Journal of Information and Data Management, 2024, 15:1, doi: 10.5753/jidm.2024.2684

NBioinfo: Establishing a Bioinformatics Core in a University-based General Hospital in South Brazil

Mariana Recamonde-Mendoza [**Hospital de Clínicas de Porto Alegre** | ***mmendoza@hcpa.edu.br***]
Gerda Cristal Villalba Silva [**Hospital de Clínicas de Porto Alegre** | ***cristal.villalba@hotmail.com***]
Thayne Woycinck Kowalski [**Hospital de Clínicas de Porto Alegre** | ***tkowalski@hcpa.edu.br***]
Otávio von Ameln Lovison [**Hospital de Clínicas de Porto Alegre** | ***olovison@hcpa.edu.br***]
Rafaela Ramalho Guerra [**Hospital de Clínicas de Porto Alegre** | ***rrguerra@hcpa.edu.br***]
Andreza Francisco Martins [**Hospital de Clínicas de Porto Alegre** | ***andfmartins@hcpa.edu.br***]
Ursula Matte [**Hospital de Clínicas de Porto Alegre** | ***umatte@hcpa.edu.br***]

Núcleo de Bioinformática, Hospital de Clínicas de Porto Alegre, Rua Ramiro Barcelos, 2350, Porto Alegre, RS, 90035-903, Brazil. Full list of authors' information is available at the end of the article.

Received: 16 May 2022 • **Published:** 16 February 2024

Abstract Bioinformatics is an indispensable discipline for current research in life and medical sciences. The increasing volume and complexity of biological data and the growing tendency for open data and data reuse projects have made computer-based analytical tools central to these research fields. However, it is an intrinsic interdisciplinary field with a multitude of skill sets required for using bioinformatics tools or undertaking research toward developing new methods. There is still a lack of skilled human resources to meet the numerous and growing application possibilities, which represents a bottleneck in many research projects. This paper reports our efforts to create the Núcleo de Bioinformática (NBioinfo, or Bioinformatics Core) at the Hospital de Clínicas de Porto Alegre (HCPA), a major public university hospital in Brazil. NBioinfo aims to serve as a hub for research and interaction in Bioinformatics and Computational Biology at HCPA, institutionally developing these areas of knowledge and promoting scientific advances triggered by bioinformatics. We briefly present our research group's history and goals, and describe our activities toward providing HCPA with competencies in these fields. We also describe the scientific and methodological challenges recently faced by our group and the advances promoted by scientific collaborations and research projects developed at NBioinfo.

Keywords: Bioinformatics, Computational Biology, Genomics, Machine Learning, Research Group, Systems Biology

1 Introduction

Bioinformatics is a fundamental research area to deal with the analytical challenges of complex and voluminous data generated in life sciences and make it possible to take full advantage of them. Since the production of the first draft of the human genome sequence by the Human Genome Project in 2000 [Collins *et al.*, 2003], we have witnessed the rapid development of cost-efficient, high-throughput technologies. This technological revolution has allowed new large-scale biology studies to be conducted in the broad range of molecular features involved in organisms' functioning and diseases, from the genome to the entire pool of transcripts and proteins, giving rise to the "omics" disciplines [Hood and Rowen, 2013; Hasin *et al.*, 2017]. The unprecedented speed and scale of biological data generation have fuelled the parallel advance of specialized computational software and methods, not only consolidating bioinformatics as an integral part of biologists' toolkit but also as a fruitful and fast-growing interdisciplinary research field [Gauthier *et al.*, 2019; Mariano *et al.*, 2020].

However, bioinformatics analyses still represent a bottleneck in many research projects despite their well-recognized importance for advancing biomedical research and positively impacting healthcare systems [Barone *et al.*, 2017]. Many challenges have been continuously reported for bioinformatics education and training due to the intrinsic interdisciplinary nature of the field [Attwood *et al.*, 2019]. Often, principal investigators (PIs) in life science research do not have adequate training in bioinformatics, thus lacking a minimum level of understanding from available resources and tools [Attwood *et al.*, 2019]. Such fundamental understanding could help them visualize potential bioinformatics applications in their research projects, enhance the critical analysis of their scientific findings, and interact with bioinformaticians and computational scientists more effectively, especially to prospect new collaborations with researchers in these fields. Other contributing factors are the lack of local expertise in bioinformatics and the limited resources to develop such capacities, mainly in low- and medium-income countries (LMICs) [Aron *et al.*, 2021]. Finally, we point out the several difficulties encountered in establishing successful interdisciplinary interfaces, integrating researchers from the different areas of knowledge involved in the field (*e.g.,* Biology, Medicine, Computer Science, Statistics) around common goals through coordinated efforts.

Previous works have reported a growing gap between data generation capacity and researchers' knowledge about how to use it effectively [Barone *et al.*, 2017]. As technology advances, the bottleneck in data analysis through bioinformatics tools will become even more critical in the coming years, demanding new actions. We have observed this phenomenon at Hospital de Clínicas de Porto Alegre (HCPA), a university-based general hospital in South Brazil. In 2013, HCPA began performing next-generation sequencing and, soon after, microarray analyses. High-throughput genomic analysis was made available to research groups with little or no training in bioinformatics. It soon became apparent that bioinformatics support was necessary to allow groups to leverage their large-scale experimental analysis results. Therefore, a small group of people with bioinformatics training started to collaborate on research projects and educational activities, aiming to disseminate bioinformatics knowledge in the hospital and its local academic community. This initiative evolved and culminated in the creation of the Núcleo de Bioinformática[1] (NBioinfo, or Bioinformatics Core) of HCPA in 2017. NBioinfo aims to act as a hub for research, interaction, and support in Bioinformatics and Computational Biology at HCPA, institutionally developing these areas of knowledge. This paper presents the history and goals of the Nbioinfo at HCPA and summarizes our academic activities and scientific interests. We also discuss the main challenges faced so far, and the prospects envisioned for the coming years.

2 Our History

HCPA is a public institution, part of the network of university hospitals of the Ministério da Educação (MEC), and academically linked to the Universidade Federal do Rio Grande do Sul (UFRGS). Located in the city of Porto Alegre, HCPA is one of the leading medical care centers in the state of Rio Grande do Sul and is also among the major public university hospitals in Brazil. Together with health care and education, research is one of the major institutional priorities, with HCPA being a recognized hub for the generation of scientific and technological knowledge. In this line, the Hospital has an excellent infrastructure, with two Research Centers, Clinical and Experimental, providing facilities totally dedicated to research activities. A survey of research numbers for HCPA in 2021 showed that the institution had 78 research groups certified in the Directory of Research Groups of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and 1392 active research projects.

Given its tradition in scientific research and the need to invest in state-of-the-art technologies to promote cutting-edge research, in 2013, HCPA began performing next-generation sequencing (NGS) analyses at the Experimental Research Center with the Ion Torrent Personal Genome Machine (PGM) platform, and shortly afterward, it acquired an Affymetrix equipment for microarray analyses. Although some of these platforms are accompanied by proprietary software that allow visualizing experimental results and basic statistics in a standard report, they have limited functionalities for data analysis and do not allow the analysis procedures to be tailored to researchers' needs. Moreover, although these software programs may facilitate obtaining basic analysis outputs to be reported in papers and academic works, they are not necessarily understandable for life science researchers who are still not familiar with concepts underlying large-scale biology studies and bioinformatics methods. Therefore, it soon became clear that to enhance the use of the technological investments made by the hospital as an incentive for research development, it would be necessary to provide bioinformatics support to the research groups interested in these analyses.

In 2014, HCPA launched a postdoctoral fellowship call, with bioinformatics as one of the priority topics. Mariana Recamonde-Mendoza was selected as a postdoctoral fellow under the supervision of Prof. Andréia Biolo from the Cardiology Service and the Laboratory of Cardiovascular Research at HCPA. As part of the guiding principles of the institutional postdoctoral program, the postdoctoral research activities should include a minimum workload of 12 hours per week to provide scientific support to the research projects linked registered at HCPA. Thus, Mariana Recamonde-Mendoza soon started interacting with several research groups, collaborating in planning, running, or interpreting bioinformatics analyses necessary for their studies. During this period, an interaction started between Mariana Recamonde-Mendoza and Ursula Matte, who at the time served as head of the Unidade de Análises Moleculares e de Proteínas (UAMP) of the Experimental Research Center.

From 2014 to 2016, a small network of researchers and graduate students with experience in bioinformatics was established at the Experimental Research Center, led by Mariana Recamonde-Mendoza and Ursula Matte. This network resulted in some scientific collaborations with different research groups and services of the HCPA, which were initially focused on assisting analyses of large-scale studies produced at the institution, such as variants analysis from DNA sequencing and analysis of transcriptomes by microarray. However, the growing number of publicly available high-quality genomic datasets, the increasingly core role taken by bioinformatics in scientific publications, and the results of the network's activities for disseminating knowledge, such as scientific seminars, soon led to increased demands for technical and scientific support in bioinformatics.

In 2016, Mariana-Recamonde Mendoza and Ursula Matte, together with Silvia Olabarriaga (AMC e-Science), Marina Siebert (HCPA), and Delva Leão (former graduate student at UFRGS), organized a short-term course entitled Bioinformatics for the Health Sector, offered to researchers and students from postgraduate programs linked to HCPA and UFRGS. The participants attended 15 hours of theoretical sessions on concepts related to NGS, variant analysis, differential expression analysis, and systems biology. The course's main goal was to introduce concepts and tools useful for life science researchers, encouraging further exploration of *in silico* approaches in research projects developed at HCPA. An important outcome of this course was the great interest manifested by participants for more institutional activities and initiatives related to bioinformatics. This feedback from the local academic community and the clear need to encourage the use of state-of-the-art methodologies to empower the research developed at the hospital have motivated the beginning of a dialogue on strategies for institutionalizing bioinformatics at the HCPA.

In 2016, Mariana Recamonde-Mendoza became an Associate Professor in the Institute of Informatics of the UFRGS,

[1] `https://sites.google.com/hcpa.edu.br/nucleodebioinformatica-hcpa`

and in 2017 she started an appointment as Research Professor at HCPA. Ursula Matte was a Researcher at HCPA between 2001 and 2014, and since 2014 she is an Associate Professor in the Genetics Department of UFRGS and a Research Professor at HCPA. In close collaboration, both researchers took the first steps towards implementing an institutional bioinformatics core in 2017, with the support from the Grupo de Pesquisa e Pós-Graduação (GPPG) of HCPA, which was coordinated by Prof. Patricia Ashton-Prolla and is responsible for managing the scientific and technological research activities in the institution. Initially, the Bioinformatics Core mainly was a "virtual lab", with some computers allocated in a shared laboratory in the Experimental Research Center, thus lacking a dedicated physical area. The group conducted many training and scientific activities in 2017 and 2018, starting with two editions of a 50-hour bioinformatics training course that combined theory and practice and covered topics from Linux and R programming to Metagenomics and Transcriptome analysis, and Systems Biology. These courses were offered to researchers, students, and professionals, and had a critical role in expanding the networking at the hospital and attracting new people to work with the group, such as Gerda Cristal Villalba Silva and Thayne Woycinck Kowalski, who later became research members of the Bioinformatics Core. In this period, we also provided support in analyzing and interpreting large-scale data obtained from biological systems for researchers who did not have such skills in their research groups.

In 2019, the creation of the NBioinfo was formally proposed within the Hospital as a 5-year institutional development project and as a research group certified in the CNPq Directory of Research Groups, both coordinated by the researchers Mariana Recamonde-Mendoza and Ursula Matte. After analyzing existing demands at the hospital and gaps in the local bioinformatics community, the main research topics defined for NBioinfo's activities were: omics data analysis, variant analysis, metagenomics, systems biology, and machine learning for biological data. The group expanded in 2020, with Andreza Martins joining the NBioinfo to strengthen the team's capacities and enhance its skills in Metagenomics. Andreza Martins has been an Associate Professor in the Department of Microbiology, Immunology and Parasitology of UFRGS since 2014 and a Research Professor at HCPA since 2020. In 2021, the NBioinfo became a shared facility linked to the Diretoria de Pesquisa (former GPPG) of HCPA, receiving a physical laboratory to conduct its activities. This achievement has certified NBioinfo's role in serving HCPA's community with bioinformatics expertise through training, consulting, or scientific collaborations.

The unique feature of the group is its truly interdisciplinary composition, as is to be expected in bioinformatics. However, establishing and consolidating these links between researchers from different areas of knowledge is not a trivial process and is one of the main challenges faced in creating bioinformatics research groups or facilities. The group's PIs, Mariana Recamonde-Mendoza, Ursula Matte, and Andreza Martins, have backgrounds in Computer Science, Genetics and Molecular Biology, and Microbiology, respectively. We also note that besides the diversity among areas of knowledge, the PIs form an intergenerational team, a characteristic that is increasingly pointed out as positive for the productivity and innovative potential of research groups. NBioinfo has broad participation of students at undergraduate and graduate levels linked to courses in Computer Science, Biological Sciences, and Health Sciences. Students have the opportunity of a very interdisciplinary training, as they are encouraged to get involved in consulting and collaborative activities promoted by NBioinfo.

Since 2016, NBioinfo had the collaboration of several postdoctoral research fellows who have helped both in the development of training and research activities promoted by the group, as well as in the interface with other researchers and services of the HCPA, namely Leandro de Mattos, Tiago Falcon, and more recently, Gerda Cristal Villalba Silva and Giovanna Giudicelli. Currently, the NBioinfo also counts on the participation of Thayne Woycinck Kowalski, a long-time collaborator who became a research member of the group. For more information about the members of the Bioinformatics Core, access the group's page in the CNPq Directory of Research Groups[2].

3 Goals and Academic Activities

We understand that Bioinformatics is a strategic field to promote scientific advances, especially in limited-resources scenarios as it is the recent research funding crisis in Brazil. Thus, NBioinfo's main goal is to provide the HCPA with competencies in the field of Bioinformatics and encourage its insertion into the institution's research projects. To achieve this goal, we recognize the importance of developing local human resources and awareness of the role and potential of Bioinformatics in our scientific community, and to continuously expand our group's expertise through active participation in research projects. Thus, the objectives guiding the activities of the NBioinfo are:

- Development of fundamental and applied research in Bioinformatics and Computational Biology, contributing to the advance of these research fields;
- Training of qualified human resources in Bioinformatics and Computational Biology;
- Assistance to researchers who need Bioinformatics in their scientific workflow through research collaborations or consulting;
- Production and dissemination of scientific and technological knowledge.

NBioinfo has its activities concentrated on three pillars: education & training, interaction & support, and research (Figure 1). These activities aim to develop bioinformatics interest and capacity at the institutional and regional levels, thus helping mitigate the important limiting factor in the field related to the lack of skilled human resources. Also, as a group born out of the union of efforts of three PIs and their research groups, NBioinfo has scientific and technological research as a primary activity. We aim to catalyze the discovery of innovative and scientifically relevant findings through the toolkit provided by bioinformatics and contribute to the

[2] http://dgp.cnpq.br/dgp/espelhogrupo/7186574396168654

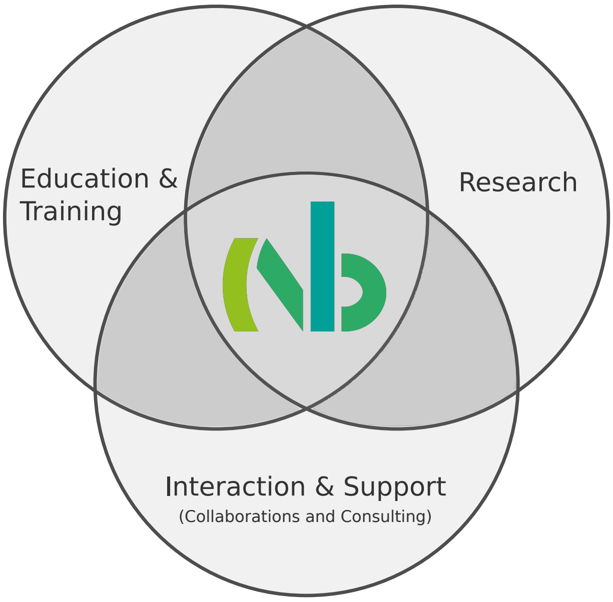

Figure 1. The three main pillars of NBioinfo's activities.

advance of the field itself. In what follows we describe our initiatives in each of the three pillars.

3.1 Education & Training

Bioinformatics skills are increasingly central to life sciences, but there is still a curricular gap in these courses regarding proper integration of statistical, mathematical, and computational training, even at the basic level. Relatively few research and higher education institutions in Brazil have incorporated bioinformatics into their academic curricula, as observed in other LMICs [Aron *et al.*, 2021], or host bioinformatics research groups. Such skills are relevant to allow students to develop necessary computational and quantitative skills and face the current big data era [Sayres *et al.*, 2018]. In contrast, computer science students often do not have the biological background needed to properly tackle biomedical scientific problems through programming. As a result, conducting data analyses and developing research in such an interdisciplinary field as bioinformatics becomes challenging due to the scarce availability of experts with the required skills and experience. Indeed, previous publications [Dragon *et al.*, 2020] have pointed out the challenges of limited personnel for the workload and the need to cover such a broad range of expertise underlying bioinformatics projects.

Since the first steps were taken by our group, when establishing a bioinformatics core at HCPA was still an idea emerging from a collaborative network, and NBioinfo did not exist as a physical laboratory, we have committed to the development of human resources in our local community. The PIs have contributed to the formation of several students at the undergraduate and graduate levels, who developed skills in bioinformatics by conducting their own projects or contributing to other research projects supported by NBioinfo. Besides, we dedicated efforts to provide training programs in a set of core bioinformatics competencies identified as relevant for HCPA's community. We promoted theoretical courses to disseminate how bioinformatics could be a useful tool in research projects developed at the institution and arouse the interest of the HCPA community in this area of research, as well as theoretical-practical courses in bioinformatics to empower more people to perform some commonly demanded analyses. These courses were offered to students, researchers, and professionals from various branches of life and medical sciences, prioritizing people affiliated with HCPA and UFRGS institutions. About 165 people were enrolled in these capacity-building courses, which we consider to have achieved the proposed objective, as many participants became self-sufficient in analyzing their data, and others later joined NBioinfo as associate members, going from students to researchers providing analytical support to other research projects.

We also continuously promote bioinformatics training and knowledge by actively organizing or participating in scientific seminars and events and sharing our expertise and experience through lectures, talks, or roundtable discussions. Among the several initiatives supported by our group, we highlight the BioIn4Girls event organized by NBioinfo in 2020. BioIn4Girls was an online event motivated by the usual underrepresentation of women as speakers and presenters in bioinformatics events. Bioin4Girls featured talks spanning seven major topics in bioinformatics, all given by female researchers. Besides, 100% of the steering committee members were women, and we sought to promote ethnicity and gender identity diversity among speakers and organizers. Talks were given in Portuguese and broadcasted by YouTube and are still available on the event's channel[3]. The event had 1705 registered attendees from 26 states of Brazil and 18 countries. As of March 31st, 2022, the BioIn4Girls' Youtube channel has more than one thousand subscribers, and the videos from the ten days of the event have 1.2k to 2.7k views each. To the best of our knowledge, BioIn4Girls was the first symposium on bioinformatics in Brazil to feature women exclusively as speakers, giving greater visibility to brilliant women researchers and their outstanding contributions to bioinformatics.

All these activities aim to contribute to a community-building process involving bioinformatics users and scientists, strengthening bioinformatics' capacities at the institutional and regional levels.

3.2 Interaction & Support

Bioinformatics is a broad field with a wide range of expertise involved, presenting possible applications in various knowledge areas and research problems. Consequently, the range of audiences who are potential users of bioinformatics tools and resources is large and diverse, with varied need profiles. In a healthcare organization like HCPA, it is natural that many research projects may greatly benefit from bioinformatics tools, whereas few groups have the expertise and independence to explore it. Moreover, while some groups may be interested in training their members to run analyses in an applied context, others may be more interested in seeking specialized assistance for this stage of their projects.

Thus, NBioinfo aims to serve as a reference for bioinformatics in HCPA, not only for promoting training activi-

[3] https://www.youtube.com/c/BioIn4Girls

ties, but also for assisting researchers from the institution, UFRGS, and eventually from external groups, in scientific matters. Our goal is to be the first point of contact within our local community, where researchers can seek assistance in planning and executing the analysis of their data and where we can stimulate cross-disciplinary interactions among institutional researchers, increasing the cooperation between the experts of the "wet lab" and "dry lab" areas. Support is provided on five main topics (omics data analysis, variants analysis, metagenomics, systems biology, and machine learning) and in all levels of investigation, from the study conception to data analysis and interpretation and results communication.

In this sense, NBioinfo operates under two forms of support to research groups: consulting services and scientific collaborations. Consultation is defined here as any guidance or assistance with methodological aspects, methods, pipelines, tools, or resources so that the researchers or their collaborators can perform bioinformatics analysis of their own. These are usually short-term interactions that aim to handle punctual issues or doubts in the course of research. Collaboration, on the other hand, is defined as the active and direct participation of NBioinfo's members in the research project, bringing significant contributions to components of this process: writing the project proposal; planning, running, and interpreting data analyses; developing customized computational methods or pipelines, or writing/revising manuscripts. This is usually a long-term interaction that often culminates in joint publications following the general principles for authorship criteria.

Initial consultation requests are submitted through a structured webform available on the HCPA's website, in which the researcher informs, among others, the topic of interest, types of analysis desired, current stage of the project, and what kind of assistance they expect from NBioinfo - either consultation or collaboration. Criteria for attribution of authorship resulting from NBioinfo's support are provided in the webform to clarify under what conditions our assistance is considered a significant contribution to the design, implementation, and analysis or presentation of the study. In cases where these criteria are not met, but NBioinfo's support was relevant for the study, we orient users to recognize NBioinfo's assistance by formal mention in the paper acknowledgments section. Very commonly, initial contacts for consulting services end up turning into collaborative projects as the conversations and ideas evolve. Consultation is free of charge for researchers and demands from academic external groups are limited to our response capacity.

Since 2018, NBioinfo has provided technical or scientific assistance to more than 80 research projects. The first interactions, between 2018 and 2020, were mainly focused on analysis of omics data and of NGS sequencing data. The HCPA offers a well-equipped technology park to run molecular analyses, through the Unidade de Pesquisa Laboratorial (UPL), including NGS sequencing with Ion S5, Ion PGM, and Thermo Fischer platforms, and microarray analysis with an Affymetrix 3000 7G scanner. The facility also counts on trained personnel to run these analyses and provide technical support. Nonetheless, microarray reagents and assays costs are still high and inaccessible to many research projects. Thus, most of the projects in the scope of omics data supported by NBioinfo aimed to leverage public data repositories, such as Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). NGS sequencing analyses were mainly requested for metagenomic approaches over 16S rDNA sequencing. More recently, we observed an increasing demand for metagenomics support, probably due to the lower costs of this analysis compared to microarray experiments and the various fields in which it has been successfully explored. Since 2021, from the 32 projects for which support was requested from our group, 53% were related to metagenomics.

We briefly exemplify some interactions with other research groups, which, inspired by their peers' publications, were elaborating ideas for bioinformatics analysis in their projects but did not have, at the moment, the knowledge or skills necessary to execute them. NBioinfo has contributed to a functional investigation of relevant genes or miRNAs selected through systematic reviews by researchers in our institution. In one study, we applied network and functional enrichment analyses to analyze genes that were candidate prognostic biomarkers for ovarian response in controlled ovarian stimulation [Eisele *et al*., 2021], while in another, we used bioinformatics tools to find the genes and biological pathways modulated by a subset of miRNAs consistently dysregulated in diabetic kidney disease [Assmann *et al*., 2018a]. We have also collaborated with studies that explored public omics databases to analyze differential expression or methylation patterns in specific conditions, such as breast cancer [Goemann *et al*., 2020] and COVID-19 [Kristem *et al*., 2021]. These *in silico* analyses were run either to complement experimental studies with samples collected at the hospital as a means to validate the findings or to test scientific hypotheses when the groups could not collect the appropriate sample data at the hospital. In all these interactions, NBioinfo was a key collaborator to enable new insights to be obtained through bioinformatics analyses.

Currently, the main bottleneck for consulting services is the limited analytical capacity of our staff for the increasing workload faced by our group. Although we intend to assist as many research groups as possible, projects are accommodated on a first-come, first-served basis, and the researchers' analysis deadlines may not be met by our team's availability. Occasionally, consultation requests were not met because they covered skills outside the group's current competencies.

3.3 Research

NBioinfo works both in fundamental and applied research in Bioinformatics and Computational Biology. Our interdisciplinary team aims to develop computational methods and tools for analyzing genomic and clinical data to help researchers shape, answer and evaluate complex, multidisciplinary questions about biological phenomena and pathophysiological mechanisms. We are engaged in research projects that aim to advance the field of bioinformatics by proposing new methods or enhanced analytical pipelines and to further bioinformatics application to other fields by developing solutions and technological innovation in contexts that demand specialized knowledge in biology and computational methods. Our main research topics will be presented in detail in Section 4.

Research is conducted in projects coordinated and executed by NBioinfo, as well as in collaboration with other research groups from HCPA, UFRGS, or external institutions. HCPA provides an interesting ecosystem in which scientific or technological research and health care activities were already successfully integrated. This environment provides a fertile ground to encourage translational bioinformatics, and NBioinfo aims to explore these frontiers in its scientific activities. Many projects developed by the group are in close interaction with life and health scientists and aim to exploit our analytical skills and the toolkit provided by Bioinformatics to allow researchers to better leverage their time, resources, and data. Common examples of scientific collaborations established by NBioinfo are with (i) projects that start with a more broad (and sometimes exploratory) research question to which we apply bioinformatics to refine hypotheses, analyzing public databases and resources and helping the investigators in selecting findings for a focused experimental investigation; and (ii) projects in which (experimental) results were already obtained and we apply bioinformatics analyses to further investigate these findings, helping in their interpretation by putting their results into a broader context or by seeking additional supporting biological evidence through integrative analysis of several related databases.

In Brazil, a limited number of bioinformatics groups or laboratories have such a close proximity to a hospital environment or clinical groups. This gives us the advantage of being near to many research groups developing high-impact research with great potential to positively influence healthcare and people well-being, who can benefit from bioinformatics to achieve even more meaningful results. Since the start of NBioinfo's activities, we have established collaborations with a number of services, laboratories and research groups from HCPA, sharing publications with researchers from the Cardiology Service, Endocrinology Service, Gynecology and Obstetrics Service, Digestive Surgery Service, Psychiatric Service, Laboratory of Genomic Medicine, among others. We also had publications with external collaborators, including international groups, like Technical University of Munich, North Carolina State University, and University of Texas at Austin - MD Anderson Cancer Hospital.

NBioinfo published a total of 37 papers in international peer-reviewed journals and 4 preprints in recognized preprint repositories from 2017 to 2022. Our group also participated in several national and international conferences with oral or poster presentations. Projects coordinated by the PIs often receive funding from state and federal funding agencies, as well as from the Hospital de Clínicas de Porto Alegre Research Fund (FIPE). Finally, we note that the group's PIs are CNPq Research Productivity Fellows (PQ) in their areas of knowledge: Ursula Matte is PQ-1D in Genetics, Andreza Martins is PQ-2 in Agricultural Science, and Mariana Recamonde-Mendoza is PQ-2 in Computer Science, demonstrating the continued impact and high-quality of their scientific production and other academic activities.

4 Research Areas

NBioinfo has concentrated its scientific efforts in five main research areas: variants analysis, omics data analysis, metagenomics, systems biology, and machine learning. In what follows, we briefly present these areas, with focus in the scientific problems we are currently working on and on challenges that we intend to pursue in the future.

4.1 Genome and Variant Analysis

High-throughput sequencing techniques, which have been called next-generation sequencing (NGS) approaches, have revolutionized the genomics field [Petersen *et al.*, 2017; Lappalainen *et al.*, 2019]. The reduced cost per base sequenced and the possibility of analyzing the whole genome in a single run have started the "omics" era, allowing several pieces of research to be accomplished in much less time than expected compared to Sanger sequencing [Lappalainen *et al.*, 2019]. The number of whole-exome sequencing (WES) and, more recently, whole-genome sequencing (WGS) have increased exponentially, providing an impressive number of data available in public databases [Petersen *et al.*, 2017].

The use of WGS to address clinical pathogens has been improving in the last few years. Although the costs are a limiting factor, the main advantage is providing much information with a unique technique. This approach is based on cultured bacteria and is conventionally applied to identification of rare bacteria, outbreaks investigation by cgMLST/MLST (multi-locus sequence typing) and to the surveillance of high-risk clone and antimicrobial resistance genes (including mobile genetic elements such as plasmids and transposons) [Cameron *et al.*, 2020]. All of these are the scope of analyses carried out in our group. In addition, in the last years, we have focused on SARS-CoV-2 genomic surveillance due to the sanitary emergence that we are experiencing. This approach is based on total nucleic acid sequencing from primary clinical specimens, using amplicons targeted to the viral genome. The generated sequences are analyzed to identify mutations, SNPs and classify the variants. Analyses of pathogens genomics still pose several challenges. Most bacterial analyses are standardized, but the study of mobile genetic elements requires pipeline customization according to microorganisms. The bacteria genome is around 4-6 Mbp, demanding a high-performance computer to align and compare plasmids and genomes and bioinformatics skills. In addition, the pipelines typically access different databases that are constantly updated. Further, sometimes it is necessary to perform the phylogenetic analyses to establish the relationship among the strains or specimens, which require a lot of computing power and time to process data [Cameron *et al.*, 2020].

In NBioinfo, WGS and WES analyses have been most requested in human genetics research. Public databases containing consortiums of exomes and genomes have been largely explored to provide better comprehension regarding variants associated with genetic disorders, including ExAc [Karczewski *et al.*, 2017], gnomAD [Karczewski *et al.*, 2020], and ABraOM [Naslavsky *et al.*, 2017]; the latter provides variants encountered in Brazilian individuals. Despite the variants being easily assessed, their pathogenicity potential is not so easily comprehended [Findlay, 2021]. Criteria for classification of the variants' pathogenicity have been provided in a guideline from the American College of Medical Genetics [Richards *et al*., 2015] and are incorporated in different databases [Kopanos *et al*., 2019]. However, several variants remain classified as Variants of Unknown Significance (VUS) [Petersen *et al*., 2017]. In addition, it is challenging to evaluate these alterations' deleteriousness in complex genetic disorders [Findlay, 2021]. Two studies from our group explored population-based genomic databases using bioinformatics pipelines to investigate variants associated with mucopolysaccharidoses (MPS). MPS are a group of lysosomal storage diseases due to deficiencies of enzymes involved in the breakdown of glycosaminoglycans. In [Borges *et al*., 2020], the prevalence of different types of MPS was estimated based on data from the ExAC and gnomAD databases, demonstrating an analytical approach that can be reproduced for other rare diseases to aid health systems in better preparing to deal with them. Moreover, motivated by the different results that may be obtained for the same gene using distinct *in silico* predictors, Borges *et al.* [Borges *et al*., 2020] compared 33 *in silico* predictors and one conservation score using two datasets of variants with different degrees of confidence to evaluate the variants of uncertain significance present in the IDUA gene, which is associated with MPS I.

Moreover, our group has participated in studies focusing on understanding the genetic alterations that might increase the susceptibility to developing Thalidomide Embryopathy (TE). Thalidomide is the only factor necessary and sufficient to cause TE; however, there is a genetic susceptibility associated [Smithells and Newman, 1992; do Amaral Gomes *et al*., 2021]. To evaluate these variants of susceptibility, Kowalski *et al.* developed a gene panel comprising the CRL4-complex genes and the transcription factors IKZF1 and IKZF3 [Kowalski *et al*., 2020b]. Thalidomide binds to this complex through CRBN and degrades other proteins, such as IKZF1 (Ikaros) and IKZF3 (Aiolos). By sequencing the coding and untranslated regions of these genes in 35 individuals with TE, the authors encountered 145 variants. When performing a heatmap score, the authors identified variants that were associated to pre-axial limb anomalies, typical of TE [Kowalski *et al*., 2020b]. A more recent study of the group has provided a comparative genomic analysis in TE [Kowalski *et al*., 2021]. The sequence of 42 genes was compared across 14 species, two being resistant to TE development. The proteins RECQL4, SALL4, CDH5, KDR, and NOS2 had several variants in unaffected species compared to the sensitive ones.

4.2 Omics Data Analysis

Omics studies aim to comprehensively measure a particular type of molecule from a biological sample, providing a holistic view of the biological system concerning the targeted molecular feature [Yamada *et al*., 2021]. Omics data - including genomics, transcriptomics, epigneomics, and proteomics - are increasingly available in publicly accessible databases. They have been successfully explored with statistical and computational methods in various scenarios to extract molecular patterns associated with conditions of interest and provide transformative insights into biological processes, events, and diseases. While the experiments were initially concentrated on one type of omics, more recently, we have observed several integrative efforts analyzing two or more types of omics for the same set of samples to increase the power in understanding the complexity underlying human health and disease [Yamada *et al*., 2021; Misra *et al*., 2019].

NBioinfo routinely conducts research aiming to leverage single- or multi-omics data to identify molecular changes underlying biological or clinical conditions of interest. Most of our recent collaborations with other research groups from our institution were in this scope and included the analysis of transcriptome, miRNome, and methylome profiles in diseases such as COVID-19 [Kristem *et al*., 2021], cancer [Goemann *et al*., 2020; Gregório *et al*., 2020; Sartor *et al*., 2019], and type 1 diabetes [Assmann *et al*., 2018b]. We have also experience in exploring the potential of meta-analysis approaches to overcome the low statistical power due to small sample size in proportion to the thousands of measured probes in omics technologies and the low reproducibility of differential analyses based on a single omics dataset because of technical variabilities. We have applied this methodology to integrate transcriptome datasets using standard statistical methods [Haas Bueno and Recamonde-Mendoza, 2020] and machine learning-based approaches developed by our group [Trevizan and Recamonde-Mendoza, 2021], and ongoing works are further exploring meta-analysis methods to integrate distinct miRnome and methylome studies. Given the increasing interest in exploring omics data from public biological databases, we prepared a comprehensive overview of human biological databases in a recent paper [Villalba and Matte, 2021]. The review includes databases and tools to search biological sequences, genes and genomes, gene expression patterns, epigenetic variation, protein-protein interactions, variant frequency, regulatory elements, and comparative analysis between human and model organisms.

Bioinformatics on omics data is also a promising approach to advance rare diseases research, although the challenges are staggering. We developed MPSBase, a database for differentially expressed genes from different public mucopolysaccharidoses (MPS) data. The database gathers 13 studies previously deposited in the GEO database (https://www.ncbi.nlm.nih.gov/geo/), providing easy access to this information through a user-friendly[4]. We believe the development of analytical and automated strategies accessible to health professionals is essential for fostering research on MPS and other diseases. In the methodological side, we also aim to further explore multi-omics analyses, which despite introducing additional analytical challenges, have the potential to explain genotypic and phenotypic heterogeneity that may not be evident from single-omic analyses [Kerr *et al*., 2020].

[4] `https://www.ufrgs.br/mpsbase/`

4.3 Metagenomics

Nowadays, the Metagenomic is the main approach to study the structure, function and diversity of microbial communities in different microbiomes [Jünemann *et al.*, 2017]. There are different techniques, but two of them are mainly used: the characterization of 16S rDNA and metagenomic shotgun. The 16S rRNA-based NGS analysis has been used to taxonomic composition and phylogenetic relationship among prokaryotes. On the other hand, the metagenomic shotgun can address other microorganisms as virus and fungi besides the function of each member in the community [Jünemann *et al.*, 2017]. These results associated with other omics technology can offer results more robust and more reliable [de Vos *et al.*, 2022].

Several studies are ongoing, all of them 16S rDNA based: characterization of gut microbiota composition after intervention (*e.g.,* effect of consumption of grape juice on the development of school children and it's association with microbiome, effects of gut microbiome on sepsis susceptibility and progression), and others on upper and lower respiratory microbiome (*e.g.,* the relationship between nasopharyngeal microbiota and COVID-19 outcomes, cystic fibrosis, the prognostic for respiratory diseases, tracheostomized children).

Recent advances in technology alongside the drastic reduction of sequencing costs that increased our capacity to generate a huge amount of microbial DNA data. From this point of view, we have observed an increase in the number of microbiome studies but the knowledge about the key points and analyses is very scarce in our scientific community. Defining the experimental design (sample size, following), adequate methodology and statistical analysis is not an easy task since the most fundamental issues in microbiome studies come from poor statistical analyses and incorrect experimental design. In addition, the output of microbiome analyses delivers an elusively ordinary matrix. These data are highly dimensional, with thousands of different taxa, sparse, with lots of zeros, and demand a wary statistical approach to generate significant results. The possibilities and the number of tools available to perform that are enormous [Callahan *et al.*, 2016; Knight *et al.*, 2018]. Moreover, the time management to improve our know-how and get skills in bioinformatics is yet a limiting factor in our service capacity.

4.4 Systems Biology

The organism's complexity and the biological pathways underlying the wide range of biochemical processes cannot be transcribed as a single feature in bioinformatics analyses. Understanding interactions among molecular or chemical features (*e.g.,* genes, proteins, transcription factors (TFs), noncoding RNAs, drugs of interest) and their association with conditions of interest has long been a goal of systems biology studies, which are gaining momentum due to the greater availability of biological data and bioinformatics resources [Pavlopoulos *et al.*, 2011]. Using this approach, we investigated the neurological impairment regarding a set of lysosomal diseases [Silva and Matte, 2022]. We applied topology analysis to search hub genes and enrichment analysis to understand the biological process and mechanisms of neuropathology in the different types of MPS. In other studies, we used systems biology to identify new ovarian response markers to controlled ovarian stimulation [Eisele *et al.*, 2021] and to construct a comprehensive TF-miRNA co-regulatory network in pathological cardiac hypertrophy that allowed new insights about key regulators, molecular mechanisms, and the interplay between microRNAs and TFs in this condition [Recamonde-Mendoza *et al.*, 2019].

Systems biology analyses are particularly challenging when evaluating embryonic development. Transcription regulation is very intrincated in embryonic development and the genetic control of proliferation and cell differentiation is more similar with the tumor cell than the normal adult one [Ma *et al.*, 2010; Shahbazi and Zernicka-Goetz, 2018] However, the number of databases comprising embryodevelopment data is limited, for protein-protein interactions (PPI), gene-gene (co-expression) and even drug-protein networks. Drug-protein networks' databases are very usual for adult cells data (*i.e.,* STITCH, ChEMBL, IntAct) [Villalba and Matte, 2021], but the knowledge regarding drug effects in embryo or fetal development is not so represented. Several drugs are known to disrupt embryo development causing congenital anomalies, being named teratogens [do Amaral Gomes *et al.*, 2021]. To better understand the teratogenic effects of anticonvulsants, a weighted gene co-expression network analysis (WGCNA) was performed in one publication of NBioinfo [Kowalski *et al.*, 2020a]. In this study, Gene Expression Omnibus (GEO) datasets were obtained, comprising only anticonvulsants exposure in human or mouse embryonic stem cells. Through this analysis, chromatin remodeling and genes were associated with Fetal Valproate Syndrome, which is induced by valproic acid use during pregnancy, whislt neurodevelopment genes were associated with Fetal Hydantoin Syndrome, caused by phenytoin exposure during prenatal development.

Regarding the wide range of methods and algorithms in systems biology, we highlight some challenges of particular interest of our group. Measuring the reliability of PPI networks is an open problem, as experimental data is usually liable to contain many spurious interactions, making difficult to reduce the noise in generated networks. Strategies to detect disease-associated nodes from biological networks also need further improvement. Reconstructing signaling pathways from PPI networks and choosing the best ontology databases are other tasks that still demand scientific attention.

4.5 Machine Learning with Biological and Medical Data

The increasing size and complexity of biological data have turned the biomedical domain into one of the most exciting and fruitful application fields for machine learning [Greener *et al.*, 2022]. Machine learning (ML) is a subfield of Artificial Intelligence that aims to fit predictive models to data or identify underlying patterns (*e.g.,* groupings) within the data that may be relevant for problem-solving by using learning algorithms. These algorithms can automatically learn patterns from data without the need for an explicit set of instructions. Although the ML field is broad and accommodates several learning paradigms, NBioinfo works mainly with supervised and unsupervised learning algorithms: while in the first, we have labeled datasets to train models to classify data or predict outcomes accurately, in the second, we do not have a label or output value for each input data and the algorithm must find by itself the clusters or associations in the given dataset. We also explore both shallow (*i.e.,* traditional) and deep learning algorithms, which extend the predictive potential of traditional ML methods by composing several layers of non-linear modules than can extract multiple levels of representation and learn very complex functions [LeCun *et al.*, 2015]. Deep learning algorithms have proven to be effective methods for many different challenging tasks within bioinformatics, although progress is still needed to fully harness their potential [Greener *et al.*, 2022; Ching *et al.*, 2018].

Our group has broad experience in exploring ML algorithms to solve several different scientific problems in bioinformatics. Various studies have been conducted with NBioinfo's collaboration in the scope of using ML algorithms to develop predictive models for clinical diagnosis or prognosis from a variety of data types, including omics, clinical, and sociodemographic variables. In [Mello *et al.*, 2020], we proposed a Support Vector Machine (SVM) classifier to discriminate solid primary tumor (TP) and adjacent normal tissue (NT) for endometrial carcinoma using lncRNAs expression profiles, and in [Brondani *et al.*, 2020], we developed a random forest model to classify patients at different stages of diabetic kidney disease according to their peptidomics profile. We have also collaborated with the development of ML models to identify high-risk drinking patterns among medical students using web survey data [Marcon *et al.*, 2021] and to support posttraumatic stress disorder staging based on clinical and sociodemographic information [Ramos-Lima *et al.*, 2022].

Besides the research related to clinical machine learning, our group has tackled the task of protein essentiality prediction using the concept of graph neural networks (GNNs), which are a class of deep learning methods designed to operate on graph-structured data [Schapke *et al.*, 2021]. We have engaged in further exploring GNNs to other node-level and edge-level prediction tasks in bioinformatics, including the discovery of cancer driver genes [Andrades and Recamonde-Mendoza, 2022] or the reverse engineering of gene regulatory networks (*e.g.,* miRNA-target interactions). Our group is also interested in leveraging feature selection methods for biomarkers discovery from omics data. A promising direction that we have recently explored is the proposal of ensemble feature selection approaches to increase the stability and accuracy of candidate diagnostic biomarkers derived from high-throughput data, such as transcriptome [Trevizan and Recamonde-Mendoza, 2021; Colombelli *et al.*, 2021]. Finally, we highlight NBioinfo's participation in an international initiative to develop a community-driven platform for minimal information standards in report AI systems for biomedical domains [Matschinske *et al.*, 2021]: the AIMe registry[5].

We believe ML algorithms have an essential role in advancing our understanding of the function of organisms and the mechanisms of diseases, extending the potential of statistical methods traditionally employed in bioinformatics pipelines. Nonetheless, we also acknowledge the need to work towards improving these algorithms concerning many open challenges that are inherent to the biomedical domain, such as the high dimensionality of omics datasets, the need to efficiently integrate multi-view (*i.e.,* multi-omics) data to enhance models' predictive power, the usual class imbalance found in clinical and omics domains, and the high susceptibility of ML models to bias introduced by common historical disparities in healthcare and genomics data. Current projects under investigation by our group aim to extend the applications of ML in bioinformatics and to tackle these issues in order to make ML models a more reliable, secure, and accurate tool for pattern detection and predictions in the field of bioinformatics and health sciences.

5 Final remarks: successes, challenges and perspectives

Bioinformatics is an inherently interdisciplinary field and has many ramifications. These characteristics require researchers' cooperation from different knowledge areas touched by bioinformatics and with distinct expertise in data analyses. Although cross-disciplinary collaborations are not an easy road, NBioinfo succeeded in aggregating a complementary combination of skilled personnel around the common goal of developing a bioinformatics expertise hub within HCPA, an important Brazilian public hospital. An analysis of our trajectory shows that many successful interactions were established with other institutional groups and that our activities enabled more researchers to leverage the potential of omics resources and bioinformatics analyses in their projects - either through our capacity-building or consultation initiatives. In times of drastic cuts to the federal science budget, the access to these resources and methodologies enables several impactful research that would not be carried out without this support. Still, many challenges remain. A critical one is the need to continuously engage and train students and researchers to deal with the multiplicity of technologies, methodologies, and applications in bioinformatics and with the growing demands faced by NBioinfo, which from time to time outnumber our ability to contribute to new projects. Through our training, consultations, and research activities, we hope to continue expanding our academic network, increasing bioinformatics outreach at institutional, regional, and national levels, and contributing to the scientific development of the field.

Acknowledgements

We acknowledge the essential support from Prof. Patricia Ashton-Prolla in creating the NBioinfo. We also acknowledge the important role that Prof. Andreia Biolo played in opening the doors of her group to Mariana Recamonde-Mendoza, which enabled such bioinformatics networking to start at HCPA. We thank Silvia Olabarriaga, Delva Leão, and Marina Siebert for their support to the first bioinformatics course offered at the HCPA. Finally, we would like to

[5] https://aime-registry.org

thank our collaborators and all the students who participated in our group or are currently under our supervision.

Authors' Information

- **Mariana-Recamonde Mendoza**, Associate Professor at the Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS) since 2016; Researcher Professor at Hospital de Clínicas de Porto Alegre (HCPA) since 2017; Member of the Bioinformatics Core/HCPA.
- **Gerda Cristal Villalba Silva**, Postdoctoral researcher at the Bioinformatics Core, HCPA, from 2021–2022. Currently she is a Postdoctoral Associate at the Single Cell Genomics Core and Department of Human Genetics, Baylor College of Medicine.
- **Thayne Woycinck Kowalski**, Biomedical scientist in the Laboratory Genetics Unit (Medical Genetics Service) at HCPA since 2023; Collaborating Professor in the Graduate Program in Genetics and Molecular Biology, UFRGS; Member of the Bioinformatics Core/HCPA.
- **Otávio von Ameln Lovison**, PhD candidate in Pharmaceutical Sciences at UFRGS since 2019; Member of the Bioinformatics Core/HCPA.
- **Rafaela Ramalho Guerra**, PhD ctudent in Pharmaceutical Sciences at UFRGS since 2021; Member of the Bioinformatics Core/HCPA.
- **Andreza Francisco Martins**, Associate Professor at the Institute of Basic Health Sciences, UFRGS, since 2014; Researcher Professor at HCPA since 2020; Member of the Bioinformatics Core.
- **Ursula Matte**, Associate Professor at the Institute of Biosciences, UFRGS, since 2014; Researcher at HCPA since 2001; Member of the Bioinformatics Core/HCPA.

References

Andrades, R. and Recamonde-Mendoza, M. (2022). Machine learning methods for prediction of cancer driver genes: a survey paper. *Briefings in Bioinformatics*. bbac062. DOI: 10.1093/bib/bbac062.

Aron, S., Jongeneel, C. V., Chauke, P. A., Chaouch, M., Kumuthini, J., Zass, L., Radouani, F., Kassim, S. K., Fadlelmola, F. M., and Mulder, N. (2021). Ten simple rules for developing bioinformatics capacity at an academic institution. *PLOS Computational Biology*, 17(12):e1009592.

Assmann, T. S., Recamonde-Mendoza, M., de Souza, B. M., Bauer, A. C., and Crispim, D. (2018a). MicroRNAs and diabetic kidney disease: Systematic review and bioinformatic analysis. *Molecular and Cellular Endocrinology*, 477:90–102.

Assmann, T. S., Recamonde-Mendoza, M., Punales, M., Tschiedel, B., Canani, L. H., and Crispim, D. (2018b). MicroRNAs expression profile in plasma from type 1 diabetic patients: Case-control study and bioinformatic analysis. *Diabetes Research and Clinical Practice*, 141:35–46.

Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., and Schneider, M. V. (2019). A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, 20(2):398–404.

Barone, L., Williams, J., and Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLoS Computational Biology*, 13(10):e1005755.

Borges, P., Pasqualim, G., Giugliani, R., Vairo, F., and Matte, U. (2020). Estimated prevalence of mucopolysaccharidoses from population-based exomes and genomes. *Orphanet Journal of Rare Diseases*, 15(1):1–9.

Brondani, L. d. A., Soares, A. A., Recamonde-Mendoza, M., Dall'Agnol, A., Camargo, J. L., Monteiro, K. M., and Silveiro, S. P. (2020). Urinary peptidomics and bioinformatics for the detection of diabetic kidney disease. *Scientific Reports*, 10(1):1–11.

Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J., and Holmes, S. P. (2016). Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research*, 5.

Cameron, A., Bohrhunter, J. L., Taffner, S., Malek, A., and Pecora, N. D. (2020). Clinical pathogen genomics. *Clinics in Laboratory Medicine*, 40(4):447–458.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., *et al.* (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387.

Collins, F. S., Morgan, M., and Patrinos, A. (2003). The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290.

Colombelli, F., Kowalski, T. W., and Recamonde-Mendoza, M. (2021). A hybrid ensemble feature selection design for candidate biomarkers discovery from transcriptome profiles. *arXiv preprint arXiv:2108.00290*.

de Vos, W. M., Tilg, H., Van Hul, M., and Cani, P. D. (2022). Gut microbiome and health: Mechanistic insights. *Gut*, 71(5):1020–1032.

do Amaral Gomes, J., Olstad, E. W., Kowalski, T. W., Gervin, K., Vianna, F. S. L., Schüler-Faccini, L., and Nordeng, H. M. E. (2021). Genetic susceptibility to drug teratogenicity: A systematic literature review. *Frontiers in Genetics*, 12:645555. DOI: 10.3389/fgene.2021.645555.

Dragon, J. A., Gates, C., Sui, S. H., Hutchinson, J. N., Karuturi, R. K. M., Kucukural, A., Polson, S., Riva, A., Settles, M. L., Thimmapuram, J., *et al.* (2020). Bioinformatics core survey highlights the challenges facing data analysis facilities. *Journal of Biomolecular Techniques: JBT*, 31(2):66.

Eisele, B., Silva, G., Bessow, C., Donato, R., Genro, V., and Cunha-Filho, J. (2021). An in silico model using prognostic genetic factors for ovarian response in controlled ovarian stimulation: A systematic review. *Journal of Assisted Reproduction and Genetics*, 38(8):2007–2020.

Findlay, G. M. (2021). Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. *Human molecular genetics*, 30:R187–R197. DOI: 10.1093/hmg/ddab219.

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N.

(2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6):1981–1996.

Goemann, I. M., Marczyk, V. R., Recamonde-Mendoza, M., Wajner, S. M., Graudenz, M. S., and Maia, A. L. (2020). Decreased expression of the thyroid hormone-inactivating enzyme type 3 deiodinase is associated with lower survival rates in breast cancer. *Scientific Reports*, 10(1):1–12.

Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55.

Gregório, C., Soares-Lima, S. C., Alemar, B., Recamonde-Mendoza, M., Camuzi, D., de Souza-Santos, P. T., Rivero, R., Machado, S., Osvaldt, A., Ashton-Prolla, P., *et al.* (2020). Calcium signaling alterations caused by epigenetic mechanisms in pancreatic cancer: from early markers to prognostic impact. *Cancers*, 12(7):1735.

Haas Bueno, R. and Recamonde-Mendoza, M. (2020). Meta-analysis of transcriptomic data reveals pathophysiological modules involved with atrial fibrillation. *Molecular Diagnosis & Therapy*, 24(6):737–751.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1):1–15.

Hood, L. and Rowen, L. (2013). The human genome project: big science transforms biology and medicine. *Genome Medicine*, 5(9):1–8.

Jünemann, S., Kleinbölting, N., Jaenicke, S., Henke, C., Hassa, J., Nelkner, J., Stolze, Y., Albaum, S. P., Schlüter, A., Goesmann, A., *et al.* (2017). Bioinformatics for NGS-based metagenomics and the application to biogas research. *Journal of biotechnology*, 261:10–23.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., *et al.* (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581:434–443. DOI: 10.1038/s41586-020-2308-7.

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B., Birnbaum, D., Consortium, T. E. A., Daly, M. J., and MacArthur, D. G. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research*, 45:D840–D845. DOI: 10.1093/nar/gkw971.

Kerr, K., McAneney, H., Smyth, L. J., Bailie, C., McKee, S., and McKnight, A. J. (2020). A scoping review and proposed workflow for multi-omic rare disease research. *Orphanet Journal of Rare Diseases*, 15(1):1–18.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., *et al.* (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422.

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Aguilera, M. A., Meyer, R., and Massouras, A. (2019). Varsome: the human genomic variant search engine. *Bioinformatics (Oxford, England)*, 35:1978–1980. DOI: 10.1093/bioinformatics/bty897.

Kowalski, T. W., Caldas-Garcia, G. B., do Amaral Gomes, J., Fraga, L. R., Schuler-Faccini, L., Recamonde-Mendoza, M., Paixão-Côrtes, V. R., and Vianna, F. S. L. (2021). Comparative genomics identifies putative interspecies mechanisms underlying Crbn-Sall4-linked thalidomide embryopathy. *Frontiers in genetics*, 12:680217. DOI: 10.3389/fgene.2021.680217.

Kowalski, T. W., do Amaral Gomes, J., Feira, M. F., Ágata de Vargas Dupont, Recamonde-Mendoza, M., and Vianna, F. S. L. (2020a). Anticonvulsants and chromatin-genes expression: A systems biology investigation. *Frontiers in Neuroscience*, 14:591196. DOI: 10.3389/fnins.2020.591196.

Kowalski, T. W., Gomes, J. d. A., Garcia, G. B. C., Fraga, L. R., Paixao-Cortes, V. R., Recamonde-Mendoza, M., Sanseverino, M. T. V., Schuler-Faccini, L., and Vianna, F. S. L. (2020b). CRL4-cereblon complex in thalidomide embryopathy: a translational investigation. *Scientific Reports*, 10(1):1–13.

Kristem, L., Recamonde-Mendoza, M., Cigerza, G. C., Khoraki, J., Campos, G. M., and Mazzini, G. S. (2021). Roux-en-y gastric bypass downregulates angiotensin-converting enzyme 2 (ACE2) gene expression in subcutaneous white adipose tissue: a putative protective mechanism against severe covid-19. *Obesity Surgery*, 31(6):2831–2834.

Lappalainen, T., Scott, A. J., Brandt, M., and Hall, I. M. (2019). Genomic analysis in the age of human genome sequencing. *Cell*, 177:70–84. DOI: 10.1016/j.cell.2019.02.032.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Ma, Y., Zhang, P., Wang, F., Yang, J., Yang, Z., and Qin, H. (2010). The relationship between early embryo development and tumourigenesis. *Journal of cellular and molecular medicine*, 14:2697–701. DOI: 10.1111/j.1582-4934.2010.01191.x.

Marcon, G., de Ávila Pereira, F., Zimerman, A., da Silva, B. C., von Diemen, L., Passos, I. C., and Recamonde-Mendoza, M. (2021). Patterns of high-risk drinking among medical students: A web-based survey with machine learning. *Computers in Biology and Medicine*, 136:104747.

Mariano, D., Ferreira, M., Sousa, B. L., Santos, L. H., and de Melo-Minardi, R. C. (2020). A brief history of bioinformatics told by data visualization. In *Brazilian Symposium on Bioinformatics*, pages 235–246. Springer.

Matschinske, J., Alcaraz, N., Benis, A., Golebiewski, M., Grimm, D. G., Heumos, L., Kacprowski, T., Lazareva, O., List, M., Louadi, Z., *et al.* (2021). The AIMe registry for artificial intelligence in biomedical research. *Nature Methods*, 18(10):1128–1131.

Mello, A. C., Freitas, M., Coutinho, L., Falcon, T., and Matte, U. (2020). Machine learning supports long noncoding rnas as expression markers for endometrial carcinoma. *BioMed Research International*, 2020.

Misra, B. B., Langefeld, C., Olivier, M., and Cox, L. A. (2019). Integrated omics: tools, advances and future approaches. *Journal of Molecular Endocrinology*, 62(1):R21–R45.

Naslavsky, M. S., Yamamoto, G. L., de Almeida, T. F., Ezquina, S. A. M., Sunaga, D. Y., Pho, N., Bozoklian, D., Sandberg, T. O. M., Brito, L. A., Lazar, M., Bernardo,

D. V., Amaro, E., Duarte, Y. A. O., Lebrão, M. L., Passos-Bueno, M. R., and Zatz, M. (2017). Exomic variants of an elderly cohort of brazilians in the abraom database. *Human Mutation*, 38:751–763. DOI: 10.1002/humu.23220.

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., and Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Mining*, 4(1):1–27.

Petersen, B.-S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D., and Franke, A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. *BMC genetics*, 18:14. DOI: 10.1186/s12863-017-0479-5.

Ramos-Lima, L. F., Waikamp, V., Oliveira-Watanabe, T., Recamonde-Mendoza, M., Teche, S. P., Mello, M. F., Mello, A. F., and Freitas, L. H. M. (2022). Identifying posttraumatic stress disorder staging from clinical and sociodemographic features: a proof-of-concept study using a machine learning approach. *Psychiatry Research*, page 114489.

Recamonde-Mendoza, M., Werhli, A. V., and Biolo, A. (2019). Systems biology approach identifies key regulators and the interplay between mirnas and transcription factors for pathological cardiac hypertrophy. *Gene*, 698:157–169.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., and Committee, A. L. Q. A. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine : official journal of the American College of Medical Genetics*, 17:405–24. DOI: 10.1038/gim.2015.30.

Sartor, I. T. S., Recamonde-Mendoza, M., and Ashton-Prolla, P. (2019). TULP3: A potential biomarker in colorectal cancer? *PLOS ONE*, 14(1):e0210762.

Sayres, M. A. W., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., Triplett, E. W., Williams, J. J., Dinsdale, E., Morgan, W. R., *et al.* (2018). Bioinformatics core competencies for undergraduate life sciences education. *PloS One*, 13(6):e0196878.

Schapke, J., Tavares, A., and Recamonde-Mendoza, M. (2021). Epgat: Gene essentiality prediction with graph attention networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Shahbazi, M. N. and Zernicka-Goetz, M. (2018). Deconstructing and reconstructing the mouse and human early embryo. *Nature cell biology*, 20:878–887. DOI: 10.1038/s41556-018-0144-x.

Silva, G. C. V. and Matte, U. (2022). Neuronetworks: Analysis of brain pathology in mucopolysaccharidoses–a systems biology approach. *Neuroscience Informatics*, 2(1):100036.

Smithells, R. W. and Newman, C. G. (1992). Recognition of thalidomide defects. *Journal of medical genetics*, 29:716–23. DOI: 10.1136/jmg.29.10.716.

Trevizan, B. and Recamonde-Mendoza, M. (2021). Ensemble feature selection compares to meta-analysis for breast cancer biomarker identification from microarray data. In *International Conference on Computational Science and Its Applications*, pages 162–178. Springer.

Villalba, G. C. and Matte, U. (2021). Fantastic databases and where to find them: Web applications for researchers in a rush. *Genetics and Molecular Biology*, 44.

Yamada, R., Okada, D., Wang, J., Basak, T., and Koyama, S. (2021). Interpretation of omics data analyses. *Journal of Human Genetics*, 66(1):93–102.