Molecular Genetics and Metabolism Reports 10 (2017) 96–99

Contents lists available at ScienceDirect

Molecular Genetics and Metabolism Reports

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

Short Communication

Comparative genomic analysis of eutherian kallikrein genes

Marko Premzl

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

ARTICLE INFO

Article history:
Received 19 December 2016
Received in revised form 21 January 2017
Accepted 21 January 2017
Available online 3 February 2017

Keywords:
Comparative genomic analysis
Gene annotations
Molecular evolution
Phylogenetic analysis
RRID:SCR_014401

ABSTRACT

The present study made attempts to update and revise eutherian kallikrein genes implicated in major physiological and pathological processes and in medical molecular diagnostics. Using eutherian comparative genomic analysis protocol and free available genomic sequence assemblies, the tests of reliability of eutherian public genomic sequences annotated most comprehensive curated third party data gene data set of eutherian kallikrein genes including 121 complete coding sequences among 335 potential coding sequences. The present analysis first described 13 major gene clusters of eutherian kallikrein genes, and explained their differential gene expansion patterns. One updated classification and nomenclature of eutherian kallikrein genes was proposed, as new framework of future experiments.

1. Brief communication

1.1. Introduction

The eutherian kallikrein genes were implicated in major physiological and pathological processes, as well as in medical molecular diagnostics [2,11,13,23]. For example, the serum level of human prostate-specific antigen abundant in seminal plasma was used as molecular marker in diagnostics of prostate cancer. The comprehensive eutherian kallikrein gene data sets established major criteria of kallikrein gene annotations, including genomic localization within single genomic kallikrein locus region, common kallikrein gene features including 5 translated exons, as well as patterns of kallikrein protein amino acid sequence similarities including invariant catalytic triad amino acid sites His, Asp and Ser and distribution of common cysteine amino acid residues [2–4,10,11,13,15–17,22,23]. Specifically, the human kallikreins comprised 15 serine peptidases, including 3 classical kallikreins KLK1 or paradigmatic pancreatic kallikrein, KLK2 or glandular kallikrein and KLK3 or prostate-specific antigen, as well as KLK4–KLK15 kallikrein-related serine peptidases. Nevertheless, future updates and revisions of comprehensive eutherian gene data sets were expected, due to the incompleteness of public eutherian genomic sequence data sets [7,8] and potential genomic sequence errors [9]. For example, the so-called lexicographical bias was described in some genomic sequence assembly programs [5] and phylogenetic analyses could be affected by potential genomic sequence errors [18]. Thus, the eutherian comparative genomic analysis protocol applicable in curation of major eutherian gene data sets was established as assistance in protection against potential genomic sequence errors [19–21]. The protocol integrated gene annotations, phylogenetic analysis and protein molecular evolution analysis with new genomics and protein molecular evolution tests into one framework of eutherian gene descriptions. For example, the protocol revised, updated and published 9 major eutherian gene data sets, including 1172 complete coding sequences deposited in European Nucleotide Archive as curated third party data gene data sets. Using eutherian comparative genomic analysis protocol and public genomic sequence assemblies [1,12,14], the present study made attempts to update and revise eutherian kallikrein *KLN* genes.

1.2. Gene annotations

The sequence alignment editor BioEdit 7.0.5.3 was used in analyses and manipulations of *KLN* nucleotide and protein sequences (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). The public eutherian genomic sequence assemblies used in gene identifications of potential *KLN* coding sequences were downloaded from National Center for Biotechnology Information (NCBI) GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/vertebrate_mammalian/), as well as NCBI's BLAST programs (ftp://ftp.ncbi.nlm.nih.gov/blast/). In addition, the public eutherian genomic sequences were downloaded from Ensembl genome browser (http://www.ensembl.org). The analyses of *KLN* gene features included direct evidence of gene annotations deposited in NCBI's nr, est_human, est_mouse and est_others databases (https://www.ncbi.nlm.nih.gov). The potential *KLN* coding sequences were tested using tests of reliability of eutherian public genomic sequences. The first test steps analysed nucleotide sequence coverages of each potential *KLN* coding sequence, using NCBI's BLAST programs and primary experimental genomic sequence reads deposited in NCBI's Trace Archive (https://www.ncbi.nlm.nih.gov/Traces/trace.cgi). Only if consensus read nucleotide sequence coverages were available for every nucleotide of each

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

http://dx.doi.org/10.1016/j.ymgmr.2017.01.009

potential *KLN* coding sequence, the potential *KLN* coding sequences were annotated as complete *KLN* coding sequences and used in analyses. Alternatively, the potential *KLN* coding sequences were described as putative *KLN* coding sequences and not used in analyses. In addition, the complete *KLN* coding sequences were deposited in European Nucleotide Archive as eutherian third party data gene data set (http://www.

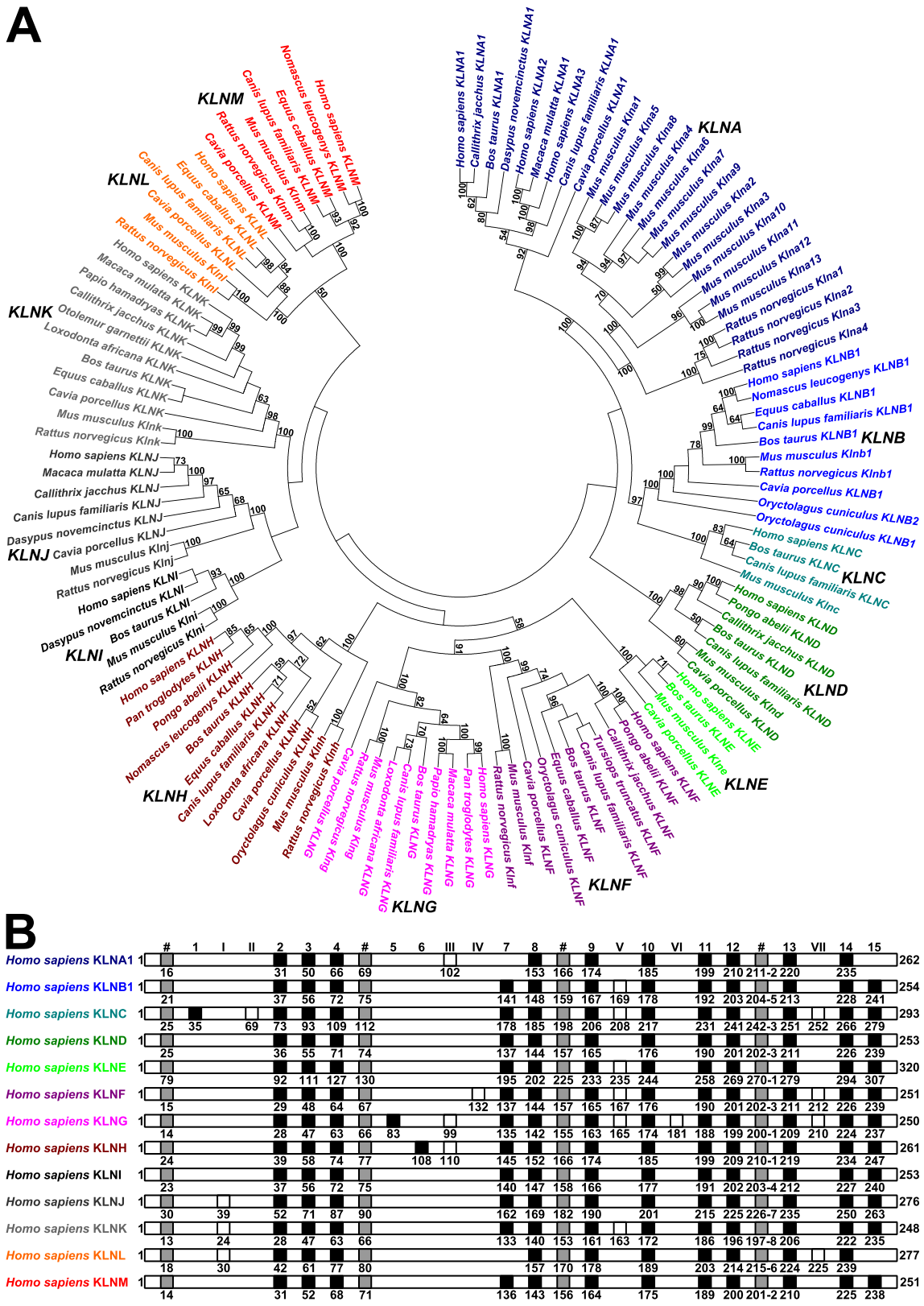

Fig. 1. (A) Phylogenetic analysis of eutherian kallikrein genes. The minimum evolution tree was calculated using maximum composite likelihood method. After 1000 replicates, the bootstrap estimates higher than 50% were shown. The major gene clusters *KLNA-KLNM* were indicated. (B) Eutherian kallikrein protein sequence alignment landmarks. Whereas the black squares labelled common cysteine amino acid residues (1-15), white squares labelled common N-glycosylation sites (I–VII) and grey squares labelled common exon-intron splice site amino acid sites (#). The numbers indicated numbers of amino acid residues.