Systematic overview of challenges and opportunities of computational data-driven research in biology

Authors and Affiliations (unsorted)

Tejasvene Ramesh

Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, 1985 Zonal Avenue, Room 713. Los Angeles, CA 90089-9121, USA

tramesh@usc.edu

ORCID: <https://orcid.org/0000-0002-5021-1338>

Karishma Chhugani

Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, 1985 Zonal Avenue, Room 713. Los Angeles, CA 90089-9121, USA

chhugani@usc.edu

ORCID: <https://orcid.org/0000-0003-1328-5028>

Vanessa D Jönsson

Departments of Computational and Quantitative Medicine and Hematology, Beckman Research Institute, City of Hope, Duarte, CA, USA

vjonsson@coh.org

ORCID: <https://orcid.org/0000-0002-6732-9607>

Twitter: vjonsson

Serghei Mangul

Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, 1540
Alcazar Street, Los Angeles, CA 90033, USA

serghei.mangul@gmail.com

ORCID: <https://orcid.org/0000-0003-4770-3443>

Twitter: smangul1

Abstract

Recent advances in omics technologies and their application in biology have impacted modern biomedical research by vastly extending the diversity, richness, and availability of data and methods across various domains. Computational data-driven research focuses on developing and applying efficient and robust computational models and methods across various types of omics datasets. Due to the fast and efficient nature of computational data driven research in analyzing large scale data, today computational data driven research becomes pervasive and has a multi-disciplinary role in biology and biomedicine. Yet, the narrow view of computational data-driven research is still widely spread, where computational scientists are often expected to have secondary roles in biomedical projects and not lead projects geared towards translational discovery. This phenomenon is partly due to the dependence on wet labs for omics data production and the role of computational researchers to focus on developing bioinformatics tools for the community rather than participating more in wet lab experiments or taking ownership in

the project. In this review, we discuss the opportunities and the pitfalls and challenges the computational data driven field faces. We discuss the shift in attitudes towards computational data driven research and the rise of computational research as an independent domain of biomedical research while addressing the significant collaborative opportunities that arise from integrating computational research with experimental and translational biology. Lastly, we share some insights of the future of computational data driven research in biology and its applications in different facets of the biomedical field.

History of computational biology

Computational biology holds a long history of more than half a century, ever since the times where personal computers were still a hypothesis and the discovery of the human genome sequence was still not known ^{1,2}. About half a century ago, computational softwares in biology was used to gather and analyze data from protein biochemistry, specifically to study protein structures and protein biology. Backdating to the early 1960s, COMPROTEIN, one of the earliest known bioinformatic software, was founded by Margaret Dayoff, using the first high-level programming language formula translation, FORTRAN, developed by International Business Machines (IBM). COMPROTEIN was primarily used to read and to determine the amino acid sequence of a primary protein structure sequenced with Edman sequencing. As a result of this discovery, the amino acid sequences of many proteins were determined. Eventually, these sequences were compiled into a library of protein sequences and published as *The Atlas of Protein Sequence and Structure*, the first biological database, in 1965 ². Around the time of the

emergence of DNA sequencing, in the late 1970s and early 1980s, Roger Staden introduced the first ever DNA sequence analyzing software, The Staden Package, dedicated to read and examine the sequences from Sanger sequencing . Since then, the availability of sequencing data started to rise, together with the accessibility of computers and development of bioinformatics softwares. In parallel in the 1980s, genome sequencing databases were developed, due to the need to archive large genome sequence data to make the information publicly available and accessible for profound data analysis and comparative genomic studies. These databases were further united into the International Nucleotide Sequence Database Collaboration, to systematize data formatting and data sharing ³. During the same era, computational biology, which was left unnoticed for years, found itself a position in modern science from the recognition of the first journal specific to bioinformatics, previously known as ‘Computer Applications in the Biosciences’ (CABIOS) and now as ‘Bioinformatics’ ^{1,4}

Modern day computational biology

In the late 20th century, the wide usage and expansion of computational power, has led to an exponential upsurge in data generation, or ‘Big data’ ¹. Until then, computational biology was highly dependent on wet labs to produce data for analysis, however with publicly available big data, computational biologists can utilize the data more independently for their research.

Such research is performed in a new type of laboratory, often referred to as a dry lab. The rapid advancement of omics technologies over the decade, has revolutionized modern biomedical

research augmenting the opportunity for more dry lab research, while vastly expanding the diversity, richness, and availability of data and methods across various domains ^{1,5}.

In contrast with the wet lab, researchers in the dry lab use the public omics data, instead of generating one. Dry lab researchers mine and reanalyze publicly available, increasingly rich large-scale open datasets, which are well positioned to make novel biological discoveries. This type of analysis is also known as the secondary analysis. Concurrently, the dry lab researchers also collaborate with wet labs to receive any evidence from wet lab experiments to perform analysis or validate their findings from their analysis in a wet lab setting.

Today, computational biology offers immense opportunities and its importance in biomedical research has grown to multiply exponentially over the decades. Despite all the opportunities computational data driven research has created, there is still a narrow view that although computational data-driven research is still extensive, computational scientists are often expected to have secondary roles in biomedical projects and are not expected to lead projects geared toward important biomedical discoveries. This phenomenon is partly due to the dependence on wet labs for omics data generation and having minimal participation in the experimental line ⁶, especially in posing the hypothesis and taking ownership in that project. In several cases, many computational biologists are recognized as a joint or second author during collaborative research and a few authors who developed the tools are left uncited.

To address some of these issues, this review discusses the opportunities, as well as the pitfalls and challenges of computational data driven research in biology. We have also reviewed to the extent at which a positive shift in attitudes towards computational data driven research from the biomedical community is observed, as the computational data driven research walks towards a path to independence from dependence on wet labs. Meanwhile, to gain more ownership of a project, some dry lab researchers take a mixed lab -combining experimental and computational lab - approach. Hence, we have shared challenges and opportunities of performing research in a purely computational lab and a mixed lab. As computational data driven research partly depends on the wet lab for data, we argue about the importance of establishing more meaningful collaborations between dry and wet lab research in novel findings and data validation. Along with that we analyze the future trends of this field and its application in biomedical research.

Opportunities of computational data driven research

Traditionally, computational researchers were focused solely on developing novel computational tools that are used to analyze the data generated in the wet lab settings. However, as the complexity of public data continues to evolve, it demands advanced computational tools to analyze the data with requirements aimed at answering specific scientific questions.

While tool development constitutes the main driver in computational data-driven research, it represents a fraction of growing research directions in this field. The applications of computational data-driven research open doors to clinicians and wet lab experimentalists to come up with new diagnostics, therapeutics, or insights to the disease. These applications include

modeling of basic features of a new assay, inferring correlation and association in large scale data, building of causal and mechanistic models, and analyzing functional big data.

As technology rapidly advances alongside reduction in the cost of computing ⁷, the methods to generate and analyze publicly available rich and diverse datasets have broadened, making opportunities of computational research increasingly prevalent ⁷. Data can be analyzed in two main ways. Firstly, the data is generated by the clinicians, and/or biomedical researchers in the wet lab scientists from the clinical assessments, clinical trials or wet lab experiments and will be analyzed by those who generated the data; this type of data analysis is primary analysis. On the other hand, data can be obtained and analyzed for further information from an existing dataset ; alternatively known as secondary analysis.

Computational research can contribute to advancing research by developing new methods and tools for primary and secondary analysis of biological data. It offers the unique opportunity to bridge analysis of datasets collected by wet labs and provide a platform to work in multidisciplinary teams- both wet and dry labs- through effective collaboration. A mutual collaboration between wet and dry labs can help to utilize an eclectic analysis and expertise, where the dry labs give inputs about the choice of tools and derive insights from the results produced by the wet lab scientists, simultaneously the wet lab scientists produce the preliminary data for computational analysis and ascertain the insights post the computational analysis . As a result, specific manuscripts have boosted the importance and impact of computational research through the development, implementation and discussion of new computational methods ⁸. Over

the past two decades, more and more studies are purely computational. For example, bioinformatics papers have accounted for more than 34% of highly cited scientific papers ⁹.

Data driven research enables bioinformatics researchers with a unique platform for creativity. Unlike wet lab research which is hypothesis driven, computational research creates hypotheses through the insights from the data interpretation. Hence, computational data driven research provides them the freedom to discover novel diseases, invent new methods and tools or improve the existing ones to enhance data analysis. In addition, computational research extensively explores methods development to data analysis at a low cost and a low dependency on external resources -for instance, animals and reagents, to name a few.

As a result, doing bioinformatics research does not need to spend a large portion of time generating data in comparison to in a wet lab; as the nature of the dry lab is more fast-paced. This expands the opportunities of computational data driven research to amalgamate more data for interpretation and collaborate with wet lab biologists to verify their findings.

Some computational scientists think that wet lab researchers may have computational knowledge limited to their domain of research. With the expertise of computational researchers, the experimental biologists can explore their data in a different perspective through the proposal of new hypotheses or improving on existing ones.

Furthermore, the scale of data analysis in wet and dry labs is different. More often than not, a number of dry lab scientists observe that commonly experimental biologists use a small sample size or biological replicates to obtain statistically significant results. However in computational research, scientists often have a chance to work and analyze data from larger sample sizes by combining data across multiple existing studies and open public datasets to obtain statistically significant results.

Pitfalls of computational data driven research

Although computational methods have provided promising frameworks for identifying new research directions, there remain challenges before considering the use of computational models to guide clinical and experimental decisions. There are many challenges which arise when managing, navigating, interpreting or deriving meaningful and accurate conclusions from big data. Some of these challenges include control over the data, accessibility of data and management of data

The main challenge that computational researchers are facing is the control over data. Data is, usually, generated by the clinicians, and/or biomedical researchers in the wet lab scientists from the clinical assessments, clinical trials or wet lab experiments. The generated data, both raw and metadata, is typically shared on the public databases such as Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO) in order for other computational biologists to be able to access the data for the secondary analysis. However, there are instances where some data are not shared or are restricted to the public ^{10–12}. Despite the vast amount of data being publicly deposited, one

of the biggest challenges that computational biologists commonly face is that they do not have much autonomy over the data which has been collected as they are not the ones who collected it. In many instances, computational biologists are not involved in planning of the project.

Data is a critical component of research; as the quality of data influences not only the analysis, interpretations, and conclusions drawn but also the choices of tools being utilized. Having access to how primary data generators collect the information, the data scientists can better understand how data is produced and potentially assist in better interpretation of data or choice of tool to be used.

Besides the control over data, there are some challenges in accessibility and management of the metadata that computational researchers face. Around five to ten years ago, installing software, examining, and manipulating a series of source code or lines of bash environment before getting to the actual command, were rather big challenges. With software libraries such as Pandas, we have overcome some of these challenges today. Today, the challenge is getting access to big data, managing the metadata, and keeping up with fast growing technology. As much as there are public databases open for data analysis, some data is still not accessible to the public¹³. As a result, computational biologists have limited information to explore and provide insights. Lack of data accessibility engenders the computational biologists to have smaller sample size in their analysis, eventually heavily impacting the quality of providing well positioned insights and discoveries.

Apart from accessibility of data, managing the metadata is also a crucial concern. Data integration and cleaning are vital for receiving normalized data for unbiased analysis. The challenge begins with cleaning and annotating the data. Often, the data is scattered, hence, researchers find discrepancies between datasets. For example, when researchers access the public repositories, it is expected that all epithelial cells and embryonic stem cells are going to be clustered together in their respective groups, but in reality, the same data are all scattered. In some cases, the data can be accessed but information in the data may be incomprehensive - due to improper annotation or incomplete details in the metadata and raw omics data^{11,12}. As the data is not generated by computational biologists, they may have difficulty integrating information from different datasets together and cleaning the data as they lack the specific details of how the experiment was conducted .

After modeling and publishing the data, the concern of privacy and confidentiality can prevail as the data has to be shared for another research group to reanalyze. Sharing of data is not only an issue of privacy and confidentiality, but the philosophical and sociological concerns¹³⁻¹⁵ of a few behind data sharing are also yet to be reviewed. Following the analysis of raw data, the storage of these big data sets is another concern. As these large dataset files are several hundred mega or gigabytes, it is important to find relevant facilities which can store this information in a cloud with large storage capacity. Equivalent to sequencing machines and molecular cloning equipment for wet labs, many dry lab biologists have to invest in computational facilities containing storage discs with sufficient storage and enormous computational capacity, meanwhile, some could rent these storage discs from cloud computing facilities.

Path towards Independence: a positive shift in perception towards computational data driven research

As many dry lab biologists are becoming more independent in developing complicated solutions to sophisticated computational problems, being at the end of the experimental pipeline is still a persistent issue. The narrow view of computational data-driven research is still widely spread, where computational scientists are often expected to have secondary roles in biomedical projects and are not expected to lead projects geared toward important biomedical discoveries. Consequently, a few dry lab biologists collaborating with experimentalists feel that their role comes in towards the end of the pipeline; to tie the loose ends of the research question ⁶. They have minimal participation in the experimental line, especially in posing the hypothesis and taking ownership in that project.

This phenomenon is in part due to the dependence on wet labs for omics data generation and has thus led computational researchers instead to focus on developing bioinformatics tools for the community. As a result of the phenomenon, many computational scientists are recognized as a joint or second author during collaborative research, with less appreciation ⁶ in spite of their notable contributions for developing methods and tools which have significantly assisted in data analysis; in some cases, the authors who developed the tools are left uncited ⁶.

However, as the accessibility of omics technologies grows, data generation and analysis becomes increasingly more complex; computational biology is likely to evolve into a principal, collaborative role ¹⁶. In the context of clinical trials, where sample size and volume are often limited, assessing tradeoffs with respect to optimizing sample usage for appropriate omics platforms, balancing cost, and assuring statistical significance is necessary to ensure success. Thus, the computational researcher's involvement early in genomics-based experimental design will be necessary to help answer complex scientific questions.

Academic infrastructure to support the independence and the principal role of computational biology in biomedical research is growing across the US ^{7,17-21}; new computational biology and biological data science departments are emerging in both university and medical institutions, signaling a transition from the reliance on "bioinformatic cores", often regarded as a service to translational biologists or clinicians, to the recognition of computational biology as an integral part of biomedical research. As biological papers become more reliant on data-driven analysis, the computational biologist will not only substantially transform into a powerful driver of biomedical research, but will also become increasingly essential.

When empowered with varied and increasingly abundant public data, computational researchers are now also well positioned to formulate competitive hypotheses. New publishing avenues are beginning to support secondary analyses as these are becoming an integral part of computational biomedical research²². By performing secondary analysis of datasets across multiple cohorts, computational researchers are gaining access to increased sample size and heterogeneity and can

capture novel signals unavailable in individual studies due a small sample size²³. Additionally, meta-analysis can capture signals that are generalizable across diverse populations. Multiple discoveries in biomedical fields were solely based on secondary analysis of public omics data^{23–26}. For example, domains of genome-wide expression profiling showed that meta-analysis was able to pinpoint a generalizable transcriptomic signature, helping to find novel biomarkers²⁷.

The perception about computational biology and its scientists has varied over the years. There have been two contrasting experiences of computational biologists in their initial days. While there are a few who felt less respected, the majority of the computational biologists were regarded for their work as not only was it unique, but also rare^{6,16}.

On the flipside, the majority of the computational biologists were highly regarded for their uniqueness in the field of expertise, yet considered as a major contributor and important complementation to the work of wet lab experimentalists. Over the years, the collaborations between the wet and dry lab scientists has strengthened the rapport and trust between scientists¹⁶. Some dry lab scientists believe that this could be due to the fact that they could provide solutions that wet lab scientists once thought was never possible; meanwhile, the dry lab scientists really observe the very deep biological knowledge which helps them in learning biology. As there are more scientists with computational biology knowledge, some experts of this field think the field is less unique, although it is still growing. They find this shift in terms of

team science nationally and internationally as a positive move on the future of computational biology.

Moreover, the perception about both the field of computational biology and the role of computational biologists has evolved amongst the biomedical community. The growing importance of computational work in biomedical research is more obvious in the sub disciplines such as biostatistics and bioinformatic methods or tools development. There are an increasing number of computational biology researchers who are first or co-first authors in biomedical studies. Over the years, many computational biologists have observed more receptiveness of the field by the biomedical community. This is evident through increasing computational projects thriving in the leading consortium integration studies and big international science projects.

Nonetheless, a handful of computational biologists still feel that the appreciation of the field is still not significant amongst the biomedical community. A few representatives from the wet lab still believe that computational biology largely relies on the data collection from the wet lab experimentalists. Furthermore, there is a thought that the data is analyzed in silos without reaching out to the primary data generator; therefore, the insights from the data may not be deep and often not taken seriously ²³. Many computational biologists believe that computational biology is given less importance compared to wet lab research based on less funding allocated to computational biology and very few computational biology researchers taking powerful roles in funding panels to make any influential decisions.

Challenges and opportunities in mixed and purely computational lab

Besides the wet or dry lab settings , many computational biology researchers would like to combine the benefits of both wet and dry lab into a single, mixed lab. As a result, these researchers suggest that they have better control over the quality of the data, making them better positioned to make novel biological discoveries. In fact, they have full control of the entire workflow from creating an experimental design, performing the experiments, collecting data to analyzing the experimental data and validating the results obtained by bioinformatics methods. Based on the analysis, scientists with wet lab expertise in a mixed lab setting can perform further experiments to validate and prove the insights provided from the computational researchers' analysis ^{28,29}.

On the flipside, there are researchers who have succeeded in building a purely computational lab as well. They largely concur with the fact that having a hybrid lab can be highly challenging. Coming from a dry lab background, it is challenging for many data scientists to learn the language of wet lab. It requires years of training to be an expert in wet lab although, acquiring knowledge in wet lab techniques can help benefit the directionality of their analysis. With effective collaborations, both the dry lab researchers and wet lab researchers can depend on each other's expertise- for instance, the dry lab are able to depend on wet labs to write protocols, perform bench work experiments and seek the expertise of the experimentalist, while feeling comfortable in sticking to their computational expertise.

Moreover with the accumulation of rich and diverse omics datasets which are publicly available, purely computational labs are able to survive independently by performing secondary analysis to extract further biological insights and analyze for new information from the publicly available large datasets ^{30,31}. Over the years, there have been many exemplary labs, performing computational biology research, spinning off start-ups and thriving in science and innovation. These labs have broken stereotypes of needing wet lab components in dry labs in order to perform state of the art research in biology. With advancement of next generation sequencing technologies, many steps which were traditionally performed in the wet lab, can now be performed in the dry lab. Biological samples can now be routinely sequenced by many companies and universities core facilities offering service for library preparation and sequencing, although collaborations are still vital to recruit patients and collect samples.

Despite widespread acceptance, there are still some members of the biomedical community that are concerned with the ethical aspect of analyzing the data generated by other researchers. To the few members of the biomedical community, the utilization of publicly available data to further analyze is often seen as looting productivity ³² of those who are data generators and they pose a threat to the disputation of existing conclusions on the data ³². It is believed that the insights from data are original only when the data is collected based on a new hypothesis rather than through the reanalysis of the public data ^{23,32}. This idea arises from the assumption that the data is owned by the primary data generators ¹⁴.

Similar to a newly developed method, tool or software that is made publicly available, publicly available data can be accessed by anyone and everyone. Data is after all the research product - sharing of data publicly and utilizing it for further processing is a key facet of scientific process, similar to how scientists cite information from peer-reviewed publications in their articles or use tools and software to analyze their datasets ^{23,33–35}. Reusability of data for secondary analysis benefits the biomedical community not only through new discoveries ³⁶, but even to uphold the robustness and validity of the preexisting findings. For example, Vioxx, an analgesic drug against rheumatoid arthritis, was considered safe according to patients in initial clinical findings. However, through reanalysis of the drug usage profile of the patients shortly after the release of drug into the market, Vioxx has been uncovered to pose potential harm, such as risk of cardiovascular events, rectifying toxicology profile of drug ^{14,34}. In fact, awards such as “Research Parasite Awards” ^{33,34} and “Research Symbiont Awards” ³³ recognize computational scientists who mainly reuse, reanalyze and share the data to provide new perspectives from the reanalysis of information, without giving co-authorship to primary data producers or sharers ³³. These awards mainly identify the importance of data reusability in promoting novel discoveries and advancements in biomedical research ^{33,34}.

Although a few are antagonistic towards secondary analysis, many members of the biomedical community share optimism of reanalyzing existing datasets. A plethora of public data for various diseases are made available in the databases daily and are accumulated over the last decades ^{31,37}. It is expected that the data availability will double every 12 to 18 months ³⁷. Moreover, the public omics data is made available after they have been used for primary analysis for biological

insights. Besides the biological insights from the primary analysis, this public data may contain some useful information about the disease, the information is left unexplored in many cases even by those who generated the data. Very often the primary data generators leave the information unexplored due to two reasons. Firstly, the data producers may not find the data useful in answering their existing hypotheses, hence, they may not further investigate anything new from the data or derive conclusions^{23,33}. Secondly, the sample size could be small, which might impede them from deriving compelling conclusions from their data- occasionally, this requires recruitment of more samples into the experiment, which can often be exorbitant²³. As a resolution, secondary analysis provides the ability to repurpose^{38,39} the data by extracting novel biological insights from the data which was not focused in primary analysis.

Establishing meaningful collaborations between computational and experimental biologists

As technology and methods advance, allowing larger scale experimental data, scientists are progressively motivated to incorporate an interdisciplinary approach to discovery. A typical collaborative scenario between wet and dry labs is described as the wet lab generates data and the dry lab conducts secondary analysis or collaborates with the wet lab to gain access to the data. In order to make novel discoveries, dry labs often rely on domain specific knowledge that will serve to guide computational results, interpretation, and validation. In order to establish meaningful collaborations between computational and experimental labs, data scientists plan their approach with their collaborators, choose a suitable collaborator, focus on reproducibility and contribution through publications.

Wet labs must generate hypotheses as the first step of their pipeline so that they can accordingly plan and order the associated materials. In contrast, dry labs are driven by the amount and quality of data availability, and as a result, create their hypothesis based on which data is available. It is for this reason that data driven research is more flexible as it is not restricted to one specific domain.

Despite multiple differences between the two types of labs, fundamentally there are a lot of similarities. Among both wet lab and dry lab approaches, experiments are iterative in nature.

While wet lab researchers may need to alter chemical measurements and biological specimens in the protocol, dry lab researchers may need to alter their script or data analysis methods multiple times. In some hybrid laboratories, the activities are intricately connected and one activity may depend on the other in order to generate and test hypotheses. Hence, when wet and dry lab researchers collaborate, it is important that they plan the approach and workflow of the experiments.

Next, it is crucial to choose suitable collaborators to establish meaningful collaborations between wet and dry labs. Computational biologists largely collaborate with both wet-lab scientists, clinicians, and computational scientists depending on the scope of the project. Computational biologists, largely collaborate with computational scientists for methods or software development or utilize the pure computationalists expertise in latest computational tools, non-parametric testing, causal influence methods, or biostatistics. To validate the results derived from computational tools, computational biologists collaborate with wet lab people. They seek

the domain experts' advice about a specific pathway, cell type behavior, or technique from the wet biologists. It is also common for computational biologists to be involved in projects associated with clinical trials to collaborate with clinicians or patients to discuss how to take computational methods from a research concept and translate them into a wet lab or into a clinical setting. Hence, it is key to choose a suitable collaborator depending on the scope of the project.

Reproducibility is another aspect which has similar requirements for both of the different labs. Dry lab researchers can reproduce data, but at the expense of needing more computational power, storage, and space to run the analysis and store the results afterwards. For example, in clinical genomics, standards of reproducibility need to be uniform in order for computational research to be able to progress into the clinics ⁴⁰. In contrast to dry lab research, reproducibility can be limited in wet lab based research as researchers must work with a small sample size and gather all essential lab equipment, animal models, chemicals in order to repeatedly carry out the experiments, resulting in increased cost of experimentation. Computational research needs to leverage the limitations of technology and deliver accurate and reproducible results. Reproducibility in computational research can be established by implementing standards of open software, open metadata, and utilizing data and tools located on archivally stable repositories ^{41,42}. Increasing reproducibility can aggrandize the reach and credibility of the findings, publication and data generators ¹⁴. Hence, reproducibility is key for accuracy and credibility of the results, which is significant for maintenance effective collaboration between wet and the dry lab researchers.

Lastly, the contributions through publications is an important factor to establish a meaningful rapport between the computational and experimental biologists. Regardless of whether it is wet or dry lab research, computational biologists largely believe it is difficult to publish a paper in general. A decade ago when bioinformatics was still relatively new, pure computational work was shared within conferences or smaller papers. Back then, publishing computational research was harder than publishing wet lab research. In some instances, the validation of data in the wet lab was vital for publishing a work.

In recent years, the field of computational biology and bioinformatics has had a growing trend of publishing pure computational work in very high impact factor journals. Computational biologists could sense an increasing support from journals, such as Genome Biology, Nature Methods, Biostatistics, to publish bioinformatic tools, algorithms, and methods. Some computational biologists suggest that the impactful work does not necessarily have to be presented in high caliber journals. Many of the significant and highly cited works were identified to be from smaller journals. Hence, the impact of work is independent of the caliber of the journal. Meanwhile, computational biologists find that hybrid integrations between scientists in wet and dry labs can give more significant findings for paper publishing. Some scientists publish two papers simultaneously, one that rigorously demonstrates the important computations used for validation and another that demonstrates the biological phenomenon. Unlike before, many computational biology scientists are able to publish works that are purely computational as well, with effective hybrid integrations.

Computational research additionally has the opportunity to expand into low resource countries as scientists in these countries can utilize online training platforms, bioinformatic tutorials, and available resources in order to expand their knowledge and skills in performing secondary analysis of publicly available data ⁷. Thus, the field of computational biology, through an established commitment to open and widely accessible platform development and data availability, not only will continue to promote cross disciplinary research but also more equitable access to the tools and data necessary for the success of computational researchers with diverse resources.

Future of computational data driven research

Transitioning into the future, genomics is likely to lead its way towards the future of computational biology. With a large number of datasets, the predictability of the model is better and the analysis will be more accurate. Computational biologists believe that genomics plays an important role in translational, clinical, and healthcare research in the future. Computational biology has a unique potential of being the bridge between the large amounts of data collected by wet labs and understanding this data. With the power of technology and robust computational methods developed by the computational biologists, the sub-discipline of genomics can pave the way for understanding diseases at a molecular level. As new types of data such as electronic ^{43,44} and personal health records ⁴⁵ emerge, the data extracted from these records will pave new areas of research or will be complementing omics data when studying the biology of disease. With effective collaborations with clinicians, and wet lab experimentalists, computational biology is

likely to flourish greatly and reach a remarkable level in the biomedical field in the future notably, in the fields of disease prevention, early diagnosis, precision medicine, drug development and treatment decisions.

Computationally-driven research has the potential to occupy a central role in assessing and guiding treatment during therapy . For example, although primary tumor biopsies can yield important genetic information that may guide initial treatment decisions, liquid biopsies aimed at measuring circulating tumor DNA (ctDNA) or circulating tumor cells (CTCs) can provide information on disease progression and metastatic potential — and highlight detailed molecular information of patient's cancer as it evolves over the course of therapy ^{46–48}. By incorporating frequently updated molecular information as part of clinical trial protocol, the combination of bioinformatics and AI based methods, computational researchers have the opportunity to develop general methods that strategize optimal treatment strategies over the disease course, towards ultimately increasing overall survival and improving patient outcome ^{49–52}. Together with increasing opportunities to access longitudinal, genomic patient data for precision care, computational data driven research in biology evolves from bioinformatic, tool-driven research activity, to an interdisciplinary research that incorporates many aspects of math and computation.

Thriving in computational data driven research in the future

Computational biology has become an increasingly integral part of modern biomedical research over the years. Effective technical, research, and communication skills are key in thriving in the

field of computational biology. As part of the technical skills, many computational biologists suggest that it is important to begin to learn how to write good code. Computational biologists should have sufficient knowledge to explain to those who are new to the field. With the working knowledge of writing code and scripts, one can understand how to manage the data and represent the data in a more interpretable manner. To elucidate and complement the data interpretation, a good understanding of statistics is equally important. Once the results are generated, the results have to be validated before any conclusions can be drawn. If there are any issues in the data or results, there can be deleterious consequences such as misinterpretation of data.

As the younger generation of computational biologist researchers begin to explore the field, the pioneers of the field believe it is essential that they ask important research questions, and seek mentorship from those already present in the field. To do so, computational biologists are required to first have a good biological background of the disease or field they are working on. As they ask questions, they need to think critically to see how sustainable their ideas are. Many experienced computational dry lab scientists quote from their experience that certain ideas may not be practical or feasible and therefore, may not always work. Instead of being rigid and giving up on the idea that does not work, the pioneers of this field suggest that the flourishing scientists should be more flexible in accepting the direction in which their data or idea is going. If required, these scientists should be open to changing direction of the research.

Thirdly, the majority of computational biologists advocate the importance of learning how to communicate effectively. There are two aspects to communication- one is to build rapport with

collaborators and the other is to communicate with the biomedical community. To succeed in the field of computational biology, scientists need to possess good interpersonal skills. They need to understand and respect the body of work of wet lab biologists. Computational biologists should be open to picking up new knowledge or technologies related to their domain. With respect and friendliness, dry lab scientists can make good collaborations with the experimentalists. Besides building strong collaborations, dry lab scientists should communicate with the biomedical community through their work. They should be able to effectively communicate the technical information, explain why their work is useful to the biomedical community, and its impact on the field in the future. Communication through publications and conferences are great opportunities for scientists to share their work and its importance to the scientific community.

Lastly, a handful of computational biologists encourage young scientists not to be afraid of being unique and offering unique perspectives. Embracing the uniqueness within the team and through collaboration can break some of the classic paradigms that people have and can promote creativity and diversity in the work the team does.

Discussion

Taken together, the review discusses unique opportunities and challenges of computational data-driven research in biology. Some dry lab researchers often identify their importance is not as valued in collaborative research. With the increasing importance of data driven research, the attitudes towards computational biology are slowly changing. Due to advancements such as open

omics data in conjunction with effective bioinformatics methods that can enable novel biological discoveries, a paradigm shift in attitude towards computational data driven research has been observed over the decades. Today, computational biology has emerged into an independent field , taking on the driver's seat from the back seat, where computational biologists can take charge of their own research. The perception about both the field of computational biology and the role of computational biologists has evolved amongst the biomedical community. With the help of wet lab collaborators who respect the art of computational biology, dry lab researchers can work with collaborators in assisting with performing experiments, writing protocols or grants to get funding.

To have more autonomy over the quality of the data for making novel biological discoveries, some of the researchers have taken a mixed lab approach of establishing both wet and dry labs. Yet, there are researchers who have also succeeded by building a purely computational lab, who focus solely on pure secondary analysis. Both types of labs have their respective opportunities and challenges which have been thoroughly explored in this review. We also address the significance of establishing meaningful collaborations between computational and experimental biologists for successful interdisciplinary biomedical research.

Veering into the future, computational data driven research has shown to be and will get increasingly popular in a few more years- especially in the fields of genomics, the regulation of artificial intelligence algorithms in the clinics and clinical assessment and guidance to treatment in the clinics. Besides the field of genomics, the regulation of artificial intelligence algorithms in

the clinics is another aspect of future computational data driven research. Despite the ubiquity and utility of artificial intelligence (AI) algorithms to automate tasks and assist in medical diagnosis and treatment design, semi- or fully autonomous, AI has not yet been broadly adopted in the clinic. Although several AI assisted medical devices have recently been approved in the areas of image-based diagnosis ^{53–55}, comprehensive FDA regulatory guidelines for the approval of AI based algorithms are currently still under development ^{56–58}. Precision or genomic based medicine is based on incorporating genomics to guide tailored treatment strategies, and cancer screening is currently at the forefront of efforts in oncology ^{59–62}. FDA approval for genomics-based testing in cancer has not only increased patient accessibility to higher standards of care, but introduced the potential to harbor significant advances in precision cancer treatment ^{63–65}.

Moving further into technological advancements, computational research becomes vital in all fields, and is not limited to biology, philosophy, sociology, to name some. Computational research is pervasive and plays a multi-disciplinary role in other fields. The applications of computational biology are believed to expand into every other field and open doors to clinicians and wet lab experimentalists to come up with new diagnostics, therapeutics, or insights to the disease. Holding a central role in treatment assessments and decisions together with increasing opportunities to access longitudinal, genomic patient data for precision care, computational data driven research in biology is evolving from bioinformatic, tool-driven research activity, to an interdisciplinary research. Such advancements in computational biology fortifies the hopes on its future - to flourish to a remarkable level, be a key game changer in every facet of biomedical research and eventually emerge into the future of biology.

References

1. Gauthier, J., Vincent, A. T., Charette, S. J. & Derome, N. A brief history of bioinformatics. *Brief. Bioinform.* **20**, 1981–1996 (2019).
2. Hagen, J. B. The origins of bioinformatics. *Nat. Rev. Genet.* **1**, 231–236 (2000).
3. Thampi, S. M. Introduction to Bioinformatics. Preprint at <https://doi.org/10.48550/arXiv.0911.4230> (2009).
4. Computer Applications In The Biosciences Volume 1, 1985.
5. Yu, U.-S., Lee, S.-H., Kim, Y.-J. & Kim, S.-S. Bioinformatics in the Post-genome Era. *BMB Rep.* **37**, 75–82 (2004).
6. Bartlett, A., Penders, B. & Lewis, J. Bioinformatics: indispensable, yet hidden in plain sight? *BMC Bioinformatics* **18**, 311 (2017).
7. Mangul, S. *et al.* How bioinformatics and open data can boost basic science in countries and universities with limited resources. *Nat. Biotechnol.* **37**, 324–326 (2019).
8. Van Noorden, R., Maher, B. & Nuzzo, R. The top 100 papers. *Nat. News* **514**, 550 (2014).
9. Wren, J. D. Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades. *Bioinformatics* **32**, 2686–2691 (2016).
10. Deshpande, D. *et al.* A comprehensive analysis of code and data availability in biomedical research. Preprint at <https://doi.org/10.31219/osf.io/uz7m5> (2021).
11. Rajesh, A. *et al.* Improving the completeness of public metadata accompanying omics studies. *Genome Biol.* **22**, 106 (2021).

12. Huang, Y.-N. *et al.* The systematic assessment of completeness of public metadata accompanying omics studies. 2021.11.22.469640 Preprint at <https://doi.org/10.1101/2021.11.22.469640> (2021).
13. Callebaut, W. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **43**, 69–80 (2012).
14. Figueiredo, A. S. Data Sharing: Convert Challenges into Opportunities. *Front. Public Health* **5**, 327 (2017).
15. Joly, Y., Dyke, S. O. M., Knoppers, B. M. & Pastinen, T. Are Data Sharing and Privacy Protection Mutually Exclusive? *Cell* **167**, 1150–1154 (2016).
16. Yanai, I. & Chmielnicki, E. Computational biologists: moving to the driver's seat. *Genome Biol.* **18**, 223 (2017).
17. UCLA Computational Medicine. <https://compmed.ucla.edu/>.
18. Computational and Quantitative Medicine.
<https://www.cityofhope.org/research/beckman-research-institute/research-departments-and-divisions/computational-and-quantitative-medicine>.
19. Data Science and Biotechnology Institute | Gladstone Institutes.
<https://gladstone.org/science/data-science-and-biotechnology-institute>.
20. Institute for Computational Medicine. *NYU Langone Health*
<https://med.nyu.edu/departments-institutes/computational-medicine/computational-medicine>.
21. Computational Medicine and Bioinformatics | Computational Medicine and Bioinformatics | Michigan Medicine. *Computational Medicine and Bioinformatics*

<https://medicine.umich.edu/dept/computational-medicine-bioinformatics> (2016).

22. Markowetz, F. All biology is computational biology. *PLOS Biol.* **15**, e2002050 (2017).
23. Data sharing and the future of science. *Nat. Commun.* **9**, 2817 (2018).
24. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).
25. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173**, 305-320.e10 (2018).
26. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).
27. Elinoff, J. M. *et al.* Meta-analysis of blood genome-wide expression profiling studies in pulmonary arterial hypertension. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **318**, L98–L111 (2020).
28. Esteki, M. Z. *et al.* HiVA: an integrative wet- and dry-lab platform for haplotype and copy number analysis of single-cell genomes. 564914 Preprint at <https://doi.org/10.1101/564914> (2019).
29. Quinn, T. A. & Kohl, P. Combining wet and dry research: experience with model development for cardiac mechano-electric structure-function studies. *Cardiovasc. Res.* **97**, 601–611 (2013).
30. Kodama, Y., Shumway, M., Leinonen, R., & on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).
31. Katz, K. *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic*

Acids Res. **50**, D387–D390 (2022).

32. Longo, D. L. & Drazen, J. M. Data Sharing. *N. Engl. J. Med.* **374**, 276–277 (2016).
33. Park, Y. & Greene, C. S. A parasite's perspective on data sharing. *GigaScience* **7**, giy129 (2018).
34. Greene, C. S., Garmire, L. X., Gilbert, J. A., Ritchie, M. D. & Hunter, L. E. Celebrating parasites. *Nat. Genet.* **49**, 483–484 (2017).
35. Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. *Nat. Rev. Genet.* **21**, 615–629 (2020).
36. Rosinger, A. Y. & Ice, G. Secondary data analysis to answer questions in human biology. *Am. J. Hum. Biol.* **31**, e23232 (2019).
37. NIH's Sequence Read Archive, the world's largest genome sequence repository: Openly accessible on AWS. *Amazon Web Services*
<https://aws.amazon.com/blogs/industries/nihs-sequence-read-archive-the-worlds-largest-genome-sequence-repository-openly-accessible-on-aws/> (2021).
38. Mangul, S. *et al.* ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. *Genome Biol.* **19**, 36 (2018).
39. Mandric, I. *et al.* Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. *Nat. Commun.* **11**, 3126 (2020).
40. Van Keuren-Jensen, K., Keats, J. J. & Craig, D. W. Bringing RNA-seq closer to the clinic. *Nat. Biotechnol.* **32**, 884–885 (2014).
41. Brito, J. J. *et al.* Recommendations to enhance rigor and reproducibility in biomedical research. *GigaScience* **9**, giaa056 (2020).

42. Mangul, S. *Interpreting and integrating big data in the life sciences*.
<https://peerj.com/preprints/27603> (2019) doi:10.7287/peerj.preprints.27603v2.
43. Ambinder, E. P. Electronic Health Records. *J. Oncol. Pract.* **1**, 57–63 (2005).
44. Hecht, J. The future of electronic health records. *Nature* **573**, S114–S116 (2019).
45. Tang, P. C., Ash, J. S., Bates, D. W., Overhage, J. M. & Sands, D. Z. Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption. *J. Am. Med. Inform. Assoc.* **13**, 121–126 (2006).
46. Chabon, J. J. *et al.* Integrating genomic features for non-invasive early lung cancer detection. *Nature* **580**, 245–251 (2020).
47. Moding, E. J. *et al.* Circulating tumor DNA dynamics predict benefit from consolidation immunotherapy in locally advanced non-small-cell lung cancer. *Nat. Cancer* **1**, 176–183 (2020).
48. Azad, T. D. *et al.* Circulating Tumor DNA Analysis for Detection of Minimal Residual Disease After Chemoradiotherapy for Localized Esophageal Cancer. *Gastroenterology* **158**, 494-505.e6 (2020).
49. West, J. *et al.* Towards Multidrug Adaptive Therapy. *Cancer Res.* **80**, 1578–1589 (2020).
50. Pritchard, J. R. *et al.* Defining principles of combination drug mechanisms of action. *Proc. Natl. Acad. Sci.* **110**, E170–E179 (2013).
51. Jonsson, V. D. *et al.* Novel computational method for predicting polytherapy switching strategies to overcome tumor heterogeneity and evolution. *Sci. Rep.* **7**, 44206 (2017).
52. Irurzun-Arana, I., McDonald, T. O., Trocóniz, I. F. & Michor, F. Pharmacokinetic Profiles Determine Optimal Combination Treatment Schedules in Computational Models of Drug

- Resistance. *Cancer Res.* **80**, 3372–3382 (2020).
53. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
54. Achenbach, S. Coronary CT angiography—future directions. *Cardiovasc. Diagn. Ther.* **7**, 43238–43438 (2017).
55. IDx-DR Overview: Close Care Gaps, Prevent Blindness. *Digital Diagnostics*
<https://www.digitaldiagnostics.com/products/eye-disease/idx-dr/>.
56. Benjamens, S., Dhunoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *Npj Digit. Med.* **3**, 1–8 (2020).
57. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback.*
<https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf> (2019).
58. Harvey, H. B. & Gowda, V. How the FDA Regulates AI. *Acad. Radiol.* **27**, 58–61 (2020).
59. Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* **183**, 818–834.e13 (2020).
60. Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1 (2008).
61. Hundal, J. *et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8**, 11 (2016).

62. O'Donnell, T. J. *et al.* MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst.* **7**, 129-132.e4 (2018).
63. FDA Approves Foundation Medicine's FoundationOne®Liquid CDx, a Comprehensive Pan-Tumor Liquid Biopsy Test with Multiple Companion Diagnostic Indications for Patients with Advanced Cancer | Foundation Medicine.
<https://www.foundationmedicine.com/press-releases/445c1f9e-6cbb-488b-84ad-5f133612b721>.
64. El Naqa, I., Haider, M. A., Giger, M. L. & Ten Haken, R. K. Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century. *Br. J. Radiol.* **93**, 20190855 (2020).
65. Dlamini, Z., Francies, F. Z., Hull, R. & Marima, R. Artificial intelligence (AI) and big data in cancer and precision oncology. *Comput. Struct. Biotechnol. J.* **18**, 2300–2311 (2020).