TYPE Perspective
PUBLISHED 31 March 2023
DOI 10.3389/fgene.2023.1106631

Check for updates

OPEN ACCESS

EDITED BY
Hari Krishna Yalamanchili,
Baylor College of Medicine, United States

REVIEWED BY
Lei Wang,
The Ohio State University, United States

*CORRESPONDENCE
Deok Ryong Kim,
✉ drkim@gnu.ac.kr

SPECIALTY SECTION
This article was submitted to ELSI in Science and Genetics,
a section of the journal
Frontiers in Genetics

RECEIVED 25 November 2022
ACCEPTED 21 March 2023
PUBLISHED 31 March 2023

CITATION
Ahmed M, Kim HJ and Kim DR (2023), Maximizing the utility of public data.
Front. Genet. 14:1106631.
doi: 10.3389/fgene.2023.1106631

Maximizing the utility of public data

Mahmoud Ahmed[1], Hyun Joon Kim[2] and Deok Ryong Kim[1]*

[1]Department of Biochemistry and Convergence Medical Sciences, Institute of Health Sciences, College of Medicine, Gyeongsang National University, Jinju, Republic of Korea, [2]Department of Anatomy and Convergence Medical Sciences, Institute of Health Sciences, College of Medicine, Gyeongsang National University, Jinju, Republic of Korea

The human genome project galvanized the scientific community around an ambitious goal. Upon completion, the project delivered several discoveries, and a new era of research commenced. More importantly, novel technologies and analysis methods materialized during the project period. The cost reduction allowed many more labs to generate high-throughput datasets. The project also served as a model for other extensive collaborations that generated large datasets. These datasets were made public and continue to accumulate in repositories. As a result, the scientific community should consider how these data can be utilized effectively for the purposes of research and the public good. A dataset can be re-analyzed, curated, or integrated with other forms of data to enhance its utility. We highlight three important areas to achieve this goal in this brief perspective. We also emphasize the critical requirements for these strategies to be successful. We draw on our own experience and others in using publicly available datasets to support, develop, and extend our research interest. Finally, we underline the beneficiaries and discuss some risks involved in data reuse.

KEYWORDS

public-data, data-reuse, data-analysis, data-sharing, reproducible-research

Introduction

The human genome project (HGP) galvanized the scientific community around an ambitious goal (Lander et al., 2001). Upon completion, the HGP produced several crucial discoveries, and a new era of research began (Gates et al., 2021). The project provided an estimate of the number of genes and a comprehensive list of their coding sequences. These developments have allowed for a shift away from single gene models and kickstarted the discipline of systems biology analysis. Furthermore, the non-coding regions came into focus, and their function began to be studied along with the variation across individuals (ENCODE Project Consortium TEP Feingold et al., 2004; Sabeti et al., 2007). More significantly, scientists developed new technologies and analytic methods during the period of the project. The cost reduction allowed many more labs to generate high-throughput datasets (Metzker, 2010). The project also served as a model for other big collaborations that generated larger datasets. These included efforts to sequence a large number of genomes from different populations across the globe (Auton et al., 2015; Wall et al., 2019; Smedley et al., 2022). Others concentrated on specific diseases, and disease models such as cancer (Weinstein et al., 2013; Subramanian et al., 2017).

Alongside these ambitious endeavors, small datasets were generated and made public and continued to accumulate in repositories. The gene expression omnibus (GEO), ArrayExpress, and sequence read archive (SRA) are just a few examples (Edgar et al., 2002; Parkinson et al., 2007; Leinonen et al., 2011). Individual labs typically use these

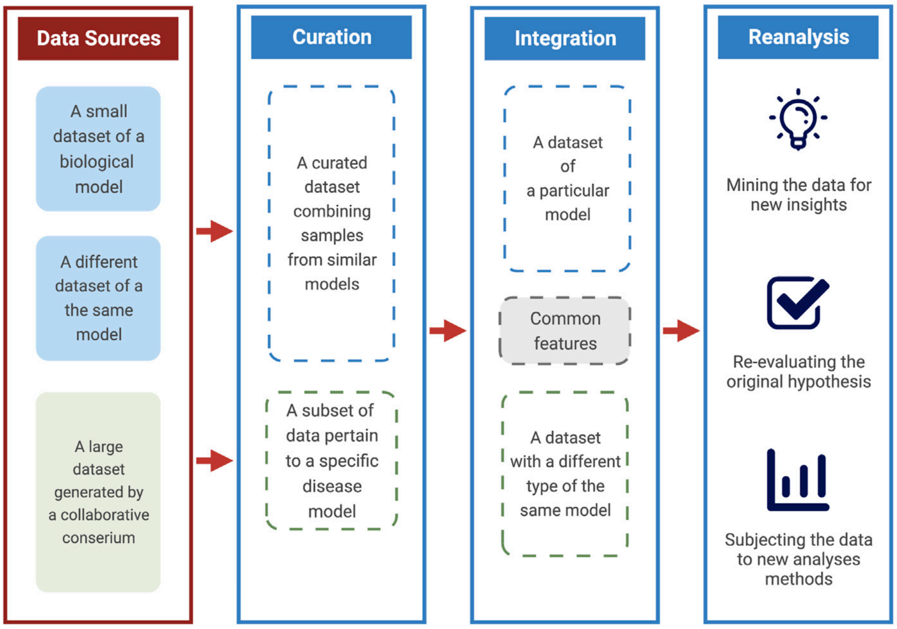
Data Sources
A small dataset of a biological model
A different dataset of a the same model
A large dataset generated by a collaborative conserium
Curation
A curated dataset combining samples from similar models
A subset of data pertain to a specific disease model
Integration
A dataset of a particular model
Common features
A dataset with a different type of the same model
Reanalysis
Mining the data for new insights
Re-evaluating the original hypothesis
Subjecting the data to new analyses methods

FIGURE 1
A scheme for maximizing the utility of public data. A diagram explaining the ways and benefits of using data from public sources. These include curating datasets of particular models, integrating different data types, and re-analysis with various goals.

repositories to document and distribute raw and processed data accompanying publications. As a result, the scientific community should consider how to use these resources for research and the public good. Efforts went into fostering the adoption of best practices to document and share data, and proper policies around accessibility (Thorogood and Knoppers, 2016; Rehm et al., 2021). Indeed, recognizing reuse as a legitimate form of research at the junior and senior levels has become more acceptable and encouraged (Duvallet, 2020; Raman, 2021). Furthermore, initiatives are being proposed to develop cloud environments to store, manage, and analyze data in effective, scalable, and secure ways (Schatz et al., 2022). Similarly, models have been developed to create an ecosystem to improve data description and hosting (Charbonneau et al., 2022).

In this short perspective, we highlight three critical areas to maximize the utility of publicly available data. We draw on ower own experience and others in using genomics and transcriptomic public datasets to support and extend our research interest. A dataset can be re-analyzed, curated, or integrated with other forms of data to enhance its utility (Figure 1). Although not an exhaustive list of possible use cases, those three strategies encapsulate a majority of research types that rely on the existence of public data. Furthermore, each has identifiable benefits and requirements. Finally, we remark on the overall benefit of reuse, the primary beneficiaries, and highlight some of its risks.

Reanalyzing primary data

High-throughput experiments generate simultaneous measurements of a large portion, if not all, of the genome. It has become a standard practice for researchers to share the raw data and documentation of how they generated it. The obvious case for reuse is to mine the dataset for insights that were not explored in the initially published studies. Investigators could focus on a particular subset of the data and analyze it in-depth. When data is available in raw or unaltered formats. The availability of the data facilitates the peer-review process where the suggested hypothesis can be examined and the analysis reproduced. Others choose to verify or refute the hypothesis presented in the original analyses by examining them independently. Both reuse cases yield additional value and benefit the wider community. We employed existing gene expression, and DNA binding datasets of preadipocytes to explore the role of autophagy during adipocyte differentiation (Ahmed et al., 2019). Our analysis focused on the subset of gene products involved in lipogenesis and autophagy. We were able to link the regulation of key autophagy genes to the transcription factors that drive the reprogramming of progenitor cells into mature adipocytes.

Any given dataset can only be analyzed in a few ways in any given study. Existing and newly developed tools can be applied to these datasets to generate new insights. Often, statistically sophisticated approaches have the potential to extract more information from the same data points. We used co-expression analysis and unsupervised learning methods to study gene product interactions in the preadipocyte differentiation dataset mentioned above (Ahmed et al., 2018). Furthermore, we deconvoluted the mixture of differentiating adipocytes into subpopulations. We suggested that mature adipocytes originate from a small fraction of the progenitors (Ahmed et al., 2021a). Researchers have employed large datasets of RNA-seq to predict genes, transcripts, and promoters (Steijger et al., 2013; Keilwagen et al., 2018; Wilson et al., 2021). Analysis methods such as chromatin segmentation

TABLE 1 Ways, benefits, and requirements to increase the utility of public data.

Strategy	Benefits	Requirements
Reanalysis	Mining the data for new insights	● Sharing the raw data
		● Documenting and sharing meta-data
	Re-evaluating the original hypothesis	● Clear presentation of how data were generated and analyzed
		● Sharing reproducible code
	Subjecting the data to new analyses methods	● Encouraging the development of new analysis methods and techniques
		● Developing new tools
Curation	increasing the sample size	Proper documentation and sharing of experimental protocols
	Filling in the gaps	● Developing an ontology to code experimental variables
		● Transparent and transferable annotations
	Improving the utility	● Funding for computational resources to generate and maintain processed data
		● Lowering the entry barrier for lab biologists
Integration	Complementing an analysis	Encouraging collaboration between dry and wet labs
	Examining a phenomenon from different angles	Generating complementary datasets (OMICS)
	Spurring the development of new methods	Funding open-source methods and tools development

are only possible when a considerable number of histone markers have been profiled using ChIP-seq (Hoffman et al., 2012). These and others are examples of research pursuits that are only possible because or enhance the utility of publicly available data.

Curating data from different sources

Researchers interested in a particular topic often use similar models and similar experimental designs. Despite being generated using different protocols, combining the datasets from separate groups could help filling the gaps in the design and *increasing the statistical power* of the analysis. For an example from our work, the cell line model 3T3-L1 has been used for years in standard time-course experiments to study adipocyte differentiation. Combining several gene expression experiments produced a dataset with numerous samples and covered more time points in the differentiation course (Ahmed and Kim, 2019). A similar problem arises when generating gene expression data with drug or genetic perturbations. Incorporating more than one dataset means including additional perturbations in the study (Ahmed et al., 2020a). The reverse is also possible by curating and annotating a subset of a larger dataset to address a specific aspect of the model or focus on a data type (Ganzfried et al., 2013).

Curation in either direction, smaller to larger, or larger to smaller datasets improves the utility of the data. One added benefit is that curators have to homogenize data from different sources and use unified terminologies. Furthermore, curators can pre-process and quality assesses large files of row data and make the data available in more accessible formats. The recount3 project accomplished just that by curating and processing thousands of RNA-seq reads files and making the results available in the form of gene counts (Wilks et al., 2021). Likewise, KnockTF is a curated gene expression dataset of transcription factors knockdown experiments (Feng et al., 2019). These two examples highlight yet another advantage of curation. Namely, it exposes the data to the scientific community beyond computational labs and makes it available and easy to use for lab biologists.

Integrating multiple types of data

Different high-throughput technologies generate data types that describe different layers of biology. Integrating data types can be beneficial to either verify or complement the observations made based on a single data type. For example, the binding of a transcription factor to the DNA of a specific region is not necessarily a claim about the function of that transcription factor. However, the likelihood that this binding is functional increases if, under the perturbation of that transcription factor, the expression of the nearest gene changes. We used both binding and gene expression data to study the interaction between adipogenic transcription factors and the autophagy genes of interest. We were able to show that a hierarchy of transcription regulators, including those controlling the differentiation program, regulates autophagy directly or indirectly through other factors (Ahmed et al., 2022). Integrating data types, in this case, allowed for identifying a phenomenon that was not otherwise obvious from gene expression or DNA-binding alone.

New methods capitalize on this idea of combining data from different sources. For example, binding and expression target analysis (BETA) infers direct target genes of transcription factors by integrating binding peaks, and gene expression changes under the factor's perturbation (Wang et al., 2013). We further extended this method to the interaction of two DNA-binding proteins as they function in cooperative or competitive ways to induce or repress a shared target (Ahmed et al., 2020b). Existing biological knowledge can also help in modeling and interpreting experimental data. The known pathways can be encoded in a network where the nodes are the biological entities, and the edges are the known interactions between them. The biological expression language (BEL) is one way to represent this knowledge in a standard computable graph (Hoyt et al., 2018). Methods such as network perturbation amplitudes (NPA) take advantage of these graphs to infer the function of the biological entities from the changes in gene expression in response to drug treatment or genetic perturbations (Martin et al., 2012; Martin et al., 2019). We used this approach to generate a database of cancer cell-specific perturbations and to screen for potential antimetastatic drugs in breast cancer (Ahmed and Kim, 2021). In one instance, we inferred the pathways that control the expression of an antimetastatic gene along with multiple drugs that target it (Ahmed et al., 2021b). We experimentally validated some of these predictions.

Discussion

In this perspective, we highlighted three avenues to maximize the utility of existing public data. Namely, re-analyzing primary data, curating data from different sources, and integrating multiple data types. We also opined the potential benefits of each strategy with examples from our work and others. We acknowledge that successful reuse places demands on the broader community regarding the documenting and sharing of data (Table 1). Other broad benefits can be accrued from this model of sharing and reuse, which we discuss next in addition to the risks some pointed out.

Our model of sharing and reusing data focuses on extracting more value from the available resources. However, we acknowledge that for this to be standard, it requires data to be documented and shared in transparent reproducible ways. In addition, resources should be available for curating, annotating, developing tools, and re-analyzing data. We similarly view these as stimulators for good: encouraging best practices of reproducible research and developing models for open science (Peng, 2011; OECD. Making Open Science a Reality, 2015). In other words, if data were generated with potential reuse, by the primary authors or others, in mind, it would be a net benefit to the community. The conditions that encourage reproducible open science are the same that foster and encourage reuse. We also believe data sharing and reuse would benefit researchers in low-resource labs and developing nations (Cheah et al., 2015). Finally, easily accessible data would facilitate and lower the entry barrier for non-computational researchers to use the extensive knowledge made possible by large datasets.

It is necessary to acknowledge the potential risks associated with the reuse of public data. Sielemann and colleagues underlined some of these and suggested solutions (Sielemann et al., 2020). They warn that user-submitted data may be of questionable quality and require substantial work to locate and obtain before reaching the point of analysis. This effort on behave of the researcher interested in reuse may be wasted as the data quality would only be assessed later in the process. Furthermore, reusing public data may produce duplicate records of the same dataset. Finally, several ethical issues arise in reuse cases. For example, not crediting the original authors may disincentive others from sharing their data and code in the future (Curty et al., 2017). Besides, funding for generating new datasets may stall because similar datasets exist or others could be repurposed (Nature, 2016).

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

MA and DK contributed to the concepts and the writing. HK edited the manuscript and supported the work.

Funding

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) of the Korea government [2020R1A2C2011416] and by the Commercializations Promotion Agency for R&D Outcomes (COMPA) grant funded by the Korea government (MSIT) [1711173796].

Acknowledgments

We thank all the lab members for their thoughtful feedback on this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmed, M., and Kim, D. R. (2021). Linps: A database for cancer-cell-specific perturbations of biological networks. *Database* 2021, baab048. doi:10.1093/database/baab048

Ahmed, M., and Kim, D. R. (2019). Modelling the gene expression and the DNA-binding in the 3T3-L1 differentiating adipocytes. *Adipocyte* 8, 401–411. doi:10.1080/21623945.2019.1697563

Ahmed, M., Lai, T. H., Hwang, J. S., Zada, S., Pham, T. M., and Kim, D. R. (2019). Transcriptional regulation of autophagy genes via stage-specific activation of cebpb and pparg during adipogenesis: A systematic study using public gene expression and transcription factor binding datasets. *Cells* 8, 1321. doi:10.3390/cells8111321

Ahmed, M., Lai, T. H., and Kim, D. R. (2021). A small fraction of progenitors differentiate into mature adipocytes by escaping the constraints on the cell structure. *Front. Cell Dev. Biol.* 9, 753042. doi:10.3389/fcell.2021.753042

Ahmed, M., Lai, T. H., Kim, W., and Kim, D. R. (2021). A functional network model of the metastasis suppressor PEBP1/RKIP and its regulators in breast cancer cells. *Cancers* 13, 6098. doi:10.3390/cancers13236098

Ahmed, M., Lai, T. H., Pham, T. M., Zada, S., Elashkar, O., Hwang, J. S., et al. (2022). Hierarchical regulation of autophagy during adipocyte differentiation. *PLoS ONE* 17, e0250865. doi:10.1371/journal.pone.0250865

Ahmed, M., Min, D. S., and Kim, D. R. (2020). Curated gene expression dataset of differentiating 3T3-L1 adipocytes under pharmacological and genetic perturbations. *Adipocyte* 9, 600–608. doi:10.1080/21623945.2020.1829852

Ahmed, M., Min, D. S., and Kim, D. R. (2020). Integrating binding and expression data to predict transcription factors combined function. *BMC Genomics* 21, 610. doi:10.1186/s12864-020-06977-1

Ahmed, M., Seok Hwang, J., Huyen Lai, T., Zada, S., Quoc Nguyen, H., Pham, T. M., et al. (2018). Co-Expression network analysis of AMPK and autophagy gene products during adipocyte differentiation. *Int. J. Mol. Sci.* 19, 1808. doi:10.3390/ijms19061808

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68. doi:10.1038/nature15393

Charbonneau, A. L., Brady, A., Czajkowski, K., Aluvathingal, J., Canchi, S., Carter, R., et al. (2022). Making common fund data more findable: Catalyzing a data ecosystem. *GigaScience* 11, giac105. doi:10.1093/GIGASCIENCE/GIAC105

Cheah, P. Y., Tangseefa, D., Somsaman, A., Chunsuttiwat, T., Nosten, F., Day, N. P. J., et al. (2015). Perceived benefits, harms, and views about how to share data responsibly: A qualitative study of experiences with and attitudes toward data sharing among research staff and community representatives in Thailand. *J. Empir. Res. Hum. Res. Ethics* 10, 278–289. doi:10.1177/1556264615592388

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., and Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *PLoS ONE* 12, e0189288. doi:10.1371/journal.pone.0189288

Duvallet, C. (2020). Data detectives, self-love, and humility: A research parasite's perspective. *Gigascience* 9, giz148. doi:10.1093/gigascience/giz148

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids Res.* 30, 207–210. doi:10.1093/nar/30.1.207

ENCODE Project Consortium TEP, Feingold, E. A., Good, P. J., Guyer, M. S., Kamholz, S., Liefer, L., et al. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science* 306, 636–640. doi:10.1126/science.1105136

Feng, C., Song, C., Liu, Y., Qian, F., Gao, Y., Ning, Z., et al. (2019). KnockTF: A comprehensive human gene expression profile database with knockdown/knockout of transcription factors. *Nucleic acids Res.* 48, D93–D100. doi:10.1093/nar/gkz881

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., et al. (2013). curatedOvarianData: Clinically annotated data for the ovarian cancer transcriptome. *Database* 2013, 2013. doi:10.1093/database/bat013

Gates, A. J., Gysi, D. M., Kellis, M., and Barabási, A. L. (2021). A wealth of discovery built on the human genome project — By the numbers. *Nature* 590, 212. doi:10.1038/d41586-021-00314-6

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* 9, 473–476. doi:10.1038/nmeth.1937

Hoyt, C. T., Domingo-Fernández, D., and Hofmann-Apitius, M. (2018). BEL commons: An environment for exploration and analysis of networks encoded in biological expression language. *Database J. Biol. databases curation* 2018, bay126. doi:10.1093/database/bay126

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., and Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinforma.* 19, 189. doi:10.1186/s12859-018-2203-5

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409. doi:10.1038/35057062

Leinonen, R., Sugawara, H., and Shumway, M. International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi:10.1093/nar/gkq1019

Martin, F., Gubian, S., Talikka, M., Hoeng, J., and Peitsch, M. C. (2019). Npa: An R package for computing network perturbation amplitudes using gene expression data and two-layer networks. *BMC Bioinforma.* 20, 451. doi:10.1186/s12859-019-3016-x

Martin, F., Thomson, T. M., Sewer, A., Drubin, D. A., Mathis, C., Weisensee, D., et al. (2012). Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC Syst. Biol.* 6, 54. doi:10.1186/1752-0509-6-54

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nat. Rev. Genet.* 11, 31. doi:10.1038/nrg2626

Nature (2016). Announcement: Where are the data? *Nature* 537, 138. doi:10.1038/537138a

OECD. Making Open Science a Reality (2015). *OECD science, technology and industry policy papers*. Berlin. Germany: OECD iLibrary.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., et al. (2007). ArrayExpress - a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* 35, D747–D750. doi:10.1093/nar/gkl995

Peng, R. D. (2011). Reproducible research in computational science. *Science* 334, 1226. doi:10.1126/science.1213847

Raman, A. T. (2021). A research parasite's perspective on establishing a baseline to avoid errors in secondary analyses. *GigaScience* 10, giab015. doi:10.1093/gigascience/giab015

Rehm, H. L., Page, A. J., Smith, L., Adams, J. B., Alterovitz, G., Babb, L. J., et al. (2021). GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* 1, 100029. doi:10.1016/j.xgen.2021.100029

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi:10.1038/nature06250

Schatz, M. C., Philippakis, A. A., Afgan, E., Banks, E., Carey, V. J., Carroll, R. J., et al. (2022). Inverting the model of genomics data sharing with the NHGRI genomic data science analysis, visualization, and informatics lab-space. *Cell Genomics* 2, 100085. doi:10.1016/j.xgen.2021.100085

Sielemann, K., Hafner, A., and Pucker, B. (2020). The reuse of public datasets in the life sciences: Potential risks and rewards. *PeerJ* 8, e9954. doi:10.7717/peerj.9954

Smedley, D., Smith, K. R., Martin, A., Thomas, E. A., McDonagh, E. M., and Cipriani, V. (2022). 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. *N. Engl. J. Med.* 385, 1868. doi:10.1530/ey.19.15.16

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Akerman, M., Alioto, T., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184. doi:10.1038/nmeth.2714

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452.e17. doi:10.1016/j.cell.2017.10.049

Thorogood, A., and Knoppers, B. M. (2016). Legal interoperability: A sine qua non for international data sharing. *Hum. Genomics* 10.

Wall, J. D., Stawiski, E. W., Ratan, A., Kim, H. L., Kim, C., Gupta, R., et al. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 576, 106–111. doi:10.1038/s41586-019-1793-z

Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., et al. (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.* 8, 2502–2515. doi:10.1038/nprot.2013.150

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi:10.1038/ng.2764

Wilks, C., Zheng, S. C., Chen, F. Y., Charles, R., Solomon, B., Ling, J. P., et al. (2021). recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* 22, 323. doi:10.1186/s13059-021-02533-6

Wilson, E. H., Groom, J. D., Sarfatis, M. C., Ford, S. M., Lidstrom, M. E., and Beck, D. A. (2021). A computational framework for identifying promoter sequences in nonmodel organisms using RNA-seq data sets. *ACS Synth. Biol.* 10, 1394–1405. doi:10.1021/acssynbio.1c00017