Genomics Data 6 (2015) 166–169

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Third party data gene data set of eutherian growth hormone genes

Marko Premzl

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

ARTICLE INFO

Article history:
Received 29 August 2015
Received in revised form 6 September 2015
Accepted 7 September 2015
Available online 11 September 2015

Keywords:
Comparative genomic analysis
Gene annotations
Molecular evolution
Phylogenetic analysis

ABSTRACT

Among 146 potential coding sequences, the most comprehensive eutherian growth hormone gene data set annotated 100 complete coding sequences. The eutherian comparative genomic analysis protocol first described 5 major gene clusters of eutherian growth hormone genes. The present updated gene classification and nomenclature of eutherian growth hormone genes integrated gene annotations, phylogenetic analysis and protein molecular evolution analysis into new framework of future experiments. The curated third party data gene data set of eutherian growth hormone genes was deposited in European Nucleotide Archive under accession numbers LM644135–LM644234.

Specifications	
Organism/cell line/tissue	35 eutherian species
Sex	N/A
Sequencer or array type	Sanger DNA sequencing method sequencers
Data format	FAS, TXT
Experimental factors	Eutherian comparative genomic analysis protocol
Experimental features	Third party data gene data set
Consent	N/A
Sample source location	N/A

1. Direct link to deposited data

Deposited data could be found here: http://www.ebi.ac.uk/ena/data/view/LM644135–LM644234.

2. Experimental design, materials and methods

The eutherian comparative genomic analysis protocol included gene annotations, phylogenetic analysis and protein molecular evolution analysis [1].

2.1. Gene annotations

The BioEdit 7.0.5.3 program was used in nucleotide and protein sequence analyses (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). The NCBI's BLAST programs were used in identification of genes in eutherian genomic sequence assemblies downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/ and ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/) [2,3]. Alternatively, the Ensembl genome browser's BLAST or BLAT web tools were used in gene identifications (http://www.ensembl.org). The gene feature analyses included direct evidence of eutherian gene annotations deposited in NCBI's nr, est_human, est_mouse and est_others databases (http://www.ncbi.nlm.nih.gov). The protocol tested potential growth hormone (*GH*) coding sequences using tests of reliability of eutherian public genomic sequences. Using NCBI's BLAST programs and primary sequence reads deposited in NCBI's Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi), the first test step analysed nucleotide sequence coverage of potential coding sequences. The potential coding sequences were described as complete coding sequences in second test step only if consensus trace sequence coverage was available for every nucleotide. Alternatively, the potential coding sequences were described as putative coding sequences. Only the complete *GH* coding sequences were deposited in European Nucleotide Archive (http://www.ebi.ac.uk/ena/about/tpa-policy) and used in phylogenetic and protein molecular evolution analyses. In gene descriptions, the guidelines of human and mouse gene nomenclature were followed (http://www.genenames.org/about/guidelines and http://www.informatics.jax.org/mgihome/nomen/gene.shtml). There were 100 complete eutherian *GH* coding sequences, among 146 potential coding sequences (Fig. 1) (Supplementary data file 1). The most comprehensive third party data gene data set of eutherian *GH* genes annotated 15 *GHA* genes, 36 *GHB* genes, 5 *GHC* genes, 39 *GHD* genes and 5 *GHE* genes. The eutherian *GHA* genes were described as prolactin *PRL* orthologues, eutherian *GHB* genes were described as growth hormone *GH* orthologues and paralogues, domesticated guinea pig *GHC* genes were first described in present work, *Ghd* genes were described as prolactin paralogues in mouse and brown rat

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

http://dx.doi.org/10.1016/j.gdata.2015.09.007

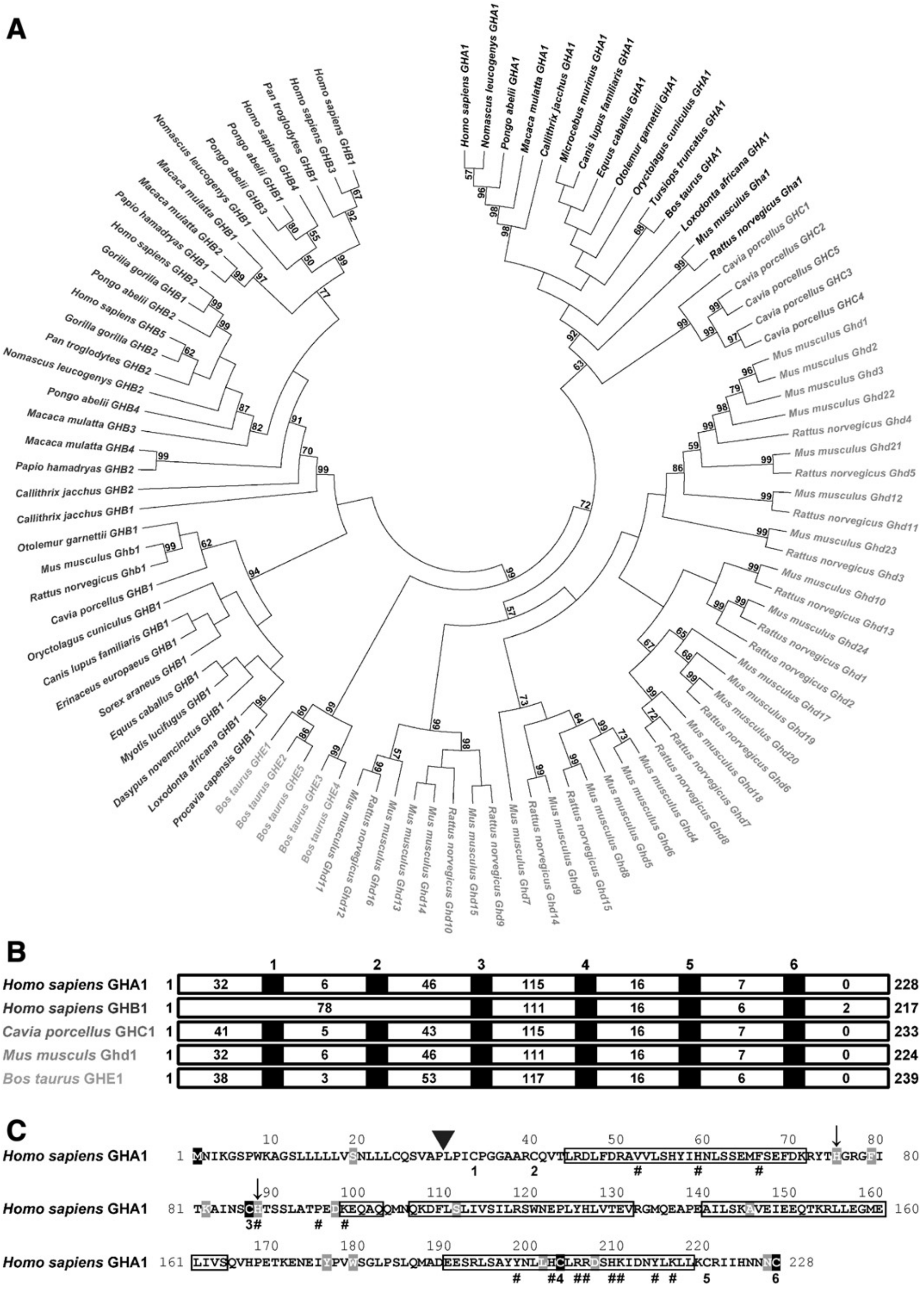

Fig. 1. (A) Phylogenetic analysis of eutherian growth hormone genes. The minimum evolution tree was calculated using maximum composite likelihood method. After 1000 bootstrap replicates, the estimates higher than 50% were shown. (B) Distribution of common cysteine amino acid residues in eutherian growth hormone proteins. The common Cys amino acid residues 1–6 were labelled using black rectangles. The numbers indicated numbers of amino acid residues. (C) Reference human GHA1 protein primary structure. The 4 invariant amino acid sites were shown using white letters on black backgrounds and 13 forward amino acid sites were shown using white letters on grey backgrounds. The common Cys amino acid residues 1–6 were labelled below reference protein amino acid sequence, as well as 14 predicted functional amino acid residues (#) [9]. The α-helical regions of human GHA1 tertiary structure 1N9D were labelled by rectangles [9]. The tertiary structure determinant amino acid sites H75 and H88 were indicated by arrows [10]. The predicted signal peptide cleavage site was indicated by black triangle.

and *GHE* genes were described as prolactin paralogues in domestic cattle [4–6]. The masking of transposable elements using RepeatMasker version open-4.0.3 was included as preparatory step in multiple pairwise genomic sequence alignments, using default settings except simple repeats and low complexity elements were not masked (sensitive mode, cross_match version 1.080812, RepBase Update 20130422, RM database version 20130422) (http://www.repeatmasker.org/). In genomic sequence alignments, the mVISTA web tool was used, using AVID alignment program and default settings (http://genome.lbl.gov/vista/index.shtml). Using ClustalW implemented in BioEdit 7.0.5.3, the common predicted promoter genomic sequence regions were aligned at nucleotide sequence level and then manually corrected. The pairwise nucleotide sequence identities of common predicted promoter genomic sequence regions were calculated using BioEdit 7.0.5.3, and used in statistical analysis (Microsoft Office Excel). The common predicted promoter genomic sequence regions of eutherian *GHA* and *GHB* genes were described (Supplementary data file 2, Supplementary data file 3). For example, among primates, the calculated patterns of average pairwise nucleotide sequence identities of common predicted promoter genomic sequence regions exceeded empirically determined cut-offs of detection of common genomic sequence regions. Whereas the average pairwise nucleotide sequence identity of primate *GHA* common predicted promoter genomic sequence regions was $\bar{a} = 0{,}872$ ($a_{\max} = 0{,}986$, $a_{\min} = 0{,}767$, $\bar{a}_{\text{ad}} = 0{,}074$) (Supplementary data file 2A, Supplementary data file 3A), average pairwise nucleotide sequence identity of primate *GHB* common predicted promoter genomic sequence regions was $\bar{a} = 0{,}844$ ($a_{\max} = 0{,}989$, $a_{\min} = 0{,}252$, $\bar{a}_{\text{ad}} = 0{,}111$) (Supplementary data file 2B, Supplementary data file 3B).

2.2. Phylogenetic analysis

The translated complete eutherian *GH* coding sequences were aligned at amino acid level using ClustalW implemented in BioEdit 7.0.5.3. Then the protein sequence alignments were manually corrected, as well as nucleotide sequence alignments (Supplementary data file 4). In phylogenetic tree calculations, the MEGA 6.06 program was used (http://www.megasoftware.net), using neighbour-joining method (default settings, except gaps/missing data treatment = pairwise deletion) (data not shown), minimum evolution method (default settings, except gaps/missing data treatment = pairwise deletion) and maximum parsimony method (default settings, except gaps/missing data treatment = use all sites) (data not shown). However, the maximum likelihood methods were not used in present analysis because their homogeneity and stationarity assumptions were not satisfied (data not shown). The pairwise nucleotide sequence identities of complete eutherian *GH* coding sequences were calculated using BioEdit 7.0.5.3, and used in statistical analysis (Microsoft Office Excel). The present work first described 5 eutherian *GHA*-*GHE* major gene clusters (Fig. 1). There were evidence of differential gene expansions in all eutherian *GH* major gene clusters, except *GHA* major gene cluster included orthologues only. For example, the present study confirmed that there were differential gene expansions of primate *GHB* paralogues [4,7], mouse and brown rat *GHD* paralogues [4,5] and domestic cattle *GHE* paralogues [4]. Of note, the present phylogenetic analysis first included completed eutherian *GH* gene data set. For example, the phylogenies of eutherian *GHA* and *GHB* major gene clusters, as well as phylogenies of domesticated guinea pig *GHC* and domestic cattle *GHE* major gene clusters were first described. The present phylogenetic analysis of primate *GHB* paralogues was in agreement with previous analyses [6,8]. In addition, the overall grouping within *Ghd* major gene cluster agreed with analysis of Soares et al. [5]. The calculated average pairwise nucleotide sequence identity of entire data set of eutherian *GH* homologues was $\bar{a} = 0{,}448$ ($a_{\max} = 0{,}995$, $a_{\min} = 0{,}224$, $\bar{a}_{\text{ad}} = 0{,}141$). Indeed, the updated and revised eutherian *GH* gene classification was confirmed by calculated patterns of pairwise nucleotide sequence identities of eutherian *GH* genes (Supplementary data file 5). First, whereas the eutherian *GHA* major gene cluster showed nucleotide sequence identities typical in comparisons between eutherian orthologues, eutherian *GHB* major gene cluster showed nucleotide sequence identities typical in comparisons between eutherian orthologues and paralogues. Next, the nucleotide sequence identities of eutherian *GHC* and *GHE* major gene clusters respectively were typical in comparisons between eutherian paralogues. However, there were calculated nucleotide sequence identity patterns of *Ghd* major gene cluster distant eutherian paralogues. Finally, there were nucleotide sequence identities of close eutherian homologues in comparisons between eutherian *GHA*, *GHC*, *Ghd* and *GHE* major gene clusters. Yet, in comparisons between eutherian *GHB* major gene cluster and other major gene clusters, there were nucleotide sequence identities of typical eutherian homologues.

2.3. Protein molecular evolution analysis

The tests of protein molecular evolution integrated patterns of nucleotide sequence similarities with protein tertiary structures. In codon usage statistic calculations, the MEGA 6.06 program was used. The ratios between observed and expected amino acid codon counts determined relative synonymous codon usage statistics (*R*). The not preferable amino acid codons with $R \le 0.7$ were TTA (0,28), TTG (0,56), CTA (0,54), ATA (0,62), GTA (0,38), TCG (0,39), CCG (0,34), ACG (0,41), GCG (0,15), TGT (0,55), CGT (0,54), CGA (0,54), AGT (0,63) and GGT (0,56). Accordingly, the reference protein sequence amino acid sites were indicated as invariant amino acid sites (invariant alignment positions), forward amino acid sites (variant alignment positions that did not include amino acid codons with $R \le 0.7$) or compensatory amino acid sites (variant alignment positions that included amino acid codons with $R \le 0.7$). The presence of preferable amino acid codons, as well as absence of not preferable amino acid codons indicated that forward amino acid sites could have major influence on protein tertiary structure and function. The DeepView/Swiss-PdbViever 4.1.0 (http://spdbv.vital-it.ch/) was used in analysis of human GHA1 tertiary structure 1N9D [9,10]. In prediction of N-terminal signal peptides, the SignalP 4.1 web tool was used, using default settings (http://www.cbs.dtu.dk/services/SignalP/). The present study first described 5 eutherian GH major protein clusters (Fig. 1). There were 6 common cysteine amino acid residues 1–6 present in eutherian GH proteins (Fig. 1B) (Supplementary data file 4). Whereas the eutherian GHB major protein cluster included 4 common Cys amino acid residues 3–6, there were 6 common Cys amino acid residues 1–6 present in other eutherian GH major protein clusters. Yet, in present eutherian GH protein data set, there were substitutions at common Cys residues 1, 2 and 5 (C33, C40 and C220 in human GHA1) but not at invariant common Cys amino acid residues 3, 4 and 6 (C87, C203 and C228 in human GHA1). Whereas the N-terminal signal peptides were predicted in all eutherian GH major protein clusters (data not shown), no invariant common potential N-glycosylation sites were found in eutherian GH major protein clusters. The present tests of protein molecular evolution included entire eutherian GH homologue data set (Fig. 1C) (Supplementary data file 4). The human GHA1 protein primary structure was used as reference protein amino acid sequence in analysis of human GHA1 tertiary structure 1N9D [9] (Supplementary data file 6). First, there were 4 invariant amino acid sites among 228 reference protein amino acid residues. For example, the invariant common Cys amino acid residues 3 and 4 (C87 and C203 in human GHA1) were implicated in disulfide linkage [10]. Second, there were 13 forward amino acid sites described in reference protein amino acid sequence. For example, the human GHA1 amino acid sites H75 and H88 were designated as major tertiary structure determinant amino acid residues [10].

3. Discussion

The eutherian *GH* third party data gene data set included genes implicated in major physiological processes [4–10]. For example, the human GH homologues were recorded in World Anti-Doping Code's Prohibited List (http://list.wada-ama.org/). The present updated gene classification and nomenclature of eutherian *GH* genes integrated gene annotations, phylogenetic analysis and protein molecular evolution analysis into new framework of future experiments.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.09.007.

References

[1] M. Premzl, Initial description of primate-specific cystine-knot prometheus genes and differential gene expansions of D-dopachrome tautomerase genes. Meta Gene 4 (2015) 118–128.

[2] R.W. Blakesley, N.F. Hansen, J.C. Mullikin, P.J. Thomas, J.C. McDowell, B. Maskeri, A.C. Young, B. Benjamin, S.Y. Brooks, B.I. Coleman, J. Gupta, S.L. Ho, E.M. Karlins, Q.L.

Maduro, S. Stantripop, C. Tsurgeon, J.L. Vogt, M.A. Walker, C.A. Masiello, X. Guan, N.I.S.C. Comparative Sequencing Program, G.G. Bouffard, E.D. Green, An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res. 14 (2004) 2235–2244.

[3] E.H. Margulies, J.P. Vinson, N.I.S.C. Comparative Sequencing Program, W. Miller, D.B. Jaffe, K. Lindblad-Toh, J.L. Chang, E.D. Green, E.S. Lander, J.C. Mullikin, M. Clamp, 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. 102 (2005) 4795–4800.

[4] M.J. Soares, The prolactin and growth hormone families: pregnancy-specific hormones/cytokines at the maternal-fetal interface. Reprod. Biol. Endocrinol. 2 (2004) 51.

[5] M.J. Soares, S.M. Alam, M.L. Duckworth, N.D. Horseman, T. Konno, D.I. Linzer, L.J. Maltais, M. Nilsen-Hamilton, K. Shiota, J.R. Smith, M. Wallis, A standardized nomenclature for the mouse and rat prolactin superfamilies. Mamm. Genome 18 (2007) 154–156.

[6] N. Petronella, G. Drouin, Gene conversions in the growth hormone gene family of primates: stronger homogenizing effects in the Hominidae lineage. Genomics 98 (2011) 173–181.

[7] R. González Alvarez, A. Revol de Mendoza, D. Esquivel Escobedo, G. Corrales Félix, I. Rodríguez Sánchez, V. González, G. Dávila, Q. Cao, P. de Jong, Y.X. Fu, H.A. Barrera Saldana, Growth hormone locus expands and diverges after the separation of New and Old World Monkeys. Gene 380 (2006) 38–45.

[8] Z. Papper, N.M. Jameson, R. Romero, A.L. Weckle, P. Mittal, K. Benirschke, J. Santolaya-Forgas, M. Uddin, D. Haig, M. Goodman, D.E. Wildman, Ancient origin of placental expression in the growth hormone genes of anthropoid primates. Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 17083–17088.

[9] C. Keeler, P.S. Dannies, M.E. Hodsdon, The tertiary structure and backbone dynamics of human prolactin. J. Mol. Biol. 328 (2003) 1105–1121.

[10] C. Keeler, M.C. Tettamanzi, S. Meshack, M.E. Hodsdon, Contribution of individual histidines to the global stability of human prolactin. Protein Sci. 18 (2009) 909–920.