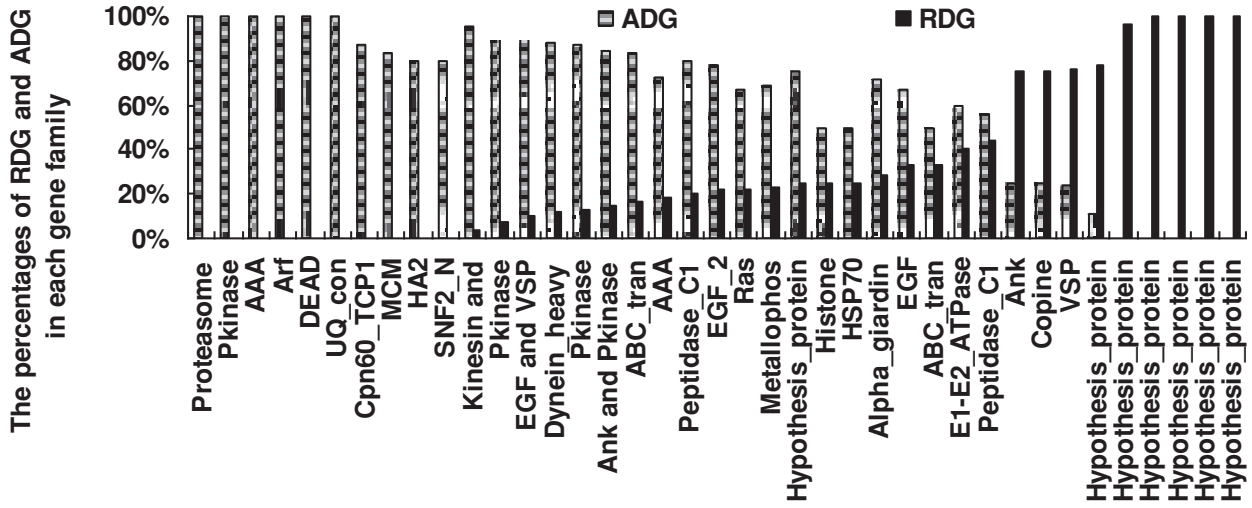

Figure 3 Rates of RDG and ADG in functional categories. The proportions of RDG and ADG in each family with more than 4 duplicated genes in *G. lamblia* were displayed. X axis means functional categories and Y axis denotes the proportion.

Discussion

Gene duplication is one of important mechanisms to provide neo-function for adaptive evolution. Systematic surveys of gene duplication in the intestinal parasite *G. lamblia* provided much more information than what we had previously expected. Although many biological machines in *G. lamblia* are considerably compact [8], more than 40% of the genes in *G. lamblia* were identified as duplicated genes in our analysis. This proportion of duplicated genes is similar with those in fly and yeast using the same parameters for BLASTP [42]. Interestingly, we found that a large number of duplicated genes were focused within two large duplicated groups. Further analysis of the two large duplicated groups indicated that there were two rounds of large scale gene duplication events in the evolutionary process of *G. lamblia*. Based on dN values of identified duplication genes, the first round of duplication events happened a long time ago. Due to the extended time of divergence among these duplicated genes, dS and even dN values would be saturated by mutations. Therefore, perhaps the accumulation of ancient duplication genes might result from a saturation of mutations at non-synonymous sites. However, the dS for most of the genes from the second round duplication events have a relatively small dS value (< 1), which could fall short of the saturation of mutations [29] (Additional file 4: the distribution of dS for the ancient and recent duplicated genes). Gene conversion would also result in very high similarities among duplicated genes [37]. Nevertheless, based on current knowledge, the mechanism for gene conversion is still unresolved in *G. lamblia* [43]. Thus, these recent duplication events might be authentic and correspond to the adaptive evolution in *G. lamblia*.

Antigenic variation is an essential mechanism allowing parasitic pathogens to escape from arrest of the host immune system. *G. lamblia* performs this variation by changing the expression of its VSP genes in a variety of hosts at different time points [23,24,44-46]. Our results showed that 74% of the genes in the second round of duplication are VSP genes. Furthermore, in comparison with other parasitic protists, we found that VSP genes expanded independently in *G. lamblia* genome (Additional file 5: gene number in each family in the five studies parasitic protists), even when considering its very close relative *Spironucleus salmonicida* [40]. A probable explanation for the recently rapid expansion of VSP genes in the genome of *G. lamblia* is that the dramatic expansion was driven by the selection of evasion from the host immune systems. Since most of the hosts for *G. lamblia* are animals, we analyzed the possible relationships between the evolution of VSPs and the evolutionary rates for orthologous genes from fish to human. Given the shorter generation time in unicellular organisms compared to the multi-cellular organisms, we inferred that at least VSP genes became expanded in *G. lamblia* after the separation of platypus from the ancestor of placental mammals. At the same time species in mammals also expanded after the divergence from platypus [39-41]. Our results indicate an interesting co-evolution pattern for the parasitic *G. lamblia* in mammals' evolutionary processes. Further, analysis at the genomic level would shed more lights on the understanding of the co-evolution between the parasite and hosts.

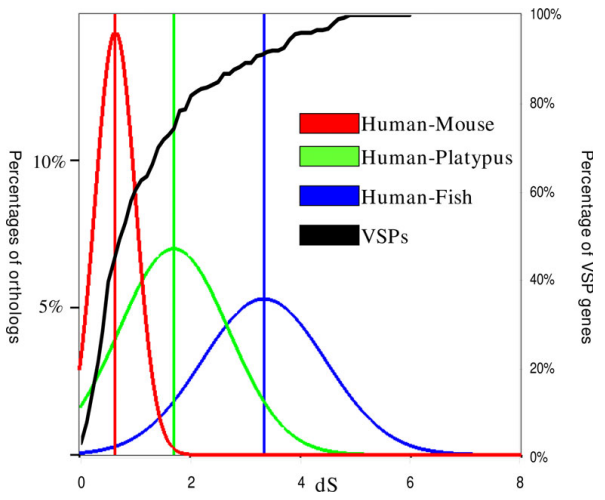

Figure 4 Comparison of dS for duplicated genes in *G. lamblia* with orthologs among animals. The red curve represents distribution of dS of orthologs between Human and Mouse; the green curve denotes distribution of dS of orthologs between Human and Platypus; the blue one displays distribution of dS of orthologs between Human and Zebrafish. The black curve is the cumulative distribution of VSP genes in *G. lamblia*.

Conclusion

Gene duplication always plays a pivotal role for the adaptive evolution of organisms under changing environments. Although *G. lamblia* is one of the most primitive eukaryotes, the origin of its parasitic lifestyle is not as long as it's surveillance. Global identification of duplicated genes in the genome of *G. lamblia* indicated that gene duplication was essential for the origin and evolution of its parasitic lifestyle. Our results advocated that the recent expansion of VSPs uniquely took place in *G. lamblia*. Comparison of the evolution of VSPs with the divergence of its hosts indicated that the rapid expansion of VSPs is consistent with the increment of its hosts. Therefore we proposed a hypothesis that the increment of *Giardia* hosts is the driving force for the rapid expansion of VSPs.

Methods

Sequence data

Protein sequences for all ORF in *Giardia lamblia* [Gla] were downloaded from GiardiaDB database [47]. Perl script was used to filter overlapped ORF which has 80% overlap on the same strand in a contig with another longer ORF. Finally, 5,986 ORFs were used to do analysis. Protein sequences for protists *Cryptosporidium parvum* [Cpa], *Entamoeba histolytica* [Ehi], *Leishmania major* [Lma] and *Plasmodium falciparum* [Pfa] were downloaded from NCBI database [48]. Protein sequences for Human, Mouse, Platypus and Zebrafish were downloaded from Ensembl [49].

Duplicated genes in *G. lamblia*

In order to detect all possible duplicated genes in *G. lamblia*, all-against-all blast search for 5,986 studied proteins in *G. lamblia* were done by BLASTP program with a loose parameter E-value < 10^{-4}. 2,403 proteins which have significant hits with another protein were defined as duplicated genes. The rest (3,583 genes) are single genes. Protein identities between each duplicated pair were used to do cluster analysis by agglomerative hierarchical clustering algorithms (AGNES) in R cluster package. Briefly, a symmetric matrix with 2,403 rows (each one represents a duplicated gene) and 2,403 columns was constructed. And then the amino acid similarities for each gene pair in the row and column were extracted from the BLASTP results. Based on this matrix, we used average method in AGNES to do cluster. Proteins will be clustered closer if they have higher protein similarities. Group I and Group II were defined based on the cluster results. Apart from Group I and II, the rest of the proteins were put together as Group III.

Evolutionary distance of duplicated genes

In order to estimate evolutionary distance of duplicated genes, we used a reverse-searching method to identify a putative parental gene for each duplicated gene. Briefly, based on BLASTP results, the duplicated pairs with the highest similarity were selected from all of duplicated pairs. For the two copies in each duplicated pair, if the first copy has higher similarities with other genes than the second copy, the first copy would be defined as parental gene for the second copy. After this, the defined daughter gene was removed from all duplicated pairs. Then we iterated this process until we can not find daughter gene at all. Finally, we identified 1,906 duplicated pairs with parental-daughter relationships. The software YN00 in the package PAML was used to estimate the synonymous distance (dS) and non-synonymous distance (dN) for each of the duplicated pairs with parental-daughter relationships[29].

Functional domain annotation and Gene family identification

Functional domains for all of the duplicated genes have been detected based on the Pfam database. The sequences for duplicated genes were used as queries to search the Pfam_fs database by the hmmpfam program in the HMMER package 2.3.2 [50]. The cut-off for the search was chosen at E-value < 0.1 according to the advice of the author for HMMER. Finally 1,767 genes were annotated as containing the known functional domain, 636 genes do not contain the known functional domain and were annotated as Hypothesis proteins. In order to study gene family expansion in *G. lamblia*, Tribe-MCL was used to do family classification among

G. lamblia and other four parasitic protists (*Cryptosporidium parvum, Entamoeba histolytica*, *Leishmania major* and *Plasmodium falciparum*).

Comparative evolution analysis of VSPs

Initially using human genes as queries, we identified all orthologous relationships of genes in human with genes in mouse, platypus and zebrafish. Synonymous substitution rates of one-to-one orthologs were used to estimate divergence among these species. The software YN00 in the package PAML was used to estimate the synonymous distance (dS) for all orthologous pairs [29]. The best hit of VSPs in human, mouse, platypus and zebrafish were identified as VSPs homologs. Then the VSPs gene and its homologs were used to construct the phylogenetic tree. All protein sequences were aligned by MUSCLE [51]. The tree was constructed by Clustalw2 after alignment [52].

Additional file 1: The list of 2,403 duplicated genes. This table listed all identified duplicated genes in *G. lamblia*.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-49-S1.XLS]

Additional file 2: List of the ratio of RDG and ADG in each gene family. This table listed the proportion of RDG and ADG in each identified gene family.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-49-S2.XLS]

Additional file 3: The phylogenetic tree of VSPs and their homologs. The amino acid similarities of VSP homologs were listed. The numbers on each branch show the similarities between human and species in the branch.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-49-S3.PDF]

Additional file 4: The distribution of dS for the ancient and recent duplicated genes. The dS distribution of all proteins in *G. lamblia* including RDG and ADG were depicted in the figure.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-49-S4.PDF]

Additional file 5: Distribution of gene families among five studied parasitic protists. In this table, we presented the distribution of gene family members among five studied parasitic protists.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-49-S5.XLS]

Acknowledgements

We thank two anonymous reviewers for helpful comments on the manuscript. We also thank Haifeng Tian, Hao Shen and Yonghai Jiang for technical assistance and discussion. We thank Crystal Conn for reading our manuscript. This work was supported by 973 program (2007CB815705), the National Natural Science Foundation of China (30830018; 30021004; 30623007), and the Knowledge Innovation Program (KSCX2-YW-R-091) to JW.

Author details

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences (CAS), Kunming, Yunnan 650223, PR China. [2]Graduate School of Chinese Academy Sciences, Beijing 100039, PR China. [3]Division of Nutritional Sciences, Cornell University, Ithaca, NY 14853, USA.

Authors' contributions

JS designed the study. HJ analyzed the data. JS, HJ, RF and JW wrote and revised the manuscript. All authors read and approved the final manuscript.

Received: 20 August 2009
Accepted: 17 February 2010 Published: 17 February 2010

References

1. Hlavsa MC, Watson JC, Beach MJ: **Giardiasis surveillance–United States, 1998-2002.** *MMWR Surveill Summ* 2005, **54(1)**:9-16.
2. Sogin ML, Gunderson JH, Elwood HJ, Alonso RA, Peattie DA: **Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia.** *Science (New York, NY)* 1989, **243(4887)**:75-77.
3. Cavalier-Smith T: **Kingdom protozoa and its 18 phyla.** *Microbiol Rev* 1993, **57(4)**:953-994.
4. Gillin FD, Reiner DS, McCaffery JM: **Cell biology of the primitive eukaryote Giardia lamblia.** *Annu Rev Microbiol* 1996, **50**:679-705.
5. Xin DD, Wen JF, He D, Lu SQ: **Identification of a Giardia krr1 homolog gene and the secondarily anucleolate condition of Giaridia lamblia.** *Molecular biology and evolution* 2005, **22(3)**:391-394.
6. Acquisti C, Kleffe J, Collins S: **Oxygen content of transmembrane proteins over macroevolutionary time scales.** *Nature* 2007, **445(7123)**:47-52.
7. Adam RD: **Biology of Giardia lamblia.** *Clinical microbiology reviews* 2001, **14(3)**:447-475.
8. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best AA, Cande WZ, Chen F, Cipriano MJ, *et al*: **Genomic minimalism in the early diverging intestinal parasite Giardia lamblia.** *Science (New York, NY)* 2007, **317(5846)**:1921-1926.
9. Ohno S: **Evolution by gene duplication.** Berlin, New York,: Springer-Verlag 1970.
10. Long M, Betran E, Thornton K, Wang W: **The origin of new genes: glimpses from the young and old.** *Nature reviews* 2003, **4(11)**:865-875.
11. Marri PR, Hao W, Golding GB: **Gene gain and gene loss in streptococcus: is it driven by habitat?.** *Molecular biology and evolution* 2006, **23(12)**:2379-2391.
12. Andersson JO, Sjogren AM, Davis LA, Embley TM, Roger AJ: **Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes.** *Curr Biol* 2003, **13(2)**:94-104.
13. Zhang J: **Evolution by gene duplication: an update.** *Trends in Ecology & Evolution* 2003, **2003(6)**:292-298.
14. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH: **Role of duplicate genes in genetic robustness against null mutations.** *Nature* 2003, **421(6918)**:63-66.
15. Gilad Y, Przeworski M, Lancet D: **Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates.** *PLoS biology* 2004, **2(1)**:E5.
16. Hittinger CT, Rokas A, Carroll SB: **Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101(39)**:14144-14149.
17. Jiang H, Liu D, Gu Z, Wang W: **Rapid evolution in a pair of recent duplicate segments of rice.** *Journal of experimental zoology Part B* 2007, **308(1)**:50-57.
18. Rice Chromosomes 11 and 12 Sequencing Consortia: **The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications.** *BMC biology* 2005, **3**:20.
19. Piskur J, Rozpedowska E, Polakova S, Merico A, Compagno C: **How did Saccharomyces evolve to become a good brewer?.** *Trends Genet* 2006, **22(4)**:183-186.
20. Jiang H, Guan W, Pinney D, Wang W, Gu Z: **Relaxation of yeast mitochondrial functions after whole-genome duplication.** *Genome Res* 2008, **18(9)**:1466-1471.
21. Zhang J, Zhang YP, Rosenberg HF: **Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey.** *Nature genetics* 2002, **30(4)**:411-415.

22. Lafay B, Sharp PM: **Synonymous codon usage variation among Giardia lamblia genes and isolates.** *Molecular biology and evolution* 1999, **16(11)**:1484-1495.
23. Nash TE, Banks SM, Alling DW, Merritt JW Jr, Conrad JT: **Frequency of variant antigens in Giardia lamblia.** *Exp Parasitol* 1990, **71(4)**:415-421.
24. Nash TE, Mowatt MR: **Variant-specific surface proteins of Giardia lamblia are zinc-binding proteins.** *Proceedings of the National Academy of Sciences of the United States of America* 1993, **90(12)**:5489-5493.
25. Prucca CG, Slavin I, Quiroga R, Elias EV, Rivero FD, Saura A, Carranza PG, Lujan HD: **Antigenic variation in Giardia lamblia is regulated by RNA interference.** *Nature* 2008, **456(7223)**:750-754.
26. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.
27. Struyf A, Hubert M, Rousseeuw PJ: **Integrating robust clustering techniques in S-PLUS.** *Computational Statistics & Data Analysis* 1997, **26(1)**:17-37.
28. Nei M: **Selectionism and neutralism in molecular evolution.** *Molecular biology and evolution* 2005, **22(12)**:2318-2342.
29. Yang Z, Nielsen R: **Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.** *Molecular biology and evolution* 2000, **17(1)**:32-43.
30. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30(7)**:1575-1584.
31. Karp G, Geer Pvd: **Cell and molecular biology: concepts and experiments.** Hoboken, NJ: John Wiley, 4 2005.
32. Elmendorf HG, Dawson SC, McCaffery JM: **The cytoskeleton of Giardia lamblia.** *International journal for parasitology* 2003, **33(1)**:3-28.
33. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG: **Dynamic evolution of the innate immune system in Drosophila.** *Nature genetics* 2007, **39(12)**:1461-1468.
34. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, *et al*: **Gene map of the extended human MHC.** *Nature reviews* 2004, **5(12)**:889-899.
35. Elde NC, Child SJ, Geballe AP, Malik HS: **Protein kinase R reveals an evolutionary model for defeating viral mimicry.** *Nature* 2009, **457(7228)**:485-489.
36. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314(5)**:1041-1052.
37. Li W-H: **Molecular evolution.** Sunderland, Mass.: Sinauer Associates 1997.
38. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science (New York, NY)* 2000, **290(5494)**:1151-1155.
39. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, *et al*: **Resolution of the early placental mammal radiation using Bayesian phylogenetics.** *Science (New York, NY)* 2001, **294(5550)**:2348-2351.
40. Andersson JO, Sjogren AM, Horner DS, Murphy CA, Dyal PL, Svard SG, Logsdon JM Jr, Ragan MA, Hirt RP, Roger AJ: **A genomic survey of the fish parasite Spironucleus salmonicida indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution.** *BMC Genomics* 2007, **8**:51.
41. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grutzner F, Belov K, Miller W, Clarke L, Chinwalla AT, *et al*: **Genome analysis of the platypus reveals unique signatures of evolution.** *Nature* 2008, **453(7192)**:175-183.
42. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, *et al*: **Comparative genomics of the eukaryotes.** *Science (New York, NY)* 2000, **287(5461)**:2204-2215.
43. Poxleitner MK, Carpenter ML, Mancuso JJ, Wang CJ, Dawson SC, Cande WZ: **Evidence for karyogamy and exchange of genetic material in the binucleate intestinal parasite Giardia intestinalis.** *Science (New York, NY)* 2008, **319(5869)**:1530-1533.
44. Nash TE: **Antigenic variation in Giardia lamblia and the host's immune response.** *Philos Trans R Soc Lond B Biol Sci* 1997, **352(1359)**:1369-1375.
45. Singer SM, Elmendorf HG, Conrad JT, Nash TE: **Biological selection of variant-specific surface proteins in Giardia lamblia.** *J Infect Dis* 2001, **183(1)**:119-124.
46. Kulakova L, Singer SM, Conrad J, Nash TE: **Epigenetic mechanisms are involved in the control of Giardia lamblia antigenic variation.** *Mol Microbiol* 2006, **61(6)**:1533-1542.
47. **GiardiaDB database.** http://www.giardiadb.org/giardiadb/.
48. **The National Center for Biotechnology Information.** http://www.ncbi.nlm.nih.gov/.
49. **Ensembl.** http://www.ensembl.org/index.html.
50. Durbin R: **Biological sequence analysis: probabilistic models of proteins and nucleic acids.** Cambridge, U.K. New York: Cambridge University Press 1998.
51. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Research* 2004, **32(5)**:1792-1797.
52. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22)**:4673-4680.

doi:10.1186/1471-2148-10-49
Cite this article as: Sun *et al.*: Gene duplication in the genome of parasitic *Giardia lamblia*. *BMC Evolutionary Biology* 2010 **10**:49.