Data in Brief 6 (2016) 208–213

Contents lists available at ScienceDirect

Data in Brief

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

Data article

Curated eutherian third party data gene data sets

Marko Premzl

Laboratory of Genomics, Centre of Animal Reproduction, 55 Heinzel St., Zagreb, Croatia

ARTICLE INFO

Article history:
Received 5 September 2015
Received in revised form
18 November 2015
Accepted 24 November 2015
Available online 11 December 2015

Keywords:
Comparative Genomic Analysis
Gene Annotations
Molecular Evolution
Phylogenetic Analysis

ABSTRACT

The free available eutherian genomic sequence data sets advanced scientific field of genomics. Of note, future revisions of gene data sets were expected, due to incompleteness of public eutherian genomic sequence assemblies and potential genomic sequence errors. The eutherian comparative genomic analysis protocol was proposed as guidance in protection against potential genomic sequence errors in public eutherian genomic sequences. The protocol was applicable in updates of 7 major eutherian gene data sets, including 812 complete coding sequences deposited in European Nucleotide Archive as curated third party data gene data sets.

Specifications Table

Subject area	Biology
More specific subject area	Genomics
Type of data	Third party data
How data was acquired	*In computo*
Data format	FAS, TXT
Experimental factors	Eutherian comparative genomic analysis protocol
	Curated gene data sets

E-mail address: Marko.Premzl@alumni.anu.edu.au

http://dx.doi.org/10.1016/j.dib.2015.11.056

Experimental features	
Data source location	N/A
Data accessibility	The original gene data sets were deposited in European Nucleotide Archive under accession numbers: FR734011-FR734074 (http://www.ebi.ac.uk/ena/data/view/FR734011-FR734074), HF564658-HF564785 (http://www.ebi.ac.uk/ena/data/view/HF564658-HF564785), HF564786-HF564815 (http://www.ebi.ac.uk/ena/data/view/HF564786-HF564815), HG328835-HG329089 (http://www.ebi.ac.uk/ena/data/view/HG328835-HG329089), HG426065-HG426183 (http://www.ebi.ac.uk/ena/data/view/HG426065-HG426183), HG931734-HG931849 (http://www.ebi.ac.uk/ena/data/view/HG931734-HG931849) and LM644135-LM644234 (http://www.ebi.ac.uk/ena/data/view/LM644135-LM644234). Data analysis is with this article.

Value of the data

- Curated gene data sets applicable in gene annotations and genome analyses.
- Curated gene data sets applicable in phylogenetic analyses.
- Curated gene data sets applicable in protein structure and function analyses.

1. Data

Undoubtedly, the eutherian comparative genomics momentum was maintained by programmatic, considerable international efforts in production, assembly and analysis of public eutherian genomic sequence data sets (Fig. 1) [1–3]. For example, the initial sequencing and analysis of human genome revised human gene data sets [4,5]. Nevertheless, these analyses were subject to future updates and revisions due to incompleteness of public eutherian genomic sequence data sets and potential genomic sequence errors [1–6]. The eutherian comparative genomic analysis protocol was proposed as guidance in protection against potential genomic sequence errors in public eutherian genomic sequences [7–12]. The protocol was established as one framework of eutherian third party data gene data set descriptions (Fig. 2). The protocol included new genomics and protein molecular evolution tests applicable in updates and revisions of 7 major eutherian gene data sets, including interferon-γ-inducible GTPase genes, ribonuclease A genes, Mas-related G protein-coupled receptor genes, lysozyme genes, adenohypophysis cystine-knot genes, macrophage migration inhibitory factor and D-dopachrome tautomerase genes and, finally, growth hormone genes (Fig. 3). The protocol discriminated major gene clusters with and without evidence of differential gene expansions. For example, the eutherian major gene clusters with no evidence of differential gene expansions could be suitable in phylogenomic analyses.

2. Experimental design, materials and methods

The eutherian comparative genomic analysis protocol included gene annotations, phylogenetic analysis and protein molecular evolution analysis [7–12] (Fig. 2). The protocol used free available eutherian genomic sequence data sets deposited in public biological databases and software.

3. Gene annotations

The gene annotations included gene identifications in eutherian genomic sequences, analyses of gene features, tests of reliability of eutherian public genomic sequences and multiple pairwise