LCGserver: A Webserver for Exploring Evolutionary Trajectory of Gene Orders in a Large Number of Genomes

Dapeng Wang,^{1,2} and Jun Yu¹

Abstract

Genes and chromosomes are highly organized; together with protein-coding sequence, gene structure at per gene level and gene order at cluster level are both variable in a context of lineages and under natural selection. How gene order and chromosome organization are related and selected remains to be illuminated. The number of newly-sequenced genomes from various taxa has been increasing rapidly, but there have not been easy-to-use web tools that allow better visualization for gene order in a large genome collection. Here, we describe a webserver, LCGserver (<http://lcgbase.big.ac.cn/LCGserver/>), for exploring evolutionary dynamics of gene orders over diverse lineages. This server provides gene order information at three levels: single gene, paired gene (a minimal cluster), and clustered gene (more than two genes). The most exclusive feature of LCGserver is alignment and visualization of neighboring genes based on orthology, allowing users to inspect all conserved and dynamic events of gene order along chromosomes in a lineage-specific manner. In addition, it categories paired genes into six patterns and identifies fully-conserved gene clusters within and among lineages.

Introduction

ONE OF THE CHIEF TASKS FOR COMPARATIVE or evolutionary genomics is to reveal similarities and differences among genomes and their genes within defined lineages in large-scale. One essential step toward this task is to identify gene orthology and orders; the minimal order appears to be two paired genes (Hurst et al., 2004; Xie et al., 2013). To extend such analysis beyond two genes at a time, we have previously developed a framework for calculating, analyzing, and visualizing cross-species data of orthologous gene pairs, which has proven to be useful in several databases, including LCGbase (Wang et al., 2012), RGKbase (Wang et al., 2013), and Plastid-LCGbase (Wang and Yu, 2015).

We here introduce a new server for user-provided data to include all advantages and benefits with regards to the functionality of our previous databases; we also expand it to bring other new characteristics that are unique as compared to other emerging relevant tools (Lopez and Samuelsson, 2011; Louis et al., 2015; Proost et al., 2012). In brief, this server is built to be user-friendly and adaptive to the increasing demands on genome-wide comparative studies at single gene, paired gene, and clustered gene levels across and within lineages based on genome assemblies of variable qualities.

Material and Methods

We provide a simple format for input files including “gene annotation file” and “orthologous file” adequate for the whole process without any other trivial parameters (Fig. 1A). The gene annotation files include 8 columns: “Category” (protein-coding or ncRNA), “Species”, “Gene Name”, “Chromosome”, “Strand”, “Start”, “End”, and “Note”. The ortholog file adopts two popular columns, “Group Name” and “Gene Name”, which are frequently used in orthologous databases, such as NCBI’s HomoloGene (<http://www.ncbi.nlm.nih.gov/homologene>) and eggNOG (Powell et al., 2014).

We emphasize that accurate definition of orthologs among massive collection of proteins is rather important for gene order determination; it is known that orthology assignment adds a major burden for storage space and computing resources, especially in the pairwise alignment of sequences. Fortunately, the current release of eggNOG database contains orthologous data from thousands of species, offering users well-defined orthologous groups directly. The core calculation component is composed of three modules: “Gene module”, “Gene pair module”, and “Gene cluster module” (Fig. 1B).

“Gene module” provides fractioning of non-orphan and orphan genes and listing of genes and gene families over

¹CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, People’s Republic of China.

²Stem Cell Laboratory, UCL Cancer Institute, University College London, London, United Kingdom.

FIG. 1. A flowchart for LCGserver design. (A) The input files; (B) The modules for calculation and analysis; and (C) The modules for statistics and visualization.

species. “Gene pair module” identifies gene pairs, and most importantly, it categorizes conserved and variable patterns and exhausts all possibilities. As a result, we define six patterns for gene pairing scheme between two species in a comparison on the basis of the degree of gene order variability from conserved to highly dynamic (Fig. 2).

“Gene cluster module” searches no more than 10 conserved gene pairs flanking an inquiry and concatenates them into one conserved gene cluster. “Statistical module” generates both tables and figures showing the landscape of genes and gene pairs in a comprehensive manner. “Visualization module” displays genome alignment assisted by single genes

FIG. 2. The six patterns of gene pair variations. (A) Fully-conserved refers to both gene orthology and transcription direction remaining exactly the same; (B) Half-conserved means that, for one gene from the pair, only one of the parameters is conserved, orthology but not transcription direction; (C) No-conserved defines orthology but varied transcription direction; (D) Two-orthologs describes a gene pair that lost the pairing position but not orthology; (E) One-ortholog is a situation where only one of the paired genes in one species has a counterpart in another species; and (F) No-ortholog indicates that both paired genes lost their mutual orthology.

FIG. 3. A snapshot of LCGserver web pages. (A) The main panel for browsing results. (B) The analysis for gene families. (C) The results for variation patterns of gene pairs. (D) The gene order display in gene pair mode. (E) The gene order display in single gene mode.

or gene pairs, which share common ancestors, where arrows and colors represent transcription direction and orthology, respectively (Fig. 1C). For well-structured web pages where all results are displayed, it provides links on the main browsing panel to list five major datasets for each species: “Gene Family”, “Gene List”, “Gene Pair”, “Gene Cluster”, and “Statistics” (Fig. 3A).

For the middle three subsections, it displays annotation information for genes and gene pairs, as well as their analysis results in tables and colored images. In addition, it also provides helpful internal links between different datasets in the results; for instance, a gene and its family members are both indicated with arrows on the images, where detailed annotations of the genes are also linked.

The server is mainly implemented in a conventional LAMP framework, together with some additional software packages. Perl and R take charge of the plentiful calculations on the back side and the combined usage of HTML and PHP is responsible for the interactivity between user and server on the front side. In particular, one main Perl script is used to generate various web pages encompassing the final results in an appropriate manner and one PHP script is used to execute and monitor all the processes, record each step, and import the data into MySQL database. All the progress is stored in a working directory respectively and independently for each user who is able to view the results at any time after the completion of the analysis.

Results and Discussion

In order to examine the performance of this server, we took notes for one typical application. We used gene annotation and

ortholog data of seven *Drosophila* subspecies from Ensembl/Ensembl Genomes (Cunningham et al., 2015) and eggNOG databases (Powell et al., 2014), respectively. The total time from uploading data to finishing all calculations was ~2 min. Taking the pair of FBgn0205304 (a member of droNOG08487 family) and FBgn0205258 (a member of droNOG11829 family) from *Drosophila virilis* as an example, in particular, 78.5% genes from this species were annotated as non-orphan genes that have at least one homologous counterpart in all input genes, and the peak in the number of gene families includes less than 10 genes per family (Fig. 3B).

When examining their conserved situation in a paired gene mode (Fig. 3D), we found that they were fully-conserved in all species except *Drosophila melanogaster*, where the gene pairing pattern was “two_orthologs” (Fig. 3C). Upon visualization in single gene mode for FBgn0205304, it was easy to see that FBgn0263206 had inserted between FBgn0028541 and FBgn0028540 (Fig. 3E). Examining gene clustering around FBgn0205304–FBgn0205258, we recognized 4 clusters having highly variable numbers of genes among the subspecies, such as 6 in *Drosophila willistoni*, 22 in *Drosophila grimshawi*, 11 in *Drosophila pseudoobscura*, and 10 in *Drosophila yakuba*.

We propose but are not limited to many potential applications based on this server in the field of comparative genomics. First, it helps to look for fundamental evolutionary rules and trajectory in genome structure with special focuses on the relationship between adjacent genes. Second, it can be used to correct genome assemblies by checking the gene and cluster orders against closely-related species and looking for co-linearity. Third, it facilitates phylogeny analysis when position information of genes and clusters in genomes are

examined together. Fourth, it promotes cross-validation between gene location and gene co-expression, as RNA-Seq technique has been providing a large number of datasets for biological studies. Last, it detects highly-conserved and lineage-conserved gene clusters, which often serve as anchorage for 2D and 3D chromosomal structures.

As a final note, for the next edition of this server, we anticipate better visualization and new modules for integrating more types of sequence variations (such as insertion-deletion and large inversion) and constructing phylogeny based on gene order.

Author Disclosure Statement

No competing financial interests exist.

References

- Cunningham F, Amode MR, Barrell D, et al. (2015). Ensembl 2015. *Nucleic Acids Res* 43, D662–669.
- Hurst LD, Pal C, and Lercher MJ. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* 5, 299–310.
- Lopez MD, and Samuelsson T. (2011). eGOB: Eukaryotic Gene Order Browser. *Bioinformatics* 27, 1150–1151.
- Louis A, Nguyen NT, Muffato M, and Roest Crollius H. (2015). Genomic update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics. *Nucleic Acids Res* 43, D682–689.
- Powell S, Forslund K, Szklarczyk D, et al. (2014). eggNOG v4.0: Nested orthology inference across 3686 organisms. *Nucleic Acids Res* 42, D231–239.
- Proost S, Fostier J, De Witte D, et al. (2012). i-ADHoRe 3.0—Fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res* 40, e11.
- Wang D, Xia Y, Li X, Hou L, and Yu J. (2013). The Rice Genome Knowledgebase (RGKbase): An annotation database for rice comparative genomics and evolutionary biology. *Nucleic Acids Res* 41, D1199–1205.
- Wang D, and Yu J. (2015). Plastid-LCGbase: A collection of evolutionarily conserved plastid-associated gene pairs. *Nucleic Acids Res* 43, D990–995.
- Wang D, Zhang Y, Fan Z, Liu G, and Yu J. (2012). LCGbase: A comprehensive database for lineage-based co-regulated genes. *Evol Bioinform Online* 8, 39–46.
- Xie B, Wang D, Duan Y, Yu J, and Lei H. (2013). Functional networking of human divergently paired genes (DPGs). *PLoS One* 8, e78896.

Address correspondence to:
Dr. Dapeng Wang
Stem Cell Laboratory
UCL Cancer Institute
University College London
London WC1E 6BT
United Kingdom

E-mail: dapeng.wang@ucl.ac.uk

or

Dr. Jun Yu
CAS Key Laboratory of Genome Sciences and Information
Beijing Institute of Genomics
Chinese Academy of Sciences
Beijing 100101
P. R. China

E-mail: junyu@big.ac.cn