Premzl *BMC Genomics* (2020) 21:542
https://doi.org/10.1186/s12864-020-06958-4

BMC Genomics

RESEARCH ARTICLE Open Access

Comparative genomic analysis of eutherian fibroblast growth factor genes

Marko Premzl

Abstract

Background: The eutherian fibroblast growth factors were implicated as key regulators in developmental processes. However, there were major disagreements in descriptions of comprehensive eutherian fibroblast growth factors gene data sets including either 18 or 22 homologues. The present analysis attempted to revise and update comprehensive eutherian fibroblast growth factor gene data sets, and address and resolve major discrepancies in their descriptions using eutherian comparative genomic analysis protocol and 35 public eutherian reference genomic sequence data sets.

Results: Among 577 potential coding sequences, the tests of reliability of eutherian public genomic sequences annotated most comprehensive curated eutherian third-party data gene data set of fibroblast growth factor genes including 267 complete coding sequences. The present study first described 8 superclusters including 22 eutherian fibroblast growth factor major gene clusters, proposing their updated classification and nomenclature.

Conclusions: The integrated gene annotations, phylogenetic analysis and protein molecular evolution analysis argued that comprehensive eutherian fibroblast growth factor gene data set classifications included 22 rather than 18 homologues.

Keywords: Gene annotations, Eutheria, Molecular evolution, Phylogenetic analysis, RRID:SCR_014401

Background

The eutherian fibroblast growth factors or FGFs were implicated as key developmental regulators [1–3]. First, the 15 paradigmatic paracrine or canonical fibroblast growth factors FGF1–10, FGF16–18, FGF20 and FGF22 were described as ligands to single-chain receptor tyrosine kinases named FGF receptors or FGFRs [2–11]. After paracrine FGF ligand and heparan sulphate glycosaminoglycan binding, the dimerized FGFRs become activated through autophosphorylation, interacting with cytosolic adaptor proteins and intracellular signaling cascades. Such transmembrane signal transduction was implicated in regulation of embryogenesis, implantation, gastrulation, body plan formation, branching morphogenesis and organogenesis, as well as in pathogeneses of human hereditary diseases including deafness, Kallmann syndrome, lacrimo-auriculo-dentodigital syndrome and different skeletal syndromes, and in tumorigenesis. Second, there were 3 endocrine fibroblast growth factors FGF19, FGF21 and FGF23 binding FGFRs and klotho protein cofactors [2, 3, 7, 12]. The endocrine FGFs were implicated in metabolism regulation including phosphate and vitamin D homeostasis, cholesterol and bile acid homeostasis and glucose and lipid homeostasis, as well as in pathogenesis of autosomal dominant hypophosphataemic rickets. Third, the 4 intracellular fibroblast growth factors named fibroblast homologous factors included FGF11 or FHF3, FGF12 or FHF1, FGF13 or FHF2 and FGF14 or FGF4 [1, 3, 13–16]. The intracellular FGFs were described as regulators of nervous system development and function including integration and encoding of complex synaptic inputs into action potential outputs in central nervous system neurons, and

Correspondence: Marko.Premzl@alumni.anu.edu.au
The Australian National University Alumni, 4 Kninski trg Sq., Zagreb, Croatia

implicated in pathogenesis of early-onset spinocerebellar ataxia. The molecular evolution and protein structure analyses indicated that eutherian FGFs folded into β-trefoil protein tertiary structures including 11 or 12 β-strands [1–3, 7, 12, 13, 17–28]. However, there were major disagreements in descriptions of comprehensive eutherian *FGF* gene data sets. Specifically, Belov and Mohammadi [2] and Beenken and Mohammadi [7] argued that bona fide eutherian FGF homologues included 18 secreted paracrine and endocrine FGFs. On the other hand, the eutherian FGF classifications by Goldfarb [1] and Ornitz and Itoh [3] included both 18 secreted FGFs and 4 intracellular FGFs.

Undoubtedly, the public eutherian reference genomic sequence data sets advanced biological and medical sciences [29–34]. Indeed, the comparative genomics momentum was maintained by considerable international efforts in production and analysis of public eutherian reference genomic sequence data sets. For example, the initial sequencing and analysis of human genome attempted to revise and update human genes, and uncover potential new drugs, drug targets and molecular markers in medical diagnostics [35, 36]. Nevertheless, due to the incompleteness of eutherian reference genomic sequence assemblies [35, 37] and potential genomic sequence errors [36, 38], future updates and revisions of public eutherian reference genomic sequence data sets were expected. Inevitably, the potential genomic sequence errors including analytical and bioinformatical errors (erroneous gene annotations, genomic sequence misassemblies) and Sanger DNA sequencing method errors (artefactual nucleotide deletions, insertions and substitutions) could compromise unquestionable utility of public eutherian reference genomic sequence data sets. For example, Gajer et al. [39] described so-called lexicographical bias in some genomic sequence assemblers. In addition, the potential genomic sequence errors affecting phylogenetic analyses [40] were observed more frequently in reference genomic sequence assemblies including lower genomic sequence redundancies [41–43]. Thus, the eutherian comparative genomic analysis protocol was established as guidance in protection against potential genomic sequence errors in public eutherian reference genomic sequence data sets [44–46]. Using public eutherian reference genomic sequence data sets, the protocol published new test of reliability of public eutherian genomic sequences using genomic sequence redundancies, and new test of protein molecular evolution using relative synonymous codon usage statistics. The protocol revised and updated 12 eutherian gene data sets implicated in major physiological and pathological processes, including 1853 published complete coding sequences. Of note, there was positive correlation between genomic sequence redundancies of 35 public eutherian reference genomic sequence data sets respectively and published complete coding sequence numbers [46].

Therefore, the present analysis attempted to revise and update comprehensive eutherian *FGF* gene data sets, and address and resolve major disagreements in their descriptions using eutherian comparative genomic analysis protocol and 35 public eutherian reference genomic sequence data sets.

Results

Gene annotations

The tests of reliability of eutherian public genomic sequences annotated 267 *FGF* complete coding sequences among 577 *FGF* potential coding sequences (Fig. 1). The most comprehensive curated eutherian *FGF* third-party data gene data set was deposited in European Nucleotide Archive under accessions: LR130242-LR130508 [47, 48] (Additional file 1).

The present study first described 8 superclusters *FGF1–8* including 22 major gene clusters of eutherian *FGF* genes, proposing their updated nomenclature (Fig. 1). The supercluster *FGF1* included 4 major gene clusters *FGF1A* (11 *FGF12* or *FHF1* genes), *FGF1B* (9 *FGF14* or *FGF4* genes), *FGF1C* (11 *FGF13* or *FHF2* genes) and *FGF1D* (15 *FGF11* or *FHF3* genes) (Additional file 2A-D). The supercluster *FGF2* included 2 major gene clusters *FGF2A* (8 *FGF2* genes) and *FGF2B* (20 *FGF1* genes) (Additional file 2E-F). The supercluster *FGF3* included 1 major gene cluster *FGF3A* (17 *FGF5* genes) (Additional file 2G). The supercluster *FGF4* included 3 major gene clusters *FGF4A* (11 *FGF20* genes), *FGF4B* (16 *FGF9* genes) and *FGF4C* (14 *FGF16* genes) (Additional file 2H-J). The supercluster *FGF5* included 4 major gene clusters *FGF5A* (14 *FGF10* genes), *FGF5B* (16 *FGF7* genes), *FGF5C* (7 *FGF3* genes) and *FGF5D* (9 *FGF22* genes) (Additional file 2 K-N). The supercluster *FGF6* included 3 major gene clusters *FGF6A* (5 *FGF18* genes), *FGF6B* (12 *FGF17* genes) and *FGF6C* (7 *FGF8* genes) (Additional file 2O-Q). The supercluster *FGF7* included 2 major gene clusters *FGF7A* (8 *FGF4* genes) and *FGF7B* (17 *FGF6* genes) (Additional file 2R-S). Finally, The supercluster *FGF8* included 3 major gene clusters *FGF8A* (12 *FGF19* genes), *FGF8B* (12 *FGF23* genes) and *FGF8C* (16 *FGF21* genes) (Additional file 2 T-V).

The present study included new genomics tests of contiguity of eutherian public genomic sequences that analysed numbers of coding exons in *FGF* genes and their relative orientation (Additional files 1 and 2). The analysis including 903 *FGF* coding exons indicated that there were no coding exon misassemblies among 267 eutherian genomic sequences harbouring *FGF* complete coding sequences. The eutherian *FGF* genes included either 5 coding exons (5 major gene clusters *FGF1A-D*

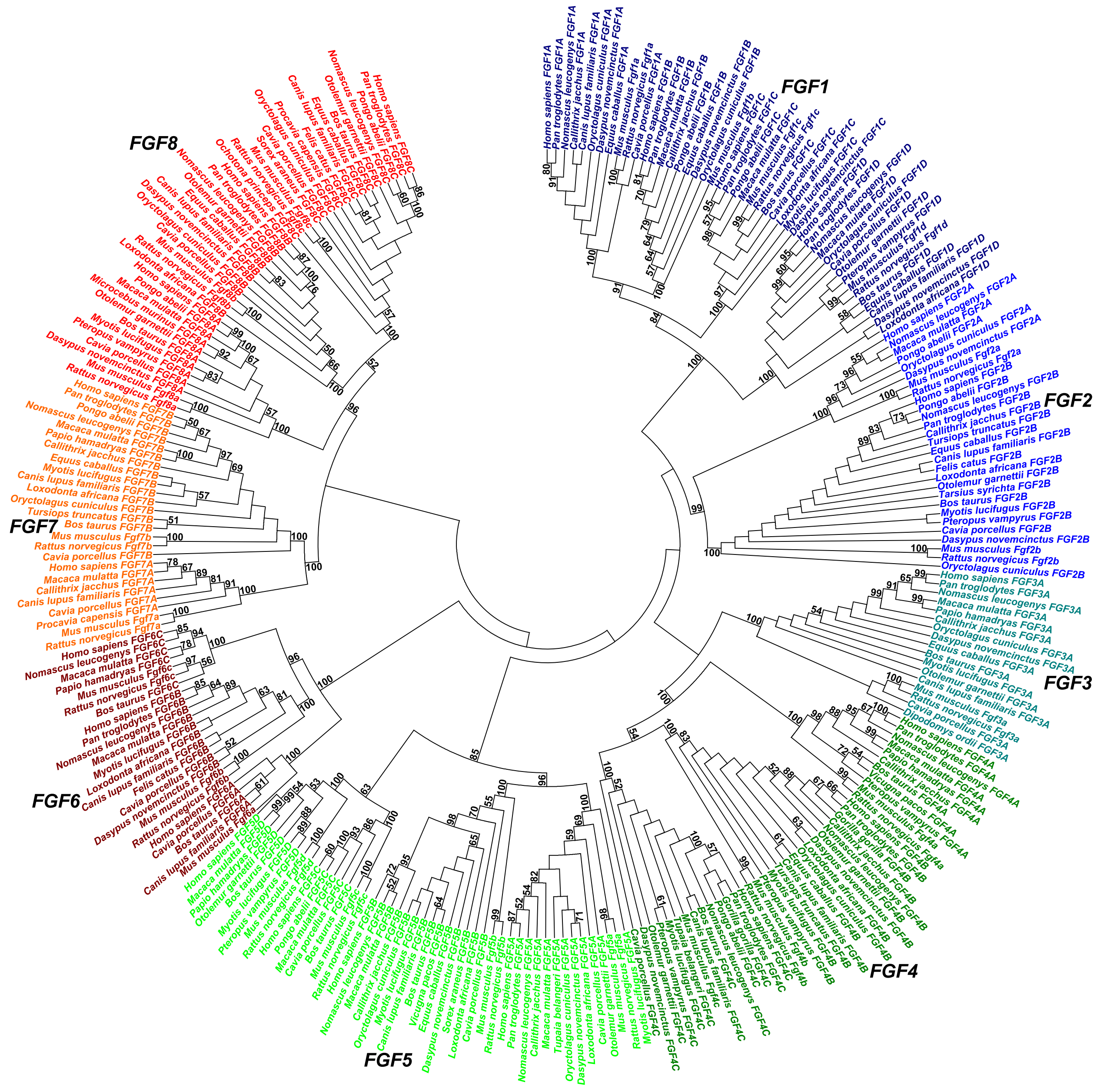


Fig. 1 Phylogenetic analysis of eutherian fibroblast growth factor genes. The minimum evolution phylogenetic tree including bootstrap estimates higher than 50% after 1000 replicates was calculated using maximum composite likelihood method. The 8 major gene superclusters *FGF1–8* were indicated

and *FGF6A*) or 3 coding exons (17 other major gene clusters). The eutherian *FGF* coding exon numbers were constant within major gene clusters, and there was no evidence of differential gene expansions indicating that 22 eutherian *FGF* major gene clusters respectively included orthologues. For example, whereas the human *FGF1A* gene included 5 coding exons along 264,215 bp (Additional file 2A), human *FGF7A* gene included 3 coding exons along 1776 bp (Additional file 2R).

Therefore, the present study annotating 22 eutherian *FGF* major gene clusters agreed with Goldfarb [1] and Ornitz and Itoh [3] but disagreed with Belov and Mohammadi [2] and Beenken and Mohammadi [7].

Phylogenetic analysis

The present minimum evolution phylogenetic tree calculations (Fig. 1) and calculations of pairwise nucleotide sequence identity patterns (Additional file 3) first classified 22 eutherian *FGF* major gene clusters among 8 superclusters *FGF1–8*. The clustering of major gene clusters *FGF1A-D* within supercluster *FGF1* agreed with subfamily *FGF11* descriptions [3, 23], Smallwood et al. [13], Ornitz and Itoh [21], subfamily *Fgf11/12/13/14* description [25] and Nam et al. [28]. The clustering of major gene clusters *FGF2A-B* within supercluster *FGF2* agreed with subfamily *FGF1* descriptions [3, 23], Smallwood et al. [13], Coulier et al. [17], Ornitz and Itoh [21], subfamily *Fgf1/2* description [25] and Nam et al. [28]. The supercluster *FGF3* description including 1 major gene cluster *FGF3A* agreed with Nam et al. [28] but disagreed with phylogenetic analyses of Ornitz and Itoh [3, 21], Coulier et al. [17] and Itoh and Ornitz [23, 25]. The clustering of major gene clusters *FGF4A-C* within supercluster *FGF4* agreed with subfamily *FGF9* descriptions [3, 23], Ornitz and Itoh [21] and subfamily *Fgf9/16/20* description [25] but disagreed with Nam et al. [28]. The clustering of major gene clusters *FGF5A-D* within supercluster *FGF5* disagreed with phylogenetic analyses of Ornitz and Itoh [3, 21], Itoh and Ornitz [23, 25] and Nam et al. [28]. The clustering of major gene clusters *FGF6A-C* within supercluster *FGF6* agreed with subfamily *FGF8* descriptions [3, 23], Ornitz and Itoh [21], subfamily *Fgf8/17/18* description [25] and Nam et al. [28]. The clustering of major gene clusters *FGF7A-B* within supercluster *FGF7* agreed with Smallwood et al. [13], Coulier et al. [17], Ornitz and Itoh [21] and Nam et al. [28] but disagreed with Ornitz and Itoh [3] and Itoh and Ornitz [23, 25]. Finally, the clustering of major gene clusters *FGF8A-C* within supercluster *FGF8* agreed with Ornitz and Itoh [21] but disagreed with Ornitz and Itoh [3], Itoh and Ornitz [23, 25] and Nam et al. [28].

Indeed, the calculations of pairwise nucleotide sequence identity patterns confirmed present phylogenetic classification of eutherian *FGF* genes (Additional file 3). The eutherian *FGF* gene data set included average pairwise nucleotide sequence identity $\bar{a} = 0{,}3$ ($a_{\max} = 1$, $a_{\min} = 0{,}115$, $\bar{a}_{\mathrm{ad}} = 0{,}094$) [1–3, 7, 12, 13, 17, 21, 23, 25–28]. Among 22 eutherian *FGF* major gene clusters respectively, there were nucleotide sequence identity patterns of very close eutherian orthologues (*FGF1A-B*, *FGF4B*), close eutherian orthologues (*FGF1C-D*, *FGF2A-B*, *FGF4A*, *FGF4C*, *FGF5B*, *FGF6A*, *FGF7B*), typical eutherian orthologues (*FGF3A*, *FGF5A*, *FGF5C-D*, *FGF6B-C*, *FGF7A*, *FGF8A*, *FGF8C*) and distant eutherian orthologues (*FGF8B*). In comparisons between eutherian *FGF* major gene clusters within superclusters, there were nucleotide sequence identity patterns of very close eutherian homologues (superclusters *FGF1–2*, *FGF4*, *FGF7*), very close and close eutherian homologues (supercluster *FGF6*), close and typical eutherian homologues (supercluster *FGF5*) and typical eutherian homologues (supercluster *FGF8*). Finally, in comparisons between eutherian *FGF* major gene clusters between superclusters, there were nucleotide sequence identity patterns of close, typical, distant and very distant eutherian homologues.

Therefore, the present phylogenetic analysis proposed updated classification of eutherian *FGF* genes.

Protein molecular evolution analysis

The protein molecular evolution analysis used protein primary structure features as major alignment landmarks in eutherian FGF protein amino acid sequence alignments, including common cysteine amino acid residues, common exon-intron splice site amino acid sites and common predicted N-glycosylation sites (Fig. 2) (Additional file 4). There were between 1 and 9 common cysteine amino acid residues included among eutherian FGF major protein clusters respectively. For example, whereas the major protein cluster FGF5D included 1 common cysteine amino acid residue, major protein cluster FGF5A included 9 common cysteine amino acid residues. There were either 4 common exon-intron splice site amino acid sites (5 major protein clusters FGF1A-D and FGF6A) or 2 common exon-intron splice site amino acid sites (17 other major protein clusters) among eutherian FGF major protein clusters respectively. Finally, there were between 0 and 2 common predicted N-glycosylation sites among eutherian FGF major protein clusters respectively.

Next, the tests of protein molecular evolution first calculated relative synonymous codon usage statistics (*R*) of eutherian *FGF* gene data set using 267 *FGF* complete coding sequences (Additional file 4), and described 20 amino acid codons including $R \leq 0{,}7$ as not preferable amino acid codons (Fig. 3a). The tests used human FGF1A protein primary structure as reference protein amino acid sequence (Fig. 3b). Among 243 human FGF1A protein amino acid residues, the tests of protein molecular evolution described 19 invariant amino acid sites, viz.: M1, C41, C55, P68, Q69, L70, K71, G72, I73, V74, T75, L77, G112, M129, G133, C145, Y159, G181 and C206, as well as 3 forward amino acid sites S101, E149 and Y208. First, the human FGF1A amino acid sites M1, L77, G133, C145 and Y159 were invariant among 267 eutherian FGF protein primary structures (except that M1 was invariant among 266 FGF protein primary structures). For example, the human FGF1A invariant amino acid sites L77, G133 and C145 were described by Goetz et al. [12, 24], Smallwood et al. [13], Coulier et al. [17], Venkataraman et al. [18], Plotnikov et al. [19] and Olsen et al. [22]. Furthermore, the human FGF1A amino acid sites G112 and M129 respectively were invariant among 21 eutherian FGF major protein clusters. For example, the human FGF1A amino acid site G112 was homologous to human FGF2B amino amino acid site G67 that was implicated in interactions between FGF2B ligand and FGFR2 receptor [19, 20]. In addition,

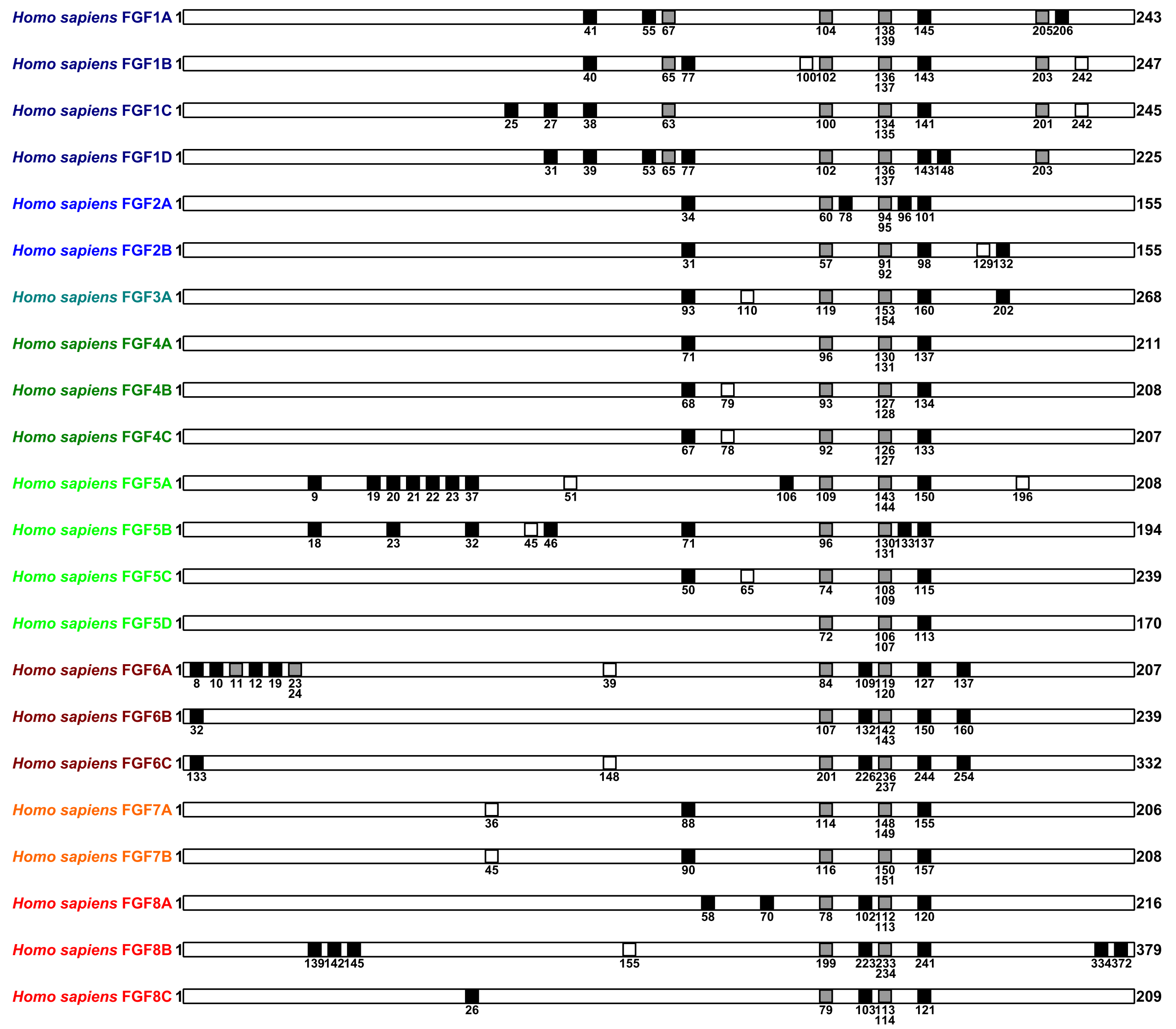

Homo sapiens FGF1A 1 41 55 67 104 138 139 145 205206 243
Homo sapiens FGF1B 1 40 65 77 100102 136 137 143 203 242 247
Homo sapiens FGF1C 1 25 27 38 63 100 134 135 141 201 242 245
Homo sapiens FGF1D 1 31 39 53 65 77 102 136 137 143148 203 225
Homo sapiens FGF2A 1 34 60 78 94 95 96 101 155
Homo sapiens FGF2B 1 31 57 91 92 98 129132 155
Homo sapiens FGF3A 1 93 110 119 153 154 160 202 268
Homo sapiens FGF4A 1 71 96 130 131 137 211
Homo sapiens FGF4B 1 68 79 93 127 128 134 208
Homo sapiens FGF4C 1 67 78 92 126 127 133 207
Homo sapiens FGF5A 1 9 19 20 21 22 23 37 51 106 109 143 144 150 196 208
Homo sapiens FGF5B 1 18 23 32 45 46 71 96 130133137 131 194
Homo sapiens FGF5C 1 50 65 74 108 109 115 239
Homo sapiens FGF5D 1 72 106 107 113 170
Homo sapiens FGF6A 1 8 10 11 12 19 23 24 39 84 109119 120 127 137 207
Homo sapiens FGF6B 1 32 107 132142 143 150 160 239
Homo sapiens FGF6C 1 133 148 201 226236 237 244 254 332
Homo sapiens FGF7A 1 36 88 114 148 149 155 206
Homo sapiens FGF7B 1 45 90 116 150 151 157 208
Homo sapiens FGF8A 1 58 70 78 102112 113 120 216
Homo sapiens FGF8B 1 139142145 155 199 223233 234 241 334372 379
Homo sapiens FGF8C 1 26 79 103113 114 121 209

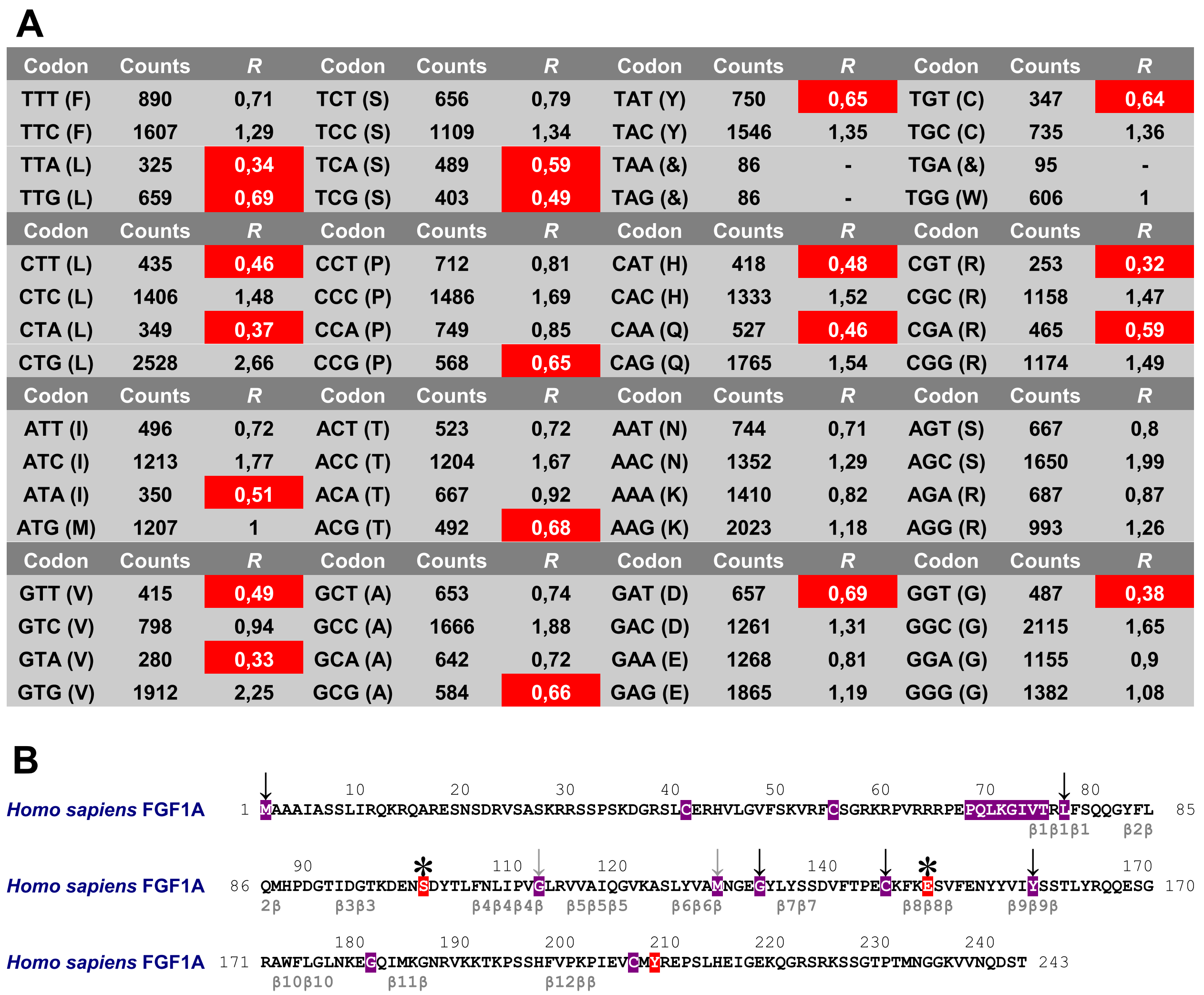
Fig. 2 Major landmarks in eutherian fibroblast growth factor protein sequence alignments. The black squares labelled common cysteine amino acid residues. The grey squares labelled common exon-intron splice site amino acid sites. The white squares labelled common N-glycosylation sites. The numbers indicated numbers of amino acid residues

the human FGF1A amino acid site G181 that was invariant among 7 eutherian FGF1–7 protein superclusters was described as first glycine amino acid residue in paracrine FGF glycine box protein amino acid sequence motif G-x(4)-G-x(2)-S/T [2]. The human FGF1A amino acid sites P68, Q69, L70, K71, G72, I73, V74 and T75 were invariant among 4 eutherian FGF1A-D major protein clusters. For example, the human FGF1A amino acid sites K71 and I73 were described as residues engaged in voltage-gated sodium channel binding [24]. Finally, the human FGF1A forward amino acid sites S101 and E149 were described among 267 eutherian FGF protein primary structures, and forward amino acid site Y208 was described among 2 eutherian FGF1–2 protein superclusters. For example, the human FGF1A forward amino acid site E149 was homologous to human FGF2A amino amino acid site E105 that was implicated in hydrogen bonding between FGF2A ligand and D3 domain of FGFR2 receptor [19, 26].

Therefore, the tests of protein molecular evolution using relative synonymous codon usage statistics described amino acid sites implicated as critical in FGF protein secondary, tertiary and quaternary structural features.

Discussion

The major disagreements in descriptions of comprehensive eutherian *FGF* gene data sets included classifications



A

Codon	Counts	*R*	Codon	Counts	*R*	Codon	Counts	*R*	Codon	Counts	*R*
TTT (F)	890	0,71	TCT (S)	656	0,79	TAT (Y)	750	0,65	TGT (C)	347	0,64
TTC (F)	1607	1,29	TCC (S)	1109	1,34	TAC (Y)	1546	1,35	TGC (C)	735	1,36
TTA (L)	325	0,34	TCA (S)	489	0,59	TAA (&)	86	-	TGA (&)	95	-
TTG (L)	659	0,69	TCG (S)	403	0,49	TAG (&)	86	-	TGG (W)	606	1
Codon	Counts	*R*	Codon	Counts	*R*	Codon	Counts	*R*	Codon	Counts	*R*
CTT (L)	435	0,46	CCT (P)	712	0,81	CAT (H)	418	0,48	CGT (R)	253	0,32
CTC (L)	1406	1,48	CCC (P)	1486	1,69	CAC (H)	1333	1,52	CGC (R)	1158	1,47
CTA (L)	349	0,37	CCA (P)	749	0,85	CAA (Q)	527	0,46	CGA (R)	465	0,59
CTG (L)	2528	2,66	CCG (P)	568	0,65	CAG (Q)	1765	1,54	CGG (R)	1174	1,49
Codon	Counts	*R*	Codon	Counts	*R*	Codon	Counts	*R*	Codon	Counts	*R*
ATT (I)	496	0,72	ACT (T)	523	0,72	AAT (N)	744	0,71	AGT (S)	667	0,8
ATC (I)	1213	1,77	ACC (T)	1204	1,67	AAC (N)	1352	1,29	AGC (S)	1650	1,99
ATA (I)	350	0,51	ACA (T)	667	0,92	AAA (K)	1410	0,82	AGA (R)	687	0,87
ATG (M)	1207	1	ACG (T)	492	0,68	AAG (K)	2023	1,18	AGG (R)	993	1,26
Codon	Counts	*R*	Codon	Counts	*R*	Codon	Counts	*R*	Codon	Counts	*R*
GTT (V)	415	0,49	GCT (A)	653	0,74	GAT (D)	657	0,69	GGT (G)	487	0,38
GTC (V)	798	0,94	GCC (A)	1666	1,88	GAC (D)	1261	1,31	GGC (G)	2115	1,65
GTA (V)	280	0,33	GCA (A)	642	0,72	GAA (E)	1268	0,81	GGA (G)	1155	0,9
GTG (V)	1912	2,25	GCG (A)	584	0,66	GAG (E)	1865	1,19	GGG (G)	1382	1,08

B
10 20 30 40 50 60 70 80
Homo sapiens FGF1A 1 MAAAIASSLIRQKRQARESNSDRVSASKRRSSPSKDGRSLCERHVLGVFSKVRFCSGRKRPVRRRPEPQLKGIVTRLFSQQGYFL 85
β1β1β1 β2β
90 110 120 140 170
Homo sapiens FGF1A 86 QMHPDGTIDGTKDENSDYTLFNLIPVGLRVVAIQGVKASLYVAMNGEGYLYSSDVFTPECKFKESVFENYYVIYSSTLYRQQESG 170
2β β3β3 β4β4β4β β5β5β5 β6β6β β7β7 β8β8β β9β9β
180 190 200 210 220 230 240
Homo sapiens FGF1A 171 RAWFLGLNKEGQIMKGNRVKKTKPSSHFVPKPIEVCMYREPSLHEIGEKQGRSRKSSGTPTMNGGKVVNQDST 243
β10β10 β11β β12ββ

Fig. 3 Tests of protein molecular evolution of eutherian fibroblast growth factors. **a** Relative synonymous codon usage statistics of eutherian *FGF* gene data set. The not preferable amino acid codons were indicated by white letters on red backgrounds. Counts, observed amino acid codon counts; *R*, relative synonymous codon usage statistics; &, stop codons. **b** Reference human FGF1A protein amino acid sequence. The 19 invariant amino acid sites were shown using white letters on violet backgrounds. Whereas the 5 amino acid sites that were invariant among 22 FGF major protein clusters were indicated by black arrows (except that M1 was invariant among 266 FGF protein primary structures), grey arrows indicated 2 amino acid sites that were invariant among 21 FGF major protein clusters respectively. The 3 forward amino acid sites were shown using white letters on red backgrounds. The stars labelled 2 forward amino acid sites described among 22 FGF major protein clusters. The positions of 12 β-strands implicated in β-trefoil protein tertiary structure were indicated below reference human FGF1A protein primary structure [22, 24]

of either 18 *FGF* genes [2, 7] or 22 *FGF* genes [1, 3]. The present analysis attempted to address and resolve these discrepancies using eutherian comparative genomic analysis protocol and public eutherian reference genomic sequence data sets [29–36, 44–46]. The advantages of eutherian reference genomic sequence data sets were well established phylogeny [29, 30, 34] and calibrated taxon sampling including genomic sequence redundancies that were applicable in tests of reliability of eutherian public genomic sequences [31–33]. Therefore, the tests of reliability of eutherian public genomic sequences annotated most comprehensive curated eutherian third-party data gene data set of *FGF* genes that included 267 complete coding sequences among 577 potential coding sequences. Second, the present study first described 8 superclusters of eutherian *FGF* genes that included 22 major gene clusters, proposing their updated nomenclature. Third, the new genomics tests of contiguity of eutherian public genomic sequences included 903 coding exons, and annotated either 3 or 5 coding exons in eutherian *FGF* genes including no evidence of differential gene expansions. Fourth, the present phylogenetic analysis proposed updated classification of eutherian *FGF* genes. Finally, the tests of protein molecular evolution using relative synonymous codon usage statistics described 19 invariant amino acid sites and 3 forward amino acid sites in reference human FGF1A protein primary structure, including amino acid residues described as critical in FGF protein secondary, tertiary and quaternary structural features. In conclusion, the present comparative genomic analysis integrating gene annotations, phylogenetic analysis and protein molecular evolution analysis argued that 22 *FGF* genes [1, 3], rather than 18 *FGF* genes [2, 7], were included in comprehensive eutherian *FGF* gene data set classifications.

Methods

Eutherian comparative genomic analysis protocol

The eutherian comparative genomic analysis protocol RRID:SCR_014401 integrated gene annotations, phylogenetic analysis and protein molecular evolution analysis with tests of reliability of eutherian public genomic sequences, tests of contiguity of eutherian public genomic sequences and tests of protein molecular evolution into one framework of eutherian gene descriptions (Fig. 4) [44–46].

Gene annotations

The protocol used gene identifications in 35 public genomic sequence assemblies, tests of reliability of eutherian public genomic sequences and new genomics tests of contiguity of eutherian public genomic sequences in eutherian *FGF* gene annotations. First, the sequence alignment editor BioEdit 7.0.5.3 was used in all analyses and manipulations of nucleotide and protein sequences [49]. The National Center for Biotechnology Information (NCBI) BLAST Genomes was used in identifications of *FGF* potential coding sequences in eutherian reference genomic sequence data sets [50–53], as well as Ensembl genome browser BLAST or BLAT tools [54, 55]. Second,

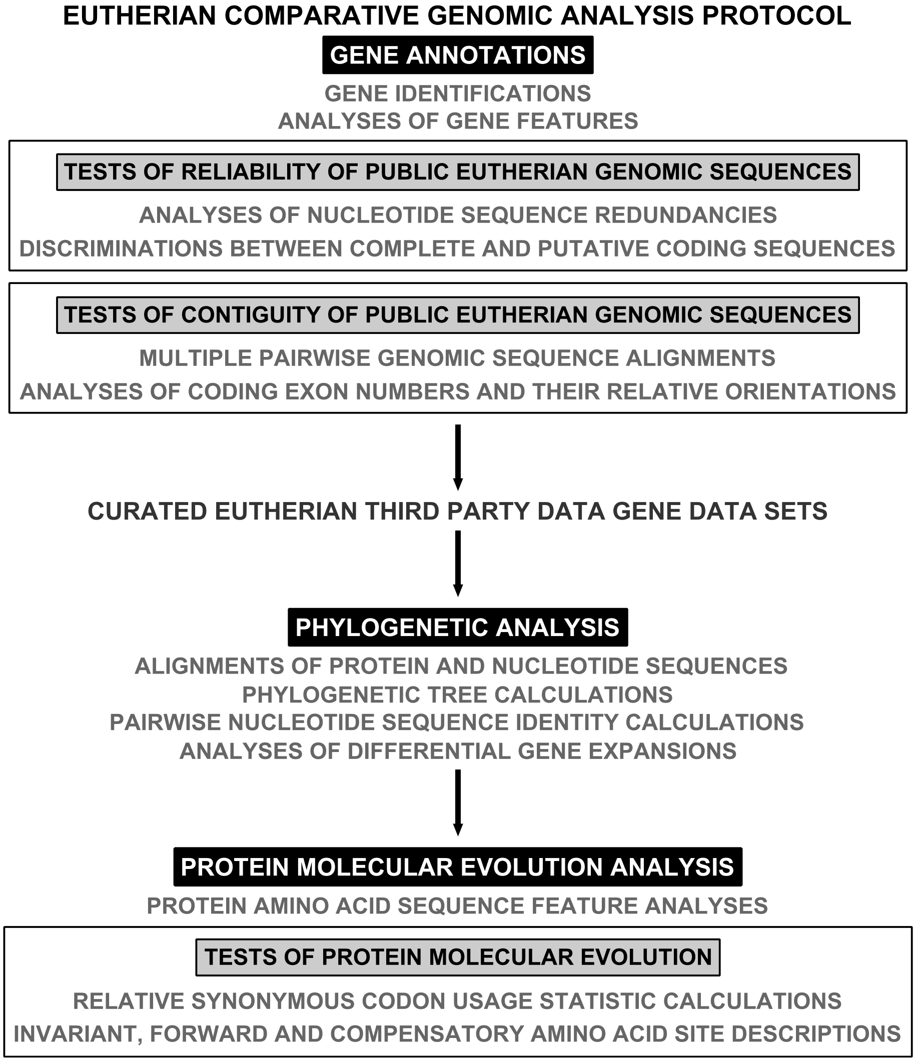

EUTHERIAN COMPARATIVE GENOMIC ANALYSIS PROTOCOL
GENE ANNOTATIONS
GENE IDENTIFICATIONS
ANALYSES OF GENE FEATURES
TESTS OF RELIABILITY OF PUBLIC EUTHERIAN GENOMIC SEQUENCES
ANALYSES OF NUCLEOTIDE SEQUENCE REDUNDANCIES
DISCRIMINATIONS BETWEEN COMPLETE AND PUTATIVE CODING SEQUENCES
TESTS OF CONTIGUITY OF PUBLIC EUTHERIAN GENOMIC SEQUENCES
MULTIPLE PAIRWISE GENOMIC SEQUENCE ALIGNMENTS
ANALYSES OF CODING EXON NUMBERS AND THEIR RELATIVE ORIENTATIONS
CURATED EUTHERIAN THIRD PARTY DATA GENE DATA SETS
PHYLOGENETIC ANALYSIS
ALIGNMENTS OF PROTEIN AND NUCLEOTIDE SEQUENCES
PHYLOGENETIC TREE CALCULATIONS
PAIRWISE NUCLEOTIDE SEQUENCE IDENTITY CALCULATIONS
ANALYSES OF DIFFERENTIAL GENE EXPANSIONS
PROTEIN MOLECULAR EVOLUTION ANALYSIS
PROTEIN AMINO ACID SEQUENCE FEATURE ANALYSES
TESTS OF PROTEIN MOLECULAR EVOLUTION
RELATIVE SYNONYMOUS CODON USAGE STATISTIC CALCULATIONS
INVARIANT, FORWARD AND COMPENSATORY AMINO ACID SITE DESCRIPTIONS

Fig. 4 Eutherian comparative genomic analysis protocol flowchart

the tests of reliability of eutherian public genomic sequences used *FGF* potential coding sequences. Using BLASTN and primary Sanger DNA sequencing information deposited in NCBI Trace Archive [51, 56], the first test steps analysed nucleotide sequence coverages of each *FGF* potential coding sequence. If consensus trace sequence coverages were available for every nucleotide, the protocol described *FGF* potential coding sequences as *FGF* complete coding sequences. However, if consensus trace sequence coverages were not available for every nucleotide, the protocol described *FGF* potential coding sequences as *FGF* putative coding sequences (not used in analyses). The protocol then deposited *FGF* complete coding sequences in European Nucleotide Archive as curated third-party data gene information [57–60]. The protocol used guidelines of human gene nomenclature [61] and guidelines of mouse gene nomenclature [62] in updated eutherian *FGF* gene classification and nomenclature. Third, the protocol used new genomics tests of contiguity of eutherian public genomic sequences in eutherian *FGF* gene annotations. Using multiple pairwise genomic sequence alignments of eutherian genomic sequences harbouring *FGF* complete coding sequences, the tests of contiguity analysed numbers of coding exons in *FGF* genes and their relative orientation. The tests discriminated between *FGF* genes not including coding exon misassemblies in eutherian genomic sequence assemblies and *FGF* genes including coding exon misassemblies. The tests used mVISTA AVID option in multiple pairwise genomic sequence alignments, using default settings [63, 64]. The empirically determined cut-offs of detection of common genomic sequence regions in pairwise alignments with base sequences (*Homo sapiens*) were 95% nucleotide sequence identity along 100 bp (*Pan troglodytes*, *Gorilla gorilla*), 90% along 100 bp (*Pongo abelii*, *Nomascus leucogenys*), 85% along 100 bp (*Macaca mulatta*, *Papio hamadryas*), 80% along 100 bp (*Callithrix jacchus*), 75% along 100 bp (*Tarsius syrichta*, *Microcebus murinus*, *Otolemur garnettii*), 65% along 100 bp (Rodentia) or 70% along 100 bp in other pairwise alignments [44–46]. In preparatory steps of multiple pairwise genomic sequence alignments, the protocol did not include masking of transposable elements in genomic sequences harbouring *FGF* complete coding sequences.

Phylogenetic analysis

The protocol used protein and nucleotide sequence alignments, calculations of phylogenetic trees, calculations of pairwise nucleotide sequence identities and analysis of differential gene expansions in phylogenetic analysis of eutherian *FGF* gene data set. First, using BioEdit 7.0.5.3, the protocol translated *FGF* complete coding sequences, and aligned them at amino acid level using ClustalW implemented in BioEdit 7.0.5.3. After manual corrections of FGF protein primary structure alignments, the *FGF* nucleotide sequence alignments were prepared accordingly. Second, the MEGA 6.06 program was used in phylogenetic tree calculations, using minimum evolution method that was applicable in phylogenetic analysis of very close, close, typical, distant and very distant eutherian *FGF* homologues (default settings, except gaps/missing data treatment = pairwise deletion and maximum composite likelihood method) [65, 66]. Third, the protocol used BioEdit 7.0.5.3 in calculations of pairwise nucleotide sequence identities of *FGF* complete coding sequences that were used in statistical analyses. The Microsoft Office Excel common statistical functions were used in calculations of pairwise nucleotide sequence identity patterns of eutherian *FGF* gene data set. Using pairwise nucleotide sequence identities of *FGF* nucleotide sequence alignments including 267 *FGF* complete coding sequences, the protocol calculated average pairwise nucleotide sequence identities ($\bar{a}$) and their average absolute deviations ($\bar{a}_{\mathrm{ad}}$), and largest ($a_{\max}$) and smallest ($a_{\min}$) pairwise nucleotide sequence identities.

Protein molecular evolution analysis

The protocol used analysis of FGF protein amino acid sequence features and tests of protein molecular evolution integrating patterns of *FGF* nucleotide sequence similarities with FGF protein primary structures in protein molecular evolution analysis. The protocol used complete *FGF* nucleotide sequence alignments in tests of protein molecular evolution, including 267 *FGF* complete coding sequences and 58,533 codons. Among eutherian *FGF* complete coding sequences, the average number of codons was 219. Using MEGA 6.06, the relative synonymous codon usage statistics were calculated as ratios between observed and expected amino acid codon counts (R = Counts / Expected counts). The protocol then described 20 amino acid codons including $R \leq$ 0,7 as not preferable amino acid codons, viz.: TTA, TTG, CTT, CTA, ATA, GTT, GTA, TCA, TCG, CCG, ACG, GCG, TAT, CAT, CAA, GAT, TGT, CGT, CGA, GGT (Fig. 3b). Finally, the protocol described reference human FGF1A 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).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06958-4.

Additional file 1. Third-party data gene data set of eutherian fibroblast growth factor genes.

Additional file 2 Multiple pairwise genomic sequence alignments of eutherian fibroblast growth factor genes. The *FGF* coding exon sequence regions in base sequences (*Homo sapiens*) were displayed as indigo rectangles, and grey arrows indicated their relative orientation (top). The genomic sequence regions including sequence identity levels above empirical cut-offs of detection of common genomic sequence regions were shown accordingly in multiple pairwise alignments.

Additional file 3. Pairwise nucleotide sequence identity patterns of eutherian fibroblast growth factor genes.

Additional file 4. Protein amino acid sequence alignments of eutherian fibroblast growth factors. The amino acid positions were labelled using white letters on black background (100% sequence identity level), white letters on dark grey background (≥ 75% sequence identity level) or black letters on grey background (≥50% sequence identity level). The 19 invariant amino acid sites were shown using white letters on violet backgrounds and 3 forward amino acid sites were shown using white letters on red backgrounds in reference human FGF1A protein primary structure (top). The stop codons were indicated by &s.

Abbreviations

FGF: Fibroblast growth factor; *FGF1–8*: Eutherian fibroblast growth factor gene superclusters

Acknowledgements

MP would like to thank manuscript reviewers on their comments and suggestions.

Author's contributions

MP conceived and conducted experiments, and prepared manuscript. The author read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The original curated third-party data gene data set including 267 eutherian *FGF* complete coding sequences was deposited in European Nucleotide Archive under accessions: LR130242-LR130508 [47]. The present study was registered under NCBI BioProject entitled "Curated eutherian third-party data gene data sets" (NCBI BioProject accession: PRJNA453891; NCBI BioSample accessions: SAMN09005565-SAMN09005599) [48, 67]. The public eutherian reference genomic sequence data sets (Additional file 1) were available in NCBI GenBank [51, 52] and Ensembl [54].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

No competing interests were declared.

Received: 29 November 2019 Accepted: 29 July 2020
Published online: 05 August 2020

References

1. Goldfarb M. Fibroblast growth factor homologous factors: evolution, structure, and function. Cytokine Growth Factor Rev. 2005;16:215–20.
2. Belov AA, Mohammadi M. Molecular mechanisms of fibroblast growth factor signaling in physiology and pathology. Cold Spring Harb Perspect Biol. 2013;5:a015958.
3. Ornitz DM, Itoh N. The fibroblast growth factor signaling pathway. Wiley Interdiscip Rev Dev Biol. 2015;4:215–66.
4. Martin GR. The roles of FGFs in the early development of vertebrate limbs. Genes Dev. 1998;12:1571–86.
5. Hogan BL. Morphogenesis. Cell. 1999;96:225–33.
6. Liu JP, Laufer E, Jessell TM. Assigning the positional identity of spinal motor neurons: rostrocaudal patterning of Hox-c expression by FGFs, Gdf11, and retinoids. Neuron. 2001;32:997–1012.
7. Beenken A, Mohammadi M. The FGF family: biology, pathophysiology and therapy. Nat Rev Drug Discov. 2009;8:235–53.
8. Makarenkova HP, Hoffman MP, Beenken A, Eliseenkova AV, Meech R, Tsau C, et al. Differential interactions of FGFs with heparan sulfate control gradient formation and branching morphogenesis. Sci Signal. 2009;2:ra55.
9. Ornitz DM, Marie PJ. Fibroblast growth factor signaling in skeletal development and disease. Genes Dev. 2015;29:1463–86.
10. Brewer JR, Mazot P, Soriano P. Genetic insights into the mechanisms of Fgf signaling. Genes Dev. 2016;30:751–71.
11. Patel VN, Pineda DL, Hoffman MP. The function of heparan sulfate during branching morphogenesis. Matrix Biol. 2017;57–58:311–23.
12. Goetz R, Beenken A, Ibrahimi OA, Kalinina J, Olsen SK, Eliseenkova AV, et al. Molecular insights into the klotho-dependent, endocrine mode of action of fibroblast growth factor 19 subfamily members. Mol Cell Biol. 2007;27:3417–28.
13. Smallwood PM, Munoz-Sanjuan I, Tong P, Macke JP, Hendry SH, Gilbert DJ, et al. Fibroblast growth factor (FGF) homologous factors: new members of the FGF family implicated in nervous system development. Proc Natl Acad Sci U S A. 1996;93:9850–7.
14. Goldfarb M. Signaling by fibroblast growth factors: the inside story. Sci STKE. 2001;2001:pe37.
15. Schoorlemmer J, Goldfarb M. Fibroblast growth factor homologous factors are intracellular signaling proteins. Curr Biol. 2001;11:793–7.
16. Goldfarb M, Schoorlemmer J, Williams A, Diwakar S, Wang Q, Huang X, et al. Fibroblast growth factor homologous factors control neuronal excitability through modulation of voltage-gated sodium channels. Neuron. 2007;55: 449–63.
17. Coulier F, Pontarotti P, Roubin R, Hartung H, Goldfarb M, Birnbaum D. Of worms and men: an evolutionary perspective on the fibroblast growth factor (FGF) and FGF receptor families. J Mol Evol. 1997;44:43–56.
18. Venkataraman G, Raman R, Sasisekharan V, Sasisekharan R. Molecular characteristics of fibroblast growth factor-fibroblast growth factor receptor-heparin-like glycosaminoglycan complex. Proc Natl Acad Sci U S A. 1999;96: 3658–63.
19. Plotnikov AN, Hubbard SR, Schlessinger J, Mohammadi M. Crystal structures of two FGF-FGFR complexes reveal the determinants of ligand-receptor specificity. Cell. 2000;101:413–24.
20. Stauber DJ, DiGabriele AD, Hendrickson WA. Structural interactions of fibroblast growth factor receptor with its ligands. Proc Natl Acad Sci U S A. 2000;97:49–54.
21. Ornitz DM, Itoh N. Fibroblast growth factors. Genome Biol. 2001;2: Reviews3005.
22. Olsen SK, Garbi M, Zampieri N, Eliseenkova AV, Ornitz DM, Goldfarb M, et al. Fibroblast growth factor (FGF) homologous factors share structural but not functional homology with FGFs. J Biol Chem. 2003;278:34226–36.
23. Itoh N, Ornitz DM. Evolution of the Fgf and Fgfr gene families. Trends Genet. 2004;20:563–9.
24. Goetz R, Dover K, Laezza F, Shtraizent N, Huang X, Tchetchik D, et al. Crystal structure of a fibroblast growth factor homologous factor (FHF) defines a conserved surface on FHFs for binding and modulation of voltage-gated sodium channels. J Biol Chem. 2009;284:17883–96.
25. Itoh N, Ornitz DM. Fibroblast growth factors: from molecular evolution to roles in development, metabolism and disease. J Biochem. 2011;149:121–30.
26. Goetz R, Mohammadi M. Exploring mechanisms of FGF signalling through the lens of structural biology. Nat Rev Mol Cell Biol. 2013;14:166–80.
27. Bertrand S, Iwema T, Escriva H. FGF signaling emerged concomitantly with the origin of Eumetazoans. Mol Biol Evol. 2014;31:310–8.
28. Nam K, Lee KW, Chung O, Yim HS, Cha SS, Lee SW, et al. Analysis of the FGF gene family provides insights into aquatic adaptation in cetaceans. Sci Rep. 2017;7:40233.
29. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. Molecular phylogenetics and the origins of placental mammals. Nature. 2001;409:614–8.
30. Wilson DE, Reeder DM. Mammal species of the world: a taxonomic and geographic reference. 3rd ed. Baltimore: The Johns Hopkins University Press; 2005.
31. Blakesley RW, Hansen NF, Mullikin JC, Thomas PJ, McDowell JC, Maskeri B, et al. An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res. 2004;14:2235–44.
32. Margulies EH, Vinson JP, Miller W, Jaffe DB, Lindblad-Toh K, Chang JL, 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.

33. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478:476–82.
34. O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, et al. The placental mammal ancestor and the post-K-Pg radiation of placentals. Science. 2013;339:662–7.
35. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
36. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931–45.
37. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766–73.
38. Mouse Genome Sequencing Consortium. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 2009;7:e1000112.
39. Gajer P, Schatz M, Salzberg SL. Automated correction of genome sequence errors. Nucleic Acids Res. 2004;32:562–9.
40. Di Franco A, Poujol R, Baurain D, Philippe H. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol Biol. 2019;19:21.
41. Hubisz MJ, Lin MF, Kellis M, Siepel A. Error and error mitigation in low-coverage genome assemblies. PLoS One. 2011;6:e17034.
42. 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.
43. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 2014;10:e1003998.
44. Premzl M. Eutherian comparative genomic analysis protocol. Nat Protoc. 2018. https://doi.org/10.1038/protex.2018.028.
45. Premzl M. Comparative genomic analysis of eutherian connexin genes. Sci Rep. 2019;9:16938.
46. Premzl M. Eutherian third-party data gene collections. Gene Rep. 2019;16:100414.
47. European Nucleotide Archive. Accessions: LR130242-LR130508. https://www.ebi.ac.uk/ena/data/view/LR130242-LR130508..
48. NCBI BioProject. Accession: PRJNA453891. https://www.ncbi.nlm.nih.gov/bioproject/453891. Accessed 27 Jul 2020.
49. BioEdit. https://bioedit.software.informer.com/. Accessed 27 Jul 2020.
50. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
51. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, et al. Database resources of the National Center for biotechnology information. Nucleic Acids Res. 2020;48:D9–16.
52. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2020;48:D84–6.
53. NCBI BLAST Genomes. https://blast.ncbi.nlm.nih.gov/Blast.cgi. Accessed 27 Jul 2020.
54. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;48:D682–8.
55. Ensembl genome browser. https://www.ensembl.org. Accessed 27 Jul 2020.
56. NCBI Trace Archive. https://www.ncbi.nlm.nih.gov/Traces/trace.cgi. Accessed 27 Jul 2020.
57. Gibson R, Alako B, Amid C, Cerdeno-Tarraga A, Cleland I, Goodgame N, et al. Biocuration of functional annotation at the European nucleotide archive. Nucleic Acids Res. 2016;44:D58–66.
58. Karsch-Mizrachi I, Takagi T, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2018;46:D48–51.
59. Amid C, Alako BTF, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J, Fan J, et al. The European nucleotide archive in 2019. Nucleic Acids Res. 2020;48:D70–6.
60. European Nucleotide Archive. https://www.ebi.ac.uk/ena/about/tpa-policy. Accessed 27 Jul 2020.
61. Guidelines of human gene nomenclature. http://www.genenames.org/about/guidelines. Accessed 27 Jul 2020.
62. Guidelines of mouse gene nomenclature. http://www.informatics.jax.org/mgihome/nomen/gene.shtml. Accessed 27 Jul 2020.
63. Poliakov A, Foong J, Brudno M, Dubchak I. GenomeVISTA--an integrated software package for whole-genome alignment and visualization. Bioinformatics. 2014;30:2654–5.
64. mVISTA. http://genome.lbl.gov/vista/index.shtml. Accessed 27 Jul 2020.
65. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35:1547–9.
66. MEGA 6.06. http://www.megasoftware.net/. Accessed 27 Jul 2020.
67. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. 2012;40:D57–63.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.