Biochemical and Biophysical Research Communications 469 (2016) 542–545

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

DLGP: A database for lineage-conserved and lineage-specific gene pairs in animal and plant genomes

Dapeng Wang [a, b, *]

[a] Stem Cell Laboratory, UCL Cancer Institute, University College London, London WC1E 6BT, UK
[b] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, PR China

ARTICLE INFO

Article history:
Received 1 December 2015
Accepted 10 December 2015
Available online 15 December 2015

Keywords:
Gene pairs
Database
Animals
Plants
Lineage

ABSTRACT

The conservation of gene organization in the genome with lineage-specificity is an invaluable resource to decipher their potential functionality with diverse selective constraints, especially in higher animals and plants. Gene pairs appear to be the minimal structure for such kind of gene clusters that tend to reside in their preferred locations, representing the distinctive genomic characteristics in single species or a given lineage. Despite gene families having been investigated in a widespread manner, the definition of gene pair families in various taxa still lacks adequate attention. To address this issue, we report DLGP (http://lcgbase.big.ac.cn/DLGP/) that stores the pre-calculated lineage-based gene pairs in currently available 134 animal and plant genomes and inspect them under the same analytical framework, bringing out a set of innovational features. First, the taxonomy or lineage has been classified into four levels such as *Kingdom*, *Phylum*, *Class* and *Order*. It adopts all-to-all comparison strategy to identify the possible conserved gene pairs in all species for each gene pair in certain species and reckon those that are conserved in over a significant proportion of species in a given lineage (e.g. *Primates*, *Diptera* or *Poales*) as the lineage-conserved gene pairs. Furthermore, it predicts the lineage-specific gene pairs by retaining the above-mentioned lineage-conserved gene pairs that are not conserved in any other lineages. Second, it carries out pairwise comparison for the gene pairs between two compared species and creates the table including all the conserved gene pairs and the image elucidating the conservation degree of gene pairs in chromosomal level. Third, it supplies gene order browser to extend gene pairs to gene clusters, allowing users to view the evolution dynamics in the gene context in an intuitive manner. This database will be able to facilitate the particular comparison between animals and plants, between vertebrates and arthropods, and between monocots and eudicots, accounting for the significant contribution of gene pairs to speciation and diversification in specific lineages.

1. Introduction

Genes are arranged linearly on the chromosomes, naturally exhibiting three typical patterns of gene pairs such as DPGs (head-to-head), CDPGs (head-to-tail) and CPGs (tail-to-tail), based on the conditions of strands of the neighboring located genes, leading to the formation of diverse degrees of gene clusters [1]. According to the evidence from comparative genomics and gene expression profiling data, it has been commonly accepted that a substantial number of gene pairs play coordinated roles in transcription and functional pathways [2,3], some of which appear to be involved in cancer or disease [4]. The underlying mechanism includes cis- and trans-regulations such as shared promoter regions, occasionally two separate promoters regulating two genes respectively [5], and the usage of common transcription factor, or modulation of antisense long non-coding RNAs with a stable form but low-level expression through multiple level interactions [6]. In addition, it is possible that distance between the neighboring genes might affect the ways of co-regulation on both genes [7,8], for example, the paired genes with a larger distance could be regulated through a huge gene network [9], reflecting the dynamics of chromatin domains. In the view of evolution, a great many of gene pairs are specific to each species but not shared by all species despite that there are similarities in the number of gene pairs and the functions of the proteins encoded by gene pairs across species in some certain

* Stem Cell Laboratory, UCL Cancer Institute, University College London, London WC1E 6BT, UK.
E-mail address: dapeng.wang@ucl.ac.uk.

http://dx.doi.org/10.1016/j.bbrc.2015.12.039
0006-291X/

clades [10–23]. Recent study revealed co-evolution pattern for neighboring genes in expression [24], with an emphasis on the effects of natural selection, reaching a global optimization of genome under spatial and temporal diversity [3]. Therefore, it is worth to characterize the conservation of gene pairs in a wide range of genomes, beneficial to the understanding of molecular mechanism of how genes cooperate in relation to speciation and diversification. The current resources with visualization modules are mostly limited to a narrow scope of taxa or a small number of representative species [25–29], and to resolve this problem, we developed a public resource dedicated to important gene pairs such as lineage-conserved and lineage-specific ones, containing 134 high-quality annotated animal and plant genomes. It conducts reasonable classification for all species and calculates the homologous gene pairs across species and effectively visualizes the gene order around the conserved gene pairs, to our best knowledge, which is the sole database for comparison spanning animals and plants at the same time so far.

2. Materials and methods

Protein sequences and gene coordinate information were retrieved from Ensembl (release-73) and Ensembl Genomes (release-20) database (ftp.ensembl.org and ftp.ensemblgenomes.org). For alternative splicing events, only the protein-coding transcript with maximum length was chosen to represent its gene. Gene homology was determined by blastp from ncbi-blast with parameters of -evalue = 1e-5 and -max_target_seqs = 500 and gene families were constructed by mcl software under −I = 1.2. In particular, we identified the existing evolutionary counterparts of each gene pair in all species in consideration of both relative orientation and homology of two neighboring situated genes, adopting an exhaustive strategy of all-to-all comparison. On the basis of the classic parent-child hierarchical classification, we readily classified all 134 species based on a 5-layer taxonomy (*Kingdom- > Phylum- > Class- > Order- > Species*) and determined lineage-conserved gene pairs by checking the child taxonomical layers of the investigated lineage (taxonomy) that must meet the minimum requirement in the numbers of different child taxa (Table 1). Furthermore, we defined lineage-specific gene pairs by excluding the gene pairs that have counterparts in the species that are not in the lineage of interest.

3. Results

In the comparison module, we provided two options for users to detect conserved gene pairs and show the conservation of them in number and fraction with reference to a large-scale level such as chromosomes (Fig. 1A). In details, we selected the top 10 chromosomes for each species in terms of the enrichment of the conserved gene pairs, and in rare conditions, we used 6 instead of 10 due to the limitation of specific species chromosome total number.

The fundamental browsing function enables the users to access the genes and gene pairs, detailing their annotation and information stored in this database, of which each gene or gene pair is linked to its gene family or gene pair family if available where it's convenient to find out the relevant family members (Fig. 1D). Most interestingly, it gives all potential lineage-conserved gene pairs arranged by four taxonomical layers such as *Order*, *Class*, *Phylum*, *Kingdom* (Fig. 1C). In particular, there is an additional column/row in the table to denote whether this lineage-conserved gene pair is a lineage-specific one (Fig. 1E). Furthermore, it creates a downloadable publication-quality image illustrating all the conserved gene pairs as well as genes in their neighborhood environment along the chromosome after inputting the gene name and choosing the appropriate species names (Fig. 1B). Specifically, each arrow means a gene, of which the orientation represents the positive/negative strand whereas the color indicates the homologous group defined in the comparison from all species chosen. In case the colors are not well distinguishable, the identification number of homologous group has been placed above each arrow (gene). Finally, the statistics section shows the number of lineage-conserved and lineage-specific gene pairs when they are categorized into three types such as DPGs, CDPGs and CPGs in the four taxonomical levels.

Table 1
The rules for defining the conserved gene pairs.

Taxonomy	Name	Number of subgroup	Threshold number
Order	Primates	10	5
Order	Rodentia	5	3
Order	Lagomorpha	2	2
Order	Carnivora	4	3
Order	Chiroptera	2	2
Order	Artiodactyla	3	3
Order	Galliformes	2	2
Order	Passeriformes	2	2
Order	Tetraodontiformes	2	2
Order	Enterogona	2	2
Order	Diptera	17	6
Order	Hemiptera	2	2
Order	Hymenoptera	3	3
Order	Lepidoptera	3	3
Order	Rhabditida	6	4
Order	Spirurida	2	2
Order	Brassicales	3	3
Order	Fabales	2	2
Order	Solanales	2	2
Order	Poales	11	5
Class	Mammalia	20	6
Class	Aves	3	3
Class	Actinopterygii	7	4
Class	Arachnida	2	2
Class	Insecta	6	4
Class	Secernentea	2	2
Class	Eudicots	5	3
Class	Monocots	2	2
Phylum	Chordata	9	5
Phylum	Mollusca	2	2
Phylum	Annelida	2	2
Phylum	Arthropoda	4	3
Phylum	Nematoda	2	2
Phylum	Angiosperms	2	2
Kingdom	Animalia	10	5
Kingdom	Plantae	4	3

3.1. Case study 1

The human DPG of ENSG00000104763 (*ASAH1*) and ENSG00000171428 (*NAT1*) is *Primates*-specific and its counterparts have been detected in other six primate species such as *Pan troglodytes*, *Gorilla gorilla*, *Pongo abelii*, *Nomascus leucogenys*, *Macaca mulatta* and *Otolemur garnettii*, but absent in all non-primates. In depth, we found out a highly stable cluster (*MTUS1* − *FGL1* − *PCM1* − *ASAH1* − *NAT1*), which is conserved in all primates shown. Besides, some species-specific variation have been observed, for instance, the insertion of ENSGGOG00000006705 between *PSD3* and *SH2D4A* in *Gorilla gorilla* (Fig. 1B).

3.2. Case study 2

As an example of *Poales*-specific gene pairs, a pair of GRMZM2G348578 (*P4H3*) and GRMZM2G047855 (*CK2*) acts as a CPG on chromosome 1 from *Zea mays*, whose family members exist in another 10 plant species such as *Sorghum bicolor*, *Setaria italica*,