BMC Genomic Data

DATA NOTE

Open Access

Revised eutherian gene collections

Marko Premzl[1,2*]

Abstract

Objectives: The most recent research projects in scientific field of eutherian comparative genomics included intentions to sequence every extant eutherian species genome in foreseeable future, so that future revisions and updates of eutherian gene data sets were expected.

Data description: Using 35 public eutherian reference genomic sequence assemblies and free available software, the eutherian comparative genomic analysis protocol RRID:SCR_014401 was published as guidance against potential genomic sequence errors. The protocol curated 14 eutherian third-party data gene data sets, including, in aggregate, 2615 complete coding sequences that were deposited in European Nucleotide Archive. The published eutherian gene collections were used in revisions and updates of eutherian gene data set classifications and nomenclatures that included gene annotations, phylogenetic analyses and protein molecular evolution analyses.

Keywords: Gene data set, Comparative genomics, Eutheria, RRID:SCR_014401

Objective

The most recent research projects in scientific field of eutherian comparative genomics included intentions to sequence every extant eutherian species genome in foreseeable future, so that future revisions and updates of eutherian gene data sets were expected [1–13]. For example, the human protein coding gene census remained unfinished: contemporary estimates included about 20,000–21,000 protein coding genes in human genome [14–27]. In addition, the proven utility of public eutherian reference genomic sequences could become compromised by potential genomic sequence errors, including analytical and bioinformatical errors, as well as Sanger DNA sequencing method errors [28–33].

Data description

Using public eutherian reference genomic sequence assemblies and free available software, the eutherian comparative genomic analysis protocol was published as guidance against potential genomic sequence errors [34–49]. The protocol included 3 major processing steps that were integrated into one framework of eutherian gene data set descriptions: gene annotations, phylogenetic analysis and protein molecular evolution analysis. The protocol published 3 original genomics and protein molecular evolution tests. First, the test of reliability of public eutherian genomic sequences used genomic sequence redundancies of public eutherian reference genomic sequence assemblies. Second, the test of contiguity of public eutherian genomic sequences used multiple pairwise genomic sequence alignments. Third, the test of protein molecular evolution used relative synonymous codon usage statistics. The protocol was made available on Protocol Exchange [44].

In aggregate, the eutherian comparative genomic analysis protocol curated 14 eutherian gene data sets implicated in major physiological and pathological processes, including 2615 published complete coding sequences that were made available in public biological databases as third-party data gene data sets [50–63] (Table 1). The curated gene data sets were deposited in European Nucleotide Archive [7–9, 12, 13] in FASTA nucleotide sequence format. The published eutherian gene

*Correspondence: Marko.Premzl@alumni.anu.edu.au

[1] The Australian National University Alumni, 4 Kninski trg Sq., Zagreb, Croatia
Full list of author information is available at the end of the article

Table 1 Overview of eutherian third-party data gene data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data set 1	Interferon-γ-inducible GTPase genes (FR734011-FR734074)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:FR734011) [50]
Data set 2	Adenohypophysis cystine-knot genes (HF564658-HF564785)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:HF564658) [51]
Data set 3	Macrophage migration inhibitory factor genes (HF564786-HF564815)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:HF564786) [52]
Data set 4	Ribonuclease A genes (HG328835-HG329089)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:HG328835) [53]
Data set 5	Mas-related G protein-coupled receptor genes (HG426065-HG426183)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:HG426065) [54]
Data set 6	Lysozyme genes (HG931734-HG931849)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:HG931734) [55]
Data set 7	Growth hormone genes (LM644135-LM644234)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:LM644135) [56]
Data set 8	Tumor necrosis factor ligand genes (LN874312-LN874522)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:LN874312) [57]
Data set 9	Globin genes (LT548096-LT548244)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:LT548096) [58]
Data set 10	Kallikrein genes (LT631550-LT631670)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:LT631550) [59]
Data set 11	Adiponectin genes (LT962964-LT963174)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:LT962964) [60]
Data set 12	Connexin genes (LT990249-LT990597)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:LT990249) [61]
Data set 13	Fibroblast growth factor genes (LR130242-LR130508)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:LR130242) [62]
Data set 14	Interferon genes (LR760818-LR761312)	FASTA (.fas)	European Nucleotide Archive (https://identifiers.org/ena.embl:LR760818) [63]

collections were used in revisions and updates of eutherian gene data set classifications and nomenclatures.

Limitations

The revisions and updates of eutherian gene data sets were contingent on primary Sanger DNA sequencing information deposited in National Center for Biotechnology Information NCBI Trace Archive [12, 13, 46, 64–66]. For example, the positive correlation was calculated between genomic sequence redundancies of 35 public eutherian reference genomic sequence assemblies respectively and curated complete coding sequence numbers.

Acknowledgements
MP would like to thank manuscript reviewers on their manuscript reviews. MP would like to express his gratitude to data analysts, producers and providers of public eutherian reference genomic sequence data sets and free available software.

Author's contributions
MP conceived and prepared manuscript. The author read and approved final manuscript.

Funding
Not applicable.

Availability of data and materials
The data described in present Data note could be freely and openly accessed in European Nucleotide Archive under accessions: 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. Please, see Table 1 and references [50–63] for details and URLs.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
No competing interests were declared.

Author details
[1]The Australian National University Alumni, 4 Kninski trg Sq., Zagreb, Croatia. [2]https://www.ncbi.nlm.nih.gov/myncbi/mpremzl/cv/130205/.

Received: 27 March 2021 Accepted: 13 July 2022
Published online: 23 July 2022

References

1. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered. 2009;100:659–74.
2. Koepfli KP, Paten B, Genome 10K Community of Scientists, O'Brien SJ. The genome 10K project: a way forward. Annu Rev Anim Biosci. 2015;3:57–111.
3. Lewin HA, et al. Earth BioGenome project: sequencing life for the future of life. Proc Natl Acad Sci U S A. 2018;115:4325–33.
4. Gibbs RA. The human genome project changed everything. Nat Rev Genet. 2020;21:575–6.
5. Green ED, et al. Strategic vision for improving human health at the forefront of genomics. Nature. 2020;586:683–92.
6. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. Nature. 2020;587:240–5.
7. Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2021;49:D121–4.
8. Cantelli G, et al. The European bioinformatics institute: empowering cooperation in response to a global health crisis. Nucleic Acids Res. 2021;49:D29–37.
9. Harrison PW, et al. The European nucleotide archive in 2020. Nucleic Acids Res. 2021;49:D82–5.
10. Howe KL, et al. Ensembl 2021. Nucleic Acids Res. 2021;49:D884–91.
11. Murphy WJ, Foley NM, Bredemeyer KR, Gatesy J, Springer MS. Phylogenomics and the genetic architecture of the placental mammal radiation. Annu Rev Anim Biosci. 2021;9:29–53.
12. Sayers EW, et al. Database resources of the National Center for biotechnology information. Nucleic Acids Res. 2021;49:D10–7.
13. Sayers EW, et al. GenBank. Nucleic Acids Res. 2021;49:D92–6.
14. Clamp M, et al. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A. 2007;104:19428–33.
15. Temple G, et al. The completion of the mammalian gene collection (MGC). Genome Res. 2009;19:2324–33.
16. Pertea M, et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. 2018;19:208.
17. Pujar S, et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res. 2018;46:D221–8.
18. Salzberg SL. Open questions: how many genes do we have? BMC Biol. 2018;16:94.
19. Mudge JM, et al. Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. Genome Res. 2019;29:2073–87.
20. Zerbino DR, Frankish A, Flicek P. Progress, challenges, and surprises in annotating the human genome. Annu Rev Genomics Hum Genet. 2020;21:55–79.
21. Zhang D, et al. Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. Sci Adv. 2020;6:eaay8299.
22. Blake JA, et al. Mouse genome database (MGD): knowledgebase for mouse-human comparative biology. Nucleic Acids Res. 2021;49:D981–7.
23. Blum M, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49:D344–54.
24. Frankish A, et al. GENCODE 2021. Nucleic Acids Res. 2021;49:D916–23.
25. Gene Ontology Consortium. The gene ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49:D325–34.
26. Tweedie S, et al. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res. 2021;49:D939–46.
27. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9.
28. Hubisz MJ, Lin MF, Kellis M, Siepel A. Error and error mitigation in low-coverage genome assemblies. PLoS One. 2011;6:e17034.
29. Prosdocimi F, Linard B, Pontarotti P, Poch O, Thompson JD. Controversies in modern evolutionary biology: the imperative for error detection and quality control. BMC Genomics. 2012;13:5.
30. Norgren RB Jr. Improving genome assemblies and annotations for non-human primates. ILAR J. 2013;54:144–53.
31. Denton JF, et al. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 2014;10:e1003998.
32. Nagy A, Patthy L. FixPred: a resource for correction of erroneous protein sequences. Database (Oxford). 2014;2014:bau032.
33. Meyer C, et al. Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes. BMC Bioinformatics. 2020;21:513.
34. Premzl M. Comparative genomic analysis of eutherian interferon-γ-inducible GTPases. Funct Integr Genomics. 2012;12:599–607.
35. Premzl M. Comparative genomic analysis of eutherian ribonuclease a genes. Mol Gen Genomics. 2014;289:161–7.
36. Premzl M. Comparative genomic analysis of eutherian mas-related G protein-coupled receptor genes. Gene. 2014;540:16–9.
37. Premzl M. Third party annotation gene data set of eutherian lysozyme genes. Genom Data. 2014;2:258–60.
38. Premzl M. Initial description of primate-specific cystine-knot Prometheus genes and differential gene expansions of D-dopachrome tautomerase genes. Meta Gene. 2015;4:118–28.
39. Premzl M. Third party data gene data set of eutherian growth hormone genes. Genom Data. 2015;6:166–9.
40. Premzl M. Curated eutherian third party data gene data sets. Data Brief. 2016;6:208–13.
41. Premzl M. Comparative genomic analysis of eutherian tumor necrosis factor ligand genes. Immunogenetics. 2016;68:125–32.
42. Premzl M. Comparative genomic analysis of eutherian globin genes. Gene Rep. 2016;5:163–6.
43. Premzl M. Comparative genomic analysis of eutherian kallikrein genes. Mol Genet Metab Rep. 2017;10:96–9.
44. Premzl M. Eutherian comparative genomic analysis protocol. Protoc Exch. 2018. https://doi.org/10.1038/protex.2018.028.
45. Premzl M. Comparative genomic analysis of eutherian adiponectin genes. Heliyon. 2018;4:e00647.
46. Premzl M. Eutherian third-party data gene collections. Gene Rep. 2019;16:100414.
47. Premzl M. Comparative genomic analysis of eutherian connexin genes. Sci Rep. 2019;9:16938.
48. Premzl M. Comparative genomic analysis of eutherian fibroblast growth factor genes. BMC Genomics. 2020;21:542.
49. Premzl M. Comparative genomic analysis of eutherian interferon genes. Genomics. 2020;112:4749–59.
50. Premzl M. Accession numbers: FR734011-FR734074. Europ Nucleotide Arch. 2012; https://identifiers.org/ena.embl:FR734011.
51. Premzl M. Accession numbers: HF564658-HF564785. Europ Nucleotide Arch. 2015; https://identifiers.org/ena.embl:HF564658.
52. Premzl M. Accession numbers: HF564786-HF564815. Europ Nucleotide Arch. 2015; https://identifiers.org/ena.embl:HF564786.
53. Premzl M. Accession numbers: HG328835-HG329089. Europ Nucleotide Arch. 2014; https://identifiers.org/ena.embl:HG328835.
54. Premzl M. Accession numbers: HG426065-HG426183. Europ Nucleotide Arch. 2014; https://identifiers.org/ena.embl:HG426065.
55. Premzl M. Accession numbers: HG931734-HG931849. Europ Nucleotide Arch. 2014; https://identifiers.org/ena.embl:HG931734.
56. Premzl M. Accession numbers: LM644135-LM644234. Europ Nucleotide Arch. 2015; https://identifiers.org/ena.embl:LM644135.
57. Premzl M. Accession numbers: LN874312-LN874522. Europ Nucleotide Arch. 2016; https://identifiers.org/ena.embl:LN874312.
58. Premzl M. Accession numbers: LT548096-LT548244. Europ Nucleotide Arch. 2016; https://identifiers.org/ena.embl:LT548096.
59. Premzl M. Accession numbers: LT631550-LT631670. Europ Nucleotide Arch. 2017; https://identifiers.org/ena.embl:LT631550.
60. Premzl M. Accession numbers: LT962964-LT963174. Europ Nucleotide Arch. 2018; https://identifiers.org/ena.embl:LT962964.
61. Premzl M. Accession numbers: LT990249-LT990597. Europ Nucleotide Arch. 2019; https://identifiers.org/ena.embl:LT990249.
62. Premzl M. Accession numbers: LR130242-LR130508. Europ Nucleotide Arch. 2020; https://identifiers.org/ena.embl:LR130242.

63. Premzl M. Accession numbers: LR760818-LR761312. Europ Nucleotide Arch. 2020; https://identifiers.org/ena.embl:LR760818.
64. Blakesley RW, et al. An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res. 2004;14:2235–44.
65. Margulies EH, et al. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci U S A. 2005;102:4795–800.
66. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478:476–82.

Publisher's Note