## Supplementary Materials

## Supplementary Note 1

For this review - besides methodological summaries - we have carefully surveyed cross-species datasets. These datasets comprise various types of data that have been already tested with the reviewed methods, including RNA-seq data, network data and cross-species ontologies. These datasets offer opportunities for data exploration, re-analysis and benchmarking. Furthermore, we introduced harmonized transcriptome databases and integrative databases: two kinds of databases compiling large amounts of biomedical knowledge and high-throughput sequencing data. These databases will also aid biologists in gaining biological insights and facilitating development of novel methods. Following is a brief summary of the dataset we curated. For details please refer to ***Supplementary file 2***.

### Gene Expression:

* **Bulk RNA-seq data:** Collections of bulk transcriptomic datasets from large-scale comparative studies. These datasets often include a vast number of samples collected during the original study. For instance, Yang et al. [[1](#2s8eyo1)] sequenced transcriptomes from zebrafish at nine different stages, covering seven major developmental periods, to characterize the transcriptome landscape of zebrafish development. In addition to study-specific datasets, some are public dataset atlases that compile samples from various public sources. Mantica et al. [[2](#17dp8vu)], for example, assembled a comprehensive transcriptomic dataset spanning eight tissues across 20 bilaterian species to investigate conserved ancestral genes crucial for bilaterian evolution. The majority of these curated datasets provide websites for users to explore the data.
* **Paired Data:** Collections of bulk transcriptomic or microarray data that are manually curated, which are matched transcriptomic profiles across species that share similar phenotypes, biological processes and disease states. For example, Brubaker et al. [[3](#3rdcrjn)] curated 36 pairs of mouse and human studies with similar disease conditions including burn, trauma, endotoxemia and sepsis.
* **Single cell RNA-seq data:** Collections of single cell transcriptomic datasets from large-scale comparative studies or those that have been used for benchmarking in methods in section “How to map agnologous cell types and cell states across species”. Single-cell datasets were collected from a wide range of sources, including both tissue-specific [[4](#26in1rg),[5](#lnxbz9)] and whole-body datasets [[6](#35nkun2),[7](#1ksv4uv)]. In addition to datasets from typical research organisms such as mice and zebrafish, we also included datasets from non-model species, such as the macaque *Macaca fascicularis* [[8](#44sinio)] and *Spongilla lacustris* [[9](#2jxsxqh)].
* **Raw Transcriptomic Repositories:** Unprocessed sequencing data or expression data processed by submitters: SRA [[10](#z337ya)], GEO [[11](#3j2qqm3)], ArrayExpress [[12](#1y810tw)]
* **Harmonized Transcriptomic Repositories:** Raw expression data has been uniformly mapped to count tables using same pipeline: refine.bio (https://www.refine.bio/), ARCHS4 [[13](#4i7ojhp)], Recount3 [[14](#2xcytpi)], Bgee [[15](#1ci93xb)], Expression Atlas[[16](#3whwml4)], Gene Expression Neublas [[17](#2bn6wsx)]

### Functional Genomics and Biomedical Concepts Integration:

* **Gene ontology:** Controlled vocabularies of genes that describe the molecular functions, biological processes, and cellular locations of genes (GO ontology [[18](#qsh70q),[19](#3as4poj)]).
* **Biomedical concepts Ontologies:** Controlled vocabularies of biomedical concepts that are harmonized and connected among research organisms (e.g. upheno (http://obofoundry.org/ontology/upheno), Uberon [[20](#1pxezwc)]).
* **Knowledge graph base of biomedical concepts:** Comprehensive integration platforms that map relationships among different biomedical concepts such as genes, variants, genotypes, phenotypes and diseases from various species (e.g. Monarch [[21](#49x2ik5)], Phenomebrowser (http://phenomebrowser.net/)).
* **Data Querying:** Integrated databases that allow for querying genes, gene sets or other types of information from multiple species and easily compare function annotations of orthologs (e.g. Alliance [[22](#2p2csry)]).
* **Data compilation and reanalysis:** Provides both searching genetic/genomic functional information across species sources from a wide array of databases as well as the ability to jointly re-analyze them together (e.g. GeneWeaver [[23](#147n2zr)], Harmonizome [[24](#3o7alnk)]).

### Network Repositories

* **Molecular Interaction Networks:** Genetic entities (i.e. genes, proteins) are represented as nodes in a network and are linked together through edges based on biologically derived association. Below is a list of network repositories that contain networks for multiple species.
  + **Experimentally-derived Interactions:** Networks created using mainly experiments such as affinity purification followed by mass spectrometry (AP-MS), genetic interactions and yeast-two-hybrid assays. Often sparse networks and gene/protein coverage can vary widely (e.g. BioGRID [[25](#23ckvvd)], IntAct [[26](#ihv636)], DIP [[27](#32hioqz)], ConsensusPathDB [[28](#1hmsyys)], STRING [[29](#41mghml)]).
  + **Prediction-derived Interactions:** Networks that contain information from prediction such as co-expression data, co-occurrence throughout evolutionary history, text-mining, etc. Often much more dense and have high gene coverage across all species. These databases may also include experimentally evidenced interactions (e.g. STRING [[29](#41mghml)], IMP [[30](#2grqrue)]).

## References

1. **Deep mRNA Sequencing Analysis to Capture the Transcriptome Landscape of Zebrafish Embryos and Larvae** Hongxing Yang, Yan Zhou, Jianlei Gu, Shuying Xie, Yao Xu, Genfeng Zhu, Lei Wang, Jiyue Huang, Hong Ma, Jihua Yao *PLoS ONE* (2013-05-20) <https://doi.org/f4xg44> DOI: [10.1371/journal.pone.0064058](https://doi.org/10.1371/journal.pone.0064058) · PMID: [23700457](https://www.ncbi.nlm.nih.gov/pubmed/23700457) · PMCID: [PMC3659048](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3659048)

2. **Evolution of tissue-specific expression of ancestral genes across vertebrates and insects** Federica Mantica, Luis P Iñiguez, Yamile Marquez, Jon Permanyer, Antonio Torres-Mendez, Josefa Cruz, Xavi Franch-Marro, Frank Tulenko, Demian Burguera, Stephanie Bertrand, … Manuel Irimia *Cold Spring Harbor Laboratory* (2022-11-16) <https://doi.org/gs2f2m> DOI: [10.1101/2022.11.14.516384](https://doi.org/10.1101/2022.11.14.516384)

3. **Computational translation of genomic responses from experimental model systems to humans** Douglas K Brubaker, Elizabeth A Proctor, Kevin M Haigis, Douglas A Lauffenburger *PLOS Computational Biology* (2019-01-10) <https://doi.org/gr8cpf> DOI: [10.1371/journal.pcbi.1006286](https://doi.org/10.1371/journal.pcbi.1006286) · PMID: [30629591](https://www.ncbi.nlm.nih.gov/pubmed/30629591) · PMCID: [PMC6343937](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6343937)

4. **Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain** Blue B Lake, Song Chen, Brandon C Sos, Jean Fan, Gwendolyn E Kaeser, Yun C Yung, Thu E Duong, Derek Gao, Jerold Chun, Peter V Kharchenko, Kun Zhang *Nature Biotechnology* (2017-12-11) <https://doi.org/gcn3bj> DOI: [10.1038/nbt.4038](https://doi.org/10.1038/nbt.4038) · PMID: [29227469](https://www.ncbi.nlm.nih.gov/pubmed/29227469) · PMCID: [PMC5951394](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5951394)

5. **A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq** Christopher Daniel Green, Qianyi Ma, Gabriel L Manske, Adrienne Niederriter Shami, Xianing Zheng, Simone Marini, Lindsay Moritz, Caleb Sultan, Stephen J Gurczynski, Bethany B Moore, … Saher Sue Hammoud *Developmental Cell* (2018-09) <https://doi.org/gd9bnb> DOI: [10.1016/j.devcel.2018.07.025](https://doi.org/10.1016/j.devcel.2018.07.025) · PMID: [30146481](https://www.ncbi.nlm.nih.gov/pubmed/30146481) · PMCID: [PMC6713459](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6713459)

6. **SUPT3H-less SAGA coactivator can assemble and function without significantly perturbing RNA polymerase II transcription in mammalian cells** Veronique Fischer, Vincent Hisler, Elisabeth Scheer, Elisabeth Lata, Bastien Morlet, Damien Plassard, Dominique Helmlinger, Didier Devys, László Tora, Stéphane D Vincent *Nucleic Acids Research* (2022-07-25) <https://doi.org/gqtzsh> DOI: [10.1093/nar/gkac637](https://doi.org/10.1093/nar/gkac637) · PMID: [35871303](https://www.ncbi.nlm.nih.gov/pubmed/35871303) · PMCID: [PMC9371916](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9371916)

7. **Stem cell differentiation trajectories in <i>Hydra</i> resolved at single-cell resolution** Stefan Siebert, Jeffrey A Farrell, Jack F Cazet, Yashodara Abeykoon, Abby S Primack, Christine E Schnitzler, Celina E Juliano *Science* (2019-07-26) <https://doi.org/ggcd2p> DOI: [10.1126/science.aav9314](https://doi.org/10.1126/science.aav9314) · PMID: [31346039](https://www.ncbi.nlm.nih.gov/pubmed/31346039) · PMCID: [PMC7104783](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7104783)

8. **Cell transcriptomic atlas of the non-human primate Macaca fascicularis** Lei Han, Xiaoyu Wei, Chuanyu Liu, Giacomo Volpe, Zhenkun Zhuang, Xuanxuan Zou, Zhifeng Wang, Taotao Pan, Yue Yuan, Xiao Zhang, … Longqi Liu *Nature* (2022-04-13) <https://doi.org/gp4kz8> DOI: [10.1038/s41586-022-04587-3](https://doi.org/10.1038/s41586-022-04587-3)

9. **Profiling cellular diversity in sponges informs animal cell type and nervous system evolution** Jacob M Musser, Klaske J Schippers, Michael Nickel, Giulia Mizzon, Andrea B Kohn, Constantin Pape, Paolo Ronchi, Nikolaos Papadopoulos, Alexander J Tarashansky, Jörg U Hammel, … Detlev Arendt *Science* (2021-11-05) <https://doi.org/g4xt> DOI: [10.1126/science.abj2949](https://doi.org/10.1126/science.abj2949) · PMID: [34735222](https://www.ncbi.nlm.nih.gov/pubmed/34735222) · PMCID: [PMC9233960](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9233960)

10. **The Sequence Read Archive: a decade more of explosive growth** Kenneth Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J Rodney Brister, Christopher O’Sullivan *Nucleic Acids Research* (2021-11-24) <https://doi.org/gn4dhc> DOI: [10.1093/nar/gkab1053](https://doi.org/10.1093/nar/gkab1053) · PMID: [34850094](https://www.ncbi.nlm.nih.gov/pubmed/34850094) · PMCID: [PMC8728234](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8728234)

11. **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository** R Edgar *Nucleic Acids Research* (2002-01-01) <https://doi.org/fttpkn> DOI: [10.1093/nar/30.1.207](https://doi.org/10.1093/nar/30.1.207) · PMID: [11752295](https://www.ncbi.nlm.nih.gov/pubmed/11752295) · PMCID: [PMC99122](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC99122)

12. **ArrayExpress--a public database of microarray experiments and gene expression profiles** H Parkinson, M Kapushesky, M Shojatalab, N Abeygunawardena, R Coulson, A Farne, E Holloway, N Kolesnykov, P Lilja, M Lukk, … A Brazma *Nucleic Acids Research* (2007-01-03) <https://doi.org/cwxjnx> DOI: [10.1093/nar/gkl995](https://doi.org/10.1093/nar/gkl995) · PMID: [17132828](https://www.ncbi.nlm.nih.gov/pubmed/17132828) · PMCID: [PMC1716725](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1716725)

13. **Massive mining of publicly available RNA-seq data from human and mouse** Alexander Lachmann, Denis Torre, Alexandra B Keenan, Kathleen M Jagodnik, Hoyjin J Lee, Lily Wang, Moshe C Silverstein, Avi Ma’ayan *Nature Communications* (2018-04-10) <https://doi.org/gc92dr> DOI: [10.1038/s41467-018-03751-6](https://doi.org/10.1038/s41467-018-03751-6) · PMID: [29636450](https://www.ncbi.nlm.nih.gov/pubmed/29636450) · PMCID: [PMC5893633](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5893633)

14. **recount3: summaries and queries for large-scale RNA-seq expression and splicing** Christopher Wilks, Shijie C Zheng, Feng Yong Chen, Rone Charles, Brad Solomon, Jonathan P Ling, Eddie Luidy Imada, David Zhang, Lance Joseph, Jeffrey T Leek, … Ben Langmead *Genome Biology* (2021-11-29) <https://doi.org/gnm7zc> DOI: [10.1186/s13059-021-02533-6](https://doi.org/10.1186/s13059-021-02533-6) · PMID: [34844637](https://www.ncbi.nlm.nih.gov/pubmed/34844637) · PMCID: [PMC8628444](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8628444)

15. **The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals** Frederic B Bastian, Julien Roux, Anne Niknejad, Aurélie Comte, Sara S Fonseca Costa, Tarcisio Mendes de Farias, Sébastien Moretti, Gilles Parmentier, Valentine Rech de Laval, Marta Rosikiewicz, … Marc Robinson-Rechavi *Nucleic Acids Research* (2020-10-10) <https://doi.org/gjvp8f> DOI: [10.1093/nar/gkaa793](https://doi.org/10.1093/nar/gkaa793) · PMID: [33037820](https://www.ncbi.nlm.nih.gov/pubmed/33037820) · PMCID: [PMC7778977](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7778977)

16. **Expression Atlas update: gene and protein expression in multiple species** Pablo Moreno, Silvie Fexova, Nancy George, Jonathan R Manning, Zhichiao Miao, Suhaib Mohammed, Alfonso Muñoz-Pomer, Anja Fullgrabe, Yalan Bi, Natassja Bush, … Irene Papatheodorou *Nucleic Acids Research* (2021-11-24) <https://doi.org/gptg5h> DOI: [10.1093/nar/gkab1030](https://doi.org/10.1093/nar/gkab1030) · PMID: [34850121](https://www.ncbi.nlm.nih.gov/pubmed/34850121) · PMCID: [PMC8728300](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8728300)

17. **Gene Expression Nebulas (GEN): a comprehensive data portal integrating transcriptomic profiles across multiple species at both bulk and single-cell levels** Yuansheng Zhang, Dong Zou, Tongtong Zhu, Tianyi Xu, Ming Chen, Guangyi Niu, Wenting Zong, Rong Pan, Wei Jing, Jian Sang, … Zhang Zhang *Nucleic Acids Research* (2021-09-30) <https://doi.org/gptg6v> DOI: [10.1093/nar/gkab878](https://doi.org/10.1093/nar/gkab878) · PMID: [34591957](https://www.ncbi.nlm.nih.gov/pubmed/34591957) · PMCID: [PMC8728231](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8728231)

18. **Gene Ontology: tool for the unification of biology** Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, JMichael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, … Gavin Sherlock *Nature Genetics* (2000-05) <https://doi.org/b9gp96> DOI: [10.1038/75556](https://doi.org/10.1038/75556) · PMID: [10802651](https://www.ncbi.nlm.nih.gov/pubmed/10802651) · PMCID: [PMC3037419](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419)

19. **The Gene Ontology knowledgebase in 2023**, Suzi A Aleksander, James Balhoff, Seth Carbon, JMichael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, … Monte Westerfield *GENETICS* (2023-03-03) <https://doi.org/gt6n4z> DOI: [10.1093/genetics/iyad031](https://doi.org/10.1093/genetics/iyad031) · PMID: [36866529](https://www.ncbi.nlm.nih.gov/pubmed/36866529) · PMCID: [PMC10158837](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10158837)

20. **Uberon, an integrative multi-species anatomy ontology** Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, Melissa A Haendel *Genome Biology* (2012) <https://doi.org/fxx6qr> DOI: [10.1186/gb-2012-13-1-r5](https://doi.org/10.1186/gb-2012-13-1-r5) · PMID: [22293552](https://www.ncbi.nlm.nih.gov/pubmed/22293552) · PMCID: [PMC3334586](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334586)

21. **The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species** Christopher J Mungall, Julie A McMurry, Sebastian Köhler, James P Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, Tom Conlin, Nathan Dunn, Mark Engelstad, … Melissa A Haendel *Nucleic Acids Research* (2016-11-29) <https://doi.org/f9v7bz> DOI: [10.1093/nar/gkw1128](https://doi.org/10.1093/nar/gkw1128) · PMID: [27899636](https://www.ncbi.nlm.nih.gov/pubmed/27899636) · PMCID: [PMC5210586](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210586)

22. **Harmonizing model organism data in the Alliance of Genome Resources**, Julie Agapite, Laurent-Philippe Albou, Suzanne A Aleksander, Micheal Alexander, Anna V Anagnostopoulos, Giulia Antonazzo, Joanna Argasinska, Valerio Arnaboldi, Helen Attrill, … Mark Zytkovicz *Genetics* (2022-02-25) <https://doi.org/gr8fcg> DOI: [10.1093/genetics/iyac022](https://doi.org/10.1093/genetics/iyac022) · PMID: [35380658](https://www.ncbi.nlm.nih.gov/pubmed/35380658) · PMCID: [PMC8982023](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8982023)

23. **GeneWeaver: a web-based system for integrative functional genomics** Erich J Baker, Jeremy J Jay, Jason A Bubier, Michael A Langston, Elissa J Chesler *Nucleic Acids Research* (2011-11-12) <https://doi.org/brqbhs> DOI: [10.1093/nar/gkr968](https://doi.org/10.1093/nar/gkr968) · PMID: [22080549](https://www.ncbi.nlm.nih.gov/pubmed/22080549) · PMCID: [PMC3245070](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245070)

24. **The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins** Andrew D Rouillard, Gregory W Gundersen, Nicolas F Fernandez, Zichen Wang, Caroline D Monteiro, Michael G McDermott, Avi Ma’ayan *Database* (2016) <https://doi.org/gf8xgq> DOI: [10.1093/database/baw100](https://doi.org/10.1093/database/baw100) · PMID: [27374120](https://www.ncbi.nlm.nih.gov/pubmed/27374120) · PMCID: [PMC4930834](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4930834)

25. **The <scp>BioGRID</scp> database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions** Rose Oughtred, Jennifer Rust, Christie Chang, Bobby‐Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, … Mike Tyers *Protein Science* (2020-11-23) <https://doi.org/gk7skv> DOI: [10.1002/pro.3978](https://doi.org/10.1002/pro.3978) · PMID: [33070389](https://www.ncbi.nlm.nih.gov/pubmed/33070389) · PMCID: [PMC7737760](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7737760)

26. **The IntAct database: efficient access to fine-grained molecular interaction data** Noemi del Toro, Anjali Shrivastava, Eliot Ragueneau, Birgit Meldal, Colin Combe, Elisabet Barrera, Livia Perfetto, Karyn How, Prashansa Ratan, Gautam Shirodkar, … Henning Hermjakob *Nucleic Acids Research* (2021-11-11) <https://doi.org/gt5bdm> DOI: [10.1093/nar/gkab1006](https://doi.org/10.1093/nar/gkab1006) · PMID: [34761267](https://www.ncbi.nlm.nih.gov/pubmed/34761267) · PMCID: [PMC8728211](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8728211)

27. **DIP: the Database of Interacting Proteins** I Xenarios *Nucleic Acids Research* (2000-01-01) <https://doi.org/bf7xcq> DOI: [10.1093/nar/28.1.289](https://doi.org/10.1093/nar/28.1.289) · PMID: [10592249](https://www.ncbi.nlm.nih.gov/pubmed/10592249) · PMCID: [PMC102387](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102387)

28. **ConsensusPathDB—a database for integrating human functional interaction networks** Atanas Kamburov, Christoph Wierling, Hans Lehrach, Ralf Herwig *Nucleic Acids Research* (2008-10-21) <https://doi.org/fj6s27> DOI: [10.1093/nar/gkn698](https://doi.org/10.1093/nar/gkn698) · PMID: [18940869](https://www.ncbi.nlm.nih.gov/pubmed/18940869) · PMCID: [PMC2686562](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2686562)

29. **The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest** Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, … Christian von Mering *Nucleic Acids Research* (2022-11-12) <https://doi.org/gs7sn3> DOI: [10.1093/nar/gkac1000](https://doi.org/10.1093/nar/gkac1000) · PMID: [36370105](https://www.ncbi.nlm.nih.gov/pubmed/36370105) · PMCID: [PMC9825434](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825434)

30. **IMP 2.0: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks** Aaron K Wong, Arjun Krishnan, Victoria Yao, Alicja Tadych, Olga G Troyanskaya *Nucleic Acids Research* (2015-05-12) <https://doi.org/f7nxbp> DOI: [10.1093/nar/gkv486](https://doi.org/10.1093/nar/gkv486) · PMID: [25969450](https://www.ncbi.nlm.nih.gov/pubmed/25969450) · PMCID: [PMC4489318](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4489318)