Eutherian comparative genomic analysis protocol

Marko Premzl (Marko.Premzl@alumni.anu.edu.au)
The Australian National University Alumni https://orcid.org/0000-0002-3362-689X

Method Article

Keywords: comparative genomics, Eutheria, gene data set, molecular evolution, RRID:SCR_014401

Posted Date: June 5th, 2020

DOI: https://doi.org/10.21203/rs.2.1502/v3

Version of Record: A version of this preprint was published at Protocol Exchange on March 12th, 2018.
See the published version at https://doi.org/10.1038/protex.2018.028.

Abstract

The eutherian genomics momentum greatly advanced biology and medicine. Nevertheless, future revisions and updates of eutherian genomic sequence data sets were expected, due to potential genomic sequence errors and incompleteness of genomic sequences. The **eutherian comparative genomic analysis protocol** was established as guidance in protection against potential genomic sequence errors in public eutherian genomic sequence assemblies. The protocol revised, updated and published **12 major eutherian gene data sets**, including **1853 complete coding sequences** deposited in European Nucleotide Archive as curated third party data gene data sets under accession numbers: FR734011-FR734074, HF564658-HF564785, HF564786-HF564815, HG328835-HG329089, HG426065-HG426183, HG931734-HG931849, LM644135-LM644234, LN874312-LN874522, LT548096-LT548244, LT631550-LT631670, LT962964-LT963174 and LT990249-LT990597.

Introduction

Background

The eutherian genomics momentum greatly advanced biological and medical sciences. For example, one major aim of initial sequencing and analysis of human genome was to **update and revise human genes**, as well as to uncover potential new drugs, drug targets, and molecular markers in medical diagnostics. Yet, **future revisions and updates of eutherian genomic sequence data sets** were expected, due to **potential genomic sequence errors** and incompleteness of reference genomic sequences. The **potential genomic sequence errors** included **Sanger DNA sequencing method errors** (artefactual nucleotide deletions, insertions and substitutions) and **analytical and bioinformatical errors** (erroneous gene annotations, genomic sequence misassemblies). Thus, the **eutherian comparative genomic analysis protocol** RRID:SCR_014401 was established as **guidance in protection against potential genomic sequence errors** in public eutherian genomic sequence assemblies [1,2,3,4,5,6,7,8,9,10,11,12,13].

Reagents

On-line resources:

Ensembl genome browser

European Nucleotide Archive

National Center for Biotechnology Information

Equipment

The present protocol required personal computer and internet connectivity.

Procedure

Eutherian comparative genomic analysis protocol

The **eutherian comparative genomic analysis protocol** RRID:SCR_014401 integrated **gene annotations, phylogenetic analysis and protein molecular evolution analysis** into one framework of eutherian gene descriptions.

Gene annotations

The eutherian **gene annotations** included **gene identifications in public reference genomic sequence assemblies, analyses of gene features, multiple pairwise genomic sequence alignments and tests of reliability of eutherian public genomic sequences**. The protocol used sequence alignment editor BioEdit in all analyses of nucleotide and protein sequences. The identifications of potential coding sequence used eutherian reference genomic sequence assemblies downloaded from National Center for Biotechnology Information's (NCBI) GenBank or Ensembl genome browser, as well as NCBI's BLAST programs and Ensembl genome browser's BLAST or BLAT web tools. In analyses of gene features, the protocol used direct evidence of gene annotations available in NCBI's nr, est_human, est_mouse and est_others databases. The protocol established **tests of reliability of eutherian public genomic sequences** that used potential coding sequences. The first test steps analysed nucleotide sequence coverages of each potential coding sequence, using NCBI's BLASTN program and primary experimental genomic sequence information deposited in NCBI's Trace Archive. The second test steps described potential coding sequences as complete coding sequences only if consensus trace coverages were available for every nucleotide in each potential coding sequence. Alternatively, the protocol described potential coding sequences as putative coding sequences (not used in analyses). The protocol used complete coding sequences in all analyses, and deposited them in European Nucleotide Archive as curated third party data gene data sets. The guidelines of human gene nomenclature and mouse gene nomenclature were used in revisions and updates of gene classifications. In multiple pairwise genomic sequence alignments, the protocol used mVISTA's AVID. In base sequences used in multiple pairwise genomic sequence alignments, transposable elements were masked by RepeatMasker. Finally, the pairwise nucleotide sequence identities of predicted promoter regions calculated using BioEdit were used in statistical analyses (Microsoft Office Excel).

Phylogenetic analysis

In **phylogenetic analyses**, the protocol included **protein and nucleotide sequence alignments, calculations of phylogenetic trees and calculations of pairwise nucleotide sequence identity patterns**. First, the complete coding sequences were translated using BioEdit, and then aligned at amino acid level using ClustalW implemented in BioEdit. After inspections and manual corrections of protein primary sequence alignments, the protocol prepared nucleotide sequence alignments. The MEGA was used in phylogenetic tree calculations. The protocol used neighbour-joining, minimum evolution, maximum parsimony and unweighted pair group method with arithmetic mean methods. The pairwise nucleotide sequence

identities of nucleotide sequence alignments calculated using BioEdit were used in statistical analyses (Microsoft Office Excel). For each nucleotide sequence alignment, the protocol calculated average pairwise identities and their average absolute deviations, as well as largest pairwise identities and smallest pairwise identities. Finally, the protocol discriminated between eutherian major gene clusters including and not including evidence of differential gene expansions.

Protein molecular evolution analysis

The protocol established **tests of protein molecular evolution** integrating patterns of nucleotide sequence similarities with protein primary structures. The protein and nucleotide sequence alignments were used in tests. First, for each nucleotide sequence alignment, the protocol calculated codon usage statistics using MEGA. The ratios between observed and expected amino acid codon counts determined relative synonymous codon usage statistics (R = Observed codon counts / Expected codon counts). Therefore, in nucleotide sequence alignments, the amino acid codons including $R \leq 0.7$ were designated as not preferable amino acid codons. The protocol then described reference protein sequence amino acid sites as invariant amino acid sites (invariant alignment positions), forward amino acid sites (variant alignment positions that did not include amino acid codons with $R \leq 0.7$) or compensatory amino acid sites (variant alignment positions that included amino acid codons with $R \leq 0.7$), using protein and nucleotide sequence alignments.

Anticipated Results

Project outcomes

The **eutherian comparative genomic analysis protocol** RRID:SCR_014401 published **1853 complete coding sequences** that included: 64 interferon-γ-inducible GTPase genes [1], 255 ribonuclease A genes [2], 119 Mas-related G protein-coupled receptor genes [3], 116 lysozyme genes [4], 128 adenohypophysis cystine-knot genes [5], 30 D-dopachrome tautomerase and macrophage migration inhibitory factor genes [5], 100 growth hormone genes [6], 211 tumor necrosis factor ligand genes [8], 149 globin genes [9], 121 kallikrein genes [10], 211 adiponectin genes [11] and 349 connexin genes [13]. These curated eutherian third party data gene data sets were applicable in gene annotations and genome analyses, phylogenetic analyses and analyses of protein molecular evolution, including initial descriptions of human genes.

References

1. Premzl M (2012) Comparative genomic analysis of eutherian interferon-γ-inducible GTPases. Funct Integr Genomics 12:599-607 DOI: 10.1007/s10142-012-0291-2

2. Premzl M (2014) Comparative genomic analysis of eutherian ribonuclease A genes. Mol Genet Genomics 289:161-167 DOI: 10.1007/s00438-013-0801-5

3. Premzl M (2014) Comparative genomic analysis of eutherian Mas-related G protein-coupled receptor genes. Gene 540:16-19 DOI: 10.1016/j.gene.2014.02.049

4. Premzl M (2014) Third party annotation gene data set of eutherian lysozyme genes. Genom Data 2:258-260 DOI: 10.1016/j.gdata.2014.08.003

5. Premzl M (2015) Initial description of primate-specific cystine-knot Prometheus genes and differential gene expansions of D-dopachrome tautomerase genes. Meta Gene 4:118-128 DOI: 10.1016/j.mgene.2015.02.005

6. Premzl M (2015) Third party data gene data set of eutherian growth hormone genes. Genom Data 6:166-169 DOI: 10.1016/j.gdata.2015.09.007

7. Premzl M (2016) Curated eutherian third party data gene data sets. Data Brief 6:208-213 DOI: 10.1016/j.dib.2015.11.056

8. Premzl M (2016) Comparative genomic analysis of eutherian tumor necrosis factor ligand genes. Immunogenetics 68:125-132 DOI: 10.1007/s00251-015-0887-5

9. Premzl M (2016) Comparative genomic analysis of eutherian globin genes. Gene Rep 5:163-166 DOI: 10.1016/j.genrep.2016.10.009

10. Premzl M (2017) Comparative genomic analysis of eutherian kallikrein genes. Mol Genet Metab Rep 10:96-99 DOI: 10.1016/j.ymgmr.2017.01.009

11. Premzl M (2018) Comparative genomic analysis of eutherian adiponectin genes. Heliyon 4:e00647 DOI: 10.1016/j.heliyon.2018.e00647

12. Premzl M (2019) Eutherian third-party data gene collections. Gene Rep 16:100414 DOI: 10.1016/j.genrep.2019.100414

13. Premzl M (2019) Comparative genomic analysis of eutherian connexin genes. Sci Rep 9:16938 DOI: 10.1038/s41598-019-53458-x

Acknowledgements

The **eutherian comparative genomic analysis protocol** RRID:SCR_014401 was established under open science research project "Comparative genomic analysis of eutherian genes".

Author: Marko Premzl PhD, The Australian National University Alumni (WFH), 4 Kninski trg Sq., Zagreb, Croatia

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 20180301FigureNatureProtocols.pdf