

Long-read sequencing reveals novel transcript diversity in *Polypterus senegalus*

Jeong Woen Shin ^a, Bo-Mi Kim ^b, Seung Hwan Lee ^c, Jun Kim ^{d,e} and Jae-Sung Rhee ^{f,g,h}

^aDepartment of Bio-AI Convergence, Chungnam National University, Daejeon, Republic of Korea; ^bDivision of Life Sciences, Korea Polar Research Institute, Incheon, Republic of Korea; ^cDivision of Animal & Dairy Science, Chungnam National University, Daejeon, Republic of Korea; ^dDepartment of Convergent Bioscience and Informatics, College of Bioscience and Biotechnology, Chungnam National University, Daejeon, Republic of Korea; ^eGraduate School of Life Sciences, College of Bioscience and Biotechnology, Chungnam National University, Daejeon, Republic of Korea; ^fDepartment of Marine Science, College of Natural Sciences, Incheon National University, Incheon, Republic of Korea; ^gResearch Institute of Basic Sciences, Core Research Institute, Incheon National University, Incheon, Republic of Korea; ^hYellow Sea Research Institute, Incheon, Republic of Korea

ABSTRACT

Polypterus senegalus is a fish species characterized by primitive traits and unique physiological-anatomical features, representing a transitional stage toward terrestrial vertebrates. Diverging from other ray-finned fishes approximately 420 million years ago, it serves as a key model for vertebrate evolutionary studies. In this study, we identified 19,295 unique *P. senegalus* transcripts using Pacific Biosciences Iso-Seq, with most classified as novel exon combinations, intron retentions, or novel start and stop codons. This high-quality, full-length transcriptome revealed previously unreported isoforms and investigated their functional roles to enhance our understanding of the genomic complexity and evolutionary traits of the species. Notably, these isoforms were identified in genes associated with *Polypterus*'s distinctive features, including limb flexibility (*ACTN3B*, *MYBPC1*), lung function (*FAM13A*, *ERBB2*, *TCF7*), and circulation (*FBN1*, *MYH11B*, *MEF2CB*). These findings offer molecular insights into early vertebrate terrestrial adaptation and the functional integration of key tissues in extant species.

ARTICLE HISTORY

Received 16 January 2025
Accepted 30 January 2025

KEYWORDS

Polypterus senegalus; PacBio Iso-Seq; isoforms; splicing patterns; biological traits

Introduction

The family Polypteridae, within the class Actinopterygii (ray-finned fishes), comprises two genera: *Polypterus* and *Erpetoichthys* (Noack et al. 1996). *Polypterus* inhabits freshwater rivers and lakes in West Africa and exhibits unique physiological traits, including aerial breathing in low-oxygen conditions and terrestrial locomotion using its pectoral fins and body (Magid et al. 1970; Lutek et al. 2022). These adaptations are believed to have diverged from those of other ray-finned fishes approximately 420 million years ago during the Devonian period (López et al. 2013). Due to its primitive characteristics, which resemble those of tetrapods, *Polypterus* serves as a valuable model for studying phenotypic innovations and the molecular mechanisms underlying tetrapod evolution (López et al. 2013).

Moreover, molecular analyses of *Polypterus* facilitate the identification of genomic regions associated with its key traits and broader genome evolution trends

(Chiu et al. 2004). The chromosome-level *Polypterus* genome, published in February 2021, provides a foundation for genomic studies of this species. With a total size of approximately 3.7 Gb and a scaffold N50 of 189.7 Mb, the *Polypterus* genome demonstrates high chromosomal accuracy and completeness (Bi et al. 2021). It has been widely used to investigate the genomic characteristics and genetic mechanisms of *Polypterus*; however, its gene annotation lacks splicing-level information for this species (Inoue et al. 2003; Kato et al. 2024; Nagashima et al. 2024).

The limb flexibility of *Polypterus* provides valuable insights into the adaptations required for early terrestrial vertebrates to transition from aquatic to terrestrial environments, enhancing mobility in response to environmental changes (Wilhelm et al. 2015). This flexibility is attributed to simple joint structures, with the actinin alpha 3b (*ACTN3B*) and myosin-binding protein

CONTACT Jun Kim  junkim@cnu.ac.kr  Department of Convergent Bioscience and Informatics, College of Bioscience and Biotechnology, Chungnam National University, Daejeon, Republic of Korea, Graduate School of Life Sciences, College of Bioscience and Biotechnology, Chungnam National University, Daejeon, Republic of Korea; Jae-Sung Rhee  jsrhee@inu.ac.kr  Department of Marine Science, College of Natural Sciences, Incheon National University, Incheon 22012, Republic of Korea Research Institute of Basic Sciences, Core Research Institute, Incheon National University, Incheon 22012, Republic of Korea Yellow Sea Research Institute, Incheon 22012, Republic of Korea

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

C1 (*MYBPC1*) genes implicated in the developmental differences underlying this trait.

Alpha-actinins (ACTN), which stabilize sarcomeric Z-lines by cross-linking actin microfilaments, play a critical role in muscle function. According to Christopher et al., five actin-related genes, including two *ACTN3* paralogs, exhibit muscle-specific expression (Holterhoff et al. 2009). Notably, approximately 16% of humans and birds lack *ACTN3*, highlighting significant physiological diversity in vertebrate muscle function (Holterhoff et al. 2009). Additionally, *MYBPC1* encodes sMyBP-C, a myosin-binding protein that regulates contraction and flexibility in slow skeletal muscle by stabilizing and supporting thick filaments (Geist and Kontogianni-Konstantopoulos 2016).

The lung of *Polypterus* serves as a key model for studying the evolutionary transition to terrestrial vertebrates (Icardo 2018). In this context, family with sequence similarity 13 member A (*FAM13A*), erb-b2 receptor tyrosine kinase 2 (*ERBB2*), and transcription factor 7 (*TCF7*) are critical factors. *FAM13A* and *ERBB2* contribute to tissue development and lung epithelial cell repair, while *TCF7* mediates T-cell differentiation and plays a role in lung infection (Sasaki et al. 2006; Tomizawa et al. 2011; Zhu et al. 2015; Chuang et al. 2017; Eisenhut et al. 2017; Chen et al. 2023).

The circulatory system of *Polypterus* regulates oxygen absorption, transport, and utilization, playing a vital role in sustaining life (Capillo et al. 2021). fibrillin 1 (*FBN1*) and myosin, heavy chain 11b, smooth muscle (*MYH11B*) are essential for smooth muscle contraction and vascular wall stability. Mutations in these genes can impair contraction, leading to vascular wall degeneration and circulatory dysfunction (Kuang et al. 2012; Hemani 2023). Myocyte enhancer factor 2cb (*MEF2CB*), expressed in the late ventricular region and regulated by FGF signaling, plays a key role in heart development, performing a function similar to that of the second heart field in mice during zebrafish cardiac development (Lazic and Scott 2011; Hinitz et al. 2012). Despite extensive research on these genes, their transcriptomic diversity at the splicing level remains unexplored in *Polypterus senegalus*.

A major challenge in transcriptomics at the splicing level is the comprehensive identification of isoforms across species. Earlier approaches relied on short-read sequencing, which often failed to capture critical splicing variants (Park and Chung 2019; Hu et al. 2021; Cao et al. 2022; Kim H, et al. 2022; Kim HS, et al. 2022; Yoo et al. 2022; Zhou et al. 2022; Abdellaoui et al. 2023; Lee et al. 2023; Lim et al. 2023). Recent advancements in long-read sequencing technologies have overcome this limitation by providing full-length transcript information (Byrne et al. 1786; Cho et al. 2014; Logsdon et al. 2020; Marx 2023). For instance, Pacific Biosciences (PacBio)

developed the Isoform Sequencing (Iso-Seq) method, which generates high-quality, full-length transcriptome data (Gonzalez-Garay 2016). This technology has facilitated the discovery of novel alternative splicing events and isoforms associated with biological, evolutionary, and pathological traits, significantly advancing our understanding of transcriptomic diversity (Gonzalez-Garay 2016; Feng et al. 2019).

This study aimed to identify previously unreported isoforms in *Polypterus senegalus* using PacBio Iso-Seq data, thereby elucidating the genomic complexity and evolutionary traits of the species. Additionally, by structurally categorizing these novel isoforms, we aimed to identify those enriched in *P. senegalus* (Figure 1). The splicing-level transcriptome data generated in this study serve as a valuable foundation for future research in vertebrate evolution and molecular biology.

Materials and methods

Sample preparation and RNA sequencing

The *P. senegalus* specimen used in this study was collected by Incheon National University and processed by DNALink, Inc. (<https://en.dnalink.com>). RNA was extracted from 12 tissues (brain, eye, gill, ovary, kidney, liver, lung, muscle, swim bladder, gut, spleen, and skin) from the same individual. The extracted RNA samples were pooled before sequencing. Complementary DNA (cDNA) was synthesized using the SMRTer PCR cDNA Synthesis Kit and DNA Template Prep Kit 1.0. Sequencing was conducted on two cells of the PacBio Sequel II platform, and the resulting data were processed using Sequel SMRT Link 8.0 software, yielding high-quality PacBio Iso-Seq data. The read lengths of the PacBio Iso-Seq raw data were assessed using Bioawk (version 20110810): `zcat ${IsoSeq_FASTQ} | bioawk -c fastx 'OFS = "," {print length($seq)}'`. The length distribution was visualized using R (version 4.2.2) with the ggplot2 package (Wickham 2106).

Novel transcript identification

The FASTA sequence file and gene annotation GTF file for the *P. senegalus* genome were downloaded from the NCBI database (Accession number: GCFXX). This genome assembly has a total size of 3.4 Gb and contains 29,359 genes. PacBio Iso-Seq reads were aligned to the *P. senegalus* genome using flair-align (version 2.0.0): `flair align -g ${Bi_FASTA} -r ${IsoSeq_FASTQ} --output ${PREFIX}` (Tang et al. 2020). Errors in splice sites from the alignment results were corrected based on the gene annotation of the *Polypterus* genome using flair-correct (version 2.0.0): `flair correct -q ${PREFIX}.bed -f`

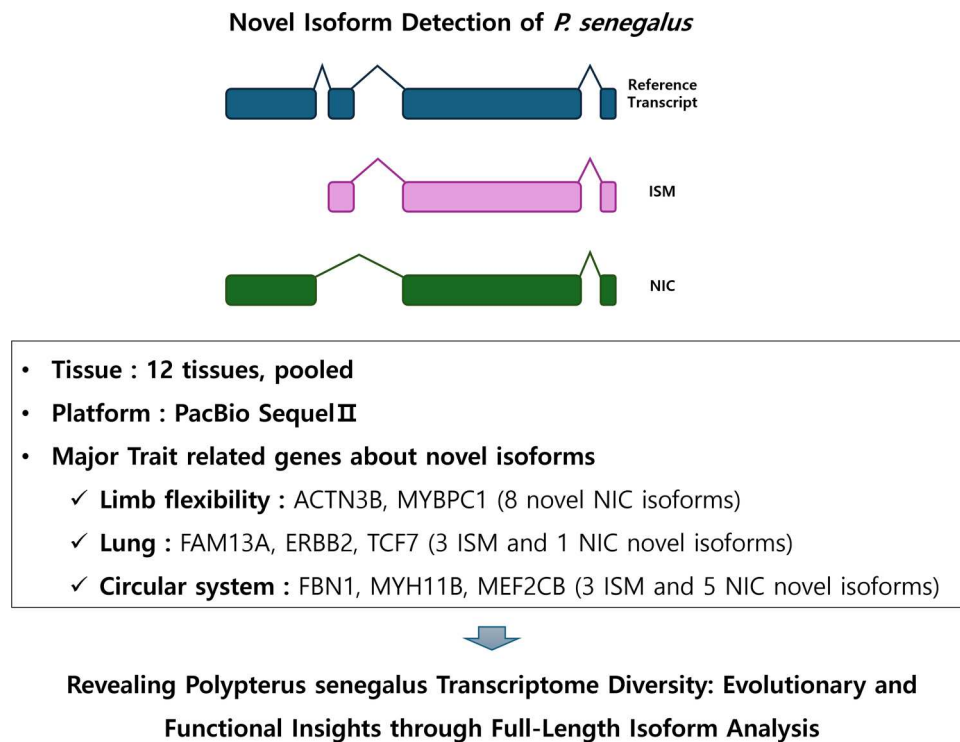


Figure 1. Graphical abstract. We aim to identify novel transcripts associated with major trait-related genes, including those involved in limb flexibility, lung function, and the circulatory system, in *Polypterus senegalus* using PacBio Iso-Seq data. Our ultimate goal is to elucidate the diversity of the *P. senegalus* transcriptome at the splicing level through full-length isoform analysis.

`${Bi_GTF} -g ${Bi_FASTA} --output ${PREFIX}` (Tang et al. 2020). Redundant transcripts corresponding to the same isoform were collapsed to generate high-confidence isoforms using flair-collapse (version 2.0.0): `flair collapse -g ${Bi_FASTA} --gtf ${Bi_GTF} -q ${PREFIX}_all_corrected.bed -r ${IsoSeq_FASTQ} --stringent --check_splice --generate_map --annotation_reliant generate --output ${PREFIX}` (Tang et al. 2020). Only full-length reads were included in this analysis. Full-length reads were defined as those covering at least 80% of the isoform and containing a minimum of 25 bp in both the first and last exons. To ensure splice site accuracy, reads were required to cover at least 4 bp of the 6 bp surrounding splice sites and have insertion mutations no greater than 3 bp.

Isoform annotation and classification

The corrected and collapsed isoforms were annotated and classified by comparing them with the gene annotation of the *Polypterus* genome using SQANTI3 (version 5.2.2): `sqanti3_qc.py ${FLAIR_ISOFORM_GTF} ${Bi_GTF} ${Bi_FASTA} --force_id_ignore -o ${PREFIX} -d ${PREFIX} --report both --isoAnnotLite` (Pardo-Palacios et al. 2024). The `--isoAnnotLite` option was applied to facilitate a detailed evaluation of isoform structural features, splicing accuracy, and novelty.

Isoforms were categorized into eight structural classes, including full splice match (FSM), incomplete splice match (ISM), novel in catalog (NIC), and novel not in catalog (NNC). FSM isoforms were further subdivided into four subtypes based on positional differences relative to the annotated transcription termination sites. ISM isoforms were classified into five subtypes depending on fragmentation, retention, or changes in exon counts. Novel isoforms were divided into five subtypes based on whether they contained known splice junctions or entirely novel splicing sites.

Protein structure prediction

Protein structures were predicted for novel isoforms of *P. senegalus* genes associated with its distinct phenotypes using AlphaFold2 and coldblood (version 1.5.5) with the following parameters: `template_mode: none, msa_mode: mmseqs2_uniref_env, pair_mode: unpaired_unpaired, model_type: auto, num_recycles: 3, recycle_early_stop_tolerance: auto, relax_max_iterations: 200, pairing_strategy: greedy, max_msa: auto, and num_seeds: 1` (Kim et al. 2025). The per-residue local distance difference test (pLDDT) was used as a confidence score for protein structure prediction, with pLDDT values visualized using different colors in the 3D models.

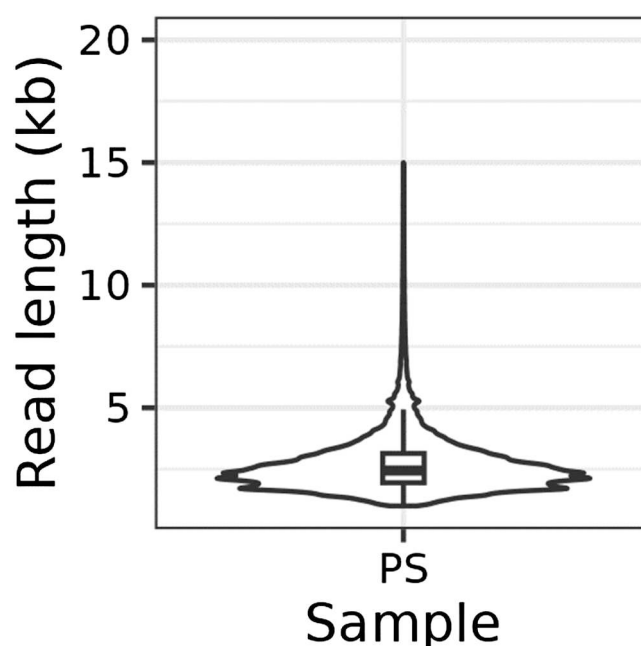


Figure 2. The length distribution of *P. senegalus* Iso-Seq Data. The Iso-Seq data generated for *P. senegalus* has an average read length of 2.74 kb, indicating its high quality and suitability for full-length transcript analysis.

Results

Iso-Seq revealed approximately 4500 novel isoforms of P. senegalus

We analyzed the transcriptome diversity of *P. senegalus* using RNA extracted from the following 12 tissues: brain, eye, gill, ovary, kidney, liver, lung, muscle, swim bladder, gut, spleen, and skin. Sequencing of the pooled RNA sample generated 2.85 Gb of data from a total of 1,037,557 reads (Figure 2). The average read length was 2.74 kb, ensuring high-quality data with sufficient read length and coverage.

Iso-Seq reads were mapped to the *P. senegalus* genome to correct splice sites and collapse redundant transcripts representing the same isoform. This process identified 19,295 unique transcripts, including 4547 novel isoforms. These transcripts corresponded to 13,431 genes, of which 12,106 were previously annotated, while 1325 represented newly discovered genes.

The 19,295 unique transcripts were categorized into the following classes: FSM (76.57%; known splice sites), ISM (6.40%; known splice sites but fragmented), NIC (8.30%; known splice sites but different combinations), Intergenic (3.09%), Antisense (2.81%), Genic Genomic (1.64%), Genic Intron (1.06%), and Fusion (0.13%) (Table 1). These findings support the accuracy of the previous gene annotation of the *P. senegalus* genome and validate the high quality of the Iso-Seq data, as most transcripts correspond to previously known isoforms categorized as FSM.

Among the 4547 novel isoforms, NIC (35.23%) and ISM (27.16%) made up the largest proportions, with 27 additional FSM transcripts identified (Table 1). This indicates that the majority of novel isoforms consisted of either novel exon combinations with known splice sites or truncated transcripts with known splice sites.

Splicing junction sites confirmed by Iso-Seq data

We further refined these categories to identify more specific splicing variants in *P. senegalus*. The FSM isoforms were predominantly classified as 'Reference match' (11,326, 76.66%), followed by 'Alternative 3' end' (15.72%), 'Alternative 3' 5' end' (5.24%), and 'Alternative 5' end' (1.04%) (Table 2). This indicates that the majority of FSM isoforms match the reference gene annotation, while the remaining isoforms primarily involve codon changes at the transcript ends. These

Table 1. Classification of transcripts in *P. senegalus*.

Transcript category	Isoforms count	Novel Isoforms count
FSM	14,775	27
NIC	1602	1602
ISM	1235	1235
Intergenic	596	596
Antisense	542	542
Genic Genomic	316	316
Genic Intron	204	204
Fusion	25	25
NNC	0	0

FSM: Full Slice Match; NIC: Novel In Catalog; ISM: Incomplete Splice Match; NNC: Novel Not in the Catalog.

Table 2. Distribution of transcript subcategories in *P. senegalus*.

FSM		ISM		NIC	
Subcategory	Count	Subcategory	Count	Subcategory	Count
Reference Match	11,326	3' fragment	388	Comb. of annot. junctions	596
Alternative 3' end	2322	Internal fragment	34	Comb. of annot. splice sites	617
Alternative 3' 5' end	774	5' fragment	632	Intron retention	350
Alternative 5' end	154	Intron retention	42	Mono-exon by intron ret.	39
Mono-exon	199	Mono-exon	139		

Comb. of annot. junctions: Combination of known junctions; Comb. of annot. splice sites: Combination of known splice sites; Mono-exon by intron ret.: Mono-exon by intron retention.

differences in codon composition could indicate true alternative isoforms or misannotations of the start or stop codons in the reference gene annotation.

The ISM isoforms were categorized as '5' fragment' (51.17%), '3' fragment' (31.42%), 'Mono-exon' (11.26%), 'Intron retention' (3.40%), and 'Internal fragment' (2.75%) (Table 2). This suggests that most ISM isoforms arise from the removal or misannotation of exons at either end of the transcripts.

The NIC isoforms were categorized as 'Combination of known splice sites' (38.51%), 'Combination of known junctions' (37.20%), 'Intron retention' (21.85%), and 'Mono-exon by intron retention' (2.43%) (Table 2). These findings indicate that most splicing sites annotated in the reference genome were validated by our Iso-Seq reads. Notably, no significant novel splicing junction sites were identified in the Iso-Seq data.

The subcategories of isoforms showed similar length distributions

The average length of all 19,295 unique transcripts was 2533 bp, with a range of 83–8790 bp. Transcripts in the 2–3 kb range were the most abundant, comprising 38.4% (7416 out of 19,295), while those in the 8–9 kb range were exclusively classified as FSM or ISM (Figure 3(a)). Novel transcripts exhibited a similar distribution, with 2099 transcripts in the 2–3 kb range. Notably, one novel transcript was identified in the 8–9 kb range (Figure 3(b)). NIC transcripts had the longest average length (2811.98 bp), while Genic Intron transcripts had the shortest (2256.47 bp) (Figure 3(c)).

The subcategories within the FSM, ISM, and NIC groups showed similar length distributions. However, FSM isoforms with 'Alternative 3' 5' end' (2579.99 bp), ISM isoforms with '3' fragment' (2643.20 bp), and NIC isoforms with 'Mono-exon by intron retention' (3029.03 bp) were slightly longer than those in other subcategories (Figure 3(d,e)).

Most of the transcripts contained 10 or more exons. The number of exons in transcripts varied, but NIC transcripts across all main categories had the highest average number of exons (12.03 exons). Conversely, antisense, genic genomic, intergenic, and genic intron transcripts contained only one exon on average (Figure 3(f)). At the subcategory level, FSM isoforms with 'Alternative 3' end' transcripts had the highest average number of exons (10.8 exons), ISM isoforms with '5' fragment' transcripts had 12.05 exons, and NIC isoforms with 'Combination of known junctions' transcripts had the most, with an average of 14.98 exons (Figure 3(g,h)).

FAM13A and ERBB2 exhibited novel protein sequences

Among the 4547 novel isoforms derived from 3783 genes, we focused on a small set of genes associated with *Polypterus*'s unique biological traits, including limb flexibility, lung function, and the circulatory system. *ACTN3B* and *MYBPC1*, which are associated with limb flexibility, yielded two and six novel isoforms, respectively. *FAM13A*, *ERBB2*, and *TCF7*, known to play roles in lung development, produced two, one, and one novel isoforms, respectively. *FBN1*, *MYH11B*, and *MEF2CB*, important genes for circulatory system development, yielded one, one, and six novel isoforms, respectively (Table 3).

Among these 20 novel isoforms, two novel isoforms of *FAM13A* (novel1199 and novel1200) and one novel isoform of *ERBB2* (novel4330), were confirmed as coding transcripts, and their amino acid sequences were obtained (Table 3). The two novel isoforms of *FAM13A* encoded identical protein sequences but differed in the lengths of their untranslated regions. These protein sequences were shorter than those of the canonical isoforms (XM_039750165.1 and XM_039750166.1) (Figure 4(a)). Similarly, the novel isoform of *ERBB2* had a shorter amino acid sequence compared to the canonical isoform (XM_039740827.1) (Figure 4(b)). The pLDDT scores calculated for these novel isoforms indicated low confidence in the predicted structures, with most regions scoring below 50.

Discussion

This study used long-read RNA sequencing data to conduct a comprehensive analysis of the *P. senegalus* transcriptome at the splicing level. The Iso-Seq data, with an average read length of 2.74 kb, provided high-quality transcriptomic information, offering a robust

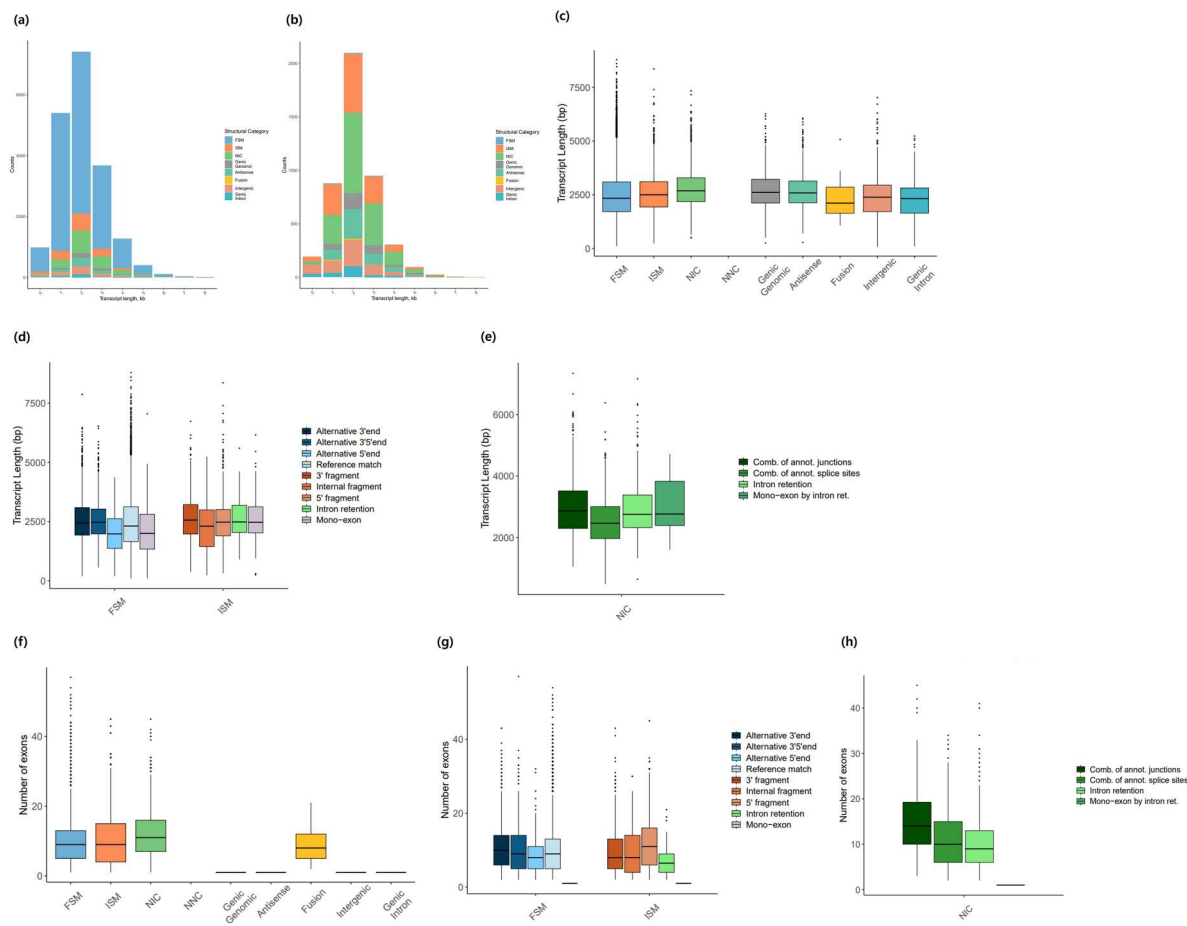


Figure 3. Read length and exon count distributions for total and novel transcripts in *P. senegalus*. (a) Distribution of structural categories by transcript length for the entire transcript set. The x-axis represents transcript length (kb), the y-axis represents the number of transcripts, and each color denotes a different category. (b) Distribution of structural categories by transcript length for novel transcripts. (c) Transcript length distribution across structural categories, with the x-axis representing structural categories and the y-axis representing transcript length (bp). Each category is color-coded. (d) Transcript length distribution for FSM and ISM subcategories. (e) Transcript length distribution for NIC subcategories. (f) Exon count distribution across structural classifications. (g) Exon count distribution for FSM and ISM subcategories. (h) Exon count distribution for NIC subcategories.

Table 3. Novel transcripts of major trait-associated genes in *P. senegalus*.

Associated trait	Gene symbol	Isoform	Chromosome	Length	Exon count	Structural category	Associated transcript	Reference length	Reference exon count
Limb flexibility	<i>ACTN3B</i>	novel3011	NC_053164.1	2752	18	NIC	novel	3001	21
		novel3012	NC_053164.1	2922	20	NIC	novel	3001	21
	<i>MYBPC1</i>	novel2186	NC_053161.1	6083	33	NIC	novel	3856	36
		novel2187	NC_053161.1	6011	32	NIC	novel	3856	36
		novel2188	NC_053161.1	5972	31	NIC	novel	3856	36
		novel2189	NC_053161.1	5966	31	NIC	novel	3856	36
		novel2190	NC_053161.1	6044	32	NIC	novel	3856	36
		novel2191	NC_053161.1	5927	30	NIC	novel	3856	36
Lung	<i>FAM13A</i>	novel1199	NC_053157.1	3313	10	ISM	XM_039750165.1	5106	25
		novel1200	NC_053157.1	3400	11	ISM	XM_039750166.1	5032	25
	<i>ERBB2</i>	novel4330	NC_053170.1	2429	16	ISM	XM_039740827.1	5478	26
	<i>TCF7</i>	novel3610	NC_053166.1	2713	11	NIC	novel	2786	12
Circular system	<i>FBN1</i>	novel3437	NC_053165.1	2665	17	ISM	XM_039771381.1	9066	59
	<i>MYH11B</i>	novel3726	NC_053166.1	6151	41	ISM	XM_039775315.1	6475	42
	<i>MEF2CB</i>	novel1861_NC	NC_053160.1	2168	1	ISM	XM_039759321.1	5301	11
		novel1979	NC_053160.1	2567	9	NIC	novel	5551	11
		novel1980	NC_053160.1	2615	10	NIC	novel	5551	11
		novel1981	NC_053160.1	2396	10	NIC	novel	5387	11
		novel1982	NC_053160.1	2348	10	NIC	novel	5387	11
		novel1983	NC_053160.1	2390	10	NIC	novel	5301	11

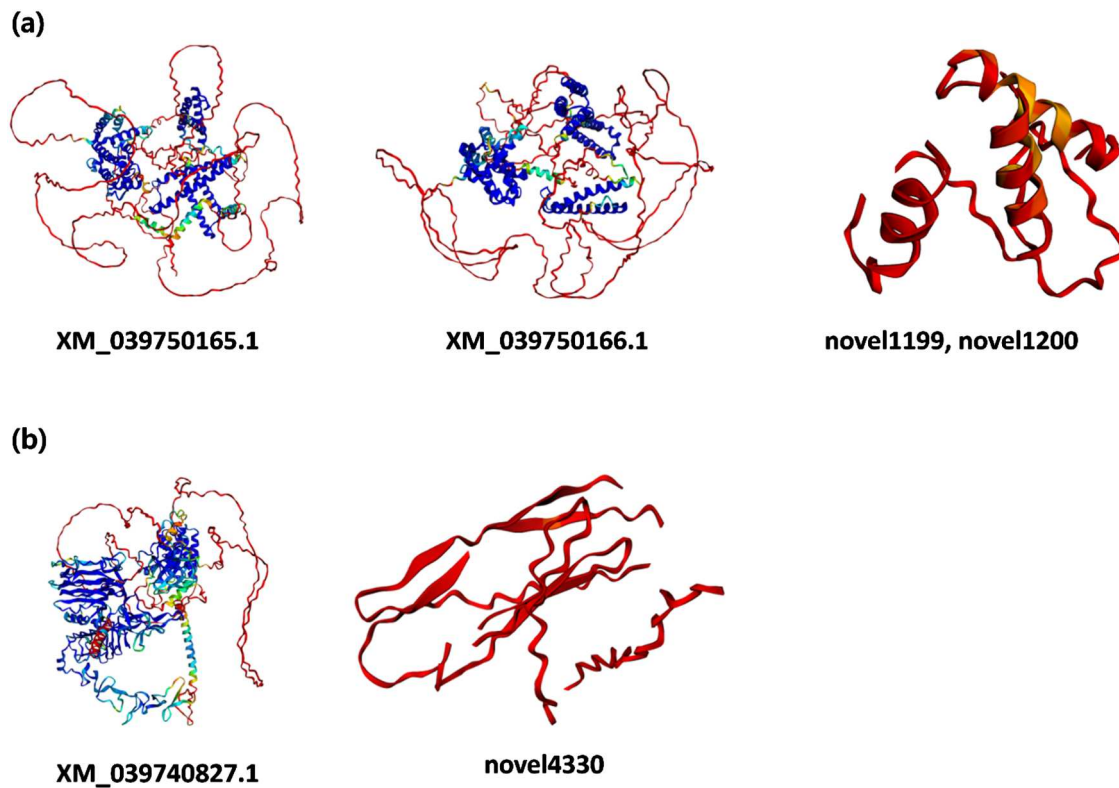


Figure 4. Predicted protein structures of the protein-coding transcripts associated with key traits in *P. senegalus*. (a) The novel isoforms novel1199 and novel1200 of *FAM13A* encode the same protein but with shorter sequences compared to the known transcripts XM_039750165.1 and XM_039750166.1. (b) The novel isoform novel4330 of *ERBB2* is significantly shorter than the known isoform XM_039740827.1.

foundation for exploring the complex gene structures and transcript diversity of *Polypterus*.

Iso-Seq revealed previously unreported isoforms associated with key phenotypes of *Polypterus*. Transcript structural classification (FSM, ISM, NIC, etc.) showed that FSM accounted for the largest proportion (approximately 76%), followed by ISM (6.40%) and NIC (8.30%). Among novel transcripts, the majority were distributed within NIC (35.23%) and ISM (27.16%), indicating that the *Polypterus* transcriptome exhibits newly identified, highly heterogeneous structures resulting from diverse splicing events. The discovery of novel NIC and ISM isoforms provides critical insights into previously uncharacterized splicing patterns and transcript regulation mechanisms. Additionally, the novel isoforms identified are associated with genes involved in the unique biological traits of *Polypterus*, including limb flexibility, the circulatory system, and lung function.

ACTN3B and *MYBPC1* are associated with limb flexibility (Holterhoff et al. 2009; Geist and Kontogianni-Konstantopoulos 2016). Two novel transcripts of *ACTN3B* were identified and classified as NIC isoforms derived from combinations of known splice sites. Additionally, six novel transcripts of *MYBPC1* were discovered, all

classified as NIC isoforms. These findings provide valuable insights into limb flexibility in *Polypterus* and the diversity of muscle function in early terrestrial adaptation. The novel NIC isoforms in *ACTN3B* and *MYBPC1* may play critical roles in regulating muscle protein structure and function, offering deeper insights into how flexibility and locomotor ability are modulated in *Polypterus* and other species. However, further validation using molecular genetics approaches is needed to confirm these findings.

FAM13A, *ERBB2*, and *TCF7* are associated with lung function and various pulmonary diseases in mammals (Sasaki et al. 2006; Tomizawa et al. 2011; Zhu et al. 2015; Chuang et al. 2017; Eisenhut et al. 2017; Chen et al. 2023). In this study, two ISM isoforms of *FAM13A* and one of *ERBB2* were identified, all classified as coding transcripts. These isoforms encoded shorter protein sequences than their known counterparts. Additionally, protein structure predictions revealed structural differences in these novel transcripts compared to known isoforms, highlighting the complexity of the *P. senegalus* transcriptome.

For *TCF7*, a novel NIC isoform was discovered. Further investigation is required to determine whether this

isoform contributes to the evolutionary adaptation of respiratory structures and pulmonary function, as well as its potential relevance to disease mechanisms in modern mammals.

For *FBN1*, *MYH11B*, and *MEF2CB*, which are associated with the circulatory system, one novel ISM isoform was identified for each gene (Lazic and Scott 2011; Kuang et al. 2012; Hemani 2023). These genes play critical roles in smooth muscle contraction, vascular stability, and overall heart and circulatory system function (Lazic and Scott 2011; Hinitz et al. 2012; Kuang et al. 2012; Hemani 2023). Further studies are necessary to investigate their gene expression patterns at the splicing level in *Polypterus*.

This study identified novel transcripts linked to the unique biological traits of *P. senegalus* using long-read RNA sequencing. These findings significantly enhance our understanding of vertebrate evolution and transcriptomic regulation, providing a comprehensive perspective on the molecular diversity and evolutionary adaptations of the species.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Korea Environmental Industry and Technology Institute through the Aquatic Ecosystem Conservation Research Program (2022003050001), funded by the Ministry of Science and ICT, South Korea and by the National Research Foundation of Korea (NRF) grant funded by the MSIT [RS-2024-00415888].

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00155857, Artificial Intelligence Convergence Innovation Human Resources Development (Chungnam National University)) and Seung Hwan Lee was supported by Chungnam National University as research leave (2023–2024).

Data availability statement

All sequencing reads produced in this study have been submitted to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject>) under accession number PRJNA1217591.

ORCID

Jeong Woen Shin  <http://orcid.org/0000-0001-7131-4080>

Bo-Mi Kim  <http://orcid.org/0000-0002-8787-8718>

Seung Hwan Lee  <http://orcid.org/0000-0003-1508-4887>

Jun Kim  <http://orcid.org/0000-0001-9753-4778>

Jae-Sung Rhee  <http://orcid.org/0000-0003-3313-8850>

References

- Abdellaoui N, Kim SY, Kim MS. 2023. Effect of TRAF6-knockout on gene expression and lncrna expression in epithelioma papulosum cyprini (EPC) cells. *Animal Cells Syst* (Seoul). 27(1):197–207. doi:10.1080/19768354.2023.2263070.
- Bi X, Wang K, Yang L, Pan H, Jiang H, Wei Q, Fang M, Yu H, Zhu C, Cai Y, et al. 2021. Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes. *Cell*. 184(5):1377–1391. doi:10.1016/j.cell.2021.01.046.
- Byrne A, Cole C, Volden R, Vollmers C. 1786. Realizing the potential of full-length transcriptome sequencing. *Philos Trans R Soc, B*. 374(1786):20190097. doi:10.1098/rstb.2019.0097.
- Cao D, Wang S, Zhang D, Zhang Y, Liu Y, Zhou H, et al. 2022. KRAB family is involved in network shifts in response to osmotic stress in camels. *Animal Cells Syst* (Seoul). 26(6):348–357. doi:10.1080/19768354.2022.2143894.
- Capillo G, Zacccone G, Cupello C, Fernandes JMO, Viswanath K, Kuciel M, Zuwala K, Guerrero MC, Aragona M, Icardo JM, Lauriano ER. 2021. Expression of acetylcholine, its contribution to regulation of immune function and O2 sensing and phylogenetic interpretations of the African butterfly fish *Pantodon buchholzi* (Osteoglossiformes, Pantodontidae). *Fish Shellfish Immunol*. 111:189–200. doi:10.1016/j.fsi.2021.02.006.
- Chen Q, Vasse GF, Nwozor KO, Bekker NJ, van den Berge M, Brandsma C-A, de Vries M, Heijink IH. 2023. FAM13A regulates cellular senescence marker p21 and mitochondrial reactive oxygen species production in airway epithelial cells. *Am J Physiol-Lung Cell Mol Physiol*. 325(4):L460–L466. doi:10.1152/ajplung.00141.2023.
- Chiu C-H, Dewar K, Wagner GP, Takahashi K, Ruddle F, Ledje C, Bartsch P, Scemama J-L, Stellwag E, Fried C, et al. 2004. Bichir *hoxa* cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res*. 14(1):11–17. doi:10.1101/gr.1712904.
- Cho H, Davis J, Li X, Smith KS, Battle A, Montgomery SB. 2014. High-resolution transcriptome analysis with long-read RNA sequencing. *PLoS One*. 9(9):e108095. doi:10.1371/journal.pone.0108095.
- Chuang JC, Stehr H, Liang Y, Das M, Diehn M, Wakelee HA, Neal JW, et al. 2017. ERBB2-Mutated metastatic non-small cell lung cancer: response and resistance to targeted therapies. *J Thorac Oncol*. 12(5):833–842. doi:10.1016/j.jtho.2017.01.023.
- Eisenhut F, Heim L, Trump S, Mittler S, Söpel N, Andreev K, Ferrazzi F, Ekici AB, Rieker R, Springel R, et al. 2017. FAM13A is associated with non-small cell lung cancer (NSCLC) progression and controls tumor cell proliferation and survival. *Oncolimmunology*. 6(1):e1256526. doi:10.1080/2162402X.2016.1256526.
- Feng S, Xu M, Liu F, Cui C, Zhou B. 2019. Reconstruction of the full-length transcriptome atlas using pacbio Iso-Seq provides insight into the alternative splicing in *Gossypium australe*. *BMC Plant Biol*. 19(1):365. doi:10.1186/s12870-019-1968-7.

- Geist J, Kontogianni-Konstantopoulos A. 2016. MYBPC1, an emerging myopathic gene: what we know and what we need to learn. *Front Physiol.* 7:410. doi:10.3389/fphys.2016.00410.
- Gonzalez-Garay ML. 2016. Introduction to isoform sequencing using pacific biosciences technology (Iso-Seq). In: Wu J, editor. *Transcriptomics and gene regulation. Translational bioinformatics, Vol 9.* , Dordrecht: Springer. doi:10.1007/978-94-017-7450-5_6.
- Hemani D. 2023. The potential protective role of FBN1 deficiency induced miRNA-146b-5p in aneurysm progression in Marfan's Syndrome (Order No. 30250058). ProQuest Dissertations & Theses Global (2833905277). <https://www.proquest.com/dissertations-theses/potential-protective-role-fbn1-deficiencyinduced/docview/2833905277/se-2>
- Hinitis Y, Pan L, Walker C, Dowd J, Moens CB, Hughes SM, et al. 2012. Zebrafish Mef2ca and Mef2cb are essential for both first and second heart field cardiomyocyte differentiation. *Dev Biol.* 369(2):199–210. doi:10.1016/j.ydbio.2012.06.019.
- Holterhoff CK, Saunders RH, Brito EE, Wagner DS. 2009. Sequence and expression of the zebrafish alpha-actinin gene family reveals conservation and diversification among vertebrates. *Dev Dyn.* 238(11):2936–2947. doi:10.1002/dvdy.22123.
- Hu T, Chitnis N, Monos D, Dinh A. 2021. Next-generation sequencing technologies: an overview. *Hum Immunol.* 82(11):801–811. doi:10.1016/j.humimm.2021.02.012.
- Icardo JM. 2018. Lungs and gas bladders: morphological insights. *Acta Histochem.* 120(7):605–612. doi:10.1016/j.acthis.2018.08.006.
- Inoue JG, Miya M, Tsukamoto K, Nishida M. 2003. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the “ancient fish”. *Mol Phylogenet Evol.* 26(1):110–120. doi:10.1016/S1055-7903(02)00331-7.
- Kato A, Pipil S, Ota C, Kusakabe M, Watanabe T, Nagashima A, Chen A-P, Islam Z, Hayashi N, Wong MK-S, et al. 2024. Convergent gene losses and pseudogenizations in multiple lineages of stomachless fishes. *Commun Biol.* 7(1):408. doi:10.1038/s42003-024-06103-x.
- Kim G, et al. 2025. Easy and accurate protein structure prediction using colabfold. *Nat Protoc* 20: 620–642. doi:10.1038/s41596-024-01060-5.
- Kim H, Choo H, Cha J, Jang M, Jeong T, Choi B-H, Lim Y, Chai H-H, Lee J, Lim D, et al. 2022. Blood transcriptome comparison between sexes and their function in 4-week Rhode Island red chickens. *Animal Cells Syst (Seoul).* 26(6):358–368. doi:10.1080/19768354.2022.2146187.
- Kim HS, Kim J, Kim J, Choi YH. 2022. Characterization of differential gene expression of broiler chicken to thermal stress in discrete developmental stages. *Animal Cells Syst (Seoul).* 26(2):62–69. doi:10.1080/19768354.2022.2059566.
- Kuang S, Kwartler CS, Byanova KL, Pham J, Gong L, Prakash SK, Huang J, Kamm KE, Stull JT, Sweeney HL, Milewicz DM, et al. 2012. Rare, nonsynonymous variant in the smooth muscle-specific isoform of myosin heavy chain, MYH11, R247C, alters force generation in the aorta and phenotype of smooth muscle cells. *Circ Res.* 110(11):1411–1422. doi:10.1161/CIRCRESAHA.111.261743.
- Lazic S, Scott IC. 2011. Mef2cb regulates late myocardial cell addition from a second heart field-like population of progenitors in zebrafish. *Dev Biol.* 354(1):123–133. doi:10.1016/j.ydbio.2011.03.028.
- Lee JH, Jeong S, Joh CS-Y, Nam HJ, Choi HS, Sserwadda H, Oh JW, Park C-G, Jin S-P, Kim HJ. 2023. Single-cell RNA sequencing identifies distinct transcriptomic signatures between PMA/ionomycin-and αCD3/αCD28-activated primary human T cells. *Genomics Inform.* 21(2):e18. doi:10.5808/gi.23009.
- Lim B, Jang M-J, Oh S-M, No JG, Lee J, Kim SE, Ock SA, Yun IJ, Kim J, Chee HK, et al. 2023. Comparative transcriptome analysis between long- and short-term survival after pig-to-monkey cardiac xenotransplantation reveals differential heart failure development. *Animal Cells Syst (Seoul).* 27(1):234–248. doi:10.1080/19768354.2023.2265150.
- Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet.* 21(10):597–614. doi:10.1038/s41576-020-0236-x.
- López JM, Perlado J, Morona R, Northcutt RG, González A. 2013. Neuroanatomical organization of the cholinergic system in the central nervous system of a basal actinopterygian fish, the senegal bichir *Polypterus senegalus*. *J Comp Neurol.* 521(1):24–49. doi:10.1002/cne.23155.
- Lutek K, Foster KL, Standen EM. 2022. Behaviour and muscle activity across the aquatic-terrestrial transition in *Polypterus senegalus*. *J Exp Biol.* 225(23):243902. doi:10.1242/jeb.243902.
- Magid AA, Vokac Z, El Din Ahmed N. 1970. Respiratory function of the swim-bladders of the primitive fish *Polypterus senegalus*. *J Exp Biol.* 52(1):27–37. doi:10.1242/jeb.52.1.27.
- Marx V. 2023. Method of the year: long-read sequencing. *Nat Methods.* 20(1):6–11. doi:10.1038/s41592-022-01730-w.
- Nagashima A, Torii K, Ota C, Kato A. 2024. *slc26a12*-a novel member of the *slc26* family, is located in tandem with *slc26a2* in coelacanth, amphibians, reptiles, and birds. *Physiol Rep.* 12(11):e16089. doi:10.14814/phy2.16089.
- Noack K, Zardoya R, Meyer A. 1996. The complete mitochondrial DNA sequence of the bichir (*Polypterus ornatipinnis*), a basal ray-finned fish: ancient establishment of the consensus vertebrate gene order. *Genetics.* 144(3):1165–1180. doi:10.1093/genetics/144.3.1165.
- Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, Estevan-Morió E, Liu T, Nanni A, McIntyre L, Tseng E, et al. 2024. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat Methods.* 21(5):793–797. doi:10.1038/s41592-024-02229-2.
- Park J, Chung Y. 2019. Identification of neoantigens derived from alternative splicing and RNA modification. *Genomics Inform.* 17(3):e23. doi:10.5808/GI.2019.17.3.e23.
- Sasaki H, Shimizu S, Endo K, Takada M, Kawahara M, Tanaka H, Matsumura A, Iuchi K, Haneda H, Suzuki E, Kobayashi Y, et al. 2006. EGFR and erbB2 mutation status in Japanese lung cancer patients. *Int J Cancer.* 118(1):180–184. doi:10.1002/ijc.21301.
- Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun.* 11(1):1438. doi:10.1038/s41467-020-15171-6.
- Tomizawa K, Suda K, Onozato R, Kosaka T, Endoh H, Sekido Y, Shigematsu H, Kuwano H, Yatabe Y, Mitsudomi T. 2011. Prognostic and predictive implications of HER2/ERBB2/neu gene mutations in lung cancers. *Lung Cancer.* 74(1):139–144. doi:10.1016/j.lungcan.2011.01.014.
- Wickham H. 2016. *Ggplot2: elegant graphics for data analysis.* New York: Springer-Verlag.

- Wilhelm BC, Du TY, Standen EM, Larsson HCE. 2015. Polypterus and the evolution of fish pectoral musculature. *J Anat.* 226(6):511–522. doi:[10.1111/joa.12302](https://doi.org/10.1111/joa.12302).
- Yoo H, Kim HJ, Yang SH, Son GH, Gim J-A, Lee HW, Kim H, et al. 2022. Gene expression profiling of the habenula in rats exposed to chronic restraint stress. *Mol Cells.* 45(5):306–316. doi:[10.14348/molcells.2022.2257](https://doi.org/10.14348/molcells.2022.2257).
- Zhou W, Wang H, Yang Y, Guo F, Yu B, Su Z. 2022. Trophoblast cell subtypes and dysfunction in the placenta of individuals with preeclampsia revealed by single-cell RNA sequencing. *Mol Cells.* 45(5):317–328. doi:[10.14348/molcells.2021.0211](https://doi.org/10.14348/molcells.2021.0211).
- Zhu Y, Wang W, Wang X. 2015. Roles of transcriptional factor 7 in production of inflammatory factors for lung diseases. *J Transl Med.* 13(1):273. doi:[10.1186/s12967-015-0617-7](https://doi.org/10.1186/s12967-015-0617-7).