

Module 6.3: Learn - Professional Development

Overview

In this section, we will talk about additional things that you will need to consider when publishing results: how authorship is determined for publications, resources used to share genomics data, and the ethics of data sharing.

Authorship

For better or for worse, publications are the currency for researchers to demonstrate their impact in the scientific community. In Module 5, we discussed the impact factor and the prestige of the journal your work is published in, but what also matters is your contribution to the work. Modern scientific research projects are rarely done by one person alone; most projects are accomplished with multiple researchers working collaboratively. Authorship order for a publication is often used as a way to communicate the overall structure of who led the work and how much each author contributed to the final product.

We want to explain how this works to you so that you can set your expectations of what is reasonable to request for authorship based on the work you do for a project. Determining authorship order can sometimes get messy, so we encourage you to discuss a plan for authorship up front so that you know what role you are being expected to play and everyone is on the same page for authorship when the work is done.

Here are the main categories of authorship status and what they mean:

1. **First author:** This is the spot reserved for the person who led the execution of the project. This is the most sought-after position on the paper because papers are typically referred to as “SoAndSo et al” when the paper is cited. If you are looking to use this paper as proof of your credentials when applying to graduate schools, fellowships, or jobs, the committee will give you more credit for publications where you are first

author. In some cases, two or more people can be given first author status if they contributed equally to the work by a symbol like an asterisk next to all the first authors and a note that says that they contributed equally.

2. **Last author:** This is the position of the supervisor or principal investigator that drove the research project and for it. This is the most sought-after position for your research advisors as these are the papers that count the goals. Typically, the last author is the supervisor for the first author. Just as there can be co-first authors, the authors if multiple principal investigators worked together equally on the published work.
3. **Corresponding author:** This person is the one who will be contacted with questions about the work. The last author is a good choice for corresponding author since the work will likely be maintained in their lab after the first author leaves. It can be the first author or anyone who has a deep understanding of the project or will be keeping resources preserved as software packages up to date.
4. **Coauthors:** This is all the other authors in the author list. In most fields, order is listed by relative contribution.
5. **Acknowledgements:** This section of a paper is for people that made small but meaningful contributions to the project, such as proofreading the manuscript or providing general resources like computing support.

Ethics of data sharing

In our research project, we are working with RNA sequencing data to measure gene expression. You have seen us working on to publish results from these placenta samples in other modules. We have included a table containing gene expression data in all samples we included in this study as supplementary material so that other scientists that seek out our data can do other analysis if they like. We did exactly this as we downloaded and reprocessed data from placenta samples to complement the data we collected from full term placentas.

It is important to share your data for several reasons:

1. **Transparency:** Just as we discussed for sharing code, it is important to provide the data so that everyone knows what you did and that you did not falsely manufacture results to mislead people into thinking you made a scientific discovery.
2. **Providing others more resources for the betterment of science and medicine:** Putting your data out there for others to use with their own expertise and methods to increase the value of your work and the potential for it to make a difference.
3. **Ethical imperative:** We have the privilege of living in a wealthy nation and have access to a wide range of resources for large-scale genomics research, but there are many groups out there that do not have access to those resources. By sharing your data, you can cut down the amount of resources needed to generate or test hypotheses.

Sometimes researchers are hesitant to share their data because they want to prevent others from “scooping” the published first because they either have more resources and credibility or because they are working more quickly on projects when the original group is playing the long game with more impactful research that takes longer to do. To share the raw data, but bogart the sample phenotype information so that you can't actually work with it in any many cases, you would have to reach out to the corresponding author of the paper that publishes results from a specific access to the data that you need for your study. In general, it is better to develop specific expertise which leads to reach out to you for more substantive collaborations but share data so others can analyze it properly with their own

This is described more in [the notes for Michael Parker's workshop on the ethics of data sharing](https://www.ncbi.nlm.nih.gov/books/NBK321546/)  [\(https://www.ncbi.nlm.nih.gov/books/NBK321546/\)](https://www.ncbi.nlm.nih.gov/books/NBK321546/).

Databases used to store and distribute primary data

Another method to share data would be to upload it to a public genomics database. It's a bit more work but has the reach of your data and putting it in the standardized format required by the database can help other researchers

Sequenced genomes used by everyone for alignment are maintained and shared by three main groups:

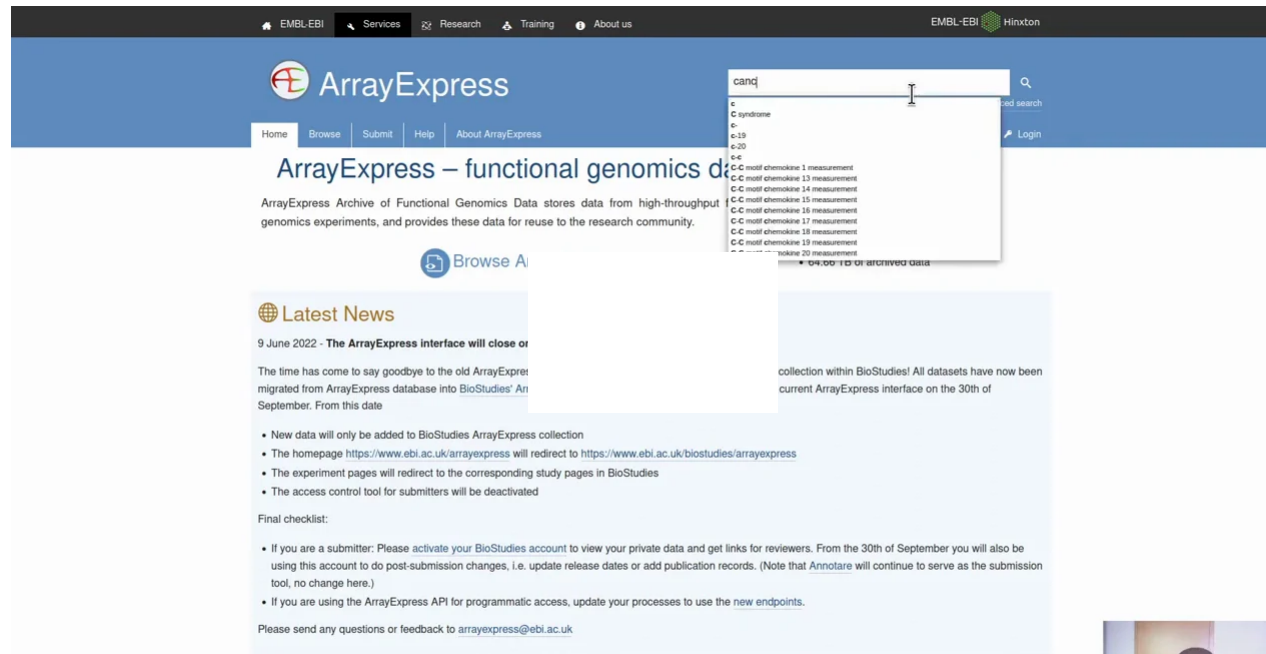
1. [National Center for Biotechnology Information \(NCBI\)](https://www.ncbi.nlm.nih.gov/)  [\(https://www.ncbi.nlm.nih.gov/\)](https://www.ncbi.nlm.nih.gov/) here in the US
2. [European Molecular Biology Laboratory \(EMBL\)](https://www.ebi.ac.uk/)  [\(https://www.ebi.ac.uk/\)](https://www.ebi.ac.uk/)
3. [DNA Databank of Japan \(DDBJ\)](https://www.ddbj.nig.ac.jp/index-e.html)  [\(https://www.ddbj.nig.ac.jp/index-e.html\)](https://www.ddbj.nig.ac.jp/index-e.html)

These are part of an international consortium that assures that the sequences are synced up, so the data inside each group can create different management systems and analysis tools to help users work with the data. Within these three groups, you will see a huge amount of databases associated to the genomes included that give information on sequence, structure, and function of genes.

We are interested in sharing gene expression data associated with specific samples from specific tissues in a specific experimental context. To do this, we would want to choose a public database that is appropriate for the type of data that includes a way to share important phenotypic data associated with the samples. The phenotype data that you can have is limited by patient privacy regulations of the study. When working with genomics data, you will rarely be given access to identifying information about the patients from the study such as name or address to maintain confidentiality during the study. You will remain impartial, but rather you will be given the study specific IDs to refer to specific samples in the data. Never share

information on public databases, but do work with your supervisor to try to share at least the phenotype data nee published results.

This video gives you an overview of how to use some of the more popular gene expression databases such as the Omnibus (GEO), Sequence Read Archive (SRA), and ArrayExpress. **Please pay close attention, and summarize progress report what you learned about common gene expression databases:**



21:55

Video. Biological Databases and downloading gene expression data.

This video introduces several public repositories for downloading gene expression data:

- (1) Gene Expression Omnibus (GEO), (2) Sequence Read Archive (SRA), (3) Array Express
- (4) Expression Atlas, (5) The Cancer Genome Atlas (TCGA), and (6) Genotype Tissue Expression Program (GTEx)

View transcript (<https://canvas.asu.edu/courses/122165/files/55755528?wrap=1>) 
(https://canvas.asu.edu/courses/122165/files/55755528/download?download_frd=1) .

The video discusses how data sets are given accession numbers to uniquely catalog the samples, the platform u expression in those samples, and the researchers and results associated with the generated data. It also shows web interface to find data sets, which help you to know what process others would go through to find your work a search for data that you might be able to use to enhance your own research.

Module 6.3 Additional Resources

- What goes into the sections of papers
 - **[HOW TO WRITE A SCIENTIFIC ARTICLE - PMC](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC34743)** ➞ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC34743>
 - **[Scitable: Scientific Papers](https://www.nature.com/scitable/topicpage/scientific-papers-13815490/)** ➞ <https://www.nature.com/scitable/topicpage/scientific-papers-13815490/>
 - **[Duke University: Scientific Writing: Sections of a Paper](https://guides.mclibrary.duke.edu/scientific)** ➞ <https://guides.mclibrary.duke.edu/scientific>
- Authorship order
 - **[How to Order Author Names and Why That Matters - Wordvice](https://blog.wordvice.com/journal-a)** ➞ <https://blog.wordvice.com/journal-a>
- Types of data access (discussed at the bottom)
 - **[Genomic Data Science Fact Sheet](https://www.genome.gov/about-genomics/fact-sheets/Genomic-Dat)** ➞ <https://www.genome.gov/about-genomics/fact-sheets/Genomic-Dat>
- Genomics databases
 - **[Genomic Data Resources: Challenges and Promises](https://www.nature.com/scitable/topicpage/ger)** ➞ <https://www.nature.com/scitable/topicpage/ger>
[challenges-and-promises-743721/](#)
- Ethics of data access
 - **[The Tension Between Data Sharing and the Protection of Privacy in Genomics Research - PMC](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4337968/)** ➞ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4337968/>