Patterns

CellPress
OPEN ACCESS

People of Data

Why are some plants taller? Researchers on the unveiling of genetic variation associated with complex quantitative phenotypes

Giovanni Marques de Castro,[1] Felipe Campelo,[2] and Francisco Pereira Lobo[3,*]
[1]Nonhuman Primate Reagent Resource, UMASS Chan Medical School, Boston, MA, USA
[2]Aston University, Birmingham, UK
[3]Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
*Correspondence: franciscolobo@ufmg.br
https://doi.org/10.1016/j.patter.2023.100774

Francisco Pereira Lobo, Giovanni Marques de Castro, and Felipe Campelo are part of an international team of collaborators that developed CALANGO, a comparative genomics tool to investigate quantitative genotype-phenotype relationships. Their *Patterns* article highlights how the tool integrates species-centric data to perform genome-wide search and detect genes potentially involved in the emergence of complex quantitative traits across species. Here, they talk about their view of data science, their experience with interdisciplinary research, and the potential applications of their tool.

Would you like to share some information about your background?

Francisco Pereira Lobo: Sure! Looking back at my scientific background, I wonder whether the disciplines that have piqued my interest can be regarded as among the immediate precursors of data science in the field of biology. I began my career in biochemistry, a field that emerges from many interdisciplinary approaches to study living beings at the molecular level. This field also strongly relies on generating, integrating, and storing molecular data and their embedded metadata into searchable databases. In fact, Margaret Dayhoff coordinated some of the earliest attempts to systematically organize molecular biological data, such as atlases and databases of protein structures, back in the 1960s and 1980s. From biochemistry I moved to bioinformatics, where I have been working in evolutionary comparative genomics for the last 20 years. As I analyzed -omics data from viruses, bacteria, and eukaryotes, I came to realize the need for general, first-principles comparative genomics tools to extract biologically relevant insights from the abundance of genomic data available, particularly for identifying genes that may contribute to the emergence of complex phenotypes.

Felipe Campelo: My background is a bit unorthodox for someone working in this field. My first degree was a BSc in electrical engineering from UFMG/Brazil. After that I moved to Japan, where I got an MSc in information science and technology and a PhD in systems science and informatics, working with the development of efficient algorithms for model-based optimization. It was only after finishing my PhD, back in 2009, that I started developing a greater interest in statistical modeling and inference, which eventually led me to work more broadly with data science, which is probably my main area of research today.

Giovanni Marques de Castro: After I finished high school, I first got in the physical and biomolecular sciences BSc from USP, but I realized it wasn't exactly what I wanted, and I dropped out to move to UFSCar, both in Brazil, where I earned a BSc in biotechnology. By the end of my BSc, I started an internship in bioinformatics at LMB-EMBRAPA, a major research institute in Brazil. I decided to join this group after a short bioinformatics course at UNICAMP/Brazil where I met Francisco Lobo, at the time a research scientist at LMB-EMBRAPA. During this internship, KOMODO2 (CALANGO's previous name), was beginning to take shape in the hands of Francisco and Jorge, with whom I share the first authorship of the *Patterns* article,[1] and at the time an MSc student under the supervision of Francisco. I followed my scientific career through an MSc in genetics at UNICAMP/Brazil. Most of my scientific efforts during this time were in RNA-seq analysis, genome assembly, and comparative genomics. Then I worked for 1.5 years as a bioinformatic technician at the same place, contributing to many projects and also helping as a teaching assistant in diverse bioinformatics courses. When Francisco moved from EMBRAPA to his current position as a professor at UFMG, I decided to join his new research group as a PhD student. My main research project was in machine learning, but I also contributed to many other projects during my time at his lab, which included the development of CALANGO.

What is the definition of data science in your opinion? What is a data scientist? Do you self-identify as one?

FPL: Data science is an interdisciplinary field that relies on computer science, statistical methods, and domain expertise to derive new knowledge from the integration of complex and varied datasets. Data scientists typically engage in a variety of tasks, including the collection, storage, processing, analysis, and interpretation of heterogeneous data types, with the ultimate goal of generating new knowledge that is tailored to a specific domain. I definitely identify as a data scientist, as demonstrated by my Twitter bio, which reads: "Human, biologist, (data) scientist."

FC: Francisco makes some pretty good points there, but I would add that data science is a bit more than the exploration of

From left to right: Dr. Giovanni Marques de Castro, Dr. Felipe Campelo, and Dr. Francisco Pereira Lobo.

questions that can be answered from data—this is, after all, something the natural sciences have been doing for hundreds of years, albeit without the analytic power provided by modern computers. Data science is also about exploring and developing new methodologically and statistically sound ways in which data can be used to answer those questions. As a data scientist, I get to work with colleagues from different areas of knowledge to co-develop strategies aimed at extracting useful, novel knowledge from large amounts of complex data.

GMC: In addition to what Francisco and Felipe said, I would complement that data science serves as a means of extracting meaning from patterns and data that might otherwise be neglected. This requires interdisciplinary knowledge to be able to process the data adequately. Data science makes it possible to uncover insights about our reality that are concealed within the data. A data scientist is an individual who is capable of using knowledge of statistics, computing, and other fields to find meaningful information from the data that cannot be easily obtained using traditional pen-and-paper methods. It was challenging to envision myself as a data scientist at first given the diverse skillset required by the field, but after learning from and dealing with so many problems in the field, I do identify as one.

What motivated you to become a data researcher? Is there anyone/anything that helped guide you on your path?

FPL: My passion for computer science began in high school when I first got my hands on a computer in a computer science class back in the 90s. I immediately started creating simple programs to solve mathematical equations. Around the same time, I was introduced to experimental biology through a science club and was fascinated by the diversity of living beings and their biological interactions, especially the increasing role of genetics and biochemistry in understanding life at the molecular level. These interests led me to pursue a career in biology while actively seeking out opportunities to learn more about the computational analysis of biological data, especially comparative genomics. I was also fortunate to have met at the same high school a lifelong friend and collaborator on several ongoing projects at the interface of biology and computation, Felipe Campelo, who critically contributed for CALANGO. In addition, my MSc and PhD supervisor, Dr. Glória Regina Franco, played a crucial role in integrating my early passion for computer science with genomics. She led one of the first bioinformatics research groups in Brazil and encouraged me to pursue my own biological questions while providing me with the freedom to explore and learn. Overall, my love for both computer science and biology, combined with the guidance of many mentors and colleagues, ultimately led me to become a data researcher in comparative genomics.

FC: I have always been passionate about science—I was one of those kids who was always collecting insects, disassembling radios, and getting my fingers burned playing with those old 1980s chemistry kits. I had the privilege of attending one of the best public high schools in Brazil back in the mid-1990s, where I was able to learn a broad range of theoretical and practical knowledge, which further sparked my interest in the sciences. That's also where Francisco and I first crossed paths and started chatting about the similarities and complementarities of our areas of interest—he was already going into biology back then, and I was at the time focused on physics and electricity. Data science came later in my career, as a natural expansion of my interest in statistical modeling and my endless curiosity to discover new, interesting things about the world.

GMC: When I finished high school and a technical course on mechatronics, I made the decision to pursue an interdisciplinary field that would allow me to apply the diverse scientific knowledge I had acquired. This led me to obtain a BSc in biotechnology. By the end of the course, I learned about how much I could do using data that was already collected and was fascinated by this idea, which was my main reason to move to bioinformatics and data science. Furthermore, I was drawn to the utilization of computers and programming in scientific research, which played a significant role in my decision to follow a career in bioinformatics. Meeting Francisco at the end of my BSc led me to start my internship at EMBRAPA, together with my friend Eiji, where I met great scientists who definitely helped me to see how amazing this area was.

Francisco, what is the role of data science in your domain/field? What advancements do you expect in data science in this field over the next 2–3 years?

FPL: I acknowledge that, like astronomy and particle physics, the biological sciences are moving rapidly toward a scenario of data-driven hypothesis due to the genomic revolution caused by decreasing costs of nucleotide sequencing. The heritable information found in genomes is intrinsically digital. Data science plays an increasingly important role in advancing our understanding of many biological processes through the generation of statistically sound, biologically meaningful knowledge from the massive amount of this genomic data and its embedded metadata. The amount

of publicly available biological sequence data already available and expected to be produced in the near future is astonishing. My research group, for instance, does not produce any new sequence data. Instead, we rely on this public data to develop new ways to extract biologically meaningful knowledge. I anticipate that data-driven approaches will become even more prevalent in the future, leading to a high demand for data scientists in the field of biology. I foresee an explosion in the use of data science to provide large-scale functional information to sequence data that remains thus far uncharacterized, with major implications for fields such as conservation biology, medicine, and agriculture.

Felipe, what do you think is the fun part of being a data scientist? And what is your advice for future data scientists?

FC: My work in data science gives me the chance to interact with brilliant colleagues from many different domains and work with them to answer immensely interesting questions in their areas of expertise. Getting to understand these problems and thinking about how to best cast them as data problems (which we can then start solving using statistical and computational tools) is what really excites me. My advice for future data scientists is to remember that data science is not a separate thing from the rest of science. Remain curious and excited, but be skeptical and treat results that look too good to be true with extreme caution. Take the time to develop a good grasp of statistical thinking (which is a great general skill to have in your toolbox) and to learn the main aspects of the problem at hand—it is an essential part of translating the scientific question into a more abstract data problem. Understanding a bit of the scientific problem under study also helps you detect and avoid subtle errors that can creep into your analysis pipeline, such as data leakage arising from dependence structures in the data, which is something that is always present to some extent in biological data.

Now, let's talk about your paper, how did this project you wrote about come to be?

FPL: Our research began with a fundamental question in biology: what genetic elements account for the vast phenotypic diversity observed across species? Answering this question required the modeling and integration of multiple species-centric data types and the development of statistical models that account for the non-independence of species data due to shared ancestry. In addition to these technical challenges, we made a significant effort to ensure our tool was widely accessible and freely available to the scientific community. Collaborations with domain experts were crucial in the development of CALANGO, as well as in the interpretation and validation of our findings. In sum, our research required a multidisciplinary and collaborative effort to develop a tool that we expect to help scientists all across biology to identify genotype-phenotype associations in the large amount of data already available in public databases.

What were the driving forces behind the project?

FPL: We have many sophisticated statistical models readily available to identify genotype-phenotype associations within a single species. These models led to the discovery of many genes associated with several complex genetic disorders and with important traits in agriculture. I previously worked as a research scientist in many projects that used these statistical models to find associations between genetic variation and phenotypic variation in species of agricultural interest. At this time I realized that there is a surprising lack of tools to search for genotype-phenotype associations while considering data from distinct species. This was particularly shocking for me because, as a biologist and a bioinformatician, I knew that the genetic and phenotypic variation observed in a single species is only a very tiny fraction of the overall variability that is present when comparing data from several species. I guess this was the moment when I realized that a tool capable of providing this information would be able to process far more data and find many potentially useful associations. As Giovanni and Felipe mentioned previously in their definitions of data science, our tool allows the analysis of data that would be otherwise neglected due to the lack of lack of proper statistical methods.

GMC: Being able to associate genotypes with phenotypes is a key question in biology and a very interesting one, in my view. Having an amazing interdisciplinary team that could help with developing, testing, and interpreting the results was also a great drive, and for me it was a very important one.

Was there a particular result that surprised you, or did you have a eureka moment? How did you react?

FPL: Plants have an astonishing phenotypic diversity, ranging from annual, herbaceous species to some of tallest and longest-living species. One of our case studies searched for genomic regions whose abundances are associated with the maximum height of flowering plants, a phenotype with more than two orders of magnitude of variation in our dataset. We found evidence that the genes coding for a reproductive system that prevents inbreeding by promoting outcrossing of genetically dissimilar organisms are independently expanded in taller species. From my perspective, the eureka moment definitely happened when we found this previously unknown reproductive strategy of these species. A motif of concern for the long-term survival of taller plants and their ecosystems is their lower evolutionary rates due to their longer reproductive times. We immediately realized that the expansion of this reproductive mechanism may counterbalance the lower evolutionary of taller species rates by increasing the genetic diversity of their offspring through the reproduction of more genetically dissimilar organisms. I remember feeling the rather unique thrill that, I believe, every scientist and every child feel when they realize, from their perspective, something new about the natural world.

How did you come to collaborate, and what's next for both of you and your team? Can we expect more collaborations?

FC: Definitely! I recently hosted a visiting PhD student from Francisco's group here at Aston, working on machine learning approaches for vaccine and epitope discovery.[2] His visit was part of our ongoing collaboration on improved computational methods for the

development of new highly specific diagnostic tests and vaccines against infectious diseases, and we have a number of exciting results to appear in the coming months.

Francisco (the lead author) and I go back a long way, and for many years we had tried to collaborate on different projects at the intersection of computational intelligence, data science, and bioinformatics, but somehow it never seemed to work out. In early 2020, we were chatting about a piece of comparative genomics software he was trying to develop at the time, and it became quite clear that it was something I could help with, so he was kind enough to let me try my hand at refining and refactoring the whole thing. That software package—then under the working title of ''KOMODO2''—went through multiple iterations of development and testing, eventually becoming the CALANGO package that we are presenting in this issue of *Patterns*.

REFERENCES

1. Hongo, J.A., de Castro, G.M., Albuquerque Menezes, A.P., Rios Picorelli, A.C., Martins da Silva, T.T., Imada, E.L., Marchionni, L., Del-Bem, L.-E., Vieira Chaves, A., Almeida, G.M.d.F., et al. (2023). CALANGO: a phylogeny-aware comparative genomics tool for discovering quantitative genotype-phenotype associations across species. Patterns *4*, 100728. https://doi.org/10.1016/j.patter.2023.100728.

2. Ashford, J., Reis-Cunha, J., Lobo, I., Lobo, F., and Campelo, F. (2021). Organism-Specific training improves performance of linear B-Cell epitope prediction. Bioinformatics *37*, 4826–4834. https://doi.org/10.1093/bioinformatics/btab536.

About the authors

Dr. Francisco Pereira Lobo is an assistant professor in genetics at Universidade Federal de Minas Gerais/Brazil (UFMG). He has a BSc in biology, an MSc in biochemistry and immunology, and a PhD in bioinformatics from UFMG. Between 2010 and 2016, he was a research scientist at EMBRAPA, a Brazilian public research company dealing with all aspects of agriculture. During this time he provided bioinformatics expertise to the analysis of -omics data from viruses, bacterial, and eukaryotic species. In 2016, he moved to his current position, where he leads a computational biology research group (LAB, Laboratory of Algorithms in Biology).

Dr. Felipe Campelo is a senior lecturer in computer science at Aston University. He has a BSc in electrical engineering from UFMG/Brazil (2003) and an MSc (information science and technology, 2006) and PhD (systems science and informatics, 2009) from Hokkaido University, Japan. Between 2010 and 2019, he was an assistant/associate professor in UFMG's Department of Electrical Engineering, and in 2019, he moved to the UK to strengthen Aston University's research and teaching in data mining/data science. Dr. Campelo is a member of the IEEE Computational Intelligence Society, the ACM SIGEVO and SIGBIO, and the Foundation for Open Access Statistics.

Dr. Giovanni Marques de Castro is a postdoc scientist at UMASS Chan Medical School at the Nonhuman Primate Reagent Resource. He has a BSc in biotechnology from UFSCar/Brazil (2013), an MSc in genetics and molecular biology from UNICAMP/Brazil (2015), and a PhD in bioinformatics from UFMG/Brazil (2021). Between 2015 and 2016, he worked as a bioinformatic technician at EMBRAPA, and in 2021, he worked for Bio Bureau as a bioinformatician. In 2022, he worked for GISAID as a data curator and software developer.