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, complete coding sequence, Eutheria, gene data set, RRID:SCR_014401

Posted Date: August 26th, 2022

DOI: https://doi.org/10.21203/rs.2.1502/v5

License: ☑️  This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

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 biological and medical sciences. Yet, future revisions and updates of eutherian genomic sequence data sets were expected, due to potential genomic sequence errors and incompleteness of genomic sequences. Importantly, the human protein coding gene census remained unfinished: contemporary estimates included \( \approx 20000-21000 \) protein coding genes in human genome, in contrast to human initial integrated gene index that included \( \approx 32000 \) known and predicted protein coding genes. The eutherian comparative genomic analysis protocol was established as one framework of eutherian gene data set descriptions. The protocol revised, updated and published 14 major eutherian gene data sets, including 2615 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, LT990249-LT990597, LR130242-LR130508 and LR760818-LR761312.

Introduction

The eutherian genomics momentum greatly advanced biomedical research. 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. However, future revisions and updates of eutherian genomic sequence data sets were expected, due to potential genomic sequence errors and incompleteness of reference genomic sequences. Specifically, 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). Importantly, the human protein coding gene census remained unfinished: contemporary estimates included \( \approx 20000-21000 \) protein coding genes in human genome, in contrast to human initial integrated gene index that included \( \approx 32000 \) known and predicted protein coding genes. 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.

Reagents

Public biological databases:

- Ensembl genome browser
- European Nucleotide Archive
- National Center for Biotechnology Information

Equipment
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. The protocol included 3 original genomics and protein molecular evolution tests, including tests of reliability of public eutherian genomic sequences using genomic sequence redundancies, tests of contiguity of public eutherian genomic sequences using multiple pairwise genomic sequence alignments and tests of protein molecular evolution using relative synonymous codon usage statistics.

1. Gene annotations

The eutherian gene annotations included gene identifications in public genomic sequence assemblies, analyses of gene features, tests of reliability of public eutherian genomic sequences and tests of contiguity of public eutherian genomic sequences.

1.1. All analyses and manipulations of nucleotide and protein sequences used sequence alignment editor BioEdit.

1.2. The eutherian reference genomic sequence data sets were accessible in National Center for Biotechnology Information's (NCBI) GenBank, as well as in Ensembl genome browser.

1.3. The identifications of potential coding sequences used public eutherian reference genomic sequence assemblies and NCBI's BLAST program including BLAST Genomes and Ensembl genome browser's BLAST or BLAT programs.

1.4. The analyses of gene features used potential coding sequences and direct evidence of eutherian gene annotations accessible in NCBI's nr, est_human, est_mouse and est_others databases.

1.5. The tests of reliability of eutherian public genomic sequences analysed potential coding sequences using good laboratory practice in Sanger DNA sequencing method. The first test steps analysed nucleotide sequence coverages of potential coding sequences using NCBI's BLAST program and processed Sanger DNA sequencing reads or traces accessible in NCBI's Trace Archive. The second test steps discriminated complete coding sequences and putative coding sequences. Specifically, the tests described potential coding sequences as complete coding sequences only if consensus trace nucleotide sequence coverages were available for every nucleotide. Alternatively, if consensus trace nucleotide sequence coverages were not available for every nucleotide, the potential coding sequences were described as putative coding sequences that were not used in analyses. For example, the good laboratory practice in Sanger DNA sequencing method exacted that minimal consensus trace nucleotide sequence coverage included 2 identical trace nucleotide sequences.
1.6. The **tests of contiguity of public eutherian genomic sequences** included multiple pairwise genomic sequence alignments. The tests used public eutherian reference genomic sequences encoding **complete coding sequences** and **mVISTA’s program AVID**. In eutherian genomic sequences, the tests analysed translated exon numbers, as well as their chimerisms and relative orders and orientations. The tests of contiguity of eutherian public genomic sequences did not use masking of transposable elements in public eutherian reference genomic sequence assemblies.

1.7. The **curated eutherian gene collections** were deposited in **European Nucleotide Archive** as third party data gene data sets. The revised and updated eutherian gene classifications and nomenclatures used **guidelines of human gene nomenclature** and **guidelines of mouse gene nomenclature**.

2. **Phylogenetic analysis**

The **phylogenetic analysis** included **protein and nucleotide sequence alignments**, **calculations of phylogenetic trees** and **calculations of pairwise nucleotide sequence identities**.

2.1. The **complete coding sequences** were translated using **BioEdit** and then aligned at amino acid level using **ClustalW** in **protein amino acid sequence alignments**. The **protein amino acid sequence alignments** were manually corrected, and **nucleotide sequence alignments** were prepared accordingly using **BioEdit**.

2.2. The **calculations of phylogenetic trees** used **nucleotide sequence alignments** and **MEGA program**.

2.3. Using **nucleotide sequence alignments**, the **pairwise nucleotide sequence identities of eutherian complete coding sequences** were calculated using **BioEdit**. The statistical analyses using Microsoft Office Excel statistical functions included calculations of average pairwise nucleotide sequence identities ($\bar{a}$) and their average absolute deviations ($\bar{a}_{ad}$), as well as largest ($a_{\text{max}}$) and smallest ($a_{\text{min}}$) pairwise nucleotide sequence identities.

3. **Protein molecular evolution analysis**

The **protein molecular evolution analysis** included **analyses of protein amino acid sequence features** and **tests of protein molecular evolution** using relative synonymous codon usage statistics.

3.1. The **protein amino acid sequence features** were annotated manually, including analyses of common cysteine amino acid residue patterns among eutherian major protein clusters.

3.2. The **tests of protein molecular evolution** using relative synonymous codon usage statistics integrated patterns of nucleotide sequence similarities with protein primary structures. Using **nucleotide sequence alignments**, the **MEGA** calculated relative synonymous codon usage statistics as ratios between observed and expected amino acid codon counts ($R = \text{Counts} / \text{Expected counts}$). The amino acid codons including $R \leq 0.7$ were described as not preferable amino acid codons. In reference protein amino acid sequences, the tests described invariant amino acid sites (invariant alignment positions), forward amino acid sites (variant alignment positions that did not include amino acid codons with $R \leq 0.7$) and
compensatory amino acid sites (variant alignment positions that included amino acid codons with $R \leq 0.7$).

**Troubleshooting**

The tests of reliability of public eutherian genomic sequences were contingent on primary Sanger DNA sequencing information deposited in NCBI's Trace Archive\(^{12,16}\). Specifically, there was positive correlation between genomic sequence redundancies of 35 public eutherian reference genomic sequence assemblies respectively and curated complete coding sequence numbers.

**Time Taken**

**Anticipated Results**

The *eutherian comparative genomic analysis protocol* RRID:SCR_014401 published 2615 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\(^11\), 349 connexin genes\(^13\), 267 fibroblast growth factor genes\(^14\), and 495 interferon genes\(^15\). These *curated eutherian third party data gene data sets* were applicable in gene annotations and genome analyses — including differential gene expansion analyses and initial descriptions of human genes, phylogenetic analyses and protein molecular evolution analyses.

**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:
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

14. Premzl M (2020) Comparative genomic analysis of eutherian fibroblast growth factor genes. BMC Genomics 21:542 DOI: 10.1186/s12864-020-06958-4

15. Premzl M (2020) Comparative genomic analysis of eutherian interferon genes. Genomics 112:4749-4759 DOI: 10.1016/j.ygeno.2020.08.029

16. Premzl M (2022) Revised eutherian gene collections. BMC Genomic Data 23:56 DOI: 10.1186/s12863-022-01071-9

**Acknowledgements**

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

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

**E-mail contact:** Marko.Premzl@alumni.anu.edu.au
The author would like to express his gratitude to data analysts, producers and providers of public eutherian reference genomic sequence data sets, as well as to thank publisher on their endorsement of open science research.

**Supplementary Files**

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

- 20210123Eutheriancomparativegenomicanalysisprotocolflowchart.pdf