This repository contains files of the entire Ribosomal Protein (RP) sequence collection, the curated entries and all gene expression information as well as scripts (shell/R/python) used for analysis of the study:

*Sequence variation, common tissue expression patterns and learning models: a genome-wide survey of vertebrate ribosomal proteins*

Konstantinos A. Kyritsis¹,², Christos A. Ouzounis²*, Lefteris Angelis³, Ioannis S. Vizirianakis¹,*

¹Laboratory of Pharmacology, School of Pharmacy, Aristotle University of Thessaloniki, GR-54124 Thessalonica, Greece

²Biological Computation & Process Laboratory, Chemical Process & Energy Resources Institute, Centre for Research & Technology Hellas, GR-57001 Thessalonica, Greece

³Department of Informatics, Aristotle University of Thessaloniki, GR-54124 Thessalonica, Greece

*correspondence: ISV ivizir@pharm.auth.gr or CAO ouzounis@certh.gr

Keywords: ribosomal proteins, tissue-specific gene expression, GTEx, vertebrate lineage
# Each directory has been compressed (*.tar.gz).
# For convenience, enumeration has been added to the files of each directory.
# To reproduce results, remove enumeration (e.g. with rename like the example below)
# before running the code:

```
~/$ tar -zxvf '3_3D_Human_Ribosome_Structure_&_Identity_Scores.tar.gz'
~/$ cd '3_3D_Human_Ribosome_Structure_&_Identity_Scores'
~/$ ls -1
3.1_Chimera_4V6X_ribosome_identity_score.com
3.2_Chimera_4V6X_ribosome_identity_score.py
3.3_Chimera_4V6X_ribosome_LSU_SSU.com
3.4_Chimera_4V6X_ribosome_LSU_SSU.py
3.5_Hsapiens_Drerio_high_identity_score_pairs.xlsx
~/$ rename 's/^.*[^.]{1}[0-9]_//' *
~/$ ls -1
Chimera_4V6X_ribosome_identity_score.com
Chimera_4V6X_ribosome_identity_score.py
Chimera_4V6X_ribosome_LSU_SSU.com
Chimera_4V6X_ribosome_LSU_SSU.py
Hsapiens_Drerio_high_identity_score_pairs.xlsx
```
Directory Contents

1_Vertebrate_Ribosomal_Protein_Sequences_Collection:

Sequence comparison analysis performed using the CAST-masked (1) collection of RP sequences for eleven representative vertebrates.

Code:

- “1.1_ribosomal_proteins_cast_masked_all_vs_all.sh”: Bash shell commands for RP sequence database creation and sequence comparisons using BLASTP (2).
- “1.2_ribosomal_protein_sequences_casta_header_annotations_heatmap.R”: R code for creating annotations regarding the number of protein sequences and their type (refseq, translated, missing, etc.) and visualization.
- “1.3_ribosomal_protein_sequences_casta_all_vs_all_mcl_clusters_analysis.R”: R code for within-cluster identity score (%) calculation and visualization. Also contains Wilcoxon rank sum test of within-cluster identity scores (%) between the small (SSU) and the large (LSU) ribosomal subunits.

Files:

- “1.4_cluster_protein_annotations_df.xlsx”: Pairwise list of common gene names for RP sequences and their respective clusters.
- “1.5_Hsapiens_Dreriio_high_identity_score_pairs.xlsx”: Data about ortholog pairs of Homo sapiens and Danio rerio RP sequences (our code: r in f_0) with the highest within-cluster identity score (%).
- “1.6_RibosomalPco_Proteins_nomenclatures.xlsx”: Pairwise list of new (3) and previously used nomenclature of RPs.
- “1.7_Biolayout_files”:
  - “1.7.1_ribosomal_proteins_sequences_casta_all_vs_all_tab_no_headers.layout”: BioLayout Express 3D (4) session file of RP sequence similarity networks and Markov-chain clustering (MCL) results (5).
  - “1.7.2_ribosomal_proteins_sequences_casta_all_vs_all_tab_no_headers.classsets”: Pairwise list of unique identifiers for RP sequences and their respective clusters, exported from BioLayout Express 3D.
- “1.8_blastp_files”:
  - “1.8.1_ribosomal_proteins_sequences_casta_prot_db”: BLASTP database created from the CAST-masked collection of RP sequences for the eleven vertebrates.
  - “1.8.2_ribosomal_proteins_sequences_casta_all_vs_all_pair.out”: BLASTP pairwise alignment output of comparisons (all-vs-all) between CAST-masked RP sequences.
  - “1.8.3_ribosomal_proteins_sequences_casta_all_vs_all_tab.out”: BLASTP tabular output of comparisons (all-vs-all) between CAST-masked RP sequences.
  - “1.8.4_ribosomal_proteins_sequences_casta_all_vs_all_tab_no_headers.txt”: BLASTP tabular output of comparisons (all-vs-all) between CAST-masked RP sequences, with ‘#’ marked rows removed.
  - “1.8.5_Ribosomal_ProteinSequences_with_unique_IDs_v6.casta”: Collection of 1083 RP sequences for eleven representative vertebrate species. Unique identifiers have been assigned to each RP sequence. Low-complexity regions have been masked using CAST (1).
  - “1.8.6_Ribosomal_Protein_Sequences_with_unique_IDs_v6_headers_only.txt”: Fasta/casta file headers (annotations) only, from RP sequence collection file.
2. Ribosomal Protein Collection vs Ensembl Vertebrate Proteomes:

Sequence comparison analysis performed using the CAST-masked collection of RP sequences as query against the eleven proteomes of the respective vertebrate species.

Code:

- “2.1_blastp_ribosomal_proteins_vertebrate_proteomes_analysis.sh”: Bash shell commands for BLASTP database creation using Ensemble proteomes of the eleven representative vertebrates (version 100) (“Vertebrate.pep.all.fa”). Moreover, commands for subsequent BLASTP sequence comparisons, using CAST-masked RP sequence collection (“Ribosomal_Protein_Sequences_with_unique_IDs_v6.casta”) as query.
- “2.2_missing_ribosomal_proteins_in_vertebrate_proteomes.R”: R code for detecting and annotating RP sequences that are lacking BLASTP hits (e-value<0.05) in at least one vertebrate proteome. Reports on missing RPs can be found in the “2.4_blastp_file/report_RPs_missing_BLASTP_hits” directory.

Files:

- “2.3_ensembl_release_100_pep_all_proteomes”
  - “2.3.1_Vertebrate.pep.all.fa”: All proteomes of the eleven representative vertebrates, concatenated in a single FASTA file. Ensembl proteomes (version 100).
- “2.4_blastp_files”:
  - “2.4.1_ribosomal_proteins_vs_vertebrate_pep_all_pair.out”: BLASTP pairwise alignment output of comparisons between CAST-masked RP sequences (query) and total proteomes of the eleven representative vertebrates.
  - “2.4.2_ribosomal_proteins_vs_vertebrate_pep_all_tab.out”: BLASTP tabular output of comparisons between CAST-masked RP sequences (query) and total proteomes of the eleven representative vertebrates.
  - “2.4.3_ribosomal_proteins_vs_vertebrate_pep_all_tab_no_headers.txt”: BLASTP tabular output of comparisons between CAST-masked RP sequences (query) and total proteomes of the eleven representative vertebrates, with ‘#’ marked rows removed.
  - “2.4.4_vertebrate_pep_all_seg.asnb”: Soft-masking information (segmasker) for the BLASTP database created from the concatenated proteomes of all eleven representative vertebrates.
  - “2.4.5_Ribosomal_Protein_Sequences_with_unique_IDs_Lengths_v6.txt”: Same as above.
  - “2.4.6_Ribosomal_Protein_Sequences_with_unique_IDs_v6_headers_only.txt”: Same as above.
  - “2.4.7_Ribosomal_Protein_Sequences_with_unique_IDs_v6.casta”: Same as above.
  - “2.4.8_vertebrate_pep_all_masked_prot_db”: BLASTP database created from the concatenated proteomes of all eleven representative vertebrates (“2.3.1_Vertebrate.pep.all.fa” file).
  - “2.4.9_report_RPs_missing_BLASTP_hits”:
    - “2.4.9.1_Ribosomal_Proteins_missing_from_Ensembl_pep_all.xlsx/.csv”: Dataset of annotations for RP sequences missing BLASTP hits (e-value<0.05) in at least one vertebrate proteome.
    - “2.4.9.2_RPs_sequences_missing_hits_organisms_table.xlsx/.csv”: Dataset with number of ortholog sequences for 19 RPs, missing from each vertebrate proteome.
3.3D_Human_Ribosome_Structure_&_Identity_Scores:
Command files (.com) and python code for annotation of Chimera-generated (7) 3D structure of the human 80S ribosome with color code based on the within-cluster identity score (%) of each RP.

- “3.1_Chimera_4V6X_ribosome_identity_score.com”: Chimera command (.com) file containing 80S ribosome 3D structure model, from entry 4V6X from PDB, annotated with within-cluster identity score color code.
- “3.2_Chimera_4V6X_ribosome_identity_score”: Chimera python (.py) file containing 80S ribosome 3D structure model, from entry 4V6X from PDB, annotated with within-cluster identity score color code.
- “3.3_Chimera_4V6X_ribosome_LSU_SSU.com”: Chimera command (.com) file containing 80S ribosome 3D structure model, from entry 4V6X from PDB, annotated for large (LSU) and small (SSU) ribosomal proteins with magenta and blue color, respectively, while eEF2 factor (red) and E-tRNA structures (green) are also colored.
- “3.4_Chimera_4V6X_ribosome_LSU_SSU.py”: Chimera python (.py) file containing 80S ribosome 3D structure model, from entry 4V6X from PDB, annotated for large (LSU) and small (SSU) ribosomal proteins with magenta and blue color, respectively, while eEF2 factor (red) and E-tRNA structures (green) are also colored.
- “3.5_Hsapiens_Drerio_high_identity_score_pairs.xlsx”: Same as above.

4_Taxon-specific_Ribosomal_Protein_Expansion_Segments:
Multiple Sequence Alignment (MSA) of RPL29/eL29, RPL14/eL14 and RPL4/uL4 vertebrate ortholog sequences for the characterization of their conserved N-terminal segment as well as their C-terminal segment of varying length.

- “4.1_Human_zebrafish_sequence_size_comparisons”:
  o “4.1.1_Homo_sapiens_Ribosomal_Proteins_Sequences_with_unique_IDs_Lengths_v6.txt”: Length (amino acids) of each sequence of Homo sapiens in RP collection.
  o “4.1.2_Danio_rerio_Ribosomal_Proteins_Sequences_with_unique_IDs_Lengths_v6.txt”: Length (amino acids) of each sequence of Danio rerio in RP collection.
  o “4.1.3_Ribosomal_Protein_Sequences_with_unique_IDs_v6.fasta”: Collection of 1083 RP sequences for eleven representative vertebrate species. Unique identifiers have been assigned to each RP sequence.
  o “4.1.4_Ribosomal_Protein_Sequences_with_unique_IDs_v6.casta”: Same as above.
  o “4.1.5_Ribosomal_Protein_Sequences_with_unique_IDs_Lengths_v6.txt”: Same as above.
  o “4.1.6_Ribosomal_Proteins_nomenclatures.xlsx”: Same as above.
  o “4.1.7_hsapiens_drerio_rp_sequence_comparison_analysis.R”: R code for comparing ortholog RP sequences (“*ref_0*”) between Homo sapiens and Danio rerio.

- “4.2_Manually_curated_RP_sequence_collection”:
  o “4.2.1_rpl4_representative_refseq_sequences”: Directory containing RPL4/uL4 ortholog sequences from the eleven representative vertebrate collection of RP sequences (FASTA files). Also contains R code for the MSA of these orthologs with MUSCLE (8) and msa (9).
  o “4.2.2_rpl14_representative_refseq_sequences”: Directory containing RPL14/eL14 ortholog sequences from the eleven representative vertebrate collection of RP sequences (FASTA files). Also contains R code for the MSA of these orthologs with MUSCLE and msa.
• “4.2.3_rpl29_representative_refseq_sequences”: Directory containing RPL29/eL29 ortholog sequences from the eleven representative vertebrate collection of RP sequences (FASTA files). Also contains R code for the MSA of these orthologs with MUSCLE and msa.

• “4.3_NCBI_ortholog_predictions”:
  - “4.3.1_RP_Ortholog_Sequences.xlsx”: Number of RPL29/eL29, RPL14/eL14 and RPL4/uL4 ortholog sequences, from NCBI Eukaryotic Genome Annotation Pipeline, used for MSA. Number of sequences for different vertebrate taxa is displayed.
  - “4.3.2_RPL4_vertebrate_orthologs_analysis”: Directory containing MSA files (FASTA format) of RPL4/uL4 ortholog sequences from NCBI Eukaryotic Genome Annotation Pipeline (10), using COBALT (11). Additionally, Jalview (12) files (.json and .png) of MSA visualizations for RPL4/uL4 ortholog sequences.
  - “4.3.3_RPL14_vertebrate_orthologs_analysis”: Directory containing MSA files (FASTA format) of RPL14/eL14 ortholog sequences from NCBI Eukaryotic Genome Annotation Pipeline (10), using COBALT. Additionally, Jalview files (.json and .png) of MSA visualizations for RPL14/eL14 ortholog sequences.
  - “4.3.4_RPL29_vertebrate_orthologs_analysis”: Directory containing MSA files (FASTA format) of RPL29/eL29 ortholog sequences from NCBI Eukaryotic Genome Annotation Pipeline (10), using COBALT. Additionally, Jalview files (.json and .png) of MSA visualizations for RPL29/eL29 ortholog sequences.
  - “4.3.5_RPs_N-terminus_CD_search”: Directory containing N-terminal conserved (within-vertebrates) regions of RPL29/eL29 (1-53), RPL14/eL14 (1-131) and RPL4/uL4 (1-341) ortholog sequences from NCBI Eukaryotic Genome Annotation Pipeline (fasta files). Conserved domains for these regions were retrieved using Conserved Domain Database (13) (.txt files).

5_Ribosomal_Protein_Expression_GTEx_Human_Tissues:

Analysis of RP gene expression profiles in human tissues using RNA-Seq data (Transcript Per Million; TPM) from GTEx project (14).

Code:

• “5.1_RPs_analysis_GTEx_data_TPM_values.R”: R code for the:
  - t-SNE dimension reduction using 83 RP gene expression profiles in 33 human tissues.
  - Separate heatmap visualization of 89 RP gene expression profiles for tissues with low (<100), moderate (<800) and large (>800) number of samples.
  - Pearson correlation analysis of 89 RP gene expression profiles in 33 human tissues.
  - Visualization of UBA52/RPL40-precursor/eL40, RPLP1/P1, RPL9/uL6 and RPS26/eS26 expression patterns in 33 human tissues.

• “5.2_MultipleClassification_MinMaxScaler_GridSearchCV.py”: Python code for the creation of four learning models (15) in order to evaluate their performance (nested cross-validation) in predicting the type of human tissue, using 83 RP gene expression profiles.

• “5.3_RPs_analysis_GTEx_data_TPM_values_MultipleClassification_GridSearchCV_Results.R”: R code for collecting and summarizing the results from the learning models’ predictions. These include the arithmetic mean and standard deviation of accuracy, F1 score and Matthew’s Correlation Coefficient (MCC).

Files:
• “5.4_multiple_classification_model_results.xlsx”: Dataset containing the arithmetic mean and standard deviation of accuracy, recall, f1 score and MCC, for each learning model.
• “5.5_RPs_GTEx_Analysis_gene_tpm.txt”: Matrix of RP gene expression (TPM) for GTEx human tissue samples.
• “5.6_GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt”: Annotation file of GTEx human tissue samples.
• “5.7_Ribosomal_Proteins_nomenclatures.xlsx”: Same as above.
• “5.8_InterMineR.Gene.Identifiers.Ribosomal.Proteins_v1.xlsx”: File containing multiple annotation identifiers for human RPs, retrieved using InterMineR (16).

• “5.9.Multiclassification_Model_Results”: Directories for the results of the four types of multiclassification learning models for predicting the type of human tissue, using 83 RP gene expression profiles in human tissue samples of GTEx:
  o “5.9.1_LogisticRegression”: Confusion matrices of three independent Logistic regression (LogisticRegression) learning models (nested cross-validation) in CSV format (Predictions of tissues are in columns and the actual tissue types are in rows). Learning models are saved in pickle (.plk) format. Values of accuracy, recall, f1 score and Matthew’s Correlation Coefficient (MCC) are reported in TSV format.
  o “5.9.2_LinearSVC”: Confusion matrices of three independent Support-vector machine, with Linear kernel (LinearSVC), learning models (nested cross-validation) in CSV format (Predictions of tissues are in columns and the actual tissue types are in rows). Learning models are saved in pickle (.plk) format. Values of accuracy, recall, f1 score and Matthew’s Correlation Coefficient (MCC) are reported in TSV format.
  o “5.9.3_SVC”: Confusion matrices of three independent Support-vector machine, with Gaussian kernel (SVC), learning models (nested cross-validation) in CSV format (Predictions of tissues are in columns and the actual tissue types are in rows). Learning models are saved in pickle (.plk) format. Values of accuracy, recall, f1 score and Matthew’s Correlation Coefficient (MCC) are reported in TSV format.
  o “5.9.4_RandomForest”: Confusion matrices of three independent Random forest (RandomForest) learning models (nested cross-validation) in CSV format (Predictions of tissues are in columns and the actual tissue types are in rows). Learning models are saved in pickle (.plk) format. Values of accuracy, recall, f1 score and Matthew’s Correlation Coefficient (MCC) are reported in TSV format.

6_Vertebrate_Species_Ribosomal_Protein_Gene_Expression:
Analysis and comparisons of RP gene expression using RNA-Seq data (corrected Reads Per Kilobase Million; cRPKM) tissues of different vertebrate species (17). Additionally, multiclassification learning models (LogisticRegression, LinearSVC, SVC and RandomForest) were trained using RP gene expression data for human tissues from GTEx and evaluated in their ability to predict the tissue type of different vertebrate species.

• “6.1_Vertebrate_RP_expression_analysis.R”: R code for comparisons and heatmap visualization of 68 RP gene expression profiles across:
  I. 6 common tissues for 5 species, i.e. maximum number of common tissues, and
  II. 3 common tissues for 7 species, i.e. maximum number of common species.
• “6.2_Ribosomal_Proteins_Vertebrate_Table.xlsx”: File containing multiple annotation identifiers for vertebrate RPs.
• “6.3_RP_vertebrate_expression.csv”: Matrix of RP gene expression (cRPKM) for vertebrate species and tissues (including Ensembl annotation identifiers for RPs).
• “6.4_Ribosomal_Proteins_nomenclatures.xlsx”: Same as above.

• “6.5_GTEx_Vertebrate_Species_Tissue_Classification”:

  **Code:**
  - “6.5.1_MultipleClassification_Preprocessing_GTEx_MinMaxScaler_Vertebrate.py”: Python code for pre-processing of training and testing RP gene expression data (filtering and Min-Max scaling) to prepare for classification.
  - “6.5.2_MultipleClassification_MinMaxScaler_GridSearchCV_GTEx_Vertebrate.py”: Python code for the creation of multiclassification learning models using RP gene expression profiles in human tissues (5-fold cross-validation) to evaluate their accuracy in predicting the tissue type of different vertebrate species.
  - “6.5.3_GTEx_Vertebrate_Species_Tissue_Classification.R”: R code for collecting and summarizing the results from the learning models’ predictions.
  - “6.5.4_binom_test_on_ML_results.R”: R code for the exact binomial tests performed in order to determine whether the learning models were able to predict with greater than random accuracy (p-value<0.1; one-tailed binomial test).

  **Files:**
  - “6.5.5_vertebrate_RP_expression.csv”: Matrix of RP gene expression (cRPKM) for vertebrate species and tissues.
  - “6.5.6_RPs_per_Species.xlsx”: Table showing for each vertebrate species the available RPs, whose gene expression profiles were used for learning model training (GTEx human tissue expression data) and testing (multiple vertebrate tissue expression data).
  - “6.5.7_Tissues_per_Species.xlsx”: Table showing for each vertebrate species the available tissue samples, used for learning model training (GTEx respective tissue samples) and testing (vertebrate respective tissue samples).
  - “6.5.8_binom_res_vertebrate_species.xlsx”: Dataset containing the results of the exact binomial tests performed in order to determine whether the learning models were able to predict with greater than random accuracy (p-value<0.1; one-tailed binomial test).
  - “6.5.9_Multiple_Classification_Model_Results.xlsx”: Dataset containing the values of accuracy, recall, f1 score and MCC for the learning models of each vertebrate species.
  - “6.5.10_GTEx_RPs_raw_TPM_expression_data.csv”: Matrix of RP gene expression (TPM) for GTEx human tissue samples in CSV format.
  - “6.5.11_Ribosomal_Proteins_nomenclatures.xlsx”: Same as above.

  **Directories:**
  - Directories containing filtered and pre-processed (Min-Max scaling) datasets for training (GTEx expression data) and testing (vertebrate expression data) in each vertebrate species (e.g. “Anolis_carolinensis”).
  - Directories containing the confusion matrices (CSV format), the learning models (pickle (.plk) format), the learning model results (TSV format) and the results of the exact binomial tests (TXT format) for each combination of vertebrate and learning model (e.g. “Anolis_carolinensis_LogisticRegression”).
7_RP_Sequence_Conservation_Expression_Profiles:

Relationship between sequence conservation (RP within-cluster identity scores (%)) and human tissue expression profiles (TPM; GTEx project) for 89 RPs.

Code:

- “7.1_RP_Sequence_Expression_Integration.R”:
  I. The creation of tanglegram visualizing the connections between the increasing, within-cluster RP identity scores and hierarchical clustering of RP Pearson correlation coefficient matrix (correlation patterns).
  II. Repeatedly splitting RPs, based on correlation patterns, to an increasing number of clusters (from 2 up to 20), based on their correlation patterns, and then comparing RP identity scores between clusters using Kruskal-Wallis rank sum test.
  III. Visualization of RP ranking based on a) within-cluster identity score and b) correlation patterns. Absolute values of rank differences are also shown using a gradient color scale (green-red). Test for Spearman correlation between paired samples was performed between the pairwise rank lists.

Files:

- “7.2_RPs_GTEx_Analysis_gene_tpm.txt”: Same as above.
- “7.3_GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt”: Same as above.
- “7.4_Hsapiens_Drerio_high_identity_score_pairs.xlsx”: Same as above.
- “7.5_Ribosomal_Proteins_nomenclatures.xlsx”: Same as above.
- “7.6_InterMineR.Gene.Identifiers.Ribosomal.Proteins_v1.xlsx”: Same as above.
References

1. Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C. and Ouzounis, C.A. (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. Bioinformatics, 16, 915–22.

2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410.

3. Ban, N., Beckmann, R., Cate, J.H., Dinman, J.D., Dragon, F., Ellis, S.R., Lafontaine, D.L., Lindahl, L., Liljas, A., Lipton, J.M., et al. (2014) A new system for naming ribosomal proteins. Curr. Opin. Struct. Biol., 24, 165–169.

4. Goldovsky, L., Cases, I., Enright, A.J. and Ouzounis, C.A. (2005) BioLayout(Java): versatile network visualisation of structural and functional relationships. Appl. Bioinformatics, 4, 71–4.

5. Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res., 30, 1575–1584.

6. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2019) Ensembl 2020. Nucleic Acids Res., 48, D682–D688.

7. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera? A visualization system for exploratory research and analysis. J. Comput. Chem., 25, 1605–1612.

8. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32, 1792–1797.

9. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. and Hochreiter, S. (2015) msa: an R package for multiple sequence alignment. Bioinformatics, 31, 3997–3999.

10. Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M. and Kitts, P. (2013) Eukaryotic Genome Annotation Pipeline. In The NCBI Handbook. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US), pp. 111–130.

11. Papadopoulos, J.S. and Agarwala, R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics, 23, 1073–1079.

12. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25, 1189–1191.

13. Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.J., Marchler, G.H., Song, J.S., et al. (2020) CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res., 48, D265–D268.

14. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013) The Genotype-Tissue Expression (GTEx) project. Nat. Genet., 45, 580–585.

15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011) Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res., 12, 2825–2830.

16. Kyritsis, K.A., Wang, B., Sullivan, J., Lyne, R. and Micklem, G. (2019) InterMineR: an R package for InterMine databases. Bioinformatics, 10.1093/bioinformatics/btz039.
17. Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Guerousov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science, 338*, 1587–93.