

Omics analyses, how to navigate through a constant data deluge

Thomas Denecker, Gaëlle Lelandais

▶ **To cite this version:**

Thomas Denecker, Gaëlle Lelandais. Omics analyses, how to navigate through a constant data deluge. 2021. hal-03429581

HAL Id: hal-03429581
https://hal.science/hal-03429581

Preprint submitted on 15 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Omics analyses, how to navigate through a constant data deluge

Thomas Denecker[1] and Gaëlle Lelandais[2,*]

[1] CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France

[2] Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CEA, CNRS, 91198 Gif-sur-Yvette, France

[*]Correspondence: gaelle.lelandais@universite-paris-saclay.fr

Running head: Omics data analysis in functional genomics projects

Keywords: omics, data, information, knowledge, visualization, analysis cycle

SUMMARY **2**

INTRODUCTION **3**

BASICS FOR DATA ANALYSIS **5**

BASICS #1: DIFFERENTIATE "DATA", "INFORMATION" AND "KNOWLEDGE" 5

BASICS #2: RECOGNIZE THE IMPORTANCE OF DATA VISUALIZATION 7

BASICS #3: THINK OF DATA ANALYSIS AS A CYCLE 7

1. *Formulation of the scientific question* *8*
2. *Collection (and cleaning) of available data* *8*
3. *Data exploration and preliminary analyses* *9*
4. *Formulation of statistical hypotheses* *9*
5. *Interpretation and conclusions* *10*

SPECIFIC STATEMENTS IN OMICS DATA ANALYSES **10**

BE AWARE OF CHERRY PICKING 10

BE AWARE OF P-HACKING 11

OUR TEN RECOMMENDED RULES **11**

CONCLUSION **14**

ACKNOWLEDGEMENT **15**

REFERENCES **15**

FIGURE CAPTIONS **18**

Summary

Omics data are very valuable for researchers in biology, but the work required to develop a solid expertise in their analysis contrasts with the rapidity with which the omics technologies evolve. Data accumulate in public databases and despite significant advances in bioinformatics software to integrate them, data analysis remains a burden for those who perform experiments. Beyond the issue of dealing with a very large number of results, we believe that working with omics data requires a change in the way scientific problems are solved. In this chapter, we explain pitfalls and tips we found during our functional genomics projects in yeasts. Our main lesson is that, if applying a protocol does not guarantee a successful project, following simple rules can help to become strategic and intentional, thus avoiding an endless drift into an ocean of possibilities.

Introduction

In 2021, there should be no need to explain that scientists' daily routines are challenged by a constant data deluge. Email messages, conferences, social media, web and smartphone alerts from around the world all bring information about new articles to read, discoveries to understand, datasets to analyze or hypotheses to evaluate. All the time. The rapidity with which data acquisition technologies evolve contrasts with the amount of work required to develop solid expertise in a domain. Researchers are struggling to find a balance between the ever-increasing flow of data they receive and the limited time they have to process and understand the data.

The terminology "data deluge" first appeared in the literature over ten years ago ***(1)***. In biology and life sciences, it has been tightly coupled to the emergence of multi-omics experimental methods, *i.e.* genomics, transcriptomics, proteomics, metabolomics, and more recently epigenomics, glycomics, lipidomics or microbiomics ***(2)***. The common element of these disciplines is the simultaneous acquisition of experimental measurements for several thousands of cellular components. Year after year, the number of technical solutions grows considerably and associated biological data accumulate in public databases. As illustrations, the website "Enseqlopdia" references more than 350 different Next Generation Sequencing (NGS) methods ***(3)*** and the scientific journal Nucleic Acids Research (NAR) publishes descriptions of more than 1600 molecular biology databases ***(4)***. And this is very likely just the tip of the iceberg. How much data has been left behind on hard drives of scientists' computers? It is not surprising that journals like BMC Research Notes have created a new category called "data notes" ***(5)*** to encourage data sharing.

Omics data are very valuable for research in biology, providing incredible opportunities to better understand cell function and the underlying genomic regulatory systems, but like any innovative technology, omics data follows the "hype cycle" ***(6)***. After a peak of inflated expectations in the 2000s (human genome project), many researchers experienced the trough of disillusionment ***(7)***. It became clear that the accumulated omics datasets cannot speak by themselves and thus, the discipline of "data science" emerged as the new hype ***(8)***. Ten years later, despite great perspectives raised by machine

learning and artificial intelligence in medical fields ***(9)***, multi-omics data analysis remains a major bottleneck in most research projects ***(10)***.

We define data analysis as the process of inspecting, cleaning, transforming and modeling a dataset, with the aim of revealing new information. In an omics project, full interpretation of the data that were generated constitutes the major expense ***(11)***. It requires a multidisciplinary team of bioinformaticians, statisticians and biologists to transform the raw data into useful biological knowledge. With this in mind, many integrated computational tools were developed to help researchers with no particular informatic skills to work on their data. The most emblematic is the Galaxy web platform ***(12)***, which is based on a tool shed gathering several thousand software packages. Clearly, technical resources exist today, as well as training programs to learn how to use them ***(13)***. Exhaustive courses in Data Science, Bioinformatics, Statistics and Computational Biology are freely available on MOOC platforms like Coursera ***(14)***, for both beginners and advanced researchers. In this context, we may wonder why data analysis remains such a burden for those who produce the data. We believe that beyond the issue of dealing with a very large number of experimental results, working with omics data requires a change in the way we solve scientific problems.

Most biologists were taught to follow the scientific method, which consists in *i*) making an observation; *ii*) formulating hypotheses to explain this observation; *iii*) designing experiments to test the hypotheses; *iv*) performing the experiments and *v*) drawing conclusions from the experiments which support or reject the hypotheses. This is a difficult process, which can last several weeks, months or even years. Therefore, how can we imagine generalizing such a method in the context of omics data? In the literature, the answer to this question instigates disagreements between researchers over how science should be properly conducted ***(15)***. More pragmatically, if a researcher does not want to spend his/her entire scientific career testing the hypotheses which can be raised after only a single omics experiment, deductive and inductive logic are both indispensable. And this is a new way of thinking especially difficult to acquire, because it is not explicitly described or even understood. In a word, there is no standard procedure.

In this chapter, our aim is to share pitfalls and tips from our experience of functional genomics in yeasts, working with multi-omics datasets ***(16–19)***. Even though our colleagues refer to us as

"bioinformaticians", we have very different scientific backgrounds and histories. Fifteen years separate our first contributions to functional genomics projects ***(20,21)***, fifteen years during which bioinformatics has changed considerably ***(22)***. From a field essentially devoted to the study of sequences (genes and proteins), bioinformatics has become a discipline centered around the data ***(Figure 1)***. We each have a different conception of what "we should be doing" to improve what "we already do". Above all, we have learned by working together that following a protocol does not guarantee a successful project. More than ever, there is no right or wrong answer, hence data analysts must be creative scientists and feel free to think outside the box. The following text is organized into three main sections. In the first one, we present three basics for data analysis. We believe they can provide a foundation for assessment of the concepts that underlie the challenges of any data analysis. In the second section, we remind the reader that "cherry picking" and "p-hacking" are two major issues they must be aware of, especially in omics data analyses. Finally in the third section, we recommend "10 simple rules", which we have experienced to be helpful in our daily work.

Basics for data analysis

Basics #1: differentiate "data", "information" and "knowledge"

Data analysis is the procedure by which new information and knowledge are discovered from the data, and hence, it helps make decisions and highlight conclusions. In this context, the words "data", "information" and "knowledge" are often connected, but it is worth considering that they refer to very different steps and aims in a data analysis ***(23)***.

Data, by definition, represents the starting point of the whole story. The term is plural, equivalent to "facts". Therefore, data are the individual facts, which are still out of any context. In biology, they rely on the raw experimental results or observations. Data can be of different types, *i.e.* numerical or textual, but they can also be images, audio recordings or videos. The diversity of data sources is certainly one of the biggest challenges for their analysis. At this stage, it is important to distinguish structured data from unstructured data ***(Figure 2)***. Structured data are easily organized in databases and can be efficiently retrieved and manipulated. Unstructured data are more inconsistent and complicated to use.

Preliminary work is required to find and/or define their underlying organization and for that, manual verification or curation is often necessary. This makes the process uncertain and time consuming. Whether structured or unstructured, the key idea behind the concept of data, is that they have no meaning.

This is the exact difference with the term “information”. Considered to be singular, information can be seen as a set of data placed in a particular context. Hence, information has a relevance to one or more people and for them, means something. Information can for instance, be that a gene is more expressed when cells are in a particular condition than it is in another condition, or that the overall organization of a network of genes has a particular topology ***(Figure 2)***. Whatever the signification is (simple or more complex), an information arises from inspection of the data, applying procedures decided by the scientist doing the data analysis. Importantly, this implies that multiple information can arise from a single dataset, depending on the processes used to inspect the data. Information therefore represents a great advance in a data analysis, but it should not be confused with the word “knowledge”.

This last idea is a key concept to understand what data analysis really is. Indeed, by definition, knowledge represents an information which is understood. It means that we know why and how the information is obtained, and we can often make predictions based on this understanding. Considering previous examples, the knowledge can be for instance, the existence of a regulatory protein which impacts the transcriptional activity of a set of genes in the studied condition and drives the overall topology of the associated gene network.

Data analysis is a very lexically challenged discipline. The terms “data”, “information” and “knowledge” are often misused or used interchangeably. To limit misunderstanding among collaborators in a project, these three terms must be formally defined and consistently used. Indeed, if one person is expecting new knowledge from an analysis whereas the other is working on collecting new information, it could be difficult to reconcile both expectations. Also, it is our experience that trying to obtain new knowledge without first generating new information is a very risky strategy, a bit like looking for a needle in a haystack.

Basics #2: recognize the importance of data visualization

At all stages of a data analysis, visualization plays a critical role. This is well illustrated by considering Anscombe's quartet ***(24)***. Whereas the four datasets have almost identical descriptive statistics (in term of individual mean, variance and pairwise correlation), they appear to have very different distributions when they are plotted on a graph ***(Figure 3A)***. Graphics are therefore very important for increasing perspectives on a set of observations.

We have often observed that data visualization is underestimated by our collaborators. The main reason is probably that we are so immersed in a world of images that we no longer notice them. But more than a simple process to "show" or to "communicate" ideas, visualization triggers a cognitive process ***(Figure 3B)***. Applied to data, visualization (often referred to as "dataviz") helps to detect potential bias in the collected datasets, to define underlying structures or to anticipate the significance of the arising information. Applied to information (often referred to as "infographics"), visualization helps to integrate heterogeneous results, compare and confront them with current knowledge in a particular field of interest. Finally, applied to knowledge, visualization helps to explain and communicate new understanding of processes. While some visualizations are quick procedures, possibly automated with the use of graphics software, others are very slow. They require deep thought to realistically synthesize and convey a complex message.

Again, visualization is above all a cerebral process. It allows us to form mental images of concepts, supports thinking, and amplifies cognition. Visualization thus gives new perspectives and drives further analyses. This is worth considering for the benefit of a data analysis.

Basics #3: think of data analysis as a cycle

Our research projects have taught us a lesson: data analysis never happens as originally planned. More complicated or time consuming than anticipated, something is always missing. What is the effect of a method parameter on the retrieved list of genes? The genes are not those expected, what shall we do? Should we change raw data normalization? These are examples of questions raised in our analyses of transcriptomics data. Writing this chapter was an opportunity for us to take a step back from our practices and explain the different steps of a more reasoned and efficient strategy.

In that respect, our main insight is that data analysis is not a linear process, but rather a cycle. The challenge for researchers is to avoid getting stuck in a perpetual cycle. We identified five different steps, which are represented in ***Figure 4*** and detailed below.

1. *Formulation of the scientific question*

This is the most important step. A clear formulation of the scientific question helps to delineate the purpose of the data analysis and to anticipate the information that will be obtained and the associated potential issues. A typical mistake is to directly target the creation of new knowledge, forgetting the information intermediate. Even if they are very interdependent notions (see previous section), keeping in mind their specificities is at this step, critical. Information is associated with "data meaning" whereas knowledge is associated with "data understanding". These are very different aims. In case of a transcriptomics data analysis for instance, it can be illustrated as follows. Imagine you are comparing gene expression between two conditions; it is very different to say "I want to know how cells adapt in response to environmental changes?" (vague question, multiple answers can be formulated) than to say "I want to know the set of genes for which normalized gene expression measurements are significantly different, with an associated type I error rate of 1% ?" (precise question associated with classical statistical procedures). Of course, the first question is more attractive from a scientific point of view. The underlying idea is clearly to create new knowledge, the goal for all researchers. But we can anticipate the first question will be very hard to fully answer, especially if several collaborators share different points of view concerning "cell adaptation". There is no such risk with the second goal. Even more, we can anticipate, based on the number of replicates available, the power of the statistical test and hence, our ability to detect genes that are truly differentially expressed ***(25)***. The results may be disappointing, but they will still be results, *i.e.* new information collected that can be associated with other information, as part of a more ambitious scientific project.

2. *Collection (and cleaning) of available data*

Being a data analyst in 2021 is a strong advantage for this step. Many sources of biological data exist and are freely available in public databases ***(4)***. It is thus possible to answer a lot of scientific questions without carrying out new experiments. Of course, the experiments of others can still be considered

imperfect and unsatisfactory, but this is not so important if some new information can be obtained from the data. A typical mistake at this step, is to collect more data than we actually need. It is important to remain focused on the scientific question only. Indeed, with too much data, the risk is to spend more time than necessary on purely technical (informatics) issues. Data cleaning is also a necessary step, whose cost increases with the amount of data collected. On this subject, interesting advice is presented here ***(26)***. To ensure good data quality, it is necessary to check validity, accuracy, completeness and consistency. This takes time but deserves to be taken into account because the quality of cleaning and data collection has a direct impact on the quality of the information, *i.e.* the adage "garbage in, garbage out" ***(27)***. In this context, being specific about what we are trying to accomplish is a great help in deciding what actions to take.

3. Data exploration and preliminary analyses

This step is our favorite. For the first time, the collected dataset is handled without any restraint regarding what is right or wrong to do. It consists in exploring the data, plotting multiple graphs, calculating many parameters, etc. We like the idea that this step is about "getting to know" the collected data. The visualization methods described above are of great interest at this stage, while ensuring the reproducibility of the process is not yet necessary (see next section). This greatly reduces the work of the data analyst and provides very appreciable freedom. A typical mistake is to spend too much time at this step and over-interpret the observations. Indeed, no solid conclusions will be drawn from these explorations and preliminary analyses, the objective of this step is only to develop intuitions, which will help for the following steps.

4. Formulation of statistical hypotheses

This is undoubtedly the most challenging part of the cycle. The problem is that we are going to subject the collected data to more complex calculations, applying procedures that involve mathematical details, often beyond our general understanding. At this step, the hard work is purely mental. It consists in converting the scientific question into relevant statistical hypotheses, prerequisite to using the right methodology. Once done, all the rest is mere computation, which simply requires statistical software. Note that at this step of the data analysis cycle, the issues of repeatability, replicability and

reproducibility must be carefully considered ***(28)***. While the previous exploration step can remain "quick and dirty" ***(29)***, this part of the analysis must be rigorous and well documented ***(30)***. A typical mistake is to confuse "scientific hypotheses" with "statistical hypotheses". Again, data analysis is a very lexically challenged discipline. As an illustration, a scientific hypothesis could be "I think that the function of this gene is required for the adaptation of cells to this environmental change", whereas a derived statistical hypothesis could be "the mean of logFC replicates in RNAseq data for this gene is different from 0". These are two very different statements. We have often encountered colleagues who were already convinced that their scientific hypothesis was true (for very good reasons which we do not question), and thus strongly rejected the results of a statistical analysis that did not go in the expected direction. Such misunderstandings lead to difficult discussions, which could be avoided through a better understanding of statistics ***(31)***. Rejecting a statistical hypothesis, based on a p-value calculation and/or a threshold for type I error rate, does not imply the scientific hypothesis is wrong ***(32)***. But the information cannot simply be ignored, on the contrary, it can help decide the course of further analyses.

5. *Interpretation and conclusions*

Finally, the last step. It is an essential time to describe the results and to interpret them in the general context of the scientific project. Expertise in the scientific field for which the data analysis is performed is very important. Links with already established knowledge are identified, discussed and may lead to new scientific questions. A new cycle of analysis can be performed and, in this way, the scientific project advances one step (or cycle) after another.

Specific statements in omics data analyses

Below, we alert the reader about two important issues, which are specific to the analysis of omics data. Although detailed explanation is beyond the scope of this chapter, we provide references for those who wish to pursue these topics.

Be aware of cherry picking

Over the past ten years, Big Data has revolutionized many areas of the life sciences ***(33)***. But is it an infinite source of new information? The answer is *a priori* yes, if appropriate analytic procedures are

applied. Otherwise, Big Data can become an important source of false discoveries. Indeed, when the mind is motivated to search through a mass of data, it always ends up discovering a path, usually a straight line, to the narrative it initially wished to find. In other words, if we are already convinced that our scientific hypothesis is true (see previous section), there is a risk of retaining from the multiple explorations of a large data set, only the information that supports the hypothesis, while rejecting (consciously or unconsciously) the information that contradicts the initial conviction. This is called "cherry picking" ***(34)***, *i.e.* the action of pointing to individual cases that confirm a particular idea, while ignoring a significant portion of related and similar cases that may contradict that idea. In this context, statistics have an important role to play, helping to discriminate observations that have a high probability of being random from those which, on the contrary, have a low probability.

Be aware of P-hacking

P-hacking relies on an inappropriate manipulation of data which favors the result to be presented as statistically significant ***(35)***. Extensively discussed in the literature (for instance ***(36)***), there are several P-hacking scenarios. In the first one, several statistical tests are performed on the same dataset (for example a parametric test and a nonparametric test), and only the results of the most significant test (with the lowest P-value) are shown. In the second one, observations are added (or removed) from a sample used as statistical test input, until the calculated P-value is below a targeted threshold (generally 0.05). In the last scenario, a single dataset composed of multiple variables is used to test multiple statistical hypotheses until one of them is significant. It is very easy to fall into the P-hacking trap, even for experienced data analysts, especially with P-hacking scenario 3, which is very close to the exploration step of the data analysis cycle described above. This is the reason why this exploration step must be followed by a rigorous validation step. To delve deeper into this subject, we suggest the very comprehensive article of ***(37)***.

Our ten recommended rules

In the last section, we list 10 rules that help us move forward, despite the constant deluge of data. We are not always able to follow them consistently, but they represent a common objective.

1. **Take the time to set your long-term goal and break it into smaller, very specific parts.**

While it is important to dream big, it is also important to define the different stages that will allow you to move forward in the desired direction. Each of these steps must be specific enough to be carried out in a simple and effective way, without too much effort.

2. **Do not bypass elementary analyses.**

Even if there is a temptation to directly apply "hype" methodologies (in 2021, this is AI), simple analyses are still very valuable. Because they are easy to understand, you can stay focused on the data rather than the underlying assumptions that guide the methodology. You also limit the risk of misunderstanding what the method can tell you. Elementary analyses are also interesting for guiding subsequent choices of more complex methodologies. Taking a little time with this can save a lot of time later. If your data is not worth considering for instance, we might not put as much energy into using it. Note here that the choice of methods (simple or complex) must always be guided to achieve the stated objective, no more, no less, hence the importance of the previous rule.

3. **Use unlimited data visualization.**

We explained in previous sections how data visualization can drive our thinking in an analysis. Be careful not to restrict your data explorations to graphics that are automatically and systematically generated by a single tool. To be fully useful, data visualization should be as flexible and broad as possible, allowing data to be viewed from very different angles. Online tools ***(38,39)*** exist to support your creativity. Keep in mind the importance of distinguishing "dataviz" (rapid process used for exploration) and "infographics" (lengthy process to support communication and help other people understand your message).

4. **Use statistics because you need them, not because others tell you to.**

Do not ask a question if you do not want to listen to the answer. This is the best advice to remember when the issue of carrying out a statistical test is discussed. Do we need statistics to convince ourselves or to convince others? We deeply believe that statistics are a powerful and indispensable tool for performing rigorous data analysis. However, poorly mastered or applied for the wrong reasons, their use can lead to borderline situations. "P-hacking" and "Cherry picking" are perfect illustrations (see previous section). Do not hesitate to consult a statistician to verify the accuracy of the analyses and keep

in mind that at the end, it is you and you alone who make the decision, statistics are only one part of your overall thinking.

5. **Choose the right time to discuss your results, neither too early nor too late.**

This is an important advice. Presenting your results (whether they are preliminary or advanced) requires clear thinking. Explaining your motivations, justifying your choices, and expressing your doubts cannot be improvised. We control how a message is sent, but we do not control how it is received. Communication therefore deserves a little preparation. Also, to be able to listen to feedback and comments, and thus identify weak points in the analysis, you cannot be in a hurry because of an impending deadline. Timing is therefore important. Finally, it is essential to take into account your audience. The less people know about your project, the more you will have to explain. This is worth considering if you are expecting the discussion to focus on a specific element of your project.

6. **Listen to the opinions of others but take full responsibility for all decisions.**

Collaboration in a multidisciplinary context can sometimes be difficult. While having different perspectives on the same issue can be very rich, it can also be crippling. At some point, therefore, it is important to be able to decide and move forward. At the end of the study, a critical review process can be applied in order to assess the relevance of the decisions that have been taken and therefore to learn from any errors.

7. **Guarantee the reproducibility of analyses but pay attention to not unnecessarily burden the process.**

The reproducibility of analyses is one of the major challenges of current bioinformatics. However, it can be arduous to test and very constraining in a situation where creativity can make all the difference. This is why we clearly separate two stages of the data analysis cycle. When exploring the data, it is important to work as freely as possible and then, only if the preliminary results are worth it, to implement good reproducibility practices. Keep in mind that the more these practices are mastered, the less the effort required to use them.

8. **Remain constantly intentional regarding where you are in a cycle.**

Working in analysis cycles has been a decisive change of habit for us. Our experience is that performing short cycles (a few days) is most effective.

9. **Take your time, but not too much, perfection is stagnation.**

You understood that to navigate through a constant deluge of data, you must be able to slow down. Not run faster and faster just to stay in the same place, like the Red Queen ***(40)***. Be careful, however, not to stop completely. Doing nothing but reading and thinking, looking for the perfect solution to the problem being addressed is ultimately useless if it does not turn into concrete action.

10. **Remember that being a data analyst is an amazing job!**

Data are nowadays everywhere and allow endless creativity for those who appreciate it. Combining computer science, statistics and biology, omics data analysis is fascinating interdisciplinary work with many attractive perspectives. Isn't data the new (s)oil?

Conclusion

In this chapter, we have shared key ideas and convictions that help us work on a daily basis. They have been developed both during the experiences of our own omics data analyses and in the context of collaborative projects with biologists, informaticians and statisticians. Overwhelmed by the constant flow of data and challenged with the methodological evolution in bioinformatics to follow the FAIR Guiding Principles ***(41)***, we started to feel "stuck", as if our work was never enough. The desire to automate our analyses was very strong at that point and we did make significant progress in that direction. While this has been helpful, it was not enough. Today, we accept that we cannot compete with the pace at which the omics technologies evolve, and we recognize that the burden represented by omics data is much too heavy to be carried alone. Again, data analysis is a complex process for which there is no universal protocol. It requires strong expertise, a lot of curiosity and above all, patience. We believe that developing one's own "know-how" is one key to success. It helps to become strategic and intentional, thus avoiding an endless drift into the ocean of possibilities. Taking full responsibility for the different choices required at each analysis step, is another key to success. It is not easy in a multidisciplinary context, where we often think that others are more legitimate to say that we are proceeding in the right direction. But in practice, we learned that if we become the person in the room who asks the right questions, we greatly help the experts to give us accurate answers. Finally, we like

the idea that data analysis is as much an art as a science[1], there are no right or wrong answers, and that it is ok to have differences in opinion. So, listening, reading, brainstorming and going one step at a time are our main recommendations, and we suggest constantly keeping in mind that if we often overestimate what we can do in a day, we also underestimate what we can do in a year.

Acknowledgement

We thank Linda Sperling for her careful proofreading of the text and her very wise advice, Bertrand Cosson and Pierre Poulain for the stimulating discussions during the DUO training program (www.https://omics-school.net/) and the members of research team "Fungal Epigenomics and Development" (I2BC, Orsay) for their daily trust. From a more personal point of view, we also thank our respective spouses, Céline Chasport (for Thomas) and Benjamin Bussière (for Gaëlle). In recent years, our concerns and questions have often occupied discussions at our family tables. Their listening and their patience were a great support. Finally, we would like to thank Frédéric Devaux, editor-in-chief of this book, for having accepted our proposal to write on this subject and thus for giving us the unique opportunity to publish this text.

References

1. Bell G, Hey T, Szalay A. Computer science. Beyond the data deluge. Science. 6 mars 2009;323(5919):1297-8.
2. Veenstra TD. Omics in Systems Biology: Current Progress and Future Outlook. Proteomics. févr 2021;21(3-4):e2000235.
3. http://enseqlopedia.com/enseqlopedia/
4. Rigden DJ, Fernández XM. The 2021 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Res. 8 janv 2021;49(D1):D1-9.
5. BMC Research Notes [Internet]. BMC Research Notes. [cité 28 juin 2021]. Disponible sur: https://bmcresnotes.biomedcentral.com/data-notes
6. Johnson AT. The technology hype cycle. IEEE Pulse. avr 2015;6(2):50.
7. Chakravarti A. Genomics is not enough. Science. 7 oct 2011;334(6052):15.

[1] https://bookdown.org/rdpeng/exdata/

8. Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. Harv Bus Rev. oct 2012;90(10):70-6, 128.

9. MacEachern SJ, Forkert ND. Machine learning for precision medicine. Genome. 22 oct 2020;1-10.

10. Krassowski M, Das V, Sahu SK, Misra BB. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Front Genet. 2020;11:610798.

11. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol. 25 août 2011;12(8):125.

12. Ostrovsky A, Hillman-Jackson J, Bouvier D, Clements D, Afgan E, Blankenberg D, et al. Using Galaxy to Perform Large-Scale Interactive Data Analyses-An Update. Curr Protoc. févr 2021;1(2):e31.

13. https://usegalaxy.org/training-material/

14. https://www.coursera.org/

15. Defining the scientific method. Editorial. Nat Methods. avr 2009;6(4):237.

16. Denecker T, Zhou Li Y, Fairhead C, Budin K, Camadro J-M, Bolotin-Fukuhara M, et al. Functional networks of co-expressed genes to explore iron homeostasis processes in the pathogenic yeast Candida glabrata. NAR Genomics Bioinforma. juin 2020;2(2):lqaa027.

17. Denecker T, Durand W, Maupetit J, Hébert C, Camadro J-M, Poulain P, et al. Pixel: a content management platform for quantitative omics data. PeerJ. 2019;7:e6623.

18. Denecker T, Lelandais G. Empowering the detection of ChIP-seq « basic peaks » (bPeaks) in small eukaryotic genomes with a web user-interactive interface. BMC Res Notes. 4 oct 2018;11(1):698.

19. Lelandais G, Denecker T, Garcia C, Danila N, Léger T, Camadro J-M. Label-free quantitative proteomics in Candida yeast species: technical and biological replicates to assess data reproducibility. BMC Res Notes. 1 août 2019;12(1):470.

20. Lucau-Danila A, Delaveau T, Lelandais G, Devaux F, Jacq C. Competitive promoter occupancy by two yeast paralogous transcription factors controlling the multidrug resistance phenomenon. J Biol Chem. 26 déc 2003;278(52):52641-50.

21. Hadjadj D, Denecker T, Maric C, Fauchereau F, Baldacci G, Cadoret J-C. Characterization of the replication timing program of 6 human model cell lines. Genomics Data. sept 2016;9:113-7.

22. Bourne PE. Is « bioinformatics » dead? PLoS Biol. mars 2021;19(3):e3001165.

23. Kempe S. The Data – Information – Knowledge Cycle [Internet]. DATAVERSITY. 2013 [cité 28 juin 2021]. Disponible sur: https://www.dataversity.net/the-data-information-knowledge-cycle/

24. Anscombe FJ. Graphs in Statistical Analysis. Am Stat. 1 févr 1973;27(1):17-21.

25. Krzywinski M, Altman N. Power and sample size. Nat Methods. 1 déc 2013;10(12):1139-40.

26. The Ultimate Guide to Data Cleaning | by Omar Elgabry | Towards Data Science [Internet]. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

27. Kilkenny MF, Robinson KM. Data quality: « Garbage in - garbage out ». Health Inf Manag J Health Inf Manag Assoc Aust. sept 2018;47(3):103-5.

28. Lewis J, Breeze CE, Charlesworth J, Maclaren OJ, Cooper J. Where next for the reproducibility agenda in computational biology? BMC Syst Biol. 15 juill 2016;10(1):52.

29. Peng RD. Exploratory Data Analysis with R [Internet]. https://bookdown.org/rdpeng/exdata/

30. Kim Y-M, Poline J-B, Dumas G. Experimenting with reproducibility: a case study of robustness in bioinformatics. GigaScience. 1 juill 2018;7(7).

31. Makin TR, Orban de Xivry J-J. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. eLife. 9 oct 2019;8.

32. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. mars 2019;567(7748):305-7.

33. Fillinger S, de la Garza L, Peltzer A, Kohlbacher O, Nahnsen S. Challenges of big data integration in the life sciences. Anal Bioanal Chem. oct 2019;411(26):6791-800.

34. Morse JM. « Cherry picking »: writing from thin data. Qual Health Res. janv 2010;20(1):3.

35. Teixeira PM. p-Hacking - A call for ethics. Pulmonology. juin 2018;24(3):207-8.

36. Dodson TB. The Problem With P-Hacking. J Oral Maxillofac Surg Off J Am Assoc Oral Maxillofac Surg. mars 2019;77(3):459-60.

37. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol. mars 2015;13(3):e1002106.

38. https://www.r-graph-gallery.com/index.html

39. https://www.python-graph-gallery.com/

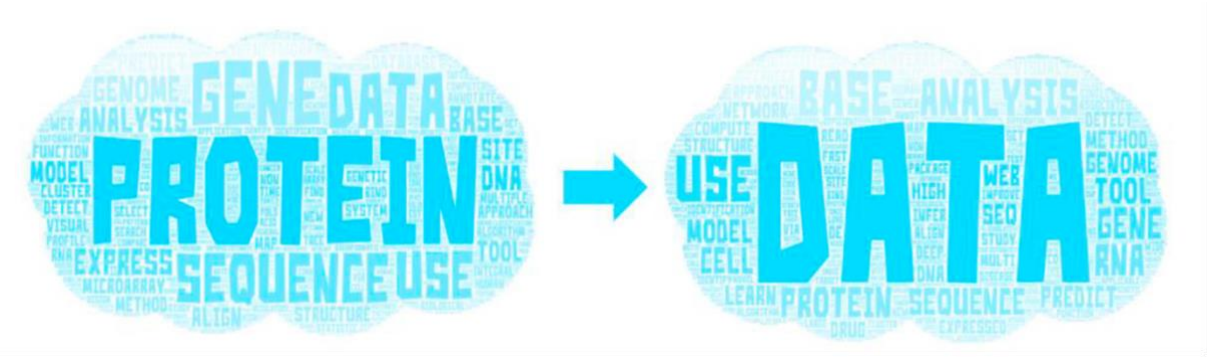
40. Red Queen: Lewis Carroll, Through the Looking Glass, 1871 (or for a modern edition: Lewis Carroll. Alice's Adventures in Wonderland and Through the Looking Glass. London: Oxford University Press, 1971)

41. Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? Eur J Hum Genet EJHG. juill 2018;26(7):931-6.

42. https://www.gastonsanchez.com/

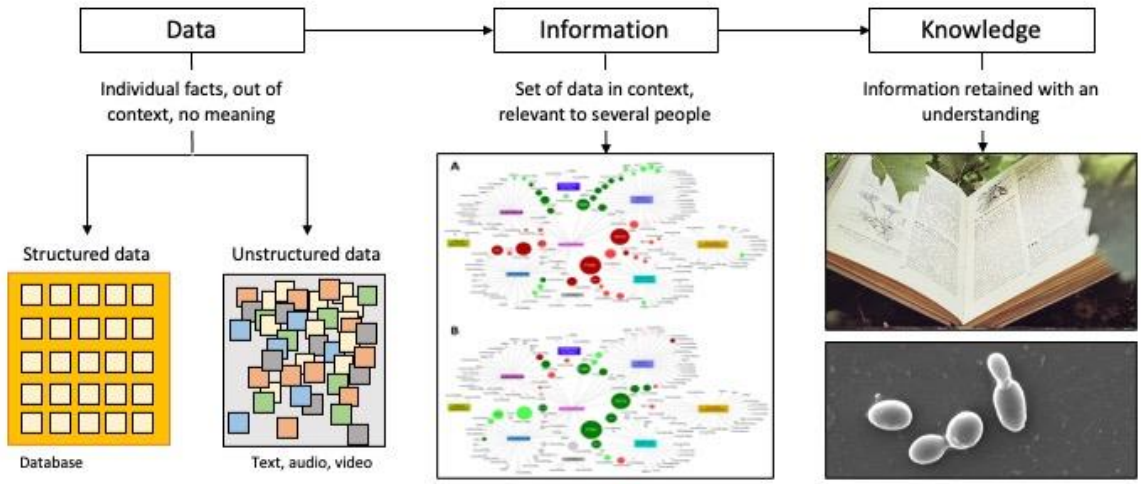
Figures

Figure 1



Word clouds from the titles of articles published in the field of "bioinformatics" between 2000 - 2005 (left) and 2015 - 2020 (right). The list of journals from which the articles were taken is: Bioinformatics, BMC Bioinformatics, Briefings in Bioinformatics and Journal of bioinformatics and computational biology. Note that the larger the word, the more frequently it is used. The period 2000 – 2005 corresponds to the time G. Lelandais prepared her PhD and 2015 – 2020 corresponds to the time T. Denecker prepared his PhD.

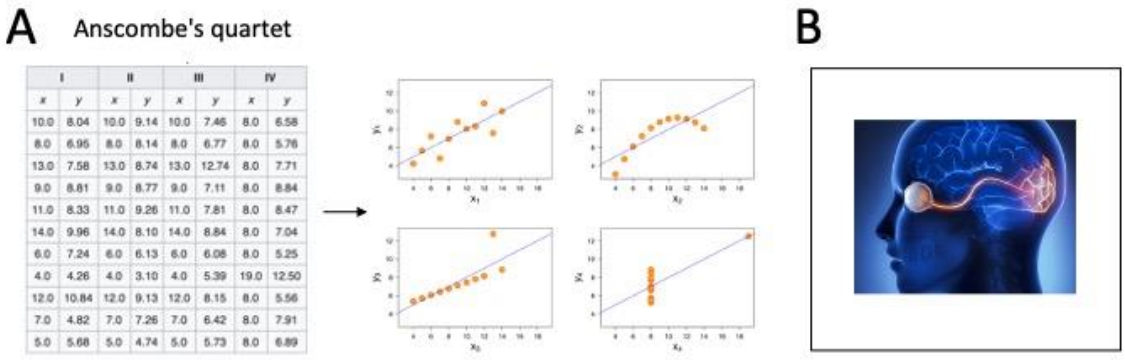
Figure 2



Illustrations of basics #1 for data analysis. The progression from data to information and knowledge is illustrated here. Data are individual facts, out of context and with no meaning. They can be either structured data or unstructured data (see the main text). Information arises from a set of data, which have been placed in a context. Information is therefore relevant to several people. In this picture, two lists of genes for which expression is activated (red color) or repressed (green color) are represented. They are

classified around functional categories. More details for this analysis can be found in ***(16)***. Knowledge is the final stage of any data analysis. It is the most challenging to achieve.

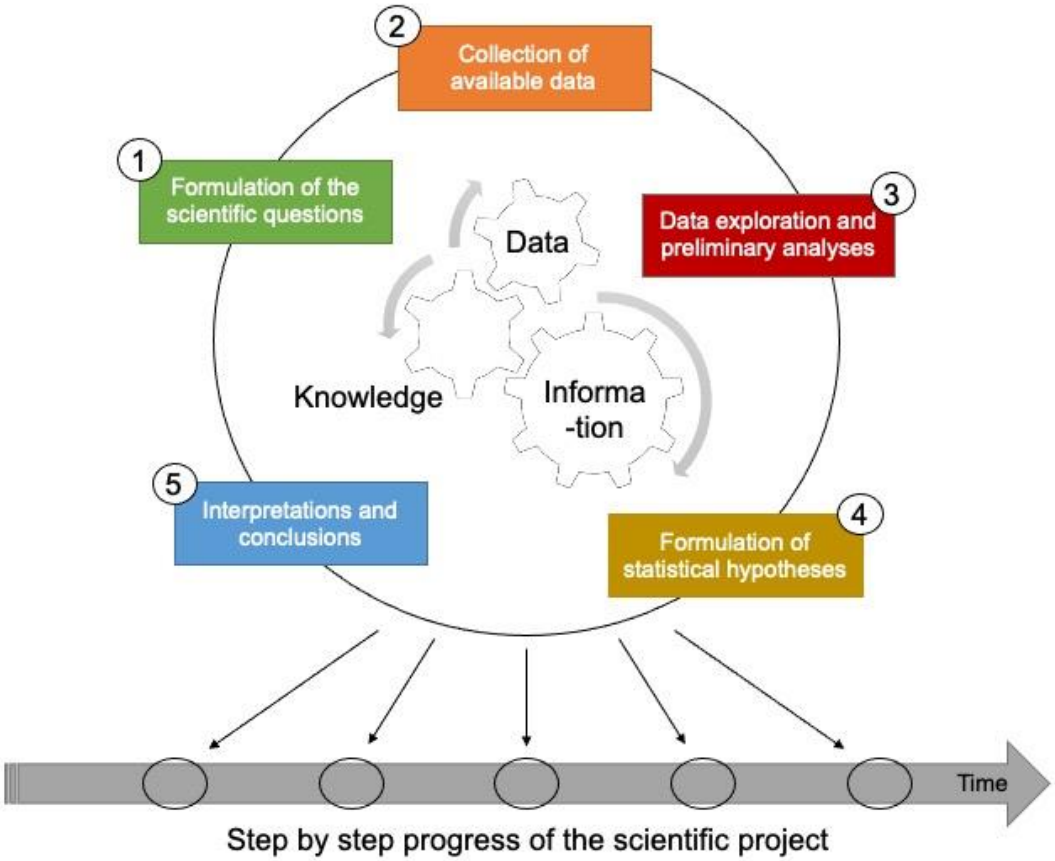
Figure 2



I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Illustrations of basics #2 for data analysis. (A) Presentation of Anscombe's quartet, which is an emblematic example of the necessity of converting numbers into graphics. Whereas all sets of data (x, y) have identical means, standard deviation and pairwise correlations, they are very different in terms of graphical shapes. (B) Illustration of the direct connection between data visualization and cognition ***(42)***.

Figure 3



Representation of the data analysis cycle that underlies the step-by-step progression of a scientific project. Five main steps are required: (1) Formulation of the scientific question; (2) Collection of available data; (3) Data exploration and preliminary analyses; (4) Formulation of statistical hypotheses and (5) Interpretations and conclusions. This last step often opens new perspectives in the project, thus initiating a new cycle of data analysis. The succession of coherent cycles allows a step-by-step progression of the project.