



Using 'collective omics data' for biomedical research training

Damien Chaussabel  and
Darawan Rinchai 
Sidra Medicine, Doha, Qatar

doi:10.1111/imm.12944

Received 9 March 2018; accepted 11 April 2018.

Correspondence: Damien Chaussabel, Sidra Medicine PO BOX 26999 Doha, Qatar.

E-mail: dchaussabel@sidra.org

Senior author: Damien Chaussabel

Summary

Systems-scale molecular profiling data accumulating in public repositories may constitute a useful resource for immunologists. It is for instance likely that information relevant to their chosen line of research be found among the more than 90,000 data series available in the NCBI Gene Expression Omnibus. Such 'collective omics data' may also be employed as source material for training purposes. This is the case when training curricula aim at the development of bioinformatics skills necessary for the analysis, interpretation or visualization of data generated on global scales. But 'collective omics data' may also be reused for training purposes to foster the development of the skills and 'mental habits' underpinning traditional reductionist science approaches. This review describes a small-scale initiative involving investigators, for the most part immunologists, having engaged in a range of training activities relying on 'collective omics data'.

Keywords: bioinformatics; genomics; transcriptomics.

Introduction

First it may be necessary to define what is meant by collective data, and more precisely 'collective omics data'. Collective omics data refers here to the vast collections of datasets generated using high throughput molecular profiling technologies, which are accumulating in public repositories. High-throughput molecular profiling technologies may include for example, next-generation sequencers or mass spectrometers. Public repositories include the NCBI Gene Expression Omnibus (GEO), which holds results from transcriptome profiling studies.¹ It is the resource that was primarily used in the context of the training activities described in this article. GEO comprises over 90 000 series or collections of profiles. One series usually corresponds to one study or publication, although in some cases large studies may be associated with several series. GEO encompasses over two million individual transcriptional profiles. Each of these profiles can include measurements of abundance of up to 50 000 transcripts for an individual sample.

Such data collections constitute valuable training material and have been recognized as a resource for teaching data science skills to biomedical researchers.²⁻⁴ Here the use of publically available large-scale profiling datasets for acquisition of basic skills by biomedical researchers is also discussed. Most of the 'collective omics data' (COD) training activities described below were undertaken by

investigators working at Sidra Medicine in Doha, Qatar. To complete the training an individual would have to publish at least three peer-reviewed papers as first author, one for each of the training modules listed below.

COD1: Reductionist interpretation of collective omics data

COD2: Creation of curated collective omics dataset collections

COD3: Re-analysis of collective omics data on a global scale

A proof of principle was recently obtained with one trainee completing all three modules.⁵⁻⁷ Five other scientists having completed at least one training module.⁸⁻¹²

Each one of the three training modules is presented in more detail below.

COD1: Reductionist investigations using collective omics data

A reductionist approach reduces a complex phenomenon into its simpler or fundamental elements. It has been the mainstay of scientific investigation until the emergence two decades ago of 'systems-scale' or omics profiling approaches, which allow simultaneous measurement and investigation of all elements constituting a given biological system. Such global approaches to scientific investigation are undeniably powerful but should complement rather than substitute to reductionist approaches.

In the first module, COD1, trainees follow a reductionist scientific approach, but rely for this exclusively on data obtained via omics profiling (Fig. 1). The interpretation focuses on one given element of the system, a transcript in the case of GEO data, while ignoring all other measurements. The perspective of the trainee is often unique because the team that generated and deposited the data almost certainly used a global approach for the analysis and interpretation of the data. Furthermore, as GEO comprises datasets from tens of thousands of studies it becomes possible to look up profiles of the same transcript in other datasets for independent validation of the initial finding and for further interpretation.

Assessment of knowledge gaps is the central notion that is explored during COD1 training. It consists of determining whether the knowledge conveyed by the data being examined is novel, and if the answer is yes, then to gain a measure of its potential impact or significance. A proof of principle study was published and can be consulted to illustrate the steps described below.⁶

With collective transcriptomic data serving as a use case:

- 1 A gene presenting with differences in transcript abundance between study or experimental groups is selected in the first step; for instance:

- a An entry from a published list of differentially expressed genes is selected (e.g. that is found in tables published in the manuscript or as a supplement); OR
 - b A transcript presenting changes in abundance across study groups identified while browsing GEO or other web applications (e.g. refs 13–15) is selected; OR
 - c A transcript presenting changes in abundance across study groups identified while carrying out a re-analysis of the entire dataset is selected.
- 2 Whether differences in transcript abundance between study or experimental groups observed for the selected gene could convey novel knowledge is ascertained in the second step.
 - a The body of peer-reviewed biomedical literature associated with the gene selected in step 1, or its product, is identified (e.g. via a PubMed query combining the official gene symbol, name and aliases) AND
 - b Concepts (biological, disease processes, tissues or cell types) relevant to the study or experimental groups for which differences in transcript abundance are identified; AND


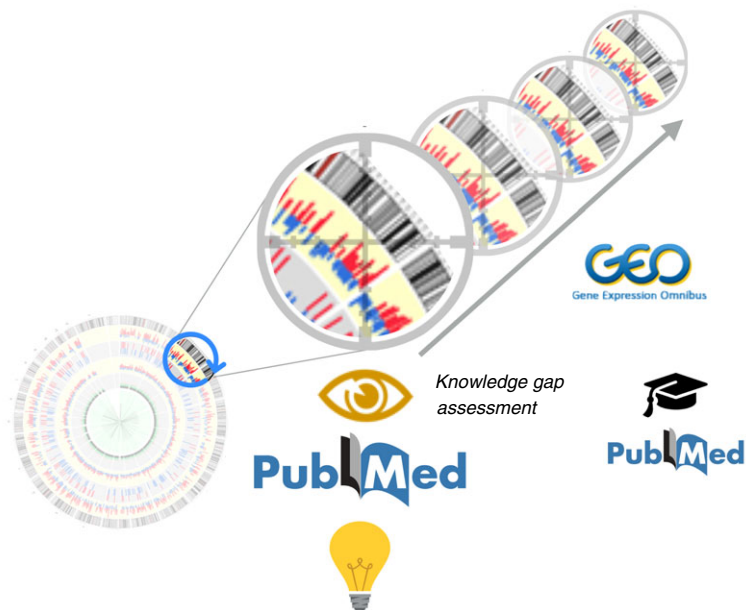
Training module	Activity	Pictogram	References
COD1	Reductionist interpretation of collective omics data		[6]

Figure 1. ‘Collective Omics Data’ training module 1 (COD1). This activity focuses on the development of skills and demonstrable experience in the conduct of reductionist biomedical research approaches. Trainees prioritize and select molecules based on potential for addressing gaps in biomedical knowledge and potential for impact (e.g. extent of clinical or conceptual advances that filling the knowledge gap in question may provide). This assessment is made by examining differences in transcript abundance between experimental groups or study groups among datasets available in the GEO repository. The current state of biomedical knowledge, as is reported in the peer-reviewed biomedical literature, is also taken into account. Expression profiles of the molecule that has been selected for further investigation can be examined in additional GEO datasets for independent validation of the initial finding or hypothesis development and testing. A graphical element used in this figure was adapted from Akula *et al.*¹⁶ Licensed obtained from Springer Nature on May 16, 2018. License number 4350660730175.



- c The overlap between the literature associated with the gene or its product (a) and concepts relevant to the study or experimental groups (b) is ascertained.

For instance: the body of literature associated with a given gene may be identified using a PubMed query combining the official symbol, name and aliases for that gene (Step 2a). This information can be found in the NCBI Entrez Gene database entry for that gene.

Concepts that may be relevant to a study or experimental group (Step 2b) may include the cell type or tissue in which transcript abundance was measured or a disease (for case-control studies) or treatment (for trials or *in vitro* experiments). For example, if measuring levels of expression in B cells of patients with multiple sclerosis pre- and post-administration of interferon- β , concepts to be examined would include: 'Multiple sclerosis', 'B cell or B lymphocyte' and 'Interferon'.

Step 2c ascertains whether absence of overlap can be observed between the literature associated with the selected gene and at least one of the relevant concepts, which would be indicative of a high likelihood of the existence of a knowledge gap.

- 3 When possible, the original observation is validated in one or multiple independent datasets, in the third step.
 - a Datasets in which identical, or similar, study or experimental group can be found are identified (e.g. measurement in disease A versus control in purified mononuclear cells rather than whole blood).
 - b Differences in transcript abundance, or the absence thereof, between study or experimental groups are recorded.
 - c The findings are reported and taken into consideration in the decision to move to the next step.
- 4 A measure of the potential significance or impact of the novel finding is obtained in the fourth step. This may be done in a variety of ways and usually relies on the subject matter expertise, experience or even intuition of the investigator. Associations or convergences will often be sought, with this step involving searching and reading relevant literature, accessing profiles for the molecule in question in other datasets, drawing of models and discussions with colleagues.

Altogether, trainees undertaking this first learning module gain experience with the key steps involved in biomedical knowledge discovery, from identification and assessment of potential knowledge gaps to independent validation of findings, hypothesis formation and testing.

In addition, trainees gain experience with accessing and browsing the data available in public repositories for individual genes or molecules of interest as well as development of advanced PubMed queries, which is another skill that will present a broader utility in most research projects.

COD2: Creation of curated collective omics dataset collections

Collections of public datasets expertly curated based on their quality and relevance to a specific subject matter can prove a useful resource, even if for a small subset of biomedical researchers sharing a similar interest. Notably, such curated dataset collection may also support the COD1 training activities described above.


In the second module, COD2, trainees create curated collective omics dataset collections that are annotated and loaded on a data browsing web application (Fig. 2). Proof of principle is provided by publication of multiple data notes by Sidra investigators.^{5,9-12}

Basic bioinformatics skills are gained as part of this training module, which are more directed to data retrieval and management tasks: dataset curation, quality control, validation, annotation of samples and sample sets, grouping; it also provides an opportunity to learn about organization, annotation and dissemination of data through data-browsing web applications.

With collective transcriptomic data serving here again as a use case:

- 1 The choice of a subject matter expertise is made by the trainee in the first step. The subject should be sufficiently narrow in scope to allow the trainee to acquire knowledge with some degree of depth.
- 2 Relevant 'data series' (data sets) are identified among the 90 000+ that are available in GEO in the second step. This involves selecting relevant keywords to use in GEO queries and restricting searches to species or technology platforms of choice. The results returned by the query are then manually curated to return a list of data sets that are deemed most relevant to the chosen subject matter.
- 3 The data sets are loaded in a data-browsing web application, for instance GXB in the third step (a description of GXB can be found elsewhere,¹⁴ and in the data notes referenced above). Study and sample information is loaded as well (i.e. the metadata). Samples are grouped, rank lists are computed and plots are customized taking experimental or study design into account (e.g. grouping samples by disease status, but also by gender or treatment groups).
- 4 Quality control of the data sets is performed in the fourth step. This step may rely on technical quality metrics, as well as biological markers (e.g. markers associated with gender, cell populations or physiological processes).

A manuscript could be prepared next, which would for instance provide information about: (i) the subject-matter of choice, (ii) the studies from which the data sets originate, (iii) quality control metrics, (iv) use cases, and (v) means by which readers can access the

Training module	Activity	Pictogram	References
COD2	Creation of curated collective omics dataset collections		[5, 8–12]

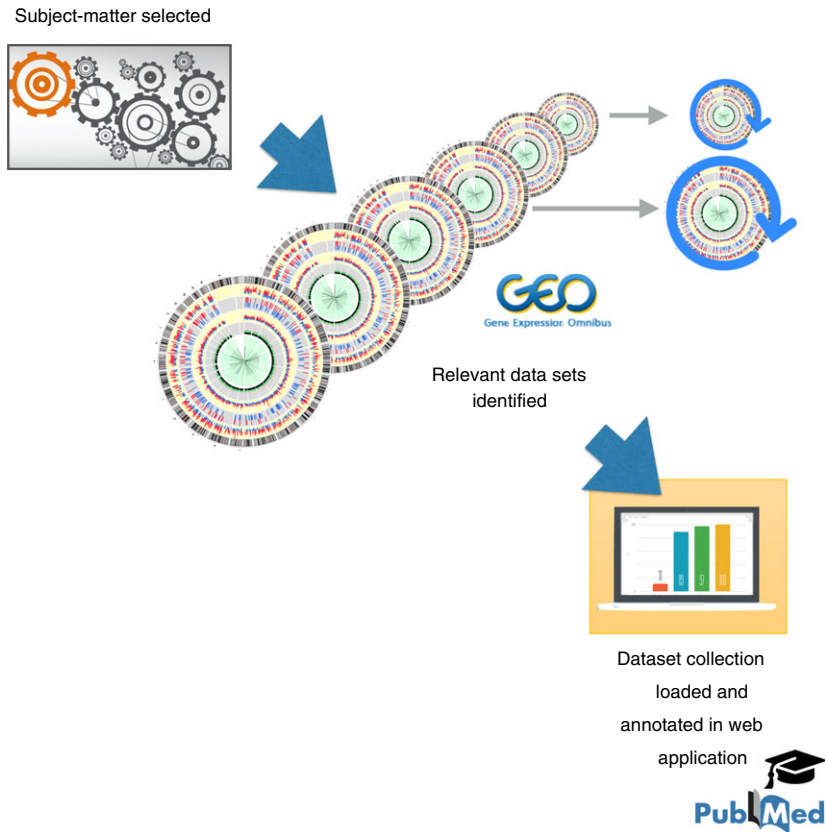


Figure 2. ‘Collective Omics Data’ training module 2 (COD2). This activity focuses on the development of skills and demonstrable experience in identification, retrieval and management of collective omics data. Trainees select a subject matter in a first step. In a second step they identify and retrieve data sets from the > 90 000 that have been deposited in GEO. Next, the selected data sets are loaded on a data-browsing application, such as GXB, which is referenced in the text. Data sets and samples are annotated and organized and this curated dataset compendium is made available publicly along with a peer-reviewed data note describing the resource. A graphical element used in this figure was adapted from Akula *et al.*¹⁶ Licensed obtained from Springer Nature on May 16, 2018. License number 4350660730175.

curated dataset collection using the data-browsing web application.


The dataset collections described in papers published by Sidra scientists covered a wide range of sometimes very narrowly focused subject matters, including: human monocyte immune biology,⁵ development and differentiation of placenta,¹⁰ functional programming of haematopoietic cells in early life,¹¹ embryonic development in healthy individuals and patients with polycystic ovary syndrome,⁹ human immunobiology of human immunodeficiency virus infection⁸ and breast cancer immune classification.¹²

In addition to data retrieval and data management skills, trainees undertaking this first learning module gain experience with various aspects of study or experimental design – through sample grouping and annotation. They also become aware of the availability of public data sets that may be used as a resource as they develop their own research projects.

COD3: Re-analysis of collective omics data on a global scale

Public omics data sets are an obvious choice as training material for the third hands-on learning module described here, COD3, which focuses on acquisition of skills necessary for analysis, interpretation and visualization of global-scale profiling data. As is the case of the other two learning modules, this activity also provides trainees with publication opportunities. These opportunities exist because different analytical strategies implemented by different groups can yield different insights. Indeed, although the validity of different strategies is usually not in question, different approaches to feature selection or functional interpretation will provide different perspectives and lead to different conclusions (Fig. 3).

The analytical approach or approaches employed in the context of COD3 training should be selected based on the desired set of skills that are to be acquired by the trainees.

Training module	Activity	Pictogram	References
COD3	Re-analysis of collective omics data on a global scale		[7]

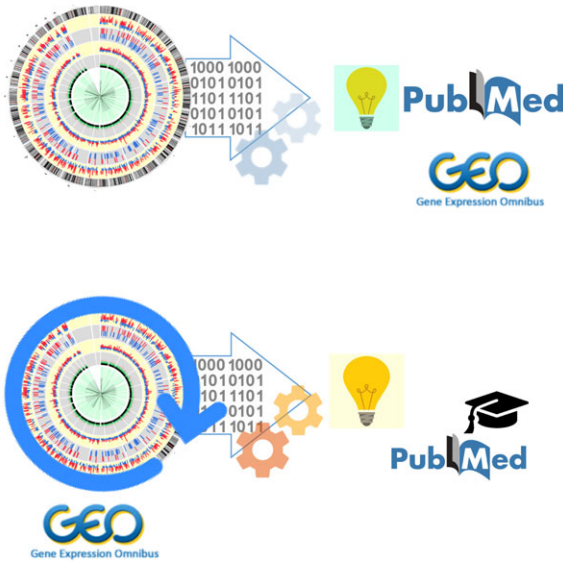


Figure 3. ‘Collective Omics Data’ training module 3 (COD3). This activity focuses on the development of skills and demonstrable experience in global scale analysis, interpretation and visualization of omics data. A dataset can yield different insights when processed, analysed and/or visualized using distinct approaches. As a result, reuse of data sets through this training activity may yield peer-reviewed publications that can serve to demonstrate the proficiency of the trainee with use of such approaches. A graphical element used in this figure was adapted from Akula *et al.*¹⁶ Licensed obtained from Springer Nature on May 16, 2018. License number 4350660730175.

The dataset selected for re-analysis should preferably not have employed the approach or approaches in question if publication is also one of the desired outcomes (i.e. an ‘analytical gap’ should be sought).

In our laboratory, training focuses on blood transcriptome analysis, and makes use of the modular repertoire analysis strategy that we have developed over the years. Rationale and methods for module repertoire construction and downstream analysis and fingerprint visualization have been described in detail previously.¹⁷ The only relevant point here is that it is an approach that was not employed by the primary investigators who conducted and published the seminal study.¹⁸ This gap between the approaches used for primary and secondary analyses is what permitted the identification of distinct but potentially complementary signatures associated with protection conferred by the RTS,S vaccine.⁷

Other alternative analytical approaches may be used as a focus for COD3 training, with data sets available in GEO including cross-sectional as well as longitudinal designs, *in vitro* and *in vivo* treatment. Curation work similar to that described in the context of COD2 but with a constitution of collections based on study design rather than subject matter would be useful in supporting COD3 training activities. Finally, meta-analysis (i.e. the combined analysis of data sets that have been generated independently) provides another means by which original findings might be obtained through re-analysis.

Hands-on training opportunities available through the COD3 training module cover advanced bioinformatics skills, such as Omics data pre-processing and quality control, Data analysis using R programming language, Omics data visualization and interpretation.

Conclusions

Proof of principle has been obtained with a few investigators at Sidra having concluded COD1, COD2 and/or COD3 training with publications in peer-reviewed journals. Workshops between 1 and 3 days in length have also been organized during which participants were presented with use cases and been given the opportunity to carry out hands-on activities associated with the COD1 training module (e.g. accessing and browsing collective omics data sets, building PubMed queries for retrieval of literature associated with a given gene, knowledge gap assessment). The format of each training module is likely to change as the programme is still in the early stages of development and there is room left for optimization at many levels.

The notion of using publically available ‘big data sets’ for training purposes is far from being novel. Education is an important component of the NIH’s Big Data to Knowledge (BD2K) initiative.^{2,3} Funds have been dedicated to the development of open educational resources, curricula or workshops, which are meant for ‘skills development in

biomedical big data science'. COD3, and to some extent COD2, relate to this goal, with data science skills being increasingly important for biomedical investigators to acquire and apply in their research. COD1, however, focuses squarely on 'old-fashioned' reductionist approaches, with the vast public collections of omics data sets simply obviating the need to generate the primary data as source material for training purposes. Another difference lies in the fact that, due to the overabundance of data, the discovery process is somewhat more opportunistic and turned towards hypothesis generation. But once the potential for addressing a gap in biomedical knowledge has been identified that process becomes hypothesis-driven and fairly similar to what most readers would have experienced in their younger days. Furthermore, downstream discovery work would be likely to require, at some point, *de novo* data generation, which was the case for our COD3 proof of principle study.⁷

Acknowledgements

Input from the participants of the training sessions and workshops was instrumental in the shaping of the training modules described in this review. Support for this project was provided in part by the Qatar National Research Fund award NPRP10-0205-170348.

Disclosure

The authors declare having no competing interest.

References

- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M *et al*. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res* 2013; **41** (Database issue):D991–5.
- Dunn MC, Bourne PE. Building the biomedical data science workforce. *PLoS Biol* 2017; **15**:e2003082.
- Garmire LX, Gliske S, Nguyen QC, Chen JH, Nemati S, Van Horn JD *et al*. The training of next generation data scientists in biomedicine. *Pac Symp Biocomput* 2017; **22**:640–5.
- Van Horn JD, Fierro L, Kamdar J, Gordon J, Stewart C, Bhattarai A *et al*. Democratizing data science through data science training. *Pac Symp Biocomput* 2018; **23**:292–303.
- Rinchai D, Boughorbel S, Presnell S, Quinn C, Chaussabel D. A curated compendium of monocyte transcriptome datasets of relevance to human monocyte immunobiology research. *F1000Res* 2016; **5**:291.
- Rinchai D, Kewcharoenwong C, Kessler B, Lertmemongkolchai G, Chaussabel D. Increased abundance of ADAM9 transcripts in the blood is associated with tissue damage. *F1000Res* 2015; **4**:89.
- Rinchai D, Presnell S, Vidal M, Dutta S, Chauhan V, Cavanagh D *et al*. Blood interferon signatures putatively link lack of protection conferred by the RTS,S recombinant malaria vaccine to an antigen-specific IgE response. *F1000Res* 2015; **4**:919.
- Blazkova J, Boughorbel S, Presnell S, Quinn C, Chaussabel D. A curated transcriptome dataset collection to investigate the immunobiology of HIV infection. *F1000Res* 2016; **5**:327.
- Mackeh R, Boughorbel S, Chaussabel D, Kino T. A curated transcriptomic dataset collection relevant to embryonic development associated with *in vitro* fertilization in healthy individuals and patients with polycystic ovary syndrome. *F1000Res* 2017; **6**:181.
- Marr AK, Boughorbel S, Presnell S, Quinn C, Chaussabel D, Kino T. A curated transcriptome dataset collection to investigate the development and differentiation of the human placenta and its associated pathologies. *F1000Res* 2016; **5**:305.
- Rahman M, Boughorbel S, Presnell S, Quinn C, Cugno C, Chaussabel D *et al*. A curated transcriptome dataset collection to investigate the functional programming of human hematopoietic cells in early life. *F1000Res* 2016; **5**:414.
- Roelands J, Decock J, Boughorbel S, Rinchai D, Maccalli C, Ceccarelli M *et al*. A collection of annotated and harmonized human breast cancer transcriptome datasets, including immunologic classification [version 2; referees: 2 approved]. *F1000Res* 2018; **6**:296.
- Ergun A, Doran G, Costello JC, Paik HH, Collins JJ, Mathis D *et al*. Differential splicing across immune system lineages. *Proc Natl Acad Sci U S A* 2013; **110**:14324–9.
- Speake C, Presnell S, Domico K, Zeitner B, Bjork A, Anderson D *et al*. An interactive web application for the dissemination of human systems immunology data. *J Transl Med* 2015; **13**:196.
- Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S *et al*. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 2009; **10**:R130.
- Akula N, Barb J, Jiang X, Wendland JR, Choi KH, Sen SK, *et al*. RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. *Mol Psychiatry*. 2014; **19**:1179–85.
- Chaussabel D, Baldwin N. Democratizing systems immunology with modular transcriptional repertoire analyses. *Nat Rev Immunol* 2014; **14**:271–80.
- Vahey MT, Wang Z, Kester KE, Cummings J, Heppner DG Jr, Nau ME *et al*. Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS,S malaria vaccine. *J Infect Dis* 2010; **201**:580–9.